Integration of Genomic and Phenomic Information in Medicine

- Big Data Approach to Medical Knowledge Discovery -

Dept. Biomedical Informatics
Tokyo Medical and Dental University
ToMMo Tohoku University
Hiroshi Tanaka

TMDU

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?

TMDU

Conventional Big Data of Japan

- NDB: National Database
  - Database of reimburse claim data
  - 2011-2012 trial use：strict MLHW inspection, only 18 applications
  - More than 6 billion data.
- Japanese Sentinel Project Database
  - DB of drug prescription, adverse effect
  - 2010- MLHW, PMDA
  - Aiming at accumulating 10 million patients' data,
  - 10 national University Hospitals' Data to PMDA
  - 2016- open for public
- Surgeons' Operation Registry Database
  - National Clinical Database
  - Academic association qualification for specialist
- DPC data
  - Diagnosis Procedure Combination,
  - Japanese version of DRG/PPS, Data are available

Dr. Kimura talked conventional Big data yesterday

TMDU

New type of Big Data emerges
Era of Big Data in Medicine

1）Clinical Information
- Conventional Clinical Information
  - Lab test, Image, Prescription etc.

2）Socio-medical Information
- Population Medicine Information
  - Epidemiological data, medical policy

3）Molecular Information
- Genome, omics information
- Personalized medicine/prevention

Conventional Medical Big data

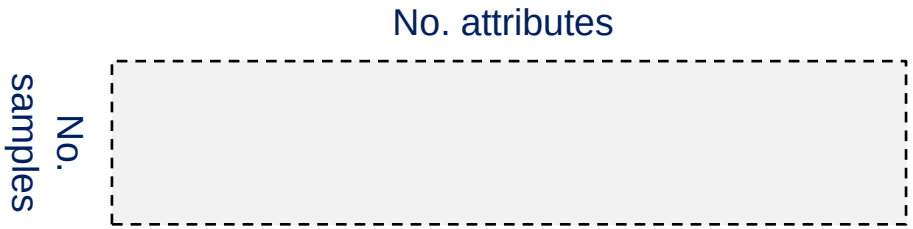
New type of Medical Big Data

TMDU

New type of Medical Big Data

- Conventional Medical Big Data
  - **“Big Small Data”**
    - For one subject(sample)
      - several tens items (attributes)
    - Big number of subjects(samples)
    - Conventional statistics framework

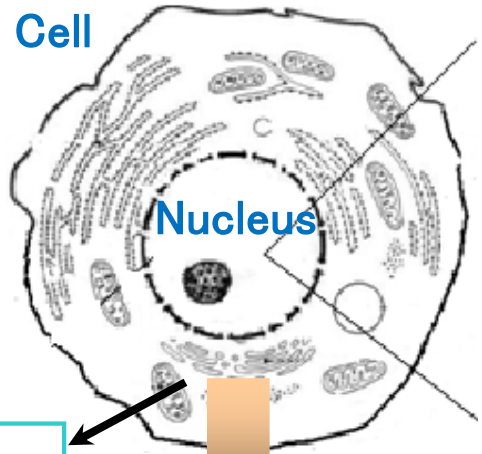
- Molecular Big Data (genome, omics)
  - **“Small Big Data”**
    - Enormously many kinds of data for one sample (patient)
    - Whole genome sequence, 100Gbp for one sequencing
    - Gene expression profiles ～40000 probes data (L.Chen)

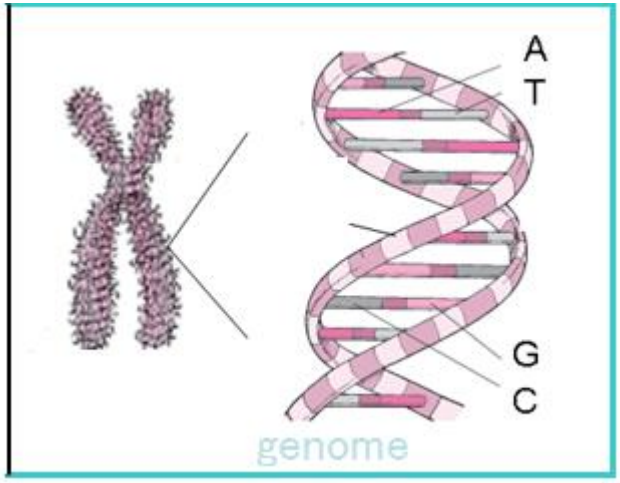


TMDU

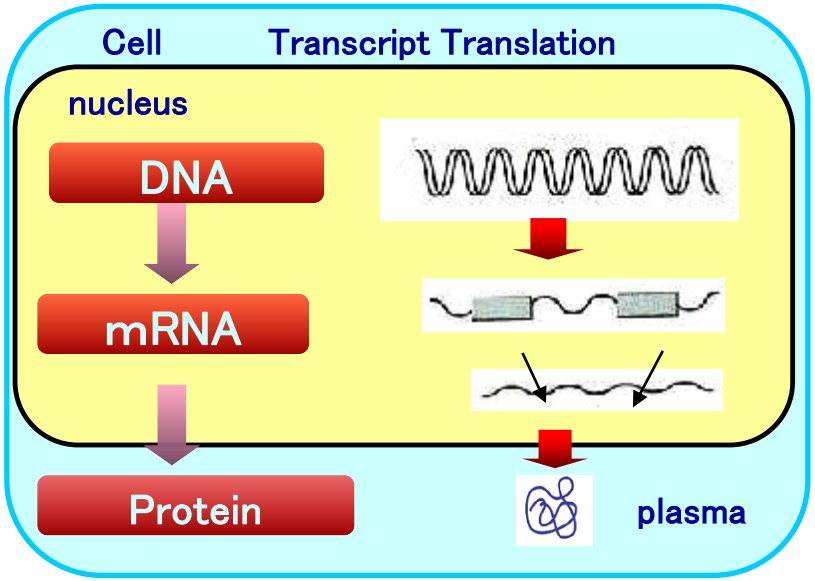
Genome and Omics

1990 Human Genome Project
2003 HGP finished





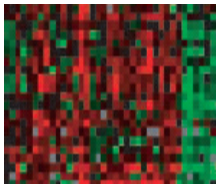
Omics
＝ −Ome (Whole)＋−ics (study)

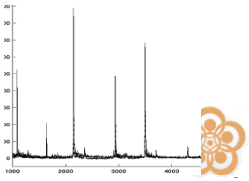


genome
genomics
transcriptome
transcriptomics
proteome
proteomics
metabolome
metabolomics

sequencer

DNA microarray





Mass Spect

TMDU

The second genome revolution

Next generation sequencer

13years⇒1day, 350 B dollar ⇒1000 dollar

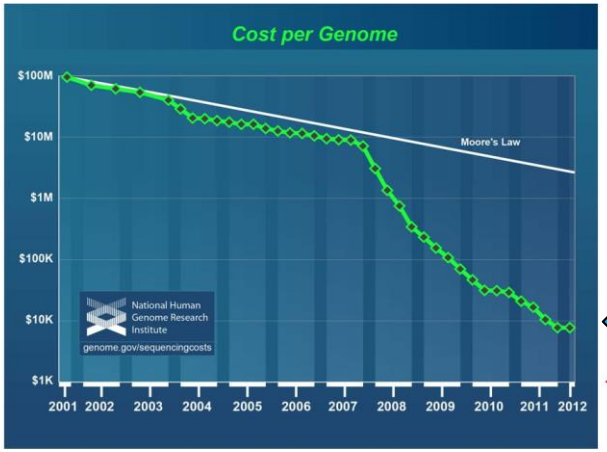
Ilumina 2500

Ion Torrent

Ilumina 2500

WGS(Whole genome sequencing)
3GB (1 person) X 30 = about 100Gbps
1 person WGS 27 hours

WES(Whole exome sequencing)
60Mb (1 person) X 100 = 6Gbps
15 persons WES for 27 hours



1000 dollar

DNA Sequencing Cost: the National Human Genome Research Institute

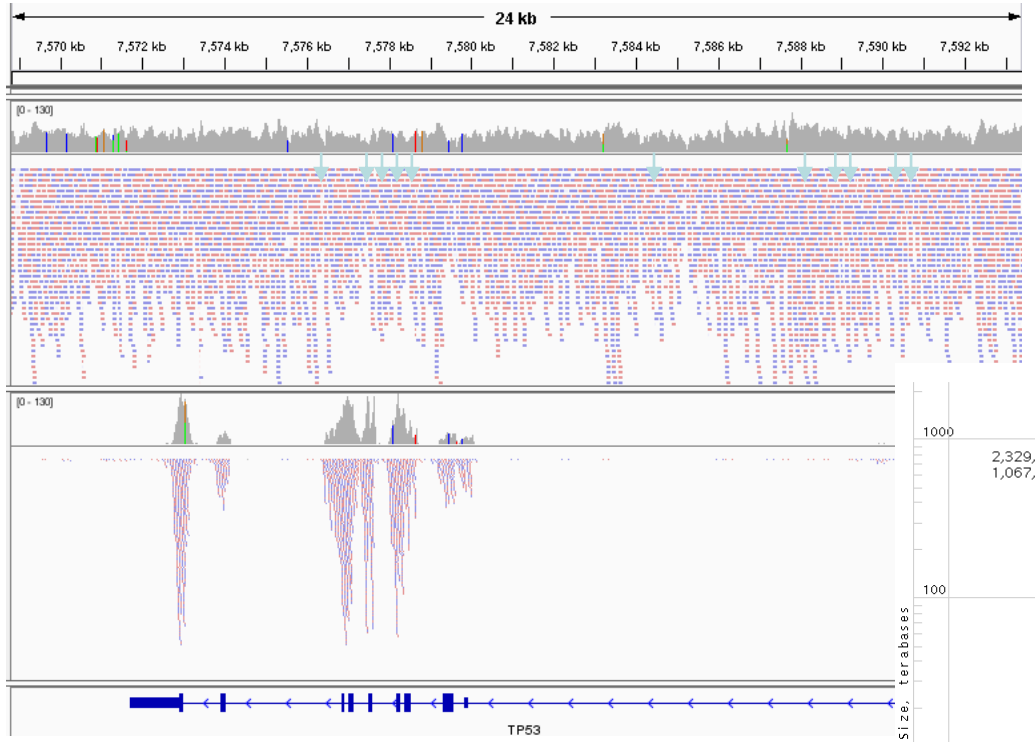
1000 dollar NGS
Ilumina Hiseq X (10set)

TMDU

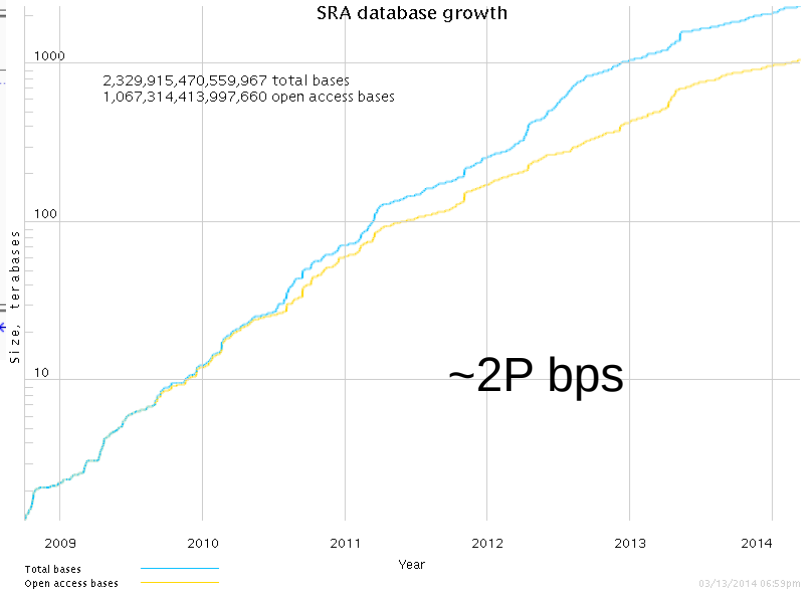
Sequence data

WGS

WES



(p53: 17chr p13.1)



~2P bps

NCBI Sequence Read Archive (SRA)
http://www.ncbi.nlm.nih.gov/Traces/sra/

TMDU

Generation of Medical Big Data

Practice of Genome medicine

Medical Big Data

Rapid advances in high-throughput biotechnology
Next generation sequencer etc.

