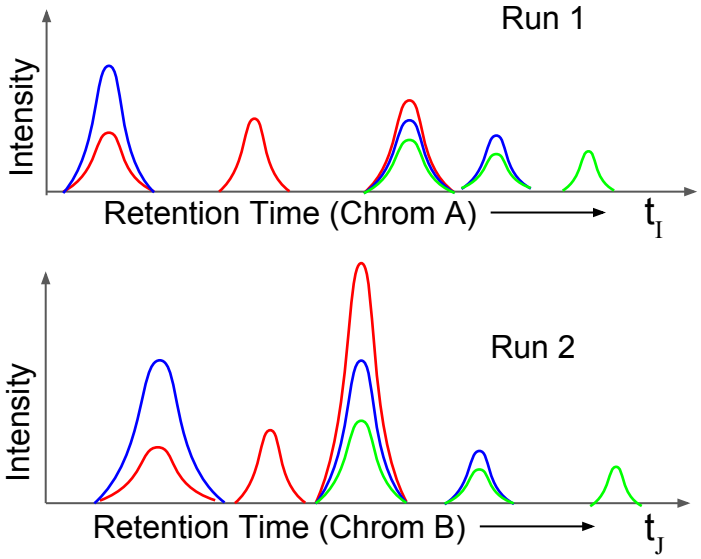
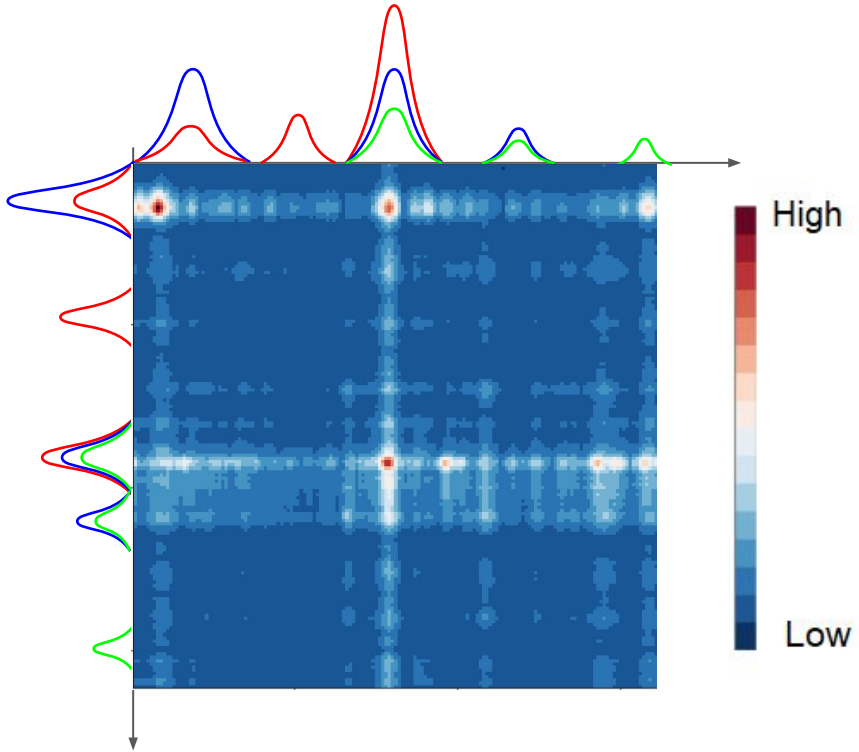
DIAlign extension: alignment of multiple runs

Irina Utkina
Rotation student

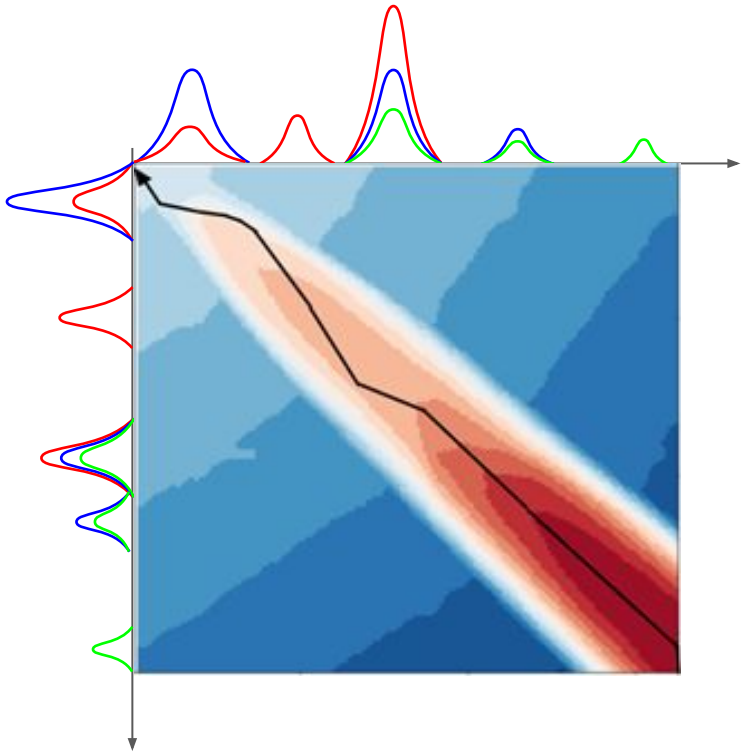
November 15, 2018

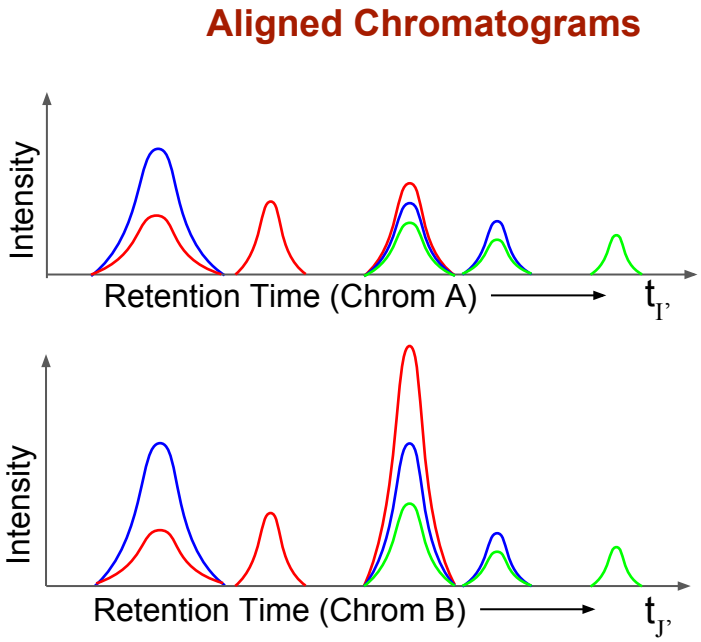
DIAlign implementation





DIAlign implementation

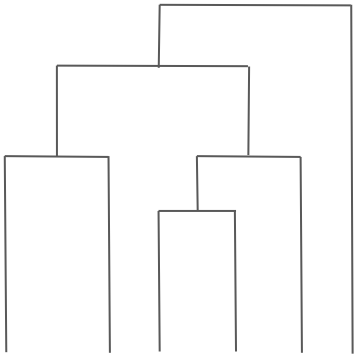


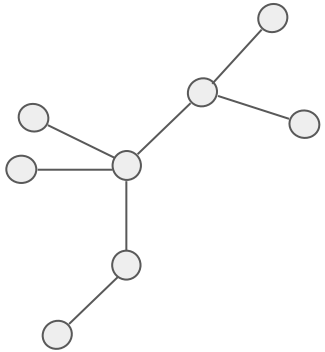


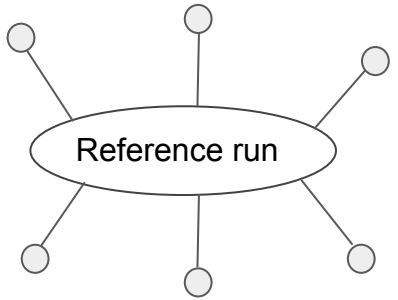
The aim of the project:

Make alignment scalable for large number of runs.

Possible solutions:



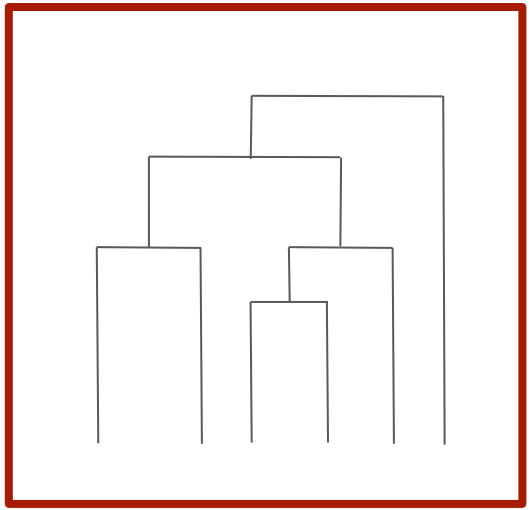


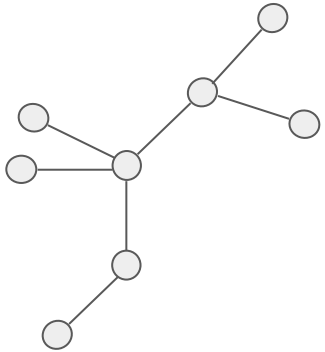


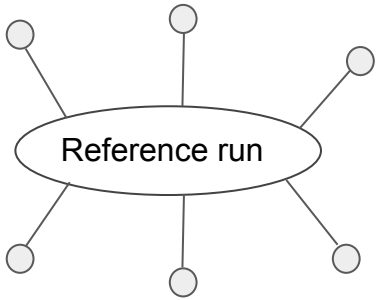
The aim of the project:

Make alignment scalable for large number of runs.

