UNIVERSITÉ GRENOBLE-ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE-ALPES

Spécialité : **Modèles, méthodes et algorithmes en biologie, santé et environnement**
Arrêté ministériel : ?

Présentée par

Keurcien LUU

Thèse dirigée par **Michael BLUM**

préparée au sein du laboratoire **Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble (TIMC-IMAG)**
et de l'école doctorale **"Ingénierie de la Santé, de la Cognition et Environnement" (EDISCE)**

Méthodes statistiques en grande dimension pour l'étude de l'adaptation biologique à l'aide de larges bases de données génomiques

Thèse soutenue publiquement le 31 octobre 2017,
devant le jury composé de :

Remerciements

The preface pretty much says it all.
Second paragraph of abstract starts here.

Préface

The preface pretty much says it all.

Table des matières

Liste des tableaux

Table des figures

Abstract

The preface pretty much says it all.
Second paragraph of abstract starts here.

Introduction

Données en grande dimension

L'accumulation de données, aussi bien en termes d'observations qu'en termes de variables, laisse à penser que le traitement de celles-ci pourrait permettre de détecter efficacement les variables qui sont responsables ou qui influencent un phénomène particulier. Cela pourrait être par exemple l'utilisation de bases de données automobiles pour prédire la durée de vie de véhicules neufs, ou encore celle de données météorologiques pour savoir s'il pleuvra ou non dans les jours qui viennent. Cette accumulation massive s'accompagne tout de même d'un phénomène bien connu en statistiques, phénomène qui porte le nom de "curse of dimensionality" (Giraud, 2014).

Chapitre 1

État de l'art

— Modèle de FLK
— Modèle de OutFLANK
— Modèle de Bayescan
— Fast PCA
— ACP en génétique des populations
— Partie III : R package pcadapt

1.1 Analyse en Composantes Principales parcimonieuse

1.2 Bootstrap ACP

1.3 Contexte

1.4 Tests multiples

1.5 Contrôle du taux de fausse découverte

Le taux de fausse découverte, correspond à la proportion de faux positifs parmi les positifs. En notant FP le nombre de faux positifs, FP le nombre de vrais positifs, on définit le taux de fausse découverte FDR par :

$$FDR = \mathbb{E}\Big[\frac{FP}{TP+FP}1_{FP+TP>0}\Big]$$

- Référence cours de Christophe Giraud

q-value, bonferroni, benjamini-hochberg La figure suivante donne les comparaisons entre les différentes procédures de correction :

Chapitre 2

Adaptation locale

Une population est dite localement adaptée à son environnement si elle a connu une évolution différente de celles qu'ont connu les autres populations de la même espèce, et ce, en réponse aux pressions sélectives auxquelles elle peut être confrontrée.

2.1 Cas d'étude utilisant pcadapt

2.2 POOP

Chapitre 3

Introgression adaptative

3.1 Qu'est-ce que l'introgression ?

Avant de s'intéresser à la notion d'introgression, intéressons-nous d'abord à celle d'hybridation. L'hybridation peut être définie comme la reproduction entre deux individus appartenant à deux espèces ou à deux populations différentes. Cette définition nous amène à nous poser deux questions. La première, relative à la notion d'espèce, est souvent sujette à controverse. La seconde concerne quant à elle la désignation de populations différentes. Qu'est-ce qui fait que deux groupes d'individus sont différents ? Harrison suggère en 1990 que deux individus issus de populations différentes doivent chacun posséder des traits héritables qui les différencient (Harrison & others, 1990).

Nous parlons d'introgression lorsqu'un certain nombre de gènes est transféré d'une population à une autre.

L'étude de régions génomiques présentant des caractéristiques d'introgression ou de divergence peut se révéler intéressante pour plusieurs raisons.

