

Convergence of coronary artery disease genes onto endothelial cell programs

Schnitzler, G.R., Kang, H., Fang, S. et al.

Nature 626, 799–807 (2024).

<https://doi.org/10.1038/s41586-024-07022-x>

Purpose

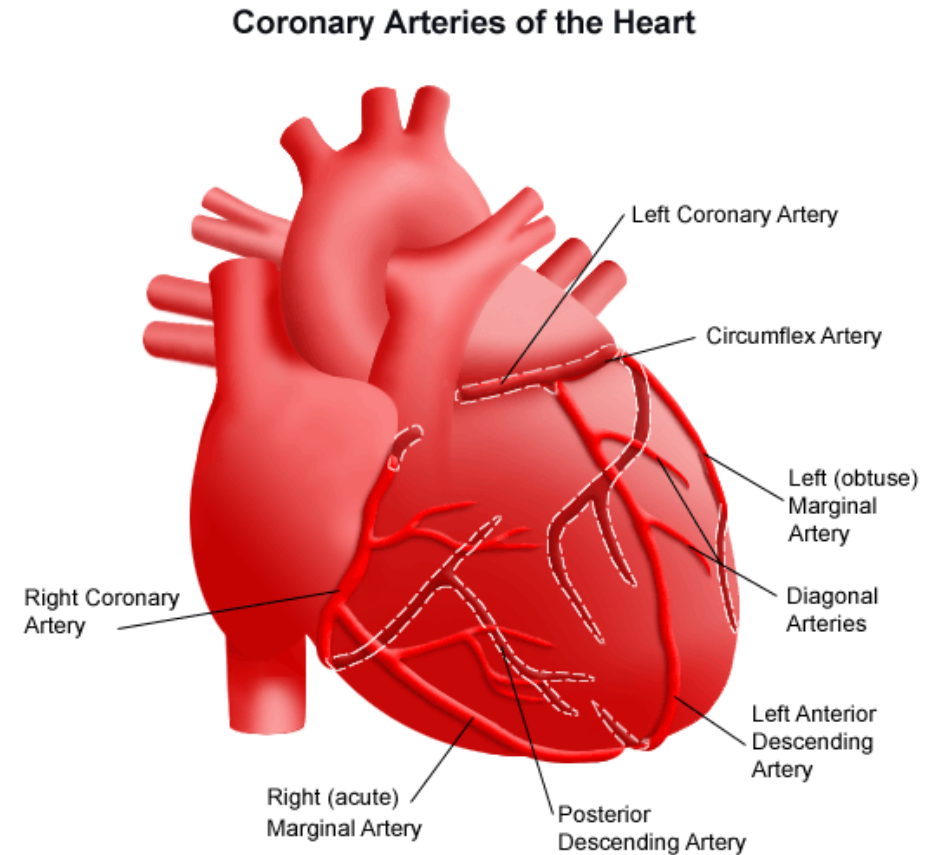
Main Target : Coronary Artery Disease (CAD)

Coronary arteries that supply blood to the heart **become narrowed or blocked.**

Common (5%) and Heritable Disease

Could be caused by **cholesterol uptake, efflux, smooth muscle cell responses.**

The causes are influenced by **genetic program at Endothelial Cell (ECs)**



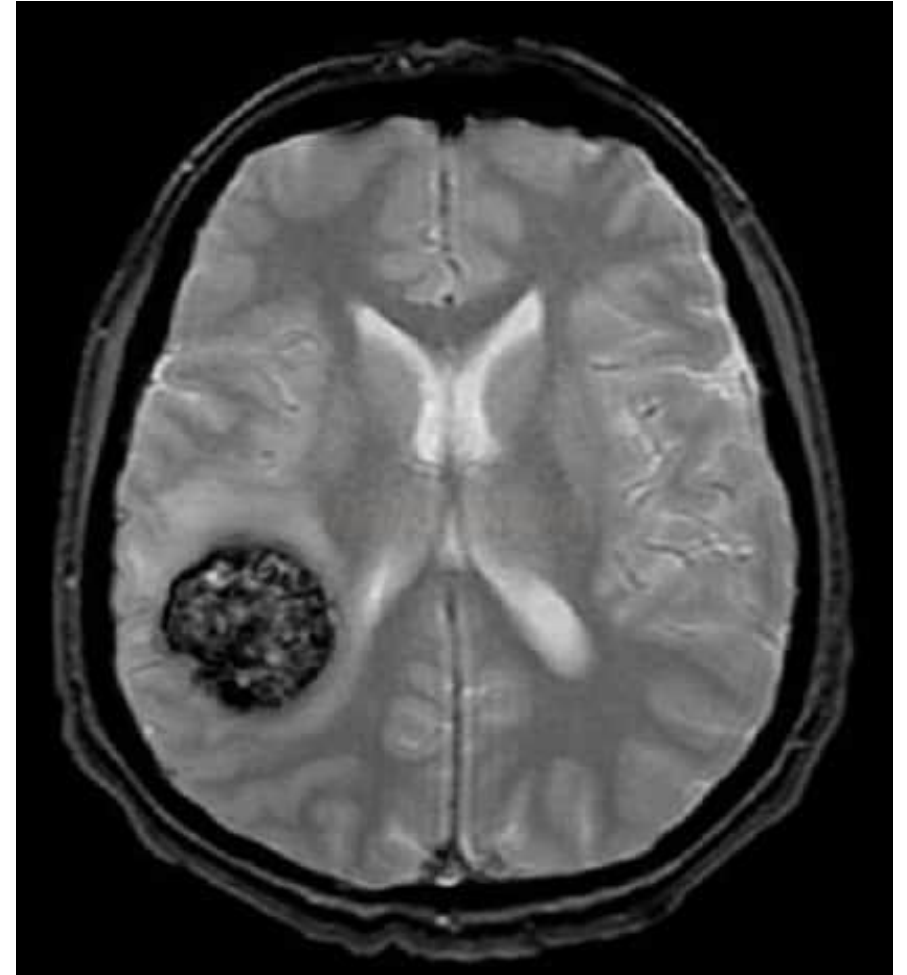
Sub Target : Cerebral Cavernous Malformation (CCM)

The lesion consists of clusters of blood vessels that are tangled together, and blood is pooled inside these vessels.

Rare but Heritable Disease

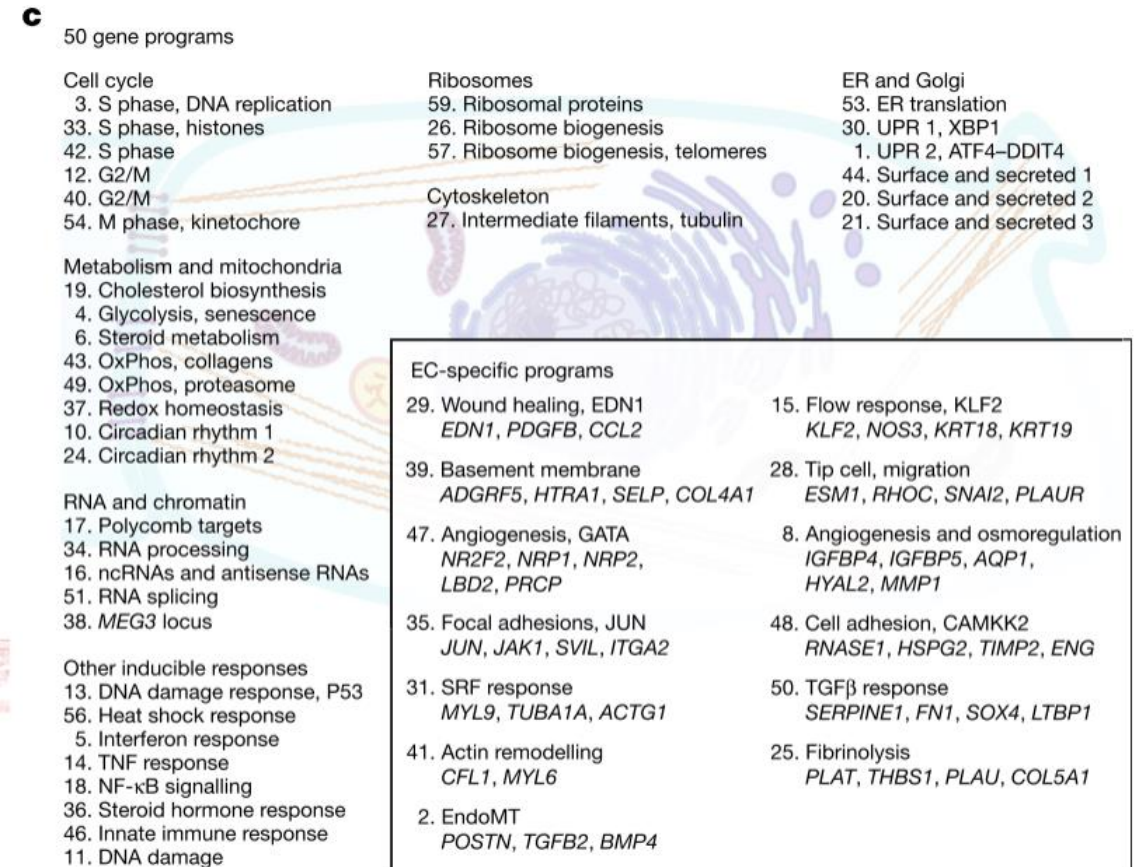
Highly Correlated with CCM1 (KRIT1), CCM2 (MGC4607) and CCM3 (PDCD10) Genes. (CCM Complex)

The authors found **the genetic connection between CCM and CAD**



What to see : Genetic Program

In this paper, the genetic program means **a collection of genes that collaborate to achieve a specific function.**



Discover genetic programs of gene associated with CAD and CCM

Why this is important?

They setup a pipeline to research link of Variant to Gene to Program.

They found the 5 programs about CAD

They found connection between CAD and CCM.

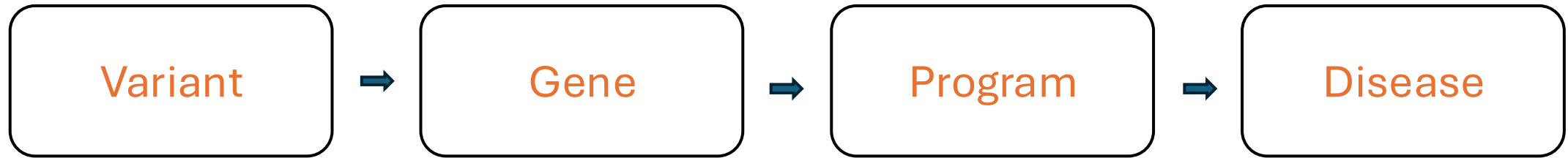
They found a gene poorly studied before but highly influential on CAD and CCM.

They studied the function of the Key genes in various way from Machine Learning to in-vivo experiment.

Methods

How to see :

V2G2P Approach (V2G2P : **V**ariant to **G**ene to **P**rogram)



First. Identify a cell type and cellular model relevant to disease.

Second. **V2G** : link target variant to genes.

Third. **G2P** : link genes to program.

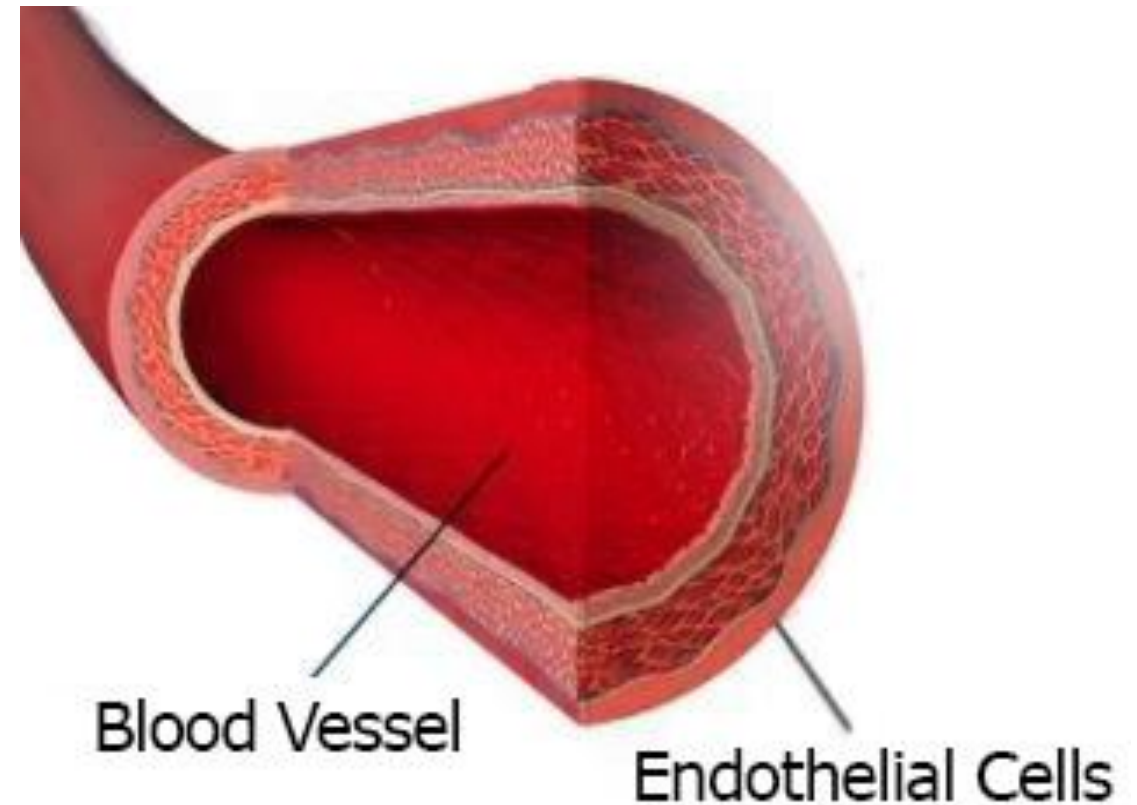
Fourth. Identify disease-associate programs.