Possible solutions:



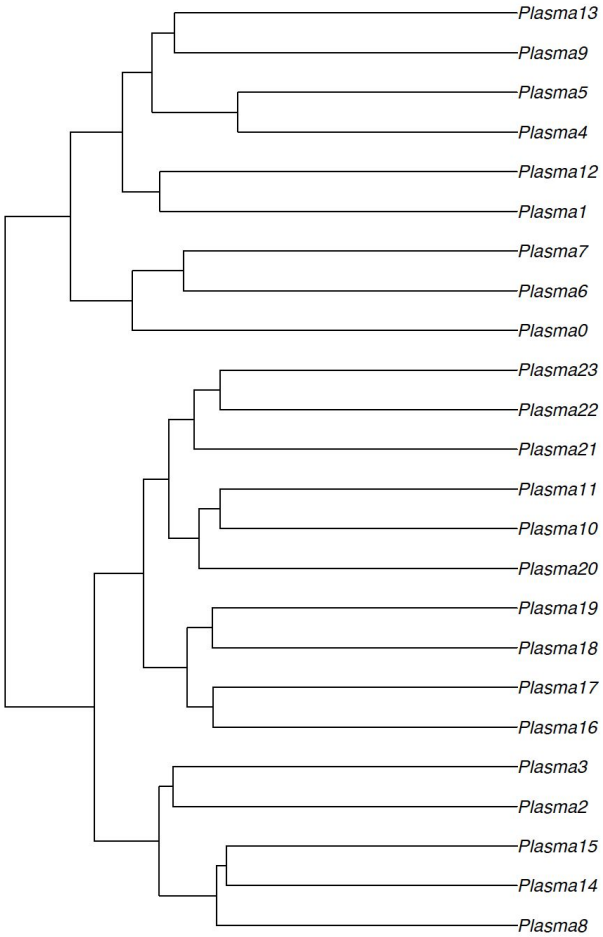




Building a hierarchical tree

1. Based on distances resulted from LOESS prediction (openSWATH) - **global tree**
2. Based on raw data - build similarity matrices for each pairwise alignment resulting in **k trees for k peptides**:
   - find a good metrics that can be used as a distance between runs
   - find an optimal approach to calculating retention time and intensity during runs fusion
   - find the best way to deal with missing values in the alignment

Global tree


Plasma13
Plasma9
Plasma5
Plasma4
Plasma12
Plasma1
Plasma7
Plasma6
Plasma0
Plasma23
Plasma22
Plasma21
Plasma11
Plasma10
Plasma20
Plasma19
Plasma18
Plasma17
Plasma16
Plasma3
Plasma2
Plasma15
Plasma14
Plasma8

Local alignment (k trees for k peptides)

1. **Find a good metrics that can be used as a distance between runs**
   - Total number of gaps
   - Median of the score matrix
   - Average maximum score
   - Distribution of normalized* maximum score at each step

Additionally, ensure that the chosen metrics is reasonable by comparison of alignments for random peptides with corresponding alignments.

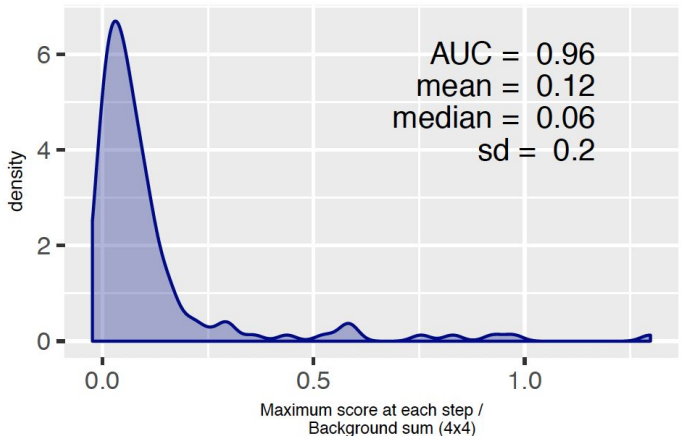
*

37	33	27	12	4
40	29	22	17	8
33	31	13	15	6
16	11	9	7	2
4	2	12	1	5

Background sum

Maximum score (at each step)

Taking into account the signal from background noise



37	33	27	12	4
40	29	22	17	8
33	31	13	15	6
16	11	9	7	2
4	2	12	1	5

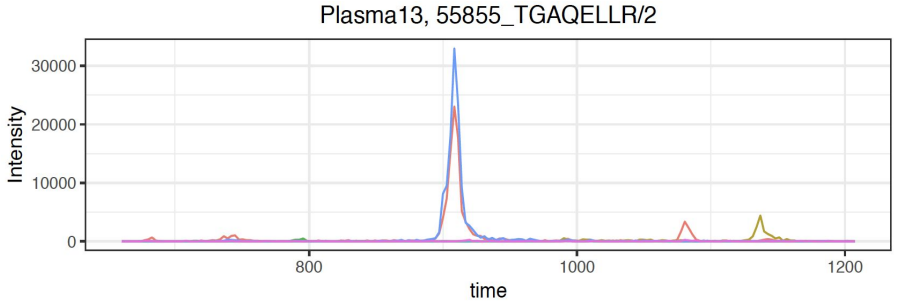
Background sum

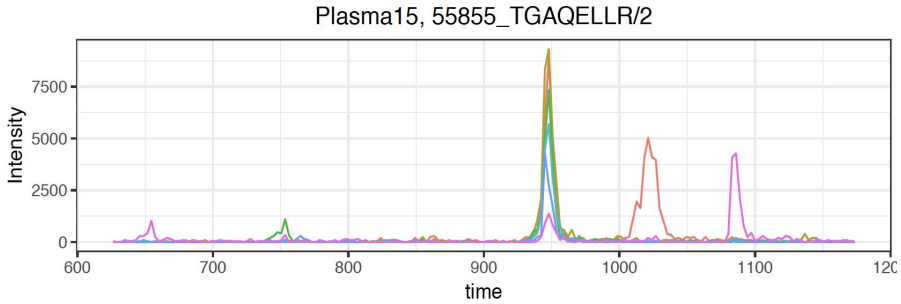
Maximum score (at each step)

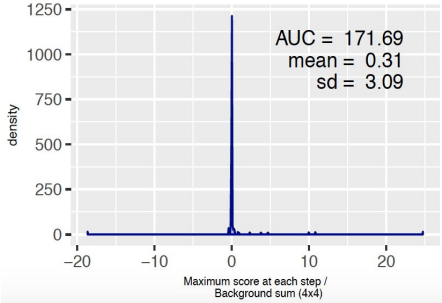
At each step we calculate: $\frac{\text{Maximum score}}{\text{Background sum (4x4)}}$

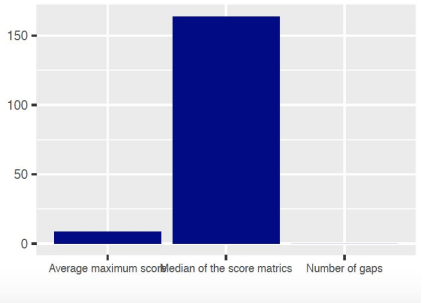
For the particular case above first ratio: 5/190 = 0.0263
second ratio: 7/362 = 0.0193

