PLS-models in Practice: sparse and sparse group extensions

Lecture 3/3

MSc Health Data Analytics – Computational epidemiology – February 11, 2021

Marc Chadeau-Hyam

m.chadeau@imperial.ac.uk

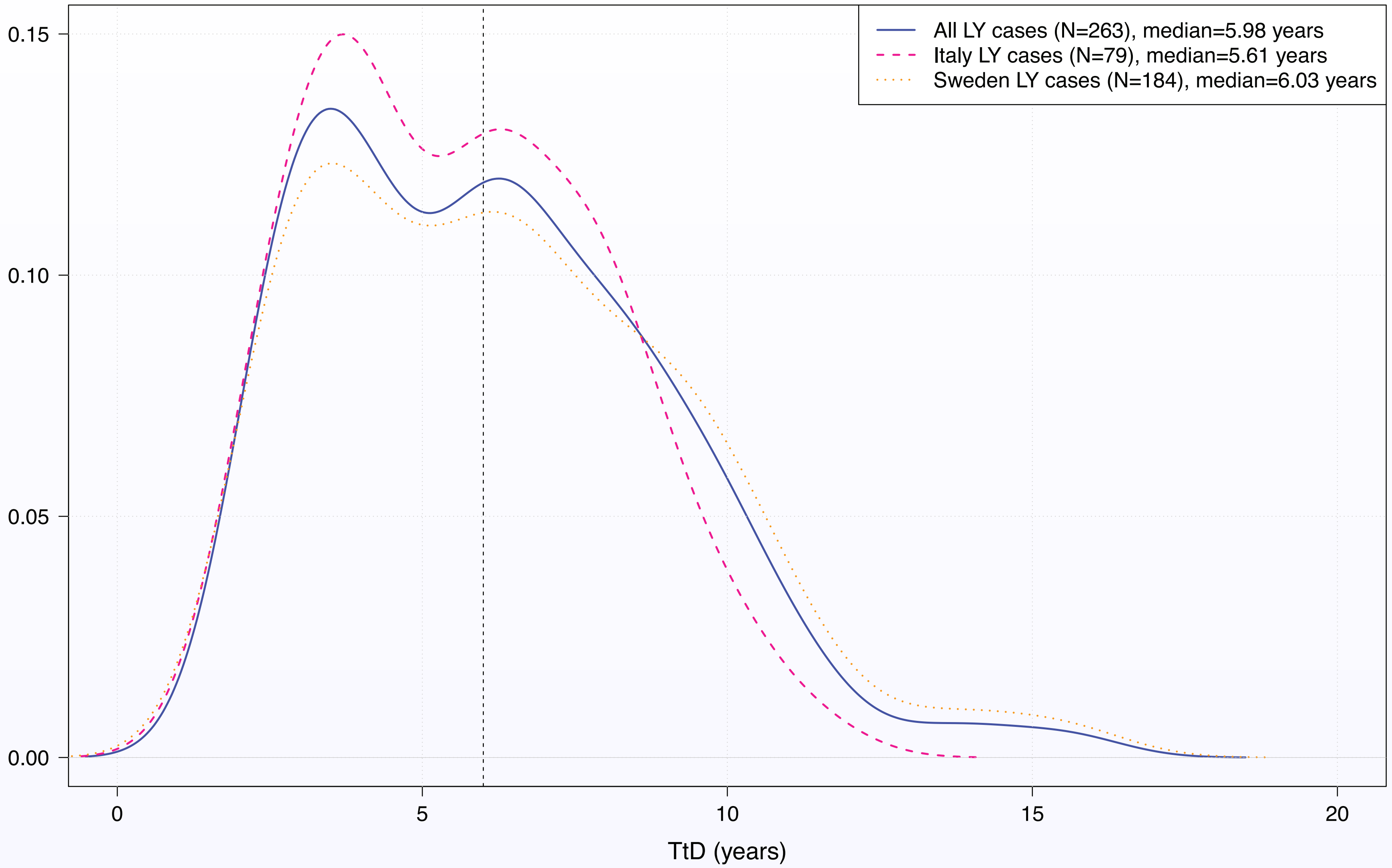
Imperial College London

Lymphoma cases by subtypes and TtD

- EnviroGenoMarkers: a multi-OMIc study of NHL
  - Two contributing cohorts: EPIC Italy, and NHSDS
  - Transcriptomics, Proteomic (N=28) data available
- Four subtypes were identified:
  - B-cell Chronic Lymphatic Leukemia (BCLL): 14.8%
  - Diffuse Large B-cell Lymphoma (DLBCL): 15.6%
  - Follicular Lymphoma (FL): 14.4%
  - Multiple Myeloma (MM): 27.4%
- Study population:

Subtype	TtD<6	TtD>6	**Total**
BCLL	15	24	**39**
DLBCL	18	23	**41**
FL	18	20	**38**
MM	42	30	**72**
Others	41	32	**73**
Total	**93**	**97**	**263**

Time to Diagnosis (TtD) distribution in LY cases



⇒ TtD=6 years is close to the median value

Benchmark screening model

- Univariate exploration of OMICs data accounting for nuisance variation
- Formulation, for individual i:
  - Variable of interest: X^i (Ca/Co)
  - Predictors: Y^i, Expression levels
  - Fixed effects: FE^i
  - Random Effect variables: u^{A^i}, where A^i are nuisance variables

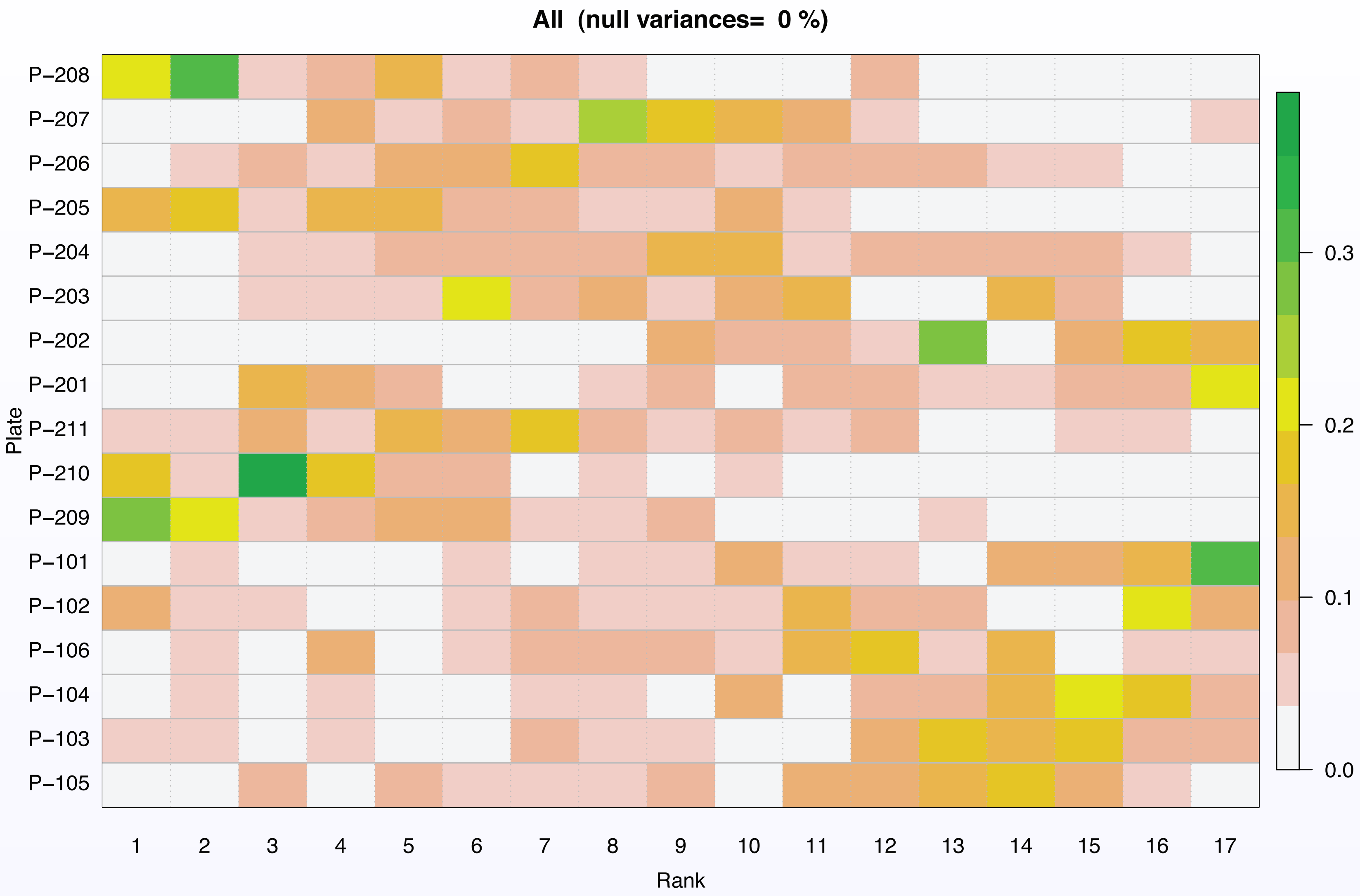
$$Y^i \sim \alpha + \beta_1 X^i + \beta_2 FE^i + u^{A^i} + \epsilon^i$$

