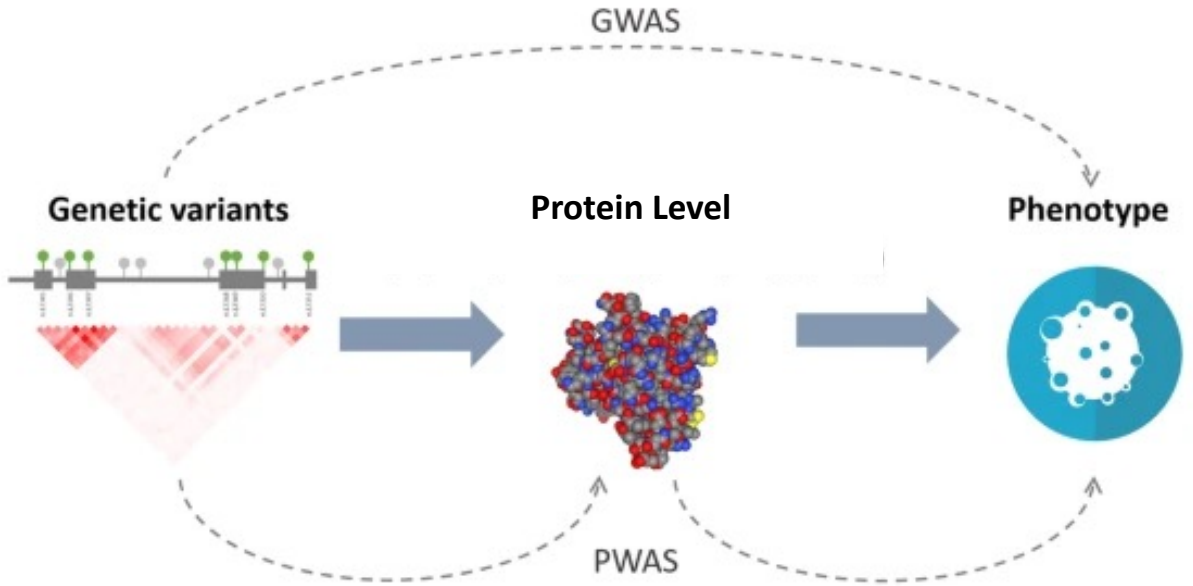
Women's Health Initiative Proteome-Wide Association Study (WHI PWAS)

Feb 18, 2022

UNC | GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

PWAS (Proteome-Wide Association Study)

