Medical Big Data and Knowledge Discovery

Tokyo Medical and Dental University
Dept. Bioinformatics, Medical Research Institute
Hiroshi Tanaka

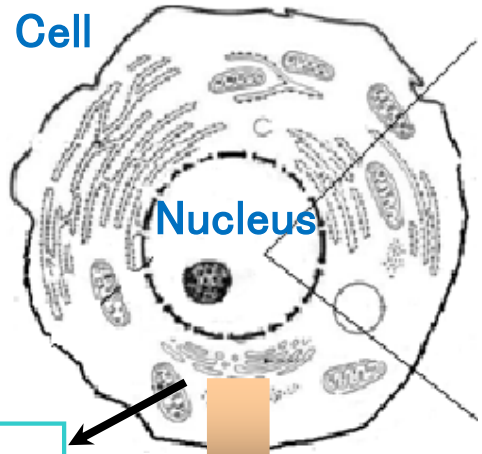
TMDU

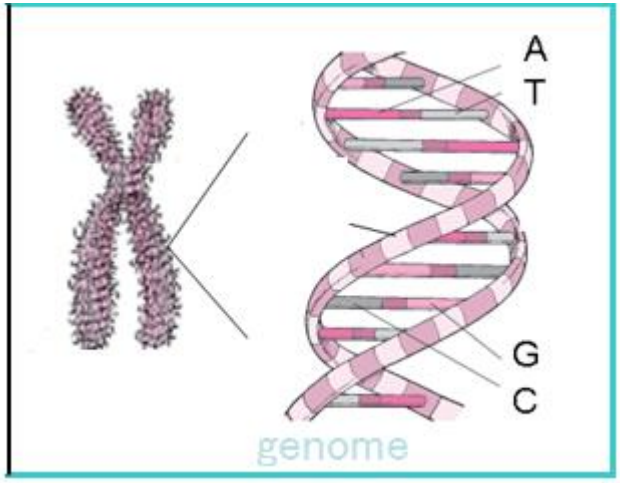
Basis of Genome/Omics

TMDU

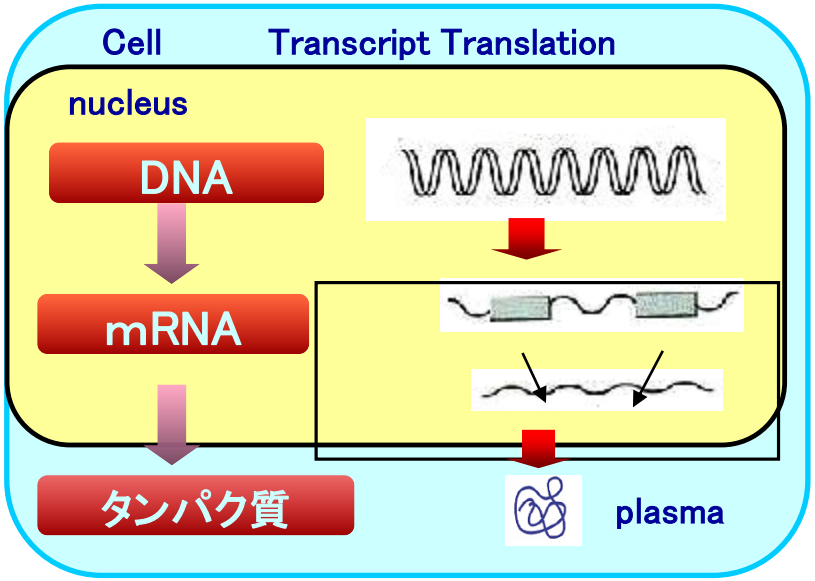
Genome and Omics

1990 Human Genome Project
2003 HGP finished





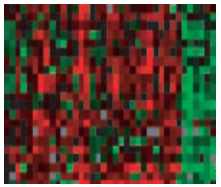
Omics
＝ −Ome（Whole）＋−ics（study）



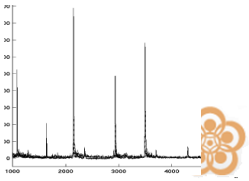
genome
genomics
transcriptome
transcriptomics
proteome
proteomics
metabolome
metabolomics

