

Big Data and Artificial Intelligence in Medicine and Drug Discovery

Hiroshi Tanaka

Biomedical Data Science

Tokyo Medical and Dental University

and

Tohoku Medical Megabank Organization,

Tohoku University



Coming ! of the era of
Big Data Medicine

In Next Decade
Framework (paradigm) of Medicine
Will be Totally Changed!!

Big Data?

Difficult to treat by conventional information processing method because it is too large, too many kinds and too frequently changing

So what is

Medical Big Data?

Big Data in Medicine

Rapid and Huge Accumulation of Big Data

- (1) **Precision Medicine** : Comprehensive **Genome-Omics data** brought by advance of biotechnology (e.g. NGS, Molecular Images)
- (2) **Genomic Biobank**: **Genomic and Environmental (exposomic)** data of Genomic Cohort participants
- (3) **mHealth**: Continuous physiological and behavioral data by **mobile Health (wearable sensor monitoring)**

Enormously **Cost Reduced**, nevertheless
High Quality Massive Data



Whole Genome seq : 13 yr, 3,500 M\$ (2003) →
1day, 1000\$ (2016)

How we should cope with this Medical Big Data

Tremendous Improvement of **Preciseness** of Medical Care
Groundbreaking Change of Medicine

New type of Big Data emerges

Medical **Big Data** Revolution

- **Clinical Conventional “Large scaled Data”**
 - Clinical Lab Tests, Prescriptions, Images
 - Ex. claim DB. Jp. Sentinel Project
 - **Socio-Medical epidemiological “Large scaled Data”**
 - Ordinary epidemiological data
 - life style, health exams, questionnaire
- Due to recent spread of “Digitalization”**

**Conventional
Medical
“Larger data”**



- **Big data of “Genome-Omics Medicine”**
 - Genome Omics Medicine
 - Due to Rapid Advance of **Clinical Sequencing**
 - **Molecular biomedical images**
- **Big Data of “Continuously monitoring biosignal”**
 - Life-course-oriented healthcare
 - Lifestyle, behavioral information, **mHealth**
 - Due to Rapid Advance of **Wearable Sensor**

**New type of
(Genuine)
Medical Big Data**

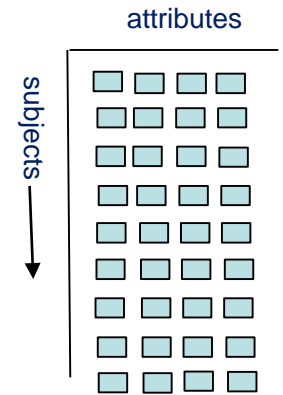
New type of Medical Big Data

Data Structure

- Conventional Medical “Big Data”

- “ n – Big Data”

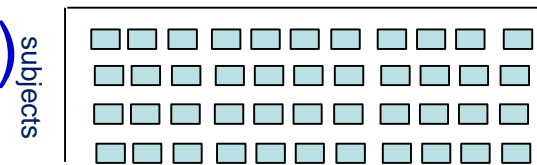
- For one subject (patient)
Num. of attributes is “Small” ($n \gg p$)
- Num. (n) of subjects (patients) is “Big”
- Conventional statistical method works well



- New type of Big Data (omics, mHealth)

- “ p - Big Data”

- Num. of attributes (p) for one subject is “Big”
- “New NP problem” ($p \gg n$)
- But Num. of subject (patients) is comparatively “Small”
- Conventional statistical method does not work well



Necessity of
New Data Science of Medicine

New type of Medical Big Data

Purpose to Collect Big Data

- Conventional Medical “Big Data”
 - **Population Medicine**
 - To **reveal** the “**collective law**” (“laws in group-level”) by collecting large number of samples
 - which can not be found by seeing each individual subject
- **New type of Big Data (genome, omics, mHealth)**
 - **Personalized (Stratified) Medicine**
 - To comprehensively **enumerate all the individualized (stratified) patterns** existing under the same name of disease; **How many individualized patterns** exists?
 - For exhaustive search, **Big number of samples** is necessary

Intention to Collect **Big Data** is Quite **Opposite**
Toward **collective** vs **individualized** pattern

Paradigm Changes

Medical Big Data Revolution Causes

- **“Population medicine”** paradigm disrupts
 - “One size fit for all” medicine is no more valid
 - Towards **“Individualized Medicine”**
 - How many **“Personalized (Stratified) Patterns” (intrinsic subtypes)** of **disease** exit under **the same name of disease**
 - How fine granularity of stratification should be?
 - **Big Data** is needed for **enumeration of these intrinsic subtypes**
- **“RCT and Evidence-based Medicine”** paradigm disrupts
 - Liberation from the “gold standard” of RCT and EBM
 - **RCT: Random (Artificial) Controlled Trials with Small-ish populations outside the Real Medical Practice**
 - These concepts are **before the discovery of “individualized medicine”** and are **no more valid**
 - **Randomization can not eliminate the difference of intrinsic subtypes** of disease unlike conventional confounding factors
 - **Towards Learning from “Real World Data”** (Disease registry, EHR big data) for clinical evaluation of drugs, devices, etc.

Big Data in Genome-Omics Medicine



Two Streams of Genome-Omics Medicine

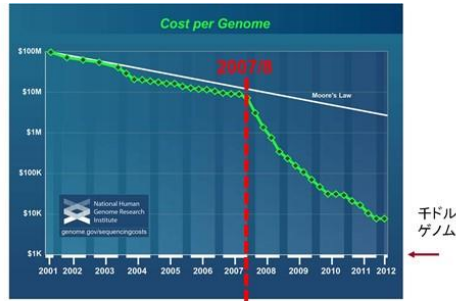
Genome Medicine in United States: **Precision Medicine**

- Surging Wave of **Rapid Clinical Implementation of Genomic medicine** (2010) shortly after “**Sequence Revolution** (2007)”
- Aiming at **dramatic improvement in therapeutic medicine** for **individual patient** by genome information
 - **POC** (Point of care) ID of **causative gene** for rare disease
 - **POC** (point of care) ID of **driver gene mutation** for cancer
 - **Preemptive PGx**: polymorphism of **drug metabolizing enzyme**

Genome Medicine in Europe: **Genomic Biobank**

- Recognition of the Value of “**Collective Genome Information**” (island) to the **Spread of Genomic Biobank** today
- Aiming at **dramatic improvement in preventive medicine** for the **general public** (a nation) by genome information: based on the concept of “welfare state”
 - **Prospective Population-based Large Genomic Cohort**
 - Prediction of **Occurrence of “Multifactorial Disease”**
 - Estimate the **interaction of genomic predisposition and environmental factors**

Genome Medicine of United States



DNA Sequencing Cost: the National Human Genome Research Institute

Sequence Revolution 2007/8

2005~ NGS 454 (LS,Roche)
2007/8~454, Solexa (Illumina),
SOLID (LT,TF)
**Sequence Revolution
Faster than Moore's law**



Illumina 2500

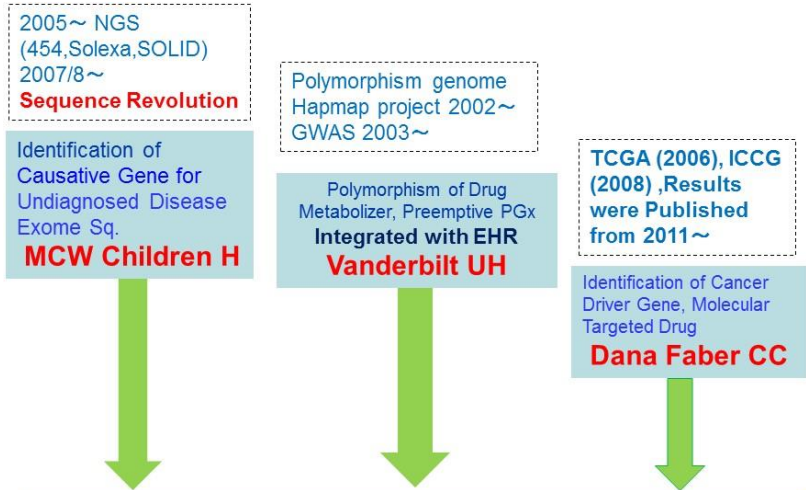
Ion Torrent

President Obama Precision Medicine Initiative



2015.1 State of the Union Address

1st term
Early adopters



**Genome Omics Medicine
Clinical Implementation**

2nd term
National project

**National project
BD2K, many consortium, WG**

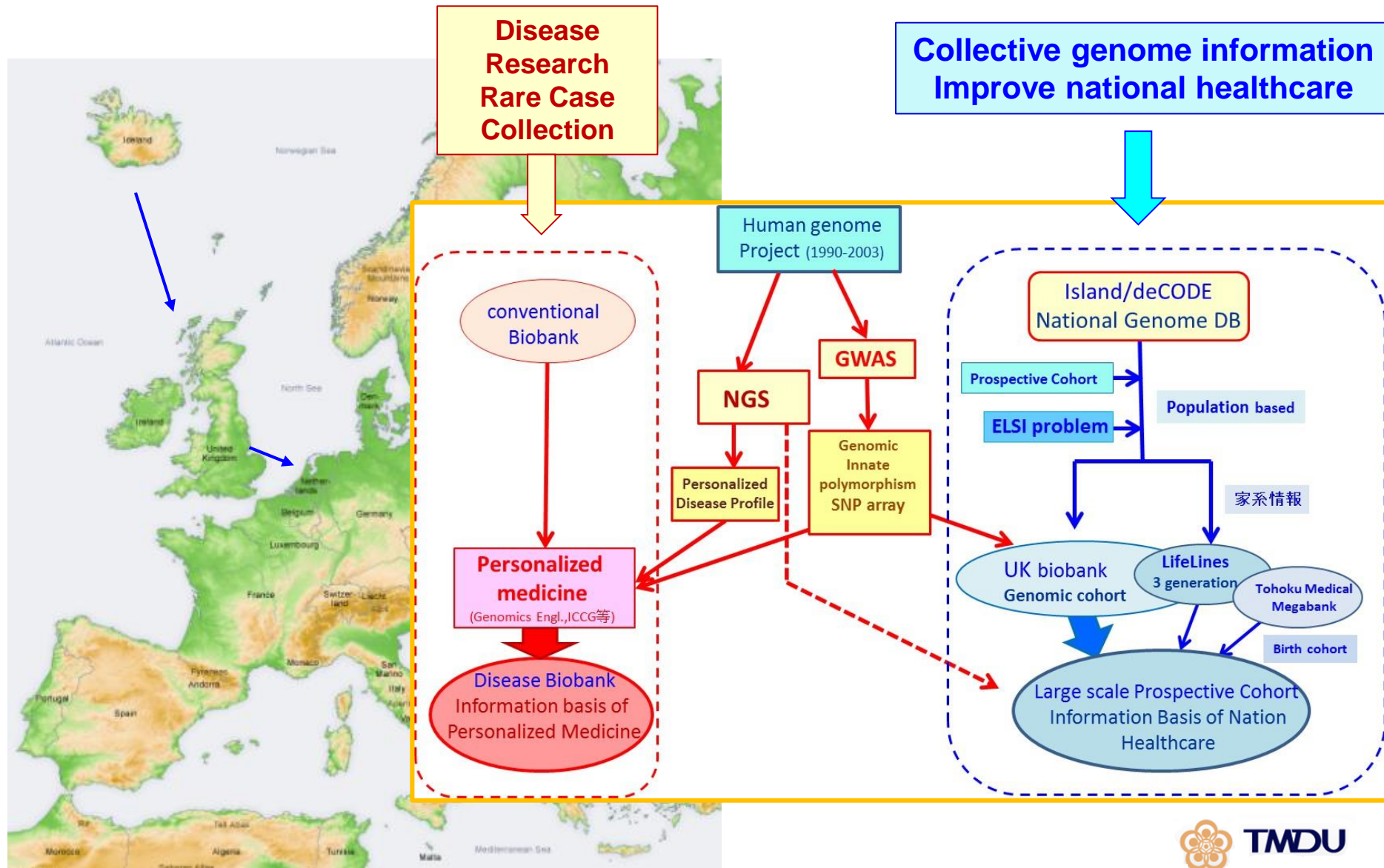
Precision Medicine Initiative
President Obama, State of Union Address

3rd term
Spread of precision medicine

Prevail of Precision Medicine
1M cohort "All of Us"
National Cancer Moonshot Project

