

# 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

Je tiens à remercier mes collègues Kevin Caye, Thomas Dias-Alves, Thomas Karaouzène et Florian Privé, avec qui j'ai partagé ces trois années de thèse et de qui j'ai beaucoup appris.



# Préface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.



# Table des matières

<b>Introduction</b> . . . . .	<b>1</b>
0.1 Données en grande dimension . . . . .	1
<b>Chapitre 1 : État de l’art</b> . . . . .	<b>3</b>
1.1 Analyse en Composantes Principales parcimonieuse . . . . .	3
1.2 Bootstrap ACP . . . . .	3
1.3 Contexte . . . . .	3
1.4 Tests multiples . . . . .	3
1.5 Contrôle du taux de fausse découverte . . . . .	3
<b>Chapitre 2 : Adaptation locale</b> . . . . .	<b>5</b>
2.1 Cas d’étude utilisant pcadapt . . . . .	5
<b>Chapitre 3 : Introgression adaptative</b> . . . . .	<b>7</b>
3.1 Qu’est-ce que l’introgression ? . . . . .	7
3.2 Coefficients de métissage globaux et locaux . . . . .	7
3.3 Introgression . . . . .	8
3.4 Lien entre Analyse en Composantes Principales et métissage global. .	8
3.5 Analyse en Composantes Principales locale . . . . .	8
3.6 Sensibilité à l’imputation des données manquantes . . . . .	9
3.7 Simulations . . . . .	9
3.7.1 Données de peupliers . . . . .	9
3.7.2 Génération aléatoire d’individus hybrides . . . . .	9
3.7.3 Méthodes de détection . . . . .	12
3.7.4 Résultats de la comparaison des logiciels . . . . .	12
3.7.5 Simulations à partir de ms et Seq-Gen . . . . .	12
<b>Conclusion</b> . . . . .	<b>15</b>
<b>Annexe A : The First Appendix</b> . . . . .	<b>17</b>
<b>Annexe B : The Second Appendix, for Fun</b> . . . . .	<b>19</b>
<b>Bibliographie</b> . . . . .	<b>21</b>





## Liste des tableaux



# Table des figures

3.1	Coefficients de métissage et ACP . . . . .	8
3.2	$\lambda = 0.001$ . . . . .	10
3.3	$\lambda = 0.01$ . . . . .	10
3.4	$\lambda = 0.1$ . . . . .	11
3.5	Alternate model . . . . .	13



# Abstract

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



# Introduction

## 0.1 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,  $TP$  le nombre de vrais positifs, on définit le taux de fausse découverte  $FDR$  par :

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

- 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 confrontée.

### 2.1 Cas d'étude utilisant pcadapt



# 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 héréditaires qui les différencient (Harrison & others, 1990).

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

### 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) :

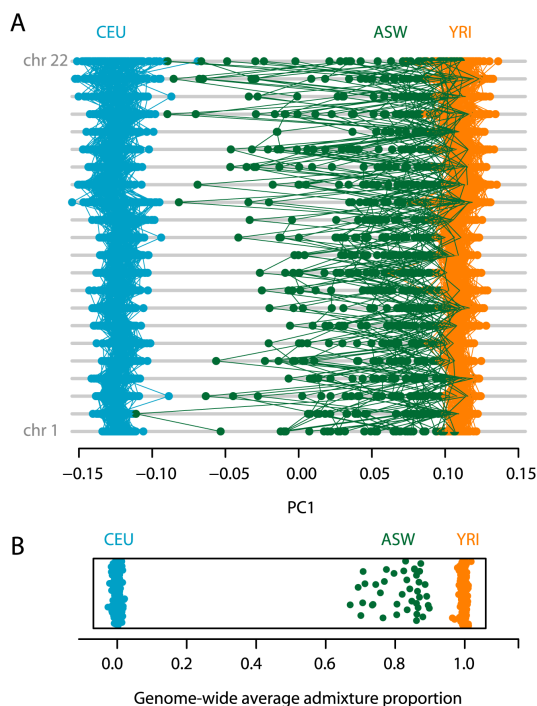


FIGURE 3.1 – Coefficients de métissage et ACP

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

## 3.7 Simulations

### 3.7.1 Données de peupliers

Le premier jeu de données est issu d'une étude d'introggression 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.