⇒ random intercept model

- Methodology: likelihood ratio test
  - Run the model with and without the variable of interest (X^i). Compare both models

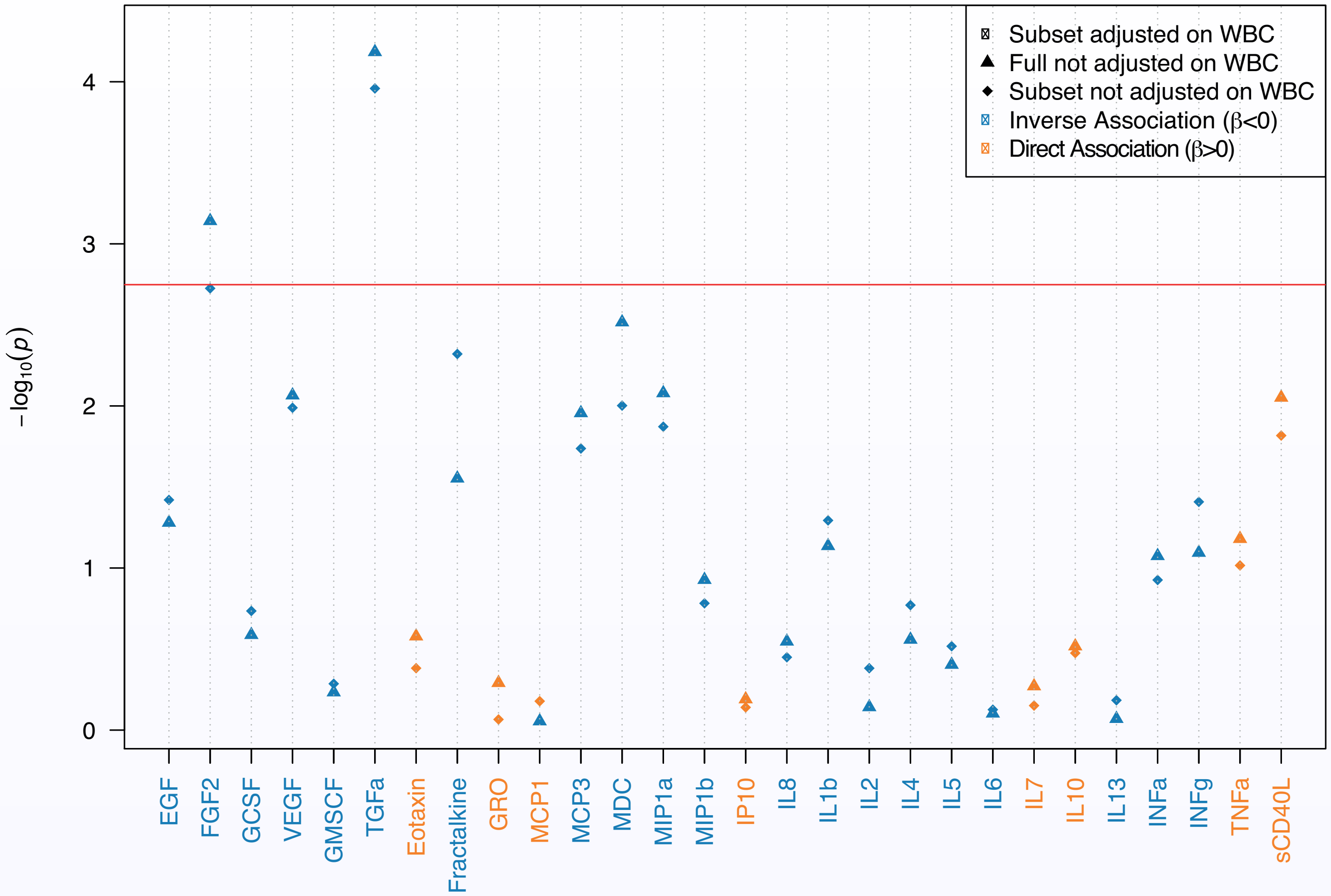
⇒ for each protein/probe we obtain a p-value testing the association between the probe and the disease status/or exposure

Proteomics: Estimation of the random effects



⇒ P-208/209/210 lead to higher variances

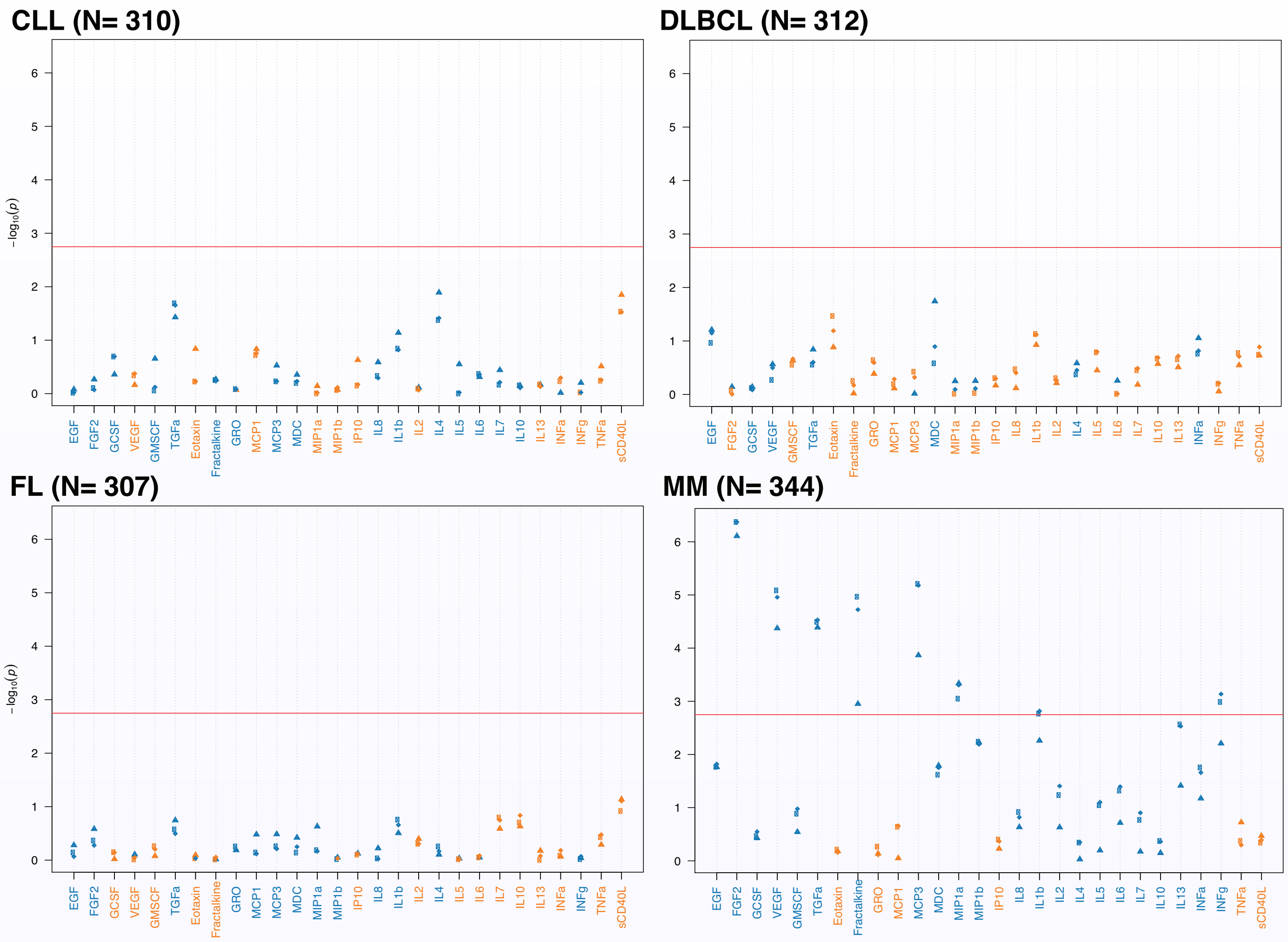
Analysis of all BCL cases



⇒ Two Bonferroni significant associations involving FGF2 & TGFα

⇒ weak effect of WBC adjustments

Histological subtype analyses


CLL (N= 310)

