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

Liste des tableaux

Table des figures

Abstract

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

Chapitre 1

thesis_pdf_updt : default

Chapitre 2

État de l'art

Chapitre 3

Adaptation locale

Chapitre 4

Introgression adaptative

4.1 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.

4.2 Coefficients de métissage globaux et locaux

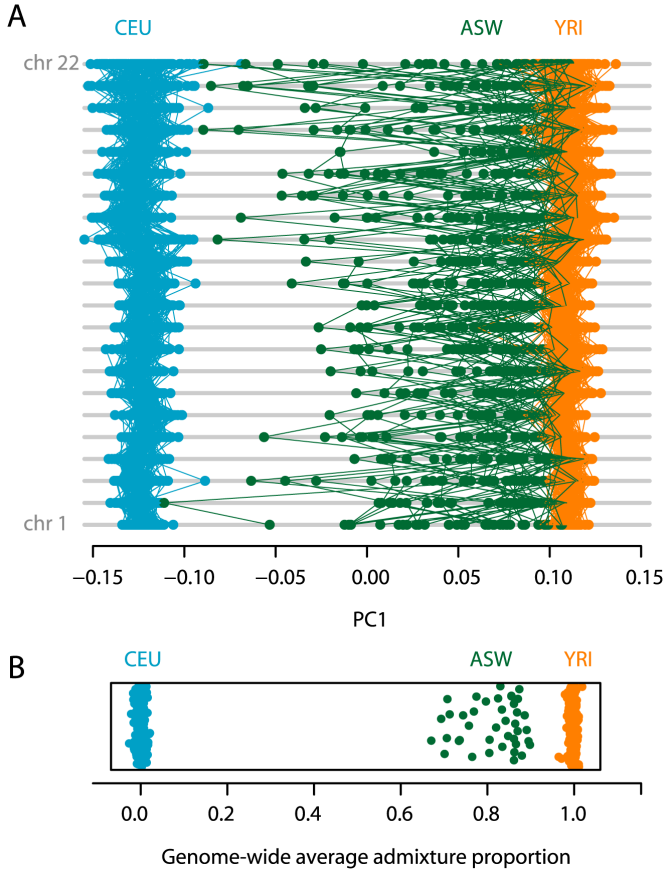
Étant données des populations ancestrales, il est possible d'estimer pour un individu donné, la proportion de son génôme provenant de chacune des populations ancestrales. Ces proportions sont connues plus communément sous le nom de *coefficients de métissage globaux*. De nombreux logiciels existent pour l'estimation de ces coefficients : STRUCTURE, ADMIXTURE (Alexander, 2009), LEA (Frichot, 2015), tess3r (Caye, 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, 2009), EILA (Yang, 2013), LAMP (Thornton, 2014), loter ou encore RFmix (Maples, 2013).

4.3 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.

4.4 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) :



Pour chacun des 22 chromosomes,

4.5 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 :

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

4.6 Sensibilité à l'imputation des données manquantes

4.7 Simulations

4.7.1 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.

4.7.2 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, 2009), en parcourant

