Vasaikar

Love tätting

2023-10-09

Introduction

This report is built on the following paper.

Title: A comprehensive platform for analyzing longitudinal multi-omics data

Journal: Nature Communications
Publication Date: March 27, 2023
Volume: 14, **Issue:** 1, **Page:** 1684
DOI: 10.1038/s41467-023-37432-w

Author: Suhas V Vasaikar[1]
Data Source: Supplementary Table 1

Data Structure

Complete Blood Count

Complete blood count was measured in all six individuals at week 1-10 (6 individuals x 10 time points).

Flow Cytometry

Flow Cytometry was performed in patient 2,4,5 and 6 at weeks 2-7 (4 individuals x 6 time points).

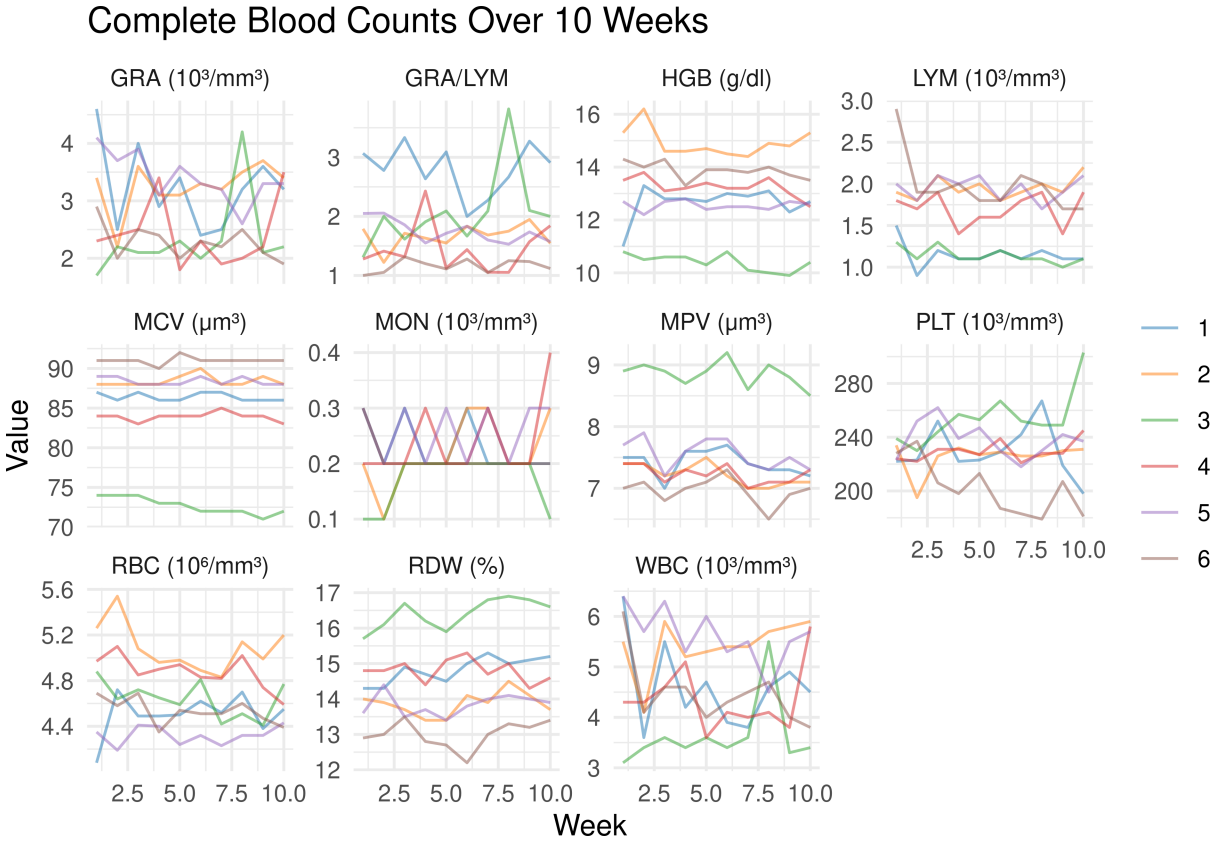
Olink

Olink was performed on plasma from all six individuals at all weeks (6 individuals x 10 time points)

Patient Metadata

donor	sex	age_group
PTID1	female	36-40
PTID2	male	36-40
PTID3	female	21-25
PTID4	male	31-35
PTID5	female	26-30
PTID6	male	36-40

Complete Blood Count



From overviewing the CBC data we can see that

- Granulocytes and WBC varies over time and between individuals. This is expected.
- One patients is anemic (#3) and the other are not. However, hemoglobin values tend to be higher in the age groups among healthy individuals. One patient is borderline to hyperglobulinemia (#2), but is probably normal as the patient is a young male. Patient #3 likely has iron deficiency anemia or thalassemia.
- MPV is distinctly raised in patient #3. This parameter is not commonly reported in Europe, but an increased MPV is commonly due to an increased production of platelets. This correlates with a state of iron deficiency, which in many cases results in overt thrombocytosis if left untreated.
- RDW is conspicuously increased in patient #3. This indicates anisocytosis (different sizes of red blood cells). It can be seen in a range of conditions. In this case, it is likely to reflect ineffective erythropoiesis of iron deficiency.
- Lymphocytes are stable over time but with some individual variation. This is expected. None have lymphocytosis, which would be an uncommon feature and invariably pathological. Two aptients have lower lymnphocyte counts, but not lymphopenia. Lymphocytes tend to decrease on inflammation.
- PATID 1 and 3 might have some underlying inflammation as indicated by granulocytes, lymphocytes and their ratio (higher GRA/LYM ratio indicates inflammation). However, the inflammation is not great and could eg. reflect obesity.
- Precision in the instrument to measure monocytes is rounded to nearest 100. This makes monocytes more of a discrete variable in this dataset.
- There are no missing values in the CBC data set.

In an overview of the CBC data, we can easily see that patient 3 is an outlier. The patient is a young female 21-25 years old, and likely to be menstruating. She is anemic, probably symptomatic on exercise but not in rest if optherwise healthy. Since she has a downward slope in hemoglobin (HGB) and mean corpuscular volume (MCV), together with a late increase in platelets (PLT), it is most likely that she has iron deficiency

anemia, rather than just thalassemia.

Patient 3 will not be excluded based on this, but her iron deficiency will have effects on other parameters that might not be readily known and hard to account for.

Among the other patients, no certain conclusions may be drawn regarding their health.

Data Cleaning

There are no obvious errors in data collection (instrument fault) that is not expected form the variance in different health states and what is normal for longitudinal blood samples. Certainly, as the number of subjects is significantly fewer than the time points and measured variables (blood types), it will be hard to conclude on outliers vs normal variation.

Outliers

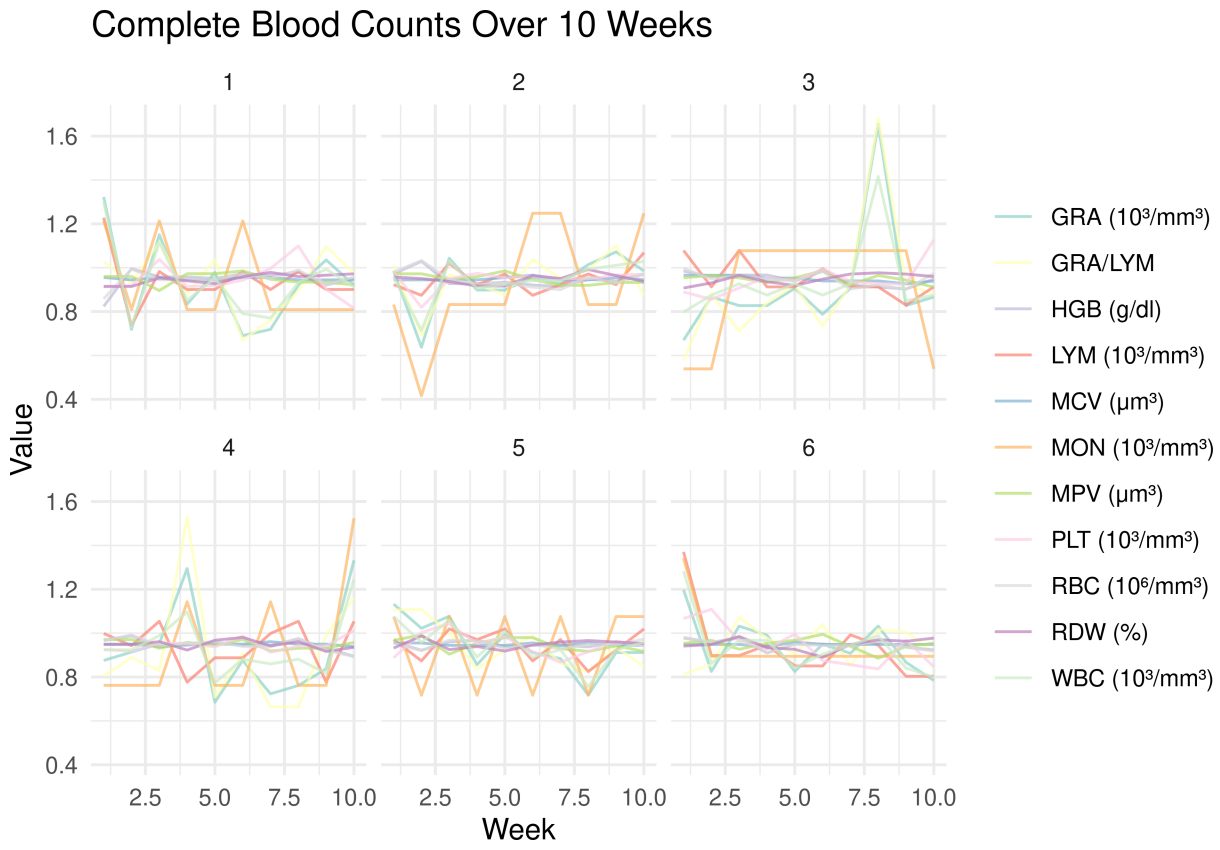


Figure 1: Values for each CBC parameter and each patient was scaled but not centeredto account for the different value ranges and individual differences.It can be seen that over time, CBC parameters are generally stable within ahealthy individual. One patient ha sanemia, which the scaling masks, but it is a slowcondition as is not expected to have a noticeable impact on individually scaled data.

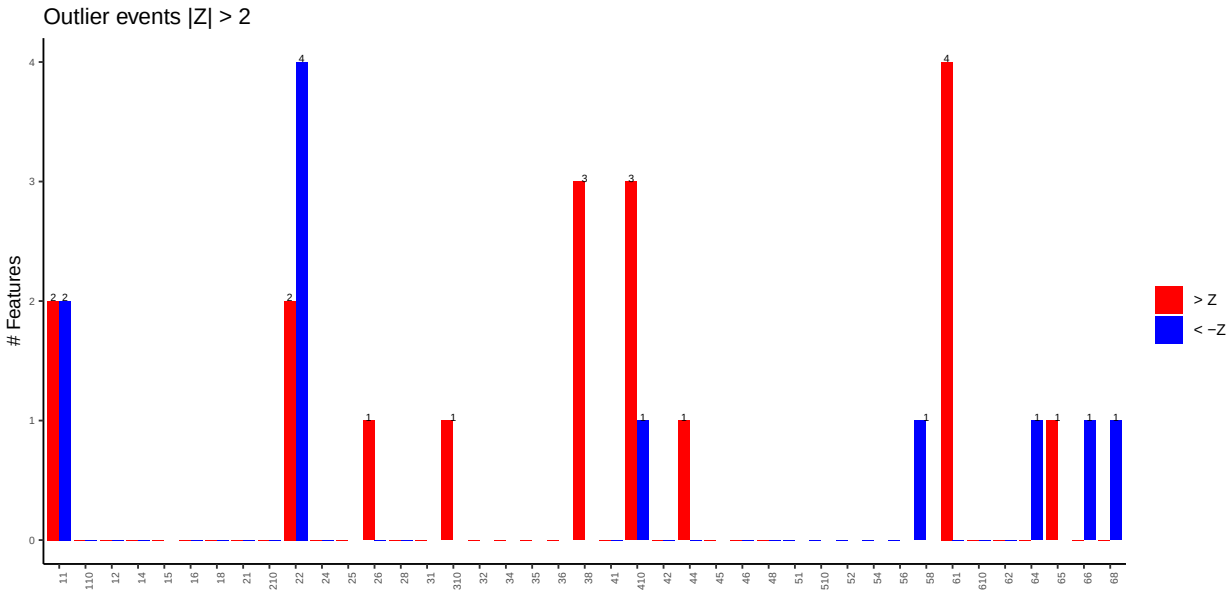


Figure 2: Outliers by Z-method. There are some measurements that areindicative of being outliers.

Partitioning of Variance

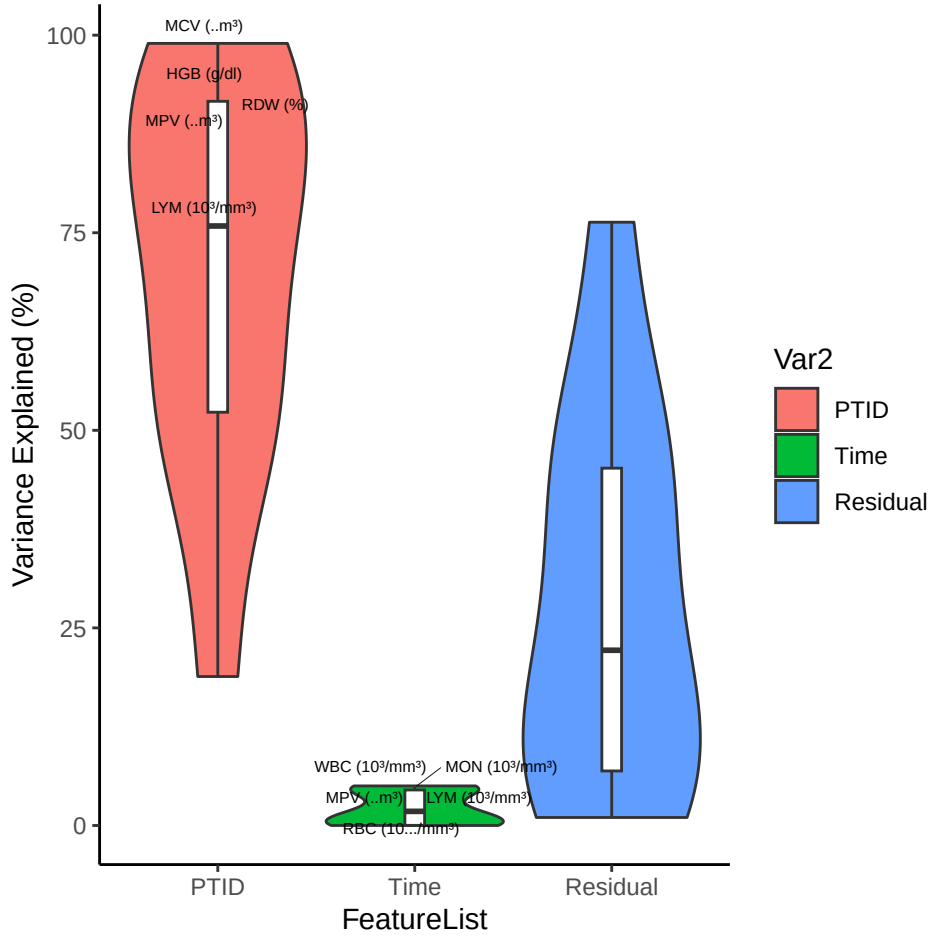


Figure 3: Variance explained by individual, time and residual variance

We can see from variance decomposition analysis that MCV, HGB, MPV, RDW, and LYM. Except for LYM, these parameters are expected to be variable in a population, but lymphocytes are generally stable across individuals and requires more perturbation of the health state to vary. In this case, two patients had lower lymphocyte counts than the others, adding excessive variation sompared to what one would expect. Additionally, patient #3 with likely iron deficiency anemia add variance to the red blood cell related parameters. Among time related parameters, even though small, WBC is typical to display more time related variance than interindividual. WBC in healthy individuals can easily change 50% from day to day and be normal, but the vairation among healthy individuals is much smaller. This is mainly explained by granulocytes, whereas monocytes and lymphocytes are generally much more stable over time.

The result is weakened but similar when removing patient #3 (not shown).

Flow Cytometry Data

For a subset of patients, a more thorough analysis of white blood cells was performed with multiparameter flow cytometry.

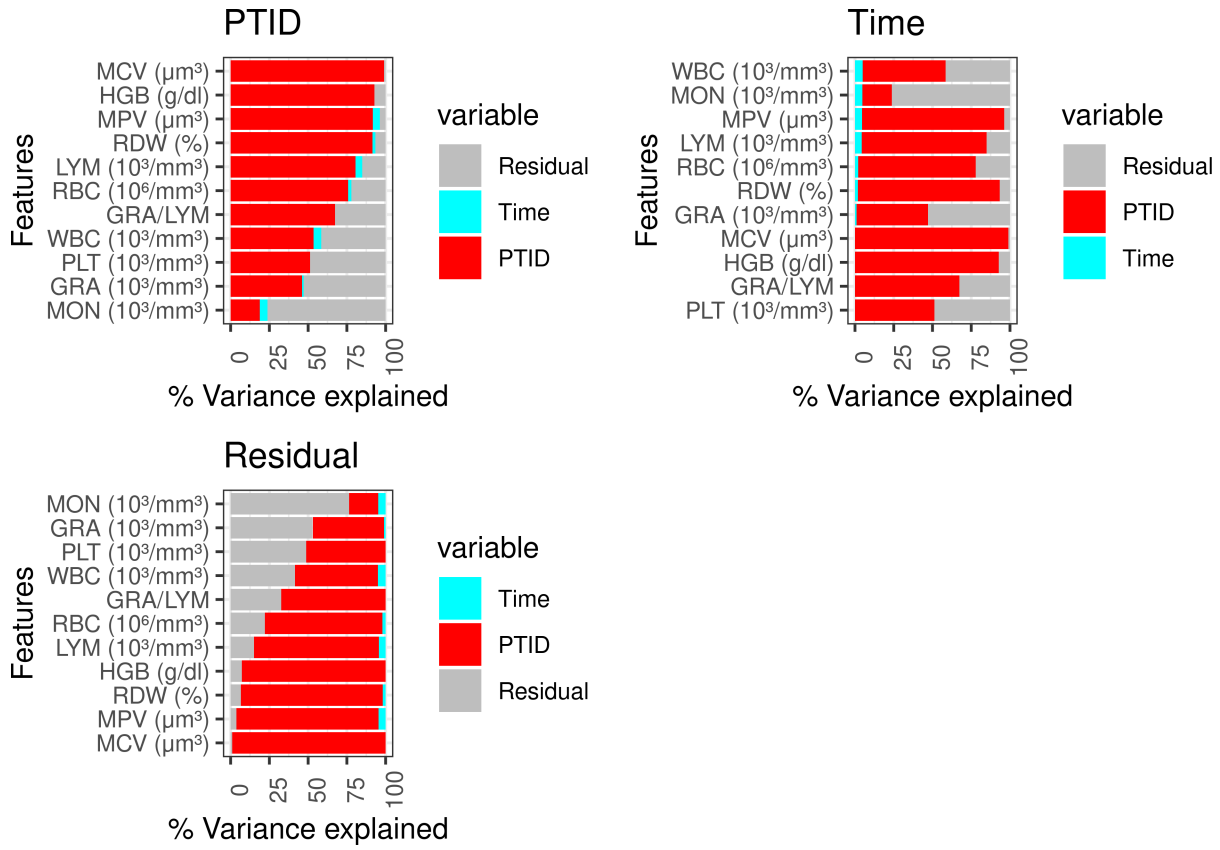


Figure 4: Variance contribution for each feature by patient, time and residuals

Data Overview

The data file is stated to be in percentages. Commonly, cell populations are either of defined parent populations or to all leukocytes. Being a percentage of all cells makes calculations easier (otherwise the sum of percentages is not 100%). It is assumed that the percentage is in relation to all living cells.

