TOHOKU
UNIVERSITY

Integration of genomic and exposomic findings to realize the precision medical care in Tohoku Medical Megabank

Division of Bioclinical Imformatics
Tohoku Medical Megabank Organization,
Tohoku University

Hiroshi Tanaka, Soichi Ogishima and Satoshi Nagaie

東北メディカル・メガバンク機構
TOHOKU MEDICAL MEGABANK ORGANIZATION

Today's Topics

- Need of Integrative Analysis and Integrative Database including genomic and exposomic data
- Tohoku Medical Megabank Integrated database "dbTMM"
- A New Disease Risk Method Taking in the Interaction of G x E

Why we need
Integrative Analysis and
Integrative Database including
genomic and exposomic
(environmental) data

Personalized Prediction/Prevention of Disease

Based on follow-up data, estimate risk of disease

One of the **Major goals** of Tohoku Medical Megabank

Specially Attacked Challenge is

To predict and prevent the occurrence of
multi-factorial (complex) disease
(Common diseases; Hypertension, Type II Diabetes)

Current **Genome Medicine** Approach

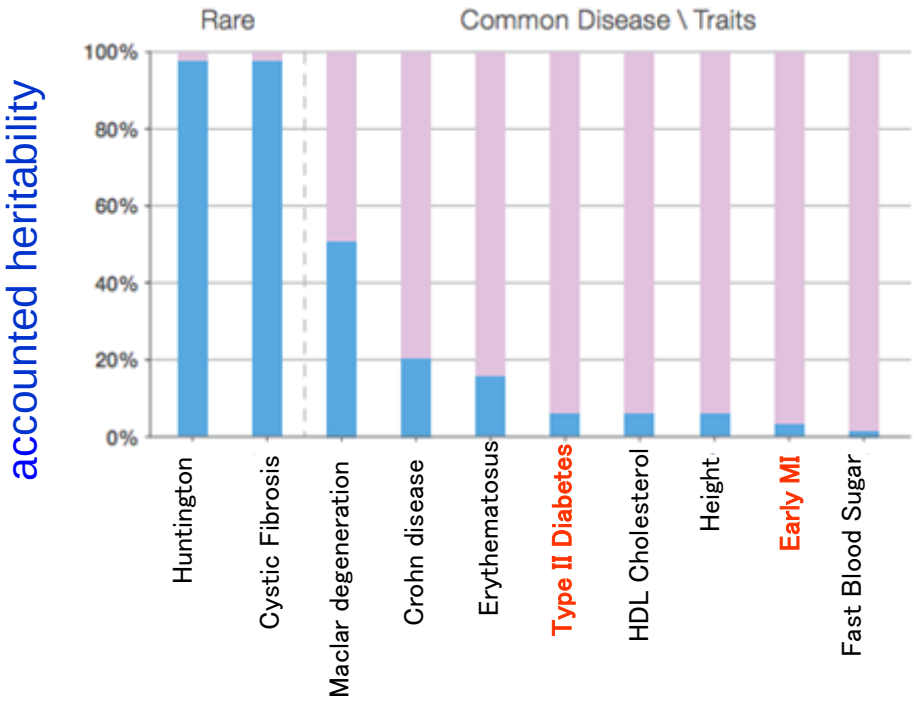
Succeeded in

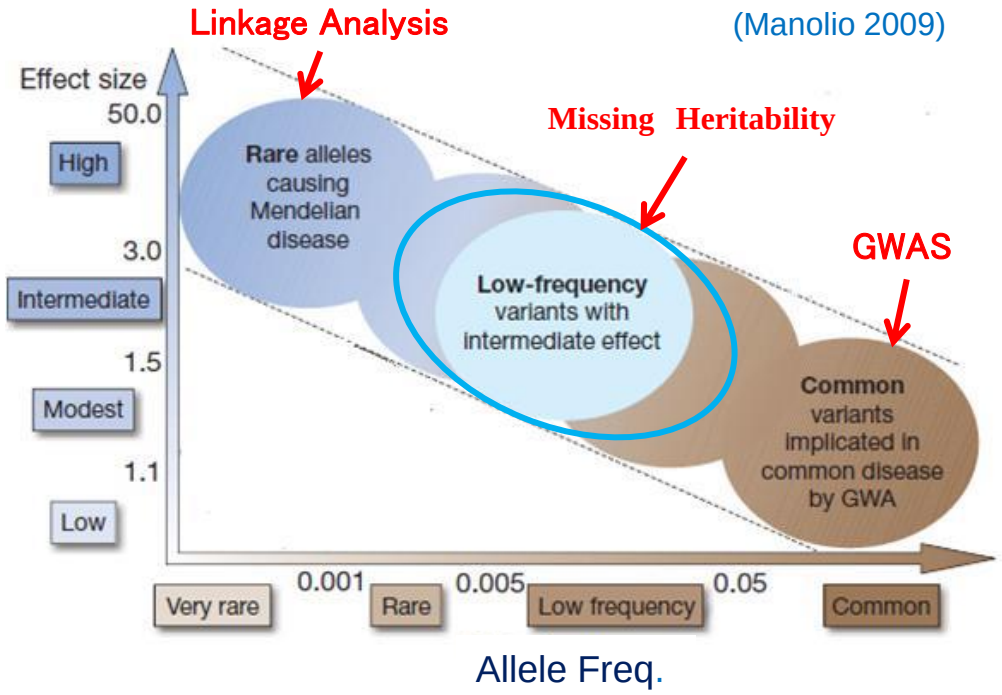
1. Identify **Causative Gene** at POC for rare/undiagnosed disease
