

# Basic Functions of AnnBuilder

Jianhua Zhang

August 22, 2004

©2003 Bioconductor

## 1 Introduction

This vignette is an overview of some of the functions that can be used to build an annotation data package. The purpose of this vignette is to provide guidance for users who are comfortable using the data package building procedures described in **ABPrimer** but would like to have more freedom in building customized data packages. First time users of **AnnBuilder** are suggested to go through the **ABPrimer** vignette before trying the code here.

Functions contained by *AnnBuilder* include:

```
> library(AnnBuilder)
> pkgpath <- .find.package("AnnBuilder")
> docFiles <- file.path(pkgpath, c("TITLE", "DESCRIPTION", "INDEX"))
> headers <- c("", "Description:\n\n", "Index:\n\n")
> footers <- c("\n", "\n", "")
> for (i in which(file.exists(docFiles))) {
+   writeLines(headers[i], sep = "")
+   writeLines(readLines(docFiles[i]))
+   writeLines(footers[i], sep = "")
+ }
```

Description:

Package: AnnBuilder

Version: 1.4.2

Date: 2004-05-04

Title: Bioconductor annotation data package builder

Author: J. Zhang <jzhang@jimmy.harvard.edu>

Maintainer: J. Zhang <jzhang@jimmy.harvard.edu>

Depends: R (>= 1.9.0), Biobase, XML, annotate  
 Description: Processing annotation data from public data repositories  
 and building annotation data packages or XML data documents  
 using the source data.  
 Keyword: annotation  
 License: LGPL  
 Packaged: Wed Jun 2 12:14:40 2004; madman  
 Built: R 1.9.1; ; 2004-08-22 18:12:57; unix

#### Index:

ABPkgBuilder	Functions that support a single API for building data packages
GEO-class	Class "GEO" represents a GEO object that reads/downloads data from the GEO web site
GO-class	Class "GO" a class to handle data from Gene Ontology
GOPkgBuilder	Functions to build a data package using GO data
GOXMLParser	Functions to read/parse the XML document of Gene Ontology data
GP-class	Class "GP" a sub-class of pubRepo to get/process data from GoldenPath
KEGG-class	Class "KEGG" a sub-class of pubRepo to get/process pathway and enzyme information
KEGGPkgBuilder	A function to make the data package for KEGG
LL-class	Class "LL" a sub-class of pubRepo to handle data from LocusLink
MeSHParser	Function to parse the XML data file from MeSH
SPPkgBuilder	A function to build a data package using Swiss-Prot protein data
UG-class	Class "UG" a sub-class of pubRepo to handle data from UniGene
YG-class	Class "YG" a sub-class of pubRepo that reads/downloads data from yeast genomic
chrLocPkgBuilder	A function to build a data package containing mappings between LocusLink ids and the chromosomal locations of genes represented by the LocusLink ids
cols2Env	Creates a environment object using data from two columns of a matrix
fileMuncher	Dynamically create a Perl script to parse a

	source file base on user specifications
fileToXML	A function to convert a text file to XML.
getChroLocation	Functions to extract data from Golden Path
getDPStats	Functions to read in the statistics about a data package
getKEGGIDNName	Functions to get/process pathway and enzyme data from KEGG
getSrcBuilt	Functions that get the built date or number of the source data used for annotation
getSrcUrl	Functions that find the correct url for downloading annotation data
getYeastData	Functions to get/process yeast genome data
homoData-class	Class "homoData"
homoPkgBuilder	Functions to build a homology data package using data from NCBI
loadFromUrl	Functions to load files from a web site
makeSrcInfo	Functions to make source information available for later use
map2LL	A function that maps LocusLink ids to other public repository ids and vice versa
print.ABQCList	Prints the quality control results for a given data package in a nice format
pubRepo-class	Class "pubRepo" a generic class for downloading/parsing data provided by various public data repositories
queryGEO	Function to extract a data file from the GEO web site
resolveMaps	Functions to obtain unified mappings for a given set of ids using various sources
sourceURLs	A data file contains urls for data available from various public repositories
unifyMappings	A function to unify mapping result from different sources
writeChrLength	Functions that creates binary files for chromosome length and organism
writeManPage	Functions that write supporting files needed by a data package
writeXMLHeader	A function to write header information to an XML file.
yeastAnn	Functions to annotate yeast genom data
yeastPkgBuilder	Functions to do a data package for yeast genome

*AnnBuilder* relies on these functions to build annotation data packages by extracting data from the following potential public data repositories.

- LocusLink([ftp://ftp.ncbi.nih.gov/refseq/LocusLink/LL\\_tmpl.gz](ftp://ftp.ncbi.nih.gov/refseq/LocusLink/LL_tmpl.gz)) to obtain mappings to LocusLink ids and annotation data.
- UniGene(<ftp://ftp.ncbi.nih.gov/repository/UniGene/Hs.data.gz>) to obtain mappings to LocusLink ids from ESTs
- GoldenPath(<http://www.genome.ucsc.edu/goldenPath/14nov2002/database>). Two data files ([refLink.txt.gz](#) and [refGene.txt.gz](#)) are used to obtain chromosomal location and orientation data
- Gene Ontology([http://www.godatabase.org/dev/database/archive/2003-02-01/go\\_200302-termdb.xml.gz](http://www.godatabase.org/dev/database/archive/2003-02-01/go_200302-termdb.xml.gz)) to obtain gene ontology terms and relationships among terms.
- KEGG(<ftp://ftp.genome.ad.jp/pub/kegg/pathways>) to obtain pathway and enzyme data for genes. Several data files may be used depending on the organism of interest.

