

Electronic supplementary material

ESM Method description

Description of double cross-validation approach for feature selection in classification and survival-regression

Method description

The following paragraph describes the double cross-validation (dCV) procedure that selects a maximally predictive subset of peptides for both classification models (e.g. islet autoantibody-positive vs. autoantibody-negative in the selection phase) and survival models (progression time in the application phase).

An outer cross-validation (CV) loop splits the dataset into an outer training and an outer test set. In an inner CV loop, the training data is again split into an inner training set and an inner test set. The aim of the inner CV loop is to determine the optimal number of peptides for predictions on the inner test set. To this end, the peptides are ranked with a support vector machine using a linear kernel and recursive feature elimination [1] (for classification) or a L1 regularization approach using a Cox lasso model (for survival) [2]. Peptides are added to a logistic regression (classification) or Cox regression (survival) model in a stepwise fashion according to this ranking, and prediction accuracy is calculated within the inner test set. All inner CV folds are averaged to determine the optimal number of peptides where maximum accuracy occurs. The ranking approach is then applied to the whole (outer) training set, and a model with the calculated optimal number of peptides is fitted. This model then predicts the outcome on the outer test set. This general procedure guarantees that model fitting is strictly performed within the outer training set, leading to unbiased prediction estimates [3].

As the measure of prediction accuracy we used the area under the curve (AUC) of the receiver operating characteristic (ROC). For the survival models we used an extension of the AUC called 'survival AUC', sAUC [4]. Variability of the (s) AUC was quantified by repeating the entire dCV approach. In order to choose the most predictive peptides, results from all outer

CV runs and repetitions were aggregated, and peptides were selected according to their selection frequency in all runs.

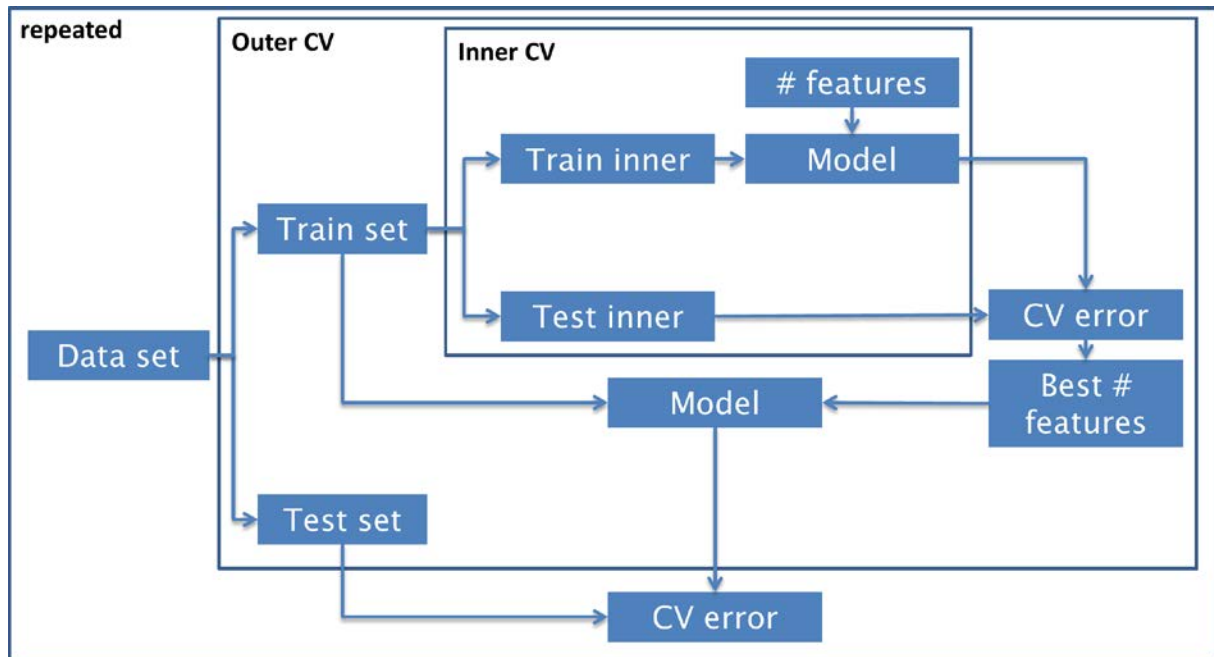


Figure for supplemental methods: Overview of the repeated dCV approach, which ensures a clean separation of training and test sets during all stages of the algorithm.

Parameter settings

In the selection phase, due to the low sample size, we used a leave-one out procedure for both the outer and inner CV loop. No variation from repetitions will be produced, so we performed this procedure just once.

In the application phase, islet autoantibody-positive vs. autoantibody-negative children, we used 5-fold cross-validation for both the outer and inner loop, with 20 repetitions of the entire procedure. For the survival analysis, we also used 5-fold cross-validation for both loops, with 50 repetitions.

ESM Tables

ESM Table 1 – Proteins and peptides analyzed by selected reaction monitoring (SRM) in the application phase

HGNC Gene Symbol	Uniprot Protein Symbol	Description	Target peptide
ALB	ALBU_HUMAN	Albumin	HPDYSVLLLR QNCELFEQLGEYK
APCS	SAMP_HUMAN	Amyloid P component, serum	IVLGQEQDSYGGK
APOA1	APOA1_HUMAN	Apolipoprotein A-I	VSFLSALEEYTK
APOA2	APOA2_HUMAN	Apolipoprotein A-II	SPELQAEAK
APOA4	APOA4_HUMAN	Apolipoprotein A-IV	LGEVNTYAGDLQK ISASAEELR
APOC4	APOC4_HUMAN	Apolipoprotein C-IV	ELLETVVNR
APOD	APOD_HUMAN	Apolipoprotein D	NILTSNNIDVK
APOE	APOE_HUMAN	Apolipoprotein E	SELEEQLTPVAEETR LEEQAQQIR
APOH	APOH_HUMAN	Apolipoprotein H	VCPFAGILENGAVR
APOM	APOM_HUMAN	Apolipoprotein M	WIYHLTEGSTDLR SLTSCCLDSK
BTD	BTD_HUMAN	Biotinidase	VDLITFDTPFAGR LSSGLVTAALYGR
C1R	C1R_HUMAN	Complement component 1, r	LFGEVTSPLFPK*
C3	CO3_HUMAN	Complement component 3	QELSEAEQATR SGSDEVQVGQQR
C4A	CO4A_HUMAN	Complement component 4A	ITQVLHFTK* GHLFLQTDQPIYNPGQR GLEEELQFSLGSK
C5	CO5_HUMAN	Complement component 5	FQNSAILTIQPK TSTSEEVCSEFYLK
C8B	CO8B_HUMAN	Complement component 8, beta	LPLEYSYGEYR
C9	CO9_HUMAN	Complement component 9	TSNFNAAISLK
CFH	CFAH_HUMAN	Complement factor H	AGEQVTYTCATYYK CTLKPCDYPDIK SSIDIENGFISESQYTYALK
CFHR4	FHR4_HUMAN	Complement factor H-related 4	VEYQCQSYELQGSK YVTCSNGDWSEPPR
CFP	PROP_HUMAN	Complement factor properdin	SISCQEIPGQQSR*
CLU	CLUS_HUMAN	Clusterin	ELDESLQVAER LFDSDPITVTVPVEVSR TLLSNLEEAK
CNDP1	CNDP1_HUMAN	Carnosine dipeptidase 1	ALEQDLPVNIK* EWVAIESDSVQPVR*

COL7A1	CO7A1_HUMAN	Collagen, type VII, alpha 1	VLCDDVICDETK
CP	CERU_HUMAN	Ceruloplasmin	HYIYIGIETTWDYASDHGEK QSEDSTFYLGERR
CPN1	CBPN_HUMAN	Carboxypeptidase N, polypeptide 1	IVQLIQDTR
CRTAC1	CRAC1_HUMAN	Cartilage acidic protein 1	GVASLFAGR
F5	FA5_HUMAN	Coagulation factor V	AADIEQQAVFAVFDENK EFNPLVIVGLSK
EFEMP1	FBLN3_HUMAN	EGF containing fibulin-like ECM protein 1	LNCEDIDECR
FN1	FINC_HUMAN	Fibronectin 1	WCHDNGVNYK WLPSSSPVTGYR VPGTSTSATLTGLTR YQCYCYGR
GAPDHS	G3PT_HUMAN	Glyceraldehyde-3-phosphate dehydrogenase	VPTPDVSVVDLTCR
GSN	GELS_HUMAN	Gelsolin	AGALNSNDAFVLK* QTQVSVLPEGGETPLFK* TGAQELLR*
HGFAC	HGFA_HUMAN	HGF activator	VANYVDWINDR
HPX	HEMO_HUMAN	Hemopexin	NFPSPVDAAFR GECQAEGVLFFQGDR
IGFALS	ALS_HUMAN	Insulin-like growth factor binding protein	NLIAAVAPGAFLGLK LAELPADALGPLQR
ITIH1	ITIH1_HUMAN	Inter-alpha-trypsin inhibitor heavy chain 1	QAVDTAVDGVFIR
ITIH2	ITIH2_HUMAN	Inter-alpha-trypsin inhibitor heavy chain 2	TEVNVLPGAK FLHVPDTEFGHFDGVPVISK
KLKB1	KLKB1_HUMAN	Kallikrein B, plasma	IYSGILNLSDITK EIIHQNYK
KNG1	KNG1_HUMAN	Kininogen 1	AATGECTATVGK YFIDFVAR
LGALS3BP	LG3BP_HUMAN	Galectin-3-binding protein	YSSDYFQAPSDYR
PCSK9	PCSK9_HUMAN	proprotein convertase subtilisin/kexin type 9	LPGTYYVVVLK
PLG	PLMN_HUMAN	Plasminogen	QLGAGSIEECAAK TPENFPCK
PPBP	CXCL7_HUMAN	Pro-platelet basic protein	EESLDSDLYAELR* NIQSLEVIGK*
PROC	PROC_HUMAN	Vitamin K-dependent protein C	TFVLNFIK WELDLDIK
PROZ	PROZ_HUMAN	Vitamin K-dependent protein Z	DFAEHLIPR
QSOX1	QSOX1_HUMAN	Quiescin Q6 sulfhydryl oxidase 1	AHFSPSNIILDFPAAGSAAR
S100A8	S10A8_HUMAN	S100 calcium binding protein A8	LLETECPQYIR
SERPINA3	AACT_HUMAN	Alpha-1-antichymotrypsin	EQLSLDDR ADLSGITGAR

SERPINF2	A2AP_HUMAN	Alpha-2-antiplasmin	LGNQEPGGQTALK*
SERPING1	IC1_HUMAN	Plasma protease C1 inhibitor	LLDSLPSDTR*
TTR	TTHY_HUMAN	Transthyretin	AADDTWEPFASGK* GSPAINVAVHVFR*

* Peptides not selected by the dCV approach during the selection phase in this study, but included based on Zhang et al. [5]

ESM Table 2 – Gene ontology (GO) enrichment of the peptides that discriminated between islet autoantibody-positive and autoantibody-negative children in univariate analyses performed in the application phase. The table lists the pathways that were significantly enriched with five or more input proteins.

