rPlant

Barb Banbury, University of Tennessee, darwinthesun@gmail.com
Kurt Michels, University of Arizona, kamichels@math.arizona.edu

June 27, 2014

Contents

1 Introduction

The iPlant Collaborative has developed many resources to deal with the emerging computational challenges facing biology. The project was initially designed to support the plant sciences, but thanks to a generic approach, can be equally used by other disciplines. Users have access to many different applications for data analysis, including clustering/network analyses, QTL mapping, sequence alignments, phylogenetic tree building, and comparative methods.

The main interface is its user-friendly Discovery Environment (`http://www.iplantcollaborative.org/discover/discovery-environment`). A command-line interface, the agave API (`http://agaveapi.co/`, (Dooley et al. 2012)), is linked to the Discovery environment. The Agave API is used for computationally intensive applications. The API is a RESTful application programming interface (Fielding 2000) that allows direct interaction with all of iPlant resources. The only way to access the API is to use *cURL* statements (Stenberg 1996), an example of a *cURL* statement will be detailed in one of the sections. The API provides access to authentication, data manipulation and storage, and job submittal via HTTPS- and command-line functions. The benefit of using the API is having programmatic access that allows advantages to power users (e.g. submitting jobs via batch files). The *rPlant* package provides a direct link between high performance resources located at the Texas Advanced Computing Center (`http://www.tacc.utexas.edu`) that the API can access and the R environment, by essentially creating wrappers around the *cURL* statements, using the *RCurl* (Lang 2007) package.

2 Getting Started

This vignette assumes you have the current version of R (R Core Team 2012). First, install and load the package. A stable release is available through CRAN (`http://cran.r-project.org/web/packages/rPlant/`) or a working repository can also be used through R-Forge (`https://r-forge.r-project.org/projects/rplant/`).

You can register as an iPlant user on their website (`http://user.iplantcollaborative.org/`) generating a unique username and password combination.

2.1 Gaining Access to the API

```
Validate(username, password, api="agave")
```

The username/password combination will be used in the `Validate` function. The `Validate` function is required for every rPlant session and needs to be the first thing executed or the session will fail. In addition, it has a four-hour expiration. *rPlant* functions will auto renew a session, thereby extending the expiration. If a session sits idle and expires, the user will not have to re-validate a session as the functions will do this automatically. The only time a user will need to use the `Validate` function is at the start of a new session.

```
> require(rPlant)
> username <- "enter your username"
> password <- "enter your secret password"
> Validate(username, password, api="agave")
```

The function checks if the username and password are valid iPlant credentials. If they aren't the above error is displayed. If the function is successful then nothing is printed. On *rPlant's* backend, a new R environment (rPlant.env) was created that stores all of the validation objects, including the user key and secret, user name and password, and token expiration. These items can be examined by using the `ls(rPlant.env)` and using the $ operator to display individual objects.

Every *rPlant* function has the option `print.curl=TRUE` or `FALSE`. This refers to `cURL` a computer software project providing a way to transfer data using various protocols, for detail on `cURL` see `http://en.wikipedia.org/wiki/CURL`. These statements (w/o the outside quotes) can be copied and pasted into a terminal in linux or unix. And if `cURL` is installed on the computer then the statements can be executed. You will see that these statements do the exact same thing as the *rPlant* functions. This is one of the big advantages of *rPlant*, it can be used on any computer (including windows) and there is no need for the user to install `cURL` on that computer, because *rPlant* uses the package `RCurl`.

Note: This package abides by the unix rule, "silence is golden". If a function is successful then no output will be displayed. If an error is attained then the error will be printed.

3 Uploading Files

3.1 UploadFile function

```
UploadFile(local.file.name, local.file.path="", filetype=NULL, print.curl=FALSE,
suppress.Warnings=FALSE)
```

The first step is to upload files onto iPlants. `UploadFile` takes a file from your computer and uploads it onto iPlants servers, it does NOT take a file from the R workspace. Objects in the workspace will need to be saved to the computer in a supported file format before uploading them.

```
> data(DNA.fasta)
> write.fasta(sequences = DNA.fasta, names = names(DNA.fasta), file.out = "DNA.fasta")
> UploadFile(local.file.name="DNA.fasta")
```

In the event that the file already exists on the iPlant server, an error will report and the file will not upload. For details on how to check contents of iPlant directories, see Section 4, and for manipulating files (like deleting, moving, sharing) see Section 5 below.
The file format for the uploaded file can be defined using `filetype`, some programs will only accept certain types of files. This can also be left as NULL, if the iPlant application doesn't require it. For the fasta file the file type is FASTA-0. See the following section for the various file types supported by the API.

3.2 Supported File Types

```
SupportFile(print.curl=FALSE)
```

There are 33 other file types supported by iPlant, use the `SupportFile` function to see all of the available file types (i.e. PHYLIP file type is "PHYLIP-0" and ClustalW is "ClustalW-1.8").

```
> SupportFile()
```

```
 [1] "2bit-0"           "ASN-0"            "BAM-0.1.2"        "Barcode-0"
 [5] "BED-0"            "BlastN-2.0"       "Bowtie-0"         "BZIP2-0"
 [9] "CEL-3"            "ClustalW-1.8"     "CSV-0"            "DOT-0"
[13] "EMBL-0"           "EXPR-0"           "FAI-0"            "FASTA-0"
[17] "FASTQ-Illumina-0" "FASTQ-Int-0"      "FASTQ-Solexa-0"   "FASTQ-0"
[21] "Genbank-0"        "GFF-2.0"          "GFF-3.0"          "GFF-3.0"
[25] "GraphML-0"        "GTF-2.2"          "HTML-4"           "HTML-5"
[29] "Newick-0"         "NEXUS-0"          "PAIR-0"           "PDB-3.2"
[33] "Phylip-0"         "PhyloXML-1.10"    "Pileup-0"         "SAI-0.1.2"
[37] "SAM-0.1.2"        "SBML-1.2"         "SBML-2.4.1"       "SBML-3.1"
[41] "Soap-PE-1"        "Soap-SE-1"        "Stockholm-1.0"    "TAB-0"
[45] "TAR-0"            "Text-0"           "VCF-3.3"          "VCF-4.0"
[49] "WIG-0"
```

4 Manipulating directories on iPlant servers

Now that the file DNA.fasta has been uploaded onto the iPlant servers we can look at the file (or at least see which directory the file is in) by using the `ListDir` function. There are a few other directory manipulation functions, they are: `MakeDir`, `ShareDir`, `PermissionsDir`, `RenameDir`, `MoveDir` and `DeleteDir`.

4.1 Listing directories

```
ListDir(dir.name, dir.path="", print.curl=FALSE, shared.username=NULL,
suppress.Warnings=FALSE)
```

Looking in the home directory you can see the "DNA.fasta" file.

```
> ListDir(dir.name="", suppress.Warnings=TRUE)
```

```
     name        type
[1,] "analyses"  "dir"
[2,] "DNA.fasta" "file"
```

Note: Some functions contain an option, `suppress.Warnings`. If you are absolutely sure that the commands you are entering are correct then to speed up the process have `suppress.Warnings=TRUE`. But be careful, if used inappropriately then files could get overwritten or the files might not even exist and you will get no warning about it.

