Détection a posteriori de structure génétique des populations hiérarchisée

Maxime Pauwels[1], Adeline Coorneart[1], Sophie Gallina[1], Cyrille Bonamy[2], Jean-François Arnaud[1]

maxime.pauwels@univ-lille1.fr adelinecoornaert@yahoo.fr sophie.gallina@univ-lille1.fr cyrille.bonamy@univ-lille1.fr jean-francois.arnaud@univ-lille1.fr

[1]: Laboratoire GEPV - UMR CNRS 8198
Université Lille1 Bât. SN2
59655 Villeneuve d'Ascq Cedex– France

[2]: Centre de ressources informatiques (CRI)
Université Lille 1
59655 Villeneuve d'Ascq Cedex– France

ENJEUX SCIENTIFIQUES

La génétique des populations s'intéresse à la distribution de la diversité génétique à l'intérieur des espèces biologiques. Elle cherche notamment à identifier des sous-ensembles, appelés populations, entre lesquels les échanges génétiques sont réduits (Hartl & Clark, 2007). La particularité de ces sous-ensembles est de présenter des patrimoine génétiques différents (**Figure 1**). Identifier ces groupes au sein d'espèces biologiques d'intérêt est un enjeu majeur lorsqu'il s'agit, par exemple, de définir des unités sur lesquelles opérer dans le cadre d'un programme de conservation.

Figure 1. L'espèce humaine *Homo sapiens* peut être divisée en un certain nombre de populations génétiquement différenciées. La société américaine de biotechnologie DNA Tribes propose d'identifier, en utilisant quelques marqueurs moléculaires sur un échantillon de votre ADN, l'origine géographique de vos ancêtres.

Plusieurs outils informatiques, dits de regroupement, utilisant la statistique bayésienne sont aujourd'hui disponibles pour déterminer *a posteriori* le nombre et les limites des populations à partir de données de génotypage moléculaire d'un échantillon d'individus.

Nous avons testé l'efficacité d'un de ces outils, implémenté dans le logiciel STRUCTURE (Pritchard *et al.*, 2000) en analysant des jeux de données simulées à l'aide du logiciel NEMO (Guillaume & Rougemont, 2006), sous deux modèles définissant une structuration hiérarchisée de la diversité génétique, c'est-à-dire lorsque un nombre déterminé de populations sont aussi regroupées en un nombre déterminé de groupes de populations génétiquement isolés (**Figure 2**).

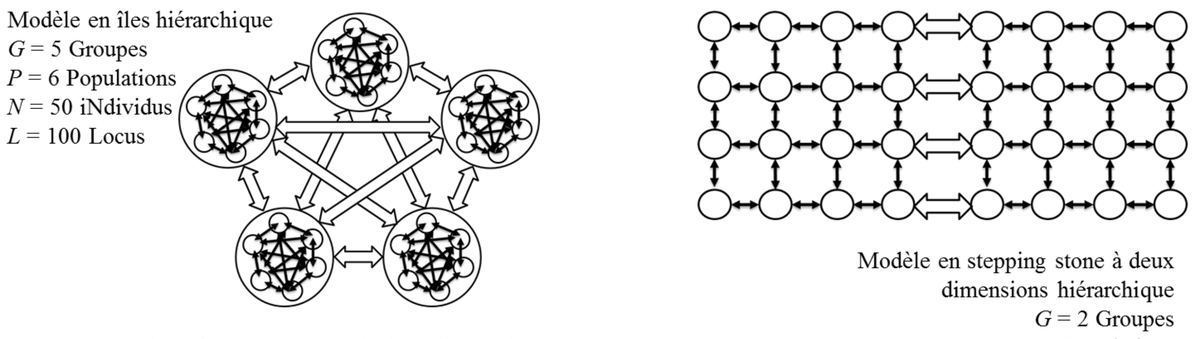

Figure 2. Modèles de simulation utilisés pour générer les données analysées. Le modèle en îles hiérarchique, à gauche, comprend 5 groupes (G, grands cercles) de 6 populations (P, petits cercles) comprenant chacune 50 individus (N). Le modèle en stepping stone à deux dimensions (à droite) comprend 2 groupes de 16 populations (petits cercles) comprenant chacune 50 individus. Dans les deux modèles, les groupes sont définis par des taux de migrations entre populations de groupes différents (flèches blanches) inférieurs aux taux de migration entre populations d'un même groupe (flèches noires). Le nombre de locus simulés (L) est égal à 100.

BESOINS EN CALCUL

Six paramétrages différents par le taux de migration entre groupes ont abouti à l'obtention de 2x6 jeux de données (**Figure 2**). Chacun des 12 jeux de données a ensuite été analysé en faisant varier le paramètre *K* (nombre de populations) à estimer de 1 à 40.

Pour chaque valeur de *K*, le logiciel a été lancé 10 fois afin d'estimer la variance statistique de la vraisemblance des paramètres correspondants (soit un total de 400 runs par jeux de données).

