QGP: Quantitative Genetics Platform. A high performance computing solution for quantitative genetics software

Misharl Monsoor, André Neau, Martin Souchal, Sylvie Nugier, François Laperruque, Eddie Iannuccelli, Pascale Le Roy, Edmond Ricard, David Robelin, Olivier Filangi

QGP : Quantitative Genetics Platform

A high performance computing solution for quantitative genetics software

Misharl Monsoor[12], André Neau[3], Martin Souchal[4], Sylvie Nugier[4], François Laperruque[5], Eddie Iannuccelli[6], Pascale Le Roy[12], Edmond Ricard[5], David Robelin[6] and Olivier Filangi[12]

[1] INRA, PEGASE Physiologie, Environnement et Génétique pour l'Animal et les Systèmes d'Elevage, UMR1348, 35590 Saint-Gilles, France

[2] INRA, PEGASE Physiologie, Environnement et Génétique pour l'Animal et les Systèmes d'Elevage, UMR1348, 35000 Rennes, France

{misharl.monsoor, pascale.leroy, olivier.filangi}@rennes.inra.fr

[3] INRA, GABI Génétique Animale et Biologie Intégrative, UMR1313, 78532 Jouy-En-Josas,France

andre.neau@jouy.inra.fr

[4] INRA, CTIG Centre de Traitement de l'Information Génétique, US0310, 78532 Jouy-En-Josas, France

{martin.souchal, sylvie.nugier}@jouy.inra.fr

[5] INRA, SAGA Station d'Amélioration Génétique des Animaux, UR0631, 31326 Castanet Tolosan, France

{francois.laperruque, edmond.ricard}@toulouse.inra.fr

[6] INRA, LGC Laboratoire de Génétique Cellulaire, UR444, 31326 Castanet Tolosan, France

{eddie.iannuccelli, david.robelin}@jouy.inra.fr

1 Background

1.1 Genotype file format and conversion

Un des besoins croissants des généticiens quantitatifs est la manipulation des fichiers de génotypage. Ces fichiers contiennent les génotypes des individus, en général de quelques centaines à quelques milliers d'individus, en chaque marqueur le long d'un génome. Les premiers développements logiciels d'analyse génétique prévoyaient l'exploitation d'informations sur des marqueurs microsatellites, avec quelques centaines de marqueurs génétiques par individu. Aujourd'hui, les puces de génotypage à haute densité peuvent produire une quantité d'information bien plus élevée, allant par exemple aujourd'hui jusqu'à huit cent mille marqueurs génétiques (Single Nucleotide Polymorphisms) sur le génome des bovins. Les fichiers de données atteignent plusieurs Gigaoctet et sont difficilement manipulables par les tableurs et logiciels (R/SAS) communément utilisés par la communauté. De plus l'automatisation des processus de vérification et de cohérence des données, caractérisée par l'exécution de milliers de processus est devenue un préalable indispensable à l'analyse des caractères complexes.

1.2 Heterogeneous computing environment

Les méthodologies et outils se sont adaptés aux nouvelles architectures utilisés dans un contexte de calcul scientifique. Dans un souci de performance, l'utilisateur est amené à utiliser plusieurs modes d'exécution : soumission de job sur une grappe de serveurs, calcul déporté sur un processeur graphique (GPU), exécution en environnement multithreadé. Ces ressources sont la plupart du temps localisées dans plusieurs infrastructures, ce qui complexifie le workflow d'exécution. L'utilisateur doit, en outre, transférer ces fichiers de serveur en serveur.

