BayPass 2.3: tutoriel de génomique d'association adapté au séquençage en pool

Jérôme OLIVARES[1]

2023-06-28

[1]INRAE, UR-1115 PSH, 228 route de l'aérodrome, 84914 Avignon, France

Contents

Résumé **3**

prérequis 3

Introduction **4**

Présentation du logiciel BayPass 2.3 **5**

Présentation générale de l'analyse : 5

Données brutes & Obtention d'un fichier Poolseq.vcf : 6

Résumé

Ce tutoriel détaille d'une part la manière de générer les fichiers d'entrées du logiciel BayPass à partir de données de séquençage en pool, d'autre part le paramétrage optimal du logiciel BayPass et en fin propose une méthode d'exploreration des résultats d'analyses produits. Un pipeline d'analyse mixant des packages sous Rstudio et des lignes de commandes Linux, est décrit afin de guider pas à pas l'utilisateur tout au long du processus depuis les données brutes jusqu'à la liste finale de loci/variants candidats. Ce tutoriel est à destination des étudiants et des bioinformaticiens débutants.

mots clefs

Logiciel BayPass, séquençage en pool, GWAS, études d'associations pangénomiques, Rstudio.

prérequis

Les commandes décrites dans cet article ont été regroupées dans un fichier au format « R markdown » (Rmd) « Poolseq_pipeline.Rmd » librement téléchargeable à l'adresse : https://github.com/Jolivares-INRAE/Download. Ce tutoriel est conçu pour décrire pas à pas les différentes étapes du fichier Rmd et permettre à l'utilisateur de les exécuter en parallèle. L'utilisateur devra avoir une connaissance basique du logiciel Rstudio et être capable d'écrire et lancer des scripts sur un cluster de calcul. Les commandes ont été rédigées sous Rstudio version 1.4.1106 couplé à R 64 bits version 4.0.5. avec toutes les librairies nécessaires à jour (Capture 1) et dans l'environnement bash/SLURM des clusters de calculs de la plateforme GenoToul de bioinformatique (GenoToul Bioinfo). Dans le cas d'une utilisation dans un autre environnement logiciel, l'utilisateur devra probablement effectuer des adaptations du code. Bien que l'essentiel des calculs de BayPass seront réalisés sur un cluster de calcul, certains de ses utilitaires seront utilisés en local sous Rstudio, la dernière version du logiciel sera donc téléchargée depuis l'adresse http://www1.montpellier.inra.fr/CBGP/software/baypass/download.html et décompressée dans un répertoire local par l'utilisateur. Dans tous les codes qui suivent l'expression « ~/path/ » sera à remplacer par les chemins personnels de l'utilisateur. Le terme de chromosome sera utilisé en références aux appellations de contigs, scaffold, ou chromosomes qui correspondent aux séquences nucléotidiques du génome de référence, plus ou moins mature, qui sera utilisé.

Introduction

Dans un contexte agronomique actuel de réduction de l'utilisation des pesticides ou de réchauffement climatique, analyser et comprendre les bases génétiques de l'adaptation des organismes aux méthodes de luttes qui leurs sont opposées ou à l'évolution de leur environnement est un enjeu majeur des recherches de ces dernières années.

Les études d'associations pangénomiques (GWAS en anglais pour genome-wide association study) adossées aux techniques de séquençage haut débit (NGS) permettant le séquençage de génome complet, sont un outil de choix pour ce type d'analyse. L'étude au niveau populationnel demandant le séquençage d'un grand nombre d'individus, les couts d'analyses étaient initialement très élevés et ces études étaient souvent réservées aux organismes dit "modèles", humain en tête. Néanmoins il a été démontré depuis, que le séquençage en pool d'individus (poolseq) c'est-à-dire en mélangeant de manière équimolaire l'ADN d'un grand nombre d'individus (50 à 100) issus d'une même population permettait non seulement de réduire drastiquement les couts puisqu'on ne réalise qu'un seul séquençage mais aussi que la découverte des points de mutations (SNP) et l'estimation de leur fréquence allélique étaient souvent plus efficaces et précises [Futschik and Schlötterer, 2010]. Parmi les logiciels à même d'analyser ces fréquences alléliques on compte Baypass [Gautier, 2015] qui évalue par une approche baysienne, la différenciation des SNP en liaison avec une covariable environnementale en tenant compte de la structure et de la parenté entre les populations en estimant la covariance (Ω) des fréquences alléliques. Baypass 2.3 a la particularité supplémentaire de pouvoir calculer un contraste des fréquences alléliques entre deux groupes de populations caractérisés par un caractère binaire, sensible ou résistant par exemple.

