Notes for pre-processing the Brown et al 2016 MEA NFA for the TCPL

Amy Carpenter

04/23/2020

Downloaded scripts from BitBucket repo

Ran MI scripts for all h5files

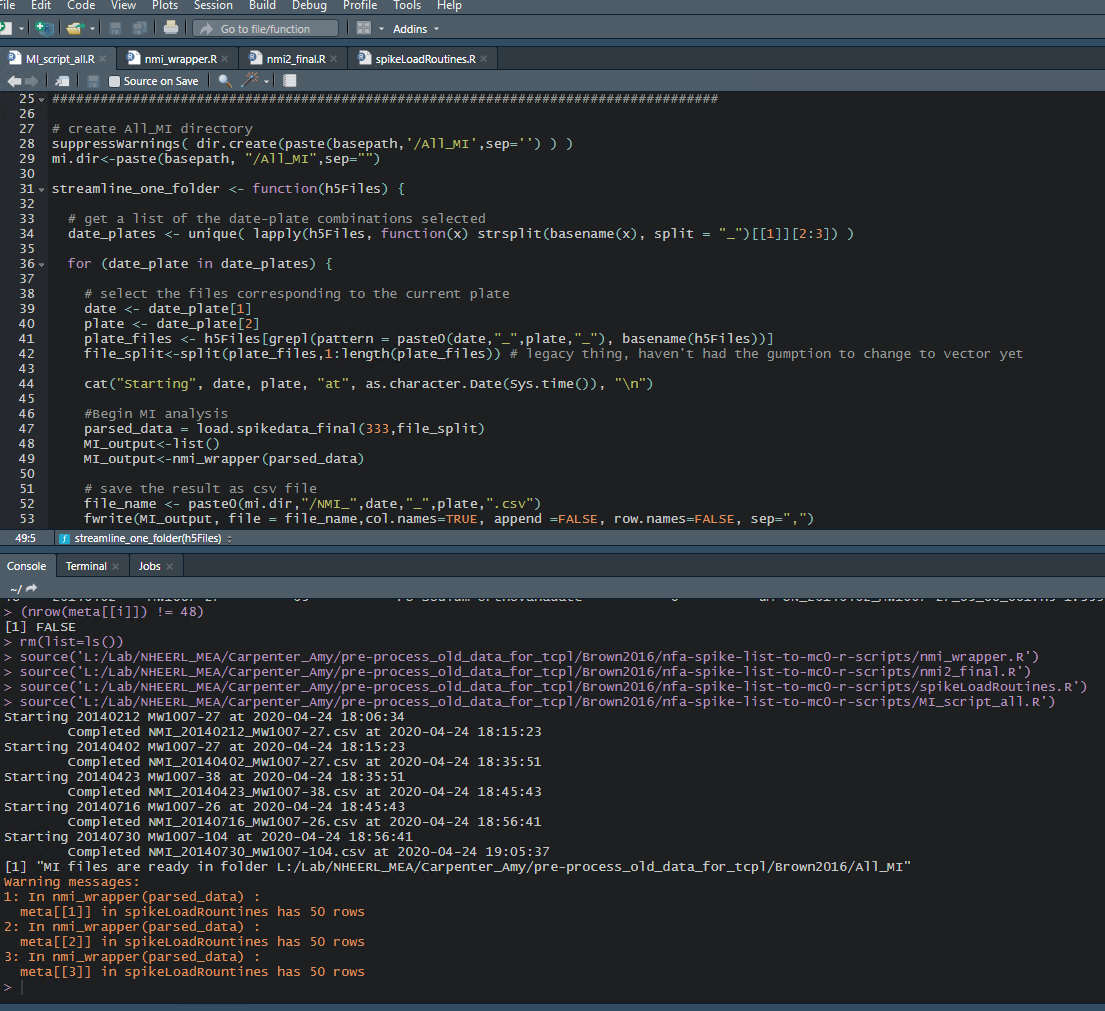
Cytotoxprep07.R should be ready to go. Just enter date and plate names as you go

However, something seems off with the first cytotox file – it says DIV 14, and it’s the wrong plate number. Can I find the correct file? What did the summary files use?

20140205 - L:\Lab\NHEERL\_MEA\PIP3 - Project\Data\Specific Aim 1\20140205 Ontogeny \*\* file says its for DIV 14...

