Note taking from paper

GWAS – a research approach identify genomic variant → identify genomic variants

↳ GWAS => statistical based?

↳ statistically linked with a risk of disease or a trait.

PredixCan => gene based?

↳ estimated components of gene expression → determine by invidual genetic profile.

↳ transcrition

↳ translation

↳ correlate the "imputed" gene expression with the phenotype* under investigation

phenotype = trait = disease*

input = assign

involed = related = relevant

etiology = reason

PredixCan → identify genes that relevant with the reson of the trait

Whole genome tissue-model

↳ train with reference transcripton data

↳ predict gene expression

→ is this low gene expression prediction work before PredixCan?

PredixCan => more understand about the biology mechanism??

=> What is the benefit of it? → less testing burden.

↳ approach to the design of follow up experience

PredixCan → detect genes related with the disease

↳ provide insight into the mechanism of the association.

GWAS → successful in identifying the susceptibility loci for complete disease

↓

What GWAS have done so far

→ conduct single variant tests → test on the variant of a single gene.

→ intergorate genome in an agnostic way?

→ make the sample size larger → better inteference.

Less successful in improve on simple approaches → GWAS??

Genome-wide significant → in stats => mahathan plot?

modest = low

→ low propotion of variance in dieseas risk.

GWAS of disease susceptibility → relevant gene affect is small

pharmacologic trait → result in lim conclusive finding on gene factor contribute to complete trait

Functional significant of discovered loci => unclear

GWAS still identify susceptibility loci

→ how to understand the biology mechanism

Gene based approach + multi marke → replace single variant test → not significant def.

↳ conduct test of assosiation on biologycally informed of SNPs

↳ reduce testing penalty

The collab of omic data → extend GWAS by addressing function gap

trait / disease under significant genetic regulatio

Gene-based study ⇒ lacking

↳ improve knowledge.

⇒ suggest PrediXCan

↳ detect trait based on gene data.

PrediXCan → identify diesease based on the mechanism of gene.

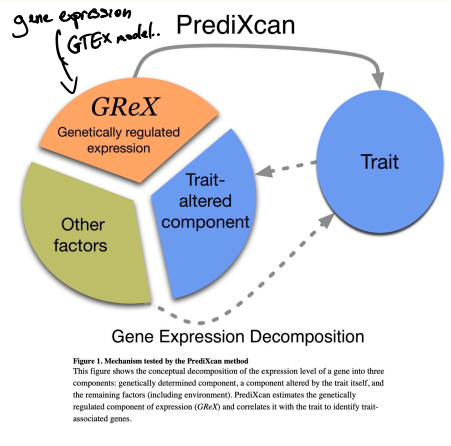


Figure 1. Mechanism tested by the PrediXcan method
This figure shows the conceptual decomposition of the expression level of a gene into three components: genetically determined component, a component altered by the trait iself, and the remaining factors (including environment). PrediXcan estimates the genetically regulated component of expression (*GReX*) and correlates it with the trait to identify trait-associated genes.

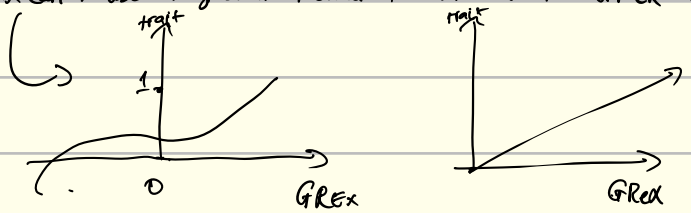
SNP = single Nucleotide Polymorphism → a type of genetic variation

Genotype impuatation work like ML.

Use prediction model to assign gene expression

↳ use PrediXCan to determine trait.