La documentation disponible de BayPass 2.3 décrit par le menu les algorithmes de fonctionnement et les différents paramètres du logiciel mais reste néanmoins, et assez logiquement, succincte sur les étapes en amont et en aval. A ma connaissance un seul tutoriel est disponible sur le web [Nielsen, 2020] et décrit de manière plus détaillée la préparation des données brutes et l'analyse en mode « poolseq » de Baypass, mais il nécessite des bases quelque peu avancées de codage sous R et en environnement bash. Si les bio-informaticiens chevronnés ne rencontreront pas de difficultés particulières, il n'en va pas forcément de même pour bon nombre d'agents que les orientations des recherches associées à la baisse des coûts de séquençage ont aiguillé vers les voies de la génomique. Cet article a pour but de détailler les différentes étapes d'une analyse de type GWAS/poolseq avec le logiciel BayPass 2.3 et d'éclairer les points qui sont habituellement peu explicités car considérés comme évident et se veut, au final, le plus proche possible de la citation de Talleyrand : « Si cela va sans le dire, cela ira encore mieux en le disant ».

Présentation du logiciel BayPass 2.3

Le logiciel BayPass est un logiciel de génomique des populations qui vise principalement à l'identification de marqueurs génétiques soumis à la sélection et/ou associés à des covariables spécifiques à la population (variables environnementales, phénotypiques, quantitatives, catégorielles…). Par une approche baysienne il évalue une **matrice Ω** de covariance des fréquences alléliques des populations résultant de leur histoire démographique. Deux manières d'estimer ces fréquences alléliques sont disponibles soit en se basant sur les génotypes référence/mutant des individus analysés soit, lorsque l'on active le « pool-seq mode », ces fréquences alléliques sont calculées en regard de la profondeur de séquençage (reads count) et pondérées par le nombre d'individus qui ont contribué à cette profondeur. C'est cette seconde approche que nous considérerons dans cet ouvrage.

BayPass propose 3 modèles statistiques d'analyse :

Le Core Model

C'est le modèle de base, il permet de calculer la matrice de covariance Ω et d'attribuer une statistique de différenciation XtX à chaque SNP, et ainsi de scanner le génome pour identifier les régions génomiques différenciées entre les populations.Le XtX est une statistique analogue au Fst mais tient compte de la co-évolution des populations grace à la matrice Ω.

Le Standard Model

Ce modèle, permet, l'orsque l'on fournit une ou plusieurs covariables (environnementales, phénotypiques…), de calculer un facteur de Bayes, ou Bayes factor en anglais, (BF) pour chaque marqueur génétique représentant la force d'association avec une covariable. C'est un modèle tout en un qui intègre le calcul des XtX et de la matrice Ω, il est adapté au faible nombre de population (< 15).

L'Auxiliairy Model

Ce modèle a une approche différente dans le calcul de la statistique BF, sans entrer dans le détail il est plus adapté au grand nombre de population (> 15), En contrepartie il nécessite que l'on fournisse une matrice Ω déjà calculée par une analyse Core Model précédente, il recalcule alors les XtX et la statistique BF.

Ces covariables évoquées doivent être distribuées en gradient entre les populations (différence de température, d'altitude…), en complément, dans le cas où la covariable étudiée serait purement binaire (sensible/résistant, gros/petit…), les modèles Standard et Auxiliaire peuvent calculer une statistique C2 qui évalue le contraste de différence des fréquences alléliques de chaque marqueur entre 2 groupes de populations.

