Should protracted speciation be incorporated in phylogenetic tree construction methods?

Richèl J.C. Bilderbeek & Rampal S. Etienne

July 28, 2016

1 Abstract

The construction of phylogenies helps us answering evolutionary biological questions. Our current phylogenetic tools ignore the fact that speciation takes time, which has an effect unknown in phylogeny reconstruction. Here, we simulate true incipient species trees and their corresponding DNA alignments, from which we measure the reconstruction of the phylogeny using a standard birth-death model. We measure the errors that the Bayesian phylogenetic software tool BEAST2 gives when recovering simulated phylogenies, for different times-to-speciate, under a range of additional parameter settings. It has been found that branch lengths are consistently and strongly underestimated for biologically relevant parameters. This research shows that protractedness is a complexity of nature that should not be ignored and should be incorporated in our phylogenetic tools.

2 Introduction

2.1 Biology: speciation takes time

2.2 The use of speciation models

2.3 First model: constant rate birth death model

Although we know that speciation takes time, we commonly ignore this when constructing a phylogeny, by chosing a constant-rate birth-death model as a speciation model. The constant-rate birth-death model (as described in for example [12]) is among the simplest speciation models, and assumes a constant speciation rate λ and constant extinction rate μ. Additionally, it assumes speciation is instantaneous. The constant-rate birth-death model is popular for its simplicity, yet has also served as a starting point for more elaborate specation models.

2.4 Other non-protracted speciation models

Other speciation models may assume that speciation rate changes in time [17], is dependent on the amount of species present [5], or is trait dependent [8]. They extend the original birth-death model by making speciation rate dependent on time, diversity or trait value. The assumption that all these extensions share is that speciation is instantaneous.

2.5 Protracted speciation model

2.5.1 Good and incipient states

The protracted speciation model [18] allows for speciation taking time. It adds an additional species state (see also figure 2): the 'incipient' stage, which a lineages has to complete before becoming a 'good' species. One view is to say that good species have achieved reproductive isolation, where incipient species are in the process of achieving this [REF]. Alternatively, an incipient species can be described as a good-species-to-be, yet not recognized as such [REF].

2.5.2 Biological mechanism of reproductive isolation

The protracted speciation model makes no assumption about the exact biolological mechanism by which reproductive isolation is acquired [7, 6, 18].

2.5.3 Effects on topology

A feature of the protracted birth-death model, is that it may result in paraphylies when the rates between good and incipient species differ [REF]. Paraphylies may be more than an experimental artifact and are claimed to be an inherent feature of nature [REF: Funk & Omland]. If the rates between good and incipient species are equal, no paraphylies are expected [REF]. This research assumes exactly this.

2.5.4 Good and incipient rates

Where the classic birth death model assumes constant speciation and extinction rates, the protracted birth-death model allows these rates to be state dependent.

The only additional parameter is the rate at which incipient species become (recognized as) good species. When setting that parameter to infinity, incipient species become good species instantanous and the model falls back to a constant-rate (non-protracted) birth-death model [REF].

2.5.5 Relationship to constant-rate birth-death model

Although the models use similar notation for their parameters, they are mathematically substantially different [REF].

There is no relationship known between the parameters of the constant-rate birth-death model and the protracted birth-death model [REF].

2.5.6 Good and incipient rates do not need to be constant

The assumption that speciation takes time is independent of the dynamics of the speciation and extinction rates. The protracted speciation model in this research assumes these rates are constant, but these rates can easily be made to depend on time, diversity or trait.

2.6 Use of a speciation model in inferring a phylogeny

Speciation models are widely used to make inferences from genetic data. There are multiple computer programs to create phylogenies and/or parameter estimates from aligned DNA sequences. BEAST2 is a widely used tool that allows for a Bayesian approach to phylogenetics[1]. BEAST2 supplies the user with multiple speciation models, that all assume instantaneous speciation.

2.7 This study

This simulation investigates the consequence of BEAST2 using instantaneous speciation, by simulating a 'true' tree that is protracted, simulate DNA alignments following that tree, and seeing how well BEAST2 can recover the original phylogeny.

