



Université Libre de Bruxelles
Service de Bioinformatique des Génomes et Réseaux (BiGRe)
Laboratory of Genome and Network Biology
<http://www.bigre.ulb.ac.be/>

Regulatory Sequence Analysis Tools (*RSAT*)

Installation guide

Jacques VAN HELDEN & the *RSAT* team

December 16, 2013

Contents

1	Quick installation guide	4
2	Description and requirements	5
2.1	Description	5
2.2	Requirements	5
2.2.1	Operating system	5
2.2.2	Perl language	5
2.2.3	Python language	5
2.2.4	Helper applications	6
3	Obtaining <i>RSAT</i> distribution	7
3.1	Installation from a compressed archive	7
4	Initializing <i>RSAT</i>	8
4.0.1	Configuring your <i>RSAT</i> server	8
4.0.2	Loading <i>RSAT</i> environment variables	8
4.1	Initializing the directories	9
4.1.1	Checking the <i>RSAT</i> path	9
5	Installing Perl modules	10
5.1	Before installing Perl modules: install the GD library	10
5.2	Automatic installation of Perl modules	10
5.3	Additional Perl modules required to support Ensembl genomes	11
6	Compiling C programs in <i>RSAT</i>	14
7	Downloading genomes	15
7.1	Original data sources	15
7.2	Requirement : wget	16
7.3	Importing organisms from the <i>RSAT</i> main server	16
7.3.1	Obtaining the list of organisms supported on the <i>RSAT</i> server	16
7.3.2	Importing a single organism	16
7.3.3	Importing a few selected organisms	17
7.3.4	Importing all the organisms from a given taxon	17

8	Testing the command-line tools	19
8.1	Testing the access to the programs	19
8.1.1	Perl scripts	19
8.1.2	Testing Perl graphical librairies	19
8.1.3	Python scripts	19
8.1.4	C programs	20
8.2	Testing genome installation	20
9	Installing third-party programs	21
9.1	Complementary programs for the analysis of regulatory sequences	21
10	Installing additional genomes on your machine	23
10.1	Adding support for Ensembl genomes	23
10.1.1	Handling genomes from Ensembl	23
10.2	Installing genomes and variations from <i>EnsEMBL</i>	24
10.2.1	<i>install-ensembl-genome</i>	24
10.2.2	Installing genomes from Ensembl genomes	25
10.2.3	Downloading variations	25
10.3	Importing genomes from NCBI BioProject	25
10.4	Importing multi-genome alignment files from UCSC	26
10.4.1	Warning: disk space requirement	26
10.4.2	Checking supported genomes at UCSC	27
10.4.3	Downloading multiz files from UCSC	27
10.5	Installing genomes from NCBI/Genbank files	27
10.5.1	Organization of the genome files	28
10.5.2	Downloading genomes from NCBI/Genbank	30
10.5.3	Parsing a genome from NCBI/Genbank	31
10.5.4	Parsing a genome from the Broad institute (MIT)	31
10.5.5	Updating the configuration file	32
10.5.6	Checking the start and stop codon composition	32
10.5.7	Calibrating oligonucleotide and dyad frequencies with <i>install-organisms</i>	33
10.5.8	Installing a genome in your own account	34
10.6	Installing genomes from EMBL files	34

1 Quick installation guide

I am aware that this installation guide is somewhat heavy. I am preparing a simplified installation procedure, which should be ready within a few days (Dec 2013). In the mean time the distribution is still coming with the detailed installation instructions. In case of trouble, don't hesitate to contact me.

2 Description and requirements

2.1 Description

This document describes the installation procedure for the software suite **Regulatory Sequence Analysis Tools** (*RSAT*), which integrates several dozens of tools to detect cis-regulatory elements in DNA sequences [?, ?, ?, ?].

2.2 Requirements

2.2.1 Operating system

RSAT is a unix-based package. It has been installed successfully on the following operating systems.

1. Linux
2. Mac OSX (latest version tested: 10.8.3)
3. Sun Solaris
4. Dec Alpha

RSAT is not compatible with any version of Microsoft Windows and we have no intention to make it compatible in a foreseeable future.

2.2.2 Perl language

Most of the programs in *RSAT* are written in Perl. Version 5.1 or later is recommended. A set of Perl modules is required, the *RSAT* package includes a script to install them automatically (see Chapter 5).

2.2.3 Python language

Some of the programs in *RSAT* are written in Python.

Python release 2.7¹ is recommended, because it contains some required libraries for remote access to external resources (UCSC genome browser).

The following Python libraries are required for various programs.

¹<http://www.python.org/getit/releases/2.7/>

- **setuptools**² is required to install other Python libraries (see installation instructions³).
- **suds**⁴ is used for accessing the SOAP interface.

2.2.4 Helper applications

wget

The program **wget**, is used to download

1. some helper programs developed by third-parties, which can be installed in **RSAT**;
2. genomes from the **RSAT**server to your local **RSAT**installation.

wget is part of linux distribution. If it is not installed on your computer, you can download it from <http://www.gnu.org/software/wget/>. An installation package for Mac OSX can be found at http://download.cnet.com/Wget/3000-18506_4-128268.html.

gnuplot

The standard version of the **RSAT**program **XYgraph** export figures in bitmapformat (png, jpeg). If you want to support vectorial drawings (pdf), which give a much better resolution for printing, you need to install the freeware software **gnuplot** (4.2 or later), which can be downloaded from <http://www.gnuplot.info/>.

git (only for developers)

For co-developers of the **RSAT**suite, the code is distributed program **git**. For external users, there is no need to use **git** since the code is distributed as a compressed archive on a Web page.

²<http://pypi.python.org/pypi/setuptools>

³<http://pypi.python.org/pypi/setuptools#installation-instructions>

⁴<https://fedorahosted.org/suds/>

3 Obtaining *RSAT* distribution

For the time being, *RSAT* is distributed as a compressed archive.

The license can be obtained from the *RSAT* web site (<http://rsat.eu/>).

3.1 Installation from a compressed archive

Download the latest version of the *RSAT* distribution. Uncompress the archive containing the programs. The archive is distributed `tar` format. The `.tar.gz` file can be uncompressed with the command *tar*, which is included in most Unix distributions.

```
tar -xpf rsat_yyyymmdd.tar.gz
```

4 Initializing *RSAT*

4.0.1 Configuring your *RSAT* server

RSAT requires to specify a set of parameters, which will be stored in three property files:

1. *RSAT_config.props*
2. *RSAT_config.mk*
3. *RSAT_config.bashrc*

In particular, *it is crucial to specify the full path of the variable RSAT*, which specifies the *RSAT* main directory.