Présentation générale de l'analyse :

La Figure 1 est une vision simplifiée des différentes étapes nécessaire à l'analyses de données poolseq. Les étapes nécessittant une importante puissance de calcul comme le variant calling ou l'analyse BayPass se déroulent soit dans l'environnement Linux du cluster de calcul, les étapes de filtrage,

manipulation de données et de résultats se font sur ordinateur local sous Rstudio. La première étape part des fichiers d'alignement au format « ***bam*** » de chaque population à analyser et consiste à effectuer une recherche de variants (variant calling) pour obtenir un fichier au format « ****vcf*** » regroupant tous les points de mutations ou SNP de toutes les populations qui sont autant de marqueurs génétiques à analyser. Ce fichier ***vcf*** sert d'entrée au package « PoolFstat » [Gautier et al., 2022] qui va permettre de filtrer les SNPs à analyser et générer les fichiers nécessaires au bon fonctionnement de BayPass mais aussi de faire une première analyse des Fst entre populations par exemple. Ces fichiers d'entrées pouvant contenir plusieurs millions de SNP, ils sont découpés en plusieurs dizaines de sous jeux de données (sub sampling) afin de réduire les temps de calculs. Une fois que BayPass a analysé tous les sous jeux de données, l'homogénéité des résultats entre eux est analysée sous Rstudio puis les résultats peuvent être regroupés, filtrés et analysés afin de déterminer les marqueurs génétiques et les régions chromosomiques d'intérêts qui seront visualisées par différents plots.

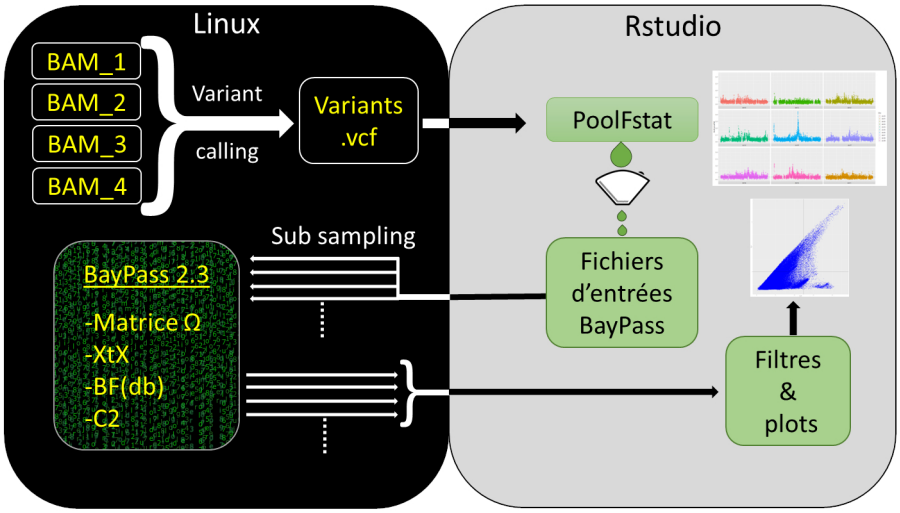

Figure 1: Pipeline d'analyse BayPass

Le pipeline d'analyse est décomposé en plusieurs étapes se déroulant soit en environnement Linux pour celles nécessitant une importante capacité de calcul, soit sous Rstudio pour le filtrage et l'analyse des résultats.

Données brutes & Obtention d'un fichier Poolseq.vcf :

