DASSAULT SYSTEMES

DS BIOVIA

INSTALLATION AND CONFIGURATION GUIDE
R SOFTWARE FOR PIPELINE PILOT SERVER 2021
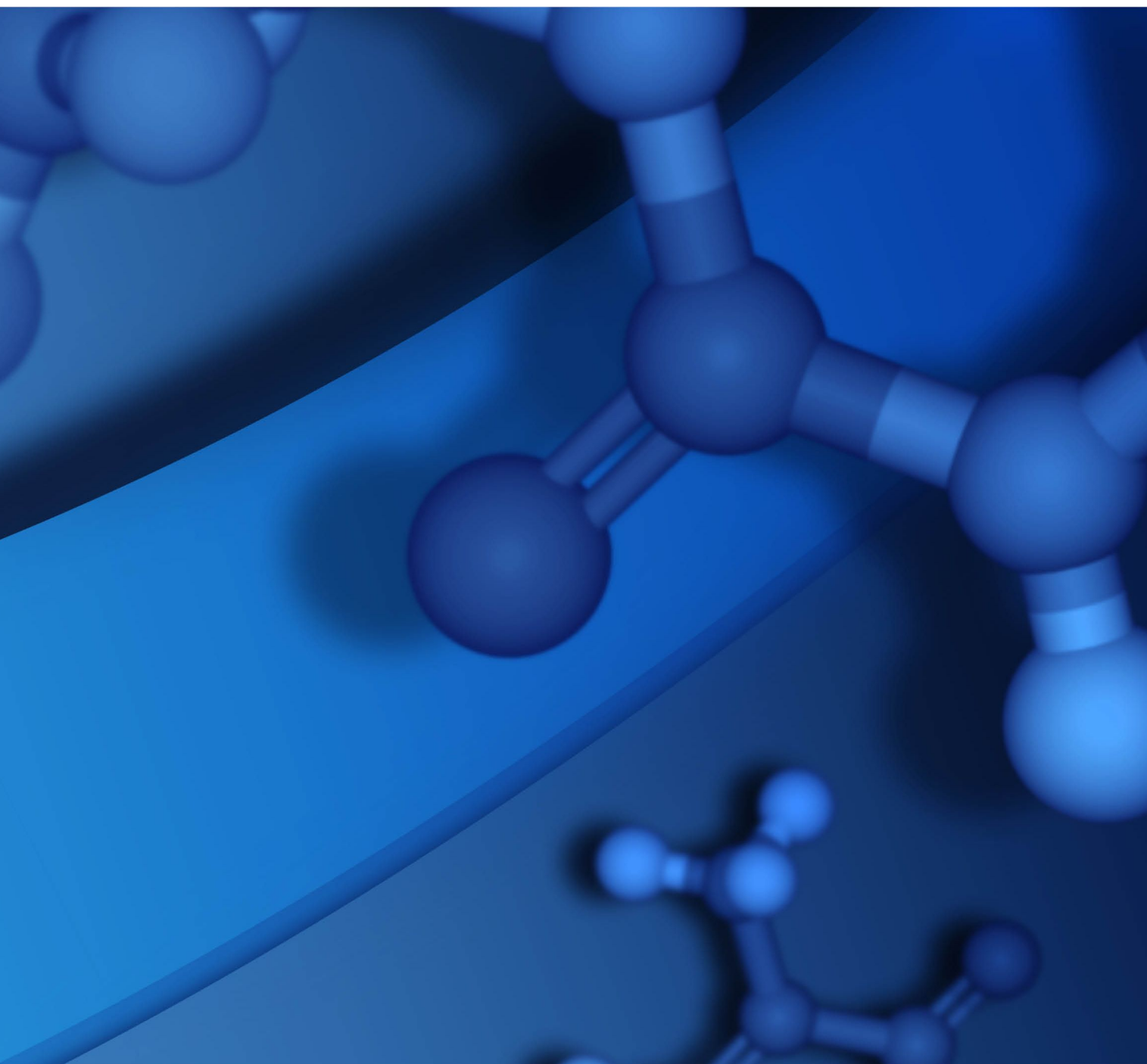
Acknowledgments and References

To print photographs or files of computational results (figures and/or data) obtained by using Dassault Systèmes software, acknowledge the source in an appropriate format. For example:

"Computational results were obtained by using Dassault Systèmes BIOVIA software programs. Pipeline Pilot Server was used to perform the calculations and to generate the graphical results."