sequencer

DNA microarray



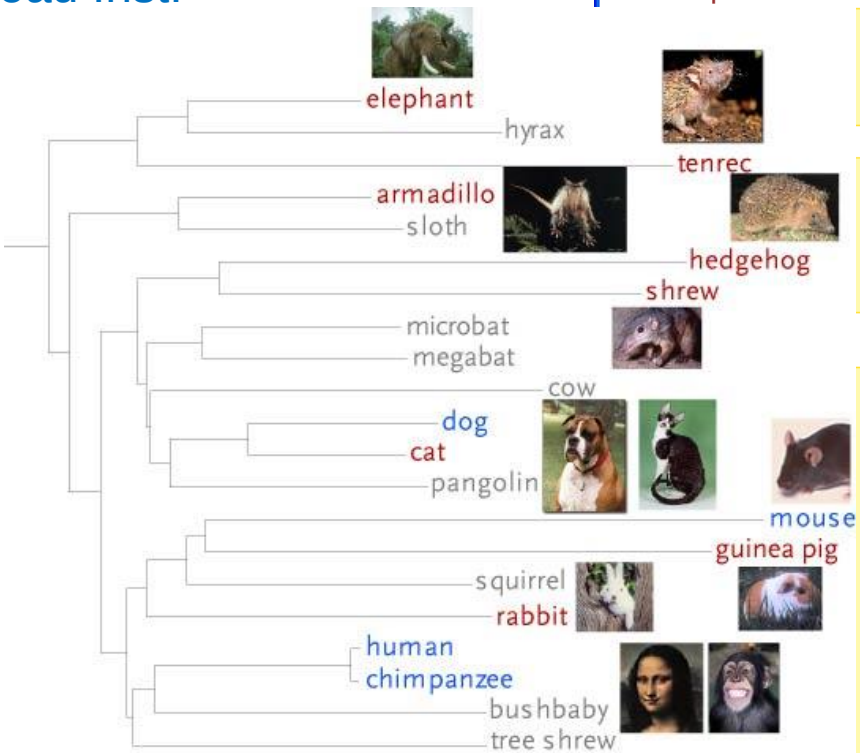
Mass Spect



TMDU

Genome of
about 1000 organisms

Broad Inst.



Ensembl Genome Browser – Mozilla Firefox

ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)

e! Ensembl

Ensembl

Ensembl release 43 - Feb 2007

HOME · BLAST · BIOMART · SITEMAP HELP

Your Ensembl

- Login or Register
- About User Accounts

Help & Documentation

- Table of Contents
- Helpdesk
- What's New
- About Ensembl
- Downloading data
- Displaying your own data
- Ensembl software

Select a species

Search Ensembl

Search: All species for Go

e.g. mouse chromosome 2 or rat X:10000..20000 or human gene BRCA2

Ensembl tools

Start a sequence search → Search Ensembl for nucleotide and peptide sequences with BLAST and SSAHA.

Mine Ensembl with BioMart → Cross-reference Ensembl datasets with BioMart, a powerful data-mining tool.

Customise Your Ensembl → Register with Ensembl to bookmark your favourite pages, customise your home page and much more!

Fetch data with the Ensembl API → Learn how to extract data from the public Ensembl database with this tutorial.

Ensembl headlines: Release 43 (February 2007)

- New Cow assembly and genebuild (*Bos taurus*)
- New species - Cat (*Felis catus*)
- New species - Tree shrew (*Tupaia belangeri*)
- New species - Hedgehog (*Erinaceus europaeus*)
- New species - Guinea Pig (*Cavia porcellus*)

More news...

Log in to see customised news · Register

Ensembl 43 | Pre! species

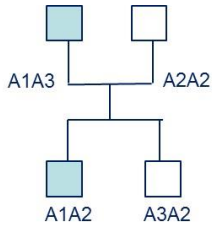
Popular genomes

- *Homo sapiens* NCBI 36 | Vega
- *Mus musculus* NCBI m36 | Vega
- *Danio rerio* Zv6 | Vega

More genomes

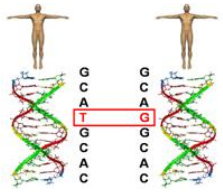
- *Aedes aegypti* AaegL1
- *Anopheles gambiae* AgamP3
- *Bos taurus* Btau_3.1 UPDATED!
- *Caenorhabditis elegans* WS160
- *Canis familiaris* CanFam 2.0
- *Cavia porcellus* cavPor2 NEW!
- *Ciona intestinalis* JGI 2
- *Ciona savignyi* CSAV 2.0
- *Dasypus novemcinctus* ARMA
- *Drosophila melanogaster* BDGP 4.3
- *Echinops telfairi* TENREC
- *Erinaceus europaeus* eriEur1 NEW!
- *Felis catus* CAT NEW!
- *Gallus gallus* WASHUC2
- *Gasterosteus aculeatus* BROAD S1
- *Loxodonta africana* BROAD E1
- *Macaca mulatta* MMUL 1.0
- *Monodelphis domestica* MonDom 4.0
- *Ornithorhynchus anatinus* Ona-5.0
- *Oryctolagus cuniculus* RABBIT
- *Oryzias latipes* HdrR

How to use genome and omics in 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～)

