Détection a posteriori de structure génétique des populations hiérarchisée

Maxime Pauwels (1), Adeline Coorneart (2), Sophie Gallina (3), Cyrille Bonamy (4), Jean-François Arnaud (5)

(1) maxime.pauwels@univ-lille1.fr, UMR-CNRS 8198
(2) adelinecoornaert@yahoo.fr, UMR-CNRS 8198
(3) sophie.gallina@univ-lille1.fr, UMR-CNRS 8198
(4) cyrille.bonamy@univ-lille1.fr, CRI – Université Lille 1
(5) jean-francois.arnaud@univ-lille1.fr, UMR-CNRS 8198

Overview:

Our study aim is to determine whether the Bayesian clustering method implemented in the STRUCTURE software is capable of detecting hierarchical population genetic structure. Using simulations under two hierarchical migration models, various sets of genotypic data were obtained, corresponding to various levels of genetic differentiation among groups of populations. Twelve datasets were analyzed with and without multiple sub-sampling of the number of genetic locus, individuals and populations. Finally, 556,800 runs of the STRUCTURE software were required for analyses. Runs were submitted to either the CRI cluster or the European grid using Perl scripts, and results were secondary analyzed on the local server.

Enjeux scientifiques, besoin en calcul, stockage et visualisation :

En génétique des populations, on définit la structure en population de la diversité génétique comme la distribution non-aléatoire de cette diversité dans différents ensembles d'individus appelés populations (Hartl & Clark, 2007). Par définition, ces populations représentent des ensembles entre lesquels les échanges génétiques par migration sont réduits, du fait de l'existence de barrières aux flux de gènes. Identifier ces groupes au sein d'espèces biologiques d'intérêt est un défi majeur lorsqu'il s'agit par exemple de définir des unités sur lesquelles opérer dans le cadre d'un programme de conservation. Plusieurs outils informatiques d'analyse bayésienne, dits "de regroupement", sont aujourd'hui disponibles pour le généticien des populations qui souhaite déterminer *a posteriori* le nombre et les limites de ces 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 l'utilisant pour analyser des jeux de données simulées 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.
Pour chacun des deux modèles, 6 paramétrages ont abouti à l'obtention de 6 jeux de données distincts, comprenant 100 marqueurs génétiques dont la diversité est distribuée dans 30 ou 32 populations regroupées en 5 ou 2 ensembles, respectivement. Chacun des douze jeux de données a ensuite était analysé à l'aide du logiciel STRUCTURE 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 correspondante (soit un total de 400 runs par jeux de données). 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. Au total, 278 400 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.

Développements, utilisation des infrastructures :

Les simulations de données ne nécessitant pas de grandes capacités de calcul, elles ont été effectuées à l'aide du logiciel NEMO (Guillaume & Rougemont, 2006) sur un ordinateur personnel, puis transférées sur le cluster du mésocentre Lillois (CRI de Lille 1). L'analyse des 12 jeux de données complets et d'un premier sous-échantillonnage des données a été réalisée sur le cluster du mésocentre Lillois. Des scripts PERL ont été développés pour lancer les différents jobs, contrôler la qualité des analyses effectuées, organiser la sauvegarde des résultats et transférer les résultats sur le serveur local de notre laboratoire. Etant donnée l'ampleur des temps de calcul nécessaire au cours de cette première étape (166 642 heures de calcul, soit 6943 jours ou 19 ans), l'analyse d'1 seul réplicat de sous-échantillonnage pour le nombre de locus, d'individus et de populations a été réalisée sur le cluster du mésocentre Lillois.
Pour réaliser les 5 réplicats de sous-échantillonnage nécessaire à ce projet, nous avons ensuite utilisé la Grille EGI, via la VO Biomed et l'Instance nationale et multi-communauté de DIRAC pour France Grilles (Arrabito *et al* 2012). Comme un ensemble de scripts PERL avait déjà été développé pour la solution « calcul sur cluster », nous avons choisi d'adapter ces scripts pour la solution « calcul sur grille via DIRAC ». Nous n'avons donc pas utilisé l'API Dirac en python. L'adaptation a principalement consisté à générer des fichier *JDL* au lieu de fichiers *PBS*, et à utiliser la commande *dirac-wms-job-submit* à la place de la commande *qsub*. Les identifiants des jobs soumis ont été stockés dans un fichier local, qui a ensuite servi à tester le statut des jobs avec la commande *dirac-wms-job-status,* à rapatrier les résultats avec *dirac-wms-job-get-output* et enfin à détruire les jobs avec *dirac-wms-job-delete*. Ces 4 primitives, d'utilisation très simples, permettent de gérer les jobs par lot. Nous avons utilisé des lots de 100 jobs pour vérifier les statuts, récupérer les résultats des N jobs / 100 qui étaient terminés puis les détruire. Cette technique diminue la charge du serveur DIRAC et donc réduit les temps de réponse. Au