Les étapes de contrôle qualité et d'alignement des données de séquençage sont largement documentés par ailleurs, et ne serons donc pas documenter ici, et nous partirons directement des fichiers d'alignement ***bam***. La première étape consiste à regrouper les fichiers ***bam*** de toutes les populations en un seul fichier puis à effectuer le variant calling avec un logiciel dédié compatible avec le séquençage en pool afin de conserver les informations de profondeur. Nous recommandons l'utilisation des Samtools [Li et al., 2009] et de Varscan 2 v2.3.6 [Koboldt et al., 2012]avec une instruction pipe entre les deux pour éviter les fichiers intermédiaires et économiser l'espace de travail Figure **??** **??** . Les commandes sont effectuées avec les paramètres de base, sauf la p-value qui est montée à 0.5 pour être le moins stringent possible à ce stade. On peut découper le travail en plusieurs chromosomes pour réduire les temps de calculs.

```
#!/bin/bash
#SBATCH --array=1-29            #création de l'array: un élément par chromosome

module load bioinfo/samtools-1.12
module load bioinfo/VarScan-2.4.2
module load bioinfo/bcftools-1.14

ls /../*.bam > BamList.txt
```

```
ls /../*.bam | sed -r 's/^.+\///' | sed -r 's/.bam//' > NameList.txt

samtools mpileup -C 50 -d 5000 -q 20 \
-r chr${SLURM_ARRAY_TASK_ID} \
-f /../ref_genome.fas -b /../BamList.txt | \
java -Xmx2G -jar $VARSCAN mpileup2cns \
--variants --min-coverage 10 \
--min-avg-qual 20 --min-var-freq 0.05 \
--p-value 0.5 --output-vcf 1 \
--vcf-sample-list NameList.txt > /../project_chr${SLURM_ARRAY_TASK_ID}.vcf

bgzip /../project_chr${SLURM_ARRAY_TASK_ID}.vcf
```

Si l'on veut découper le travail en chromosomes, il est indispensable de travailler sur des fichiers bam correctement indexés et d'utiliser l'option -r/–region qui tire profit de cet index.

IMPORTANT : Les chromosomes sexuels ayant une évolution historique différente des autosomes il conviendra, lorsque cela est possible, de les analyser à part.

```
knitr::include_graphics("C:/Users/Olivares/Documents/R/Git_Work/BayPass_Tutorial/images/Analyse2.jpg")
```

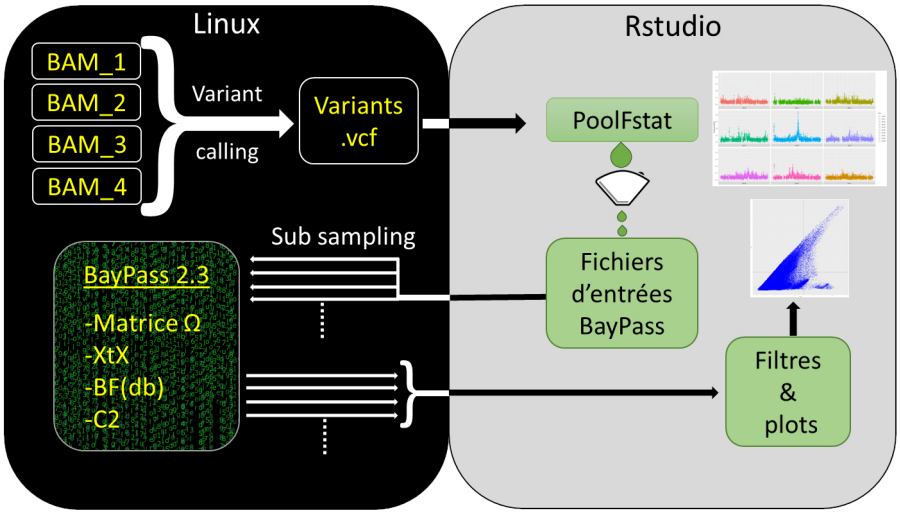

Figure 2: My caption

Bibliography

Andreas Futschik and Christian Schlötterer. The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics*, 186(1):207–218, September 2010. ISSN 1943-2631. doi: 10.1534/genetics.110.114397. URL https://academic.oup.com/genetics/article/186/1/207/6063740.

Mathieu Gautier. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics*, 201(4):1555–1579, December 2015. ISSN 1943-2631. doi: 10.1534/genetics.115.181453.

Mathieu Gautier, Renaud Vitalis, Laurence Flori, and Arnaud Estoup. f-statistics estimation and admixture graph construction with Pool-Seq or allele count data using the R package poolfstat. 2022.

Daniel C. Koboldt, Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, March 2012. ISSN 1088-9051. doi: 10.1101/gr.129684.111. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.129684.111.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/btp352. URL https://academic.oup.com/bioinformatics/article/25/16/2078/204688.

Erica S. Nielsen. Pool-Seq Analyses: PoolFstat & BayPass, 2020. URL https://esnielsen.github.io/post/pool-seq-analyses-poolfstat-baypass/.