이학박사 학위논문 최종심사 – 2018.12.20

Statistical Method Development for Genetic Association Analyses of Dichotomous Phenotypes with Related Samples and its Application to Genetic Studies

Wonji Kim

Interdisciplinary Program of Bioinformatics
Seoul National University

Contents

- Outline of the thesis
- Updates after the 1st defense
- Revision of the thesis

Outline of the thesis

Chapter 1

Introduction

Chapter 2

Application of Genome-wide Association Study and Fine-mapping for Independent Samples (Summited to *European Respiratory Journal*)

Chapter 3

Selecting Cases and Controls for Genome-wide Association Studies Using Family Histories of Disease (Published in *Statistics in Medicine*)

Chapter 4

Heritability Estimation of Dichotomous Phenotypes Using a Liability Threshold Model on Ascertained Family-based Samples (Summited to *Genetic Epidemiology*)

Chapter 5

Summary and Conclusions

Updates after the 1st defense

Chapter 1

Introduction

Chapter 2

Application of Genome-wide Association Study and Fine-mapping for Independent Samples (Summited to *European Respiratory Journal*)

Chapter 3

Selecting Cases and Controls for Genome-wide Association Studies Using Family Histories of Disease (Published in *Statistics in Medicine*)

Chapter 4

Heritability Estimation of Dichotomous Phenotypes Using a Liability Threshold Model on Ascertained Family-based Samples (Summited to *Genetic Epidemiology*)

Chapter 5

Summary and Conclusions

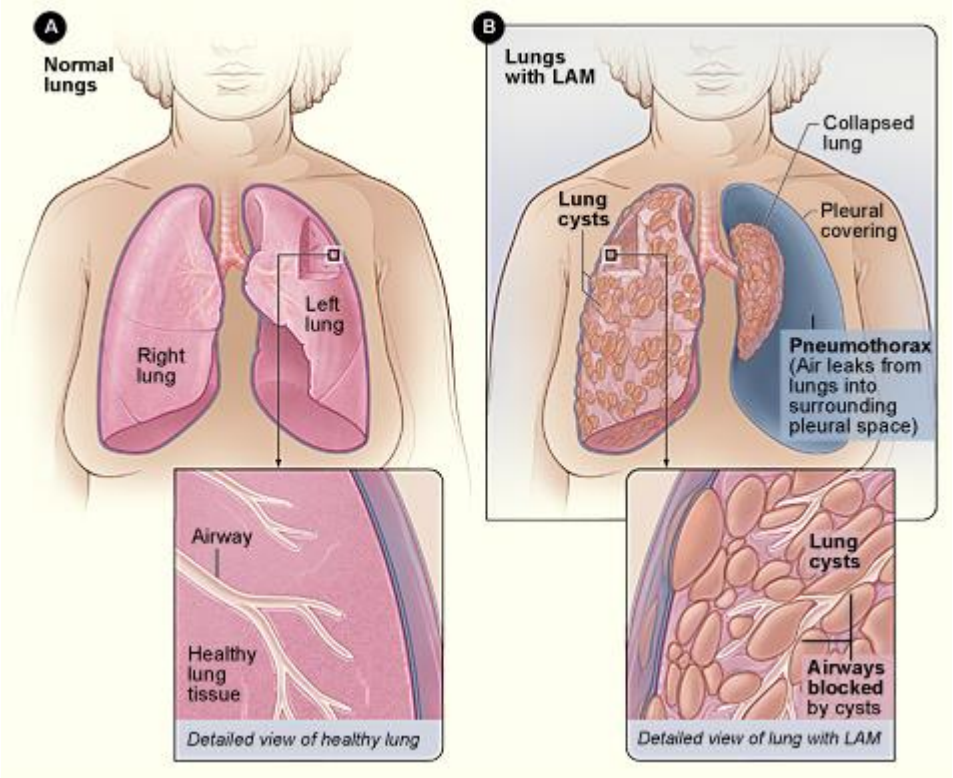
Chapter 2

Application of Genome-wide Association Study and Fine-mapping for Independent Samples

Introduction

- Lymphangioleiomyomatosis (LAM) is a rare disease affecting women of child-bearing age, usually in their 30s and 40s.

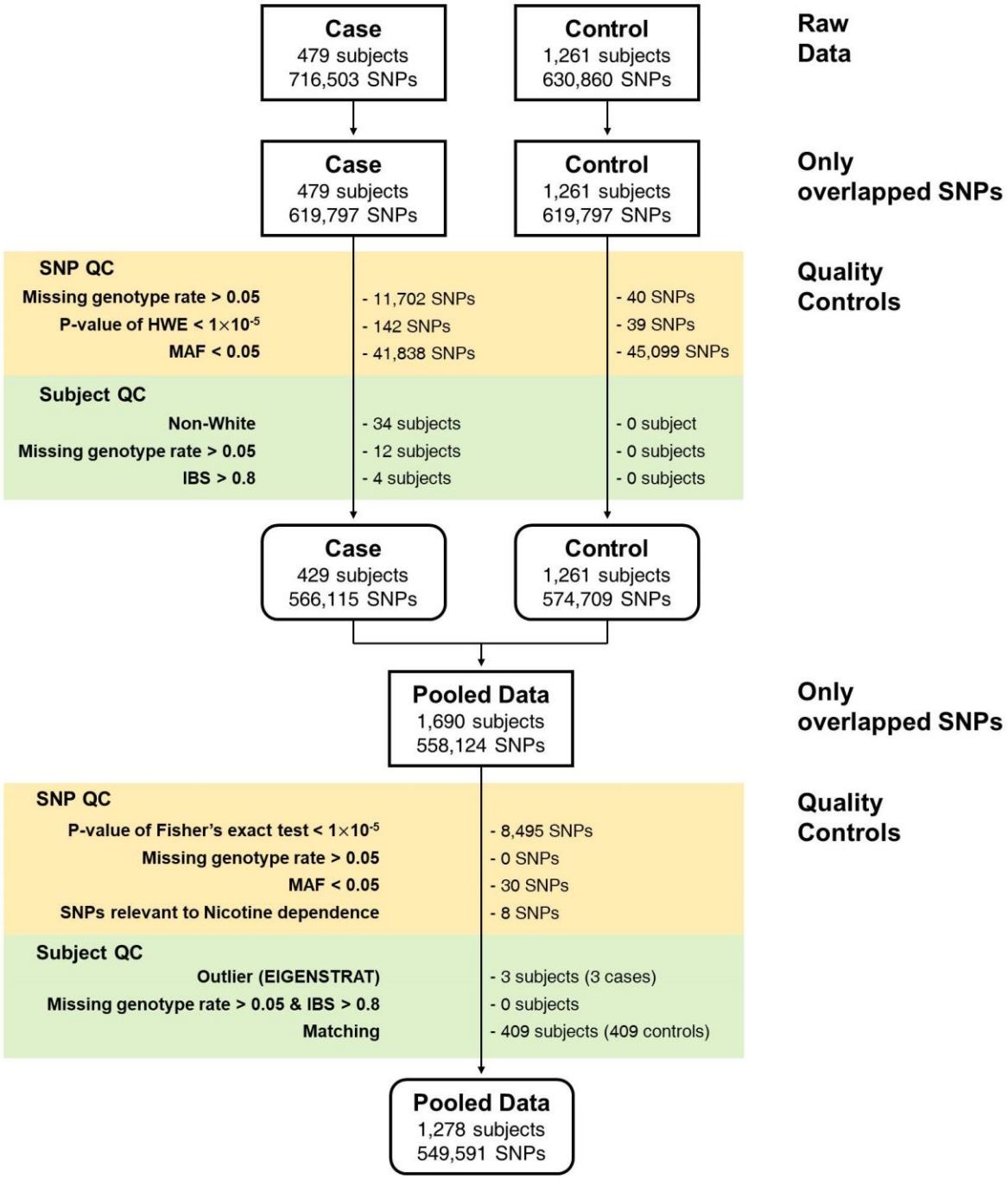
- **sporadic LAM (S-LAM) :** for patients with LAM not associated with tuberous sclerosis complex (TSC),
  TSC-LAM : LAM associated with TSC which is due to mutations in either *TSC1 or TSC2*.



- **Hypothesis :** DNA sequence variants outside of *TSC2*/*TSC1* might be associated with S-LAM

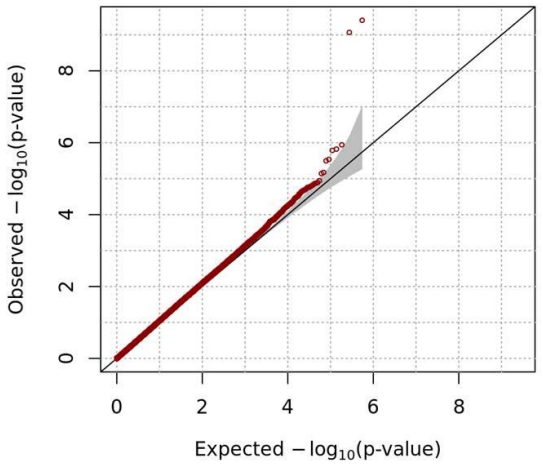
Workflow of GWAS

- **Case**
  S-LAM patients from 14 countries
- **Control**
  Healthy women from COPDGene dataset
- **Method**
  I performed conditional logistic regression (CLR) using matched cases and controls (1:2).
  I used two PC scores with two largest eigenvalues and age for matching.