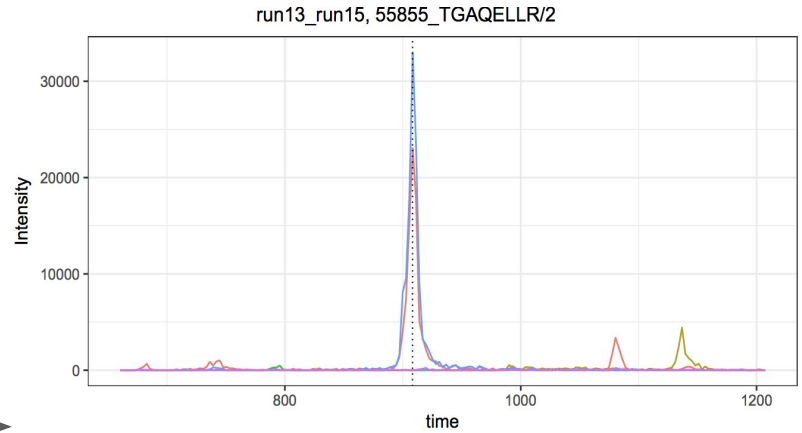
Example of a good alignment

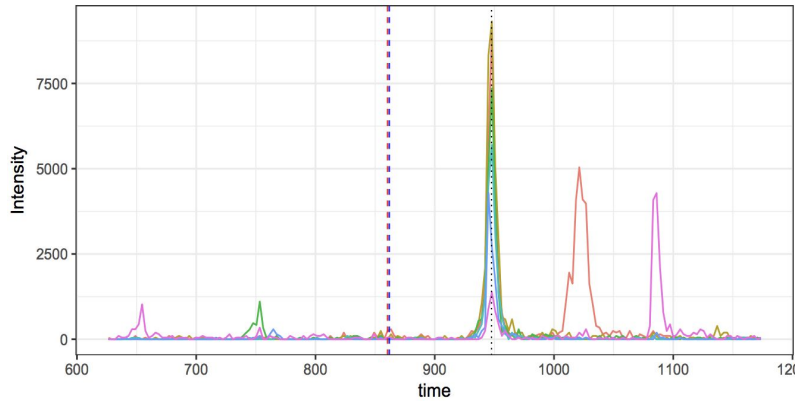




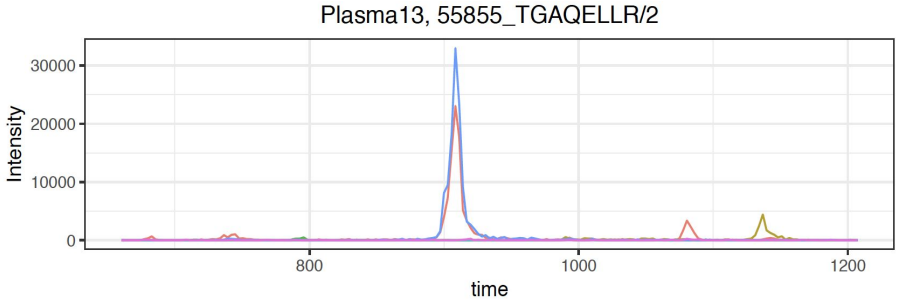


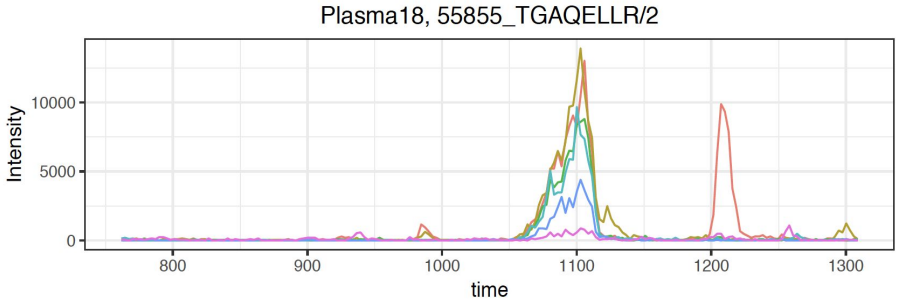


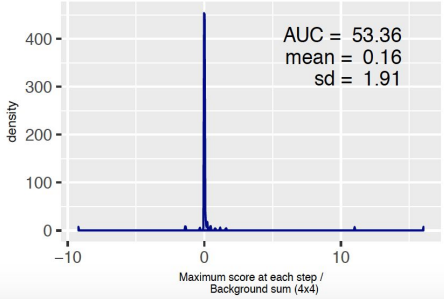


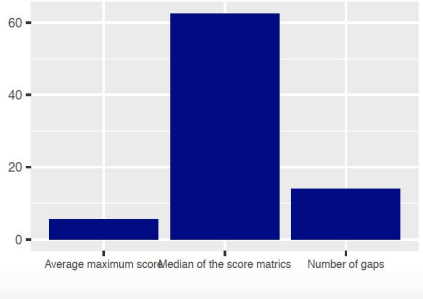


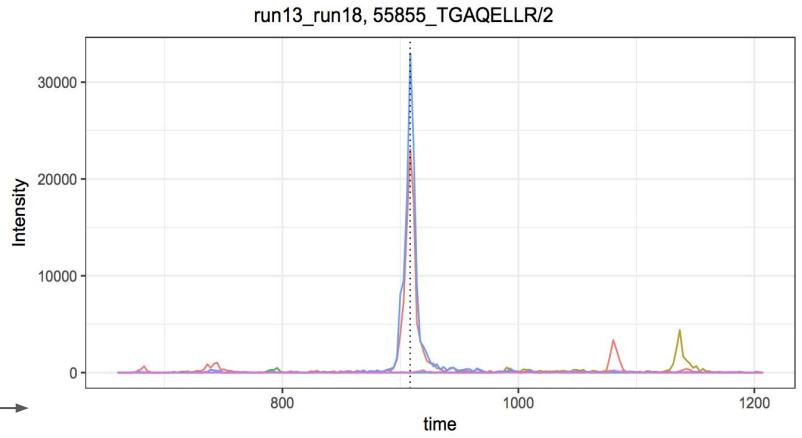
Example of a bad alignment

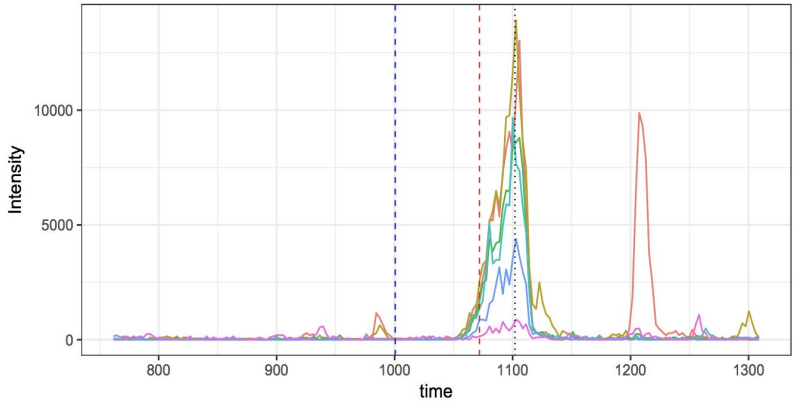




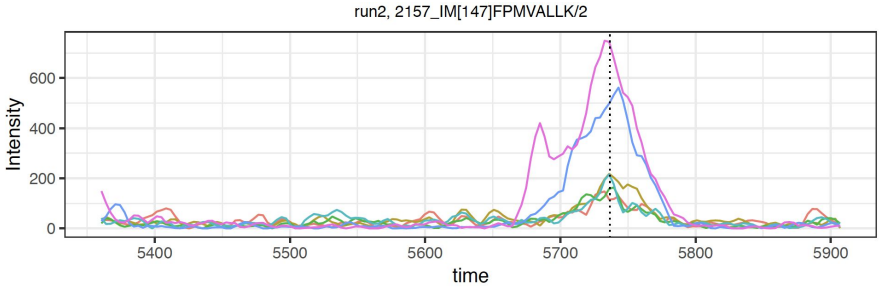


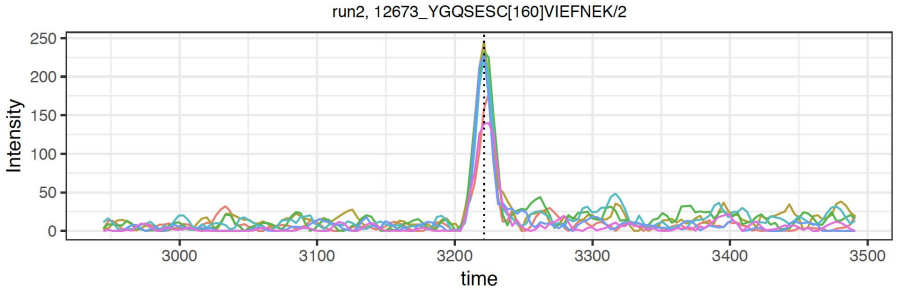