The simplest way to update the configuration file is to run the following script.

```
## Enter in RSAT distribution folder
cd rsat

## Run the configuration script
perl perl-scripts/configure_rsat.pl
```

Alternatively, you can edit the files with a text editor of your choice.

4.0.2 Loading *RSAT* environment variables

The configuration script has created a bash file *RSAT_config.bashrc* in the *RSAT* distribution directory (folder *rsat*).

This file should be loaded each time you enter a session. There are several alternative ways to do this.

1. Source the file manually at each new session

```
## Load the RSAT environment variables
source RSAT_config.bashrc
```

2. Copy the content of this file in the bash configuration file in your home directory (*/.bashrc* or */.bash_profile*). *RSAT* environment variables will then be loaded for you at each connexion.
3. If you dispose of admin rights, you can copy the content of this file in the main bash configuration file:
 - in Linux Centos, you can create a file */etc/profile.d/rsat.bashrc*;

- Mac OSX, copy it to the main bashrc file (*/etc/bashrc*).

RSAT environment variables will then be automatically loaded for each user of this computer.

4.1 Initializing the directories

In addition to the programs, the installation of rsat requires the creation of a few directories for storing data, access logs (for the web server), and temporary files.

The distribution includes a series of make scripts which will facilitate this step. You just need go to the rsat directory, and start the appropriate make file.

```
cd $RSAT ;  
make -f makefiles/init_rsat.mk init
```

4.1.1 Checking the RSAT path

The **RSAT** programs should now be included in your path. To check if this is done properly, just type:

```
random-seq -l 350
```

If your configuration is correct, this command should return a random sequence of 350 nucleotides.

Don't worry if you see a warning looking like this:

```
; WARNING The tabular file with the list of supported organism cannot be read  
; WARNING Missing file [RSAT_PARENT_PATH]/rsat/public_html/data/supported_organisms.tab
```

This warning will disappear as soon as you download the first organism in **RSAT**.

5 Installing Perl modules

Some Perl modules are required for the graphical tools of *RSAT*, and for some other specific programs. The perl modules can be found in the Comprehensive Perl Archive Network (<http://www.cpan.org/>), or can be installed with the command *cpan*.

5.1 Before installing Perl modules: install the GD library

The Perl module *GD.pm* requires prior installation of the *GD* library.

- On *Linux* systems, this library can be installed with the package manager of the distribution. for example:
 - *apt-get* for Ubuntu
 - *yum* for Centos
 - *yast* for Suze
 - ...
- On *Mac OSX* systems, the installation of the GD library can be done with the program *brew*.

5.2 Automatic installation of Perl modules

The simplest way to install all the required Perl modules is to type the command below. *Beware:* this command *sudo* requires administrator rights on the computer. If you don't have the root password, please consult your system administrator.

```
## Acquire the system administrator rights
sudo bash;

## Define the RSAT environment variable.
## You must replace [PARENT_PATH] by the full directory of the folder
## where rsat directory has been installed
export RSAT=[PARENT_PATH]/rsat;

cd $RSAT

## Display the list of Perl modules that will be installed
```

```
make -f makefiles/install_rsatz.mk perl_modules_list;

## Print the command that will be used to install the Perl modules (just for checking)
make -f makefiles/install_rsatz.mk perl_modules_cmd;

## Install the Perl modules
make -f makefiles/install_rsatz.mk perl_modules_install;
```

Beware, ***cpan*** will frequently ask you to confirm the installation steps. you should thus check the CPAN process and answer "yes" at each prompt.

The following command enables to install all the Perl modules in a somewhat risky, but less cumbersome way. It relies on the command ***yes*** to automatically answer each question by a carriage return, which will lead ***cpan*** to chose the default option.

```
## Install the Perl modules
make -f makefiles/install_rsatz.mk perl_modules_install;
```

In case some modules would not be properly installed with the above commands, you can try installing them manually (the list of required modules is listed in the next section).

5.3 Additional Perl modules required to support Ensembl genomes

This section is required only if you intend to use the ***RSAT*** programs interfaced to the Ensembl database. Since 2008, a series of ***RSAT*** programs support a direct access to the Ensembl database in order to ensure a convenient access to genomes from higher organisms [?].

- ***supported-organisms-ensembl***
- ***ensembl-org-info***
- ***retrieve-ensembl-seq.pl***
- ***get-ensembl-genome.pl***

Those programs require to install a few Perl libraries as well as a MySQL client on your machine.

The first requirement is the ***BioPerl*** module, which has in principle been installed in Chapter 5). The MySQL client should also have been installed in Chapter 5).

To obtain Ensembl ¹.

```
## Make sure you start from the right directory
cd $RSAT

## Display the parameters for installing Ensembl API (in particular,
```

¹Full instructions at http://useast.ensembl.org/info/docs/api/api_cvs.html

```
## the version for Ensembl and EnsemblGenomes).
##
## Check the number of the latest release on the respective web sites.
##   Ensembl: http://www.ensembl.org/index.html
##   EnsemblGenomes: http://ensemblgenomes.org/
make -f makefiles/install_software.mk install_ensembl_api_param

## Install the ensembl library
make -f makefiles/install_software.mk install_ensembl_api

## Notes: you need to enter the following passwords for the CVS
## servers.
## - For Ensembl:  CVSUSER.
## - For bioperl:  cvs
```

Note that there may be incompatibilities between successive versions of the Ensembl API. The install script includes a parameter `ENSEMBL_VERSION` to specify the version ("branch") of the Ensembl API distribution. Moreover, there are different release numbers of the “historical” Ensembl database, and for the EnsemblGenomes databases (Bacteria, Fungi, Plants, Metazoa).

In addition, there are dependencies between releases. So, EnsemblGenomes 20 is compatible with Ensembl version 73, whereas Ensembl has already released its version 74. In order to install an API compatible with Ensembl and EnsemblGenomes, we recommend always check the latest releases of both databases on the EnsemblGenomes web page (<http://www.ensemblgenomes.org>) and adapt the following command accordingly.

```
## Install the ensembl library with a specific branch number.
make -f makefiles/install_software.mk install_ensembl_api \
    ENSEMBL_VERSION=73
```

The installation script will print out a series of modifications of the `PERL5LIB` variable, that should be added to your `bashrc` file in order to provide support for Ensembl Perl API.

