**Pipeline for Processing PBM Data**

All of the perl scripts are located in the following directory

DIR: ~/perl = /projectnb/siggers/perl\_master

\*\*is ‘siggers’ the only user able to execute the scripts below? Or is there a group that users belong to that can run the processing? Asking in case of downstream permissions\*\*

**Make an analysis file -** An analysis file relates the spot coordinates with their DNA sequences. **Each time you design a new PBM array an analysis file needs to be constructed**. This file can be reused for any future analyses with the same array.

**Steps to make analysis file:**

1. Download design files from Agilent SureDesign site

<https://earray.chem.agilent.com/suredesign/>

* 1. To sign in login in with the following credentials:
     1. Username: [tsiggers@bu.edu](mailto:tsiggers@bu.edu)

Password: SDPotte!5

* + 1. To find your array design go to find designs  CGH  My Designs  look for your PBM ID  select and download all files as .zip files (they will be toggled in a dialog box)
    2. On the SCC, copy the .zip files to a working directory \*\* is the working directory here arbitrary, or does the user follow some guideline of what’s a ‘legal’ working directory? \*\*
  1. To unzip the design folder, use the following line of code

>> unzip 0XXXX.zip \*\* In terms of the file system, do all of these zip files get unzipped in the same directory, /project/siggers/perl/PBM/ ?\*\*

1. Use the PERL scripts to generate **make\_analysis\_file.com**
   1. You will need 3 files:
      1. File 1 – FOO\_046030\_DNAfront\_BCbottom\_foo.txt \*\* is the “046030” generally supposed to be the pbm ID of a given design? Also, looking at the example below, what is the significance of “\_D\_” in the filename? Does it pertain to a certain property of the design? I only ask since I anticipate doing string manipulations down the line and want to know what it specifically means.\*\*

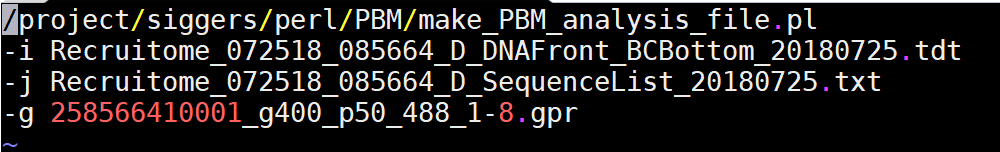
txt?

* + 1. File 2 – FOO\_D\_SequenceList\_XXX.txt

Significance/where does this number come from? Will the number of scans always be in the filename(i.e. if I split this gpr filestring by ‘\_’, will the #scans be the 4th value always)?

* + 1. File 3 – one of the \*.gpr files from your image analysis

**Make\_analysis\_file.com** looks like this:

**>>perl ~/perl/PBM/make\_analysis\_file.com.pl**

**-i FOO\_046030\_DNAfront\_BCbottom\_foo.txt**

**-j FOO\_D\_SequenceList\_XXX.txt**

**-g FOO\_1-4.gpr**

**<<EOF**

1. To generate your analysis file, run the following line of code:

**>> perl /project/siggers/perl/MISC/run\_comfile\_alert.pl -com make\_analysis\_file.com > ID\_49519\_4x180k\_NfkB\_Genomic\_analysis.txt**

**\*\*How does the naming system work for this output file? Which directory is the output being written to?\*\***

Note – you do not have to submit this to the cluster to run it.

What the output should look like:



NOTE: all the files for generating the analysis file all need to be in the same directory!

**Masliner for PBM scans at multiple intensities** - Masliner is run when you need to ‘stitch’ scans together that were performed at multiple intensities. This happens when the dynamic range of the PBM spots is so large that at some laser power/gain settings you can just see the low intensity spots, but the high intensity spots are already saturated. In this situation you perform multiple scans at different intensities (keep the laser power at 50% and change the gain) and combine them computationally using masliner. It is based on the observation that fluorescence values change linearly with increasing scan intensity.

**Steps for Masliner:**

1. Go to your working directory where you transferred all your .gpr files from all your scans

**Cd /projectnb/siggers/pbmdata/gpr/Recruitome/v2\_array/258566410001** (the last parts of the path will be whatever your working folder is \*\*from gpr onward?\*\*)

1. Make separate directories for your 488 and 635 scans since they have to be processed separately \*\*Is there a naming convention for such directories(e.g. “…./488/ ..../635/)?\*\*
2. **Make experiment\_description.txt file**. This will have all your 488 gpr files separated by chamber. Each section will be divided by the following header:

\*\*which parts of these headers are user-dependent(e.g. does the user say pbm=1,concentration=100, etc.)?\*\*

Pbm=1

Concentration=100

Cy3=FOO

To move each file into the experiment\_description.txt file, use the following line of code:

**Ls -1 25……\_g\*1-8\* > experiment\_description.txt**

**For all the subsequent gpr files that you are adding to the experiment\_description.txt file, use >> to concatenate**

1. If copy or make a com file to run the masliner program using the que. Call this file masliner.com

**perl /project/siggers/perl/GENEPIX/masliner\_list.pl**

**-i experiment\_description.txt**

1. Submit the job to the cluster:

**Perl /project/siggers/perl/MISC/run\_comfile\_alert.pl –com masliner.com –qsub masliner.com**

*\*\*if you are making the* ***make\_analysis.txt file for 647 files****, you have to run the following line of code first. This line of code converts all the 647 in the column names to 488.*