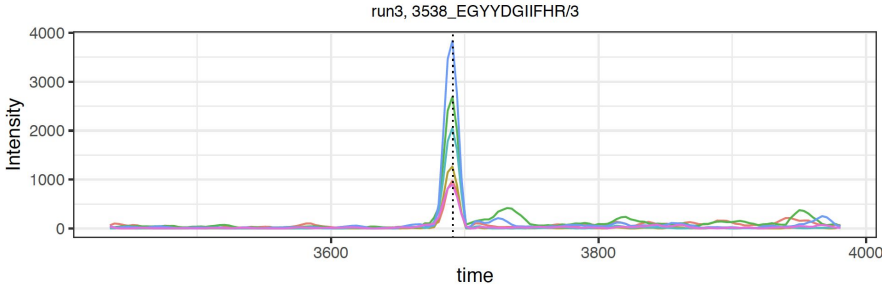


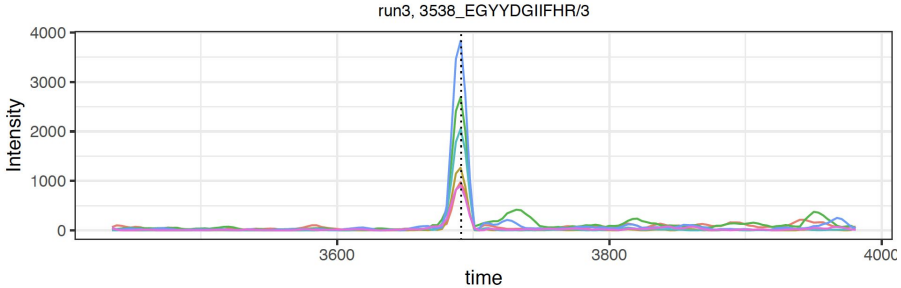


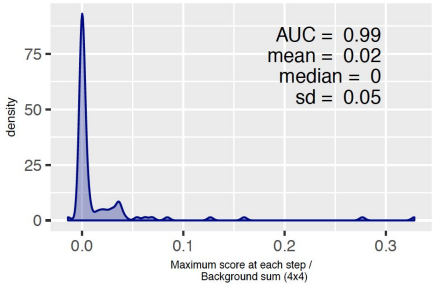
Alignments of random peptides

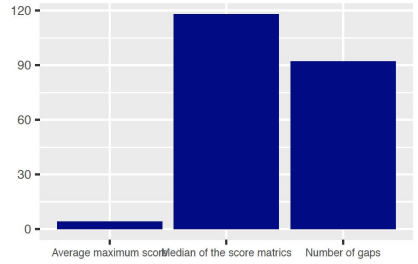


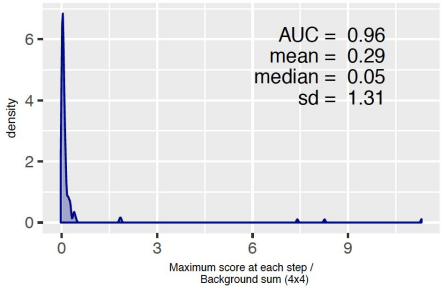


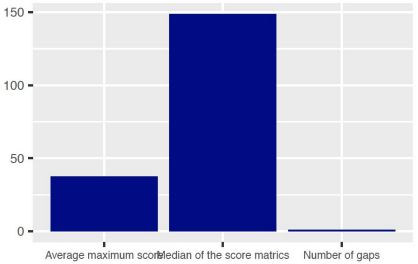










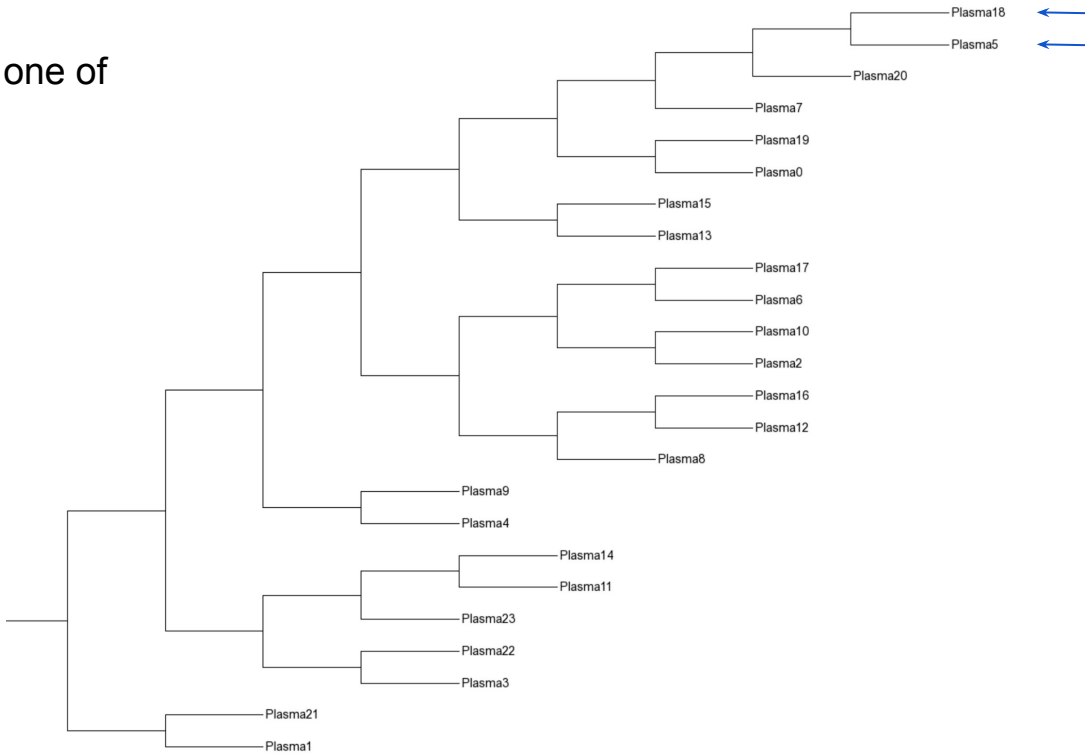


Chosen distance metrics:

$$\frac{(1 + \text{Total number of Gaps})}{\text{AUC}}$$

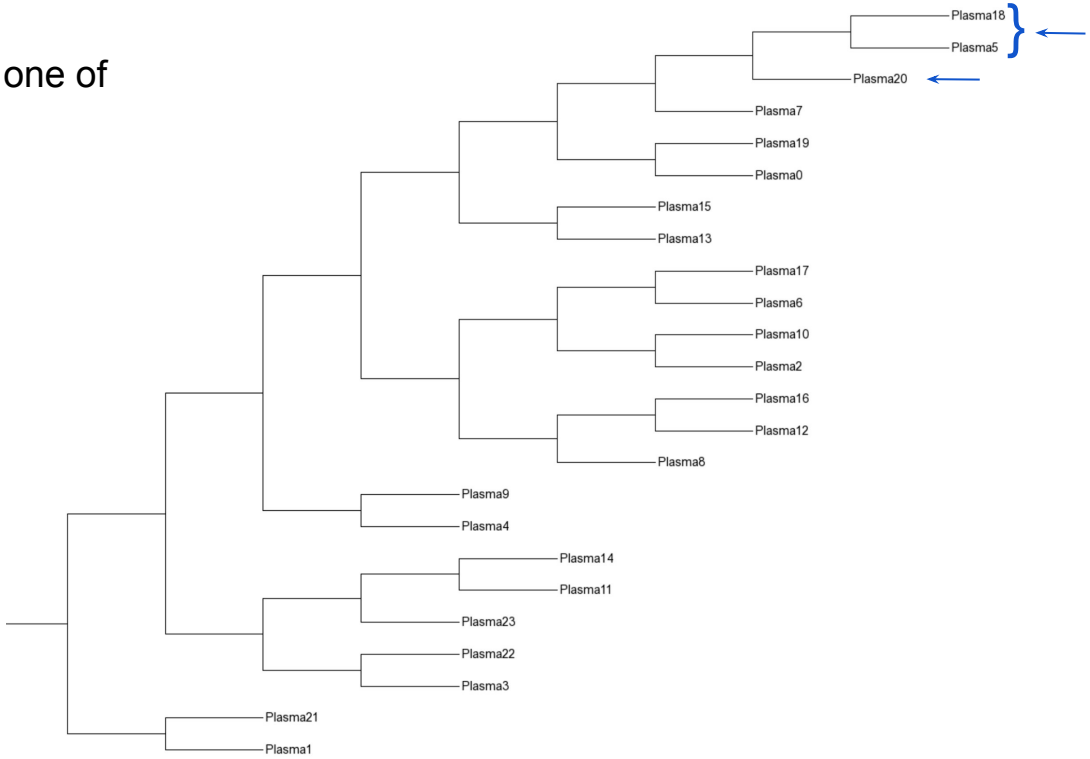
Local alignment (k trees for k peptides)