There is no missing data in the dataset.

Table 2: Leakage of cells outside any target gate can be seen to be small and almost all identified leukocytes have been accounted for. (continued below)

PTID5W2	PTID5W3	PTID5W4	PTID5W5	PTID5W6	PTID5W7	PTID6W2	PTID6W3
96.81	97.25	98.14	97.74	97.42	98.09	97.99	97.77

Table 3: Table continues below

PTID6W4	PTID6W5	PTID6W6	PTID6W7	PTID2W2	PTID2W3	PTID2W4	PTID2W5
97.72	97.73	97.99	98.21	98.07	98.09	97.89	98.16

PTID2W6	PTID2W7	PTID4W2	PTID4W3	PTID4W4	PTID4W5	PTID4W6	PTID4W7
97.68	97.27	97.75	97.48	97.66	97.23	97.51	97.93

PCA Analysis

PCA analysis is concurrent with plotting of the coefficient of variation. Individual 2 and 4 contributes most to variation, but in different cell types.

OLink

The Olink data is assumed to be already normalized. There is no possibility to normalise the data as indicators of plates and experimental runs are absent in the raw data.

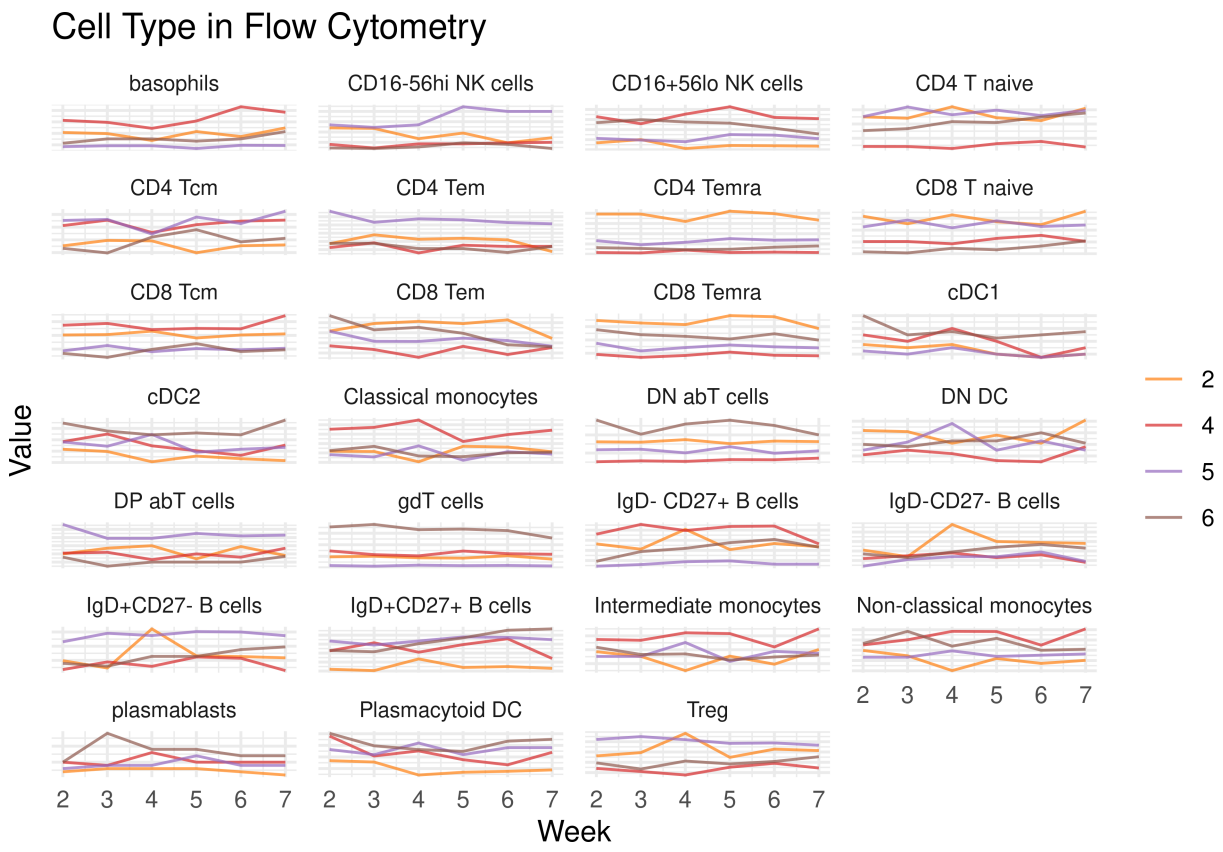


Figure 5: Longitudinal alalysis on each cell type from flow cytometry.

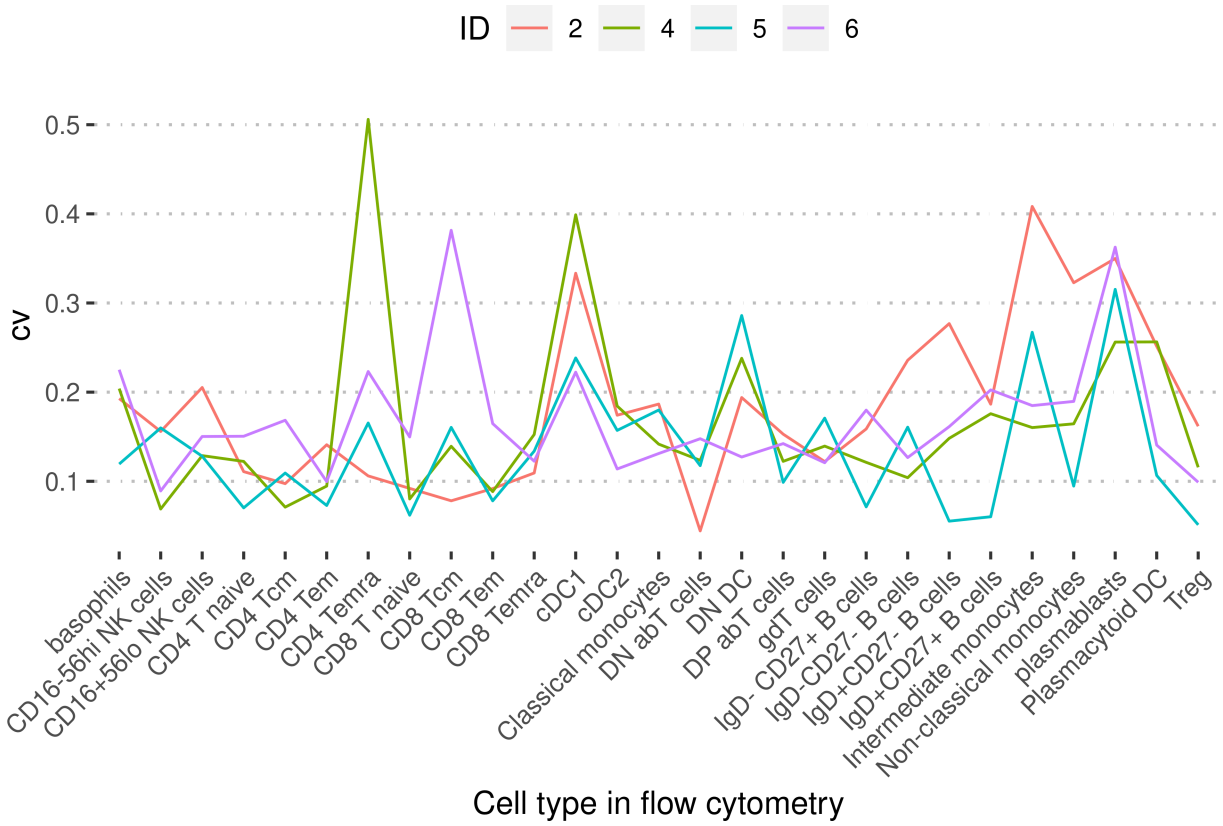


Figure 6: Coefficient of variation for each cell type and ID over all measured weeks.It can be seen that 2 and 4 contributes most variation, but in different cell types.

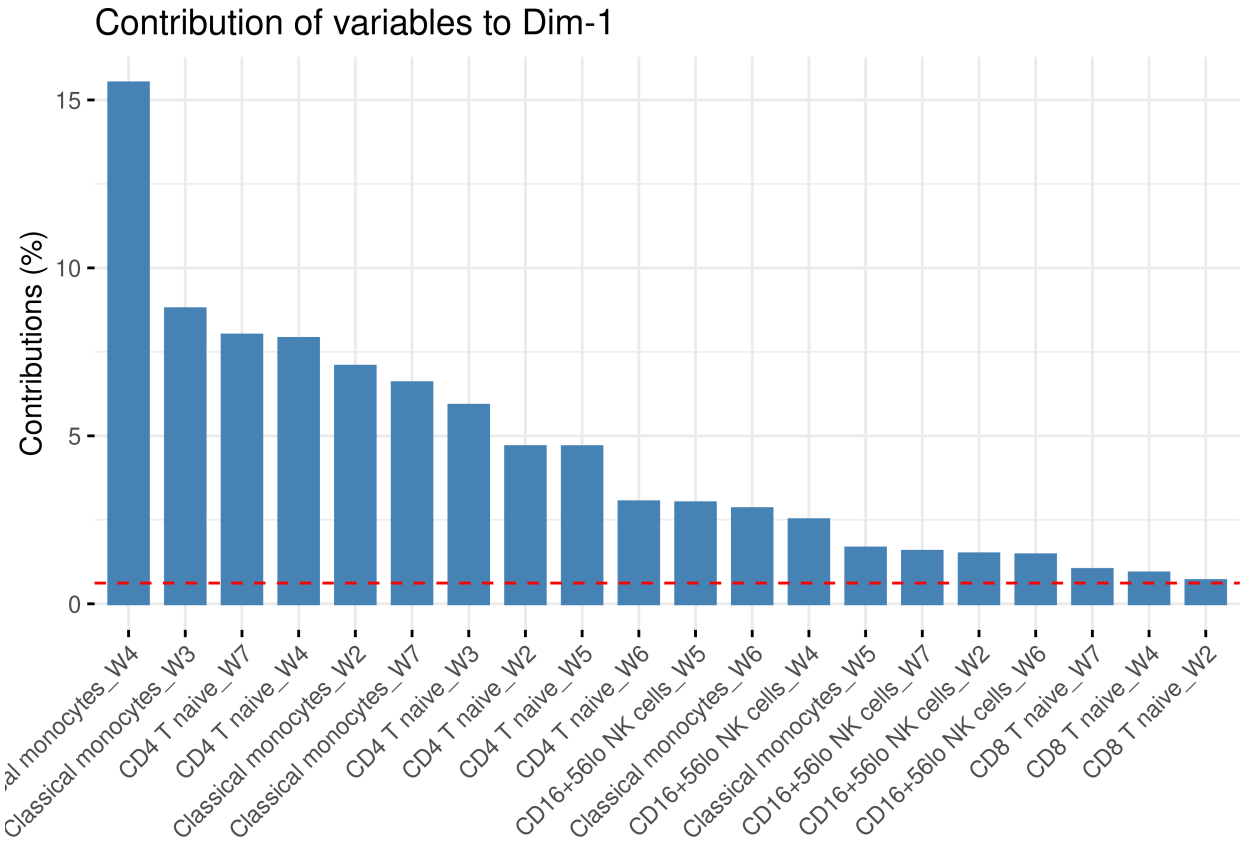


Figure 7: Variable contribution to principal component 1.Innate immune cells, mainly monocytes, and unprimed adaptive immune cells,particularly naive CD4 T-cells and NK-cells, can be seen to contribute most to variation

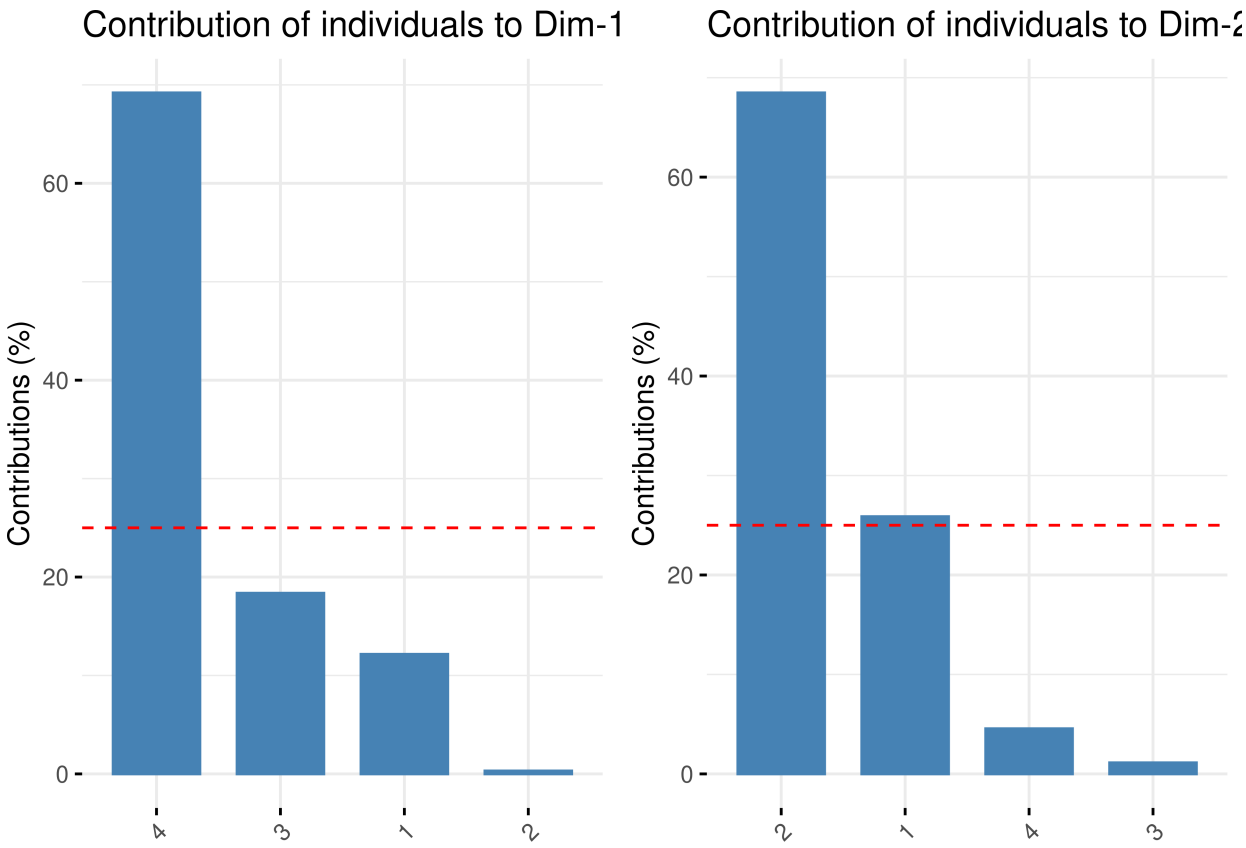


Figure 8: Individual contribution to component 1 and 2. It can be seen that individual 4 is the principal cause of overall variation, and that individual 2 have a principally orthogonal variation to individual 4.

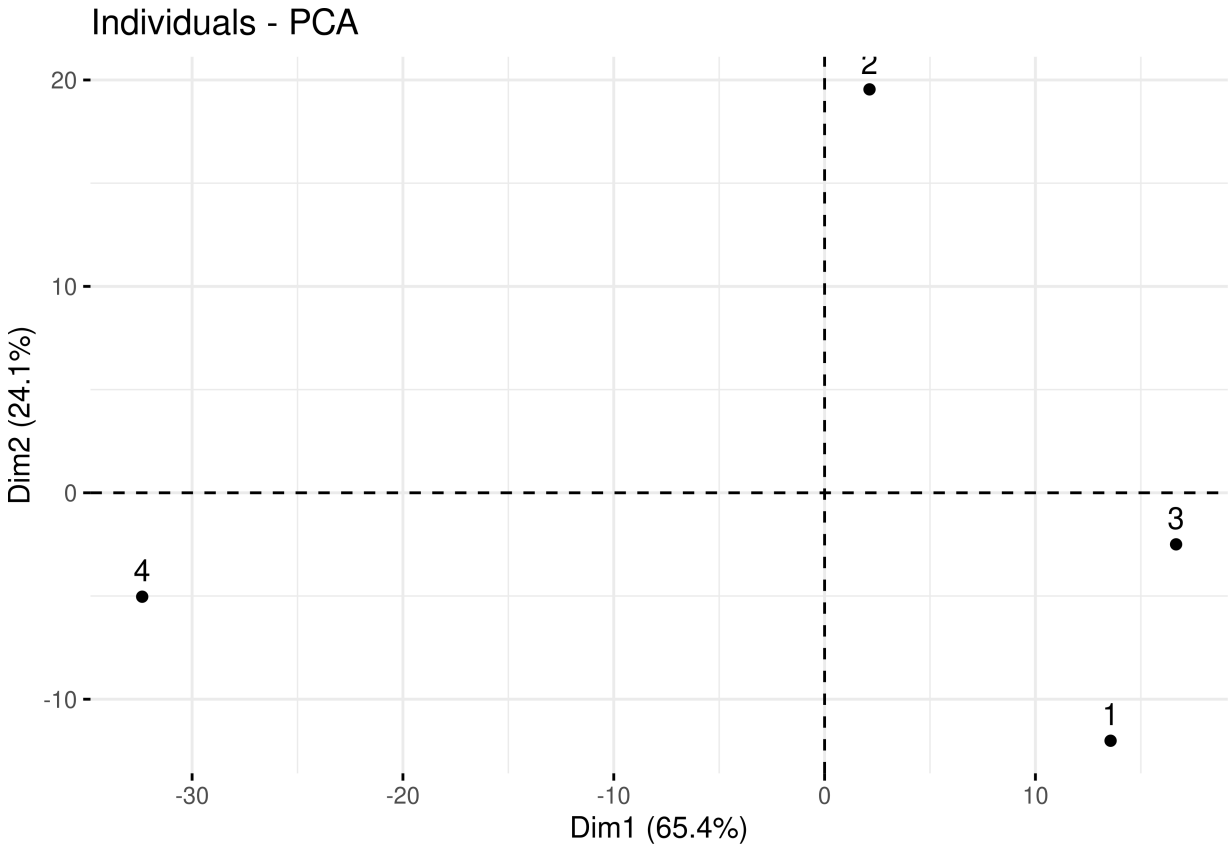


Figure 9: Plot of Individuals in component 1 and 2.Individual 4 and 2 are different from the group in different ways.Individual 1 and 3 are similar.