2.8 Novelty

This study is the first to measure the error made when acknowledging that speciation takes time, yet using one of the many tools that ignores this fact.

Part of the parameter space will create 'true' trees that are not protracted. This will result in a test of BEAST2 against itself, which is, to the best of our knowledge, not been done, probably due to the extensive computations that are needed.

2.9 Analysis

The 'true' and inferred trees are compared with the nLTT statistic, which is a novel and efficient summary statistic, that has been proven to have better performance than the gamma statistics and phylogenetic in Approximate Bayesian Computation [11] (yet on par with the likelihood). The nLTT statistics is the summed absolute difference between two Lineages-Through-Time plot lines that are normalized to let both time and number of lineages range from zero to one. Additionally, this statistic allows to pinpoint the timespan that contributed most to its value.

2.10 Expectations

It is expected that for higher protractedness (thus deviation from BEAST2 its assumptions), the error will increase. Trees are expected to be inferred better, when there is more information available to construct them, thus we expect less

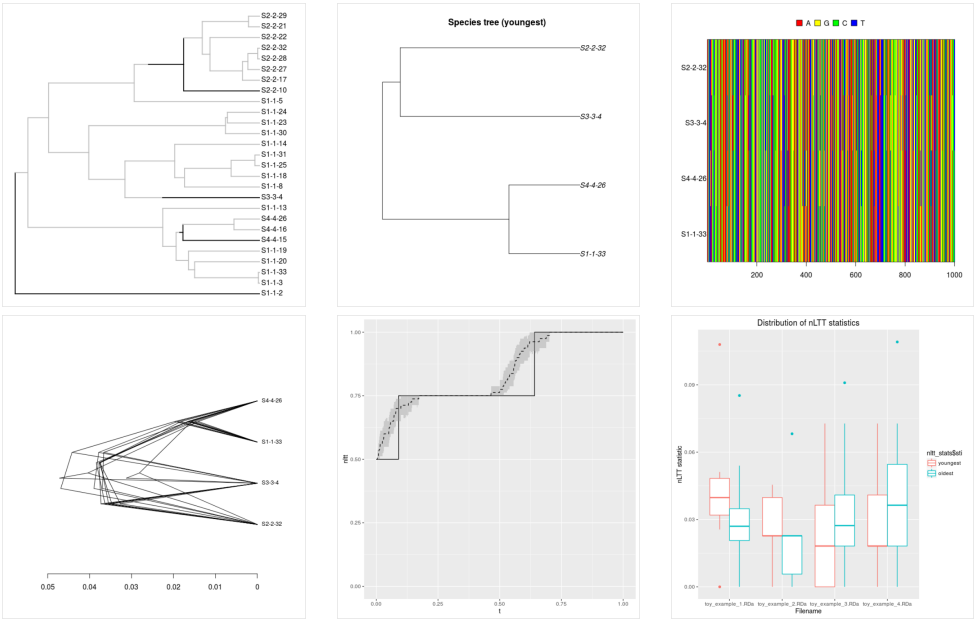

Figure 1: Workflow of the experiment: top-left: per parameter set, the creation of an incipient species tree. Top-center: constructing two species trees with oldest and youngest representative per species (only youngest shown). Top-right: constructing n_a alignments per species trees. Bottom-left: creating n_b posteriors per alignment. Bottom-center: extract the nLTT curve of the species tree (thick line) and of the n_s posteriors. Bottom right: compare the nLTT statistics between all parameters

error when there are more taxa and when the simulation DNA alignments are longer. The error is expected to be biggest close to the present, as there will be the most incipient species present.

It is unknown, however, under which biological settings this error is relevant.

2.11 Preview results

This study shows that [TODO: put result here]

3 Methods

3.1 Parameter space

This study investigates five parameters in all their combinations (see table 1 for the symbols and descriptions). For each point in parameter space, n_i incipient species trees are simulated as the 'true' incipient species tree. After constructing

Symbol	Description
b_g	Speciation initiation rate of a good species (per lineage, per million year)
b_i	Speciation initiation rate of an incipient species (per lineage, per million year)
λ	Speciation completion rate (per lineage, per million year)
μ_g	Extinction rate of a good species (per lineage, per million year)
μ_i	Extinction rate of an incipient species (per lineage, per million year)
t_c	Crown age (million years)
r	Mutation rate (?base pairs changed per duplication)
l_a	DNA alignment length (base pairs)
n_a	Number of alignments per species tree
n_b	Number of BEAST2 runs per alignment
n_i	Number of incipient species trees per parameter set
n_s	Number of MCMC samples
i_s	MCMC sampling interval

Table 1: Parameter descriptions

two species trees from it (see section 3.3), n_a alignments are similated per species tree. Per alignment, n_b BEAST2 runs are performed.

