Rapport projet 19/12

Au cours de ce rapport, nous cherchons à relater et faire le bilan de ses trois premières semaines de projet de dernière année d'études, qui s'intitule "Cause toujours, tu m'intéresses ! Quel rôle causal jouent les liens inter-omiques dans les phénotypes complexes". Pour essayer de répondre à cette problématique, une idée qui va nous guider tout au long du projet est de transformer nos données génomiques en variables latentes et regarder si à partir de ses variables latentes, on observe un effet causal avec des variables latentes de métabolomes ou de microbiote.

Le plan de ce rapport suit à peu près l'ordre chronologique de nos travaux. Nous avons d'abord fait de la recherche bibliographique puis, nous avons cherché un côté plus pratique en trois points. Premièrement, nous avons cherché à simuler les données omiques, deuxièmement nous avons cherché à faire de la réduction de dimensions sur les données omiques, enfin nous avons exploré différents modèles pour notre analyse.

I. Bibliographie

Durant notre exploration bibliographique, nous avons étudié plusieurs grandes thématiques afin de répondre à notre problématique. Nous avons donc lu des articles scientifiques et des thèses autour de l'inférence causale, qui utilisent des DAG (Directed Acyclic Graph) et/ou des SEM (Structural Equation Model).

A. Généralités de l'inférence causale

Le premier article que nous avons lu est *Causal inference with latent outcomes (Stoetzer et al).* Il nous a appris que, lorsque l'on cherche à étudier des effets causaux sur des variables latentes, on doit faire attention sur deux aspects : l'identification de la causalité et des problèmes de mesures. Les erreurs de mesures peuvent être (mal) interprétées comme une structure causale sous-jacente, biaisant ainsi les conclusions après analyse.

Nous avons également lu *Causal Effects Based on Latent Variable Models (Axel Mayer)* qui nous apprend qu'un modèle statistique n'est pas nécessairement causal. Le modèle devient causal uniquement si des hypothèses fortes sont vérifiées comme ici l'absence de biais de la fonction des effets conditionnels.

L'article *Causal Inference for Latent Outcomes Learned with Factor Models (Landy et al.)* montre comment utiliser des modèles factoriels (du type analyse factorielle) pour définir des variables latentes et ensuite raisonner causalement sur ces variables plutôt que sur chaque variable observée séparément.
Pour notre projet, cet article offre un cadre théorique pour construire des facteurs latents multi-omiques (génomique, métabolomique, microbiote) et les interpréter comme des "outcomes latents" d'intérêt biologique. Et formuler des modèles

causaux (DAG/SEM) entre ces facteurs, ce qui peut nous aider pour relier proprement réduction de dimension et inférence causale sans tomber dans la “causal salad”.

De plus, nous avons lu une thèse *Découverte causale sur des jeux de données classiques et temporels. Application à des modèles biologiques. (Franck Simon),* qui nous a permis d’identifier “le défi de la découverte causale” qui “réside dans la conservation des liens directs qui reflètent une certaine compréhension de la nature, du processus de génération des données, tout en rejetant les fausses interactions qui sont des conséquences indirectes des relations réelles”. Il nous a aussi fourni une définition de la causalité claire : “Le concept de causalité, que l’on appelle communément relation de cause à effet peut être décrit comme l’influence par laquelle un événement, un processus, un état ou un objet (une cause) contribue à la production d’un autre événement, processus, état ou objet (un effet) où la cause est en partie responsable de l’effet, et l’effet dépend en partie de la cause. En général, un processus a de nombreuses causes, qui sont également considérées comme des facteurs causaux, et toutes se situent dans son passé. Un effet peut à son tour être la cause ou le facteur causal de nombreux autres effets, qui se situent tous dans son avenir.”

Enfin, nous avons regardé une vidéo “*Science Before Statistics : Causal Inference”* Richard McElreath, qui nous met en garde sur ce qui appelle la “causal salad” qu’il définit comme une pratique des statistiques où on utilise des outils statistiques standards pour tenter de répondre à des questions causales, sans avoir établi au préalable un modèle causal formel. On effectue des statistiques non-causales, mais on interprète les résultats comme s’ils expliquaient des causes et des effets. L’auteur nous donne même une recette de la salade causale :

1. Prendre une question vague
2. Trouver toutes les variables qui semblent liées au sujet ou que l'on a sous la main.
3. Mettre toutes ces variables ensemble dans une régression multiple (comme on mélangerait des ingrédients dans une salade) et laisser le logiciel "trier" le tout.
4. Prétendre qu'il n'y a pas de facteurs de confusion, d'erreurs de mesure ou de problèmes de données manquantes.
5. Espérer que les p-values ou l'AIC sélectionnent le bon modèle scientifique.

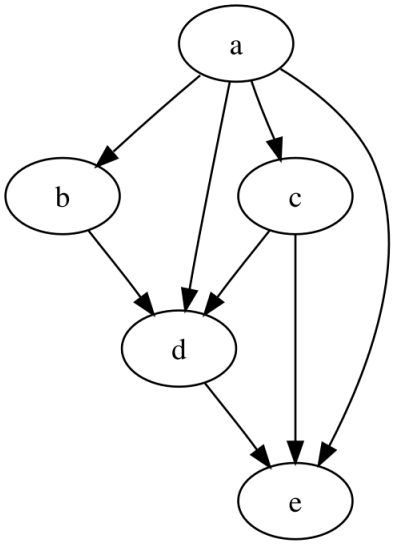
Faire une “causal salad” est à éviter, car les modèles statistiques ne voient que des associations, des corrélations et non des causes. Sans un modèle théorique préalable (comme un graphe causal ou DAG) pour justifier quelles variables inclure ou exclure, ajouter des variables de contrôle au hasard peut souvent créer des biais au lieu de les corriger.

