

Hickory:
A Package for Analysis of
Population Genetic Data
v1.1

Kent E. Holsinger and Paul O. Lewis
Department of Ecology & Evolutionary Biology, U-3043
University of Connecticut
Storrs, CT 06269-3043

Copyright © 2003-2007 Kent E. Holsinger and Paul O. Lewis
Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled “GNU Free Documentation License”.

Contents

Preface	ii
Acknowledgments	iv
1 Installation	1
1.1 Windows	1
1.1.1 Uninstalling	2
1.2 Linux	2
1.2.1 Uninstalling	3
1.3 Compiling Hickory from source	3
2 Using Hickory	4
2.1 The Graphical User Interface	4
2.1.1 File	4
2.1.2 Data	5
2.1.3 Analyses	5
2.1.3.1 Log files	6
2.1.4 Plots	6
2.1.5 Tools	6
2.2 The Command-Line Interface	7
2.3 Getting data into Hickory	8
2.3.1 General features of NEXUS data files	8
2.3.2 The <code>alleles</code> block	10
2.3.3 The <code>hickory</code> block	12
3 Interpreting the Output	14
3.1 Interpreting the θ statistics	18
3.2 Parameters of the posterior distributions	20

3.3	Model choice	21
3.4	Estimating f from dominant markers	22
3.4.1	Estimating θ without estimating f	23
4	Producing plots	24
5	Posterior comparisons	26
6	Revision history	28
7	Implementation details	31
8	GNU Free Documentation License	32
0.	Preamble	32
1.	Applicability and definitions	33
2.	Verbatim copying	35
3.	Copying in quantity	35
4.	Modifications	36
5.	Combining documents	38
6.	Collections of documents	39
7.	Aggregation with independent works	39
8.	Translation	40
9.	Termination	40
10.	Future revisions of this license	40
	Literature Cited	42

Preface

This documentation is less preliminary than it was in earlier versions, but it is still far from complete. We expect to expand it considerably as we extend the features available in `Hickory`. Eventually we hope to include a more general discussion of F -statistics, outlining some of the different ways in which they are interpreted.

This version of the software should be reasonably easy to use and stable, but it isn't complete. If you want check convergence using formal convergence diagnostics, for example, you'll need to save log files and use a separate package (see below). A future version of `Hickory` is likely to include some of the standard convergence diagnostics, but we can't promise when that version will appear.

This version of `Hickory` can produce figures similar to those in Holsinger et al. (2002), but they aren't as nice as those that can be produced with R. We include the R code we used to generate figures for Holsinger et al. (2002).¹ If you know R (or S), the code should be reasonably self-explanatory. If you don't know R, you're free to ask for advice, but we may not have time to help you.

While we believe the numerical and simulation routines in `Hickory` are stable and accurate,² statisticians with much more experience than we have always pointed out that **MCMC is dangerous**. We urge you to try several runs with your data to ensure that the results you obtain are consistent (within the bounds of rounding and stochastic error). Should you have a data set in which separate runs give different results, the Markov chain is probably not converging. If you've changed the default sampler parameters

¹R is a freely available statistical package whose syntax and grammar is very similar to S-Plus. Binaries and source code (in C) are available at <http://www.r-project.org/>.

²This is *mostly* true. See the comments on estimates of f from dominant markers in §3.4.

(`burnin = 5000`, `sample = 100,000`, `thin = 20`), change back to the default and try several more runs. If the problem persists, please send us a copy of your data so we can investigate the source of the problem. If you're really paranoid,³ you'll want to do more formal checks on convergence of your results. For these, you'll need to have `Hickory` generate a log file and use an external analytical routine. We have found Bayesian Output Analysis (<http://www.public-health.uiowa.edu/boa/>) very useful. You'll need to know a little `R` (or `S`) to use BOA, but you don't need to know much.

In earlier versions of this program we reported Bayes factors, which are intended to be used in choosing among models. There are good theoretical reasons for using Bayes factors in choosing among models, but our investigations failed to find a way in which we could produce numerically stable Bayes factor estimates. Thus, we have removed them, and we recommend that you use the Deviance Information Criterion (DIC; Spiegelhalter et al. 2002) as a model choice criterion (see §3.3).

Note on numerical differences in v1.0.5 and later

In versions 1.0.4 and earlier we used single-component Metropolis-Hastings sampling for all parameters. In versions 1.0.5 and later we use slice sampling for all univariate parameters. We use Metropolis-Hastings sampling only for multivariate parameters, i.e., allele frequencies at multi-allelic loci.

When we converted our sampling routines we also improved our ability to sample from beta and Dirichlet distributions when the allele frequencies are extreme. So far we have seen only very minor effects on parameter estimates in most cases. Table 1 shows the results for the sample data sets distributed with `Hickory`.⁴ You'll see that the differences are very small.⁵ Not surprisingly, the one case in which we have encountered a large effect was for a dataset in which there is very little observed heterozygosity, because in this case we are now able to sample from the corresponding beta and Dirichlet distributions more accurately. In this case, the numerical error caused allele frequencies to be overly smoothed, resulting in a substantial underestimate of θ . If you notice any apparent discrepancies, please let us know so that we

³And being paranoid about MCMC analyses is not a bad thing.

⁴Results for v1.1 were produced without `Estimate theta-y` unchecked to make them directly comparable to those from v1.0.4 and earlier.

⁵`theta-II` in v1.0.5 and later corresponds to `theta-B` in v1.0.4 and earlier. See §3.1 for more information.

Dataset	f			θ^1		
	mean	(2.5%, 50%, 97.5%)		mean	(2.5%, 50%, 97.5%)	
dominant.nex	0.5334	(0.0324, 0.5442, 0.9744)		0.0503	(0.0464, 0.0374, 0.1701)	
	0.5544	(0.0448, 0.5638, 0.9828)		0.0497	(0.0018, 0.0353, 0.1708)	
mosquito.nex	0.0639	(0.0040, 0.0590, 0.1535)		0.1226	(0.0420, 0.1107, 0.2705)	
	0.0645	(0.0042, 0.0592, 0.1545)		0.1199	(0.0421, 0.1071, 0.2656)	
worknisw.nex	0.0326	(0.0019, 0.0297, 0.0811)		0.0290	(0.0105, 0.0262, 0.0649)	
	0.0325	(0.0022, 0.0302, 0.0784)		0.0288	(0.0103, 0.0265, 0.0617)	

¹theta-B in v1.0.4. theta-II in v1.1.

Table 1: Comparison of numerical results from the sample data sets for v.1.0.4 and v1.1. The first line for each dataset reports results from a full model analysis in v1.1 in which θ_y is not estimated. The second reports results from v1.0.4. `nburnin = 5000`, `nsample = 25000`, `thin = 5` for both sets of analyses.

can investigate.

Acknowledgments

The graphical user interface in `Hickory` is built using the `wxWidgets` library (<http://www.wxwidgets.org>).

Routines for MCMC sampling, including slice sampling of univariate parameters and Metropolis-Hastings sampling of multivariate parameters, uses facilities provided by `mcmc++` (<http://darwin.eeb.uconn.edu/mcmc++/mcmc++.html>).

Routines for reading NEXUS files make use of the NEXUS Class Library (NCL: <http://hydrodictyon.eeb.uconn.edu/ncl/>).

Development of `Hickory` is supported in part by a grant from the National Institutes of Health, National Institute of General Medical Sciences, 1R01-GM068449-01A1.