3.1.1 Parameter range

The range of parameters needed to be tuned carefully. The parameter values for the speciation-initiation rates b, extinction rate μ and phylogeny crown age t_c need to be balanced, as $t_c b \gg \mu$ results in overly taxon-rich phylogenies, where $t_c \mu \gg b$ results in extinction of all lineages.

It was chosen to follow the same values as [6], as there the same tuning was needed, also allowing comparison between the results.

We also follow the simplication of [6] to assume that species can both (1) give rise to new species, and (2) go extinct, independent of species status (thus $b_i = b_g$ and$\mu_i = \mu_g$). This also ensures the absence of paraphylies [REF again].

An additional contraint followed is that a full simulation (for one point in parameter space) should be completed in 240 hours.

Table 2 shows an overview of all parameter values used.

3.1.2 Fall back to classic birth-death model

When setting speciation completion rate close to infinity, the protracted speciation model falls back to a constant-rate (non-protracted) birth-death model. This allows to measure the performance of BEAST2 with all its assumptions satisfied.

Parameters	Values
$b = b_g = b_i$	0.1, 0.5, 1.0
λ	0.1, 0.3, 1.0, 10^6
$\mu = \mu_g = \mu_i$	0.0, 0.1, 0.2, 0.4
t_c	15
r	10^{-1},10^{-2},10^{-3}
l_a	10^3,10^4,10^5
n_a	2
n_b	2
n_i	1
n_s	10^3
i_s	10^3

Table 2: Parameters used

3.2 Creating incipient species trees

A protracted speciation model assumes that species have at least two states: a species is either a good or incipient species. A good species is a species recognized as such, where an incipient species is not yet. Both good and incipient species can generate new incipient species, at the speciation-initiation rates b_g and b_i respectively. An incipient species can become a good species at the speciation-completion rate λ. Both good and incipient species can go extinct at rates μ_i and μ_g respectively (see also figure 2).

This research uses the constant-rate protracted speciation model, which assumes all these rates are constant in time.

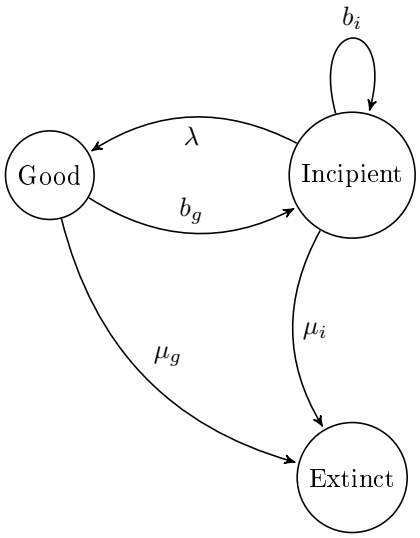

Figure 2: The states and transitions of a species. b_i: speciation-initiation rate of incipient species. b_g: speciation-initiation rate of good species. λ: speciation completion rate. μ_i: extinction rate of incipient species. μ_g: extinction rate of good species. Figure after Etienne et al, 2014, Evolution

From a parameter combination, a 'true' protracted constant-rate birth-death incipient species tree is simulated in the R programming language [16], using the PBD package [4].

3.3 Creating species trees

A species tree is a phylogeny with one lineage per species. Assuming instantaneous speciation, every lineage represents a species per definition. For protracted speciation one needs to define which subspecies is chosen to represent a species.

One such way is to randomly sample one lineage per species. For the sampling being representative, this approach would have to be repeated multiple times, which is computationally infeasable. Instead, we chose deterministically the two most extreme species trees: of each species, we pick the subspecies that has been present longest (the oldest) and shortest (the youngest).