You should also check the specification of ensembl paths in the `props` file.

```
ensembl=[RSAT_PARENT_PATH]/rsat/lib/ensembl/modules
compara=[RSAT_PARENT_PATH]/rsat/lib/ensembl-compara/modules
variation=[RSAT_PARENT_PATH]/rsat/lib/ensembl-variation/modules
```

You also need to define the URL of the Ensembl database in that configuration file:

```
## EnSEMBL host
## Used by the EnSEMBL-accessing tools (retrieve-ensembl-seq,
## get-ensembl-genome).
## URL of the server for the EnSEMBL DB. By default, the
## main ensembl server is called, but a local server can be specified.
ensembl_host=ensemldb.ensembl.org
```

Notes:

1. to access EnSEMBL versions above 47, you need port 5306 to be opened on your machine. This might require an intervention of your system administrator of your network in order to ensure that the Firewall accepts this port.

Detailed information about the Ensembl libraries can be obtained on the Ensembl web site (²).

²http://www.ensembl.org/info/using/api/api_installation.html

6 Compiling C programs in *RSAT*

Some of the tools available in *RSAT* (*info-gibbs*, *matrix-scan-quick*, *count-words*) are written in the *C* language. The distribution only contains the sources of these tools, because the binaries are operating system-dependent. The programs can be compiled in a very easy way.

```
cd $RSAT;  
make -f makefiles/init_rsate.mk compile_all
```

This will compile and install the following programs in the directory *\$RSAT/bin*.

7 Downloading genomes

RSAT includes a series of tools to install and maintain the latest version of genomes.

The most convenient way to add support for one or several organisms on your machine is to use the programs ***supported-organisms*** and ***download-organism***.

Beware, the complete data required for a single genome may occupy several hundreds of Mb, because *RSAT* not only stores the genome sequence, but also the oligonucleotide frequency tables used to estimate background models, and the tables of BLAST hits used to get orthologs for comparative genomics. If you want to install many genomes on your computer, you should thus reserve a sufficient amount of space.

7.1 Original data sources

Genomes supported on *RSAT* were obtained from various sources.

Genomes can be installed either from the *RSAT* web site, or from their original sources.

- NCBI/Genbank (<ftp://ftp.ncbi.nih.gov/genomes/>) was the primary source for installing genomes on *RSAT*. Genomes are downloaded from the ftp site and installed locally on the *RSAT* server by parsing the .gbk files.
- The EBI genome directory (<ftp://ftp.ebi.ac.uk/pub/databases/genomes/Eukaryot>) contains supplementary genomes, which can be downloaded and installed on the *RSAT* server by parsing files in embl format.
- UCSC (<http://genome.ucsc.edu/>) for the multi-genome alignment files (multiz) used by ***peak-footprints***.
- Since 2008, ENSEMBL (<http://www.ensembl.org/>) genomes are supported by special tools (***retrieve-ensembl-seq***, ***supported-organisms-ensembl***), that remotely address queries to the Ensembl database.
- Since 2013, genomes can be downloaded and installed on *RSAT* servers, using the tool ***install-ensembl-genome***. Once installed, ensembl genomes can be queried with the same tools as the other genomes installed on *RSAT* servers (***retrieve-seq***, ***gene-info***, ...).

Other genomes can also be found on the web site of a diversity of genome-sequencing centers.

7.2 Requirement : wget

The download of genomes relies on the application **wget**, which is part of linux distribution¹.

wget is a “web aspirator”, which allows to download whole directories from ftp and http sites. You can check if the program is installed on your machine.

```
wget --help
```

This command should return the help pages for **wget**. If you obtain an error message (“command not found”), you need to ask your system administrator to install it.

7.3 Importing organisms from the *RSAT*main server

The simplest way to install organisms on our *RSAT*site is to download the *RSAT*-formatted files from the web server. For this, you can use a web aspirator (for example the program **wget**).

Beware, the full installation (including Mammals) requires a large disk space (several dozens of Gb). You should better start installing a small genome and test it before processing to the full installation. We illustrate the approach with the genome of our preferred model organism: the yeast *Saccharomyces cerevisiae*.

7.3.1 Obtaining the list of organisms supported on the *RSAT*server

By default, the program **supported-organisms** returns the list of organisms supported on your local *RSAT*installation. You can however use the option `-server` to obtain the list of organisms supported on a remote server.

```
supported-organisms-server
```

The command can be refined by restricting the list to a given taxon of interest.

```
supported-organisms-server -taxon Fungi
```

You can also ask additional information, for example the date of the last update and the source of each genome.

```
supported-organisms-server -taxon Fungi -return last_update,source,ID
```

7.3.2 Importing a single organism

The command

```
download-organism
```

¹For Linux: <http://www.gnu.org/software/wget/>; for Mac OSX http://download.cnet.com/Wget/3000-18506_4-128268.html

allows you to download one or several organisms.

Beware, the complete data for a single genome may occupy several tens of Megabytes (Bacterial genomes) or a few Gigabases (Mammalian). Downloading tenomes thus requires a fast Internet connection, and may take time. If possible, please download genomes during the night (European time).

As a first step, we recommend to download the genome of the yeast *Saccharomyces cerevisiae*, since this is the model organism used in our tutorials.

```
download-organism -v 1 -org Saccharomyces_cerevisiae
```

In principle, the download should start immediately. *Beware*, the data volume to be downloaded is important, because the genome comes together with extra files (blast hits with other genoems, oligonucleotide and dyad frequencies). Depending on the network bandwidth, the download of a genome may take several minutes or tens of minutes.

After the task is completed, you can check if the configuration file has been correctly updated by typing the command.

```
supported-organisms
```

In principle, the following information should be displayed on your terminal.

```
Saccharomyces_cerevisiae
```

You can also add parameters to get specific information on the supported organisms.

```
supported-organisms -return ID,last_update
```

7.3.3 Importing a few selected organisms

The program **download-organism** can be launched with a list of organisms by using iteratively the option `-org`.

```
download-organism -v 1 -org Escherichia_coli_K12 -org Mycoplasma_genitalium_G37_uid5770
```

Note: genome names may change with time, since genome centers are occasionally adding new suffixes for genomes. The organism names indicated after the option `-org` should belong to the list of supported organisms collected with the command **supported-organisms -server**.

7.3.4 Importing all the organisms from a given taxon

For comparative genomics, it is convenient to install on your server all the organisms of a given taxon. For this, you can simply use the option `-taxon` of **download-organism**.