DLBCL (N= 312)

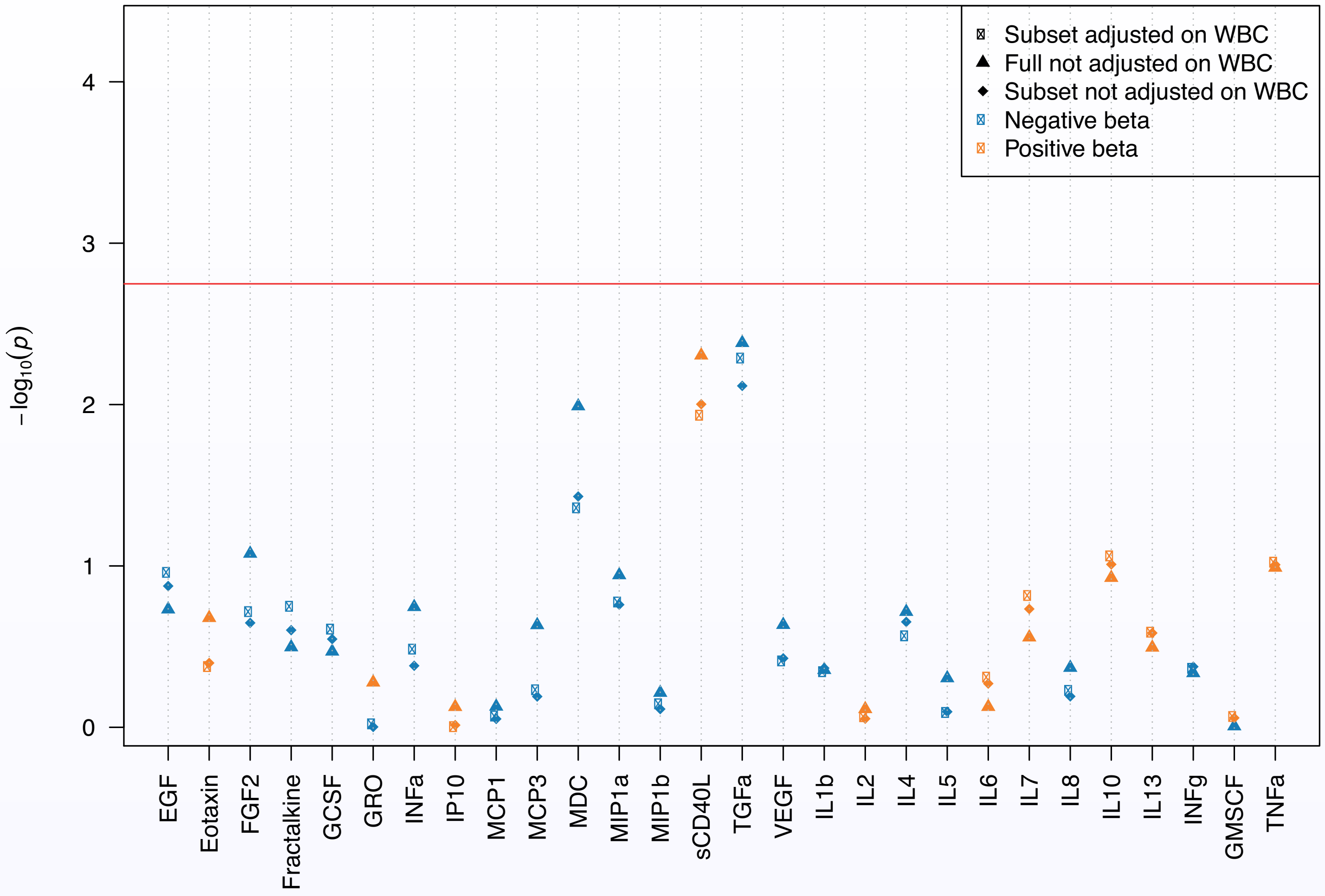
FL (N= 307)

MM (N= 344)

⇒ 8 (strong) and inverse associations for MM

⇒ no association for the other subtypes

All BCL excluding MM



⇒ Both BCL-related associations lose significance upon exclusion of MM cases

⇒ MM may have driven the BCL associations

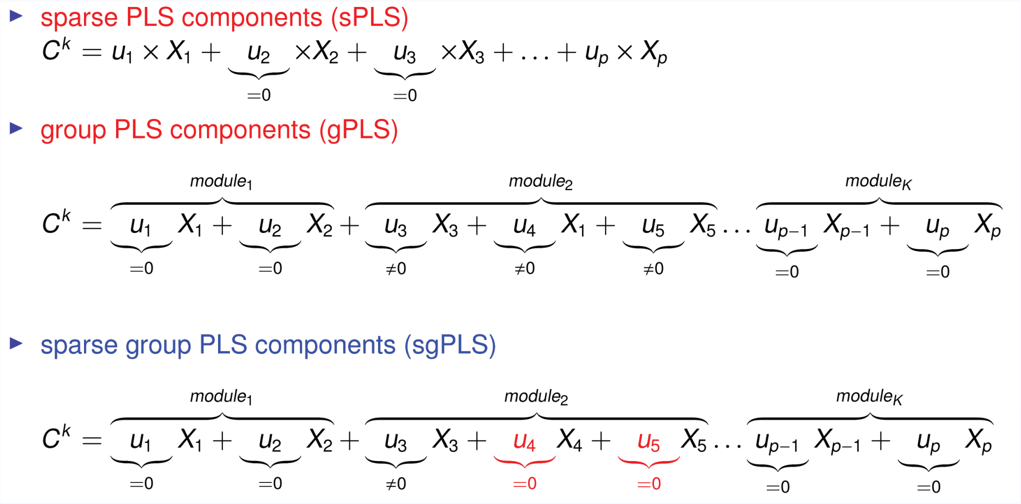
PLS analyses: Rationale and plan

- The 28 proteins can be classified in three functional groups
  - Growth Factors (N=6)
  - Chemokines (N=10)
  - Cytokines (N=12)
- Research questions
  - Do proteins jointly concur to BCL (and subtypes) onset?
  - Is the functional grouping relevant to the disease?
  - Are there groups (and proteins within each group) more associated to disease?

One Million $ question: ⇒ How can we use PLS?

PLS analyses: Rationale and plan

- The 28 proteins can be classified in three functional groups
  - Growth Factors (N=6)
  - Chemokines (N=10)
  - Cytokines (N=12)
- Additional versions of PLS:
  - Sparsity achieved through penalisation
  - Grouping signals a priori (e.g. pathways, genes)



- sparse PLS components (sPLS)

$$C^k = u_1 \times X_1 + \underbrace{u_2}_{=0} \times X_2 + \underbrace{u_3}_{=0} \times X_3 + \ldots + u_p \times X_p$$

- group PLS components (gPLS)

$$C^k = \overbrace{\underbrace{u_1}_{=0} \; X_1 + \underbrace{u_2}_{=0} \; X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} \; X_3 + \underbrace{u_4}_{\neq 0} \; X_1 + \underbrace{u_5}_{\neq 0} \; X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} \; X_{p-1} + \underbrace{u_p}_{=0} \; X_p}^{module_K}$$

- sparse group PLS components (sgPLS)

$$C^k = \overbrace{\underbrace{u_1}_{=0} \; X_1 + \underbrace{u_2}_{=0} \; X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} \; X_3 + \underbrace{u_4}_{=0} \; X_4 + \underbrace{u_5}_{=0} \; X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} \; X_{p-1} + \underbrace{u_p}_{=0} \; X_p}^{module_K}$$

PLS analyses: Rationale and plan

- The 28 proteins can be classified in three functional groups
  - Growth Factors (N=6)
  - Chemokines (N=10)
  - Cytokines (N=12)
- Research questions
  - Do proteins jointly concur to BCL (and subtypes) onset?
  - Is the functional grouping relevant to the disease?
  - Are there groups (and proteins within each group) more associated to disease?