3.4 Creating a DNA alignment

From each species tree, n_a DNA alignments were simulated following a Jukes-Cantor nucleotide subsititution model using the phangorn R package [19].

In creating DNA sequence alignments from the phylognies, a mutation rate r and DNA sequence length l_a need to be balanced to contain sufficient information. As a starting point, we chose the DNA sequence length to be 1kb, as

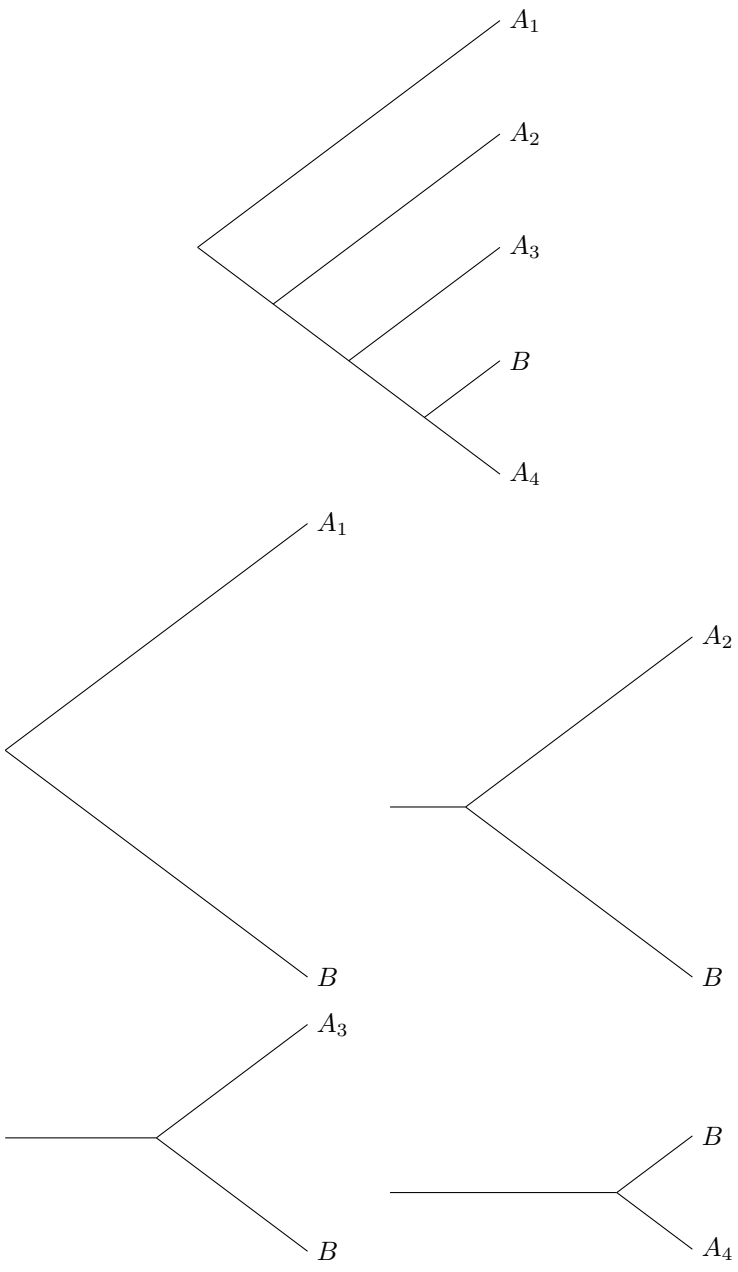

Figure 3: Example incipient species tree and four possible sampled species trees. All trees are drawn with the same scale.