Data Overview

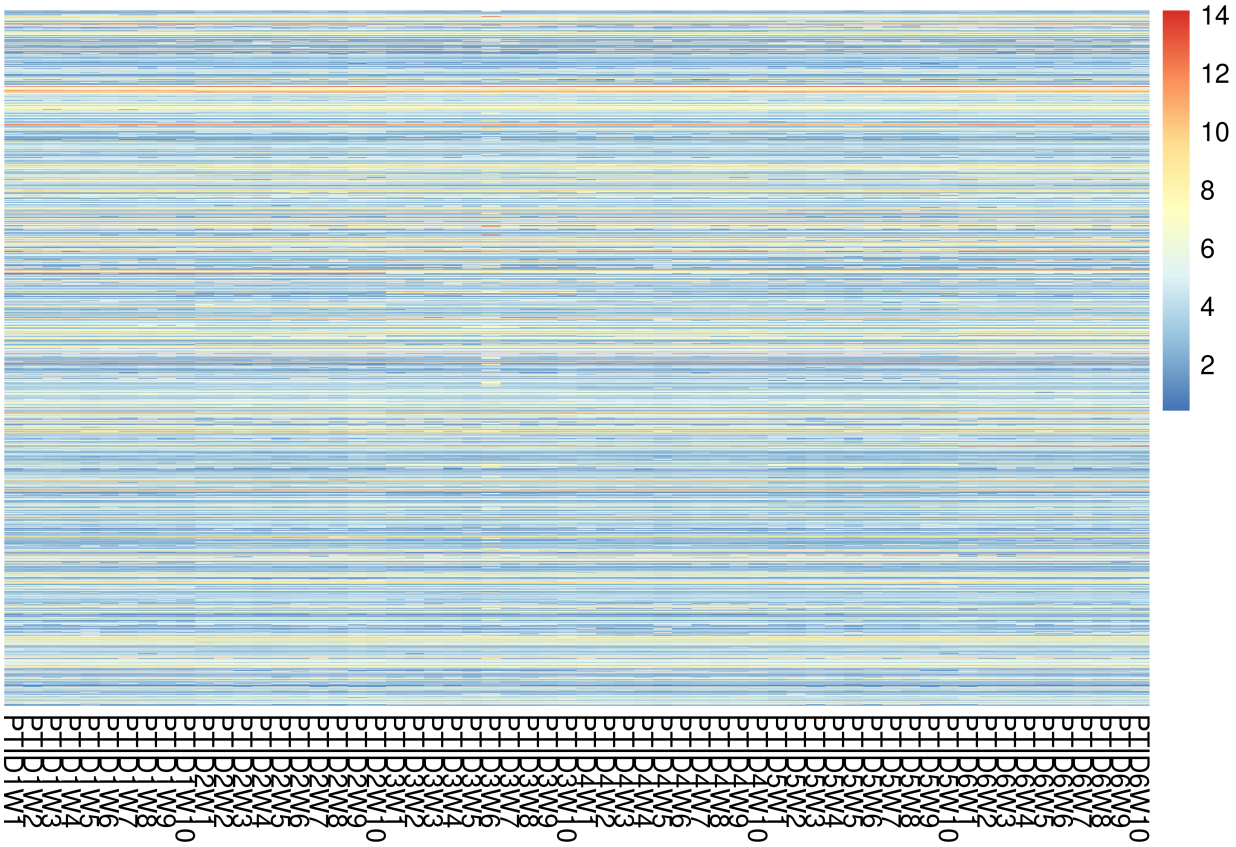


Figure 10: Heatmap of values from Olink. 1156 proteins was measured across 6 individuals atdifferent time points. No obvious outliers can be seen among the individual time points(no column formations), rather protein abundance is related to the specific protein

Outliers

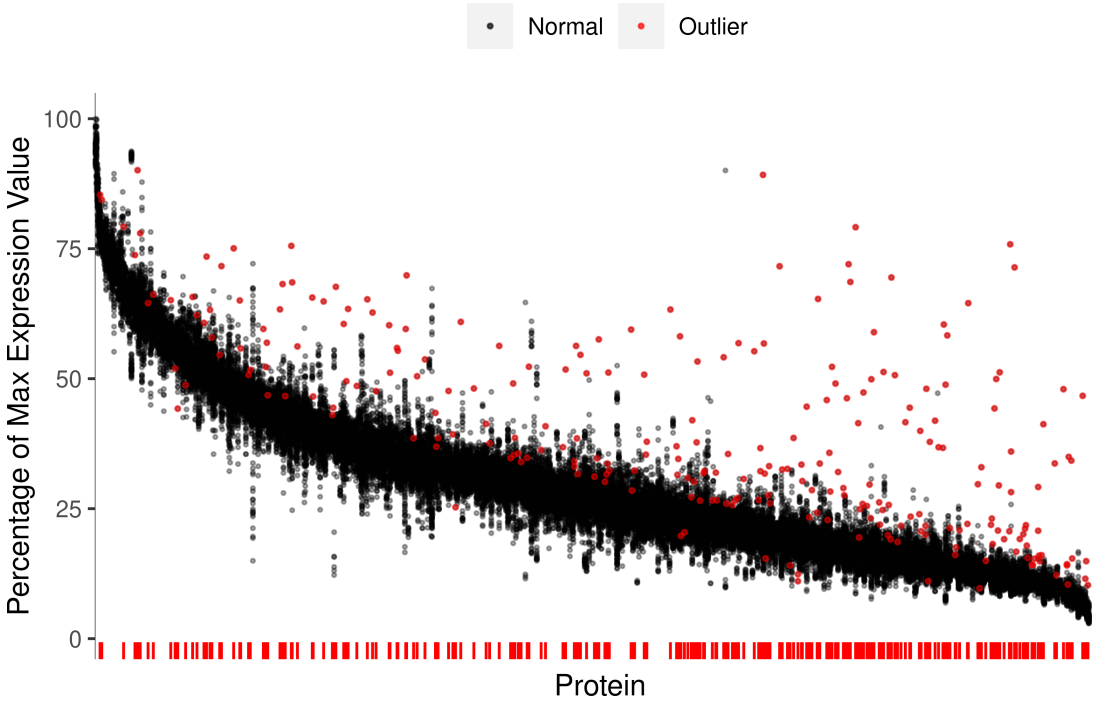


Figure 11: Olink measurements of protein abundance as percentage ofoverall max value (black dots). Red dots indicate outliers (3 z scores).It can be seen that outliers increase as the mean expression decreases.A rug plot visually aids identifying the density of missing values

It can be seen that outliers are mainly concentrated to patient 3 at day 6. The cause for this might be technical, but the overall picture of the data is similar to others and might simply reflect inflammation from a common cold or other ailment bound to happen to someone during a longitudinal study like this. The outliers will not be removed, but it is good to know that outliers have a clear relationship to this day and patient.

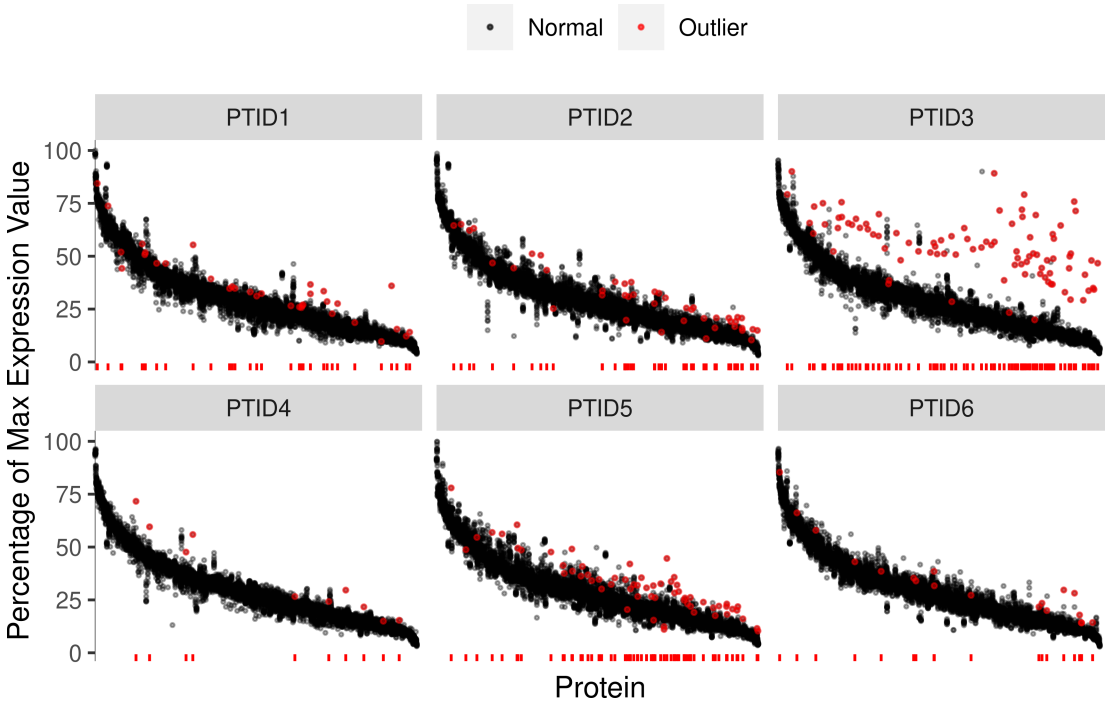


Figure 12: It can be seen that especially PTID 3 and 5 contribute outliers

Missing Values

Missingness is similar among all patients and time points in distribution, with an increase towards lower protein abundance.

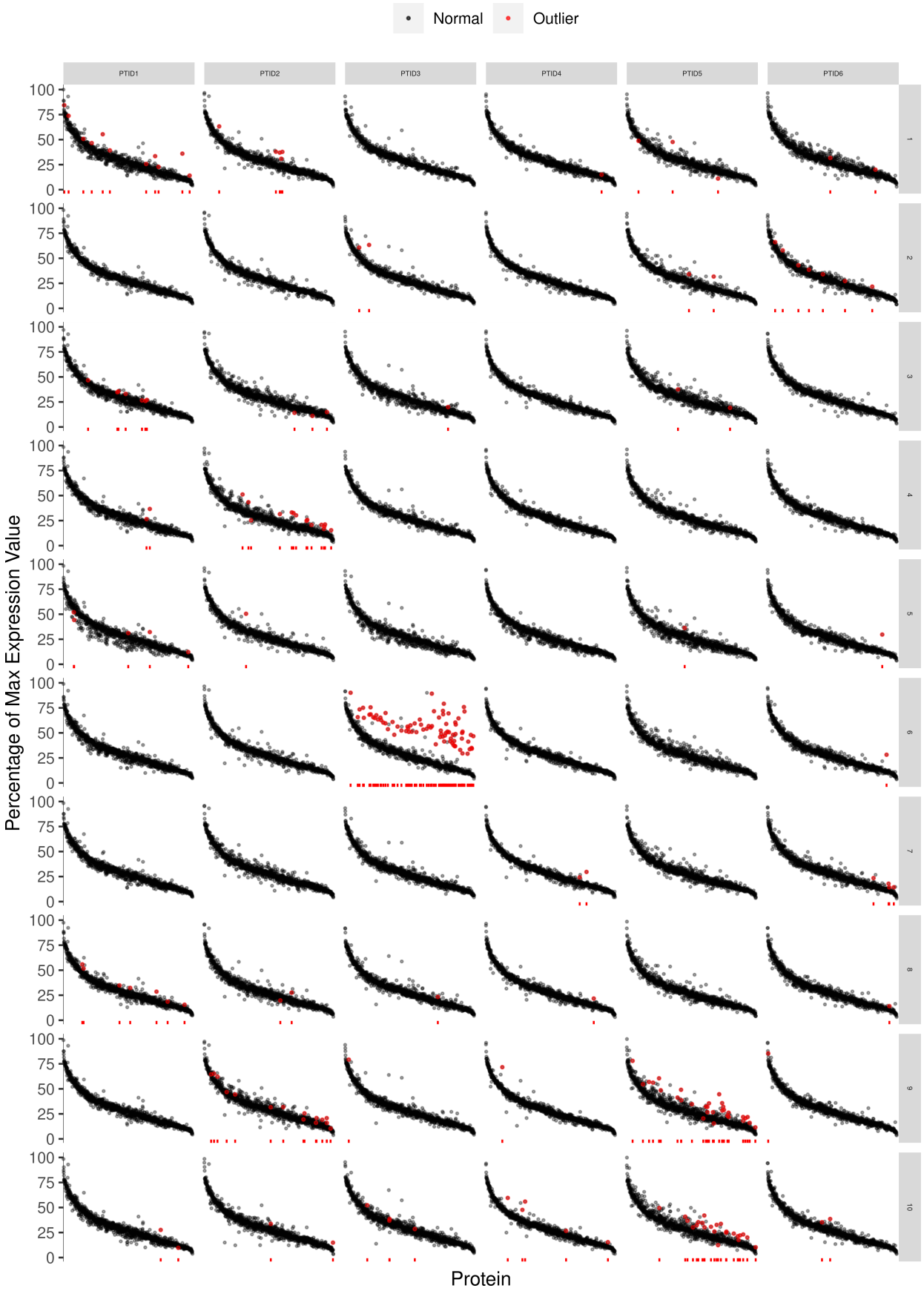


Figure 13: Looking at all the data, we can see that is mainly PTID3 at day 6that contributes outliers

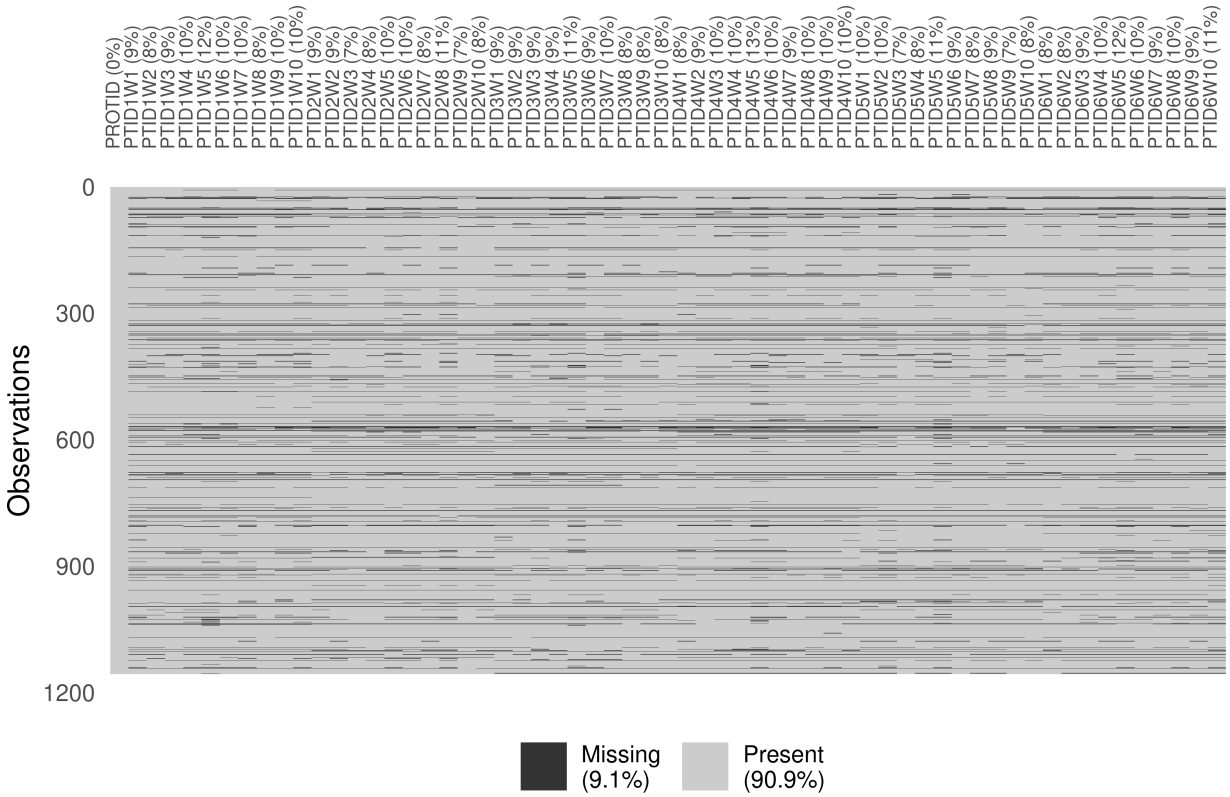


Figure 14: Missing values in Olink data. Missingness seem to be related to certain proteins,but no obvious relation to individual or time point can be seen

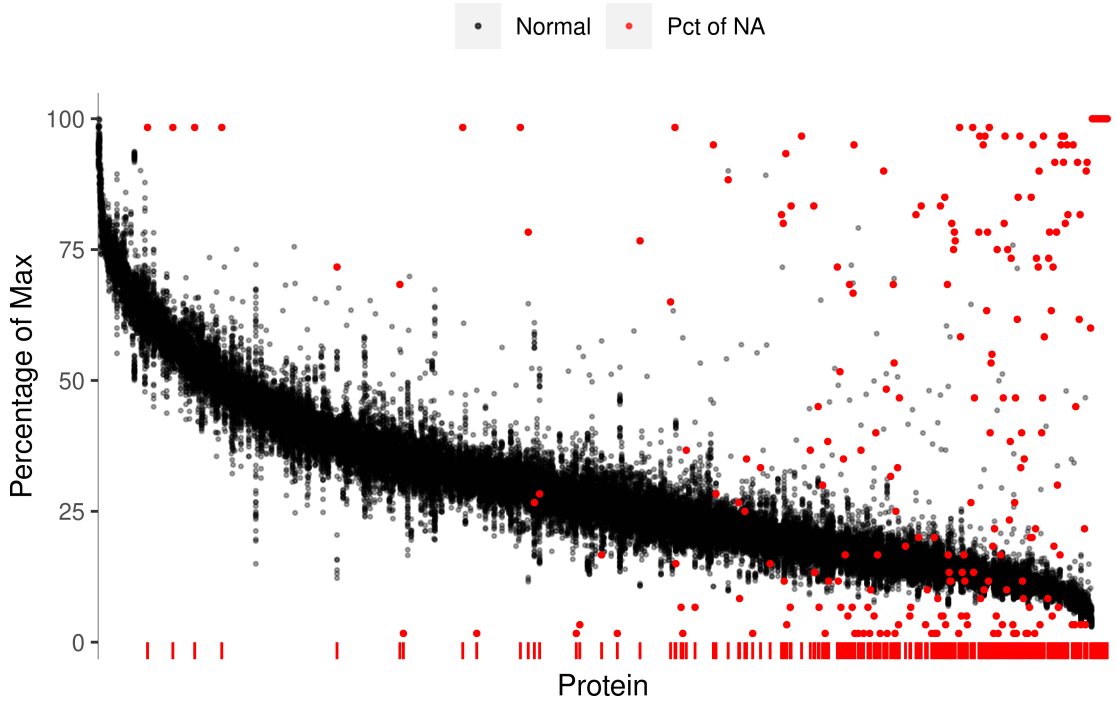


Figure 15: NA values can be seen to increase with decreasing protein abundance.The percentage of NA values among 60 measurements of a protein is shown in red.The rug plot indicates that more than 10% of measurements for a protein is NA

Figure 16: Facet plot of pct of missing values by expression level of proteins.The pattern of increasing missingness is similar across all measurements.