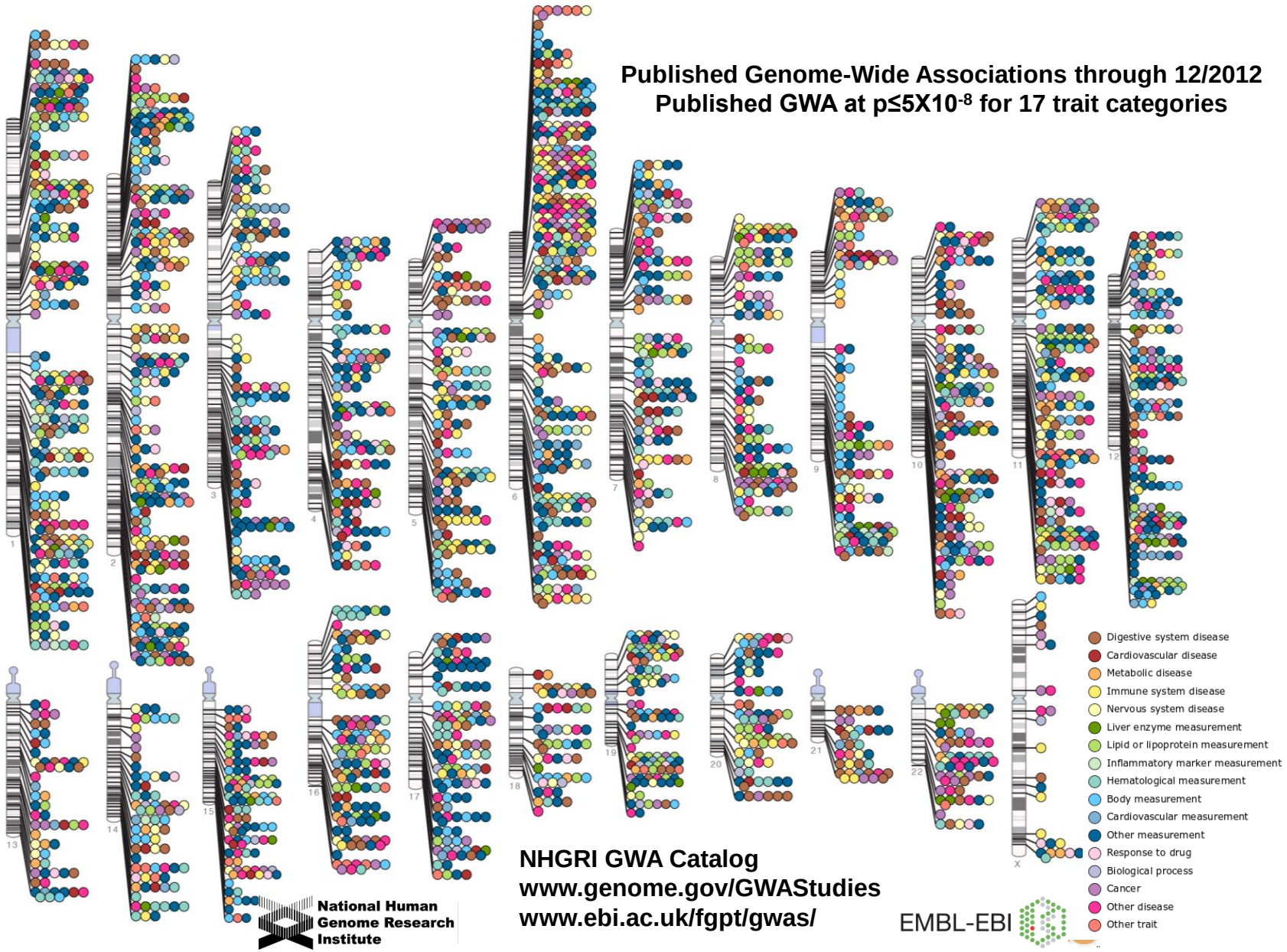
- Based on and **direct use** of **"acquired omics profile"**
- Aiming at **"Predictive/Preemptive medicine"**