Dassault Systèmes may grant permission to republish or reprint its copyrighted materials. Requests should be submitted to Dassault Systèmes Customer Support, either by visiting https://www.3ds.com/support/ and clicking **Call us** or **Submit a request**, or by writing to:

Dassault Systèmes Customer Support
10, Rue Marcel Dassault
78140 Vélizy-Villacoublay
FRANCE

Contents

Contents

Getting Started

About R Software

R is a free software environment for statistical computing and graphics, provided by the R Project for Statistical Computing at https://www.r-project.org/. R compiles and runs on a variety of platforms, including Windows and Linux. Some collections require R Software and other add-on packages. This software typically resides on a server that runs Pipeline Pilot or another server that is accessible to Pipeline Pilot.

Who should read this is guide

This guide is for system administrators who need to install and properly configure R Software on a server to ensure that all Pipeline Pilot collections are properly supported.

Guidelines for R Software Support

Since you may have more than one collection that requires R software, the information below summarizes the tasks required to install and configure R and other required R packages on your server.

IMPORTANT! Be sure to install and configure the software in the order described.

Do this:	Details:	Done?
1. Determine all requirements for your collections	In addition to using a supported version of the basic R software, you will also need to install additional R packages, (which are different for each collection).	☐
2. Download and install R software	Instructions are available for both Windows and Linux.	☐
3. Configure Pipeline Pilot for R Support	The Pipeline Pilot Server needs to know where the R executable is located, the name of the server running R (if you are using multiple servers). In addition, you can add global properties on the Pipeline Pilot Server to support some collections.	☐
4. Install all other required R packages for each collection	Many collections provide preconfigured protocols that you can run to automatically download and install other required packages. You can also install them manually.	☐

Requirements

> **Requirements:**
> - To install and configure R and any other required packages, you need to know how to run R commands.
>
> **Required Pipeline Pilot Permissions:**
> - *Platform/Administration/Logon* permissions
> - Group membership of *Platform/Administrators*
> - Read permission for all folders under Components and Protocols

R Software-dependent Collections

The following collections require R software:

- Analytical Instrumentation
- Gene Expression
- Mass Spec for Proteomics
- Next Generation Sequencing
- Plate Data Analytics
- R Statistics

> **IMPORTANT!** Not all collections support the same versions of R software, and many collections also require that additional third-party R packages are installed to work with R. To ensure that your server is configured with the proper software, review the following list of requirements for R and other packages.

Supported R Software Versions

Collection	Supported R Software Versions
Analytical Instrumentation	3.0.x 3.1.x 3.2.x 3.3.x 3.4.x 3.5.x
Gene Expression	3.5.x (supported with Bioconductor 3.7)
Mass Spec for Proteomics	3.5.x (supported with Bioconductor 3.7)
Next Generation Sequencing	3.5.x

Collection	Supported R Software Versions
Plate Data Analytics	2.9.x 2.10.x 2.12.x 2.15.x 3.0.x 3.1.x 3.2.x 3.3.x 3.4.x 3.5.x
QSAR Workbench	2.15.x 3.0.x 3.1.x 3.2.x 3.3.x 3.4.x 3.5.x
R Statistics	2.15.x 3.0.x 3.1.x 3.2.x 3.3.x 3.4.x 3.5.x

Note: On SuSE Linux 64-bit, graphical output is not supported for R Statistics components.

R package-dependent collections

The following packages are also required for some collections:

Collection	Required R Package
Analytical Instrumentation	TIMP, deSolve
Gene Expression	Bioconductor
Mass Spec for Proteomics	Bioconductor XCMS
Next Generation Sequencing	CNV-Seq
Plate Data Analytics	drc
R Statistics	Miscellaneous R packages

Tip: To streamline your installation, some of these packages can be installed automatically by running a Pipeline Pilot protocol. For further details, see Installing R Packages Automatically.

