Integration
Self-Study Exercises

Bioconductor Team

Fred Hutchinson Cancer Research Center

17-18 February, 2011

1 Introduction

These excrcises pull together the various aspects of data manipulation and class construction and methods development covered in this the Advanced *R* course.

The goal is to compute linkage disequilibrium on the set of snp data we have. We will create a new method for the .cld wrapper for linkage disequlibrium. This method will be used to iterate through the snp data

Exercise 1
The pupose of this excercise is to create a method for the .cld wrapper function. The method should accept a file name and process data as the `getCols` function does. Additionally, it should accept any arguments that the wrapper function `.cld.R` does.

Question 1

- *Create a generic for .cld. Call the generic `cld`.*
- *Create a method for `cld`.*

Solution:

```
setGeneric("cld", signature="x",
    function(x, first, last, width = 5) standardGeneric("cld")
)

setMethod("cld", "GWASdata",
    function(x, first, last, width = 5)
    {
        data <- getCols(x, first, last)
        colnames(data) <- getSnps(metadatapath(x))$snp_id[first:last]
        .cld(data, width = width)
    }
```

)

Exercise 2
Write a function that performs composite linkage disequilibrium on the snpData.nc file. Iterate through the data in chunks of 10000 snps with width of 100. Filter out snps that are in high linkage disequlibrium by removing pairs that have results > 0.8. The remaining snps will be annotated in the next excercise.

One of the challenges of this excercise is to construct your function such that linkage disequilibrium is computed for all snps in the dataset (except the last 'width' number of snps). Note that of the snps input to the `.cld` function, linkage disequilibrium results are only computed for (# snps - width) snps. This is because the number of linkage disequlibrium comparisions made are equal to the `width` argument. Thus if 100 snps are input with a width of 5, the ouptut will be a matrix of the first 95 snps (one row per snp) by 5 columns. The 5 columns represent the comparisions with the five snps to the right of the snp.

At your disposal you have the following

- *cld method you wrote in the previous excercise*
- *fapply function from the `EfficientR` portion of the course, found in AdvancedR2001/script/efficient.R*

Solution:

```
> library(StudentGWAS)
> library(AdvancedR2011Data)
> functionPath <- system.file("script", "efficient.R",
+                             package="AdvancedR2011")
> source(functionPath)
> dataPath <- system.file("extdata", "snpData.nc",
+                             package="AdvancedR2011Data")
> metadataPath <- system.file("extdata", "metadata.sqlite",
+                                 package="AdvancedR2011Data")
> data <- GWASdata(dataPath, metadataPath)
> snps <- matrix(sample((0:2), replace=TRUE), nrow=10, ncol=6)
> StudentGWAS:::.cld(snps, width=3)
> res <- cld(snps, first=1, last=25, width=3)
> LD <- function(data, idx, width = 5) {}
>
```

Exercise 3
Marc's annotation excercise here. The output of my function above will be a vector of snp rs names.