Two Million $ question: which models????

PLS analyses: Rationale and plan

- The 28 proteins can be classified in three functional groups
  - Growth Factors (N=6)
  - Chemokines (N=10)
  - Cytokines (N=12)
- Research questions
  - Do proteins jointly concur to BCL (and subtypes) onset? – (s)PLS
  - Is the functional grouping relevant to the disease? – gPLS
  - Are there groups (and proteins within each group) more associated to disease? – sgPLS

Two Million $ response

PLS analyses: Rationale and plan

- The 28 proteins can be classified in three functional groups
  - Growth Factors (N=6)
  - Chemokines (N=10)
  - Cytokines (N=12)
- Research questions
  - Do proteins jointly concur to BCL (and subtypes) onset? – (s)PLS
  - Is the functional grouping relevant to the disease? – gPLS
  - Are there groups (and proteins within each group) more associated to disease? – sgPLS
- Analytical Plan: all PLS variants to analyse
  - All BCL
  - Each subtype separately
  - In cases only: the time to diagnosis

gPLS: Penalty function and calibration

- For each component:

$$\min_{||u||=1,\ ||v||=1} \sum_{k=1}^{K} \sum_{l=1}^{L} \underbrace{||X^{(k)^T} Y^{(l)} - u^{(k)} v^{(l)^T} ||_F^2}_{\text{covariances between } k^{th} \text{ and } l^{th} \text{ block}} + P_{\lambda_1}(u) + P_{\lambda_2}(v)$$

where

$$P_{\lambda_1}(u) = \lambda_1 \sum_{k=1}^{K} \sqrt{p_k}\ \underbrace{||u^{(k)}||_2}_{\substack{\text{loadings of} \\ k^{th} \text{ block in X}}} \qquad P_{\lambda_2}(v) = \lambda_2 \sum_{l=1}^{L} \sqrt{q_l}\ \underbrace{||v^{(l)}||_2}_{\substack{\text{loadings of} \\ l^{th} \text{ block in Y}}}$$

- Penalisation adapts to the number of variables in each group (p_k, q_l)
- Calibration: Number of selected groups in X and Y via cross-validation using MSEP

sgPLS: Penalty function and calibration

- For each component:

$$\min_{||u||=1,\ ||v||=1} \sum_{k=1}^{K} \sum_{l=1}^{L} \underbrace{||X^{(k)^T} Y^{(l)} - u^{(k)} v^{(l)^T}||_F^2}_{\text{covariances between } k^{th} \text{ and } l^{th} \text{ block}} + P_{\lambda_1}(u) + P_{\lambda_2}(v)$$

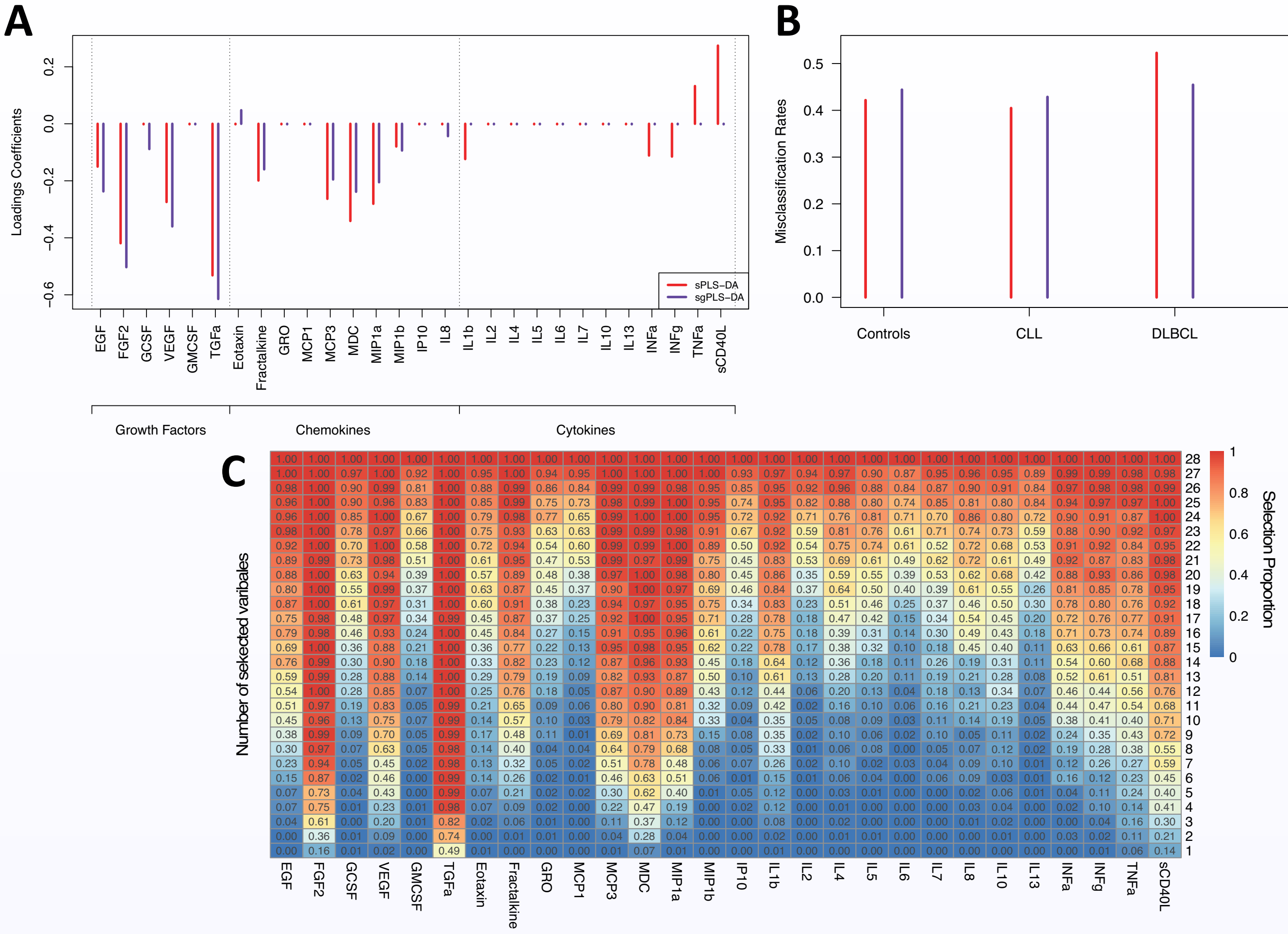
- Adding a LASSO penalty within each group:

$$P_{\lambda_1}(u) = \lambda_1 \sum_{k=1}^{K} \sqrt{p_k}||u^{(k)}||_2 + \underbrace{\alpha_1 \lambda_1 ||u||_1}_{\substack{\text{sparsity} \\ \text{in X}}}$$

$$P_{\lambda_2}(v) = \lambda_2 \sum_{l=1}^{L} \sqrt{q_l}||v^{(l)}||_2 + \underbrace{\alpha_2 \lambda_2 ||v||_1}_{\substack{\text{sparsity} \\ \text{in Y}}}$$

- Calibration: Number of selected groups in X and Y via cross-validation and the components sparsity parameter (not the number of variables)