Downloading and Installing R Software

Accessing a mirror site

1. From your web browser, go to the R Project for Statistical Computing at http://www.r-project.org.
2. Select the CRAN link (left side under "Download").
3. From the CRAN Mirrors page that opens, locate a mirror site close to you, and click the appropriate link.

Windows installation

1. From the **Download and Install R** section (for pre-compiled binary distributions of the base system), select **Download R for Windows**.
2. Select **base**.
3. Perform <u>one</u> of the following tasks:

   If 3.4.x is the latest version:

   ➤ Save `R-3.4.x-win.exe` to disk.

   If version 3.4.x is not the latest version:

   a. Select the **Previous releases** link.
   b. Choose version 3.4.x or an earlier supported version on that page.
   c. Select the appropriate installer (`.exe` file).
   d. Save the file to disk.
4. Double-click the downloaded file to begin installing.

Linux installation

1. Select **Source code of older versions of R is available here**.
2. Select **R-3**.
3. To download the file, select the compressed tar file for the R release (for example, `R-3.4.x.tar.gz`).
4. For unpacking R and installing on Linux, follow the instructions in the *R Installation and Administration Guide* (available on the R Software web site). After extracting the `.tar` file, an HTML version of this document (`R-admin.html`) is available in the "doc/html" subdirectory of the R archive.

> **Note:** For R graphics to work, "devel" versions of the X11, JPEG, and PNG libraries must be installed on your Linux server. To determine if they are present, enter the following command:
> `rpm -qa | grep devel`
> If they are not available, ask your Linux system administrator to install them.

Configuring Pipeline Pilot for R Support

Use the Pipeline Pilot Administration Portal (Admin Portal) to define global variables, so your server knows where to find your local R installation. You might also need to define additional globals required by some collections. You can do this before or after installing the R packages required by your collections.

Adding the R Global Property

All collections that require R must define a **global property** that indicates the location of the R executable (R_Exe).

1. Log into the Admin Portal and go to **Setup > Global Properties**.

2. In the **Custom Global Protocol Properties** section, configure your R installation as follows:

   ■ In **Property**, enter the name of the R executable (for example, R_exe).

   ■ In **Value**, enter the location of the R executable (for example, C:\Program Files\R\R-3.4.0\bin\x64\R.exe or /usr/local/r-3.4.0/bin/R).

3. Click **Add**. The R configuration is listed as a new property in the Admin Portal.
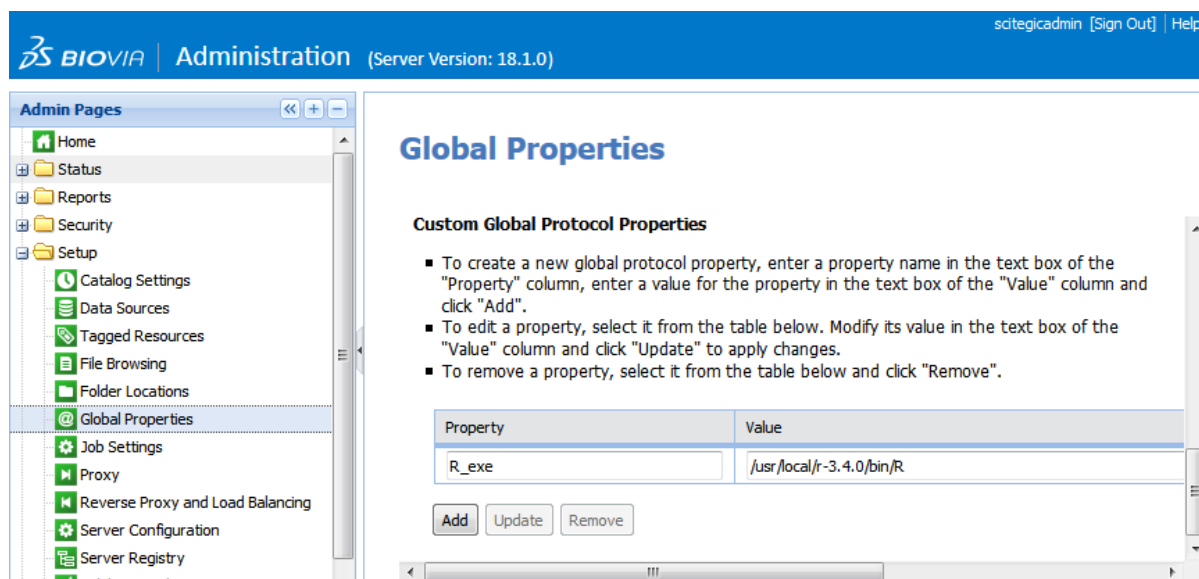
Setting custom global protocol properties in the Admin Portal to support R Software