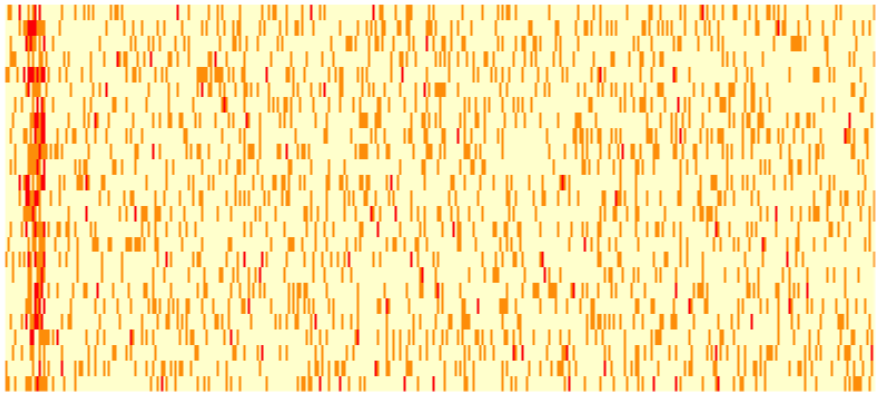


FIGURE 4.1 – $\lambda = 0.001$

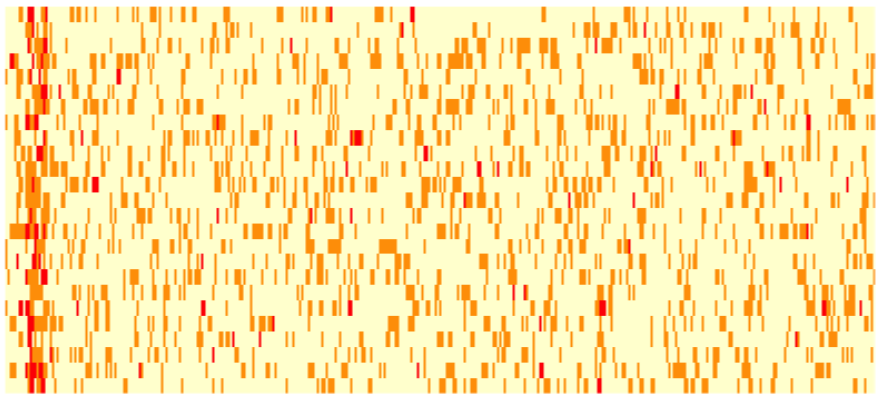


FIGURE 4.2 – $\lambda = 0.01$

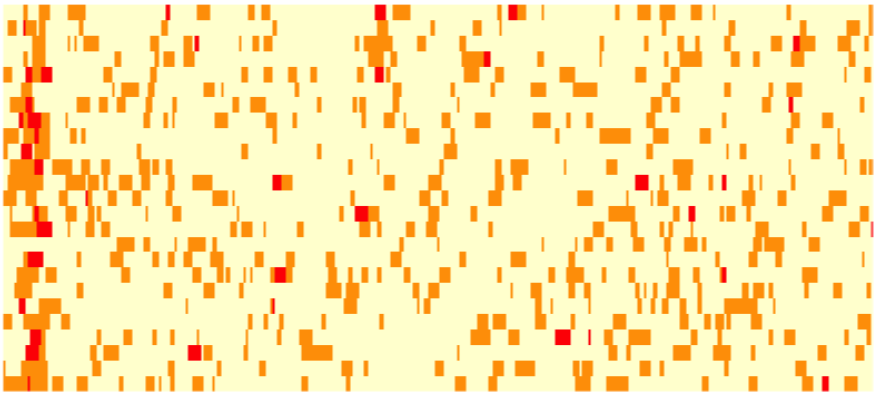


FIGURE 4.3 – $\lambda = 0.1$

```
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(data.table::fread(paste0(path, output.name, ".map"),
                                        data.table = FALSE))
H1 <- as.matrix(data.table::fread(paste0(path, output.name, "_H1"),
                                  data.table = FALSE))
H2 <- as.matrix(data.table::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
```