2. Identify **Driver mutation of Cancer** for Molecular Targeted Drug
3. **Preemptive PGx** based on identifying Molecular Polymorphism of Drug Metabolizing Enzyme

But

Totally **Ineffective** for multifactorial complex disease

Ineffectiveness of Current Genomic Method





- Limitation of "**single genetic cause approach**"
  - Explore the single genetic cause of disease (**single** gene or polymorphism) without any reference to **effects of interaction with other genes**.
  - **Genomic Big data** (due to **p>>n problem**: ordinary statistics does not work because) makes the **multivariate analysis (using more than two SNPs) substantially impossible**.
- **Missing Heritability** might be due to not involving the interaction terms
  - Interaction **among genes** : "epistasis" (genes on the same pathway), GxG
  - Interaction **between genes and environment**: G x E (x means interaction)

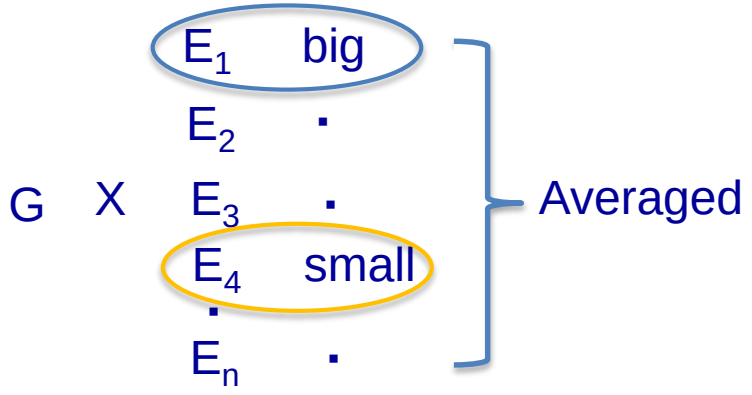
Interaction between gene and environment

Except rare monogenetic disease, most of diseases result from a complex interaction between an individual's genetic make-up and the environmental (exposomic) agents that he or she is exposed to.