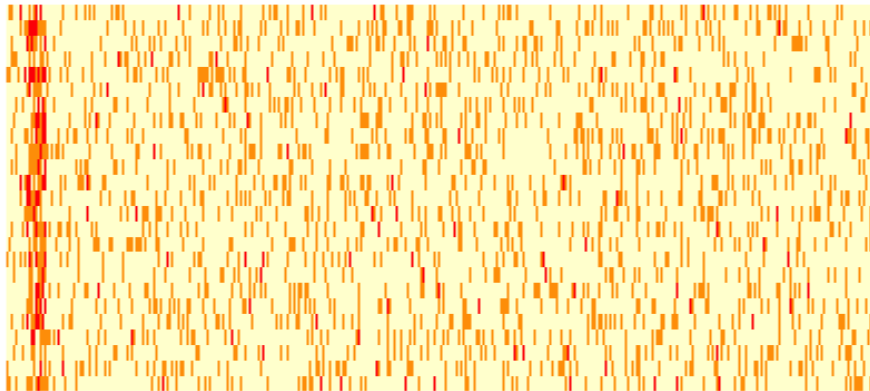
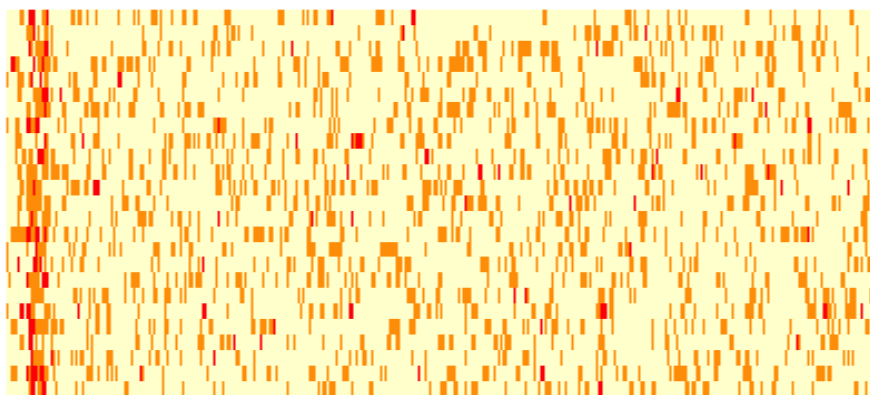
Chacune des simulations est constituée de 50 haplotypes de la souche continentale, de 50 haplotypes de la souche boréale, ainsi que de 50 haplotypes d'individus hybrides générés à partir des haplotypes ancestraux. Ces haplotypes ancestraux ont été estimés à l'aide du logiciel Beagle. À partir des positions 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  $\tau_r$ , 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 :

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

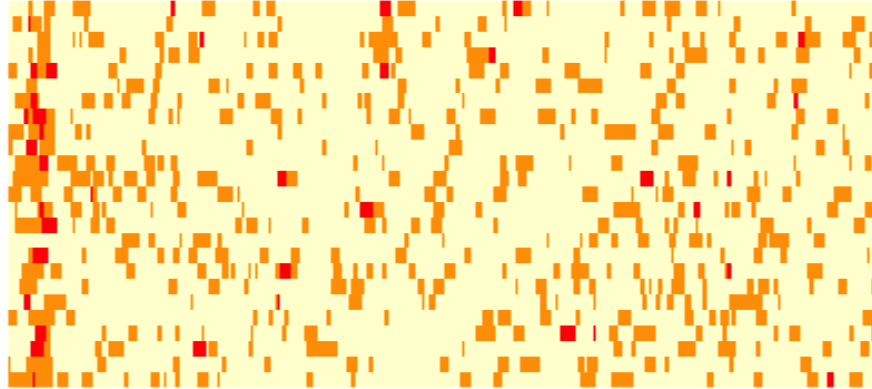
Dans ce scénario, les simulations ont été produites en utilisant un taux de recombinaison génétique moyen  $\tau_r$  de 0.05 centiMorgans par million de paire de bases, correspondant à la valeur utilisée par les auteurs de l'étude avec le logiciel RASPBerry (*Recombination via Ancestry Switch Probability*). À 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 événements de recombinaison. Pour ce faire, nous utilisons le modèle décrit dans (Price et al., 2009), en parcourant

FIGURE 3.2 –  $\lambda = 0.001$ FIGURE 3.3 –  $\lambda = 0.01$



FIGURE 3.4 –  $\lambda = 0.1$ 

```

path <- "~/Documents/thesis/git/simulations/introgression/"
output.name <- "populus"
recombinationRate <- 0.05 # in Morgans per Megabase
nSNP <- 50000
ancstrl.1 <- 1
ancstrl.2 <- 3
hyb <- 4
intro.size <- 500
global.ancestry <- 0.1
inverted.ancestry <- 0.5

info.map <- as.matrix(fread(paste0(path, output.name, ".map"),
                           data.table = FALSE))
H1 <- as.matrix(fread(paste0(path, output.name, "_H1"),
                     data.table = FALSE))
H2 <- as.matrix(fread(paste0(path, output.name, "_H2"),
                     data.table = FALSE))
n.hyb <- ncol(H1) / 2

### Introgression region
idx <- sample(1:nSNP, size = 1)
beg.reg <- max(1, idx - intro.size)
end.reg <- min(nSNP, idx + intro.size)
intro.reg <- beg.reg:end.reg

