Data Processing Worflow

By Euchie

Summary of Steps from Raw Signals to Analyzable Concentrations:

- TRACE METALS:
  - Raw data processing **BEFORE** using R ***(slides 4 - 8)***
  - Raw data processing using R
    - **Step 1:** Calculating standard concentration. ***(slides 10 - 14)***
    - **Step 2:** Concentration calculation and dataset manipulation. ***(slides 15 - 18)***
    - **Step 3:** Adding environmental and biological data to concentration dataset ***(slide 19)***

- ORGANIC COMPOUNDS:
  - Raw data processing **BEFORE** using R ***(slides 22 - 26)***
  - Raw data processing using R
    - **Step 1:** Cleaning and Modifying Datasets. ***(slides 29)***
    - **Step 2:** Calculating Concentrations in MG/KG. ***(slides 30)***
    - **Step 3:** Creating final concentration dataset. ***(slides 31)***

FOR TRACE METALS

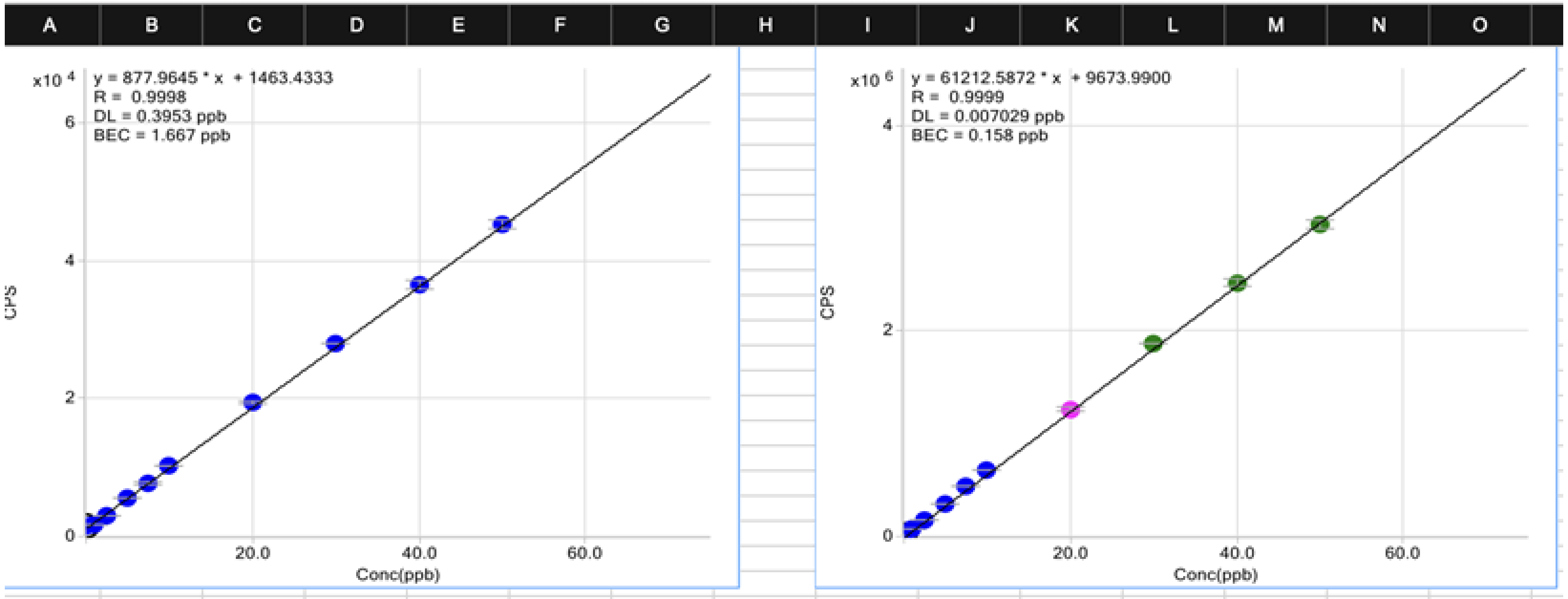
PROCESSING RAW DATA BEFORE USING R

PROCESS INTRODUCTION

- About 30-40 samples (one batch) were analyzed per week.
- Inksac and muscle samples were analyzed first, for their respective batches, due to the presumed reduced sediments in the digested samples to avoid cross-contamination during chemical analysis.
- Calibration curves were constructed for every batch analyzed each week.
  - Calibration curves are used to quantify the instrumental response of an analyte (Trace metals, e.g Fe), and to predict the concentration of the analyte in a sample.
  - Inductively Coupled Plasma Mass Spectrometry (ICP-MS).

PROCESSING THE RAW DATA BEFORE R

- After instrumental analysis the raw data is received in an excel file.
- The file has two sheets
  - **Sheet 1** has the plotted calibration curves for all ten trace metals



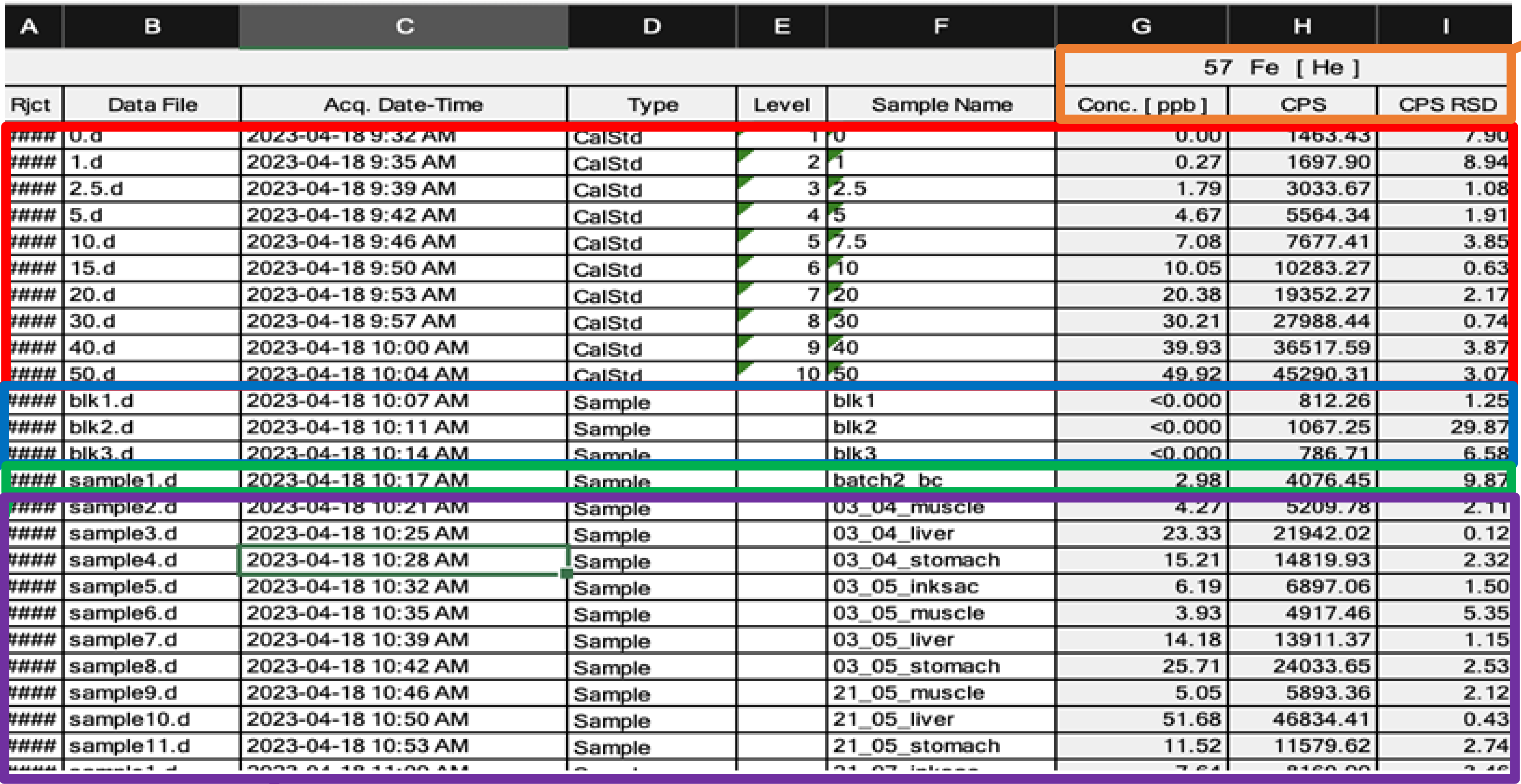
Snippet showing two out of the ten calibration graphs with the ten calibration standards in each graph.