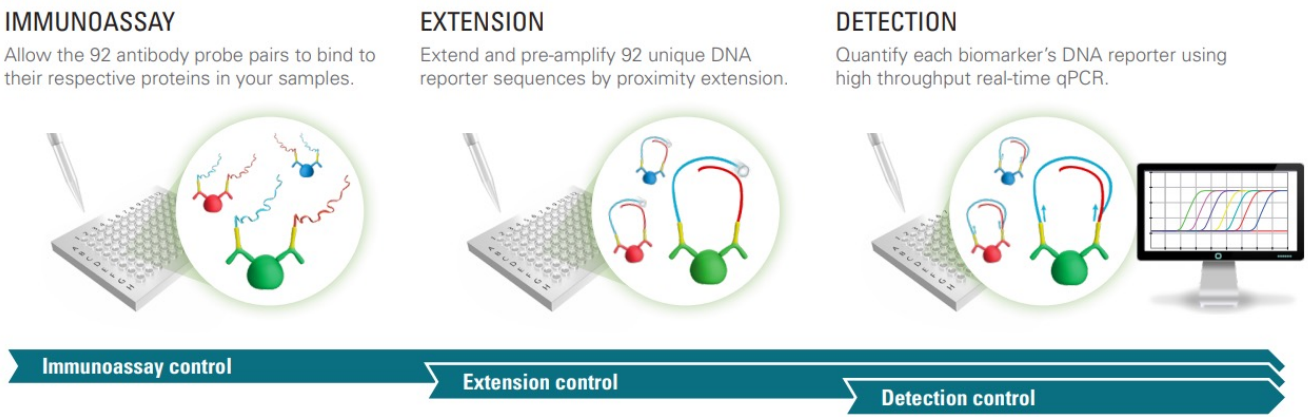


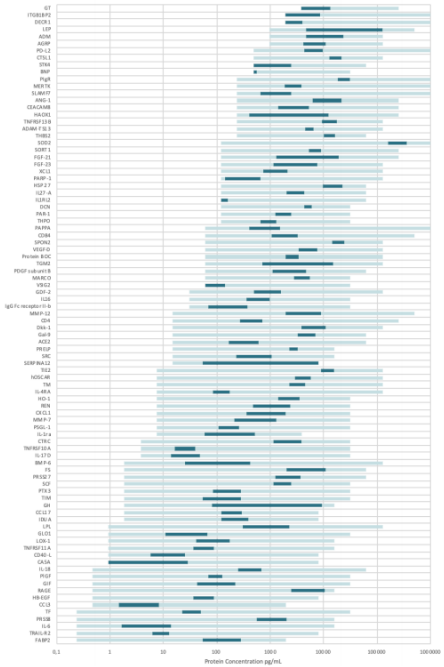
Women's Health Initiative (WHI) Proteomic Data

- Subjects: 1133 women aged between 65-95 years old (multiple self reported race/ethnicities)
- Genotype: TOPMed Whole Genome Sequencing (Freeze 9b, Phase 2, sequencing center: BROAD)
- Protein: 552 protein level data
- Samples were included only if they had complete data for all covariates and protein data.
- In total, 1002 samples were included in the analysis.

WHI Proteomics Data – Protein (OLINK)

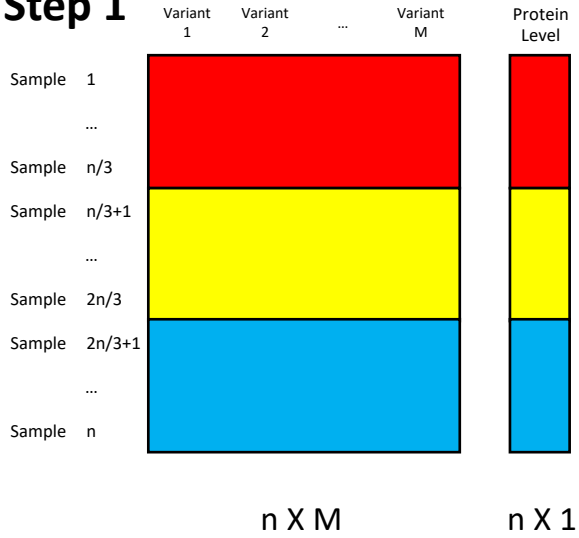
- 6 panels (Cardiometabolic, Cardiovascular II, Cardiovascular III, Inflammation, Neurology, Onlcology III)
- 552 proteins total (92 proteins per each panel)
- 1002 WHI samples distributed over 16 plates (~63 per each)





Methods – Proposed method

Step 1



Step 2: GWAS

Protein Level ~ Age + Plate + Single Variant

Step 3: Elastic Net Training

Protein Level ~ Significant Variants (pQTLs)

Step 4: Model Testing

Correlation (Observed Protein Level, Predicted Protein Level)

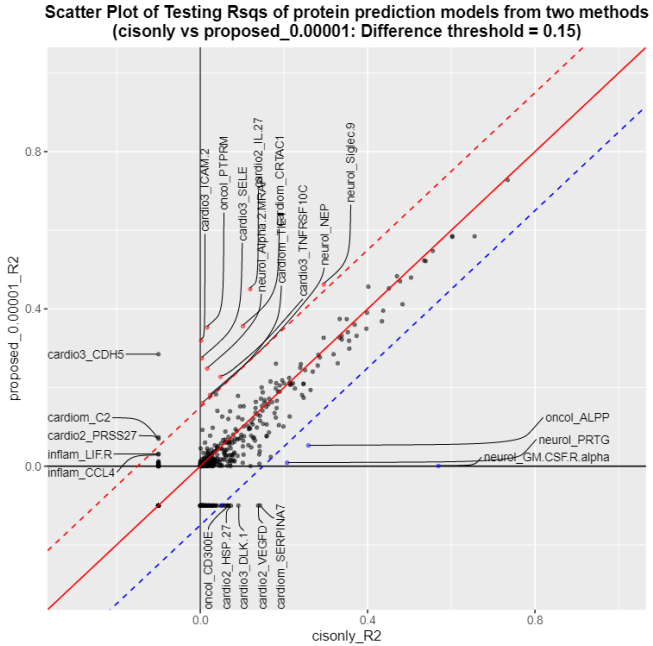
Step 5: Association

Phenotype ~ Predicted Protein Level
(New data, protein data not available)