PrediXCan : use regression method to work with GReX data

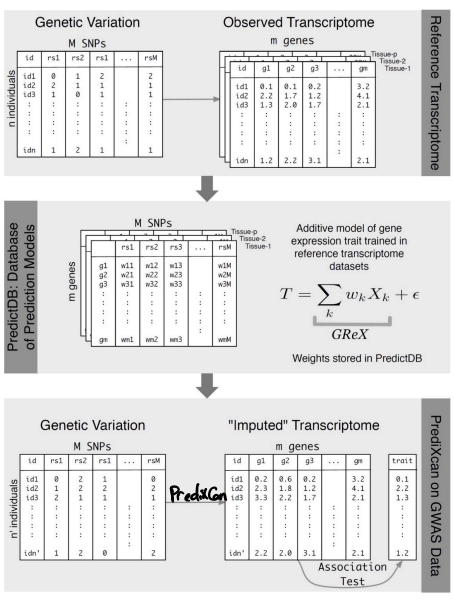


attenuation: decrease

Aware that the attenuation bias because error in prediction

↳ does not affect the analysis

Do more about GReX



SNPs → gen expression → GReX

↓

a ML model

Additive model: build a data base of prediction model

⇒ unclear

PrediXCan contribute to psychiatric + neurodevelopmental disorder

↳ identify gene target for therapy

↳ gene based, better understand the mechanism → direct insight

Pros: less test burden

more straight forward for information prior + grouping of functional unit

No transcriptome data needed? ⇒ less complex

the result is simplified

Multiple tissue can be evaluate

Can be mass apply

The prediction model was build in the DGN whole blood cohort using LASSO, elastic net ($\alpha = 0.5$), variant p-value

↳ a regression model

↓

the risk of developing trait

Asses: Judge lasso

↳ 10-fold cross-validation → use for large data set

↳

test	—	—	—
—	test	—	—
—	—	test	—
—	—	—	test

Estimate the heritability of gene expression in DGN attributable to SNPs using a mixed-effect model:

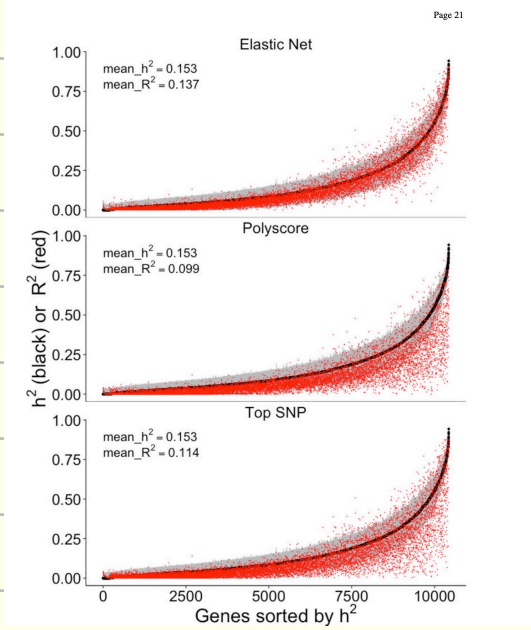
↳ only using local SNPs

↳ benchmark for prediction R^2

↳ compare the benchmark with R^2

↳ .153 & .137 ⇒ close

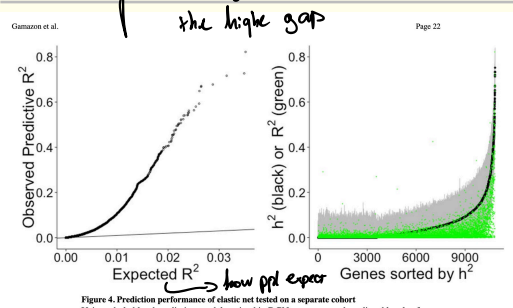
Benchmark = Standard



Test the prediction model train by DGN blood

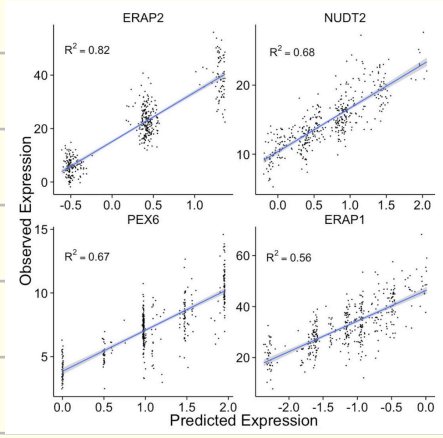
↳ predict gene expression level

↳ the higher expect the higher gap



↳ how ppl expect

Figure 4. Prediction performance of elastic net tested on a separate cohort
Using whole blood prediction models trained in DGN, we compared predicted levels of expression with observed levels on lymphoblastoid cell lines from the 1000 Genomes project. RNA-sequenced data (n=421) on these cell lines have been made publicly available by the GEUVADIS consortium. Left panel shows the squared correlation, R^2, between predicted and observed levels plotted against the null distribution of R^2 Right panel shows prediction performance (R^2 of GReX vs. observed expression in green) compared to GEUVADIS gene expression heritability (h^2) estimates (black with 95% confidence interval in gray).

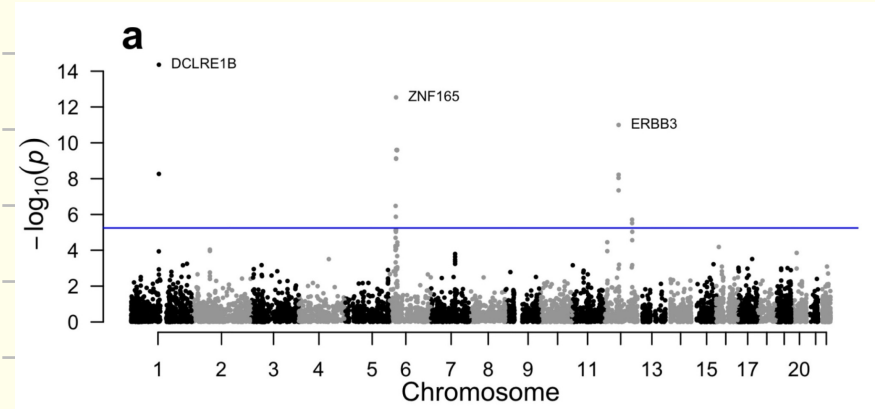


is the ≠ in value due to the lack of data? or the variant of data?

Apply PrediXcan into WTCCC

↳ correlated gene expression of 867 gene → identify 41 association with 5 diesease