PROCESSING THE RAW DATA BEFORE R
(see excel snippet showing one trace metal (Fe) on next slide)

- **Sheet 2** has:

  - The first ten rows are the calibration standards used to make the calibration curve.
  - The rows below these ten rows are the instrument wash cycles labelled as wash/blk.
  - After the rows with the wash cycles comes the blank control. That is used to check for the background noise in the solution after trace metal extraction from the samples.
    - The solution used to wash the instrument is 2% nitric acid which is also used as the internal standard.
    - The solution used for blank control has 2% nitric acid and a diluted version of the strong acid used to extract the trace metals from the sample.
  - After the blank control are the 30-40 analyzed samples.
    - the samples are named using Area_ID number_Tissue (e.g. 65_01_muscle)
  - The columns are arranged in this manner.
    - there are 10 trace metals.
    - each trace metal are further sub-divided into 3 columns namely:
      - Concentration (ppb) of the sample
      - Count per seconds (CPS). how many ions of the specific trace metal that hit the detection plate in the instrument per second.
      - Counts per seconds relative standard deviation (CPS RSD). The rate percentage at which the hit the detection plate.

SNIPPPET OF RAW DATA EXCEL FILE AFTER INSTRUMENT ANALYSIS



A	B	C	D	E	F	G	H	I
						57 Fe [He]		
Rjct	Data File	Acq. Date-Time	Type	Level	Sample Name	Conc. [ppb]	CPS	CPS RSD
####	0.d	2023-04-18 9:32 AM	CalStd	1	0	0.00	1463.43	7.90
####	1.d	2023-04-18 9:35 AM	CalStd	2	1	0.27	1697.90	8.94
####	2.5.d	2023-04-18 9:39 AM	CalStd	3	2.5	1.79	3033.67	1.08
####	5.d	2023-04-18 9:42 AM	CalStd	4	5	4.67	5564.34	1.91
####	10.d	2023-04-18 9:46 AM	CalStd	5	7.5	7.08	7677.41	3.85
####	15.d	2023-04-18 9:50 AM	CalStd	6	10	10.05	10283.27	0.63
####	20.d	2023-04-18 9:53 AM	CalStd	7	20	20.38	19352.27	2.17
####	30.d	2023-04-18 9:57 AM	CalStd	8	30	30.21	27988.44	0.74
####	40.d	2023-04-18 10:00 AM	CalStd	9	40	39.93	36517.59	3.87
####	50.d	2023-04-18 10:04 AM	CalStd	10	50	49.92	45290.31	3.07
####	blk1.d	2023-04-18 10:07 AM	Sample		blk1	<0.000	812.26	1.25
####	blk2.d	2023-04-18 10:11 AM	Sample		blk2	<0.000	1067.25	29.87
####	blk3.d	2023-04-18 10:14 AM	Sample		blk3	<0.000	786.71	6.58
####	sample1.d	2023-04-18 10:17 AM	Sample		batch2_bc	2.98	4076.45	9.87
####	sample2.d	2023-04-18 10:21 AM	Sample		03_04_muscle	4.27	5209.78	2.11
####	sample3.d	2023-04-18 10:25 AM	Sample		03_04_liver	23.33	21942.02	0.12
####	sample4.d	2023-04-18 10:28 AM	Sample		03_04_stomach	15.21	14819.93	2.32
####	sample5.d	2023-04-18 10:32 AM	Sample		03_05_inksac	6.19	6897.06	1.50
####	sample6.d	2023-04-18 10:35 AM	Sample		03_05_muscle	3.93	4917.46	5.35
####	sample7.d	2023-04-18 10:39 AM	Sample		03_05_liver	14.18	13911.37	1.15
####	sample8.d	2023-04-18 10:42 AM	Sample		03_05_stomach	25.71	24033.65	2.53
####	sample9.d	2023-04-18 10:46 AM	Sample		21_05_muscle	5.05	5893.36	2.12
####	sample10.d	2023-04-18 10:50 AM	Sample		21_05_liver	51.68	46834.41	0.43
####	sample11.d	2023-04-18 10:53 AM	Sample		21_05_stomach	11.52	11579.62	2.74
####	sample1.d	2023-04-18 11:00 AM	Sample		21_07_inksac	7.61	8160.00	2.46

Trace metal

- Concentration (ppb)
- CPS
- CPS RSD

Calibration curve

Wash cycles

Blank control

Squid samples

PROCESSING RAW DATA BEFORE R

Preliminary checks were made:

- To check if squid IDs were entered correctly.
- To check or make changes for efficient processing in R.
- To check if any comments were added in terms of processing.

FOR TRACE METALS

PROCESSING RAW DATA USING R

(***See: 1-Data_Preprocessing/Rscripts/Raw_trace_metals_data_to_preprocessed_data.R***)

PROCESSING RAW DATA USING R

Step 1: Calculating the Standard concentration for each batch:

- **Sheet 2** is saved as a CSV file, which is then used to calculate the standard concentrations.
- The standard concentrations represent the concentrations of the calibration standards which will be used later to compute the, **slope**, **intercept**, **limit of detection (LOD)** and **limit of quantification (LOQ)** using **average intensity**. This is calculated using the first ten rows of the CSV file (**typically from row 3-12**).
- First the **slope** and **intercept** are obtained which will later be used in Step 4, to calculate the actual concentration ppb for each sample.