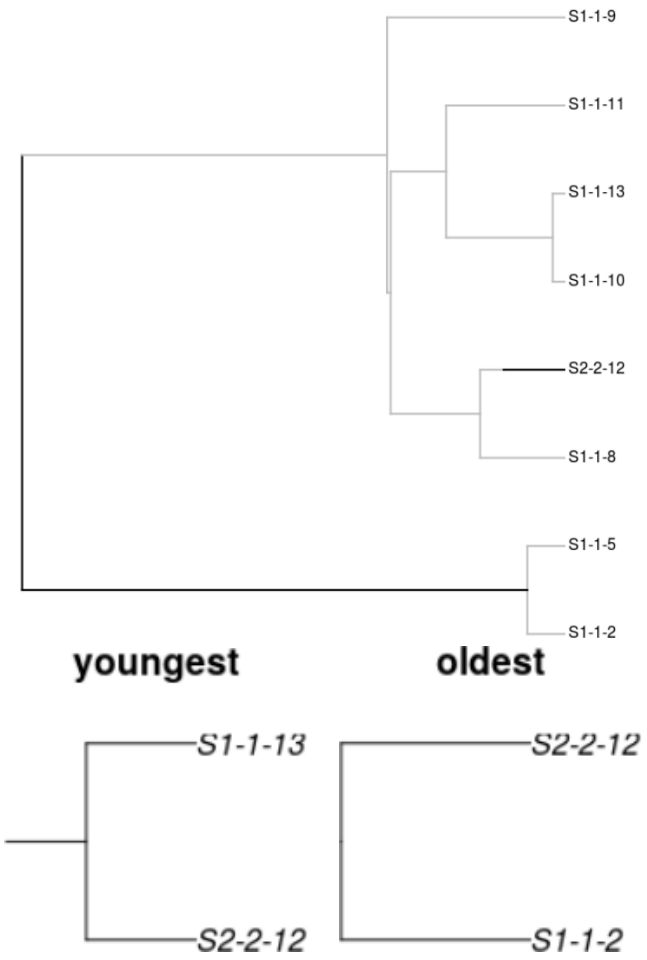

Figure 4: Example incipient species tree and its two species trees. Because the simulation algorithm, the oldest lineage has the lowest subspecies index (the third number), where the youngest lineage has the highest subspecies index. Top: incipient species trees, where black lines depict a good species and grey lines an incipient species. Species names are in the form 'Sx-x-y', where 'x-x' is the genus-species names (like *Rattus rattus*) and 'y' is the subspecies index. Bottom-left: the species tree with the youngest subspecies to represent a species. In this case S1-1-13 represents species S1-1. Bottom-right: the species tree with the oldest subspecies to represent a species. In this case S1-1-2 represents species S1-1.

this a common gene sequence length, but we also included one and two orders of magnitude bigger. The mutation rate is chosen to match a mutation rate as in nature [TODO: find that value].

3.5 Creating a BEAST2 posterior

From these DNA alignments, a posterior containing phylogenies and parameter combinations were constructed using the BEAST2 software package [1]. Per alignment, n_b BEAST2 runs were performed to see if both runs result in similar posteriors. The MCMC created by BEAST has n_s states sampled at an interval of i_s states.

The BEAST2 priors used were as follows:

For the site model, the default parameters were used: which is the Jukes-Cantor substitution model.

For the clock model, the default parameters are used, which is a strict clock prior, with a clock rate of 1.0.

For the tree prior, the 'Birth Death Model' was selected and its parameters kept at the default uniform birth-rate (range 0-10^5, initial value 1) and default uniform birth-rate (range 0-1, initial value 0.5).

3.6 Analysis

Tools used The raw data is analyzed by multiple bash (`www.gnu.org/software/bash`) scripts and the R programming language [16], using the packages rBEAST [13] to process BEAST2 output files, ape [14], phangorn [19] and ggplot2 [20] for plotting.

Measuring error For each simulation, each of its two species trees is compared with its posterior species trees. This comparison is done by calculating the nLTT statistic. Each species tree generated $n_a n_b n_s$ posterior trees resulting in a distribution of that many nLTT statistics.

3.7 Peripherals

3.7.1 Software

The full workflow of this research is put in a package called 'Cer2016' and is hosted at GitHub (`http://github.com/richelbilderbeek/Cer2016`). Using GitHub is good practice [15] and helps to improve transparency [9]. The code follows all advice given by the devtools [22], lintr [?]and goodpractice [2] packages at their default settings. The code is tested by the testthat [21] and testit [23] R packages. Code coverage, whick correlates with code quality[3], is above 90%, as measured by the covr [10] R package and visualized by Codecov, `https://codecov.io/`. Continuous integration is serviced by Travis CI, `www.travis-ci.org`.