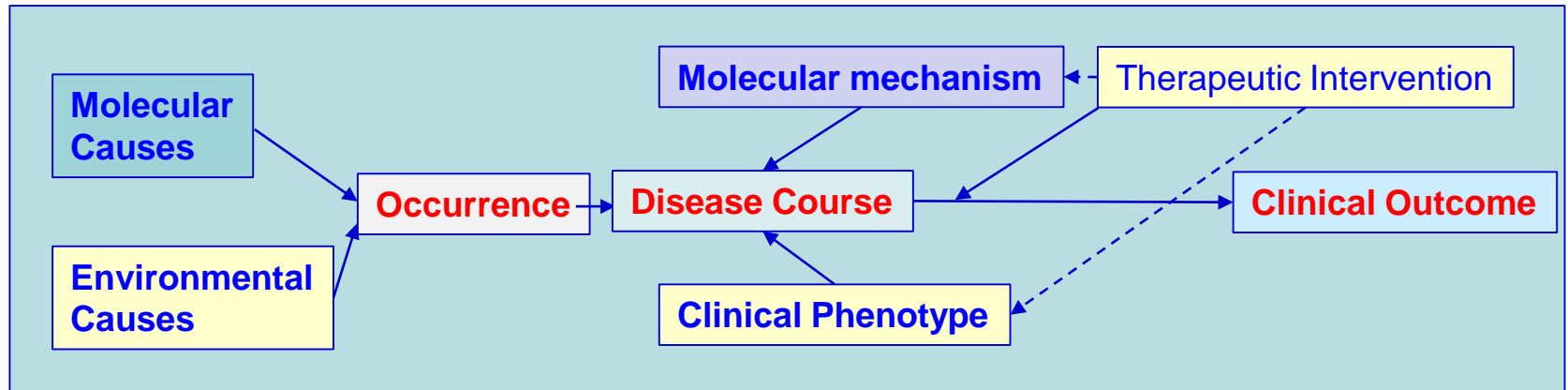
2007
2009
2010
2011
2012
2013
2014
2015
2016
2017

Genome Medicine of Europe



Challenge of Big Data Medicine I

Disambiguation of corresponding “**non-genomic**” information



Ontology of disease course

Molecular Mechanism of Disease Occurrence and Progression

Disambiguation of corresponding **exposomic and phenotypic** information

Environmental factors **Clinical Phenotype** Therapeutic Intervention



eMERGE and PheKB

phase I (2007-2011)

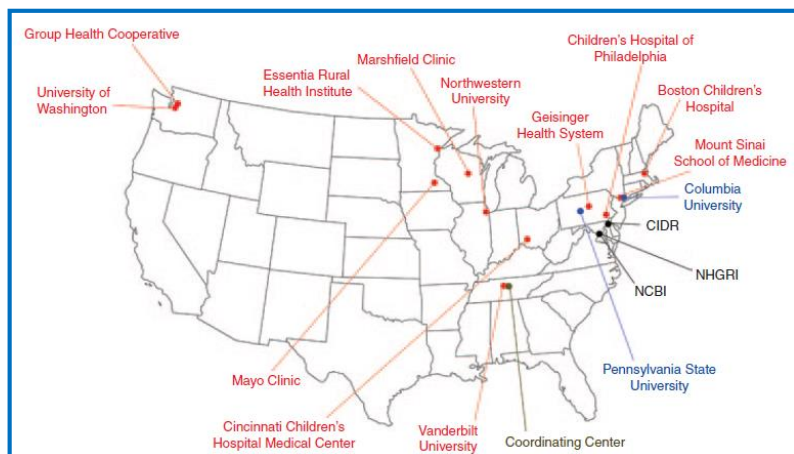
- **Phenotyping from EMR**
 - genomic discovery and genomic medicine implementation research.
- **EMR-based GWAS**
 - Each with its own **biorepository** (DNA etc) linked to phenotypic data contained within **EMRs**
- **eMERGE-I: 5 Institutes, PheKB**

phase II (2011-2015)

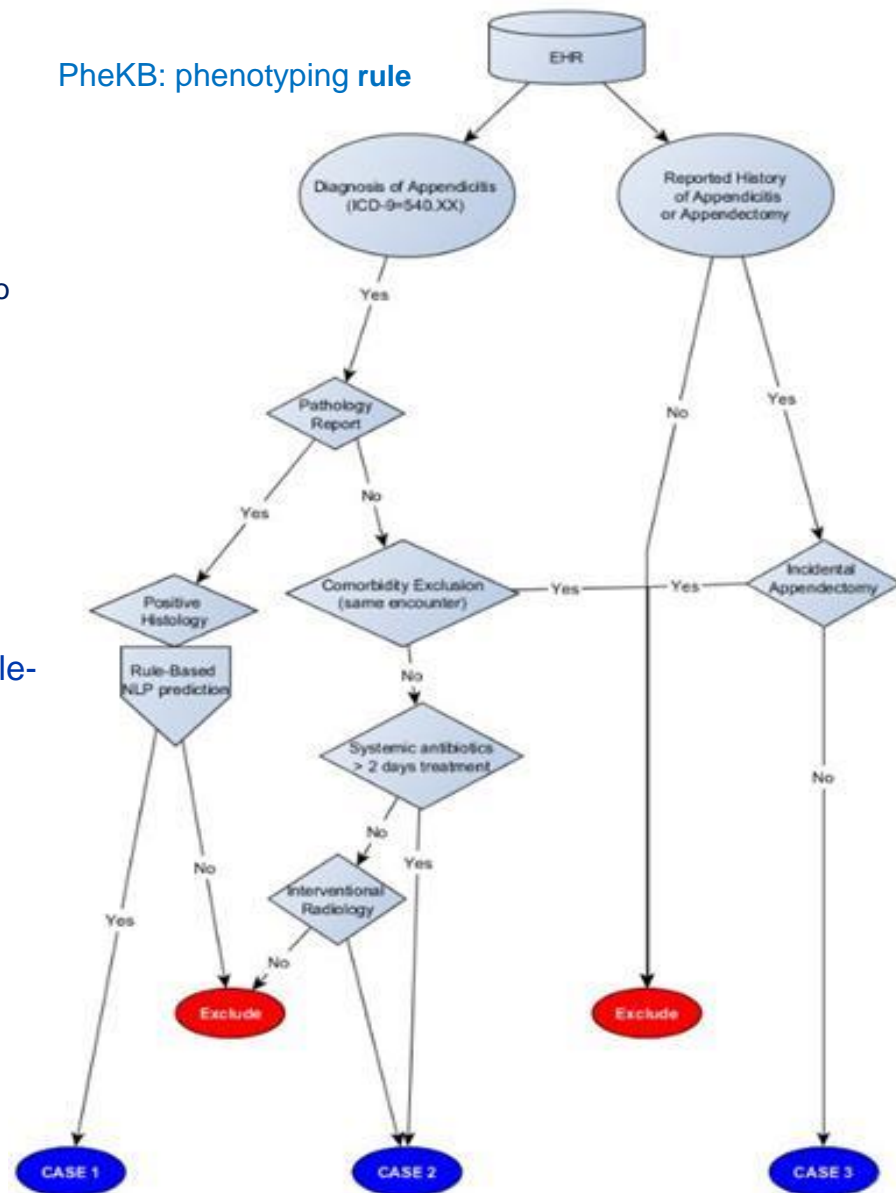
- **Integration of Genomic Information into EMR** (Clinical Implementation)
- PGx implementation in EMR
- Return of (Genomic) Result (RoR)

Phase III (2015~2019)

- explore the potential of whole-genome and whole-exome



PheKB: phenotyping rule



Challenge of Big Data Medicine II

Contraction Methodology to extract the Intrinsic Information Structure

Medical Big Data

➔ **Hyper Multi-dimensional Correlation Network of Data**

Clinical Genome Medicine

many to many relationship

Comprehensive Molecular Information

Genome, multi-omics

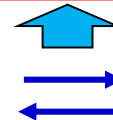


Clinical Phenome

clinical signs, lab test, medical image

Genomic Biobank

Disease Occurrence



Genetic Disposition/Molecular Mechanism

SNV, disease network dysregulation

Exposomic Factors

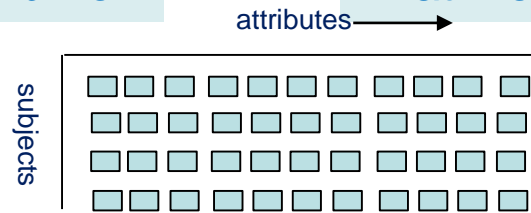
life style, environmental factors

Data Principle of Big Data

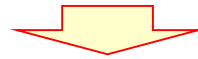
Challenge: num. attributes (p) \gg num. subjects (n)

p : may be billion

n : at most, tens thousand



If this huge number of attributes are independent, we can not do anything



Big Data · Sparse Assumption

Big data is intrinsically determined by the latent variables, number of which is less than number of subjects

principle of compositionality

Big data is hierarchically composed of nested structure

Multi-dimensional medical big data should be contrasted to intrinsic structure

Distinguishably Effective Method

AI, Deep Learning

- Limitation of Machine Learning

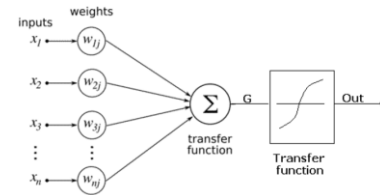
- “Supervised learning”

- Construct AI by providing the feature and answer

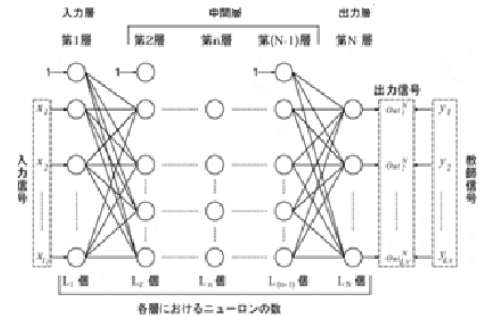
- Deep Learning revolution

- “Unsupervised learning”