Before doing this, it is wise to check the number of genomes that belong to this taxon on the server.

```
## Get the list of organisms belonging to the order "Enterobacteriales" on the server
supported-organisms -taxon Enterobacteriales -server

## Count the number of organisms
supported-organisms -taxon Enterobacteriales -server | wc -l
```

In Dec 2013, there are 203 Enterobacteriales supported on the *RSAT* server. Before starting the download, you should check two things:

1. Has your network a sufficient bandwidth to ensure the transfer in a reasonable time ?
2. Do you have enough free space on your hard drive to store all those genomes ?

If the answer to both questions is “yes”, you can start the download.

```
download-organism -v 1 -taxon Enterobacteriales
```

8 Testing the command-line tools

8.1 Testing the access to the programs

8.1.1 Perl scripts

From now on, you should be able to use the perl scripts from the command line. To test this, run:

```
random-seq -help
```

This should display the on-line help for the random sequence generator.

```
random-seq -l 200 -n 4
```

Should generate a random sequence of 200 nucleotides.

8.1.2 Testing Perl graphical librairies

RSAT includes some graphical tools (***feature-map*** and ***XYgraph***), which require a proper installation of Perl modules.

GD.pm Interface to Gd Graphics Library.

PostScript::Simple Produce PostScript files from Perl.

To test if these modules are available on your machine, type.

```
feature-map -help
```

If the modules are available, you should see the help message of the program *feature-map*. If not, you will see an error message complaining about the missing librairies. In such a case, ask your system administrator to install the missing modules.

8.1.3 Python scripts

The *RSAT* distribution includes some Python scripts. To test if they are running correctly, you can try the proram ***random-motif***.

```
random-motif -l 10 -c 0.85 -n 3
```

This command will generate 3 position-specific scoring matrix (PSSM) of 10-columns with 85% conservation of one residue in each column.

8.1.4 C programs

You can test the correct installation of the C programs with the following command.

```
random-seq -l 10000 -n 10 | count-words -l 2 -v 1 -2str -i /dev/stdin
```

The first program (***random-seq***) is a Perl script, which generates a random sequence. The output is directly piped to the C program ***count-words***, which computes the frequencies and occurrences of each dinucleotide.

8.2 Testing genome installation

We will now test if the genomes are correctly installed. You can obtain the list of supported organisms with the command:

```
supported-organisms
```

If this command returns no result, it means that genomes were either not installed, or not correctly configured. In such a case, check the directories in the *data/genomes* directory, and check that the file *data/supported_organisms.pl*.

Once you can obtain the list of installed organisms, try to retrieve some upstream sequences. You can first read the list of options for the ***retrieve-seq*** program.

```
retrieve-seq -help
```

Select an organism (say *Saccharomyces cerevisiae*), and retrieve all the start codons with the following options :

```
retrieve-seq -org Saccharomyces_cerevisiae -featype CDS \  
-type upstream -from 0 -to +2 -all \  
-format wc -nocomment
```

This should return a set of 3 bp sequences, mostly ATG (in the case of *Saccharomyces cerevisiae* at least). We can combine ***retrieve-seq*** and ***oligo-analysis*** to check the frequencies of trinucleotides found at the start positions of all yeast genes.

```
retrieve-seq -org Saccharomyces_cerevisiae -featype CDS \  
-type upstream -from 0 -to +2 -all \  
| oligo-analysis -l 3 -lstr -return occ,freq -v 1 -sort
```

9 Installing third-party programs

9.1 Complementary programs for the analysis of regulatory sequences

The **RSAT** distribution only contains the programs developed by the **RSAT** team.

A few additional programs, developed by third parties, can optionally be integrated in the package. All third-party programs may be located in the directory *bin* of the **RSAT** distribution.

In order to add functionalities to **RSAT**, install some or all of these programs and include their binaries path *\$RSAT/bin*. If you are not familiar with the installation of unix programs, ask assistance to your system administrator.

Some of those can be downloaded and installed automatically using the makefile *install_rsat.mk*. Before doing this, you must make sure that the program **wget** (this program is supported on Linux ¹ and Mac OSX ² systems).

You can then run the following commands to install some of the third-party programs that are complementary to **RSAT**.

```
cd $RSAT;  
make -f makefiles/install_software.mk install_ext_apps
```

Some other third-party programs will require a manual installation (in particular, **vmatch** and **mkvtree**).

vmatch and **mkvtree** : developed by Stefan Kurtz, are used by the program **purge-sequences**, to mask redundant sequences that bias motif discovery statistics.

seqlogo : developed by Thomas D. Schneider, is used by the programs **convert-matrix**, **compare-matrices**, **peak-motifs**, **matrix-quality** and a few others, to generate logos. **seqlogo** is the command-line version of **WebLogo**³.

Download the source code archive and uncompress it. Copy the following files to the directory *bin* of your **RSAT** distribution: *seqlogo*, *logo.pm*, *template.pm* and *template.eps*.

seqlogo requires a recent version of **gs** (ghostscript⁴) to create PNG and PDF output, and **ImageMagic's convert**⁵ to create GIFs.

¹<http://www.gnu.org/software/wget/>

²http://download.cnet.com/Wget/3000-18506_4-128268.html

³<http://weblogo.berkeley.edu/>

⁴<http://www.ghostscript.com/>

⁵<http://www.imagemagick.org/>

Program	author	URL
vmatch+mkvtree	Stefan Kurtz	http://www.vmatch.de/
seqlogo	Thomas Sneider	http://weblogo.berkeley.edu/
patser	Jerry Hertz	ftp://ftp.genetics.wustl.edu/pub/stormo/Consensus/
consensus	Jerry Hertz	ftp://ftp.genetics.wustl.edu/pub/stormo/Consensus/
meme	Tim Bailey	http://meme.sdsc.edu/
MotifSampler	Gert Thijs	http://www.esat.kuleuven.ac.be/~thijs/download.html

Table 9.1: Programs from other developers which are complementary to the *RSAT* package.

matrix-based pattern discovery : several third-party pattern discovery programs can be optionally called from some *RSAT* task managers (e.g. *multiple-family-analysis*, *peak-motifs*).

- *meme* (Tim Bailey)
- *consensus* (Jerry Hertz)
- *MotifSampler* (Gert Thijs)
- *gibbs* (Andrew Neuwald)

Their installation is not properly required for *RSAT* functioning, but it may be convenient to install them in order to compare the results returned by alternative motif discovery approaches on the same data sets.