B. DAG

Les Diagrammes acycliques causaux (DAG) permettent une représentation visuelle éclairée des relations causales entre différentes variables qu'elles soient inhérentes au jeu de données ou latentes. Chaque nœud ou bulle d'un DAG représente une variable et chaque arête avec ou sans direction représente un lien causal.

Les DAG sont conditionnées par la règle de d-séparation, on entend par là qu'un graphique DAG doit être acyclique donc les liens causaux ne peuvent pas former de cycle causé.

Si deux nœuds (variables) x et y sont d-séparés par une quantité S de nœuds alors les variables aléatoires correspondantes vx et vy sont conditionnellement indépendantes en considérant la quantité de variable aléatoire VS. Pour chaque distribution, existe un DAG avec lequel on peut représenter visuellement la liste des indépendances conditionnelles entre les distributions.



Dans notre projet, le DAG pourrait servir de colonne vertébrale à l'analyse causale en rendant explicites les hypothèses faites déjà implicitement sur les relations entre génome, métabolome et microbiote.

Concrètement, nous pourrions proposer un graphe où les SNPs influencent certains métabolites, qui à leur tour affectent des phénotypes d'intérêt, tandis que le microbiote joue un rôle de médiateur ou de modulateur externe.
Un tel DAG nous permettrait ensuite :

- d'identifier graphiquement quels ensembles de variables il faut ajuster (critère de back-door) pour estimer proprement l'effet causal des SNPs sur les métabolites ou des métabolites sur les phénotypes,

- de vérifier que votre réduction de dimension (ACP, regroupement de métabolites corrélées, facteurs latents) ne mélange pas des rôles causaux incompatibles dans une même variable latente
- et de guider le choix et l'interprétation des modèles ultérieurs (path analysis, SEM, algorithme de Peter–Clark) en évitant la "causal salad" décrite par McElreath, c'est-à-dire l'ajout aveugle de covariables sans justification causale claire.

Une proposition de DAG raisonnable pour nos données pourrait avoir la structure suivante (niveau "familles" de variables) :

- Génome (SNP) → Métabolome → Génotype
- Génome (SNP) → Génotype
- Microbiote ↔ Métabolome → Génotype

En blocs, cela donnerait un schéma avec un bloc G : variables génomiques (SNPs), un bloc M : variables metabolomiques (ou facteurs latents issus de la réduction de dimension), bloc B : variables microbiotiques (ASV ou facteurs latents):

On peut donc proposer les arêtes suivantes :

- G → M : les SNPs influencent les niveaux de métabolites (régulation génétique des voies metabolomiques);
- B → M : le microbiote peut modifier certaines voies metabolomiques (ex. métabolites produites /dégradées par les microbes).
- G → B (optionnel) : la plante peut "façonner" son microbiote via des exsudats racinaires génétiquement déterminés, ce qui crée un chemin G → B → M → Y.

C. Sem

Pour avoir une meilleure compréhension des SEM (Structural Equation Model) nous avons d'abord lu *Structural equations with latent variables Kenneth A.Bollen* qui nous apprend que les SEM peuvent être vus comme des équations de régressions avec des hypothèses qui autorisent les erreurs de mesures dans les variables explicatives et la variable à expliquer.

Nous avons aussi lu l'article *Structural Equation Modeling (Stein et al.)* qui présente le Structural Equation Model (SEM) comme cadre pour modéliser des relations complexes entre variables observées et variable latente. Le SEM utilise régression, analyse factorielle et path analysis. Il apporte surtout une sorte de feuille de route pratique : spécifier un modèle structurel (pour nous : génome → métabolome → phénotype, avec éventuellement le microbiote et des facteurs latents), puis vérifier l'ajustement global via la matrice de variance-covariance, et enfin tester la force et la significativité des chemins causaux.

D. Présentation du jeu de données

Notre jeu de données est issu de mesures observationnelles transmises par Anouk Zancarini. Il contient des mesures biologiques de la plante de colza et de ses interactions avec les sols : le microbiote du colza ; ces données devraient nous permettre d'éclairer les liens causaux entre le métabolisme, le génotype et le microbiote (interaction avec les sols). Ce jeu de données est divisé en 3 parties :

1. **Métabolomique** : matrice de 242 observations de 79 variables correspondant à différentes métabolites présentes dans le fonctionnement biologique du colza (ex : Fructose, Glucose, Citrate …). Les données metabolomiques sont contenues dans des variables quantitatives exprimées en concentration du métabolite sur la matière sèche (micromol/g).
2. **Microbiotique** : matrice de 242 observations de 14 815 variables correspondant à différents microorganismes présents dans l'environnement racinaire de la plante de colza : son microbiome. Ces mesures sont sous la forme d'Amplicon Sequence Variants (ASV) : il s'agit de séquences individuelles d'ADN récupérées à partir de l'analyse d'un gène marqueur, cette méthode permet la différenciation très fine entre deux individus biologique, à une paire de base près. Ces mesures sont contenues dans des variables quantitatives, sans unité, car elles traduisent un comptage, un pourcentage d'abondance relative du gène marqueur dans la séquence.
3. **Génomique** : matrice de 242 observations de 28 017 gènes de la plante elle-même. Les données génomiques sont contenues dans des variables qualitatives (0/1/2) : des SNPs (Single Nucleotide Polymorphism ou Polymorphisme Nucléotidique) qui traduisent la variation d'un nucléotide pour chaque individu mesuré. Ces variables ont 3 modalités : soit un polymorphisme sur un chromosome, sur l'autre chromosome, ou sur les deux (le colza étant diploïde).