Fifth. Study the genes in disease associated programs.

First. Identify a cell type and cellular model relevant to disease.

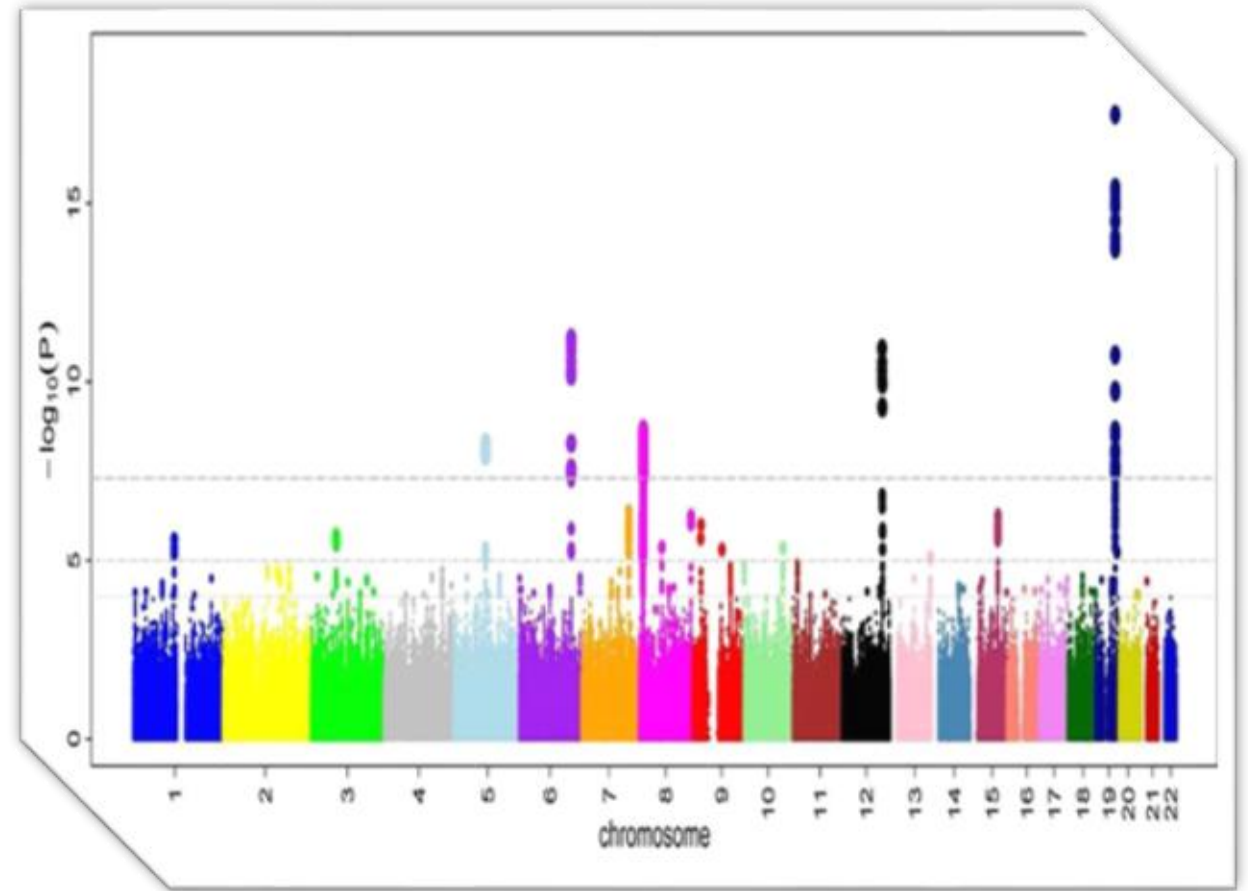
The genetic activity in **ECs** (Endothelial Cells) is one of the major cause of CAD

They use telomerase-immortalized human aortic ECs (**teloHAECs**)



Second. V2G : link target variant to genes

From non-lipid CAD GWAS,
pick 228 variant signal

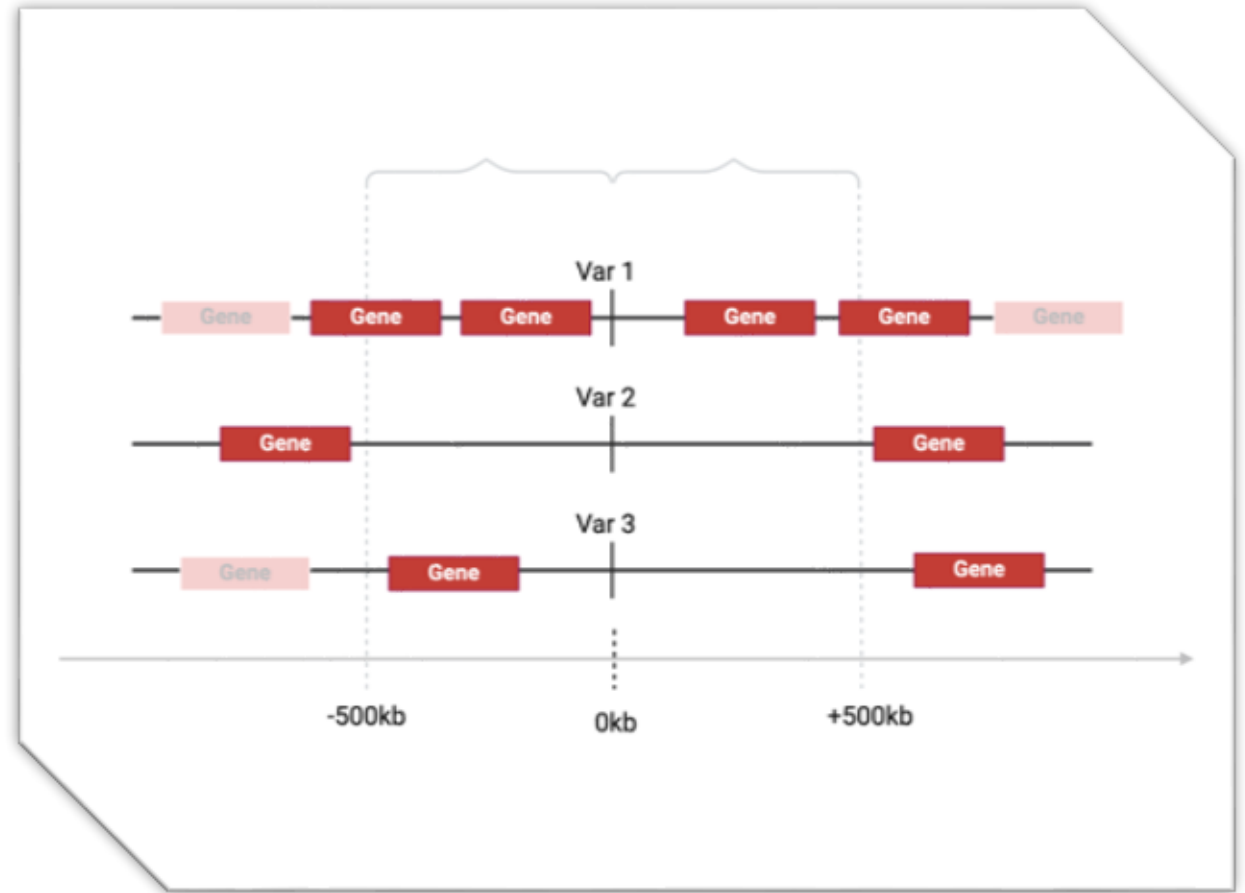


Second. V2G : link target variant to genes

From non-lipid CAD GWAS,
pick 228 variant signal

For each variant, pick genes satisfying
following conditions along variant loci

- Closest genes on either side
- Every gene in ± 500 kb at the loci



Second. V2G : link target variant to genes

From non-lipid CAD GWAS,
pick 228 variant signal

For each variant, pick genes satisfying
following conditions along variant loci

- Closest genes on either side
- Every gene in ± 500 kb at loci

Setup a dictionary linking each variant
with genes



Second. V2G : link target variant to genes

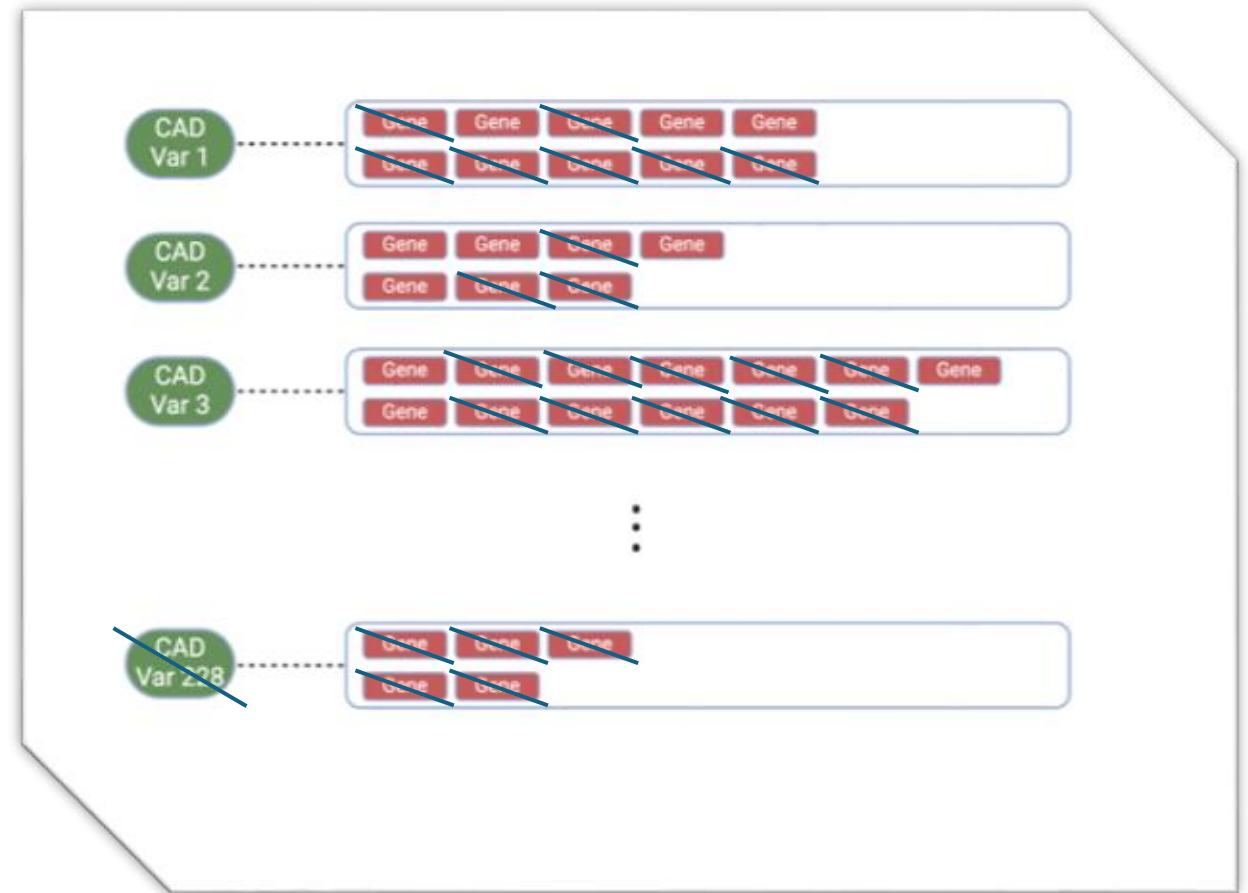
From non-lipid CAD GWAS,
pick 228 variant signal

For each variant, pick genes satisfying
following conditions along variant loci

- Closest genes on either side
- Every gene in ± 500 kb at loci

Setup a dictionary linking each variant
with genes

Using the Activity-by-Contact (ABC) Model,
filter out the genes.