The following users have offered helpful suggestions, supplied data sets that helped us to identify problems, or asked questions that helped us improve the documentation for `Hickory`:

Rafael González Albaladejo

Laurie Alexander

Luke Barrett Kiran Kumar Battula

Erika Baus

Roger Butlin

Tin Chi Solomon Chak

Caroline Chong

Simon Creer

Serge J. Edmé
Joanna Freeland
Katherine Dunbar
Nathan Haig
Mark Hershkovitz
Yeng Wen-Hua
Beate Hub
Hania Lada

P.L.M. Lee
Roberta Mason-Gamer
Jaime Pinñera
Elie Poulin
Lisa Wallace
Annegret Werzner
Bayazit Yunusbayev

Chapter 1

Installation

Since v0.8 of `Hickory` we have distributed two pre-compiled versions of `Hickory`: one with a graphical user interface and one for use from the command line. Most users will probably prefer the GUI, but the command-line version is more convenient if you have a large data set that you want to run in the background or if you want to do some exploratory simulations. We use the command-line version in our own simulation work with `Hickory`. The analytical routines use exactly the same code in both versions. Only the user interface is different.

1.1 Windows

We recommend using `setup.exe` to install `Hickory` unless you encounter problems. This approach provides a standard Windows installation interface, including adding `Start Menu` shortcuts for starting the GUI version of `Hickory` and providing `uninstall` information.

By default, `Hickory` will be installed under

```
C:/Program Files/hickory-v1.1
```

but you can select any destination that you want.¹

- A complete installation of `Hickory` will include both GUI and command-line executables, in addition to complete source code (ex-

¹Notice that we use version-specific subdirectories for the installation. This allows you to keep old versions around in case you want to cross-check analyses between versions.

cept for the `MCMC++`, `NEXUS`, and `wxWidgets` libraries), documentation, and sample data files.

- A **compact** installation will include only the GUI executable, documentation, and sample data files.

If you install from the ZIP file, be sure to use the appropriate options to preserve the directory structure. After unZIPping, you will have all the files found in a complete installation with `setup.exe`.

1.1.1 Uninstalling

If you use `setup.exe` for installation, be sure to use the **Start Menu** shortcut to uninstall. If you installed from a ZIP file, all you need to do is to delete the relevant directories. In either case, be sure that you've saved any of your data elsewhere first.²

1.2 Linux

If you're running Linux, installation is trivial. Put the `tar.gz` file in a directory *above* the one where you want `Hickory`, and

```
tar -zxvf hickory-v1.1.tar.gz
```

You'll find the executables in `hickory-v1.1` under your current directory, and the source code, datafiles, and documentation in the obvious subdirectories of that. You may want to add the `Hickory` directory to your `PATH` in one of your startup files, or you can simply provide the full path when starting an analysis. Or if you want to make `Hickory` available to all users on your system, just copy the binaries to an appropriate directory (e.g., `/usr/local/bin`).

²It's probably a good idea to save your data somewhere else anyway. You can always point `Hickory` to it after it's started.

1.2.1 Uninstalling

Hickory doesn't make any changes outside of its own directory, so `rm -rf hickory-v1.1` from the appropriate directory will remove any evidence that it was ever there. If you the binaries elsewhere, just `rm -rf hickory hickory-nogui tools` in the appropriate directory.

1.3 Compiling Hickory from source

Right now, compiling Hickory from source is a bit of a chore, especially if you're having trouble with `wxWidgets`. In a future version we expect to provide a `configure` script that will automate building binaries on systems that support it.³ If you'd like to try it now, e-mail Kent for suggestions and help.⁴

³Note to Mac users: Once we have `configure` script support, it shouldn't be too difficult to build Hickory under OS X.

⁴And if you compile it under OS X, which *should* be possible, we'd be delighted to distribute OS X binaries for you and to give you credit for producing them.

Chapter 2

Using Hickory

2.1 The Graphical User Interface

The graphical user interface makes it easy to choose your data sets and analyses. There are only small differences between the Linux and Windows versions, so we don't discuss them separately. The only significant difference is in how you start **Hickory**: To start **Hickory** under Windows, double click the **Hickory** icon or select it from the **Start Menu** in the **Hickory Program Group**.¹ Under Linux start it from a command shell or a desktop shortcut like any other program.²

2.1.1 File

Once **Hickory** opens, click on **File**, and select **Open**. Navigate to the **Samples** directory. Click on it, and select **dominant.nex**. A new window (a new panel under Linux) will open with a **NEXUS** format data file for a simple data set with 5 loci and 6 populations. The format is relatively easy to understand, especially if you've used the Lewis and Zaykin **GDA** program before. Like all **NEXUS** files it starts with **#NEXUS** on the first line. The section of the file between **[!** and **]** is a comment that explains part of the data format specific to dominant markers. The files **mosquito.nex** and **workn1sw.nex** illustrate

¹You can also start it from a command shell, if you're so inclined.

²You'll have to make the desktop shortcut yourself. We don't do it automatically. But if you're running Linux, you can probably figure out how to make your own shortcut if you're so inclined.

the formats used for co-dominant marker data.³

2.1.2 Data

If you change the data file you've opened from the **File** menu, save it and open the **Hickory** output window. Click on **Data->Load from file** and select **dominant.nex** again. you'll see the comment at the head of the file flash by, then some messages about reading blocks of the NEXUS data file, a report of any monomorphic loci that will not be included in the analysis, and (if all goes well) a final message: "Data from file 'dominant.nex' read and stored."⁴ With that we're ready to proceed to analyses.

2.1.3 Analyses

Click on **Analyses**, and you'll see several options. **Hickory** will automatically recognize whether your data are from dominant markers or co-dominant markers.⁵ If you have dominant marker data, all analysis options will be available. If you have co-dominant marker data, the **Run f-free model** will be grayed out.⁶ If you select the **Run full model** option you'll get results similar to those presented in Holsinger et al. (2002).⁷ You can save the output using the **File** menu, or you can print using the standard print dialogs, which are also found under the **File** menu.

Starting with v1.1 you'll also find an option labeled **Estimate thetaY**. This option is checked by default. As described in Song et al. (2006), by

³For more details see §2.3.

⁴To **Hickory** a monomorphic locus is one in which every individual in every population is homozygous for the same allele. Future versions may allow user selectable criteria for polymorphism, e.g. > 5% frequency in the sample.

⁵Well, it's automatic in the sense that it assumes the data is co-dominant unless you tell it otherwise with the **dominant** keyword in the alleles block.

⁶We can't see any reason why anyone would choose not to estimate f with co-dominant marker data. Doing so will improve the estimate of θ^B and the estimate of f with co-dominant markers doesn't suffer from the identifiability problem that plagues dominant markers (§3.4).

⁷If you were to run the *Platanthera leucophaea* data (which would be hard, since we don't provide it), you'd see that the reported estimate for θ^B is different. In the course of trying to track down the problem with estimating f (see §3.4) we found a small numerical problem with our original routines. Surprisingly, the error has little detectable effect on the bias or root mean-squared error statistics we reported because the direction of the difference was random (Holsinger and Wallace 2004).

modeling the distribution of allele frequencies across loci we are able to estimate the correlation among populations (ρ in Fu et al. 2003). With this option checked you'll get an estimate of ρ and a couple of new estimates of θ (see Chapter 3 for details).