total, les 3 scripts pour 1) soumettre les jobs, 2) rapatrier les résultats et 3) détruire les jobs représentent 250 lignes de code Perl, donc un très faible investissement en temps de développement. Par ailleurs, nous avons choisi d'utiliser directement une version exécutable pour processeur 64bits du programme Structure, sans chercher à le recompiler sur les nœuds de la grille, car ce programme n'utilise que des librairies standards. Environ 2% des jobs lancés ont retourné un statut d'erreur.

Outils, le cas échéant complémentarité des ressources, difficultés rencontrées :

Les besoins informatiques pour notre projet concernant principalement des calculs paramétriques, le mesocentre Lillois nous semblait bien adapté. Cela nous a permis de réaliser une première étape dans ce projet en analysant un premier réplicat de ré-échantillonnage. Cependant, l'analyse complète de tous les réplicats ne pouvait pas être réalisée avec les seuls ressources du mésocentre. L'ingénieur responsable du mésocentre nous a alors orienté vers l'utilisation de l'instance nationale DIRAC et nous a fourni une aide précieuse pour les aspects administratifs de demande d'entrée dans la grille, le choix et l'installation des outils et des certificats et enfin une aide technique pour l'utilisation des ressources via des scripts PERL de base pour la soumission et la récupération des résultats. Nous avons ensuite adapté ces scripts afin de lancer les commandes par lots pour la récupération des résultats, ce qui a réduit le temps pour nous et probablement allégé la consommation de ressources sur l'instance DIRAC. Le mésocentre Lillois teste actuellement la possibilité de faire des soumissions de jobs paramétriques, nous prévoyons donc d'utiliser cette technique dès qu'elle sera validée par le mésocentre.

Résultats scientifiques :

Les résultats permettent de discuter la puissance de détection d'une structuration génétique de la diversité génétique par le logiciel de regroupement bayesien STRUCTURE en fonction du niveau d'isolement génétique entre les populations ou groupes de populations et de la quantité d'information disponible (nombre de locus, d'individus, de populations). A travers la comparaison des résultats obtenus à partir des différents jeux de données, nous avons pu obtenir une lecture plus instruite et donc plus critique des représentations graphiques des résultats d'analyse. Nous avons pu montrer que le logiciel était particulièrement sensible aux faibles niveaux de différentiation entre population/groupes de populations et à la quantité d'information analysées, notamment le nombre de locus.

Perspectives :

L'analyse effectuée, parce qu'elle nécessitait l'exploitation de capacités de calcul importantes, était inimaginable sans les ressources fournies par la grille. Par conséquent, notre protocole a initialement été défini en réduisant les modalités de tests des capacités du logiciel STRUCTURE, notamment par exemple, en limitant les paramétrages possibles du logiciel. Le significatif gain de temps associé à l'utilisation de la grille rend envisageable une complémentation des résultats obtenus par une multiplication des conditions d'analyses de jeux de données génétiques simulés.

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**