Activity-by-Contact (ABC) Model

Model which give a connectivity score between Gene and Element (enhancer, snp.. etc)

$$ABC_{E,G} = \frac{A_E \times C_{E,G}}{\sum_{e \in Neighbor(G, 5Mb)} A_e \times C_{e,G}}$$

A_E : The Strength of E-th Enhancer

$C_{E,G}$: The distance from E-th Enhancer to G-th Gene

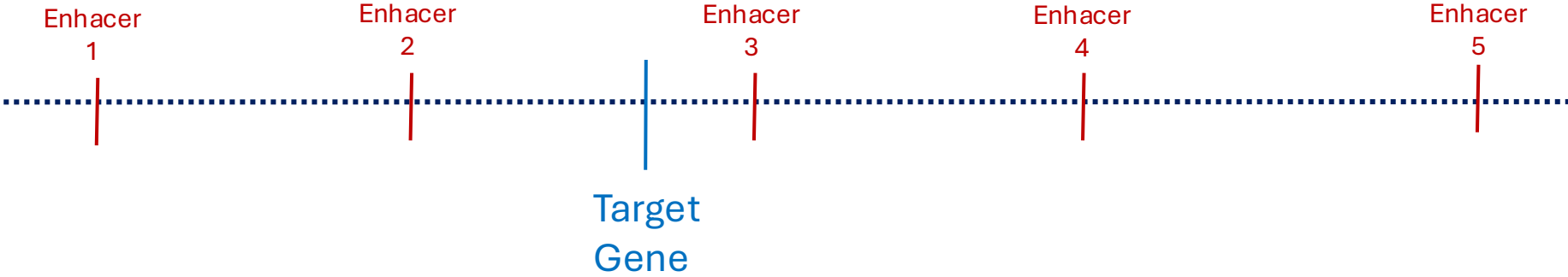
In this paper, they used ABC scores from other references
(`engothelial_cell_of_umbilical_vein_Zhang2013`)

Activity-by-Contact (ABC) Model

Enhancer	Enhancer 1	Enhancer 2	Enhancer 3	Enhancer 4	Enhancer 5
A (Power)	1	2	3	4	5

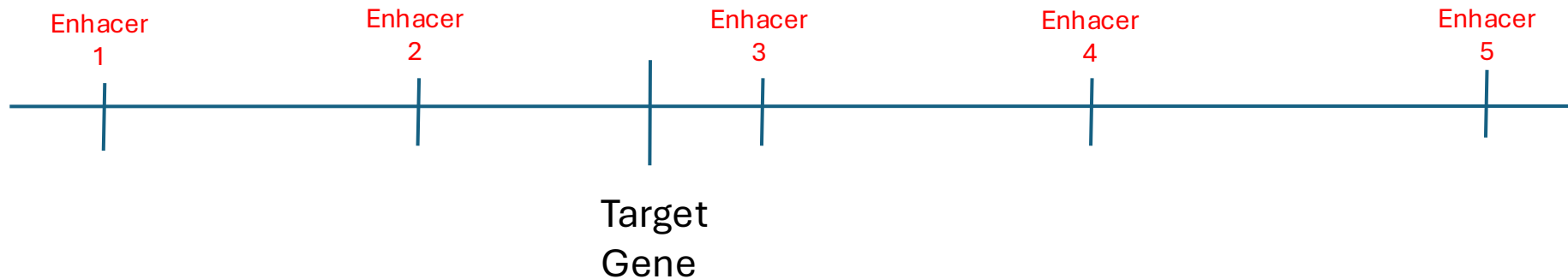
Activity-by-Contact (ABC) Model

Enhancer	Enhancer 1	Enhancer 2	Enhancer 3	Enhancer 4	Enhancer 5
A (Power)	1	2	3	4	5
C (Weight with distance)	2	4	5	3	1



Activity-by-Contact (ABC) Model

Enhancer	Enhancer 1	Enhancer 2	Enhancer 3	Enhancer 4	Enhancer 5
A (Power)	1	2	3	4	5
C (Weight with distance)	2	4	5	3	1
A * C	2	8	15	12	5
ABC score	0.05 (2/42)	0.17 (8/42)	0.36 (15/42)	0.29 (12/42)	0.12 (5/42)



Activity-by-Contact (ABC) Model



As a result, they got 254 of Genes linked with 125 CAD GWAS Variant.

Third. G2P : link genes to specific program.

Generating

- CRIPRi Perturbed-Seq

Processing

- Consensus Nonnegative Matrix Factorization (cNMF)

Interpreting

- Annotation with manual curation and supplementary analysis

Third. G2P : link genes to specific program.

CRISPRi Perturbed-Seq

It's the technique to study how expression of genes are changed after suppressing the activity of a specific gene.

The total number of cells is 214,449

Total 37,637 gRNA are used

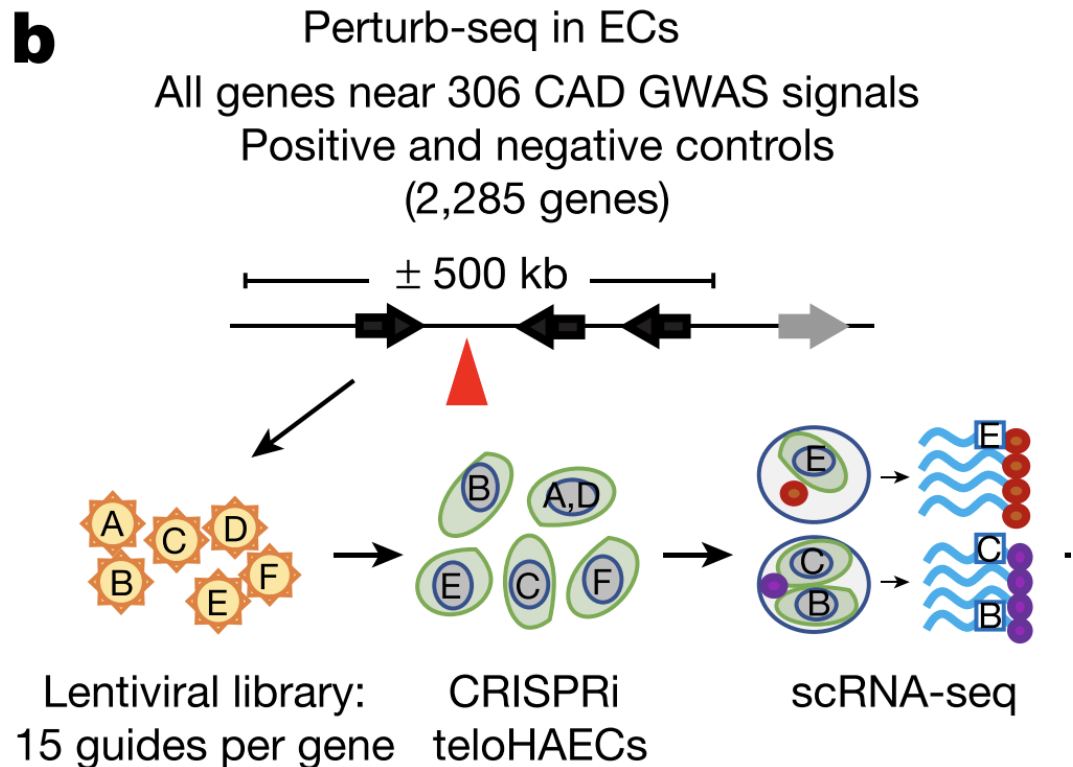
1,661 Genes near all 306 CAD GWAS

624 Control genes

100 Random gRNAs

(15 x (1,661+624) + 1000).

b



Third. G2P : link genes to specific program.

	gRNA 1	gRNA 2	...	gRNA 2285	Gene 1	Gene 2	...	Gene 20k
Cell A	Perturb	Control	...	Control	3	0	...	2
Cell B	Perturb	Control	...	Control	2	0	...	3
Cell C	Control	Perturb	...	Control	0	2	...	1
Cell D	Control	Perturb	...	Control	0	5	...	8
Cell E	Control	Control	...	Control	1	0	...	3
Cell F	Control	Control	...	Control	0	0	...	1
Cell G	Control	Perturb	...	Control	1	5	...	4
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...



Comp 1	Comp 2	...	Comp 60
0.03	0.12	...	0.12
0.01	1.32	...	1.32
0.12	2.13	...	2.13
1.32	0.03	...	0.03
2.11	0.01	...	2.13
2.13	0.12	...	0.12
0.42	1.32	...	1.32
...	...	...	...
...	...	...	...

By using the Consensus Non-negative Matrix Factorization (cNMF), they compress the 20k gene expression information into 60 component factors.

cf) Factor Analysis

Extracting abstract general factors containing much larger set of information.

ID	Q1	Q2	Q3	Q4	Q5
A	1	4	1	5	7
B	2	3	2	4	7
C	3	2	2	5	7
D	4	5	2	2	7
E	5	6	3	3	7
F	6	7	1	1	7
G	7	8	4	2	7

Q1. How many hours do you spend running each week?

Q2. How many hours do you spend swimming each week?

Q3. How many hours do you spend reading each week?

Q4. How many hours do you spend studying each week?

Q5. How many hours do you work in each day?

cf) Factor Analysis

Extracting abstract general factors containing much larger set of information.

ID	Q1	Q2	Q3	Q4	Q5
A	1	4	1	5	7
B	2	3	2	4	7
C	3	2	2	5	7
D	4	5	2	2	7
E	5	6	3	3	7
F	6	7	1	1	7
G	7	8	4	2	7

Q1. How many hours do you spend running each week?

Q2. How many hours do you spend swimming each week?

Q3. How many hours do you spend reading each week?

Q4. How many hours do you spend studying each week?

Q5. How many hours do you work in each day?

Hobby Factor

Job Factor

Factor Score of A = (1+4+1, 5+7) = (6, 12)

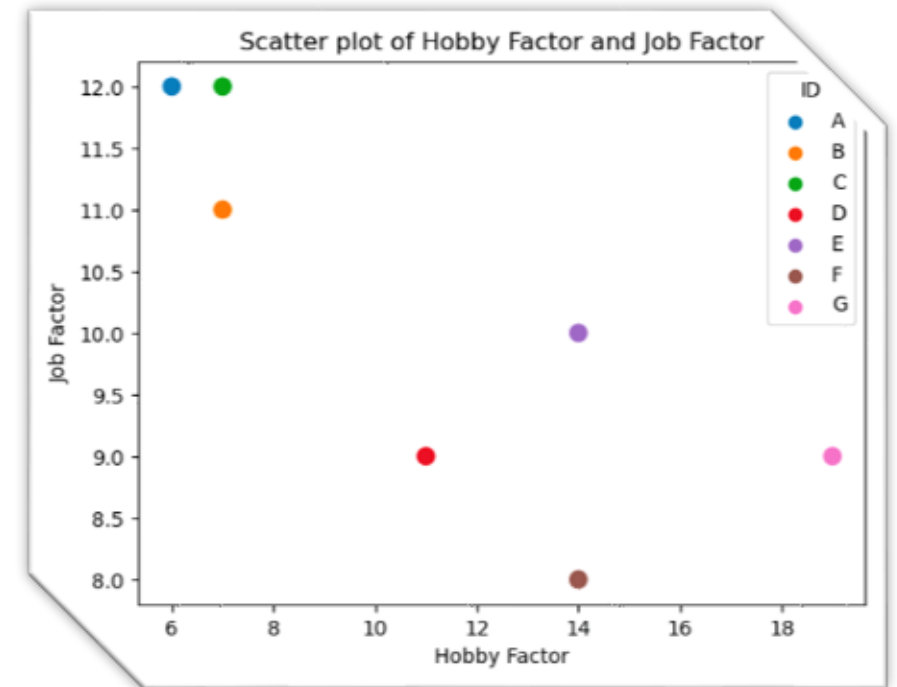
Factor Score of G = (7+8+4, 2+7) = (19, 9)

cf) Factor Analysis

This is the most basic and easiest way to do the Factor analysis

ID	Q1	Q2	Q3	Q4	Q5
A	1	4	1	5	7
B	2	3	2	4	7
C	3	2	2	5	7
D	4	5	2	2	7
E	5	6	3	3	7
F	6	7	1	1	7
G	7	8	4	2	7

ID	Hobby	Job
A	6	12
B	7	11
C	7	12
D	11	9
E	14	10
F	14	8
G	19	9



cf) Factor Analysis

We can make more plexible factor using the weights

Hobby Factor

Q1. How long do you run in a week?
X 0.9

Q2. How long do you swim in a week?
X 0.8

Q3. How long do you read a book in a week?
X 0.5

Q4. How long do you study in a week?
X -0.6

Q5. How long do you work in a week?
X -0.4

Intellectual Factor

Q1. How long do you run in a week?
X -0.3

Q2. How long do you swim in a week?
X -0.3

Q3. How long do you read a book in a week?
X 0.5

Q4. How long do you study in a week?
X 0.7

Q5. How long do you work in a week?
X 1.0

$$\text{Factor Score} = (Q1 * 0.9 + Q2 * 0.8 + Q3 * 0.5 + Q4 * -0.6 + Q5 * -0.4 , \\ Q1 * -0.3 + Q2 * -0.3 + Q3 * 0.5 + Q4 * 0.7 + Q5 * 1.0)$$

cf) Factor Analysis

This process could be expressed with Matrix

ID	Q1	Q2	Q3	Q4	Q5
A	1	4	4	5	7
B	2	3	3	4	7
C	3	2	4	5	7
D	4	5	2	2	7
E	5	6	3	3	7
F	6	7	1	1	7
G	7	8	4	2	7



ID	Hobby	Job
Q1	0.9	-0.3
Q2	0.8	-0.3
Q3	0.5	0.5
Q4	-0.6	0.7
Q5	-0.4	1.0



ID	Hobby	Job
A	0.8	4.5
B	1.6	5.3
C	1.4	3.5
D	5.4	4.7
E	6.2	7.3
F	8.1	4.3
G	11.1	4.9

$$\text{Factor Score} = (\text{Q1} * 0.9 + \text{Q2} * 0.8 + \text{Q3} * 0.5 + \text{Q4} * -0.6 + \text{Q5} * -0.4, \\ \text{Q1} * -0.3 + \text{Q2} * -0.3 + \text{Q3} * 0.5 + \text{Q4} * 0.7 + \text{Q5} * 1.0)$$

cf) Factor Analysis

This process could be expressed with Matrix

ID	Q1	Q2	Q3	Q4	Q5
A	1	4	4	5	7
B	2	3	3	4	7
C	3	2	4	5	7
D	4	5	2	2	7
E	5	6	3	3	7
F	6	7	1	1	7
G	7	8	4	2	7



ID	Hobby	Job
Q1	0.9	-0.3
Q2	0.8	-0.3
Q3	0.5	0.5
Q4	-0.6	0.7
Q5	-0.4	1.0



ID	Hobby	Job
A	0.8	4.5
B	1.6	5.3
C	1.4	3.5
D	5.4	4.7
E	6.2	7.3
F	8.1	4.3
G	11.1	4.9

$$O \times W = F$$

cf) Factor Analysis

This process could be expressed with Matrix

ID	Q1	Q2	Q3	Q4	Q5
A	1	4	4	5	7
B	2	3	3	4	7
C	3	2	4	5	7
D	4	5	2	2	7
E	5	6	3	3	7
F	6	7	1	1	7
G	7	8	4	2	7



ID	Hobby	Job
A	0.8	4.5
B	1.6	5.3
C	1.4	3.5
D	5.4	4.7
E	6.2	7.3
F	8.1	4.3
G	11.1	4.9



	Hobby	Job
Q1	0.9	-0.3
Q2	0.8	-0.3
Q3	0.5	0.5
Q4	-0.6	0.7
Q5	-0.4	1.0

-1

$$O \times W = F \Rightarrow O \approx F \times W^{-1} \Rightarrow O \approx F \times L$$

cf) Factor Analysis

F - Factor Score Matrix :

Matrix containing the Factors Values of each Samples (Cell, Variant)

L – Factor Loading Matrix :

Matrix containing the relation between Factors and the Original Features

Most of the case, the Factor Loading is Correlation between Original Features and Factors

$$\text{Cov}(O, F) = \text{Cov}(FL + E, F) = L^t \text{Cov}(F, F) + \text{Cov}(F, E) = L^t$$

We can interpret any FA result based on this.

cf) Factor Analysis

This process could be expressed with Matrix

ID	Q1	Q2	Q3	Q4	Q5
A	1	4	4	5	7
B	2	3	3	4	7
C	3	2	4	5	7
D	4	5	2	2	7
E	5	6	3	3	7
F	6	7	1	1	7
G	7	8	4	2	7



ID	Hobby	Job
A	0.8	4.5
B	1.6	5.3
C	1.4	3.5
D	5.4	4.7
E	6.2	7.3
F	8.1	4.3
G	11.1	4.9



	Q1	Q2	Q3	Q4	Q5
Hobby	.6	.7	.5	-.3	-.5
Job	-.3	-.2	-.4	.6	.7

This approach is arbitrary and manual.