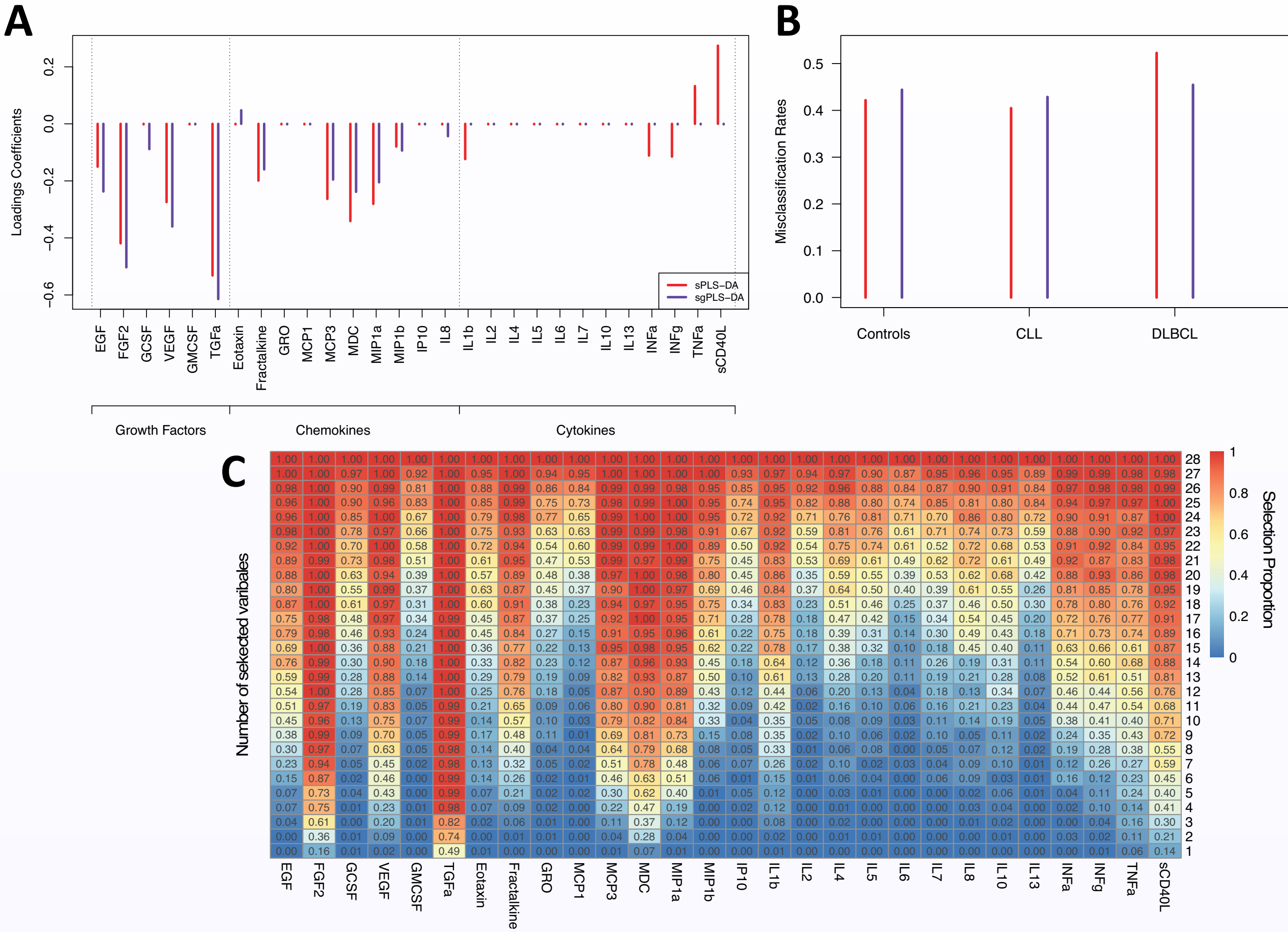
PLS analyses: All BCL



EGF	FGF2	GCSF	VEGF	GMCSF	TGFa	Eotaxin	Fractalkine	GRO	MCP1	MCP3	MDC	MIP1a	MIP1b	IP10	IL1b	IL2	IL4	IL5	IL6	IL7	IL8	IL10	IL13	INFa	INFg	TNFa	sCD40L	
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	28
1.00	1.00	0.97	1.00	0.92	1.00	0.95	1.00	0.94	0.95	1.00	1.00	1.00	1.00	0.93	0.97	0.94	0.97	0.90	0.87	0.95	0.96	0.95	0.89	0.99	0.99	0.98	0.98	27
0.98	1.00	0.90	0.99	0.81	1.00	0.88	0.99	0.86	0.84	0.99	0.99	0.98	0.95	0.85	0.95	0.92	0.96	0.88	0.84	0.87	0.90	0.91	0.84	0.97	0.98	0.98	0.99	26
0.96	1.00	0.90	0.96	0.83	1.00	0.85	0.99	0.75	0.73	0.98	0.99	1.00	0.95	0.74	0.95	0.82	0.88	0.80	0.74	0.85	0.81	0.80	0.84	0.94	0.97	0.97	1.00	25
0.96	1.00	0.85	1.00	0.67	1.00	0.79	0.98	0.77	0.65	0.99	1.00	1.00	0.95	0.72	0.92	0.71	0.76	0.81	0.71	0.70	0.86	0.80	0.72	0.90	0.91	0.87	1.00	24
0.98	1.00	0.78	0.97	0.66	1.00	0.75	0.93	0.63	0.63	0.99	0.99	0.98	0.91	0.67	0.92	0.59	0.81	0.76	0.61	0.71	0.74	0.73	0.59	0.88	0.90	0.92	0.97	23
0.92	1.00	0.70	1.00	0.58	1.00	0.72	0.94	0.54	0.60	0.99	0.99	1.00	0.89	0.50	0.92	0.54	0.75	0.74	0.61	0.52	0.72	0.68	0.53	0.91	0.92	0.84	0.95	22
0.89	0.99	0.73	0.98	0.51	1.00	0.61	0.95	0.47	0.53	0.99	0.97	0.99	0.75	0.45	0.83	0.53	0.69	0.61	0.49	0.62	0.72	0.61	0.49	0.92	0.87	0.83	0.98	21
0.88	1.00	0.63	0.94	0.39	1.00	0.57	0.89	0.48	0.38	0.97	1.00	0.98	0.80	0.45	0.86	0.35	0.59	0.55	0.39	0.53	0.62	0.68	0.42	0.88	0.93	0.86	0.98	20
0.80	1.00	0.55	0.99	0.37	1.00	0.63	0.87	0.45	0.37	0.90	1.00	0.97	0.69	0.46	0.84	0.37	0.64	0.50	0.40	0.39	0.61	0.55	0.26	0.81	0.85	0.78	0.95	19
0.87	1.00	0.61	0.97	0.31	1.00	0.60	0.91	0.38	0.23	0.94	0.97	0.95	0.75	0.34	0.83	0.23	0.51	0.46	0.25	0.37	0.46	0.50	0.30	0.78	0.80	0.76	0.92	18
0.75	0.98	0.48	0.97	0.34	0.99	0.45	0.87	0.37	0.25	0.92	1.00	0.95	0.71	0.28	0.78	0.18	0.47	0.42	0.15	0.34	0.54	0.45	0.20	0.72	0.76	0.77	0.91	17
0.79	0.98	0.46	0.93	0.24	1.00	0.45	0.84	0.27	0.15	0.91	0.95	0.96	0.61	0.22	0.75	0.18	0.39	0.31	0.14	0.30	0.49	0.43	0.18	0.71	0.73	0.74	0.89	16
0.69	1.00	0.36	0.88	0.21	1.00	0.36	0.77	0.22	0.13	0.95	0.98	0.95	0.62	0.22	0.78	0.17	0.38	0.32	0.10	0.18	0.45	0.40	0.11	0.63	0.66	0.61	0.87	15
0.76	0.99	0.30	0.90	0.18	1.00	0.33	0.82	0.23	0.12	0.87	0.96	0.93	0.45	0.18	0.64	0.12	0.36	0.18	0.11	0.26	0.19	0.31	0.11	0.54	0.60	0.68	0.88	14
0.59	0.99	0.28	0.88	0.14	1.00	0.29	0.79	0.19	0.09	0.82	0.93	0.87	0.50	0.10	0.61	0.13	0.28	0.20	0.11	0.19	0.21	0.28	0.08	0.52	0.61	0.51	0.81	13
0.54	1.00	0.28	0.85	0.07	1.00	0.25	0.76	0.18	0.05	0.87	0.90	0.89	0.43	0.12	0.44	0.06	0.20	0.13	0.04	0.18	0.13	0.34	0.07	0.46	0.44	0.56	0.76	12
0.51	0.97	0.19	0.83	0.05	0.99	0.21	0.65	0.09	0.06	0.80	0.90	0.81	0.32	0.09	0.42	0.02	0.16	0.10	0.06	0.16	0.21	0.23	0.04	0.44	0.47	0.54	0.68	11
0.45	0.96	0.13	0.75	0.07	0.99	0.14	0.57	0.10	0.03	0.79	0.82	0.84	0.33	0.04	0.35	0.05	0.08	0.09	0.03	0.11	0.14	0.19	0.05	0.38	0.41	0.40	0.71	10
0.38	0.99	0.09	0.70	0.05	0.99	0.17	0.48	0.11	0.01	0.69	0.81	0.73	0.15	0.08	0.35	0.02	0.10	0.06	0.02	0.10	0.05	0.11	0.02	0.24	0.35	0.43	0.72	9
0.30	0.97	0.07	0.63	0.05	0.98	0.14	0.40	0.04	0.04	0.64	0.79	0.68	0.08	0.05	0.33	0.01	0.06	0.08	0.00	0.05	0.07	0.12	0.02	0.19	0.28	0.38	0.55	8
0.23	0.94	0.05	0.45	0.02	0.98	0.13	0.32	0.05	0.02	0.51	0.78	0.48	0.06	0.07	0.26	0.02	0.10	0.02	0.03	0.04	0.09	0.10	0.01	0.12	0.26	0.27	0.59	7
0.15	0.87	0.02	0.46	0.00	0.99	0.14	0.26	0.02	0.01	0.46	0.63	0.51	0.06	0.01	0.15	0.01	0.06	0.04	0.00	0.06	0.09	0.03	0.01	0.16	0.12	0.23	0.45	6
0.07	0.73	0.04	0.43	0.00	0.99	0.07	0.21	0.02	0.02	0.30	0.62	0.40	0.01	0.05	0.12	0.00	0.03	0.01	0.00	0.00	0.02	0.06	0.01	0.04	0.11	0.24	0.40	5
0.07	0.75	0.01	0.23	0.01	0.98	0.07	0.09	0.02	0.00	0.22	0.47	0.19	0.00	0.02	0.12	0.00	0.01	0.00	0.00	0.00	0.03	0.04	0.00	0.02	0.10	0.14	0.41	4
0.04	0.61	0.00	0.20	0.01	0.82	0.02	0.06	0.01	0.00	0.11	0.37	0.12	0.00	0.00	0.08	0.00	0.02	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.04	0.16	0.30	3
0.00	0.36	0.01	0.09	0.00	0.74	0.00	0.01	0.01	0.00	0.04	0.28	0.04	0.00	0.00	0.02	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.03	0.02	0.11	0.21	2
0.00	0.16	0.01	0.02	0.00	0.49	0.01	0.00	0.00	0.00	0.01	0.07	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.14	1