HomoloGene A data file provided by <ftp://ftp.ncbi.nih.gov/pub/HomoloGene/> will be used to extract mappings between LocusLink ids and HomoloGene ids.

The urls with date components may change when the maintainers update the data. However, *AnnBuilder* has the ability to figure out the latest updates and use the corresponding data for annotation as long as the current path structure of the urls remain. Source data will be downloaded from the urls given.

Each of the public data repositories is represented as an object of a S4 class. Common methods for an object include `readData` that reads in data line by line and `parseData` that parses data based on the instructions given in a segment of Perl code. In both cases, data are downloaded from the source url and then processed locally.

As data from the aforementioned data sources are usually large, truncated versions of the corresponding data will be used to ensure reasonable speed. These files have already been stored in Bioconductor web site. Thus, the source urls will be different for a real annotation project.

## Getting Source Data

Suppose we are interested in annotating genes on Affymetrix HG\_U95av2 gene chip. A file containing a column for Affymetrix probe ids and another for mappings to GenBank accession numbers can be produced based on the data file provided by Affymetrix and then used as the base to extract annotation data from different data sources. The base file has to be saved as a text file with the two columns separated by a delimiter (e. g.

a tab - "␣"). Here we just create a truncated one on the fly and store it in the current working directory.

```
> geneNMap <- matrix(c("32468_f_at", "D90278;M16652", "32469_at",
+ "L00693", "32481_at", "AL031663", "33825_at", "X68733", "35730_at",
+ "X03350", "36512_at", "L32179", "38912_at", "D90042", "38936_at",
+ "M16652", "39368_at", "AL031668"), ncol = 2, byrow = TRUE)
> write.table(geneNMap, file = "geneNMap", sep = "\t", quote = FALSE,
+ row.names = FALSE, col.names = FALSE)
```

We can see that the file has two columns for Affymetrix probe ids and the matching GenBank accession numbers:

```
> geneNMap

      [,1]      [,2]
[1,] "32468_f_at" "D90278;M16652"
[2,] "32469_at"   "L00693"
[3,] "32481_at"   "AL031663"
[4,] "33825_at"   "X68733"
[5,] "35730_at"   "X03350"
[6,] "36512_at"   "L32179"
[7,] "38912_at"   "D90042"
[8,] "38936_at"   "M16652"
[9,] "39368_at"   "AL031668"
```

The first step to annotating these probe ids in the base file is to map them to LocusLink ids and then use mapped LocusLink ids as the point of linkage to other annotation data provided by various data sources. As Affymetrix probe ids (probes for other platform as well) may be mapped to LocusLink ids through LocusLink and UniGene (and other sources), each of which can be complementary to each other, we may want to get the mappings from all the available sources and then combine the results to ensure completeness. *Annbuilder* has a unifying mechanism that allows users to unify mapping information from different sources to obtained a combined result that is assumed to be more reliable.

In this vignette, we first would like to map the probes to LocusLink ids using data from both LocusLink and UniGene. The following code creates objects for LocusLink and UniGene with parsers needed to parse the source data file for mapping Affymetrix probe ids in baseF to LocusLink ids. *AnnBuilder*.

```
> makeSrcInfo()
> llUrl <- "http://www.bioconductor.org/datafiles/wwwsources/T11_tmpl.gz"
> ugUrl <- "http://www.bioconductor.org/datafiles/wwwsources/Ths.data.gz"
```

```
> ll <- LL(srcUrl = llUrl, parser = file.path(pkgpath, "scripts",
+       "gbLLParser"), baseFile = "geneNMap")
> ug <- UG(srcUrl = ugUrl, parser = file.path(pkgpath, "scripts",
+       "gbUGParser"), baseFile = "geneNMap", organism = "human")
```

Again, the urls used in the example are for demonstration purpose only. ll and ug objects also take a parser as an argument. A parser is a segment of a Perl script that contains instructions on how the data source will be parsed and how the output will be generated. Please refer to the documents for pubRepo for detailed information on parsers and the objects for various public data repositories. Each object has a function named `parseData` that can be invoked to obtain the parsed data. `parseData` has a parameter - `fromWeb` that should be set to `FALSE` if the source data has been stored locally. The following code needs human intervention under windows and is therefore turned off. Copying the code chunk and then pasting into an R session under windows should work.

```
> if (.Platform$OS.type != "windows") {
+   llMapping <- parseData(ll, fromWeb = TRUE)
+   colnames(llMapping) <- c("PROBE", "LL")
+   ugMapping <- parseData(ug, fromWeb = TRUE)
+   colnames(ugMapping) <- c("PROBE", "UG")
+ }
```

The parsed data from LocusLink and UniGene are:

```
> if (.Platform$OS.type != "windows") {
+   llMapping
+   ugMapping
+ }
```

	PROBE	UG
32468_f_at	"32468_f_at"	"1084;63036"
32469_at	"32469_at"	"10;1084"
32481_at	"32481_at"	"7051"
35730_at	"35730_at"	"125"
36512_at	"36512_at"	"1084"
38912_at	"38912_at"	"10;NA"
38936_at	"38936_at"	"63036"
39368_at	"39368_at"	"NA"

Please note the differences between the mappings from the two sources and some of the Affymetrix probe ids can be mapped to multiple LocusLink ids and ";" is used to separate multiple mappings in such cases.