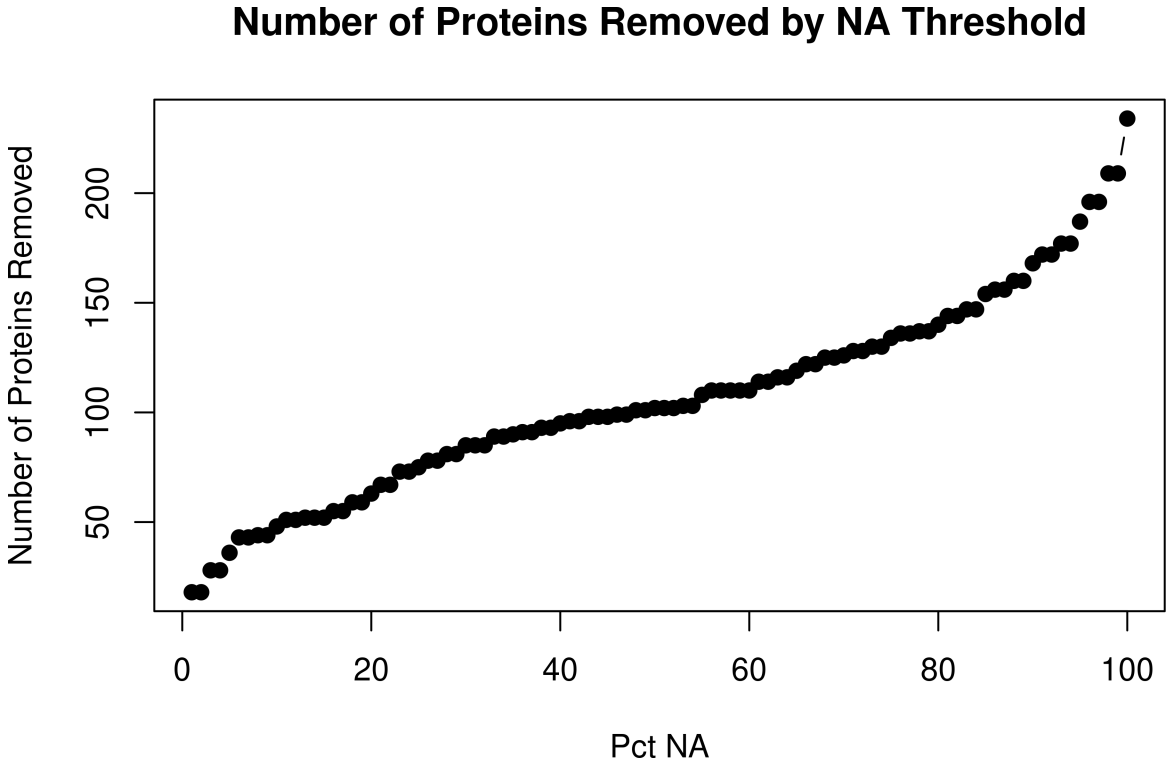
Table 5: Proteins with more than 70 pct missing values

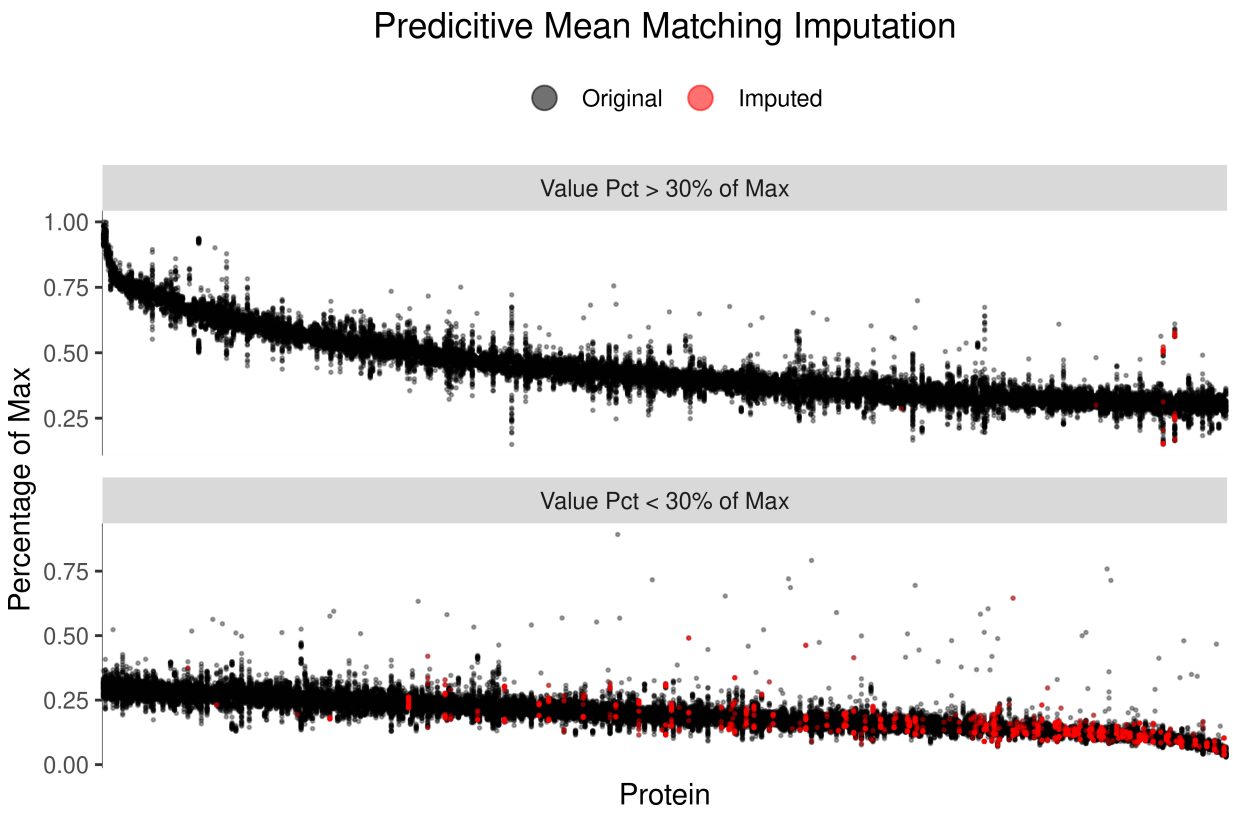
GBP2	IL2RB	CD28	METAP1	ARNT
CEACAM3	SIRT5	IL17F	CNPY2	CLSPN
NXPH1	PREB	WASF3	ACTN4	DCTN2
CTF1	EREG	HSP90B1	GLRX	EDIL3
CRX	PADI2	DPP7	SOD1	EIF5A
ECE1	LYPD1	KPNA1	DGKZ	IL13
DIABLO	DUSP3	REG3A	JUN	IL1A
ING1	TSLP	CSNK1D	RASA1	IL2
RCOR1	LTBP2	STXBP3	AHCY	IL20
IL24	FKBP4	S100P	TNFAIP8	IL33
GSTP1	GGA1	VASH1	TCL1B	LCN2
EPHA10	FXYD5	IQGAP2	LYAR	LTA4H
ITGB1BP1	LIF	EGLN1	RGS8	MAPT
IL4	NQO2	ATP6AP2	CAMKK1	NINJ1
CCL7	AGR3	RAB6A	NCLN	PPM1B
PDP1	DEFA1	BIRC2	ITGAM	SEPTIN9
ARTN	ANXA4	TPMT	APBB1IP	TPT1

Outlier Removal and Imputation

Proteins with high percentages of NA values will be removed. NA-values, or lack of protein detection, can clearly be seen to increase as the abundance of a protein decreases.

We remove proteins with more than 70% missing values as no meaninfgul imputation can be done. The cut off is heuristic, but gioded by the plot on missing values and expression data. A plot of how many proteins are removed at different NA thresholds (1-100%) indicates that roughly 70% is where the linear domain is lost.



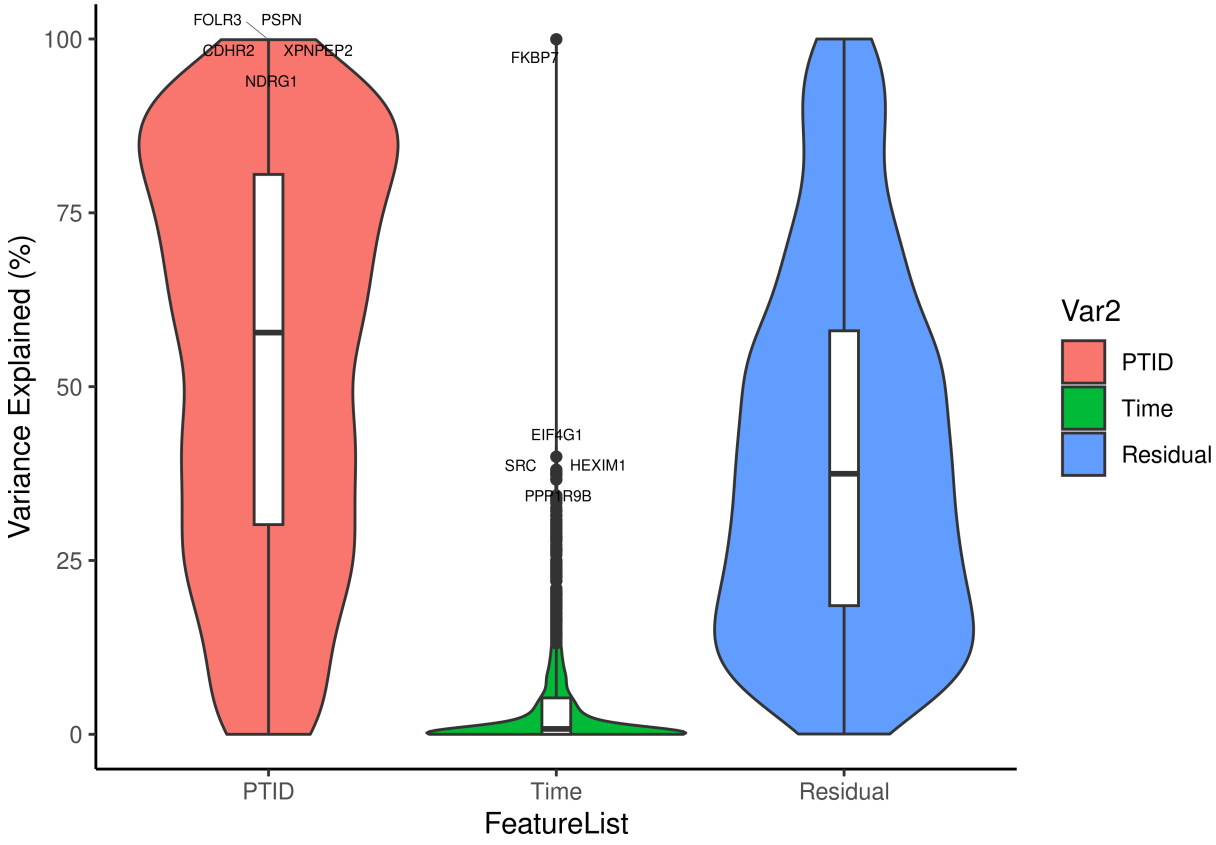


Values in red are imputed and overlap well with original data

Data was imputed with predictive mean matching PTID and Week were used as class and random variables with the package mice. Imputation mostly affects proteins with low expression values. The imputed values align well with the observed data.

The imputed values will be carried forward in analysis.

PALMO


FOLR3
PSPN
CDHR2
XPNPEP2
NDRG1
FKBP7
EIF4G1
SRC
HEXIM1
PPP1R9B
Variance Explained (%)
100
75
50
25
0
PTID
Time
Residual
FeatureList
Var2
PTID
Time
Residual

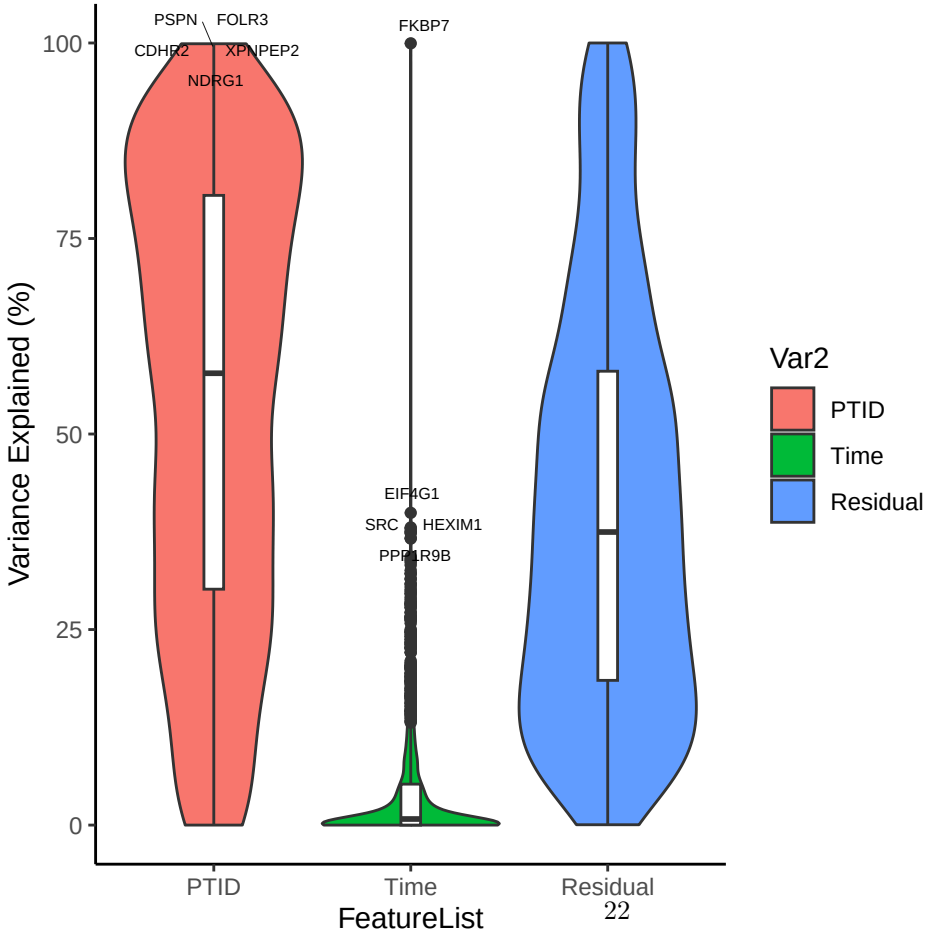
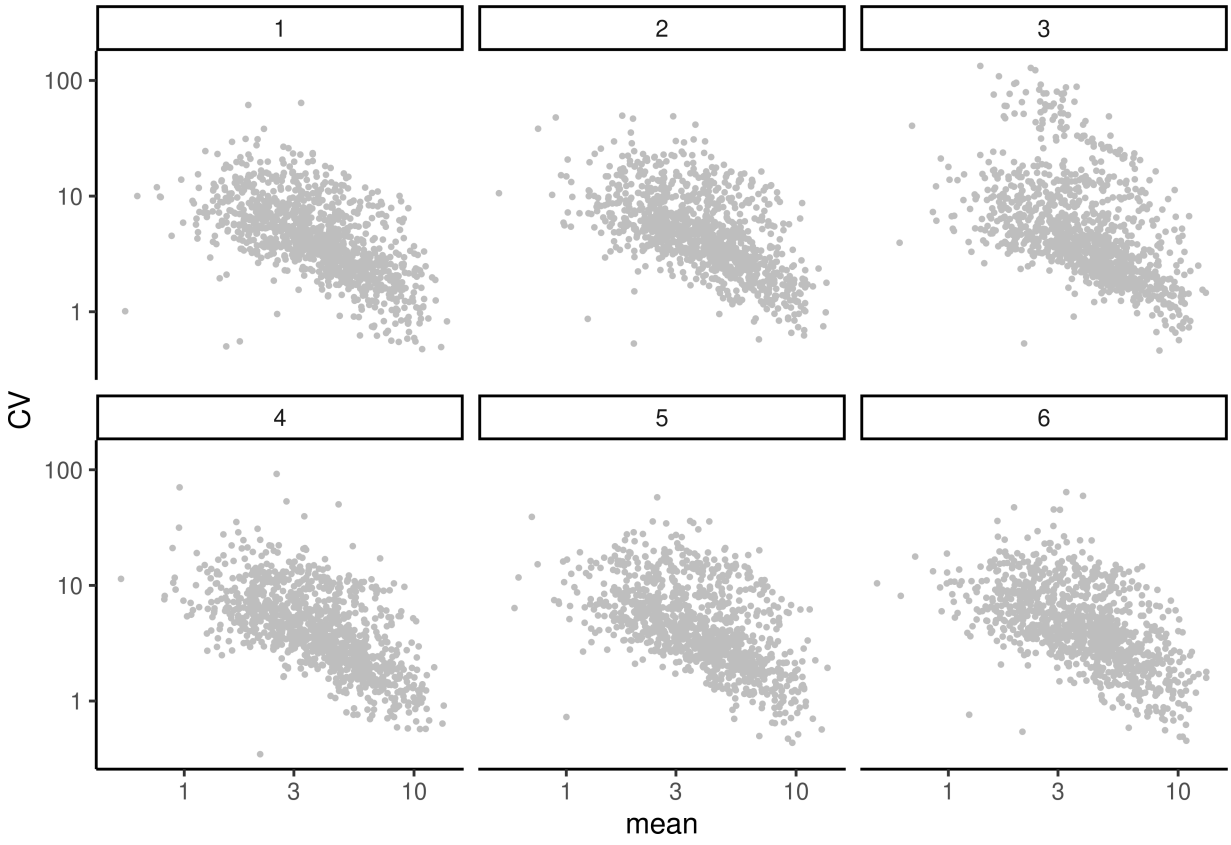

PSPN
FOLR3
CDHR2
XPNPEP2
NDRG1
FKBP7
EIF4G1
SRC
HEXIM1
PPP1R9B
Variance Explained (%)
100
75
50
25
0
PTID
Time
Residual
FeatureList
Var2
PTID
Time
Residual

Table 6: Proteins with highest inter-donor variation

Protein	Mean	Median	sd	Max	PTID	Time	Residual	n imputed
FOLR3	9.3	7.5	2.7	13.3	99.9	0.0	0.1	0
PSPN	3.8	4.3	1.3	5.6	99.3	0.1	0.6	0
CDHR2	2.3	2.0	0.9	4.4	99.1	0.1	0.8	1
XPNPEP2	8.9	9.1	0.7	9.7	98.7	0.0	1.3	0
NDRG1	3.6	2.6	1.9	7.4	98.5	0.0	1.5	0
IL17C	3.2	2.7	1.2	6.0	98.4	0.1	1.5	9
CSF2RA	5.4	5.7	0.7	6.1	98.3	0.0	1.7	0
MICA	4.3	4.5	0.6	4.8	97.7	0.0	2.3	0
ICAM5	6.0	6.0	0.5	6.8	97.4	0.0	2.6	0
LILRB5	3.9	3.8	0.8	5.1	97.4	0.0	2.6	0

Table 7: Proteins with highest time variation. Caution as 37/60 values for FKBP7 was imputed

Protein	Mean	Median	sd	Max	PTID	Time	Residual	n imputed
FKBP7	1.7	1.3	1.8	10.1	0.0	100.0	0.0	37
EIF4G1	5.2	5.3	0.7	6.9	11.3	39.9	48.8	0
SRC	7.4	7.5	0.5	8.1	18.3	38.1	43.6	0
HEXIM1	5.4	5.4	0.5	7.0	11.7	37.9	50.4	0
PPP1R9B	4.2	4.3	0.6	5.8	20.4	37.6	42.0	0
METAP2	5.1	5.1	0.4	6.0	17.2	37.5	45.4	0
MYO9B	2.2	2.2	0.4	3.1	28.7	37.3	34.0	0
DNAJB1	4.3	4.3	0.5	5.4	29.3	36.7	34.0	0
PRDX1	1.7	1.7	0.2	2.1	38.9	36.6	24.5	0
PLXNA4	6.8	6.8	0.7	8.7	14.7	34.3	51.0	0



