PARAMO pipeline

***P**hylogenetic **A**ncestral **R**econstruction of **A**natomy by **M**apping **O**ntologies*

PARAMO pipeline requires three initial pieces of data: character matrix, dated phylogeny, and anatomy ontology. Herein, we use a of 9 characters and 87 species from the large-scale phylogeny of Hymenoptera M. J. Sharkey et al. (2012), this dataset was slightly modified for the demonstrative purpose. For reconstructing character histories, we use the dated phylogeny of Klopfstein et al. (2013), and for characters-ontology linking, we use Hymenoptera Anatomy Ontology (HAO, Yoder et al. 2010). In this demonstration, we are interested in constructing the amalgamated characters for the three leveles of amalgamation (=anatomical hierarchy): anatomical dependecies (AD), body regions (BRs) and entire phenotype (EF). At the BR level, three main body regions are considered – "head", "legs" and "wings".

STEP 1. Initial character matrix

Our initial character matrix consists of a sample of 9 characters scored for 87 taxa of Hymenoptera. These characters are taken from the large Hymenoptera dataset of M. J. Sharkey et al. (2012) and slightly modified for the demonstrative purpose. The nexus file of the intial matrix can be found at `STEP_1/Step1_matrix.nex` and viewed using, for example, Mesquite. Below, in describing characters, the follwoing notation is used $C_{\#}\{S_1, S_2, ...\}$ where $C_{\#}$ stands for a character ID and $S_1, S_2, ...$ stand for character states. Let us have a look at the character report.

```
##   CHAR_ID CHAR*                            CHAR_STATEMENT   STATE_0
## 1      C1     3            Notch on medial margin of eye    absent
## 2      C2    23                        Position of labrum  anterior
## 3      C3    25                                   Labrum    absent
## 4      C4   353  Forewing  costal and radial vein fusion not_fused
## 5      C5   363  Hind wing  subcostal  SC  vein, absent         no
## 6      C6   363 Hind wing  subcostal  SC  vein, present        yes
## 7      C7   380          Inner posterior mesotibial spur    simple
## 8      C8   381              Foretibial apical sensillum   present
## 9      C9   382              Metatibial apical sensillum   present
##                      STATE_1        DEPENDENCY
## 1                    present                no
## 2                  posterior   C2{0,1} < C3{1}
## 3                    present   C3{1} > C2{0,1}
## 4 fused_along_their_lengths                no
## 5                        yes C5{0,1} <>C6{1,0}
## 6                         no C5{0,1} <>C6{1,0}
## 7    modified into a calcar                no
## 8                     absent                no
## 9                     absent                no
##                                                                                CODING NOTE
## 1
## 2                                                                 state "-" means dependency on C3
## 3
## 4
## 5 Char363 from Sharkey_2012 is recoded into binary on pupose to demonstrate the synchronous dependecy
## 6
## 7
## 8
## 9
```

Note, the two pairs of characters in the matrix $\{C_2, C_3\}$ and $\{C_5, C_6\}$ are subjected to anatomical dependencies.

- $C_2\{0,1\}$ is hierrachically (anatomically) dependent on $C_3\{0\}$. This dependecy is indicated by $<$ or $>$ depending on the direction of the dependecy. The hierrachical dependecy means that states $C_2\{0,1\}$ appear immediately as C_3 swithches to the state $C_3\{0\}$.
- C_5 and C_6 are subjected to synchronous changes, which means tha the states of this characters are mutually exclusive and hence dependent because one trait is coded using absent/present coding. The synchronous dpenedecy is indicated as $<>$; the notation $C_5\{0,1\} <> C_6\{1,0\}$ means – if C_5 is $\{0\}$ then C_6 is $\{0\}$, and if C_5 is $\{1\}$ then C_6 is $\{1\}$.

Step 2. Incorporating anatomical dependencies: constructing amalgamations at the AD level

The pairs of anatomically depenent characters – C_2, C_3 and C_5, C_6 have to be apropriately amalgamated into single characters to adeuqtely model the dependecies. The amalgamation produces the following (see the text in the article):

- $C_3 \oplus C_2 = C_{3,2}\{00, 01, 10, 11\}$. $C_{\{}3, 2\}$ is coded in the matrix as $C_{3,2}\{0\&1, 0\&1, 2, 3\}$.
- C_5 and C_6 are combinded into $C_{\{}5, 6\}$. The synchronous dependency between these characters has to be eliminated that gives the character $C_{5,6}$ without changing the state pattern.