4. (Optional) If R software is installed on a remote server (a different Pipeline Pilot server machine than your primary one), add the following global properties:

Property	Value
R_Run_ Remote	True
R_Server	Set the value to the name of the remote Pipeline Pilot Server running R in the format "<server name>:<port>"

Global Properties for Analytical Instrumentation

A component is available in the Analytical Instrumentation collection called *Extract Pure Components (TIMP)*. This component requires the TIMP and deSolve packages and can automatically download and install them for you, if you configure the following global properties in the Administration Portal:

Property	Value
`AIC_R_Downloads_Allowed`	`True`
`AIC_R_Library`	Provide the required location of the library. To use the default R library location, leave this value blank.

Global Properties for Bioconductor-dependent Collections

For collections that require Bioconductor (Gene Expression and Mass Spec for Proteomics), you may need to add the following optional global properties:

Property	Value
`BIO_R_Library`	Path to a particular directory for R to store Bioconductor packages (if you do not want R to install these packages in the default directory).
`BIO_R_Downloads_Allowed`	`True` by default to allow components to automatically download new Bioconductor packages. To prohibit automatic download and installation, set this value to `False`.

If you are installing R on multiple Pipeline Pilot Servers and running parallel subprotocols (the Gene Expression components that use Bioconductor are preconfigured to do this), set the following global properties:

Property	Value
`BIO_R_Run_Remote`	`True` (enables remote execution).
`BIO_R_Server`	Set the value to the name of the server running R in the format `<server name>:<port>`.
`BIO_R_Processes`	Number of remote servers to use.

IMPORTANT! If `BIO_R_Run_Remote` is set to `True` or if the above `BIO_R_*` global properties are not defined, the server runs R locally by default.

Tip: You can also use a *Custom Manipulator (PilotScript)* component to define global properties for Bioconductor packages. Use the following PilotScript expressions for this purpose:
```
@R_exe:= "/usr/local/bin/R";
@BIO_R_Run_Remote := true;
@BIO_R_Server := "ppprod:9944";
@BIO_R_Processes := 1;
@BIO_R_Downloads_Allowed := true;
```

Installing Required R Packages

After you install R and configure the Admin Portal to support it, you need to install additional R packages for each collection that uses R.

> **Tip:** Some collections provide protocols that you can run to automatically download and install the packages. You might also need to manually download and install packages for some collections.

Installing R Packages Automatically

If you have adequate permissions on your server, you can run **preconfigured installation protocols** that install the required packages for the following collections:

Collection	Protocol
Analytical Instrumentation	■ **Reaction Kinetics with TIMP:** Installs the required TIMP and deSolve packages required by *Extract Pure Components (TIMP)* component. (*Protocols > Examples > Analytical Instrumentation*) ■ **Extract Pure Components (TIMP):** Automatically installs the required TIMP and deSolve packages, if you first add a global variable (`AIC_R_Downloads_Allowed`) in the Admin Portal. (*Components > Laboratory > Analytical Instrumentation > Manipulators*) For further details, see Global Properties for Analytical Instrumentation.
Gene Expression	■ **Check Bioconductor Installation:** Installs the required Bioconductor packages. (*Protocols > Examples > Gene Expression > Utilities*) **Note:** This installation is directed by some R-related global properties. See Global properties for Bioconductor-dependent collections for further details.
Mass Spec Proteomics	■ **Check Bioconductor and XCMS Installation:** Installs the required Bioconductor packages. (*Protocols > Examples > Mass Spec Proteomics > Utilities*) **Note:** This installation is directed by some R-related global properties. See Global properties for Bioconductor-dependent collections for further details.
R Statistics	■ **Check R Installation:** Installs the required R packages. (*Protocols > Examples > R Statistics*)

Permissions for Automatic Installation

If you are running without impersonation, the Apache user needs permissions for installing software. If running with impersonation, the user running the protocol needs to have the required permissions to install the packages.