Methods – Other methods for predicting Protein Level

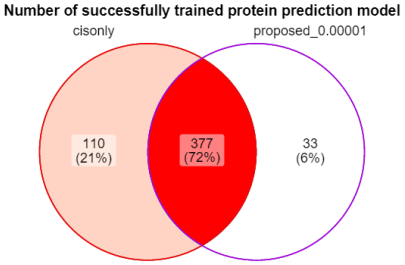
	Prediction Model	Remark
Proposed	GWAS variant selection + EN	Trans pQTLs + Cis pQTLs 1/3 GWAS + 1/3 Training + 1/3 Testing * Different thresholds for defining pQTLs
PrediXcan	EN	Cis only (Protein Coding gene $\pm$1Mbp) 2/3 Training + 1/3 Testing
BGW-TWAS	TWAS + Bayesian variant selection	Trans + Cis, handled by Bayesian variant selection (different priors) 2/3 Training + 1/3 Testing

Results – Proposed(pQTL thresh = 1E-4) vs. PrediXcan(cis-only)



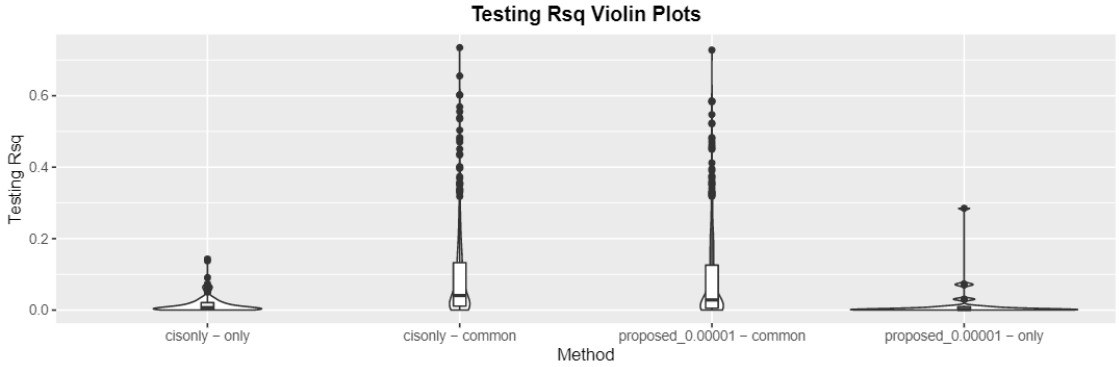
Correlation b/w R2	0.9365
Proposed > Cis-only	241
Proposed < Cis-only	148

Results – Proposed(pQTL thresh = 1E-4) vs. PrediXcan(cis-only)