GO-Term	GO-Term id	P-value	Adjusted p-value	# Proteins (observed)	# Proteins (expected)	# Proteins (total)	List of observed proteins
Lipoprotein metabolic process	GO:0042157	2.18E-03	0.0E+00	6	1.57	18	APOD, APOA2, ALB, APOM, APOE, APOA4
Metabolic process	GO:0008152	1.39E-03	0.0E+00	26	20.43	235	C3, TTR, QSOX1, FN1, CP, C4A, APOC4, APOD, PROZ, APOA2, KNG1, C5, ITIH1, C8B, EFEMP1, ALB, APOM, ITIH2, APOE, CLU, APOA4, SERPINF2, CPN1, HPX, BTB, C9
Lipid homeostasis	GO:0055088	1.76E-03	1.0E-03	5	1.04	12	APOC4, APOA2, APOM, APOE, APOA4
Organic hydroxy compound transport	GO:0015850	2.18E-03	1.0E-03	6	1.57	18	APOA2, C5, APOM, APOE, CLU, APOA4
Regulation of protein metabolic process	GO:0051246	1.67E-03	1.0E-03	17	9.48	109	C3, FN1, C4A, APOD, APOA2, KNG1, C5, ITIH1, C8B, APOM, ITIH2, APOE, CLU, APOA4, SERPINF2, HPX, C9
Regulation of primary metabolic process	GO:0080090	4.67E-03	1.0E-03	18	11.22	129	C3, FN1, C4A, APOD, APOA2, KNG1, C5, ITIH1, C8B, EFEMP1, APOM, ITIH2, APOE, CLU, APOA4, SERPINF2, HPX, C9
Protein metabolic process	GO:0019538	1.82E-03	1.0E-03	22	14.78	170	C3, QSOX1, FN1, C4A, APOD, PROZ, APOA2, KNG1, C5, ITIH1, C8B, EFEMP1, ALB, APOM, ITIH2, APOE, CLU, APOA4, SERPINF2, CPN1, HPX, C9
Reverse cholesterol transport	GO:0043691	3.36E-04	2.0E-03	5	0.78	9	APOA2, APOM, APOE, CLU, APOA4
Humoral immune response mediated by circulating immunoglobulin	GO:0002455	2.09E-03	2.0E-03	7	2.09	24	C3, C4A, C5, C8B, CLU, HPX, C9
Lipid localization	GO:0010876	1.77E-03	2.0E-03	8	2.61	30	C3, APOC4, APOD, APOA2, APOM, APOE, CLU, APOA4
Lipid metabolic process	GO:0006629	2.62E-03	2.0E-03	9	3.39	39	C3, TTR, APOC4, APOD, APOA2, ALB, APOE, CLU, APOA4
Organic substance metabolic process	GO:0071704	2.13E-03	2.0E-03	25	19.04	219	C3, TTR, QSOX1, FN1, C4A, APOC4, APOD, PROZ, APOA2, KNG1, C5, ITIH1, C8B, EFEMP1, ALB, APOM, ITIH2, APOE, CLU, APOA4, SERPINF2, CPN1, HPX, BTB, C9
Cholesterol transport	GO:0030301	5.53E-03	3.0E-03	5	1.30	15	APOA2, APOM, APOE, CLU, APOA4
Regulation of protein processing	GO:0070613	2.77E-03	3.0E-03	12	5.57	64	C3, FN1, C4A, KNG1, C5, ITIH1, C8B, ITIH2, APOE, CLU, SERPINF2, C9
Regulation of humoral immune response	GO:0002920	5.30E-03	4.0E-03	6	1.83	21	C3, C4A, C5, C8B, HPX, C9
Chemical homeostasis	GO:0048878	3.19E-03	4.0E-03	9	3.48	40	CP, APOC4, APOA2, KNG1, C5, APOM, APOE, APOA4, HPX

Negative regulation of protein metabolic process	GO:0051248	2.77E-03	4.0E-03	12	5.57	64	C3, C4A, APOD, KNG1, C5, ITIH1, APOM, ITIH2, APOE, CLU, APOA4, SERPINF2
Protein processing	GO:0016485	5.66E-03	4.0E-03	12	6.00	69	C3, FN1, C4A, KNG1, C5, ITIH1, C8B, ITIH2, APOE, CLU, SERPINF2, C9
Regulation of proteolysis	GO:0030162	3.85E-03	5.0E-03	10	4.26	49	C3, FN1, C4A, KNG1, C5, ITIH1, ITIH2, APOE, CLU, SERPINF2
Protein maturation	GO:0051604	6.46E-03	5.0E-03	12	6.09	70	C3, FN1, C4A, KNG1, C5, ITIH1, C8B, ITIH2, APOE, CLU, SERPINF2, C9
Lipid transport	GO:0006869	5.61E-03	6.0E-03	7	2.43	28	APOC4, APOD, APOA2, APOM, APOE, CLU, APOA4
Organism metabolic process	GO:0044710	8.60E-03	6.0E-03	20	13.83	159	C3, TTR, QSOX1, CP, C4A, APOC4, APOD, APOA2, KNG1, C5, C8B, ALB, APOM, APOE, CLU, APOA4, SERPINF2, HPX, BTD, C9
Primary metabolic process	GO:0044238	5.40E-03	6.0E-03	24	18.26	210	C3, TTR, QSOX1, FN1, C4A, APOC4, APOD, PROZ, APOA2, KNG1, C5, ITIH1, C8B, EFEMP1, ALB, APOM, ITIH2, APOE, CLU, APOA4, SERPINF2, CPN1, HPX, C9
Negative regulation of cellular metabolic process	GO:0031324	6.46E-03	7.0E-03	12	6.09	70	C3, C4A, APOD, KNG1, C5, ITIH1, APOM, ITIH2, APOE, CLU, APOA4, SERPINF2
Homeostatic process	GO:0042592	4.89E-03	8.0E-03	11	5.13	59	QSOX1, CP, APOC4, APOA2, KNG1, C5, ALB, APOM, APOE, APOA4, HPX
Inflammatory response	GO:0006954	8.59E-03	8.0E-03	11	5.48	63	C3, FN1, C4A, APOD, APOA2, KNG1, C5, C8B, APOE, SERPINF2, C9
Negative regulation of metabolic process	GO:0009892	7.82E-03	8.0E-03	13	7.04	81	C3, C4A, APOD, APOA2, KNG1, C5, ITIH1, APOM, ITIH2, APOE, CLU, APOA4, SERPINF2
Negative regulation of macromolecule metabolic process	GO:0010605	8.37E-03	9.0E-03	12	6.26	72	C3, C4A, APOD, KNG1, C5, ITIH1, APOM, ITIH2, APOE, CLU, APOA4, SERPINF2
Cellular metabolic process	GO:0044237	6.96E-03	9.0E-03	22	15.91	183	C3, TTR, QSOX1, FN1, C4A, APOD, PROZ, APOA2, KNG1, C5, ITIH1, EFEMP1, ALB, APOM, ITIH2, APOE, CLU, APOA4, SERPINF2, CPN1, HPX, BTD
Complement activation, classical pathway	GO:0006958	8.72E-03	1.0E-02	6	2.00	23	C3, C4A, C5, C8B, CLU, C9
Adaptive immune response	GO:0002250	8.54E-03	1.0E-02	7	2.61	30	C3, C4A, C5, C8B, CLU, HPX, C9

ESM Table 3 - Characteristics of children stratified into low, medium, and high risk score groups according to tertiles of risk score for progression from islet autoantibody seroconversion to clinical type 1 diabetes.

	Low risk score	Medium risk score	High risk score
Children; n	24	23	23
Risk score ^a ; median (range)	-0.744 (-1.942, -0.339)	-0.016 (-0.319, 0.390)	0.781 (0.395, 2.100)
5-year diabetes risk ^b ; % (95% CI)	8.3 (2.2, 29.4)*	34.8 (19.2, 57.7)*	78.3 (60.1, 92.1)*
Age, years; median (IQR) ^c	5.7 (3.7, 7.1)**	3.1 (2.5, 3.8)**	2.5 (1.7, 3.0)**
HLA genotype; n (%)			
<i>DR3/DR4-DQ8</i>	5 (20.8)	4 (17.4)	5 (26.3)
<i>DR4-DQ8/DR4-DQ8</i>		5 (21.7)	4 (21.1)
<i>DR3/DR3</i>		2 (8.7)	2 (10.5)
<i>DR4-DQ8/x^d</i>	11 (45.8)	8 (34.8)	5 (26.3)
<i>DR3/x^d</i>	4 (16.7)	2 (8.7)	2 (10.5)
<i>x/x^d</i>	4 (16.7)	2 (8.7)	1 (5.3)
Islet autoantibodies; n (%) ^c			
1 antibody positive	3 (12.5)	3 (13.0)	4 (17.4)
2 antibodies positive	14 (58.3)	3 (13.0)	8 (34.8)
3 or 4 antibodies positive	7 (29.2)	17 (74.0)	11 (47.8)
IAA positive	18 (75.0)	20 (87.0)	20 (87.0)
GADA positive	18 (75.0)	16 (69.6)	16 (69.6)
IA-2A positive	9 (37.5)	15 (65.2)	10 (43.5)
ZnT8A positive	14 (58.3)	14 (63.6) ^e	11 (57.9) ^f