The mappings obtained from the two sources are then unified to obtain a comprehensive mapping between Affymetrix probe ids and LocusLink ids. The unified mappings are saved in a file show below:

```
> if (.Platform$OS.type != "windows") {
+   base <- matrix(scan("geneNMap", what = "", sep = "\t", quote = "",
+     quiet = TRUE), ncol = 2, byrow = TRUE)
+   colnames(base) <- c("PROBE", "ACC")
+   merged <- merge(base, llMapping, by = "PROBE", all.x = TRUE)
+   merged <- merge(merged, ugMapping, by = "PROBE", all.x = TRUE)
+   unified <- resolveMaps(merged, trusted = c("LL", "UG"), srcs = c("LL",
+     "UG"))
+   unified
+ }
```

```
[1] "/BIOCPkg/lib/AnnBuilder/temp/tempFile60e33853"
```

In the above code, "LL" has been identified as the trusted source meaning that when the two sources provide conflicting mappings, the one from LocusLink will be used. The unified mapping has four columns with the first one for Affymetrix probe ids, second for GenBank accession numbers, third for mappings to LocusLink ids, and forth for the number of sources that agreed with the mappings.

```
> if (.Platform$OS.type != "windows") {
+   read.table(unified, sep = "\t", header = FALSE)
+ }
```

	V1	V2	V3	V4
1	32468_f_at	D90278;M16652	1084	1
2	32469_at	L00693	10	1
3	32481_at	AL031663	7051	1
4	33825_at	X68733	12	1
5	35730_at	X03350	125	1
6	36512_at	L32179	10	1
7	38912_at	D90042	10	1
8	38936_at	M16652	63036	1
9	39368_at	AL031668	NA	4

The unified mappings can then be used as the base file to parse the data from LocusLink to obtain annotation data for each of the Affymetrix probe ids. To do so, we need to assign a new parser that processes the data from LocusLink to get annotation data including gene name, chromosomal location, and so on.

```

> if (.Platform$OS.type != "windows") {
+   parser(ll) <- file.path(.path.package("AnnBuilder"), "scripts",
+     "llParser")
+   baseFile(ll) <- unified
+   annotation <- parseData(ll, ncol = 15, fromWeb = TRUE)
+   colnames(annotation) <- c("PROBE", "ACCNUM", "LOCUSID", "UNIGENE",
+     "GENENAME", "SYMBOL", "CHR", "MAP", "PMID", "GRIF", "SUMFUNC",
+     "GO", "OMIM", "NM", "NP")
+ }

```

The annotation data obtained has 12 columns for the elements indicated by the column names. Let us view the chromosomal number of the Affymetrix probe ids.

```

> if (.Platform$OS.type != "windows") {
+   annotation[, c("PROBE", "LOCUSID")]
+ }

```

	PROBE	LOCUSID
32468_f_at	"32468_f_at"	"1084"
32469_at	"32469_at"	"10"
32481_at	"32481_at"	"7051"
33825_at	"33825_at"	"12"
35730_at	"35730_at"	"125"
36512_at	"36512_at"	"10"
38912_at	"38912_at"	"10"
38936_at	"38936_at"	"63036"
39368_at	"39368_at"	"NA"

Other annotation data can be obtained from other sources. In this vignette, we try to get data from GoldenPath for chromosomal location and orientation and Gene Ontology for ontology terms and relations among terms. As usual, we create the objects with truncated data from Bioconductor rather than the actual web site. Two source data files (Tlink.txt.gz and TGene.txt.gz) have to be downloaded/unzipped from GoldenPath (<http://www.genome.ucsc.edu/goldenPath/10april2003/database/>) in order to obtain the chromosome location data. We only have to provide the url under unix as the system knows how to get the latest version of the two files.

```

> if (.Platform$OS.type != "windows") {
+   gpUrl <- "http://www.bioconductor.org/datafiles/wwwsources/"
+   goUrl <- "http://www.bioconductor.org/datafiles/wwwsources/Tgo.xml"
+   gp <- GP(srcUrl = gpUrl, organism = "human")
+   go <- GO(srcUrl = goUrl)
+ }

```



To get the chromosomal data from GoldenPath with the actual url, one only needs to call a function called `getStrand` by typing "`strand <- getStrand(gp)`" where `gp` is the object for goldenPath with correct url. In this vignette, however, we take a somewhat different approach to get the data as we are using a truncated set of data from a dummy.

```
> if (.Platform$OS.type != "windows") {
+   strand <- getChroLocation(srcUrl(gp), gpLinkNGene(TRUE),
+     fromWeb = TRUE)
+ }
```

The data processed are then merged with the annotation we previously obtained.

```
> if (.Platform$OS.type != "windows") {
+   annotation <- merge(annotation, strand, by = "LOCUSID", all.x = TRUE)
+ }
```

