Retour d'expérience sur l'utilisation des moyens de calcul et de stockage du Mésocentre de Calcul Scientifique Intensif de l'Université de Lille

Benjamin Grenier-Boley*[1]

[1]Inserm U1167 – Inserm : U1167 – France

Résumé

Notre équipe, au sein de l'unité Inserm UMR-1167 RID-AGE située à l'Institut Pasteur de Lille, a pour principal objectif la caractérisation des facteurs de risque génétiques de la maladie d'Alzheimer. Pour cela, l'équipe réalise des analyses pangénomiques d'épidémiologie génétique qui considèrent de plus en plus de patients (dizaine de milliers, voire plusieurs centaines de milliers) sur de plus en plus de variations génétiques (plusieurs millions voire centaines de millions) grâce à la réduction des coûts de production des données de génotypage et de séquençage due aux évolutions technologiques mais aussi grâce aux multiples collaborations avec de nombreux partenaires internationaux réunis au sein de consortiums. Cela nécessite donc d'avoir accès à un stockage dimensionné avec des moyens de calculs appropriés.
Premièrement, le stockage doit être fiable et performant mais surtout pérenne. En effet, les données génétiques générées doivent être stockées indéfiniment de par leur coût de production, la quantité limitée de l'ADN qui a été utilisée et de par leur réanalyse au cours du temps par de nouvelles méthodes. Par ailleurs, certains de nos collaborateurs faisant partie de nos consortiums doivent pouvoir accéder à leurs données et les analyser dans un souci de mutualisation et d'uniformisation.

Deuxièmement, les moyens de calcul doivent être proches du stockage afin de limiter les multiples transferts d'une telle volumétrie. De plus, malgré l'explosion des calculs sur GPUs, la très grande majorité de nos pipelines et de nos analyses utilise encore des CPUs. D'autre part, nos calculs sont la plupart du temps facilement parallélisables en découpant chaque étape en de nombreuses tâches indépendantes (par exemple par région chromosomique). Cela nous a donc dirigé naturellement vers l'utilisation des supercalculateurs, ou clusters HPC (High Performance Computing), qui disposent de nombreuses ressources CPUs et mémoire.
Ces besoins spécifiques aussi bien au niveau du stockage que du calcul nous ont obligés à nous tourner vers des solutions externes à l'unité et à l'Institut. En analysant les offres des très grands centres de calcul pilotés par le Grand Equipement National de Calcul Intensif (GENCI), nous avons estimé que le fonctionnement en projets mais surtout la non pérennisation des données ne répondaient pas à tous nos besoins. C'est pourquoi nous avons choisi d'utiliser les services proposés par le Mésocentre de Calcul Scientifique Intensif de l'Université de Lille en termes de stockage et de calcul mais aussi d'y investir depuis maintenant 10 ans. Ainsi, le but de cette présentation serait d'en faire un retour d'expérience.

*Intervenant