Building Annotation Packages with pdInfoBuilder for Use with the oligo Package

Benilton Carvalho

March 27, 2009

1 Introduction

The oligo package offers support to multiple types of microarrays produced by Affymetrix and NimbleGen. The package will successfully read in CEL (Affymetrix) and XYS (NimbleGen) files, as long as the associated annotation package is already installed on the user's system.

The user must note that the annotation packages built for the affy package are **not** compatible with oligo. To have an annotation package that is compatible with oligo, one must use the pdInfoBuilder package.

This document shows examples on how to create such annotation packages for different platforms. After the package is created, the user must install it and not just copy it to the library tree.

2 The General Strategy for Building Annotation Packages with pdInfoBuilder

Building annotation packages with pdInfoBuilder depends on the followings files:

- Array design file: CDF (Affymetrix Expression), NDF (NimbleGen), BPMAP (Affymetrix Tiling) or PGF+CLF (Affymetrix Exon ST or Gene ST);
- Positions file: POS (NimbleGen - Tiling);
- Template of intensity file: CEL (Affymetrix) or XYS (NimbleGen);
- Probe sequence file: TAB (Affymetrix Expression)
- Probeset annotation file: PROBESET.CSV (Affymetrix Exon/Gene)

3 Loading the package

```
R> library(pdInfoBuilder)
```

Type	Package
50K Xba	pd.mapping50k.xba240
50K Hind	pd.mapping50k.hind240
250K Sty	pd.mapping250k.sty
250K Nsp	pd.mapping250k.nsp
SNP 5.0	pd.genomewidesnp.5
SNP 6.0	pd.genomewidesnp.6

Table 1: List of packages for SNP chips

4 Affymetrix SNP Array

The annotation packages for any Affymetrix SNP chip is available from BioConductor.

5 Affymetrix HT-HGU133

For this particular array, the user must have access to three components: CDF, CEL (which will provide information on the array geometry) and probe sequence file (TAB-delimited).

```
R> baseDir <- "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyHTHGU133P"
R> (cdf <- list.files(baseDir, pattern = ".CDF",
     full.names = TRUE))

[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyHTHGU133P/HT_HG-U133_Plus_PM.

R> (cel <- list.files(baseDir, pattern = ".CEL",
     full.names = TRUE)[1])

[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyHTHGU133P/Human_PM_TestData.A

R> (tab <- list.files(baseDir, pattern = "_tab",
     full.names = TRUE))

[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyHTHGU133P/HT_HG-U133_Plus_PM.

R> seed <- new("AffyExpressionPDInfoPkgSeed",
     cdfFile = cdf, celFile = cel,
     tabSeqFile = tab, author = "Benilton Carvalho",
     email = "bcarvalh@jhsph.edu",
     biocViews = "AnnotationData",
     genomebuild = "NCBI Build 36",
     organism = "Human", species = "Homo Sapiens",
     url = "http://www.biostat.jhsph.edu/~bcarvalh")
R> makePdInfoPackage(seed, destDir = ".")
```

```
=============================================
Building annotation package for Affymetrix Expression array
CDF..............:  HT_HG-U133_Plus_PM.CDF
CEL..............:  Human_PM_TestData.A01.CEL
Sequence TAB-Delim:  HT_HG-U133_Plus_PM.probe_tab
=============================================
Parsing file: HT_HG-U133_Plus_PM.CDF ... OK
Parsing file: Human_PM_TestData.A01.CEL ... OK
Parsing file: HT_HG-U133_Plus_PM.probe_tab ... OK
Getting information for featureSet table... OK
Getting information for pm/mm feature tables ... OK
Combining probe information with sequence information ... OK
Getting sequence information for AFFX probes ...OK
Creating package in ./pd.ht.hg.u133.plus.pm
Inserting 54715 rows into table "featureSet"... OK
Inserting 519517 rows into table "pmfeature"... OK
Inserting 180 rows into table "mmfeature"... OK
Inserting 16943 rows into table "bgfeature"... OK
Counting rows in bgfeature
Counting rows in featureSet
Counting rows in mmfeature
Counting rows in pmfeature
Creating index idx_bgfsetid on bgfeature ... OK
Creating index idx_bgfid on bgfeature ... OK
Creating index idx_pmfsetid on pmfeature ... OK
Creating index idx_pmfid on pmfeature ... OK
Creating index idx_fsfsetid on featureSet ... OK
Saving XDataFrame object for PM.
Saving XDataFrame object for BG.
Done.
```