2.1.3.1 Log files

If you want to run convergence diagnostics on your analyses or if you want to do posterior comparisons of the parameters (see §5), you'll want to select **Sample log file**. You'll also need to produce the log files if you want to use R to produce plots rather than using the routines in Hickory. You can also read the log files into Tracer (<http://evolve.zoo.ox.ac.uk/software.html?id=tracer>) for an alternative analysis of the posterior. You'll see a check next to the option after it's been selected, and each sample of f and other parameters from the posterior will be written to a log file. The log file will be found in the same directory as your NEXUS data file, and it will be given a name based on the name of your data file with `.txt` substituted for `.nex` and `-full`, `-fzero`, `-thetazero`, or `-ffree` inserted as appropriate. If a file with that name is found in the directory, new log files are numbered sequentially from 01 to 99.

2.1.4 Plots

After the analyses are finished, you can plot by selecting the appropriate options under the **Plot** menu. Each analysis will be plotted in a separate window, and analyses you haven't done will be grayed out. See §4 for details.

2.1.5 Tools

If you have existing log files, whether from the current or previous sessions, you can select two of them using the **Compare posteriors** menu item.⁸ Hickory will report the mean and 95% credible intervals for each parameter. If the datasets have different parameters, then you'll see **NA** reported for comparisons that are "not applicable."

⁸Currently the only menu item available. This is where you'll find the convergence diagnostics in some later version of Hickory.

2.2 The Command-Line Interface

If you're old enough to have used command-line programs under DOS or if you're familiar with command-line interfaces for programs under Unix or Linux, the command line interface will be familiar:

```
hickory-nogui [-afhlntz] <infile.nex> [outfile.txt]
```

```
-a --all          do all analyses
-f --full         do full analysis
-F --ffree       f free analysis
-h --help        help (this message)
-l --log         keep log file(s)
-n --nothetay    don't estimate correlation
-t --thetazero   do theta=0 analysis
-z --fzero       do f=0 analysis
```

Thus, to do all analyses of the data in `sample/dominant.nex`, keep a log file of each, and save the results in `sample/dominant.txt`, simply enter

```
hickory-nogui --all --log sample/dominant.nex sample/dominant.txt
```

or

```
hickory-nogui -al sample/dominant.nex sample/dominant.txt
```

at the command prompt. Notice that single-letter options can be combined. Because neither of these examples specify `-n` (or `-nothetay`), the among population correlation will be estimated.

To do all the relevant analyses of the data in `sample/mosquito.nex` and have the results appear on the terminal enter

```
hickory-nogui --full --fzero --thetazero smaple/mosquito.nex
```

or

```
hickory-nogui -fzt sample/mosquito.nex
```

at the command prompt. As you would expect, command-line redirection can be used to redirect output to a file. If you specify `--all` and you have co-dominant data, only the `full`, `f = 0`, and `theta = 0` models will be analyzed, and you'll get a message reminding you that the `f free` model is not implemented for co-dominant marker data.

It is not currently possible to produce posterior plots from the command line. It is, however, possible to do posterior comparisons. Simply enter

```
tools <filename_1.txt> <filename_2.txt>
```

at the command prompt. The results will be printed to standard output, or you can use command-line redirection to send the output to a file.

2.3 Getting data into Hickory

If you've used GDA, PAUP*, MacClade, Mesquite or another program that uses NEXUS format data files. The format of the data files will look very familiar. Hickory requires only an `alleles` block; the `hickory` block is optional. If you're satisfied with the program defaults for priors and MCMC parameters, you don't need a `hickory` block. If you're not sure what the default parameters are, erase (or comment out) the `hickory` block in a data file, read it into Hickory, and select `Show sample info` under the `Analyses` menu (or run the command-line version without specifying any analyses).

2.3.1 General features of NEXUS data files

A simple example of a NEXUS file is provided below to facilitate discussion of the data file format.

```
#nexus
```

```
begin alleles; [comments are surrounded by square brackets]
  dimensions newpops npops=2 nloci=3;
  format labels missing=? separator=/;
  locusallelelabels
    1 'pgi 1',
    2 'pgi 2',
    3 adh
```

```

;
matrix
  Embudo:
    indiv_1 A/A 100/110 slow/fast
    indiv_2 A/A 75 / 90 slow/slow
    indiv_3 A/a 75/100 fast/slow
    indiv_4 A/A 100/100 fast/fast,
  Black_Mesa:
    1 a/a 110/100 fast/slow
    2 a/A 75/100 slow/slow
    3 a/a 100/100 fast/fast
;
end;

```

Most of the features of NEXUS data files can be illustrated using this simple example. NEXUS files are free-format, which means that the entire file could conceivably consist of a single, long line of text. It does not matter to *Hickory* where you break lines (as long as you don't split up a keyword or the name of a locus, allele or population), nor does it matter to *Hickory* if you use one space or a dozen spaces to separate the individual words (tokens) in the file. Tokens may be casually defined as sequences of characters separated by whitespace (e.g., spaces, carriage returns, line feeds, tabs, etc.). Exceptions to this definition are common, however. Special punctuation characters may also delineate adjacent tokens in the absence of whitespace; for example, in the genotype designation A/a (locus `pgi 1` in individual `indiv.2` of the population `Embudo`, above) there are three distinct tokens, because the forward slash (/) character serves as a single-character token. Other common single-character tokens include the equal sign (=), semicolons (;), colons(:), and commas (,). Tokens may also contain whitespace! If a token begins with a single or double quote character, then every character until the next, matching quote character will be treated as a single token. This is useful as a means of putting blank spaces inside population or locus labels. For example, if you want a population label to contain an embedded space, you simply enclose it in single quotes. This was done for the locus names 'pgi 1' and 'pgi 2' in the example above. An alternative approach is to use the underscore character (_) where you want the space to appear. The underscore character approach was used in the labels for individuals in the Embudo population above (e.g., `indiv_1`). Underscores show up as blank spaces when

Hickory reports results.

NEXUS *blocks* are made up of *commands*, which always end in a semi-colon (;). The `alleles` block in the example is composed of a `begin` command, a `dimensions` command, a `format` command, a `locusallelelabels` command, a `matrix` command, and an `end` command.

2.3.2 The alleles block

Those of you familiar with GDA will notice certain strong similarities between the example above and the one used as an example in the documentation for GDA. There are a few important differences, however. First, the block name is `alleles` in Hickory whereas it is `gdadata` in GDA. The `alleles` block is designed to be similar to the newer versions of traditional NEXUS blocks (e.g. the newer `characters` block vs. the older `data` block). Second, it is important to add the `newpops` keyword between `dimensions` and `npops`. Like other newer NEXUS blocks, the `alleles` block allows the taxa (in this case populations) to be described in a separate `taxa` block preceding the `alleles` block. While it is unlikely that anyone would wish to separate the population names into a separate `taxa` block, this behavior is allowed and the `newpops` keyword is thus needed to let the program know that the names of the populations will be listed within the `alleles` block. Third, the `gdadata` block assumes by default that labels will be provided for each individual, whereas the `alleles` block assumes that, by default, no labels will be provided. Thus, you should specify the keyword `labels` in the `format` command if you are providing individual labels. Again, this change was made to make the `alleles` block consistent with the newer NEXUS blocks now recognized by programs such as PAUP*, MacClade, and Mesquite.

In the `dimensions` command, `nloci` specifies the number of loci scored and `npops` is the number of populations scored. While the number of populations and loci must be specified in the `dimensions` command, but not the number of individuals per population. The end of the data for one population is signified by *either* a comma *or* the semicolon indicating the end of the `matrix` command itself. NEXUS files are for the most part not case-sensitive by default. A big exception is in the `matrix` command, where (by default) an allele named `A` is treated as being distinct from `a`.