Relative Risk= Complex Interaction between G and E

- Neither additive ($G \oplus E$), nor multiplicative ($G \otimes E$)
- <(G,E) Combination - Specific> Effect

The reason why Relative Risk of SNP (GWAS) is so small (1.1～1.3), combinatory effects are **averaged** on the side of **all the environment factors**



Example of <combination-specific> relative risk

- Typical Example: Interaction of genomic and environmental factor
  - Nether **additive**, nor **multiplicative**
- Colon cancer RR study
  - Survey in Hawaii
    - Le Marchand 2001
  - **E**: Smoking, Well-done red meat
  - **G**: CYP1A2, NAT2

relative risk of disease occurrence is

Combination - specific

		CYP1A2 Phenotype ≦Median		CYP1A2 Phenotype ＞Median	
		Likes rare/medium meat	Likes well-done meat	Likes rare/medium meat	Likes well done meat
Non-Smoker	NAT2 Slow	1	1.9	0.9	1.2
	NAT2 Rapid	0.9	0.8	0.8	1.3
Ever-Smoker	NAT2 Slow	1	0.9	1.3	0.6
	NAT2 Rapid	1.2	1.3	0.9	8.8

L. Le Marchand, JH. Hankin, LR. Wilkens, et alCombined Effects of Well-done Red Meat, Smoking, and Rapid N-Acetyltransferase 2 and CYP1A2 Phenotypes in Increasing Colorectal Cancer Risk, Cancer Epidemiol. Biomarkers Prev 2001;10:1259-1266

Necessity of Integrated Database

Thus, we should **comprehensively** inquire **the effects** on disease occurrence by **every combination of genomic and exposomic (environmental) factors**.

Tohoku Medical Megabank (TMM) collects vast kinds of data
Genomic, Exposomic (environmental) and Phenomic data

Not confined to the above purpose, but for the various comprehensive analysis of genomic cohort,

Integrated Database is necessary, which

- Store genomic, exposomic/phenomic data in **same data space**
- Can **transversally** retrieve the subjects (participants) satisfying both genomic and exposomic/phenomic conditions

As the Basis for **personalized prediction/prevention** of disease

- Taking into account all kinds of participant data and **temporal relationship** for **life course analysis**

New Disease Risk Method

Taking in the Interaction of G x E

New Disease Risk Method
Taking in the Interaction of G x E

The risk of disease is GxE combination-specific

Thus, we should **comprehensively** inquire **the effects** on disease occurrence by **every combination of genomic and exposomic (environmental) factors**.

At the first step, evaluate the effect of **one to one** relation

<one of genetic variates> <one of expsomic factors >

$\mathbf{G_i}$ x $\mathbf{E_j}$

To collect the result (RR or p-value) of each G_i x E_j, Risk distribution for overall GxE combination is obtained as the 2 dimensional landscape of RR or p-value

This is the first step of GxE risk estimation method. We will proceed to deal with more plural GxE factors