10 Installing additional genomes on your machine

The easiest way to install genomes on your machine is to download them from the main *RSAT* server, as indicated in the Chapter “Downloading genomes” (Chap. 7 of the installation guide).

In some cases, you may however wish to install a genome by yourself, because this genome is not supported on the main *RSAT* server. For this, you can use the programs that we use to install new genomes on the main *RSAT* server.

10.1 Adding support for Ensembl genomes

In addition to the genomes imported and maintained on your local *RSAT* server, the program *retrieve-ensembl-seq* allows you to retrieve sequences for any organism supported in the Ensembl database (<http://ensembl.org>).

For this, you first need to install the Bioperl and Ensembl Perl libraries (see section 5.3).

10.1.1 Handling genomes from Ensembl

The first step to work with Ensembl genomes is to check the list of organisms currently supported on their Web server.

```
supported-organisms-ensembl
```

You can then get more precise information about a given organism (build, chromosomes) with the command *ensembl-org-info*.

```
ensembl-org-info -org Drosophila_melanogaster
```

Sequences can be retrieved from Ensembl with the command *retrieve-ensembl-seq*.

You can for example retrieve the 2kb sequence upstream of the transcription start site of the gene *PAX6* of the mouse.

```
retrieve-ensembl-seq.pl -org Mus_musculus -q PAX6 \  
-type upstream -feattype mrna -from -2000 -to -1 -nogene -rm \  
-alltranscripts -uniqseqs
```

Options

- `-type upstream` specifies that we want to collect the sequences located upstream of the gene (more precisely, upstream of the mRNA).

- `-featype mrna` indicates that the reference for computing coordinates is the mRNA. Since we collect upstream sequences, the 5' most position of the mRNA has coordinate 0, and upstream sequences have negative coordinates. Note that many genes are annotated with multiple RNAs for different reasons (alternative splicing, alternative transcription start sites). By default, the program will return the sequences upstream of each mRNA annotated for the query gene.
- `-nogene clip` the sequences to avoid overlapping the next upstream gene.
- `-rm repeat` masking (important for pattern discovery). Repetitive sequences are replaced by N characters.

10.2 Installing genomes and variations from *EnsEMBL*

10.2.1 *install-ensembl-genome*

The program *install-ensembl-genome* downloads the complete genomic sequence of a given organism from the *EnsEMBL* Web site, and installs it on the local *RSAT* site. It also installs the descriptions of genomic features (genes, CDS, mRNAs, ...), and the variations.

As usually, the complete help message can be obtained with the option `-help`.

```
## Get the description of the program + all options
install-ensembl-genome -help
```

Before installing a genome, it is generally a good idea to know which genomes are available. For this, use the option `-available_species`.

```
## Retrieve the list of supported species on EnsEMBL
install-ensembl-genome -v 1 -available_species
```

Note: inter-individual variations are available for a subset only of the genomes available in *EnsEMBL*. The option `-available_species` indicates, for each species, the availability (genome, features, variations). Obviously, the analysis of regulatory variations only makes sense for the genomes documented with variations.

We can now download the complete genomic sequence for the species of our choice. For the sake of space and time economy, we will use a small genome for this manual: the budding yeast *Saccharomyces cerevisiae*.

Beware: some installation steps take a lot of time, in particular the installation of genomic features, because the *EnsEMBL* interface requires to send individual queries for each gene separately. The full installation can thus take several hours. This is not a big issue, since installing a genome is not a daily task, but it is worth knowing that the whole process requires a continuous connection during several hours.

```
## Install the genome sequences for a selected organism
install-ensembl-genome -v 2 -species Saccharomyces_cerevisiae
```

The download time depends on genome size, on the speed of your internet connection, and on the number of genes.

10.2.2 Installing genomes from Ensembl genomes

The historical *EnsEMBL* project ¹ was focused on vertebrate genomes + a few model organisms (*Saccharomyces cerevisiae*, *Drosophila melanogaster*, ...).

A more recent project called *EnsemblGenomes*² extends the *EnsEMBL* project to a wider taxonomic space.

The program *install-ensembl-genome* supports the installation of genomes from *EnsEMBL* as well as *EnsemblGenomes*. By default, it opens a connection to the historical *EnsEMBL* database, but the option `-ensembl_genomes` enables to install genomes from the new project *EnsemblGenomes*.

```
## Get the list of available species from the extended project
## EnsemblGenomes
install-ensembl-genome -v 2 -available_species -ensembl_genomes \
    -o available_species_at_EnsemblGenome.txt
```

You can then locate your genome of interest in the file *available_species_at_EnsemblGenome.txt*, and start the installation (don't forget the option `-ensembl_genomes`).

```
## Install Escherichia coli (strain K12 MG1665) from EnsemblGenomes
install-ensembl-genome -v 2 -ensembl_genomes \
    -species Escherichia_coli_str_k_12_substr_mg1655
```

10.2.3 Downloading variations

The program *download-ensembl-variations* downloads variations from the *EnsEMBL* Web site, and installs it on the local *RSAT* site.

This program relies on *wget*, which must be installed beforehand on your computer.

```
## Retrieve the list of supported species in the Ensembl variation database
download-ensembl-variations -v 1 -available_species
```

We can now download all the variations available for the yeast.

```
## Retrieve the list of supported species in the Ensembl variation database
download-ensembl-variations -v 1 -species Saccharomyces_cerevisiae
```

10.3 Importing genomes from NCBI BioProject

The BioProject database hosts the results of genome sequencing and transcriptome projects.

1. Open a connection to the Bioproject Web site
<http://www.ncbi.nlm.nih.gov/bioproject>
2. Enter a query to select the organism of interest. E.g. `ostreococcus+tauri[orgn]`

¹<http://www.ensembl.org/>

²<http://ensemblgenomes.org/>

3. If the organism genome has been sequenced, you should see a title “Genome Sequencing Projects” in the record. Find the relevant project and open the link.
For example, for *Ostreococcus tauri*, the most relevant project is PRJNA51609
<http://www.ncbi.nlm.nih.gov/bioproject/51609>
4. Take note of the Accession of this genome project: since a same organism might have been sequenced several times, it will be useful to include this Accession in the suffix of the name of the file to be downloaded.
5. On the left side of the page, under Related information, click the link “Nucleotide genomic data”. This will display a list of Genbank entries (one per contig).
6. **Important:** we recommend to create one separate directory per organism, and to name this directory according to the organism name followed by the genome project Accession number. For example, for *Ostreococcus tauri*, the folder name would be *Ostreococcus_tauri_PRJNA51609*.