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

Disease Genes



Genome Omics Medicine and medical Big Data

TMDU

The second genome revolution

Next generation sequencer

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

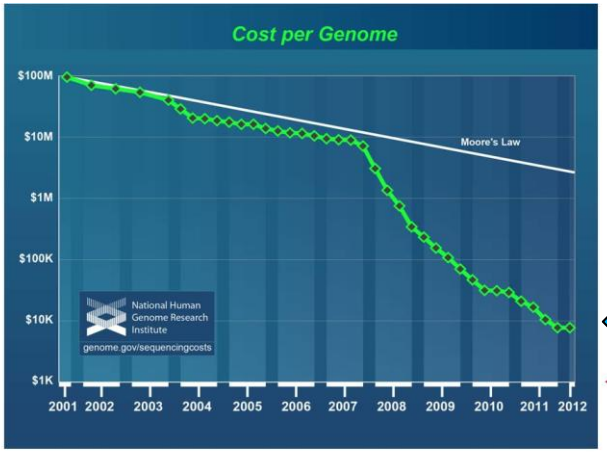
Ilumina 2500

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

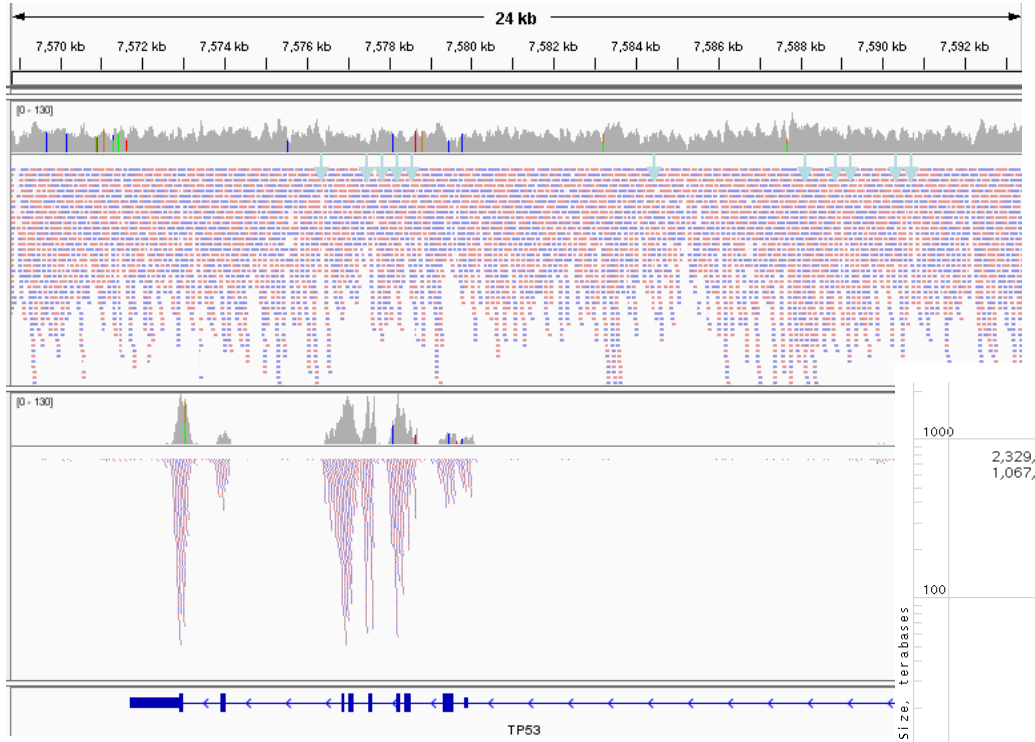
1000 dollar NGS
Ilumina Hiseq X (10set)

TMDU

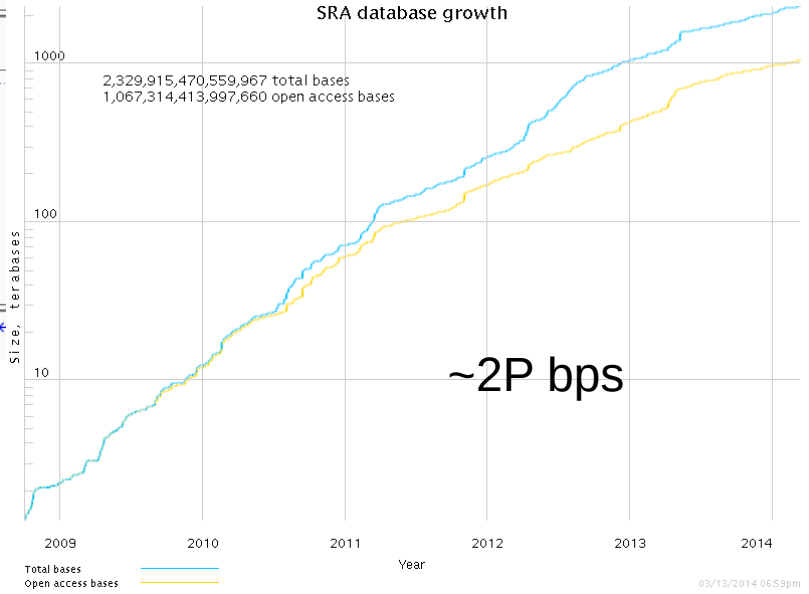
Sequence data

WGS

WES



(p53: 17chr p13.1)



~2P bps

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

TMDU

Rapid Spread of Genome/Omics medicine

TMDU

Clinical Implementation of 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

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 ctr
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