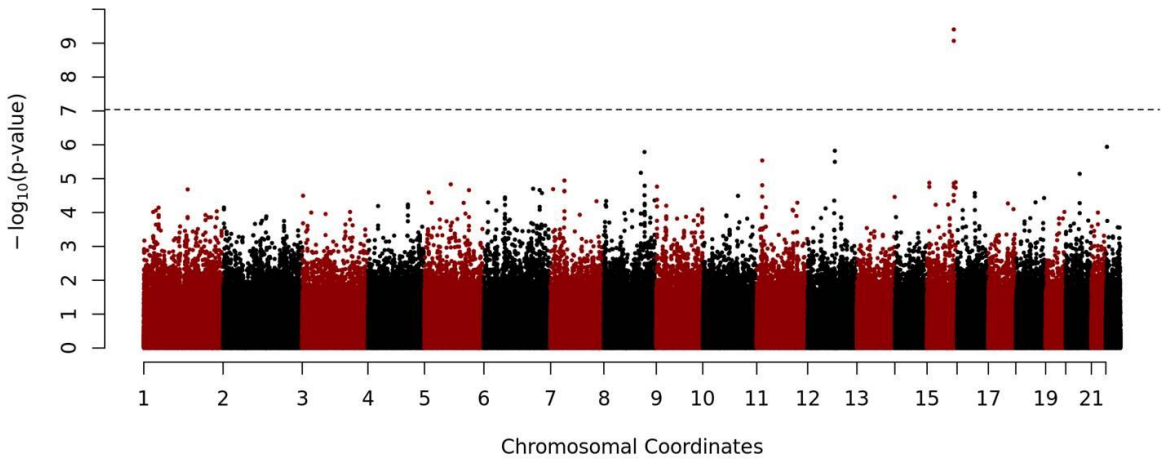


Result of GWAS

a)



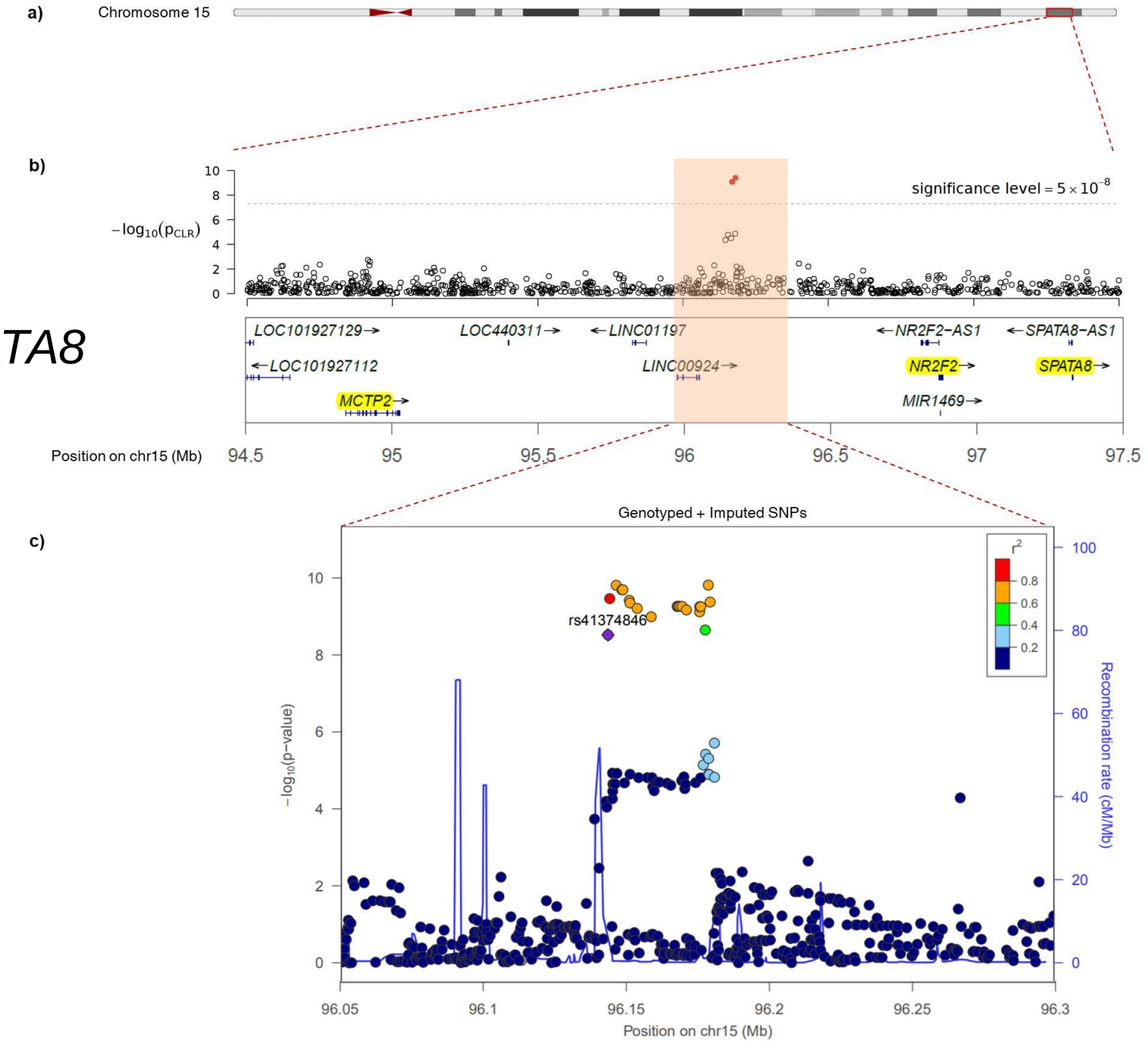
b)



	rs4544201	rs2006950
Chromosome	15q26.2	15q26.2
SNP position (hg19)	96167827	96179390
Minor / Major alleles	A / G	A / G
Minor allele frequency		
S-LAM	0.1655	0.1420
Control	0.2750	0.2529
Genotype counts (AA / AG / GG / Missing)		
S-LAM	16 / 108 / 299 / 3	11 / 99 / 316 / 0
Control	62 / 343 / 444 / 3	58 / 315 / 479 / 0
Discovery data		
Odds ratio		
Original	0.4916	0.4732
Bias adjusted	0.5677	0.5315
P-value	8.51×10^{-10}	3.92×10^{-10}
Replication data		
Odds ratio	0.3288	0.2731
P-value	4.32×10^{-5}	1.56×10^{-5}