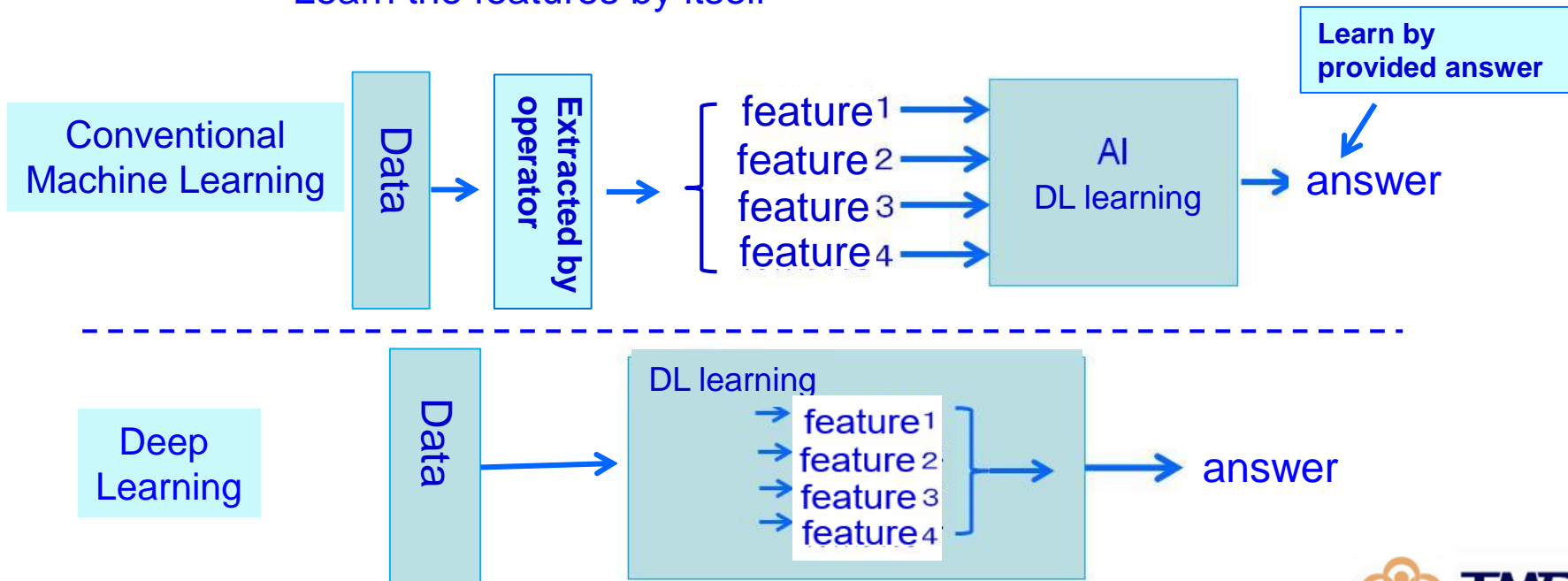
- Learn the features by itself



Neural information element

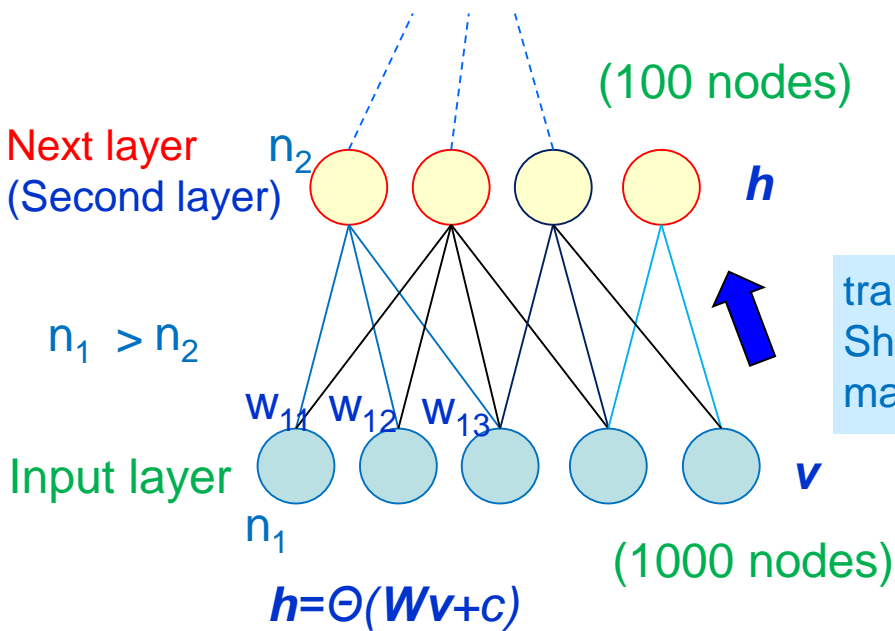


Multiple Layer neuro-network



Revolutionary point of DL Autoencoder

- Principle of **autoencoder**: Learn specific **intrinsic features** of the big data
- Restore the **node values of input layer** from the **node values of next layer** where the number of nodes is decreasing compared with **input layer**.
 - **Intrinsic features** should be **explored** so that the input layer to be **recovered** as same as possible
 - discover **intrinsic features**

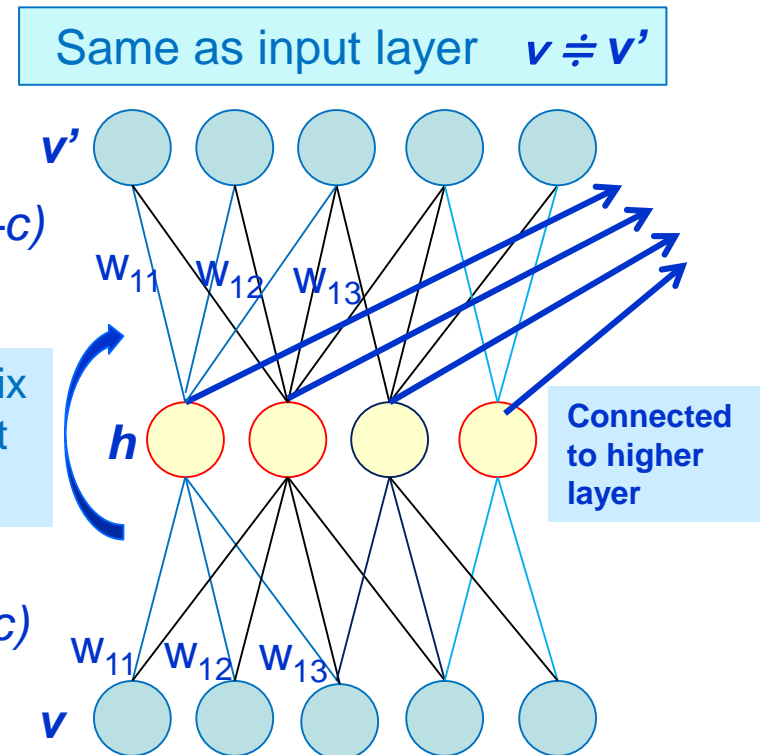


$$v' = \Theta(W^t h + c)$$

$$W' = W^t$$

transposed matrix
Share the weight matrix

$$h = \Theta(Wv + c)$$

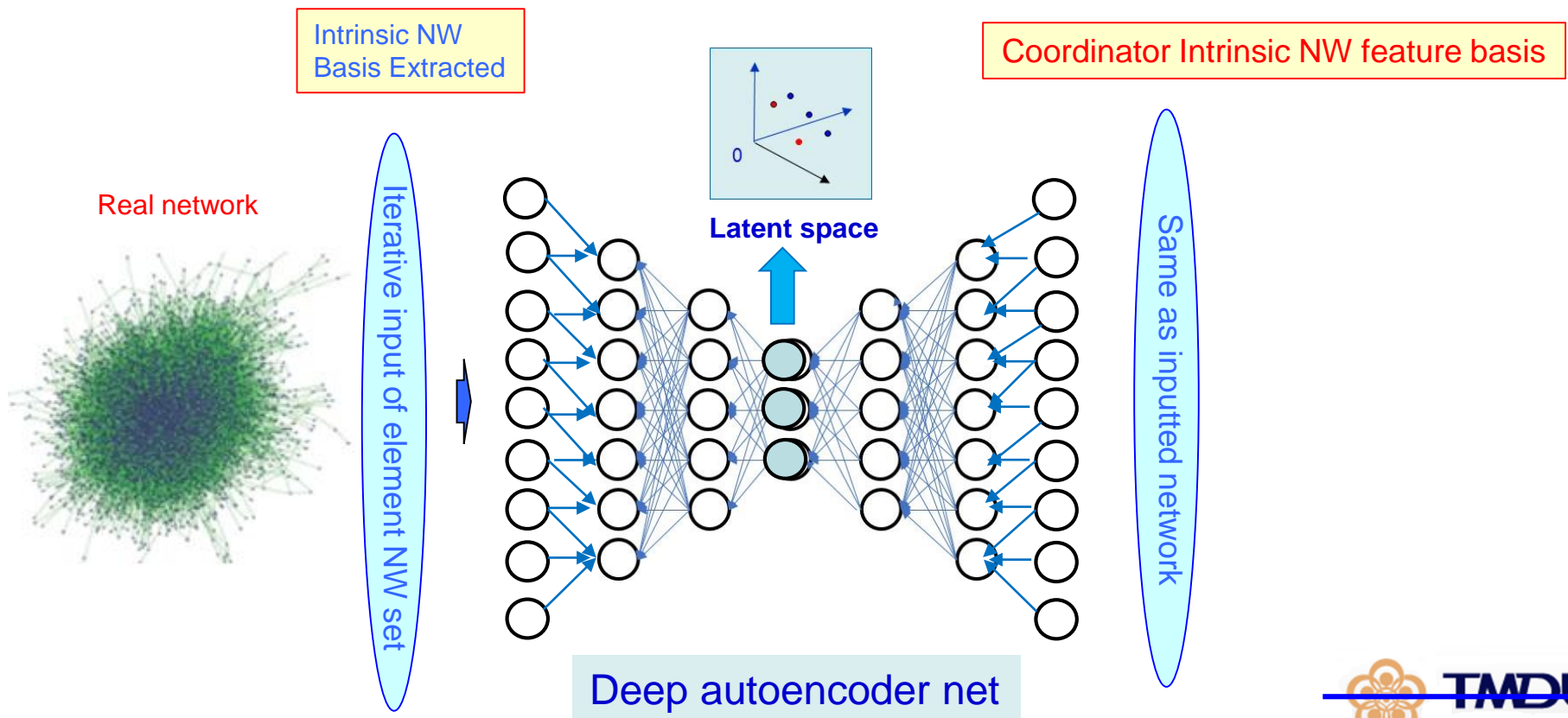


Deep Autoencoder Network

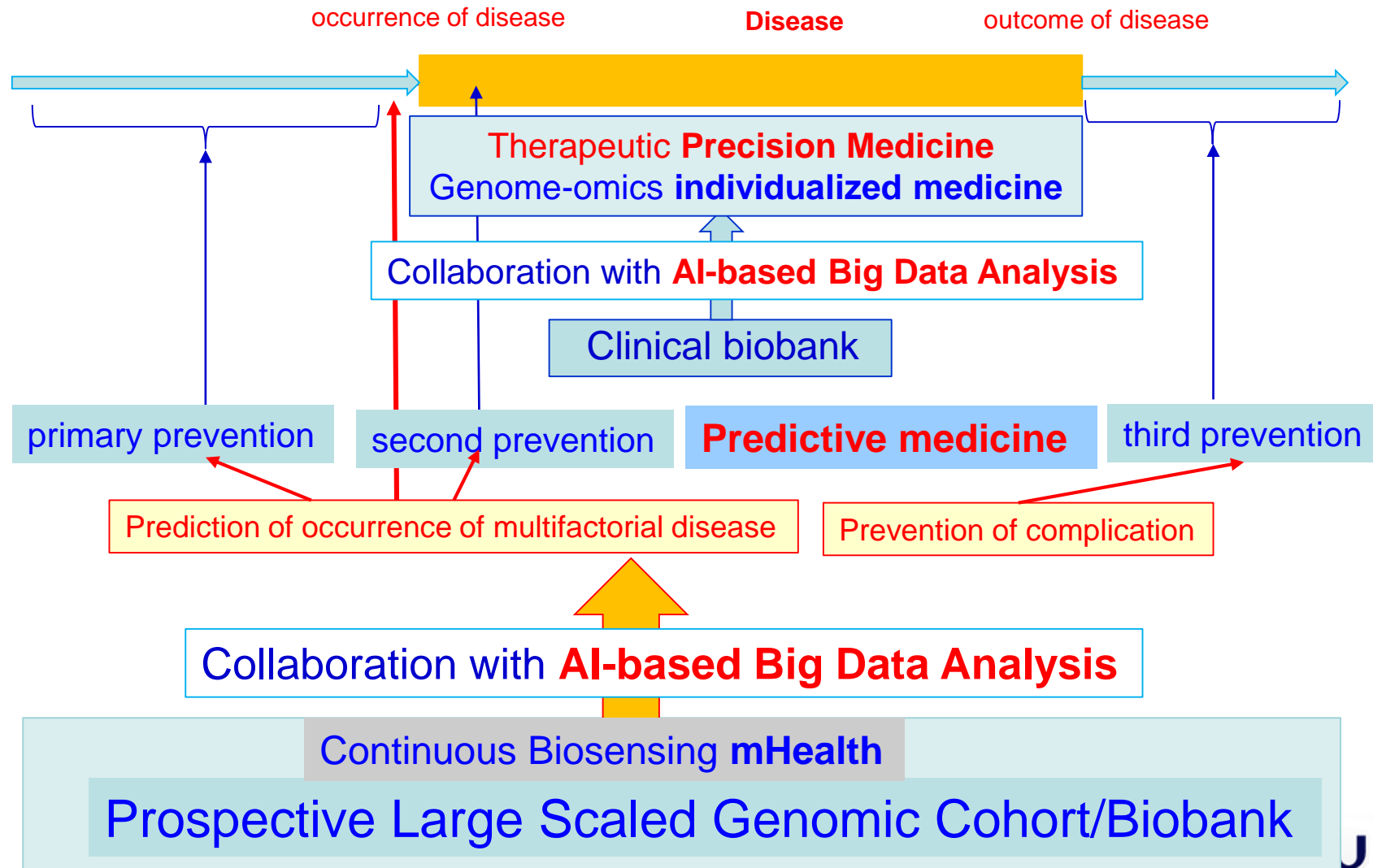
- Deep Learning-based Correlation Network Contraction

Multi-dimensional correlation network information structure
⇒ Contract to be composed of a few network variables

- Projection of data to be composed of intrinsic bases by nonlinear contraction. Contraction to “latent space”