Hierarchical tree for one of the peptides



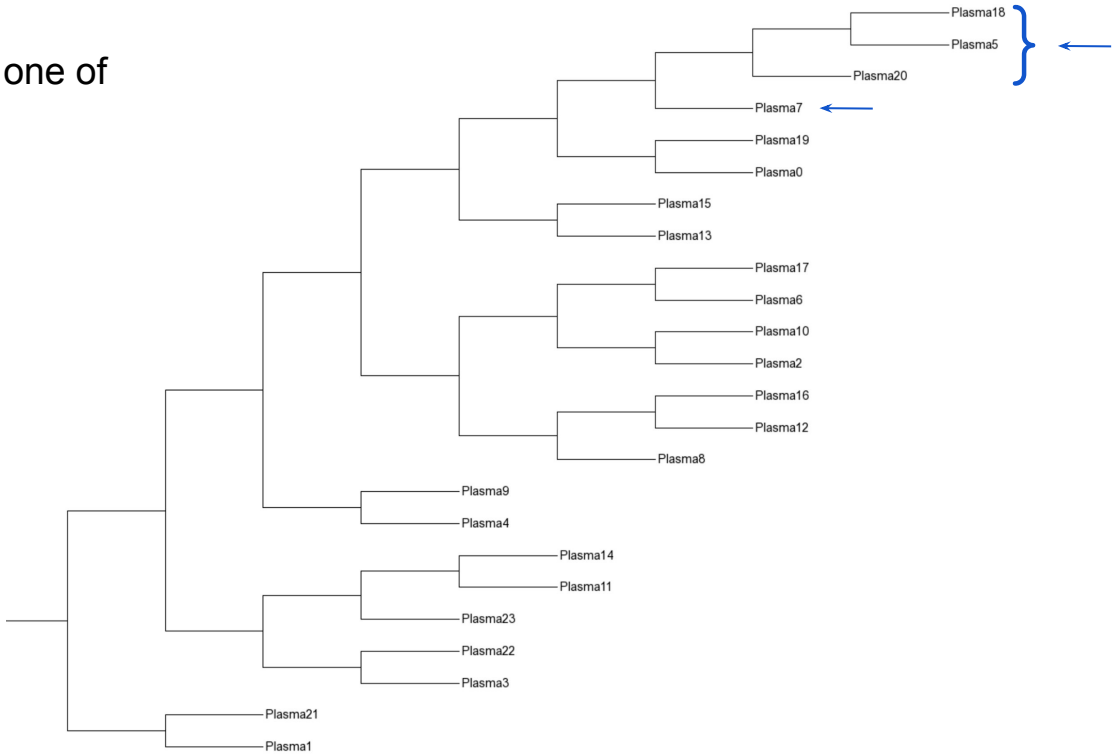
Local alignment (k trees for k peptides)

Hierarchical tree for one of the peptides



Local alignment (k trees for k peptides)

Hierarchical tree for one of the peptides



Local alignment (k trees for k peptides)

2. Find an optimal approach to calculating intensity during the merge

- Mean of intensities
- Median of intensities
- Weighted mean of intensities

3. Find the best way to deal with missing values in the alignment

- Relabel timepoints of the alignment considering all NAs, but recalculate the intensities of the shifts in a different way
- Relabel timepoints of the alignment considering only terminal NAs and skipping all the intermediate NAs
- Consider all terminal NAs and randomly skip some intermediate NAs

  Resample the vector of intensities back to 195 timepoints after each step of alignment and merge.

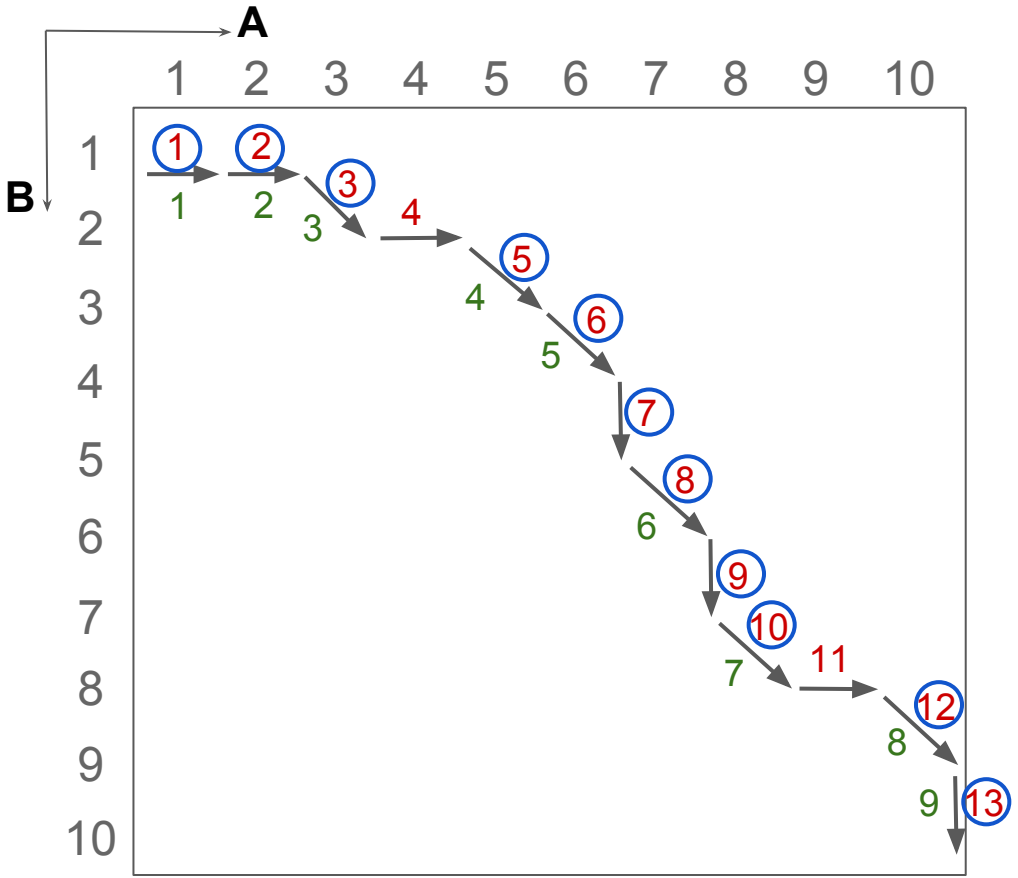
Models of intensity recalculation

Model 1: consider all NAs, calculate intensities as follows:

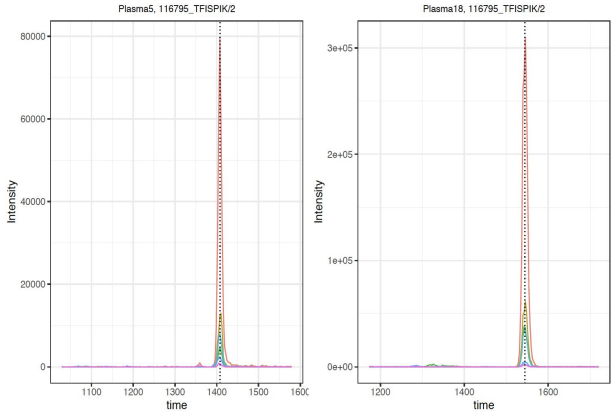
$Int_1 = (Int_{A1} + Int_{B1})/2$
$Int_2 = (Int_{A2} + Int_{B1})/2$
$Int_3 = (Int_{A3} + Int_{B2})/2$
$Int_4 = ((Int_{B2} + Int_{B3})/2 + Int_{A4})/2$
$Int_5 = (Int_{A5} + Int_{B3})/2$

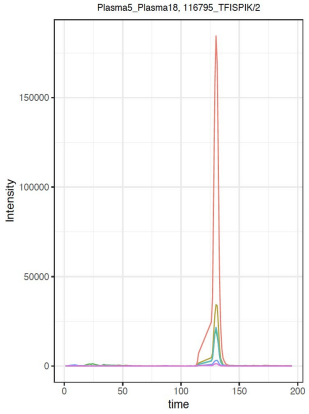
Model 2: consider only terminal NAs

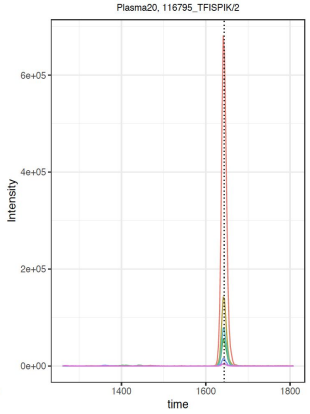
Model 3: randomly skip some NAs

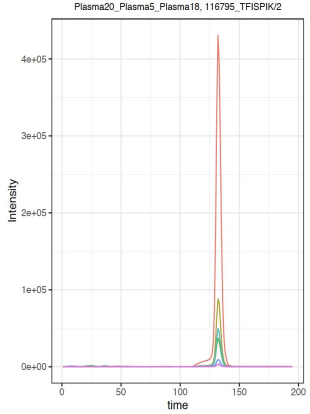


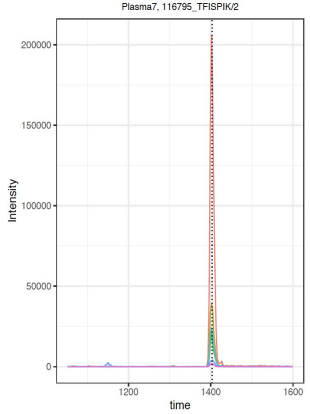
Merging steps 1-3: runs 7-20-5-18 (model 1)