The recoding of dependent characters constructs the amalgamated characters at the AD level. If a character does not display any dependencies then we treat it as correctly amalgamated at the AD level by default. The new matrix and character report of the characters amalgamated at the AD level can be found in `STEP_2/Step2_matrix.nex` and `STEP_2/Char_info_step_2.csv` respectively. Let's have a look into them.

```
CH2<-read.csv(file="STEP_2/Char_info_step_2.csv", header = T, as.is=T, check.names=F)
CH2
```

```
##   CHAR_ID  CHAR*                          CHAR_STATEMENT          STATE_0
## 1      C1      3            Notch on medial margin of eye           absent
## 2    C3,2 25, 23              Labrum + Position of labrum absent, anterior
## 3      C4    353  Forewing  costal and radial vein fusion        not_fused
## 4    C5,6    363 Hind wing  subcostal  SC  vein, present           present
## 5      C7    380          Inner posterior mesotibial spur           simple
## 6      C8    381              Foretibial apical sensillum          present
## 7      C9    382              Metatibial apical sensillum          present
##                     STATE_1           STATE_2            STATE_3
## 1                   present
## 2         absent, posterior present, anterior present, posterior
## 3 fused_along_their_lengths
## 4                    absent
## 5    modified into a calcar
## 6                    absent
## 7                    absent
```

```
# kable(CH2, caption = 'Anatomical characters. CHAR*- char id in Sharkey et al. (2012) ')%>%
#   kable_styling(full_width = F, font_size=11) %>%
#   column_spec(1, bold = T)

# characters matrix
MT<-read.csv("STEP_4/matrix.csv", header = T, row.names=1, as.is=T, check.names=F)
MT
```

```
##               C1 C3-2 C4 C5-6 C7 C8 C9
```

```
## Acanthochalcis  0    3  0    1  0  0  1
## Aleiodes        1    3  1    1  0  ?  ?
## Anacharis       0    3  0    1  0  ?  ?
## Archaeoteleia   0    3  0    1  0  ?  ?
## Athalia         0    3  0    1  0  0  1
## Aulacus         0    3  0    1  0  1  1
## Australomymar   ?    3  ?    ?  ?  ?  ?
## Austroserphus   0    3  0    1  0  ?  ?
## Belyta          0    3  0    1  0  ?  ?
## Brachygaster    0    ?  0    1  0  1  1
## Cales           0    3  0    1  0  ?  ?
## Cephalcia       ?    3  ?    ?  ?  ?  ?
## Cephalonomia    0  0&1  0    1  1  ?  ?
## Cephus          0    3  0    1  0  1  1
## Ceraphron       0    ?  ?    1  1  ?  ?
## Chiloe          ?    3  ?    ?  ?  ?  ?
## Cirrospilus     0    3  0    1  0  ?  ?
## Cleonymus       0    3  0    1  0  ?  ?
## Coccobius       ?    3  0    1  ?  ?  ?
## Coccophagus     0    3  0    1  0  ?  ?
## Corynis         ?    3  ?    ?  ?  ?  ?
## Decameria       ?    3  ?    ?  ?  ?  ?
## Diplolepis      0    3  0    1  0  ?  ?
## Doryctes        0    3  1    1  0  ?  ?
## Dusona          0    ?  1    1  0  ?  ?
## Eurytoma        0    ?  0    1  0  ?  ?
## Evania          0    3  0    1  0  1  1
## Evaniella       0    ?  0    1  0  1  1
## Foersterella    ?    3  ?    ?  ?  ?  ?
## Gasteruption    0    3  0    1  0  1  1
## Gonatocerus     0    3  0    1  0  ?  ?
## Hartigia        ?    3  ?    ?  ?  ?  ?
## Helorus         0    3  0    1  0  1  1
## Heteroperreyia  0    3  ?    ?  ?  ?  ?
## Ibalia          0    3  0    1  0  0  1
## Ismarus         0    ?  0    1  ?  ?  ?
## Isostasius      0    ?  0    1  0  ?  ?
## Labena          1    ?  1    1  0  ?  ?
## Lagynodes       0    3  ?    ?  0  ?  ?
## Lymeon          0    3  1    1  0  ?  ?
## Maaminga        0    3  0    1  0  ?  ?
## Macroxyela      0    2  0    0  0  1  1
## Megalyra        0    3  0    1  0  0  0
## Megaspilus      0    ?  1    1  0  ?  ?
## Megastigmus     0    3  0    1  0  ?  ?
## Megischus       0    3  0    1  ?  ?  ?
## Melanips        0    3  0    1  0  ?  ?
## Metapolybia     1    3  0    1  0  1  1
## Monoctenus      0    3  0    1  0  1  1
## Monomachus      0    3  0    1  ?  ?  ?
## Mymaromma       0    3  ?    1  0  ?  ?
## Nasonia         0    3  0    1  0  ?  ?
## Notofenusa      0    ?  0    1  0  1  1
## Onycholyda      0    3  0    0  0  1  1
```

```
## Orthogonalys    0    3  0    1  0  1  1
## Orussobaius     0    3  ?    ?  ?  ?  ?
## Orussus         0    3  0    1  0  ?  ?
## Pantolytomyia   0    3  0    1  ?  ?  ?
## Parnips         0    ?  0    1  0  ?  ?
## Pelecinus       0    3  0    1  0  ?  ?
## Periclistus     0    3  0    1  0  1  1
## Pimpla          0    ?  1    1  0  ?  ?
## Pison           1    3  0    1  0  ?  ?
## Poecilopsilus   0    3  0    1  0  ?  ?
## Pristaulacus    0    3  0    1  0  ?  ?
## Proctotrupes    ?    3  0    1  0  1  1
## Proplatygaster  0    3  0    1  0  ?  ?
## Pseudofoenus    0    3  0    1  0  ?  ?
## Psilocharis     ?    3  ?    ?  ?  ?  ?
## Rhopalosoma     1    ?  1    1  0  1  1
## Rhysipolis      1    3  1    1  0  ?  ?
## Ropronia        0    3  0    1  0  ?  ?
## Runaria         ?    3  ?    ?  ?  ?  ?
## Sapyga          1  0&1  0    1  0  ?  ?
## Schlettererius  0    3  0    1  ?  0  1
## Sirex           ?    3  ?    ?  ?  ?  ?
## Stangeella      0    3  0    1  0  0  0
## Sterictiphora   ?    3  ?    ?  ?  ?  ?
## Syntexis        0    3  0    1  0  1  1
## Telenomus       0    ?  ?    ?  ?  ?  ?
## Tenthredo       ?    3  ?    ?  ?  ?  ?
## Tremex          0    3  0    1  0  1  1
## Vanhornia       0    3  0    1  ?  ?  ?
## Wroughtonia     0    3  1    1  0  ?  ?
## Xiphydria       0    3  0    1  0  ?  1
## Xyela           ?    2  0    0  0  1  1
## Zagryphus       0    ?  1    1  0  ?  ?
```