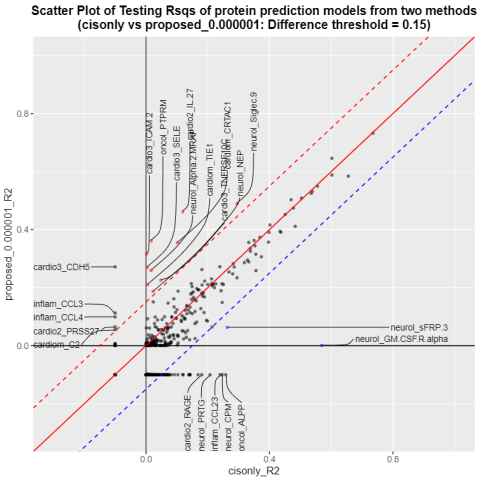


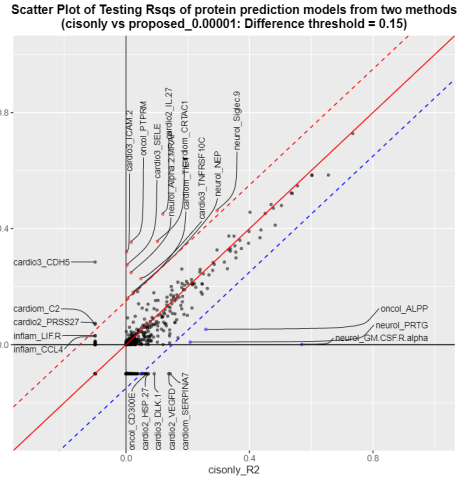
Testing Rsq Summary statistics

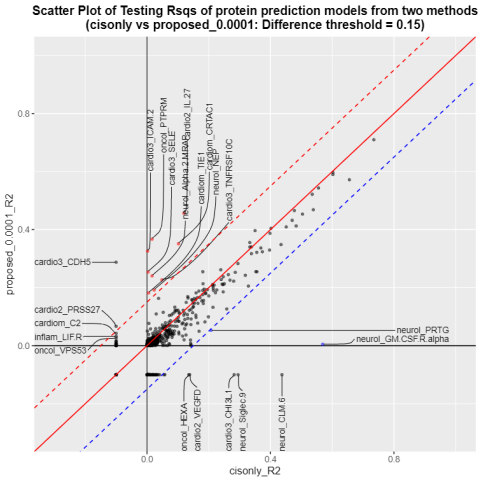
	cisonly – only	cisonly – common	proposed_0.00001 – common	proposed_0.00001 – only
Min.	0.0001133	0.0001098	0.0001198	0.0001104
1st Qu.	0.001807	0.01119	0.005346	0.0004372
Median	0.007024	0.04075	0.0285	0.002055
Mean	0.01717	0.0961	0.08882	0.01729
3rd Qu.	0.02105	0.1328	0.1259	0.009186
Max.	0.1428	0.7347	0.728	0.2849

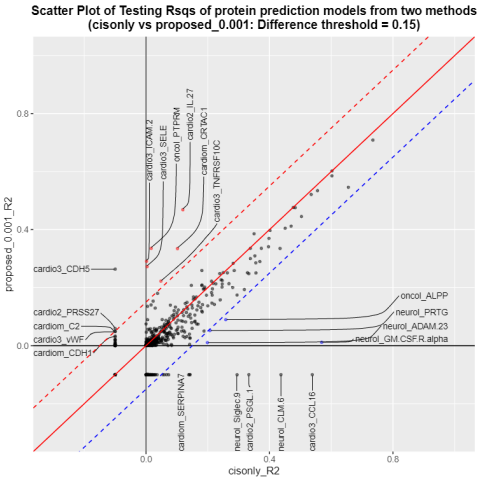


Results – Proposed(other pQTL threshs) vs. PrediXcan(cis-only)

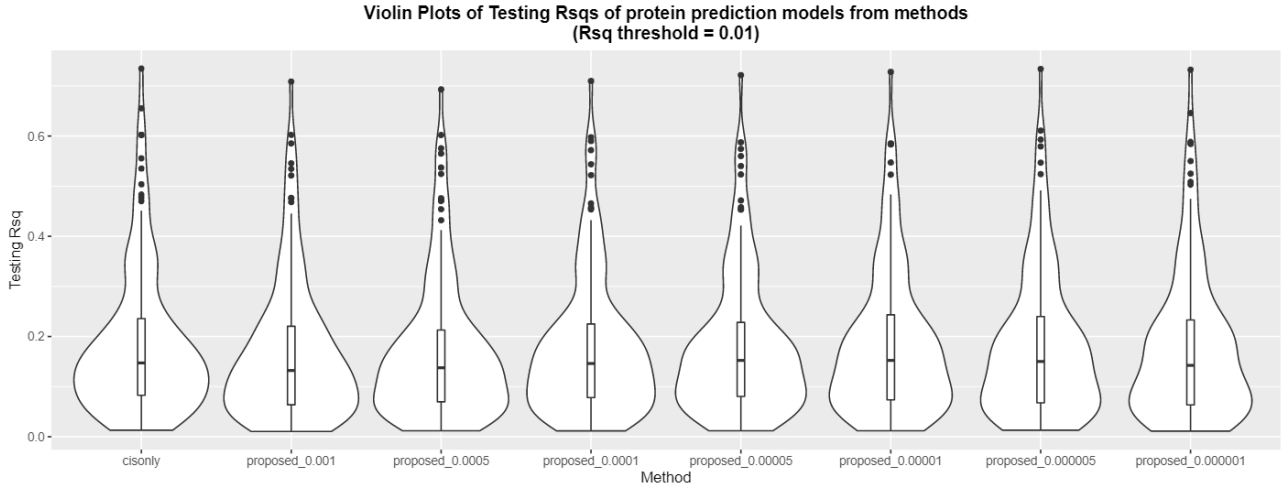








Results – Proposed(other pQTL threshs) vs. PrediXcan(cis-only)