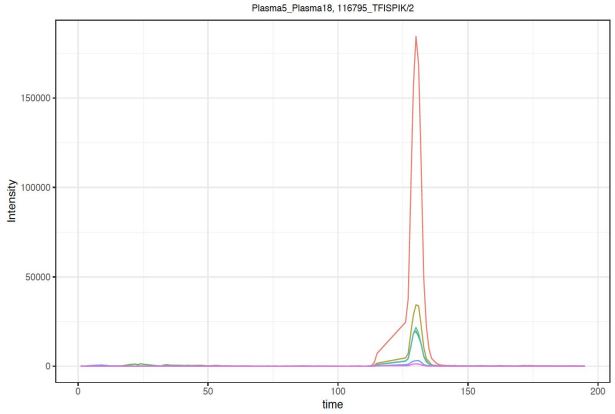


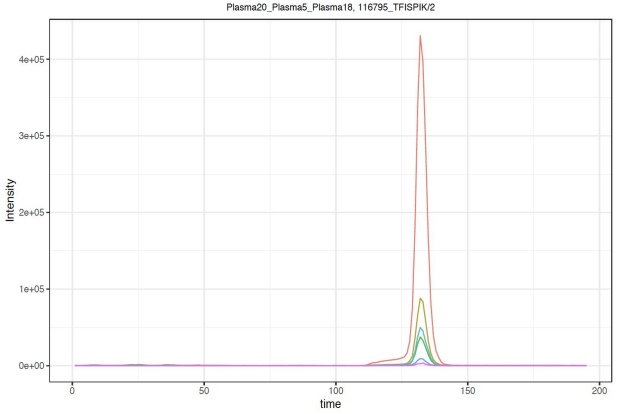


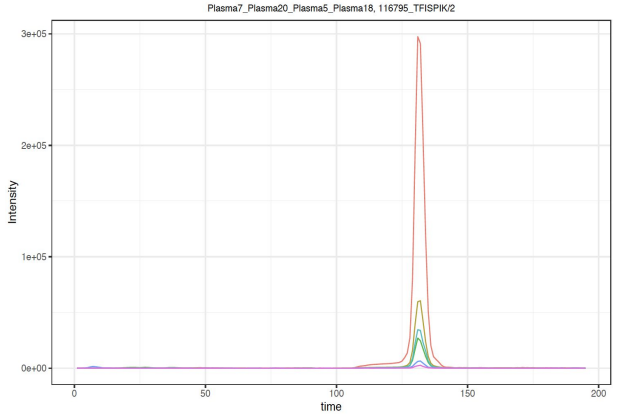




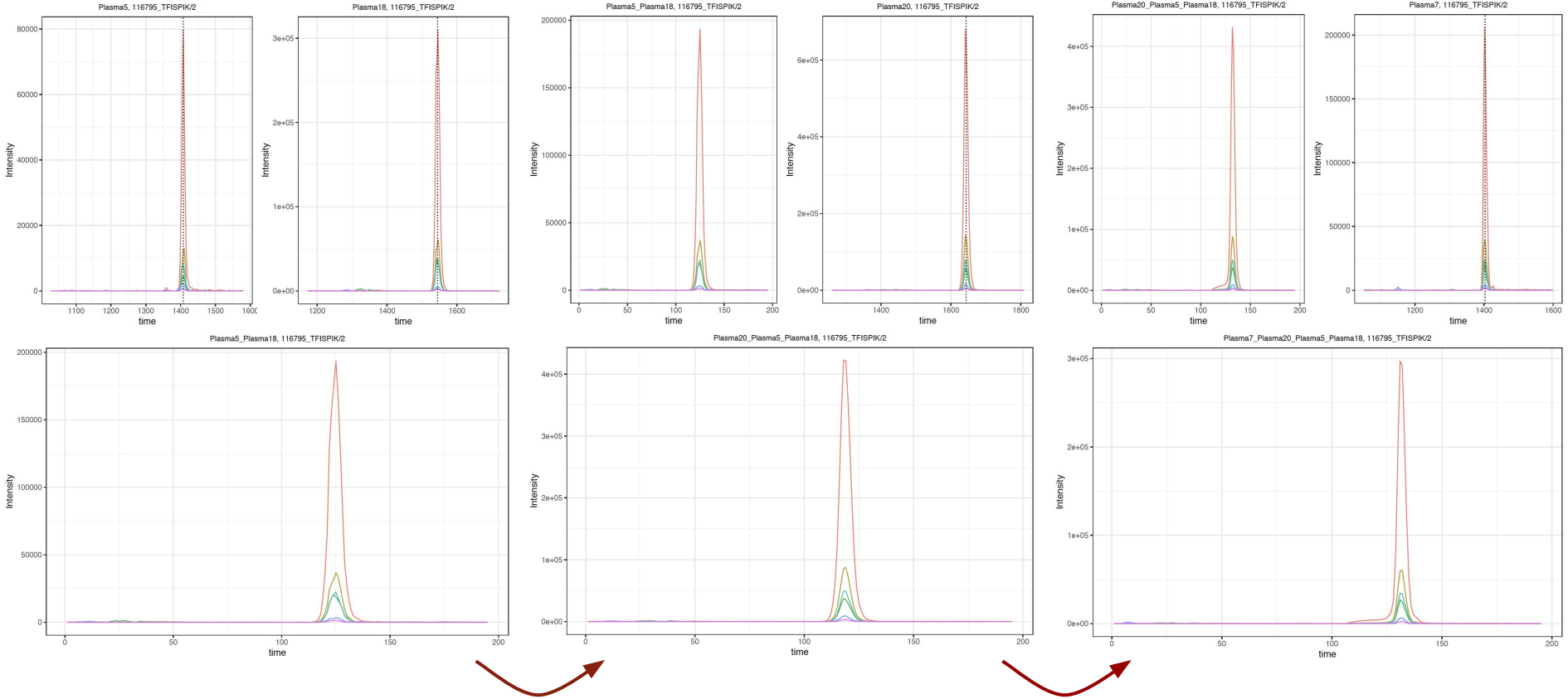




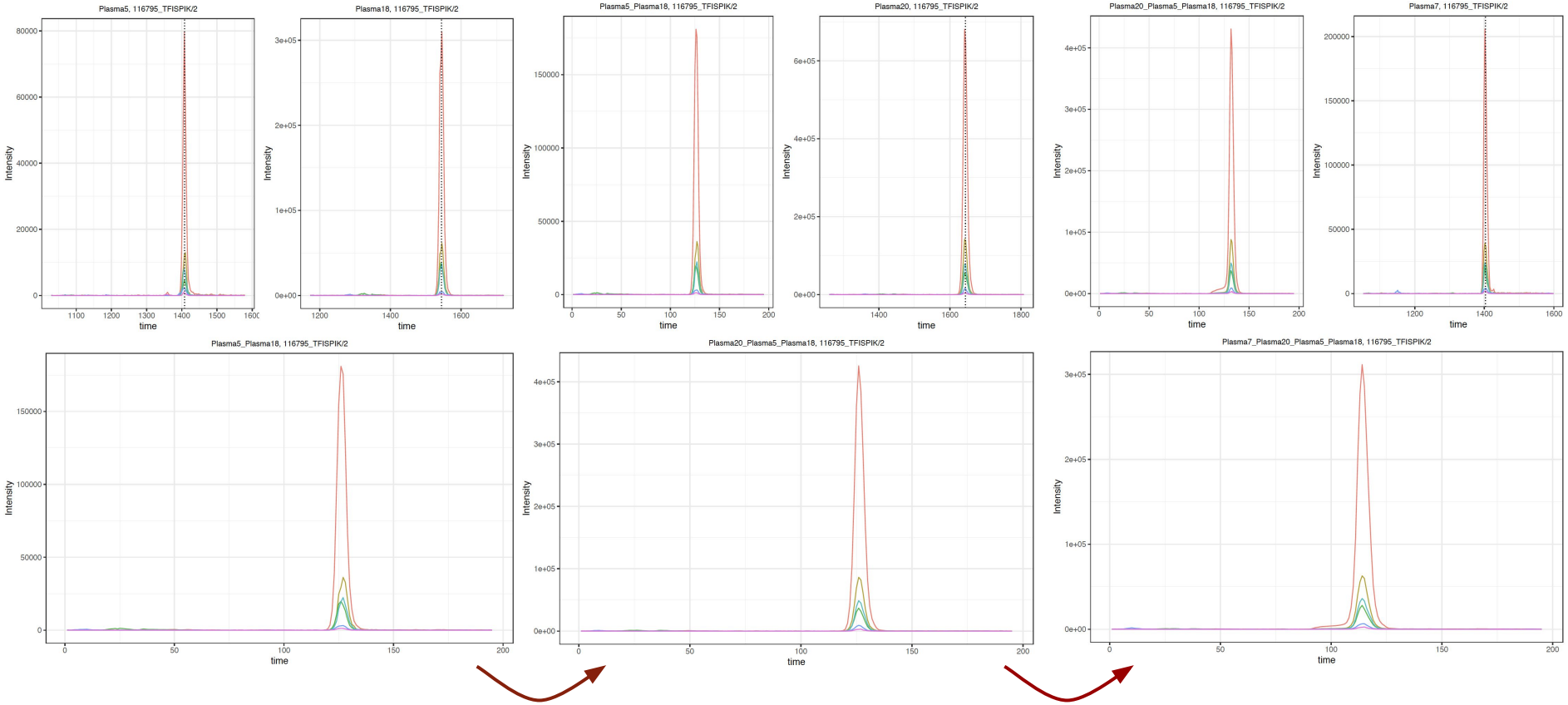




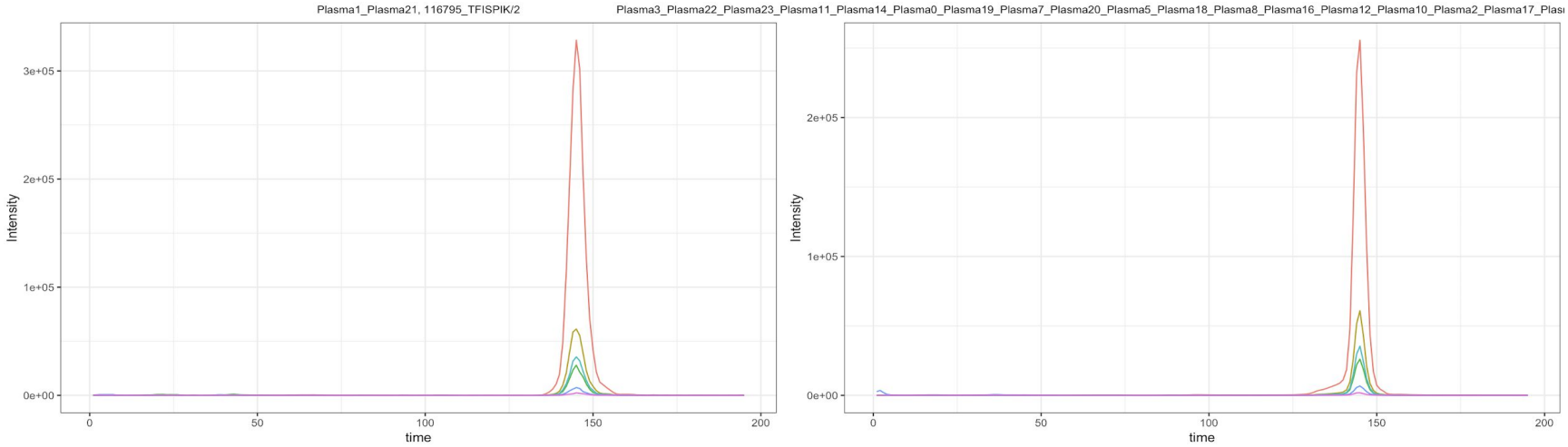
