Merging the datasets

First, outer join rbm_tube with final_rbm_data on the sample ids. This join seems ok -- all final_rbm_data has 553 row and so does the left join with rbm_tube. Therefor there are no issues with rbm_tube key. Next, outer join the 553 immunoassay rectangle with the 554 cpl_rosetta on the mouse ids and we end up with 555 rows. Let's look at these two extra rows.

mouse_id <chr>	sample_id <chr>	Apo A1 (Apolipoprotein A1) <dbl>	Beta-2 Microglobulin <chr>
Maouse#3236	US-1140333	38.1	1.0600000000000001

mouse_id <chr>	sample_id <chr>	Apo A1 (Apolipoprotein A1) <dbl>	Beta-2 Microglobulin <chr>
Mouse#3236	*NA*	*NA*	*NA*

We see that there's a mis-spelling of 'mouse' in one of the columns. It's clear that it was a mistake to keep the 'mouse' string in the id. Going back and only extracting the number from the id columns in my script will solve this issue.

After changing this, since final_rbm_data has 553 row and cpl_rosetta has 554, we still expect to have an extra row when we do an outer join on the mouse_ids. After changing the mouse_ids into integers, this is now the case, noting that mouse 3597 is the row present in cpl_rosetta that is not present in final_rbm_data. It appears we have fixed this issue.

vWF (von Willebrand Factor) <chr>	Mouse <chr>	NEFA mEq/L <chr>	LDL mg/dL <chr>	HDL mg/dL <chr>	Total CHOL. mg/dL <chr>
NA	tid28315_Mouse3597_F2_Female	not received	not received	not received	not received

Merging in complete_f2 (554 rows) on this dataset by mouse_id maintains a row count of 554, so it appears there are no issues with this dataset.

Pheno_lipomics has 682 rows. An outer join and a filtering by NA values reveals that there is one row that is in the merged dataset that is not present in pheno_lipomics. That row has mouse_id 3611, and browsing pheno_lipomics the issue on the left is revealed (recall the spreasheet weas sorted by Mouse ID when imported) – a typo. To me, it is obvious that this value is a mistake and it is appropriate to manually correct this value within the R script. Not doing so leads to the loss of information for this mouse, and leads to an extra, dangling row that will be confusing to deal with later on.

	Mouse ID	SEX
604	3604.0	M
605	3605.0	F
606	3606.0	F
607	3607.0	F
608	3608.0	F
609	3609.0	M
610	3610.0	M
611	9.5	M
612	3612.0	M

Finally, we merge in necropsy_tracking (554 rows). With the current dataset at 682 rows after the outer join with pheno_lipomics, we outer join necropsy tracking and observe that no extra rows are added. This does confirm there are no issues, however an outer join with another dataset with 554 rows, cpl_rosetta, confirms that the mice in the necropsy_tracking dataset are the same as those in the cpl_dataset.

With this knowledge in hand, I've decided to proceed with the assignment using only the data that has complete information across all

the datasets – i.e. the 553 mice that have full observed data across all of the 5 datasets. To do this, I duplicate the operations above, performing left joins instead of outer joins. This results in 553 mice and 123 columns of data.

Data cleaning issues

Next, we need to look at these 123 columns. Visual inspection of the excel spreadsheets revealed that many of the columns have majority-numerical data that have missing or high/low values coded with text. All of these columns need to be re-formatted into numerical columns. My efforts will be partitioned into the datasets the variables originated as, the task was more manageable this way.

Final_rbm_data

The first rule of thumb to use, is that if R's data reading/importing function successfully recognized and imported a column as numeric, it should be fine. This means that there is no random text/NA coding in the column, though it does not ensure that there are not incorrectly entered values/outliers. This type of quality check should be performed in a different, later part of the analysis.

I find the table() and class() functions in R to be very helpful to filter out columns that are numeric, concentrate on those that are not, and then see what non-numeric values exist in the columns. Then, remove the non-numeric values, check if the column now only contains numeric values, if so, cast as numeric, and repeat until all columns are correctly casted.

A pervasive issue in this file is the use of '<LOW>' to encode that I presume to be immeasurably low data values. I'll re-code all instances of '<LOW>' as NA. If I were working with an investigator, I would make sure to ask whether it would be appropriate to change all '<LOW>' values to 0 instead. Doing this and re-casting all character columns with no non-numeric characters to numeric did 95% of the work. Three columns, Fibrinogen, Clusterin, and Growth Hormone, had a '>' prepended to a single observation, which was simple to fix manually.

Rbm_tube

Nothing needed to be done here. Issues with mouse_id typos have already been fixed earlier in the project.

Cpl_rosetta

Missing data and low observations were coded as 'not received', 'QNS', and 'less than 2.00'. These values were replaced as NAs and the columns were cast as numeric.

Complete_f2

Some errant '-', '?', and 'NA's in one of the columns were removed for NAs.

Pheno_lipomics_bleeds

'no plasma','-','missed', and 'dead' popped up, and were removed for NAs

necropsy_tracking

One final repeat of the process discussed above.