	cisonly	proposed_0.001	proposed_0.0005	proposed_0.0001	proposed_0.00005	proposed_0.00001	proposed_0.000005	proposed_0.000001
Min.	0.0131	0.01038	0.01179	0.0116	0.01189	0.01173	0.01308	0.01113
1st Qu.	0.08265	0.06367	0.06961	0.07812	0.0803	0.07359	0.06749	0.06343
Median	0.1472	0.1321	0.1376	0.1461	0.1523	0.1523	0.1501	0.1424
Mean	0.185	0.1689	0.172	0.1782	0.1812	0.1852	0.1829	0.18
3rd Qu.	0.2357	0.2203	0.2128	0.225	0.2283	0.2433	0.2398	0.233
Max.	0.7347	0.7087	0.6931	0.7099	0.7216	0.728	0.734	0.7323

Results – Top proteins with higher proposed R2

Protein	Proposed R2†	Cis-only R2	BGW-TWAS R2	chr	begin	end
cardio2_IL.27	0.456	0.119	0.108	16	28499362	28512051
cardio3_SELE	0.253	0.003	0.010	1	169722640	169764705
cardio3_TNFRSF10C	0.226	0.048	0.253	8	23102921	23117445
cardiom_CRTAC1	0.351	0.101	0.165	10	97865000	98030828
neurol_Alpha.2.MRAP	0.240	0.016	- ‡	4	3563612	3532446

† pQTL threshold of 1E-4 was used
‡ BGW-TWAS method did not converge

Results – Top proteins with higher proposed R2

Protein	Coding region	Top GWAS pQTL	Gene
cardio2_IL.27	16:28499362-28512051	19:4236999:G:A	EBI3
cardio3_SELE	1:169722640-169764705	9:133273983:A:G	ABO
cardio3_TNFRSF10C	8:23102921-23117445	19:43648948:A:G	PLAUR
cardiom_CRTAC1	10:97865000-98030828	3:186675758:T:C	HRG, LOC105374258
neurol_Alpha.2.MRAP	4:3563612-3532446	6:31322175:G:A	Non-coding (HLA-B region)*

* tangled with many diseases (**Pleiotropic)**

Results – Cardiom_CRTAC1(10:97865000-98030828)

Chr	Position	Dosage allele	Other allele	Effect
chr10	97604412	T	C	-0.008275998
chr10	97605446	A	C	-0.004166192
chr10	97799838	G	T	-0.031561578
chr10	97865562	C	T	-0.241755278
chr10	97874320	T	A	-0.031657292
chr10	97882605	C	A	-0.001671401
chr10	98009631	T	C	-0.010379587
chr10	98597841	C	T	-0.001523264