STEP 3. Linking anatomical characters with ontology

Having initial characters properly coded to account for anatomical dependecies, let's move on character-ontology linking. The Table below shows the Hymenoptera characters linked with the terms of Hymenoptera Anatomy Ontology HAO. This table will be used in *"Retrieve all characters"* (*RAC*) query that retrives all characters asscoaited with an input ontology term.

```
AN<-read.csv(file="STEP_3/Char_annotation.csv", header = T, as.is=T, check.names=F)
AN
```

```
##   CHAR_ID CHAR_ID2  CHAR*                          CHAR_STATEMENT
## 1      C1        1      3            Notch on medial margin of eye
## 2    C3,2      3,2 25, 23              Labrum + Position of labrum
## 3      C4        4    353  Forewing  costal and radial vein fusion
## 4    C5,6      5,6    363 Hind wing  subcostal  SC  vein, present
## 5      C7        7    380          Inner posterior mesotibial spur
## 6      C8        8    381              Foretibial apical sensillum
## 7      C9        9    382              Metatibial apical sensillum
##        HAO_ID HAO_ID_NAME
```

```
## 1 HAO:0000234     cranium
## 2 HAO:0000639  mouthparts
## 3 HAO:0000351   fore wing
## 4 HAO:0000400   hind wing
## 5 HAO:0001351   mesotibia
## 6 HAO:0000350  fore tibia
## 7 HAO:0000631   metatibia
```

To run *RAC*, we use `ontologyIndex` package and a set of preccoked `R` functions located in `PARAMO_functions.R`. For our demonstrative purposes *RAC* is supposed to work with the BR and EF levels of amalgamation; remember that, at the BR level, we condider three body regions "head", "wings" and "legs". So, let's test our query. First of all, we need to make character-ontology to be a part of the ontolgy graph.

```
library("ontologyIndex")
```

```
## Warning: package 'ontologyIndex' was built under R version 3.5.2
```

```
source("R_PARAMO/PARAMO_functions.R")

# opening HAO file ("BFO:0000050" is part_of relatinship)
ONT<-get_OBO("STEP_3/HAO.obo", extract_tags="everything", propagate_relationships = c("BFO:0000050", "is

# let's create  "annot" list of anotations from the annotation table
char_id<-paste0("CHAR:", AN$CHAR_ID2)
annot<-set_names(table2list(AN[,c(2,5)]), char_id)

# next we make the annotations to be the part of the ontology object ONT
ONT$terms_selected_id<-annot
```

