Présentation de Iuliana Ionita-Laza du 03 February 2017:

# Contexte : fonction dans le génome :

Le contexte général est l’interprétation des études d’association pour des pathologies complexes – et, au hasard, disons les résultats d’association génome entier (GWAS) à partir de données de caractérisation fonctionnelle (par exemple ENCODE).

Les données de type ENCODE ont apporté beaucoup d’information quant au potentiel fonctionnel des régions de l’ADN. Des séries d’expériences « biochimiques » ont permis d’interroger l’état de chaque région chromosomique – chromatine ouverte, traces de méthylation etc …

Le projet ENCODE s'est donné pour objectif d'interpréter les séquences du génome humain, y compris les gènes, les transcrits et les régions régulatrices de la transcription. Il intègre pour cela de multiples technologies et approches pour les découvrir.

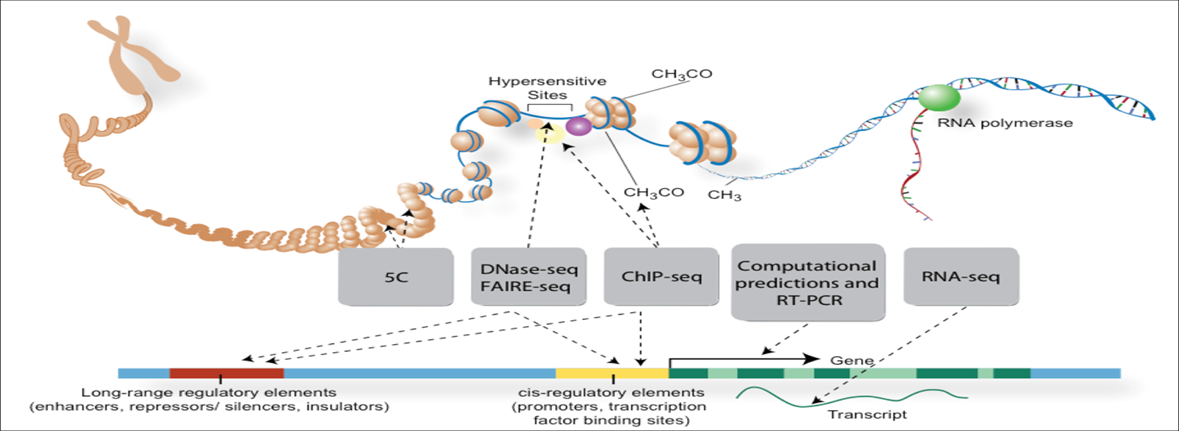


Figure 1: Expériences ENCODE

*Représentation schématique et principales expériences (dans les boîtes grises) visant à caractériser des éléments fonctionnels. Ces expériences sont menées dans > 100 tissus/lignées cellulaires.*

Bien entendu, la définition des régions fonctionnelles et de leur nature (*promoteur, enhancer* etc …) est rarement certaine. Nous pourrions plutôt considérer que nous avons un score par localisation– plus le score est haut, plus la région a des chances d’être fonctionnelle. C’est d’ailleurs dans ce contexte que Dan Graur avait critiqué de façon assez forte ENCODE puisque si on choisit un seuil suffisamment bas, on peut dire que 100% du génome est fonctionnel. Toutefois, en utilisant des scores, il est possible d’avoir une idée «probabiliste» du potentiel fonctionnel d’une région de l’ADN.

La présentation de Mme Ionita-Laza abordera la problématique de définition d’un score spécifique pour chaque variant dans chaque tissu. Il est possible d’avoir cette probabilité à partir des résultats des différentes expériences (Figure 1).

## Pourquoi :

Pour pouvoir combiner les résultats statistiques d’une étude d’association (les SNPs associés) et cette annotation. Par exemple, on peut imaginer que dans le cas du diabète de type 1, il y ait un excès de SNPs associés dans des régions « Immune enhancers » (Figure 2).