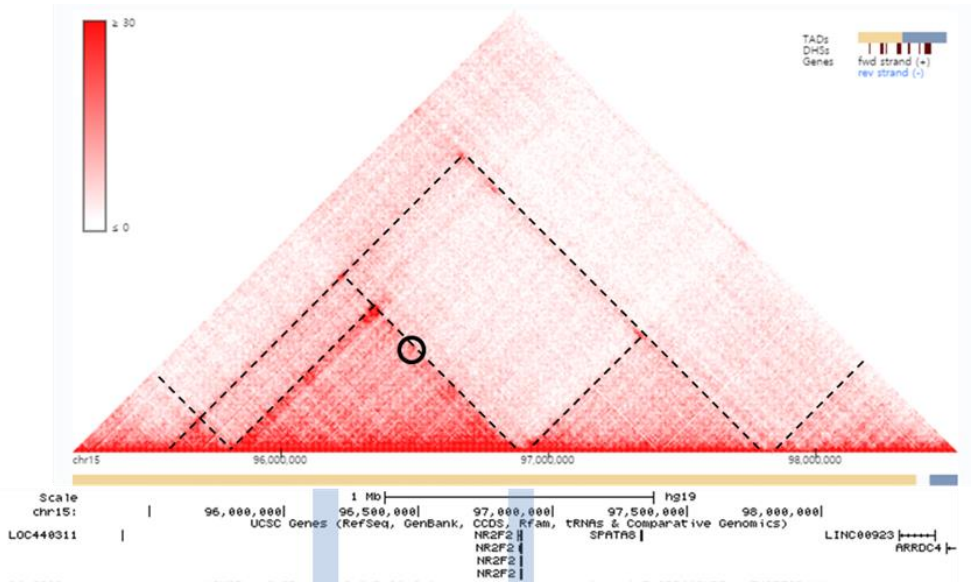
Regional Plot

- Protein coding genes nearby GWAS signals : *MCTP2, NR2F2, SPATA8*

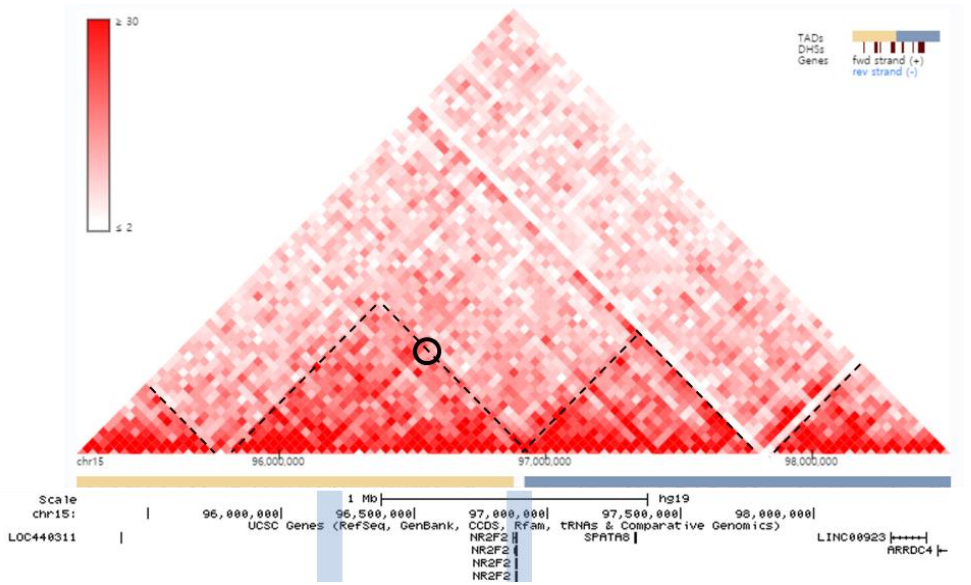


TADs

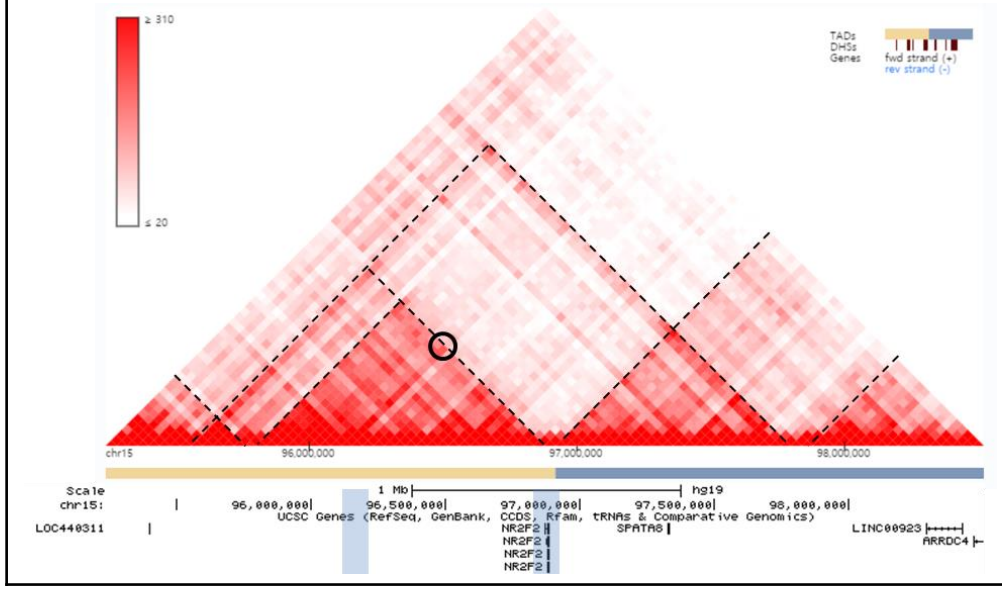
IMR90



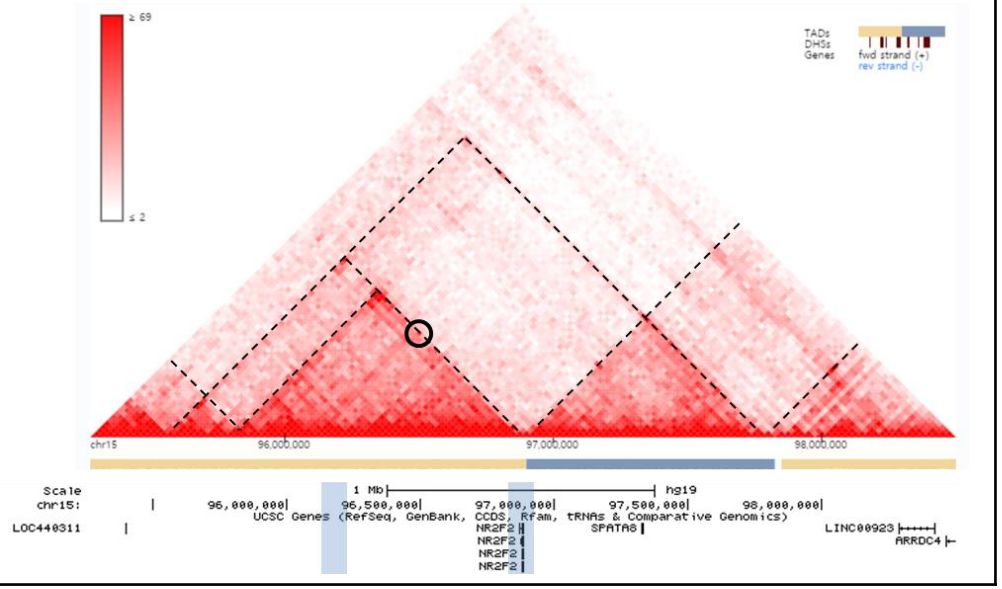
Lung



h1-MSC cell

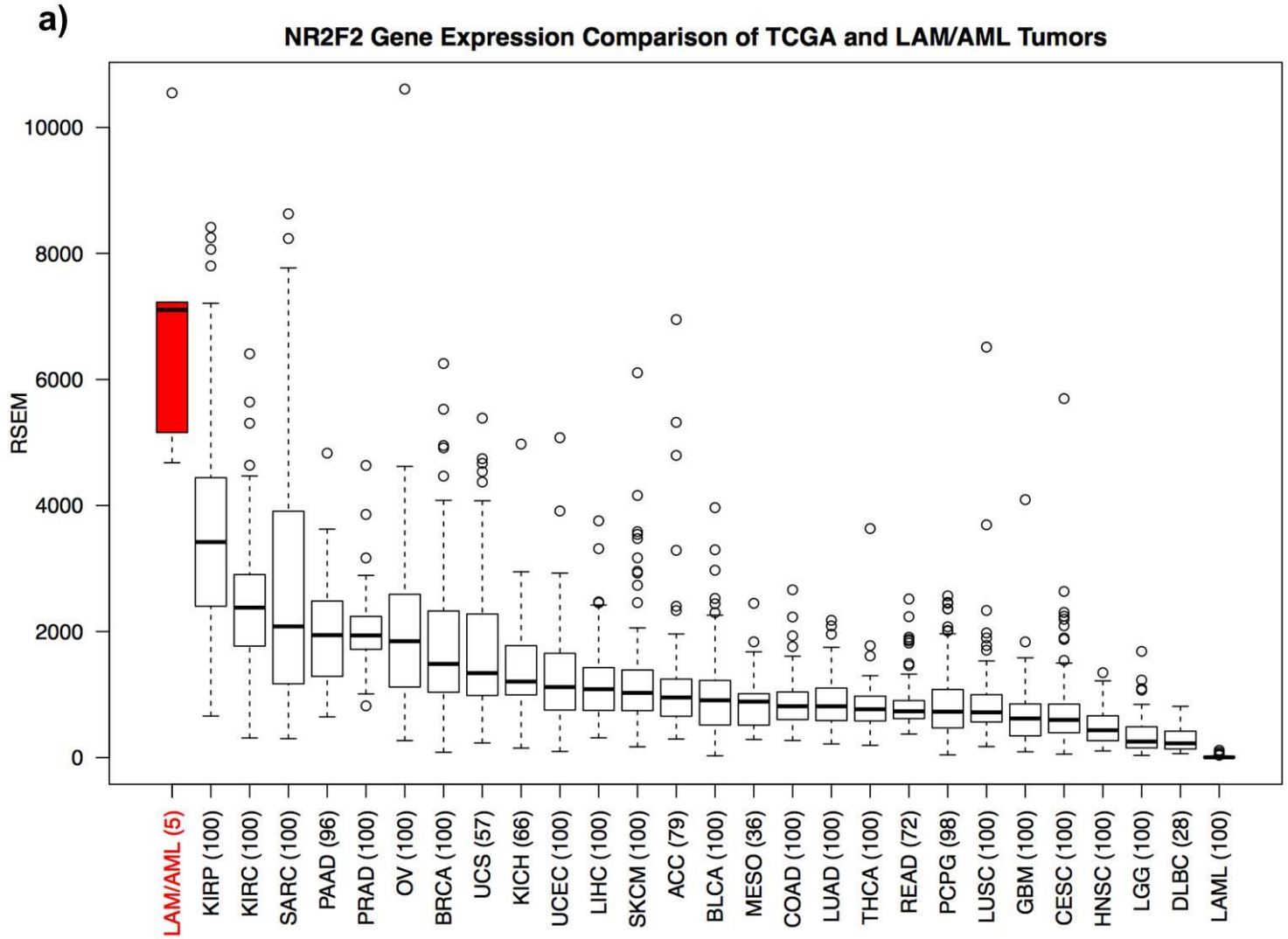


HUVEC



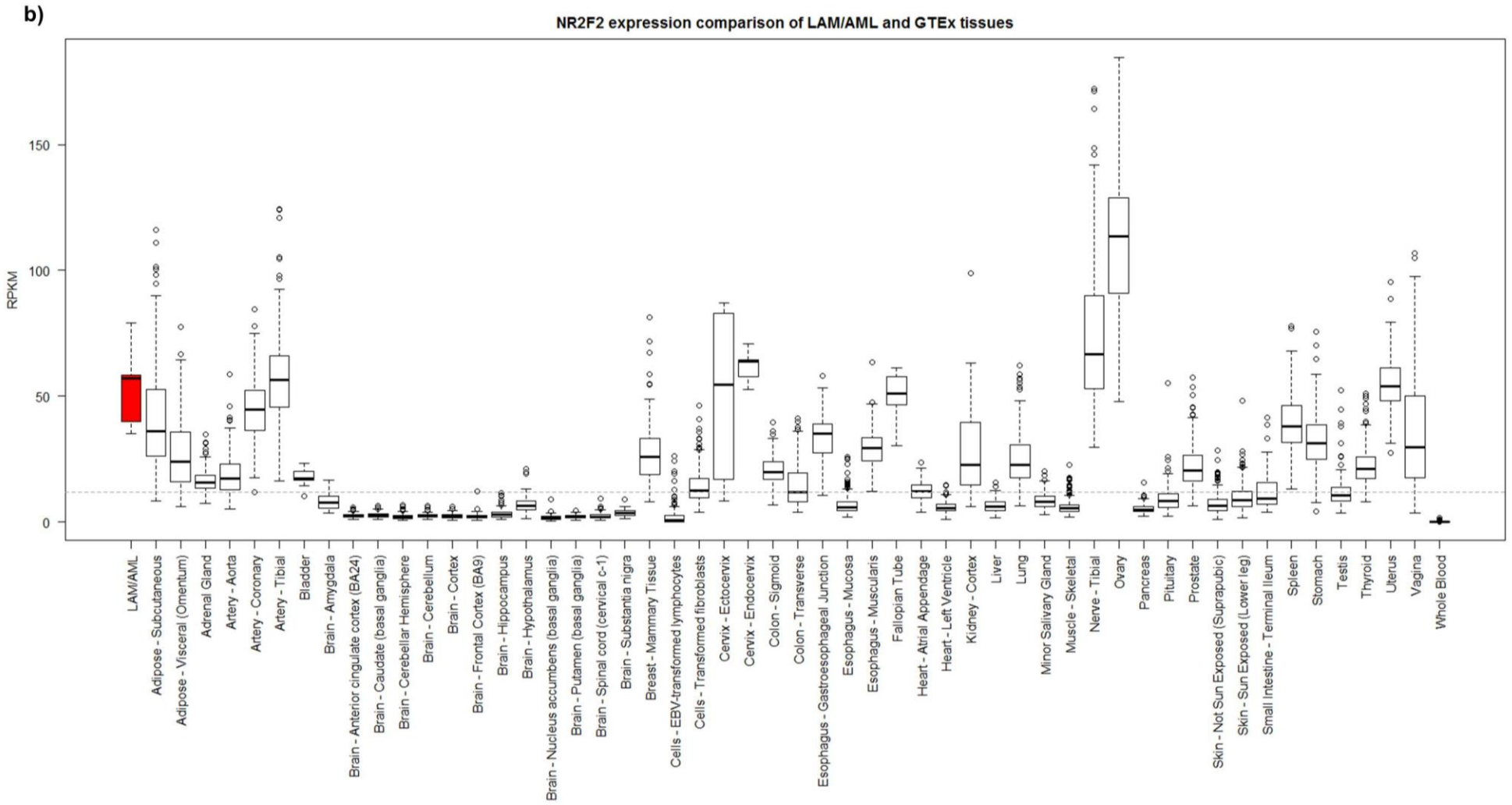
Comparison of NR2F2 expression

- with cancer tissues (TCGA)



Comparison of NR2F2 expression

- with normal tissues (GTEx)

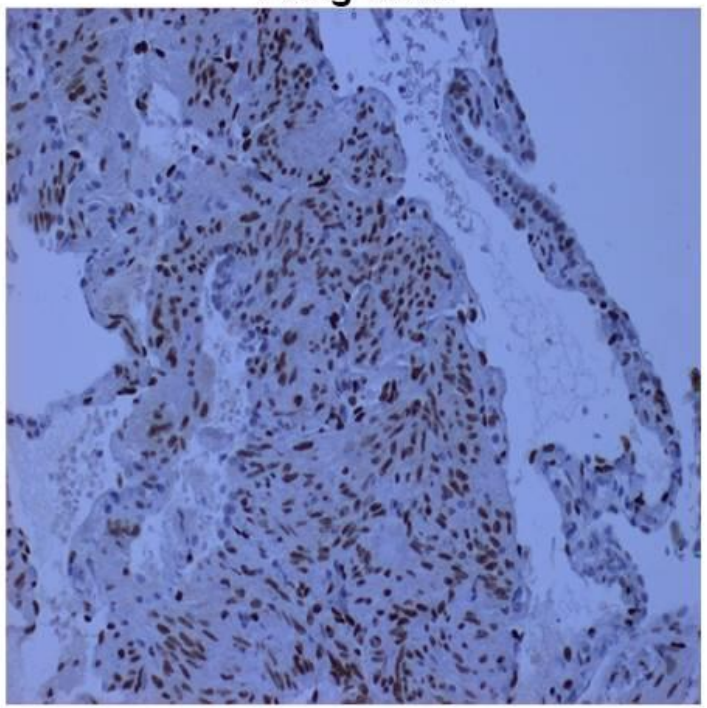


Immunohistochemistry for *NR2F2* in LAM/AML

- Strong nuclear staining is seen in lung LAM cells (a) and Kidney AML cells (b)

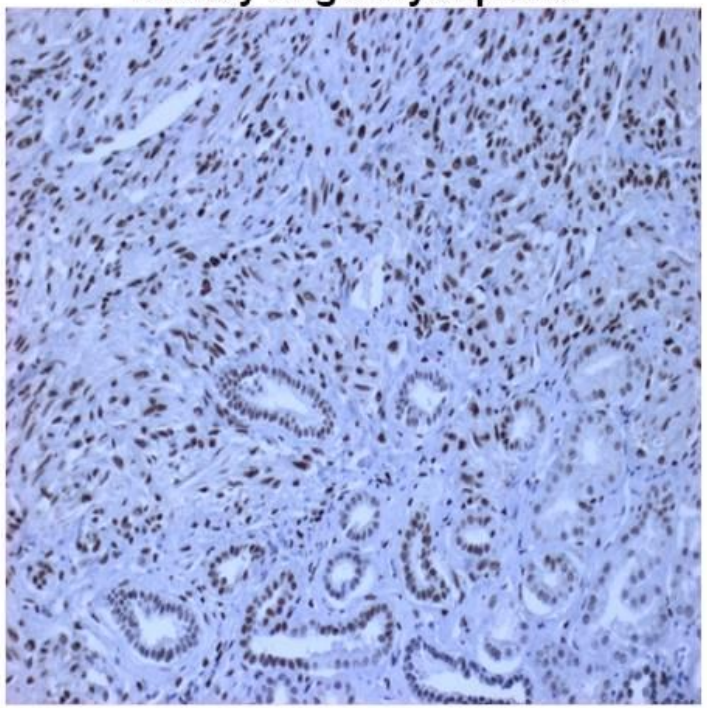
a)

Lung LAM

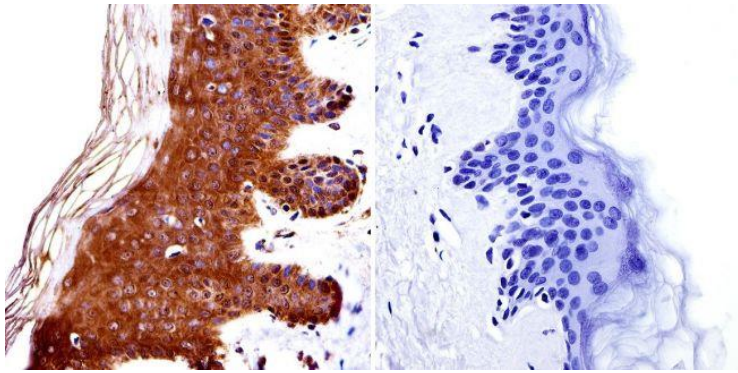


b)

Kidney angiomyolipoma



참고



Conclusion

- In this study, I conducted a GWAS in a large cohort of S-LAM subjects.
- Two intergenic SNPs, rs4544201 and rs2006950, were identified in a 34kb LD block on chromosome 15, that met genome-wide significance for association with LAM.
- The nearest protein-coding gene is ***NR2F2***, 700kb away, and consideration of chromatin TADs in this region indicates that only *NR2F2* is in/on the border of the TAD region containing the SNPs showing association with S-LAM in four relevant cells/tissues.
- *NR2F2* is highly expressed in LAM and angiomyolipoma by RNA-Seq analysis in comparison to large cancer and normal tissue data sets.
- *NR2F2* shows high expression with nuclear localization in both LAM and AML by IHC.