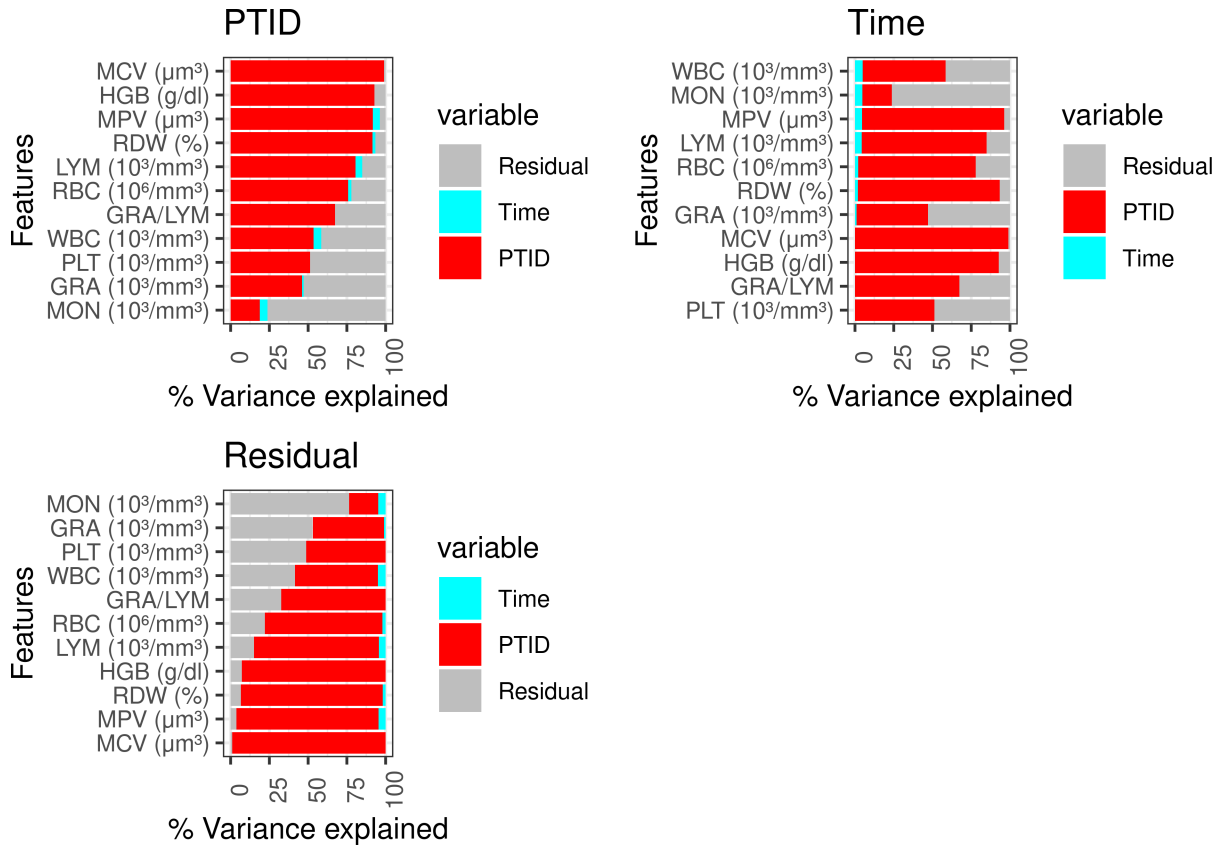


Figure 17: Variance Feature Plots

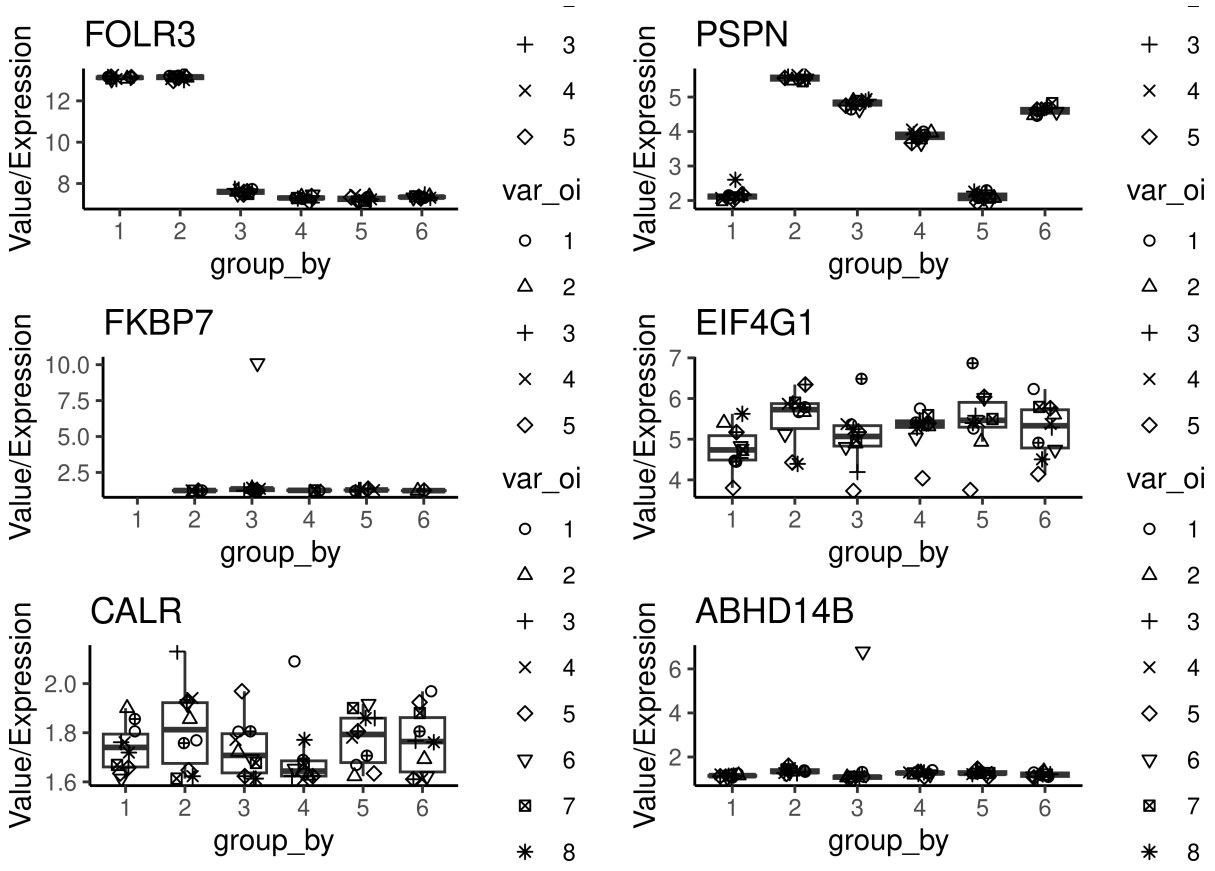
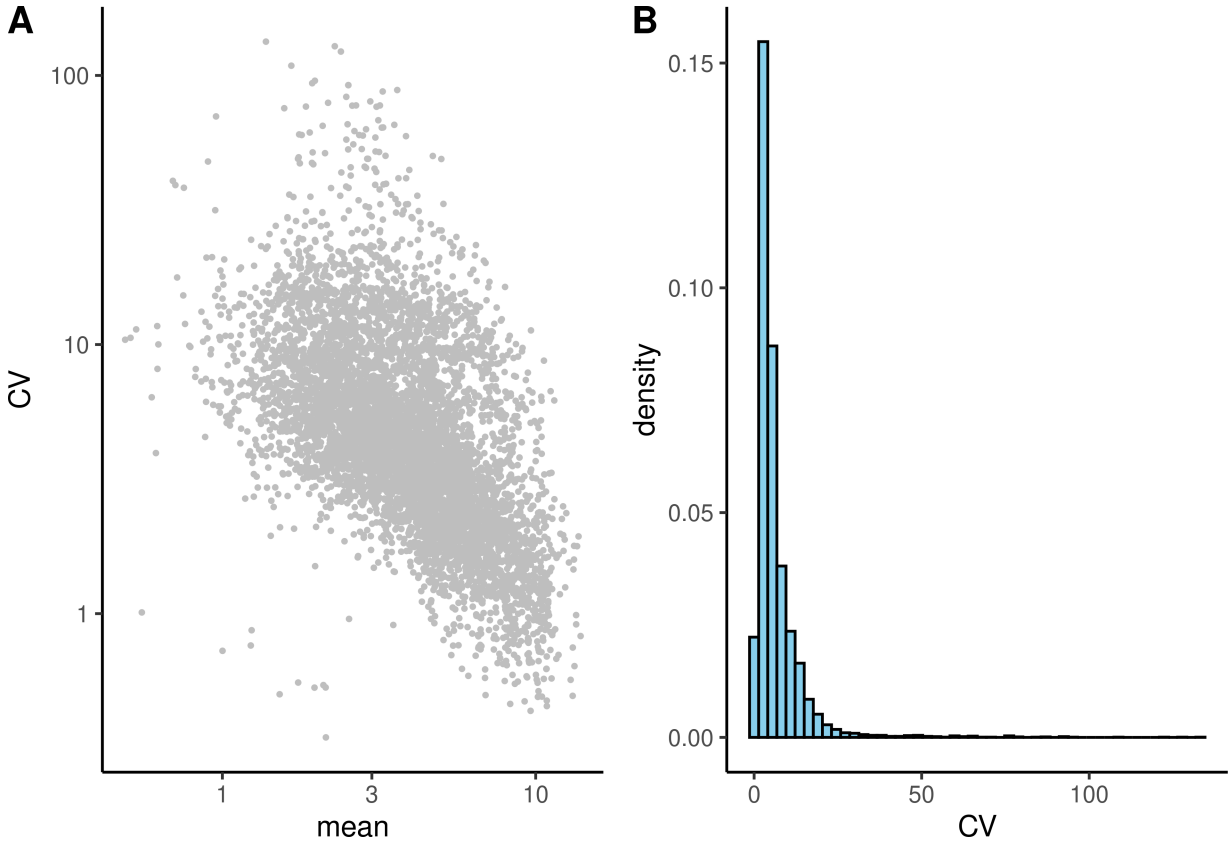


Figure 18: Protein Feature Plots. FOLR3 and PSPN can readily be seen to have high inter-donorvariation and low intra-donor variation. FBP7 had many imputed values. EIF4G1 was top in variationover time. ABHD14 and CALR had high residual variance




```
## [1] "output/sample_cv_split_000001.pdf" "output/sample_cv_split_000002.pdf"
## [3] "output/sample_cv_split_000003.pdf" "output/sample_cv_split_000004.pdf"
## [5] "output/sample_cv_split_000005.pdf" "output/sample_cv_split_000006.pdf"
```