Data from yet some other sources can be processed and merged to the annotation data following the same procedures as described above. When data from all the desired sources have been obtained, an XML data document and an R data package can be produced for easy distribution and usage. The following code generates an XML document containing the annotation data. `textttmultC` is used to identify elements (columns in data object `textttannotation`) that have many to one relations to probe ids. When the data were parsed by the parser, ";" was used to separate multiple entries for these elements (e.g. a probe id may be related to several PubMed ids as shown by 256352;63254;4687264). `texttttypeC` identifies elements that also have many to one relation to probe ids. Additionally, they also contain another attribute that are appended to the end of the mapped value with "@" as a separator (e.g. GO:0001234@E;GO:0004875@NR). These elements need to be dealt with differently.

```
> if (.Platform$OS.type != "windows") {
+   multC <- c("PMID", "CHR", "OMIM")
+   typeC <- c("GO", "CHRLoc")
+   XMLOut <- file.path(getwd(), "test.xml")
+   fileToXML(targetName = "hgu95a", outName = XMLOut, inName = annotation,
+     colNames = "", idColName = "PROBE", multColNames = multC,
+     typeColNames = typeC, isFile = FALSE, version = "1.0.0")
+ }
```

The contents of the XML document generated are shown below.

```
> if (.Platform$OS.type != "windows") {
+   readLines(XMLOut)
+ }
```

```

[1] "<?xml version = \"1.0\" encoding = \"UTF-8\" standalone = \"yes\"?>"
[2] "<!DOCTYPE AnnBuilder: SYSTEM \"http://www.bioconductor.org/datafiles/dtds/annota"
[3] "<AnnBuilder:Annotate xmlns:AnnBuilder = 'http://www.bioconductor.org/AnnBuilder/"
[4] "<AnnBuilder:Attr>"
[5] "<AnnBuilder:Target value = \"hgu95a\"/>"
[6] "<AnnBuilder:DateMade value = \"Sun Aug 22 18:15:04 2004\"/>"
[7] "<AnnBuilder:Version value = \"1.0.0\"/>"
[8] "<AnnBuilder:SourceFile url = \"ftp://ftp.ncbi.nih.gov/refseq/LocusLink/LL_tmpl.g"
[9] "<AnnBuilder:SourceFile url = \"http://www.godatabase.org/dev/database/archive/20"
[10] "<AnnBuilder:SourceFile url = \"ftp://ftp.genome.ad.jp/pub/kegg/pathways\" built"
[11] "<AnnBuilder:SourceFile url = \"http://www.genome.ucsc.edu/goldenPath/hg17/databa"
[12] "<AnnBuilder:SourceFile url = \"ftp://ftp.ncbi.nih.gov/repository/UniGene/Hs.data"
[13] "<AnnBuilder:Entryid value = \"LocusLink identifier\"/>"
[14] "<AnnBuilder:Element value = \"PROBE\" describ = \"Generic identifier\"/>"
[15] "<AnnBuilder:Element value = \"ACCNUM\" describ = \"GenBank accession number\"/>"
[16] "<AnnBuilder:Element value = \"UNIGENE\" describ = \"UniGene cluster identifier a"
[17] "<AnnBuilder:Element value = \"GENENAME\" describ = \"Gene Description used for g"
[18] "<AnnBuilder:Element value = \"SYMBOL\" describ = \"Symbol used for gene reports\"/"
[19] "<AnnBuilder:Element value = \"CHR\" describ = \"Chromosome assignment\"/>"
[20] "<AnnBuilder:Element value = \"MAP\" describ = \"Cytoband location of gene \"/>"
[21] "<AnnBuilder:Element value = \"PMID\" describ = \"PubMed unique identifier\"/>"
[22] "<AnnBuilder:Element value = \"GRIF\" describ = \"PubMed unique identifier\"/>"
[23] "<AnnBuilder:Element value = \"SUMFUNC\" describ = \"A brief summary of the funct"
[24] "<AnnBuilder:Element value = \"GO\" describ = \"Gene Ontology identifier\"/>"
[25] "<AnnBuilder:Element value = \"OMIM\" describ = \"MIM (Mendelian Inheritance in M"
[26] "<AnnBuilder:Element value = \"NM\" describ = \"RefSeq accession for a mRNA recor"
[27] "<AnnBuilder:Element value = \"NP\" describ = \"RefSeq accession for a protein re"
[28] "<AnnBuilder:Element value = \"CHRLOC\" describ = \"Chromosomal location \"/>"
[29] "</AnnBuilder:Attr>"
[30] "<AnnBuilder:Data>"
[31] "<AnnBuilder:Entry id=\"32469_at\" describ = \"PROBE\">"
[32] "\t<AnnBuilder:Item name=\"LOCUSID\" value=\"10\" />"
[33] "\t<AnnBuilder:Item name=\"ACCNUM\" value=\"L00693\" />"
[34] "\t<AnnBuilder:Item name=\"UNIGENE\" value=\"Hs.2\" />"
[35] "\t<AnnBuilder:Item name=\"GENENAME\" value=\"N-acetyltransferase 2 (arylamine N-"
[36] "\t<AnnBuilder:Item name=\"SYMBOL\" value=\"NAT2\" />"
[37] "\t<AnnBuilder:Item name=\"MAP\" value=\"NA\" />"
[38] "\t<AnnBuilder:Item name=\"GRIF\" value=\"NA\" />"
[39] "\t<AnnBuilder:Item name=\"SUMFUNC\" value=\"Arylamine N-acetyltransferase 2; N-"
[40] "\t<AnnBuilder:Item name=\"NM\" value=\"NM_000015\" />"
[41] "\t<AnnBuilder:Item name=\"NP\" value=\"NP_000006\" />"
[42] "\t<AnnBuilder:Item name=\"PMID\" value=\"8460648\" />"