4.2 Making directories

```
MakeDir(dir.name, dir.path="", print.curl=FALSE, suppress.Warnings=FALSE)
```

The following function will make a directory hello in the home directory.

```
> MakeDir(dir.name="hello")
```

Again making the directoy "all" in the "hello" directory.

```
> MakeDir(dir.name="all", dir.path="hello")
```

Here we make another directory "robots" in the "all" directory. Note how the `dir.path` needs to be constructed, and how the `dir.name` and `dir.path` are related. All of the functions have this same format.

```
> MakeDir(dir.name="robots", dir.path="hello/all")
```

We can look inside the "hello/all/robots" directory and see that there is nothing in there.

```
> ListDir(dir.name="robots", dir.path="hello/all", show.hidden=TRUE)
```

```
     name type
```

4.3 Sharing Directories

```
ShareDir(dir.name, dir.path="", shared.username, read=TRUE, execute=TRUE,
print.curl=FALSE, suppress.Warnings=FALSE)
```

A really nice feature of *iPlant* is the file sharing feature. As was said in the introduction one of iPlant's goals was to work with very large data sets. And when data sets are too large to send via email then a sharing feature is absolutely necessary. There are in fact two share functions, one for sharing a single file (`ShareFile`) and the other (`ShareDir`) for sharing an entire directory. When sharing a directory, all files contained within will be shared.

In this sample we share the "all" directory.

```
> ShareDir(dir.name="all", dir.path="hello", shared.username="kamichels")
```

In the above example I share something with user "henryl". I can also view files and directories that have been shared with me.

```
> ListDir(dir.name="data", dir.path="", shared.username="phyllisl")
```

```
     name            type
[1,] "muscle3.fasta" "file"
```

Several other functions use the `shared.username` option. For example, the `SubmitJob` function and the wrapper functions. A job can be run on files that are shared with you (see Section 7).

4.4 Checking Permissions on Directories

```
PermissionsDir(dir.name, dir.path="", print.curl=FALSE, suppress.Warnings=FALSE)
```

When using the sharing feature of rPlant, a user needs a function which will check the permissions of a directory. This includes the users with whom the directory is shared and what the permissions are included. Below confirms the sharing performed earlier with "kamichels".

```
> PermissionsDir(dir.name="all", dir.path="hello")

     Name  Username    Permissions
[1,] "all" "kamichels" "All"
```

4.5 Renaming Directories

```
RenameDir(dir.name, new.dir.name, dir.path="", print.curl=FALSE,
suppress.Warnings=FALSE)
```

This function renames a directory.

```
> RenameDir("robots", "tools", "hello/all")
```

And you can see that it has been changed.

```
> ListDir("all", "hello")

     name    type
[1,] "tools" "dir"
```

Note: When the directory is renamed, permissions on sharing will have to be redone.

4.6 Moving Directories

```
MoveDir(dir.name, dir.path="", end.path="", print.curl=FALSE,
suppress.Warnings=FALSE)
```

This function moves a directory. The following code will move the directory 'tools' from 'hello/all' to the home directory. Verified below.

```
> MoveDir("tools", "hello/all", end.path="")
> ListDir("")

     name        type
[1,] "analyses"  "dir"
[2,] "hello"     "dir"
[3,] "tools"     "dir"
[4,] "DNA.fasta" "file"
```

Note: When the directory is moved, permissions on sharing will have to be redone.

4.7 Deleting Directories

`DeleteDir(dir.name, dir.path="", print.curl=FALSE, suppress.Warnings=FALSE)`

This function deletes a directory and all of the subdirectories.

```
> DeleteDir("tools")
> ListDir("")
```

```
     name        type
[1,] "analyses"  "dir"
[2,] "hello"     "dir"
[3,] "DNA.fasta" "file"
```

5 Manipulating files on iPlant servers

The file manipulation tools available in this package are very similar to directory manipulation tools. The file manipulation functions are: `ShareFile`, `PermissionsFile`, `RenameFile`, `MoveFile` and `DeleteFile`.

5.1 Sharing Files

`ShareFile(file.name, file.path="", shared.username, read=TRUE, execute=TRUE, print.curl=FALSE, suppress.Warnings=FALSE)`

As described in the `ShareDir` function, a really nice feature of iPlant is the file sharing feature. This function shares one file at a time.

The following code will share the file DNA.fasta with phyllisl.

```
> ShareFile(file.name="DNA.fasta", shared.username="phyllisl")
```

5.2 Checking Permissions on a File

`PermissionsFile(file.name, file.path="", print.curl=FALSE, suppress.Warnings=FALSE)`

When using the sharing feature of rPlant, a user needs a function which will check the permissions of a file. This includes the users with whom the file is shared and what the permissions are included. Below confirms the sharing performed earlier with "phyllisl".

```
> PermissionsFile(file.name="DNA.fasta")
```

```
     Name        Username   Permissions
[1,] "DNA.fasta" "phyllisl" "All"
```

5.3 Moving Files

`MoveFile(file.name, file.path="", end.path="", print.curl=FALSE, suppress.Warnings=FALSE)`

This function moves the file from one directory to another.

```
> MoveFile("DNA.fasta", end.path="hello/all")
> ListDir("all", "hello")
```

```
     name        type
[1,] "DNA.fasta" "file"
```

The move took the file DNA.fasta from the home directory into the "hello/all" directory.

Note: When the file is moved, permissions on sharing will have to be redone.

5.4 Renaming Files

```
RenameFile(file.name, new.file.name, file.path="", print.curl=FALSE,
suppress.Warnings=FALSE)
```

This function renames a single file.

```
> RenameFile("DNA.fasta", "lp.fasta", "hello/all")
> ListDir("all", "hello")

     name       type
[1,] "lp.fasta" "file"
```

Note: When the file is renamed, permissions on sharing will have to be redone.

5.5 Deleting Files

```
DeleteFile(file.name, file.path="", print.curl=FALSE, suppress.Warnings=FALSE)
```

This function deletes a single file in the specified directory.

```
> DeleteFile("lp.fasta", "hello/all")
> ListDir("all", "hello")

     name type
```

The file lp.fasta is no longer in the "hello/all" directory.

6 Applications

The real power in the *rPlant* package is to have the programmatic access to the phylogenetic tools/applications that are available in the API. *rPlant* can be used to interact with any of the API applications.