So, some people disagree with the weights

So, we cannot do it with 20,000 genes or 4,000,000 variants

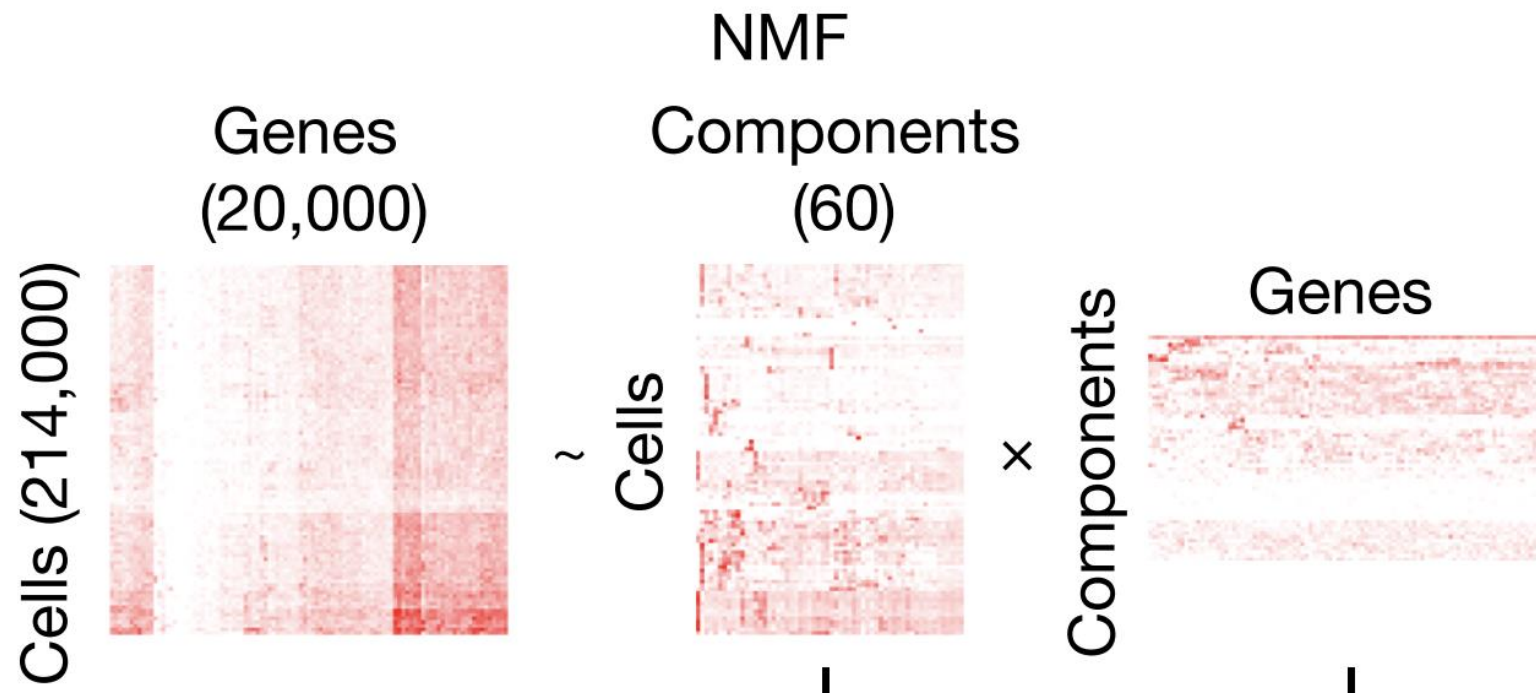
Therefore, we need to

Set up rational & integrated criteria, not an arbitrary criteria.

Do it automatically, not a manually.

There are wide-admitted methods to do it :

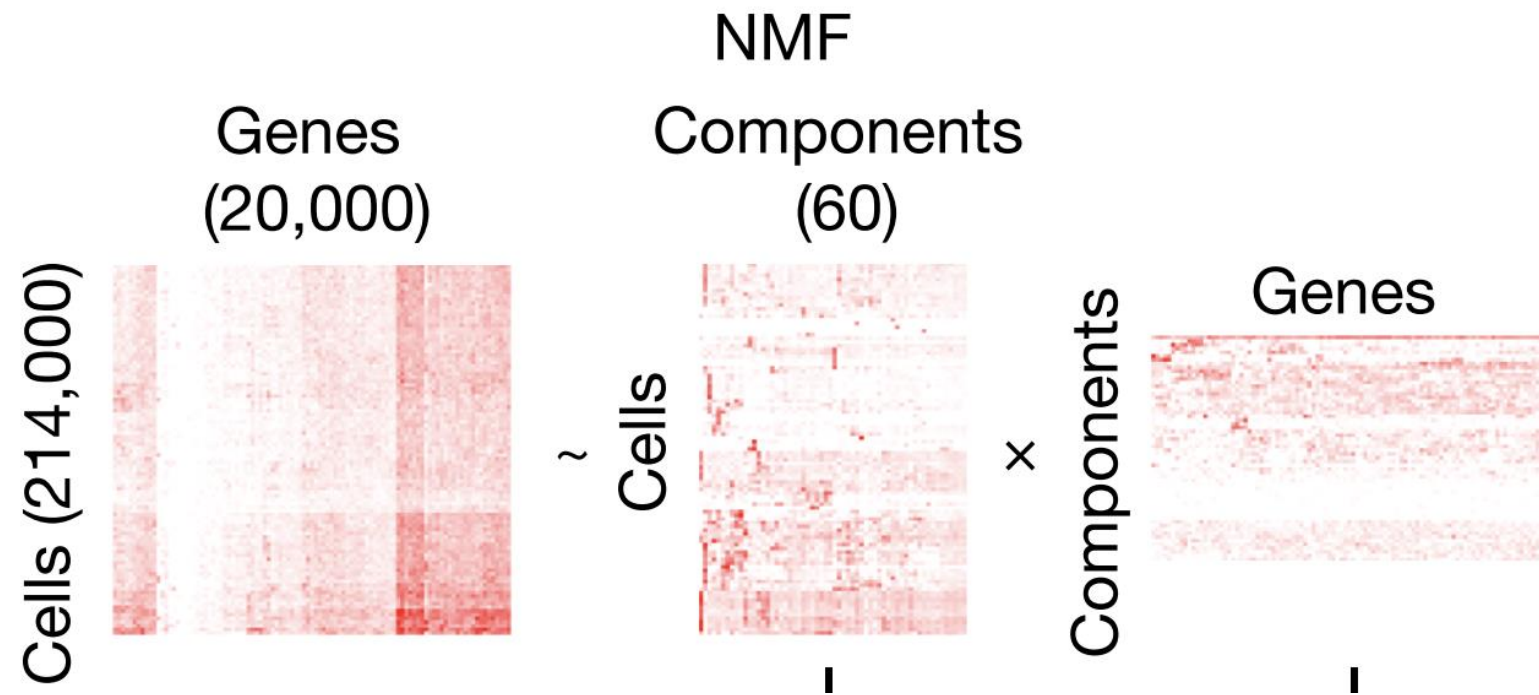
PCA, NMF, MCA, CCA, SIR, SAVE, CR, DR, PHD, MAVE, kPCA, sPCA, iPCA
etc..



The property of NMF

- Every components become positive
- Easy to interpret
- Making the sparse weight (Clustering Effect)

	Gria1	Grin1	Anp32e	Psd95	Kmt5a	Camkk2	Rere	Slc17a7	Mef2c	Znf592
NMF 1	.8	.9	.0	.7	.0	.6	.0	.8	.0	.0
PCA 1	.7	.1	-.3	-.5	-.7	.3	.7	.6	.4	.4



The property of NMF

- Every components become positive
- Easy to interpret
- Making the sparse weight (Clustering Effect)

	Gria1	Grin1	Anp32e	Psd95	Kmt5a	Camkk2	Rere	Slc17a7	Mef2c	Znf592
Excitatory synaptic genes	.8	.9	.0	.7	.0	.6	.0	.8	.0	.0
PCA 1	.7	.1	-.3	-.5	-.7	.3	.7	.6	.4	.4

Third. G2P : link genes to specific program.

For each programs..

1. Finding 300 Co-regulated Genes of the program (=Factor) based on the Factor Loading.
2. Using the GLM, finding (zero to 35) Regulators of the program (Effective gRNA)
3. Now, we have a set of genes per each program.
4. Based on the set, naming the programs. (They use manual curation, analysis of TF motifs, and gene set enrichment)
5. Based on the previous study about gene activities on ECs, select 13 EC-specific programs

c

50 gene programs

Cell cycle

- 3. S phase, DNA replication
- 33. S phase, histones
- 42. S phase
- 12. G2/M
- 40. G2/M
- 54. M phase, kinetochore

Metabolism and mitochondria

- 19. Cholesterol biosynthesis
- 4. Glycolysis, senescence
- 6. Steroid metabolism
- 43. OxPhos, collagens
- 49. OxPhos, proteasome
- 37. Redox homeostasis
- 10. Circadian rhythm 1
- 24. Circadian rhythm 2

RNA and chromatin

- 17. Polycomb targets
- 34. RNA processing
- 16. ncRNAs and antisense RNAs
- 51. RNA splicing
- 38. MEG3 locus

Other inducible responses

- 13. DNA damage response, P53
- 56. Heat shock response
- 5. Interferon response
- 14. TNF response
- 18. NF- κ B signalling
- 36. Steroid hormone response
- 46. Innate immune response
- 11. DNA damage

Ribosomes

- 59. Ribosomal proteins
- 26. Ribosome biogenesis
- 57. Ribosome biogenesis, telomeres

Cytoskeleton

- 27. Intermediate filaments, tubulin

ER and Golgi

- 53. ER translation
- 30. UPR 1, XBP1
- 1. UPR 2, ATF4-DDIT4
- 44. Surface and secreted 1
- 20. Surface and secreted 2
- 21. Surface and secreted 3

EC-specific programs

- | | |
|--|--|
| 29. Wound healing, EDN1
<i>EDN1, PDGFB, CCL2</i> | 15. Flow response, KLF2
<i>KLF2, NOS3, KRT18, KRT19</i> |
| 39. Basement membrane
<i>ADGRF5, HTRA1, SELP, COL4A1</i> | 28. Tip cell, migration
<i>ESM1, RHOC, SNAI2, PLAUR</i> |
| 47. Angiogenesis, GATA
<i>NR2F2, NRP1, NRP2, LBD2, PRCP</i> | 8. Angiogenesis and osmoregulation
<i>IGFBP4, IGFBP5, AQP1, HYAL2, MMP1</i> |
| 35. Focal adhesions, JUN
<i>JUN, JAK1, SVIL, ITGA2</i> | 48. Cell adhesion, CAMKK2
<i>RNASE1, HSPG2, TIMP2, ENG</i> |
| 31. SRF response
<i>MYL9, TUBA1A, ACTG1</i> | 50. TGF β response
<i>SERPINE1, FN1, SOX4, LTBP1</i> |
| 41. Actin remodelling
<i>CFL1, MYL6</i> | 25. Fibrinolysis
<i>PLAT, THBS1, PLAUI, COL5A1</i> |
| 2. EndoMT
<i>POSTN, TGFB2, BMP4</i> | |

d

EC-specific programs

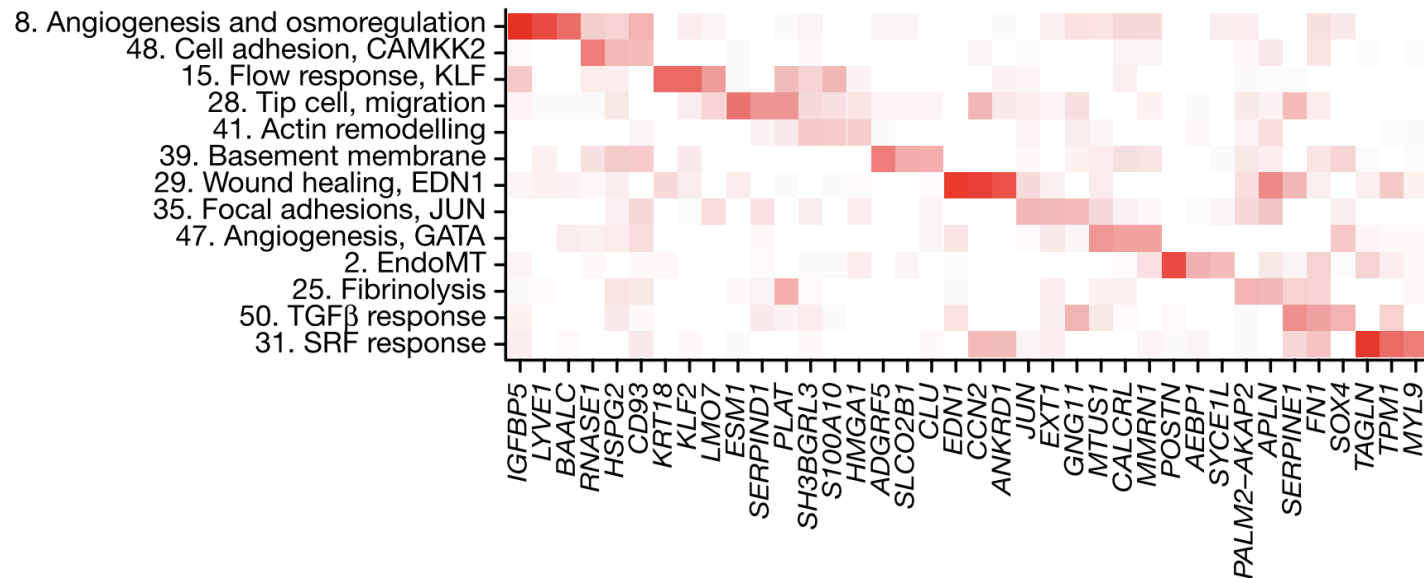
Co-regulated genes

Program gene

z score (specificity)