GE-WAS Analysis Flow

G_i Genomic Data

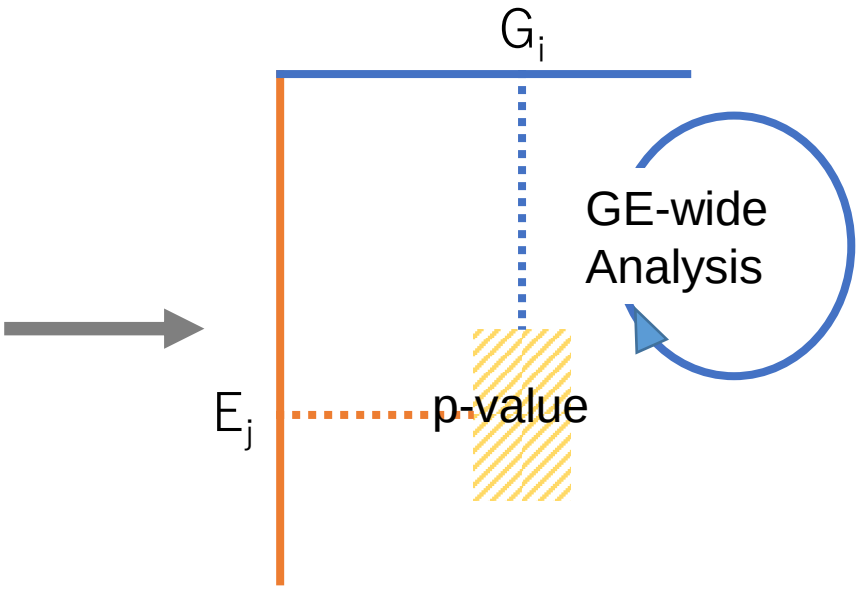
E_j Environmental Data

T Disease Phenotype

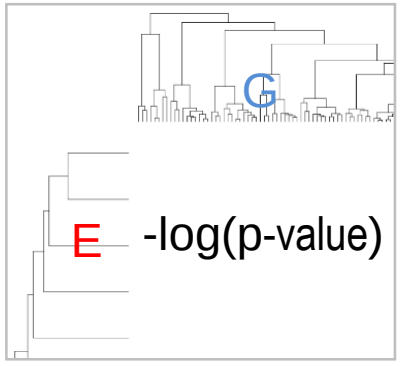
Count

$T_{Control}$		G_i			
T_{Case}		aa	aA	AA	A
E_j	0	N^1_{ij}	N^3_{ij}	N^5_{ij}	$^5_{ij}$
	1	N^2_{ij}	N^4_{ij}	N^6_{ij}	$^6_{ij}$

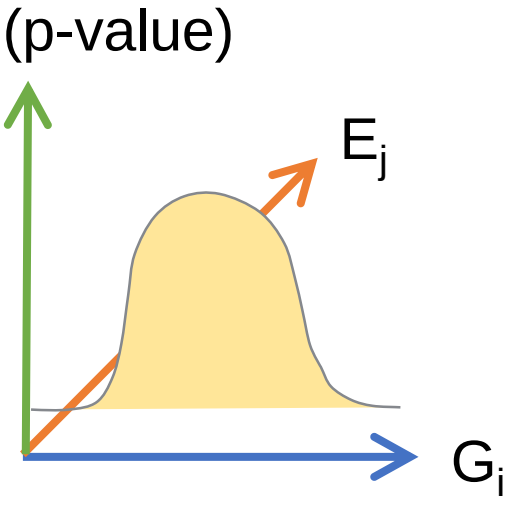
Cochran-Mantel-Haenszel Chi-Squared Test



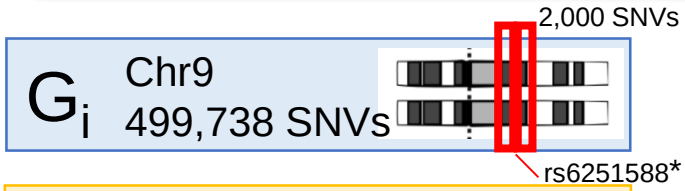
G・E Clustering



3D Plot



GE-WAS Analysis Flow



G_i Chr9 499,738 SNVs

E_j Laboratory Test 41 variables (normal/abnormal)