Pour tester la puissance de l'analyse dans des situations où la quantité d'informations apportées est réduite, les analyses ont ensuite été réalisées après sous-échantillonnage aléatoire, pour chaque jeu de données, du nombre de locus (3 réductions), du nombre d'individus par population (2 réductions) et du nombre de populations par groupe (1 réduction) Chaque sous-échantillonnage a été effectué 5 fois pour estimer l'effet statistique du sous-échantillonnage (**Figure 3**).

Au total, 1 113 600 runs du logiciel du STRUCTURE ont été nécessaire pour analyser les données (complètes et réduites) de chacun des 2 modèles de simulations. Les résultats d'analyse ont ensuite été rapatriés sur serveur local de notre laboratoire, compilés, reformatés et synthétisés pour permettre différentes représentations graphiques (**Figure 4**).

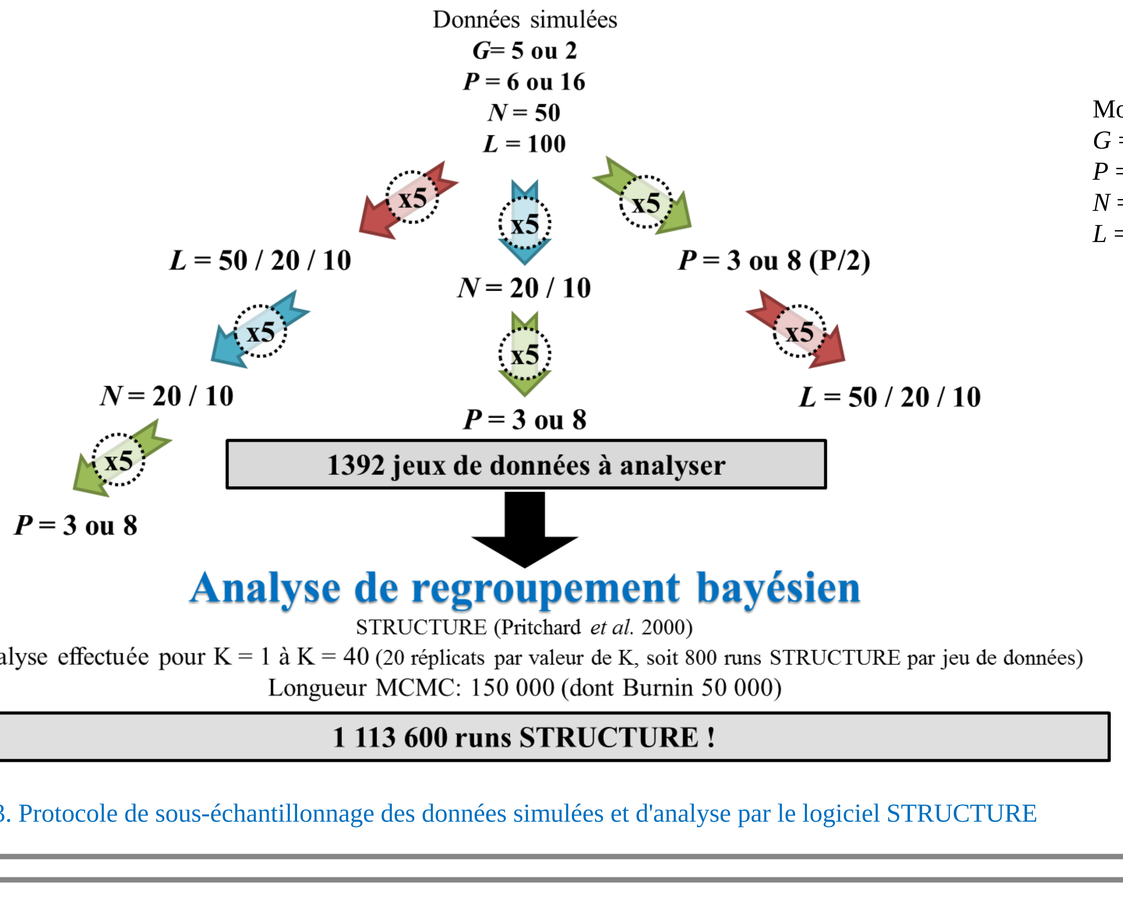

Figure 3. Protocole de sous-échantillonnage des données simulées et d'analyse par le logiciel STRUCTURE