6 Affymetrix Tiling Array

```
R> baseDir <- "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyTiling"
R> (bpmap <- list.files(baseDir, pattern = ".bpmap",
     full.names = TRUE))
```

```
[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyTiling/Hs35b_P02R_v01-3_NCBIv
```

```
R> (cel <- list.files(baseDir, pattern = ".CEL",
     full.names = TRUE)[1])
```

```
[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyTiling/GSM178873.CEL"
```

```
R> seed <- new("AffyTilingPDInfoPkgSeed",
     bpmapFile = bpmap, celFile = cel,
     author = "Benilton Carvalho",
     email = "bcarvalh@jhsph.edu",
     biocViews = "AnnotationData",
     genomebuild = "NCBI Build 34",
     organism = "Human", species = "Homo Sapiens",
     url = "http://www.biostat.jhsph.edu/~bcarvalh")
R> makePdInfoPackage(seed, destDir = ".")
```

```
============================================
Building annotation package for Affymetrix Tiling array
BPMAP:  Hs35b_P02R_v01-3_NCBIv34.bpmap
CEL..:  GSM178873.CEL
============================================
Parsing file: Hs35b_P02R_v01-3_NCBIv34.bpmap ... OK
Getting geometry from CEL file... OK
Getting PMs...OK
Getting MMs...OK
Getting background probes...OK
Getting sequences...OK
Creating package in ./pd.hs35b.p02r.v01
Inserting 7 rows into table "chrom_dict"... OK
Inserting 6020293 rows into table "pmfeature"... OK
Inserting 1774 rows into table "mmfeature"... OK
Inserting 37687 rows into table "bgfeature"... OK
Counting rows in bgfeature
Counting rows in chrom_dict
Counting rows in mmfeature
Counting rows in pmfeature
Creating index idx_bgfid on bgfeature ... OK
Creating index idx_pmfid on pmfeature ... OK
Saving XDataFrame object for PM.
Saving XDataFrame object for BG.
Done.
```

7 Affymetrix Exon ST Array

```
R> library(pdInfoBuilder)
R> baseDir <- "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyExon"
R> (pgf <- list.files(baseDir, pattern = ".pgf",
     full.names = TRUE))
```

```
[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyExon/HuEx-1_0-st-v2.r2.pgf"
```

```
R> (clf <- list.files(baseDir, pattern = ".clf",
     full.names = TRUE))
```

```
[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyExon/HuEx-1_0-st-v2.r2.clf"
```

```
R> (prob <- list.files(baseDir, pattern = ".probeset.csv",
     full.names = TRUE))

[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyExon/HuEx-1_0-st-v2.na27.hg18

R> seed <- new("AffyExonPDInfoPkgSeed",
     pgfFile = pgf, clfFile = clf,
     probeFile = prob, author = "Benilton Carvalho",
     email = "bcarvalh@jhsph.edu",
     biocViews = "AnnotationData",
     genomebuild = "NCBI Build 36",
     organism = "Human", species = "Homo Sapiens",
     url = "http://www.biostat.jhsph.edu/~bcarvalh")
R> makePdInfoPackage(seed, destDir = ".")

=============================================
Building annotation package for Affymetrix Exon ST Array
PGF.....:  HuEx-1_0-st-v2.r2.pgf
CLF.....:  HuEx-1_0-st-v2.r2.clf
Probeset:  HuEx-1_0-st-v2.na27.hg18.probeset.csv
=============================================
Parsing file: HuEx-1_0-st-v2.r2.pgf ... OK
Parsing file: HuEx-1_0-st-v2.r2.clf ... OK
Creating initial table for probes...OK
Creating dictionaries... OK
Parsing file: HuEx-1_0-st-v2.na27.hg18.probeset.csv ... OK
Creating probeset -> gene table... OK
Creating genes table... OK
Creating package in ./pd.huex.1.0.st.v2
Inserting 100 rows into table "chrom_dict"... OK
Inserting 5 rows into table "level_dict"... OK
Inserting 8 rows into table "type_dict"... OK
Inserting 1625370 rows into table "fset2gene"... OK
Inserting 114281 rows into table "gene"... OK
Inserting 1425647 rows into table "featureSet"... OK
Inserting 5344479 rows into table "pmfeature"... OK
Inserting 37687 rows into table "bgfeature"... OK
Counting rows in bgfeature
Counting rows in chrom_dict
Counting rows in featureSet
Counting rows in fset2gene
Counting rows in gene
Counting rows in level_dict
Counting rows in pmfeature
Counting rows in type_dict
Creating index idx_bgfsetid on bgfeature ... OK
Creating index idx_bgfid on bgfeature ... OK
Creating index idx_pmfsetid on pmfeature ... OK
Creating index idx_pmfid on pmfeature ... OK
Creating index idx_fsfsetid on featureSet ... OK
Creating index idx_fs2gfsetid on fset2gene ... OK
Creating index idx_fs2ggid on fset2gene ... OK
Creating index idx_genegid on gene ... OK
Saving XDataFrame object for PM.
Saving XDataFrame object for BG.
Done.
```

