

# Tout ce que vous avez toujours voulu savoir sur les tests multiples pour données génomiques sans jamais oser le demander

Café chouquettes statistique proposé par David Causeur

La biologie des systèmes a bénéficié ces dernières années d'une accessibilité croissante aux technologies de mesure à haut débit de l'expression d'un génome. Une grande diversité de méthodes statistiques adaptées à la forte hétérogénéité et à la grande dimension des données résultant de ces technologies permet aujourd'hui d'associer des gènes à l'expression de certains caractères phénotypiques d'intérêt. L'analyse des variations de transcription est l'objectif le plus souvent cité du développement de ce que l'on a appelé la statistique génomique mais une part importante de ce corpus de méthodes est transposable aux analyses d'association à grande échelle, à celles du protéome, du métabolome, ...

C'est le cas de la toute première étape de l'analyse de données à haut débit, qui consiste à réduire le trop vaste objet d'étude au sous-ensemble de taille plus raisonnable constitué de ses éléments liés de manière significative à une expression phénotypique ou une condition expérimentale. Cette opération que la plupart des génomiciens connaissent sous le nom d'analyse différentielle n'est rien d'autre pour le statisticien qu'une procédure de tests multiples, pas différente dans ses objectifs et dans ses grands principes de celle présentée dans tous les bons manuels d'introduction à la statistique depuis le début du XX<sup>ème</sup> siècle. Toutefois, face à l'impuissance de ce vieux pilier des cours de statistique à répondre aux attentes des génomiciens aux prises avec des données à haut débit, les procédures de test multiples se sont progressivement modifiées. Se voulant au plus près des besoins de chacun, les méthodes se sont multipliées dans la littérature biogéno-informatico-statistique et dans les logiciels, générant ainsi une cascade de questions nouvelles pour le biologiste : FDR ou FWER ? Bonferroni ou Benjamini-Hochberg ? Step-up ou step-down ? Quelqu'un peut-il me donner l'adresse mail d'un statisticien compétent et disponible ?

Cet exposé se veut un espace d'échanges de connaissances autour d'un café à destination de ceux qui se sentent un peu perdus dans la jungle des procédures de tests multiples, qui veulent faire part de leur expérience, de leur préférence pour telle ou telle méthode, et surtout qui cherchent une réponse (aussi claire que possible) aux questions mentionnées ci-dessus.