6.1 Listing Applications

```
ListApps(description=FALSE, print.curl=FALSE)
```

This function returns a sorted list of the newest versions of the public applications that are available via the Foundation API. These applications are ones that can be used with the `SubmitJob` function (see Section 7).

```
> ListApps(description=FALSE)
```

```
 [1] "Adjustp-0.0.1u1"
 [2] "AllpathsLG_lonestar-44837u1"
 [3] "bismark-0.7.4u1"
 [4] "bismark_genome_preparation-0.7.4u1"
 [5] "bismark_methylation_extractor-0.7.4u1"
 [6] "blastx-stampede-ncbi-db-2.2.26u3"
 [7] "buildicm-1.0.0u1"
 [8] "bwa-lonestar-0.5.9u3"
 [9] "ClustalW2-2.1u1"
[10] "clustalw2Dispatcher-1.0.13100u1"
[11] "clustalw2-lonestar-2.1u2"
[12] "diginorm-1.0.0u1"
[13] "dnalc-cuffdiff-lonestar-2.1.1u3"
[14] "dnalc-cuffdiff-stampede-2.1.1u3"
[15] "dnalc-cuffdiff-test-2.1.1u1"
[16] "dnalc-cufflinks-lonestar-2.1.1u2"
[17] "dnalc-cufflinks-stampede-2.1.1.1u1"
[18] "dnalc-cufflinks-stampede-2.1.1u2"
[19] "dnalc-cuffmerge-lonestar-2.1.1u1"
[20] "dnalc-cuffmerge-stampede-2.1.1u1"
[21] "dnalc-fastqc-lonestar-0.10.1u1"
[22] "dnalc-fastqc-stampede-0.10.1u1"
[23] "dnalc-fastx-lonestar-0.0.13.2u1"
[24] "dnalc-fastx-stampede-0.0.13.2u2"
[25] "dnalc-fxtrim-lonestar-0.0.13.2u1"
[26] "dnalc-fxtrim-lonestar-0.0.13.3u1"
[27] "dnalc-fxtrim-stampede-0.0.13.2u1"
[28] "dnalc-fxtrim-stampede-0.0.13.3u1"
[29] "dnalc-tophat-lonestar-2.0.8u2"
[30] "dnalc-tophat-stampede-2.0.11.1u1"
[31] "dnalc-tophat-stampede-2.0.11u1"
[32] "dnalc-tophat-stampede-2.0.8.4u1"
[33] "dnalc-tophat-stampede-2.0.8u2"
[34] "EMMAX-0.0.1u1"
[35] "EMMAX2-0.0.1u1"
[36] "FaST-LMM-1.09u1"
[37] "fasttreeDispatcher-1.0.0u1"
[38] "forward-regression-0.0.1u1"
[39] "gapcloser-1.12u1"
[40] "gatk-1000bulls-geno-lonestar-1.00u1"
[41] "GeneSeqer-5.0u2"
[42] "GenomeSelection-0.0.1u2"
[43] "glimmer-1.0.0u1"
[44] "GMAP_stampede-121212u1"
[45] "GSNAP_lonestar-121212u1"
[46] "GSNAP_stampede-121212u2"
[47] "gth-lonestar-1.0u1"
[48] "head-stampede-5.97u2"
[49] "idbaUD-1.0.0u2"
[50] "interproscan-5.44.0u1"
[51] "macs-ranger-1.4-1.4u1"
[52] "mafft-7.113u1"
[53] "mafftDispatcher-1.0.13100u1"
[54] "mafft-lonestar-6.864u1"
[55] "MergeG2P-0.0.1u1"
[56] "metagenemark-1.00u3"
[57] "metaphlan-lonestar-1.6.0u4"
[58] "metavelvet-1.0.0u1"
[59] "mga-1.0.0u1"
```

```
[60] "MLMM-0.0.1u1"
[61] "MrBayesmpi_basic-3.2.1u1"
[62] "Muscle-3.8.31"
[63] "Muscle-3.8.32u4"
[64] "muscle-lonestar-3.8.31u2"
[65] "newbler-2.6.0u1"
[66] "NPUTE-0.0.2u1"
[67] "NumericalTransform-0.0.1u1"
[68] "oases-0.2.08u1"
[69] "phylip-dna-parsimony-lonestar-3.69u2"
[70] "phylip-protein-parsimony-lonestar-3.69u2"
[71] "plink-1.07u1"
[72] "prodigal-1.0.0u1"
[73] "quicktree-dm-lonestar-1.1u2"
[74] "quicktree-tree-lonestar-1.1u2"
[75] "raxml-lonestar-7.2.8u1"
[76] "ray-2.2.0u1"
[77] "scarf-1.00u1"
[78] "soapdenovo-1.05u1"
[79] "soapdenovo-2.04u1"
[80] "soapdenovo_trans-1.0u1"
[81] "spa-1.0.0u1"
[82] "STRUCTURE-2.3.4u3"
[83] "STRUCTURE-2.3.5u1"
[84] "STRUCTURE2TASSEL-0.0.1u1"
[85] "TASSEL4-GLM-0.0.1u1"
[86] "tasselDispatcher-1.0.13350u1"
[87] "TASSEL(GLM)-0.0.1u1"
[88] "TASSEL(MLM)-0.0.1u1"
[89] "TNRS4GWAS-0.0.2u1"
```

Note: As said above, these applications are PUBLIC applications. Applications in the API are split into two categories, public and private. Private applications are ones that are developed and tested and changed. Only the user who created the private application can use it. The other category is public applications. After a private application has gone through extensive testing, then the application can be published and it becomes a public application which is available to all iPlant users. In the API a public application is labeled by adding the suffix 'u1' to it. The '1' is referred to as the version number, so if a public application is fixed and republished the suffix becomes 'u2'.

6.2 Application Information

`GetAppInfo(application, return.json=FALSE, print.curl=FALSE)`

The `GetAppInfo` function returns the application with a short description and the input/output filetypes.

```
> GetAppInfo("velveth-1.2.07u1")

$Description
[1] "Genome assembler for short sequencing reads, first stage."

$Application
[1] "velveth-1.2.07u1" "Public App"       "Newest Version"

$Information
      kind     id               fileType/value
```

```
 [1,] "input"  "reads6"         "fasta-0"
 [2,] "input"  "reads3"         "fasta-0"
 [3,] "input"  "reads5"         "fasta-0"
 [4,] "input"  "reads2"         "fasta-0"
 [5,] "input"  "reads4"         "fasta-0"
 [6,] "input"  "reads1"         "fasta-0"
 [7,] "output" "format1"        "string"
 [8,] "output" "format5"        "string"
 [9,] "output" "format2"        "string"
[10,] "output" "strandSpecific" "string"
[11,] "output" "kmer"           "string"
[12,] "output" "Output"         "string"
[13,] "output" "format4"        "string"
[14,] "output" "format3"        "string"
      details
 [1,] "Sequences:"
 [2,] "Sequences:"
 [3,] "Sequences:"
 [4,] "Sequences:"
 [5,] "Sequences:"
 [6,] "Sequences:"
 [7,] "sequence file format, library 1"
 [8,] "sequence file format, library 5"
 [9,] "sequence file format, library 2"
[10,] "strand specific"
[11,] "kmer size"
[12,] "Name for output directory"
[13,] "sequence file format, library 4"
[14,] "sequence file format, library 3"
```

The `GetAppInfo` function returns a list of information about the application that is needed for the `SubmitJob` function. The first element gives a short description of the application. The second element in the list gives information on the application including its use permissions (public vs. private) and whether it is the newest version. The third element in the list is a matrix with four columns of information: kind, id, file type or value, and any details. In the example above, first column ('kind') states there are six inputs for this app, the 'id' column names those inputs as 'reads5', 'reads3', etc. There are also eight parameters for the app, such as 'format2', 'kmer', etc. The third column ('Type/value') returns the type of file the application is expecting if it is input or it returns the type of input necessary for the application parameters, common ones are string, boolean, etc. The last column gives brief details on each input.

7 Submitting Jobs in the rPlant package

7.1 Submitting Job

```
SubmitJob(application, file.path="", file.list=NULL, input.list, args.list=NULL,
job.name, nprocs=1, private.APP=FALSE, suppress.Warnings=FALSE, shared.username=NULL,
print.curl=FALSE, email=TRUE)
```

An important benefit of using *rPlant* is the ability to create batch-scripted files that automate job submittal and retrieval. For example, a user could submit parallel alignment jobs of different gene regions or multiple jobs with the same data and different parameter values. The results could then be automatically downloaded upon completion.

The following function is the main way to submit a job to the iPlant server, and can be used for any iPlant application. You can also submit jobs via the wrapper functions (for example, `Muscle()`, see Section 9), which call upon the `SubmitJob` function internally.

```
> UploadFile(local.file.name="DNA.fasta", filetype="FASTA-0")
> ListDir("")
> myJobM <- SubmitJob(application="Muscle-3.8.32u4", file.list=list("DNA.fasta"),
+                     input.list=list("stdin"), args.list=list(c("arguments",
+                     "-phyiout -center -cluster1 upgma")), job.name="Muscle")

     name        type
[1,] "analyses"  "dir"
[2,] "hello"     "dir"
[3,] "DNA.fasta" "file"

Job submitted.
You can check your job using CheckJobStatus(55041)
```

Several important argument definitions are listed below, but can also be found in the help files: `input.list`: This argument defines what kind of input you are passing the application. You can get application information from the `GetAppInfo` function (`GetAppInfo("Muscle-3.8.32u4")$Information`). In this example, the 'kind' column states there is one input for this app, and the 'id' column names that input as 'stdin'. Input types change from application to application.