8 Affymetrix Gene ST Array

```
R> library(pdInfoBuilder)
R> baseDir <- "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyGene"
R> (pgf <- list.files(baseDir, pattern = ".pgf",
     full.names = TRUE))
```

```
[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyGene/HuGene-1_0-st-v1.r4.pgf"
```

```
R> (clf <- list.files(baseDir, pattern = ".clf",
     full.names = TRUE))
```

```
[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyGene/HuGene-1_0-st-v1.r4.clf"
```

```
R> (prob <- list.files(baseDir, pattern = ".probeset.csv",
     full.names = TRUE))
```

```
[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/AffyGene/HuGene-1_0-st-v1.na27.2.
```

```
R> seed <- new("AffyGenePDInfoPkgSeed",
     pgfFile = pgf, clfFile = clf,
     probeFile = prob, author = "Benilton Carvalho",
     email = "bcarvalh@jhsph.edu",
     biocViews = "AnnotationData",
     genomebuild = "NCBI Build 36",
     organism = "Human", species = "Homo Sapiens",
     url = "http://www.biostat.jhsph.edu/˜bcarvalh")
R> makePdInfoPackage(seed, destDir = ".")
```

```
============================================
Building annotation package for Affymetrix Gene ST Array
PGF.....:  HuGene-1_0-st-v1.r4.pgf
CLF.....:  HuGene-1_0-st-v1.r4.clf
Probeset:  HuGene-1_0-st-v1.na27.2.hg18.probeset.csv
============================================
Parsing file: HuGene-1_0-st-v1.r4.pgf ... OK
Parsing file: HuGene-1_0-st-v1.r4.clf ... OK
```

```
Creating initial table for probes...OK
Creating dictionaries... OK
Parsing file: HuGene-1_0-st-v1.na27.2.hg18.probeset.csv ... OK
Creating probeset -> gene table... OK
Creating genes table... OK
Creating package in ./pd.hugene.1.0.st.v1
Inserting 125 rows into table "chrom_dict"... OK
Inserting 5 rows into table "level_dict"... OK
Inserting 8 rows into table "type_dict"... OK
Inserting 1011778 rows into table "fset2gene"... OK
Inserting 87374 rows into table "gene"... OK
Inserting 257430 rows into table "featureSet"... OK
Inserting 764885 rows into table "pmfeature"... OK
Inserting 818005 rows into table "f2fset"... OK
Inserting 16943 rows into table "bgfeature"... OK
Counting rows in bgfeature
Counting rows in chrom_dict
Counting rows in f2fset
Counting rows in featureSet
Counting rows in fset2gene
Counting rows in gene
Counting rows in level_dict
Counting rows in pmfeature
Counting rows in type_dict
Creating index idx_bgfsetid on bgfeature ... OK
Creating index idx_bgfid on bgfeature ... OK
Creating index idx_pmfid on pmfeature ... OK
Creating index idx_f2fsfid on f2fset ... OK
Creating index idx_f2fsfsetid on f2fset ... OK
Creating index idx_fsfsetid on featureSet ... OK
Creating index idx_fs2gfsetid on fset2gene ... OK
Creating index idx_fs2ggid on fset2gene ... OK
Creating index idx_genegid on gene ... OK
Saving XDataFrame object for PM.
Saving XDataFrame object for BG.
Done.
```