Le caractère observationnel de nos données nous permet de considérer et de respecter plusieurs hypothèses clés dans la partie causale de notre étude. En effet, on distingue données expérimentales où on organise une expérience en mesurant un groupe traité versus un groupe contrôle où les individus sont choisis et données observationnelles, c'est-à-dire des mesures "non-interventionnelles" : sans implication du sujet, on mesure juste.

Les données observationnelles répliquent donc la "vérité" du monde réel tel qu'on l'a mesuré, ce qui induit une richesse structurelle où les outcomes varient librement sans influence du protocole.

Au sein de notre jeu de données, sont présents des cases vides (NAs), que nous avons décidé de supprimer. Néanmoins, nous avons uniquement supprimé les colonnes (variables : gènes, ou métabolites, ou ASV microbiotique) et non pas les individus (mesures). Cela nous permet de conserver une plus grande quantité de mesures, moins nombreuses que les variables.

II. Matériel et méthode

A. Réduction de dimension

Notre jeu de données est extrêmement large, avec plus de 28 000 variables pour le génome, cela pose plusieurs problèmes à considérer pour l'analyse causale. Une de ces considérations est le coût computationnel d'analyses sur un tel jeu de données, en effet construire un simple modèle de régression linéaire, pénalisé ou non, ou une ANOVA avec toutes les variables serait contre productif tant cela prendrait du temps. De plus, l'objectif principal de l'inférence causale n'est pas d'expliquer les simples corrélations entre les variables mais découvrir des vraies causalités entre les variables, et cela peut passer par des variables latentes/induites dans le jeu de données.

Nous avons donc un enjeu de réduire ou grouper les variables ensemble afin de faciliter l'analyse des données et permettre d'identifier des chemins de causalités entre les variables génomiques et les deux autres familles de variables (metabolomique et microbiotique).

Dans cette première partie de l'étude, nous nous concentrons sur la famille de variables metabolomiques.

La réduction de dimension peut être représentée en deux familles distinctes : l'extraction de caractéristiques et la sélection de caractéristiques. L'extraction de caractéristiques permet de choisir et de sortir des sous-ensembles de variables originales sans les modifier et en supprimant les variables non retenues. L'espace reste donc le même, mais avec moins d'axe, cela entraîne alors une perte d'information, mais améliore l'interprétabilité des caractéristiques.

La sélection de caractéristiques permet de transformer les variables et de les projeter sur un nouvel espace de plus faible dimension. Les variables originales sont donc contenues dans les variables représentatives de ce nouvel espace, des variables latentes. Ces variables latentes sont de nouvelles variables qui expliquent en partie chaque variables originales. Dans le cadre de l'ACP ces latentes sont les Composantes Principales et sont donc des combinaisons linéaires des variables originales. On conserve alors une quantité fixée de variables latentes, par exemple

20, qui explique une grande partie de toutes les variables originales (dans notre cas 28000) ce qui permet de mieux représenter les caractéristiques du jeu de données. Les variables latentes issues de cette sélection sont donc plus informatives que les variables originales individuelles, mais plus difficilement interprétables.

1. FA (Vadim)

Pour compléter l'ACP, nous avons exploré l'analyse factorielle comme autre approche de réduction de dimension sur les données génomiques. L'idée est de supposer l'existence de quelques facteurs latents continus qui génèrent la structure de corrélation observée entre les gènes, chaque gène étant écrit comme une combinaison linéaire de ces facteurs plus un terme d'erreur spécifique.

Dans la pratique, nous avons ajusté des modèles factoriels avec un nombre de facteurs choisi pour capturer la majorité de la variance commune, puis étudié les charges factorielles pour repérer des groupes de gènes partageant un même "profil" latent, ce qui est cohérent avec la présence de voies biologiques ou de modules metabolomiques.

Ce type de représentation nous semble particulièrement adapté pour la suite causale, car il fournit des variables latentes interprétables (facteurs) pouvant servir de nœuds dans un futur DAG ou modèle d'équations structurelles, tout en filtrant le bruit idiosyncratique des métabolites individuelles.