Integration of Big Data Medicine into life-course oriented healthcare



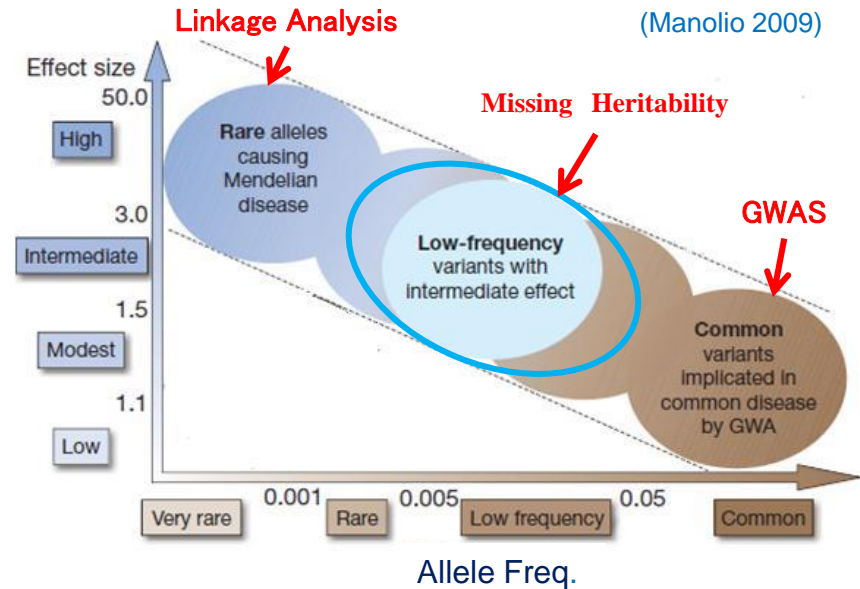
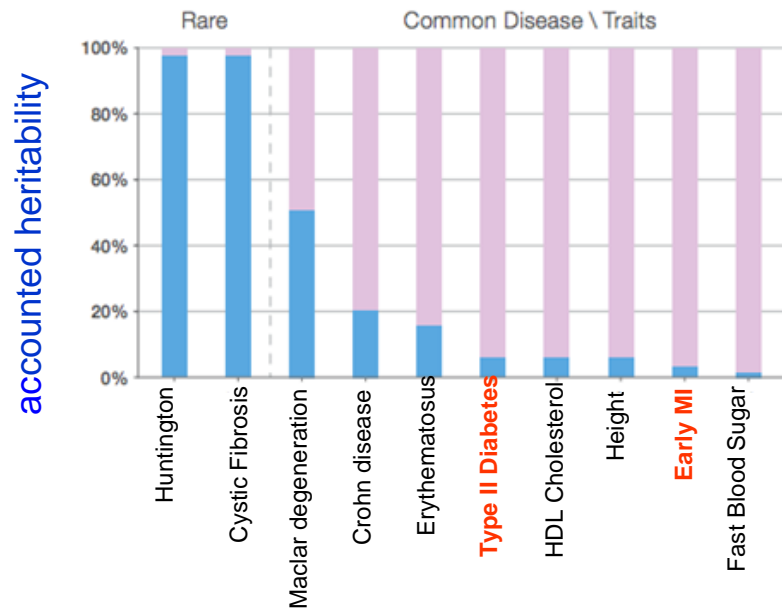
Future Big Data Medicine

- Genome Medicine, Genomic Biobank and mHealth are integrated in
- **Life-course oriented healthcare**
 - Understand Individual in **his Totality** with respect to **Overall Susceptibility of Contacting Diseases** through Person's Whole Life
 - 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

Toward Understanding of Multifactorial Disease

-- Interaction between Genomic and
Exposomic Factors--

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)

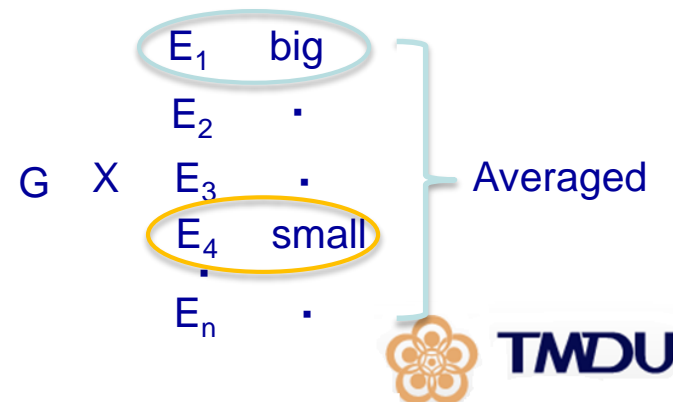
Interaction between gene and environment

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

		CYP1A2 Phenotype \leq 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

relative risk of disease occurrence is

Combination - specific

L. Le Marchand, JH. Hankin, LR. Wilkens, et al Combined 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

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

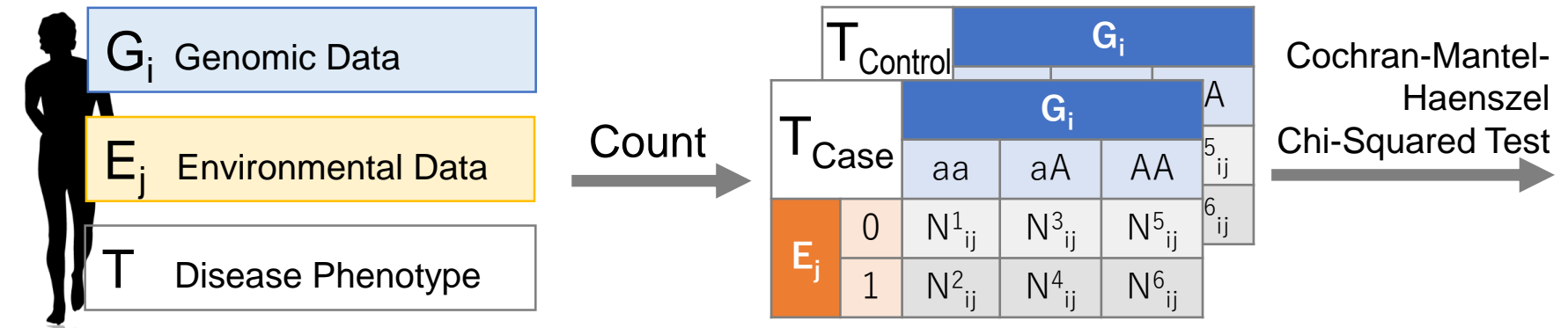
<one of genetic variates> <one of exposomic factors >
G_i x **E_j**

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

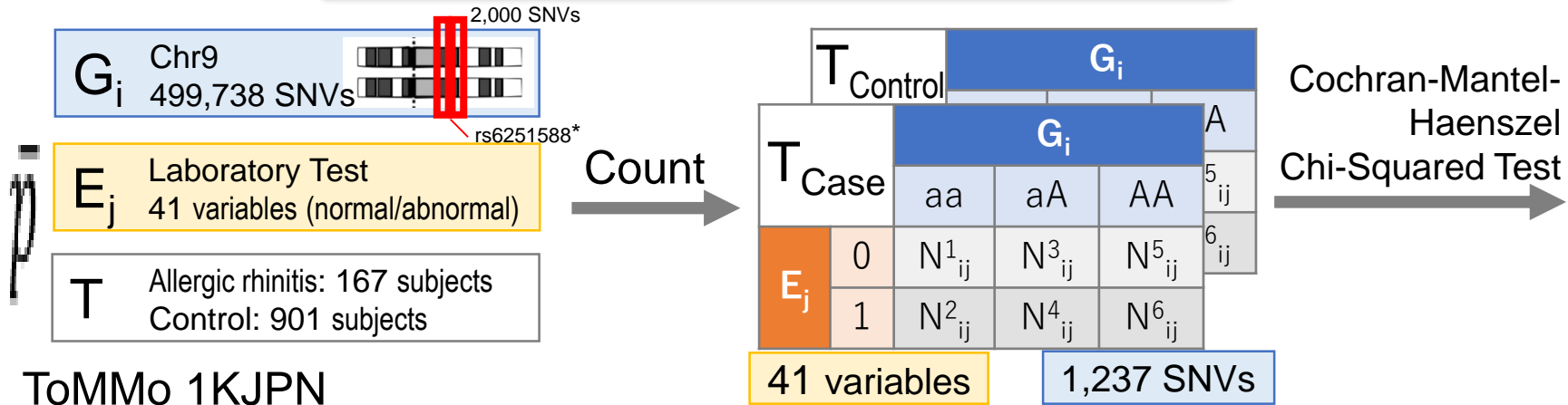
This is the **first step** of **G x E risk estimation method**. We will proceed to deal with **more plural G x E factors**



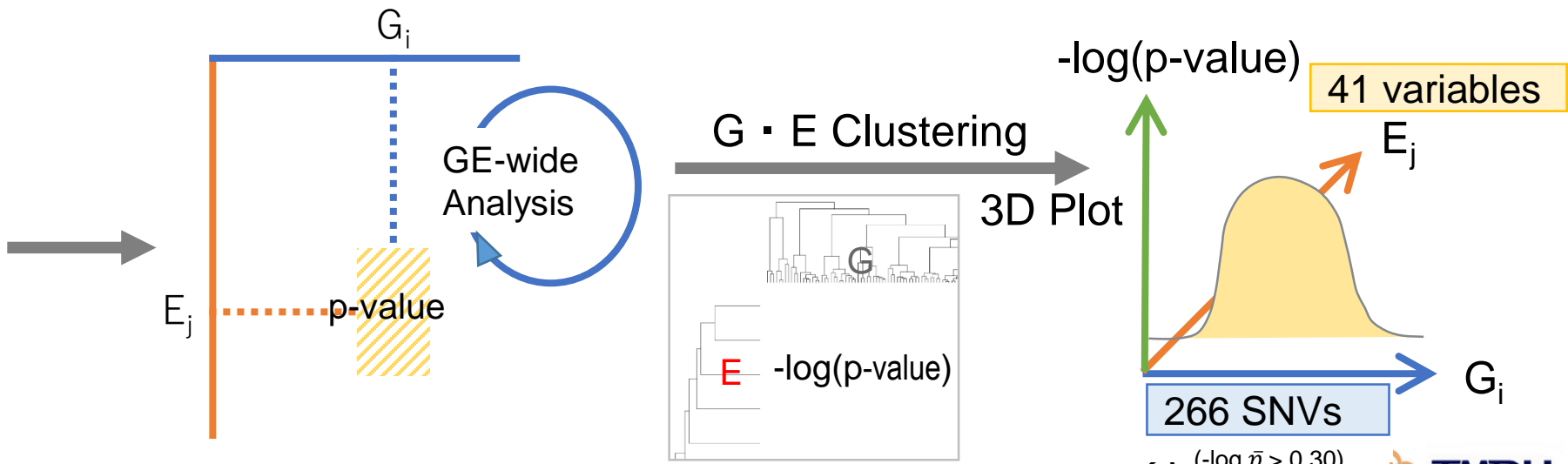
Our Risk Analysis Flow



GE-WAS Analysis Flow



ToMMo 1KJPN



*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”

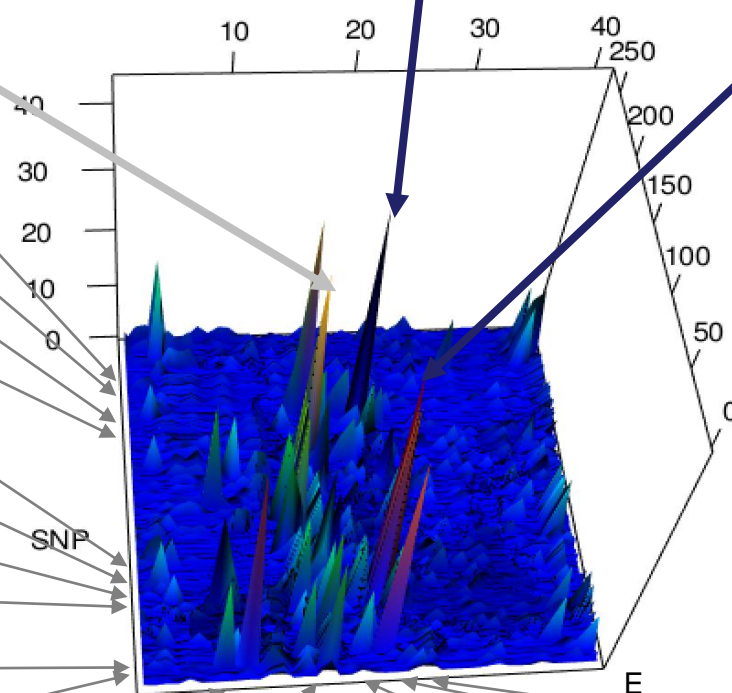
-log p=26.7
 G: rs2381416
 E: Candida albicans

Hay Fever
 (Schröder
 2016)

-log p=35.7
 G: rs549716210
 E: Dog dander

-log p=43.4
 G: rs549868264
 E: Milk

- rs186758850
- rs192922406
- rs549716210
- rs2381416
- rs549868264
- rs146693626
- rs146898930
- rs568525828
- rs191520260
- rs76741691



Grey alder Common ragweed Candida Dog dander Milk
 ハンノキ ブタクサ イヌ皮屑(フケ)

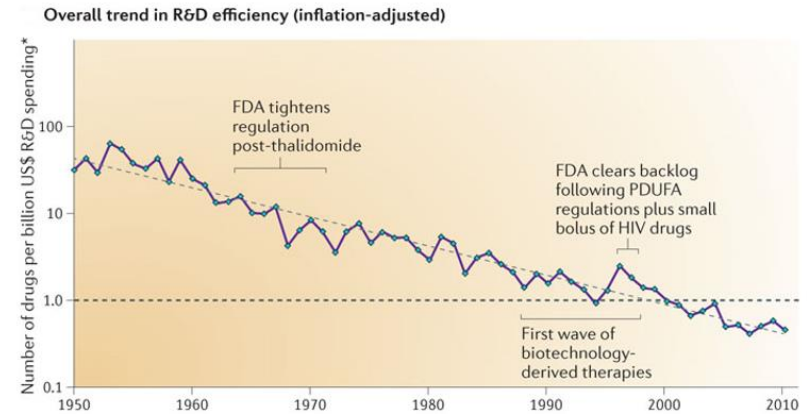
Discussion

- 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.
- We are applying Deep Learning to all the combination of .SNPs and exposomic factors to obtain the essential relations albeit the vast amount of number of combination.