All figures of this article can be reproduced with the start of a single command.

3.7.2 Hardware

Computions were performed on the Peregrine HPC cluster from the University of Groningen.

4 Results

5 Discussion

The PBD model is just one implementation of speciation taking time

Used the simplest version of the PBD model A constant-rate protracted birth-death model is used, this research could easily be modified to compare other protracted birth-death models, for example, a diversity-dependent protracted birth-death model. There has been now work done on the diversity-dependent protracted birth-deatg model yet.

Biological implication of assuming equal speciation-initiation and extinction rates

Allopatric speciation only The protracted speciation model matches best to the allopatric mode of speciation, in which populations are split into two equal pieces, without gene flow between parental and this incipient species.

This research assumes that speciation is allopatric, because in the simulation of the DNA alignments, there is no gene flow anymore between two taxa of still-the-same species.

BEAST2 is just one of many tools

Meager coverage of parameter space

nLTT statistic is just one of many summary statistics

References

[1] Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014.

[2] Gabor Csardi. *goodpractice: Advice on R Package Building*, 2016. R package version 1.0.0.

[3] Fabio Del Frate, Praerit Garg, Aditya P Mathur, and Alberto Pasquini. On the correlation between code coverage and software reliability. In *Software Reliability Engineering, 1995. Proceedings., Sixth International Symposium on*, pages 124–132. IEEE, 1995.

[4] Rampal S. Etienne. *PBD: Protracted Birth-Death Model of Diversification*, 2015. R package version 1.1.

[5] Rampal S Etienne, Bart Haegeman, Tanja Stadler, Tracy Aze, Paul N Pearson, Andy Purvis, and Albert B Phillimore. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20111439, 2011.

[6] Rampal S Etienne, Hélène Morlon, and Amaury Lambert. Estimating the duration of speciation from phylogenies. *Evolution*, 68(8):2430–2440, 2014.

[7] Rampal S Etienne and James Rosindell. Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Systematic Biology*, 61(2):204–213, 2012.

[8] Richard G FitzJohn, Wayne P Maddison, and Sarah P Otto. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic biology*, 58(6):595–611, 2009.

[9] Krzysztof J Gorgolewski and Russell Poldrack. A practical guide for improving transparency and reproducibility in neuroimaging research. *bioRxiv*, page 039354, 2016.

[10] Jim Hester and Willem Ligtenberg. *covr: Test Coverage for Packages*, 2016. R package version 2.1.0.9000.

[11] Thijs Janzen, Sebastian Höhna, and Rampal S Etienne. Approximate bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nltt. *Methods in Ecology and Evolution*, 6(5):566–575, 2015.

[12] Sean Nee, Robert M May, and Paul H Harvey. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 344(1309):305–311, 1994.

[13] olli0601. *rBEAST*.

[14] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

[15] Yasset Perez-Riverol, Laurent Gatto, Rui Wang, Timo Sachsenberg, Julian Uszkoreit, Felipe Leprevost, Christian Fufezan, Tobias Ternent, Stephen J Eglen, Daniel SS Katz, et al. Ten simple rules for taking advantage of git and github. *bioRxiv*, page 048744, 2016.

[16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[17] Daniel L Rabosky and Irby J Lovette. Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? *Evolution*, 62(8):1866–1875, 2008.

[18] James Rosindell, Stephen J Cornell, Stephen P Hubbell, and Rampal S Etienne. Protracted speciation revitalizes the neutral theory of biodiversity. *Ecology Letters*, 13(6):716–727, 2010.

[19] K.P. Schliep. phangorn: phylogenetic analysis in r. *Bioinformatics*, 27(4):592–593, 2011.

[20] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[21] Hadley Wickham. testthat: Get started with testing. *The R Journal*, 3(1):5–10, 2011.

[22] Hadley Wickham and Winston Chang. *devtools: Tools to Make Developing R Packages Easier*, 2016. R package version 1.12.0.9000.

[23] Yihui Xie. *testit: A Simple Package for Testing R Packages*, 2014. R package version 0.4.