0

0.005

**Program 8.****Angiogenesis and osmoregulation****Coregulated Genes**

- IGFBP5
- LYVE1
- BAALC

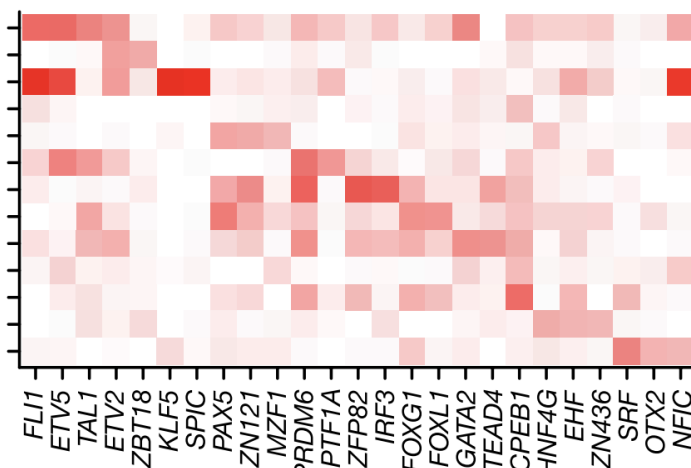
TF motifs

- FLI1
- ERV5
- TAL1

TF motifs in enhancers

 $-\log_{10}(\text{FDR})$

0 1 2 3 4



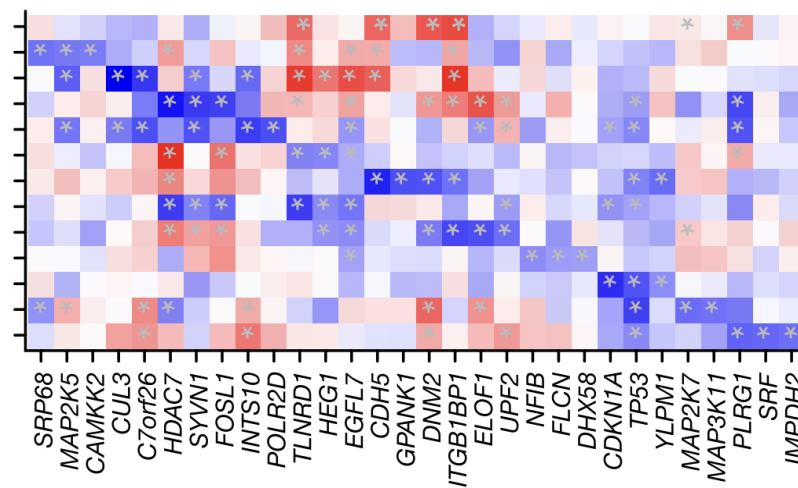
Regulators

 $\log_2(\text{fold change})$

-1

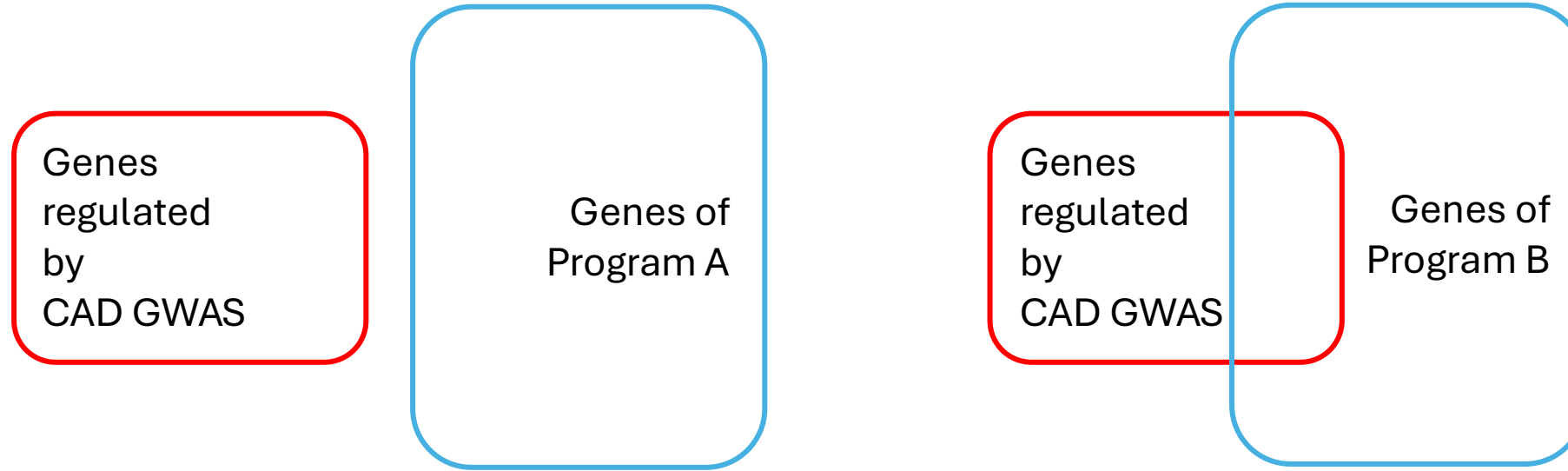
0

1

**Regulators**

- SRP68
- MAP2k5
- CAMKK2

Fourth. Identify disease-associate programs.

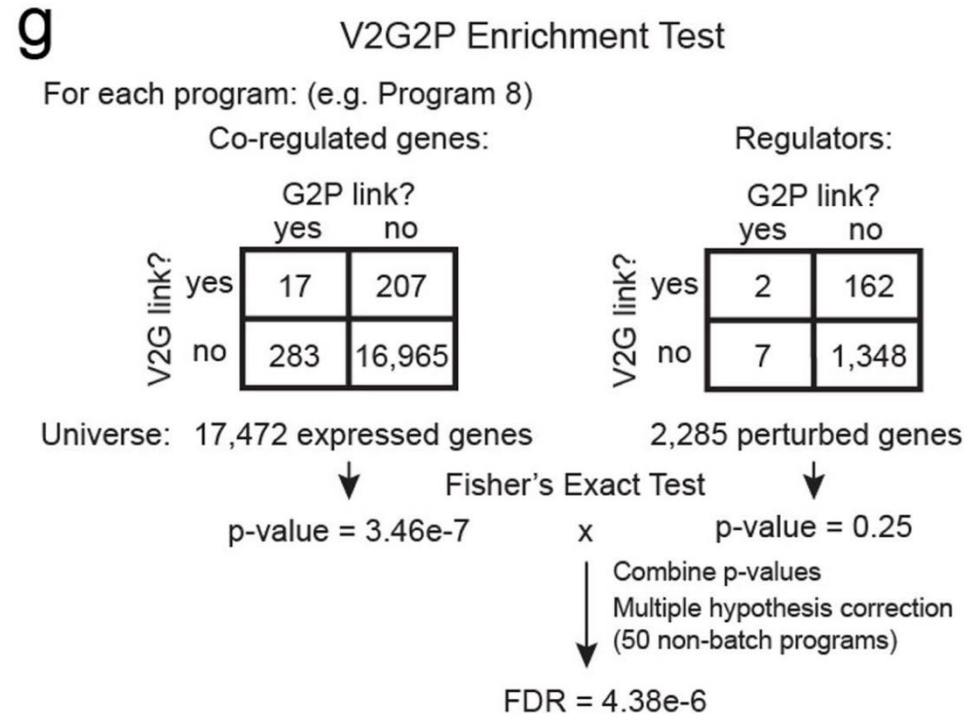


Remark

- V2G process contains the CAD information but G2P is not
- Essentially, the results of two process are sets of Genes per variant (V2G maps) and sets of Genes per programs (G2P maps).

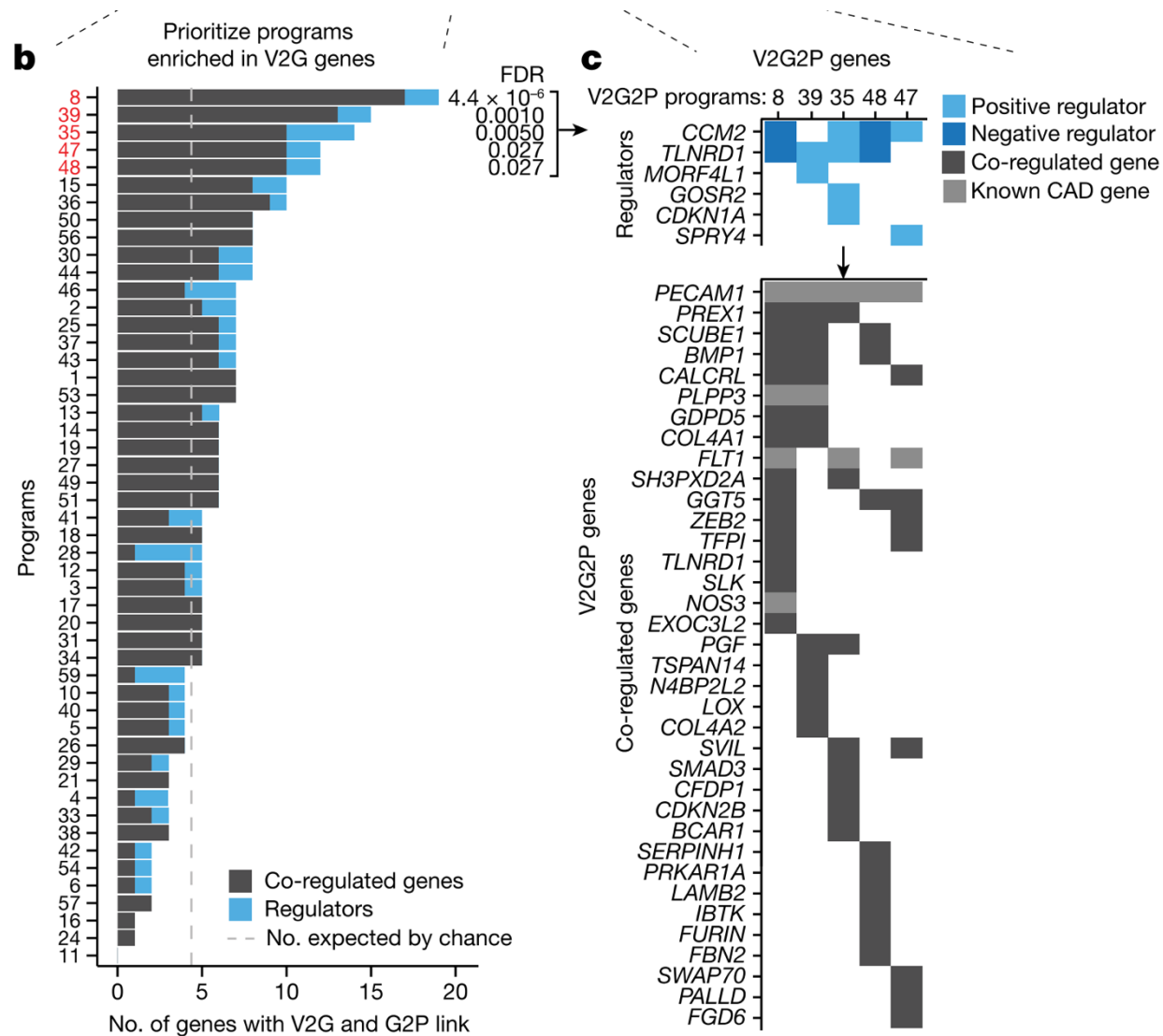
So, by connecting the two process, we can find programs which is associated genes associated with CAD. It's called V2G2P programs

Fourth. Identify disease-associate programs.