Now, we can construct and query the vectors of HAO terms that correspond to the focal BRs and EF.

```
# BR level
level2<-set_names(c("HAO:0000397", "HAO:0001089", "HAO:0000494"), c("head", "wings", "legs") )

# EF level
level3<-set_names(c("HAO:0000012"), c("whole_organism") )

# we use get_descendants_chars to get the set of all anatomcal characters that descend from a particula
#get_descendants_chars(ONT, annotations="manual", terms="HAO:0000012")

# now we can query all characters for the level 2 and 3 using the Ontology
L2<-lapply(level2, function(x)
  get_descendants_chars(ONT, annotations="manual", terms=x)  )
"Level 2"
```

```
## [1] "Level 2"
```

```
L2
```

```
## $head
## [1] "CHAR:1"   "CHAR:3,2"
##
## $wings
## [1] "CHAR:4"   "CHAR:5,6"
##
## $legs
## [1] "CHAR:7" "CHAR:8" "CHAR:9"
```

```
L3<-lapply(level3, function(x)
  get_descendants_chars(ONT, annotations="manual", terms=x)  )
"Level 3"
```

```
## [1] "Level 3"
```

```
L3
```

```
## $whole_organism
## [1] "CHAR:1"   "CHAR:3,2" "CHAR:4"   "CHAR:5,6" "CHAR:7"   "CHAR:8"
## [7] "CHAR:9"
```

STEP 4. Inference: linking characters with models and tree

At this step, we need to construct data files for analysing the set of our seven individual characters (obtained at Step 2) using RevBayes. Three files have to be created for each character: (1) chraracter file, (2) RevBayes script, and (3) tree file that is shared accross all chracters. The process of file cretion can be automitized using the following scripts.

```
# reading chracter matrix
MT<-read.csv("STEP_4/matrix.csv", header = T, row.names=1, as.is=T, check.names=F)

# creating chracter files using the matrix
#setwd("~/Documents/Recon-Anc_Anat/Supplementary_materials/STEP_4/RevBayes/data")
for (i in 1:ncol(MT))
{
  C.rev<-MT[,i]
  C.rev<-gsub("&", " ", C.rev)

  out<-cbind(rownames(MT), C.rev)
  write.table(file=paste0(colnames(MT[i]), ".char"), out, quote=F, sep=" ",
              row.names=F, col.names=F)
}

# write Rev file for the two-state characters
setwd("~/Documents/Recon-Anc_Anat/Supplementary_materials/STEP_4/RevBayes/")

# For constructing .Rev files we use the procooked template "PARAMO2_templ.Rev"
fl.in  <- readLines("PARAMO2_templ.Rev")

for (i in 1:ncol(MT))
{
  fl.in  <- readLines("PARAMO2_templ.Rev")
  fl.in  <- gsub(pattern = "@analysis_name@", replace = paste0(colnames(MT[i])),
                 x = fl.in)
  fl.in <- gsub(pattern = "@chrs_2_read@", replace = paste0("data/", colnames(MT[i]), ".char"), x = fl.

  cat(file=paste0(colnames(MT[i]), ".Rev"), sep="\n", fl.in)

}

# write Rev file for dependent four-state character C3-2
setwd("~/Documents/Recon-Anc_Anat/Supplementary_materials")
```

```
# I use precooked set of functions for constracting SMM from Tarasov (2019)
source("STEP_4/SMM_functions.R")

###################################
# same SMMs as for the tail color problem
###################################
char.state<-c("a", "p")
rate.param<-c(1, 1)
TL<-init_char_matrix(char.state, rate.param, diag.as=0)
char.state<-c("r", "b")
rate.param<-c(1, 1)
COL<-init_char_matrix(char.state, rate.param, diag.as=0)

#SMM-ind
TC.ind<-comb2matrices(TL, COL, controlling.state=NULL, name.sep="", diag.as="")
TC.ind
in.rev<-Mk_Rev(TC.ind)
cat(in.rev) # COPY the output and insert in Rev template PARAMO2_templ.Rev
#cat(in.rev, file="STEP_4/input_Rev.txt") # or save this outout to a file and then copy to Rev template
```

Now having created the files for the inference, we run RevBayes. Each RevBayes output consists of four files: log file, ancestral character state reconstruction (asr), and stochastim maps (stm).