T Allergic rhinitis: 167 subjects Control: 901 subjects

ToMMo 1KJPN

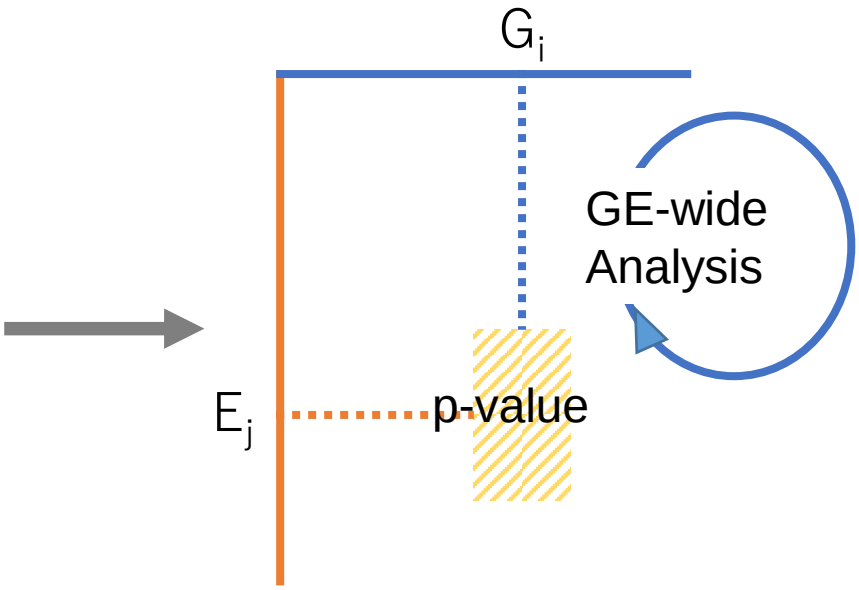
Count

$T_{Control}$		G_i			
T_{Case}		aa	aA	AA	A
E_j	0	N^1_{ij}	N^3_{ij}	N^5_{ij}	$^5_{ij}$
	1	N^2_{ij}	N^4_{ij}	N^6_{ij}	$^6_{ij}$

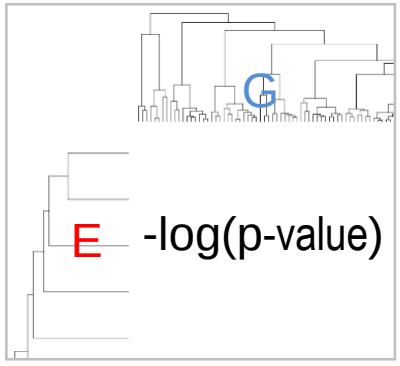
41 variables

1,237 SNVs

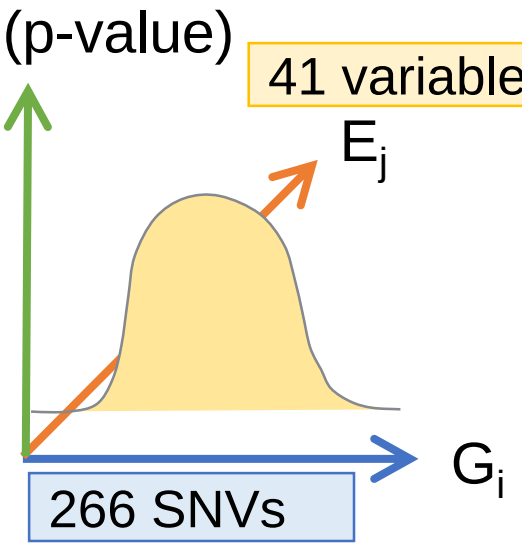
Cochran-Mantel-Haenszel Chi-Squared Test



G・E Clustering



3D Plot



41 variables

266 SNVs

(-log $\bar{p}$ > 0.30)

*Sakashita, Clin Exp Allergy, 2008

Laboratory Test Items

No.	Test Name	No.	Test Name	No.	Test Name	No.	Test Name	No.	Test Name
1	NT-proBNP	11	Common silver birch	21	Crab	31	Leukocyte count	41	IgE
2	Albumin/Cre	12	House dust mite	22	Milk	32	erythrocyte count		
3	Albumin	13	House dust	23	Beaf	33	hemoglobin content		
4	Creatinine (blood)	14	Penicillium notatum	24	Egg white	34	hematocrit		
5	Timothy	15	Candida albicans	25	Peanut	35	mean red cell volume		
6	Sweet vernal grass	16	Cat dander	26	Antibody concentration	36	concentration concentration		
7	Cocksfoot	17	Dog dander	27	Urea nitrogen	37	blood platelet count		
8	Common ragweed	18	Cultivated wheat	28	Uric Acid	38	lymph corpuscle		
9	Mugwort	19	Rice	29	Glucose	39	Acidocyte		
10	Grey alder	20	Shrimp	30	glycoalbumin	40	neutrophil		