2. MOFA

Au cours de nos recherches, nous avons découvert une page github présentant un modèle d'analyse factoriel appelé MOFA (Multi-Omic Factor Analysis). Nous pensions que ce modèle serait adapté à notre projet ,car ses créateurs considèrent que MOFA est "une généralisation polyvalente et statistiquement rigoureuse de l'analyse en composantes principales aux données multi-omiques". MOFA fonctionne ainsi : "[1]À partir de plusieurs matrices de données contenant des mesures de divers types de données omiques sur des ensembles d'échantillons identiques ou se chevauchant, MOFA infère une représentation interprétable de faible dimension sous la forme de quelques facteurs latents."[2]. Nous pensions donc avoir trouvé la méthode parfaite qui permettait à la fois de réduire la dimension de nos données génomiques en des variables latentes et d'inférer un effet de cause à effet, car ces "facteurs appris représentaient la source motrice de la variation à travers les modalités de données" . Cependant pour construire ses variables latentes MOFA fait une combinaison linéaire de nos trois tableaux omiques. Or, cela ne

[1] https://biofam.github.io/MOFA2/

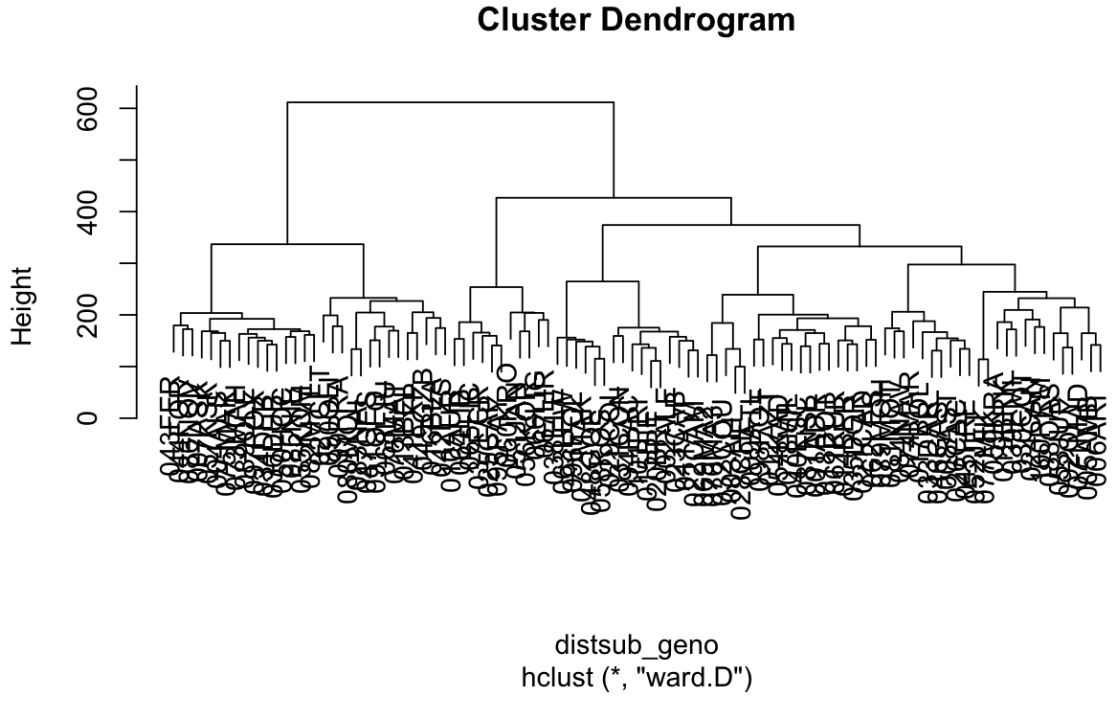
[2] https://biofam.github.io/MOFA2/

correspond pas du tout à la façon dont on veut résoudre le problème. Puisqu'on veut montrer un effet causal entre des données génomiques et metabolomiques (par exemple) on veut donc que notre variable latente ne soit composée uniquement de données génomiques. Même si un facteur à une source de variation motrice élevée en génomique avec MOFA, il reste une combinaison linéaire de génomes, métabolomes et microbiome. On n'utilisera donc pas MOFA pour la suite de notre projet.

3. HC (Vadim)

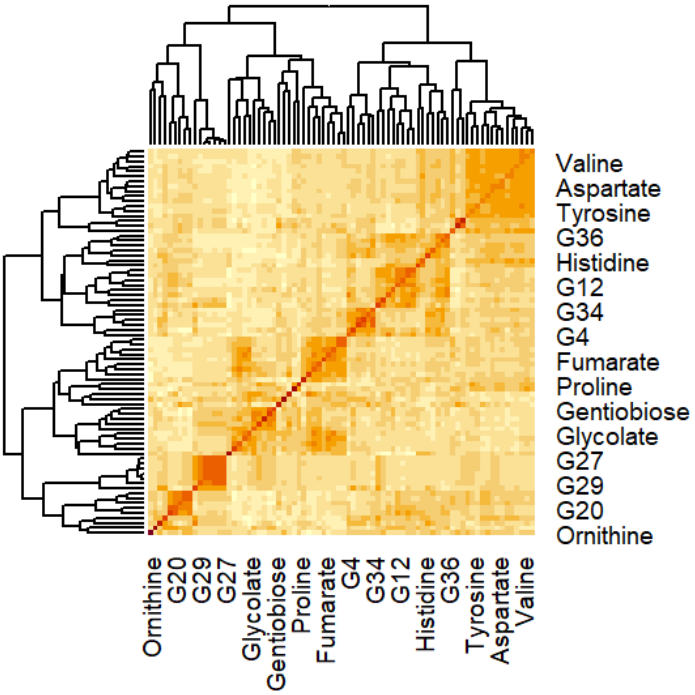
En parallèle, nous avons utilisé un clustering hiérarchique (HC) sur les métabolites afin d'identifier des groupes de variables au comportement similaire, dans une démarche similaire à celle de réduction de dimension. Nous avons construit une matrice de distance entre génomes (par exemple en transformant la corrélation en distance) puis appliqué un algorithme de regroupement hiérarchique agglomératif, ce qui produit un dendrogramme où des gènes fortement corrélés se retrouvent dans les mêmes sous-arbres.