Merging steps 1-3: runs 7-20-5-18 (model 2)

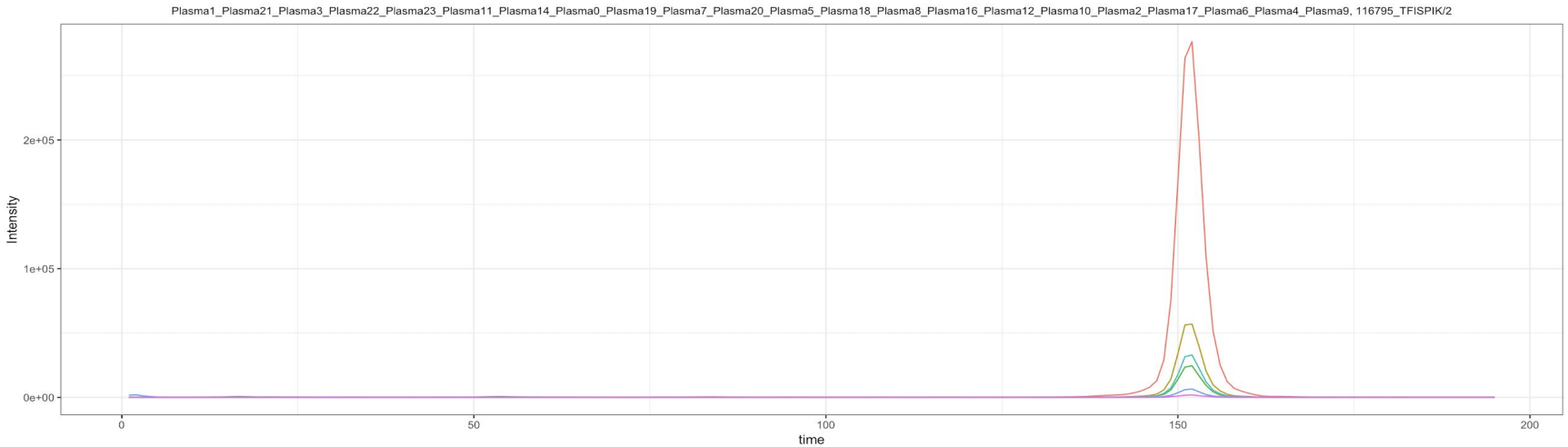


Merging steps 1-3: runs 7-20-5-18 (model 2)

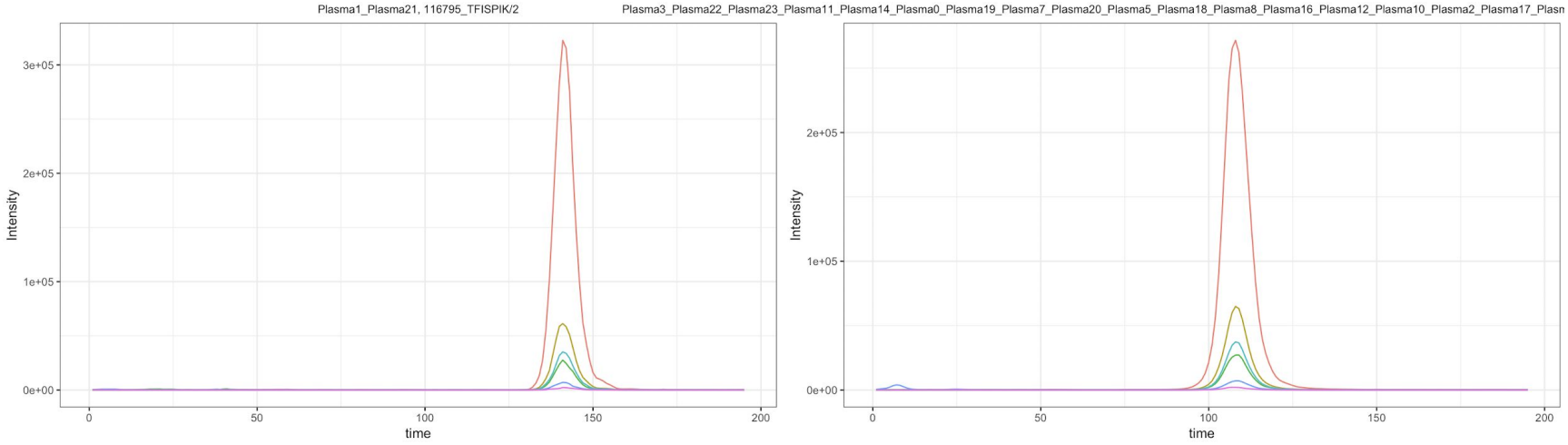


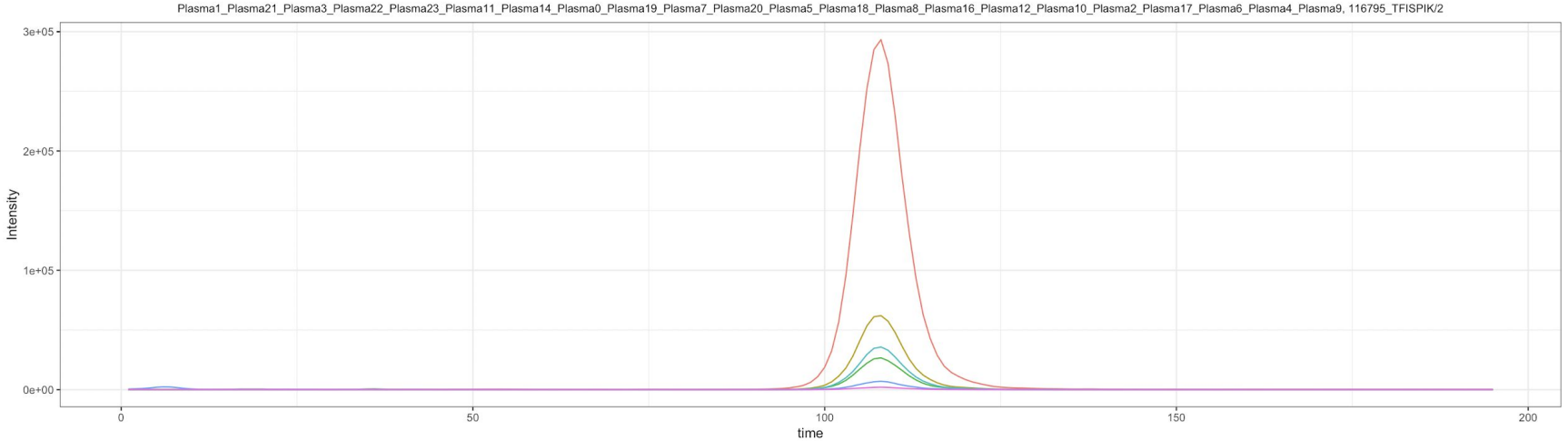
The very last merging step (model 1)