So, they used fisher-exact test between V2G and G2P links : Test about "Whether the genes in the programs are also regulated by CAD variants"

> Got 5 CAD-associated Programs among all 60 programs



CAD Associated Programs

8. Angiogenesis and osmoregulation
IGFBP4, IGFBP5, AQP1, HYAL2, MMP1
39. Basement membrane
ADGRF5, HTRA1, SELP, COL4A1
35. Focal adhesions, JUN
JUN, JAK1, SVIL, ITGA2
48. Cell adhesion, CAMKK2
RNASE1, HSPG2, TIMP2, ENG
47. Angiogenesis, GATA
NR2F2, NRP1, NRP2, LBD2, PRCP

CCM2 and TLNDR1 is main regulator of CAD programs

Cf) Supplementary Analyses backing up the 5 programs

1. All five programs are **EC-specific** programs (each program include at least one EC CAD Genes).
2. All five programs were significantly enriched for CAD heritability based on **MAGMA** analysis. (And 2 were enriched by **S-LDSC**)
3. The 41 genes which is in the 5 programs and CAD V2G at the same time were highly ranked at an independent gene prioritization method (**PoPS**)
4. Nine of the 41 genes have previously been found to affect atherosclerosis and vascular barrier integrity by the peoples

Results

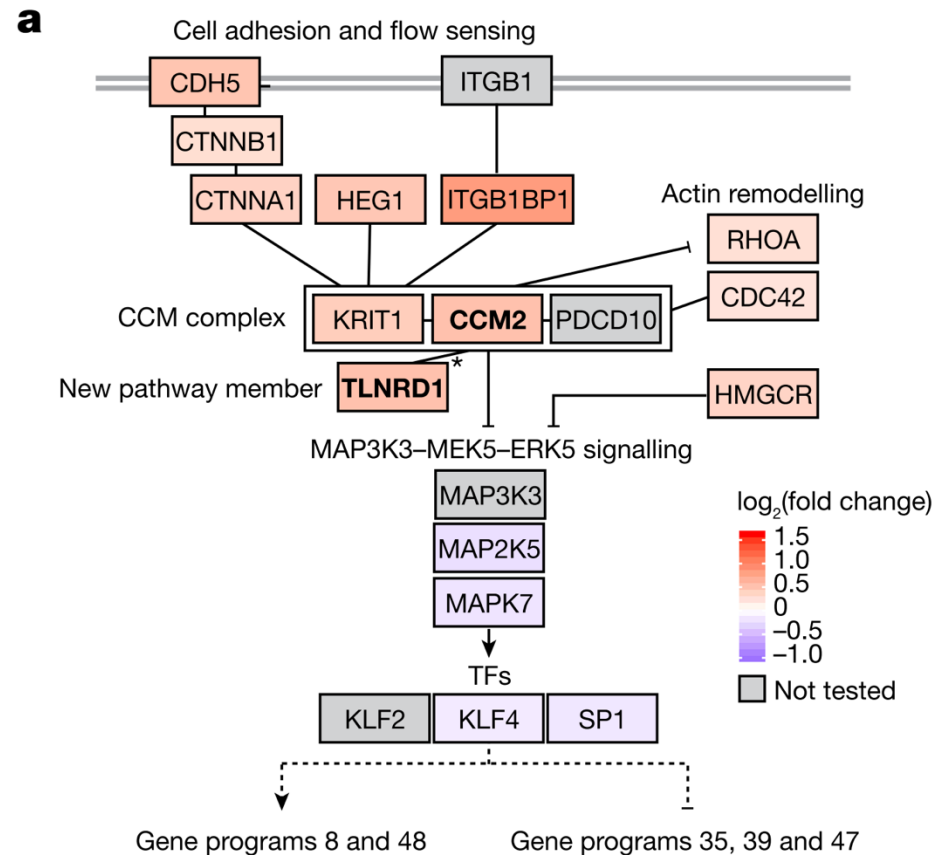
The key genes of V2G2P – CCM2 & TLNRD1

CCM2

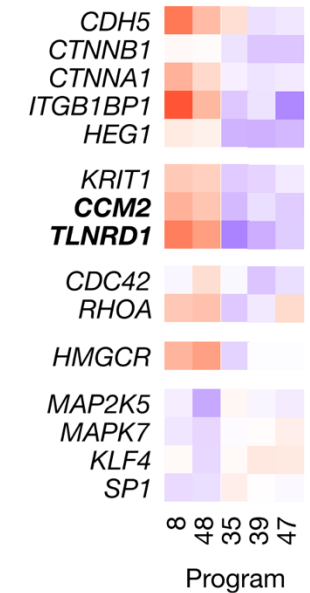
A known member of CCM complex

TLNRD1

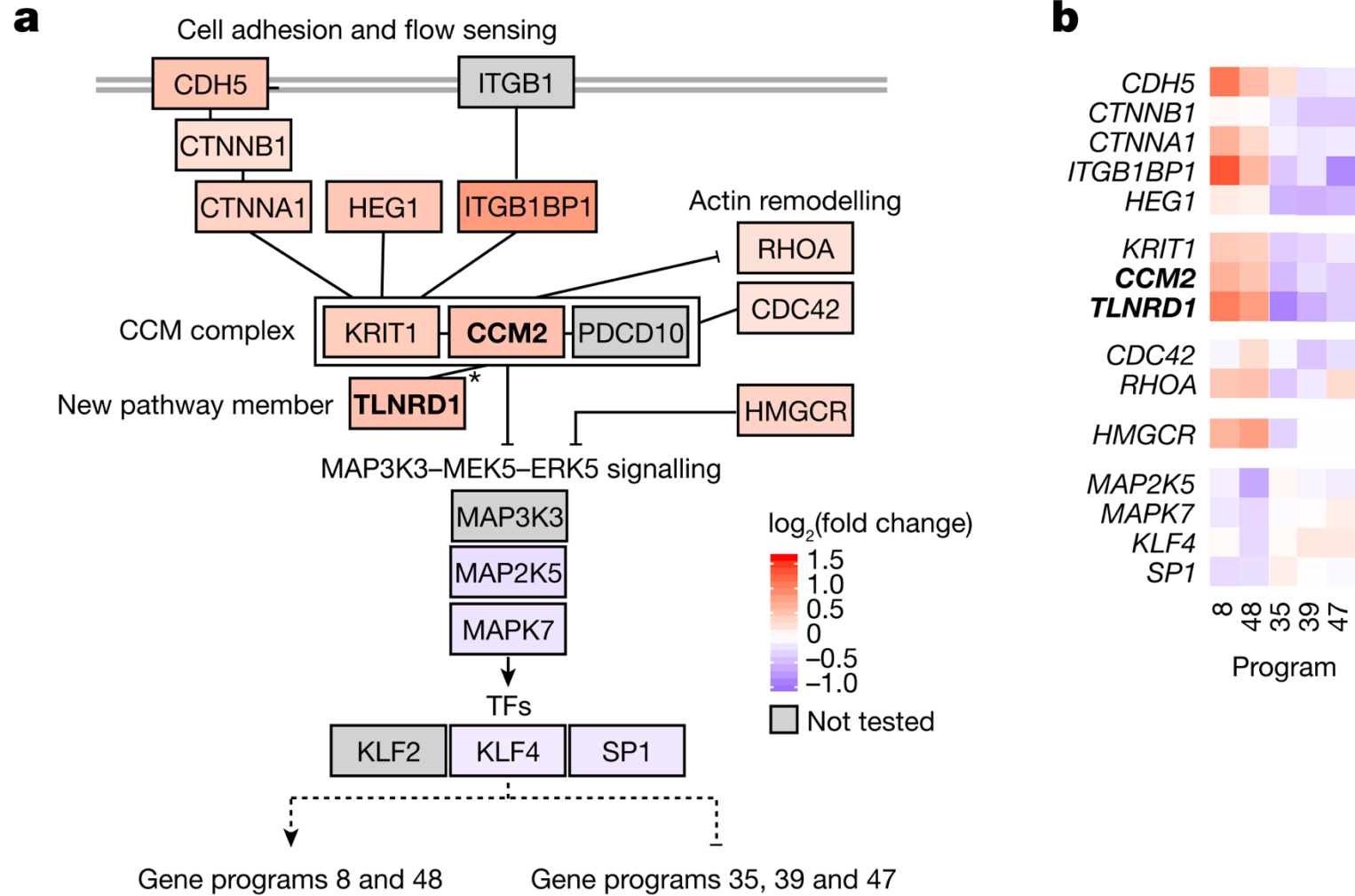
A perviously poorly studied gene with no known function in ECs



b

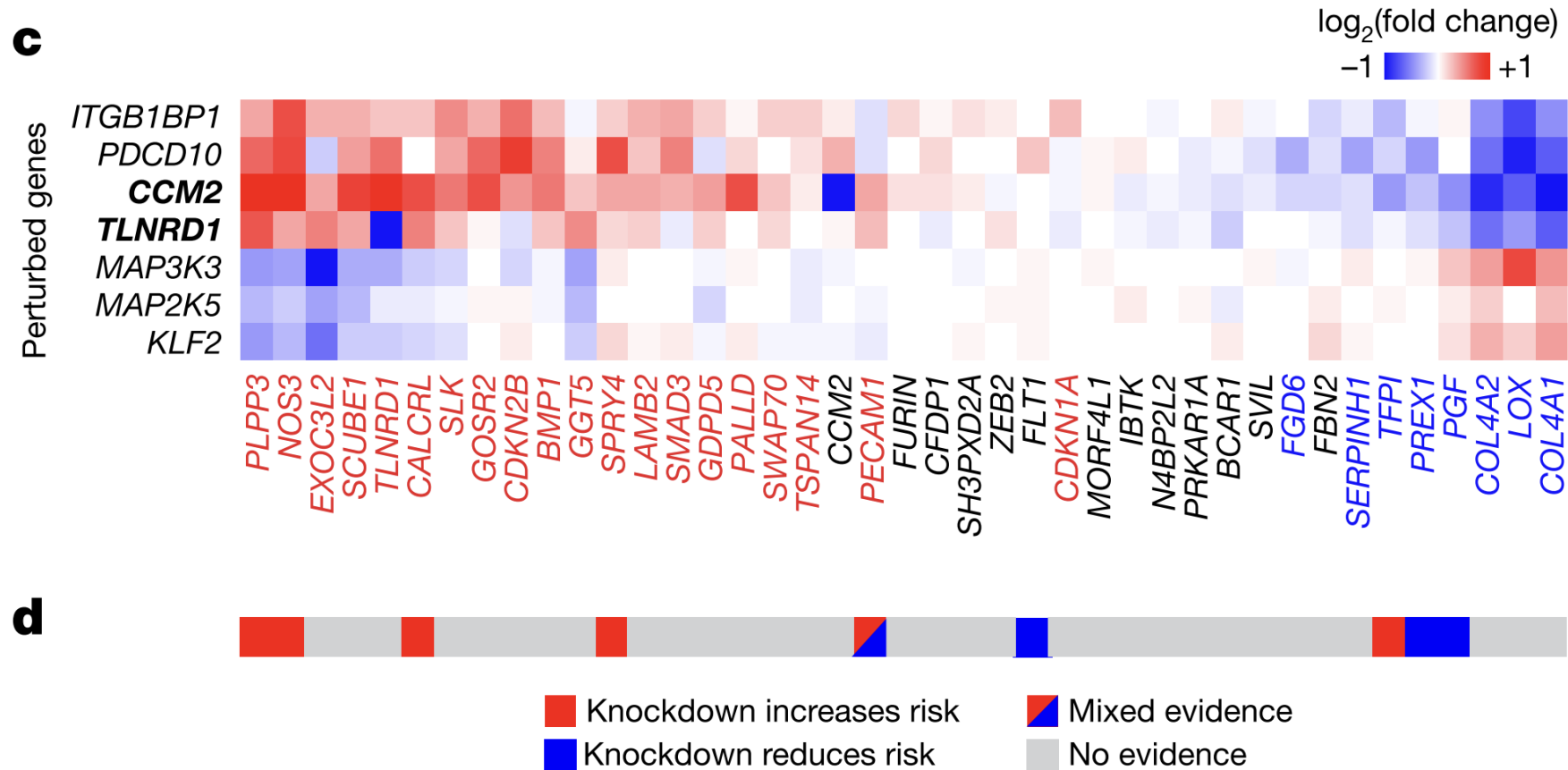


'Rare loss-of-function mutations in proteins forming the CCM complex can cause rare monogenic vascular malformations through effects on microvascular ECs, including activation of MAP3K3-MEK5-ERK5 signalling to KLF2 and KLF4.'