- sPLS mainly selects variables in GF an chemokines groups
- Two cytokines proteins selected with larger loadings (TNF-α, sCD40)
- sgPLS selects the two group with more non zero loadings

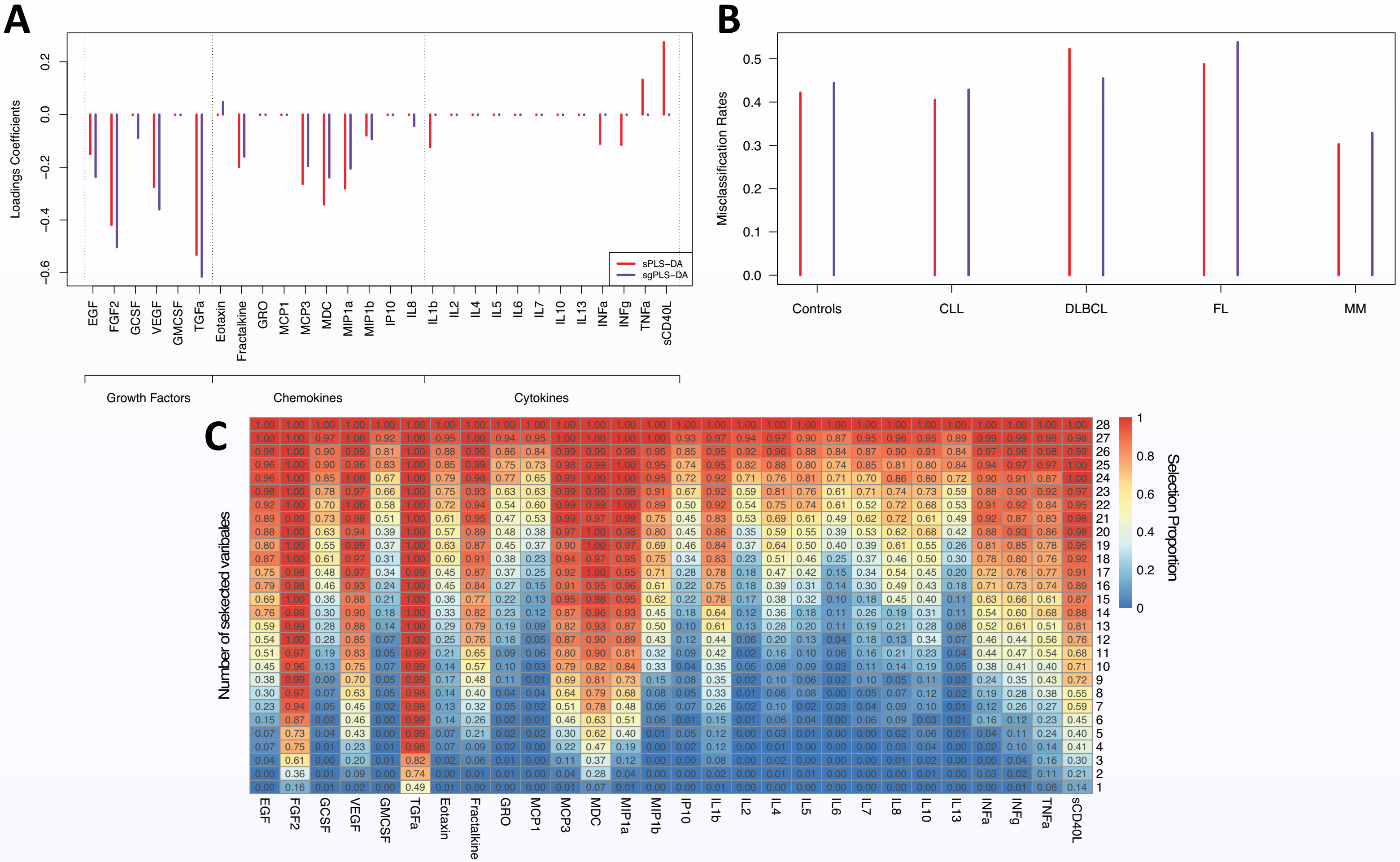
PLS analyses: All BCL



	EGF	FGF2	GCSF	VEGF	GMCSF	TGFa	Eotaxin	Fractalkine	GRO	MCP1	MCP3	MDC	MIP1a	MIP1b	IP10	IL1b	IL2	IL4	IL5	IL6	IL7	IL8	IL10	IL13	INFa	INFg	TNFa	sCD40L
28	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
27	1.00	1.00	0.97	1.00	0.92	1.00	0.95	1.00	0.94	0.95	1.00	1.00	1.00	1.00	0.93	0.97	0.94	0.97	0.90	0.87	0.95	0.96	0.95	0.89	0.99	0.99	0.98	0.98
26	0.98	1.00	0.90	0.99	0.81	1.00	0.88	0.99	0.86	0.84	0.99	0.99	0.98	0.95	0.85	0.95	0.92	0.96	0.88	0.84	0.87	0.90	0.91	0.84	0.97	0.98	0.98	0.99
25	0.96	1.00	0.90	0.96	0.83	1.00	0.85	0.99	0.75	0.73	0.98	0.99	1.00	0.95	0.74	0.95	0.82	0.88	0.80	0.74	0.85	0.81	0.80	0.84	0.94	0.97	0.97	1.00
24	0.96	1.00	0.85	1.00	0.67	1.00	0.79	0.98	0.77	0.65	0.99	1.00	1.00	0.95	0.72	0.92	0.71	0.76	0.81	0.71	0.70	0.86	0.80	0.72	0.90	0.91	0.87	1.00
23	0.98	1.00	0.78	0.97	0.66	1.00	0.75	0.93	0.63	0.63	0.99	0.99	0.98	0.91	0.67	0.92	0.59	0.81	0.76	0.61	0.71	0.74	0.73	0.59	0.88	0.90	0.92	0.97
22	0.92	1.00	0.70	1.00	0.58	1.00	0.72	0.94	0.54	0.60	0.99	0.99	1.00	0.89	0.50	0.92	0.54	0.75	0.74	0.61	0.52	0.72	0.68	0.53	0.91	0.92	0.84	0.95
21	0.89	0.99	0.73	0.98	0.51	1.00	0.61	0.95	0.47	0.53	0.99	0.97	0.99	0.75	0.45	0.83	0.53	0.69	0.61	0.49	0.62	0.72	0.61	0.49	0.92	0.87	0.83	0.98
20	0.88	1.00	0.63	0.94	0.39	1.00	0.57	0.89	0.48	0.38	0.97	1.00	0.98	0.80	0.45	0.86	0.35	0.59	0.55	0.39	0.53	0.62	0.68	0.42	0.88	0.93	0.86	0.98
19	0.80	1.00	0.55	0.99	0.37	1.00	0.63	0.87	0.45	0.37	0.90	1.00	0.97	0.69	0.46	0.84	0.37	0.64	0.50	0.40	0.39	0.61	0.55	0.26	0.81	0.85	0.78	0.95
18	0.87	1.00	0.61	0.97	0.31	1.00	0.60	0.91	0.38	0.23	0.94	0.97	0.95	0.75	0.34	0.83	0.23	0.51	0.46	0.25	0.37	0.46	0.50	0.30	0.78	0.80	0.76	0.92
17	0.75	0.98	0.48	0.97	0.34	0.99	0.45	0.87	0.37	0.25	0.92	1.00	0.95	0.71	0.28	0.78	0.18	0.47	0.42	0.15	0.34	0.54	0.45	0.20	0.72	0.76	0.77	0.91
16	0.79	0.98	0.46	0.93	0.24	1.00	0.45	0.84	0.27	0.15	0.91	0.95	0.96	0.61	0.22	0.75	0.18	0.39	0.31	0.14	0.30	0.49	0.43	0.18	0.71	0.73	0.74	0.89
15	0.69	1.00	0.36	0.88	0.21	1.00	0.36	0.77	0.22	0.13	0.95	0.98	0.95	0.62	0.22	0.78	0.17	0.38	0.32	0.10	0.18	0.45	0.40	0.11	0.63	0.66	0.61	0.87