Spread of Clinical Sequencing in Hospital

WGS 100Gb

WES 6Gb

Tens of hospitals in US practice

Accumulation of genome/omics data

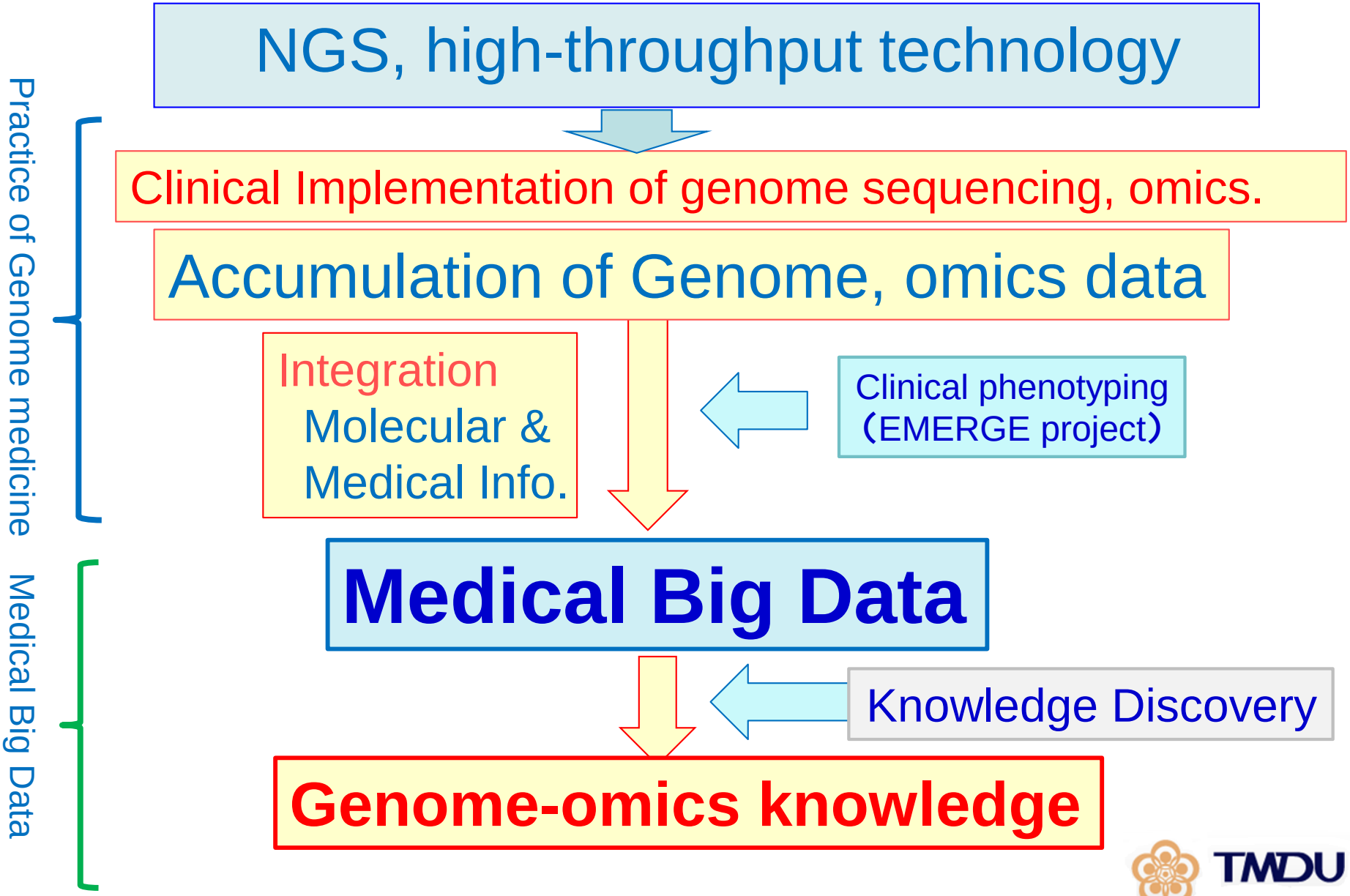
Clinical Information Integration

Clinical Phenotyping

Medical Big Data

TMDU

Genome omics medicine and Big Data



TMDU

Medical Big Data

Big Data for Healthcare, Drug Discovery

- Healthcare, Medicine
  - Personalized Medicine,
    - Genome omics medicine・Precision Medicine
    - Large scale Biobank, disease cohort
  - Personalized Prevention
    - population biobank cohort spreading all over the world
- Drug Discovery
    - Drug discovery/repositioning
      by connectivity map and gene expression profle DB
  - In silico screening

TMDU

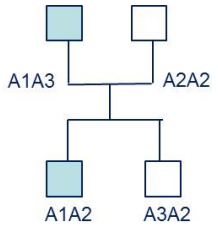
Personalized Medicine
Clinical Implementation in United States
Genome/Omics Medicine

More than 20 hospitals have implemented Genome/Omics medicine

Institution	Major Projects
MC Wisconsin	Using whole genome sequencing to establish diagnosis in patients with currently undiagnosed genetic disorders
Mount Sinai	• CYP2C19 testing for antiplatelet rx post percutaneous coronary intervention • Personalized decision support for CVD risk management incorporating genetic risk info
Northwestern	Using pharmacogenomics evidence (from GWA genotyping) to guide prescriptions in primary care and assess risk for other conditions such as HFE/hemochromatosis
Cleveland Clinic	Tumor-based screening for Lynch syndrome, endometrial cancer
UCSD	• Screening for actionable mutations in malignant gliomas and glioblastomas for biomarker based RCTs • Targeted rx (such as RET inhibitor) of metastatic solid tumors based on tumor mutation status
Morehouse	• Exome sequencing of 1200 early onset severe African American hypertension cases and 1200 controls
Duke	• Computer-based family hx collection and CDS tool with 1-yr follow-up for perceptions, attitudes, behaviors related to thrombosis and breast, ovarian, and colon cancer • SLCO1B1*5 genotyping and statin adherence • Effect of genetic risk info on anxiety and adherence in T2DM

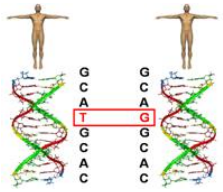
Institution	Major Projects
Alabama	Planning stages for projects in risk assessment, pharmacogenetic analysis, identification of families for further research
Baylor	Whole exome and whole genome sequencing in Mendelian disorders to improve diagnosis
Geisinger	• Selection for gastric bypass surgery vs other wt loss means based on genetic variants predictive of long-term benefit from surgery • IL28B variants and response to hepatitis C treatment • KRAS and BRAF mutational analysis in thyroid cancer patients
Ohio State	• Personalized genomic med study of CHF and HTN pts randomized to genetic counseling vs usual care • CYP2C19 testing in interventional cardiovascular procedures for clopidogrel
Harvard	Whole genome sequencing with integration in EMR and CDS; pilot of 3 patients to start
U Penn	Genotyping for assessment of MI risk in Preventive Cardiology program
St. Jude's	Pre-emptive PGx genotyping in children
Vanderbilt	Pre-emptive PGx genotyping for clopidogrel, warfarin, or high-dose simvastatin
U Maryland	Develop and apply evidence-based gene/drug guidelines that allow clinicians to translate genetic test results into actionable medication prescribing decisions
Mayo	• PGx driven selection/dosing of antidepressants • CYP2C19 genotyping for antiplatelet rx post PCI
Inter-Mountain	Tumor-based screening for Lynch syndrome

Personalized Medicine



1st generation "Genomic Medicine"(1990～)

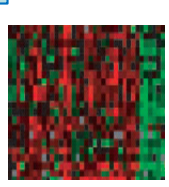
- Human genome ～0.5% different, mutation /polymorphism, **SNPs**
- Based on the **inborn** (germline) **individual differences of genome**
- Aiming at **"Personalized medicine"**



- Estimation of "constitutional risk" of contracting disease
  - **disease causative gene** for genetic disease,
  - **disease susceptibility gene** for " common disease (hypertension, Diabetes) SNP
  - No treatment for genetic disease, **low genotype relative risk** for common disease
- Personalized medication based on pre-diagnosis of drug response
  - Pharmacogenomics (PGx) diagnosis of different individual response to drug

2nd generation "Omics-based Medicine"(2000～)

- Based on and **direct use** of **"acquired omics profile"**
- Aiming at **"Predictive/Preemptive medicine"**



gene expression

- Using **omics profile of disease** (gene expression profile, etc)
  - Diseases due to **acquired somatic** cell mutation /alternation
  - It changes depending on disease stage and sites ("molecular phenome")
- Estimation of degree of on-going state of disease progression
  - Discover of **disease subtype** based on "omics profile", ex. breast cancer
  - Directly related to **prognosis** or **early detection** of disease more precise than clinico-pathological findings