Disease	Gene	Evidence	Chr	TSS	PrediXcan Z-Statistic	PrediXcan P-value	No. of SNPs in Predictor	Cross Validated Prediction R^2
RA	*DCLRE1B*	V	1	114,447,763	−6.68	2.46E-11	4	0.0388
RA	*PTPN22*	G	1	114,356,433	5.67	1.44E-08	32	0.0795
BD	*PTPRE*		10	129,705,325	4.94	7.71E-07	38	0.0355
CD	*ATG16L1*	G	2	234,118,697	6.37	1.94E-10	20	0.0638
CD	*IL23R*	G	1	67,632,083	5.23	1.74E-07	38	0.0378
CD	*APEH*	V	3	49,711,435	5.14	2.77E-07	31	0.1164
CD	*ZNF300*	G ‡	5	150,273,954	−4.98	6.29E-07	34	0.0387
CD	*NKD1*	G	16	50,582,241	−4.91	8.91E-07	43	0.0693
CD	*BSN*	G ‡	3	49,591,922	−4.68	2.89E-06	39	0.2336
CD	*GPX1*	V	3	49,394,609	−4.62	3.87E-06	28	0.0211
CD	*SLC22A5*	G ‡	5	131,705,444	−4.54	5.75E-06	42	0.6356
HT	*KCNN4*		19	44,270,685	−4.7	2.62E-06	81	0.4655
T1D	*DCLRE1B*	V	1	114,447,763	−7.84	4.34E-15	4	0.0388
T1D	*ZNF165*	M	6	28,048,753	7.3	2.92E-13	19	0.0374
T1D	*ERBB3*	G	12	56,473,641	−6.81	1.01E-11	9	0.2206
T1D	*EGFL8*	H	6	32,132,360	6.33	2.52E-10	36	0.0558
T1D	*C6orf136*	H	6	30,614,816	−6.33	2.52E-10	15	0.0137
T1D	*HCG27*	H	6	31,165,537	−6.33	2.52E-10	81	0.3721
T1D	*GTF2H4*	H	6	30,875,961	6.33	2.52E-10	69	0.0982
T1D	*DDR1*	H	6	30,844,198	6.33	2.52E-10	48	0.1427
T1D	*AGER*	H	6	32,148,745	−6.33	2.52E-10	39	0.0502
T1D	*POU5F1*	H	6	31,130,253	6.33	2.52E-10	45	0.2874
T1D	*ATP6V1G2*	H	6	31,512,239	6.33	2.52E-10	95	0.2543
T1D	*TUBB*	H	6	30,687,978	6.33	2.52E-10	56	0.0295
T1D	*AIF1*	H	6	31,582,961	6.33	2.52E-10	34	0.039
T1D	*CYP21A2*	H	6	32,006,042	−6.33	2.52E-10	80	0.229
T1D	*LSM2*	H	6	31,765,173	6.33	2.52E-10	31	0.0317
T1D	*VARS2*	H	6	30,876,019	6.33	2.52E-10	87	0.3628
T1D	*APOM*	H	6	31,620,193	−6.33	2.52E-10	58	0.0699
T1D	*DDAH2*	H	6	31,694,815	−6.33	2.52E-10	32	0.1943
T1D	*NCR3*	H	6	31,556,672	−6.33	2.52E-10	79	0.2548
T1D	*ZSCAN16*	M	6	28,092,338	6.16	7.37E-10	34	0.0291
T1D	*ZKSCAN4*	M	6	28,212,401	6.15	7.73E-10	17	0.0991
T1D	*PTPN22*	G	1	114,356,433	5.83	5.41E-09	32	0.0795
T1D	*RPS26*	G ‡	12	56,435,637	5.82	6.00E-09	23	0.0719
T1D	*GDF11*	V	12	56,137,064	−5.75	9.11E-09	39	0.0341
T1D	*SUOX*	G ‡	12	56,390,964	−5.47	4.49E-08	50	0.1339
T1D	*BTN3A2*	M	6	26,365,387	−5.11	3.30E-07	49	0.7662
T1D	*PRSS16*	M	6	27,215,480	4.83	1.34E-06	31	0.1639
T1D	*FAM109A*	V	12	111,798,339	−4.76	1.94E-06	17	0.0665
T1D	*SH2B3*	G	12	111,843,752	4.67	3.05E-06	26	0.0368



x: represent genomic location

y: $-\log_{10}(P)$ => smaller p => stronger association

the horizontal line: significan threshold

↳ determine by ppl?

ppl take gene from multiple disease

Discussion

Gene expression → phenotype between genetic variation & higher level phenotype

PrediXcan → use the information derived from gene data to map trait-association loci

└→ test the mediating effect of gene expression level by qualifying the relation between genetically expression and the traits.

How to implement: developed prediction model using large scale data base

→ choose elastic net model. Performe similar to LASSO

Combination of KNN and elastic net can improve the peformance

PrediXcan provide the direction of effect → opportunities for therapie

Prediction model can be applied into large bio database to gain knowledge of side effect of a drug of a specific gene

Benefit of direction effect: improve the interpretation of sequence analyze

Bonferroni correction: a stats methode use reduce the change of false positive when performing multihypothesis test

Causal: original

Method:

Additive model for trait

$$y_g = \sum_k w_{kjg} X_k + \epsilon$$

| ⇒ is it combined multiple tissuse to lead to a trait?

y_g: trait of gene g.

W_{kjg}: effect size → estimated using elastic net & LASSO

X_k: number of reference allen of marke l

Calculate the proportion of the variance of gene expression

$$Y = Xb + G_{local} + e$$

Y: gene expression trait

Estimation of GReX

$$GReX_g = \sum_k \hat{w}_{kjg} X_k$$

| is the process repeated?