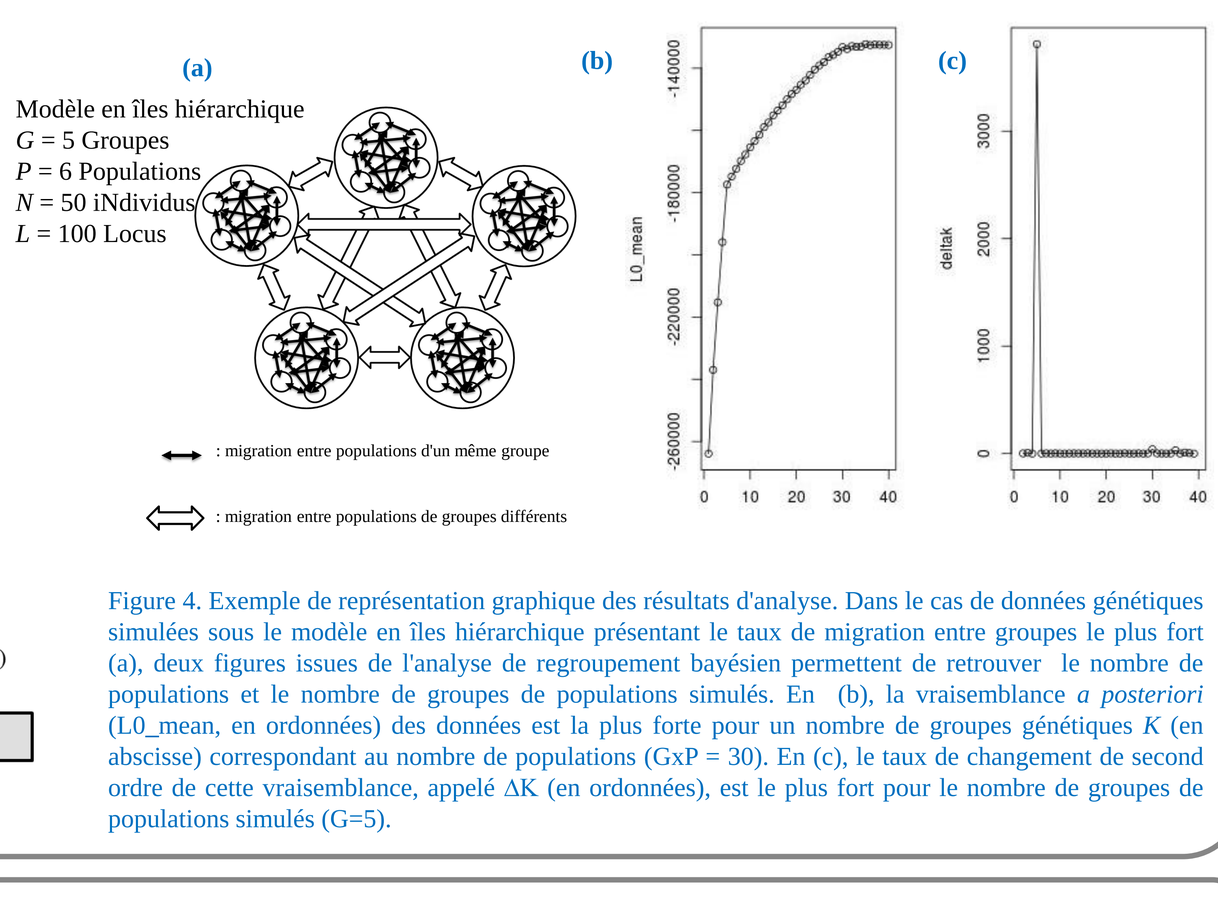

Figure 4. Exemple de représentation graphique des résultats d'analyse. Dans le cas de données génétiques simulées sous le modèle en îles hiérarchique présentant le taux de migration entre groupes le plus fort (a), deux figures issues de l'analyse de regroupement bayésien permettent de retrouver le nombre de populations et le nombre de groupes de populations simulés. En (b), la vraisemblance *a posteriori* (L0_mean, en ordonnées) des données est la plus forte pour un nombre de groupes génétiques *K* (en abscisse) correspondant au nombre de populations (GxP = 30). En (c), le taux de changement de second ordre de cette vraisemblance, appelé ΔK (en ordonnées), est le plus fort pour le nombre de groupes de populations simulés (G=5).

DÉVELOPPEMENT

Cluster du mésocentre Lillois

Simulation des 12 jeux de données

Analyse de un réplicat de sous-échantillonnage

Scripts Perl:
- Transfert des données
- Contrôle de qualité des résultats
- Analyse et synthèse des résultats

Soumission des jobs [*qsub - PBS*]

Temps de calcul: 125.000 jobs – 4 mois

Grille de Calcul Européenne

Simulation des 12 jeux de données

Analyse de 5 réplicats de sous-échantillonnage

Scripts Perl:
- Soumission des jobs [dirac-wms-job-submit – JDL
- Récupération des résultats [direc-wms-job-status, dirac-wms-job-get-output]
- Destruction des jobs [dirac-wms-job-delete]
- Contrôle de qualité des résultats
- Analyse et synthèse des résultats

Soumission sans re-compilation
~ 2% de jobs en échec

Temps de calcul: 1.113.6 jobs – 3 semaines

RÉFÉRENCES

Guillaume F, Rougemont J. 2006. Nemo: An evolutionary and population genetics programming framework. Bioinformatics 22(20): 2556-2557.

Hartl DL, Clark AG. 2007. Principles of population genetics. Sunderland, Massachusetts: Sinauer Associates, Inc.

Pritchard JK et al. 2000. Inference of population structure using multilocus genotype data. Genetics 155(2): 945-959.

Arrabito L et al. 2012. Instance nationale et multi-communauté de DIRAC pour France Grilles. Journées scientifiques mésocentre et France grilles 1-3 octobre 2012 Paris