Chr	Position	Dosage allele	Other allele	Effect	Chr	Position	Dosage allele	Other allele	Effect	Chr	Position	Dosage allele	Other allele	Effect
chr1	53412053	G	C	0.021374624	chr4	8045444	T	A	0.029475358	chr10	97882636	T	G	0.000909407
chr1	55408611	C	T	0.013881464	chr4	14102900	A	G	-0.018981542	chr10	97882974	C	T	-0.012202091
chr1	58736135	T	C	-0.000498083	chr4	69455575	A	G	0.008843145	chr10	97884723	A	ATAG	0.042092083
chr1	98444139	G	T	0.027224516	chr4	69457268	G	A	0.009782561	chr10	113971745	C	T	-0.021123189
chr1	175652655	C	T	0.008686621	chr4	138038574	A	G	-0.006400009	chr11	125239338	G	A	-0.019257107
chr1	210187356	A	C	-0.012936621	chr4	138039670	A	G	-0.004603097	chr11	130004032	T	G	0.063899215
chr1	210191636	T	C	-0.003891127	chr4	138040631	A	T	-0.01549694	chr11	133939137	C	T	-0.019393803
chr1	210197506	A	G	-0.011685119	chr4	138061340	G	A	-0.004893897	chr12	52651219	G	A	0.042219022
chr1	210198956	T	C	-0.012431236	chr4	147836315	A	C	0.053054811	chr12	52672132	T	C	0.037542118
chr1	210199110	T	C	-0.011235327	chr5	2307561	C	T	0.018186659	chr12	129728881	C	A	-0.064542702
chr2	45223627	A	C	-0.091296837	chr5	3679607	A	G	0.040055449	chr12	132929135	C	T	0.024423865
chr2	51391916	A	G	-0.028901559	chr5	19978017	C	T	0.029478138	chr14	24267755	T	C	-0.031641478
chr2	98383912	A	T	0.022118694	chr5	57849947	G	A	-0.035265686	chr14	24267997	A	T	-0.050932782
chr2	98384351	G	A	0.017783897	chr5	71702226	T	C	-0.038315173	chr14	24273475	C	G	-0.038152002
chr2	98384385	G	C	0.021066046	chr5	115747837	C	T	-0.010415121	chr15	58392768	T	G	-0.001182611
chr2	141583289	C	G	0.014812876	chr5	158037168	T	C	0.000669803	chr16	12826640	T	C	0.011353429
chr2	183579761	G	T	-0.030848933	chr6	99030934	A	G	-0.020057847	chr16	26866855	C	A	-0.056092685
chr3	71785567	G	C	-0.046403818	chr7	20421953	C	T	-0.015412669	chr16	28522310	G	T	0.052656064
chr3	186642021	T	G	0.065854479	chr7	26354919	A	AT	0.025536226	chr16	52386662	T	TA	0.007440686
chr3	186642620	T	G	0.0250685	chr8	82874999	A	C	-0.009067646	chr16	88548200	A	T	0.007508193
chr3	186671770	T	C	0.084961575	chr8	82879220	G	T	-0.001036412	chr17	13724863	C	T	-0.031508496
chr3	186672039	A	G	0.023289661	chr9	96693878	T	C	0.022032428	chr17	53771447	G	T	0.024336922
chr3	186675758	T	C	0.209109032	chr10	43376514	T	C	0.040981711	chr17	55435546	A	G	0.022114884
chr3	186675997	G	A	0.208746359	chr10	60030165	C	T	-0.068096726	chr17	73949037	A	G	-0.001669923
chr3	186677647	C	T	0.165547227	chr10	97865562	C	T	-0.491256074	chr18	3332475	C	A	0.000870523
chr3	186678827	T	C	0.16290186	chr10	97880706	C	T	0.011660173	chr18	57400128	C	T	-0.005441302
chr4	1556563	A	G	-0.05088851	chr10	97881054	C	T	0.07497584	chr18	69192703	C	G	0.001089696
										chrX	88712818	C	T	-0.002679584

Results – Cardio3_SELE (1:169722640-169764705)

Chr	Position	Dosage allele	Other allele	Effect
chr1	168722746	C	T	0.015189428
chr1	168740549	C	T	-0.0125724
chr1	168744173	A	G	-0.0017921
chr1	168759234	C	T	0.021001928
chr1	168762414	G	T	0.008713327
chr1	168765219	G	A	0.030086561
chr1	169632718	C	T	-0.065893932
chr1	169646042	C	A	-0.006577659
chr1	170224861	G	A	-0.024104233
chr1	170226746	T	A	-0.03716496

ABO locus: pQTL for E-selectin (9:133250401-133276024)

Chr	Position	Dosage allele	Other allele	Effect	Chr	Position	Dosage allele	Other allele	Effect
Chr1	118415878	A	G	-0.042514378	chr9	20125261	C	T	0.01589
Chr1	182354697	C	A	-0.00157013	chr9	20125396	C	T	0.000869
chr2	179643783	T	G	-0.019339605	chr9	35144751	G	A	-0.00714
Chr3	110310996	A	G	-0.002775586	chr9	35146875	G	A	-0.0205
Chr3	110311538	G	A	-0.001705695	chr9	133166937	T	A	-0.02982
Chr3	157140511	G	A	-0.007264246	chr9	133257761	T	C	-0.0286
chr3	157190525	G	T	-0.024312382	chr9	133261703	A	G	0.00184
chr3	157205024	A	C	-0.024188674	chr9	133263362	GCGCCCACCACTA	G	-0.27175
chr4	12479999	C	A	-0.036385733	chr9	133266456	T	C	0.108349
chr4	12511332	G	C	0.008626011	chr9	133266790	C	A	0.02558
chr4	181329067	A	G	0.02317239	chr9	133268030	G	A	0.014179
chr5	11386566	T	C	0.032817029	chr9	133271001	T	C	0.025298
chr5	34928242	T	G	-0.044098048	chr9	133271018	T	TAAGAC	0.121981
chr5	34931494	C	A	-0.068833002	chr9	133271745	T	C	0.000713
chr5	117887452	C	G	0.042829487	chr9	133273813	C	T	0.028952
chr5	155268614	T	G	-0.008019795	chr9	133273983	A	G	0.128335
chr5	155268764	T	C	-0.001829788	chr9	133274293	AC	A	-0.03813
chr5	162483682	C	G	0.018964532	chr9	133274295	A	T	-0.03873
chr5	162485210	G	A	0.01702547	chr9	133274414	A	G	0.041961
chr5	173251980	T	C	0.001743214	chr9	133279427	T	C	0.081206
chr6	166219009	A	G	0.017561279	chr9	133317947	G	A	0.077365
chr7	66172694	C	T	0.010034199	chr9	134936064	G	A	0.013608
chr8	139152871	T	C	-0.001814465	chr12	7519002	A	T	0.021549
chr9	10515244	C	T	-0.034061173	chr15	27576869	G	T	-0.07486
chr9	20111606	T	C	0.041389426	chr16	26255328	T	A	0.006537
chr9	20123970	C	T	0.003156217					