PROCESSING RAW DATA USING R

Step 1: Calculating the Standard concentration for each batch (Continued):

- The **slope** was calculated using linear regression and the **intercept** is actually represented by the **average intensity** of the 0ppb calibration standard.

- For the trace metals that have concentrations between 0 and 1 ppb. The first 5 calibration standards were used to do the regression. This is because the more points you include, the less accurate the concentrations will be, especially at low concentrations. For the trace metals that have high concentrations all the calibration standards were used to run the linear regression.

PROCESSING RAW DATA USING R

Step 1: Calculating the Standard concentration for each batch (Continued):

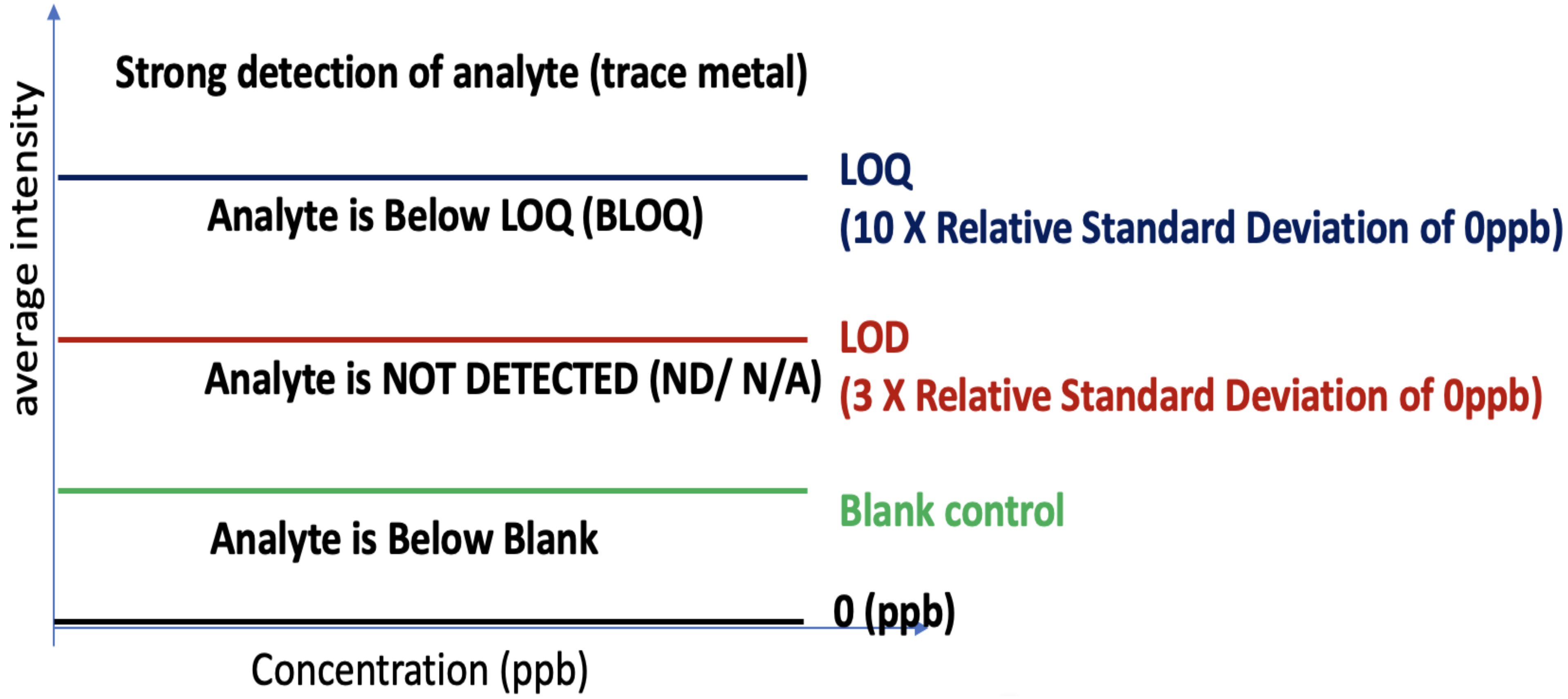
LOD and LOQ are calculated.

- LOD = **Average intensity of the trace metal (CPS) in the 0ppb calibration standard + 3 X (The relative standard deviation of the trace metal in the 0ppb calibration standard (CPS RSD) X 0.01 X Average intensity of the trace metal in the 0ppb calibration standard (CPS))**
- LOQ **= Average intensity of the trace metal (CPS) in the 0ppb calibration standard + 10 X (The relative standard deviation of the trace metal in the 0ppb calibration standard (CPS RSD) X 0.01 X Average intensity of the trace metal in the 0ppb calibration standard (CPS))**
- The LOD and LOQ will later be used to quantify the concentration (ppb) for each sample in Step 2.

PROCESSING RAW DATA USING R

Step 1: Calculating the Standard concentration for each batch:

- After the LOD and LOQ are calculated sampleconcentrations are classified using below schema:


Strong detection of analyte (trace metal)
LOQ
(10 X Relative Standard Deviation of 0ppb)
Analyte is Below LOQ (BLOQ)
LOD
(3 X Relative Standard Deviation of 0ppb)
Analyte is NOT DETECTED (ND/ N/A)
Blank control
Analyte is Below Blank
0 (ppb)
average intensity
Concentration (ppb)

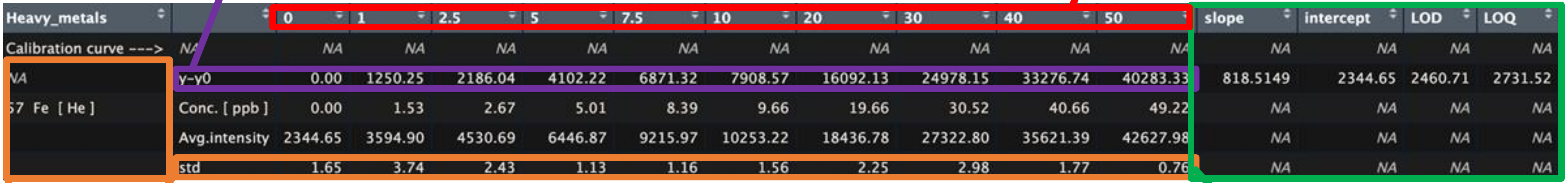
PROCESSING RAW DATA USING R

Step 1: Calculating the Standard concentration for each batch (Continued):

Final product from **Step 1**: (Year)Std_concentration_ppb.csv

# Check point 1: Chose one metal in output (**std_conc dataset)** to check if calculations are correct.

Snippet showing first 5 rows from std_conc dataset in R



Heavy_metals		0	1	2.5	5	7.5	10	20	30	40	50	slope	intercept	LOD	LOQ
Calibration curve --->	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	y-y0	0.00	1250.25	2186.04	4102.22	6871.32	7908.57	16092.13	24978.15	33276.74	40283.33	818.5149	2344.65	2460.71	2731.52
57 Fe [He]	Conc. [ppb]	0.00	1.53	2.67	5.01	8.39	9.66	19.66	30.52	40.66	49.22	NA	NA	NA	NA
	Avg.intensity	2344.65	3594.90	4530.69	6446.87	9215.97	10253.22	18436.78	27322.80	35621.39	42627.98	NA	NA	NA	NA
	std	1.65	3.74	2.43	1.13	1.16	1.56	2.25	2.98	1.77	0.76	NA	NA	NA	NA