14	0.76	0.99	0.30	0.90	0.18	1.00	0.33	0.82	0.23	0.12	0.87	0.96	0.93	0.45	0.18	0.64	0.12	0.36	0.18	0.11	0.26	0.19	0.31	0.11	0.54	0.60	0.68	0.88
13	0.59	0.99	0.28	0.88	0.14	1.00	0.29	0.79	0.19	0.09	0.82	0.93	0.87	0.50	0.10	0.61	0.13	0.28	0.20	0.11	0.19	0.21	0.28	0.08	0.52	0.61	0.51	0.81
12	0.54	1.00	0.28	0.85	0.07	1.00	0.25	0.76	0.18	0.05	0.87	0.90	0.89	0.43	0.12	0.44	0.06	0.20	0.13	0.04	0.18	0.13	0.34	0.07	0.46	0.44	0.56	0.76
11	0.51	0.97	0.19	0.83	0.05	0.99	0.21	0.65	0.09	0.06	0.80	0.90	0.81	0.32	0.09	0.42	0.02	0.16	0.10	0.06	0.16	0.21	0.23	0.04	0.44	0.47	0.54	0.68
10	0.45	0.96	0.13	0.75	0.07	0.99	0.14	0.57	0.10	0.03	0.79	0.82	0.84	0.33	0.04	0.35	0.05	0.08	0.09	0.03	0.11	0.14	0.19	0.05	0.38	0.41	0.40	0.71
9	0.38	0.99	0.09	0.70	0.05	0.99	0.17	0.48	0.11	0.01	0.69	0.81	0.73	0.15	0.08	0.35	0.02	0.10	0.06	0.02	0.10	0.05	0.11	0.02	0.24	0.35	0.43	0.72
8	0.30	0.97	0.07	0.63	0.05	0.98	0.14	0.40	0.04	0.04	0.64	0.79	0.68	0.08	0.05	0.33	0.01	0.06	0.08	0.00	0.05	0.07	0.12	0.02	0.19	0.28	0.38	0.55
7	0.23	0.94	0.05	0.45	0.02	0.98	0.13	0.32	0.05	0.02	0.51	0.78	0.48	0.06	0.07	0.26	0.02	0.10	0.02	0.03	0.04	0.09	0.10	0.01	0.12	0.26	0.27	0.59
6	0.15	0.87	0.02	0.46	0.00	0.99	0.14	0.26	0.02	0.01	0.46	0.63	0.51	0.06	0.01	0.15	0.01	0.06	0.04	0.00	0.06	0.09	0.03	0.01	0.16	0.12	0.23	0.45
5	0.07	0.73	0.04	0.43	0.00	0.99	0.07	0.21	0.02	0.02	0.30	0.62	0.40	0.01	0.05	0.12	0.00	0.03	0.01	0.00	0.00	0.02	0.06	0.01	0.04	0.11	0.24	0.40
4	0.07	0.75	0.01	0.23	0.01	0.98	0.07	0.09	0.02	0.00	0.22	0.47	0.19	0.00	0.02	0.12	0.00	0.01	0.00	0.00	0.00	0.03	0.04	0.00	0.02	0.10	0.14	0.41
3	0.04	0.61	0.00	0.20	0.01	0.82	0.02	0.06	0.01	0.00	0.11	0.37	0.12	0.00	0.00	0.08	0.00	0.02	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.04	0.16	0.30
2	0.00	0.36	0.01	0.09	0.00	0.74	0.00	0.01	0.01	0.00	0.04	0.28	0.04	0.00	0.00	0.02	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.03	0.02	0.11	0.21
1	0.00	0.16	0.01	0.02	0.00	0.49	0.01	0.00	0.00	0.00	0.01	0.07	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.14

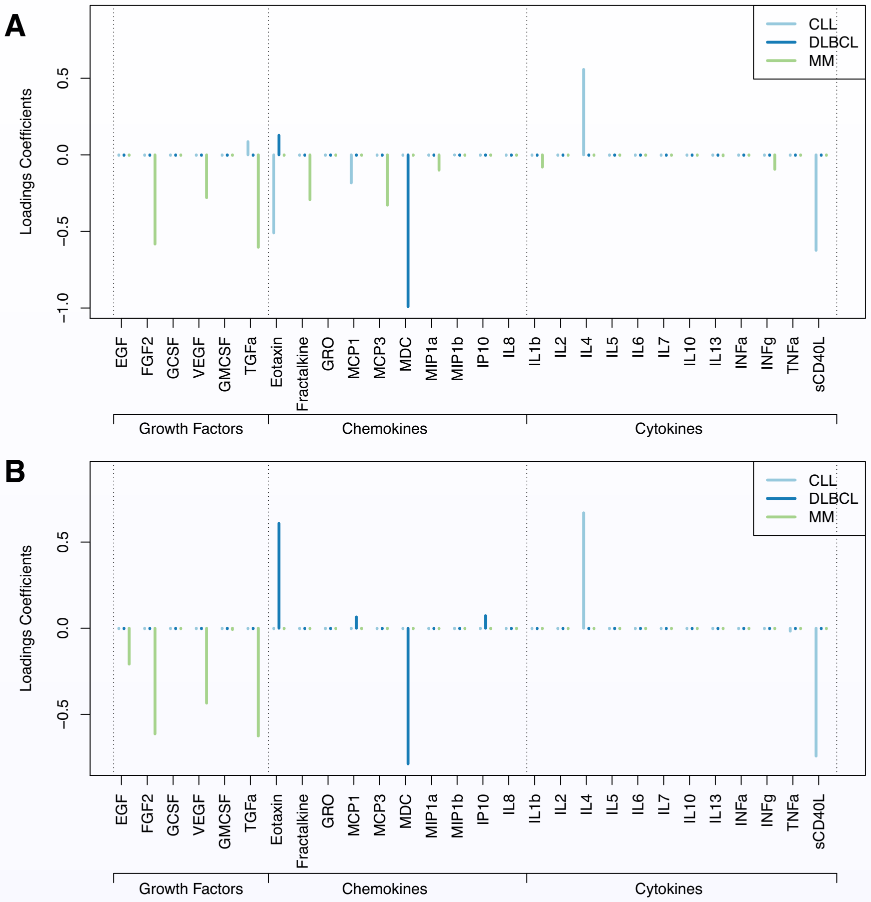
- sPLS and sgPLS yield comparable misclassification rates (unimportant exclusion of cytokines)
- Better misclassification rates for MM