The genotype for a diploid locus at a co-dominant locus takes the form `A/a`. In this example, the genotype represents a heterozygote having both allele `A` and allele `a`. The forward slash (/) symbol serves as a separator

symbol, making it clear which pairs of alleles belong to each locus. If the loci represent dominant genetic markers, do not use the slash and specify only one allele for each locus.

The phenotype for a dominant locus can be specified in several different ways. The keyword `dominant` signals that the data is for a dominant marker. By default, the dominant phenotype is scored as 1 (or +) or 0 (-). Should you want to use a different convention for scoring your phenotypes Hickory accepts the following alternatives:

```
dominant;  
dominant all:1;  
dominant all:0;  
dominant 1,2,3:1 4,5:0;  
dominant 1-3:1 4-5:0;
```

The first line is equivalent to the default option in which only the keyword `dominant` is specified. The second would have the dominant phenotype at all loci scored as 0 and the recessive phenotype scored as 1. The final two lines have the dominant phenotype at the first three loci scored as 1 and the dominant phenotype at the last two loci scored as 0.⁹

Be sure that the data for each locus are separated by one or more spaces. Specifically, an individual scored at five loci with dominant phenotype at loci 1, 3, and 5 should be written as

```
1 0 1 0 1
```

not as

```
10101
```

Notice that as of v1.0.4 “+” and “-” are allowed as allele labels (“+” corresponding with 1, and “-” corresponding with 0). Thus, the phenotype above could also be specified as

```
+ - + - +
```

⁹Most users will probably be satisfied with the default option. We provide the options simply because we think its better to err on the side of having more flexibility rather than less.

The `missing=?` in the `format` command specifies the character to be used for missing data. A missing diploid genotype would thus be specified as `?/?`, with a separate missing data symbol replacing each of the two alleles at this locus for this particular individual.

The `locusallelelabels` command provides a way to give names to loci. The `locusallelelabels` command is optional, but if it is present, it should comprise a comma-separated list of locus number, locus name pairs. No comma should follow the final locus name. Like the `matrix` command, the terminating semicolon takes the place of the final comma.

The `locusallelelabels` command is optional: loci will simply be numbered beginning with 1 if this command is absent. The `locusallelelabels` command also provides a way to have `Hickory` do some error checking for you. Notice how, for the locus `adh`, two allele names are provided (`slow` and `fast`). If in the matrix section `Hickory` finds an allele name other than either `slow` or `fast` for locus `adh`, it will stop parsing the data file and report it as an error. Thus, `Hickory` would find two errors in the sample data file above: the first is `Slow` (individual `indiv_3` of population `Embudo`) and the second is `solw` (individual 2 of the `Black Mesa` population). If valid allele names are not provided in a `locusallelelabels` command, `Hickory` simply interprets each distinct allele name as a separate allele. Were it not for the `locusallelelabels` command, the above example data set would be interpreted as having four alleles (not two) for the `adh` locus!

The `mosquito.nex` and `worknisw.nex` data files illustrate some additional options for data input. Both of these files are for co-dominant markers. Again, the data format should be relatively self-explanatory.

2.3.3 The hickory block

The `hickory` block is where you can set parameters governing the MCMC sampler. **Make sure that you understand what you're doing before you fiddle with anything here.** In fact, you can leave this segment of the data file out completely and use the default values, which is probably a good idea unless (a) you have good prior information on f or θ that you want to incorporate into your analysis or (b) are very sure that you want to change the default MCMC parameters.

If you want to use prior information on f , $\theta^{(II)}$,¹⁰ or on the mean allele

¹⁰Refer to Chapter 3 for a description of the new notation. It is, unfortunately, a little

frequencies, to change the length of the burn-in or sample, or to change the frequency of thinning, this is where you do it. The parameters should be self-explanatory, except for `alphaF`, `betaF`, `alphaTheta`, `betaTheta`, `alphaPi`, and `betaPi`. These are the parameters of a Beta distribution used as a prior for f , θ_x , and the mean allele frequencies.¹¹ If you don't know how to adjust those parameters to reflect the prior information you want to use, you should probably consult someone who understands the relationship between the parameters of a beta distribution and its mean, variance, and quantiles. Get that person to help you select appropriate values for the parameters. Briefly, setting `alphaF` and `betaF` both to 1 yields a Beta(1,1) prior for f , which is equivalent to a Uniform(0,1) prior. Increasing both `alphaF` and `betaF` (but keeping them equal) creates a symmetrical prior density that peaks at 0.5 and becomes more concentrated around 0.5 with higher values of the two parameters. Setting `alphaF=betaF=1000` would produce a very strong prior that would override the information in even a large dataset, and force estimates of f to be very close to 0.5. Without strong prior knowledge otherwise, we suggest using the default values, which produce “flat” (“vague”) priors.

If you want to see estimates of the mean frequency of every allele at every (polymorphic) locus and of the allele frequency of every allele at every locus in every populations specify `set reportFrequencies`. Setting this option will consume a lot more memory for data sets with a large number of loci. In most cases, we're not really interested in the allele frequencies anyway. That's why the default is not to report them. If you `set reportFrequencies`, the results (including 95% credible intervals) will be reported in the table after f and the θ s and before the within population diversity estimates.

In earlier versions of Hickory it was possible to change some options that affected the performance of the MCMC sampler. We have removed that option. All univariate parameters are sampled directly from the full conditional distribution using a slice sampler, and we use an adaptive routine to optimize the proposal distribution for multivariate parameters using a Metropolis-Hastings sampler.

bit complicated.

¹¹You can undoubtedly guess which is which.

Chapter 3

Interpreting the Output

This is what the output from a run with the data in `worknisiw.nex` looks like with `Estimate thetaY` selected and including results from all models, i.e., this is what you get when you `Do all analyses`.

Reading "HICKORY" block...

Finished with "HICKORY" block.

Summary of data now stored in memory

Number of loci: 2

Number of polymorphic loci: 2

Number of populations: 10

Data from file "C:\projects\bayes-fst\hickory\sample\worknisiw.nex" read and stored.

Sample characteristics:

Locus	Baboquivari	Chukut Kuk	Guachi	Gu Vo	Hickiwan	Pisinimo
MN locus (1)						
M/M	41	18	55	20	36	16
M/N	27	8	24	24	16	8
N/N	3	3	1	4	3	4
Ss locus (2)						
S/S	18	5	15	7	5	5
S/s	27	11	45	23	20	12
s/s	26	13	20	18	30	11

Locus	Schuk Toak	Sells	Sif Oidak	San Xavier
MN locus (1)				
M/M	43	128	27	41
M/N	21	52	21	16
N/N	4	7	3	3
Ss locus (2)				
S/S	11	28	9	27

S/s	48	80	28	25
s/s	9	79	14	8

Sampler characteristics:

Setting	Value
nBurnin	5000
nSample	25000
thin	5
alphaF	1.00
betaF	1.00
alphaTheta	1.00
betaTheta	1.00

Full model...

Posterior summary...