Calibration standards

y-y0 = Avg. intensity of 1ppb to 50ppb –Avg. intensity of 0ppb

CPS RSD

Trace metal

Slope is calculated from regression
Intercept is Avg. intensity of 0ppb
LOD and **LOQ** for trace metal (given in CPS)

PROCESSING RAW DATA USING R

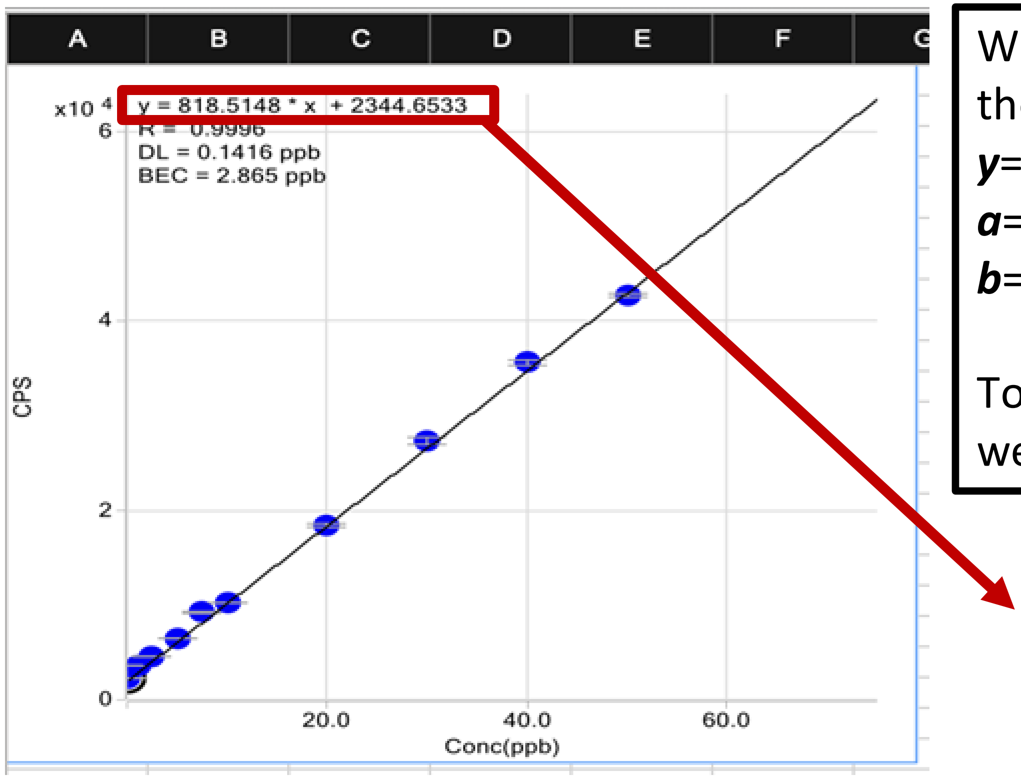
Step 2: Concentration calculation and dataset manipulation

- Getting the dataset into the format needed for statistical analysis.
- The dry weight dataset is first formatted to be able to be added to the concentration dataset.
  - The dry weight for each sample is added to dataset to help calculate the final concentration mg/kg.

PROCESSING RAW DATA USING R

Step 2: Concentration calculation and dataset manipulation (Continued)

- Calculation of concentration ppb (**Conc_ppb**) see red-outlined column in next slide



When finding the Concentration ppb (**Conc_ppb**) we first use the **transformation function** (***y=ax+b***) and then solve for ***X***.
y= Avg.intensity of sample (e.g 03_29_muscle, see next slide)
a= slope from std_conc dataset (see slide 14 for example)
b= intercept from std_conc dataset (see slide 14 for example)

To solve for ***X:***
we first subtract ***b*** from ***y*** and divide by ***a***

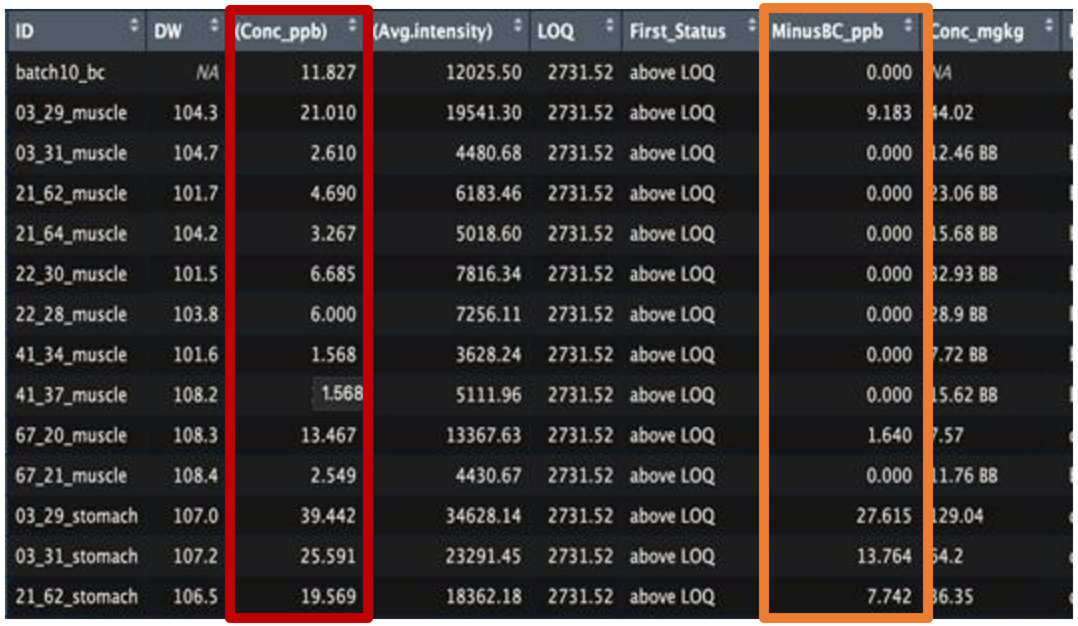
Transformation function used in calibration curve from ICP-MS

Snippet of Fe calibration curve from ICP-MS

PROCESSING RAW DATA USING R

Step 2: Concentration calculation and dataset manipulation (Continued)

- Calculating **MinusBC_ppb**:



ID	DW	(Conc_ppb)	(Avg.intensity)	LOQ	First_Status	MinusBC_ppb	Conc_mgkg
batch10_bc	NA	11.827	12025.50	2731.52	above LOQ	0.000	NA
03_29_muscle	104.3	21.010	19541.30	2731.52	above LOQ	9.183	44.02
03_31_muscle	104.7	2.610	4480.68	2731.52	above LOQ	0.000	12.46 BB
21_62_muscle	101.7	4.690	6183.46	2731.52	above LOQ	0.000	23.06 BB
21_64_muscle	104.2	3.267	5018.60	2731.52	above LOQ	0.000	15.68 BB
22_30_muscle	101.5	6.685	7816.34	2731.52	above LOQ	0.000	32.93 BB
22_28_muscle	103.8	6.000	7256.11	2731.52	above LOQ	0.000	28.9 BB
41_34_muscle	101.6	1.568	3628.24	2731.52	above LOQ	0.000	7.72 BB
41_37_muscle	108.2	1.568	5111.96	2731.52	above LOQ	0.000	15.62 BB
67_20_muscle	108.3	13.467	13367.63	2731.52	above LOQ	1.640	7.57
67_21_muscle	108.4	2.549	4430.67	2731.52	above LOQ	0.000	11.76 BB
03_29_stomach	107.0	39.442	34628.14	2731.52	above LOQ	27.615	129.04
03_31_stomach	107.2	25.591	23291.45	2731.52	above LOQ	13.764	54.2
21_62_stomach	106.5	19.569	18362.18	2731.52	above LOQ	7.742	36.35

Snippet showing first 14 rows from in R output

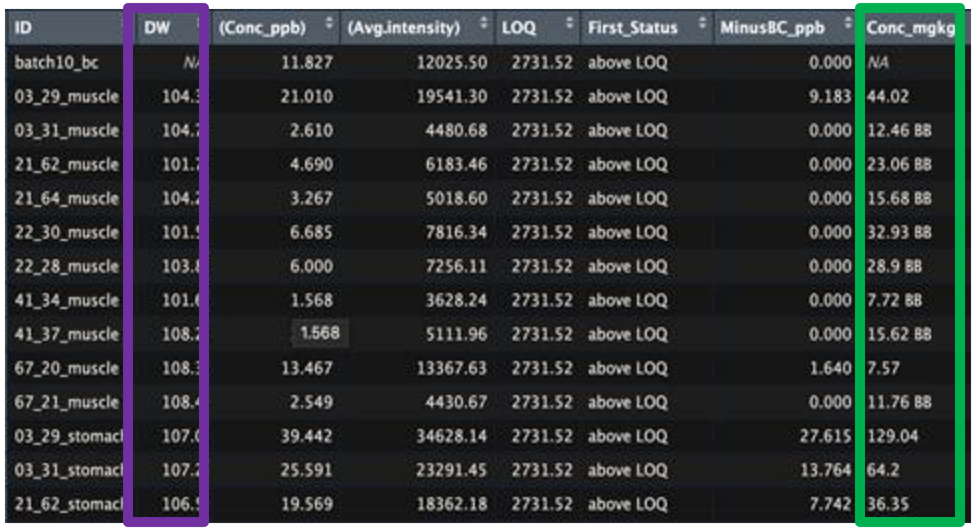
When finding the **MinusBC_ppb**:
We first compare **sample Conc_ppb** (e.g 03_29_muscle) with **Conc_ppb of blank control** (e.g batch10_bc = 11.87):

- If **sample Conc_ppb** > **Conc_ppb of blank control** then **sample Conc_ppb** - **Conc_ppb of blank control** (e.g 21.010-11.827 = 9.183) and results placed in **MinusBC_ppb.**
- If **sample Conc_ppb < Conc_ppb of blank control** then "0" is placed in **MinusBC_ppb.**

PROCESSING RAW DATA USING R

Step 2: Concentration calculation and dataset manipulation (Continued)

- Calculating **Conc_mg/kg**:
- # Check point 2 : choose at least 3 rows (top, middle and bottom) to check calculations.



ID	DW	(Conc_ppb)	(Avg.intensity)	LOQ	First_Status	MinusBC_ppb	Conc_mgkg
batch10_bc	NA	11.827	12025.50	2731.52	above LOQ	0.000	NA
03_29_muscle	104.3	21.010	19541.30	2731.52	above LOQ	9.183	44.02
03_31_muscle	104.7	2.610	4480.68	2731.52	above LOQ	0.000	12.46 BB
21_62_muscle	101.7	4.690	6183.46	2731.52	above LOQ	0.000	23.06 BB
21_64_muscle	104.2	3.267	5018.60	2731.52	above LOQ	0.000	15.68 BB
22_30_muscle	101.5	6.685	7816.34	2731.52	above LOQ	0.000	32.93 BB
22_28_muscle	103.8	6.000	7256.11	2731.52	above LOQ	0.000	28.9 BB
41_34_muscle	101.6	1.568	3628.24	2731.52	above LOQ	0.000	7.72 BB
41_37_muscle	108.2	1.568	5111.96	2731.52	above LOQ	0.000	15.62 BB
67_20_muscle	108.3	13.467	13367.63	2731.52	above LOQ	1.640	7.57
67_21_muscle	108.4	2.549	4430.67	2731.52	above LOQ	0.000	11.76 BB
03_29_stomach	107.0	39.442	34628.14	2731.52	above LOQ	27.615	129.04
03_31_stomach	107.2	25.591	23291.45	2731.52	above LOQ	13.764	64.2
21_62_stomach	106.5	19.569	18362.18	2731.52	above LOQ	7.742	36.35

Snippet showing first 14 rows from R output

When finding the **Conc_mg/kg**:

- We first multiply the Concentration (ppb) in **MinusBC_ppb** by 100 and 5 then divide by the dry weight (**DW/1000**) and divide again by 1000. (e.g 9.183*100*5)/(104.3/1000)/1000= 44.02
- If **MinusBC_ppb =** 0, it means that only noise was detected. In that case the above formula was used to calculate **Conc_mg/kg** directly from the **Conc_ppb** without subtracting it from blank control. Letters like **'BB'** (below blank) or **'BLOQ'** were added to identify these concentrations (subject to change).

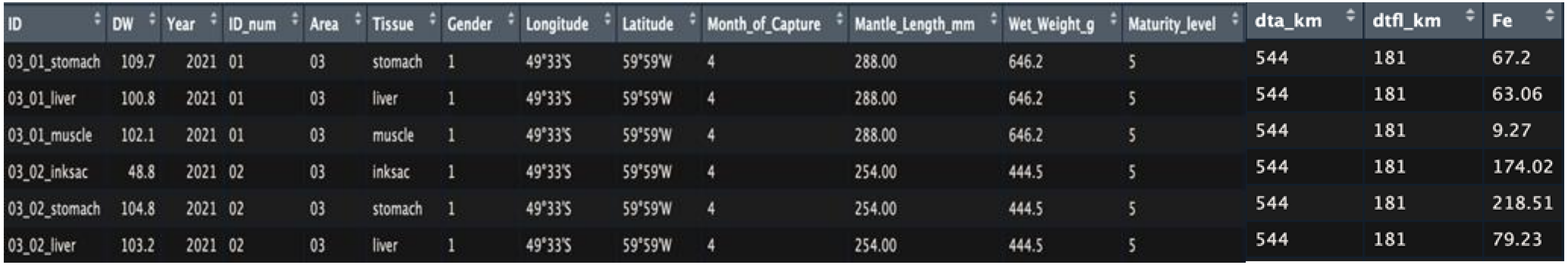
NB. 100= how many times the sample was diluted and 5 is the volume of the extraction solvent used in each sample.