9 NimbleGen Expression Array

```
R> library(pdInfoBuilder)
R> baseDir <- "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/NgsExpression"
R> (ndf <- list.files(baseDir, pattern = ".ndf",
     full.names = TRUE))
```

```
[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/NgsExpression/HG18_60mer_expr.ndf
```

```
R> (xys <- list.files(baseDir, pattern = ".xys",
     full.names = TRUE)[1])
```

```
[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/NgsExpression/9868701_532.xys"
```

```
R> seed <- new("NgsExpressionPDInfoPkgSeed",
     ndfFile = ndf, xysFile = xys,
     author = "Benilton Carvalho",
     email = "bcarvalh@jhsph.edu",
     biocViews = "AnnotationData",
     genomebuild = "NCBI Build 36",
     organism = "Human", species = "Homo Sapiens",
     url = "http://www.biostat.jhsph.edu/~bcarvalh")
R> makePdInfoPackage(seed, destDir = ".")
```

```
=============================================
Building annotation package for Nimblegen Expression Array
NDF:  HG18_60mer_expr.ndf
XYS:  9868701_532.xys
=============================================
Parsing file: HG18_60mer_expr.ndf ... OK
Parsing file: 9868701_532.xys ... OK
Merging NDF and XYS files ...OK
Preparing contents for featureSet table ...OK
Preparing contents for bgfeature table ...OK
Preparing contents for pmfeature table ...OK
Creating package in ./pd.hg18.60mer.expr
Inserting 24000 rows into table "featureSet"... OK
Inserting 71998 rows into table "pmfeature"... OK
Counting rows in featureSet
Counting rows in pmfeature
Creating index idx_pmfsetid on pmfeature ... OK
Creating index idx_pmfid on pmfeature ... OK
Creating index idx_fsfsetid on featureSet ... OK
Saving XDataFrame object for PM.
Done.
```

10 NimbleGen Tiling Array

```
R> library(pdInfoBuilder)
R> baseDir <- "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/NgsTiling"
R> (ndf <- list.files(baseDir, pattern = ".ndf",
     full.names = TRUE))
```

```
[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/NgsTiling/2006-07-18_HG18_RefSeq_
```

```
R> (pos <- list.files(baseDir, pattern = ".pos",
     full.names = TRUE))

[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/NgsTiling/2006-07-18_HG18_RefSeq_

R> (xys <- list.files(baseDir, pattern = ".xys",
     full.names = TRUE)[1])

[1] "/thumper/ctsa/snpmicroarray/organizing/pdInfoVignette/NgsTiling/92204_532.xys"

R> seed <- new("NgsTilingPDInfoPkgSeed",
     ndfFile = ndf, xysFile = xys,
     posFile = pos, author = "Benilton Carvalho",
     email = "bcarvalh@jhsph.edu",
     biocViews = "AnnotationData",
     genomebuild = "HG 18", organism = "Human",
     species = "Homo Sapiens", url = "http://www.biostat.jhsph.edu/~bcarvalh")
R> makePdInfoPackage(seed, destDir = ".")

============================================
Building annotation package for Nimblegen Tiling Array
NDF:  2006-07-18_HG18_RefSeq_promoter.ndf
POS:  2006-07-18_HG18_RefSeq_promoter.pos
XYS:  92204_532.xys
============================================
Parsing file: 2006-07-18_HG18_RefSeq_promoter.ndf ... OK
Parsing file: 2006-07-18_HG18_RefSeq_promoter.pos ... OK
Parsing file: 92204_532.xys ... OK
Creating package in ./pd.2006.07.18.hg18.refseq.promoter
Inserting 18029 rows into table "featureSet"... OK
Inserting 386230 rows into table "pmfeature"... OK
Inserting 3077 rows into table "bgfeature"... OK
Counting rows in bgfeature
Counting rows in featureSet
Counting rows in pmfeature
Creating index idx_bgfsetid on bgfeature ... OK
Creating index idx_bgfid on bgfeature ... OK
Creating index idx_pmfsetid on pmfeature ... OK
Creating index idx_pmfid on pmfeature ... OK
Creating index idx_fsfsetid on featureSet ... OK
Saving XDataFrame object for PM.
Saving XDataFrame object for BG.
Done.
```