Conclusion

- LAM occurs nearly exclusively in women, and estrogen levels influence LAM development and progression.
- siRNA knockdown of ERα (Estrogen Receptor) in MCF-7 breast cancer cells decreased *NR2F2* expression, while treatment with estradiol increased its expression.
- This interaction between ERα and *NR2F2* may also play a role in LAM development.
- *NR2F2* has not previously been implicated in LAM, and these novel and unexpected findings will hopefully lead to better understanding of the pathogenesis of this often progressive and lethal lung disorder.

Chapter 4

Heritability Estimation of Dichotomous Phenotypes Using a Liability Threshold Model on Ascertained Family-based Samples

Notations

- i : family, $i = 1, \ldots, n$
- j : individual, $j = 1, \ldots, n_i$, $N = \sum_{i=1}^{n} n_i$
- $\mathbf{X}_{ij}$: environmental or genetic effects for subject j in family i
- U_{ij} : random effect including polygenic effect and random effect for subject j in family i
- L_{ij} : liability for subject j in famliy i
- c : threshold value determined from the prevalence of the disease
- Y_{ij} : the dichotomous phenotypes for subject j in family i

$$Y_{ij} = I\left(L_{ij} > c\right)$$

- Once Y_{ij} is observed, we can infer the range of the liability $\left(a_{ij}, b_{ij}\right)$, which are the lower and upper bound of L_{ij} respectively.

Notations

- Vector form

$$\mathbf{L_i} = \begin{pmatrix} L_{i1} \\ \vdots \\ L_{in_i} \end{pmatrix}, \mathbf{X_i} = \begin{pmatrix} \mathbf{X}_{i1} \\ \vdots \\ \mathbf{X}_{in_i} \end{pmatrix}, \mathbf{U_i} = \begin{pmatrix} U_{i1} \\ \vdots \\ U_{in_i} \end{pmatrix}, \mathbf{Y_i} = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \mathbf{a_i} = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{in_i} \end{pmatrix} \text{ and } \mathbf{b_i} = \begin{pmatrix} b_{i1} \\ \vdots \\ b_{in_i} \end{pmatrix}$$

and

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_1 \\ \vdots \\ \mathbf{L}_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix}, \mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \vdots \\ \mathbf{U}_n \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_n \end{pmatrix}, \mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{pmatrix}$$

Disease model

- We assumed that liability scores are normally distributed as follows,

$$\mathbf{L}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{U}_i, \mathbf{L}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$$

where $\mathbf{U}_i \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_i)$ and $\boldsymbol{\Sigma}_i = h^2\boldsymbol{\Phi}_i + (1-h^2)\mathbf{I}_{n_i}$.

- Joint probability density function (pdf) of the complete data

$$f(\mathbf{Y}, \mathbf{L}) = f(\mathbf{Y}|\mathbf{L})f(\mathbf{L}) = f(\mathbf{L})I(\mathbf{a} < \mathbf{L} < \mathbf{b})$$

- If we denote the parameters of interest as $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, h^2)^t$, then

$$l(\boldsymbol{\theta};\ \mathbf{Y}, \mathbf{L}) = \sum_{i=1}^{n}\left[-\frac{n_i}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{L}_i - \mathbf{X}_i\boldsymbol{\beta})^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{L}_i - \mathbf{X}_i\boldsymbol{\beta})\right].$$

Disease model

- When Families are ascertained due to probands,

$$\mathbf{Y}^P = \begin{pmatrix} Y_1^P \\ \vdots \\ Y_n^P \end{pmatrix},\ \mathbf{Y}^{NP} = \begin{pmatrix} \mathbf{Y}_1^{NP} \\ \vdots \\ \mathbf{Y}_n^{NP} \end{pmatrix} \text{ and } \mathbf{Y} = \begin{pmatrix} \mathbf{Y}^P \\ \\ \mathbf{Y}^{NP} \end{pmatrix}$$

- The conditional likelihood

$$f(\mathbf{Y}^{NP}|\mathbf{Y}^P;\boldsymbol{\theta}) = \frac{f(\mathbf{Y};\boldsymbol{\theta})}{f(\mathbf{Y}^P;\boldsymbol{\theta})}$$

- The log of the conditional likelihood

$$\log f(\mathbf{Y}^{NP}|\mathbf{Y}^P;\boldsymbol{\theta}) = l(\boldsymbol{\theta};\ \mathbf{Y}) - l(\boldsymbol{\theta};\ \mathbf{Y}^P)$$

Disease model

- Global lower bound of $\log f(\mathbf{Y}^{NP}|\mathbf{Y}^{P};\boldsymbol{\theta})$

$$\log f(\mathbf{Y}^{NP}|\mathbf{Y}^{P};\boldsymbol{\theta}) \geq \mathcal{F}(\boldsymbol{\theta}) - \mathcal{G}(\boldsymbol{\theta}).$$

  where $\mathcal{F}(\boldsymbol{\theta})$ is the lower bound of $l(\boldsymbol{\theta};\ \mathbf{Y})$ and $\mathcal{G}(\boldsymbol{\theta})$ is the upper bound of $l(\boldsymbol{\theta};\ \mathbf{Y}^{P})$.

- At $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, $\mathcal{F}(\boldsymbol{\theta})$ can be obtained by:

$$\mathcal{F}(\boldsymbol{\theta}) = E_{\mathbf{L}|\mathbf{Y},\boldsymbol{\theta}^{(k)}}\left(l(\boldsymbol{\theta};\ \mathbf{Y},\mathbf{L})\right) + H\left(f\left(\mathbf{L}|\mathbf{Y},\boldsymbol{\theta}^{(k)}\right)\right)$$

  where $H(\cdot)$ is the entropy.
- $\mathcal{G}(\boldsymbol{\theta})$ for $l(\boldsymbol{\theta};\ \mathbf{Y}^{P})$ can be defined as $l(\boldsymbol{\theta};\ \mathbf{Y}^{P}) + \text{constant}$.

Disease model

- Therefore,

$$\mathcal{F}(\boldsymbol{\theta}) - \mathcal{G}(\boldsymbol{\theta}) = E_{\mathbf{L}|\mathbf{Y},\boldsymbol{\theta}^{(k)}}\left(l(\boldsymbol{\theta};\, \mathbf{Y},\mathbf{L})\right) - l(\boldsymbol{\theta};\, \mathbf{Y}^P) + \text{constant.}$$

Expectation step of EM algorithm for randomly selected families

Generalized linear model

- $l(\boldsymbol{\theta};\, \mathbf{Y}^P)$ is simply given by:

$$l(\boldsymbol{\beta};\, \mathbf{Y}^P) = \sum_{i=1}^{n} l(\boldsymbol{\beta}; Y_i^P) = \sum_{i=1}^{n} [Y_i^P \alpha_i - \log(1 + e^{\alpha_i})] \text{ where } \alpha_i = \log\frac{\mu_i}{1-\mu_i}$$

$$\mu_i = E(Y_i^P) = \Pr(Y_i^P = 1) = \Pr(L_i^P > c) = 1 - \Phi(c - \mathbf{X}_i^P\boldsymbol{\beta})$$

Conditional Expected Score Test (CEST)

- By Fisher (1925),

$$\frac{\partial l(\boldsymbol{\theta};\, \mathbf{Y})}{\partial \boldsymbol{\theta}} = E_{\mathbf{L}|\mathbf{Y}}\left[\frac{\partial l(\boldsymbol{\theta};\, \mathbf{Y}, \mathbf{L})}{\partial \boldsymbol{\theta}}\right].$$

- The conditional expected score (CES) for family *i*

$$\mathbf{S}_i = E_{\mathbf{L}|\mathbf{Y}}\begin{bmatrix}\frac{\partial l_i(\boldsymbol{\theta};\, \mathbf{Y}, \mathbf{L})}{\partial \boldsymbol{\beta}} \\ \frac{\partial l_i(\boldsymbol{\theta};\, \mathbf{Y}, \mathbf{L})}{\partial h^2}\end{bmatrix} = \begin{bmatrix}\mathbf{X}_i^t \boldsymbol{\Sigma}_i^{-1} \mathbf{B}_i - \mathbf{X}_i^t \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i \boldsymbol{\beta} \\ -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Phi}_i - \mathbf{I}_{n_i}\right)\right) - \frac{1}{2}\mathrm{tr}(\mathbf{C}_i \mathbf{A}_i) + \boldsymbol{\beta}^t \mathbf{X}_i^t \mathbf{C}_i \left(\mathbf{B}_i - \frac{1}{2}\mathbf{X}_i \boldsymbol{\beta}\right)\end{bmatrix}$$

where $\mathbf{A}_i = E_{\mathbf{L}|\mathbf{Y}}(\mathbf{L}_i \mathbf{L}_i^t)$, $\mathbf{B}_i = E_{\mathbf{L}|\mathbf{Y}}(\mathbf{L}_i)$ and $\mathbf{C}_i = \partial \boldsymbol{\Sigma}_i^{-1} / \partial h^2$.