```

```

[43] "\t<AnnBuilder:Item name=\"PMID\" value=\"8102597\" />"
[44] "\t<AnnBuilder:Item name=\"PMID\" value=\"7915226\" />"
[45] "\t<AnnBuilder:Item name=\"PMID\" value=\"7773298\" />"
[46] "\t<AnnBuilder:Item name=\"PMID\" value=\"2734109\" />"
[47] "\t<AnnBuilder:Item name=\"PMID\" value=\"2340091\" />"
[48] "\t<AnnBuilder:Item name=\"PMID\" value=\"1968463\" />"
[49] "\t<AnnBuilder:Item name=\"PMID\" value=\"1676262\" />"
[50] "\t<AnnBuilder:Item name=\"PMID\" value=\"1381364\" />"
[51] "\t<AnnBuilder:Item name=\"PMID\" value=\"13\" />"
[52] "\t<AnnBuilder:Item name=\"CHR\" value=\"8\" />"
[53] "\t<AnnBuilder:Item name=\"OMIM\" value=\"NA\" />"
[54] "\t<AnnBuilder:Item name=\"GO\" value=\"GO:0004060\" type=\"E\"/>"
[55] "\t<AnnBuilder:Item name=\"CHRLoc\" value=\"+118524\" type=\"10\"/>"
[56] "</AnnBuilder:Entry>"
[57] "<AnnBuilder:Entry id=\"36512_at\" describ = \"PROBE\">"
[58] "\t<AnnBuilder:Item name=\"LOCUSID\" value=\"10\" />"
[59] "\t<AnnBuilder:Item name=\"ACCNUM\" value=\"L32179\" />"
[60] "\t<AnnBuilder:Item name=\"UNIGENE\" value=\"Hs.2\" />"
[61] "\t<AnnBuilder:Item name=\"GENENAME\" value=\"N-acetyltransferase 2 (arylamine N-"
[62] "\t<AnnBuilder:Item name=\"SYMBOL\" value=\"NAT2\" />"
[63] "\t<AnnBuilder:Item name=\"MAP\" value=\"NA\" />"
[64] "\t<AnnBuilder:Item name=\"GRIF\" value=\"NA\" />"
[65] "\t<AnnBuilder:Item name=\"SUMFUNC\" value=\"Arylamine N-acetyltransferase 2; N-"
[66] "\t<AnnBuilder:Item name=\"NM\" value=\"NM_000015\" />"
[67] "\t<AnnBuilder:Item name=\"NP\" value=\"NP_000006\" />"
[68] "\t<AnnBuilder:Item name=\"PMID\" value=\"8460648\" />"
[69] "\t<AnnBuilder:Item name=\"PMID\" value=\"8102597\" />"
[70] "\t<AnnBuilder:Item name=\"PMID\" value=\"7915226\" />"
[71] "\t<AnnBuilder:Item name=\"PMID\" value=\"7773298\" />"
[72] "\t<AnnBuilder:Item name=\"PMID\" value=\"2734109\" />"
[73] "\t<AnnBuilder:Item name=\"PMID\" value=\"2340091\" />"
[74] "\t<AnnBuilder:Item name=\"PMID\" value=\"1968463\" />"
[75] "\t<AnnBuilder:Item name=\"PMID\" value=\"1676262\" />"
[76] "\t<AnnBuilder:Item name=\"PMID\" value=\"1381364\" />"
[77] "\t<AnnBuilder:Item name=\"PMID\" value=\"13\" />"
[78] "\t<AnnBuilder:Item name=\"CHR\" value=\"8\" />"
[79] "\t<AnnBuilder:Item name=\"OMIM\" value=\"NA\" />"
[80] "\t<AnnBuilder:Item name=\"GO\" value=\"GO:0004060\" type=\"E\"/>"
[81] "\t<AnnBuilder:Item name=\"CHRLoc\" value=\"+118524\" type=\"10\"/>"
[82] "</AnnBuilder:Entry>"
[83] "<AnnBuilder:Entry id=\"38912_at\" describ = \"PROBE\">"
[84] "\t<AnnBuilder:Item name=\"LOCUSID\" value=\"10\" />"