TMDU

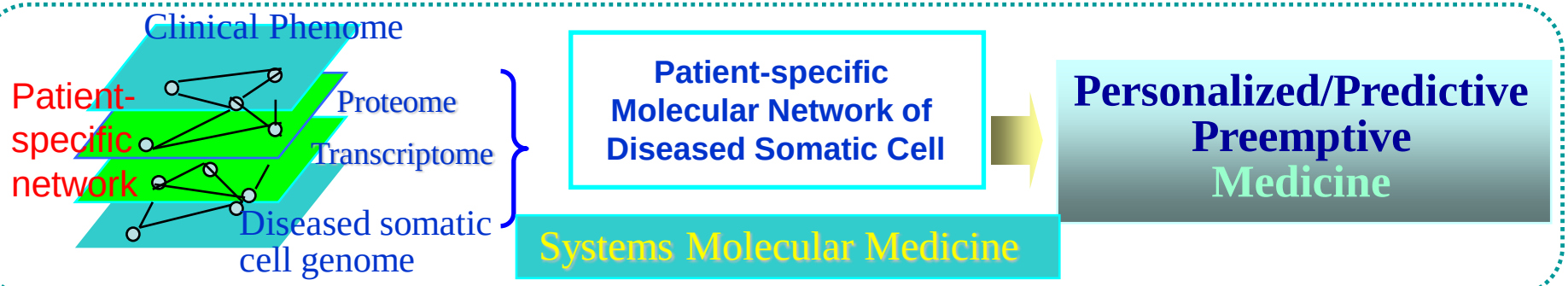
The Third Generation of Molecular Medicine

- "Systems Molecular Medicine"
  - Methodological basis: **Disease systems biology**
  - Understand disease as an unified system
  - Disease is system distortion of molecular network
  - "Pathway centric" diagnosis/therapy, not "molecule centric"
  - Main approach is Using omics profile to identify the patient –specific dysregulated (distorted) pathway branches

Molecular systems medicine

systems pathology

- Application to Cancer medicine
  - "**Cancer systems biology**"
  - More than 10 institutes for cancer systems biology in US



TMDU

Major Areas of Genome/Omics Medicine is mainly first generation (genomic medicine)

1. Identification of unknown disease causative gene at the point of clinical routine practice
   Wisconsin Univ. Baylor Medical Colleage
2. Identification of cancer driver mutation
   Mayo Clinic, MD Anderson cancer center
3. Identification of well-known disease causative gene
   BRCA1/2 etc.
4. Identification of polymorphism of drug metabolizing enzyme (EMR implementation)
   Vanderbilt Univ. ・Mayo Clinic

Wiscon

Genome sequencing program, Patient Section

Baylor Medical Colleage

Whole genome laboratory In-house, Seq

Vanderbilt

EMR

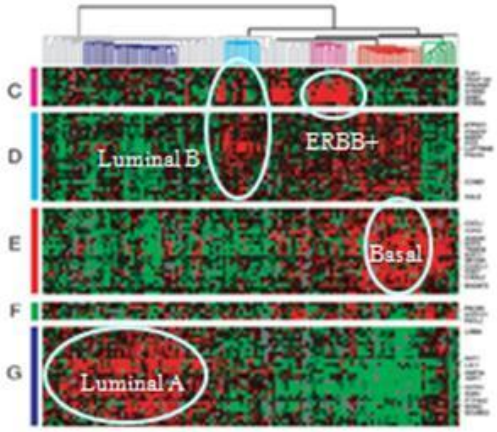
Genome/Omics medicine in Japan
- trial stage-

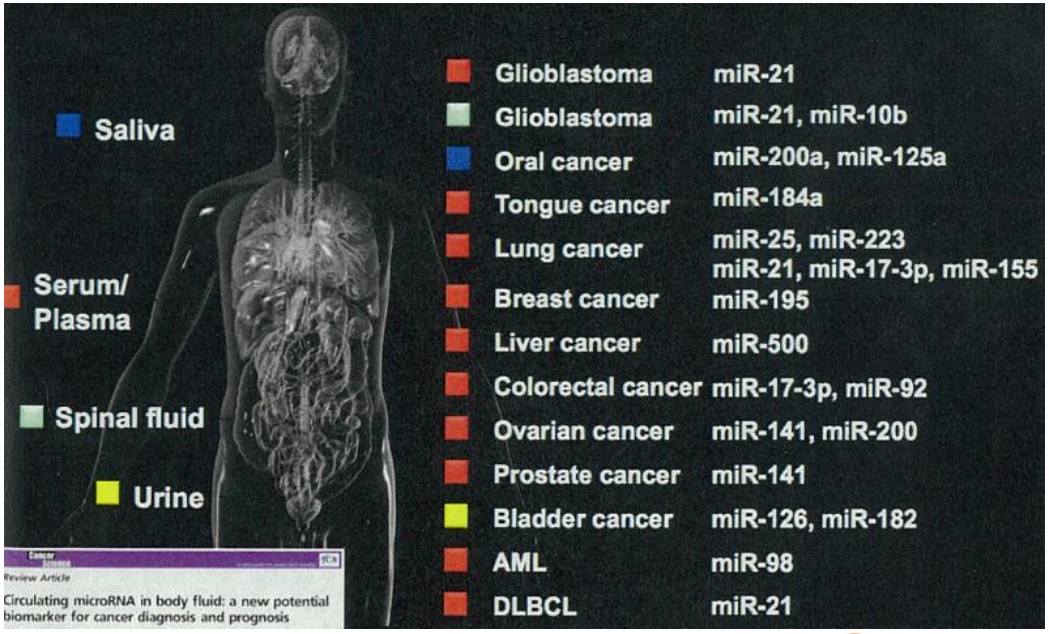
- National Cancer Center: Hospital East
  - Research Center for Innovative Oncology (2014 ~)
  - Targeted sequence to find driver mutation of cancer
  - Allocate a patient to the clinical trial for anticancer molecular target drug
  - Supported by research fund
- Sizuoka Cancer Center
  - HOPE project (High-tech Omics-based Patient Evaluation)
  - Multi-omics based evaluation technology for driver mutation of cancer
  - Supported by research fund
- The University of Tokyo: Center for Genome Medicine
  - Identify genomic cause of intractable disease
  - Genetic counseling and reference
  - Research fund and Patient's own expense
- Juntendo University Hospital
  - Personalized medication based on polymorphism of drug metabolizing enzyme (preparing)

TMDU

Omics measurement

- Gene expression profile is established in
  - Breast Cancer Intrinsic classification
  - Prediction micro arrat
  - mammaprint (70 genes) oncotype D (25 genes)
- microRNA, exosome
  - Excellular RNA（exRNA）
  - 84% correct, Salve test for spleen cancer
  - National Cancer Center：NEDO 5yr.project（7.9 B¥）
    - Serum miRNA, miRNA chip, Biobank
- Liquid Biopsy
  - Circulating Tumor Cell
  - Circulating miRNA
  - Circulating DNA
  - Exosome
  - Cancer metabolome



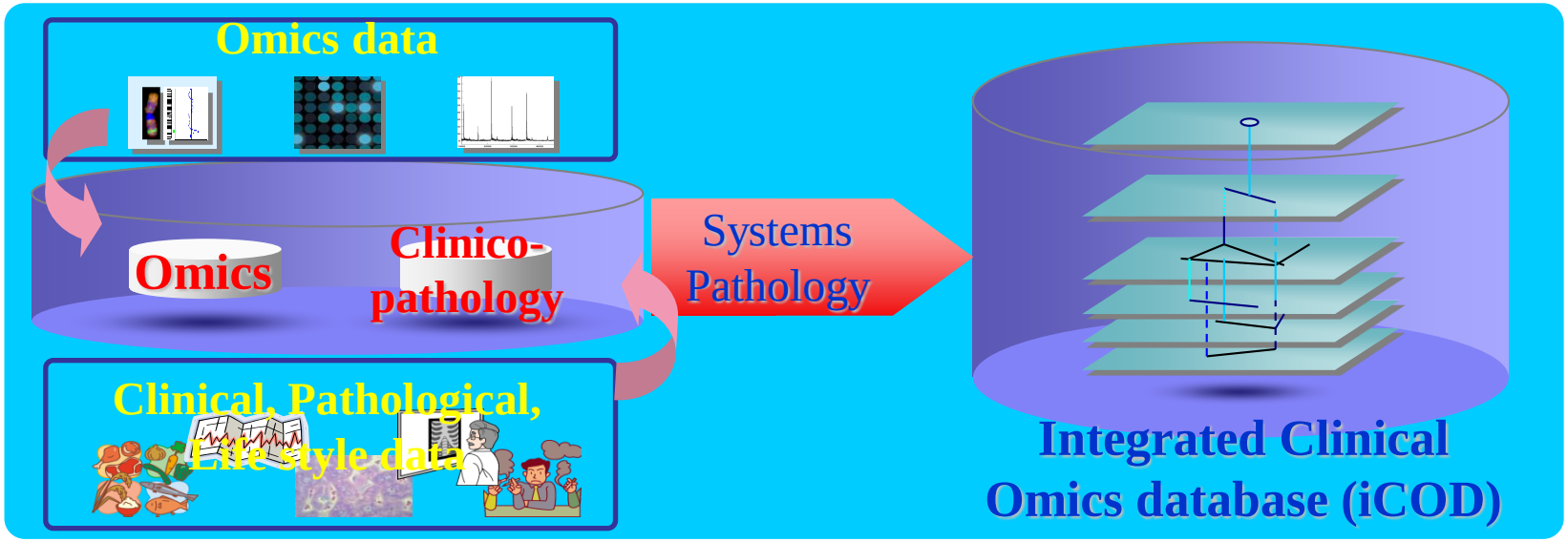


Kosaka,et al Cancer Sci. 101, 2087-2092,2010, Ochiai Mod.