PROCESSING RAW DATA USING R

Step 3: Adding environmental and biological data to concentration dataset.

- The squid information from the fishing vessel and lab were added to the concentration dataset.
- The current batch is appended to the previous batches
- # Check point 3: choose at least 3 rows (top, middle and bottom) to check calculations.

Snippet showing first 6 rows from R output.



ID	DW	Year	ID_num	Area	Tissue	Gender	Longitude	Latitude	Month_of_Capture	Mantle_Length_mm	Wet_Weight_g	Maturity_level	dta_km	dtfl_km	Fe
03_01_stomach	109.7	2021	01	03	stomach	1	49°33'S	59°59'W	4	288.00	646.2	5	544	181	67.2
03_01_liver	100.8	2021	01	03	liver	1	49°33'S	59°59'W	4	288.00	646.2	5	544	181	63.06
03_01_muscle	102.1	2021	01	03	muscle	1	49°33'S	59°59'W	4	288.00	646.2	5	544	181	9.27
03_02_inksac	48.8	2021	02	03	inksac	1	49°33'S	59°59'W	4	254.00	444.5	5	544	181	174.02
03_02_stomach	104.8	2021	02	03	stomach	1	49°33'S	59°59'W	4	254.00	444.5	5	544	181	218.51
03_02_liver	103.2	2021	02	03	liver	1	49°33'S	59°59'W	4	254.00	444.5	5	544	181	79.23

PROCESSING RAW DATA USING R

NB. The total number of rows after combining all the batches **DO NOT** include the samples that were not measured due to insufficient tissue amount, specifically those squids with small inksacs.

- For 2019 trace metals analysis no ink sacs were measured
- For 2020 and 2021 trace metals analysis not all ink sacs were analyzed due to this issue.

FOR ORGANIC COMPOUNDS

PROCESSING RAW DATA BEFORE USING R

PROCESS INTRODUCTION

- About 40-90 samples (one batch) were analyzed per month.
- Calibration curves were constructed for every organic compound (4) analyzed.
    - Calibration curves were used to quantify the instrumental response of an analyte (organic compound, e.g Metolachor), and to predict the concentration of analyte in a sample by using the Area ratio and the concentration ppb.
    - LC-MS coupled with Electrospray Ionization Mass Spectrometry (ESI-MS) were used

PROCESSING THE RAW DATA BEFORE R

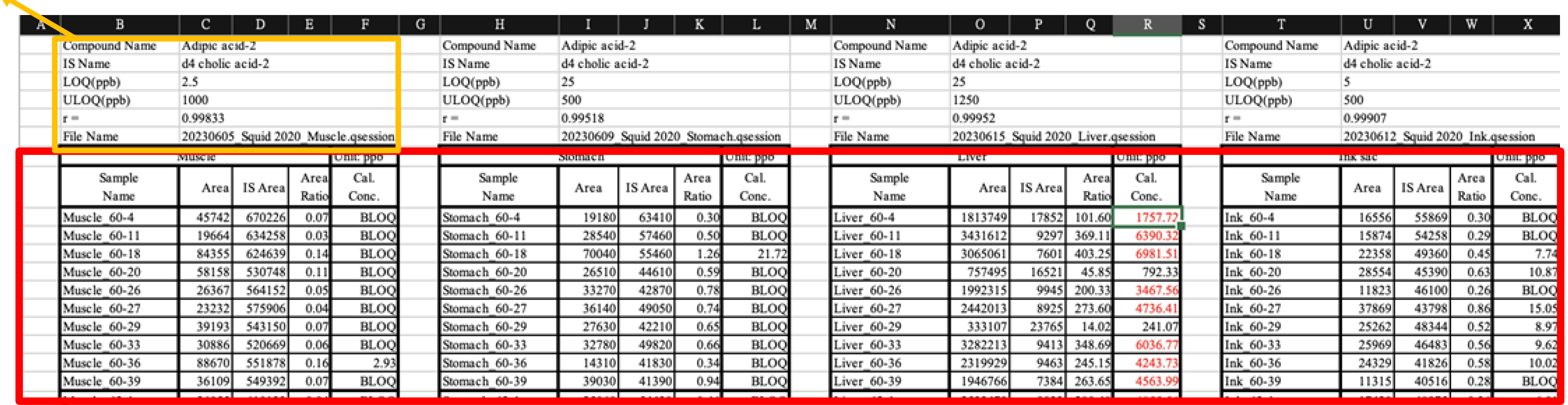
- After instrumental analysis the raw data is received in an excel file.
- The file has 7 sheets.
  - The beginning 2 or 3 sheets such as Calibration Curve, LC-MS condition and **IS-ratio** are mainly for reference.
  - The 4 sheets that come after, are each an organic compound with all the samples that were measured for that organic compound.
  - Each sheet has four tables (one for each tissue) with the samples listed under.

    Snippet on next slide explaining table structure for organic compound (Adipic acid)

  - At the top of each table reads the (batch information): name of the **organic compound**, the **internal standard(IS name)** used, the **LOQ** and **upper LOQ** for that organic compound within that specific tissue, the **R2 score** and the file name.
  - The calculated concentration was already given in ppb as compared to trace metals and is in the “Calc. Conc” column.

SNIPPPET OF RAW DATA EXCEL FILE FOR organic compoundS AFTER INSTRUMENT ANALYSIS

Batch information for muscle



Compound Name	Adipic acid-2	Adipic acid-2	Adipic acid-2	Adipic acid-2
IS Name	d4 cholic acid-2	d4 cholic acid-2	d4 cholic acid-2	d4 cholic acid-2
LOQ(ppb)	2.5	25	25	5
ULOQ(ppb)	1000	500	1250	500
r =	0.99833	0.99518	0.99952	0.99907
File Name	20230605_Squid 2020_Muscle.qsession	20230609_Squid 2020_Stomach.qsession	20230615_Squid 2020_Liver.qsession	20230612_Squid 2020_Ink.qsession