3.2 Coefficients de métissage globaux et locaux

Étant données des populations ancestrales, il est possible d'estimer pour un individu donné, la proportion de son génôme provenant de chacune des populations ancestrales. Ces proportions sont connues plus communément sous le nom de *coefficients de métissage globaux*. De nombreux logiciels existent pour l'estimation de ces coefficients : STRUCTURE, ADMIXTURE (Alexander, Novembre, & Lange, 2009), LEA (Frichot & François, 2015), tess3r (Caye, Deist, Martins, Michel, & François, 2016). En complément à cette information globale, il peut être intéressant de déterminer sur des portions plus petites du génôme, de la même manière que dans le cas global, les proportions venant de telle ou telle population ancestrale pour chacune de ces portions. Nous parlons dans ce cas de *coefficients de métissage locaux*. Encore une fois, plusieurs logiciels ont été proposés dans le but d'estimer ces coefficients : Hapmix (Price et al., 2009), EILA (Yang, Li, Buu, & Williams, 2013), LAMP (Thornton & Bermejo, 2014), loter ou encore RFmix (Maples, Gravel, Kenny, & Bustamante, 2013).

3.3 Introgression

L'introgression peut être détectée de différentes façons. Une première approche consiste à utiliser les *coefficients de métissage locaux*. Les méthodes mentionnées plus haut estiment ces coefficients pour chaque individu, permettant de calculer à partir de ceux-ci des coefficients de métissage locaux pour chaque population.

3.4 Lien entre Analyse en Composantes Principales et métissage global.

L'un des premiers articles à établir un lien entre l'ACP et les coefficients de métissage global fut sur l'interprétation généalogique de l'ACP de Gil McVean (McVean, 2009) :

Coefficients de métissage et ACP

Pour chacun des 22 chromosomes,

3.5 Analyse en Composantes Principales locale

Notant p le nombre de marqueurs génétiques, i un entier compris entre 1 et p, et x_i la position génétique (en Morgans) ou la position physique (en paires de bases) du i-ème marqueur génétique. Nous définissons pour cet entier i la fenêtre W_i^T de taille T et centrée en i :

$$W_i^T = \{j \in [|1, p|], |x_i - x_j| \leq T/2\}$$

3.6 Sensibilité à l'imputation des données manquantes

3.6.1 Méthodes de détection

Etat de l'art

Scénario à flux de gènes

La statistique D de Patterson La statistique D de Patterson (Durand, Patterson, Reich, & Slatkin, 2011) demeure aujourd'hui la méthode la plus utilisée pour détecter la trace de flux de gènes dans une population. La méthode repose sur l'observation de motifs portant les noms $ABBA$ et $BABA$, en référence aux différents types de généalogie possible pour un site nucléotidique.

$$D = \frac{\sum_i C_{ABBA}(i) - C_{BABA}(i)}{\sum_i C_{ABBA}(i) + C_{BABA}(i)}$$

RNDmin [Descriptif de RNDmin]

Pour comprendre la récente méthode proposée par (Rosenzweig, Pease, Besansky, & Hahn, 2016), il est nécessaire de définir un certain nombre de statistiques dont $RND_{\min}$ dérive.

— d_{xy} est la distance de Hamming entre la séquence X de la population 1 et la séquence Y de la population 2. Ainsi, si $x = (x_i)_{1 \leq i \leq n}$ et $y = (y_i)_{1 \leq i \leq n}$, alors :

$$d_{xy} = \text{Card}(\{i \in [|1, n|] \mid x_i \neq y_i\})$$

Cette statistique est ainsi définie pour une paire de séquences (x, y). Pour quantifier la dissimilarité entre deux ensembles de séquences, deux approches sont possibles. La première consiste à calculer la distance moyenne d_{XY}. De par sa définition, cette distance présente néanmoins le défaut d'être peu sensible aux épisodes récents d'introgression (Geneva, Muirhead, Kingan, & Garrigan, 2015). En effet, les faibles valeurs de d_{xy} correspondant à des évènements de divergence récents peuvent voir leur influence diminuée en cas de présence d'évènements de divergence plus anciens. Pour pallier à ce problème, considérer la distance minimale entre les deux ensembles de séquences (Joly, McLenachan, & Lockhart, 2009) constitue une solution intéressante.