TMDU

iCOD – integrated Clinical Omics DB

- Bridging the molecular omics information and clinical/pathological, life style information
- Government commissioned project of Integrative database with more than 800 cases based on the concept of "omics-based systems pathology"



Shimokawa ,K., Tanaka, H. et.al,
iCOD: an integrated clinical omics database based on the systems-pathology view of disease. ***BMC Genomics***. 11: S19. (2010)

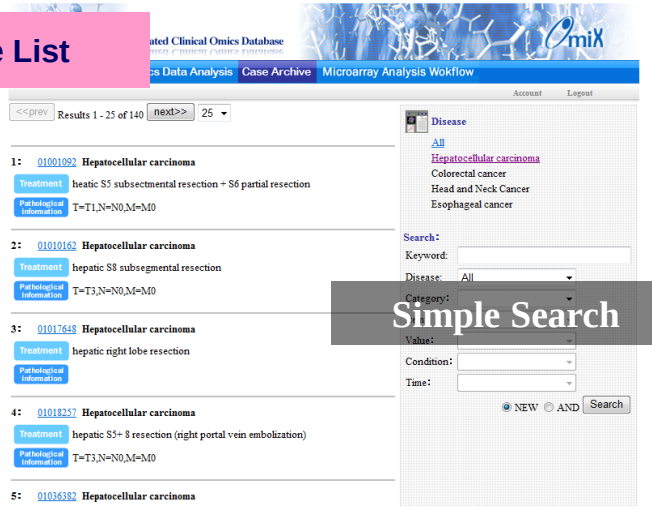
TMDU

Screen shots of iCOD

Top page

Case Archive

Case List

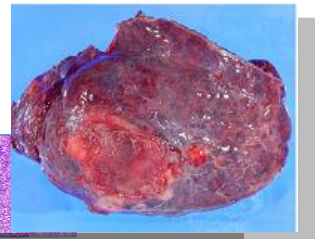


Simple Search

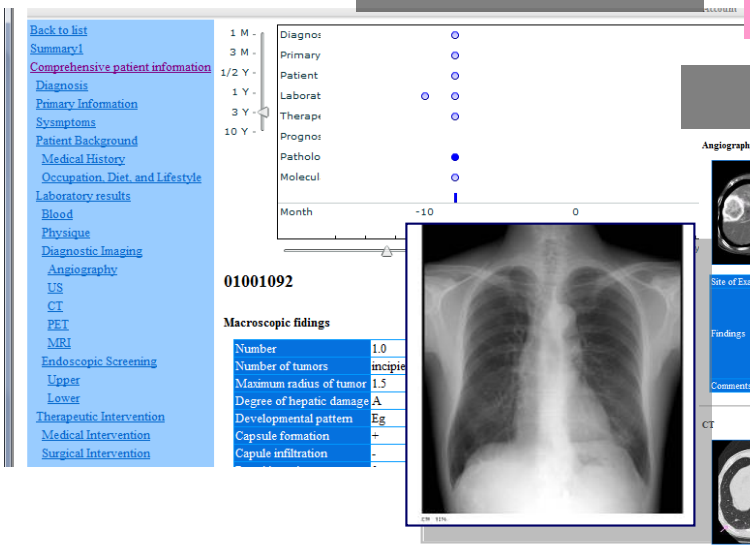
timeline view

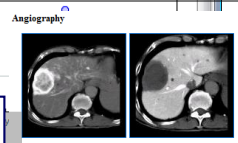
Case Details

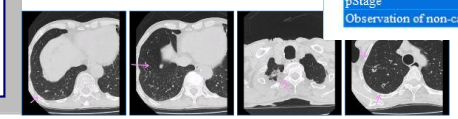
Pathological data



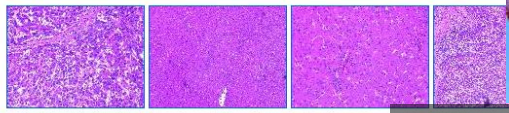
Imaging data







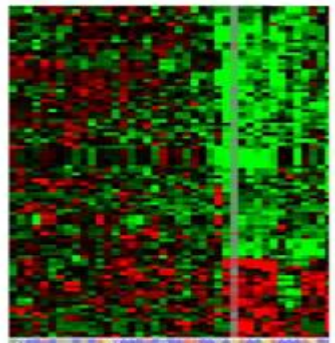
Tissue findings



Diagnosis	Hepatocellular carcinoma
Differentiation	poor
Grade	compact
Developmental pattern	eg
Capsule formation	+
Capsule infiltration	+
Portal invasion	0
Hepatic vein invasion	3
Portal vein/ hepatic vein invasion	+
Hepatic artery invasion	0
Bile duct invasion	0
SM	-
pStage	IVA
Observation of non-cancerous region	CH

Molecular Data

Sampling	surgical
Sampling Location	cancerous region
Storage at Nod al Temperature	0:01



Site of Examination	lung
Findings	Funicular shadow or soft tissue image with indication of bronchiectasis observed in the right pulmonary apex. Two cavities of nodular shadows with bronchiectasis, the size of which are both less than 11mm present in the right S1. 2 nodular shadows the size less than 7mm observed in the base segment of the right lung. Funicular shadows observed in the right S2, S6, and the base segment of the left lung. Bronchiectasis also present in S6, possibly cured inflammation. No pleural effusion, pleural thickening and calcification observed in the left lung, possibly cured inflammation. Lymphadenopathy of the mediastinal hilus uncertain. No artery enlargment. No adrenal gland mass.

http://omics.tmd.ac.jp/

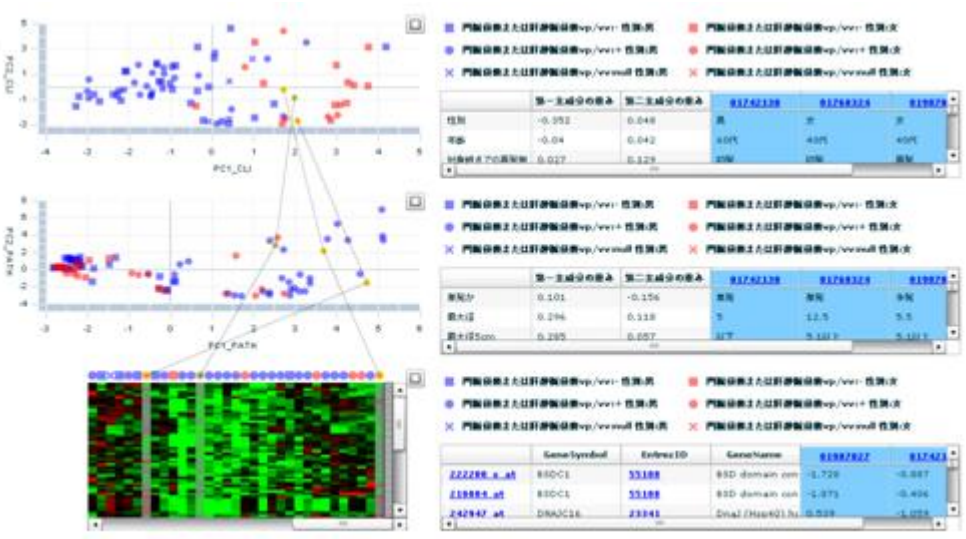
Analysis between molecular and of clinical phenotypes in iCOD