4.7.3 Résultats de la comparaison des logiciels

```
setwd("~/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
N <- 10
pop <- c(rep(ancstrl.1, ncol(H1) / 2),
         rep(ancstrl.2, ncol(H2) / 2),
         rep(4, n.hyb))
results <- data.frame(Software = c("pcadapt", "eila", "RFMix"), Power = c(0, 0, 0), FDR
info.map <- as.matrix(data.table::fread(paste0(path, output.name, ".map"),
                                        data.table = FALSE))

compute.fdr = function(list, ground.truth){
  if (length(list) == 0){
    x <- 0
  } else {
    x <- sum(!(list %in% ground.truth)) / length(list)
  }
  return(x)
}

compute.power = function(list, ground.truth){
  if (length(ground.truth) == 0){
    warning("The list of true positives is empty.")
  } else {
    x <- sum(list %in% ground.truth) / length(ground.truth)
  }
  return(x)
}

for (n.simu in 21:30){
  dir.name <- paste0("RFMix_v1.5.4/simu", n.simu, "/")

  input.pcadapt <- as.matrix(data.table::fread(paste0(dir.name, "simu.pcadapt"), data.ta
  input.eila <- simulate::eila_from_pcadapt(input.pcadapt, pop, anc1 = ancstrl.1, anc2 =
  param <- read.table(paste0(dir.name, "/parameters.txt"))
```

```
gt <- (param$begin):(param$end)

### run pcadapt
wsize <- 1000
mmaf <- 0.01
nomap <- 1:nSNP
maf <- pcadapt::cmpt_minor_af(input.pcadapt, 2)
proxy.map <- info.map[1:nSNP]
filtered.map <- nomap[maf >= mmaf]
stat.pcadapt <- pcadapt::scan.intro(input.pcadapt, K = 1, pop = pop,
                            ancstrl.1 = ancstrl.2,
                            ancstrl.2 = ancstrl.1,
                            admxd = hyb,
                            min.maf = mmaf,
                            window.size = wsize,
                            ploidy = 2,
                            side = "middle",
                            map = nomap)

### run eila
obj.eila <- EILA::eila(admixed = input.eila$admixed, anc1 = input.eila$anc1,
                       anc2 = input.eila$anc2, position = info.map[, 1], lambda
loc.anc.eila <- simulate::haplo_to_ancestry(obj.eila$local.ancestry, 1)

### run rfmix
allele <- paste0("./simu", n.simu, "/rfmix_alleles.txt")
classes <- paste0("./simu", n.simu, "/rfmix_classes.txt")
markerLocation <-  paste0("./simu", n.simu, "/rfmix_markerLocation.txt")
output <- paste0("simu", n.simu, "/output_simu", n.simu)
window.rfmix <- 0.00002
command <- paste("python2.7 RunRFMix.py PopPhased", allele,  classes,  markerLoc
setwd("~/Documents/thesis/git/simulations/introgression/RFMix_v1.5.4/")
system(command = command)
setwd("~/Documents/thesis/git/simulations/introgression/")
aux.rfmix <- simulate::rfmix.local.ancestry(paste0("RFMix_v1.5.4/simu", n.simu,
loc.anc.rfmix <- simulate::haplo_to_ancestry(aux.rfmix, 1)

### FDR
interp <- approx(filtered.map, stat.pcadapt[[1]], 1:nSNP)
sd.pcadapt <- sd(interp$y, na.rm = TRUE)
list.pcadapt <- which(interp$y > 3)
results[1, 3] <- results[1, 3] + compute.fdr(list.pcadapt, gt) / N
results[1, 2] <- results[1, 2] + compute.power(list.pcadapt, gt) / N
```

```
  sd.eila <- sd(loc.anc.eila, na.rm = TRUE)
  stat.eila <- (loc.anc.eila - mean(loc.anc.eila)) / sd.eila
  list.eila <- which(stat.eila > 3)
  results[2, 3] <- results[2, 3] + compute.fdr(list.eila, gt) / N
  results[2, 2] <- results[2, 2] + compute.power(list.eila, gt) / N

  sd.rfmix <- sd(loc.anc.rfmix, na.rm = TRUE)
  stat.rfmix <- (loc.anc.rfmix - mean(loc.anc.rfmix)) / sd.rfmix
  list.rfmix <- which(stat.rfmix > 3)
  results[3, 3] <- results[3, 3] + compute.fdr(list.rfmix, gt) / N
  results[3, 2] <- results[3, 2] + compute.power(list.rfmix, gt) / N
}


ggres <- data.frame(Software = rep(c("pcadapt", "eila", "RFMix"), 2), Stat = rep(0, 6),
                    Percent = rep(0, 6))
ggres$Stat[1:3] <- results$Power * 100
ggres$Stat[4:6] <- results$FDR * 100
ggres$Percent <- as.numeric(format(ggres$Stat, digits = 2))
p0 <- ggplot(ggres, aes(x = Software, y = Stat, fill = as.factor(Type)))
p0 <- p0 + ggtitle(expression(lambda == 1)) + ylab("")
p0 <- p0 + geom_bar(stat = "identity", position = position_dodge(width = 0.9))
p0 <- p0 + guides(fill = guide_legend(title = ""))
p0 <- p0 + geom_text(aes(label = Percent), position = position_dodge(width = 0.9),
                     color = "white", vjust = 1.4, size = 5)
p0 <- p0 + theme_bw() + theme(axis.text = element_text(size = 15),
                              axis.title = element_text(size = 15, face = "bold"),
                              title = element_text(size = 15, face = "bold"),
                              legend.text = element_text(size = 15),
                              legend.key.height = unit(1, "line"),
                              legend.key.width = unit(3, "line")
)
print(p0)
```