2 Implementation

2.1 Galaxy : A Web unified interface

La plate-forme QGP (Quantitative Genetics Platform, https://qgp.jouy.inra.fr) a pour objectif de répondre à ces besoins. QGP implémente son environnement d'exécution sur le framework Galaxy [1,2,3]

qui fournit un ensemble d'outils d'analyse, de manipulation et de visualisation des données génomiques pour la communauté bio-informatique. Galaxy est capable de gérer des fichiers de séquence de grande taille comme le transfert, la suppression ou l'insertion de lignes ou de colonnes. Ces fichiers étant comparables aux fichiers de genotypage utilisés par les outils de génétique quantitative, les outils de manipulation de fichiers restent accessibles de la plate-forme QGP. Galaxy est conçu pour exécuter des chaînes de traitements prédéfinies (Workflow) et applicables aux nouveaux lots de données.

2.2 A variety of quantitative genetics software tools

Les logiciels hébergés par la plate-forme sont implémentés pour plusieurs types de serveurs de calcul et adressent plusieurs problèmes de génétique quantitative dans le domaine animal. L'analyse de la variabilité génétique des caractères complexes dans des populations expérimentales (QTLMap [4]), l'évaluation génétique (GS3, GABayes), l'analyse de parenté (PEDIG [5]), la détection d'incompatibilité de génotypes (Mendelsoft [6]). Un ensemble d'outils pour la conversion de format et la génération de graphiques sont également accessibles. Galaxy est capable de soumettre des exécutions de jobs sur des serveurs, distants et hétérogènes. Cette fonctionnalité permet de déployer des implémentations spécifiques, telles que les implémentations de QTLMap sur GPU [7], mais aussi des parallélisations naïves (indépendances des vérifications des génotypes aux marqueurs, reconstruction d'haplotypes par famille de demi ou plein frères,...) de façon simple et transparente à l'utilisateur.

3 Conclusion

L'utilisation de données sur des séquences complètes est d'ores et déjà évoquée pour l'analyse des caractères ou l'évaluation génétique des animaux d'élevage. Par ailleurs, il est probable que les phénotypes seront bientôt également difficiles à gérer, en quantité et en diversité, l'exemple du traitement des données de transcriptome ou de métabolome étant déjà une réalité pour beaucoup de généticiens. Le choix du framework Galaxy, imaginé à l'origine pour gérer des données de séquençage, permettra donc de suivre plus aisément cette évolution attendue du volume des données à traiter conjointement. De plus, il permet de profiter des evolutions de Galaxy (adaptation aux serveurs de calcul émergents, outils générique pour la manipulation de fichiers). Enfin, le choix de cet outil devrait également faciliter les connexions, devenues indispensables pour la biologie intégrative, entre le monde de la bioinformatique, où la culture est celle de l'intégration des connaissances et de la visualisation ou de l'extraction des résultats, et celui de la génétique quantitative, où la culture est celle de la modélisation et du calcul scientifique.

References

[1] Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010 Aug 25;11(8):R86.

[2] Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. "Galaxy: a web-based genome analysis tool for experimentalists". Current Protocols in Molecular Biology. 2010 Jan; Chapter 19:Unit 19.10.1-21.

[3] Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. "Galaxy: a platform for interactive large-scale genome analysis." Genome Research. 2005 Oct; 15(10):1451-5.

[4] Filangi.O , Moreno C. , Gilbert H. , Legarra A. , Le Roy P. , Elsen JM. *QTLMap, a software for QTL detection in outbred populations.* 9th World Congress on Genetics Applied to Livestock Production, Leipzig. 2010.

[5] Boichard D., 2002. Pedig : a fortran package for pedigree analysis suited to large populations. 7th World Congress on Genetics Applied to Livestock Production, Montpellier, 19-23 août 2002, paper 28-13.

[6] Givry S.,Palhiere I.,Vitezica Z.G., Schiex T., Mendelian error detection in complex pedigree using weighted constraint satisfaction techniques, Workshop on Constraint Based Methods for Bioinformatics, Spain, Octobre 2005

[7] Chapuis G., Filangi O., Leroy P., Elsen JM., Lavenier D., *GPU Accelerated QTLMap*, The 15Th QTL-MAS Workshop, Rennes, 2011