```

### 3.7.3 Méthodes de détection

Etat de l'art

Scénario à flux de gènes

La statistique  $D$  de Patterson

RNDmin

BDF

Scénario de métissage

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

### 3.7.4 Résultats de la comparaison des logiciels

Dans ce paragraphe, nous allons comparer notre statistique de test à un ensemble de statistiques implémentées dans le package R *PopGenome* : la statistique  $D$  de Patterson, RNDmin (Rosenzweig, Pease, Besansky, & Hahn, 2016) et BDF (Pfeifer & Kapan, 2017).

### 3.7.5 Simulations à partir de `ms` et `Seq-Gen`

FIGURE AVEC LES DIFFERENTS ARBRES P1, P2, P3 ET O

Décortiquons la ligne de commande `ms` permettant de générer les séquences de nucléotides.

```
./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 #background model
```

- 200 : nombre d'individus.
- 4 : nombre de populations dans lesquelles se répartissent les individus.
- 50 50 50 50 : répartition des individus dans chacune des populations.
- -ej 1 2 1 : 1 unité de temps auparavant, la population 2 fusionne avec la population 1.
- -ej 2 3 1 : 2 unités de temps auparavant, la population 3 fusionne avec la population 1 (qui est désormais composée des populations 1 et 2).

- -ej 3 4 1 : 3 unités de temps auparavant, la population 4 fusionne avec la population 1 (constituée des populations 1, 2 et 3).
- -es 0.1 2 0.8 : 0.1 unité de temps auparavant, la population 2 perd une partie de ses individus en faveur d'une population autre que celles déjà présentes (population 5). Pour chaque individu initialement présent dans la population 2, sa probabilité d'y rester est 0.80.
- -ej 0.1 5 3 : 0.1 unité de temps auparavant, la population 5 fusionne avec la population 3.

Pour utiliser *ms*, il faut définir les évènements depuis le temps présent. Pour comprendre le scénario démographique, il est plus aisé de reprendre les évènements dans l'ordre chronologique.

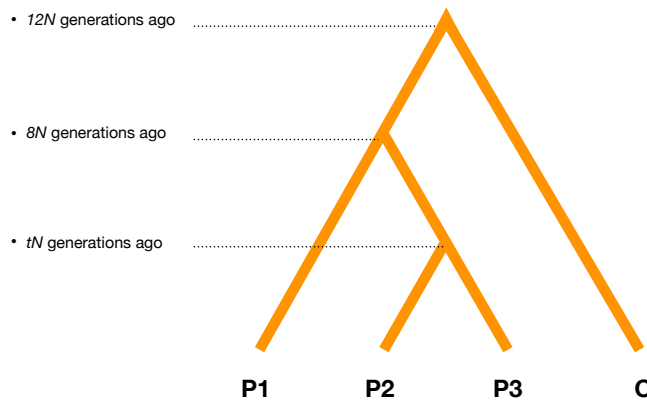


FIGURE 3.5 – Alternate model

Nous observons sur la figure 3.5,



# 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.



# Annexe A

## 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 readability and/or setup.

In the main Rmd file

```
# This chunk ensures that the thesisdown package is  
# installed and loaded. This thesisdown package includes  
# the template files for the thesis.  
if(!require(devtools))  
  install.packages("devtools", repos = "http://cran.rstudio.com")  
if(!require(thesisdown))  
  devtools::install_github("ismayc/thesisdown")  
library(thesisdown)
```

In Chapter ?? :

```
# This chunk ensures that the thesisdown package is  
# installed and loaded. This thesisdown package includes  
# the template files for the thesis and also two functions  
# used for labeling and referencing  
#opts_chunk$set(cache=TRUE)  
  
if(!require(devtools))  
  install.packages("devtools", repos = "http://cran.rstudio.com")  
if(!require(dplyr))  
  install.packages("dplyr", repos = "http://cran.rstudio.com")  
if(!require(ggplot2))  
  install.packages("ggplot2", repos = "http://cran.rstudio.com")  
if(!require(ggplot2))  
  install.packages("bookdown", repos = "http://cran.rstudio.com")  
if(!require(thesisdown)){  
  library(devtools)
```

```
  devtools::install_github("ismayc/thesisdown")
}  
library(thesisdown)  
library(data.table)  
library(pcadapt)  
library(EILA)  
library(simulate)
```



## **Annexe B**

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

*demiology*, 38(S1).

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