^a Risk score based on concentrations of 3 peptides representing Hepatocyte growth factor activator, Complement factor H and Ceruloplasmin, and age

^b Cumulative progression to clinical diabetes within 5 years from autoantibody seroconversion

^c Status at time of sample used for peptide measurement

^d *x*, non-*DR3* and non-*DR4-DQ8*

^e ZnT8A status not available for 1 child

^f ZnT8A status not available for 4 children

* Likelihood-ratio test, $P < 0.0001$

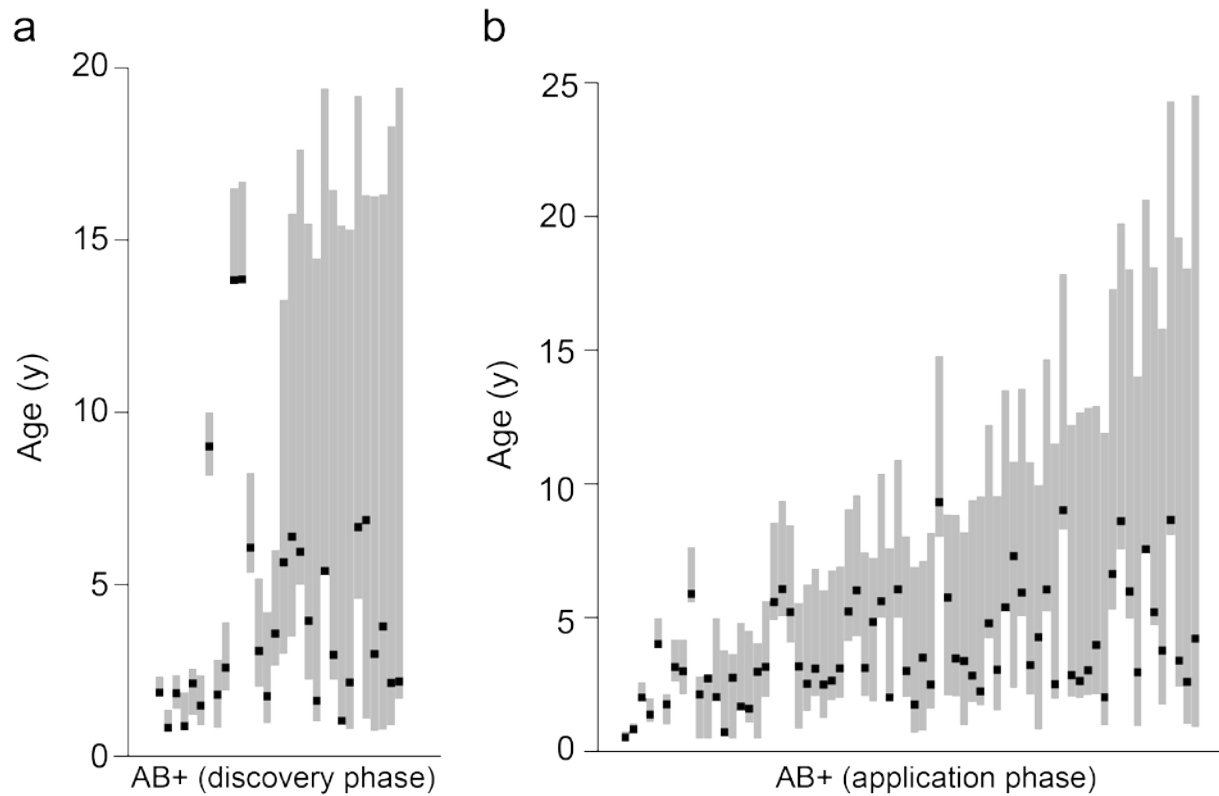
** Kruskal-Wallis test, $P < 0.0001$

ESM References

- [1] Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Mach Learn* 46: 389-422
- [2] Simon N, Friedman JH, Hastie T, Tibshirani R (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *2011* 39: 13
- [3] Laimighofer M, Krumsiek J, Buettner F, Theis FJ (2016) Unbiased Prediction and Feature Selection in High-Dimensional Survival Regression. *Journal of Computational Biology*
- [4] Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ (2011) On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine* 30: 1105-1117
- [5] Zhang Q, Fillmore TL, Schepmoes AA, et al (2013) Serum proteomics reveals systemic dysregulation of innate immunity in type 1 diabetes. *J Exp Med* 210: 191-203

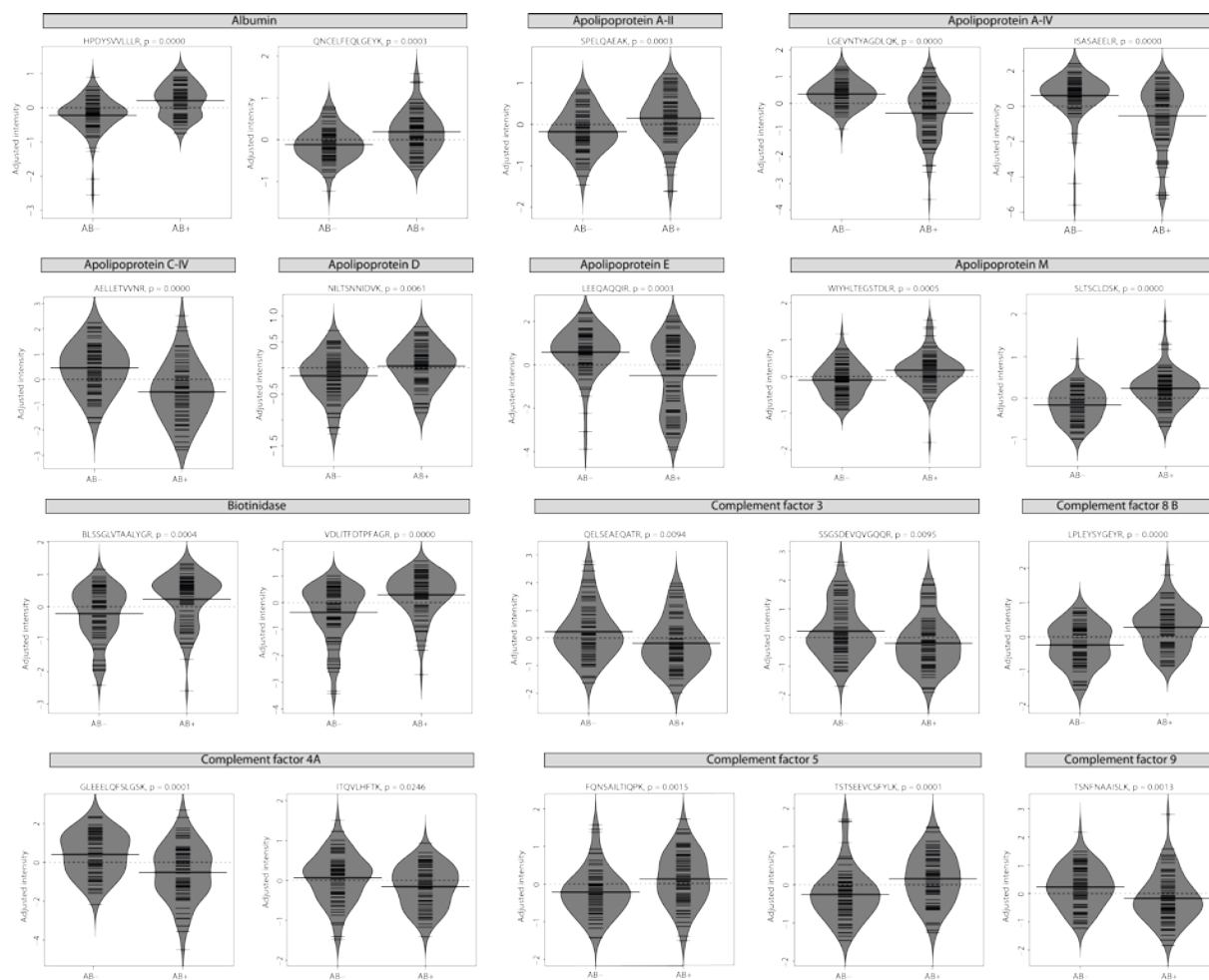
ESM Figures

ESM Figure 1 – Age in years (y) of seroconversion (lower boundary of grey bars), age in years (y) at sample for set 1 in selection phase (black square), and age of type 1 diabetes diagnosis or last follow up (upper boundary of grey bars) of autoantibody-positive (AB+) children. a: discovery phase (n=30). b: application phase (n=70).