GxE Landscape of “Allergic rhinitis“

-log p=26.7

Hay Fever
(Schröder 2016)

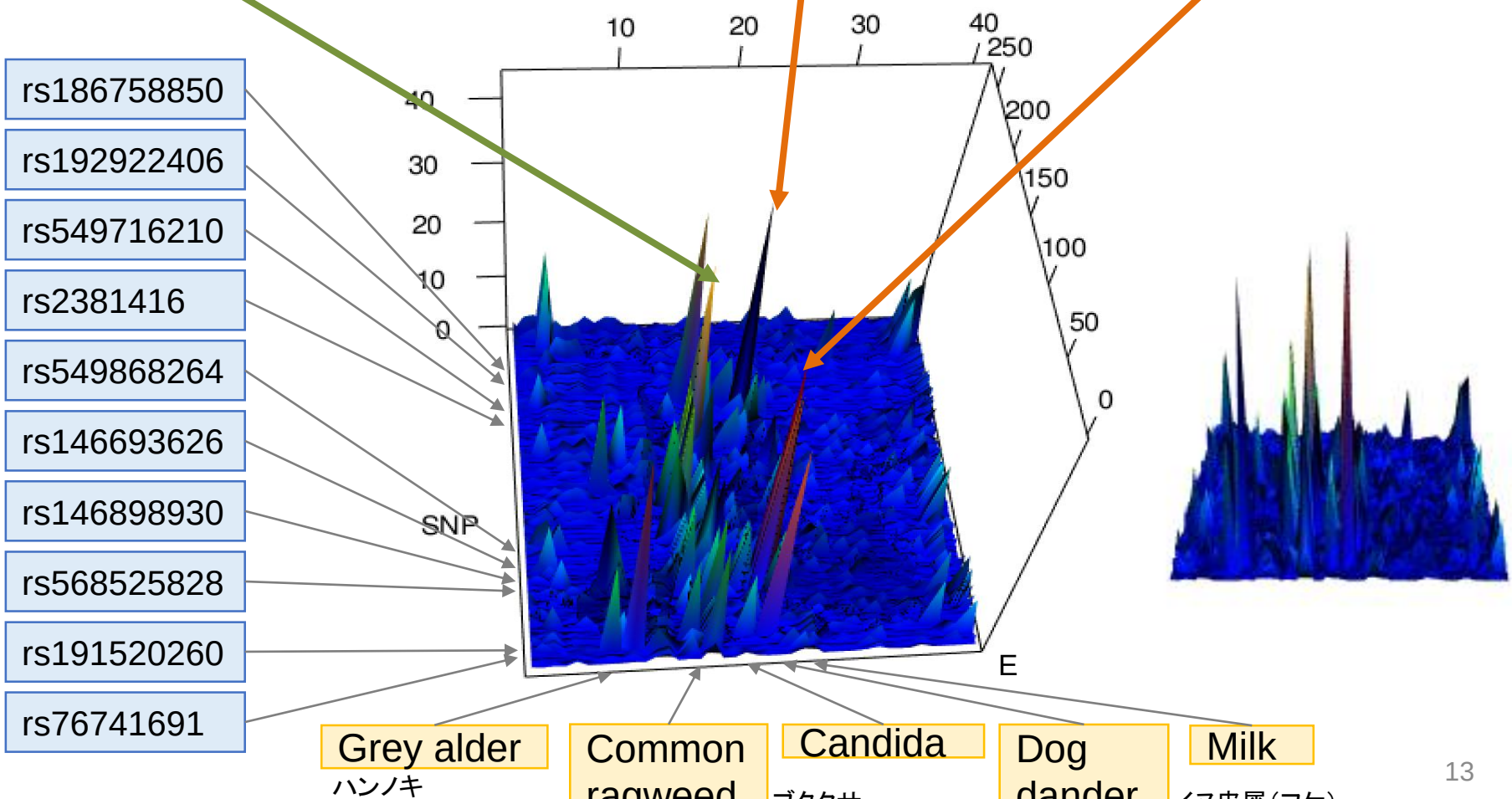
G: rs2381416
E: Candida albicans

-log p=35.7

G: rs549716210
E: Dog dander

-log p=43.4

G: rs549868264
E: Milk



Discussin of Our Method

- Our method comprehensively calculates the significance levels of p_{ij} for every <one genetic factor G_i> x <one exposomic factor E_j> contingency table.
- From the result of our example, effect of GxE is found to be **combination – specific**.
- For designated SNP, some exposomic factor produce a large effect whereas other factor does not.
- "**Relative risk of SNP for contracting disease**" the previous GWAS's concept is found to be

  quite misleading false idea.

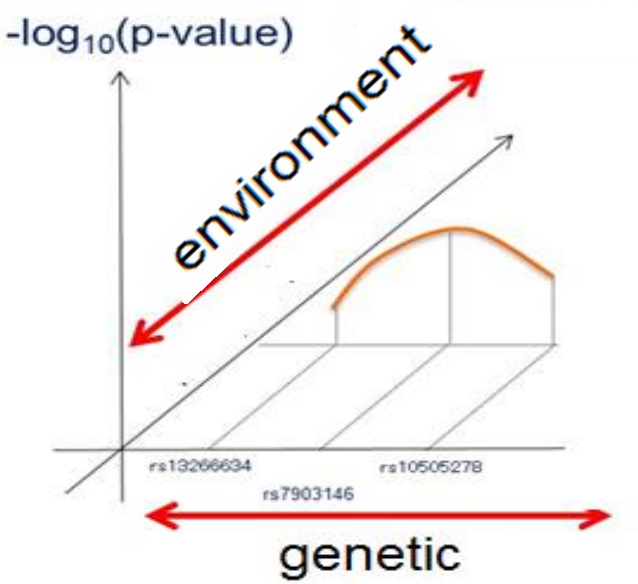
Ultimate Purpose of population-type Biobank "Medicine for Person"