Conditional Expected Score Test (CEST)

- The observed Fisher information matrix is given by

$$\hat{I}(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\mathbf{S}_i \mathbf{S}_i^t) - \frac{1}{n} \left(\sum_{i=1}^{n} \mathbf{S}_i \right) \left(\sum_{i=1}^{n} \mathbf{S}_i^t \right)$$

and it is equivalent to

$$\hat{I}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{i}_{\boldsymbol{\beta}} & \boldsymbol{i}_{\boldsymbol{\beta} h^2} \\ \boldsymbol{i}_{h^2 \boldsymbol{\beta}} & i_{h^2} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \left(\mathbf{S}_{\boldsymbol{\beta} i} \mathbf{S}_{\boldsymbol{\beta} i}^t \right) - \mathbf{S}_{\boldsymbol{\beta}} \mathbf{S}_{\boldsymbol{\beta}}^t / n & \sum_{i=1}^{n} \left(\mathbf{S}_{\boldsymbol{\beta} i} S_{h^2 i} \right) - \mathbf{S}_{\boldsymbol{\beta}} S_{h^2} / n \\ \sum_{i=1}^{n} \left(S_{h^2 i} \mathbf{S}_{\boldsymbol{\beta} i} \right) - S_{h^2} \mathbf{S}_{\boldsymbol{\beta}}^t / n & \sum_{i=1}^{n} \left(S_{h^2 i}^2 \right) - S_{h^2}^2 / n \end{pmatrix}$$

Conditional Expected Score Test (CEST)

- Score statistics for H_0: $\boldsymbol{\beta} = \mathbf{0}$

$$\mathbf{S}_{\boldsymbol{\beta}}^{\boldsymbol{t}} \left\{ \boldsymbol{i}_{\boldsymbol{\beta}} - \boldsymbol{i}_{\boldsymbol{\beta}\widehat{h^2}} i_{\widehat{h^2}}^{-1} \boldsymbol{i}_{\widehat{h^2}\boldsymbol{\beta}} \right\}^{-\mathbf{1}} \mathbf{S}_{\boldsymbol{\beta}} \sim \chi^2(df = p) \text{ under } H_0: \boldsymbol{\beta} = \mathbf{0}$$

- Score statistics for H_0: $h^2 = 0$

$$\mathbf{S}_{h^2}^{\boldsymbol{t}} \left\{ i_{h^2} - \boldsymbol{i}_{h^2\widehat{\boldsymbol{\beta}}} \boldsymbol{i}_{\widehat{\boldsymbol{\beta}}}^{-1} \boldsymbol{i}_{\widehat{\boldsymbol{\beta}}h^2} \right\}^{-\mathbf{1}} \mathbf{S}_{h^2} \sim \frac{\mathbf{1}}{\mathbf{2}} \cdot \mathbf{0} + \frac{\mathbf{1}}{\mathbf{2}} \cdot \chi^2(df = 1) \text{ under } H_0: h^2 = 0$$

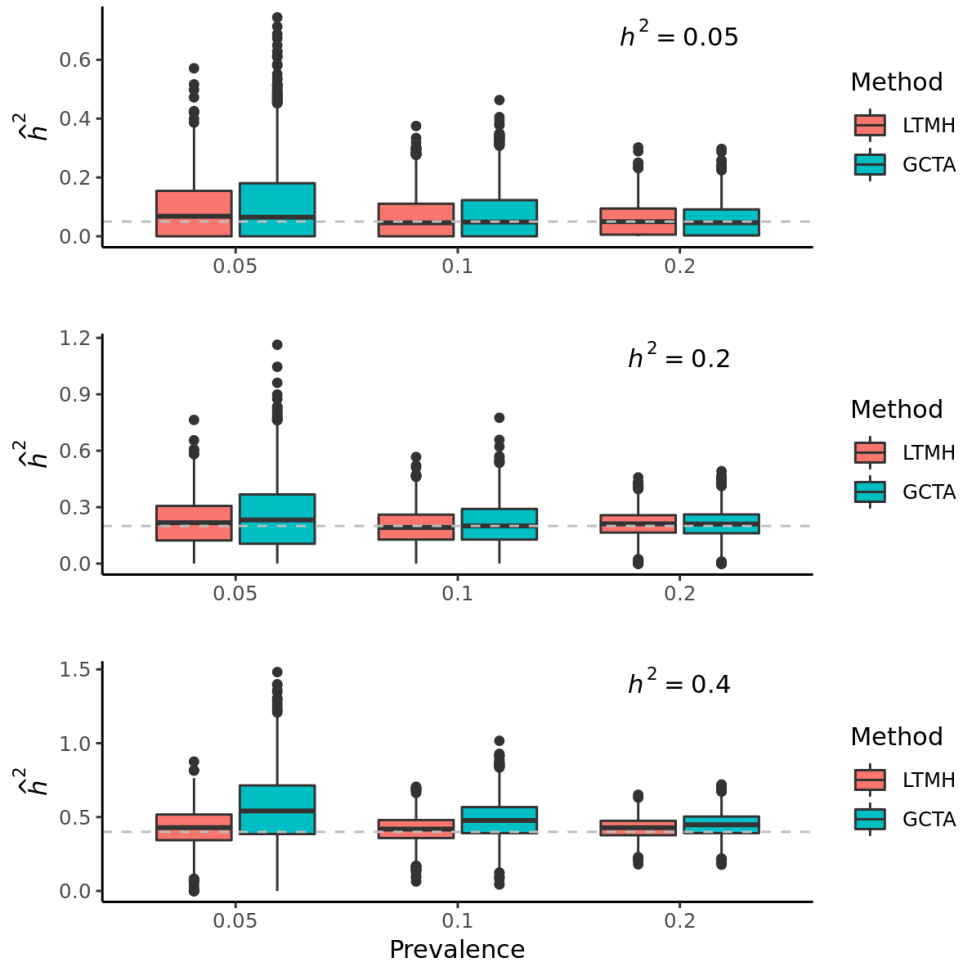
Simulation study

- Simulation settings
  - One causal SNP with MAF of 0.2 & h_a^2 : 0.005 (β=0.1253).
  - No environmental effect
  - h^2 : 0.05, 0.2 and 0.4
  - q : 0.05, 0.1 and 0.2
- Simulation studies were conducted under two different scenarios where **families were either randomly selected (scenario 1)** or **ascertained with probands (scenario 2)**.
  - For Scenario 1, 500 families were generated.
  - For Scenario 2, 50,000 families were generated and 500 probands were selected from affected individuals.
- All results were compared with GCTA.

Simulation study (2,000 replicates)

- **Accuracy of $\widehat{\beta}$ and $\widehat{h^2}$ from randomly selected families (scenario 1)**

Heritability	Prevalence	LTMH		GCTA
		β	h^2	h^2
0.05	0.05	0.1226 (0.0223)	0.0933 (0.0971)	0.1105 (0.1303)
	0.1	0.1281 (0.0181)	0.0660 (0.0716)	0.0734 (0.0828)
	0.2	0.1277 (0.016)	0.0584 (0.0538)	0.0563 (0.0539)
0.2	0.05	0.1267 (0.0223)	0.2184 (0.1282)	0.2511 (0.1852)
	0.1	0.1239 (0.0190)	0.1950 (0.0993)	0.2111 (0.1219)
	0.2	0.1285 (0.0164)	0.2106 (0.0725)	0.2115 (0.0775)
0.4	0.05	0.1309 (0.0229)	0.4324 (0.1313)	0.5546 (0.2437)
	0.1	0.1276 (0.0225)	0.4230 (0.1315)	0.4825 (0.1377)
	0.2	0.1286 (0.0189)	0.4181 (0.0950)	0.4486 (0.085)



Simulation study (2,000 replicates)

- **Accuracy of $\widehat{\beta}$ and $\widehat{h^2}$ from ascertained families (scenario 2)**

Heritability	Prevalence	LTMH		GCTA
		β	h^2	h^2
0.05	0.05	0.1335 (0.0193)	0.0474 (0.0376)	1.72×10^{-6} (4.47×10^{-7})
	0.1	0.1233 (0.0181)	0.0336 (0.0339)	1.96×10^{-6} (2.01×10^{-7})
	0.2	0.1194 (0.0144)	0.0304 (0.0287)	1.83×10^{-6} (3.78×10^{-7})
0.2	0.05	0.1234 (0.0199)	0.2018 (0.0437)	1.01×10^{-6} (9.18×10^{-8})
	0.1	0.1257 (0.0135)	0.2086 (0.0342)	0 (0)
	0.2	0.1239 (0.0153)	0.1692 (0.0407)	1.01×10^{-6} (7.40×10^{-8})
0.4	0.05	0.1358 (0.0189)	0.4004 (0.0449)	0 (0)
	0.1	0.1167 (0.0144)	0.3868 (0.0339)	0 (0)
	0.2	0.1186 (0.0150)	0.4090 (0.0444)	0 (0)

Simulation study (2,000 replicates)

- **Type-1 error and power estimates of the proposed test for $H_0: h^2 = 0$ under scenario 1 and scenario 2**