Three Layered Map

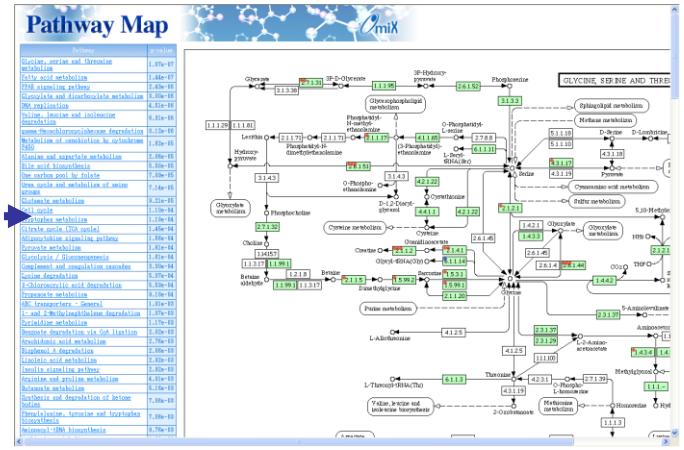
Clinical Level

Pathology Level

Molecular Level

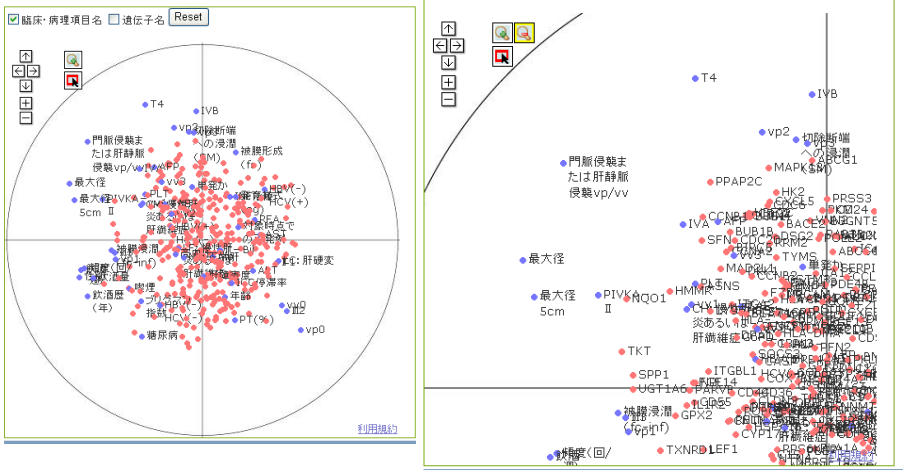


Pathway map



Transcriptome mapped on KEGG

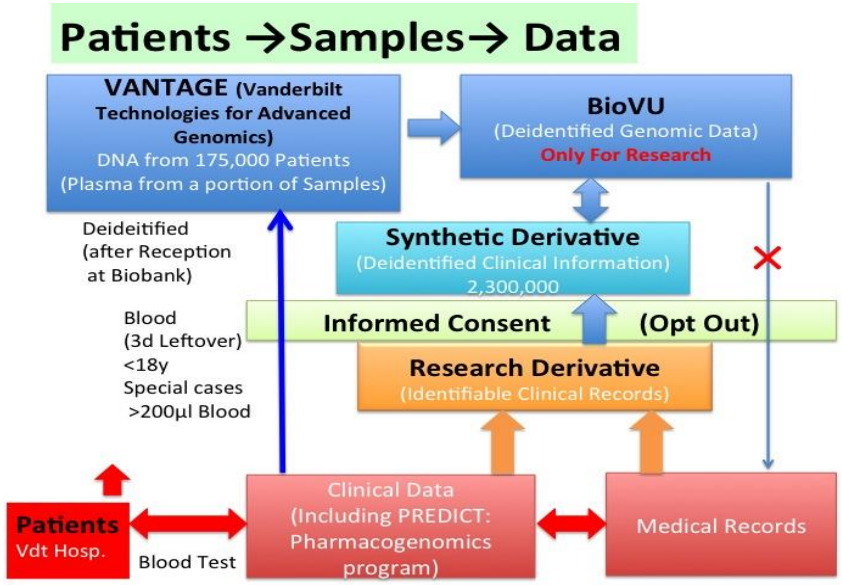
Pathome-genome Map



Integrated Clinical Omics Systems is an Institutional LHS

Aiming at realization of Personalized medicine

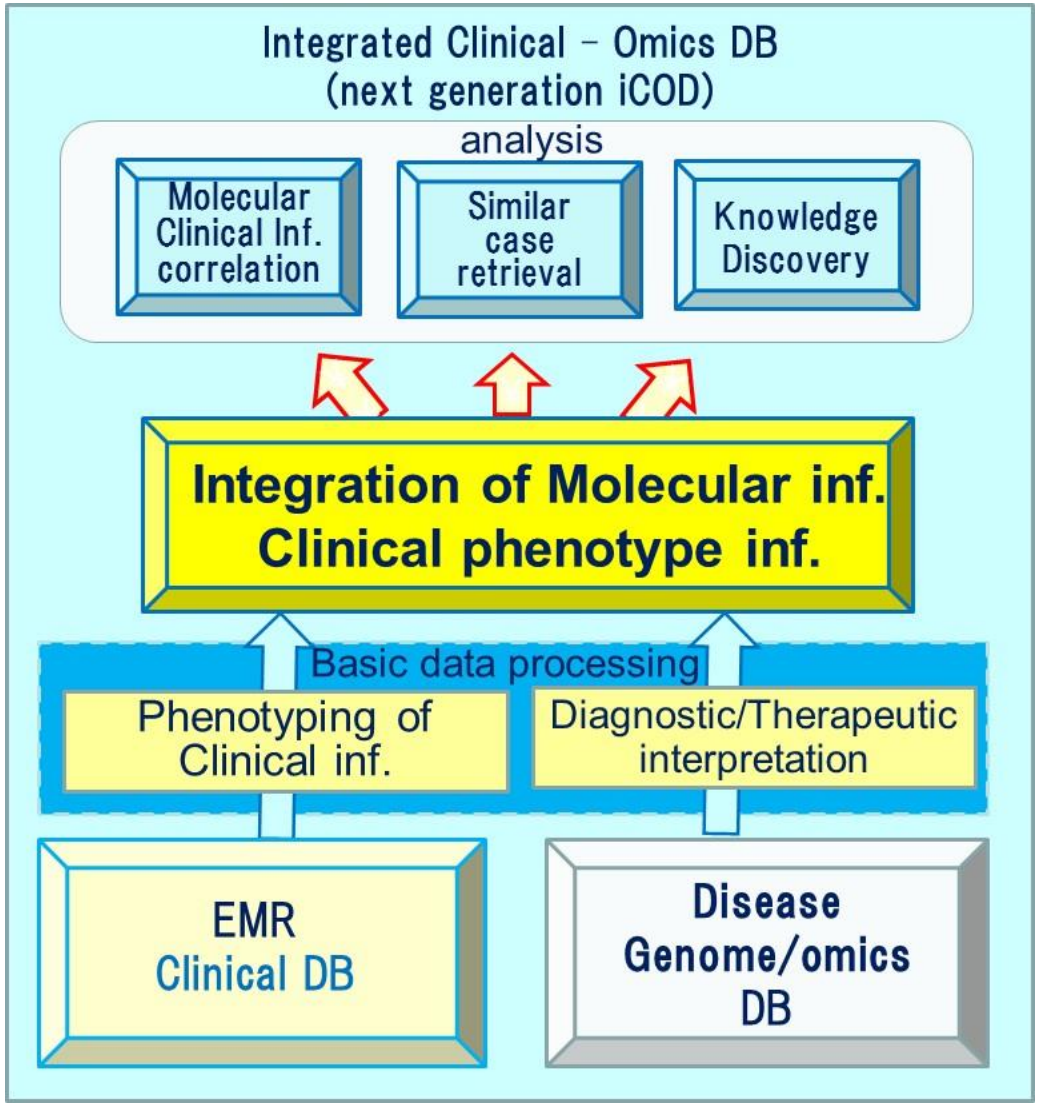
- Learn the molecular-clinical phenotype information relation
- Extract clinical knowledge to feedback to clinical practice
- Develop genome EMR system
- Similar to Vanderbilt system



Vanderbilt Univ. Hospital system

TMDU

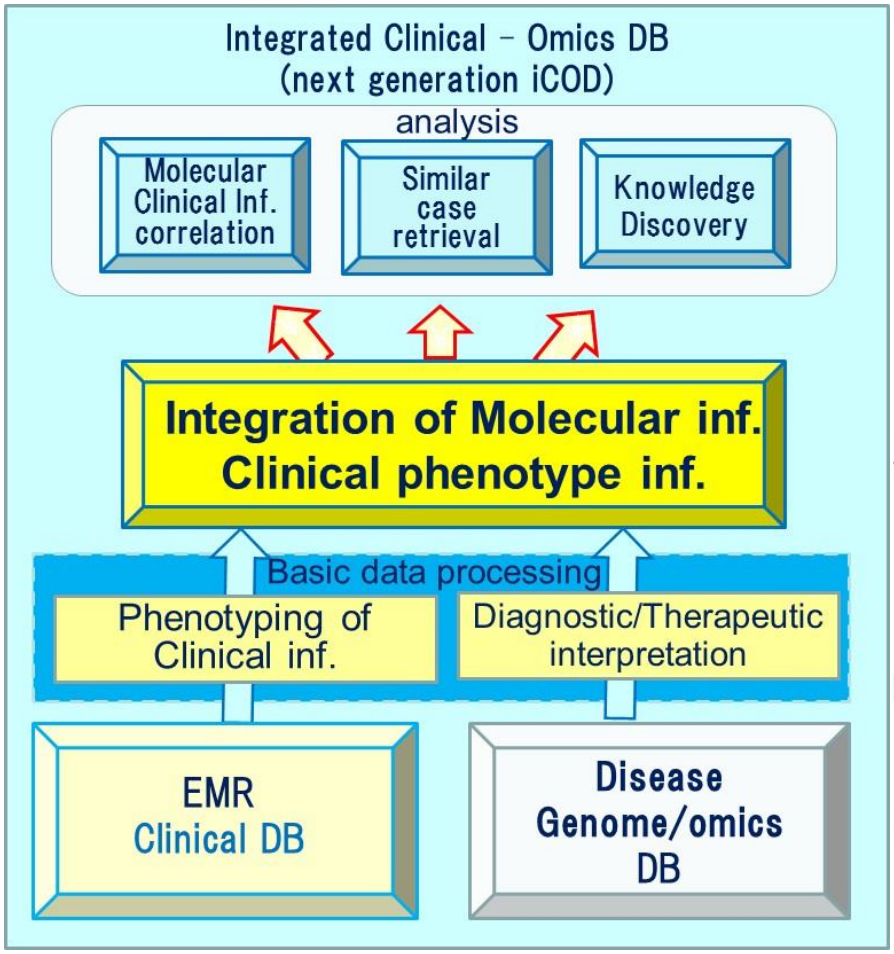
Basic DB Structure for Genome/Omics Medicine, Integrated DB



1. Molecular (disease omics) and clinical phenotypic data integrated to form the unified profiles of disease by network representation
2. Each kind of data must be preprocessed before integration
   Phenotyping of clinical data
   Interpretation of genome/omics data
3. From integrated representation of unified disease profile, we
   1) molecular-clinical phenotype correlation analysis
   2) similar patient case retrieval case-based inference
   3) clinical knowledge discovery

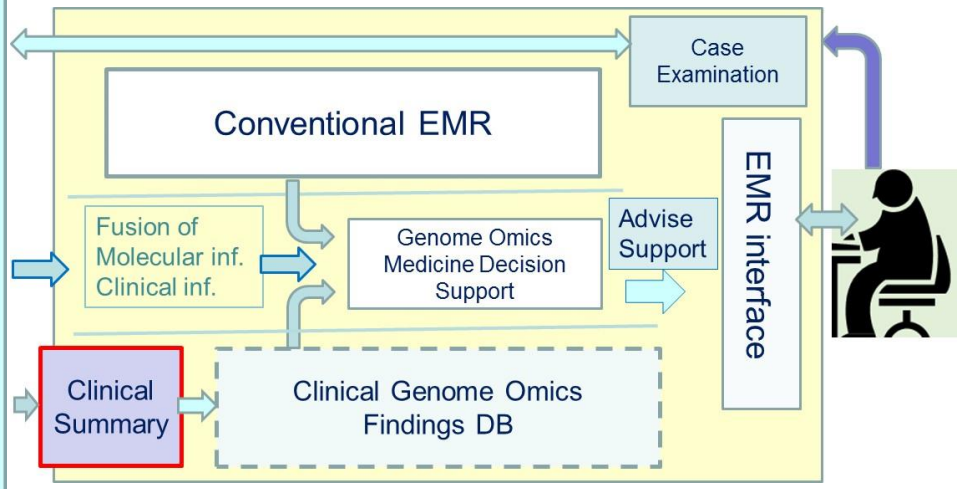
TMDU

Basic Structure of genomic EMR



Genomic EMR consists of two parts

1. Integrated clinical omics database
2. Hospital implemented system of Genomic EMR



Making New type of Medical Big Data

TMDU

NIH

"Big Data to Knowledge" (BD2K)initiative

- Previous Project: "Biomedical Information Science and Technology Initiative (BISTI)"
- BD2K: Big Data to Knowledge Initiative 2013 start
  - WG on Data and Informatics for Advisory Committee to the Director (ACD) of NIH
    - several focused **workshops**, calls for **proposals** for centers of excellence, for a data discovery index, for training programs,
    - **Associate Director of Data Sciences-**--New Position
  - Francis Collins： "lead an NIH-wide priority initiative to take better advantage of the exponential growth of biomedical research datasets, which is an area of critical importance to biomedical research. The era of '**Big Data**' has arrived, and it is vital that the NIH play a major role in coordinating access to and analysis of many different data types that make up this revolution in biological information."
  - http://bd2k.nih.gov

NIH “Emerge Project”

- The Electronic Medical Records and Genomics (*eMERGE*) Network
  - National Human Genome Research Institute (NHGRI) – funded consortium
  - Developing methods and best practices for the utilization of the electronic medical record (EMR) as a tool for genomic research.
  - nine groups: each with its own biorepository (DNA etc) linked to phenotypic data contained within EMRs.