A. CCM pathway map

B. Effect of the genes in CCM pathway to CAD programs



C. Effect of the genes in CCM pathway to genes in CAD programs

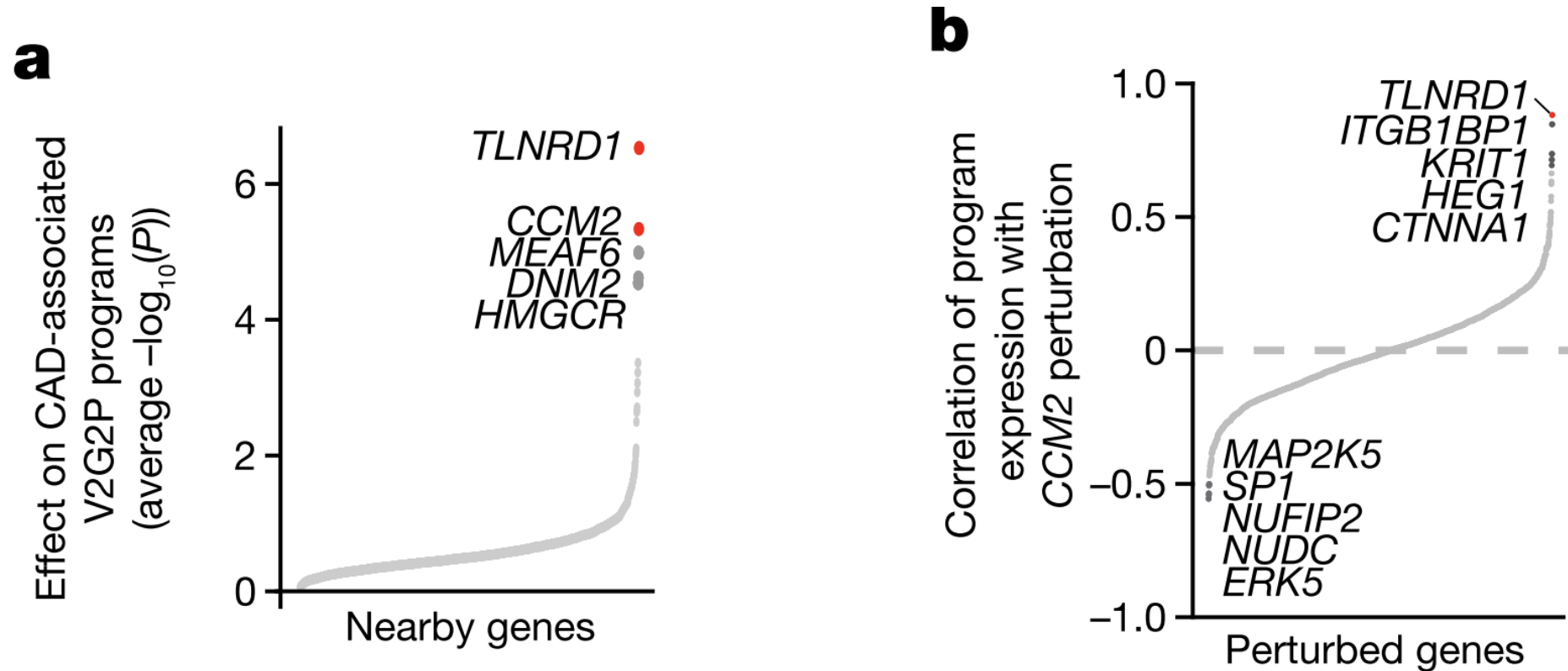
- 28 genes among 41 genes are affected by CCM genes

D. Direction of CAD Risk per Genes by previous research

Conclusion

1. **CAD programs are connected with CCM Pathway**

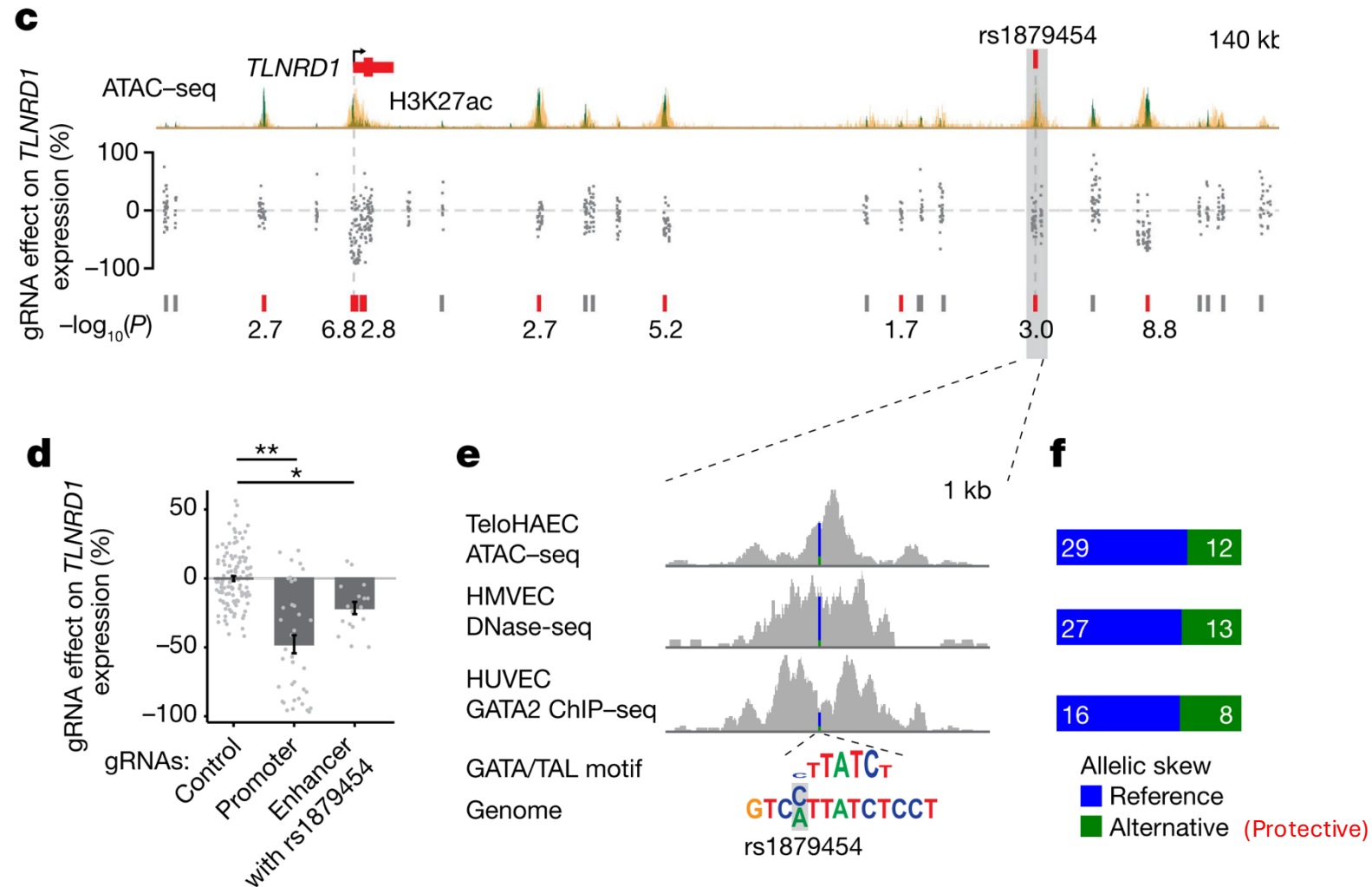
Studies on the TLNRD1 on Transcriptomics



A. Significance of genes on 5 CAD V2G2P Programs

B. Correlation between perturbation effect of CCM2 and others

This results say that TLNRD1 is highly related with CCM and CAD



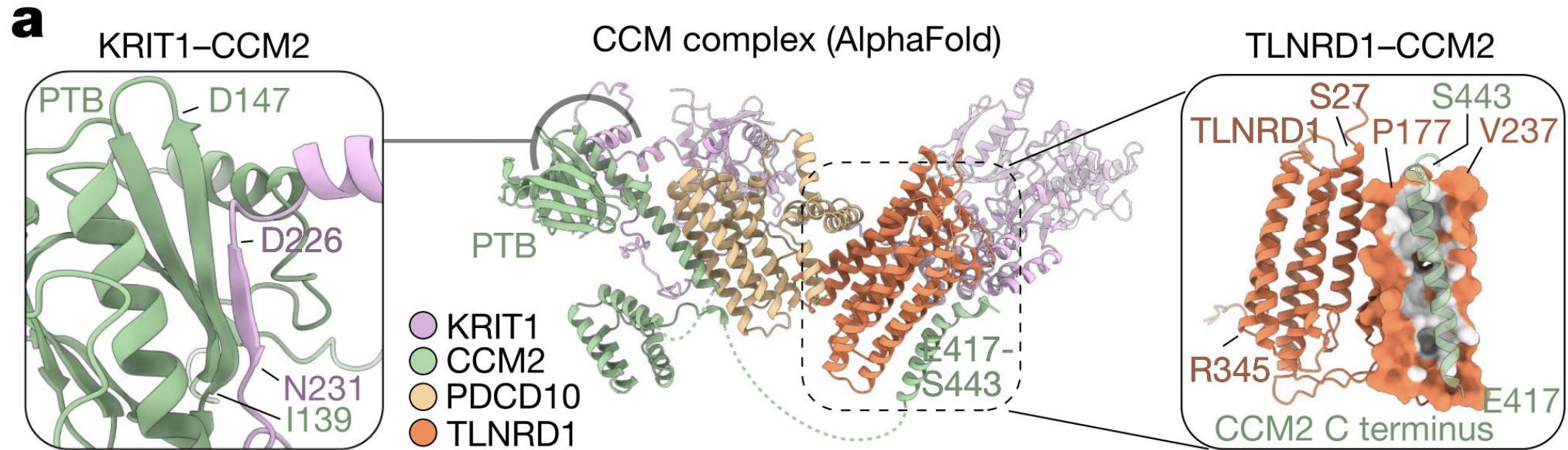
C,D. Effect of knockout of rs1879454 on CRISPERi-FlowFISH
E,F. ATAC-seq,DNase-seq and ChIP-seq at the rs1879454

This results shows rs1879454 works to suppress the *TLNRD1* and decrease the risk of CAD

Conclusion

1. **CAD programs are connected with CCM Pathway**
2. **TLNRD1 is highly related with CCM2 and it can decrease the risk CAD.**

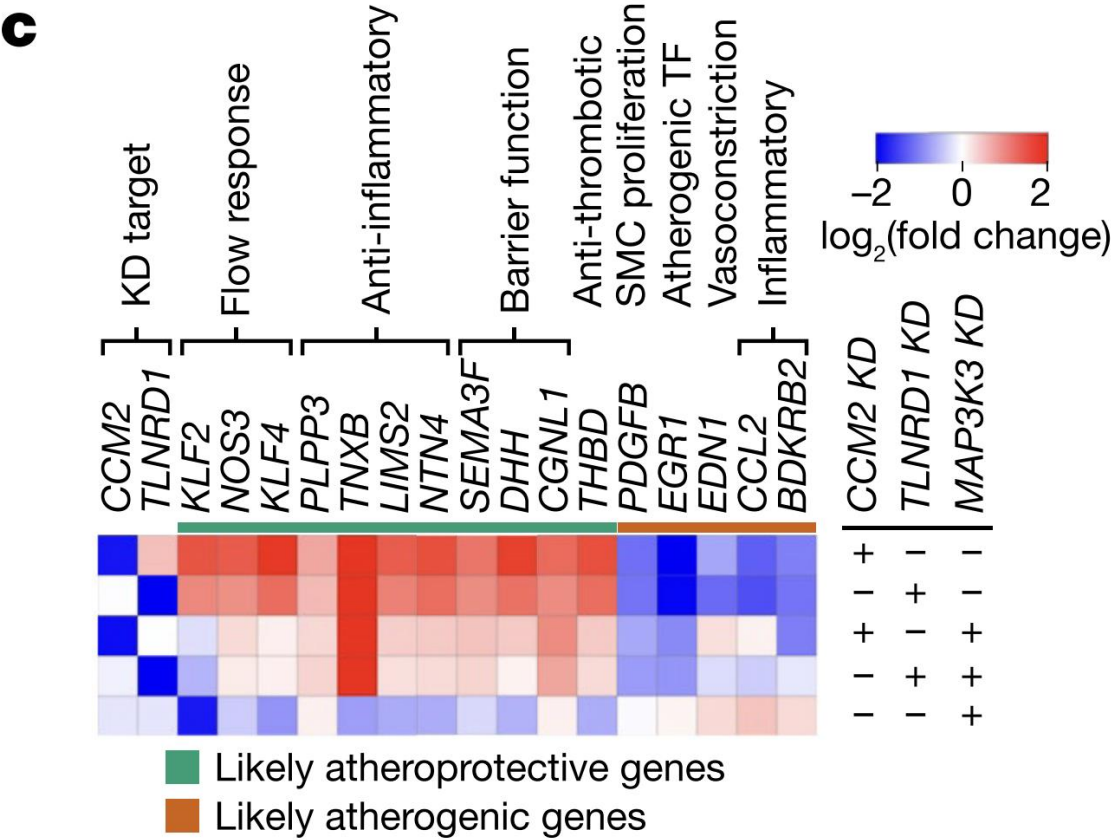
Studies on the TLNRD1 in protein



A. Potential Protein Complex (Quaternary Structure) predicted by Alpha Fold2.3 multimer

This results shows the protein level interaction between CCM complex and TLNRD1

Studies on the TLNRD1 in CCM Pathway



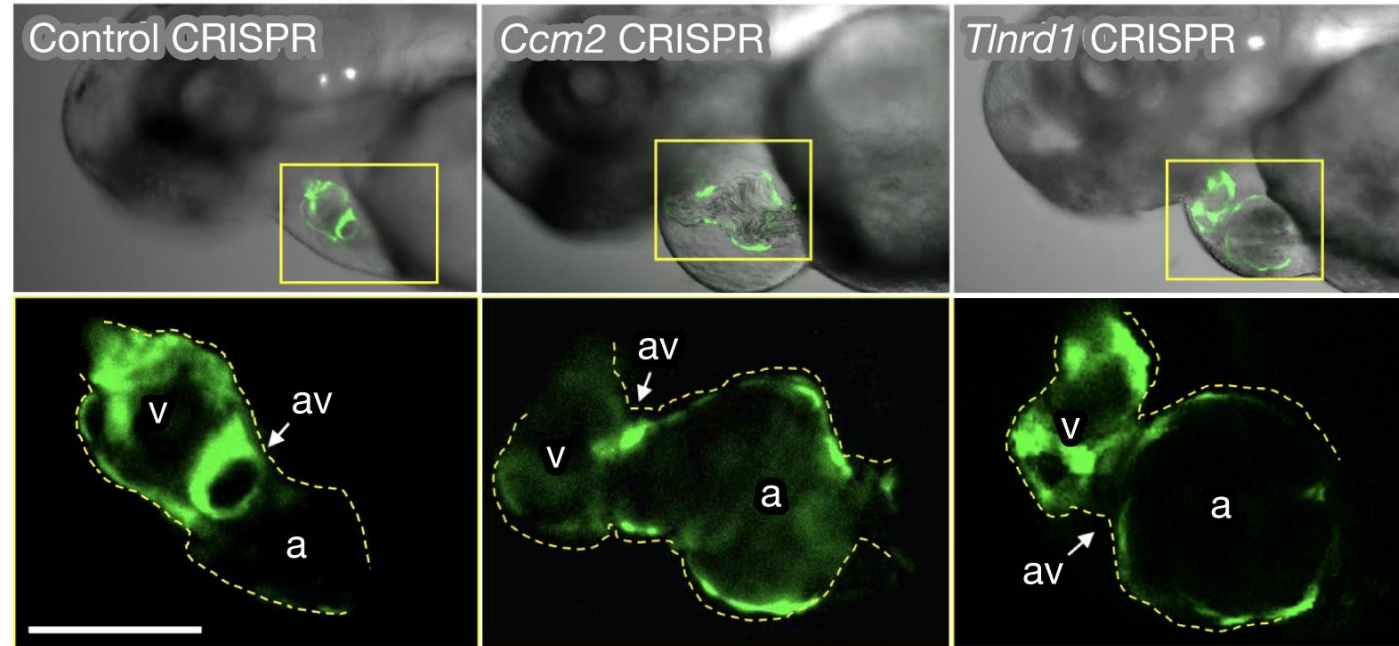
C. The effects of knockout of CCM2 and TLNRD1