[Expliquer pourquoi on définit d_{out} et RND]

En définissant $d_{out} = \frac{1}{2}(d_{XO} + d_{YO})$, il est possible de définir de la même façon RND_{min} :

$$RND_{\min} = \frac{d_{\min}}{d_{\text{out}}}$$

Pour récapituler, $RND_{\min}$ est une statistique robuste aux variations de taux de mutation et qui reste sensible aux récents évènements d'introgression. En pratique, l'introduction de d_{out} requiert ainsi la donnée d'une population ancestrale commune aux deux populations d'intérêt.

Bdf

Analyse Linéaire Discriminante

Régression linéaire, régression logistique, forêts aléatoires et importance des variables

Régression locale, package mgcv, locfit, Backward selection strategy

ACP locale et espace de formes

3.7 Simulations

3.7.1 Données de peupliers

Le premier jeu de données est issu d'une étude d'introgression adaptative chez les peupliers d'Amérique du Nord (Suarez-Gonzalez, 2016). La simulation d'haplotypes d'individus admixés est effectuée à partir des deux populations ancestrales qui y sont présentes. La première, *Populus Balsamifera*, est une espèce de peupliers qui peuple le nord du continent nord-américain, d'Est en Ouest, et se trouve exposée à des conditions climatiques peu clémentes.La seconde, *Populus Trichocarpa*, est principalement localisée en Californie, et bénéficie d'un climat continental.

Chacune des simulations est constituée de 50 haplotypes de la souche continentale, de 50 haplotyêpes de la souche boréale, ainsi que de 50 haplotypes d'individus hybrides générés à partir des haplotypes ancestraux. Ces haplotypes ancestraux ont été estimés à l'aide du logiciel Beagle. A partir des positions en paires de base, une carte de recombinaison génétique est générée en utilisant le taux de recombinaison moyen chez le peuplier. Le taux de recombinaison, noté τ_r, correspond au nombre moyen de paires de bases à parcourir pour qu'ait lieu un épisode de recombinaison génétique, *i.e.*, notant L la longueur du chromosome en Morgans (M), et N_{bp} le nombre de paires de bases le constituant, le taux de recombinaison génétique pour ce chromosome est donné par la relation :

$$\tau_r = \frac{L}{N_{bp}}$$

Dans ce scénario, les simulations ont été produites en utilisant un taux de recombinaison génétique moyen τ_r de 0.05 centiMorgans par million de paire de bases, correspondant à la valeur utilisée par les auteurs de l'étude avec le logiciel RASPberry (*Recombination via Ancestry Switch Probability*). A partir de la donnée de la position physique en paires de bases ainsi que du taux de recombinaison moyen, nous générons une carte de recombinaison génétique adaptée à nos simulations.

3.7.2 Génération aléatoire d'individus hybrides

Pour simuler un individu métissé, il est d'abord nécessaire de simuler l'emplacement des évènements de recombinaison. Pour ce faire, nous utilisons le modèle décrit dans (Price et al., 2009), en parcourant

$\lambda = 0.001$
$\lambda = 0.01$
$\lambda = 0.1$

3.7.3 Simulations à partir de ms et Seq-Gen

Dans le scénario d'introgression via flux de gènes, nous nous inspirons des modèles de simulation décrits dans (Martin & Jiggins, 2015). Ces modèles sont largement repris dans la littérature pour l'évaluation de statistiques telles que RND_{min} (Rosenzweig et al., 2016) et Bd_f (Pfeifer & Kapan, 2017). Chaque simulation est constituée de 100 individus. Un individu est généré en concaténant un certain nombre de séquences de nucléotides, d'une longueur fixée à 5000 paires de bases par séquence. Chacune de ces séquences est elle-même simulée suivant un modèle neutre ou alternatif. Le modèle neutre décrit un scénario démographique classique de populations divergentes. Le modèle alternatif décrit quant à lui un scénario légèrement différent, et servira à caractériser les séquences *introgressées*. Les lignes de commande *ms* permettant de générer les séquences de nucléotides pour le modèle neutre ainsi que pour le modèle alternatif sont données ci-dessous :