This convention will facilitate the further steps of installation, in particular the parsing of genbank-formatted files with the program ***parse-genbank.pl***.
7. In the top corner of the page, click on the Send to link and activate the following options.
Send to > File > Genbank full > Create file
Save the file in the organism-specific directory described in the previous step.
8. You can now parse the genome with the program ***parse-genbank.pl***. Note that ***parse-genbank.pl*** expects files with extension .gbk or .gbk.gz (as in the NCBI genome repository), whereas the BioProject genome appends the extension .gb. You should thus use the option -ext gb.

```
parse-genbank.pl -v 2 -i Ostreococcus_tauri_PRJNA51609 -ext gb
```

After parsing, run the program ***install-organism*** with the following parameters (adapt organism name).

```
install-organism -v 2 -org Ostreococcus_tauri_PRJNA51609 \  
-task config,phylogeny,start_stop,allup,seq_len_distrib \  
-task genome_segments,upstream_freq,oligos,dyads,protein_freq
```

10.4 Importing multi-genome alignment files from UCSC

10.4.1 Warning: disk space requirement

The UCSC multi-genome alignment files occupy a huge disk space. The alignments of 30 vertebrates onto the mouse genome (mm9 multiz30) requires 70Gb. If you intend to offer

support for multi-genome alignments, it might be safe to acquire a separate hard drive for this data.

The complete data set available at UCSC in April 2012 occupies 1Tb in compressed form, and probably 7 times more once uncompressed. For efficiency reasons, it is necessary to uncompress these files for using them with the indexing system of **peak-footprints**.

10.4.2 Checking supported genomes at UCSC

As a first step, we will check the list of supported genomes at the UCSC Genome Browser.

```
supported-organisms-ucsc
```

Each genome is associated with a short identifier, followed by a description. For example, several versions of the mouse genome are currently available.

```
mm10 Mouse Dec. 2011 (GRCm38/mm10) Genome at UCSC
mm9 Mouse July 2007 (NCBI37/mm9) Genome at UCSC
mm8 Mouse Feb. 2006 (NCBI36/mm8) Genome at UCSC
mm7 Mouse Aug. 2005 (NCBI35/mm7) Genome at UCSC
```

10.4.3 Downloading multiz files from UCSC

Multi-genome alignments at UCSC are generated with the program **multiz**, which produces files in a custom text format called *maf* for Multi-Alignment file.

We show hereafter the command to download the mm9 version of the mouse genome, and install it in the proper directory for **peak-footprints** (*\$RSAT/data/UCSC_multiz*).

```
download-ucsc-multiz -v 1 -org mm9
```

Beware: the download of all the multi-species alignments can take several hours for one genome.

The program will create the sub-directory for the mm9 genome, download the corresponding compressed multiz files (files with extension *.maf.gz*), uncompress them, and call **peak-footprint** with specific options in order to create a position index, which will be further used for fast retrieval of the conserved regions under peaks.

10.5 Installing genomes from NCBI/Genbank files

In the section 7, we saw that the genomes installed on the main **RSAT**server can easily be installed on your local site. In some cases, you would like to install additional genomes, which are not published yet, or which are not supported on the main **RSAT**server.

If your genomes are available in Genbank (files .gbk) or EMBL (files .embl) format, this can be done without too much effort, using the installation tools of **RSAT**.

The parsing of genomes from their original data sources is however more tricky than the synchronization from the **RSAT**server, so this procedure should be used only if you need to install a genome that is not yet supported.

If this is not your case, you can skip the rest of this section.

10.5.1 Organization of the genome files

In order for a genome to be supported, *RSAT* needs to find at least the following files.

1. organism description
2. genome sequences
3. feature tables (CDS, mRNA, ...)
4. lists of names/synonyms

From these files, a set of additional installation steps will be done by *RSAT* programs in order to compute the frequencies of oligonucleotides and dyads in upstream sequences.

If you installed *RSAT* as specified above, you can have a look at the organization of a supported genome, for example the yeast *Saccharomyces cerevisiae*.

```
cd ${RSAT}/public_html/data/genomes/Saccharomyces_cerevisiae/genome
ls -l
```

As you see, the folder *genome* contains the sequence files and the tables describing the organism and its features (CDSs, mRNAs, ...). The *RSAT* parser exports tables for all the feature types found in the original genbank file. There are thus a lot of distinct files, but you should not worry too much, for the two following reasons:

1. *RSAT* only requires a subset of these files (basically, those describing organisms, CDSs, mRNAs, rRNAs and tRNAs).
2. All these files can be generated automatically by *RSAT* parsers.

Organism description

The description of the organism is given in two separate files.

```
cd ${RSAT}/public_html/data/genomes/Saccharomyces_cerevisiae/genome
ls -l organism*.tab
```

```
more organism.tab
```

```
more organism_names.tab
```

1. *organism.tab* specifies the ID of the organism and its taxonomy. The ID of an organism is the TAXID defined by the NCBI taxonomical database, and its taxonomy is usually parsed from the .gbk files (but you may need to specify it yourself in case it would be missing in your own data files).
2. *organism_name.tab* indicates the name of the organism.

Genome sequence

A genome sequence is composed of one or more contigs. A contig is a contiguous sequence, resulting from the assembly of short sequence fragments obtained during the sequencing. When a genome is completely sequenced and assembled, each chromosome comes as a single contig.

In *RSAT*, the genome sequence is specified as one separate file per contig (chromosome) sequence. Each sequence file must be in raw format (i.e. a text file containing the sequence without any space or carriage return).

In addition, the genome directory contains one file indicating the list of the contig (chromosome) files.

```
cd $RSAT/data/genomes/Saccharomyces_cerevisiae/genome/

## The list of sequence files
cat contigs.txt

## The sequence files
ls -l *.raw
```

Feature table

The *genome* directory also contains a set of feature tables giving the basic information about gene locations. Several feature types (CDS, mRNA, tRNA, rRNA) can be specified in separate files (*cds.tab*, *mrna.tab*, *trna.tab*, *rrna.tab*).

Each feature table is a tab-delimited text file, with one row per feature (cds, mrna, ...) and one column per parameter. The following information is expected to be found.

1. Identifier
2. Feature type (e.g. ORF, tRNA, ...)
3. Name
4. Chromosome. This must correspond to one of the sequence identifiers from the fasta file.
5. Left limit
6. Right limit
7. Strand (D for direct, R for reverse complement)
8. Description. A one-sentence description of the gene function.

```
## The feature table
head -30 cds.tab
```

Feature names/synonyms

Some genes can have several names (synonyms), which are specified in separate tables.