Chaque "bloc" identifié dans la matrice trouve ainsi un écho dans la hiérarchie de clusters, ce qui nous permet de proposer des groupes de gènes cohérents du point de vue statistique et biologiquement plausibles (à vérifier avec Anouk Zancarini). Ces clusters pourront ensuite être condensés en variables latentes (par exemple via des moyennes ou premières composantes principales par groupe) qui serviront de candidats naturels pour les futurs chemins causaux entre génome, métabolome et microbiote. Cette méthode de HC puis ACP a été testée, malheureusement sans résultat cohérent ni interprétable.



4. Var-Cov

Finalement, nous avons opté d'analyser la structure de notre jeu de données en se basant sur des matrices de variances-covariance et de corrélation.



Voici la matrice de corrélation de notre jeu de données métabolomiques. On voit quelques carrés le long de la diagonale, une dizaine environ, qui représente des groupes de métabolite qu'on peut associer ensemble, car elles ont des comportements similaires (quand l'une à une concentration qui monte, les autres ont tendance à avoir une concentration qui monte aussi). Nous pensons que c'est une bonne piste pour former nos variables latentes. Pour l'instant, la meilleure piste que nous avons, consiste donc, en regroupant des ensembles de métabolites selon leur corrélation.

B. Modèles et analyses

1. Manova

Une première idée, que nous avions eue pour notre modèle causal était de relier les variables latentes génomiques et métabolomiques par une simple régression linéaire. Etant donné qu'on considérait plusieurs variables explicatives avec plusieurs variables réponses un modèle linéaire ne paraissait pas suffisante donc nous avons eu recours à une autre méthode qui nous paraissait assez intuitive qui un Modèle linéaire multiple ou MLM sur lequel on fait donc une MANOVA. Une

MANOVA est une extension de l'ANOVA où on a plusieurs variables réponses. Dans l'exemple ci-dessous, la MANOVA ajuste simultanément une équation de régression linéaire pour chacun des 4 métabolites.

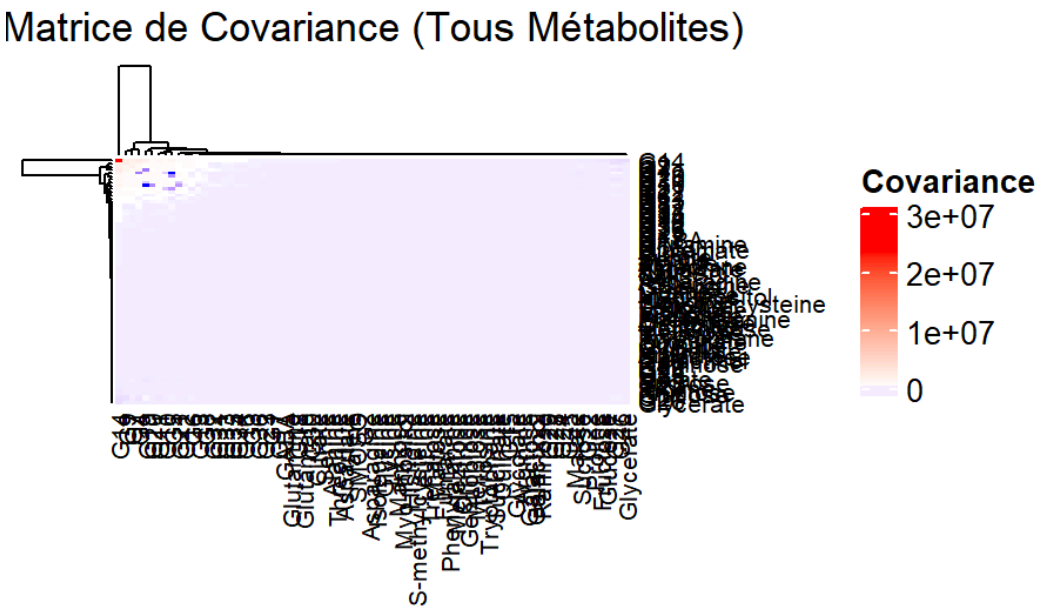
exemple pour le premier groupe de snp-metabo avec un lien causal

- G7 ~ SNP1 +SNP2+SNP3+SNP4
- G8 ~ SNP1 +SNP2+SNP3+SNP4
- G11 ~ SNP1 +SNP2+SNP3+SNP4
- G12 ~ SNP1 +SNP2+SNP3+SNP4

L'intérêt de la MANOVA ici contrairement à 4 ANOVAs réside dans le fait de pouvoir faire un seul test pour voir si les effets sont globalement significatifs à travers l'ensemble tout en prenant en compte les corrélations entre G7, G8, G11, G12. Le problème ici, c'est qu'une ANOVA ou une MANOVA ne teste "que" les effets des variables explicatives sur une variable réponse, il n'y a pas de notion de causalité dans le modèle. L'erreur qu'on a faite ici ressemble à ce que Richard McElreath essayait de nous mettre en garde : c'est une causal salad ! Nous n'utiliserons donc pas de MANOVA puisqu'il n'y a pas de notion de causalité.