STEP 5. Ontology-informed amalgamation of the stochastic maps

Our goal is to construct the amalgamated characters for the AD, BR and EF levels of anatomical hierarchy. The AD level exibits the individual stochastic maps obtained at the previous step. At this step, we will constrauct characters for BR and EF levels using ontology-informed amalgamatiion of the stochastic maps. At the BR level, three main body regions are considered – "head", "legs" and "wings".

Before starting ontology-informed amalgamation of characters, let us first fix potential issues with data maipulation. To amalgamate maps, they first have to be descritezed – each tree branch is split into small bins, whereas each bin indicates the state of a charatcer. This dicretization facilitates character amalgamation but may subtantially increseas memory usage in `R` if map samples and trees are large. To make the computations efficient, we put each stochastic map in a seprate `.rds` file, then we put all of these files belonging to the same character into a seprate zip archive. This trick avoids use of bunch of separate files and, at the same time, allows getting access to the individual maps upon demand.

```
library("phytools")
# we use a set of precooked functions to work with stoch. maps
source("R_PARAMO/Functions_Discr_maps.R")
# let's make character list

c=paste0("C", AN$CHAR_ID2)
c<-sub(",", "-", c  )

# dir to write and read files
dirW= ("STEP_5/Discr_maps/")
dirR= ("STEP_4/RevBayes/output/")

#####################################
# Read a sample of 100 maps from .stm files and save them in the poper format .stmR
#####################################
```

```
for (i in 1:length(c))
{
  tree<-read_Simmap_Rev(paste0(dirR, c[i], ".stm"),
                        start=400, end=500,
                        save = NULL) %>% read.simmap(text=., format="phylip")


  write.simmap(tree, file=paste0(dirW, c[i], ".stmR"))
}
##########

#####################################
# Read stmR, discretize maps, and save each map as a separate rds file;
# in turn all rds file for a chracter are stored
# in a zip archive
#####################################

for (i in 1:length(c))
  {
  # read in undesritezed trees
  print(paste0("Reading ", c[i]))
  sim=read.simmap(file=paste0(dirW, c[i], ".stmR"), format="phylip")

  # descritize trees by looping over sample and saving as rds

  for (j in 1:length(sim)){
    tryCatch({

      print(paste0("Descritizing tree ", j))

      ## errors with na

      ##

      ##### make trees equal with template
      sim.d<-make_tree_eq(tree.tmp.final, sim[[j]], round=5)
      ###

      #sim.d<-discr_Simmap_all(sim[[j]], 1000)
      sim.d<-discr_Simmap_all(sim.d, 1000)

      saveRDS(sim.d, file =  paste0(dirW,c[i], "_", j, ".rds") )

    }, error=function(e){
      cat("ERROR :",conditionMessage(e), "\n")
      #errors<-rbind(errors, c(ii,jj))
    }  )

  }

  # putting rds files into archive
  files<-paste0(dirW, c[i], "_", c(1:length(sim)), ".rds")
  zip(paste0(dirW, c[i], ".zip"), files=files)
```

```
  file.remove(files)

}

# close connections
 showConnections (all=T)
 closeAllConnections()
#########################
```

Now having the stochastic maps converted to the proper format, we can start their ontology-guided amalgamation.

```
source("R_PARAMO/Functions_Stack_maps.R")

# dir to write and read files
dirW= ("STEP_5/Discr_maps/")
dirR= ("STEP_4/RevBayes/output/")
#############
# Level 2 stacks - Body regions
#############
level2
cc<-lapply(L2, function(x) sub("CHAR:", "C", x) )
cc<-lapply(cc, function(x) sub(",", "-", x) )


L2.maps<-vector("list", length(L2))
names(L2.maps)<-names(L2)

# batch stacking

for (i in 1:length(L2.maps))
{
  map<-paramo(cc[[i]], ntrees=1, dirW=dirW)
  L2.maps[[i]]<-map
}

#############
# Level 3 stacks - Entire phenotype
#############
level3
cc3<-lapply(L3, function(x) sub("CHAR:", "C", x) )
cc3<-lapply(cc3, function(x) sub(",", "-", x) )


L3.maps<-vector("list", length(L3))
names(L3.maps)<-names(L3)

# batch stacking
for (i in 1:length(L3.maps))
{
  map<-paramo(cc3[[i]], ntrees=10, dirW=dirW)
  L3.maps[[i]]<-map
}
```

```
library("phytools")
```

```
## Loading required package: ape
```