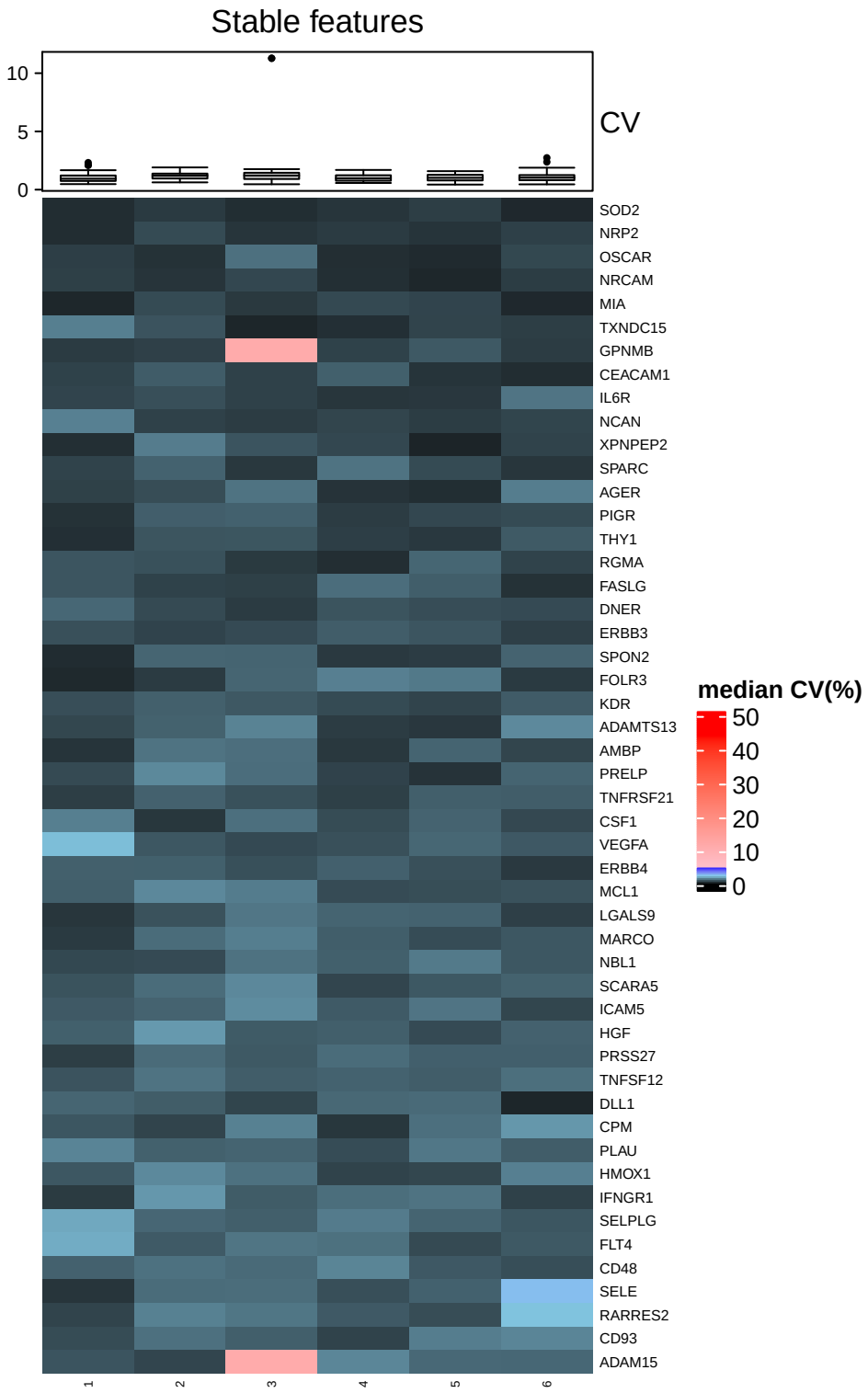


Figure 19: Stable Proteins

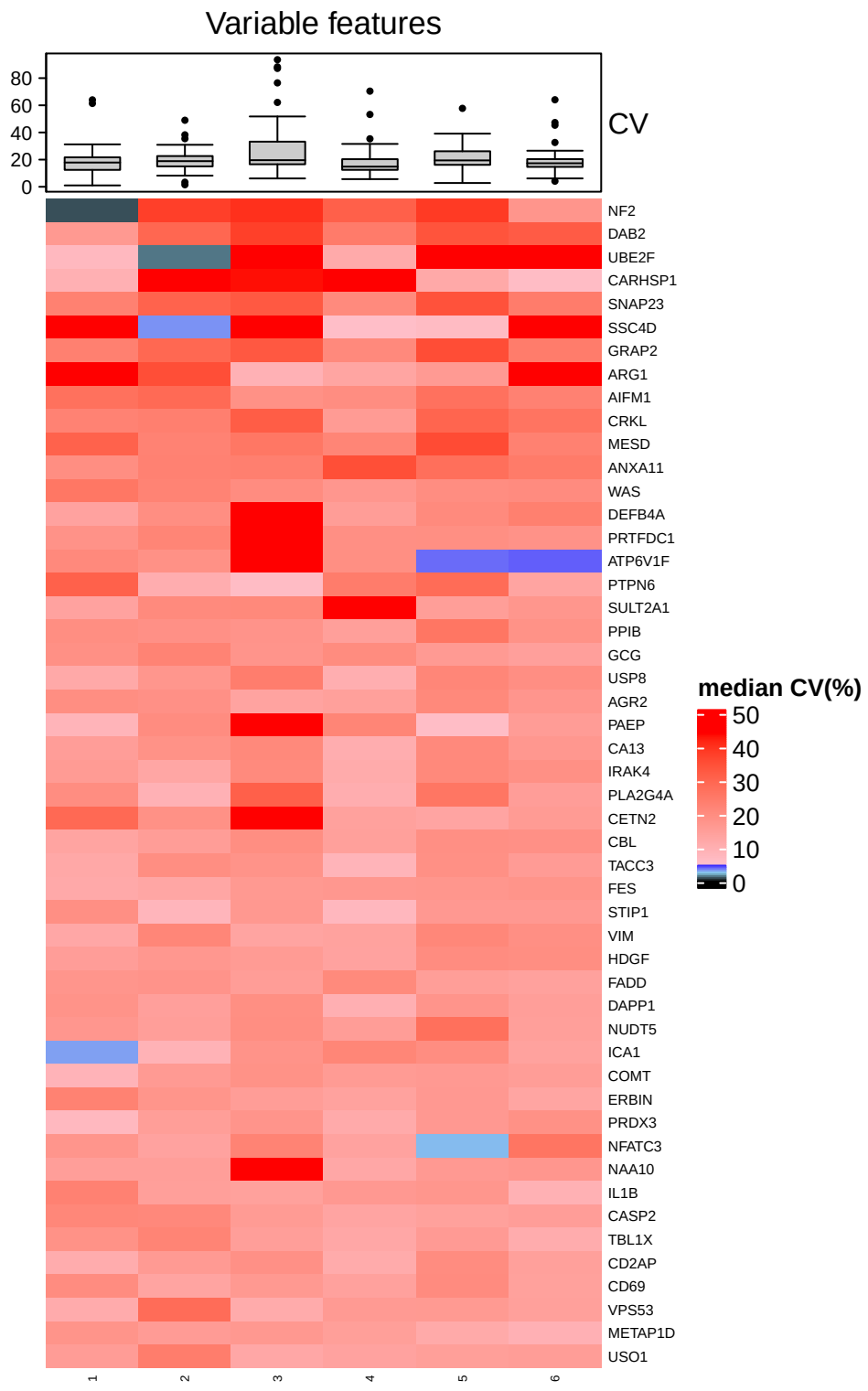


Figure 20: Variable Proteins

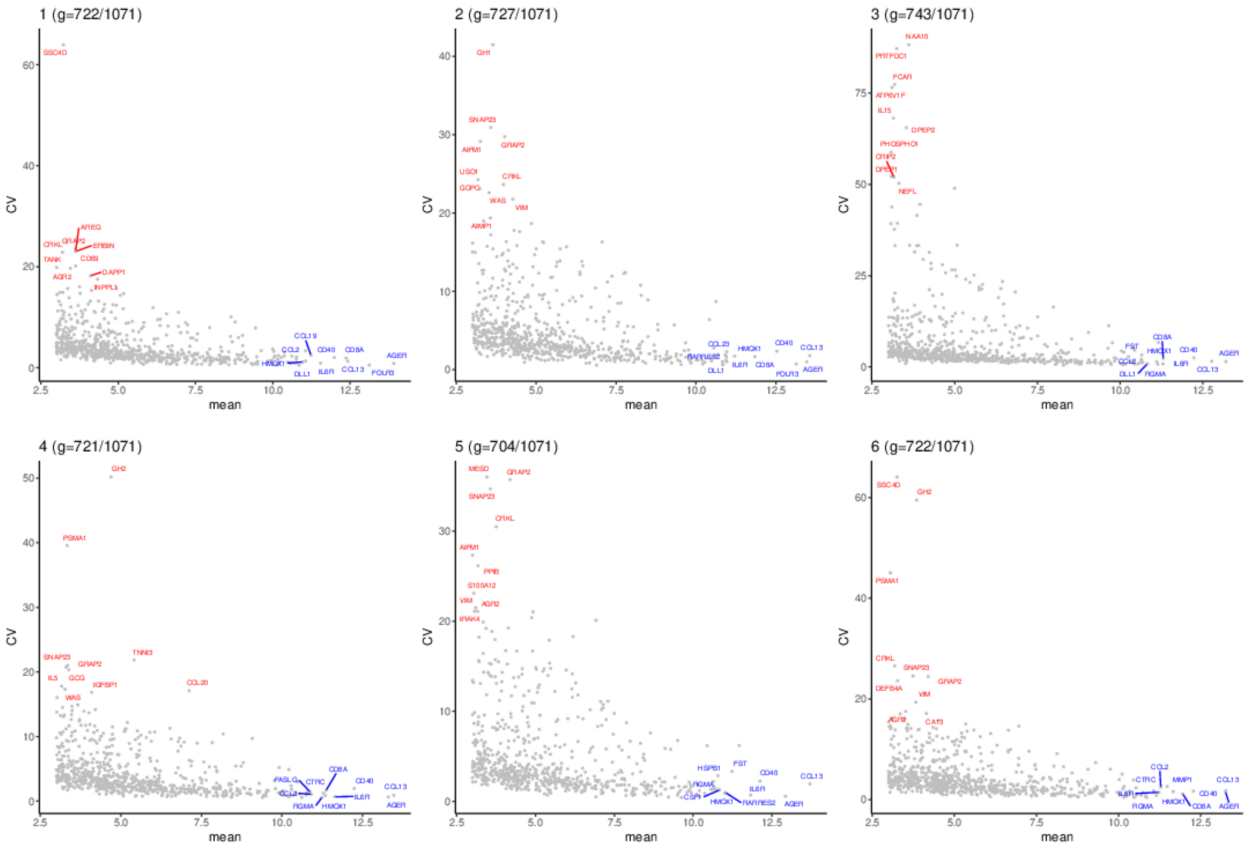


Figure 21: Sample CV plots

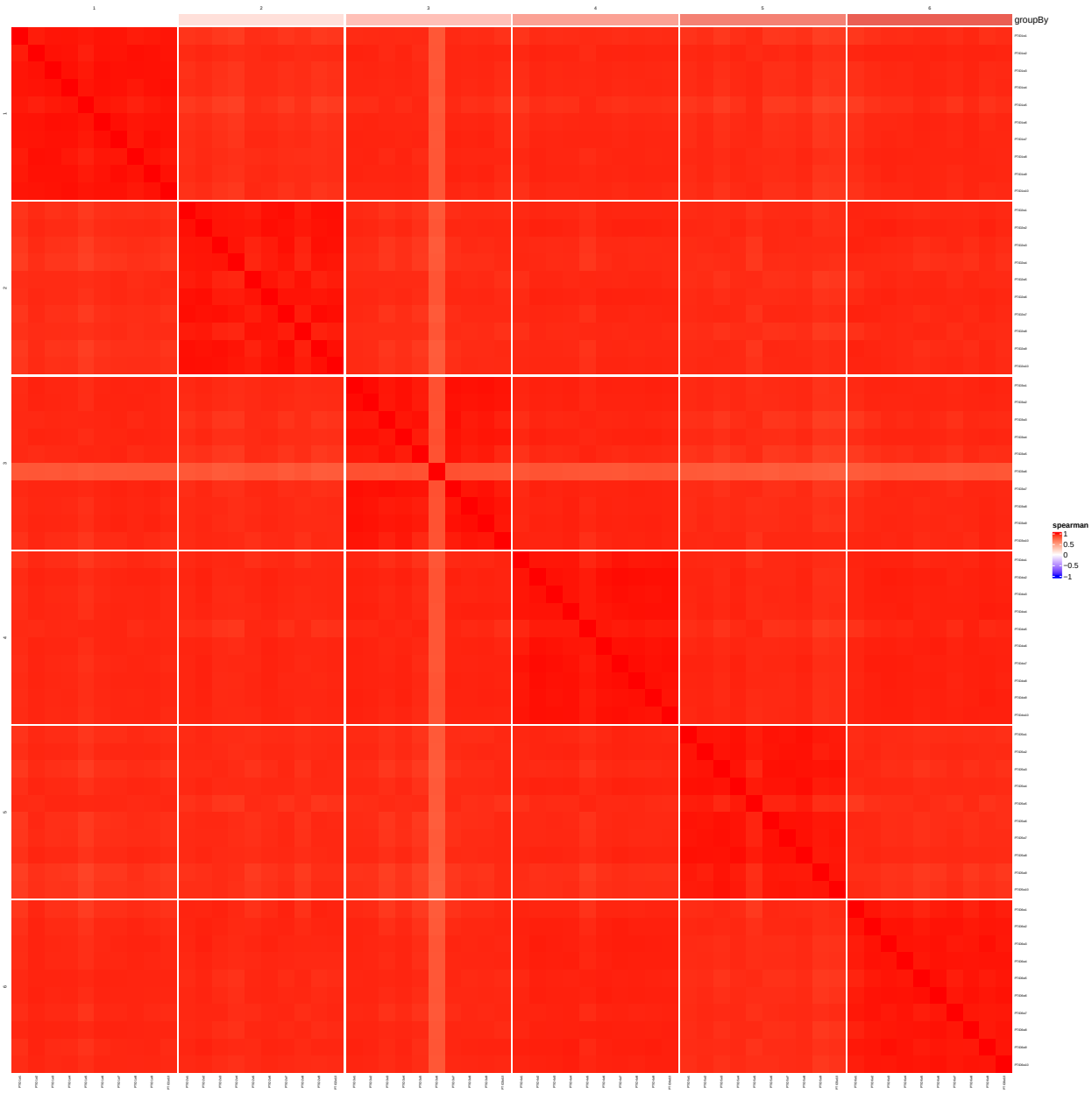


Figure 22: Sample autocorrelation.It can be readily seen that there is a high autocorrelation among samples.

t-SNE

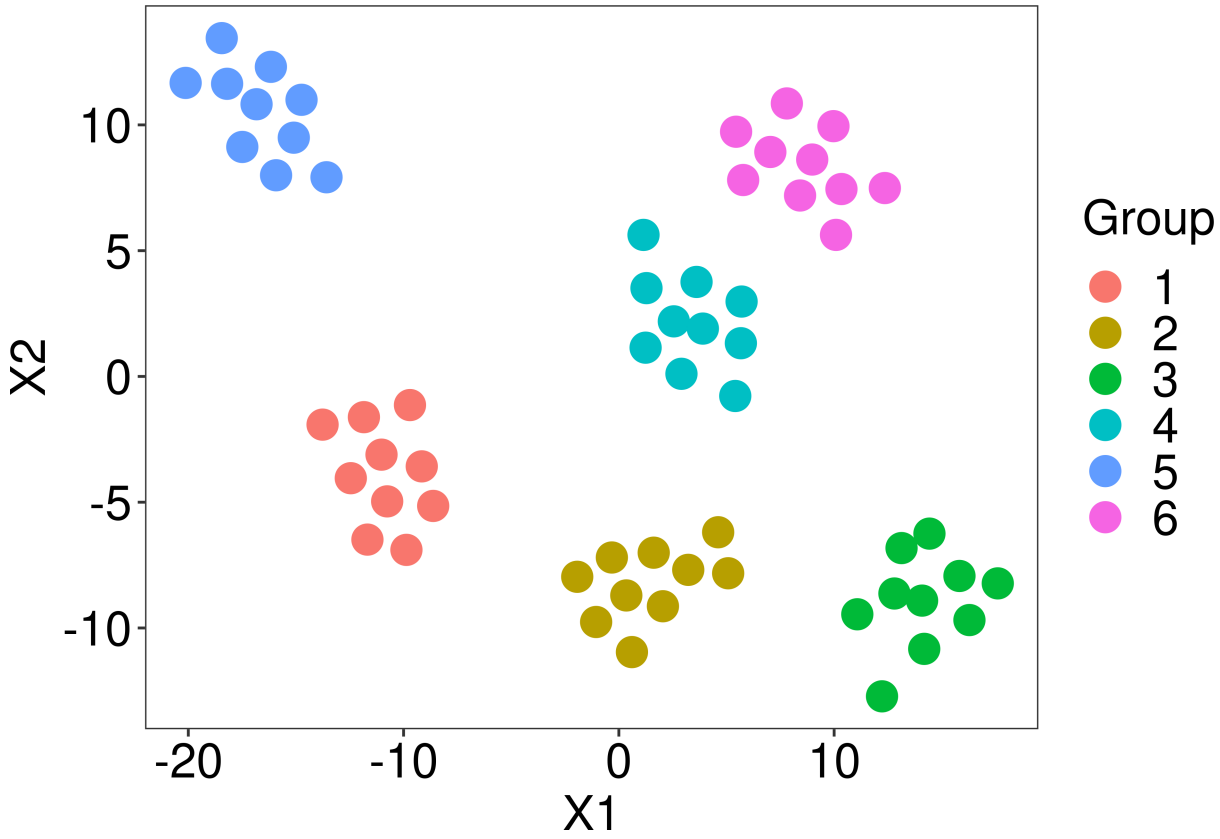


Figure 23: t-SNE of proteome. Data clusters per subject

PCA

PCA was performed on Olink data to check for meaningful decomposition. However, decomposition failed to find much variance coverage in the first axes. Only 30% were covered in the two first axes.

MOFA

MOFA is a package to do factor analysis generalised to multi-omic data. An addition has been added to allow covariates, such as time. The data set from CBC, immunophenotype and proteome will be analysed together and weeks added as a covariate. Data was scaled as they have inherently different ranges.

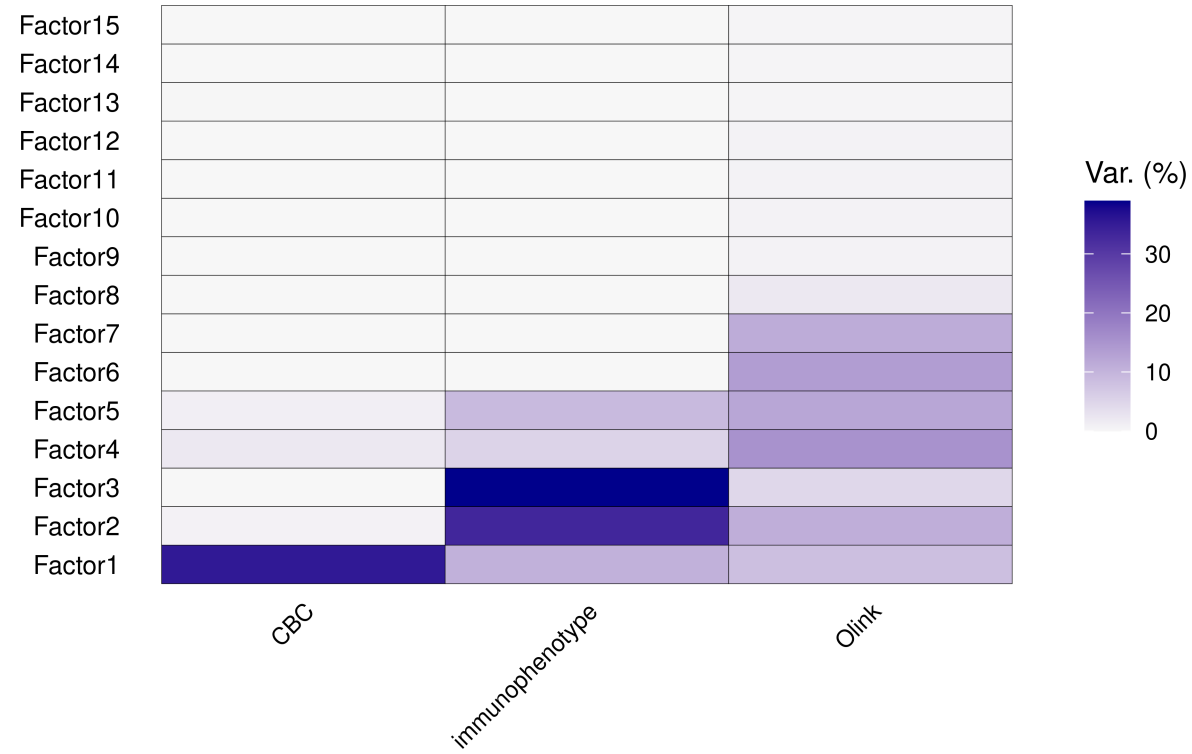


Figure 24: Variance explained in MOFA Analysis. CBC and and immunophenotype canbe seen to have features that concentrate, indicating a simple structure forvariance, whereas olink data is distributed over many factors

Table 8: Weight of the covariate time to each factor. It can be seen that time and the other features (mainly factor 1-8) are orthogonal

	x
Factor1	0.0000000
Factor2	0.0000000
Factor3	0.0000000
Factor4	0.0000000
Factor5	0.0000000
Factor6	0.0000000
Factor7	0.3244583
Factor8	0.1048001
Factor9	0.8867957
Factor10	0.6575238
Factor11	0.9288854
Factor12	0.0000000
Factor13	0.7949425
Factor14	0.0000000
Factor15	0.8096951

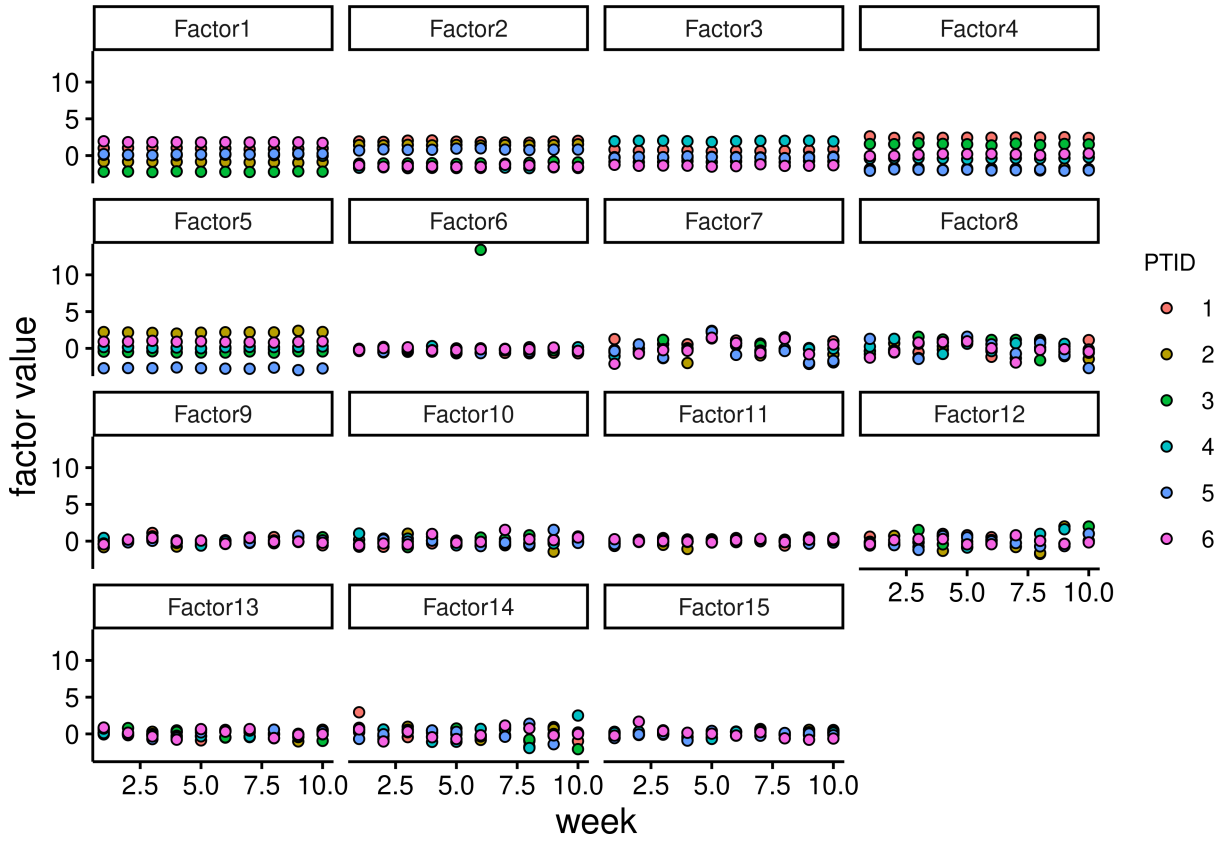


Figure 25: The effect of the covariate time on factor weights.The effect of time is somewhat evenly distributed.In factor 4, PTID 5 can be seen to be a bit of an outlier. The result is expected, asthere was no intervention and no expectation to see a difference at different weeks

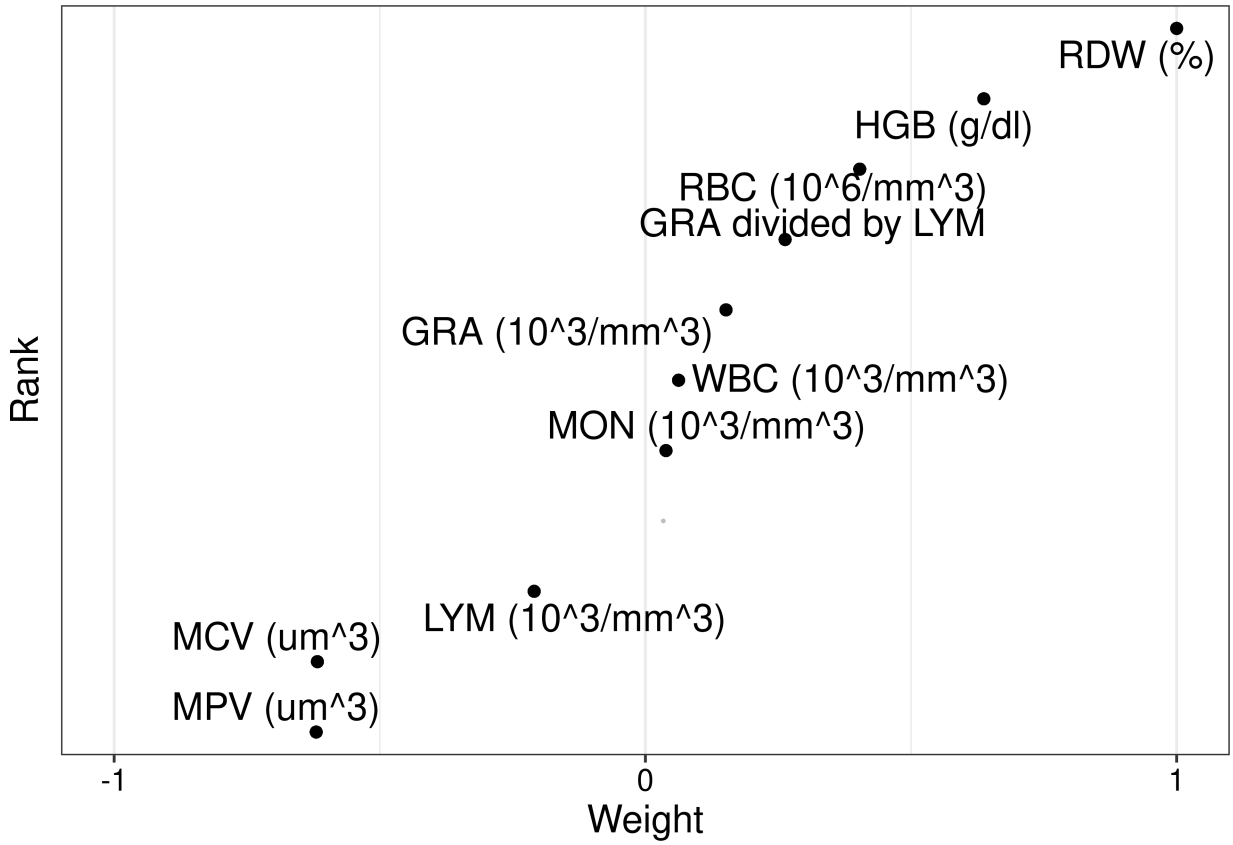


Figure 26: Weights for factor 3 in CBC samples (that captured most variance).PLT and MCV are most prominent. PLT is expected and has a normal high variance.MCV likely reflect PTID 3 that had anemia.