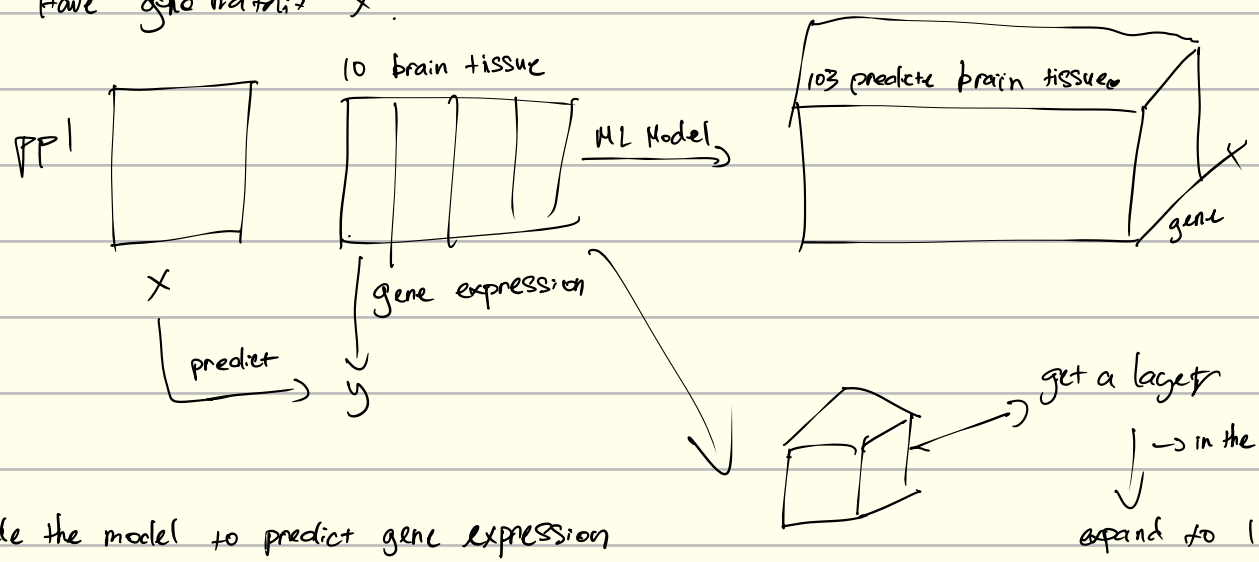
What I have gain?

- More understand about the paper
  - ↳ know what is the model do, how it work?
  - ↳ learn a lot of this (biology, statistics,..)
- Question:
  - ↳ Could you mind to repeat what you want to do with this?
    - ↳ Applied into a differ disease database?
      - ↳ brain disease, ...
  - ↳ Do you have an expected deadline of this project? Mine is 1 week before the class start
    - ↳ easier to break it down.
  - ↳ What step I need to do next?
    - ↳ where can I get the data.

First Step. Prev

Start with genotype infor (GTeX)

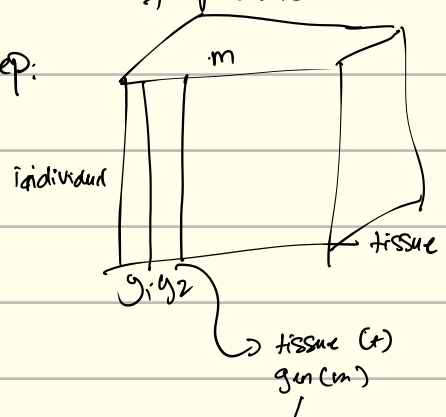
Have geno matrix X.

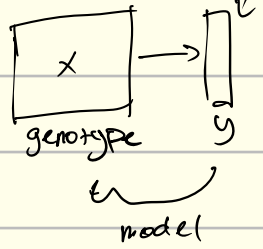


Made the model to predict gene expression from 103

Expaned version of prediXcan => have more detail of brain tissue

Next step:



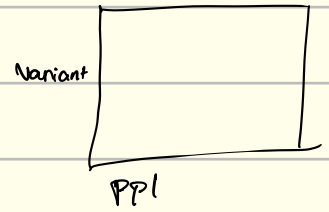


repeat for every y.

model

newperson → gene expression + addition infor

Meet with other ppl's next week → how to download the data.

the gene have a location

Make 1 model for every combination of gene.