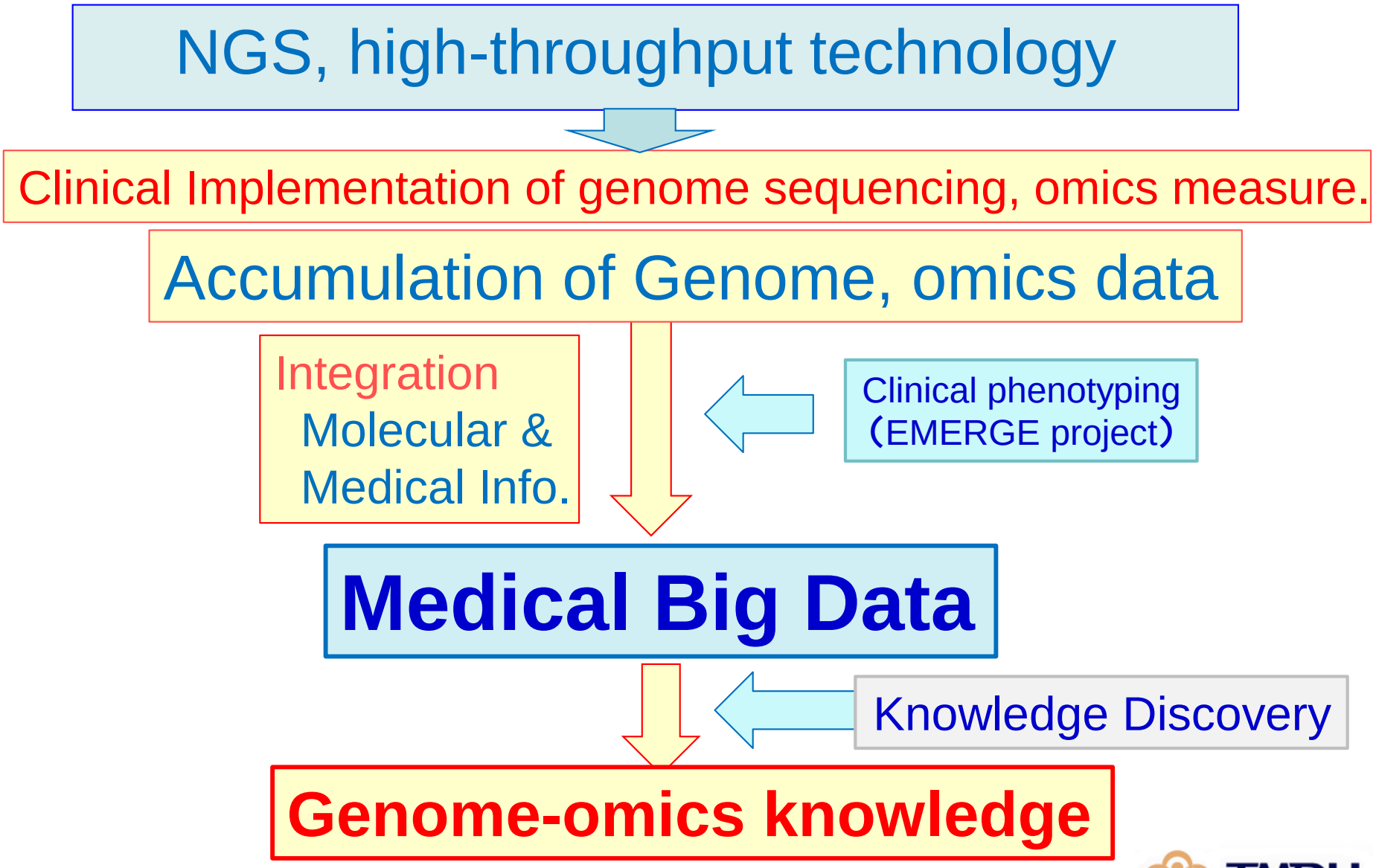
Whole genome laboratory In-house, Seq

Vanderbilt

EMR

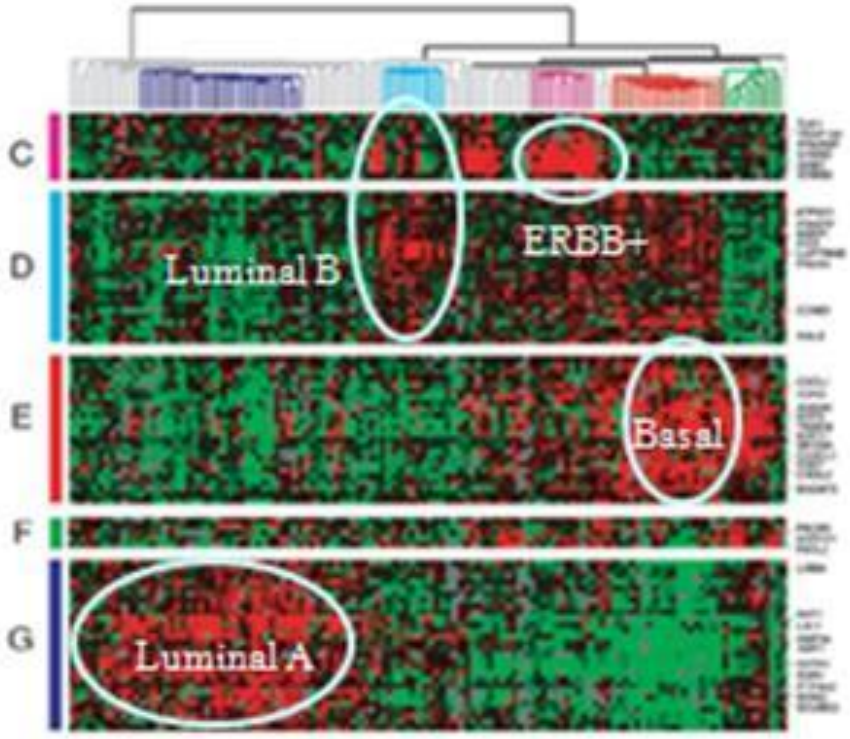
TMDU

Genome omics medicine and Big Data



TMDU

Omics measurement



Breast Cancer
Intrinsic classification

Prediction tool
mammaprint (70 genes)
oncotype D (25 genes)

intrinsic分類	ER	PgR	HER2	予後
Luminal A	ER(+) and/or PgR(+)		(−)	予後良好
Luminal B	ER(+) and/or PgR(+)		(+)*	A型より不良
HER2 enriched	(−)	(−)	(+)	予後不良
Basl-like	(−)	(−)	(−)	予後不良

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

Medical Big Data

Big Data for
Healthcare, Drug Discovery

- Healthcare, Medicine
  - Personalized Realization of Genome omics medicine・Precision Med
  - Large scale Biobank, disease cohort
  - Personalized Prevention: population cohort