04/24/2020

The MI file for the first plate compiled smoothly with the first run. However, I got an error message with the second plate that the meta data from the h5 file was 50 rows (instead of 48). The rows corresponding to A1 and A2 had been duplicated. I determined that this was indeed from the h5 file, and not from the script. I adjusted nmi\_wrapper2.R to take unique(meta[[i]]) to remove duplicate rows. I added a warning whenever the meta data had more than 48 rows before taking the unique() function.



Cytotoxicity status:

Based on the percent-of-control values presented in the file All\_SA1\_Cytotox\_Results\_Table\_pip3\_paper1.xlsx, the percent values in the Summary files agree. This indicates that the values in this tab are the Alamar Blue values, even though the section headings are e.g. “Mean Firing Rate” and “Active Electrodes”. I can use the cytotox\_prep06 function for 1 file to pull the data.

Interestingly, some of the cytotoxicity files calculated the blank-corrected value from more than 3 blank wells. This would be another reason to use the blank-corrected values as they are presented in the Summary files.

For consistency within my own data, and for agreement with the source data used to create All\_SA1\_Cytotox\_Results\_Table\_pip3\_paper1.xlsx, I think I should pull all of the data from these summary files, using whatever blank-correction method they chose to do for each plate.

I made a few edits to the script cytotox\_prep06.R: commented out the LDH sections, and set the tag phrase for the beginning index of the value map to 19 rows below the word “column” (instead of searching for the tag phrase “Corrected for blank”)

This was the output after running the script

> source('L:/Lab/NHEERL\_MEA/Carpenter\_Amy/pre-process\_old\_data\_for\_tcpl/Brown2016/nfa-spike-list-to-mc0-r-scripts/cytotox\_prep06.R')

[1] "no master chem list found, using treatment names from input data"

MW1007-26Summary.xlsx AB

MW1007-27Summary.xlsx AB

[1] "some blank-corrected values are negative. Setting these to zero"

MW1007-27Summary.xlsx AB

[1] "some blank-corrected values are negative. Setting these to zero"

MW1007-38Summary.xlsx AB

MW1007-26Summary.xlsx AB

MW1007-104Summary.xlsx AB

Brown2016\_cytotoxicity.csv is ready

I want to test and confirm that a – I gathered the correct values, and b – these are the correct raw data files to use.

Okay so… for some reason, the values are not actually the same (see compare\_cyto\_poc\_data.csv and All\_SA1\_Cytotox\_Results\_Table\_pip3\_paper1.xlsx). Possibilities:

- because I used the mean aggregate function?

- actually using different source data files or recordings?

- different POC function?

To do:

* Find out what is wrong with cyto data and fix
* Calculate the AUCs!
* Find the spids and replace (only 6 compounds!)
* Change wllq->0 where needed (hopefully no where)

Things to correct down the line:

* Will need to change DIV col for MI data
* Check that the data rows are in reasonable order for MI. Maybe plot and make sure it looks right? I’m a little concerned that h5 file might be different…
* Will need to change apid in cytotox data
* Remove Glypho, Van, other data from cytotox
* Renaming compound in cyto (note bis1 has multiple names)

04/27/2020

Possible reasons why POC cytotox values are different:

* We used different source data/raw data readings
* Maybe their agg function was median instead of mean?
* I scraped the wrong values from the summary files
* They did not calculate the summary files the same way that I did

Found the prob:

Not all of the blank-corrected files are the same distance from the tag phrase. I should have known…

Try getting the corrected for blank values for 3 time found of “Row” (then immediately check if raw data is right for each!)

I officially don’t know where else to look for the spid’s. Will need to email someone, probably.

Hmm… I know that we have tested all of the compounds as some point (yes, in Frank et al). Can I assume that the spids are the same? Actually probably yes. (Man, would be nice if one of the databases had the supplier, catalogue number, etc info, so that I could verify is the spid matches my records)

We have in Shafer\_103 list: Acetminophen, Loperamide (under long name)

Shafer\_42 list: Domoic acid, mevastatin, sodium orthovanadate, Bis1!