Parameter	Mean	s.d.	2.5%	50%	97.5%
f	0.0323	0.0211	0.0018	0.0296	0.0799
theta-I	0.3135	0.1674	0.0848	0.2792	0.7006
theta-II	0.0292	0.0141	0.0108	0.0260	0.0624
theta-III	0.0216	0.0064	0.0104	0.0211	0.0351
theta-Y	0.2924	0.1768	0.0550	0.2553	0.7208
rho	0.8989	0.0958	0.6318	0.9285	0.9921
hs[1]	0.4222	0.0178	0.3861	0.4228	0.4552
hs[2]	0.4139	0.0265	0.3590	0.4151	0.4620
hs[3]	0.3922	0.0181	0.3567	0.3921	0.4265
hs[4]	0.4483	0.0185	0.4082	0.4496	0.4807
hs[5]	0.3780	0.0253	0.3266	0.3787	0.4243
hs[6]	0.4302	0.0249	0.3767	0.4318	0.4737
hs[7]	0.4171	0.0179	0.3806	0.4176	0.4504
hs[8]	0.3812	0.0139	0.3537	0.3813	0.4077
hs[9]	0.4342	0.0187	0.3957	0.4348	0.4684
hs[10]	0.3936	0.0217	0.3493	0.3939	0.4336
Hs	0.4111	0.0074	0.3964	0.4113	0.4253
Ht	0.4207	0.0074	0.4060	0.4209	0.4350
Gst-B	0.0228	0.0068	0.0110	0.0223	0.0373
Dbar	64.1454				
Dhat	47.954				
pD	16.1914				
DIC	80.3368				

Parameters for f

alpha: 2.235579
beta: 66.98078
Ie: 2.591778
H-d: 0.00816

Parameters for theta-I

alpha: 2.09349
beta: 4.585288
Ie: 0.425793
H-d: 0.005348

Analysis started: Mon Apr 9 16:40:36 2007

Analysis finished: Mon Apr 9 16:42:50 2007
 Elapsed time: 00:02:14

f=0 model...

Posterior summary...

Parameter	Mean	s.d.	2.5%	50%	97.5%
theta-I	0.3143	0.1698	0.0824	0.2798	0.7165
theta-II	0.0295	0.0140	0.0106	0.0268	0.0641
theta-III	0.0219	0.0064	0.0105	0.0215	0.0353
theta-Y	0.2960	0.1778	0.0544	0.2622	0.7146
rho	0.8989	0.0961	0.6297	0.9300	0.9916
hs[1]	0.4225	0.0176	0.3873	0.4230	0.4561
hs[2]	0.4148	0.0257	0.3615	0.4158	0.4613
hs[3]	0.3919	0.0178	0.3557	0.3923	0.4273
hs[4]	0.4484	0.0183	0.4098	0.4492	0.4800
hs[5]	0.3776	0.0247	0.3256	0.3783	0.4235
hs[6]	0.4318	0.0237	0.3820	0.4329	0.4746
hs[7]	0.4176	0.0176	0.3827	0.4180	0.4506
hs[8]	0.3810	0.0139	0.3542	0.3812	0.4076
hs[9]	0.4344	0.0188	0.3956	0.4352	0.4686
hs[10]	0.3929	0.0217	0.3481	0.3934	0.4329
Hs	0.4113	0.0071	0.3973	0.4115	0.4248
Ht	0.4210	0.0071	0.4069	0.4211	0.4350
Gst-B	0.0231	0.0068	0.0110	0.0227	0.0375
Dbar	63.69				
Dhat	47.8741				
pD	15.8158				
DIC	79.5058				

Parameters for theta-I

alpha: 2.036017
 beta: 4.441807
 Ie: 0.413921
 H-d: 0.005223

Analysis started: Mon Apr 9 16:42:50 2007
 Analysis finished: Mon Apr 9 16:44:51 2007
 Elapsed time: 00:02:01

theta=0 model...

Posterior summary...

Parameter	Mean	s.d.	2.5%	50%	97.5%
f	0.0441	0.0238	0.0043	0.0422	0.0940
theta-I	0.3076	0.1746	0.0664	0.2726	0.7153
hs[1]	0.4119	0.0067	0.3988	0.4119	0.4251
hs[2]	0.4119	0.0067	0.3988	0.4119	0.4251
hs[3]	0.4119	0.0067	0.3988	0.4119	0.4251
hs[4]	0.4119	0.0067	0.3988	0.4119	0.4251
hs[5]	0.4119	0.0067	0.3988	0.4119	0.4251
hs[6]	0.4119	0.0067	0.3988	0.4119	0.4251
hs[7]	0.4119	0.0067	0.3988	0.4119	0.4251
hs[8]	0.4119	0.0067	0.3988	0.4119	0.4251
hs[9]	0.4119	0.0067	0.3988	0.4119	0.4251
hs[10]	0.4119	0.0067	0.3988	0.4119	0.4251

Hs	0.4119	0.0067	0.3988	0.4119	0.4251
Ht	0.4119	0.0067	0.3988	0.4119	0.4251

Dbar	113.408				
Dhat	110.69				
pD	2.71715				
DIC	116.125				

Parameters for f
alpha: 3.231368
beta: 69.978865
Ie: 2.415261
H-d: 0.012469

Analysis started: Mon Apr 9 16:44:51 2007
Analysis finished: Mon Apr 9 16:46:29 2007
Elapsed time: 00:01:38

Most of this should be pretty self explanatory – except for the lines dealing with θ . They’re explained below (§3.1). Another part that may be new to you, unless you’ve played with WinBUGS or an earlier version of Hickory, are the lines that follow each table. DIC is a model-choice criterion analogous to Akaike’s Information Criterion, which is explained below (§3.3). The lines headed **Parameters for θ** or **parameters for f** are also explained below (§3.2). Starting with v0.8 we report the starting time, ending time, and total time for each of the analyses.

The first table simply provides a count of the number of individuals with each genotype in the population. The number after the locus names refers to the locus numbers used internally, and corresponds with the locus numbers if you **reportFrequencies**. Only polymorphic loci are used in the analysis. With dominant markers the format is similar:

Population	Locus 2	Locus 3	Locus 4
-----	-----	-----	-----
Pop 1	9/10	1/8	4/9
Pop 2	6/8	5/8	3/8
Pop 3	5/6	3/6	2/6
Pop 4	6/6	2/6	3/6
Pop 5	8/8	4/8	1/7
Pop 6	7/7	4/7	2/7

The number of dominant alleles at a particular locus in a particular population appears before the slash, and the total number of individuals scored for

that locus appears after the slash. Notice that only the counts for polymorphic loci are reported.

What you're probably most interested in is what's in the next part of the table, the one headed `Full model . . . f` is an estimate of F_{is} , the inbreeding within populations. The number in the column headed `Mean` is the best single estimate for F_{is} . The numbers in the `2.5%` and `97.5%` columns correspond to the upper and lower bounds of the 95% credible interval. Credible intervals play a role in Bayesian statistics similar to confidence intervals in classical statistics. There's a good chance that the real value of F_{is} is between these two bounds.

3.1 Interpreting the θ statistics

In versions of `Hickory` prior to v1.1, we reported only a single estimate of θ , which we called `Theta-B`. As described in Fu et al. (2003) and Song et al. (2006), there are two statistics that one might think of as corresponding to Wright's F_{st} :

$$\begin{aligned}\theta^{(I)} &= \frac{\sigma_{p(t)}^2}{\mu_p(1 - \mu_p)} \\ \theta^{(III)} &= \text{E} \left(\frac{(1/K) \sum_k (p_k(t) - \bar{p}(t))^2}{\bar{p}(t)(1 - \bar{p}(t))} \right)\end{aligned}$$

$\theta^{(I)}$ corresponds to a scaled allele frequency variance, where the variance is measured across evolutionary time (assuming that the underlying stochastic evolutionary process has a stationary distribution) or across evolutionary replicates. μ_p is the mean allele frequency at stationarity. Notice that it is, by assumption, equal in all populations if allelic variation is neutral, provided that the mutational dynamics are the same across all populations. $\theta^{(I)}$ corresponds directly to Wright's F_{st} .