- MAP3K3 is known to be a key gene in the CCM Complex Pathway

This result shows the pathway level coherence between CCM2 and TLNRD1

Studies on the TLNRD1 in VIVO

d



D. Effect of knockdown of CCM2 and TLNRD1 on Zebrafish.

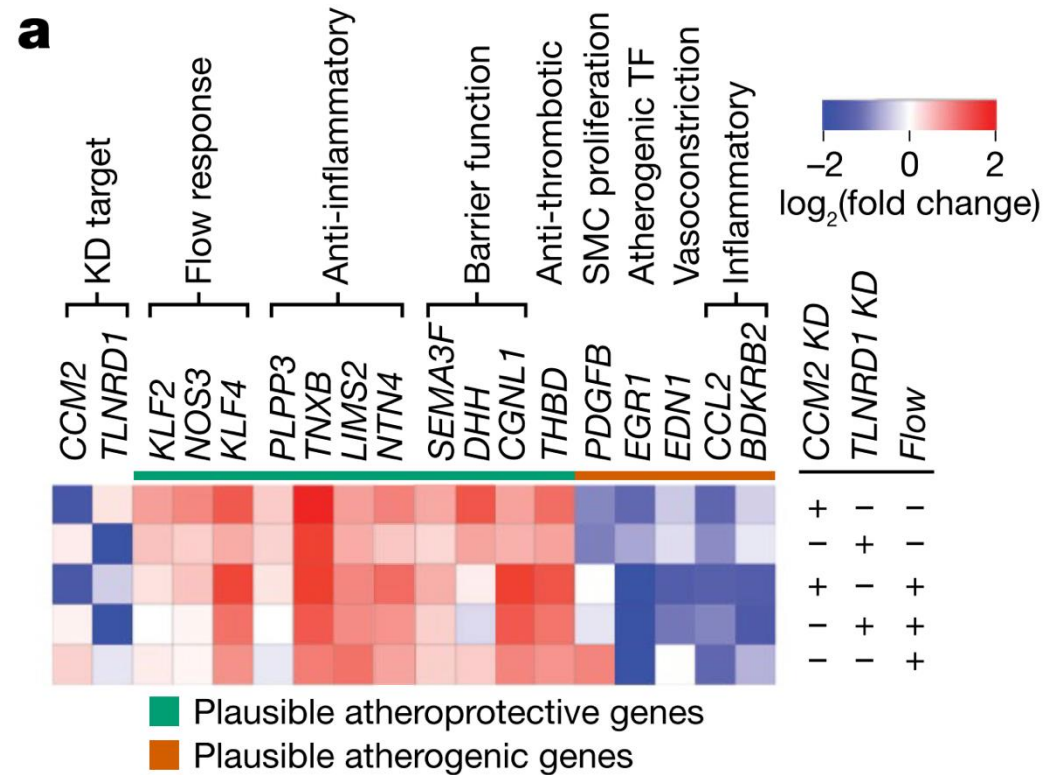
a: atrium, v:ventricle, av: atrioventricular valve

- Both of them had vascular defects, including posterior cardinal vein dilation and increased vascular permeability to red dextran particles

Conclusion

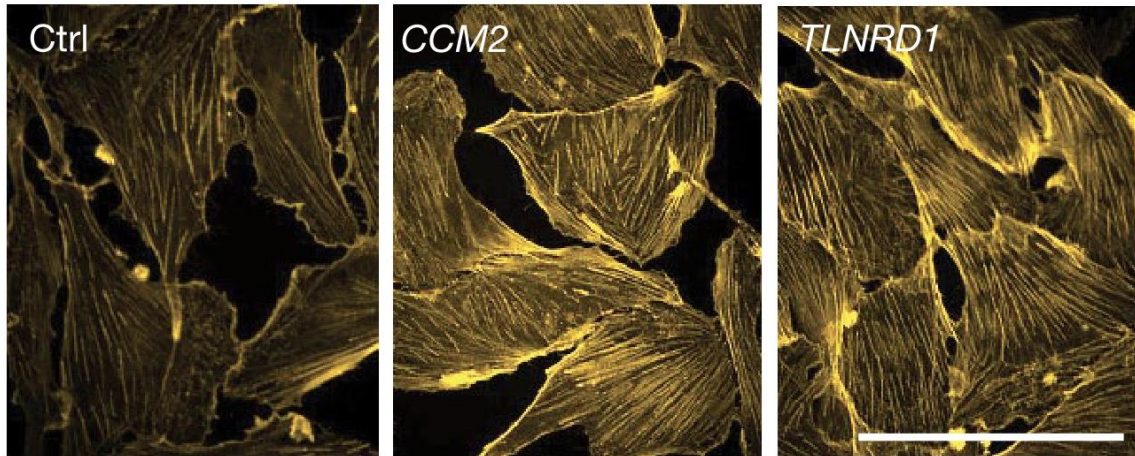
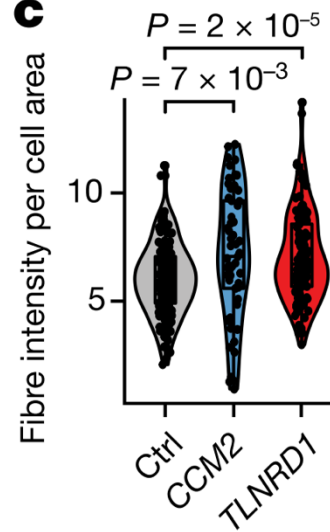
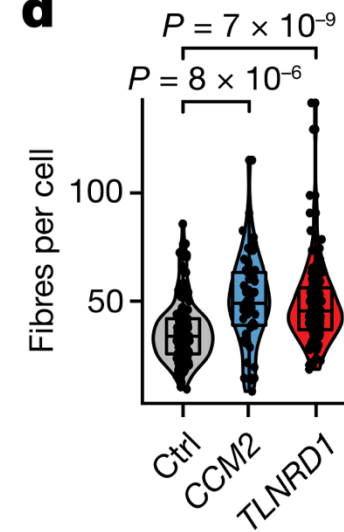
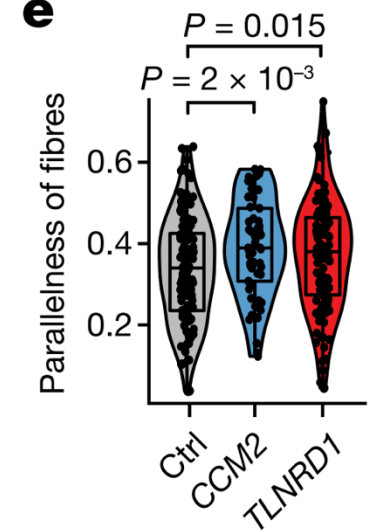
- 1. CAD programs are connected with CCM Pathway**
- 2. TLNRD1 is highly related with CCM2 and it can decrease the risk CAD.**
- 3. TLNRD1 have a potential to be included at CCM complex**

Studies on the TLNRD1 in function



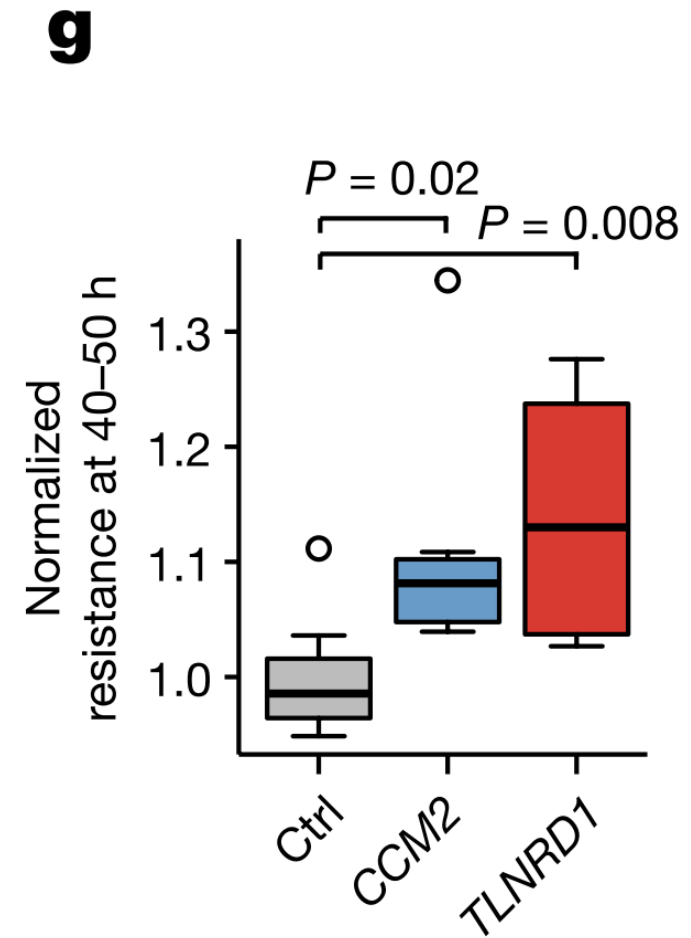
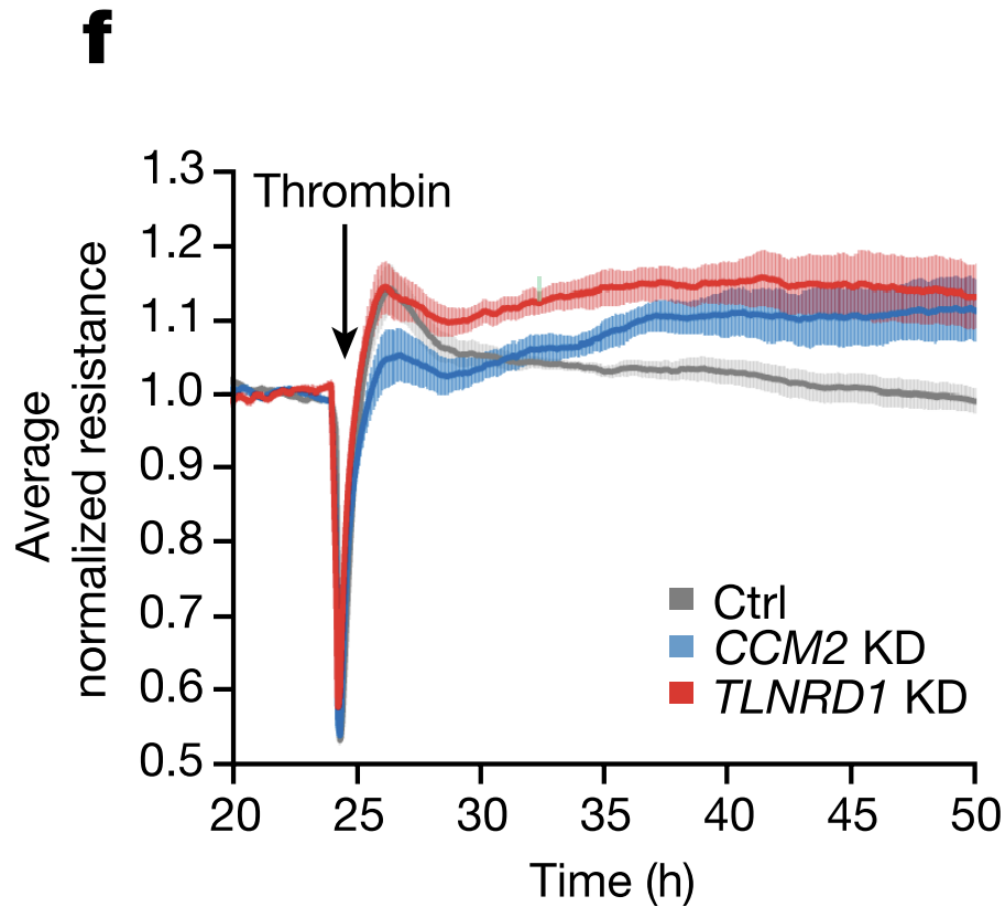
A. Effect of knockdown of CCM2 and TLNRD1 on key genes.

- Flow means laminar shear stress. Flow – means static state
- The upregulated genes are plausible atheroprotective genes and the downregulated genes are atherogenic genes.

b**c****d****e**

B,C,D,E. Effect of knockdown of CCM2 and TLNRD1 on teloHAECs.

- Knock out of the genes increase the number and parallel arrangement of actin stress



F,G. Resistance of EC measured by trans-endothelial electrical resistance (TEER)

Conclusion

- 1. CAD programs are connected with CCM Pathway**
- 2. TLNRD1 is highly related with CCM2 and it can decrease the risk CAD.**
- 3. TLNRD1 have a potential to be included at CCM complex**
- 4. Downregulation of TLNRD1 and CCM2 can improve the Barrier of the ECs**

Why it is published at Nature?

The size of the Experiment

- 214,449 Cells and 37,637 gRNA

The variety of the Methods

- Statistical techniques (MAGMA, S-LDSC and PoPs), Machine Learning (NMF, Alphafold) and Experiment (CRIPRI-Perturbed-seq, TEER)

The significance of the Results

(But it takes 1.5 years to be accepted..)