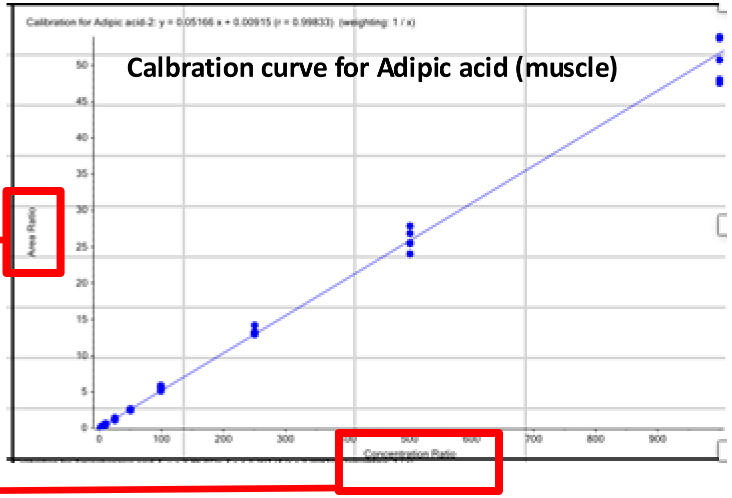
Muscle				Unit: ppb
Sample Name	Area	IS Area	Area Ratio	Cal. Conc.
Muscle_60-4	45742	670226	0.07	BLOQ
Muscle_60-11	19664	634258	0.03	BLOQ
Muscle_60-18	84355	624639	0.14	BLOQ
Muscle_60-20	58158	530748	0.11	BLOQ
Muscle_60-26	26367	564152	0.05	BLOQ
Muscle_60-27	23232	575906	0.04	BLOQ
Muscle_60-29	39193	543150	0.07	BLOQ
Muscle_60-33	30886	520669	0.06	BLOQ
Muscle_60-36	88670	551878	0.16	2.93
Muscle_60-39	36109	549392	0.07	BLOQ

Stomach				Unit: ppb
Sample Name	Area	IS Area	Area Ratio	Cal. Conc.
Stomach_60-4	19180	63410	0.30	BLOQ
Stomach_60-11	28540	57460	0.50	BLOQ
Stomach_60-18	70040	55460	1.26	21.72
Stomach_60-20	26510	44610	0.59	BLOQ
Stomach_60-26	33270	42870	0.78	BLOQ
Stomach_60-27	36140	49050	0.74	BLOQ
Stomach_60-29	27630	42210	0.65	BLOQ
Stomach_60-33	32780	49820	0.66	BLOQ
Stomach_60-36	14310	41830	0.34	BLOQ
Stomach_60-39	39030	41390	0.94	BLOQ

Liver				Unit: ppb
Sample Name	Area	IS Area	Area Ratio	Cal. Conc.
Liver_60-4	1813749	17852	101.60	1757.72
Liver_60-11	3431612	9297	369.11	6390.32
Liver_60-18	3065061	7601	403.25	6981.51
Liver_60-20	757495	16521	45.85	792.33
Liver_60-26	1992315	9945	200.33	3467.56
Liver_60-27	2442013	8925	273.60	4736.41
Liver_60-29	333107	23765	14.02	241.07
Liver_60-33	3282213	9413	348.69	6036.77
Liver_60-36	2319929	9463	245.15	4243.73
Liver_60-39	1946766	7384	263.65	4563.99

Ink sac				Unit: ppb
Sample Name	Area	IS Area	Area Ratio	Cal. Conc.
Ink_60-4	16556	55869	0.30	BLOQ
Ink_60-11	15874	54258	0.29	BLOQ
Ink_60-18	22358	49360	0.45	7.74
Ink_60-20	28554	45390	0.63	10.87
Ink_60-26	11823	46100	0.26	BLOQ
Ink_60-27	37869	43798	0.86	15.05
Ink_60-29	25262	48344	0.52	8.97
Ink_60-33	25969	46483	0.56	9.62
Ink_60-36	24329	41826	0.58	10.02
Ink_60-39	11315	40516	0.28	BLOQ

1. Samples arranged by tissue.
2. **Area Ratio** = **Area/IS Area**
3. **Concentration (ppb)/ Cal. Conc.** = **Area ratio** plotted on the calibration curve then compared with **LOQ(ppb)** above.
    - If **Cal. Conc** < **LOQ(ppb)** then **Cal. Conc** = BLOQ.
    - If **Cal. Conc** > **LOQ(ppb)** then **Cal. Conc** = Concentration.
    - If **NO AREA** was detected then **Cal. Conc = NA.**



Calbration curve for Adipic acid (muscle)

Area ratio

Concentration (ppb)

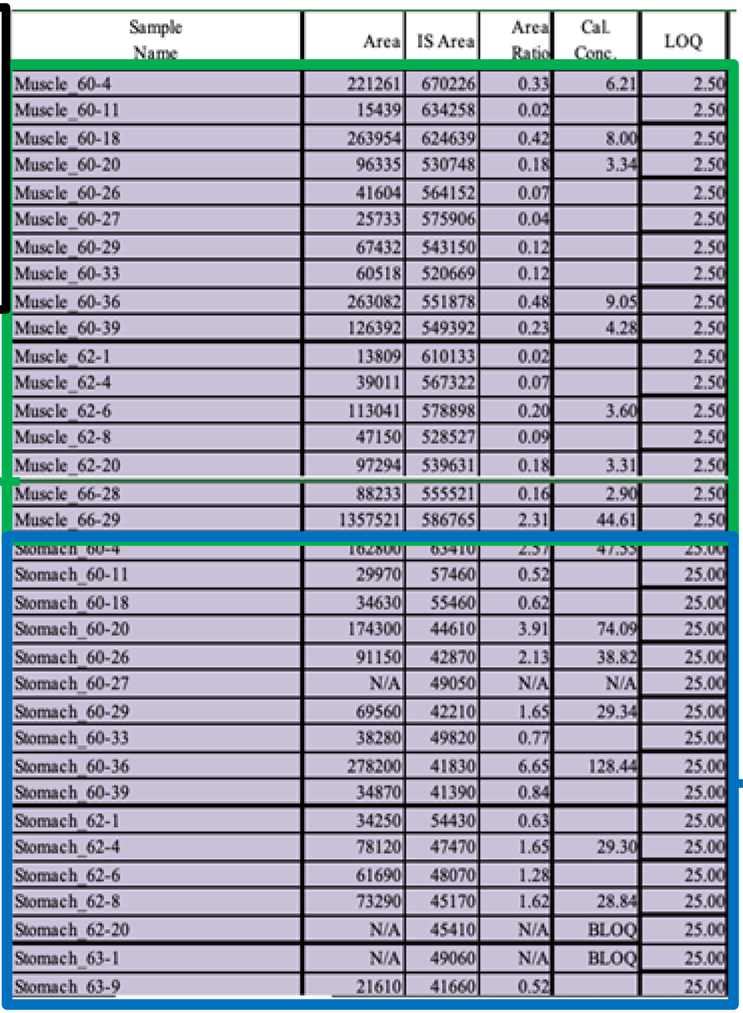
PROCESSING RAW DATA BEFORE R

Preliminary checks were changes made:

- To check if squid IDs were entered correctly.
- To check or make changes for efficient processing in R.
- To check if any comments were added in terms of processing.
- To add numerical LOQ column.