$\theta^{(III)}$ corresponds to a scaled allele frequency variance, where the variance is measured among contemporaneous populations.¹ Estimating $\theta^{(III)}$ requires that you know the number of populations exchanging genes, which

¹ $\bar{p}(t) = (1/K) \sum_k p_k(t)$ is the mean allele frequency in the set of populations at time t , $p_k(t)$ is the allele frequency in population k at time t , and K is the number of populations that are exchanging genes.

	mosquito.nex	worknism.nex
$\theta^{(I)}$	0.1905 (0.0730, 0.3951)	0.3100 (0.0818, 0.7226)
$\theta^{(II)}$	0.1165 (0.0419, 0.2575)	0.0290 (0.0105, 0.0632)
$\theta^{(III)}$	0.0433 (0.0248, 0.0646)	0.0214 (0.0101, 0.0352)

Table 3.1: Estimates of $\theta^{(I)}$, $\theta^{(II)}$, and $\theta^{(III)}$ for two sample data sets distributed with Hickory.

will usually be larger than the number of populations from which you actually have samples and will usually be unknown. If K is reasonably large (more than 15-20), however, then

$$\theta^{(III)} \approx \theta^{(II)} = \frac{\theta^{(I)}(1 - \rho)}{1 - \theta^{(I)}\rho} .$$

By default we provide estimates of $\theta^{(I)}$, $\theta^{(II)}$, $\theta^{(III)}$, and ρ . The estimate of $\theta^{(III)}$ assumes that samples have been taken from every population that is exchanging genes. If the number of populations sampled is small, $\theta^{(II)}$ and $\theta^{(III)}$ may be quite different from one another. In general, $\theta^{(II)}$ is the better measure of differentiation among contemporaneous populations unless you have sampled from all populations that are potentially exchanging genes.

Notice that in Table 3.1 the estimates of $\theta^{(I)}$ are larger than those of $\theta^{(II)}$ and $\theta^{(III)}$. That's expected, because $\theta^{(II)}$ and $\theta^{(III)}$ reflect only the differentiation among contemporaneous populations. Because we expect populations to be partially coupled through gene flow, allele frequencies are correlated. As a result, allele frequency variation among populations at any one point in time is smaller than variation in any one population over time. Notice also that $\theta^{(II)}$ and $\theta^{(III)}$ have reasonably similar magnitudes in the `worknism.nex` data set, which has 10 populations, but that they are very different in the `mosquito.nex` data set, which has only two populations.

$\theta^{(I)}$ is the best estimate of Wright's F_{st} . $\theta^{(II)}$ is the best estimate of the proportion of genetic diversity due to differences among contemporaneous populations, a statistic similar to Nei's G_{st} . Unlike the Bayesian version of G_{st} described below, it treats populations as a random sample from all populations that could have been sampled. Unless you have reason to ignore stochasticity in the underlying evolutionary process, $\theta^{(II)}$ will provide a better estimate of differentiation among contemporaneous populations than `Gst-B`.

Should you decide that you do not want to estimate the among-population

correlation, uncheck the option to “**Estimate theta-Y**” before your analysis. You will then receive an estimate *only* for $\theta^{(I)}$. In our experience estimates of $\theta^{(II)}$ are relatively insensitive to whether or not θ_y is estimated, but the differences between $\theta^{(I)}$ and $\theta^{(II)}$ (or $\theta^{(III)}$) can be substantial. In general, we recommend estimating θ_y .

Whether to base interpretations on $\theta^{(I)}$ or $\theta^{(II)}$ depends on the purposes of your study. If your study is intended to provide some insight into the extent of evolutionary connection among populations, $\theta^{(I)}$ is probably the most appropriate statistic, because it is the most closely related to Wright’s F_{st} .² It measures the variability of allele frequencies within a single population across time. Unfortunately, $\theta^{(I)}$ is not the parameter estimated by typical approaches to analysis of F -statistics, e.g., Weir & Cockerham’s approach. Such approaches estimate $\theta^{(II)}$, which is a measure of the amount of genetic differentiation among populations. If your study is intended primarily to say something about the extent of genetic differentiation among contemporaneous populations, then $\theta^{(II)}$ is not only the most relevant statistic, it is also directly comparable to estimates of F_{st} based on Weir & Cockerham’s approach.

The lines labeled **hs** [] are estimates of genetic diversity (defined as average panmictic heterozygosity) within each population. **Hs** is the average of **hs** [] across populations, **Ht** is the panmictic heterozygosity based on mean allele frequencies, and **Gst-B** is a Bayesian analog of Nei’s G_{st} (see Holsinger 1999 for details; see above for a brief discussion on differences from $\theta^{(II)}$).

3.2 Parameters of the posterior distributions

The posterior distributions for f and θ lie on $[0, 1]$. There are two convenient ways to summarize the posterior distribution for these parameters. We can either report the posterior mean and standard deviation, or we can report the parameters of a Beta distribution that provides a good approximation to the posterior distribution. **alpha** and **beta** are those parameters. We recommend that users report either the standard deviation or these parameters for f and θ . In that way, future investigators can use earlier results to provide informative priors on the parameters should they choose to do so.

²And to $1/(4N_e(m + \mu) + 1)$, if you’re willing to assume mutation-migration-drift stationarity and a finite-island model of migration.

I_e is a measure of the “information” provided about a parameter by the data. The larger the value of I_e , the more information that is provided. $H-d$ is the Hellinger distance between the simulated posterior distribution and the Beta distribution that approximates it. It is interpreted as the percentage of non-overlap between the distributions. For identical distributions $H-d = 0$. For completely non-overlapping distributions $H-d = 1$. We have yet to see a data set with $H-d > 0.008$ for either f or θ^B , which indicates that a Beta distribution provides an excellent approximation to the posterior distributions of these parameters.

Holsinger and Wallace (2004) provide details on the calculation of I_e and $H-d$. Future versions of `Hickory` will allow users to simulate data sets. We expect such simulations to assist sampling designs by identifying the combination of number of populations, sample size within populations, and number of loci that provides the most information about the parameters of interest. I_e provides an appropriate measure for such investigations.

3.3 Model choice

Spiegelhalter et al. (2002) introduced a convenient method for choosing among alternative statistical models. The Deviance Information Criterion they propose (DIC) is similar in spirit to Akaike’s Information Criterion in that it takes account both of how well a particular model fits the data and of how many parameters are required to do it.

Table 3.2 summarizes the DIC statistics for the three analyses presented above. $Dbar$ is a measure of how well the model fits the data (smaller values are better). DIC also takes into account the approximate number of parameters being estimated (pD).³ Models with the smaller DIC are preferred, and a difference of less than 5 or 6 units among models is small enough that there isn’t strong evidence favoring one model over another. Thus, with the Workman and Niswander data, we have compelling evidence that there are genetic differences among populations (a difference of about 36 units), but no evidence that the genotype frequencies within populations departs from Hardy-Weinberg expectations (a difference of only about 1 unit).

As described in Holsinger and Wallace (2004), you’ll want to pay attention not only to DIC, but also to $Dbar$ and pD in deciding among models. Although

³If you don’t understand why we only know the number of parameters being estimated approximately, feel free to ask, but be forewarned: the answer’s a bit complicated.

Model	Dbar	Dhat	pD	DIC
Full	64.0822	48.0605	16.0217	80.1040
$f = 0$	63.4415	47.8337	15.6078	79.0492
$\theta^B = 0$	113.3748	110.6938	2.6810	116.0557

Table 3.2: DIC statistics for three models applied to the co-dominant marker data set in `worknisw.nex`

models with the smaller DIC are generally preferred, the model with a better fit to the data (smaller `Dbar`) may be preferred when the DIC difference is primarily a result of differences in model dimension, `pD`. In the Workman and Niswander data, the Full model and the $f = 0$ model are little different in any of the DIC components, indicating that there is no reason to prefer the Full model to the one with $f = 0$. In other words, we have no evidence of inbreeding in these populations. The $f = 0$ model is, however, strongly preferred to the $\theta^B = 0$ model, indicating that there are differences in allele frequency among the populations.