PLS analyses: All BCL



- Assessing the sensitivity to calibration via stability analyses
- The largest loadings are the first and most frequently selected (sPLS)

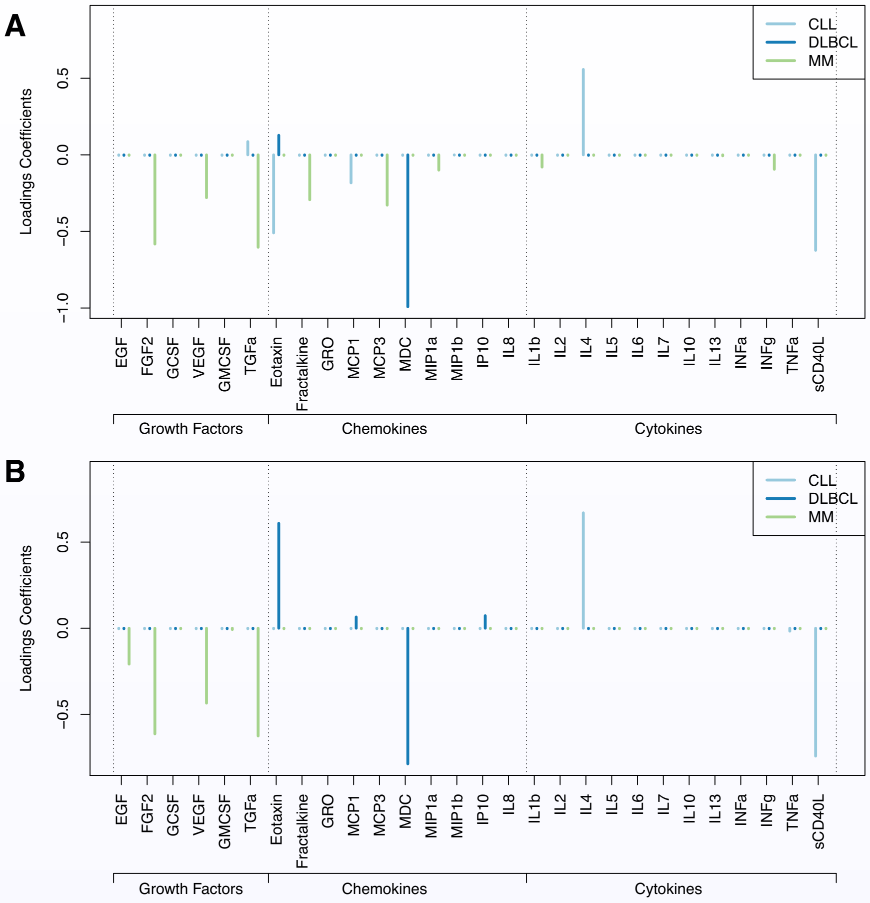
PLS analyses: subtype analyses



sPLS analyses select:

- **MM**: proteins mainly in chemokines and growth factors
- **CLL**: chemokines and cytokines (though only 2/12 proteins)
- **DLBCL**: 2 chemokines are selected

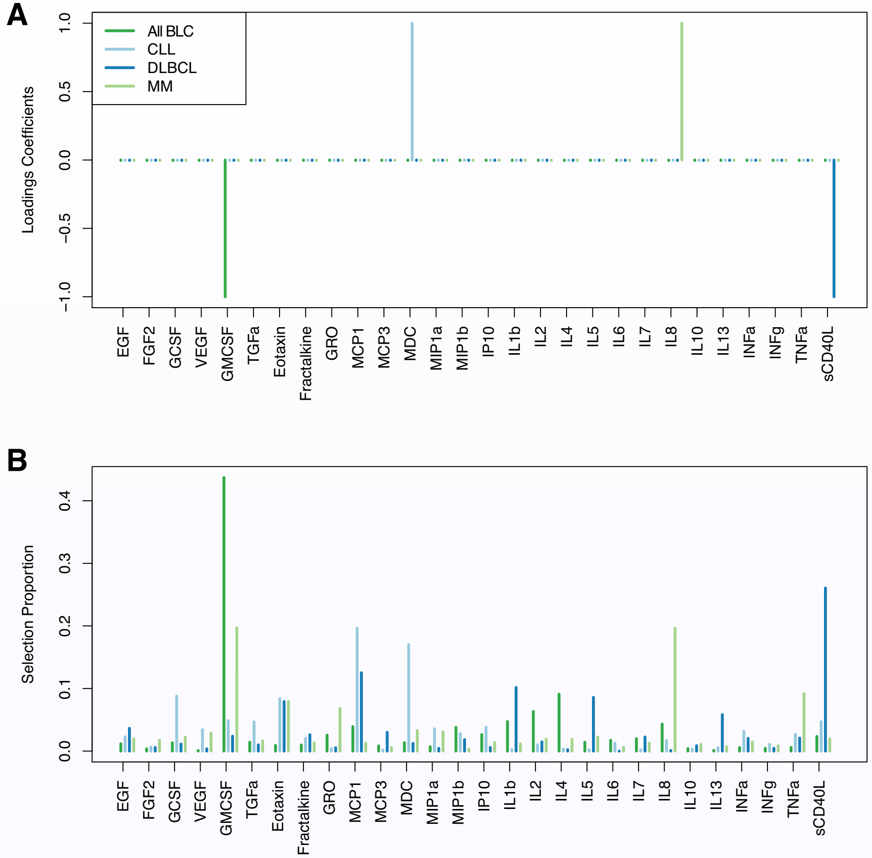
PLS analyses: subtype analyses



sgPLS analyses select:

- **MM**: growth factors and within the group the same variables as sPLS
- **CLL**: cytokines and both the sPLS proteins
- **DLBCL**: Chemokines are selected (including the the 2 sPLS proteins)

Cases-only analyses



sPLS analyses of TtD (continuous outcome) selects:

- A single and specific protein for each subtype
- For all BCL, GMCSF is mostly selected, for other subtypes, several candidates compete
- For DLBCL, sCD40 seem to be more frequently selected.

Wrap-up summary

- PLS analyses were able to identify associations the were not detected by univariate models

  ⇒ these potential markers were supported by external biological evidence

- Inclusion of groups allows to account for correlations across proteins and select the most informative sets of predictors

  ⇒ contribution to the sparsity and interpretability of the results

- Limitation: sensitivity to the grouping strategy

  ⇒ grouping is defining a prior hypothesis

- Extensions:
  - s-g-sg-PLS can accommodate large block of data (e.g. gene expression)
  - OMICs integration via sgsPLS
  - Computational Optimisation: bigPLS

Acknowledgments

Imperial College London

- M Chadeau-Hyam
- B Bodinier
- M Karimi

Universiteit Utrecht

- R Vermeulen
- L Portengen
- F Saberi

- Collaborators: B Liquet (University of Pau)
- Financial Support:
  - EU FP7 EXPOsOMICS
  - CR-UK Population Research Committee project ’Mechanomics’