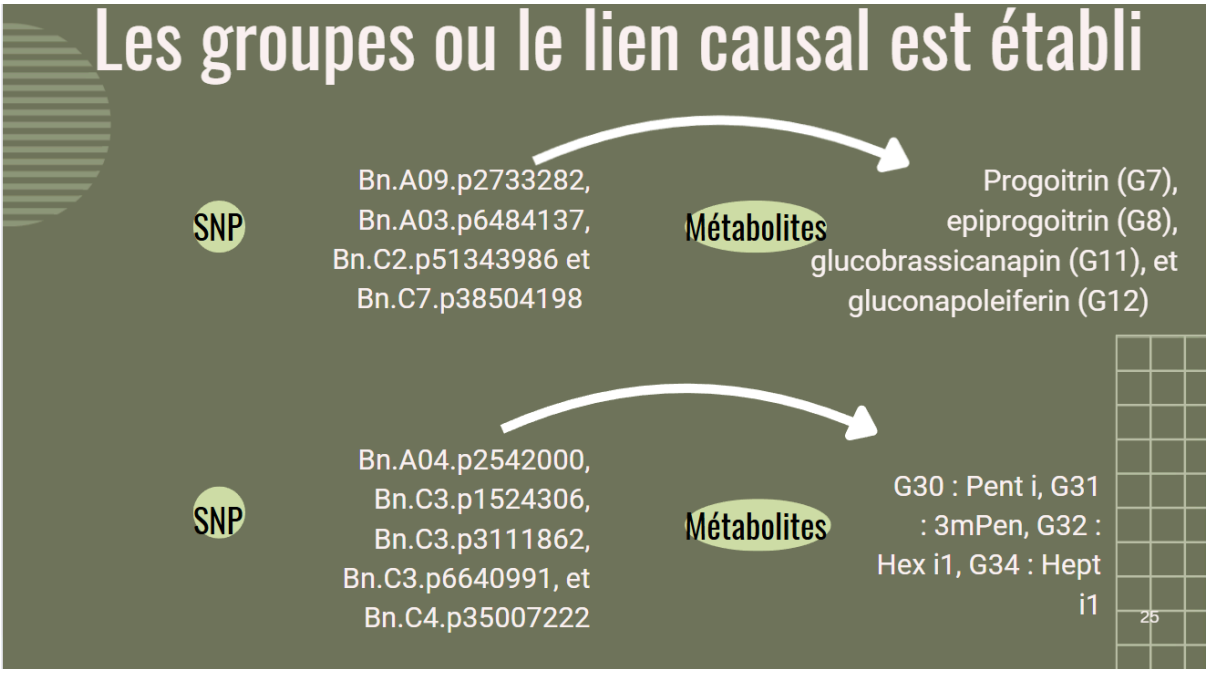
2. Var-Cov

Pour essayer de mieux comprendre la structure de nos données, nous avons étudié la matrice de variance covariance et de corrélation des génomes comme vous avez pu le voir plus tôt.



On voit que G14 est un outlier avec beaucoup de variance, mais on n'a pas d'informations sur la structure comme on pourrait avoir avec la matrice de corrélation.

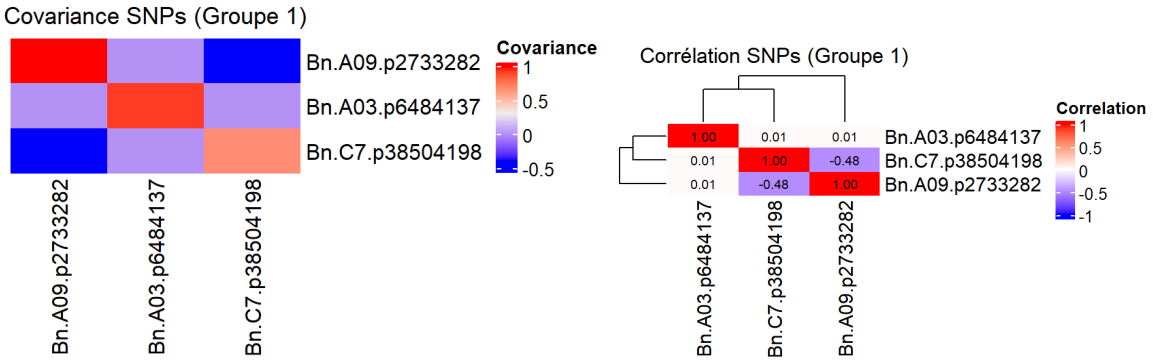
De plus, Anouk Zancarini nous a envoyé 2 groupes de SNPs et métabolomes dont on a trouvé un effet causal par GWAS (Genome-Wide Association Study).



Nous avons donc fait des matrices de variance covariance et de corrélation de ces groupes pour mieux comprendre leur fonctionnement. Il est important de noter que nous ne disposons pas de tousxses SNP dans notre jeu de données donc on représentera uniquement ceux du jeu de données.