- Drug Discovery
    - Drug discovery -- Precompetitive
    - Drug reposioning
  - In silico screening

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 と 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

Discovery in Drug Repositioning

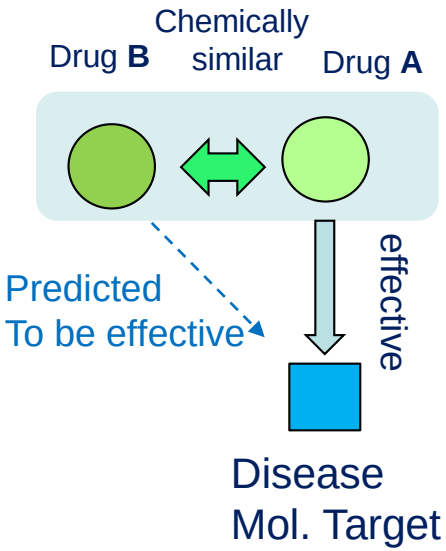
TMDU

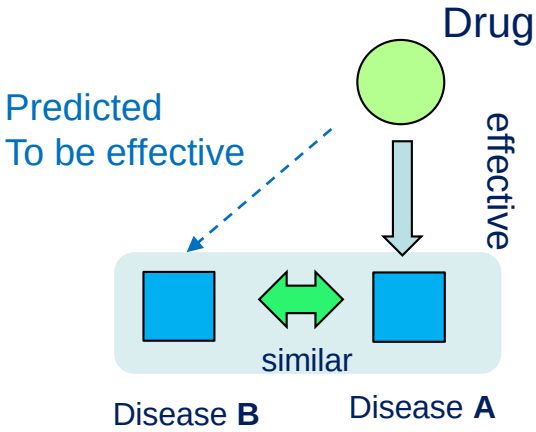
Needs for Drug Repositioning

- R&D expense increasing but N of drug decreasing
- Drug repositioning:
  - Other names, drug repurposing, drug re-profiling, therapeutic switching and drug re-tasking
  - the application of known drugs and compounds to new indications (i.e., new diseases).
- To use already approved drug
  - Safety and toxicity is already confirmed
  - Low cost and faster development