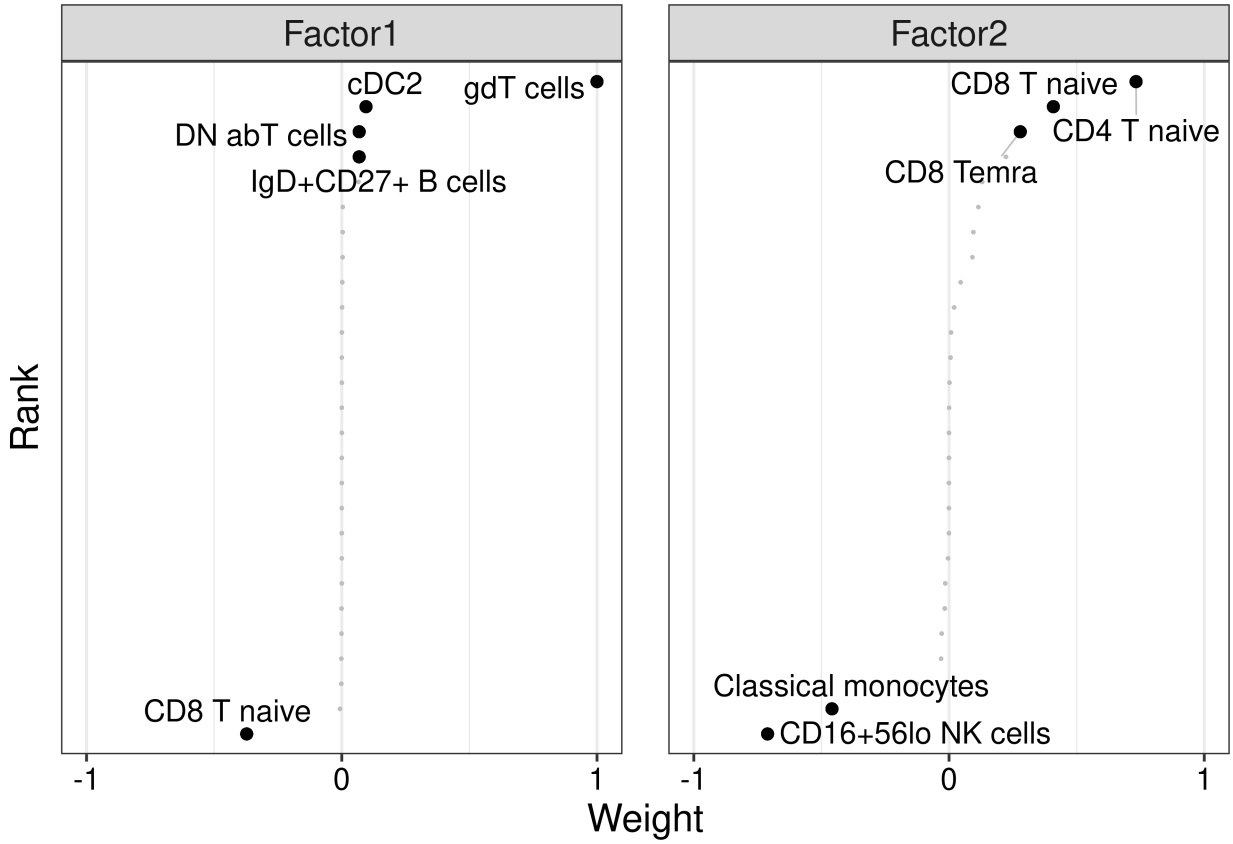


Figure 27: Top weights in immunohenotyping factor 1 and 2. The result is indicativeof the inter-individual difference in these subsets, but they are generally stablewithin an individual as indicated from the lack of covariate (time) effect

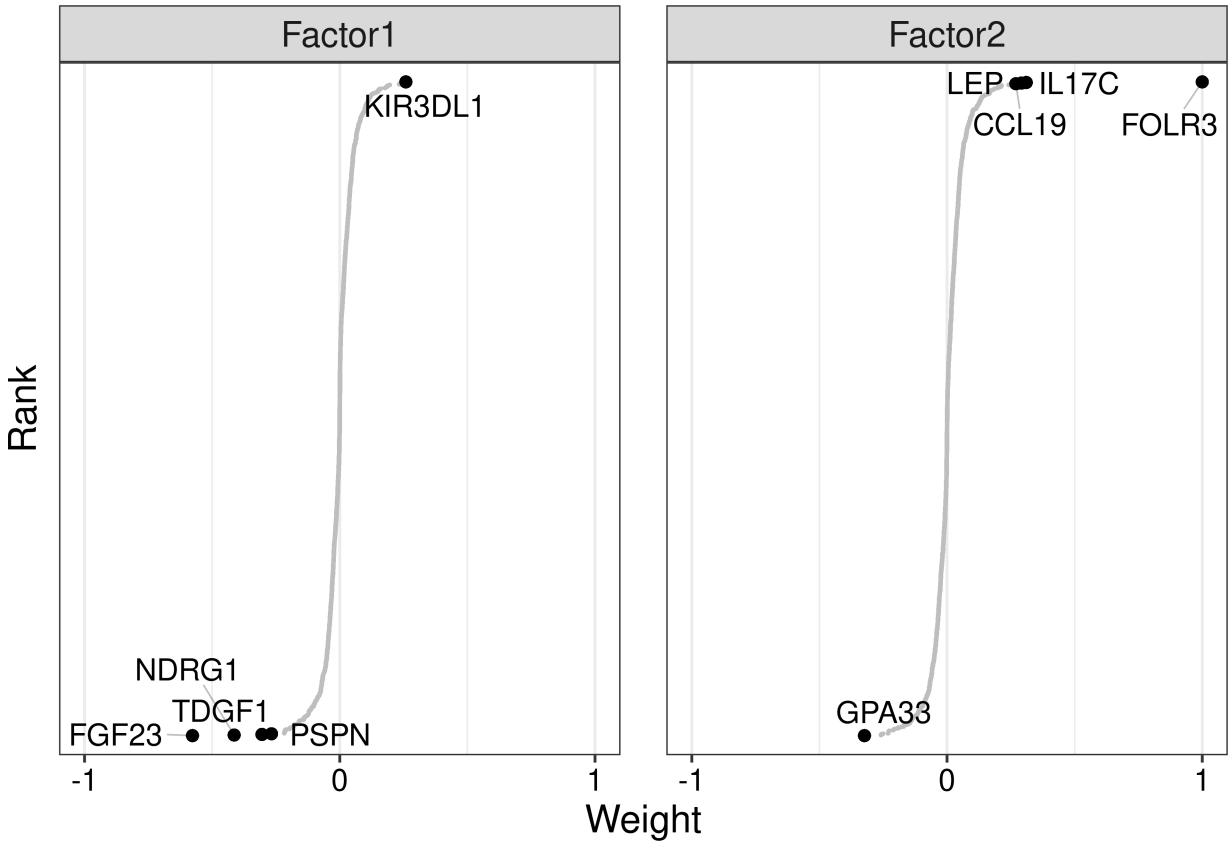


Figure 28: Top features to factor 1 and 2

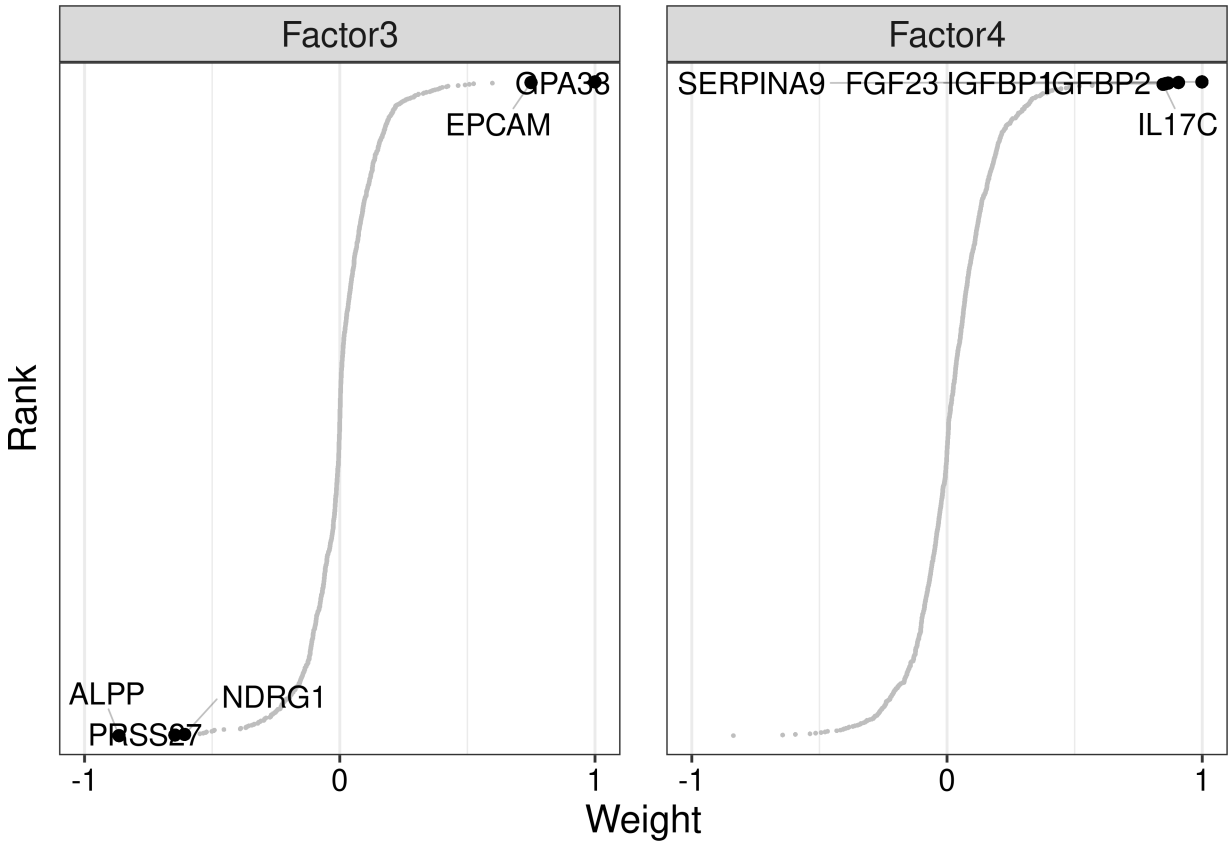


Figure 29: Top features to factor 3 and 4

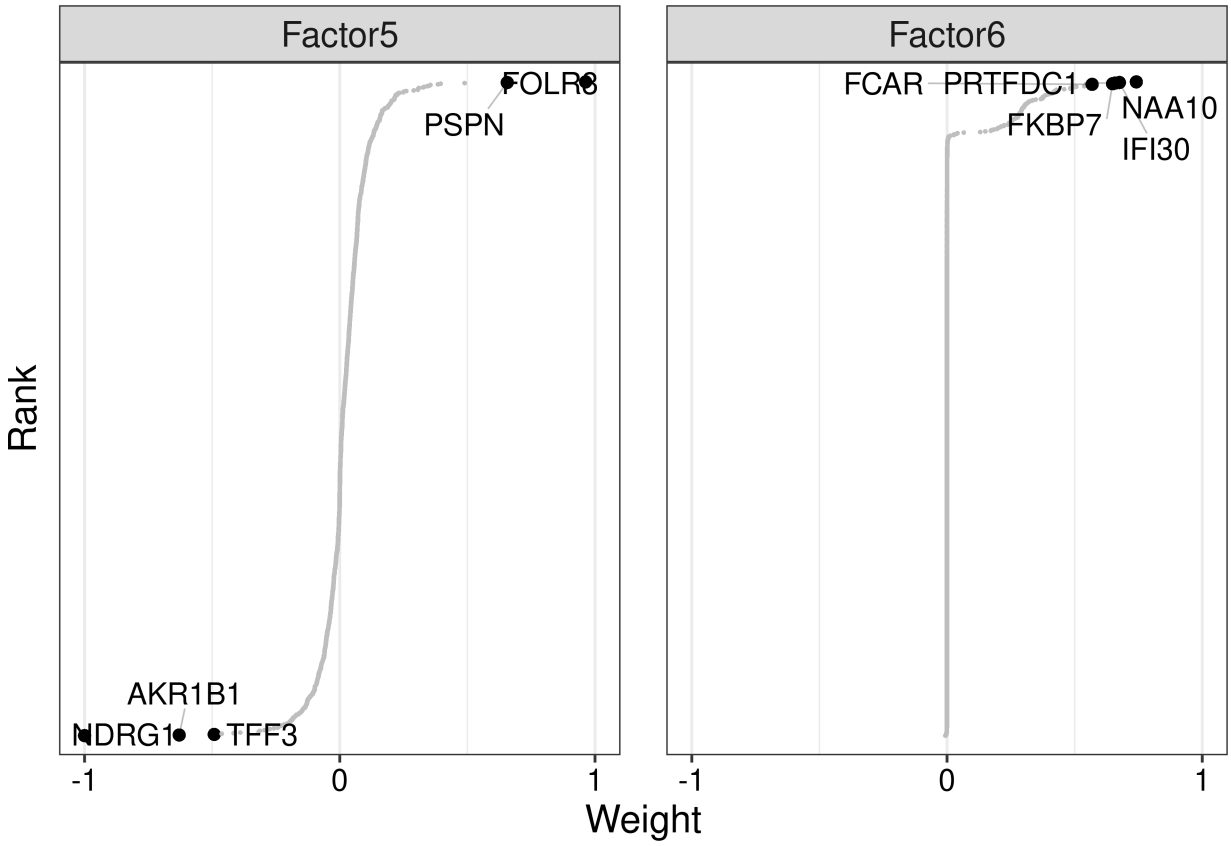


Figure 30: Top features to factor 5 and 6

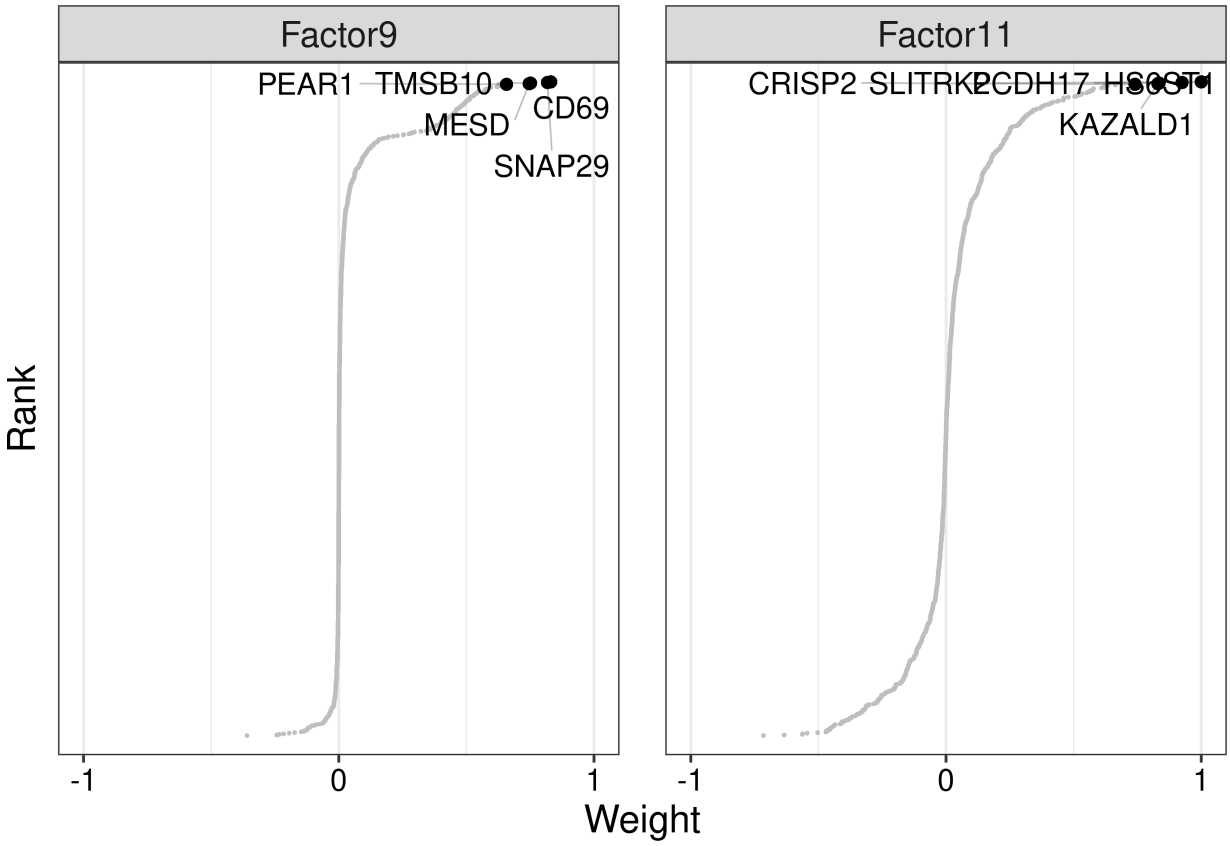


Figure 31: Top features to factor 9 and 11 where time had most effect

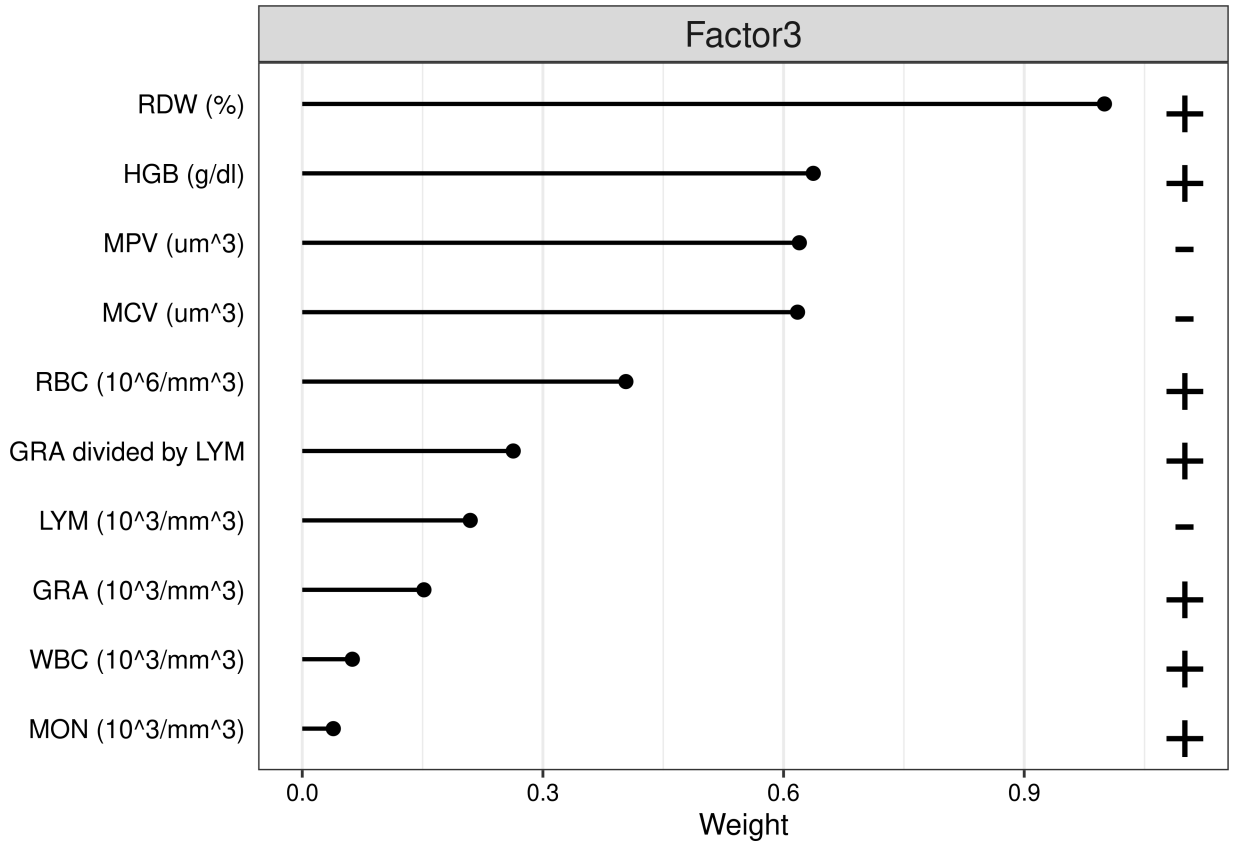


Figure 32: Example of lollipop chart of top weights in factor 3 in CBC.Individual weights are easier to compare than previous plot. Signs to the rightindicate direction of weight