Possible workarounds include:

■ Set permissions on the default R Library directory to allow your user account to read and write packages to the library and then run the protocol.

■ Use a different directory that is readable by your user account and run the protocol.

- You can install the XCMS package manually for the Mass Spec Proteomics collection and the required Bioconductor packages for the Gene Expression collection.

Setting Permissions on the Default R Library Directory

The default library path depends on the configuration options specified when your version of R was compiled, linked, and installed. You can determine the path to the default library for your R installation by typing .Library on the R command line.

For example:

```
> .Library
[1]  "/usr/local/R-3.1.2/library"
```

Using a Different Directory

You can use a different directory that is readable by your user account. Choose a directory in your home directory (for example, /home/scitegicuser/rlib). In the Admin Portal, set the library path in the BIO_R_Library global property to this path (see Configuring Pipeline Pilot for R Support for details). After setting the BIO_R_Library to the desired path, Gene Expression and Mass Spec Proteomics components use that directory for installing and loading Bioconductor libraries.

> **Note:** Any libraries that are installed in the default location are still available to R. You can continue to use any R packages installed in that location.

Installing R Packages Manually

You can manually install all required R packages for your collections. The basic R software must first be installed and configured on your server before manually installing these packages.

> **IMPORTANT!** After installing the required packages, configure the Pipeline Pilot server to support R if you did not do so previously. For details, see Configuring Pipeline Pilot for R Support.

Analytical Instrumentation

In addition to the basic R installation, the following R packages are required by the *Extract Pure Components (TIMP)* component included with the Analytical Instrumentation collection:

- TIMP
- DeSolve

Installing R packages

After completing the basic R installation, use the following R commands to install the required additional packages:

```
install.packages("TIMP")
install.packages("deSolve")

options(download.file.method = "wget")
install.packages("https://cran.r-project.org/src/contrib/data.table_
1.10.4.tar.gz", repos=NULL, type="source")
install.packages("https://cran.r-
project.org/src/contrib/Archive/xgboost/xgboost_0.4-4.tar.gz", repos=NULL,
type="source")
```

Gene Expression

The Gene Expression collection requires R software with specific Bioconductor packages installed on the server. R software must be installed before installing Bioconductor packages.

To install the basic Bioconductor package:

➢ Go to http://master.bioconductor.org/install/.

Required Bioconductor packages

The following Bioconductor packages are required by some Gene Expression components and example protocols:

- affy
- affydata
- affyio
- affyPLM
- affyQCReport
- amap
- annaffy
- annotate
- AnnotationDbi
- Biobase
- biomaRt
- Biostrings
- bitops
- caTools
- convert
- ctc
- DBI

- DynDoc
- gcrma
- gdata
- genefilter
- geneplotter
- GEOquery
- GO.db
- goProfiles
- goTools
- gplots
- gtools
- hgu133acdf
- hgu133a.db
- hgu133b.db
- hgu133plus2.db
- hgu133plus2cdf
- hgu95av2.db

- IRanges
- KEGG.db
- limma
- makecdfenv
- marray
- multtest
- plier
- preprocessCore
- RColorBrewer
- RCurl
- RSQLite
- simpleaffy
- vsn
- XML
- xtable

Note: Because some Bioconductor packages contain source code that must be built locally, the Linux development tools and packages must be installed on your system. For Red Hat systems, the required RPM packages should be present in your Red Hat installation.