Scenario 1

Heritability	Prevalence	Significance level		
		0.01	0.05	0.1
0	0.05	0.0015	0.0115	0.0285
	0.1	0.0050	0.0480	0.1020
	0.2	0.0015	0.0200	0.0505
0.2	0.05	0.0485	0.2260	0.3990
	0.1	0.3420	0.6730	0.8055
	0.2	0.6210	0.8675	0.9405
0.4	0.05	0.4575	0.8190	0.9050
	0.1	0.9395	0.9930	0.9960
	0.2	1.0000	1.0000	1.0000

Scenario 2

Heritability	Prevalence	Significance level		
		0.01	0.05	0.1
0	0.05	0.0000	0.0025	0.0100
	0.1	0.0005	0.0045	0.0095
	0.2	0.0000	0.0075	0.0215
0.2	0.05	0.4735	0.8110	0.9185
	0.1	0.8520	0.9660	0.9850
	0.2	0.8155	0.9540	0.9855
0.4	0.05	1.0000	1.0000	1.0000
	0.1	1.0000	1.0000	1.0000
	0.2	1.0000	1.0000	1.0000

Simulation study (2,000 replicates)

- **Type-1 error and power estimates of the proposed test for $H_0: \beta = 0$ under scenario 1**

h_a^2	Heritability	Prevalence	Significance level		
			0.01	0.05	0.1
0	0.2	0.1	0.0155	0.0661	0.1023
		0.2	0.0120	0.0560	0.0900
	0.4	0.1	0.0060	0.0480	0.0940
		0.2	0.0130	0.0580	0.1020
0.005	0.2	0.1	0.1303	0.3372	0.4713
		0.2	0.4460	0.6800	0.7980
	0.4	0.1	0.2740	0.5340	0.6640
		0.2	0.3540	0.6000	0.7180

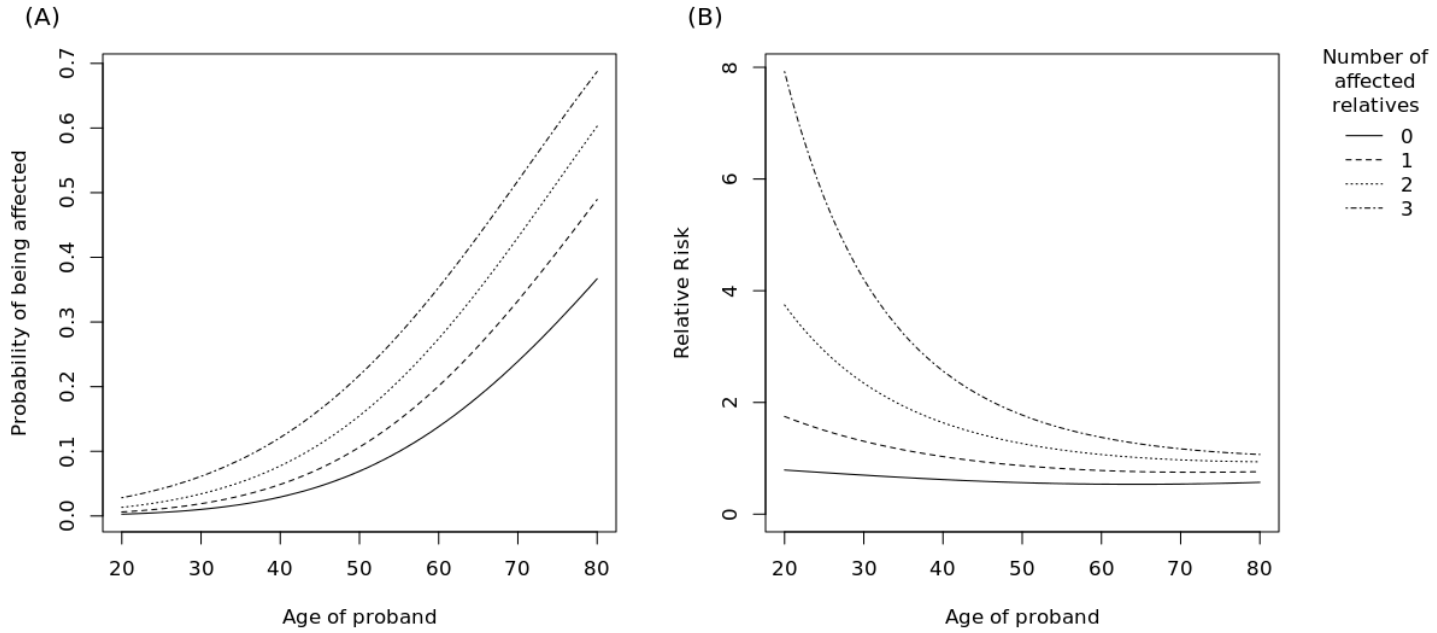
Real data analysis

- Applications of LTMH and CEST to T2D
  - Seoul national university hospital recruited 681 probands affected to T2D.
  - This study preferentially included T2D patients with a positive family history of T2D in first-degree relatives.
  - After excluding subjects without age information, we have 648 probands and 4,149 non-probands.

	Proband	Non-proband
Disease status		
T2D†	648 (100%)	1,115 (26.87%)
Normal	0 (0%)	3,034 (73.13%)
Sex		
Male	308 (47.53%)	2,058 (49.6%)
Female	340 (52.47%)	2,091 (50.4%)
Age	55.44 (10.7)	47.56 (16.09)
Relationship of relatives with proband		
Parents		329 (7.93%)
Sibling		2,457 (59.22%)
Offspring		1,363 (32.85%)

Real data analysis

- Applications of LTMH and CEST to T2D
  - Estimated heritability of T2D was 29.44% (P-value = 1.20×10^{-5}).
  - This finding is slightly overestimated in comparison to other determinations of heritability estimates for T2D (26%) using the ACE model based on twin data (Poulsen, P., et al, 1999)
  - The coefficient estimate for non-standardized age was 0.051.

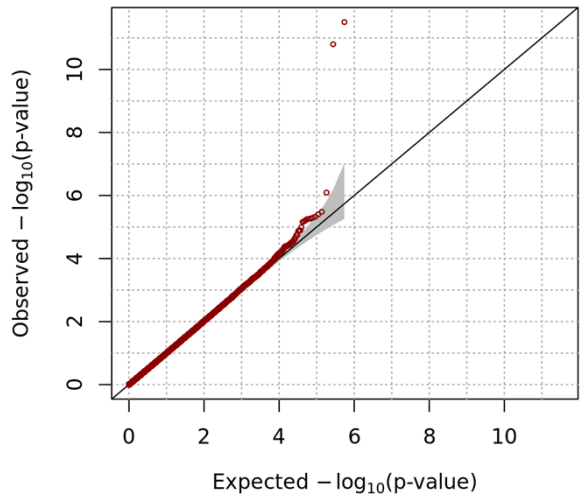


Real data analysis

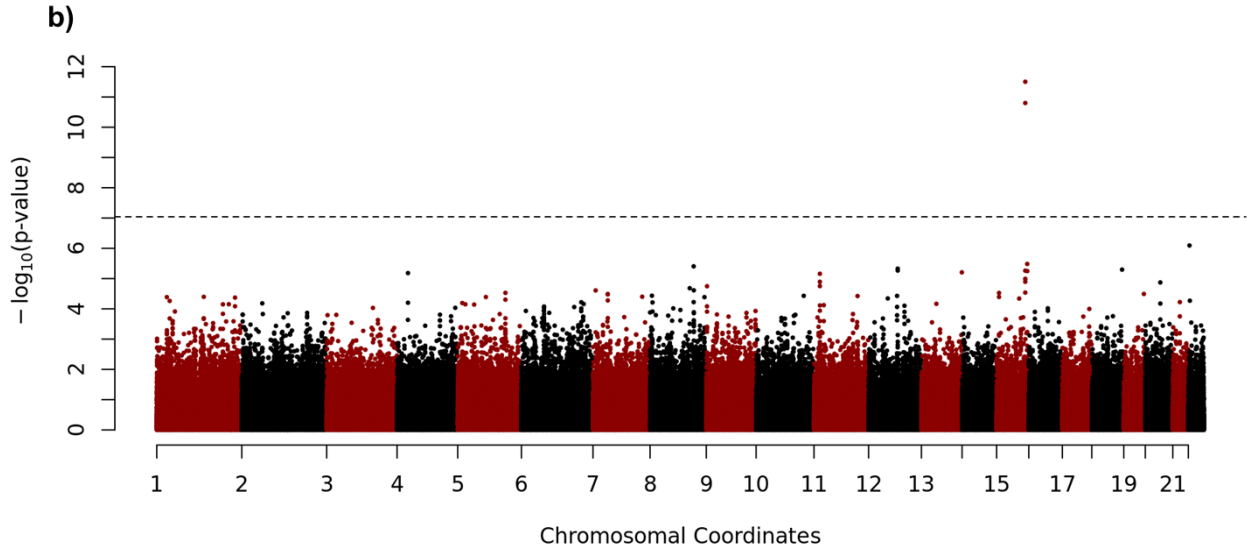
- Applications of CEST to S-LAM disease
  - Using matched cases and controls for CLR (Chapter 2), each pair of one case and two controls was regarded as if a family having relatedness structure of genetic relationship matrix.

