Preface

> For many years, population genetics was an immensely rich and powerful theory with virtually no suitable facts on which to operate [...] Quite suddenly the situation has changed [...] and facts in profusion have been poured into the hopper of this theory machine. [...] The entire relationship between the theory and the facts needs to be reconsidered.
>
> —Lewontin [159]

Sometimes, some researchers may say that they started to work on a topic or issue incidentally before they spent a lot of their time and energy on it. This could be said about my involvement in population genomics. My early training in population genetics was very light, and my background in genomics (a very new field when I was student) was even lighter. My long-time interest in evolution and my investment in R started in late 1990s have eventually led me to focus more and more of my scientific interest onto population genomics. Backing on my experience with ape, I started the development of pegas which was first released in May 2009. Developing an R package for evolutionary analysis was not a new thing at this time and several colleagues had initiated similar projects. The idea of this book emerged in part from discussions with some of these colleagues. At this time high-throughput sequencing (HTS) was just beginning its breakthrough and we were just starting to foresee the eventual impacts of this technological revolution. The hackathon held at the National Evolutionary Synthesis Center (NESCent) in Durham, USA, in March 2015 was for me a great opportunity to develop new tools to handle HTS data in pegas. The idea of *Population Genomics With R* then grew progressively to become a book project after discussions with John Kimmel started in September 2017.

There are three main driving ideas behind this book. The first one is to consider all types of population genetic and genomic data, from the simplest genetic data to the most large-scale genomic 'big data'. The second idea is to provide a single, common computing environment to address a wide range of questions or tackle a wide range of analyses with population genetic and genomic data. The third idea is to promote the use of free and open source software. In the progress of writing, I found out that statisticians and developers in genomics use R more frequently than I thought. After having defended the use of R for more than two decades, this is a fact that certainly provides me some satisfaction.

The basic materials of *Population Genomics With R* are the R packages listed in Chaper 1. Clearly, this is not an exhaustive list of computing resources for population genomics. I have tried as much as possible to consider packages which are operational and integrate in the general framework of population genomics outlined above. Theferore, I avoided to mention packages that are clearly not maintained (e.g., orphaned packages on CRAN) or appeared to not work correctly. During my research, I have certainly missed some packages that should have been included in this book. As an example, DECIPHER, a package distributed on BioConductor for managing very large databases of sequence data, should have been cited in several chapters of this book. On the other hand, I did not consider packages which are not distributed on a server or which are too specialized: these include several R packages developed to analyze human populations which are available only on request to their authors—although it is not clear how this way to distribute software conflicts with the free and open source software framework which I tried to follow.

Writing this book was a very progressive process and I benefited from critical and very helpful comments on early drafts from Olivier François, Santhosh Girirajan, Sarah Hendricks, and several anonymous reviewers. Hilmar Lapp invited me to the hackathon held at NESCent in 2015 where I had one of the most stimulating week of work and development: thanks to him and to the colleagues and friends who were there too. Thibaut Jombart shared a lot of great discussions on many occasions: thanks to him for organizing and inviting me to another hackathon in London. I had the opportunity to give several workshops on the packages I develop: these events are very rewarding experiences and I want to thank particularly Frédéric Chiroleu, Soledad De Esteban-Trivigno, Jérôme Goudet, and Nicolas Salamin for their invitations. Many thanks to Agnès Mignot for her full support to develop my research while she was head of the Institute of Evolutionary Sciences in Montpellier. I am very grateful to John Kimmel for giving me the oppotunity to write another book on using R to analyze evolution. Robin Lloyd Starkes enthusiastically worked out all the practical aspects handling my manuscript. Finally, I am grateful to my wife Sinta and my daughter Laure for their permanent support.

Bangsaen

Emmanuel Paradis
March 2020

Symbol Description

$\mathbb{E}$	expectation.
H	heterozygosity.
h	haplotype diversity.
K	number of populations, groups, or clusters.
k	number of alleles for a locus.
$\mathcal{L}$	likelihood.
N	population size.
N_e	effective population size defined as the number of alleles transmitted to the next generation.
n	sample size (number of individuals or of alleles); number of rows in a table or in a matrix.
p	number of variables; number of columns in a table or in a matrix.
p_i	proportion of allele i in the population ($i = 1, \ldots, k$).
$\hat{p}_i$	estimate of p_i.
$\tilde{p}_i$	predicted value of p_i.
$\Pr$	probability.
μ	mean; mutation rate.
π	nucleotide diversity.
σ	standard-deviation.
σ^2	variance.
Σ	variance-covariance matrix.
Θ	the population genetic parameter defined as $\Theta = 2N_e\mu$.