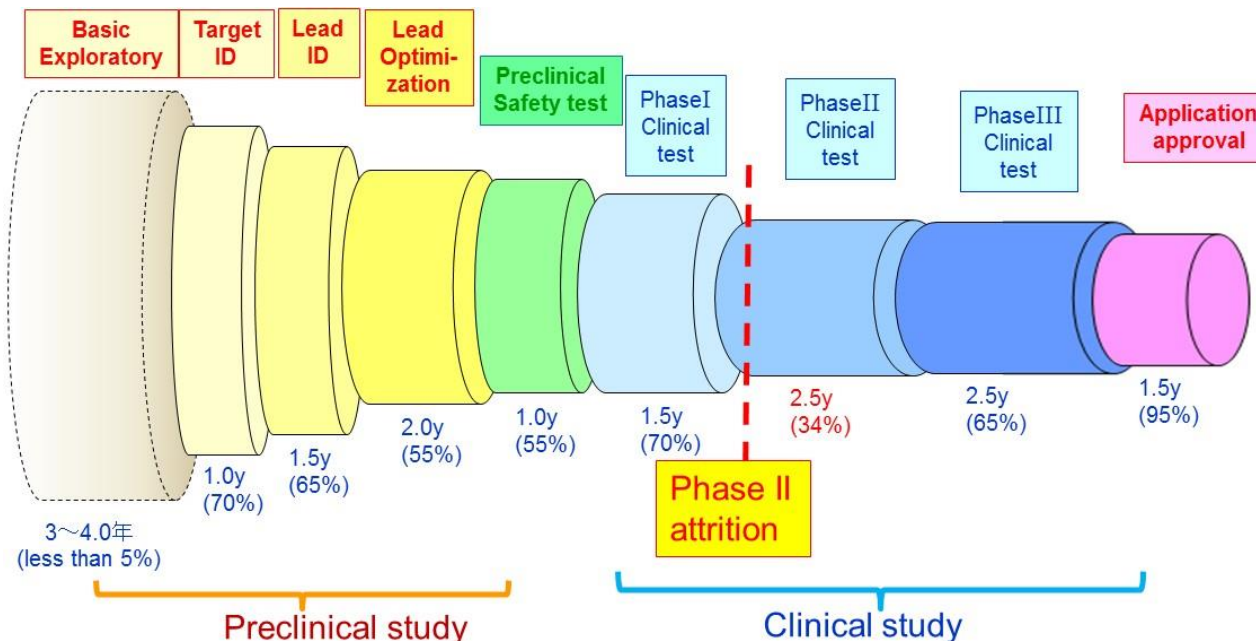
AI-based Drug Discovery

Current Situation of Drug Discovery

- Rapid increase of R&D expenditure
 - More than 1B \$ for one marketed drug
- Decrease of success rate
 - now about 1/20,000~1/30,000
 - Remarkable Drop Between non-clinical and clinical test (**phase II attrition**)
- **Clinical Predictability**
 - At as early as possible stage,
Estimation of clinical efficacy and toxicity
- **Efficient measures**
 - Use Disease-specific iPS cell
 - Use of **Human Bio Big Data** in early stage

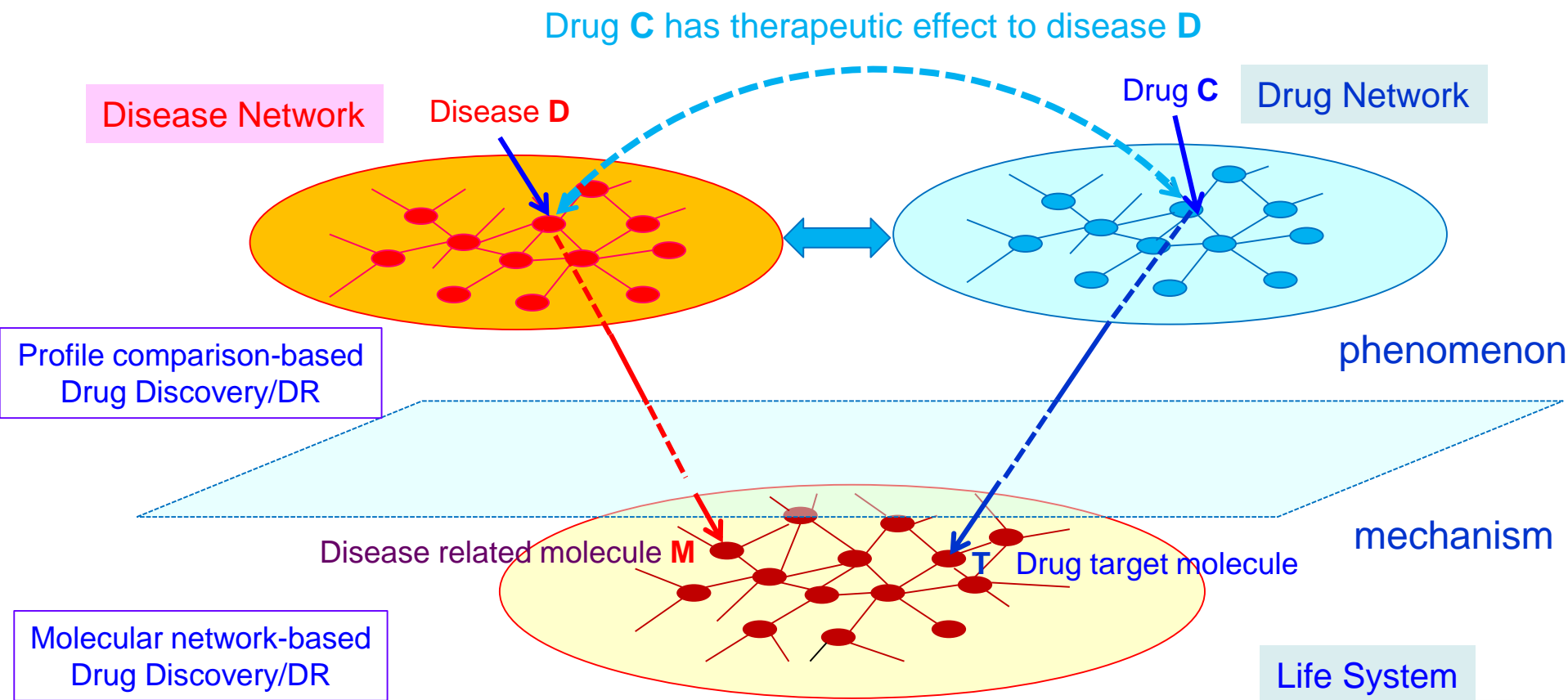


Nature Reviews Drug Discovery (2012)



Basic structure of profile-based computational drug discovery

Framework of Triple-layer disease and drug network

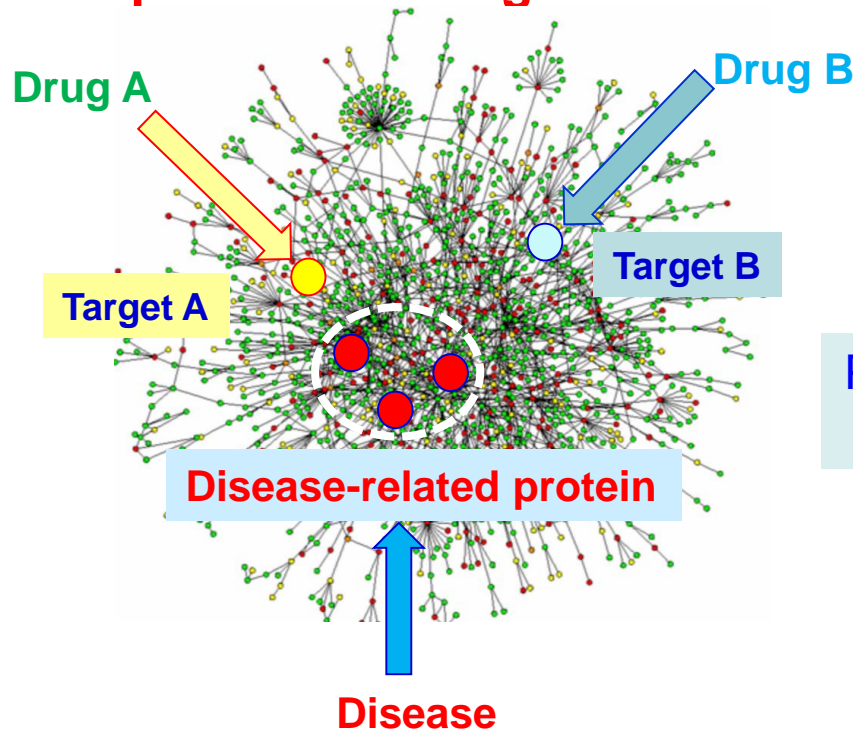


DR: Drug Repositioning: is the application of known drugs (compounds) to treat new indications (i.e., new diseases)

Common Platform of Drug Discovery/DR

Protein-Protein interaction network (PPIN)

- **Common Platform bionetwork**: mediating disease and drug action
- **Protein-protein interaction network (PPIN)** as common platform
- **Disease**: Scaffolding in PPIN: **Disease-related protein** (gene)
- **Drug** : Scaffolding in PPIN: **Drug Target protein**
- Based on **the distance (proximity)** between **Disease-related protein** and **target protein**,
the impact of the drug is measured