3.4 Estimating f from dominant markers

In Holsinger et al. (2002) we suggested that reliable estimates of F_{is} can be obtained from dominant marker data. We are now much less certain that is the case. A user of `v0.6` brought a data set to our attention in which it is biologically unreasonable to think that F_{is} is large, yet estimates derived from `Hickory` suggest that it is around 0.9. Kent spent several weeks in late June and early July, 2002 trying to identify a bug or numerical glitch in the program. He found a small glitch, but even when he substituted robust code for Gibbs sampling `Hickory` produced unreasonable estimates of f for this data set. Unfortunately, the problem seems to be most pronounced when a large number of loci are available, and we haven't been able to reproduce it in our simulated data sets. It also appears to arise primarily (only?) when the sample size within populations is fairly small, say less than 10 individuals.

The problem seems intrinsically related to the weak identifiability of f with dominant markers. As we explain in Holsinger et al.(2002), we obtain information about f with dominant markers only because we assume the same f and θ across loci and because the pattern of variation among

populations constrains the reasonable values of f only indirectly through its constraints on θ .

As a result, we urge you to regard estimates of F_{is} derived from dominant markers with extreme caution, especially if the sample size within populations is small. If you get an estimate of F_{is} that seems consistent with what you expect based on other information, e.g., allozymes, microsatellites, or general knowledge of the mating system, you could regard the estimates from `Hickory` as further evidence for inbreeding (or the lack of it) in your populations. If your estimate of F_{is} is inconsistent with other sources of information, however, we urge you to discount the estimate `Hickory` gives you unless it can be verified through other means.

3.4.1 Estimating θ without estimating f

Because estimates of f derived from dominant marker data may be unreliable, we provide another way of estimating θ . If you select `Run f free model` from the `Analyses` menu, the sampler will not attempt to estimate f . Instead it will choose values of f at random from its prior distribution while estimating other parameters during the MCMC run. Estimates of θ and other parameters obtained in this way incorporate all of the uncertainty in the prior of f . The effect is several-fold:

1. You don't have to worry that unreasonable estimates of f are affecting your estimate of θ .
2. The posterior mean of θ will be drawn toward the value that would be obtained assuming that f was fixed at its prior mean. Thus, if you assume a non-informative prior in an outbred population, the posterior mean of θ will be pulled toward the value you would get if you assumed a fixed value of $f = 0.5$. Fortunately, the posterior mean of θ does not appear to be too greatly affected by the prior on f when there are a large number of polymorphic loci in the sample.
3. The credible interval for θ will be broader than if you estimate f . (So if you think `Hickory` is giving you reasonable values for f , you may want to estimate θ using the full model.)
4. If you have prior information on reasonable values of f , you can set that prior information and sample estimates of θ given that prior instead of one that assumes a uniform distribution.

Chapter 4

Producing plots

Beginning with version 1.0 of `Hickory` plots of the posterior distribution and the sample trace can be produced without leaving the program (Figure 4.1). The `Plots` menu will have options available to plot separate results for each of the analyses that have been run. By default the posterior distribution plotted is for the Beta distribution that approximates the posterior. The `Include kernel density` option is a toggle that includes a plot of a Gaussian kernel density estimate for the posterior. The kernel density is drawn in blue. You can save the plot to a file or print it using the appropriate options under the `File` menu. Supported file types for saving are: BMP, JPEG, PNG, PCX, PNM, and XPM.¹

We also provide a small file of functions for R (<http://www.r-project.org>) that will produce figures like those we present in Holsinger et al. (2002) and Holsinger and Wallace (2004). To produce these figures you have to activate the `Sample log file` option on the `Analyses` menu, or specify the log file option on the command line. Doing that will produce a file named `<basename>-full.txt` if you run the `full` model in the same directory as the `NEXUS` file that you use for input and `<basename>` is the name of your `NEXUS` file without the `.nex` extension.² If that file already exists, it will create a new one numbering sequentially from 01 to 99, e.g., `<basename>-full-01.txt`.

To produce figures using R, start R, `source()` the file named

¹TIFF may show up as an option, but it is not supported. If you use a “.tif” extension, you’ll get an error message, and you’ll have to try again.

²It will produce `<basename>-fzero.txt` and `<basename>-thetazero.txt` if you do all three analyses.

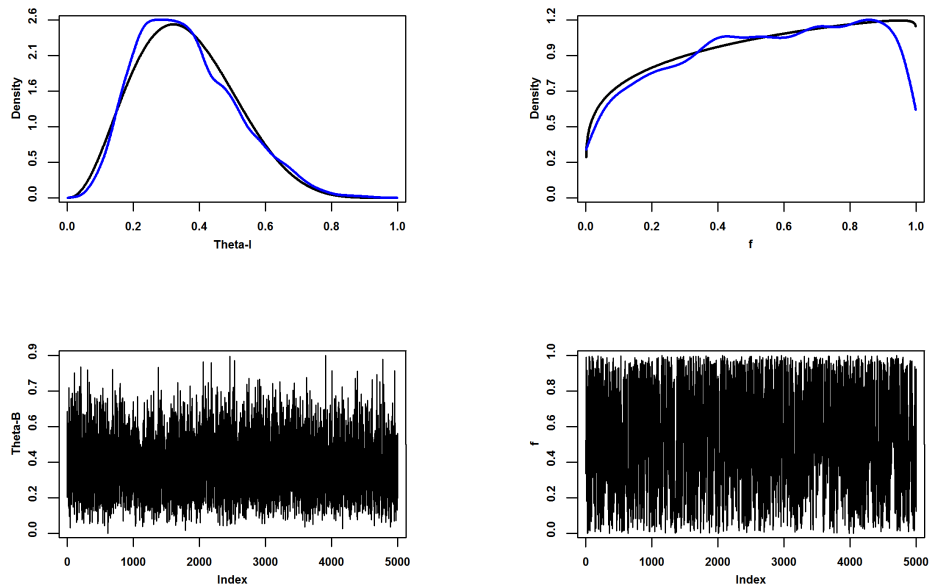


Figure 4.1: Posterior distributions and trace plots for the data in `dominant.nex`.

`plot-results.R` in the `build` directory of the distribution, and type `plot.results()` at the command prompt. You'll be asked for the name of the data file to analyze. Enter `<basename>-full.txt`. You'll get two sets of plots, one like those you see in Holsinger et al.(2002) and another that shows a simultaneous 95% credible region (the minimum convex polygon containing 95% of the posterior density of the MCMC sample) and all points in the MCMC sample. You'll see the results on two scales, one where the axes are f and θ^B and one where the results are $\log(f/(1-f))$ and $\log(\theta^B/(1-\theta^B))$.

Chapter 5

Posterior comparisons

One advantage of a Bayesian approach to analyzing population structure is the relative ease with which certain hypotheses can be analyzed. For example, a variety of powerful and sophisticated approaches exist for determining whether F_{is} or F_{st} is different from 0. It is nonetheless very difficult to determine whether two data sets, both of which have $F_{st} > 0$, have different values of F_{st} . If the analyses are done in **Hickory**, however, each analysis provides a full posterior distribution for θ^B . All we need to do is to take paired random samples from the posterior distribution of θ^B for each data set and calculate the difference for each pair. Then we have the posterior distribution of the difference between θ^B for each analysis. If the 95% credible interval for that difference includes 0, we have no evidence that the estimates are different. Otherwise, we have evidence that the amount of among-population differentiation is greater in one sample than in another.