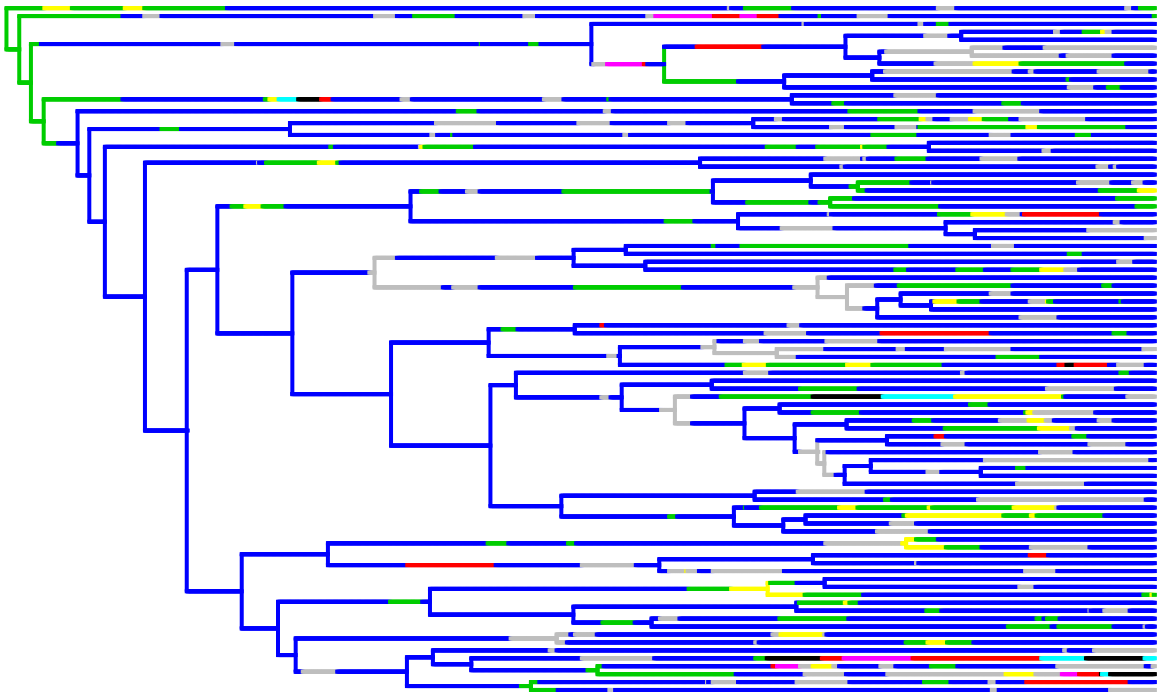
```
## Loading required package: maps
```

```
#########
# BR level
########

# plot one stochastic maps for the head character
plotSimmap(L2.maps$head[[1]], pts=F,ftype="off", ylim=c(0,100) )
```

```
## no colors provided. using the following legend:
##        00        01        02        03        10        11        12
##   "black"     "red"  "green3"    "blue"    "cyan" "magenta"  "yellow"
##        13
##    "gray"
```

```
title("\n Head character")
```