“Medical BigData”

- eMERGE consortium
- CSER consortium
  - “Clinical Sequencing Exploratory Research” NHGRI
  - explore the potential of whole-genome and whole-exome sequencing to generate new knowledge and improve patient outcomes
  - Many of the issues are also relevant to the eMERGE consortium (designated liaison)

Medical Big Data

Genome ＋ Clinico - Environmental（EHR)

＋

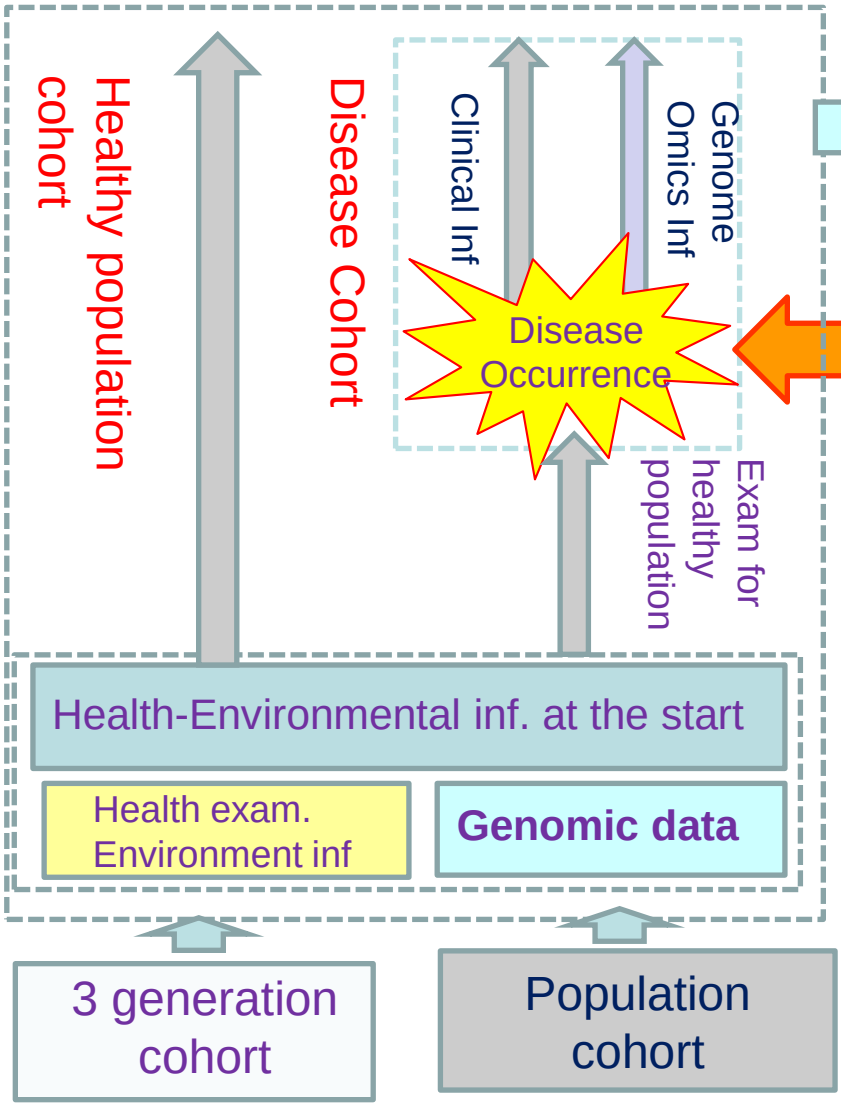
Learning System

Big Data and Learning system

- Learning system: ASCO (American Society of Clinical Oncology)
- The ASCO **CancerLinQ** initiative
  - focused on building a “learning health system” composed of a knowledge-generating computer network
  - collect and analyze cancer care data from millions of patient visits and expert guidelines
  - feed the knowledge back to providers at the point of care
  - Pilot prototype in 2013
    - every patient’s experience to help inform future cancer care would help drive the advent of personalized medicine
    - a 170,000-record prototype Production version by 2015
    - For any given tumor type, database of 10,000 to 20,000 patients, and with 50 to 100 common tumor types, records of at least one million patients
    - uses statistical functions and an artificial neural network to learn, structure, and map data fields
- Cancer centers and IBM Watson
  - Memorial Sloan-Kettering Cancer Center（MSKCC）
    - The Oncology Expert Adviser software (OEA)
  - New York Genome Center
    - Glioblastoma as a target

TMDU

Personalized Prevention
Prospective Population Biobank



ToMMo Biobank

Medical Information from Regional healthcare network (MMWIN)

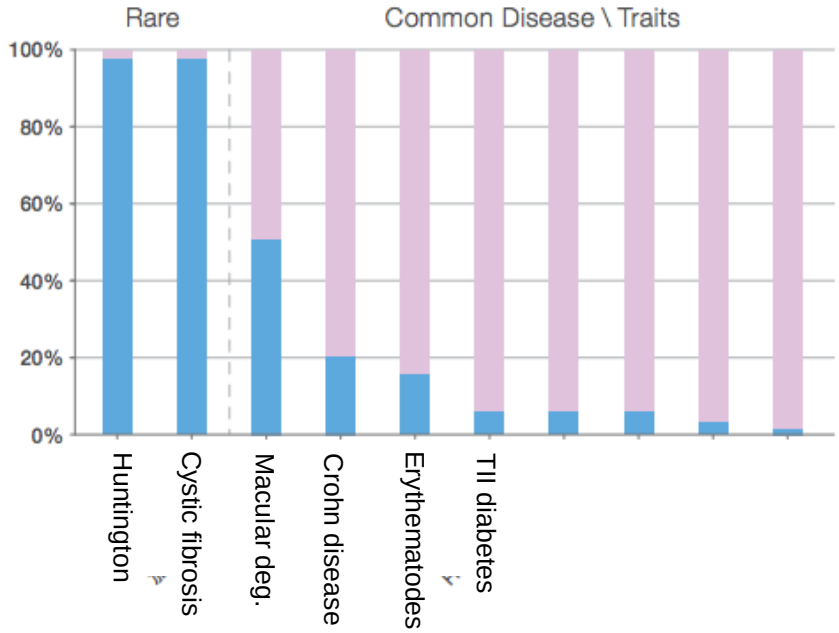
As LHS Aims at

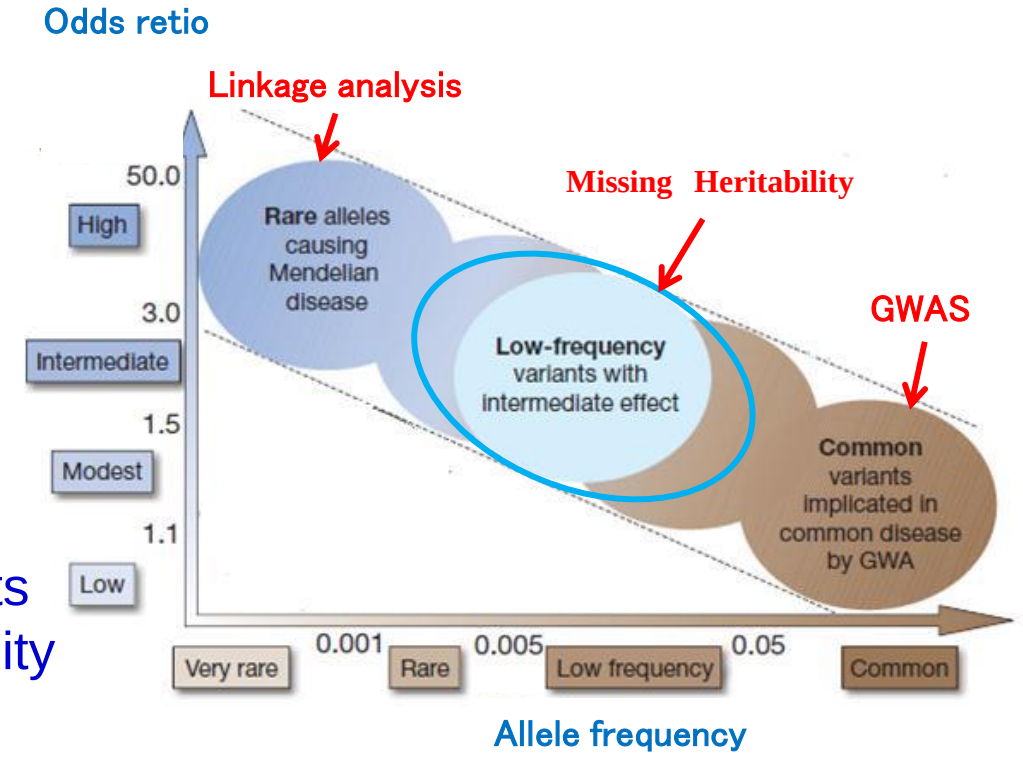
Personalized Prevention

Prospective Biobank

TMDU

Missing Heritability and GxE interaction





Linkage analysis and GWAS results not enough to explain all the heredity of disease

Our view

limitation to ascribe disease cause to a single variant

Gene-gene interaction

Pathway-integrated polygenic effects

Gene-Environment interaction

Teri A. Manolio, Francis S. Collins et al. Finding the missing heritability of complex diseases, vol 461, 8 October 2009

TMDU

Idiosyncratic Effect of Combination of G x E factors

- Interaction between genetic factor and environmental factor
- Relative risk of colon cancer in Hawai
  - **Relative risk is not the multiplication, idiosyncratic** Effect

		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

HCA (hetero cyclic amine) Carcinogen

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

TMDU

GxE interaction In PTSD

- Serotonin transporter
  - SLC6A4（STin2, 5-HTTLPR, rs25531）variation, effects PTSD
- Follow-up study on North iIlinois University Gunfight Incident
  - Arch Gen Psychiatry. 2012 Jan;69(1):89-97.
- 5-HTTLPR genotype L/L and anxiety sensitivity and childhood trauma by GxE interaction
  - Depress Anxiety. 2011 Dec 21;28(12):1048-57.