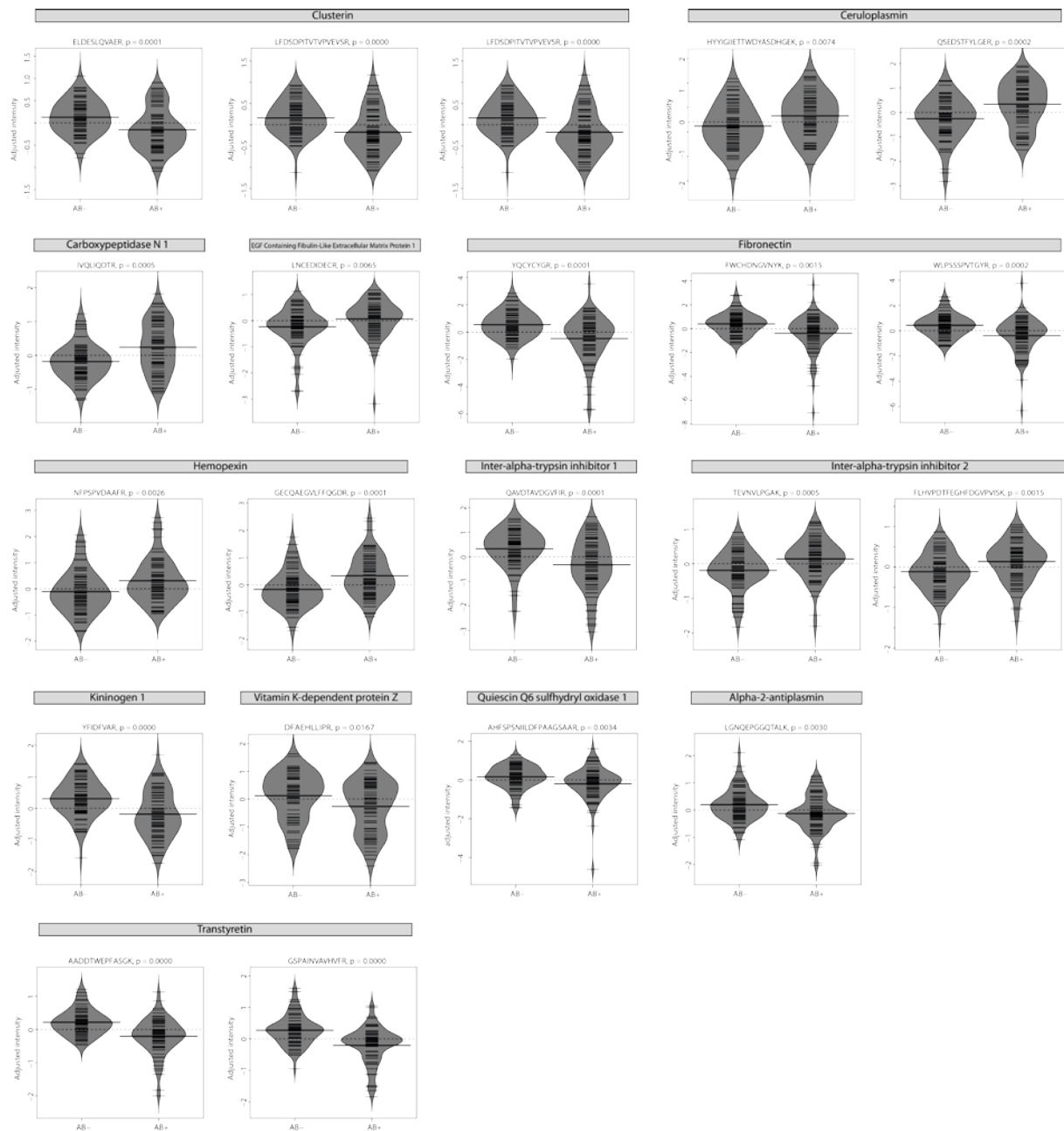


ESM Figure 2 – Violin plots of the adjusted SRM measurement intensities of the 41 peptides representing 26 proteins that were significantly different between islet autoantibody-positive (AB+) and autoantibody-negative children (AB-) according to univariate Wilcoxon rank sum tests performed in the application phase.