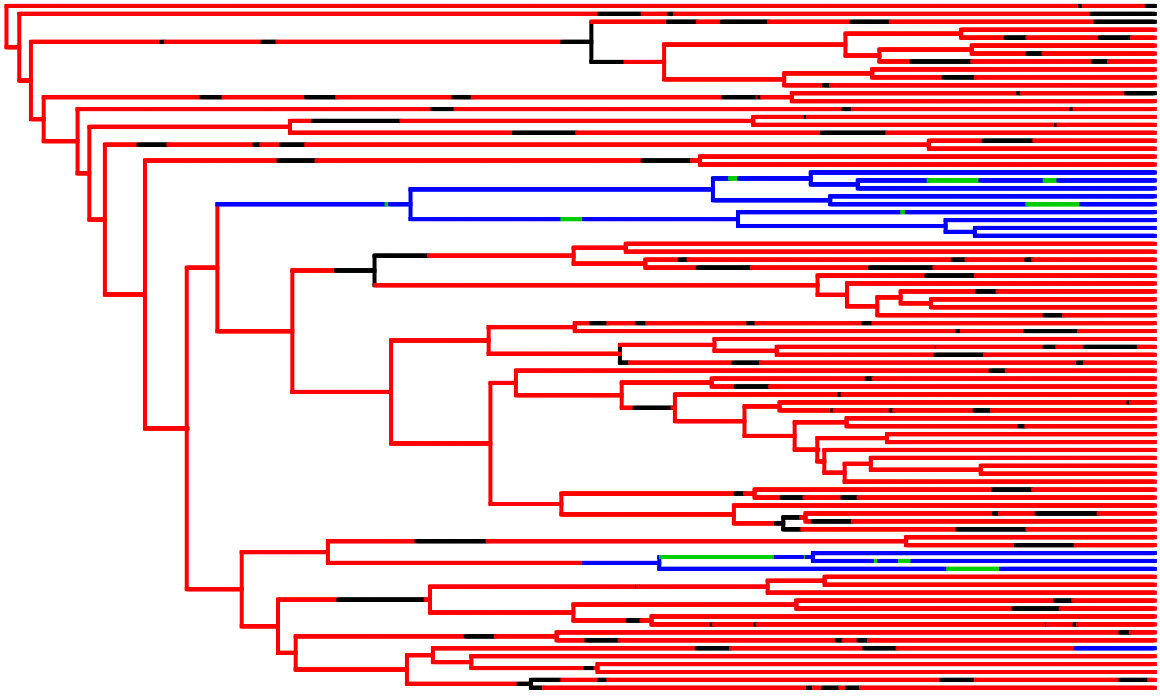



```
# plot one stochastic maps for the wings character
plotSimmap(L2.maps$wings[[1]], pts=F,ftype="off", ylim=c(0,100) )
```

```
## no colors provided. using the following legend:
##       00       01       10       11
##  "black"    "red" "green3"   "blue"
```

```
title("\n Wings character")
```

Wings character

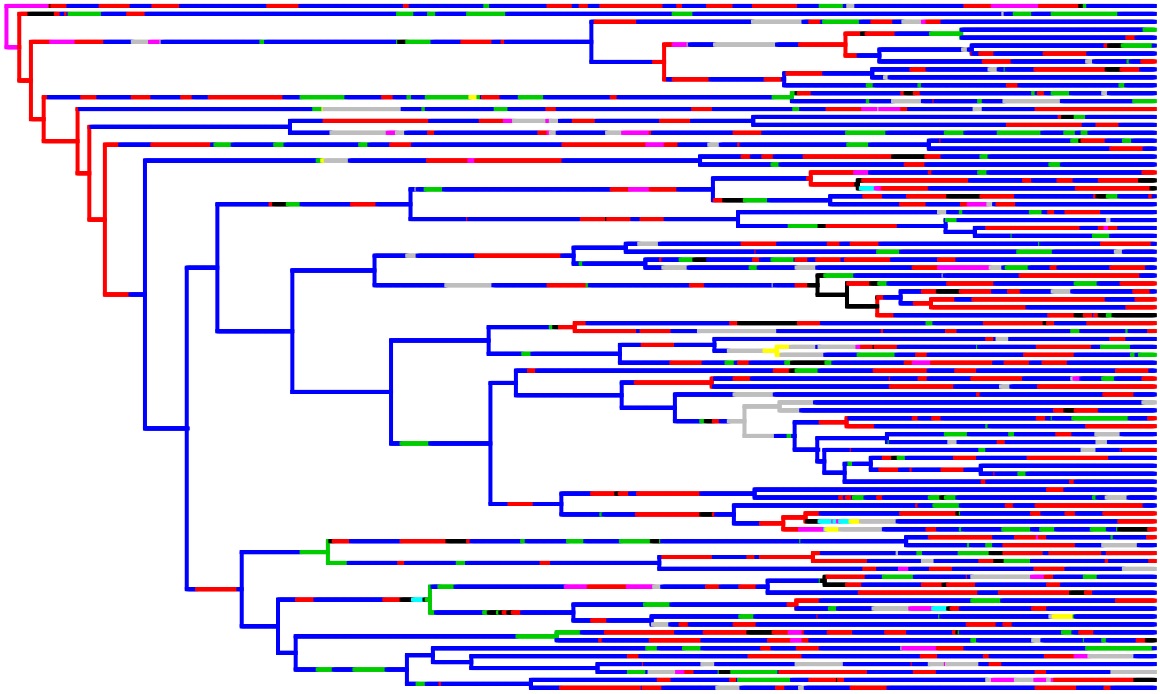


```
# plot one stochastic maps for the legs character
plotSimmap(L2.maps$legs[[1]], pts=F,ftype="off", ylim=c(0,100) )
```

```
## no colors provided. using the following legend:
##       000       001       010       011       100       101       110
##   "black"     "red"  "green3"    "blue"    "cyan" "magenta"  "yellow"
##       111
##    "gray"
```

```
title("\n Legs character")
```