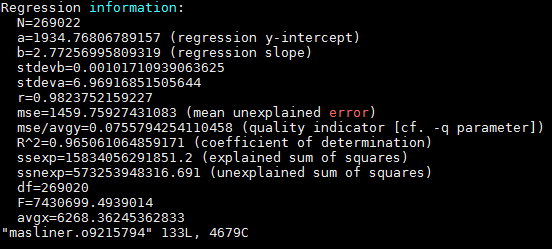
***Perl /project/siggers/perl/GENEPIX/change\_fluor.pl –i gpr.list***

***\*\* 647 scans? When does this usually occur, and why convert to “488”? \*\****

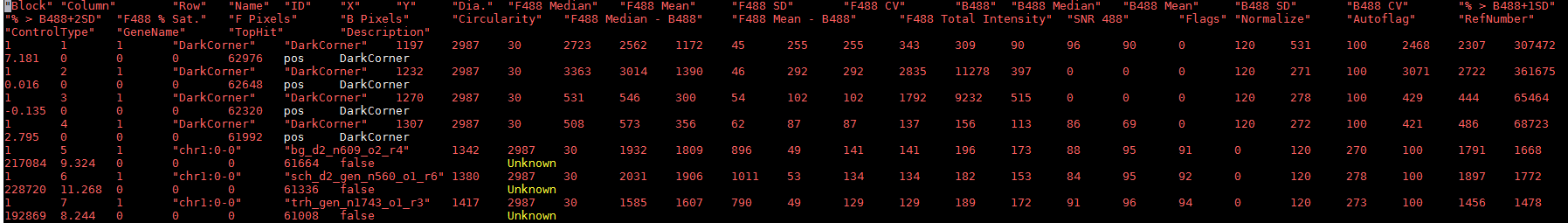
*After running this, repeat the steps above \*\*From masliner 1 onward?\*\*.*

1. The masliner will generate files that start with madj\_. It will also output an error and output file. Check the output file for the R2 value to make sure it is within an acceptable range before moving forward. In the event that the R2 value is low, try to omit some of your scans taken at lower gains. \*\*Where do the output and error files respectively get written to? \*\*

Example of a masliner.com output file:



Example of a madj\_\* file:



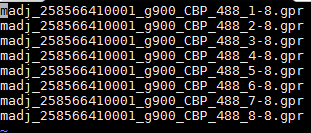
**Spatial Detrending** - This spatial de-trending process normalizes different regions of the scan to a common median. In other words, if a portion of the scan was generally dimmer (this happens) we can adjust the intensities for the spots in this area up a little. The approach is based on the idea that on average, sufficiently large areas of the scan should have the same median values. For this use the masliner adjusted (madj\_\*) files. Just use the highest scan intensity

1. Make a **madj\_gpr.list file**.

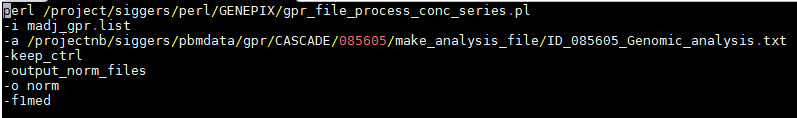
To move all your madj files to this list file, use the following line of code:

**Ls -1 madj\_\* > madj\_gpr.list**

**\*\*which directory? The ..../v2\_array/giant\_number/ ?\*\***



1. Copy or make a com file for the run. Call this file **process\_custom\_probes.com \*\*screencap below is the structure of said file?\*\***



Description of some of the flags/options\*\*what are the other flags? Are those optional?\*\*

-cy3\_gpr.list : *List of cy3 gpr files if you want to flag bad spots based on cy3 scans*

**-keep\_ctrl :** *This will keep the Ctrl/Cusotm probe sequences*

**-output\_norm\_files :** *Needed for ctrl/genomic pbms, as opposed to universal pbms*

**-**o norm : *A prefix of your choice for the normalized version of the files*

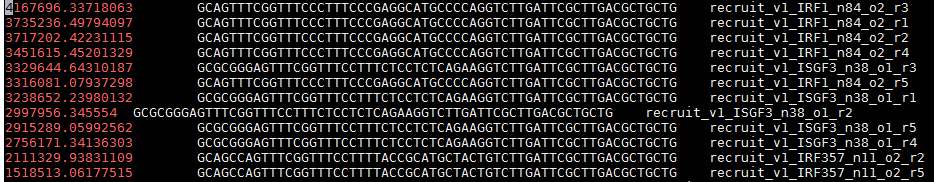
**-f1med :** *Uses the spot median values and not background subtracted values*