TABLE 4.1 – Correlation of Inheritance Factors for Parents and Child

Factors	Correlation between Parents & Child	Inherited
Education	-0.49	Yes
Socio-Economic Status	0.28	Slight
Income	0.08	No
Family Size	0.18	Slight
Occupational Prestige	0.21	Slight

We can also create a link to the table by doing the following : Table 4.1. If you go back to [Loading and exploring data] and look at the `kable` table, we can create a reference to this max delays table too : Table **??**. The addition of the `(\#tab:inher)` option to the end of the table caption allows us to then make a reference to Table `\@ref(tab:label)`. Note that this reference could appear anywhere throughout the document after the table has appeared.

4.8 Figures

If your thesis has a lot of figures, *R Markdown* might behave better for you than that other word processor. One perk is that it will automatically number the figures accordingly in each chapter. You'll also be able to create a label for each figure, add a caption, and then reference the figure in a way similar to what we saw with tables earlier. If you label your figures, you can move the figures around and *R Markdown* will automatically adjust the numbering for you. No need for you to remember ! So that you don't have to get too far into LaTeX to do this, a couple **R** functions have been created for you to assist. You'll see their use below.

In the **R** chunk below, we will load in a picture stored as `reed.jpg` in our main directory. We then give it the caption of "Reed logo", the label of "reedlogo", and specify that this is a figure. Make note of the different **R** chunk options that are given in the R Markdown file (not shown in the knitted document).

```
include_graphics(path = "figure/reed.jpg")
```

FIGURE 4.4 – Reed logo

Here is a reference to the Reed logo : Figure 4.4. Note the use of the `fig:` code here. By naming the **R** chunk that contains the figure, we can then reference that figure later as done in the first sentence here. We can also specify the caption for the figure via the R chunk option `fig.cap`.

Below we will investigate how to save the output of an **R** plot and label it in a way similar to that done above. Recall the `flights` dataset from Chapter **??**. (Note that we've shown a different way to reference a section or chapter here.) We will next explore a bar graph with the mean flight departure delays by airline from Portland for 2014. Note also the use of the `scale` parameter which is discussed on the next page.

```
flights %>% group_by(carrier) %>%
  summarize(mean_dep_delay = mean(dep_delay)) %>%
  ggplot(aes(x = carrier, y = mean_dep_delay)) +
  geom_bar(position = "identity", stat = "identity", fill = "red")
```