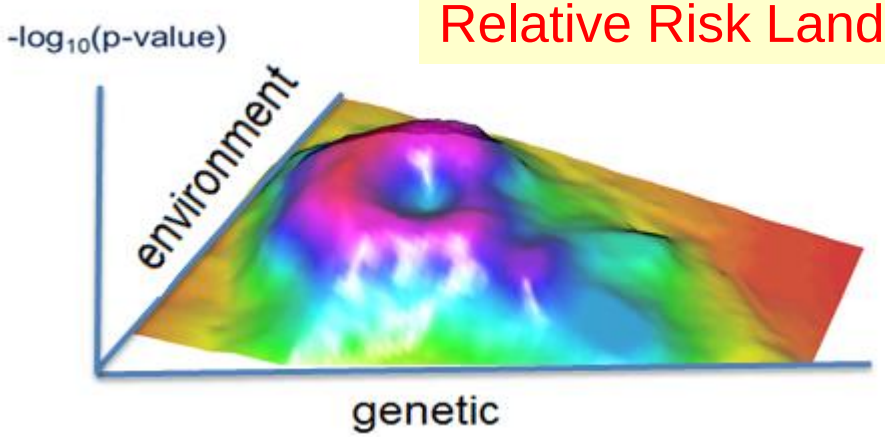
- "Medicine for Disease"
  - Has been the main concept of modern medicine.
  - Remarkable scientific progress since 19c latter (Koch-Pastour's Bacteria Pathology, Virphew Cell pathology formed the basis)
  - Concept fit for "disease biobank"
- "Medicine for Person"
  - Understand Person in **his Totality** with respect to Overall Susceptibility of Contacting Diseases through Person's Whole Life
  - Modern Medicine begins to bring in this concept
    1. throughout **total life span of his life**
       - "from uterus to grave"; DOHaD theory, life course healthcare
    2. throughout **total ecosystem he lives in**
       - Gut Microbiome as mediator between environment factor and biosystem, basis of various diseases
  - Concept much fit for " population biobank"