— Modèle neutre :

```
./ms 200 1 -I 4 50 50 50 50 -ej 1 2 1 -ej 2 3 1 -ej 3 4 1
-r 50 5000 -T
```

Modèle neutre. $12N$ générations auparavant, premier épisode de divergence donnant naissance à P1 et à O. $8N$ générations auparavant, deuxième épisode de divergence voyant l'apparition de P3. $4N$ générations auparavant, dernier épisode de divergence et apparition de P2.

— Modèle alternatif :

```
./ms 200 1 -I 4 50 50 50 50 -ej 1 2 1 -ej 2 3 1 -ej 3 4 1
-es 0.1 2 0.8 -ej 0.1 5 3 -r 50 5000 -T
```

Modèle alternatif. $12N$ générations auparavant, premier épisode de divergence donnant naissance à P1 et à O. $8N$ générations auparavant, deuxième épisode de divergence voyant l'apparition de P3. $4N$ générations auparavant, dernier épisode de divergence et apparition de P2. t unités de temps auparavant, épisode de flux de gènes de P3 vers la population P2.

La variable f présente en figure **??** représente le taux d'introgression, elle quantifie la proportion d'haplotypes présents dans la population P2 et qui proviennent de la population P3. Ainsi, une valeur de f égale à 1 reviendrait à simuler un épisode de divergence de la population P3, duquel découlerait la naissance de la population P2. La valeur de f utilisée ci-dessus est 0.2, signifiant que 20% des haplotypes présents dans la population 2 sont issus de la population P3.

3.7.4 Résultats de la comparaison des logiciels

```
im.df <- data.frame(methods = c("Bdf", "D", "f_d", "pcadapt", "RNDmin"),
                    input = c("genoytpes", "genoytpes", "genoytpes", "genoytpes",
                    nb_of_pop = c("4", "4", "4", ">= 3", "3"),
                    outgroup = c("Oui", "Oui", "Oui", "Oui", "Oui"))

knitr::kable(im.df,
             col.names = c("Statistique",
                           "Format",
                           "Nombre de populations",
                           "Dont Outgroup"),
             format = "latex")
```

Statistique	Format	Nombre de populations	Dont Outgroup
Bdf	genoytpes	4	Oui
D	genoytpes	4	Oui
f_d	genoytpes	4	Oui
pcadapt	genoytpes	>= 3	Oui
RNDmin	haplotypes	3	Oui

Scénario de métissage Nous comparons ici notre méthode au logiciel RFMix destiné à la détermination de coefficients de métissage local.

10 generations
100 generations
1000 generations

```
library(magrittr)
source("R/summarySE.R")

res.tp <- read.table("data/results_admixture_setting_1000_tp.txt", header = TRUE)
res.fp <- read.table("data/results_admixture_setting_1000_fp.txt", header = TRUE)

power <- 100 * res.tp$true_positives / 5
res.pow <- cbind(res.tp, power)
colnames(res.pow) <- c(colnames(res.tp), "value")
fdr <- 100 * res.fp$false_positives / (res.tp$true_positives + res.fp$false_positi
fdr[is.na(fdr)] <- mean(fdr, na.rm = TRUE)
res.fdr <- cbind(res.fp, fdr)
colnames(res.fdr) <- c(colnames(res.fp), "value")

pdf <- summarySE(res.pow, measurevar = "value", groupvars = c("method", "delta_p")
```

```
## ------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dp
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------

##
## Attachement du package : 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
fdf <- summarySE(res.fdr, measurevar = "value", groupvars = c("method", "delta_p"))

p0 <- dplyr::bind_rows(
  mutate(pdf, stat = "Power"),
  mutate(fdf, stat = "False Discovery Rate")
) %>% mutate(stat_f =  factor(stat, levels=c("Power", "False Discovery Rate"))
) %>% ggplot(aes(x = delta_p, y = value, colour = method)) +
  geom_errorbar(aes(ymin = value - se, ymax = value + se),
                width =.05, size = 1.5) +
  ylim(0, 105) + facet_grid(~stat_f, margins = "am") + ylab("") + xlab(expression(atop(
  geom_line(size = 2) + geom_point(size = 1.5) +
  scale_x_reverse() +
  theme_bw() + guides(colour = guide_legend(title = "Method")) +
  theme(axis.text = element_text(size = 20, face = "bold"),
        axis.title.x = element_text(margin = margin(t = 20, r = 0, b = 0, l = 0)),
        axis.title=element_text(size = 20, face = "bold"),
        title = element_text(size = 20, face = "bold"),
        strip.text = element_text(size=20, face = "bold"),
        legend.text = element_text(size = 20),
        legend.key.height = unit(3, "line"),
        legend.key.width = unit(3, "line"))
print(p0)
```