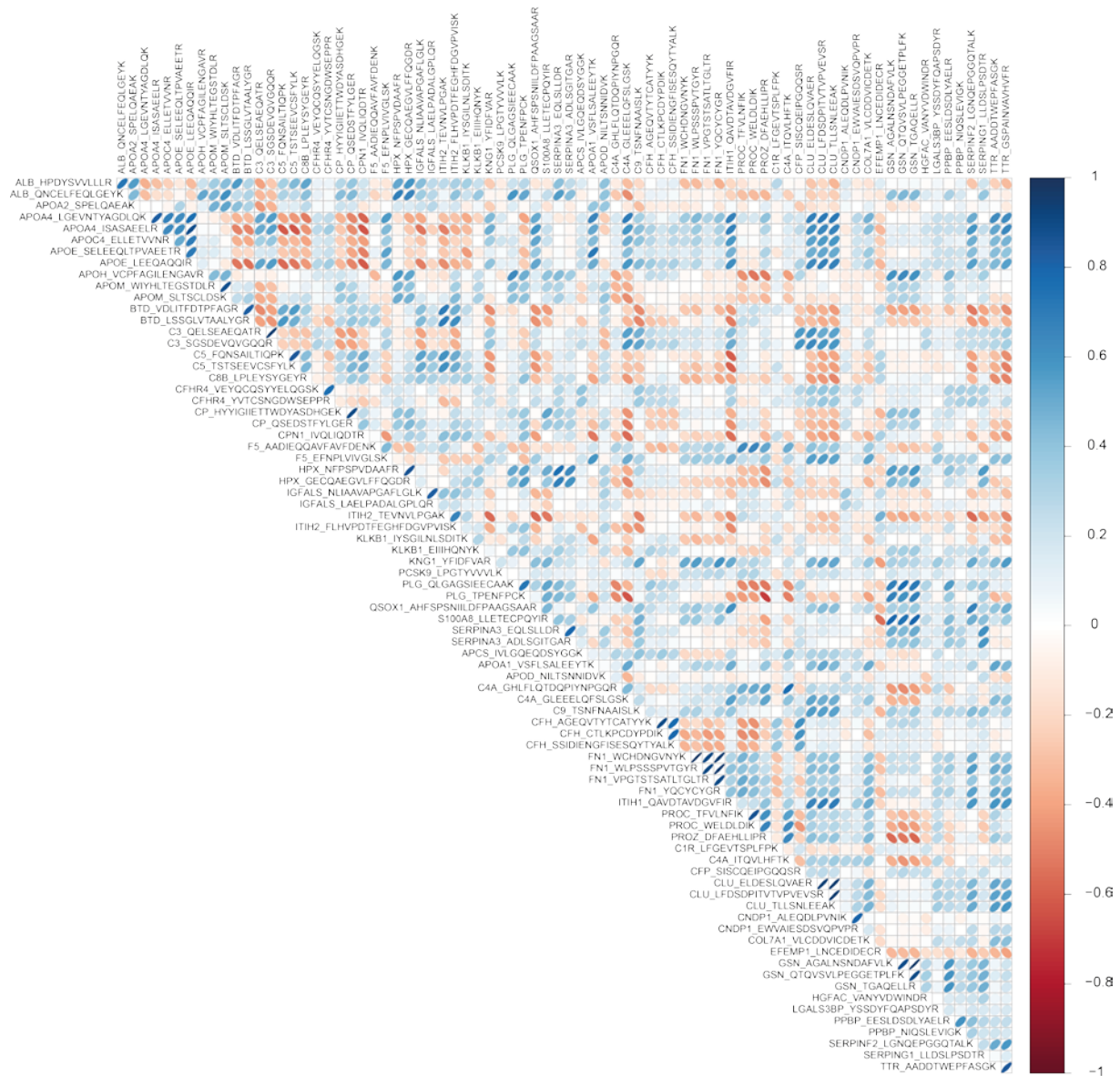
Part1



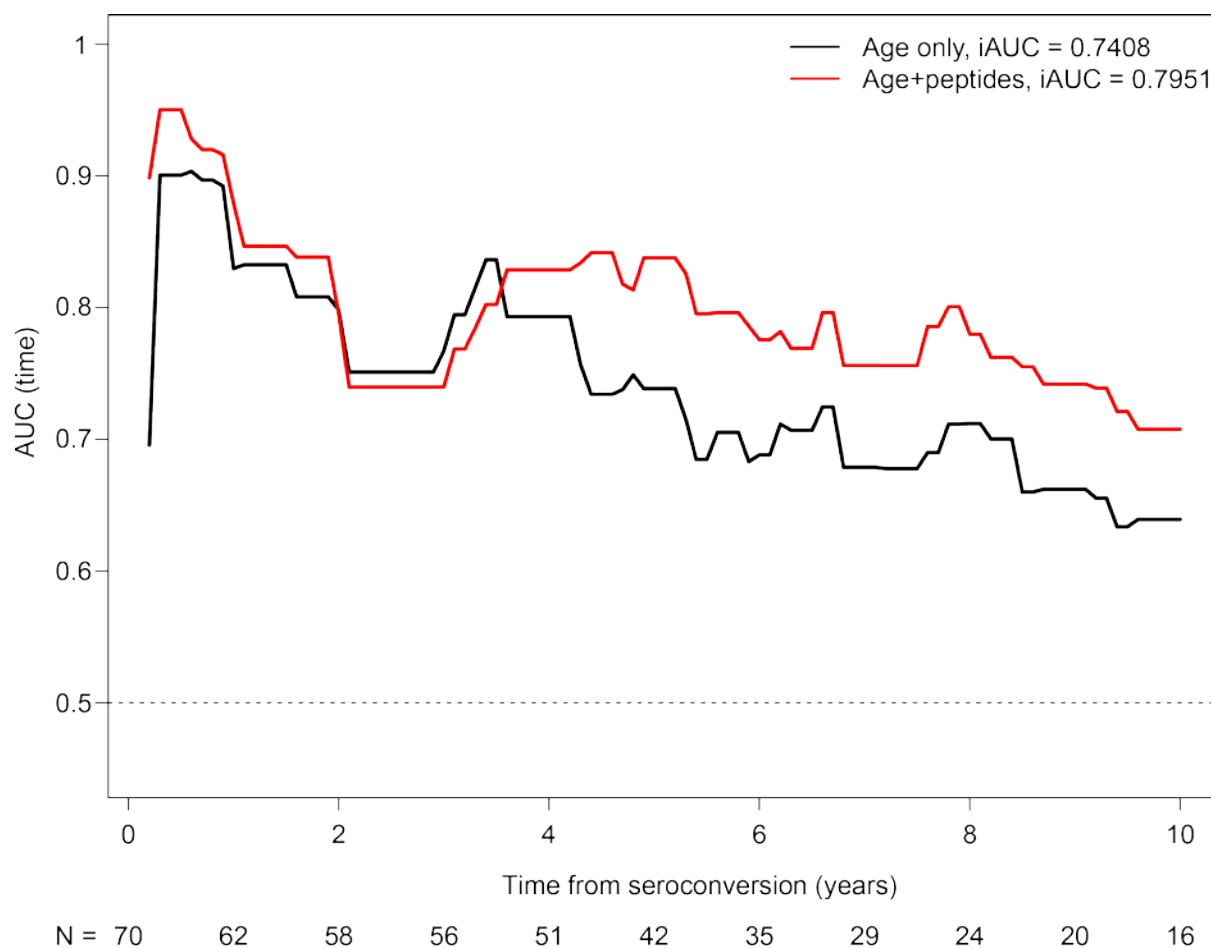
Part 2



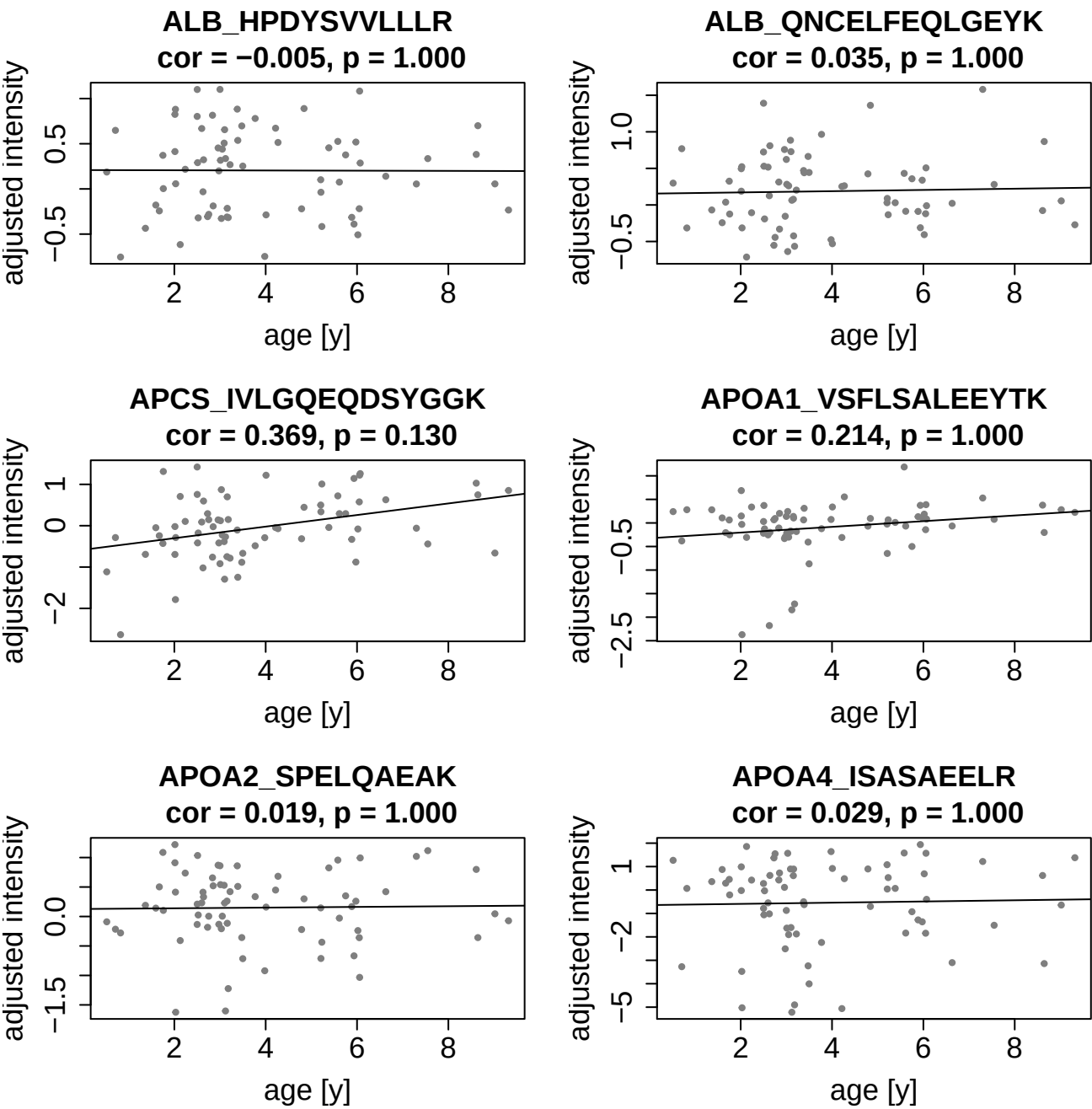
ESM Figure 3 - Pearson's correlation plot of peptides discriminating between islet autoantibody-positive and autoantibody-negative children in the application phase. Correlations are displayed as ellipses, depending on the strength of correlation. Circles indicate no correlation and very sharp ellipses indicate high correlations between the two peptides. Blue and red shapes indicate positive and negative correlations, respectively.

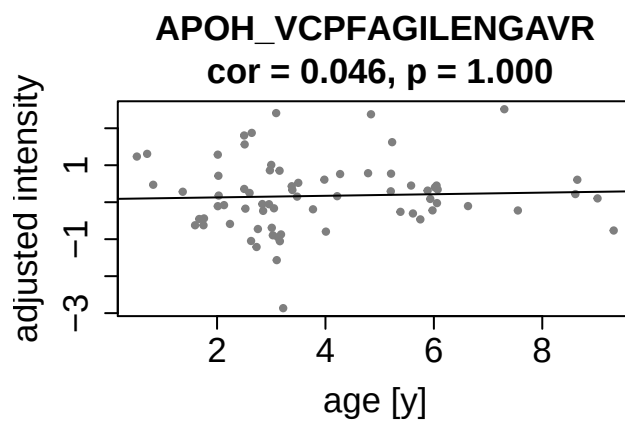
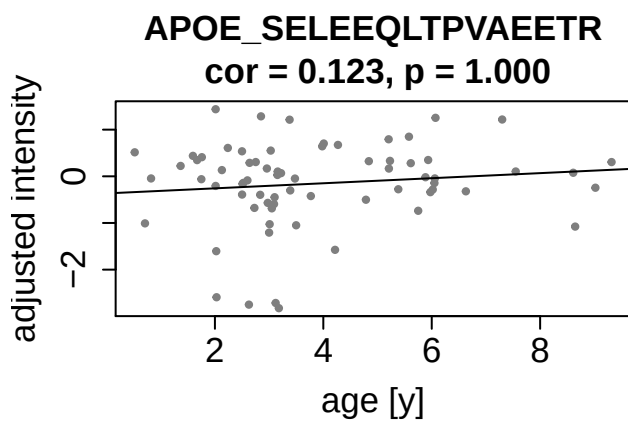
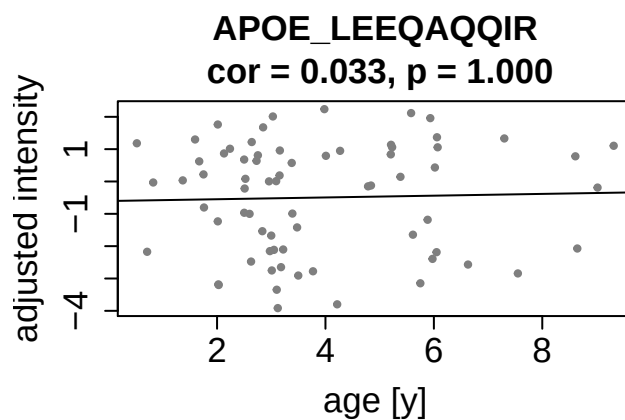
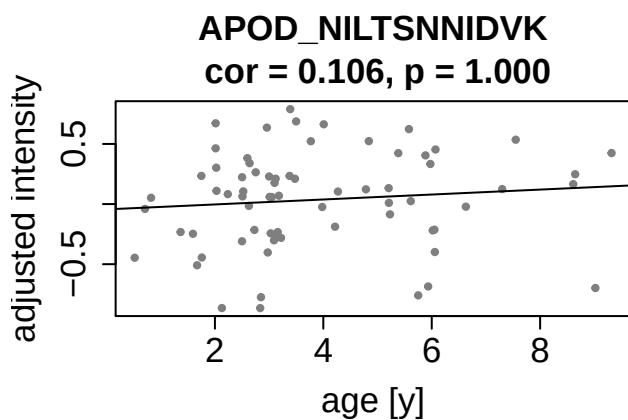
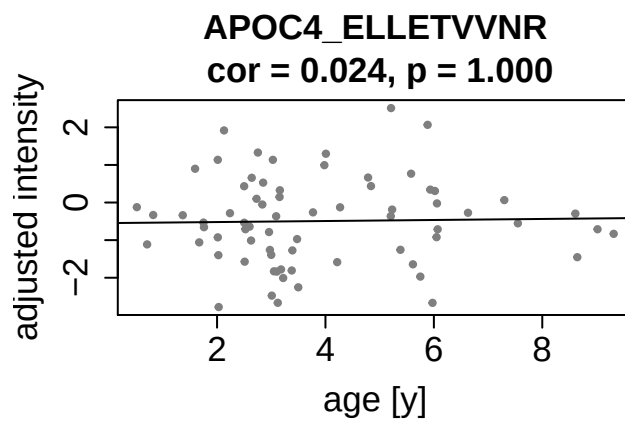
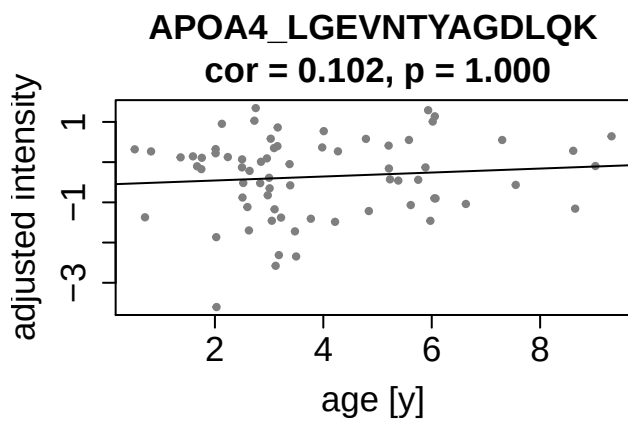