Suppose, for example, that we want to determine whether estimating f from the data has an effect on our estimate of θ^B in the Workman and Niswander data. Run the Full and $f = 0$ analyses of the data, making sure the **Sample log file** option is checked before we begin. After the analyses are done, select **Compare posteriors** from the **Tools** menu. Choose **worknisw-full.txt** and **worknisw-fzero.txt** from the file dialog boxes. Results similar to these will appear in the **Hickory** output window:

Comparing results as

```
C:\projects\bayes-fst\hickory\sample\worknisw-full.txt
- C:\projects\bayes-fst\hickory\sample\worknisw-fzero.txt
-----
```

```
f: NA
theta: -7.822e-05 (-0.03923,0.04041)
```

The comparison is done in the direction indicated in the output. Parameters in the second file selected (`worknisw-fzero.txt`) are subtracted from those in the first file selected (`worknisw-full.txt`). The NA on the line with f means that the comparison is “not applicable” (because f isn’t estimated in the $f = 0$ model). The difference in estimates of θ^B between the two analyses is less than 10^{-4} , and the 95% credible interval overlaps zero almost symmetrically. We can confidently conclude that estimating f has no detectable impact on our estimate of how genetic variation is partitioned among populations.¹

¹This should not be surprising, since f is small ($\hat{f} = 0.0326$) in the full model.

Chapter 6

Revision history

- Changes from v1.0.4 to v1.1
 - Use slice sampling for all univariate parameters.
 - Estimate among-population correlation by default.
 - Make log files Tracer compatible (<http://evolve.zoo.ox.ac.uk/software.html?id=tracer>).
 - Eliminate (faulty) option to save plots as Postscript. Substitute option to save in supported bitmap formats (BMP, PNG, JPEG, PNM, PCX, XPM).

- Changes from v1.0 to v1.0.4
 - Alleles Block
 - * Added ability to specify a fixed set of alleles for any given locus in the `locusallelelabels` command. After the locus name, a backslash symbol can be followed by a space-delimited list of valid allele names. If any allele in the matrix differs from this list of valid alleles, an error message is generated. This allows typographic errors to be caught. For example, if no list of valid allele names is provided, entering 'solw' in the data matrix when 'slow' was intended will result in a new (rare) allele 'solw' being recognized.
 - * You can use +/- for presence absence of dominant marker bands instead of scoring them as 1/0.

- Polymorphic loci
 - * The algorithm used for identifying polymorphic loci in dominant marker data has been corrected. It now recognizes as polymorphic those loci in which some populations are monomorphic for band presence and others are monomorphic for band absence.¹
- H_S calculations for co-dominant markers
 - * An error in the algorithm for calculating H_S with co-dominant markers has been corrected. Because G_{ST} is calculated from H_S , reported values of G_{ST} were also incorrect.²
- DIC calculations
 - * A numerical overflow that sometimes results in NaN being reported in the DIC statistics has been corrected.
- Changes from v0.8 to v1.0
 - Added posterior plots.
 - Added posterior comparisons.
 - Added report of parameter estimates for posterior densities, of the information provided by the data about parameters, and of the Hellinger distance between the posterior density and a beta density selected to approximate it.
 - Added description of implementation details to documentation.
 - Allow user specification of proposal parameters.
- Changes from v0.7 to v0.8
 - Added analytical routines for co-dominant marker data.
 - Removed Bayes factor calculations.
- Changes from v0.6 to v0.7

¹The code for identifying polymorphic loci with co-dominant markers has not been changed.

²This error does not affect estimates from dominant marker data.

- Hickory block of NEXUS data file
 - * Added `set alphaPi` and `set betaPi` to specify prior for mean allele frequencies (default `alphaPi = betaPi = 0`).
 - * Added `set prnwidth` and `set colwidth` to adjust tabular output.
 - * Added `set reportFrequencies` to report mean allele frequencies at every locus and allele frequencies at every locus in every population.³
 - * Added `set estimatePi` and `reportPi` (**EXPERIMENTAL**) to estimate priors for mean allele frequencies.
- Analyses menu
 - * Added `Run f free model` to estimate θ^B while allowing f to follow its prior distribution.

³As if you really wanted that much information.

Chapter 7

Implementation details

Hickory uses a combination of slice sampling (for univariate parameters) and a single-component Metropolis-Hastings sampler (for multivariate parameters, i.e., allele frequencies). The Metropolis-Hastings sampler uses a Dirichlet proposal density. The proposal mechanism is documented in `MCMC++`. See the documentation there for details.

Chapter 8

GNU Free Documentation License

Version 1.2, November 2002

Copyright © 2000,2001,2002 Free Software Foundation, Inc.

59 Temple Place, Suite 330, Boston, MA 02111-1307 USA

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

0. Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document “free” in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondly, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of “copyleft”, which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used

for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. Applicability and definitions

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The “Document”, below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as “you”. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A “Modified Version” of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A “Secondary Section” is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document’s overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The “Invariant Sections” are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The “Cover Texts” are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A “Transparent” copy of the Document means a machine-readable copy,

represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not “Transparent” is called “Opaque”.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, \LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The “Title Page” means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, “Title Page” means the text near the most prominent appearance of the work’s title, preceding the beginning of the body of the text.

A section “Entitled XYZ” means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as “Acknowledgements”, “Dedications”, “Endorsements”, or “History”.) To “Preserve the Title” of such a section when you modify the Document means that it remains a section “Entitled XYZ” according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. Verbatim copying

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. Copying in quantity

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin dis-

tribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. Modifications

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
- D. Preserve all the copyright notices of the Document.
- E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

- G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- H. Include an unaltered copy of this License.
- I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
- N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
- O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the

Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version’s license notice. These titles must be distinct from any other section titles.

You may add a section Entitled “Endorsements”, provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. Combining documents

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled “History” in the various original documents, forming one section Entitled “History”; likewise combine any sections Entitled “Acknowledgements”, and any sections Entitled “Dedications”. You must delete all sections Entitled “Endorsements”.

6. Collections of documents

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. Aggregation with independent works

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an “aggregate” if the copyright resulting from the compilation is not used to limit the legal rights of the compilation’s users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document’s Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. Translation

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled “Acknowledgements”, ”Dedications”, or “History”, the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. Termination

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. Future revisions of this license

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See <http://www.gnu.org/copyleft/>.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License “or any later version” applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the

Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

Addendum: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

Copyright © Year Your name

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled “GNU Free Documentation License”.

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the “with...Texts.” line with this:

with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.

Literature Cited

- Holsinger, K. E. 1999. Analysis of genetic diversity in geographically structured populations: a Bayesian perspective. *Hereditas* 130:245–255.
- Holsinger, K. E., P. O. Lewis, and D. K. Dey. 2002. A Bayesian method for analysis of genetic population structure with dominant marker data. *Molecular Ecology* 11:1157–1164.
- Holsinger, K. E., and L. E. Wallace. 2004. Bayesian approaches for the analysis of population genetic structure: an example from *Platanthera leucophaea* (Orchidaceae). *Molecular Ecology* 13:887–894.
- Song, S., D. K. Dey, and K. E. Holsinger. 2006. Hierarchical models with migration, mutation, and drift: implications for genetic inference. *Evolution* 60:1–12.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B* 64:483–689.