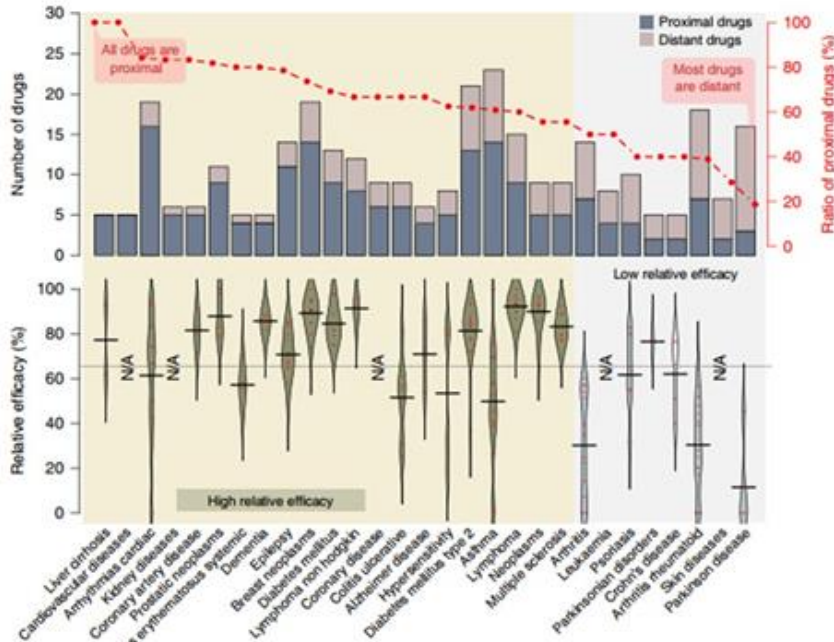
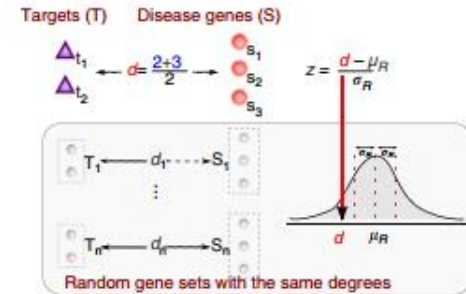
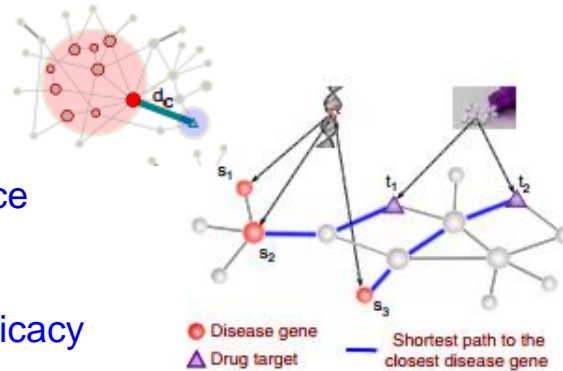


Protein-protein Interaction
Network (PPIN)

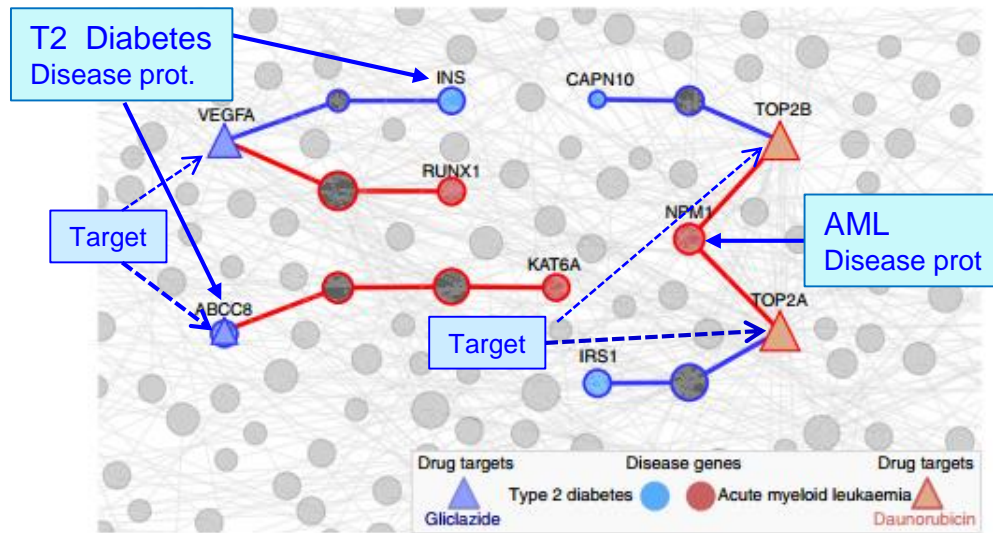
Proximity between Drug and Disease at PPIN

Relative Proximity Index d_c :

- Distance between Target and the nearest protein among disease-related protein module
- The distance is normalized among the distance of the molecules in same context
 $z < -0.15 \Rightarrow$ proximate
- closest measure d_c : best index to measure efficacy



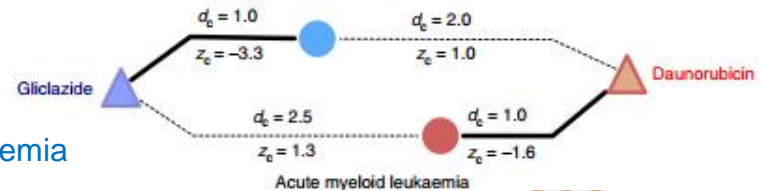
(Guney, Barabasi, 2016, Nat. Com)



Drug - disease proximity

Type 2 diabetes

Average distance: about 2 rinks



AML: acute myeloid leukemia

Need for Learning

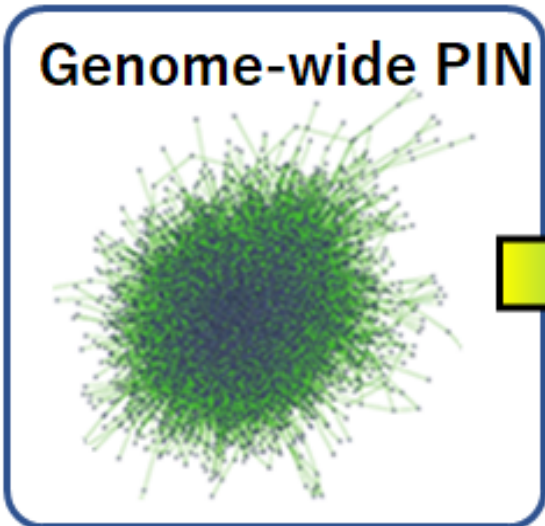
- We are **still missing in understanding** of the necessary conditions for molecule to be effective to disease
- We should find these conditions by **learning from the succeeded <disease-drug-target molecule> combinations**
- **Artificial Intelligence (AI)**, specially **Deep Learning** is now the most powerful method

Our Approach

- **By using deep learning and genome-wide protein interaction network,**
- **We build a computational framework to predict potential Drug Target genes and**
- **Repositionable drugs for Alzheimer's disease.**

Our computational workflow

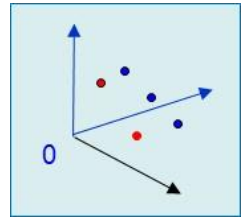
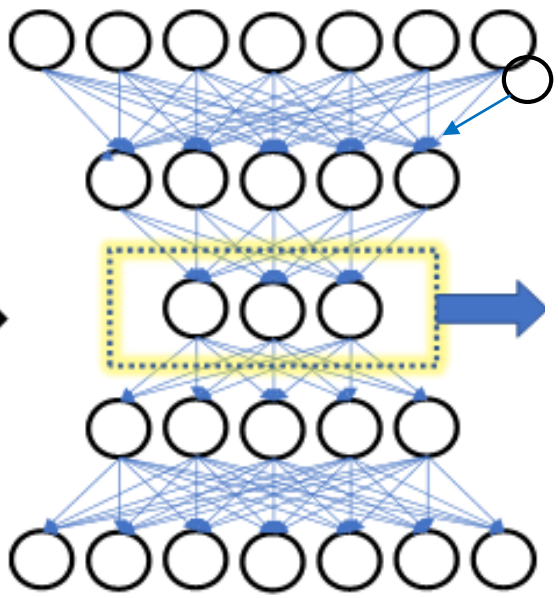
Step 1: Input data



Step 2: Feature Engineering

Feature engineering by “**deep autoencoder**” and a state-of-the-art feature selection algorithm

Dimensional reduction by “**deep autoencoder**”



Latent space

Restoration Accuracy between Deep Learning and SVD (singular value decomposition)

For a certain protein, the connections are described by adjacency vector;

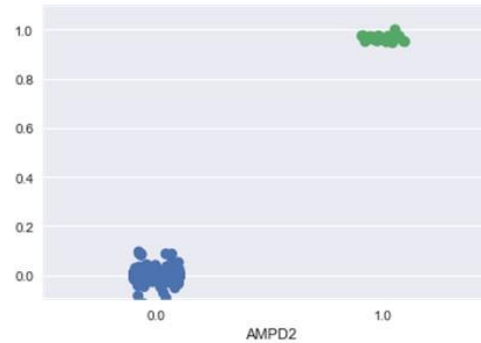
$(0,0,0,1,0,1,0,\dots)$,

where

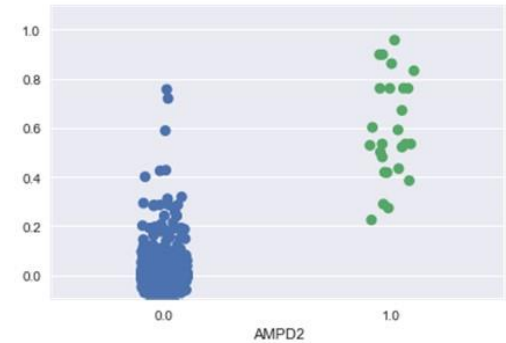
$0_{(i)}$: not connected to i th node

$1_{(i)}$: connected to i th node

AMPD2 (adenosine monophosphate deaminase 2)
degree=26

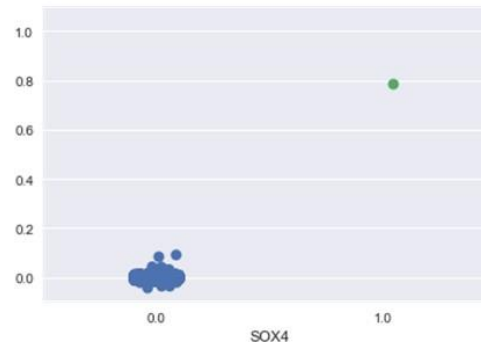


Autoencoder

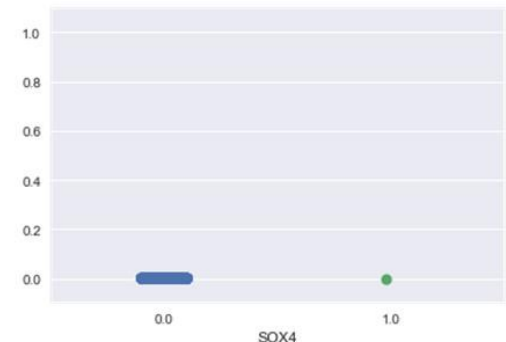


SVD

SOX4 (SRY-box 4)
degree=1



Autoencoder



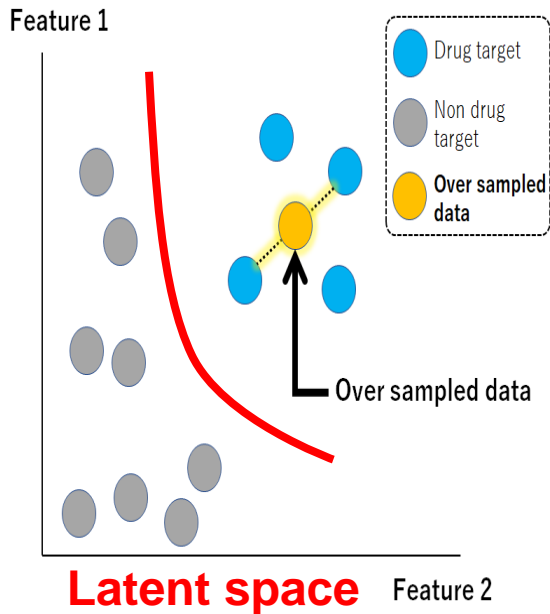
SVD

N=8,502

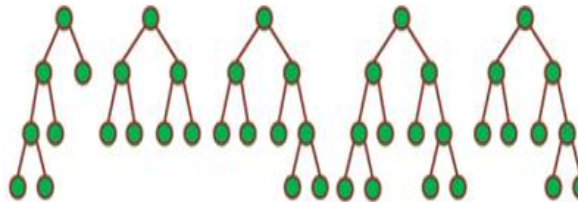
Step 3: Classifier model

A binary classifier model to target prioritization by **state-of-the-art machine learning algorithms**