Discussion

- Subset of proteins with trans variants showed higher prediction quality
- Future directions: **Step 5. Association study of Predicted Protein Level and phenotype on external data without observed protein level**
- 7 phenotypes in the rest of the WHI samples without measured protein:
  - Coronary Heart Disease (CHD), Congestive Heart Failure (CHF), Myocardial Infarction (MI), Deep Vein Thrombosis (DVT), Hemorrhagic Stroke (STRKHEMO), Ischemic Stroke (STRKISCH), Pulmonary Embolism (PE)
- Other TWAS methods (FUSION, TIGAR)
- Polygenic risk score approach

References:

1. Brandes, N., Linial, N. & Linial, M. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. *Genome Biol* **21,** 173 (2020).

2.Jonathan D. Mosley., Mark D. Benson., Robert E. Gerszten., Thomas J. Wang. et al. Probing the Virtual Proteome to Identify Novel Disease Biomarkers. *Circulation* Vol. 138, No. 22 (2018)

3. Justin M. Luningham., Junyu Chen., Shizhen Tang., Philip L. De Jager., David A. Bennett., Aron S. Buchman., Jingjing Yang. Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *AJHG* Vol. 107, Issue 4, 714-726 (2020)

4. Gamazon, E., Wheeler, H., Shah, K. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091–1098 (2015).

Supplementary Slides

Methods – Proposed method Details

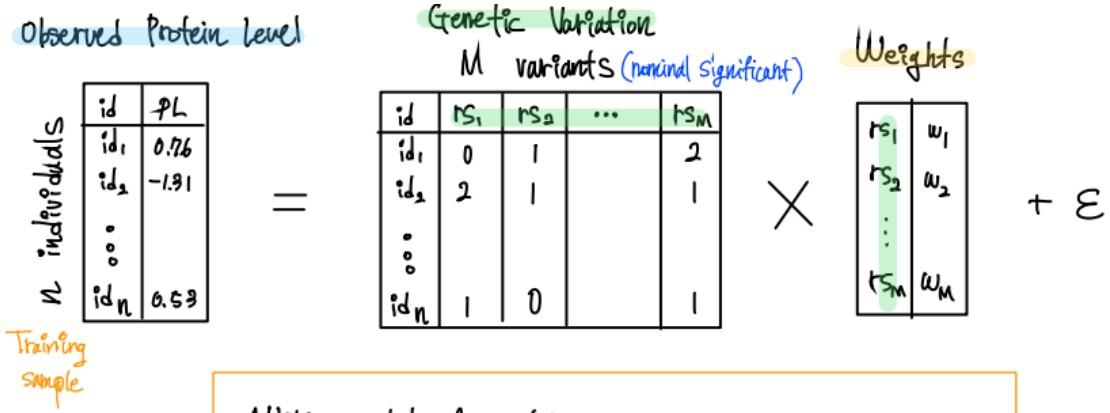
Step 2: Proteomic GWAS to select nominally significant variants

Protein Level ~ Age + GWAS Array + plate + single variant

- Inverse Normalize raw protein level
- EPACTS v3.3.0 EMMAX (Efficient Mixed Model Association eXpedited) test accounting for sample structure including population structure and hidden relatedness
- Variants with nominal significant p-value < 1E-04 were selected for the next step (pQTL)