Next:

Remake cytotox file, confirm it is correct

Checkout datases again to see if can get any info on the chem suppliers, to verify if spids are correct

04/28/2020

Confirmed with Tim to use same spids as in Frank 2017 data set

**Ran cytotox\_prep06.R**

(This script was adjusted to accommodate this data set, namely it does not look for cytotoxicity data, and it looks for the value map based on the 3rd occurrence of the tag phrase “Row” in the sheet)

Input files: all files in folder for cytotoxicity data

Output file: Intermediate\_Output/Brown2016\_cytotoxicity.csv

The POC check – seems so close, but there are still some diff’s. Sodium ortho, where that data coming from? Especially last culture… none of those values are the same.

I could just open each of the 6 files, check the first and last blank-corrected value, make sure I am getting it, then move on – ya, that is probs the move

**Calculating parameter info**

Ran create\_ont\_csv.R with all files in h5files folder as input. Output files will be saved in the folder *prepared\_data*. Create\_burst\_ont\_Data.R was adjusted to match plate files by both date and plate SN, instead of just by plate, because of the re-used plates.

**Ran comb\_summary.R for prepared data and mutual information**

**For plate with missing DIV 9 recording**

Used the script add\_DIV9\_as\_mean.R to add DIV 9 values for Mw1007-104 based on the mean in the corresponding trt and dose wells in all of the other plates in this data set.

Input files: prepared\_data/Brown2016\_all\_prepared\_data.csv, All\_MI/Brown2016\_MI.csv

Output files: Intermediate\_Output/Brown2016\_ all\_prepared\_data\_added\_DIV9.csv, Intermediate\_Output/Brown2016\_ MI\_added\_DIV9.csv

**AUC calculation**

Ran Burst\_parameter\_to\_AUC.R

Added 2 lines of code to remove “DIV” letters from DIV column in source files, which occurred because of differently named h5 files

Input files: Intermediate\_Output/Brown2016\_ all\_prepared\_data\_added\_DIV9.csv, Intermediate\_Output/Brown2016\_ MI\_added\_DIV9.csv

Output file: Brown2016\_AUC.csv

**Cytotoxicity**

Ran cytotox\_prep06.R with all of the Summary files in the Cytotoxicity data folder as input. Output file named Intermediate\_outpute/Brown2016\_cytotoxicity.csv.

I checked the output of a test file from this script against the A1 and F8 wells for the blank-corrected AB data values in all 6 of the Summary files. I confirmed that he script collected the correct blank-corrected values for the correct wells.

Based on the output from the script, the following plates had some negative values that were set to zero:

MW1007-27, MW1007-27

**Creating long file**

Ran the script tcpl\_MEA\_AUC\_dev.R

Input files: Intermediate\_output/ Brown2016\_AUC.csv , Intermediate\_output/Brown2016\_cytotoxicity.csv

Output file: Brown2016\_longfile.csv

Other Settings:

Set default control treatment to DMSO. Set control treatment to Water for Acetaminophen, Domoic Acid, and Sodium Orthovanadate, based on Table 1 in paper.

Based on the following output, apid’s for the re-used plates have been re-assigned as follows:

source('L:/Lab/NHEERL\_MEA/Carpenter\_Amy/pre-process\_old\_data\_for\_tcpl/Brown2016/nfa-spike-list-to-mc0-r-scripts/tcpl\_MEA\_dev\_AUC.R')

The plate MW1007-26 was used in multiple culture dates ( 20140205 20140716 ).

[1] "MW1007-26, 20140205 is now assigned to MW1007-26a"

[1] "MW1007-26, 20140716 is now assigned to MW1007-26b"

The plate MW1007-27 was used in multiple culture dates ( 20140212 20140402 ).

[1] "MW1007-27, 20140212 is now assigned to MW1007-27a"

[1] "MW1007-27, 20140402 is now assigned to MW1007-27b"

**Assign spids**

Ran the script assign\_spids.R. Used the SPIDs found in the 2 files in the folder SPID\_Maps. These files are the same files used to assign spids for the Frank et al 2017 data set. Compounds were matched based on the CASRN’s found in table 1 in the Brown et al 2016 paper.

Input file: Brown2016\_longfile.csv

Output file: Brown2016\_longfile\_withspids.csv

Lingering concerns:

* Check that the data rows are in reasonable order for MI. Maybe plot and make sure it looks right? I’m a little concerned that h5 file might be different…
* Confirm that DIV 9 addition stuff looks okay
* Remove any of the wells mentioned in the one file?