1. ID. This must be one identifier found in the feature table
2. Synonym
3. Name priority (primary or alternate)

```
## View the first row of the file specifying gene names/synonyms
head -30 cds_names.tab
```

Multiple synonyms can be given for a gene, by adding several lines with the same ID in the first column.

```
## An example of yeast genes with multiple names
grep YFL021W cds_names.tab
```

10.5.2 Downloading genomes from NCBI/Genbank

The normal way to install an organism in *RSAT* is to download the complete genome files from the NCBI ³, and to parse it with the program *parse-genbank.pl*.

However, rather than downloading genomes directly from the NCBI site, we will obtain them from a mirror ⁴ which presents two advantages?

- Genome files are compressed (gzipped), which strongly reduces the transfer and storage volume.
- This mirror can be queried by *rsync*, which facilitates the updates (with the appropriate options, *rsync* will only download the files which are newer on the server than on your computer).

RSAT includes a makefile to download genomes from different sources. We provide hereafter a protocol to create a download directory in your account, and download genomes in this directory. Beware, genomes require a lot of disk space, especially for those of higher organisms. To avoid filling up your hard drive, we illustrate the protocol with the smallest procaryote genome to date: *Mycoplasma genitalium*.

```
## Creating a directory for downloading genomes in your home account
cd $RSAT
mkdir -p downloads
cd downloads
```

```
## Creating a link to the makefile which allows you to dowload genomes
ln -s $RSAT/makefiles/downloads.mk ./makefile
```

³<ftp://ftp.ncbi.nih.gov/genomes/>

⁴bio-mirror.net/biomirror/ncbigenomes/

We will now download a small genome from NCBI/Genbank.

```
## Downloading one directory from NCBI Genbank
cd $RSAT/downloads/
make one_genbank_dir NCBI_DIR=Bacteria/Mycoplasma_genitalium
```

We can now check the list of files that have been downloaded.

```
## Downloading one directory from NCBI Genbank
cd $RSAT/downloads/
ls -l ftp.ncbi.nih.gov/genomes/Bacteria/Mycoplasma_genitalium/
```

RSAT parsers only use the files with extension *.gbk.gz*.

You can also adapt the commande to download (for example) all the Fungal genomes in a single run.

```
## Downloading one directory from NCBI Genbank
cd $RSAT/downloads/
make one_ncbi_dir NCBI_DIR=Fungi
```

You can do the same for Bacteria, or for the whole NCBI genome repository, but this requires several Gb of free disk space.

10.5.3 Parsing a genome from NCBI/Genbank

The program ***parse-genbank.pl*** extract genome information (sequence, gene location, ...) from Genbank flat files, and exports the result in a set of tab-delimited files.

```
parse-genbank.pl -v 1 \
-i $RSAT/downloads/ftp.ncbi.nih.gov/genomes/Bacteria/Mycoplasma_genitalium
```

10.5.4 Parsing a genome from the Broad institute (MIT)

The website <http://www.broad.mit.edu/> offers a large collection of genomes that are not available on the NCBI website. We wrote a specific parser for the Broad files.

To this, download the following files for the organism of interest : the supercontig file, the protein sequences and the annotation file in the GTF format.

These files contain sometimes too much information that should be removed.

This is an example of the beginning of the fasta file containing the protein traduction. In this file, we should remove everything that follows the protein name.

```
>LELG_000001 | Lodderomyces elongisporus hypothetical protein (translation) (1085 aa)
MKYDTAAQLSLINPQTLKGLPIKPFPLSQPVFVQGVNNDTKAITQGVFLDVTVHFISLPA
IILYLHEQIPVGQVLLGLPFQDAHKLSIGFTDDGDKRELRFNRANGNIHKFPIRYDGDSNYH
IDSFPTVQVSQTVVIPPLSEMLRPAFTGSRASEDDIRYFVDQCAEVSDVFYIKGGDPGRL
```

This is an example of the beginning of the fasta file containing the contigs. In this file, we should remove everything that follows the name of the contig.

```
>supercontig_1.1 of Lodderomyces elongisporus
AAGAGCATCGGGCAAATGATGTTTTTCAGTCCATCAATGTGATGGATCTGATAGTTGAAG
GTCCTGATGAAGTTCAACCATTTGTAAACCCGATTTACAAAGTGTGAATTATCGAGTGGT
TTATTCATCACAAAGGACAAGAGCTTTGTTGGTTGACAGAGATGTTTTGCAGAAAGCCCTT
AAGGATGGTATTGCCTTGTTCAAGAAGAAACCAGTTGTTACTGAAGTAAATCTGACGACC
```

This is an example of the beginning of the GTF file containing the contigs annotation. We should rename the contig name so that it corresponds to the fasta file of contig. To this, we will remove the text in the name of the contig (only keep the supercontig number) and add a prefix.

```
supercont1.1%20of%20Lodderomyces%20elongisporus LE1_FINAL_GENECALL start_codon
322 324 . + 0 gene_id "LELG_00001"; transcript_id "LELT_00001";
supercont1.1%20of%20Lodderomyces%20elongisporus LE1_FINAL_GENECALL stop_codon
3574 3576 . + 0 gene_id "LELG_00001"; transcript_id "LELT_00001";
supercont1.1%20of%20Lodderomyces%20elongisporus LE1_FINAL_GENECALL exon 322
3576 . + . gene_id "LELG_00001"; transcript_id "LELT_00001";
supercont1.1%20of%20Lodderomyces%20elongisporus LE1_FINAL_GENECALL CDS 322 3573
. + 0 gene_id "LELG_00001"; transcript_id "LELT_00001";
```

We use the parse ***parse-broad-mit***.

```
parse-broad-mit.pl -taxid 36914 -org Lodderomyces_elongisporus \
-nuc_seq lodderomyces_elongisporus_1_supercontigs.fasta \
-gtf lodderomyces_elongisporus_1_transcripts.gtf \
-gtf_remove 'supercont' \
-gtf_remove '%20of%20Lodderomyces%20elongisporus' \
-contig_prefix LELG_ -nuc_remove supercontig_ \
-nuc_remove ' of Lodderomyces elongisporus' \
-aa lodderomyces_elongisporus_1_proteins.fasta -aa_remove '.*'
```

This will create the raw files, the feature files and the protein sequence file.