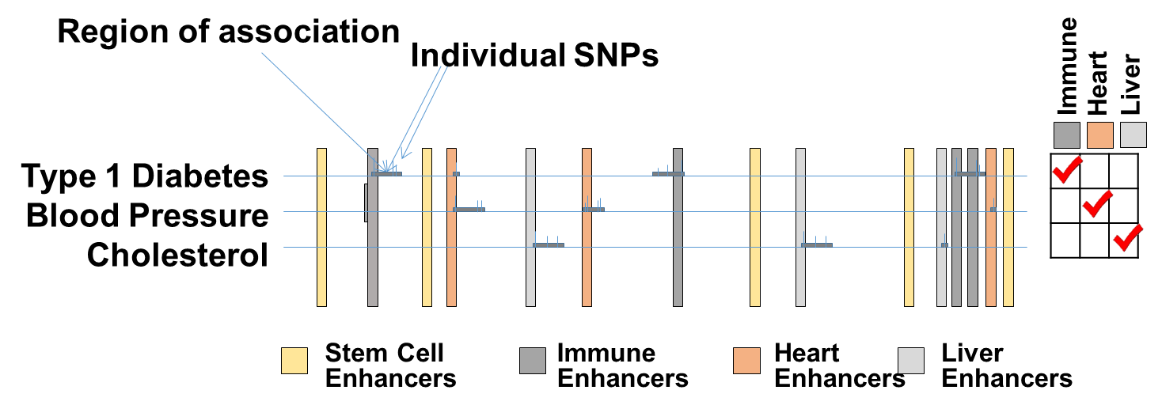


Figure 2: Combinaison GWAs / Annotations

*Ici, les SNPs associés à la pression sanguine sont souvent trouvés dans des régions de type enhancer dans le cœur, alors que les SNPs (et régions d’association) associés au taux de cholesterol se retrouvent dans des régions qui agissent comme enhancer dans le foie.*

# Méthode Statistique:

Ce niveau de complexité des données nécessite l’utilisation d’outils adaptés.

## La méthode:

La méthode Latent Dirichlet Assignation est un outil – au même titre que la régression linéaire, par exemple – qui va permettre d’analyser ces données qui ont une structure complexe. La partie la plus compliquée, mais aussi la plus intéressante à exploiter, est la corrélation entre tissus proches.

L’objectif final est d’avoir à chaque locus (paire de base) pour chaque tissu-catégorie cellulaire, une probabilité d’appartenir à un type fonctionnel. Dans le travail présenté vendredi, il y a 9 types fonctionnels (je ne comprends pas pourquoi il y a 3 silent : sans doute lié aux différents type d’hétérochromatine) :

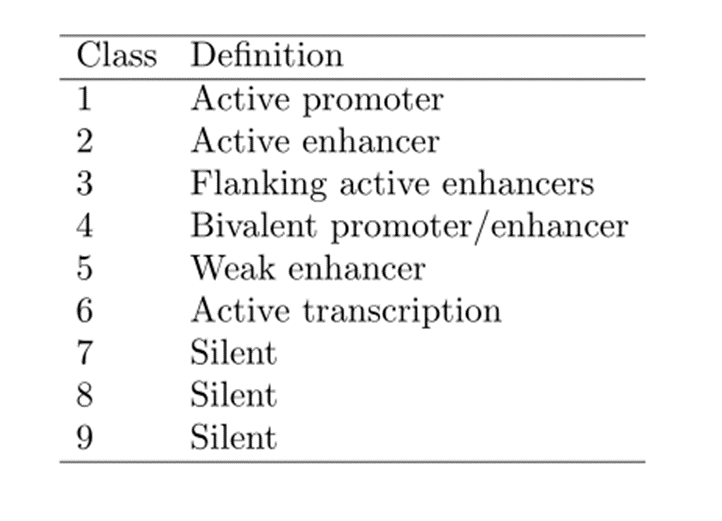


Figure 3: Classes fonctionnelles

## Illustration de la méthode LDA sur un exemple simple:

Soient 5 phrases:

I ate a banana and spinach smoothie for breakfast

I like to eat broccoli and bananas.

Chinchillas and kittens are cute.

My sister adopted a kitten yesterday.

Look at this cute hamster munching on a piece of broccoli.

Nous voudrions savoir combien de “sujets” sont traités dans ces 5 phrases et, pour chaque phrase, quel % de chaque sujet

## Solution

Si nous essayons de résoudre à la main ce problème nous avons l’impression qu’il n’y a que deux sujets dans ces phrases (sujets A et B).

Les phrases 1 et 2 s’occupent à 100% d’un seul sujet A – nous allons voir lequel. Les deux phrases suivantes semblent être dédiées à un autre sujet B, toujours à 100%. Enfin, dans la dernière phrase, nous avons deux sujets, A et B, en proportions égales.

En regardant la catégorie des mots, on en déduit que : Le **sujet A** est la nourriture et le **sujet B** les **animaux domestiques**.

Nous avons : comparé les mots dans un phrase et trouvé à chaque fois s’ils étaient «proches» (vis-à-vis d’un sujet) et quels % de la phrases ces mots proches occupaient-ils. Cela signifie qu’en même temps nous avons défini dans notre tête des sujets et les mots qui y sont attachés (en faisant quand même en sorte que ces sujets et les mots afférents soient partagés par les autres).

Au final, nous avons (sujet **A en rouge** et **sujet B en vert**):

I ate a **banana** and **spinach smoothie** for **breakfast. 100%**

I like to **eat broccoli** and **bananas**. **100%**

**Chinchillas** and **kittens** are cute. **100%**

My sister adopted a **kitten** yesterday. 100%

Look at this cute **hamster** munching on a piece of **broccoli**. **50%**, **50%**

**Nous avons fait une Assignation par Allocation de Dirichlet Latente.**

La méthode LDA doit estimer ces groupes de façon itérative, en ayant moins d’informations préalables mais en se basant sur les ressemblances.