04/30/2020

**Checking on the Mutual information, and all h5 files, because of the error/warning I saw when I preparing the MI data with the meta data having 50 rows instead of 48**

See the script check\_h5file\_length.R.

Confirmed that only h5 files from DIV 5, 7, and 9 on plate 1007-27 20140212 had anything other than exactly 48 rows per meta info (well, treatment, dose, and units). I confirmed that taking the unique() of the meta info does produce a table with 48 rows without any re-arranging.

In the second half of check\_h5file\_length.R, I confirm that the individual electrode name indexes do not appear off. It seems much more likely that only the external well, treatment, dose, etc. data just got duplicated. (the axion data would not be messed up, I highly doubt it and I don’t think it is reasonable to expect me to check that). And because we subset by a unique list of well.names, no well data is duplicated in the output h5File. See script for more details. I feel good about this now!

**Deciding what to do about the file that says certain cytotoxicity wells were excluded**

There are 4 percent of control values that were excluded in the analysis in the file All\_SA1\_Cytotox\_Results\_Table\_pip3\_paper1.xlsx

Let’s confirm if those percent of control values actually match what is in summary files (confirmed the first 2 Bis 1 at conc of 3 only 23%, and Acet at conc of 3 only 0%)

Arguments to include this data regardless:

* The note in the file All\_SA1\_Cytotox\_Results\_Table\_pip3\_paper1.xlsx just says “percentages in red have been excluded from the analysis”. It doesn’t give a reason to indicate why
  + I want to remove wells if there was something truly wrong (e.g., if incorrectly dosed, there was a spill, etc.) If the value is just mysteriously large, then I think we should still include it.
* It seems semi-common for AB assays to randomly have a few 0 values. This is a big problem that we should address. I feel like it is better to include the data to let the full variance of the assay be known, rather than try to hide it
* Why decided not to exclude any of the cytotox data for PFAS
* We are keeping values with POC 0% for Loperamide at high concentrations
* TCPL is designed to handle outliers
  + E.g. for Acetaminophen at conc 3, the values are 92, 9, 103 (POC values). I highly expect that the point of 0 might increase the RMSE, but it would not change the curve fit. And thus would not affect the hit call.

Not seeing any notes in the folders in Specific Aim 1. But they would probably be in the physical lab notebook regardless.

Reasons to exclude the values in red

* Consistency with what was done before, what has been published
  + But Tim asked me to re-run everything from the h5 files, because he wanted it to be updated

I am deciding to not exclude the four wells mentioned. I will mention that to Tim. I also want to clarify what our policy will be going forward, so that I don’t have to himmy and haw over this every time.

**Comparing old vs newly calculated data values with correlation plots**

See figs/correlation\_plots\_newly\_calculated\_vs\_original\_parameter\_values\_Brown2016.pdf correlation plots.

For the most part, the values are quite similar. The most difference occurs with r (the correlation coefficient). That makes sense because I updated how that is calculated (filtering by active electrode). Even still, these values are quite similar except for 3 old balls. That is expected.

Still, I am a bit concerned that that are several differences for the rest of the endpoints. I just would like to know what has changed

You know, other people might have tweaked the scripts since this data was calculated in 2016 (or earlier). Yes, that is very likely. And that is why Tim wanted me to re-run the pipeline anyhow.

These correlation plots confirm that nothing is wildly different without reason. And I have confirmed that the duplicated wells in the h5files would not be causing any of this.

Checked if there are any values that are NA in original but not new calculation (or vice versa). Output of script confirmed no.

[1] "meanfiringrate has 0 values that are NA in original but not in new."

[1] "meanfiringrate has 0 values that are NA in new but not in original."

[1] "burst.per.min has 0 values that are NA in original but not in new."

[1] "burst.per.min has 0 values that are NA in new but not in original."

[1] "mean.isis has 0 values that are NA in original but not in new."

[1] "mean.isis has 0 values that are NA in new but not in original."

[1] "per.spikes.in.burst has 0 values that are NA in original but not in new."

[1] "per.spikes.in.burst has 0 values that are NA in new but not in original."

[1] "mean.dur has 0 values that are NA in original but not in new."