10.5.5 Updating the configuration file

After having parsed the genome, you need to perform one additional operation in order for ***RSAT*** to be aware of the new organism: update the configuration file.

```
install-organism -v 1 -org Mycoplasma_genitalium -task config

## Check the last lines of the configuration file
tail -15 $RSAT/data/supported_organisms.pl
```

From now on, the genome is considered as supported on your local ***RSAT*** site. You can check this with the command ***supported-organisms***.

10.5.6 Checking the start and stop codon composition

Once the organism is found in your configuration, you need to check whether sequences are retrieved properly. A good test for this is to retrieve all the start codons, and check whether

they are made of the expected codons (mainly ATG, plus some alternative start codons like GTG or TTG for bacteria).

The script ***install-organism*** allows you to perform some additional steps for checking the conformity of the newly installed genome. For example, we will compute the frequencies of all the start and stop codons, in order to check that gene locations were correctly parsed.

```
install-organism -v 1 -org Mycoplasma_genitalium -task start_stop

ls -l $RSAT/data/genomes/Mycoplasma_genitalium/genome/*start*

ls -l $RSAT/data/genomes/Mycoplasma_genitalium/genome/*stop*
```

The stop codons should be TAA, TAG or TGA, for any organism. For eucaryotes, all start codons should be ATG. For some procaryotes, alternative start codons (GTG, TGG) are found with some genome-specific frequency.

```
cd $RSAT/data/genomes/Mycoplasma_genitalium/genome/

## A file containing all the start codons
more Mycoplasma_genitalium_start_codons.wc

## A file with trinucleotide frequencies in all start codons
more Mycoplasma_genitalium_start_codon_frequencies

## A file containing all the stop codons
more Mycoplasma_genitalium_stop_codons.wc

## A file with trinucleotide frequencies in all stop codons
more Mycoplasma_genitalium_stop_codon_frequencies
```

10.5.7 Calibrating oligonucleotide and dyad frequencies with ***install-organisms***

The programs ***oligo-analysis*** and ***dyad-analysis*** require calibrated frequencies for the background models. These frequencies are calculated automatically with ***install-organism***.

```
install-organism -v 1 -org Debaryomyces_hansenii \
    -task allup,oligos,dyads,upstream_freq,protein_freq
```

Warning: this task may require several hours of computation, depending on the genome size. For the ***RSAT*** server, we use a PC cluster to regularly install and update genomes. As the task ***allup***, is a prerequisite for the computation of all oligonucleotide and dyad frequencies, it should be run directly on the main server before computing oligo and dyad frequencies on the nodes of the cluster. We will thus proceed in two steps. Note that this requires a PC cluster and a proper configuration of the batch management program.

```
## Retrieve all upstream sequences
## Executed directly on the server
```

```
install-organism -v 1 -org Debaryomyces_hansenii \
    -task allup

## Launch a batch queue for computing all oligo and dyad frequencies
## Executed on the nodes of a cluster
install-organism -v 1 -org Debaryomyces_hansenii \
    -task oligos,dyads,upstream_freq,protein_freq -batch
```

10.5.8 Installing a genome in your own account

In some cases, you might want to install a genome in your own account rather than in the *RSAT* folder, in order to be able to analyze this genome before putting it in public access.

In this chapter, we explain how to add support for an organism on your local configuration of *RSAT*. This assumes that you have the complete sequence of a genome, and a table describing the predicted location of genes.

First, prepare a directory where you will store the data for your organism. For example:

```
mkdir -p $HOME/rsat-add/data/Mygenus_myspecies/
```

Once you have this information, start the program ***install-organism***. You will be asked to enter the information needed for genome installation.

Updating your local configuration

- Modify the local config file
- You need to define an environment variable called `RSA_LOCAL_CONFIG`, containing the full path of the local config file.

10.6 Installing genomes from EMBL files

RSAT also includes a script ***parse-embl.pl*** to parse genomes from EMBL files. However, for practical reasons we generally parse genomes from the NCBI genome repository. Thus, unless you have a specific reason to parse EMBL files, you can skip this section.

The program ***parse-embl.pl*** reads flat files in EMBL format, and exports genome sequences and features (CDS, tRNA, ...) in different files.

As an example, we can parse a yeast genome sequenced by the “Genolevures” project ⁵.

Let us assume that you want to parse the genome of the species *Debaryomyces hansenii*.

Before parsing, you need to download the files in your account,

- Create a directory for storing the EMBL files. The last level of the directory should be the name of the organism, where spaces are replaced by underscores. Let us assume that you store them in the directory `$RSAT/downloads/Debaryomyces_hansenii`.

⁵<http://natchaug.labri.u-bordeaux.fr/Genolevures/download.php>

- Download all the EMBL file for the selected organism. Save each name under its original name (the contig ID), followed by the extension `.embl`)

We will check the content of this directory.

```
ls -l $RSAT/downloads/Debaryomyces_hansenii
```

On my computer, it gives the following result

```
CR382133.embl
CR382134.embl
CR382135.embl
CR382136.embl
CR382137.embl
CR382138.embl
CR382139.embl
```

The following instruction will parse this genome.

```
parse-embl.pl -v 1 -i $RSAT/downloads/Debaryomyces_hansenii
```

If you do not specify the output directory, a directory is automatically created by combining the current date and the organism name. The verbose messages will indicate you the path of this directory, something like *\$HOME/parsed_data/embl/20050309/Debaryomyces_hansenii*.

You can now perform all the steps above (updating the config file, installing oligo- and dyad frequencies, ...) as for genomes parsed from NCBI.

Installing a genome in the main *RSAT* directory

Once the genome has been parsed, the simplest way to make it available for all the users is to install it in the *RSAT* genome directory. You can already check the genomes installed in this directory.

```
ls -l $RSAT/data/genomes/
```

There is one subdirectory per organism. For example, the yeast data is in *\$RSAT/data/genomes/Saccharomyces_cerevisiae/*. This directory is further subdivided in folders: *genome* and *oligo-frequencies*.

We will now create a directory to store data about *Debaryomyces_hansenii*, and transfer the newly parsed genome in this directory.

```
## Create the directory
mkdir -p $RSAT/data/genomes/Debaryomyces_hansenii/genome

## Transfer the data in this directory
mv $HOME/parsed_data/embl/20050309/Debaryomyces_hansenii/* \
  $RSAT/data/genomes/Debaryomyces_hansenii/genome

## Check the transfer
ls -ltr $RSAT/data/genomes/Debaryomyces_hansenii/genome
```