```

```

[85] "\t<AnnBuilder:Item name=\"ACCNUM\" value=\"D90042\" />"
[86] "\t<AnnBuilder:Item name=\"UNIGENE\" value=\"Hs.2\" />"
[87] "\t<AnnBuilder:Item name=\"GENENAME\" value=\"N-acetyltransferase 2 (arylamine N-
[88] "\t<AnnBuilder:Item name=\"SYMBOL\" value=\"NAT2\" />"
[89] "\t<AnnBuilder:Item name=\"MAP\" value=\"NA\" />"
[90] "\t<AnnBuilder:Item name=\"GRIF\" value=\"NA\" />"
[91] "\t<AnnBuilder:Item name=\"SUMFUNC\" value=\"Arylamine N-acetyltransferase 2; N-
[92] "\t<AnnBuilder:Item name=\"NM\" value=\"NM_000015\" />"
[93] "\t<AnnBuilder:Item name=\"NP\" value=\"NP_000006\" />"
[94] "\t<AnnBuilder:Item name=\"PMID\" value=\"8460648\" />"
[95] "\t<AnnBuilder:Item name=\"PMID\" value=\"8102597\" />"
[96] "\t<AnnBuilder:Item name=\"PMID\" value=\"7915226\" />"
[97] "\t<AnnBuilder:Item name=\"PMID\" value=\"7773298\" />"
[98] "\t<AnnBuilder:Item name=\"PMID\" value=\"2734109\" />"
[99] "\t<AnnBuilder:Item name=\"PMID\" value=\"2340091\" />"
[100] "\t<AnnBuilder:Item name=\"PMID\" value=\"1968463\" />"
[101] "\t<AnnBuilder:Item name=\"PMID\" value=\"1676262\" />"
[102] "\t<AnnBuilder:Item name=\"PMID\" value=\"1381364\" />"
[103] "\t<AnnBuilder:Item name=\"PMID\" value=\"13\" />"
[104] "\t<AnnBuilder:Item name=\"CHR\" value=\"8\" />"
[105] "\t<AnnBuilder:Item name=\"OMIM\" value=\"NA\" />"
[106] "\t<AnnBuilder:Item name=\"GO\" value=\"GO:0004060\" type=\"E\"/>"
[107] "\t<AnnBuilder:Item name=\"CHRLOC\" value=\"+118524\" type=\"10\"/>"
[108] "</AnnBuilder:Entry>"
[109] "<AnnBuilder:Entry id=\"32468_f_at\" describ = \"PROBE\">"
[110] "\t<AnnBuilder:Item name=\"LOCUSID\" value=\"1084\" />"
[111] "\t<AnnBuilder:Item name=\"ACCNUM\" value=\"D90278;M16652\" />"
[112] "\t<AnnBuilder:Item name=\"UNIGENE\" value=\"NA\" />"
[113] "\t<AnnBuilder:Item name=\"GENENAME\" value=\"NA\" />"
[114] "\t<AnnBuilder:Item name=\"SYMBOL\" value=\"NA\" />"
[115] "\t<AnnBuilder:Item name=\"MAP\" value=\"NA\" />"
[116] "\t<AnnBuilder:Item name=\"GRIF\" value=\"NA\" />"
[117] "\t<AnnBuilder:Item name=\"SUMFUNC\" value=\"NA\" />"
[118] "\t<AnnBuilder:Item name=\"NM\" value=\"NA\" />"
[119] "\t<AnnBuilder:Item name=\"NP\" value=\"NA\" />"
[120] "\t<AnnBuilder:Item name=\"PMID\" value=\"NA\" />"
[121] "\t<AnnBuilder:Item name=\"CHR\" value=\"NA\" />"
[122] "\t<AnnBuilder:Item name=\"OMIM\" value=\"NA\" />"
[123] "NULL"
[124] "\t<AnnBuilder:Item name=\"CHRLOC\" value=\"+865077\" type=\"10\"/>"
[125] "</AnnBuilder:Entry>"
[126] "<AnnBuilder:Entry id=\"33825_at\" describ = \"PROBE\">"

```

```

[127] "\t<AnnBuilder:Item name=\"LOCUSID\" value=\"12\" />"
[128] "\t<AnnBuilder:Item name=\"ACCNUM\" value=\"X68733\" />"
[129] "\t<AnnBuilder:Item name=\"UNIGENE\" value=\"NA\" />"
[130] "\t<AnnBuilder:Item name=\"GENENAME\" value=\"NA\" />"
[131] "\t<AnnBuilder:Item name=\"SYMBOL\" value=\"NA\" />"
[132] "\t<AnnBuilder:Item name=\"MAP\" value=\"NA\" />"
[133] "\t<AnnBuilder:Item name=\"GRIF\" value=\"NA\" />"
[134] "\t<AnnBuilder:Item name=\"SUMFUNC\" value=\"Alpha-1-antichymotrypsin; member of"
[135] "\t<AnnBuilder:Item name=\"NM\" value=\"NM_001085\" />"
[136] "\t<AnnBuilder:Item name=\"NP\" value=\"NP_001076\" />"
[137] "\t<AnnBuilder:Item name=\"PMID\" value=\"9880565\" />"
[138] "\t<AnnBuilder:Item name=\"PMID\" value=\"8244391\" />"
[139] "\t<AnnBuilder:Item name=\"PMID\" value=\"6687683\" />"
[140] "\t<AnnBuilder:Item name=\"PMID\" value=\"6606438\" />"
[141] "\t<AnnBuilder:Item name=\"PMID\" value=\"6556193\" />"
[142] "\t<AnnBuilder:Item name=\"PMID\" value=\"6547997\" />"
[143] "\t<AnnBuilder:Item name=\"PMID\" value=\"3492865\" />"
[144] "\t<AnnBuilder:Item name=\"PMID\" value=\"3485824\" />"
[145] "\t<AnnBuilder:Item name=\"PMID\" value=\"3260956\" />"
[146] "\t<AnnBuilder:Item name=\"PMID\" value=\"32\" />"
[147] "\t<AnnBuilder:Item name=\"CHR\" value=\"NA\" />"
[148] "\t<AnnBuilder:Item name=\"OMIM\" value=\"NA\" />"
[149] "\t<AnnBuilder:Item name=\"GO\" value=\"GO:0005209\" type=\"NR\"/>"
[150] "\t<AnnBuilder:Item name=\"GO\" value=\"GO:0006953\" type=\"NR\"/>"
[151] "\t<AnnBuilder:Item name=\"GO\" value=\"GO:0004866\" type=\"NR\"/>"
[152] "\t<AnnBuilder:Item name=\"GO\" value=\"GO:0004867\" type=\"E\"/>"
[153] "\t<AnnBuilder:Item name=\"CHRLoc\" value=\"-845602\" type=\"10\"/>"
[154] "</AnnBuilder:Entry>"
[155] "<AnnBuilder:Entry id=\"35730_at\" describ = \"PROBE\">"
[156] "\t<AnnBuilder:Item name=\"LOCUSID\" value=\"125\" />"
[157] "\t<AnnBuilder:Item name=\"ACCNUM\" value=\"X03350\" />"
[158] "\t<AnnBuilder:Item name=\"UNIGENE\" value=\"NA\" />"
[159] "\t<AnnBuilder:Item name=\"GENENAME\" value=\"NA\" />"
[160] "\t<AnnBuilder:Item name=\"SYMBOL\" value=\"NA\" />"
[161] "\t<AnnBuilder:Item name=\"MAP\" value=\"NA\" />"
[162] "\t<AnnBuilder:Item name=\"GRIF\" value=\"NA\" />"
[163] "\t<AnnBuilder:Item name=\"SUMFUNC\" value=\"NA\" />"
[164] "\t<AnnBuilder:Item name=\"NM\" value=\"NA\" />"
[165] "\t<AnnBuilder:Item name=\"NP\" value=\"NA\" />"
[166] "\t<AnnBuilder:Item name=\"PMID\" value=\"NA\" />"
[167] "\t<AnnBuilder:Item name=\"CHR\" value=\"NA\" />"
[168] "\t<AnnBuilder:Item name=\"OMIM\" value=\"NA\" />"