computational DR: two approaches

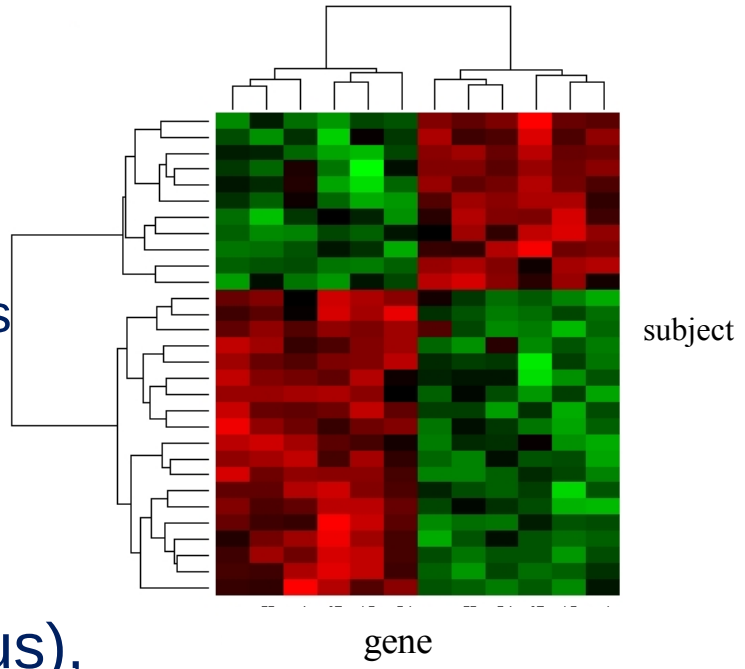
- Drug-based (drug-centric)
  - Based on similarity of Chemical function and characteristics
    - ① Chemical structure similarity
    - ② Gene expression profile (GEP) when drug is administered
- Disease-based(disease-centric)
  - Based on disease similarity
    - ① Sharing of disease causative genes
    - ② Similarity of GEP
- Fusion of the above two





GEP omics utilization

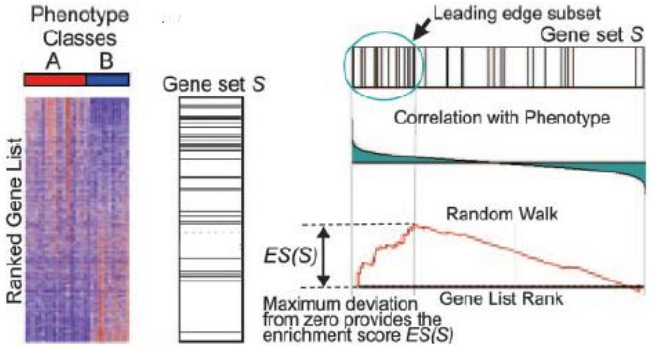
- Drug-induced GEP or Siginficantly Differential Expression (SDE)
  - CMAP：Connectivity Map
    - Broad Institute,1309 chimicals,
    - MCF7,PC5 5 Cell-line, 7000 GEPs
    - Signatures
    - DB：query Sig, order rank
- Disease-associated SDE
  - GEO（gene expression omnibus),
    - NCBI　25000 experiments,
    - 70000 GEPs
    - ArrayExpressEBI



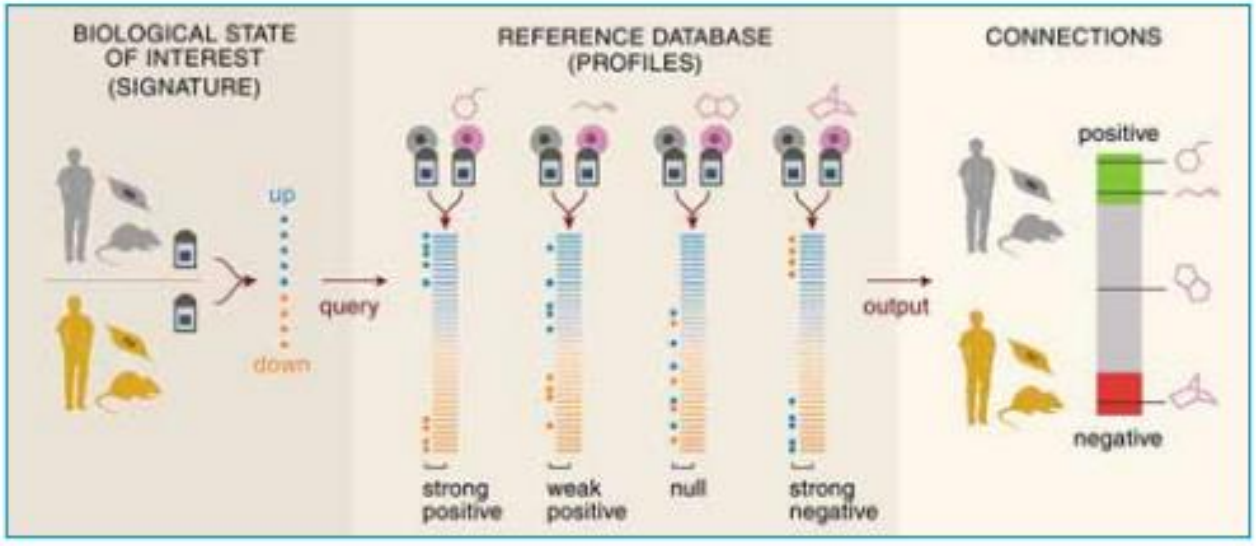
DR based on GEP (1)