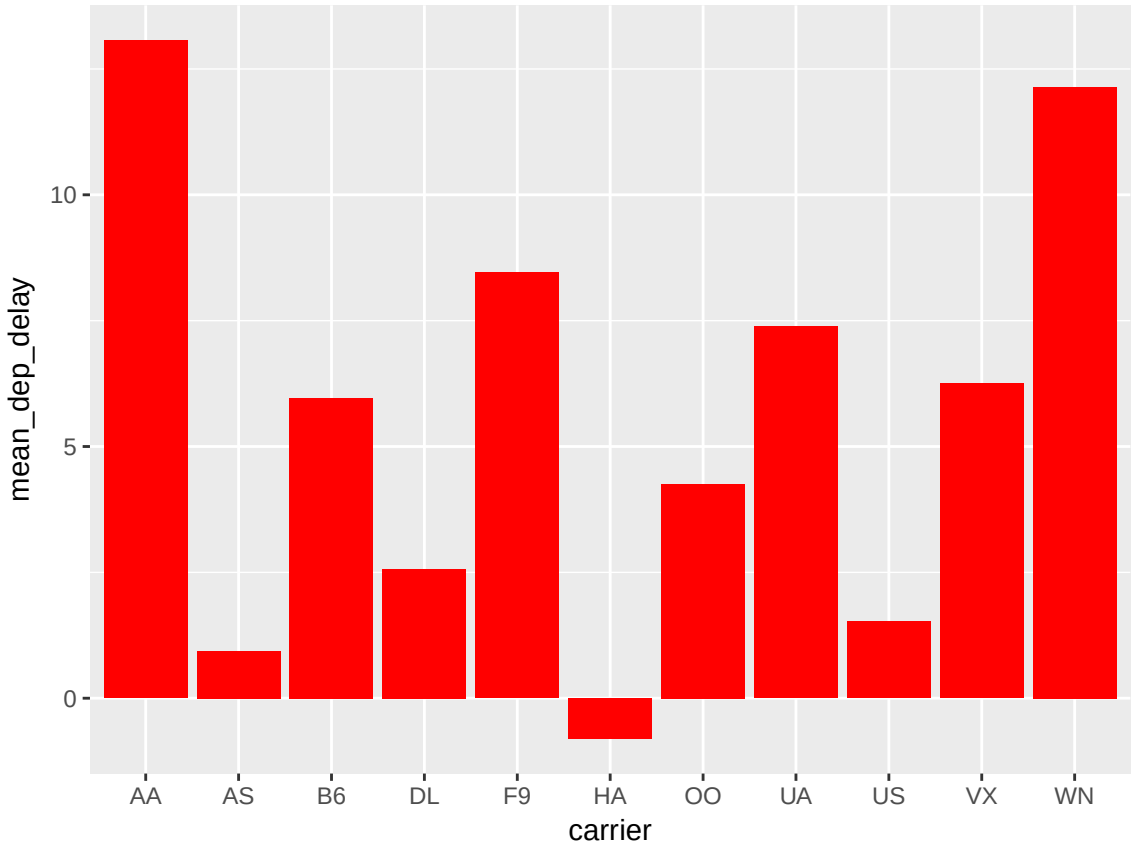


FIGURE 4.5 – Mean Delays by Airline

Here is a reference to this image : Figure 4.5.

A table linking these carrier codes to airline names is available at `https://github.com/ismayc/pnwflights14/blob/master/data/airlines.csv`.

Next, we will explore the use of the `out.extra` chunk option, which can be used to shrink or expand an image loaded from a file by specifying `"scale= "`. Here we use the mathematical graph stored in the "subdivision.pdf" file.

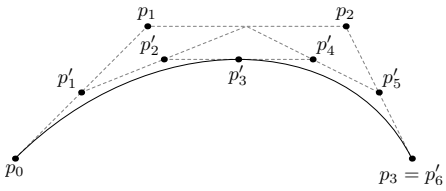


FIGURE 4.6 – Subdiv. graph

Here is a reference to this image : Figure 4.6. Note that `echo=FALSE` is specified so that the **R** code is hidden in the document.

More Figure Stuff

Lastly, we will explore how to rotate and enlarge figures using the `out.extra` chunk option. (Currently this only works in the PDF version of the book.)

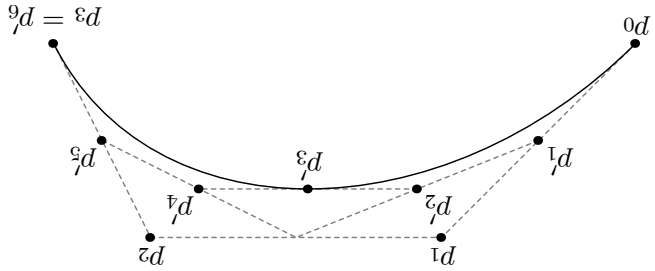


FIGURE 4.7 – A Larger Figure, Flipped Upside Down

As another example, here is a reference : Figure 4.7.

4.9 Footnotes and Endnotes

You might want to footnote something. [1] The footnote will be in a smaller font and placed appropriately. Endnotes work in much the same way. More information can be found about both on the CUS site or feel free to reach out to `data@reed.edu`.