Installing R packages

You can use the following R script to install the packages manually:

```
source("http://bioconductor.org/biocLite.R")
biocLite("affy", lib=.Library)
biocLite("affydata", lib=.Library)
biocLite("affyio", lib=.Library)
biocLite("affyPLM", lib=.Library)
biocLite("affyQCReport", lib=.Library)
biocLite("amap", lib=.Library)
biocLite("annaffy", lib=.Library)
biocLite("annotate", lib=.Library)
biocLite("AnnotationDbi", lib=.Library)
biocLite("Biobase", lib=.Library)
```

```
biocLite("biomaRt", lib=.Library)
biocLite("Biostrings", lib=.Library)
biocLite("bitops", lib=.Library)
biocLite("caTools", lib=.Library)
biocLite("convert", lib=.Library)
biocLite("ctc", lib=.Library)
biocLite("DBI", lib=.Library)
biocLite("DynDoc", lib=.Library)
biocLite("gcrma", lib=.Library)
biocLite("gdata", lib=.Library)
biocLite("genefilter", lib=.Library)
biocLite("geneplotter", lib=.Library)
biocLite("GEOquery", lib=.Library)
biocLite("GO.db", lib=.Library)
biocLite("goProfiles", lib=.Library)
biocLite("goTools", lib=.Library)
biocLite("gplots", lib=.Library)
biocLite("gtools", lib=.Library)
biocLite("hgu133acdf", lib=.Library)
biocLite("hgu133a.db", lib=.Library)
biocLite("hgu133b.db", lib=.Library)
biocLite("hgu133plus2.db", lib=.Library)
biocLite("hgu133plus2cdf", lib=.Library)
biocLite("hgu95av2.db", lib=.Library)
biocLite("IRanges", lib=.Library)
biocLite("KEGG.db", lib=.Library)
biocLite("limma", lib=.Library)
biocLite("makecdfenv", lib=.Library)
biocLite("marray", lib=.Library)
biocLite("multtest", lib=.Library)
biocLite("plier", lib=.Library)
biocLite("preprocessCore", lib=.Library)
biocLite("RColorBrewer", lib=.Library)
biocLite("RCurl", lib=.Library)
biocLite("RSQLite", lib=.Library)
biocLite("simpleaffy", lib=.Library)
biocLite("vsn", lib=.Library)
biocLite("XML", lib=.Library)
biocLite("xtable", lib=.Library)
```

> **Tip:** Load the CDF and annotation packages for your additional Affymetrix chips. When manually installing packages, be sure to install all packages in advance so the reader components in this collection do not automatically download chip definitions for new microarray types.

Mass Spec for Proteomics

The Mass Spec for Proteomics collection requires R software with specific Bioconductor packages installed on the server. Be sure to install all required software in the following order:

1. R software

2. Bioconductor

3. XCMS

Installing Bioconductor

To install the basic Bioconductor package:

➢ Go to http://master.bioconductor.org/install/.

Installing the XCMS package

To download and install on Windows:

1. Start R.

2. Enter the following:

```
source("http://bioconductor.org/biocLite.R")
biocLite("xcms")
```

3. To check whether the XCMS package can be loaded successfully, enter `library(xcms)`.

Note: This collection requires installations of both R and the Bioconductor XCMS package. To install the necessary XCMS package automatically, run the *Check XCMS Installation* protocol located in the "Utilities" folder for the MSP example protocols. The version of XCMS will coordinate with your version of R.

To download and install on Linux:

1. Install the **NetCDF library**. Your Linux distributions may already have the library. Otherwise, follow the steps below to install it:

   a. Download a suitable binary release for Linux from NetCDF-3 Binary Distributions.

   b. Unpack the file.

   c. Copy all files under the `include` folder to `/usr/local/include`.

   d. Copy all files under the `lib` folder to `/usr/local/lib`.

2. Start R.

3. Enter the following:

```
source("http://bioconductor.org/biocLite.R")
biocLite("xcms")
```

4. To check whether the XCMS package can be loaded successfully, enter `library(xcms)`.

Note: If `biocLite("xcms")` fails because XCMS cannot be built upon the netCDF library, you need to download the netCDF source code from NetCDF-3 Distributions and build the library on your machine. You may also need to add a symbolic link in `/usr/lib` (for Linux 32) or */usr/lib64* (for Linux 64) to `/usr/local/lib/libnetcdf.so.4`.