SNP	CLR	CEST
rs4544201	8.51×10^{-10}	1.58×10^{-11}
rs2006950	3.92×10^{-10}	3.14×10^{-12}

a)



b)



Conclusion

- In this study, I proposed a new method to estimate the heritability of a dichotomous trait based on the Liability Threshold Model for ascertained family-based samples.
- A simulation study demonstrated that LTMH generally provides more accurate estimates of heritabilities than does GCTA, and the differences between these methods are substantial in the context of ascertained families.
- Statistical performance of CEST analysis were also assessed. Despite of conservative property of CEST, I found that such inflation does not affect the statistical power of our analysis but certain modifications such as bootstrapping are necessary.

Conclusion

- However, there is a limitation that the proposed method is computationally intensive when the family size is large.
- The most significant computational burden arises from the calculation of conditional expectation in the E-step of the EM algorithm.
- Therefore, the computational burden can be reduced by reducing the number of iterations for the EM algorithm or by approximating the moment of the multivariate truncated normal.
- Despite several limitations, our proposed method successfully enabled heritability estimation of dichotomous traits in ascertained families, and this method may provide a promising strategy to estimate the narrow-sense heritability of various diseases.

Revision

예비심사 지적 사항 수정 보고

1. **Introduction에 연구의 목적을 조금 더 구체적으로 명시할 것**
   - (5p) Chapter 1.3 The purpose of this study 추가하였습니다.
   - 그 외에 각 chapter별로 Introduction에 각 연구 별 연구목적 구체적으로 명시하였습니다.

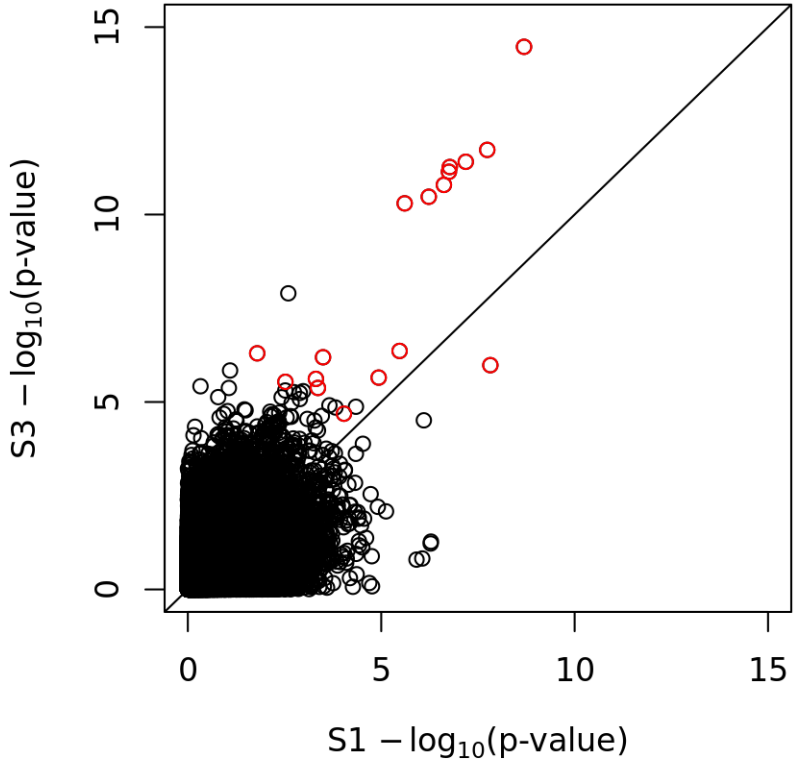
2. **Chapter 3의 바탕이 된 게재된 논문을 key reference로 넣을 것**
   - (49p) 아래의 문구 삽입하였습니다.
     "This chapter was published in Statistics in Medicine as a partial fulfillment of Wonji Kim's Ph.D program."
   - (50p) reference 삽입하였습니다. (reference 81)

3. **Chapter 3의 Conditional expectation에 왜 selection bias가 들어가지 않는지 표시할 것**
   - (84p) 아래의 문구 삽입하였습니다.
     "Moreover, the use of subjects with extreme phenotypes in GWAS is not the case for selection bias because the choice of subjects is based on phenotype, not on genotype."

예비심사 지적 사항 수정 보고

4. **Chapter 3의 real data 분석에서 각 방법 별 성능 비교를 위하여 Pairwise P-value plot을 그릴 것**
   - (83p) 두 방법의 P-value를 이용한 산점도 그림을 아래와 같이 추가하였습니다.

Figure 3.9 Scatter plot for P-values of GWAS of type 2 diabetes using S1 and S3. Red dots indicate significant SNPs when all subjects are used for GWAS.



예비심사 지적 사항 수정 보고

5. **Chapter 4에서 기존의 연구와 본인의 연구의 구분을 명확히 할 것**
   - 무작위로 추출된 가족 혹은 case-control 연구에 대한 유전율 추정 알고리즘은 기존에 잘 개발이 되어 있으나 ascertained family에 대한 연구는 활발하게 이루어지지 않았습니다.

   - 본 논문의 연구 주제는 이분형 표현형의 유전율 추정을 다루고 있으며, 특히 proband에 의하여 분석에 참여하게 된 ascertained family들에 대한 유전율 추정 알고리즘을 주 연구주제로 하고 있습니다. 본 연구의 방법을 통하여 기존의 알고리즘에서 발생했던 편차를 효과적으로 보정할 수 있었습니다.

   - (124p) 이러한 내용을 4.4 Discussion의 첫 번째 단락에 서술하였습니다.

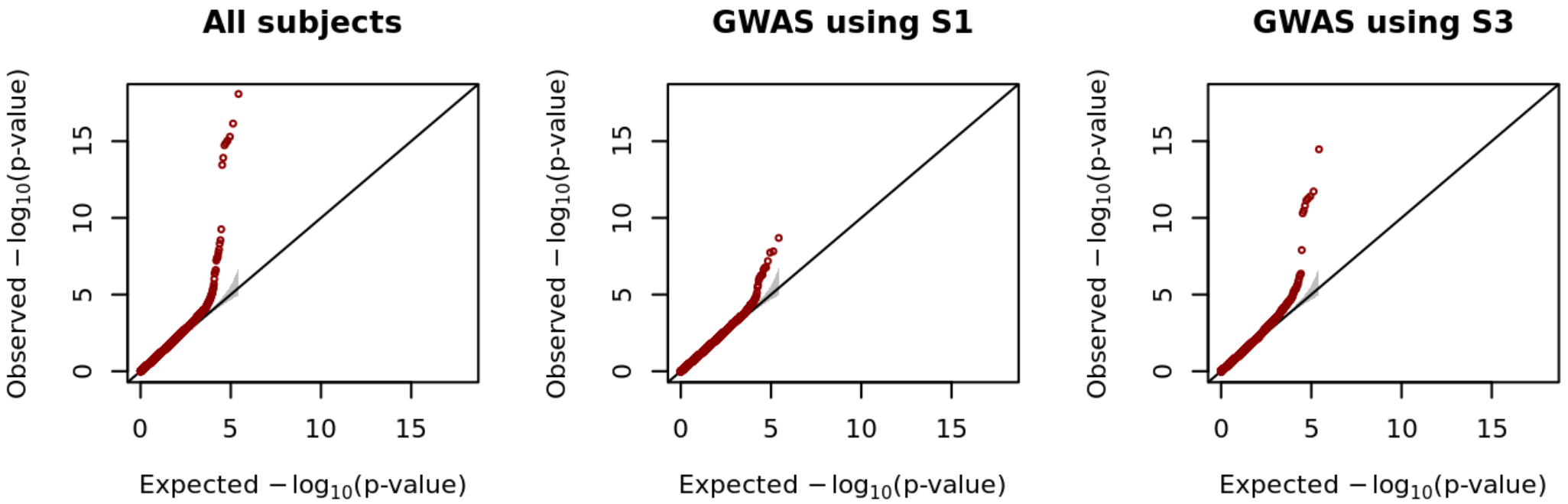
예비심사 지적 사항 수정 보고

6. **Chapter 4의 Score test의 variance 추청과 관련하여 Bootstrap 방법을 사용하면 계산량이 너무 증가할 것. 다른 대안을 찾아볼 것**
   - Bootstrapping 방법은 계산량이 너무 많아지는 단점이 존재하기 때문에 실질적 적용에 한계가 있었습니다.
   - 따라서 본 연구에서는 Information matrix의 유사값을 이용하여 variance를 추정하였고, discussion에 bootstrap의 사용 가능성에 대하여 명시하였습니다. (104p, 124p)

예비심사 지적 사항 수정 보고

6. QQ plot scale 맞추기

– (Figure 3.2, 3.3, 3.4, 3.6) 그림 수정 하였습니다.



THANK YOU

학위논문 심사 서류

- 모든 교수님들께서 작성 및 날인 해주셔야 하는 서류
  - 구술고사 성적표 (서식 10)
  - 투표용지 (서식 11)
  - 박사학위 논문심사 결과표 (서식 13)
  - 인준지

- 심사위원장 교수님께서 추가적으로 작성 및 날인해주셔야 하는 서류
  - 박사학위 논문 예비심사 결과보고 (서식 9)
  - 구술고사 성적표 (서식 10)
  - 투표용지 (서식 11)
  - 박사학위 논문심사요지 (서식 12)