`file.list`: Similar to `input.list`, the `file.list` argument defines which files are being passed to the application. The named file must be on the DE within the `file.path` and be formatted to the correct specification (for example, `GetAppInfo("Muscle-3.8.32u4")$Information`). If it the file types don't match then the application will fail.

`args.list`: The `args.list` is where application flagging options can be entered. These typically change default options. Using information from the `GetAppInfo` function, the 'kind' column states there is one parameter for this app, the 'id' column gives the name of that parameter is "arguments", and the "fileType/value" column tells me it is a string. This is where the fourth column 'details' comes in handy; it tells me that the parameter input is "program arguments and options", which means it can accept a string. For example: `args.list=list(c(arguments, "-phyiout -center -cluster1 upgma"))`

The `args.list` is a list that is as long as the number of parameters (so length 1 in the Muscle example), that means there as many vectors as there are parameters. All vectors are of the same length, two in this example. In the first position, is the name of the parameter, "arguments", and in the second position is the value of that parameter, "-phyiout -center -cluster1 upgma". In this example, it is a string of command line flags.

The function `SubmitJob` will return a list of two objects (in this example, `myJobM`). The first object, is the job number and the second is the job name, both are important information for retrieving results.

If job submittal is successful, then the function automatically creates the folder "analyses" within a user's cloud (if it did not previously exist). If the job finishes, then a folder is created within the analyses folder that is named the job name.

7.2 Submitting a job with a shared file

Jobs can also be submitted from files stored in other user's cloudspace that are shared. In the below example, a file that had been previously shared (Section 3), a job can be submitted using that file.

```
> myJobS <- SubmitJob(application="Muscle-3.8.32u4", file.list=list("muscle3.fasta"),
+                     file.path="data", shared.username="phyllisl",
+                     args.list=list(c("arguments", "-fastaout")),
+                     input.list=list("stdin"), job.name="MuscleShare")

Job submitted.
You can check your job using CheckJobStatus(55042)
```

8 Checking Job Status and Retrieving Job output

Once the job is submitted, it is assigned a job identification number (job.id) that we can use to check the status and download any results files. The job.id is returned with the SubmitJob function, so if you create an object when you submit a job then you can use that object as an identifier as well. Otherwise, you can copy/paste a job ID into any of the functions as a character string. The job number is used in a few rPlant functions including, `CheckJobStatus`, `KillJob`, `ListJobOutput`, `RetrieveJob` and `DeleteJob`. If you need to get job IDs from older jobs, you can use the function `GetJobHistory()`.

8.1 Checking job status

```
CheckJobStatus(job.id, print.curl=FALSE)
```

This function checks the status of a job on the iPlant servers.

Table 1: Possible Outputs for `CheckJobStatus()`

Stages
PENDING
STAGING_INPUTS
CLEANING_UP
ARCHIVING
STAGING_JOB
FINISHED
KILLED
FAILED
STOPPED
RUNNING
PAUSED
QUEUED
SUBMITTING
STAGED
PROCESSING_INPUTS
ARCHIVING_FINISHED
ARCHIVING_FAILED

```
> CheckJobStatus(myJobS[[1]])
```

```
[1] "PENDING"
```

8.2 Killing Job

If you don't want a job to run any more, then simply stop it by using the `KillJob` function. Generally this is used if a job is `QUEUED`, it must be stopped first to delete it.

```
> KillJob(myJobM[[1]])
```

8.3 Listing job status

```
ListJobOutput(job.id, print.curl=FALSE, print.total=TRUE)
```

This function lists the output files from a finished job. For example, these files are the output from our MUSCLE example above.

```
> ListJobOutput(myJobS[[1]])
```

```
[1] "fasta.aln"
[2] "muscle3.fasta"
[3] "muscleshare_2014-06-27_17-04-45955-55042.err"
[4] "muscleshare_2014-06-27_17-04-45955-55042.out"
```

8.4 Looking at Job History

```
GetJobHistory(return.json=FALSE, print.curl=FALSE)
```

This function displays the entire job history for the user. This is an easy way to grab old job IDs in order to retrieve files or check the status of a set of jobs. You can see that one finished and one was stopped.

```
> GetJobHistory()
```

```
    job.id  job.name                                application
job "55042" "MuscleShare_2014-06-27_17-04-45.955" "Muscle-3.8.32u4"
job "55041" "Muscle_2014-06-27_17-04-38.045"      "Muscle-3.8.32u4"
    status
job "ARCHIVING_FINISHED"
job "STOPPED"
```

8.5 Retrieve job files

```
RetrieveJob(job.id, file.vec=NULL, print.curl=FALSE, verbose=FALSE)
```

One very handy thing about the *rPlant* package is the ability to download the files directly from the iPlant servers to your computer. The following downloads all of the output files.

```
> RetrieveJob(myJobS[[1]])
```

The files have been downloaded into a new directory within your working directory. If you want to download select files at a time, then these will need to be defined within the file.vec argument (for example: `RetrieveJob(myJobM, file.vec=c("fasta.aln"))`)

8.6 Delete job

`DeleteJob(job.id, print.curl=FALSE, ALL=FALSE)`

After the job has been submitted, the results downloaded, and you have no need for the job anymore, you can use the `DeleteJob` function to delete the job. The nice thing about this function is that not only will it delete the job number from the job history but it will also delete the job folder and all contents in the analyses folder in the user's cloudspace.

```
> DeleteJob(myJobS[[1]])
```

You also have the option to erase ALL job history from a user's past. This may be useful after a round of testing.

```
> DeleteJob(ALL=TRUE)
```

9 Submitting Jobs With Wrappers

We have ten dedicated wrapper functions for iPlant applications that will ease submitting jobs. These wrappers will use many application defaults and/or change flags into wrapper function arguments. Of course, if a user needs more flexibility in flagging options, then they can still submit jobs using the `SubmitJob` function.

Of the iPlant applications that have dedicated wrappers, there is a clear bias towards phylogenetic applications because the authors are evolutionary biologists and regularly use these programs. Writing wrapper functions is not programmatically difficult, but it does require familiarity with the individual programs and their associated data sets. We would like to encourage any users who are using programs without wrappers to submit patches adding wrapper functions or request to be a developer. You can make these feature requests at the R-Forge site: `https://r-forge.r-project.org/tracker/?group_id=1328`.