Scénario à flux de gènes Dans ce paragraphe, nous comparons notre statistique de test à un ensemble de statistiques implémentées dans le package R *PopGenome* : la statistique D de Patterson, RNDmin (Rosenzweig et al., 2016) et BDF (Pfeifer & Kapan, 2017).

Chapitre 4

Aspect computationnel

Dans cette partie, nous nous intéresserons brièvement à l'aspect computationnel des méthodes qui ont été présentées dans les chapitres précédents. Le développement d'outils logiciels destinés à l'exploration de données génétiques volumineuses requiert qu'une attention toute particulière soit portée à l'utilisation des ressources de calcul.

Blockwise computation of covariance matrix Random SVD Storage in binary format Memory-mapping

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info : the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

(APPENDIX) Appendix

Chapitre 5

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

Chapitre 6

The Second Appendix, for Fun

Bibliographie

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664.

Caye, K., Deist, T. M., Martins, H., Michel, O., & François, O. (2016). TESS3 : Fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources*, *16*(2), 540–548.

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252.

Frichot, E., & François, O. (2015). LEA : An r package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*(8), 925–929.

Geneva, A. J., Muirhead, C. A., Kingan, S. B., & Garrigan, D. (2015). A new method to scan genomes for introgression in a secondary contact model. *PloS One*, *10*(4), e0118621.

Giraud, C. (2014). *Introduction to high-dimensional statistics* (Vol. 138). CRC Press.

Harrison, R. G., & others. (1990). Hybrid zones : Windows on evolutionary process. *Oxford Surveys in Evolutionary Biology*, *7*, 69–128.

Joly, S., McLenachan, P. A., & Lockhart, P. J. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*, *174*(2), E54–E70.

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix : A discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, *93*(2), 278–288.

Martin, J. W. D., Simon H., & Jiggins, C. D. (2015). Evaluating the use of abba–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, 244–257.

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, *5*(10), e1000686.

Pfeifer, B., & Kapan, D. D. (2017). Estimates of introgression as a function of pairwise

distances. *BioRxiv*, 154377.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., ... Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, *5*(6), e1000519.

Rosenzweig, B. K., Pease, J. B., Besansky, N. J., & Hahn, M. W. (2016). Powerful methods for detecting introgressed regions from population genomic data. *Molecular Ecology*, *25*(11), 2387–2397.

Suarez-Gonzalez, et a., Adriana. (2016). Genomic and functional approaches reveal a case of adaptive introgression from populus balsamifera (balsam poplar) in p. trichocarpa (black cottonwood). *Molecular Ecology*, 2427–2442.

Thornton, T. A., & Bermejo, J. L. (2014). Local and global ancestry inference and applications to genetic association analysis for admixed populations. *Genetic Epidemiology*, *38*(S1).

Yang, J. J., Li, J., Buu, A., & Williams, L. K. (2013). Efficient inference of local ancestry. *Bioinformatics*, *29*(21), 2750–2756.