```

```

[169] "NULL"
[170] "\t<AnnBuilder:Item name=\"CHRLOC\" value=\"-784967\" type=\"10\"/>"
[171] "</AnnBuilder:Entry>"
[172] "<AnnBuilder:Entry id=\"38936_at\" describ = \"PROBE\">"
[173] "\t<AnnBuilder:Item name=\"LOCUSID\" value=\"63036\" />"
[174] "\t<AnnBuilder:Item name=\"ACCNUM\" value=\"M16652\" />"
[175] "\t<AnnBuilder:Item name=\"UNIGENE\" value=\"NA\" />"
[176] "\t<AnnBuilder:Item name=\"GENENAME\" value=\"NA\" />"
[177] "\t<AnnBuilder:Item name=\"SYMBOL\" value=\"NA\" />"
[178] "\t<AnnBuilder:Item name=\"MAP\" value=\"NA\" />"
[179] "\t<AnnBuilder:Item name=\"GRIF\" value=\"NA\" />"
[180] "\t<AnnBuilder:Item name=\"SUMFUNC\" value=\"NA\" />"
[181] "\t<AnnBuilder:Item name=\"NM\" value=\"NA\" />"
[182] "\t<AnnBuilder:Item name=\"NP\" value=\"NA\" />"
[183] "\t<AnnBuilder:Item name=\"PMID\" value=\"NA\" />"
[184] "\t<AnnBuilder:Item name=\"CHR\" value=\"NA\" />"
[185] "\t<AnnBuilder:Item name=\"OMIM\" value=\"NA\" />"
[186] "NULL"
[187] "\t<AnnBuilder:Item name=\"CHRLOC\" value=\"+845627\" type=\"10\"/>"
[188] "</AnnBuilder:Entry>"
[189] "<AnnBuilder:Entry id=\"32481_at\" describ = \"PROBE\">"
[190] "\t<AnnBuilder:Item name=\"LOCUSID\" value=\"7051\" />"
[191] "\t<AnnBuilder:Item name=\"ACCNUM\" value=\"AL031663\" />"
[192] "\t<AnnBuilder:Item name=\"UNIGENE\" value=\"NA\" />"
[193] "\t<AnnBuilder:Item name=\"GENENAME\" value=\"NA\" />"
[194] "\t<AnnBuilder:Item name=\"SYMBOL\" value=\"NA\" />"
[195] "\t<AnnBuilder:Item name=\"MAP\" value=\"NA\" />"
[196] "\t<AnnBuilder:Item name=\"GRIF\" value=\"NA\" />"
[197] "\t<AnnBuilder:Item name=\"SUMFUNC\" value=\"NA\" />"
[198] "\t<AnnBuilder:Item name=\"NM\" value=\"NA\" />"
[199] "\t<AnnBuilder:Item name=\"NP\" value=\"NA\" />"
[200] "\t<AnnBuilder:Item name=\"PMID\" value=\"NA\" />"
[201] "\t<AnnBuilder:Item name=\"CHR\" value=\"NA\" />"
[202] "\t<AnnBuilder:Item name=\"OMIM\" value=\"NA\" />"
[203] "NULL"
[204] "NULL"
[205] "</AnnBuilder:Entry>"
[206] "<AnnBuilder:Entry id=\"39368_at\" describ = \"PROBE\">"
[207] "\t<AnnBuilder:Item name=\"LOCUSID\" value=\"NA\" />"
[208] "\t<AnnBuilder:Item name=\"ACCNUM\" value=\"AL031668\" />"
[209] "\t<AnnBuilder:Item name=\"UNIGENE\" value=\"NA\" />"
[210] "\t<AnnBuilder:Item name=\"GENENAME\" value=\"NA\" />"