Among the wrappers there are three which do alignments: `Muscle`, `Mafft` and `ClustalW`. The alignments will do both protein and nucleotide. Also make sure that the taxon names in the sequence files do not contain tabulators, carriage returns, spaces, ":", ",", ")", "(", ";", "]", "[".

```
> data(PROTEIN.fasta)
> write.fasta(sequences=PROTEIN.fasta, names=names(PROTEIN.fasta), file.out="PROTEIN.fasta")
> UploadFile(local.file.name="PROTEIN.fasta", filetype="FASTA-0")
```

9.1 Muscle

`Muscle(file.name, file.path="", job.name=NULL, args=NULL, version="Muscle-3.8.32u4", print.curl=FALSE, aln.filetype="PHYLIP_INT", shared.username=NULL, suppress.Warnings=FALSE)`

MUSCLE is a program for creating multiple alignments of amino acid or nucleotide sequences. A range of options is provided that give you the choice of optimizing accuracy, speed, or some compromise between the two. The manual is also available here: `http://www.drive5.com/muscle/muscle_userguide3.8.html`

```
> myJobMuDP <- Muscle("DNA.fasta", aln.filetype="PHYLIP_INT", job.name="muscleDNAphyINT")
```

```
Job submitted.
You can check your job using CheckJobStatus(55043)
Result file: phylip_interleaved.aln
```

```
> myJobMuDF <- Muscle("DNA.fasta", aln.filetype="FASTA", job.name="muscleDNAfasta")
```

```
Job submitted.
You can check your job using CheckJobStatus(55044)
Result file: fasta.aln
```

```
> myJobMuDPP <- Muscle("DNA.fasta", aln.filetype="PHYLIP_PARS", job.name="muscleDNAphyPARS")
```

```
Job submitted.
You can check your job using CheckJobStatus(55045)
Result file: phylip_pars.aln
```

```
> myJobMuDPS <- Muscle("DNA.fasta", aln.filetype="PHYLIP_SEQ", job.name="muscleDNAphySEQ")
```

```
Job submitted.
You can check your job using CheckJobStatus(55046)
Result file: phylip_sequential.aln
```

```
> myJobMuDC <- Muscle("DNA.fasta", aln.filetype="CLUSTALW", job.name="muscleDNAclustalw")
```

```
Job submitted.
You can check your job using CheckJobStatus(55047)
Result file: clustalw.aln
```

```
> myJobMuDM <- Muscle("DNA.fasta", aln.filetype="MSF", job.name="muscleDNAmsf")
```

```
Job submitted.
You can check your job using CheckJobStatus(55048)
Result file: msf.aln
```

```
> myJobMuPP <- Muscle("PROTEIN.fasta", aln.filetype="PHYLIP_INT", job.name="musclePROTEINphyINT")
```

```
Job submitted.
You can check your job using CheckJobStatus(55049)
Result file: phylip_interleaved.aln
```

```
> myJobMuPF <- Muscle("PROTEIN.fasta", aln.filetype="FASTA", job.name="musclePROTEINfasta")
```

```
Job submitted.
You can check your job using CheckJobStatus(55050)
Result file: fasta.aln
```

```
> myJobMuPPP <- Muscle("PROTEIN.fasta", aln.filetype="PHYLIP_PARS", job.name="musclePROTEINphyPARS")
```

```
Job submitted.
You can check your job using CheckJobStatus(55051)
Result file: phylip_pars.aln
```

```
> myJobMuPPS <- Muscle("PROTEIN.fasta", aln.filetype="PHYLIP_SEQ", job.name="musclePROTEINphySEQ")
```

```
Job submitted.
You can check your job using CheckJobStatus(55052)
Result file: phylip_sequential.aln
```

```
> myJobMuPC <- Muscle("PROTEIN.fasta", aln.filetype="CLUSTALW", job.name="musclePROTEINclustalw")
```

```
Job submitted.
You can check your job using CheckJobStatus(55053)
Result file: clustalw.aln
```

```
> myJobMuPM <- Muscle("PROTEIN.fasta", aln.filetype="MSF", job.name="muscleDNAmsf")
```

```
Job submitted.
You can check your job using CheckJobStatus(55054)
Result file: msf.aln
```

MUSCLE outputs six alignments: fasta.aln (http://en.wikipedia.org/wiki/FASTA_format), phylip_sequential.aln, phylip_interleaved.aln, phylip_pars.aln (http://www.bioperl.org/wiki/PHYLIP_multiple_alignment_format), clustalw.aln (http://meme.nbcr.net/meme/doc/clustalw-format.html) and msf.aln (http://en.wikipedia.org/wiki/MSF).

9.2 Mafft

Mafft(file.name, file.path="", type="DNA", print.curl=FALSE, version="mafftDispatcher-1.0.13100u1", args=NULL, job.name=NULL, aln.filetype="FASTA", shared.username=NULL, suppress.Warnings=FALSE)

MAFFT is a multiple sequence alignment program for unix-like operating systems. It offers a range of multiple alignment methods, L-INS-i (accurate; for alignment of about 200 sequences), FFT-NS-2 (fast; for alignment of about 10,000 sequences), etc. See http://mafft.cbrc.jp/alignment/software/. The manual is also available here: http://mafft.cbrc.jp/alignment/software/manual/manual.html.

```
> myJobMaDF <- Mafft("DNA.fasta", job.name="mafftDNAfasta")

Job submitted.
You can check your job using CheckJobStatus(55055)
Result file: mafft.fa

> myJobMaDC <- Mafft("DNA.fasta", aln.filetype="CLUSTALW", job.name="mafftDNAclustalw")

Job submitted.
You can check your job using CheckJobStatus(55056)
Result file: mafft.fa

> myJobMaPF <- Mafft("PROTEIN.fasta", type="PROTEIN", job.name="mafftPROTEINfasta")

Job submitted.
You can check your job using CheckJobStatus(55057)
Result file: mafft.fa

> myJobMaPC <- Mafft("PROTEIN.fasta", type="PROTEIN", aln.filetype="CLUSTALW",
+                    job.name="mafftPROTEINclustalw")

Job submitted.
You can check your job using CheckJobStatus(55058)
Result file: mafft.fa
```

MAFFT outputs two alignments (both named: mafft.fa): FASTA (http://en.wikipedia.org/wiki/FASTA_format) and CLUSTALW (http://meme.nbcr.net/meme/doc/clustalw-format.html).

9.3 ClustalW

ClustalW(file.name, file.path="", type="DNA", job.name=NULL, version="ClustalW2-2.1u1", print.curl=FALSE, args=NULL, aln.filetype="PHYLIP", shared.username=NULL, suppress.Warnings=FALSE)

An approach for performing multiple alignments of large numbers of amino acid or nucleotide sequences is described. The method is based on first deriving a phylogenetic tree from a matrix of all pairwise sequence similarity scores, obtained using a fast pairwise alignment algorithm. See details on `http://www.clustal.org/clustal2/`.

```
> myJobCWDP <- ClustalW("DNA.fasta", job.name="clustalwDNAphylip")

Job submitted.
You can check your job using CheckJobStatus(55059)
Result file: clustalw2.fa

> myJobCWDC <- ClustalW("DNA.fasta", aln.filetype="CLUSTALW", job.name="clustalwDNAclustalw")

Job submitted.
You can check your job using CheckJobStatus(55061)
Result file: clustalw2.fa

> myJobCWDN <- ClustalW("DNA.fasta", aln.filetype="NEXUS", job.name="clustalwDNAnexus")

Job submitted.
You can check your job using CheckJobStatus(55062)
Result file: clustalw2.fa

> myJobCWDGCG <- ClustalW("DNA.fasta", aln.filetype="GCG", job.name="clustalwDNAgcg")

Job submitted.
You can check your job using CheckJobStatus(55063)
Result file: clustalw2.fa

> myJobCWDGDE <- ClustalW("DNA.fasta", aln.filetype="GDE", job.name="clustalwDNAgde")

Job submitted.
You can check your job using CheckJobStatus(55064)
Result file: clustalw2.fa

> myJobCWDPIR <- ClustalW("DNA.fasta", aln.filetype="PIR", job.name="clustalwDNApir")

Job submitted.
You can check your job using CheckJobStatus(55065)
Result file: clustalw2.fa

> myJobCWPP <- ClustalW("PROTEIN.fasta", type="PROTEIN", job.name="clustalwPROTEINphylip")

Job submitted.
You can check your job using CheckJobStatus(55066)
Result file: clustalw2.fa

> myJobCWPC <- ClustalW("PROTEIN.fasta", type="PROTEIN", aln.filetype="CLUSTALW",
+                       job.name="clustalwPROTEINclustalw")

Job submitted.
You can check your job using CheckJobStatus(55067)
Result file: clustalw2.fa

> myJobCWPN <- ClustalW("PROTEIN.fasta", type="PROTEIN", aln.filetype="NEXUS",
+                       job.name="clustalwPROTEINnexus")

Job submitted.
You can check your job using CheckJobStatus(55068)
Result file: clustalw2.fa

> myJobCWPGCG <- ClustalW("PROTEIN.fasta", type="PROTEIN", aln.filetype="GCG",
+                       job.name="clustalwPROTEINgcg")
```

```
Job submitted.
You can check your job using CheckJobStatus(55069)
Result file: clustalw2.fa
```

```
> myJobCWPGDE <- ClustalW("PROTEIN.fasta", type="PROTEIN", aln.filetype="GDE",
+                         job.name="clustalwPROTEINgde")
```

```
Job submitted.
You can check your job using CheckJobStatus(55070)
Result file: clustalw2.fa
```

```
> myJobCWPPIR <- ClustalW("PROTEIN.fasta", type="PROTEIN", aln.filetype="PIR",
+                         job.name="clustalwPROTEINpir")
```

```
Job submitted.
You can check your job using CheckJobStatus(55071)
Result file: clustalw2.fa
```