1. Run the com file:

**Perl /project/siggers/perl/MISC/run\_comfile\_alert.pl –com process\_custom\_probes.com –qsub customprobes**

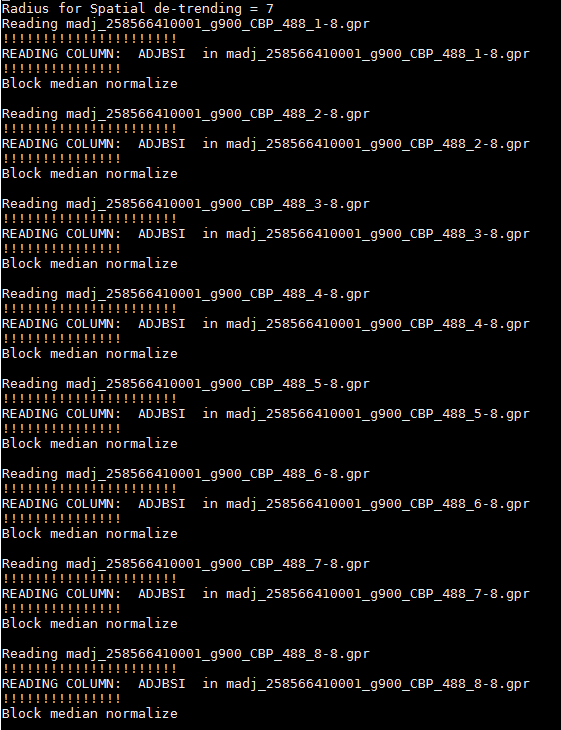
1. Process\_custom\_probes.com will output files are called norm\_madj\*

Example file:



It will also output an error and output file. It is normal for the error file to have contents.

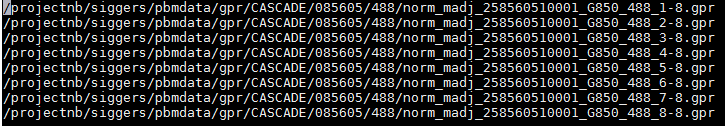
Example error file:



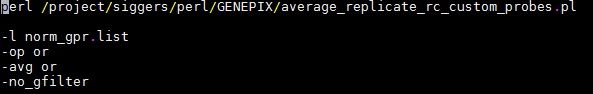
**Averaging probe intensities over replicates and orientations.**

1. Make a **norm\_gpr.list** file that lists **with full directory paths**, the normalized, masliner-adjusted (norm\_madj\*) files.

Example:



1. Make a com file called **average\_probes.com**



* Op or – means that the output files will contain ‘or’ in the front of the name (or\_norm\_madj\*)\*\*significance of ‘or’?\*\*
* Avg or – means that it will be averaged over replicates
* No\_gfilter – turns of G-filter – requires that o1 and o2 match

You can also run the following averaging \*\*when would one want to use one averaging method over another?\*\*:

-op r – will output files that start with o1 or o2 (o1match\_o\_norm\_madj\* or o2match\_o\_norm\_madj\*)

- avg r – will average over different probes (ie. Different orientations). This is a good way to check for consistency.

-op br – will output files that start with o1o2top\_br\_norm\_madj\*

- avg br – will average over the background

1. Run the com file:

**Perl /project/siggers/perl/MISC/run\_comfile\_alert.pl –com average\_probes.com –qsub avg**

Note: each type of averaging needs to be done separately.

**Make Data matrix of all GPR files – for analysis and plotting**

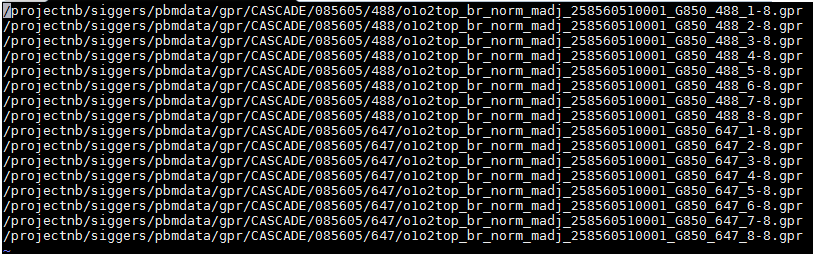
1. Make **list files for each type of file**. If you do different types of averaging, it is best to make separate lists and data matrices. This **file must contain the complete path to the files**.

If you averaged by br – br\_gpr.list

If you average by o – o1 (or o2)\_gpr.list

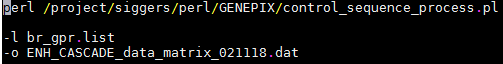
If you averaged over or – or\_gpr.list

Example:



1. Make a com file called make\_datamatrix.com. (I make separate com files for each type of file, see above).

Example:



1. The output file will be a .dat file. The name of the file will be whatever your specified in the com file\*\*general naming convention, then, is foo\_[ID].dat?\*\*.
2. Run the com file:

**Perl /project/siggers/perl/MISC/run\_comfile\_alert.pl –com make\_datamatrix.com –qsub matrix**

Note: To generate each matrix, the com file needs to be run separately.

Col 1 : probe sequence

Col 2 : reverse-complement of probe sequence

Col 3 : number indicated number of ‘data’ columns

Col 4 : Data column 1 – chamber 1

Col 5 : Data column 2 - chamber 2