SMOTE algorithm to build a training data



Xgboost algorithm to build a binary classifier



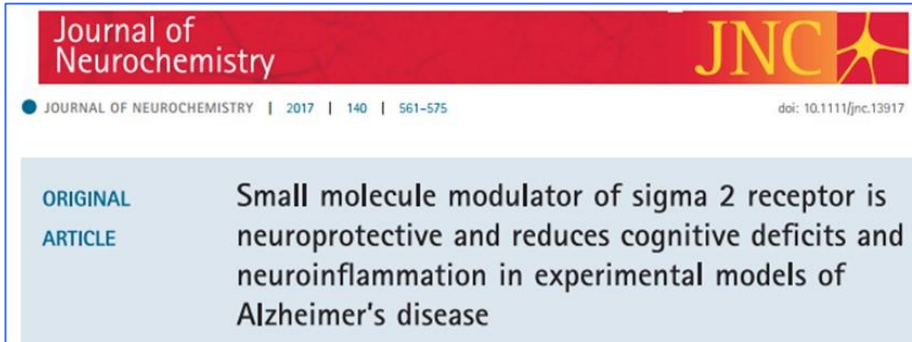
Step 4: Target prioritization

Scores for potential targets

Gene	Score (mean probability)
GRASP	0.982971499
PGRMC1	0.98234516
GPM6A	0.98234516
NRP2	0.975193546
PFKM	0.972127568
DLGAP2	0.953659343
CD81	0.941095327
IQGAP1	0.926867425
TROVE2	0.916886333
TOP3B	0.915745595
TJP1	0.914564961
PDGFB	0.914082375
SETD2	0.905462331
CFLAR	0.900456515
PROS1	0.883435477
SIT1	0.879989294
SIGLEC7	0.879989294
SHC2	0.879989294

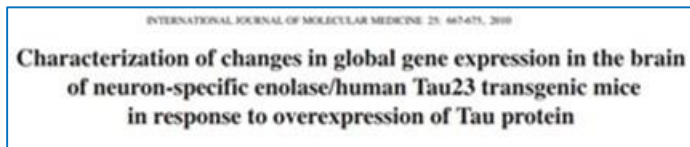
Correspondent with Wet research

PGCM1 : progesterone receptor membrane 1

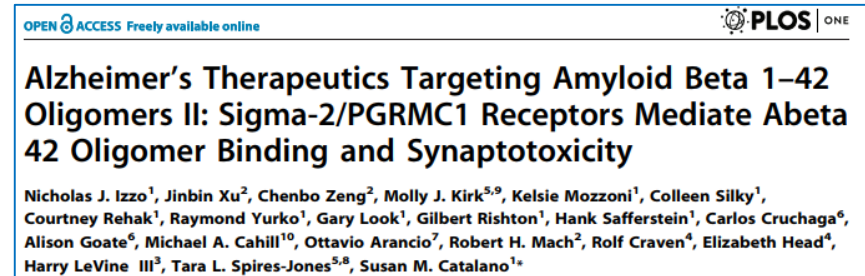
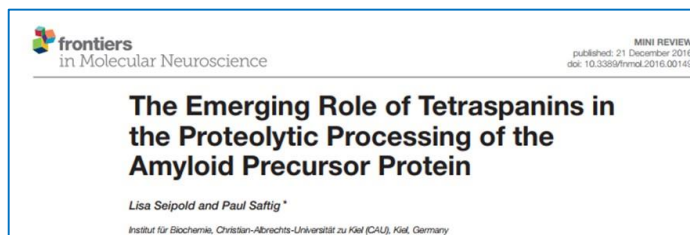


神経保護の効果 (neuroprotective)認知不全・炎症に治療効果

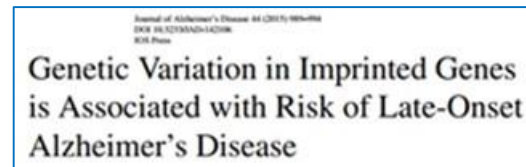
GPM6A : Glycoprotein M6A



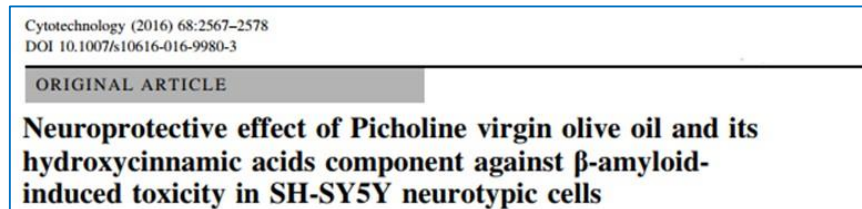
CD81:Tetraspanins family



DLGAP2 : DLG-Associated Protein 2



PFKM: Phosphofruktokinase



GRASP	PIK3C2B	PKIA
PGRMC1	NEU3	PFKP
GPM6A	SLC25A38	PAN2
NRP2	TNFSF12	GLUD1
PFKM	ADRA1B	DNM3
DLGAP2	DPM2	ITGA5
CD81	NLRP12	RILPL2
IQGAP1	NLRC4	MAEA
TROVE2	UIMC1	NCDN
TOP3B	IL8	DGCR14
TJP1	VAV1	PACSIN3
PDGFB	ARHGEF1	CD46
SETD2	WISP2	NIT1
CFLAR	PRKCE	ICAM4
PROS1	TBXA2R	GNA13
SIT1	TSPAN4	STK40
SIGLEC7	EPHB4	ROGDI
SHC2	LOC63920	CDH10
SH2D1A	PSEN1	WSB2
	SPOCK3	PHPT1
	TSP0	
	SLC4A1	

By using the **AI-based method**, we successfully predict potential **drug targets** (more than 100 genes) for Alzheimer's disease.

Example,

SLC25A38 (APPOPTOSIN)

SLC25A38 increases in the brain from Alzheimer's disease patients as well as from infarct patients. Further, SLC25A38 downregulation is likely to inhibit apoptosis induced by Bax/BH3l and neuronal death induced by A β /glutamate.

[← Previous](#)

[Next →](#)

Featured Article | Articles, Cellular/Molecular

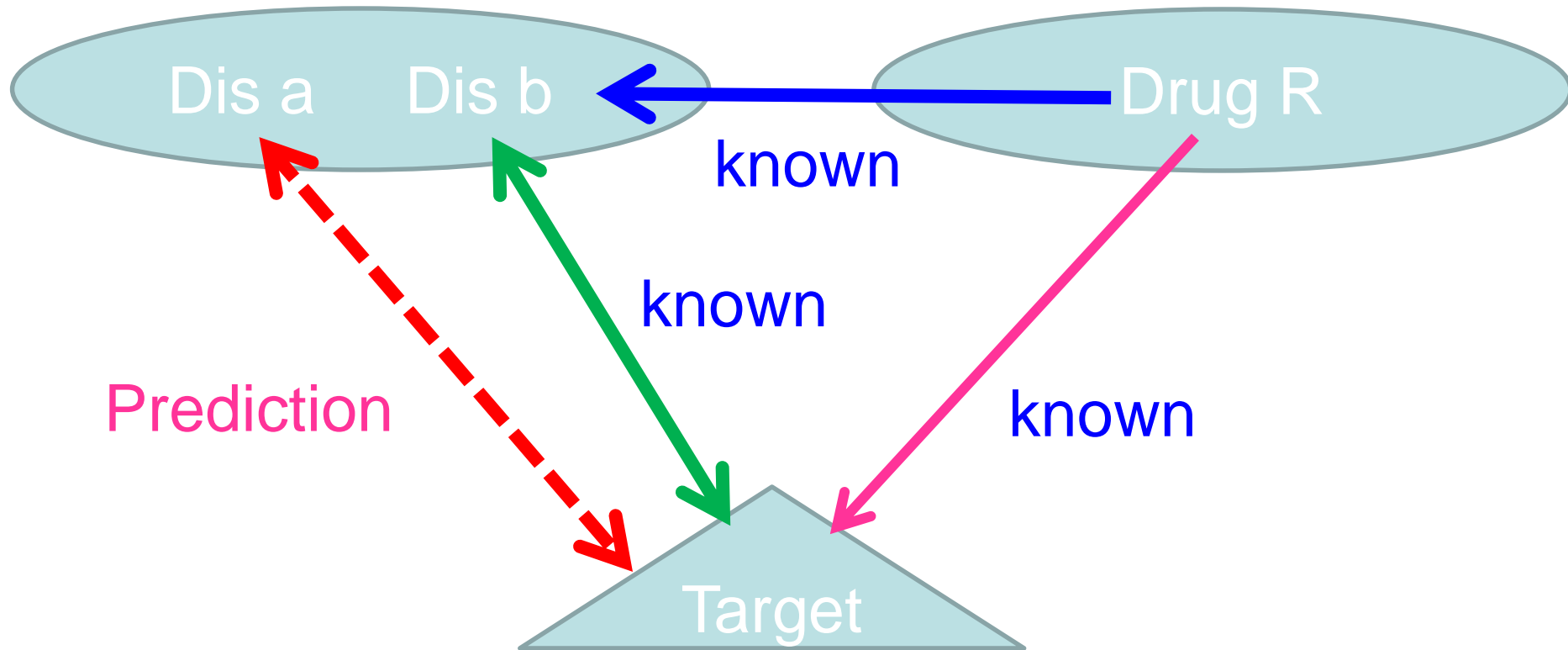
Appoptosin is a Novel Pro-Apoptotic Protein and Mediates Cell Death in Neurodegeneration

Han Zhang, Yun-wu Zhang, Yaomin Chen, Xiumei Huang, Fangfang Zhou, Weiwei Wang, Bo Xian, Xian Zhang, Eliezer Masliah, Quan Chen, Jing-Dong J. Han, Guojun Bu, John C. Reed, Francesca-Fang Liao, Ye-Guang Chen, and Huaxi Xu

Journal of Neuroscience 31 October 2012, 32 (44) 15565-15576; DOI: <https://doi.org/10.1523/JNEUROSCI.3668-12.2012>



If predicted target for disease A is known drug-target of drug R for disease B, the drug R may be repositionable drug for disease A.



Potential (predicted) repositionable drugs for Alzheimer's disease

repositionable drug	target	# of target	category
Tamoxifen	PRKCB PRKCE PRKCG ESRRG	4	Anti-Estrogens; Antineoplastic Agents; Antineoplasti
Mianserin	SLC6A4 DRD3 OPRK1 ADRA1B	4	Adrenergic Agents; Adrenergic alpha-Antagonists; A
Amitriptyline	SLC6A4 OPRK1 ADRA1B OPRM1	4	
Dextromethorphan	SLC6A4 PGRMC1 OPRM1 OPRK1	4	Alkaloids; Antitussive Agents; Central Nervous Syste
Mirtazapine	OPRK1 ADRA1B DRD3 SLC6A4	4	Adrenergic Agents; Adrenergic alpha-Antagonists; A
Tramadol	OPRM1 OPRK1 SLC6A4	3	Alcohols; Amines; Analgesics; Analgesics, Opioid; C
Zinc	MPG SERPINA1 SERPIND1	3	Acetates; Acetic Acid; Acids; Acids, Acyclic; Acids, N
Amoxapine	SLC6A4 DRD3 ADRA1B	3	Adrenergic Agents; Adrenergic Uptake Inhibitors; Al
Etorphine	OPRM1 OPRK1 OPRL1	3	Alkaloids; Analgesics; Analgesics, Opioid; Central N
Tapentadol	OPRM1 OPRK1 SLC6A4	3	Analgesics; Analgesics, Opioid; Benzene Derivatives
Loxapine	ADRA1B DRD3 SLC6A4	3	Antipsychotic Agents; Antipsychotic Agents (First Ge
Pethidine	OPRK1 OPRM1 SLC6A4	3	Acids, Heterocyclic; Adjuvants; Adjuvants, Anesthesi
Talampanel	GRIA1	1	Benzazepines; Heterocyclic Compounds; Heterocycli
Etanercept	FCGR3B	1	Amino Acids, Peptides, and Proteins; Analgesics; A
Vitamin E	PRKCB	1	Antioxidants; Benzopyrans; Chemical Actions and Us
N-[(2R)-2-benzyl-4-(hydroxyamino)-4-	LTA4H	1	
Adalimumab	FCGR3B	1	Amino Acids, Peptides, and Proteins; Anti-Inflamm
ALPHA-HYDROXYFARNESYLPHOSPH	FNTB	1	Alcohols; Fatty Alcohols; Hydrocarbons; Lipids; Orga

Example,

The two FDA-approved drugs, **adalimumab and etanercept**, may be most promising candidates, because they are inhibitors of TNF-alpha (a key cytokine to regulate immune response) and overexpression of TNF-alpha cause inflammation in various organs, especially in central nerve system.

MedGenMed *Medscape General Medicine*

MedGenMed. 2006; 8(2): 25.
Published online 2006 Apr 26.