`ClustalW` outputs six alignments (all named: clustalw.fa): CLUSTALW `http://meme.nbcr.net/meme/doc/clustalw-format.html`, PHYLIP_INT `http://www.bioperl.org/wiki/PHYLIP_multiple_alignment_format`, NEXUS `http://en.wikipedia.org/wiki/Nexus_file`, GCG `http://www.genomatix.de/online_help/help/sequence_formats.html#GCG`, GDE `http://www.cse.unsw.edu.au/~binftools/birch/GDE/overview/GDE.file_formats.html`, and PIR `http://www.bioinformatics.nl/tools/crab_pir.html`.

9.4 FastTree

```
Fasttree <- function(file.name, file.path="", job.name=NULL, args=NULL, type="DNA",
model=NULL, gamma=FALSE, stat=FALSE, print.curl=FALSE, version="fasttreeDispatcher-
1.0.0u1", shared.username=NULL, suppress.Warnings=FALSE)
```

`FastTree` infers approximately-maximum-likelihood phylogenetic trees from alignments of nucleotide or protein sequences. See `http://meta.microbesonline.org/fasttree/`

```
> myJobFaDMuP <- Fasttree("phylip_interleaved.aln", file.path=paste("analyses/",myJobMuDP[[2]],
+                         sep=""), job.name="fasttreeMUSCLEdnaPHY")
```

```
Job submitted.
You can check your job using CheckJobStatus(55072)
```

```
> myJobFaDCWP <- Fasttree("clustalw2.fa", file.path=paste("analyses/",myJobCWDP[[2]], sep=""),
+                         job.name="fasttreeCLUSTALWdnaPHY")
```

```
Job submitted.
You can check your job using CheckJobStatus(55073)
```

```
> myJobFaDMuF <- Fasttree("fasta.aln", file.path=paste("analyses/",myJobMuDF[[2]], sep=""),
+                         job.name="fasttreeMUSCLEdnaFASTA")
```

```
Job submitted.
You can check your job using CheckJobStatus(55074)
```

```
> myJobFaDCWF <- Fasttree("mafft.fa", file.path=paste("analyses/",myJobMaDF[[2]], sep=""),
+                         job.name="fasttreeMAFFTdnaFASTA")
```

```
Job submitted.
You can check your job using CheckJobStatus(55075)
```

```
> myJobFaPMuP <- Fasttree("phylip_interleaved.aln", file.path=paste("analyses/",myJobMuPP[[2]],
+                         sep=""), type="PROTEIN",
+                         job.name="fasttreeMUSCLEproteinPHY")
```

```
Job submitted.
You can check your job using CheckJobStatus(55076)
```

```
> myJobFaPCWP <- Fasttree("clustalw2.fa", type="PROTEIN", file.path=paste("analyses/",myJobCWPP[[2]],
+                         sep=""), job.name="fasttreeCLUSTALWproteinPHY")
```

```
Job submitted.
You can check your job using CheckJobStatus(55077)
```

```
> myJobFaPMuF <- Fasttree("fasta.aln", file.path=paste("analyses/",myJobMuPF[[2]], sep=""),
+                         type="PROTEIN", job.name="fasttreeMUSCLEproteinFASTA")
```

```
Job submitted.
You can check your job using CheckJobStatus(55078)
```

```
> myJobFaPCWF <- Fasttree("mafft.fa", file.path=paste("analyses/",myJobMaPF[[2]], sep=""),
+                         type="PROTEIN", job.name="fasttreeMAFFTproteinFASTA")
```

```
Job submitted.
You can check your job using CheckJobStatus(55079)
```

`Fasttree` outputs trees in Newick format `http://en.wikipedia.org/wiki/Newick_format`. The placement of the root is not biologically meaningful. The local support values are given as names for the internal nodes, and range from 0 to 1, not from 0 to 100 or 0 to 1,000. If all sequences are unique, then the tree will be fully resolved (the root will have three children and other internal nodes will have two children). If there are multiple sequences that are identical to each other, then there will be a multifurcation. Also, there are no support values for the parent nodes of redundant sequences.

9.5 RAxML (Randomized Accelerated Maximum Likelihood)

```
RAxML(file.name, file.path="", job.name=NULL, type="DNA", model=NULL, bootstrap=NULL,
algorithm="d", multipleModelFileName=NULL, args=NULL, numcat=25, nprocs=12,
version="raxml-lonestar-7.2.8u1", print.curl=FALSE, shared.username=NULL,
substitution_matrix=NULL, empirical.frequencies=FALSE, suppress.Warnings=FALSE)
```

`RAxML` is a program for sequential and parallel Maximum Likelihood based inference of large phylogenetic tress. It has originall been derived from from fastDNAml which in turn was derived from Joe Felsentein's dnaml which is part of the PHYLIP package. See `http://sco.h-its.org/exelixis/oldPage/RAxML-Manual.7.0.4.pdf` for details.

```
> myJobRDMuP <- RAxML("phylip_interleaved.aln", file.path=paste("analyses/",myJobMuDP[[2]],
+                     sep=""), job.name="raxmlMUSCLEdnaPHY")
```

```
Job submitted.
You can check your job using CheckJobStatus(55080)
```

```
> myJobRDCWP <- RAxML("clustalw2.fa", file.path=paste("analyses/",myJobCWDP[[2]], sep=""),
+                     job.name="raxmlCLUSTALWdnaPHY")
```

```
Job submitted.
You can check your job using CheckJobStatus(55081)
```

```
> myJobRPMuP <- RAxML("phylip_interleaved.aln", file.path=paste("analyses/",myJobMuPP[[2]],
+                     sep=""), type="PROTEIN", job.name="raxmlMUSCLEproteinPHY")
```

```
Job submitted.
You can check your job using CheckJobStatus(55082)
```

```
> myJobRPCWP <- RAxML("clustalw2.fa", file.path=paste("analyses/",myJobCWPP[[2]], sep=""),
+                     type="PROTEIN", job.name="raxmlCLUSTALWproteinPHY")

Job submitted.
You can check your job using CheckJobStatus(55083)
```

For this application there are numerous output files. See pg 16-17 of the manual for complete details. RAxML outputs trees in Newick format `http://en.wikipedia.org/wiki/Newick_format`.