4.10 Bibliographies

Of course you will need to cite things, and you will probably accumulate an armful of sources. There are a variety of tools available for creating a bibliography database (stored with the .bib extension). In addition to BibTeX suggested below, you may want to consider using the free and easy-to-use tool called Zotero. The Reed librarians have created Zotero documentation at `http://libguides.reed.edu/`

1. footnote text

`citation/zotero`. In addition, a tutorial is available from Middlebury College at `http://sites.middlebury.edu/zoteromiddlebury/`.

R Markdown uses *pandoc* (`http://pandoc.org/`) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard LaTeX requires. To cite references in your thesis (after creating your bibliography database), place the reference name inside square brackets and precede it by the "at" symbol. For example, here's a reference to a book about worrying : (Molina & Borkovec, 1994). This `Molina1994` entry appears in a file called `thesis.bib` in the `bib` folder. This bibliography database file was created by a program called BibTeX. You can call this file something else if you like (look at the YAML header in the main .Rmd file) and, by default, is to placed in the `bib` folder.

For more information about BibTeX and bibliographies, see our CUS site (`http://web.reed.edu/cis/help/latex/index.html`) [2]. There are three pages on this topic : *bibtex* (which talks about using BibTeX, at `http://web.reed.edu/cis/help/latex/bibtex.html`), *bibtexstyles* (about how to find and use the bibliography style that best suits your needs, at `http://web.reed.edu/cis/help/latex/bibtexstyles.html`) and *bibman* (which covers how to make and maintain a bibliography by hand, without BibTeX, at `http://web.reed.edu/cis/help/latex/bibman.html`). The last page will not be useful unless you have only a few sources.

If you look at the YAML header at the top of the main .Rmd file you can see that we can specify the style of the bibliography by referencing the appropriate csl file. You can download a variety of different style files at `https://www.zotero.org/styles`. Make sure to download the file into the csl folder.

Tips for Bibliographies

— Like with thesis formatting, the sooner you start compiling your bibliography for something as large as thesis, the better. Typing in source after source is mind-numbing enough ; do you really want to do it for hours on end in late April ? Think of it as procrastination.

— The cite key (a citation's label) needs to be unique from the other entries.

— When you have more than one author or editor, you need to separate each author's name by the word "and" e.g. `Author = {Noble, Sam and Youngberg, Jessica},`.

— Bibliographies made using BibTeX (whether manually or using a manager) accept LaTeX markup, so you can italicize and add symbols as necessary.

— To force capitalization in an article title or where all lowercase is generally used, bracket the capital letter in curly braces.

— You can add a Reed Thesis citation [3] option. The best way to do this is to use the phdthesis type of citation, and use the optional "type" field to enter "Reed thesis" or "Undergraduate thesis."

2. Reed College (2007)
3. Noble (2002)

4.11 Anything else ?

If you'd like to see examples of other things in this template, please contact the Data @ Reed team (email `data@reed.edu`) with your suggestions. We love to see people using *R Markdown* for their theses, and are happy to help.

Conclusion

Chapitre 5

The First Appendix

References

Alexander, D. (2009). *Fast model-based estimation of ancestry in unrelated individuals.*

Caye, K. (2016). *TESS3 : Fast inference of spatial population structure and genome scans for selection.*

Frichot, É. (2015). *LEA : An r package for landscape and ecological association studies.*

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

Maples, B. K. (2013). *RFMix : A discriminative modeling approach for rapid and robust local-ancestry inference.*

McVean, G. (2009). A genealogical interpretation of principal components analysis.

Molina, S. T., & Borkovec, T. D. (1994). The Penn State worry questionnaire : Psychometric properties and associated characteristics. In G. C. L. Davey & F. Tallis (Eds.), *Worrying : Perspectives on theory, assessment and treatment* (pp. 265–283). New York : Wiley.

Noble, S. G. (2002). *Turning images into simple line-art* (Undergraduate thesis). Reed College.

Price, A. L. (2009). *Sensitive detection of chromosomal segments of distinct ancestry in admixed populations.*

Reed College. (2007, march). LaTeX your document. Retrieved from `http://web.reed.edu/cis/help/LaTeX/index.html`

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. (2014). *Local and global ancestry inference, and applications to genetic association analysis for admixed populations.*

Yang, J. J. (2013). *Efficient inference of local ancestry.*