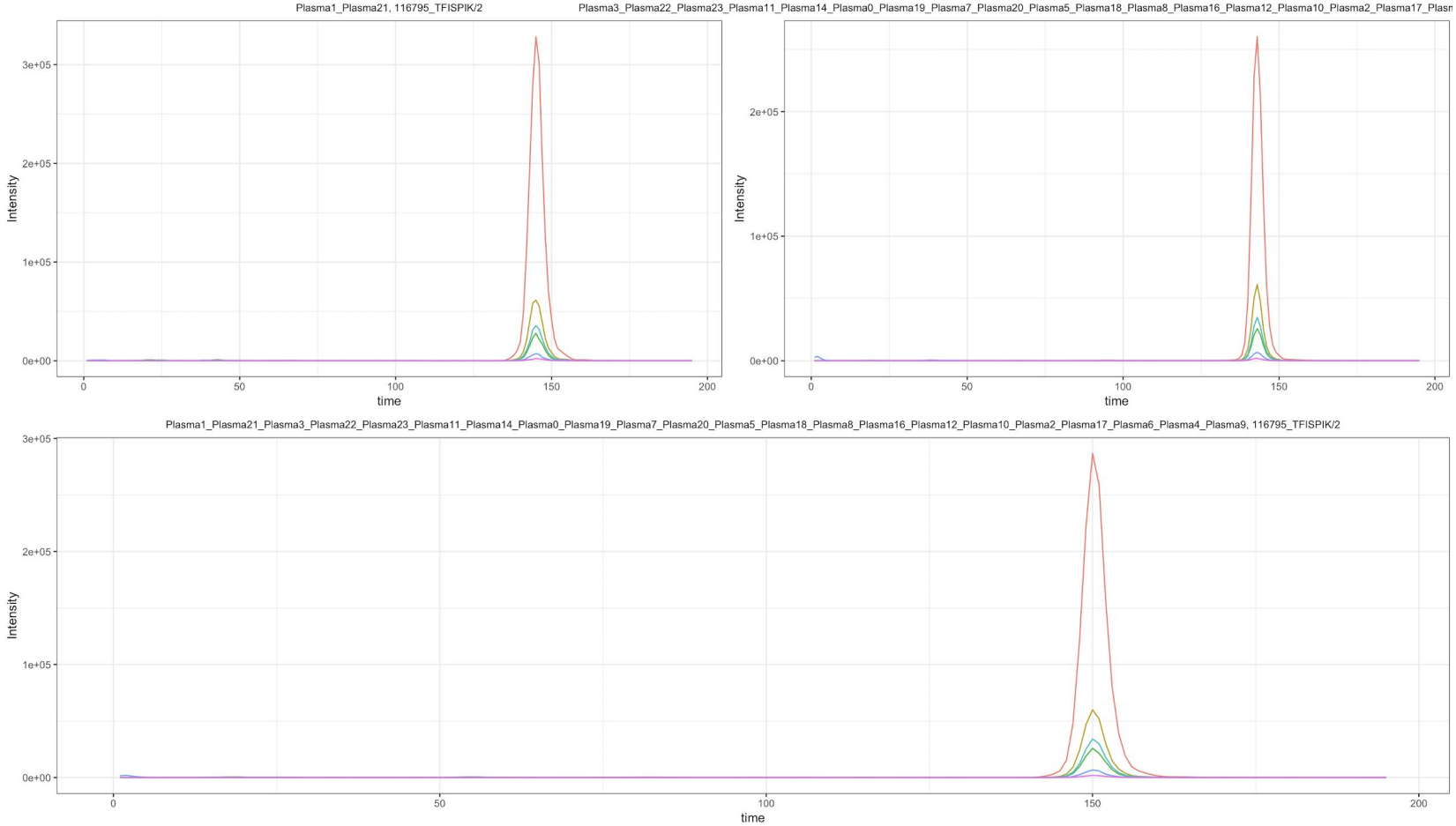


The very last merging step (model 2)





The very last merging step (model 3)



Future work

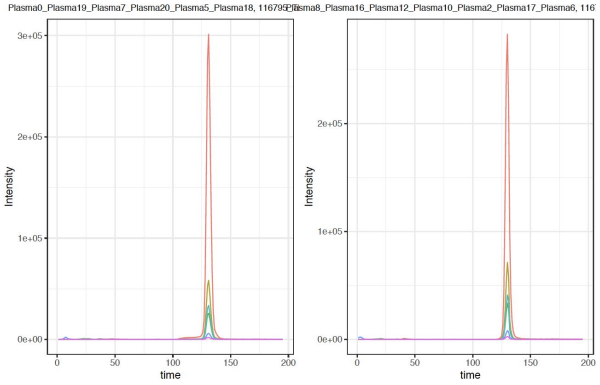
- Improve calculation of the retention time and intensity during the merge
- Try reference-based approach supported by CV error of regression (LOESS)
- Optimize the model of handling NAs and resampling procedure
- Finally start fill in the lab notebook

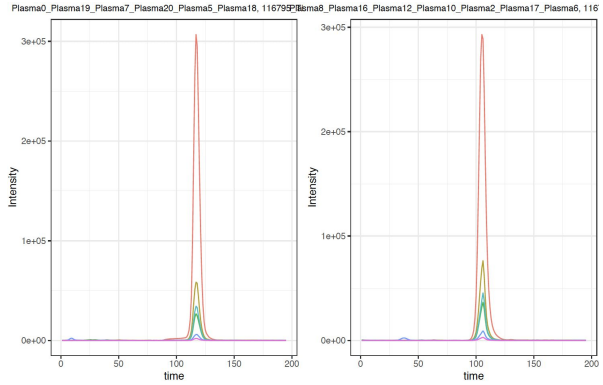
Acknowledgments

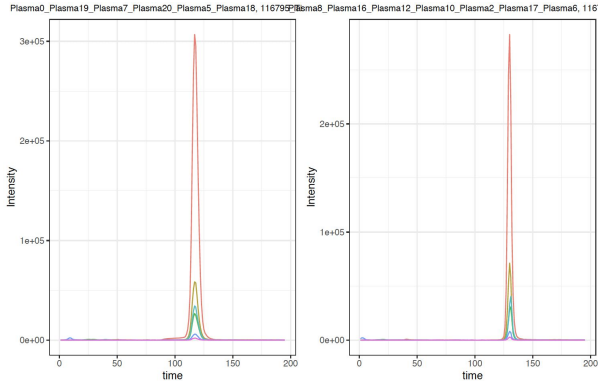
Hannes
Shubham
Annie
Leon
Ron
Charlotte
Gordon
Oliver

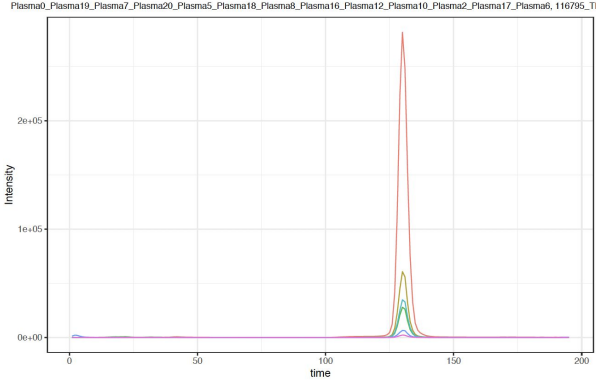
Thank you! :)

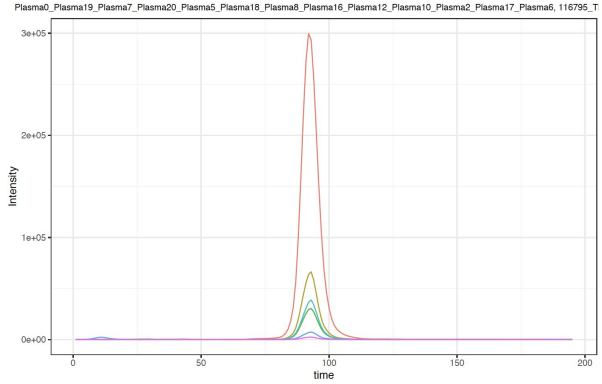
Merging step #11

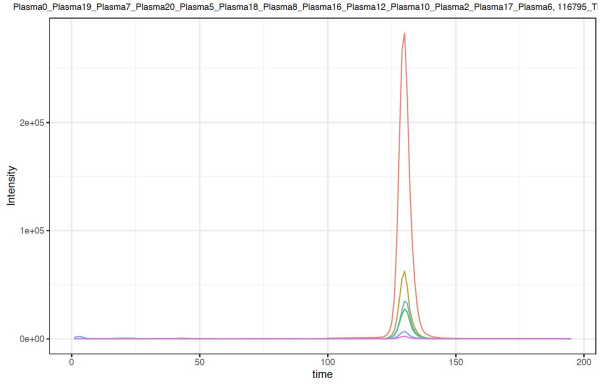












Model 1 (keep all NAs)

Model 2 (skip intermediate NAs)

Model 3 (randomly skip intermediate NAs)