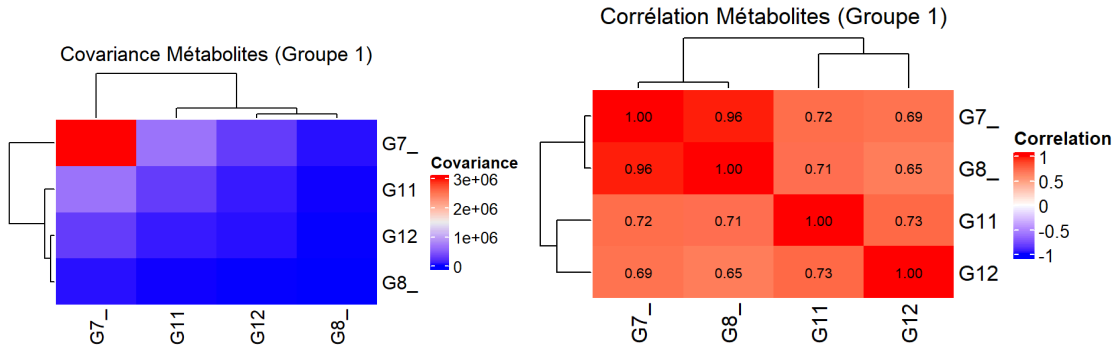
<u>Pour le groupe 1 (celui du haut):</u>

Les SNP



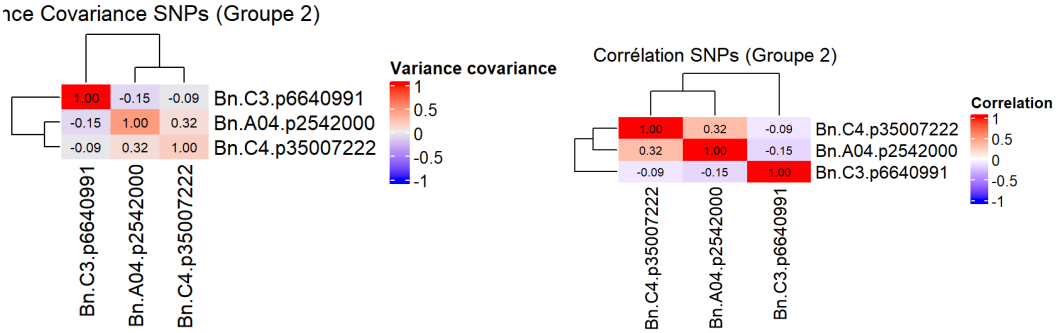
On voit une importante décorrélation 0.48 entre deux SNPs.

Les métabolites



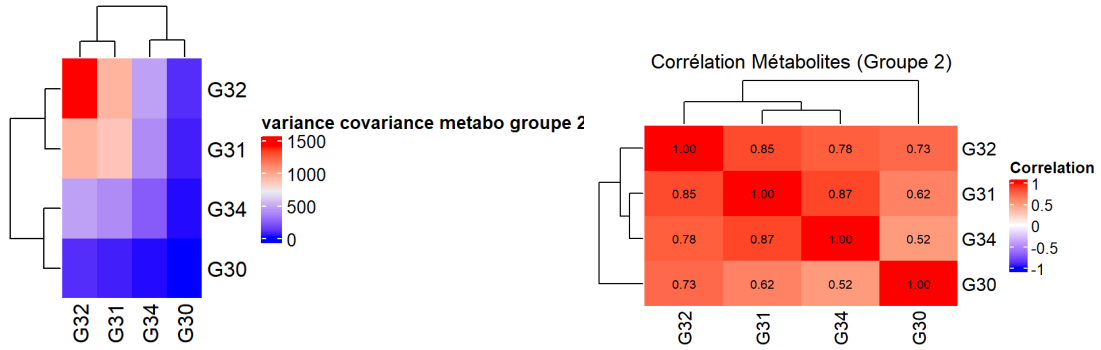
Ici, avec la matrice de corrélation, on peut voir que G7 et G8 forment une sorte de duo au comportement similaire. La matrice de variance covariance nous montre que G7 est une métabolite qui varie beaucoup

<u>Pour le groupe 2 (celui du bas):</u>
Pour les SNP



Ici encore on n'a que 3 SNP. Et on voit une légère corrélation 0.32 entre deux SNPs.

Pour les métabolites :



Ici, on voit que les quatre métabolites sont très liés et que G32 varie beaucoup.

Après avoir présenté ses résultats nous avons appris que GWAS est une méthode biaisée et donc nous ne devons pas trop nous attarder sur ses groupes où l'effet causal est "avéré", puisqu'il ne l'est pas tant que ça.

C. Simulation de données

Après l'exploration des stratégies de réduction de dimension et de modèles d'analyse de la causalité, on essaie de simuler les données réelles selon nos hypothèses, pour pouvoir tester des situations et DAG spécifiques.

Nous avons proposé dans un premier temps une simulation des données génomiques G et des métabolites M.

1. Génération de G

On commence par créer un vecteur de taille 28 000 de fréquences alléliques tirées d'une loi uniforme entre 0.05 et 0.5. On tire alors dans une loi binomiale pour chaque individu (250) et pour chaque SNP notre valeur 0, 1 ou 2 pour remplir la matrice de taille 250x28000 qui correspond donc à notre matrice G génomique simulée. On peut sélectionner par exemple 30 colonnes qui seront nos SNP causaux.

2. Génération de M

Nous allons simuler M selon un modèle linéaire simple M (250 x 80) = beta x G_causal. Beta (30 x 80) représente alors les liens "causaux" simulés entre la couche génomique et la couche des métabolites.

3. Génération de Beta

Beta représente la vérité dans notre modèle de relation causale. Il représente l'effet global de la causalité (direction du vecteur) et contient pour chaque SNP contient une information du métabolite impacté et la force de ce lien.