Legs character



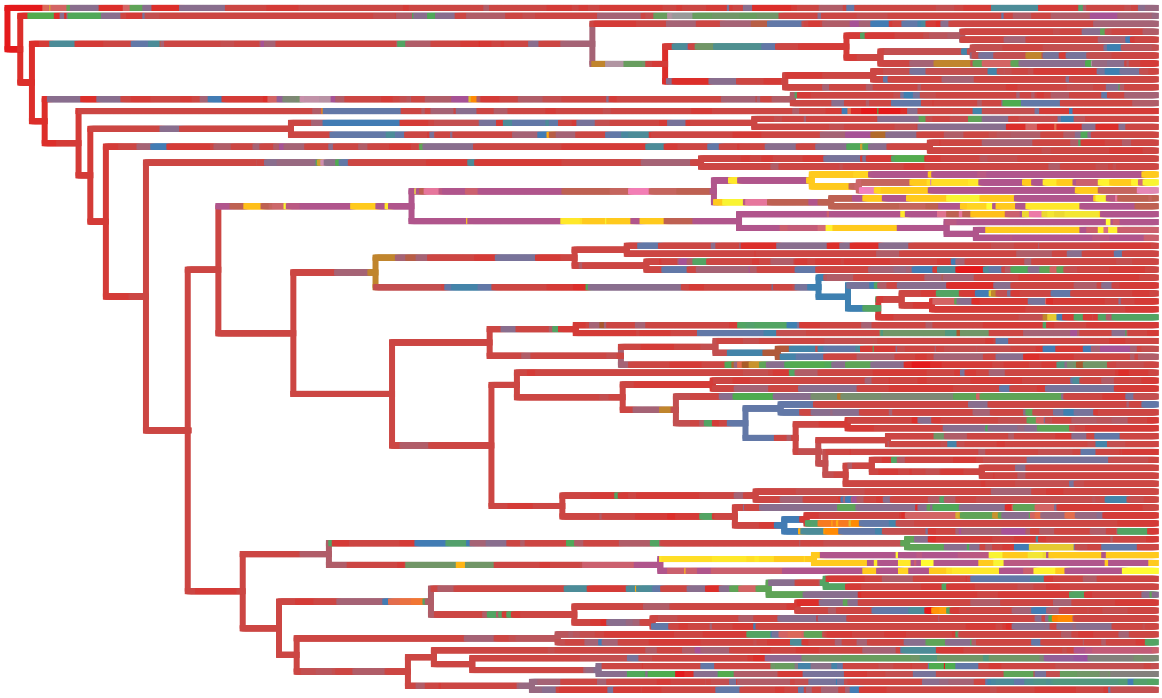
```
#########
# EF level
#########

# plot one stochastic maps for the entire phenotype character
# first, let's define color pallette for the characters sonce it contains many states
library("RColorBrewer")
tmm<-L3.maps$whole_organism[[1]]
lapply(tmm$maps, names) %>% unlist %>% unique->states
# number of states in the character
#length(states)

hm.palette <- colorRampPalette(brewer.pal(9, 'Set1'), space='Lab')
color<-hm.palette(length(states))

plotSimmap(tmm, setNames(color, states),  lwd=3, pts=F,ftype="off", ylim=c(0,100))
title("\n Entire Phenotype character")
```

Entire Phenotype character



References

Klopfstein, Seraina, Lars Vilhelmsen, John M Heraty, Michael Sharkey, and Fredrik Ronquist. 2013. "The Hymenopteran Tree of Life: Evidence from Protein-Coding Genes and Objectively Aligned Ribosomal Data." *PLoS One* 8 (8). Public Library of Science: e69344.

Sharkey, Michael J, James M Carpenter, Lars Vilhelmsen, John Heraty, Johan Liljeblad, Ashley PG Dowling, Susanne Schulmeister, et al. 2012. "Phylogenetic Relationships Among Superfamilies of Hymenoptera." *Cladistics* 28 (1). Wiley Online Library: 80–112.

Yoder, Matthew J, Istvan Miko, Katja C Seltmann, Matthew A Bertone, and Andrew R Deans. 2010. "A Gross Anatomy Ontology for Hymenoptera." *PloS One* 5 (12). Public Library of Science: e15991.