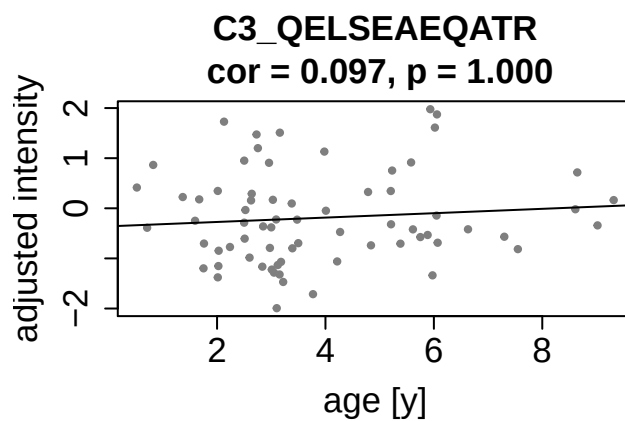
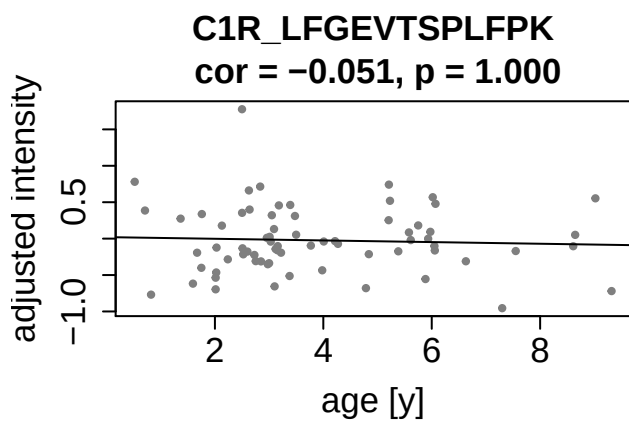
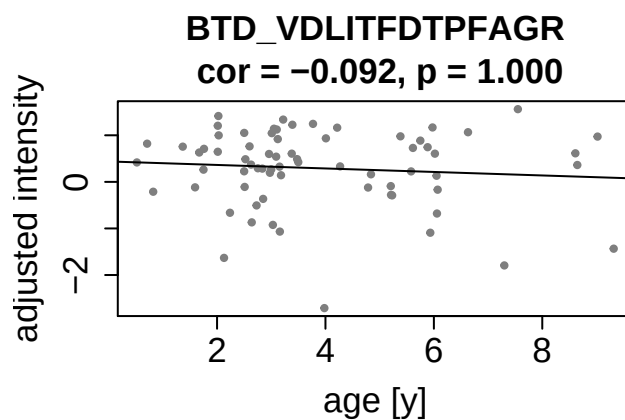
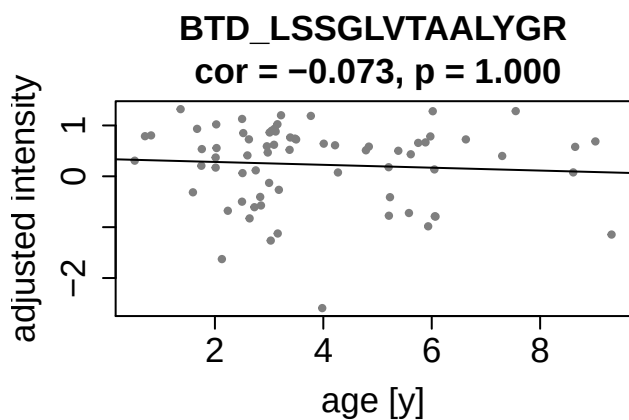
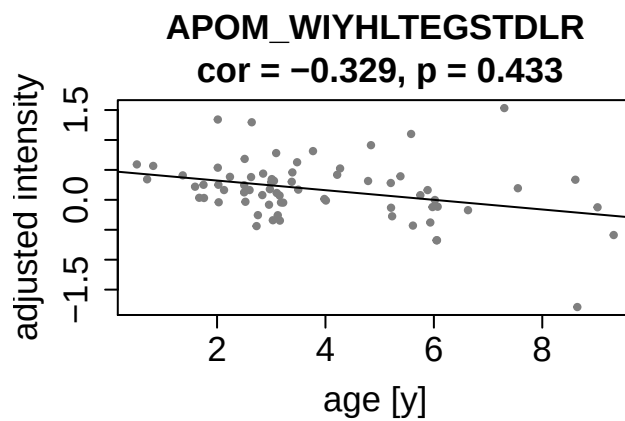
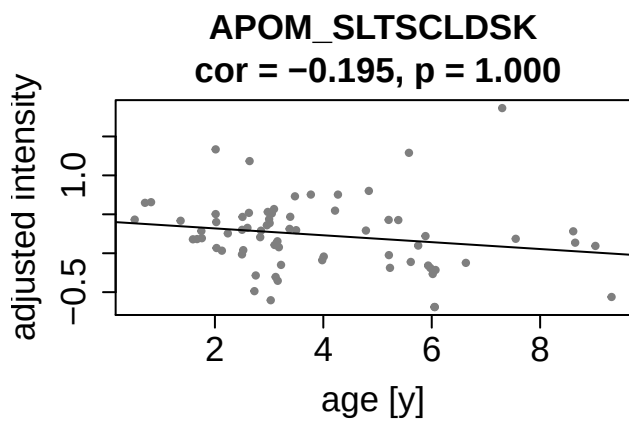
ESM Figure 4 - Comparison of the discrimination performance (measured as survival AUC) of the progression score over time with a score consisting only of age. The red line indicates the survival AUC of the combined score of Hepatocyte growth factor activator, Complement factor H, and Ceruloplasmin, and age, whereas the black line shows the survival AUC of the score of age alone over time. The integrated AUC (iAUC) over time indicates the integral over the survival AUC curve in time up to 10 years. The number of children still at risk is shown below the time axis.

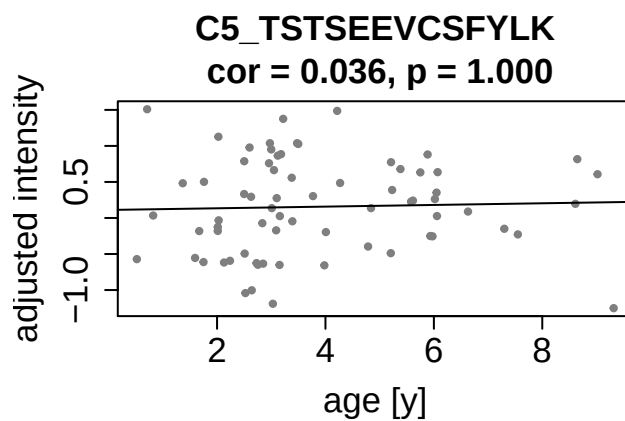
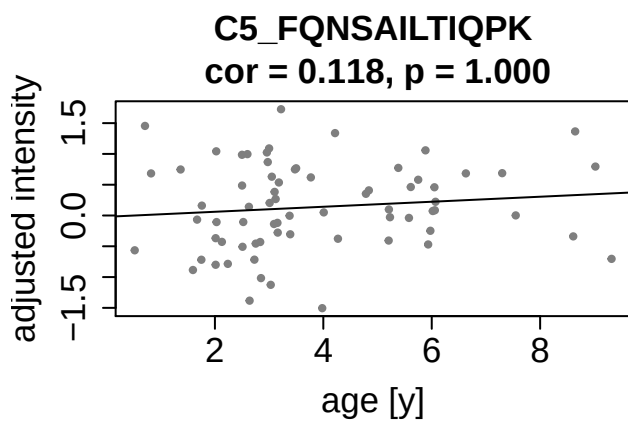
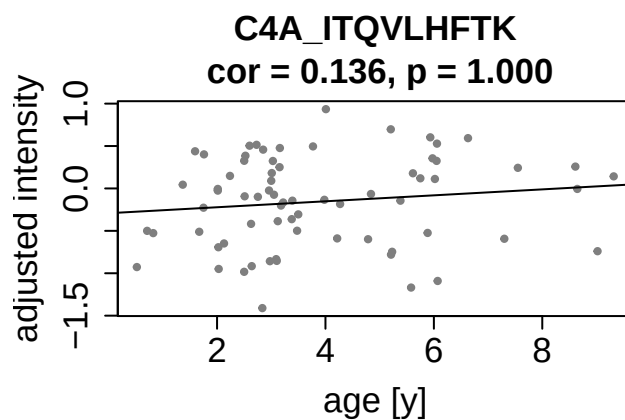
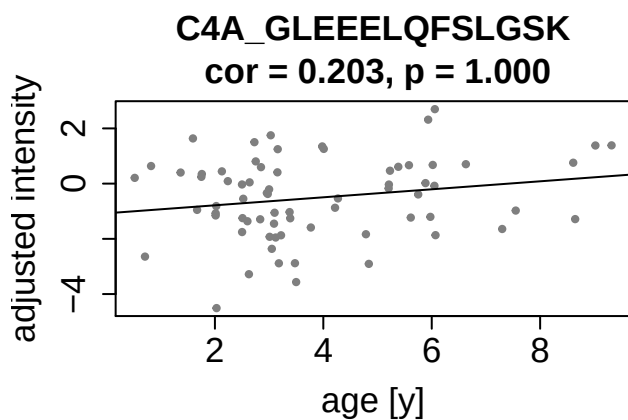
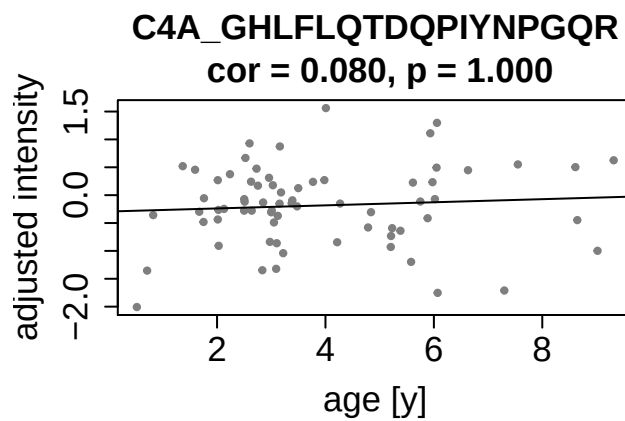
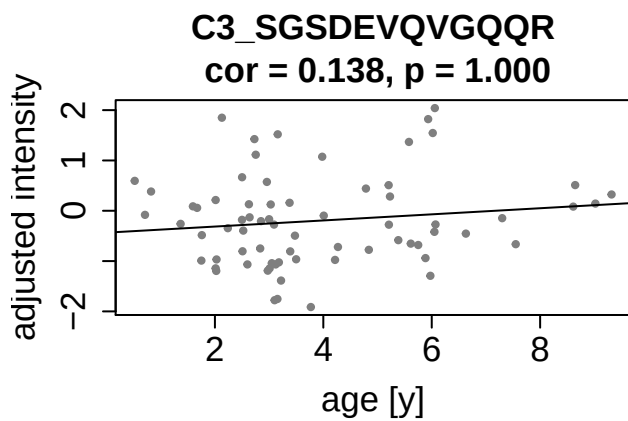