[1] "mean.dur has 0 values that are NA in new but not in original."

[1] "mean.IBIs has 0 values that are NA in original but not in new."

[1] "mean.IBIs has 0 values that are NA in new but not in original."

[1] "nAE has 0 values that are NA in original but not in new."

[1] "nAE has 0 values that are NA in new but not in original."

[1] "nABE has 0 values that are NA in original but not in new."

[1] "nABE has 0 values that are NA in new but not in original."

[1] "ns.n has 0 values that are NA in original but not in new."

[1] "ns.n has 0 values that are NA in new but not in original."

[1] "ns.peak.m has 0 values that are NA in original but not in new."

[1] "ns.peak.m has 0 values that are NA in new but not in original."

[1] "ns.durn.m has 0 values that are NA in original but not in new."

[1] "ns.durn.m has 0 values that are NA in new but not in original."

[1] "ns.percent.of.spikes.in.ns has 0 values that are NA in original but not in new."

[1] "ns.percent.of.spikes.in.ns has 0 values that are NA in new but not in original."

[1] "ns.mean.insis has 0 values that are NA in original but not in new."

[1] "ns.mean.insis has 0 values that are NA in new but not in original."

[1] "ns.durn.sd has 0 values that are NA in original but not in new."

[1] "ns.durn.sd has 0 values that are NA in new but not in original."

[1] "ns.mean.spikes.in.ns has 0 values that are NA in original but not in new."

[1] "ns.mean.spikes.in.ns has 0 values that are NA in new but not in original."

[1] "r has 0 values that are NA in original but not in new."

[1] "r has 0 values that are NA in new but not in original."

[1] "cv.time has 0 values that are NA in original but not in new."

[1] "cv.time has 0 values that are NA in new but not in original."

[1] "cv.network has 0 values that are NA in original but not in new."

[1] "cv.network has 0 values that are NA in new but not in original."

I feel confident using the newly calculated values.

**Evaluating the added DIV 9 values for MW1007-104**

Looking at plots in MW1007-104\_comparison\_all\_wells\_with\_updates.pdfMW1007-104\_comparison\_control\_wells\_with\_updates.pdf

Note that all other plates in the culture were used to generate the DIV 9 values.

Ideas for improvements:

* Take the median instead of the mean
* For control wells, get the median of all control wells (those rows are still named by the compound in that row, so would be better to rename those data rows to “control”, then calculate median by that treatment name)

I will try using the median, and setting control values in a group together.

Prepared data files and AUC files with DIV 9 values added by mean or by median were created (and named as such). Boxplot visualization were created with the scripts in *supplemental\_scripts* .

Comparing by mean and by median additions (plus control by all controls, instead of by trt row)

Results vary only sliiiighty. Hardly at all. Range of boxplot values is perhaps slightly smaller with the by median values (which makes sense).

I want to move forward with the by-median values.

Comparing AUC values in 1007-104 with and without DIV 9, how it looks

Things that look okay:

Mean firing rate

Burst rate

Per burst interspike interval

IBI mean

Per burst spike percent

Percent of spikes in bursts

Mean burst duration

nAE

nABE

Network spike number

Newtwork spike pead

Mean network spike duration

Percent of spikes in network spikes

Interval between network spikes

SD of network spike duration

Mean # of spikes in network spikes

Correlation coefficient

Not sure the mean is an improvement

Mutual information - new values are definitely smaller. But that doesn’t mean better or worse

I think these added DIV 9 values by median (and median of all controls) for all plates is acceptable. Will move forward with these values

Summary of files and scripts:

prepared\_data/Brown2016\_all\_prepared\_data.csv, All\_MI/Brown2016\_MI.csv -> *add\_DIV9\_as\_median.R* ->

Intermediate\_output/Brown2016\_all\_prepared\_data\_added\_DIV9\_median.csv, Intermediate\_output/Brown2016\_MI\_added\_DIV9\_median.csv -> *burst\_parameter\_to\_AUC.R* ->

Brown2016\_AUC.csv, + Intermediate\_output/Brown2016\_cytotoxicity.csv -> *tcpl\_MEA\_dev\_AUC.R* ->

Brown2016\_longfile.csv -> *assign\_spids.R* ->

Intermediate\_output/Brown2016\_longfile\_withspids.csv