Plate Data Analytics

The *R Dose Response Fit (DRC)* component in the Plate Data Analytics collection requires R software and an additional R package on the server. R software must be installed before installing the package.

To install the required package, start R and run the following command:

```
install.packages("drc")
```

Next Generation Sequencing

The *Detect CNV (CNV-seq) for Region* of the Next Generation Sequencing collection requires R software and the CNV-Seq R package.

To install CNV:

1. On the command line, go to `<pps_install>/apps/scitegic/nextgenseq/lang/R/cnv-seq`.

2. Run the command:

```
R CMD INSTALL cnv/
```

The *Detect CNV (CNV-seq) for Region* component can run in parallel by setting NGS global properties such as `NGS_Run_Remote` and `NGS_Server`. In this case, all the servers specified by `NGS_Server` must follow the above steps to have CNV R package installed.

R Statistics

The R Statistics components in the Analytics and Machine Learning collection require R software with specific packages installed on the server. R software must be installed before installing the required packages specified below.

Required Packages

In addition to the basic R installation, the following R packages are required by some R Statistics components:

- AlgDesign
- apcluster
- brglm (optional)
- car
- cluster
- data.table (required by xgboost)
- deepnet
- e1071
- FrF2 (on Linux, requires that R is built with Tcl/Tk support)
- glmpath (optional)
- igraph
- impute
- mda
- mixexp
- pls
- randomForest
- Rtsne
- sas7bdat (optional; needed for the *Read a sas7bdat File with R Custom Script* example protocol)
- scatterplot3d
- superpc
- xgboost

Xvfb

On Linux servers, R Statistics components require Xvfb, which is part of Xwindows. For certain Linux distributions, Xvfb may not be part of the standard installation. If so, you will need to download it from a Linux resource site.

Alternatively, you can avoid the requirement for Xvfb by building R with support for Cairo graphics, first making sure all required libraries (for example, Cairo and Pango devel libraries) are present on your system.

Installing R Packages

After completing the basic R installation, use the following R commands to install the required additional packages:

```
install.packages("AlgDesign")
install.packages("apcluster")
install.packages("brglm")
install.packages("car")
install.packages("cluster")
install.packages("data.table")
install.packages("deepnet")
install.packages("e1071")
install.packages("FrF2")
install.packages("glmpath")
install.packages("igraph")
```

To install "impute" in the latest versions of R, you must get the package from Bioconductor as follows:

```
source(http://bioconductor.org/biocLite.R)
biocLite("impute", suppressUpdates=TRUE)

install.packages("mda")
install.packages("mixexp")
install.packages("pls")
install.packages("randomForest")
install.packages("Rtsne")
install.packages("sas7bdat")
install.packages("scatterplot3d")
install.packages("superpc")
install.packages("xgboost")
```

Note: glmpath is an optional package for the *Learn R Generalized Linear Model* component. For R 2.6.1 and later versions that run on Windows, you must compile this package yourself if you want to use it.

IMPORTANT!
The xgboost package is present on CRAN for R 3.1.x and R 3.3.x and above, but not for R 3.2.x. If you are running R 3.2.x, you need to do the following to install xgboost and data.table:

■ On Windows, download xgboost_0.4-4.zip and use the "Install package from local zip file" menu option.

■ On Linux, run the following commands from the R console:

```
options(download.file.method = "wget")
install.packages("https://cran.r-project.org/src/contrib/data.table_
1.10.4.tar.gz", repos=NULL, type="source")
install.packages("https://cran.r-
project.org/src/contrib/Archive/xgboost/xgboost_0.4-4.tar.gz",
repos=NULL, type="source")
```

Dassault Systèmes Support Resources

For additional resources or to contact Dassault Systèmes Customer Support, visit the Support portal:

https://www.3ds.com/support/

From this portal, you can:

- Call or email Dassault Systèmes Customer Support
- Submit a request
- Download installers
- Access hardware and software requirements
- Access Knowledge Base
- Access Communities and Twitter feeds