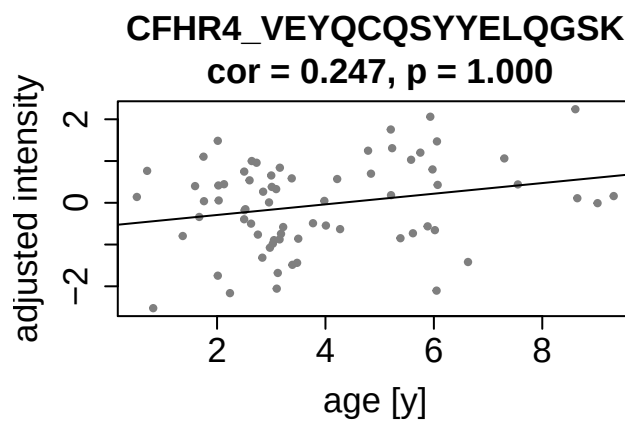
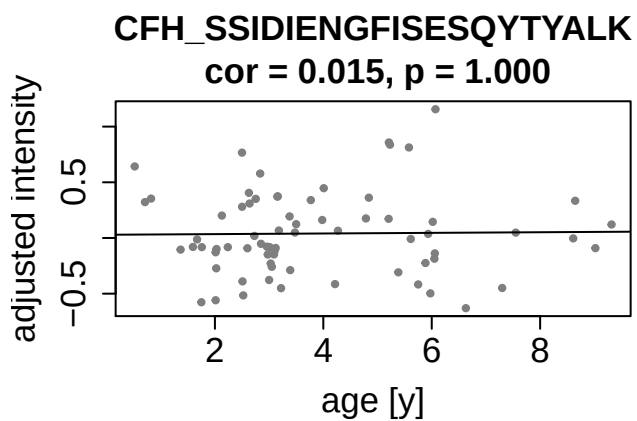
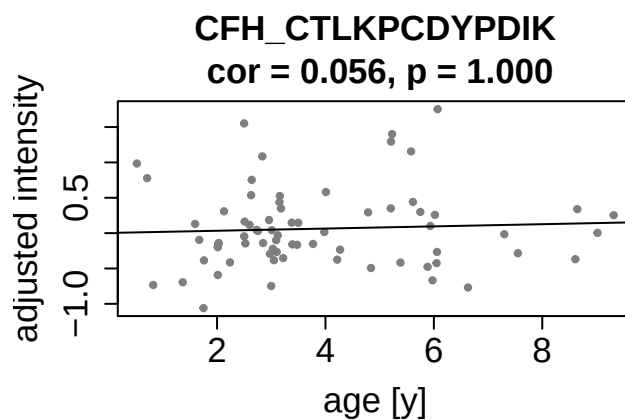
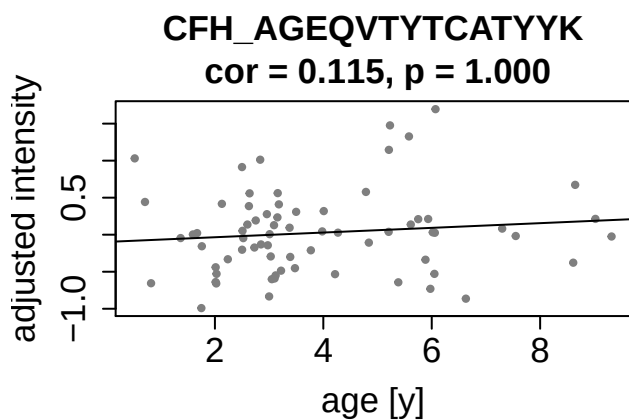
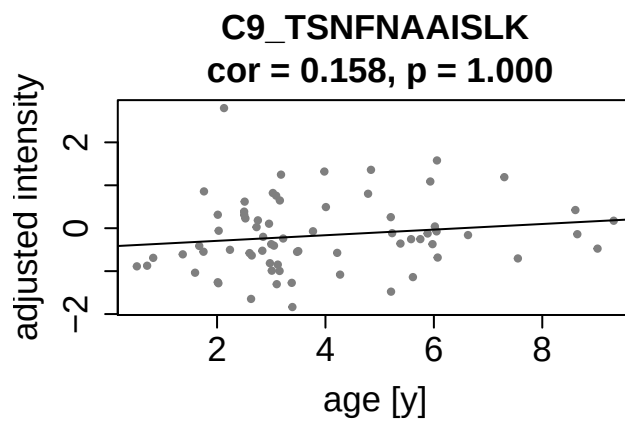
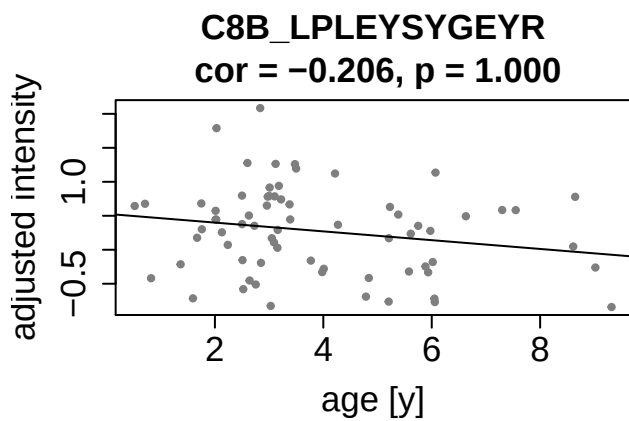
ESM Figure 5 - Correlations between peptide intensities and age. Cor indicates the Pearson's correlation coefficient of every peptide of every protein and p the corresponding p-value of the two sided test statistics.

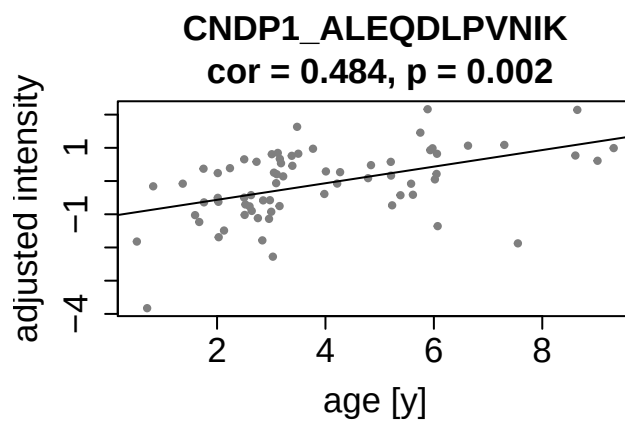
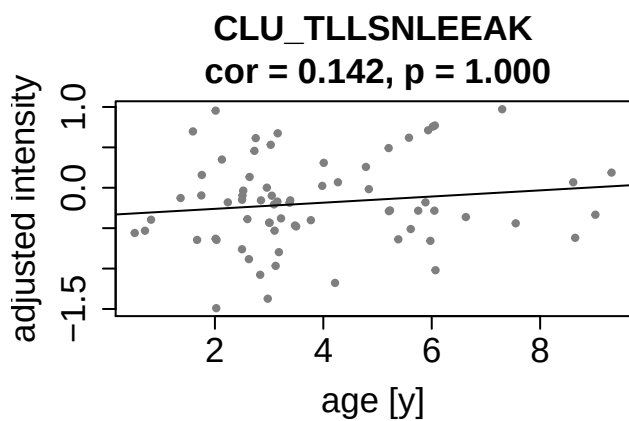
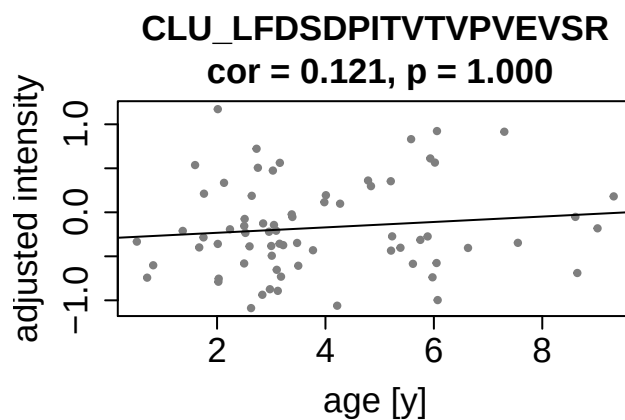
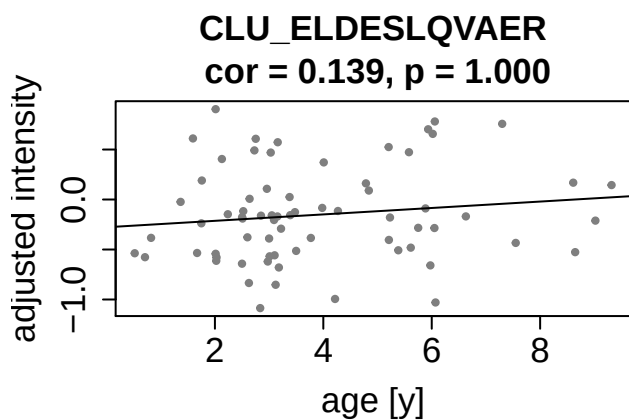
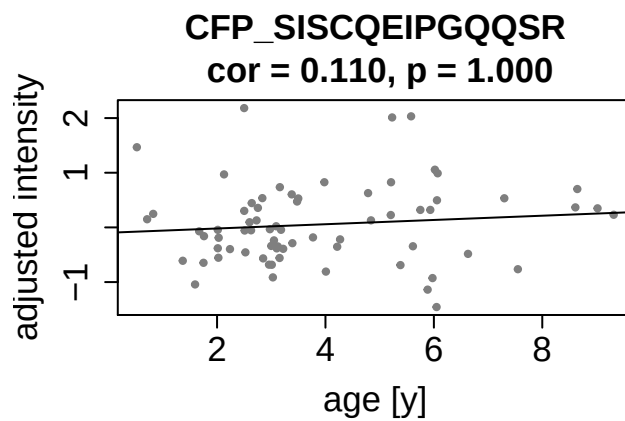
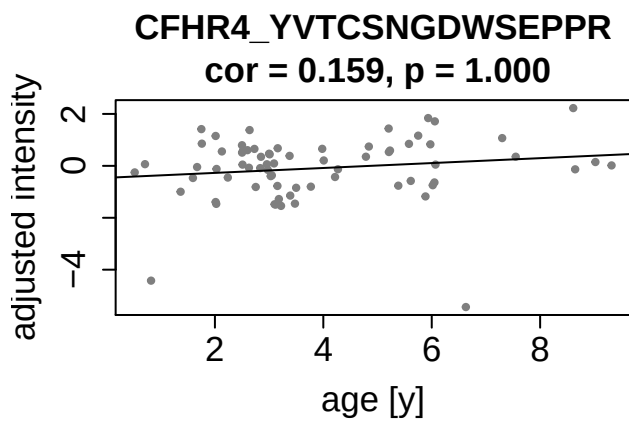