Thank you for your kind attention

Personalized prevention

Combination-Specific Effect of Combination of GxE factors



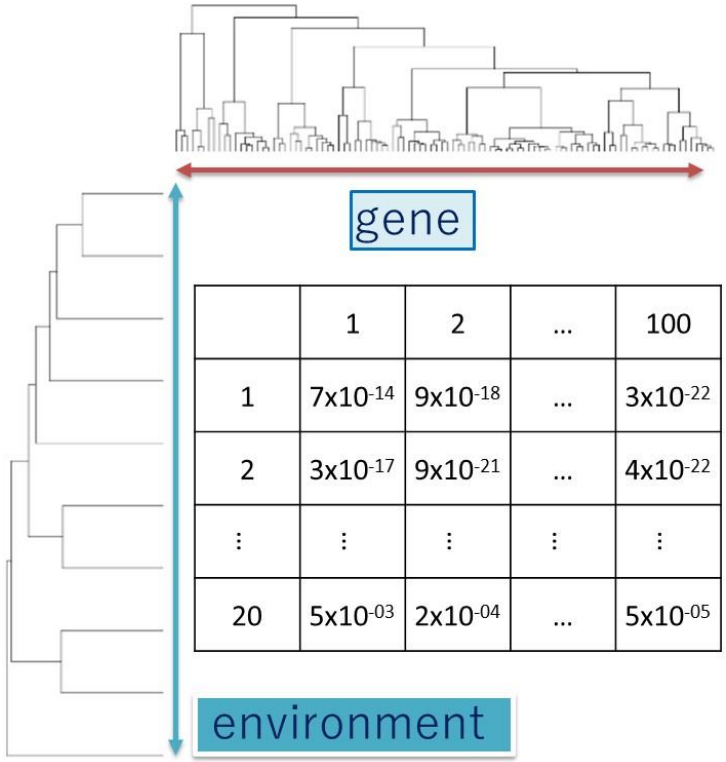


Relative Risk Landscape

Each of variables (horizontal: genes, Virtual: environment factors) are rearranged by hierarchical clustering

p value list

Environment factors	gene 1	2	…	100
1	7×10^{-14}	9×10^{-18}	…	3×10^{-22}
2	5×10^{-03}	2×10^{-04}	⋯	5×10^{-05}
⋮	⋮	⋮	⋮	⋮
20	3×10^{-17}	9×10^{-21}	…	4×10^{-22}



	1	2	…	100
1	7×10^{-14}	9×10^{-18}	…	3×10^{-22}
2	3×10^{-17}	9×10^{-21}	…	4×10^{-22}
⋮	⋮	⋮	⋮	⋮
20	5×10^{-03}	2×10^{-04}	…	5×10^{-05}

Calculation of p value of each table

Cochran-Mantel-Haenszel table

population		Disease (+)		Disease (-)	
		Env1 (+)	**Env1** (–)	**Env1** (+)	**Env1** (–)
Gene1	0 (aa)	n_{00}	n_{01}	n_{00}	n_{01}
	1 (aA)	n_{10}	n_{11}	n_{10}	n_{11}
	2 (AA)	n_{20}	n_{21}	n_{20}	n_{21}

P value for G1x E1 ➡ D

Gene set

Environment factors

p	1	2	…	100
1	$7x10^{-14}$	$9x10^{-18}$	…	$3x10^{-22}$
2	$5x10^{-03}$	$2x10^{-04}$	⋯	$5x10^{-05}$
⋮	⋮	⋮	⋮	⋮
20	$3x10^{-17}$	$9x10^{-21}$	…	$4x10^{-22}$

Gene allele X Environment = risk of Disease