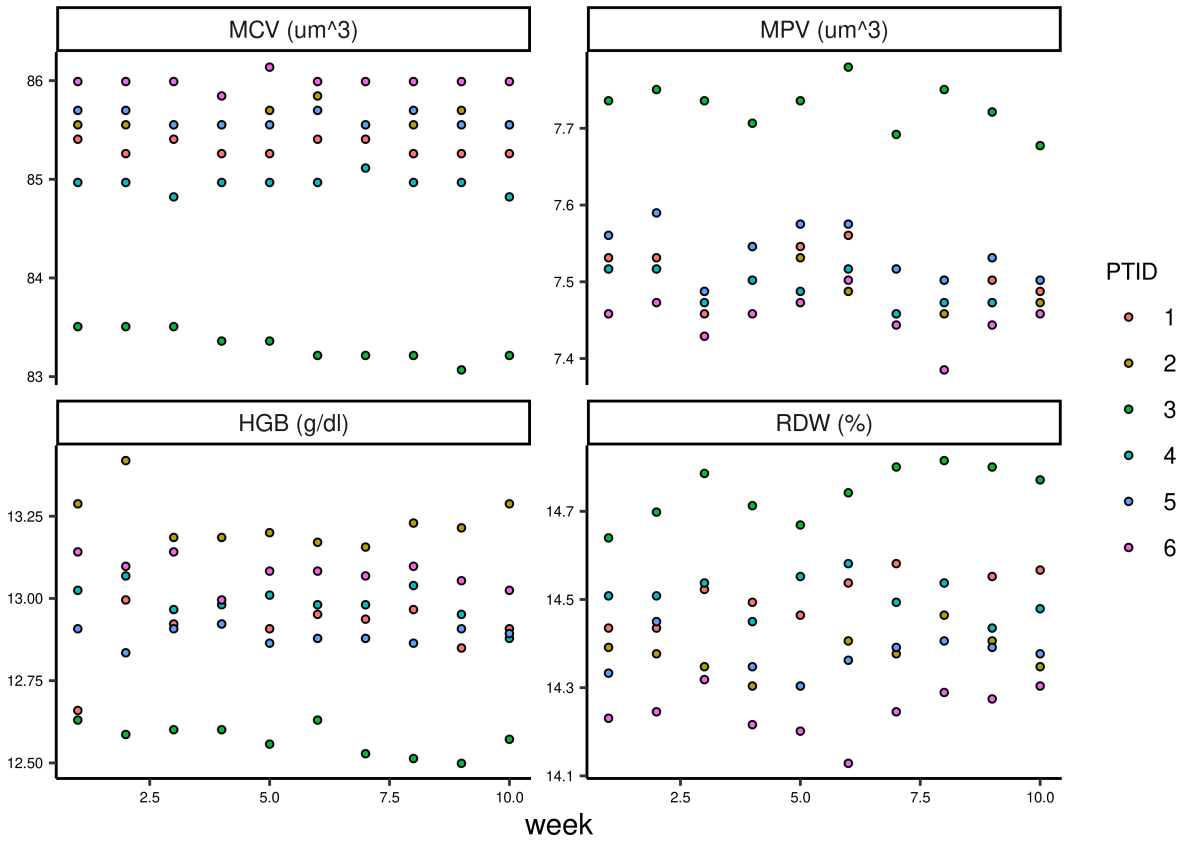


Figure 33: Plotting data vs covariates colored for PTID we can easily see thatHGB, MCV and RDW are due to the anemic PTID 3. PLT is due to somewhat lowerplatelets in PTID 6 (still normal)

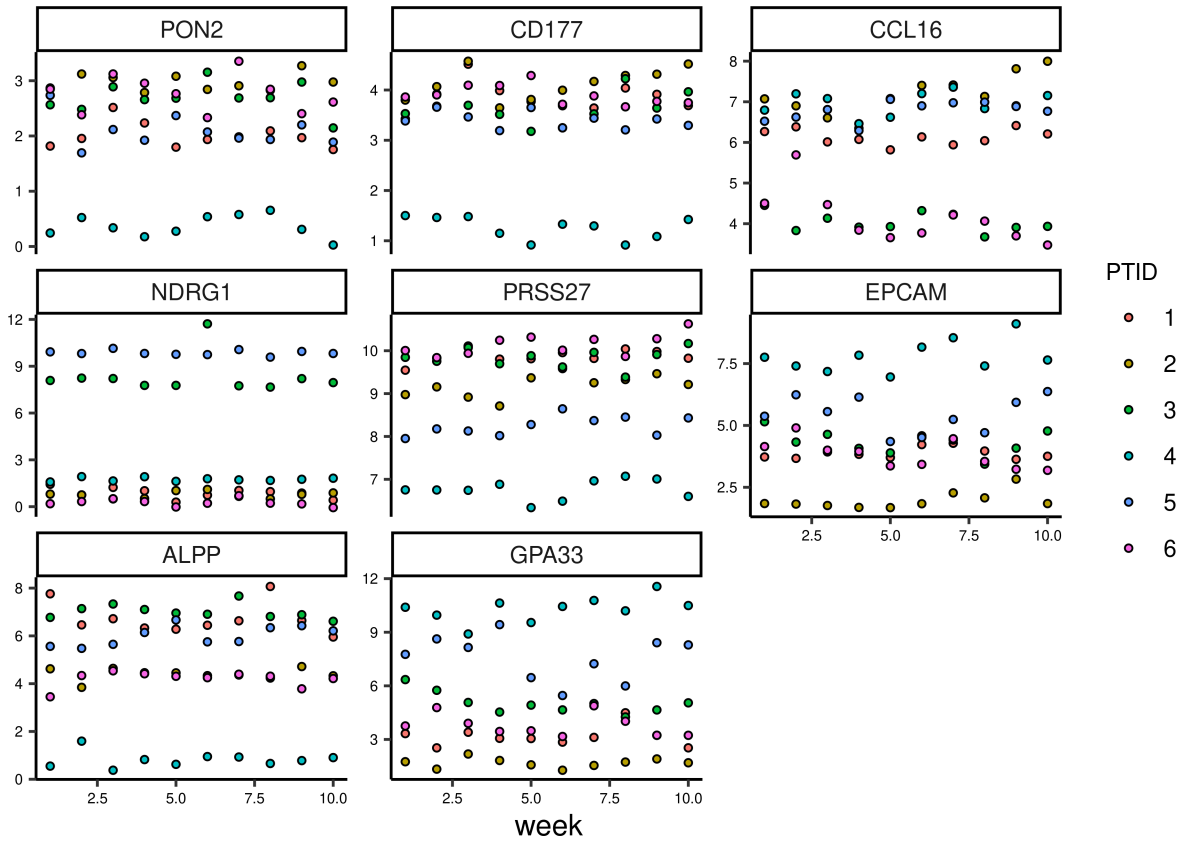


Figure 34: Top proteins in factor 3 (where CBC had most variance). PTID3 can be seento segregate in this factor as well, but other subjects do as well.GH2, GSAP, FGF23 might be related to that subject's anemic state

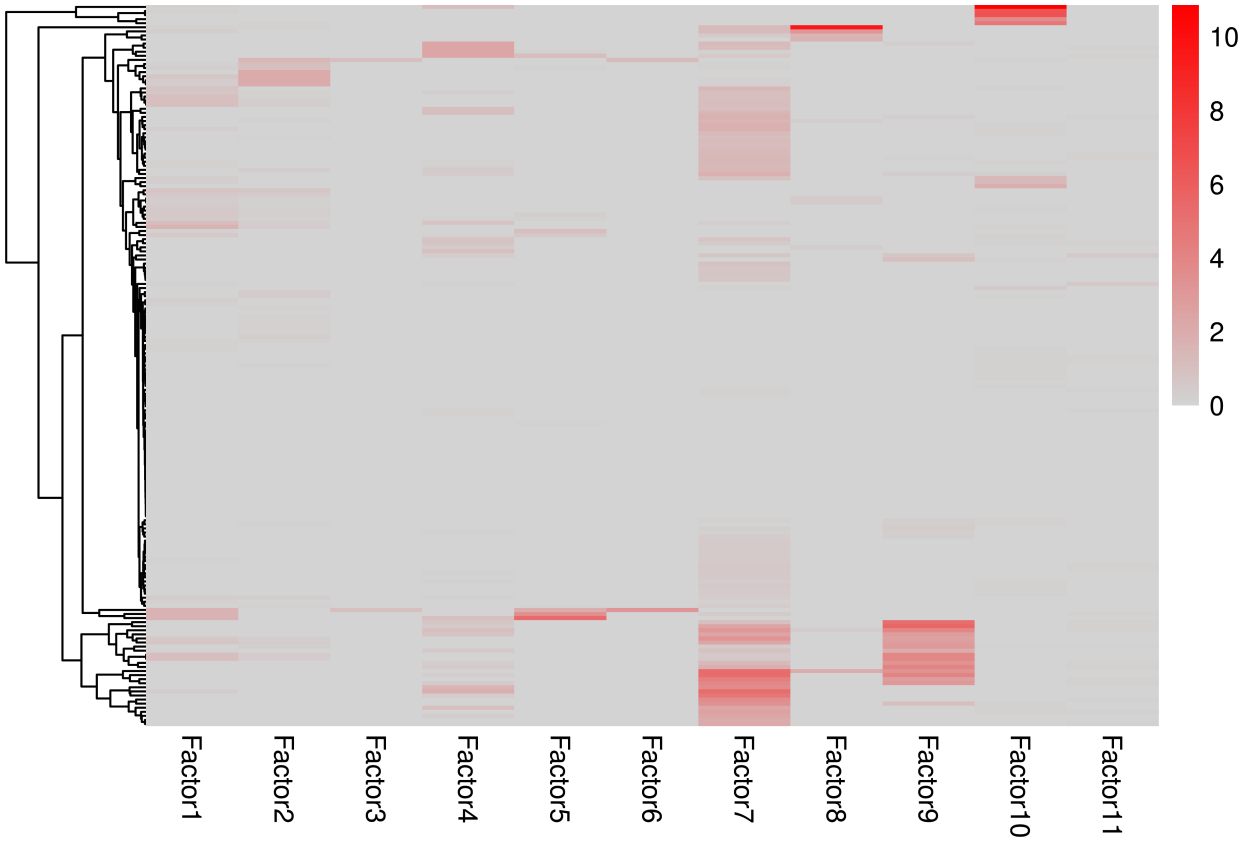


Figure 35: Heatmap och enrichment analysis of proteins against MSigDB.Factor 7 and 10 seem to capture most proteins

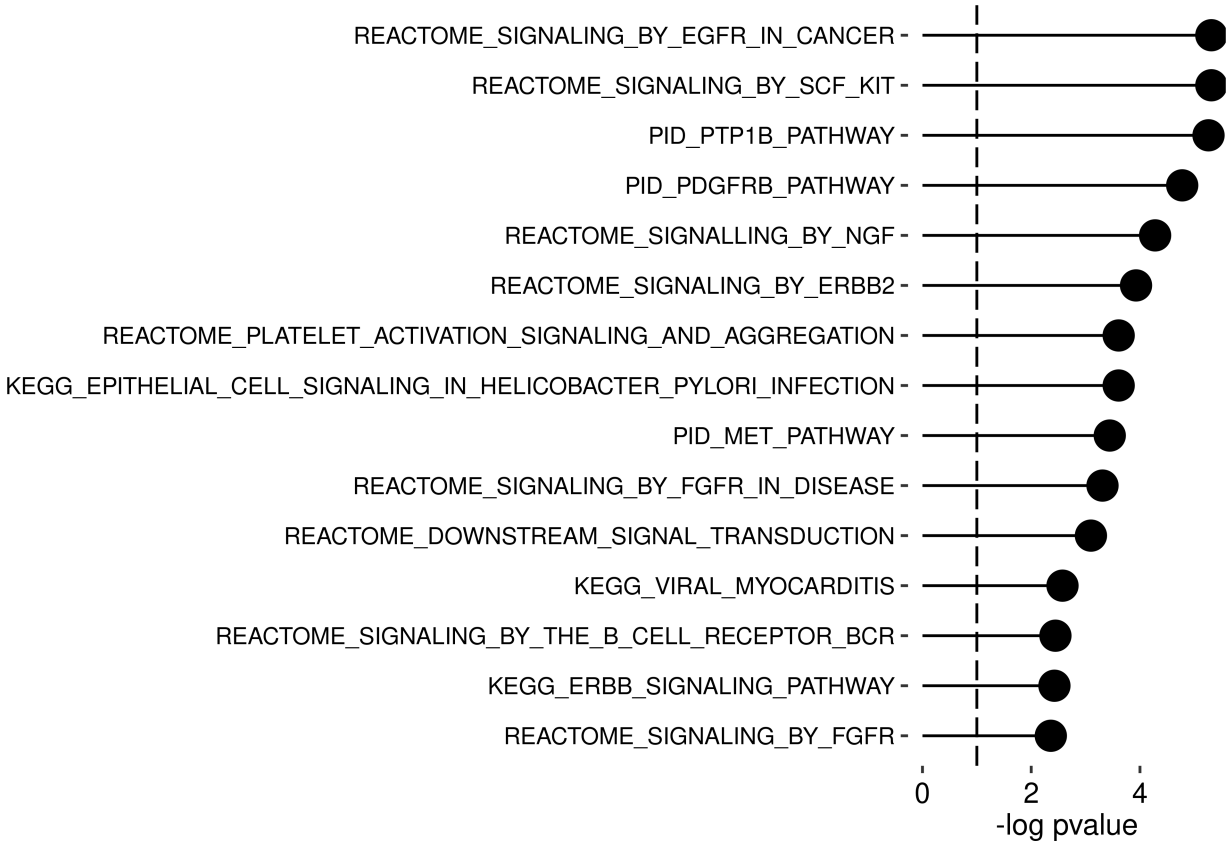


Figure 36: Factor 7 protein enrichment for MSigDB pathways.

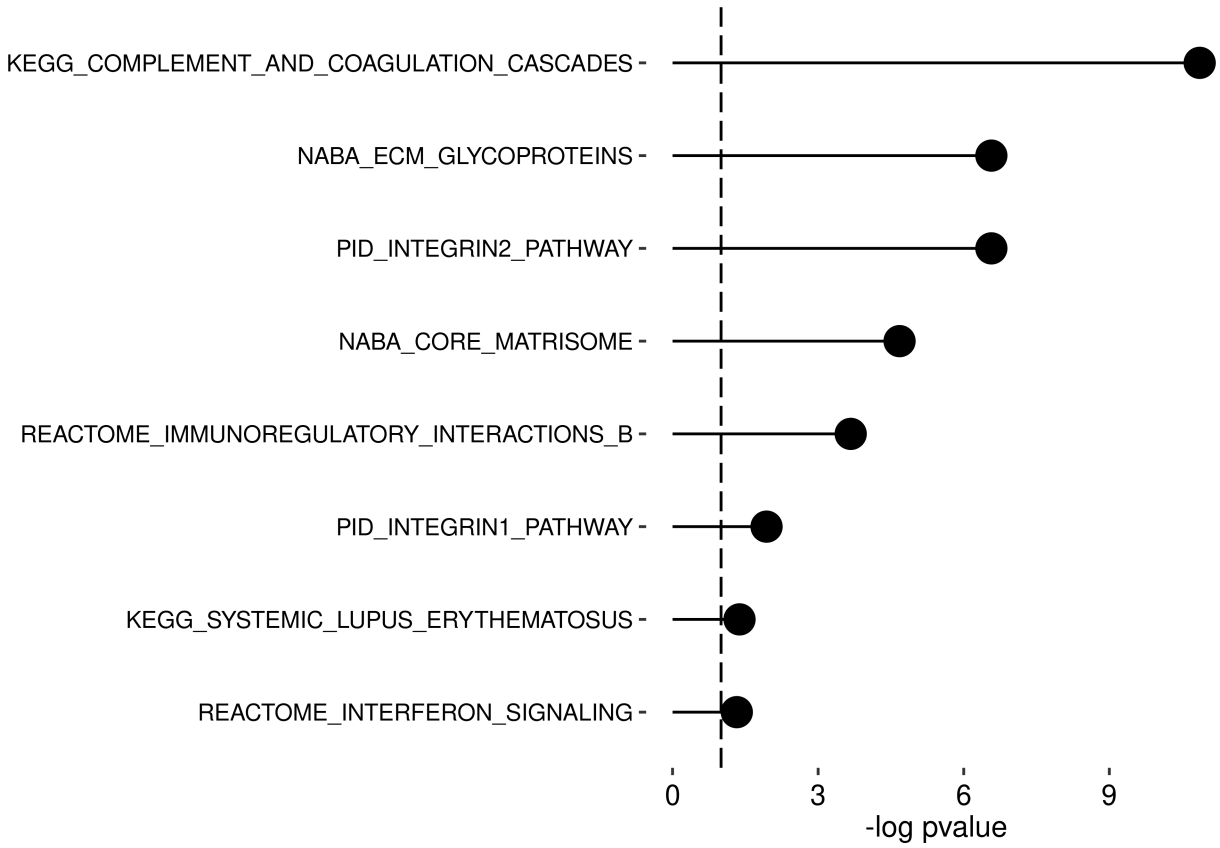


Figure 37: Factor 10 protein enrichment for MSigDB pathways.Pathways clipped to 40 characters

Final Remarks

The data set included multimodal data on healthy subjects. The different data sets had a large range in features. As there is no intervention, no specific hypothesis can be made with regards to the longitudinal follow up as they are deemed to be by chance or unknown effects from activities of normal life not captured in the study. The data, however, show that in all three domains, CBC, immunophenotype and plasma proteome, most variance is caused by different basal levels in different subjects, and quite stable over time. One subject was anemic, contributing a sustained difference in certain CBC parameters and might explain some of the differences seen in proteomic data.