9.6 PHYLIP-Parsimony 3.69

`PHYLIP_Pars(file.name, file.path="", job.name=NULL, type="DNA", print.curl=FALSE, shared.username=NULL, suppress.Warnings=FALSE)`

`PHYLIP` is a free package of programs for inferring phylogenies. It is distributed as source code, documentation files, and a number of different types of executables. The web page: `http://evolution.genetics.washington.edu/phylip/doc/main.html`, by Joe Felsenstein of the Department of Genome Sciences and the Department of Biology at the University of Washington, contain information on PHYLIP. PHYLIP (the PHYLogeny Inference Package) is a package of programs for inferring phylogenies (evolutionary trees). Methods that are available in the package include parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees.

```
> myJobPDMuPP <- PHYLIP_Pars("phylip_pars.aln", file.path=paste("analyses/",myJobMuDPP[[2]],
+                            sep=""), job.name="phylipMUSCLEdnaPHYpars")

Job submitted.
You can check your job using CheckJobStatus(55084)

> myJobPPMuPP <- PHYLIP_Pars("phylip_pars.aln", file.path=paste("analyses/",myJobMuPPP[[2]],
+                            sep=""), type="PROTEIN", job.name="phylipMUSCLEproteinPHYpars")

Job submitted.
You can check your job using CheckJobStatus(55085)
```

`PHYLIP Parsimony` outputs trees in Newick format `http://en.wikipedia.org/wiki/Newick_format`.

9.7 Genome Wide Association Study models

Upload the sample files. They are in the transposed PLINK format `http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#tr`.

```
> data(geno_test.tfam)
> write.table(geno_test.tfam, file = "geno_test.tfam", row.names=FALSE,
+             col.names=FALSE, quote=FALSE, sep="\t")
> UploadFile(local.file.name="geno_test.tfam")
> data(geno_test.tped)
> write.table(geno_test.tped, file = "geno_test.tped", row.names=FALSE,
+             col.names=FALSE, quote=FALSE, sep="\t")
> UploadFile(local.file.name="geno_test.tped")
```

9.8 PLINK Conversion

```
PLINKConversion(file.list="", file.path="", output.type="-recode", job.name=NULL,
shared.username=NULL, print.curl=FALSE, version="plink-1.07u1", suppress.Warnings=FALSE)
```

This function converts the standard PLINK file formats (Regular (ped/map), Transposed (tped/tfam), and Binary (bed/bim/fam)) to various other PLINK file formats.

```
> myJobPLINKCT <- PLINKConversion(file.list=list("geno_test.tfam","geno_test.tped"), job.name="PCT",
+                                 out.basename="plinkout")

Job submitted.
You can check your job using CheckJobStatus(55086)

> myJobPLINKCR <- PLINKConversion(file.list=list("plinkout.map","plinkout.ped"),
+                                 file.path=paste("analyses/", myJobPLINKCT[[2]], sep=""),
+                                 output.type="--recode --transpose", job.name="PCR")

Job submitted.
You can check your job using CheckJobStatus(55087)
```

There are many output files possible, `http://pngu.mgh.harvard.edu/~purcell/plink/reference.shtml#output`

9.9 PLINK

```
PLINK(file.list="", file.path="", job.name=NULL, association.method="-assoc",
no.sex=TRUE, args=NULL, print.curl=FALSE, multi.adjust=TRUE, version="plink-1.07u1",
shared.username=NULL, suppress.Warnings=FALSE)
```

PLINK is an open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner, check `http://pngu.mgh.harvard.edu/~purcell/plink/` for details.

```
> myJobPLINKT <- PLINK(file.list=list("geno_test.tfam","geno_test.tped"), job.name="PLINKT")

Job submitted.
You can check your job using CheckJobStatus(55088)

> myJobPLINKR <- PLINK(file.list=list("plinkout.map","plinkout.ped"),
+                      file.path=paste("analyses/", myJobPLINKCT[[2]], sep=""), job.name="PLINKR")

Job submitted.
You can check your job using CheckJobStatus(55089)
```

There are many output files possible, `http://pngu.mgh.harvard.edu/~purcell/plink/reference.shtml#output`

9.10 FaST-LMM (Factored Spectrally Transformed Linear Mixed Models)

```
FaST_LMM(input.file.list="", ALL.file.path="", print.curl=FALSE, sim.file.list=NULL,
pheno.file.name=NULL, mpheno=1, args=NULL, covar.file.name=NULL, job.name=NULL,
version="FaST-LMM-1.09u1", shared.username=NULL, suppress.Warnings=FALSE)
```

FaST-LMM (Factored Spectrally Transformed Linear Mixed Models) is a program for performing genome-wide association studies (GWAS) on large data sets. FaST-LMM is described more fully at http://www.nature.com/nmeth/journal/v8/n10/abs/nmeth.1681.html, and also at http://fastlmm.codeplex.com/

```
> myJobFaST_LMMT <- FaST_LMM(input.file.list=list("geno_test.tfam","geno_test.tped"),
+                            job.name="FaST_LMMT")

Job submitted.
You can check your job using CheckJobStatus(55090)

> myJobFaST_LMMR <- FaST_LMM(input.file.list=list("plinkout.map","plinkout.ped"),
+                            ALL.file.path=paste("analyses/", myJobPLINKCT[[2]], sep=""),
+                            job.name="FaST_LMMR")

Job submitted.
You can check your job using CheckJobStatus(55091)
```

Not all information on the FaST-LMM model is here, see the FaST-LMM website http://fastlmm.codeplex.com/, or the FaST-LMM manual for more information.

10 Creating workflows

Finally, each of these steps can be combined to generate multi-step analyses. This has the benefit of reducing errors that can occur when manually running each application and, more importantly, ensures that results are reproducible. In the following example, a user starts with unaligned sequences on her or his local computer and ends with aligned sequences and a phylogenetic tree with all applications running on the iPlant servers.

10.1 Workflow One

This first workflow takes an amino acid fasta file, uses MUSCLE to get a PHYLIP_PARS alignment type. A couple things about this alignment; it is only available from MUSCLE and this alignment is very specific to the PHYLIP 3.69 model. The PHYLIP model then produces a tree. Note: this is the only way to do this workflow.

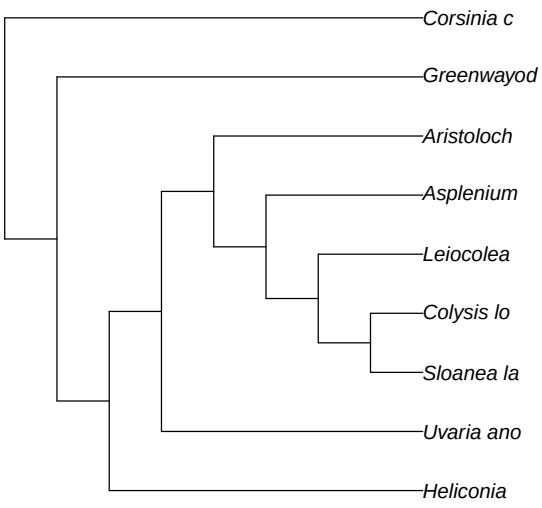
```
> myJobW1MP <- Muscle("PROTEIN.fasta", aln.filetype="PHYLIP_PARS", job.name="muscleWORKFLOW1protein")

Job submitted.
You can check your job using CheckJobStatus(55092)
Result file: phylip_pars.aln

> Wait(myJobW1MP[[1]], minWait, maxWait)
> myJobW1PPP <- PHYLIP_Pars("phylip_pars.aln", file.path=paste("analyses/",myJobW1MP[[2]],
+                           sep=""), type="PROTEIN", job.name="phylipWORKFLOW1protein")

Job submitted.
You can check your job using CheckJobStatus(55093)

> Wait(myJobW1PPP[[1]], minWait, maxWait)
> RetrieveJob(myJobW1PPP[[1]], c("outtree.nwk"))
> read.tree(paste(getwd(), myJobW1PPP[[2]], "outtree.nwk", sep="/")) -> Tree
```



10.2 Workflow Two

The second workflow takes the same amino acid fasta file and this time it uses `Mafft` to get a `FASTA` alignment type. `MUSCLE` also can output a fasta alignment. The `FastTree` model is then used to make the tree.