Identification of Gene-Environment Interaction related to disease development

simulation data by GENS2* / ToMMo** data

*Gene-Environment iNteraction Simulator 2
**Tohoku Medical Megabank Organization

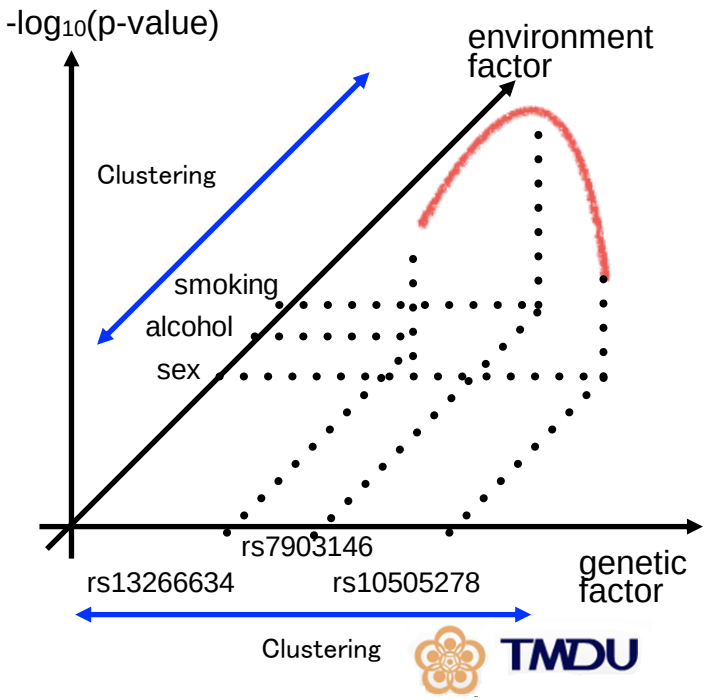
2 x 3 Contingency Table

Environment

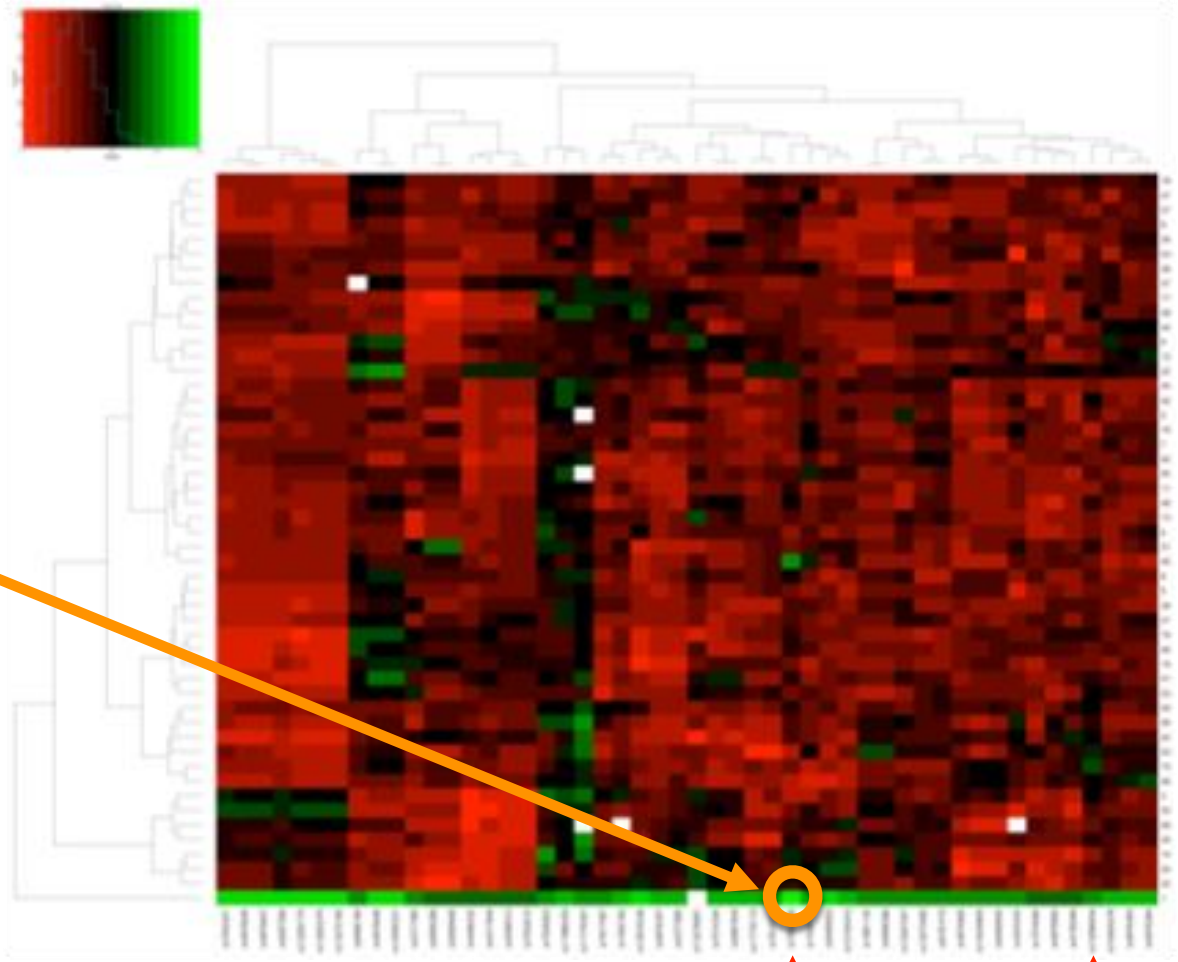
Genotype	affected			unaffected		

Calculate P value

Visualize P value based on genotype and environment



G x E interaction



Future Trends of Health Information System

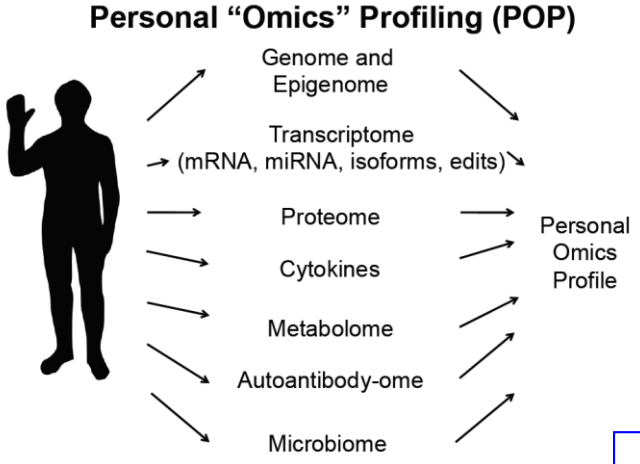
Two Major Trends

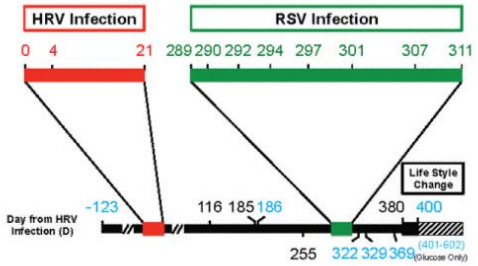
- Becoming Wider and Comprehensive
  - **Regional health information organization** (Regional HIE like MMWIN)
  - Information network to share the patient data, now about 150 RHIOs in Japan
  - Becomes wider: 2nd medical zone (cities) to 3rd medical zone (whole prefecture) to regional block of prefectures
  - Japanese government announce: **Until 2008, Regional Health Information Network cover the whole Japan**
- For nation-wide integration of RHIOs
  - "Minimum set of medical information items to share everywhere" (Japanese nation-wide EHR)
  - Nation-wide common healthcare ID for all the people

Two Major Trends

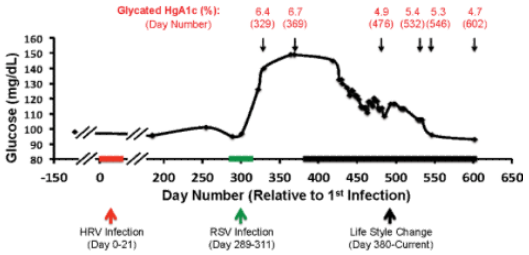
- Individualization (Personalization)
- Health information system must be individualized to support “Life-long healthcare”
- By two ways
  - (1) Omics biomarkers to find disease occurring beforehand
  - (2) daily life monitoring of physiological quantities

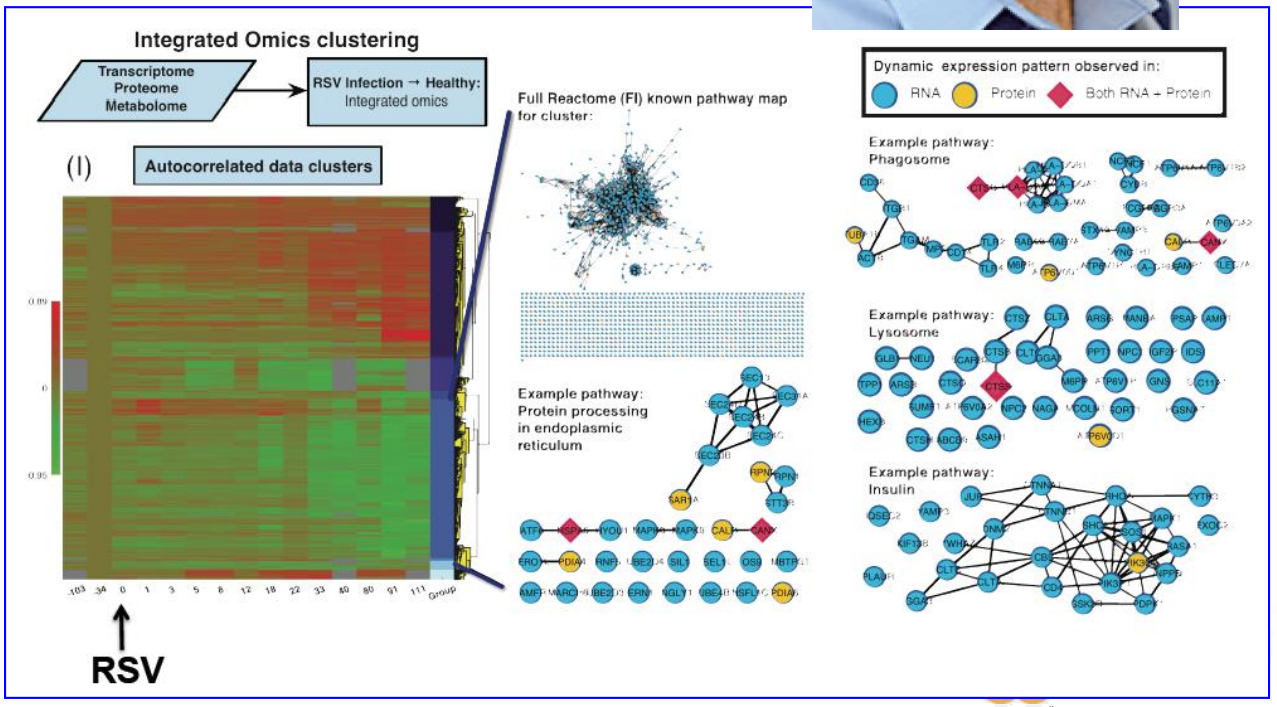
iPOP (integrated Personal Omics Profiling)





Fourier Transform to
Time series of omics profile
To detect perturbation for disease

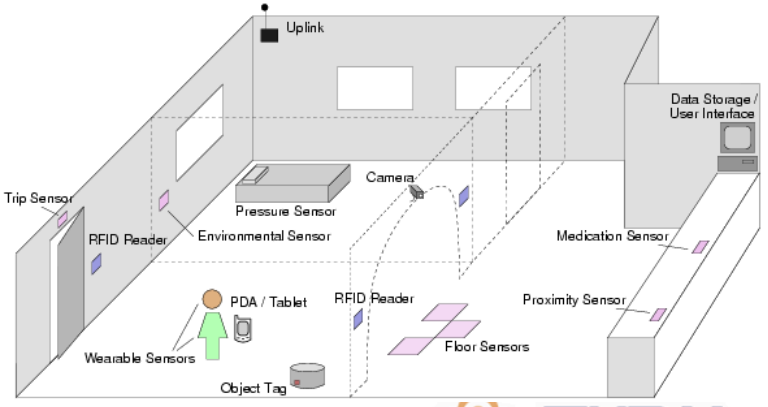




Life-long healthcare and PHR

- Quanfied Self
  - Movement from San Francisco area
  - Combination of wearable computer and monitoring sensors to observe physiological quantities
- COI Tohoku Univ. and Toshiba
  - collaborate to develop sensing
- mHealth (mobile healthcare)
  - Continua consortium & so on