Methods – Proposed method Details

Step 3. EN model training on nominally significant variants



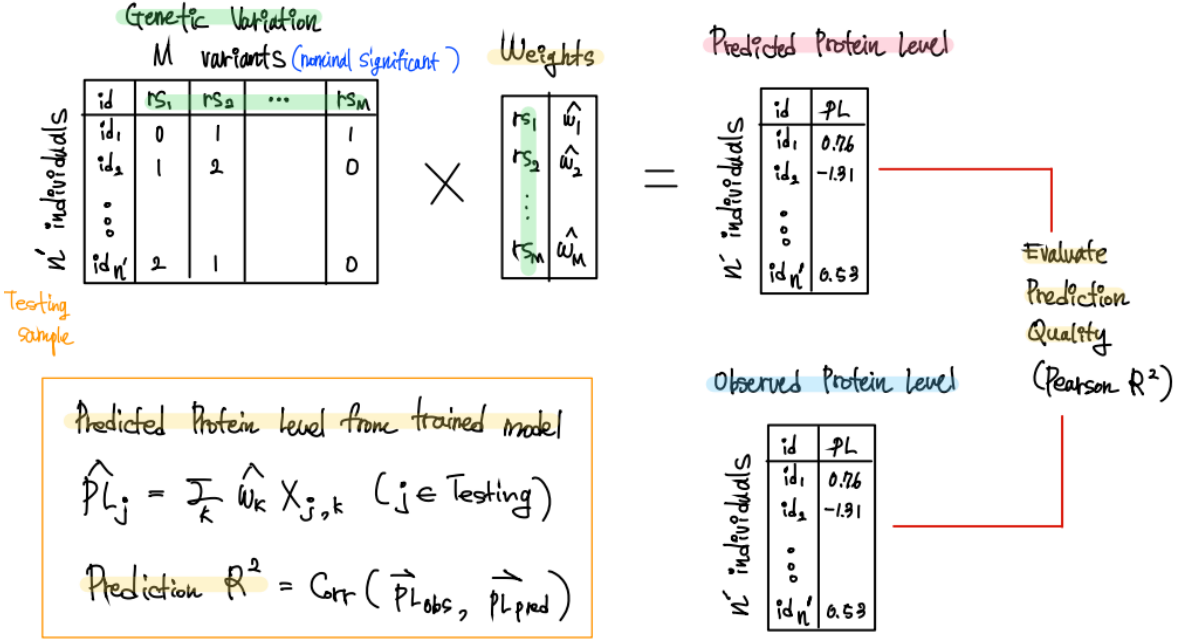
Additive model of protein abundance level trained in reference data set

$$PL_i = \sum_k w_k x_{i,k} + \varepsilon_i \quad (i \in \text{Training})$$

$x_{i,k}$: variant dosage

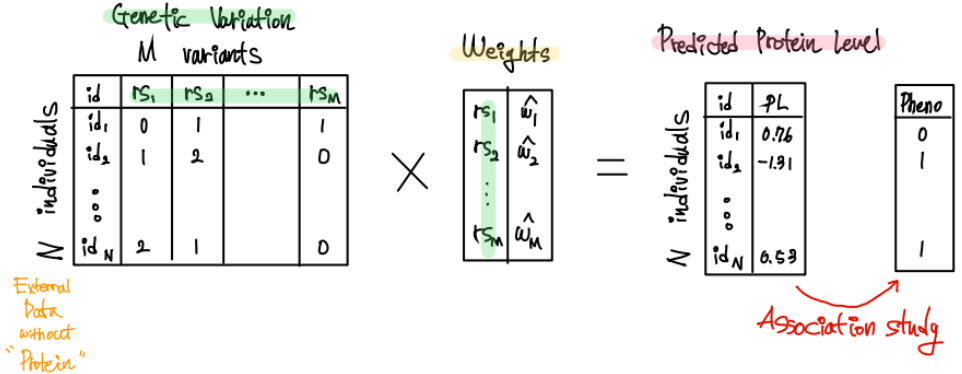
Methods – Proposed method Details

Step 4. Model evaluation on testing samples



Methods – Proposed method Details

Step 5 (Future). Association study of Predicted Protein Level and phenotype



Association study of phenotype by Predicted Protein Level

$$\text{logit}[P(Y_i=1)] = \beta_0 + \beta_1 PC_i + \beta_2 Age_i + \beta_3 \widehat{PL}_i$$

$(i \in \text{External Data})$