```

```

[211] "\t<AnnBuilder:Item name=\"SYMBOL\" value=\"NA\" />"
[212] "\t<AnnBuilder:Item name=\"MAP\" value=\"NA\" />"
[213] "\t<AnnBuilder:Item name=\"GRIF\" value=\"NA\" />"
[214] "\t<AnnBuilder:Item name=\"SUMFUNC\" value=\"NA\" />"
[215] "\t<AnnBuilder:Item name=\"NM\" value=\"NA\" />"
[216] "\t<AnnBuilder:Item name=\"NP\" value=\"NA\" />"
[217] "\t<AnnBuilder:Item name=\"PMID\" value=\"NA\" />"
[218] "\t<AnnBuilder:Item name=\"CHR\" value=\"NA\" />"
[219] "\t<AnnBuilder:Item name=\"OMIM\" value=\"NA\" />"
[220] "NULL"
[221] "NULL</AnnBuilder:Entry>"
[222] "</AnnBuilder:Data>"
[223] "</AnnBuilder:Annotate>"
[224] ""

```

To generate an R data package containing the annotation data, we first save the data as R environment objects with key-value pairs. Each of the annotation element will be saved as a separate environment with probe ids as keys and the corresponding annotation as values (e. g. gene name) or vice versa. To do so, we save each of the environment objects in a common environment and then dump the contents of the common environment to the destination using an existing R function `textttpackage.skeleton`. We have to do this in a few steps as the mappings between probe ids and annotation elements may have different structure as previously mentioned.

```

> if (.Platform$OS.type != "windows") {
+   pkgName <- "test"
+   workEnv <- new.env(hash = TRUE, parent = NULL)
+   createEmptyDPkg("test", getwd(), force = TRUE)
+   for (i in getUniColNames()) {
+     env <- new.env(hash = TRUE, parent = NULL)
+     multiassign(as.vector(annotation[, "PROBE"]), sapply(annotation[,
+       i], splitEntry), env)
+     assign(paste(pkgName, i, sep = ""), env, workEnv)
+   }
+   for (i in intersect(colnames(annotation), getMultiColNames())) {
+     env <- new.env(hash = TRUE, parent = NULL)
+     multiassign(as.vector(annotation[, "PROBE"]), sapply(annotation[,
+       i], twoStepSplit), env)
+     assign(paste(pkgName, i, sep = ""), env, workEnv)
+   }
+   for (i in c("GO", "CHRLOC")) {
+     env <- new.env(hash = TRUE, parent = NULL)
+     multiassign(as.vector(annotation[, "PROBE"]), sapply(annotation[,

```

```
+           i], splitEntry), env)
+       assign(paste(pkgName, i, sep = ""), env, workEnv)
+   }
+ }
```

When we have all the annotation data we want saved as environment objects in a common environment, we can call `textttpackage.skeleton` to create a data package with data files stored in correct subdirectories in the package.

```
> if (.Platform$OS.type != "windows") {
+   for (i in ls(workEnv)) {
+       save(list = i, file = file.path(getwd(), "test", "data",
+           paste(i, ".rda", sep = "")), envir = workEnv)
+   }
+ }
```

The data package is stored in the current working directory under the name `texttttest`.

```
> if (.Platform$OS.type != "windows") {
+   list.files(file.path(getwd(), "test"))
+ }
```

```
[1] "R"      "data" "man"
```

As can be seen, the data package contains all the required elements of a normal R package and can be installed in the same way as an R package. The annotation data are all stored as rda files in the data directory.

```
> if (.Platform$OS.type != "windows") {
+   list.files(file.path(getwd(), "test", "data"))
+ }
```

```
[1] "testACCNUM.rda"  "testCHR.rda"      "testCHRLOC.rda"   "testGENENAME.rda"
[5] "testGO.rda"      "testGRIF.rda"     "testLOCUSID.rda"  "testMAP.rda"
[9] "testNM.rda"      "testNP.rda"       "testOMIM.rda"     "testPMID.rda"
[13] "testSUMFUNC.rda" "testSYMBOL.rda"   "testUNIGENE.rda"
```

Each of the rda files contains key and value pairs with the key being Affymetrix probe ids and value being the annotation element in this case.

The last step is to write the needed documentations and statistic data for quality control purpose. The following code can be used to generate the required documentations for the data package. Some part of the code may fail as the urls used may subject to changes by maintainers of the web sits in the future.



```

> if (.Platform$OS.type != "windows") {
+   writeAccessory(pkgName = "test", pkgPath = getwd(), organism = "human",
+     version = "1.1.1", author = list(author = "your name",
+       maintainer = "youremail@net.com"))
+   getDPStats(baseF = "geneNMap", pkgName = "test", pkgPath = getwd(),
+     saveList = TRUE)
+   writeMan4QC("test", pkgPath = getwd())
+ }

```

Now, we can clean up the mess we have left.

```

> if (.Platform$OS.type != "windows") {
+   unlink(c(unified, XMLOut, "geneNMap", "test.xml", "testByNum.xml"))
+   unlink(file.path(getwd(), "test"), TRUE)
+ }

```