ECG; EEG; Skin Conductivity; EVG



TMDU

Life-long (life-course) healthcare

- Goal of daily life monitering
  - Life-long healthcare: prediction of disease occurence
- Precognition of disease by molecular biomarker
  - preemptive medicine
  - Liquid biopsy
    circulating RNA,DNA, precognition of cancer, Alzheimer disease
  - DIY genomics
  - Integrated Personal Omics Profile

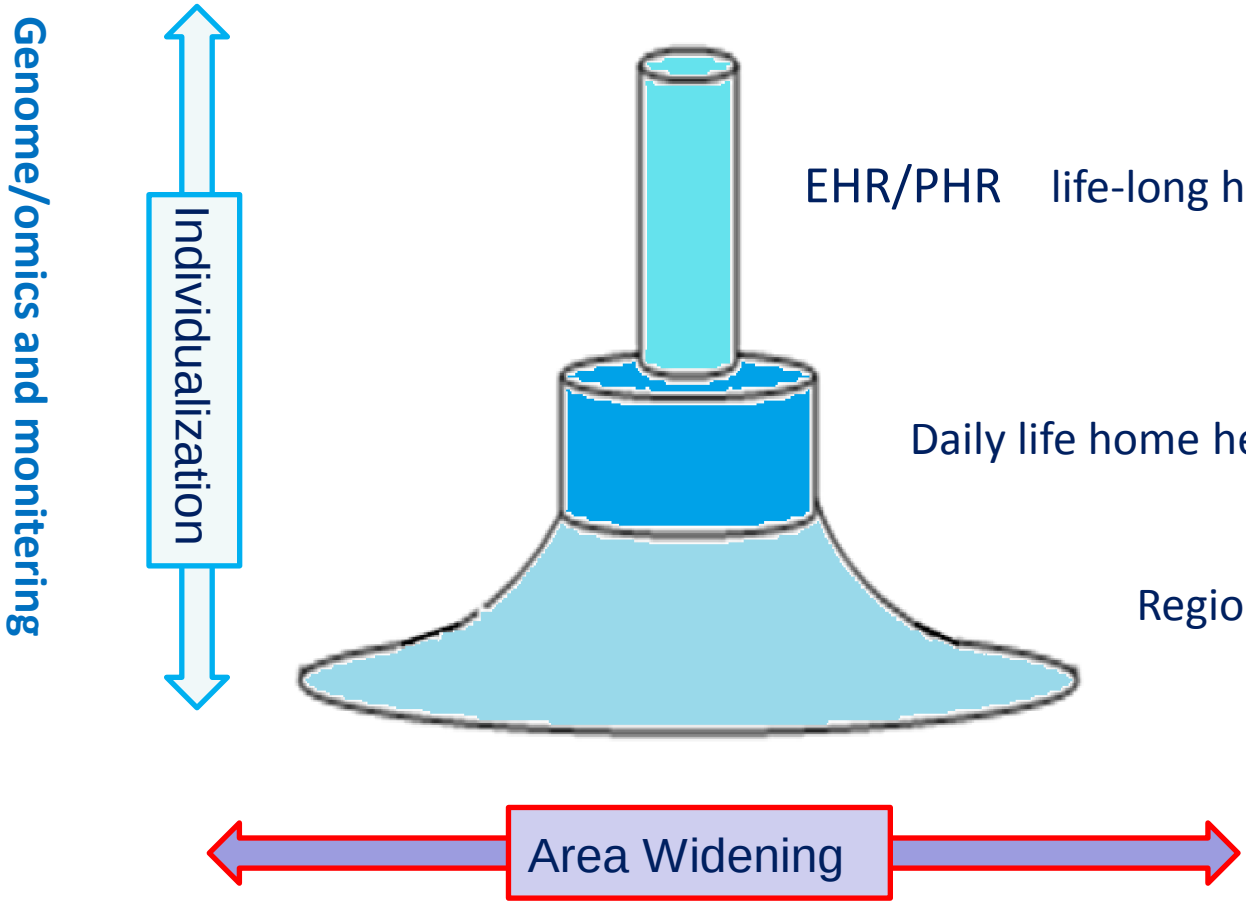
Nanopour sequencer

Preemptive medicine (Zerhouni 2005)
By making use of precise molecular knowledge to detect disease before symptoms are manifest, and intervening before disease can strike.

Reactive medicine → proactive medicine

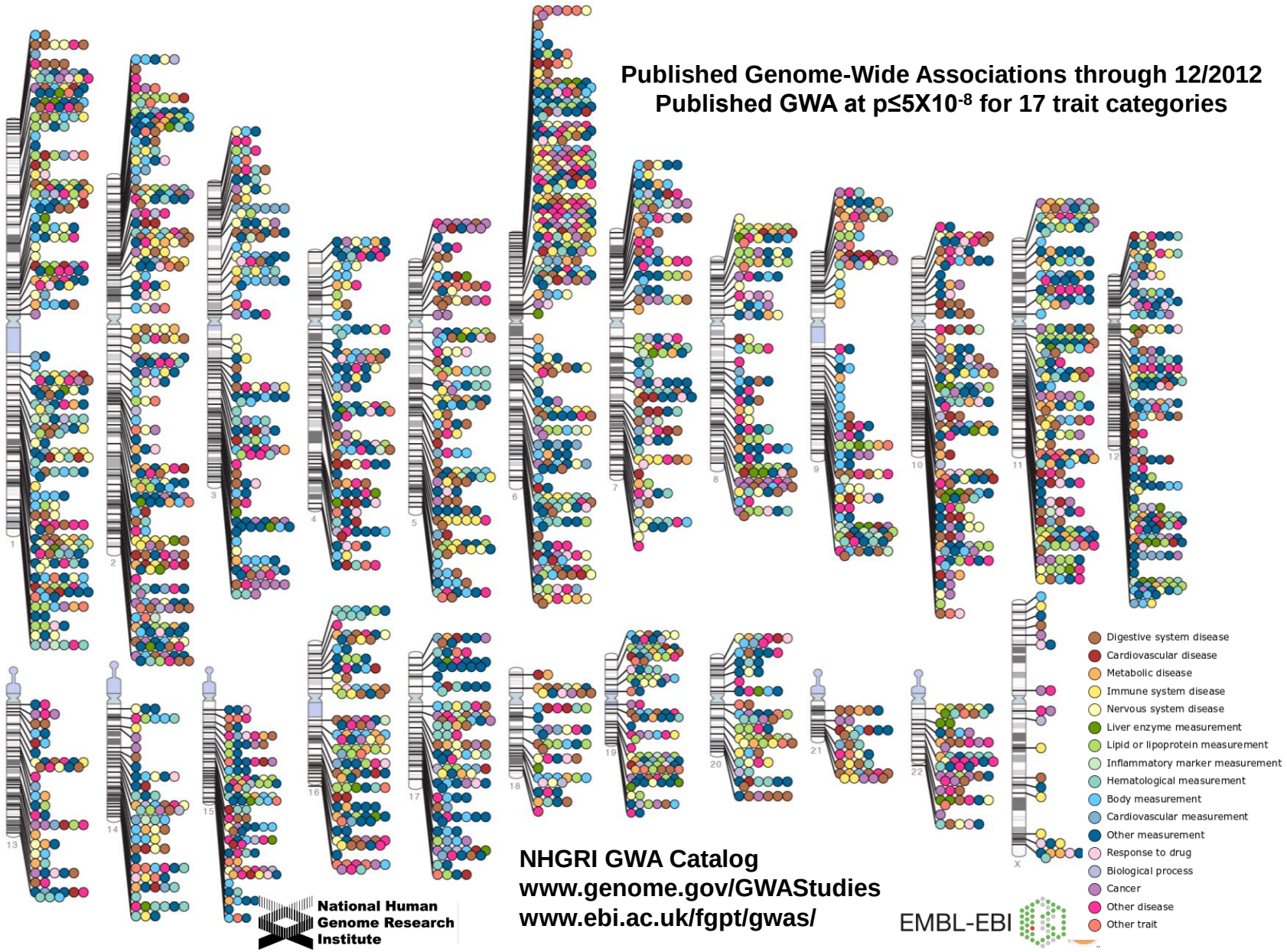
TMDU

Future of Health System



Thank you for kind attention

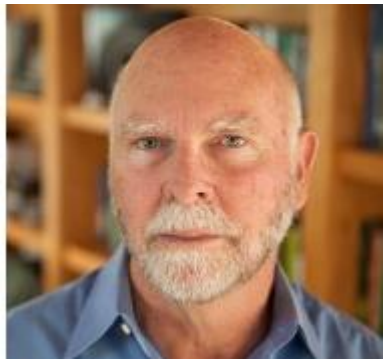
Disease Genes



Biobankとゲノムコホート

- **バイオバンクの目的･機能の変化**
  - 従来は再生医療ための生体標本や臨床研究の資料保存、
  - 近年はゲノム情報の収集が加わる
  - **ゲノム/オミックス個別化医療、創薬**の情報基盤
    - **疾患型BioBank**：全国的・全世界規模で症例の分子(ゲノム）情報とそれに対応できる臨床症例の収集。疾患ゲノムコホート. 臨床治験DBなども
  - **個別化予防**の情報基盤
    - **Population型BioBank**：前向きコホートで健常人の分子情報（ゲノム）と環境情報を集めて追跡するゲノム・コホート
- **最近の動向**
  - **UK biobank**：
    - 50万人の健常者の健診・血液を集め、
      その健康医療状況を追跡する
  - **BBMRI**（Biobank/Biomole. Res. Infra.）　BioVUの構成
    - 250以上の欧州BioBankを統合
- **わが国のBiobank計画**
  - 東北メディカルメガバンク、
    Biobank Japan, 6NC 疾患コホート等

米国民間ゲノムデータベース



- Craig Ventor “Human Longevity Inc.”
  - 健康・長寿（健康寿命伸長）
  - ゲノム科学、幹細胞治療 （Haririと共に）
  - 初期資本7000億円医療費削減、HiseqX 5sets
  - 一年40000ゲノム（幼児から老人まで、患者・健常者も）収集し最大のゲノムDBを作る、臨床情報も収集
  - 腸内細菌も含む　一日５人のヒト全ゲノム
  - がん（Moores Cancer Centerと提携)、糖尿病、認知症などの成人疾患に
  - Och F (機械学習の専門家）が加わる
- Google Xプロジェクト“Baseline”
  - 健康に関する尺度発見
  - Conrad AのもとにDuke大学やStanford大学が協力
  - 現在175名、先制医療的なバイオマーカ探し、今後拡大

TMDU

ゲノム・オミックス医療を支える情報システム

- 統合臨床オミックスデータベース
  - 各病院に装備するか、センター方式で病院群が共有するか、２つの方式がある
  - 地域医療連携と連動する方式も考えられる
  - Population（健常人）型 Biobankとも連携が考えられるが、暫くは疾患コホートBiobankとの連携の方が容易である。
- 臨床表現型とゲノム・オミックス情報の相関関係の表現方式が重要である

TMDU