Pour générer beta, on sélectionne pour chaque SNP causal un nombre k au hasard de métabolites qui vont être impactés puis on tire au hasard k métabolites associés au SNP. Enfin on tire dans une loi normale (0.2, 0.1) l'effet que le SNP va avoir sur chacun des métabolites.

4. Réduction de dimension

On réduit M par ACP, ce qui donne notamment les coordonnées des individus dans l'espace des variables latentes, ou loadings L_40 (250 x 40) par exemple si on garde 40 dim. On veut regarder comment se comportent les liens de causalité entre M_reduc et G, ce qui pose la question de la représentation de M_reduc.

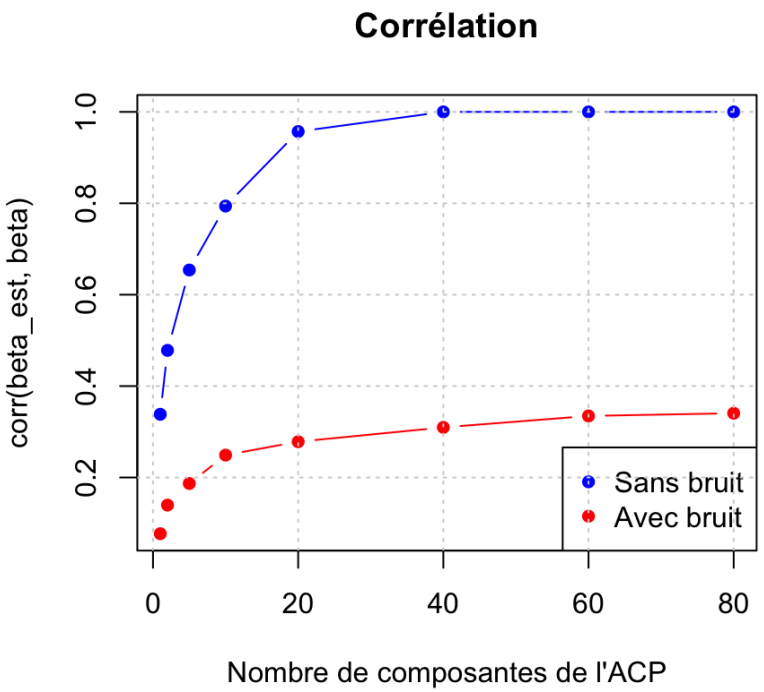
Une première idée consiste à re projeter M dans un espace 250 x 80. Pour cela on récupère L_40 (250 x 40) que l'on multiplie par les contributions de chaque variable dans cet espace réduit (matrices des vecteurs propres t(V) (40 x 80)).

On récupère donc une matrice projetée, qui comporte les informations uniquement des 40 premières composantes, étant données qu'elle est reconstruite à partir uniquement des coordonnées des individus sur les 40 composantes principales, et des contributions des variables sur ces mêmes composantes.

5. Etude de la causalité

Une première approche consiste alors à faire des régressions linéaire variable par variable ou métabolite par métabolite entre M_proj (250 x 80) et G_causal (250 x 30). On effectue 80 régressions qui donnent à chaque fois une vecteur alpha de taille (30 x 1). Ainsi on peut reconstruire une matrice beta_chapeau (30 x 80) par construction de chacun des alpha.

On peut alors étudier l'écart entre beta_chapeau et beta. Une première approche consiste à étudier la corrélation et le MSE, en fonction de la sélection du nombre de composantes principales :



Le bruit ajouté ici est simplement additionné à M à partir de cette matrice : noise <- matrix(rnorm(N * K, 0, 1), N, K).

On note une très grande sensibilité au bruit et récupération à partir de 20 dimensions de pratiquement toute l'information causale que l'on recherche.

6. Discussions

Plusieurs zones d'ombres subsistent dans cette première proposition de simulation. a) le nombre de métabolites associés à chaque SNP si il est trop grand est un problème car l'événement recherché est assez rare. On préférera k<2 plutôt dans un premier temps plutôt que 1<k<5 proposé initialement. Cela donne au final une matrice beta plus sparse dont on cherchera d'ailleurs à quantifier la "sparsité". b) l'utilisation de la matrice M_proj peut poser certains problèmes. Dans notre cas on un rapport entre n les individus et p les variables assez équitables, la perte et le gain d'information causale ne pourrait n'être pas forcément dû à la structure de la causalité mais juste à la structure instable de M_proj. c) La mesure de la corrélation répond-elle à notre problématique ? C'est sans doute le questionnement principal qui se dégage de la simulation. Dans notre cas, la corrélation mesure un effet global de causalité : "est-ce que les rapports d'intensité entre les différents niveaux de causalité sont conservés ?"

III. Pistes pour l'inférence
   A. Path Analysis
   B. Peter Clark Algorithm
   C.

Découverte (Exploratoire), il cherche "qui cause qui",
L'algorithme PC est une méthode de découverte. Vous ne lui donnez pas le modèle, c'est lui qui vous le donne.

Sources

https://biofam.github.io/MOFA2/

Multi-Omics Factor Analysis — a framework for unsupervised integration of multi-omics data sets (Arguelaguet et al.)
Multi-omics profiling of mouse gastrulation at single-cell resolution (Arguelaguet et al.)