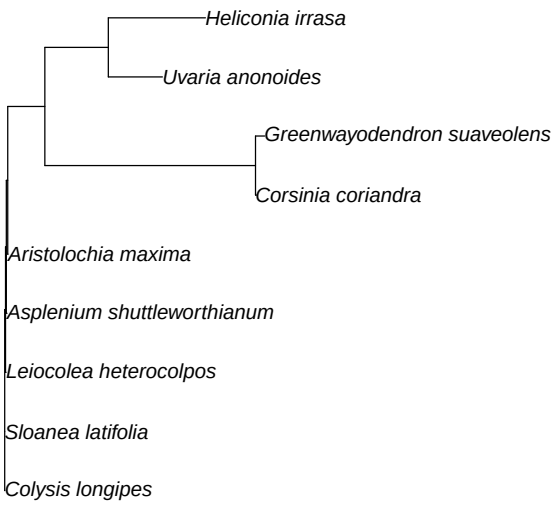
```
> myJobW2CWD <- Mafft("PROTEIN.fasta", type="PROTEIN", job.name="mafftPROTEINfasta")

Job submitted.
You can check your job using CheckJobStatus(55094)
Result file: mafft.fa

> Wait(myJobW2CWD[[1]], minWait, maxWait)
> myJobW2FaD <- Fasttree("mafft.fa", type="PROTEIN", file.path=paste("analyses/",myJobW2CWD[[2]],
+                        sep=""), job.name="fasttreeCLUSTALWfasta")

Job submitted.
You can check your job using CheckJobStatus(55095)

> Wait(myJobW2FaD[[1]], minWait, maxWait)
> RetrieveJob(myJobW2FaD[[1]], c("fasttree.nwk"))
> read.tree(paste(getwd(), myJobW2FaD[[2]], "fasttree.nwk", sep="/")) -> Tree
```



10.3 Workflow Three

The third workflow is again dealing with `FastTree`. `FastTree` can take either a FASTA alignment or a phylip interleaved alignment as inputs. Now `ClustalW` takes a nucleotide fasta file to get a `PHYLIP INTERLEAVED` alignment type. `MUSCLE` also can output that alignment. The `FastTree` model is then used to make the tree.

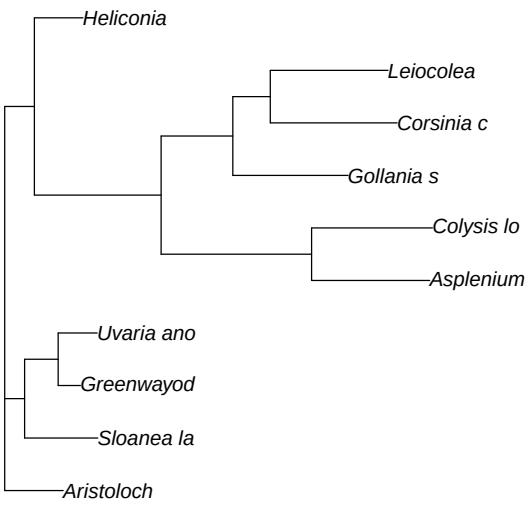
```
> myJobW3MuP <- ClustalW("DNA.fasta", job.name="clustalwDNAfasta")

Job submitted.
You can check your job using CheckJobStatus(55096)
Result file: clustalw2.fa

> Wait(myJobW3MuP[[1]], minWait, maxWait)
> myJobW3FaP <- Fasttree("clustalw2.fa", file.path=paste("analyses/",myJobW3MuP[[2]], sep=""),
+                        job.name="fasttreeMUSCLEdna")

Job submitted.
You can check your job using CheckJobStatus(55097)

> Wait(myJobW3FaP[[1]], minWait, maxWait)
> RetrieveJob(myJobW3FaP[[1]], c("fasttree.nwk"))
> read.tree(paste(getwd(), myJobW3FaP[[2]], "fasttree.nwk", sep="/")) -> Tree
```



10.4 Workflow Four

The fourth workflow is using `RAxML`. `MUSCLE` takes a nucleotide fasta file to get a `PHYLIP INTERLEAVED` alignment type. `ClustalW` also can output that alignment. The `RAxML` model is then used to make the tree.

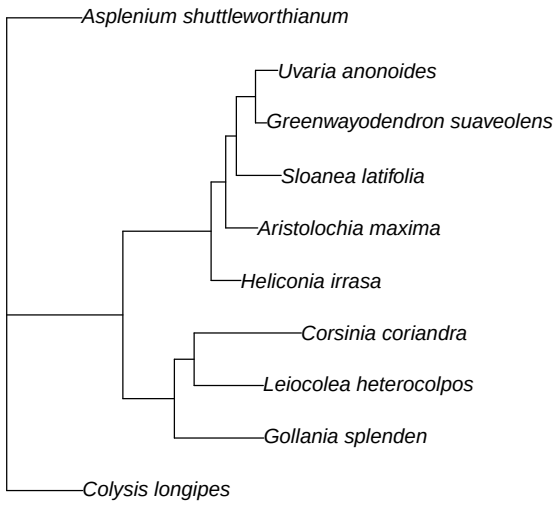
```
> myJobW4MuP <- Muscle("DNA.fasta", aln.filetype="PHYLIP_INT", job.name="muscleWORKFLOW4dna")

Job submitted.
You can check your job using CheckJobStatus(55098)
Result file: phylip_interleaved.aln

> Wait(myJobW4MuP[[1]], minWait, maxWait)
> myJobW4RP <- RAxML("phylip_interleaved.aln", file.path=paste("analyses/",myJobW4MuP[[2]], sep=""),
+                    job.name="raxmlWORKFLOW4dna")

Job submitted.
You can check your job using CheckJobStatus(55099)

> Wait(myJobW4RP[[1]], minWait, maxWait)
> RetrieveJob(myJobW4RP[[1]], c("RAxML_bestTree.nwk"))
> read.tree(paste(getwd(), myJobW4RP[[2]], "RAxML_bestTree.nwk", sep="/")) -> Tree
```


Asplenium shuttleworthianum
Uvaria anonoides
Greenwayodendron suaveolens
Sloanea latifolia
Aristolochia maxima
Heliconia irrasa
Corsinia coriandra
Leiocolea heterocolpos
Gollania splenden
Colysis longipes

References

Dooley, R., Vaughn, M., Stanzione, D., Terry, S., and Skidmore, E. "Software-as-a-Service: The iPlant Foundation API." In *5th IEEE Workshop on Many-Task Computing on Grids and Supercomputers* (2012).
URL `http://datasys.cs.iit.edu/events/MTAGS12/p07.pdf`

Fielding, R. T. "Architectural styles and the design of network-based software architectures." Ph.D. thesis, University of California (2000).

Lang, D. T. "R as a Web Client–the RCurl package." *Journal of Statistical Software, http://www. jstatsoft. org* (2007).

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2012). ISBN 3-900051-07-0.
URL `http://www.R-project.org/`

Stenberg, D. "1.47. 1 Available under license." *Open Source Used In Cisco Unified Communications Manager 8.6. 2a*, 5:395 (1996).