PROCESSING THE RAW DATA BEFORE R

After receiving raw data from Chem dept. the tables in each sheet are then arranged on top of each other (long format) to be able to be processed in R



Muscle

Stomach

Sample Name	Area	IS Area	Area Ratio	Cal. Conc.	LOQ
Muscle_60-4	221261	670226	0.33	6.21	2.50
Muscle_60-11	15439	634258	0.02		2.50
Muscle_60-18	263954	624639	0.42	8.00	2.50
Muscle_60-20	96335	530748	0.18	3.34	2.50
Muscle_60-26	41604	564152	0.07		2.50
Muscle_60-27	25733	575906	0.04		2.50
Muscle_60-29	67432	543150	0.12		2.50
Muscle_60-33	60518	520669	0.12		2.50
Muscle_60-36	263082	551878	0.48	9.05	2.50
Muscle_60-39	126392	549392	0.23	4.28	2.50
Muscle_62-1	13809	610133	0.02		2.50
Muscle_62-4	39011	567322	0.07		2.50
Muscle_62-6	113041	578898	0.20	3.60	2.50
Muscle_62-8	47150	528527	0.09		2.50
Muscle_62-20	97294	539631	0.18	3.31	2.50
Muscle_66-28	88233	555521	0.16	2.90	2.50
Muscle_66-29	1357521	586765	2.31	44.61	2.50
Stomach_60-4	162800	63410	2.57	47.55	25.00
Stomach_60-11	29970	57460	0.52		25.00
Stomach_60-18	34630	55460	0.62		25.00
Stomach_60-20	174300	44610	3.91	74.09	25.00
Stomach_60-26	91150	42870	2.13	38.82	25.00
Stomach_60-27	N/A	49050	N/A	N/A	25.00
Stomach_60-29	69560	42210	1.65	29.34	25.00
Stomach_60-33	38280	49820	0.77		25.00
Stomach_60-36	278200	41830	6.65	128.44	25.00
Stomach_60-39	34870	41390	0.84		25.00
Stomach_62-1	34250	54430	0.63		25.00
Stomach_62-4	78120	47470	1.65	29.30	25.00
Stomach_62-6	61690	48070	1.28		25.00
Stomach_62-8	73290	45170	1.62	28.84	25.00
Stomach_62-20	N/A	45410	N/A	BLOQ	25.00
Stomach_63-1	N/A	49060	N/A	BLOQ	25.00
Stomach_63-9	21610	41660	0.52		25.00

Snippet of Adipic acid table arranged in long format for processing in R

FOR ORGANIC COMPOUND

PROCESSING RAW DATA USING R

(***See: 1-Data_Preprocessing/Rscripts/ Raw_organic_compounds_data_to_preprocessed_data.R***)

PROCESSING RAW DATA USING R

The sheets in the raw data excel file are first converted to CSV using a chunk of code in the R script (***lines 372-384***) and saved as a folder in the current working directory.

- The folder, immediately after processing in R, is saved as the current year_month_day_time of its processing.
- That folder is then loaded into R from the working directory and the sheets within the folder are saved into a list
- Sample separation data (dry weight dataset) was loaded as ‘dry_weight’.

PROCESSING RAW DATA USING R

Step 1: CLEANING AND MODIFYING DATASETS :

- Each concentration dataset in the list were cleaned and modified:
    - Remove extra letters added to the sample names.
    - Remove columns that are not needed (e.g. 'Dilu_Factor’ etc..)
    - Remove rows that are not needed for statistical analysis (e.g calibration standards)
    - Replace “-” with “_”.
    - Replace concentrations that are ‘BLOQ’ with numerical values.
    - Add extra columns.
    - Rearrange samples name so sample IDs can resemble that of trace metals.
    - Rename columns.
    - Formatted sample IDs in each CSV file need to resemble those in trace metals dataset (60_04_muscle).
- Check point 1: data_manipulation1 function output was printed (top, middle and bottom rows checked)
- Check point 2: Data_manipulation2 function output was printed (top, middle and bottom rows checked)

PROCESSING RAW DATA USING R

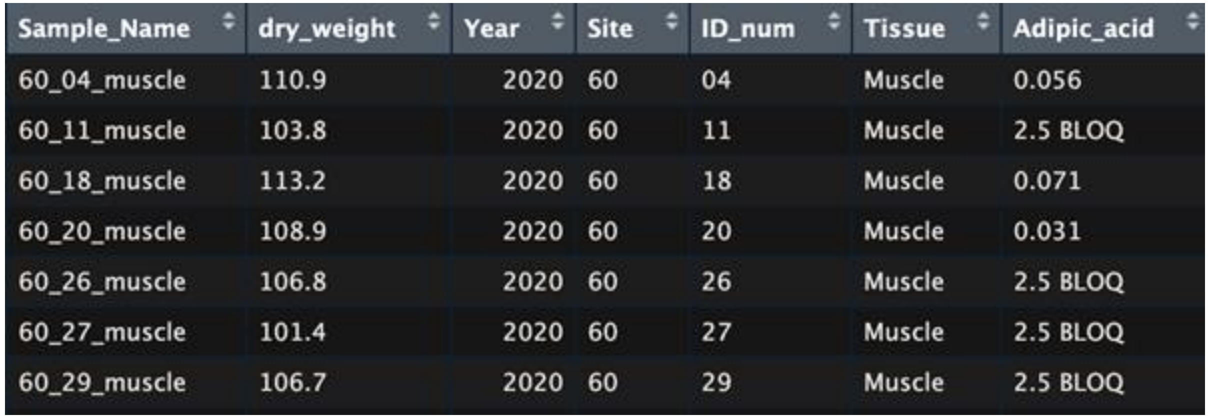
Step 2: CALCULATING CONCETRATION MG/KG:

- The sample ID column for dry weight dataset was formatted to resemble that of the concentration dataset (e.g. 60_04)
- The dry weight for each sample (row) in each concentration dataset within the list was added.
- The **concentration** for each sample (row) in each concentration dataset within the list was simultaneously calculated.
    - The concentrations with the exception of samples that show BLOQ were processed as concentration mg/kg.(***formula= (sample Calculated concentration (ppb)/sample dry weight)*1000/1000***
    - Those concentrations that were BLOQ were processed as the LOQ (ppb) for that tissue + "BLOQ" (e.g 250 BLOQ) (subject to change)
    - Those concentrations that were N/A were processed as "0" since no concentration was detected.
- Check point 3: add_dry_weight function output was printed (top, middle and bottom rows checked)

PROCESSING RAW DATA USING R

Step 3: CREATING FINAL CONCENTRATIONS DATASET:

- More manipulation was done on this dataset later to have it resemble the trace metals dataset. A few more columns were added and some columns were renamed.



Sample_Name	dry_weight	Year	Site	ID_num	Tissue	Adipic_acid
60_04_muscle	110.9	2020	60	04	Muscle	0.056
60_11_muscle	103.8	2020	60	11	Muscle	2.5 BLOQ
60_18_muscle	113.2	2020	60	18	Muscle	0.071
60_20_muscle	108.9	2020	60	20	Muscle	0.031
60_26_muscle	106.8	2020	60	26	Muscle	2.5 BLOQ
60_27_muscle	101.4	2020	60	27	Muscle	2.5 BLOQ
60_29_muscle	106.7	2020	60	29	Muscle	2.5 BLOQ

Snippet of final processed organic compounds dataset in R

Check point 4: organic_compound_concentration_dataset was printed (top, middle and bottom rows checked)