PMCID: PMC1785182

TNF-alpha Modulation for Treatment of Alzheimer's Disease: A 6-Month Pilot Study

[Edward Tobinick](#), MD, Assistant Clinical Professor of Medicine, [Hyman Gross](#), MD, Clinical Professor of Neurology, [Alan Weinberger](#), MD, Associate Clinical Professor of Medicine/Rheumatology, and [Hart Cohen](#), MD, FRCP, Associate Clinical Professor of Medicine/Neurology




[CNS Drugs](#)

November 2016, Volume 30, [Issue 11](#), pp 1111-1120

Treatment for Rheumatoid Arthritis and Risk of Alzheimer's Disease: A Nested Case-Control Analysis

Authors

[Authors and affiliations](#)

Richard C. Chou , Michael Kane, Sanjay Ghimire, Shiva Gautam, Jiang Gui

Future strategies and trends

- Big Data era of **genomic medicine and drug discovery**
- Contracting multidimensional network by Deep Learning
 - Apply to big data medicine
 - Correlative network structure of **comprehensive molecular information – clinical phenotype** in genome medicine
 - Disease onset and **genetic – environment factor** in biobank
- AI drug discovery has now ready to be realized
- We are now starting “Big data medicine/AI drug discovery consortium of Japan” to promote the project, coordinated by pharmaceutical company, IT company and medical institution

The Second Generation of Genome Medicine

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

Developmental Origin of Health and Disease (DOHaD)

- **Netherland Famine**

- The end of World War II, Blockade by Nazis
About a half year
- Fetus during the famine
- When became adult, contract obesity, T2D,
- Baker Hypothesis, UK increase of MI



Netherland
famine (1944)

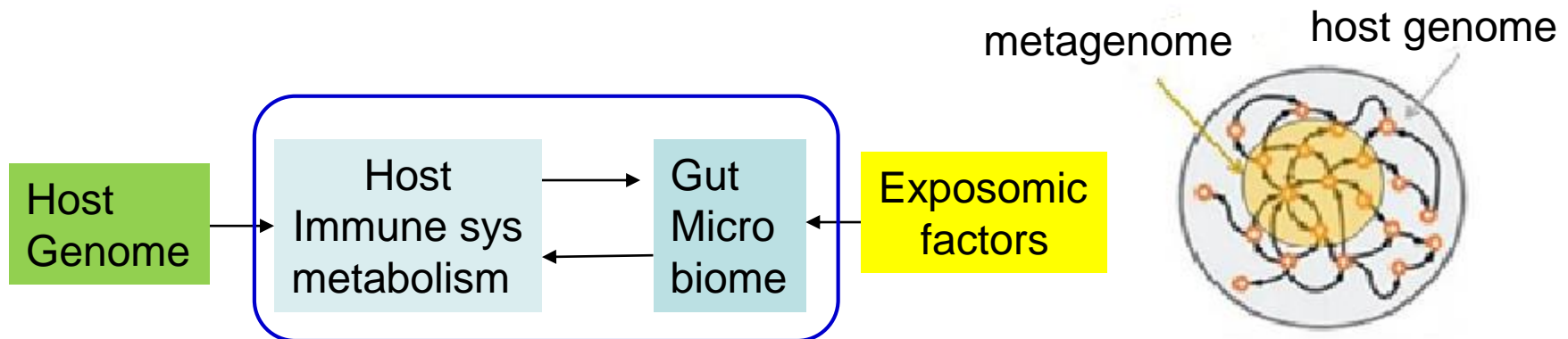
- **Epigenetic Mechanism**

- Excessive undernutrition : in liver **PPAR α / γ** (**thrifty gene**) decline of methylation, gene expression starts
- epigenetic change is reversible, short term change, long term memory to next generation

Environment factor → Epigenetic change → Change of gene expression → Disease occurrence

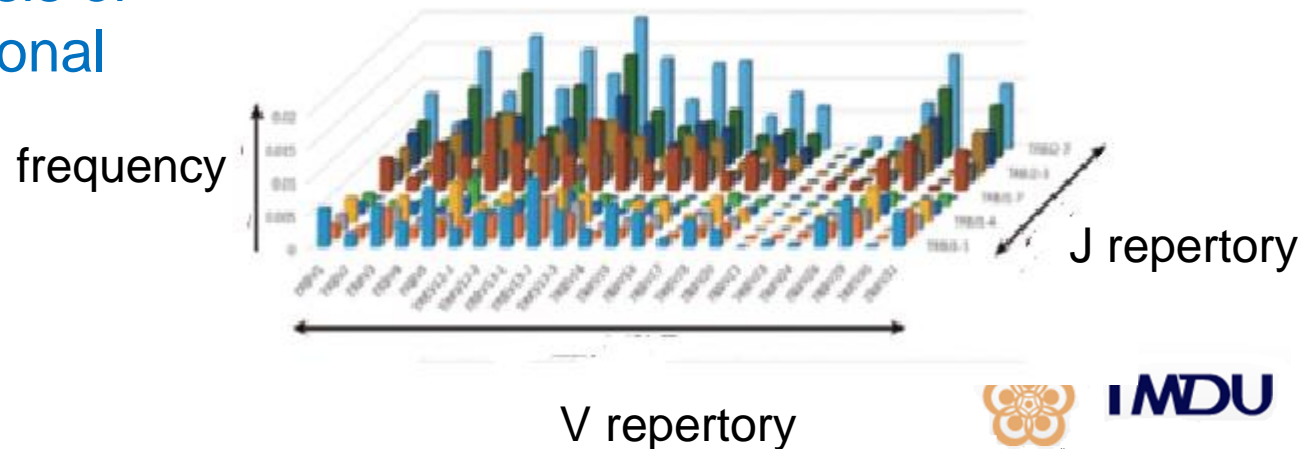
Interstinal microbiome : hologenome

- **Major exposomic factor of Disease**
 - gut microbiome: **one of most biggest environmental factor**
 - About 1000 species, 100, trillion, 1 ~ 1.5kg, substaional organ
 - Num. genes, **half million**, totally one million gense
- **Immune, inflammation, interaction against mucosal immunity cell**
- **Dietary fiber** : metabolized by gut microbiome: “short chain fatty acid” energy source
- Metabolite of gut microbiome (short chain fatty acid, TMAO) interact with host



Immunome

- Variable and Complimentary Region (CDR 3) : DNA/RNA
- next generation sequence
- Immuno-Repertory
 - Total Profile of TCR
 - Three dimensional display of V(D)J
 - Total Distribution is changed instantly with perturbation
 - Frequency of VDJ use
 - Change of diversity
 - Due to disease or aging
- Clinical sequencing
- Feature analysis of three dimensional distribution



The second generation of genomic medicine

- Interaction with Environmental factors
 - Extention of Clinical Sequencing
- Sequence Disease “**Meta Omics**”
 - Epigenome
 - Microbiome
 - Immunome
- Sequencing clinical meta-omics gives the essential information of multi-factorial disease

Thank you for kind attention



Learning Health System

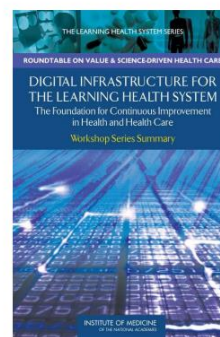
From Discovery of Biological knowledge to Clinical Implementation: 17 yr

While practicing healthcare, acquire the new knowledge

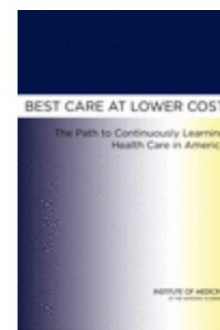
- IOM: “Clinical Data as a Basic Staple of Health Learning”
- “Data obtained from routine medical practice is the Key to support LHS”
Sharing and learning data improves Health care system
- RCT: **Gold standard**, but conducting **outside the ordinary healthcare systems**.
- Is RCT representing the patient group, healthcare is actually directed to
- RCT takes a time and cost
- Effective knowledge accelerate data accumulation

IOM(Institute of Medicine) report
2007, proposed as the paradigm
replacing EBM/RCT

*Digital Infrastructure for the
Learning Health System: The
Foundation for Continuous
Improvement in Health and
Health Care*



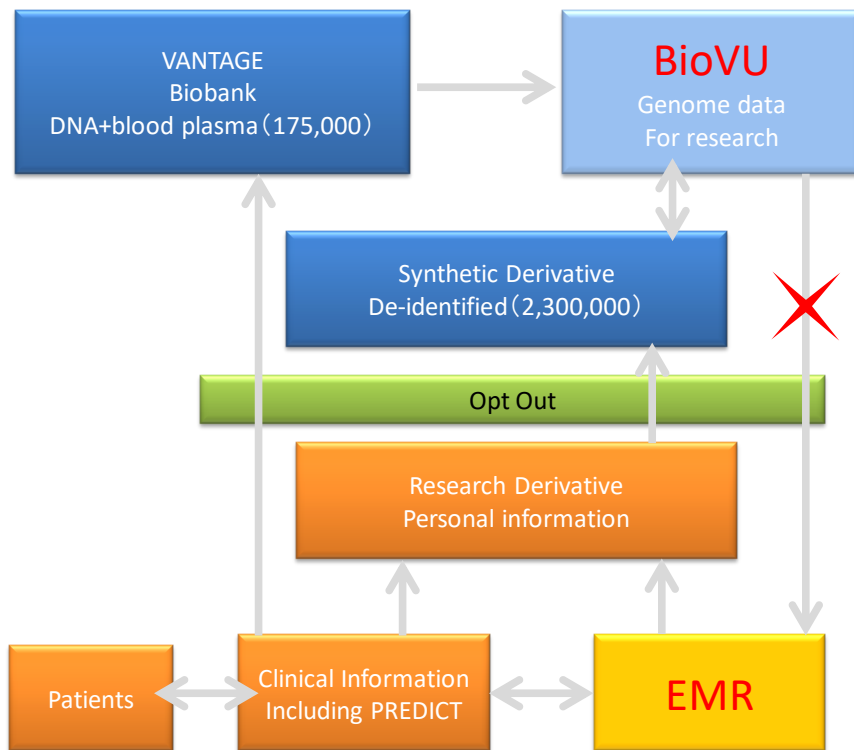
*Best Care at Lower Cost: The Path to
Continuously Learning Health Care in
America*



Typical example of LHS

Integration of Genomic and Clinical information

BioVU Vanderbilt UH



EMR

Synthetic Derivative :

De-identified EMR information
Opt out (2,300,000 records)

Biobank and Genome Analysis

BioVU :

Genome (DNA)
Information Integration with
Synthetic Derivative

VANTAGE Core :

175,000 specimen,
DNA extracted from blood,
Genomic analysis

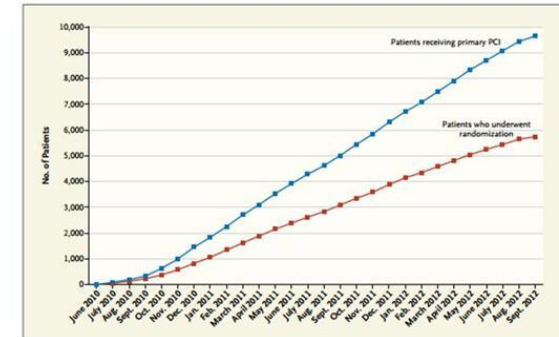
EBM changes to BDM (Big Data based Medicine) Paradigm Shift of Clinical Research

- **Disparity between RCT Study Population and Real World Data**
 - **Impossible in reality** to make study population including **all the stratified (personalized) patterns** of disease
 - Current clinical research study population is in “**artificial environment**” outside real world data
- **Directly use Big Real World Data**
 - No need for **unbiased sampling** from population
 - Because Big Real World Data is very close to population data
 - But still exist the **bias and confounder** (causality) problem

Possible Solution

Registry-based Clinical Randomized Trial

- Advantage to use “Real World Data” and the rigorous “Randomization” is fused
 - Thrombus Aspiration in ST-Elevation Myocardial Infarction in Scandinavia (**TASTE**)
 - first trial of RRCT with cost 50 \$ per participant
 - Large scaled trial build on already-existing high quality registry
- **RRCT process**
 - Select the **study population** from the **disease registry** where already exist much of clinical information (7244 MI patients)
 - **Randomized allocation of study and control drug** to selected population among registry
 - **End point of trial** is observed by registry.



Rapid Randomization in the TASTE Trial, with Enrollment of Most Patients Receiving Primary Percutaneous Coronary Intervention (PCI). Adapted from the Institute of Medicine (www.iom.edu/-/media/Files/Activity%20Files/Quality/VSR7/L57%20Workshop/Presentations/Granger.pdf). The incremental cost of the Thrombus Aspiration in ST-Elevation Myocardial Infarction in Scandinavia (TASTE) trial was \$300,000, or \$50 for each participant who underwent randomization.