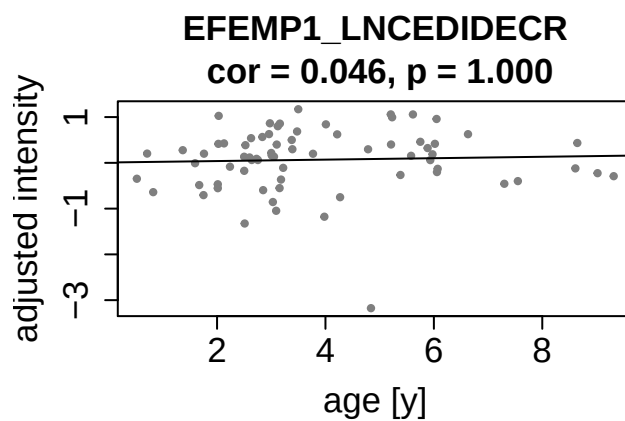
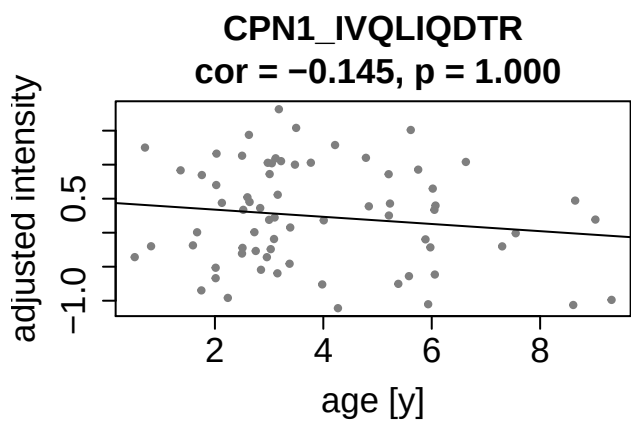
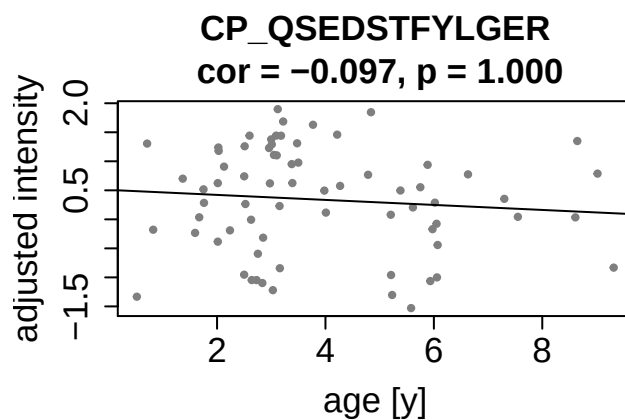
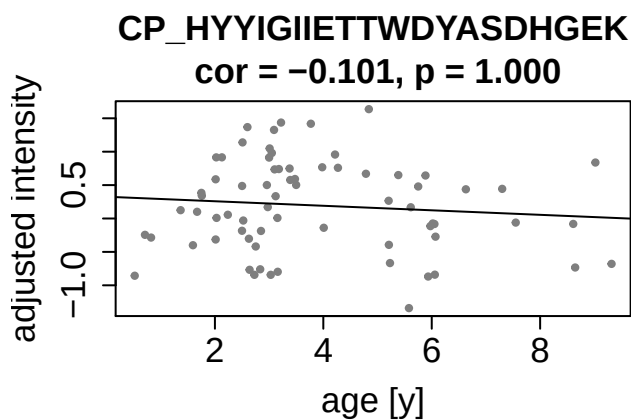
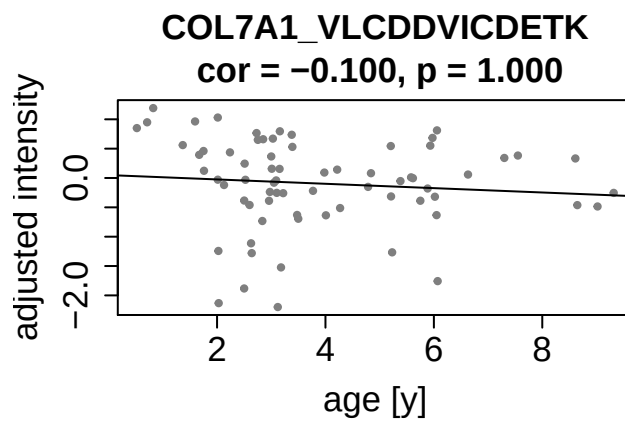
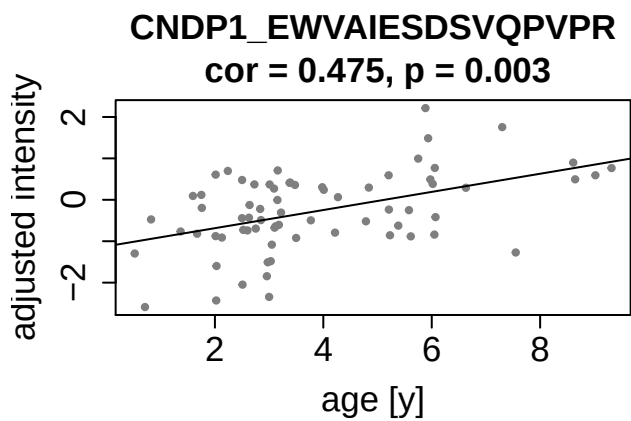




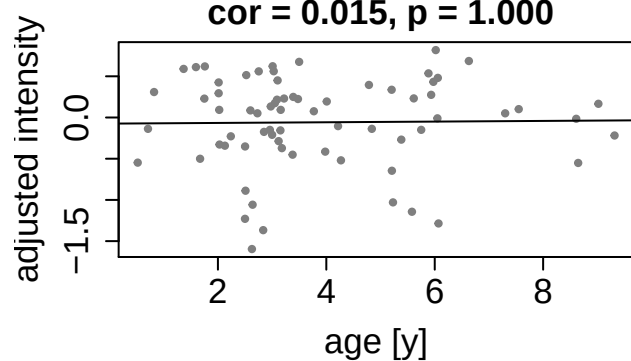




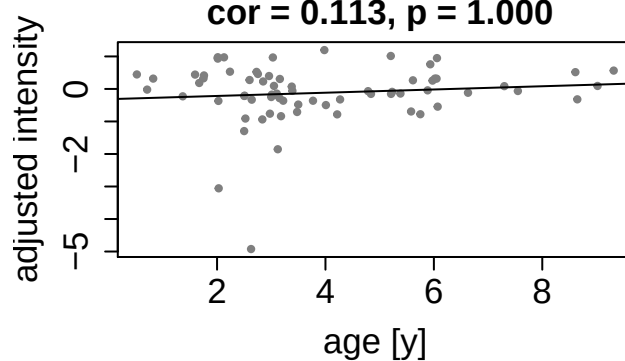




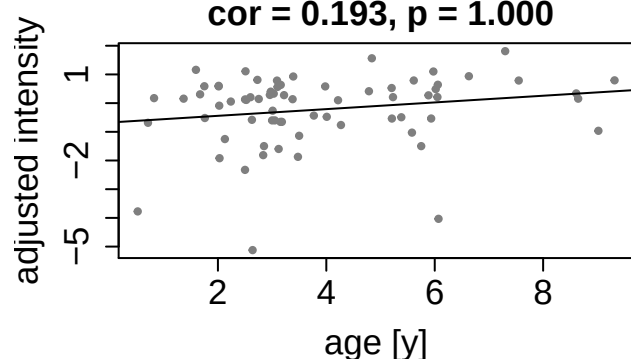
F5_AADIEQQAVFAVFDENK
cor = 0.015, p = 1.000



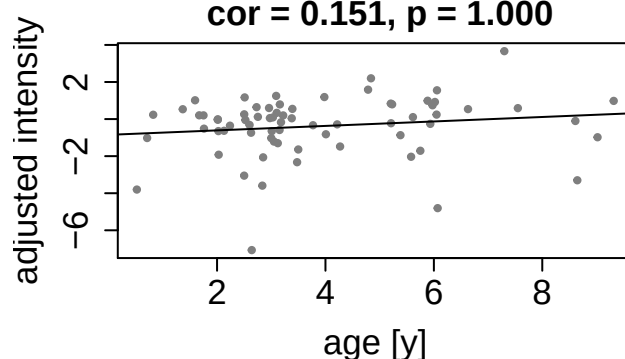
F5_EFNPLVIVGLSK
cor = 0.113, p = 1.000



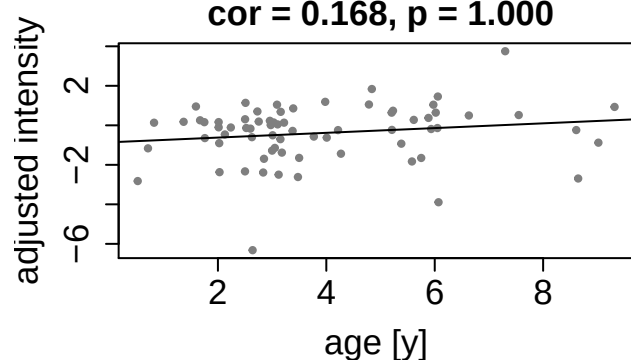
FN1_VPGTSTSATLTGLTR
cor = 0.193, p = 1.000



FN1_WCHDNGVNYK
cor = 0.151, p = 1.000



FN1_WLPSSSPVTGYR
cor = 0.168, p = 1.000



FN1_YQCYCYGR
cor = 0.189, p = 1.000

