README

Jessica Bonnie

10/06/2015

ScatterPlots

The most common and complicated scatter plots drawn during the QC process are those that illustrate PCA. I've written different scripts depending on the goal of the graph. In each case, a certain preparation of the data in necessary in order to color the graph appropriately.

Assuming that the principle components were calculated using KING, the table will be 26 columns with no titles. The first six columns will be: *FID*, *IID*, *Father*, *Mother*, *Sex*, *Status*. The next 20 columns will be PC1-PC20.

Throughout this **README**, the calls to the code in this directory will be proceeded with `${scatterfolder}` which refers to the path to this folder, currently set to the following location with this line:

```
scatterfolder=/h2/t74/cphgdesk/share/cphg_OnengutLab/Jessica/Farewell
```

HapMap Projections

There are a few different versions of HapMap projection graph scripts. One of the key differences relates to which population of samples is meant to be highlighted. In either case, the HapMap samples must be given a status of "1" with the study samples given a status of "2" in column 6 during prior to projection. This is part of Wei-Min's usual process, and it makes things much easier.

Single Cohort

If the projected PCs of a single study population are meant to be graphed on top of the HapMap samples which have been colored by ethnicity, a 27th column must be added containing the names of the HapMap populations and the name of the single study. The HapMap population file which I stole from Rany is included in this folder for easy reference. In the code below `${projfile}` is the projection output from KING, `${study}` is the name of the single study to be graphed, and `${study}_hapmap3pc.txt` is the table that will be passed to the R script.

```
popfile=/h2/t74/cphgdesk/share/cphg_OnengutLab/Jessica/Farewell/relationships_w_pops_041510.txt

LANG=en_EN join -1 2 -2 1  <(awk '$6 == 1' $projfile | LANG=en_EN sort -k2,2) <(awk '{print $2,$7}' ${p

awk -v study=${study} '$6 == 2 {print $0, study}' $projfile >> ${study}_hapmap3pc.txt
```

Single Graph: PC1 vs PC2 This script takes either **4** or **8** arguments. The first four, required, arguments are **(1)** the path of the prepared file, **(2)** the name of the study, **(3)** the name of the chip (for the title), and **(4)** the name/path to the psfile (the **.ps** will be added).The four additional, optional, additional arguments, `${xmin}`, `${xmax}`, `${ymin}`, `${ymax}` refer to the boundary lines that should be drawn onto the graph to indicate which samples are outliers.

```
R CMD BATCH "--args ${study}_hapmap3pc.txt ${study} ${chip} ${graphout} ${xmin} ${xmax} ${ymin} ${ymax}
```

Scatter Plot Matrix This script takes the same first four arguments from the last one. It draws three pages of scatter plot grids. It can be used with the same file as before or with a different projection file (such as CEU+TSI) as long as the file is prepared the same way.

```
R CMD BATCH "--args ${study}_hapmap3pc.txt ${study} ${chip} ${graphout}" ${scatterfolder}/PCA_SCATTER_J
```

Multiple Cohorts

If multiple cohorts are meant to be graphed against HapMap samples with different colors assigned to each cohort, the file must be prepared in the same manner as in the previous case, with the 27th column listing the Cohort or whatever else must be colored by, with the HapMap Samples listed as "HapMap" in that column in the same table with a "**1**" in the status.

Unprojected PCA

If you want to draw PCA graphs colored by anything, just put the names of the groups (population, cohort, removed/unremoved, etc.) in the 27th column. No need to worry about the status value at all.

```
R CMD BATCH "--args ${unprojfile} ${study} ${chip} ${graphout}" ${scatterfolder}/PCA_SCATTER_PLAIN.R
```