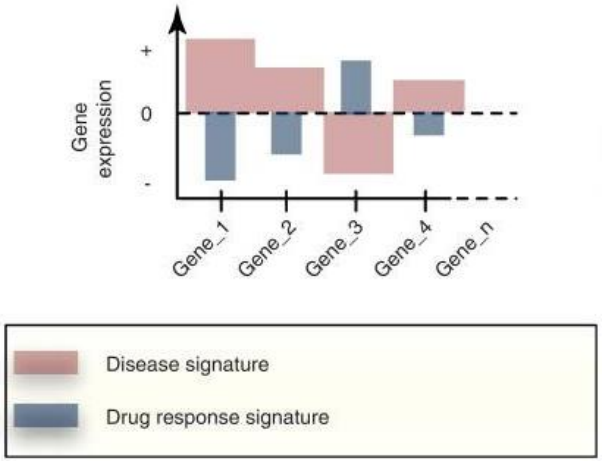
signature reversion method

- Drug-specific GEP signature
- Disease-specific GEP signature
- Negatively correlated
- Non-parametric correlation coeffiecient
  - Gene Set Enrichment Analysis（GSEA）: ES score
- Example IBD⊂anti-convulsant topiramate,



GSEA

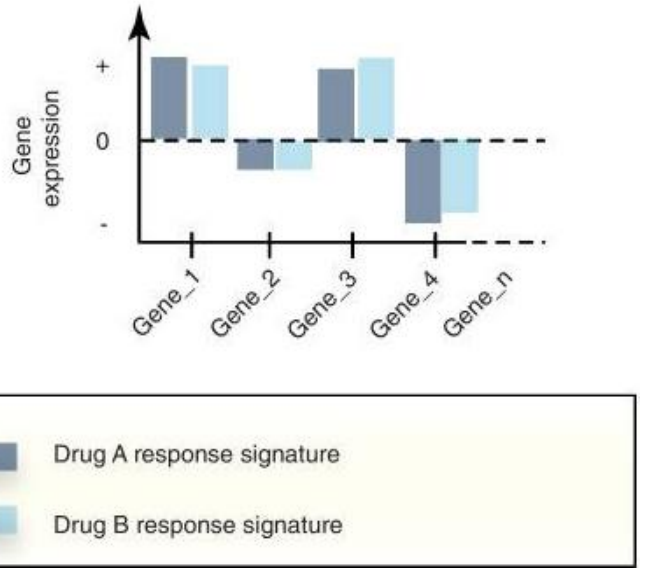




TMDU

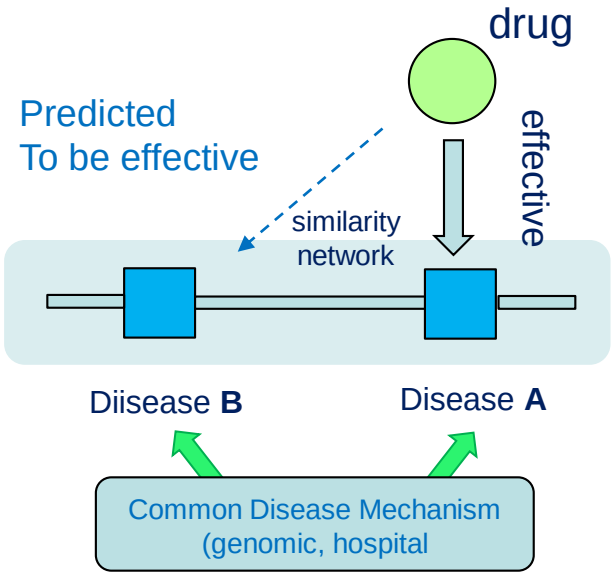
DR based on GEP (2)

- guilt-by-association :
- Drug-drug
  - Connectivity map
    similar GEP drug estimated by GSE
  - Also search for neibourhood
  - Antimalaria clone disease
- Drug-disease
  - Drug specific SNG Disease specific S
    SNG similar
  - Non-parametric correlation positive
  - Toxicity and side effect possibility inc



DR based on disease network

- System of disease classification: nosology
  - Phenotypic classification of disease by Linne, more than 300 years
  - Disease classification base on the difference of Disease Occurring Mechanism by Genomics-Omics
  - Disease Network: similarity network among diseases
  - Which genomic or omic mechanism is adopted



疾患の成立機序における主要機序

- 疾患関連遺伝子型（第一世代型）
  - 原因遺伝子、疾患感受性遺伝子の変異・多型が主要発症機序
- 疾患オミックス型（第２世代型）
  - 疾患オミックスプロファイルの変容が主要発症機序
  - Transdisease omics
- 疾患分子ネットワーク型（第３世代型）
  - 分子ネットワークの歪みが主要発症機序
  - がんなどで遺伝子型（肺腺がん等）でない通常のがん

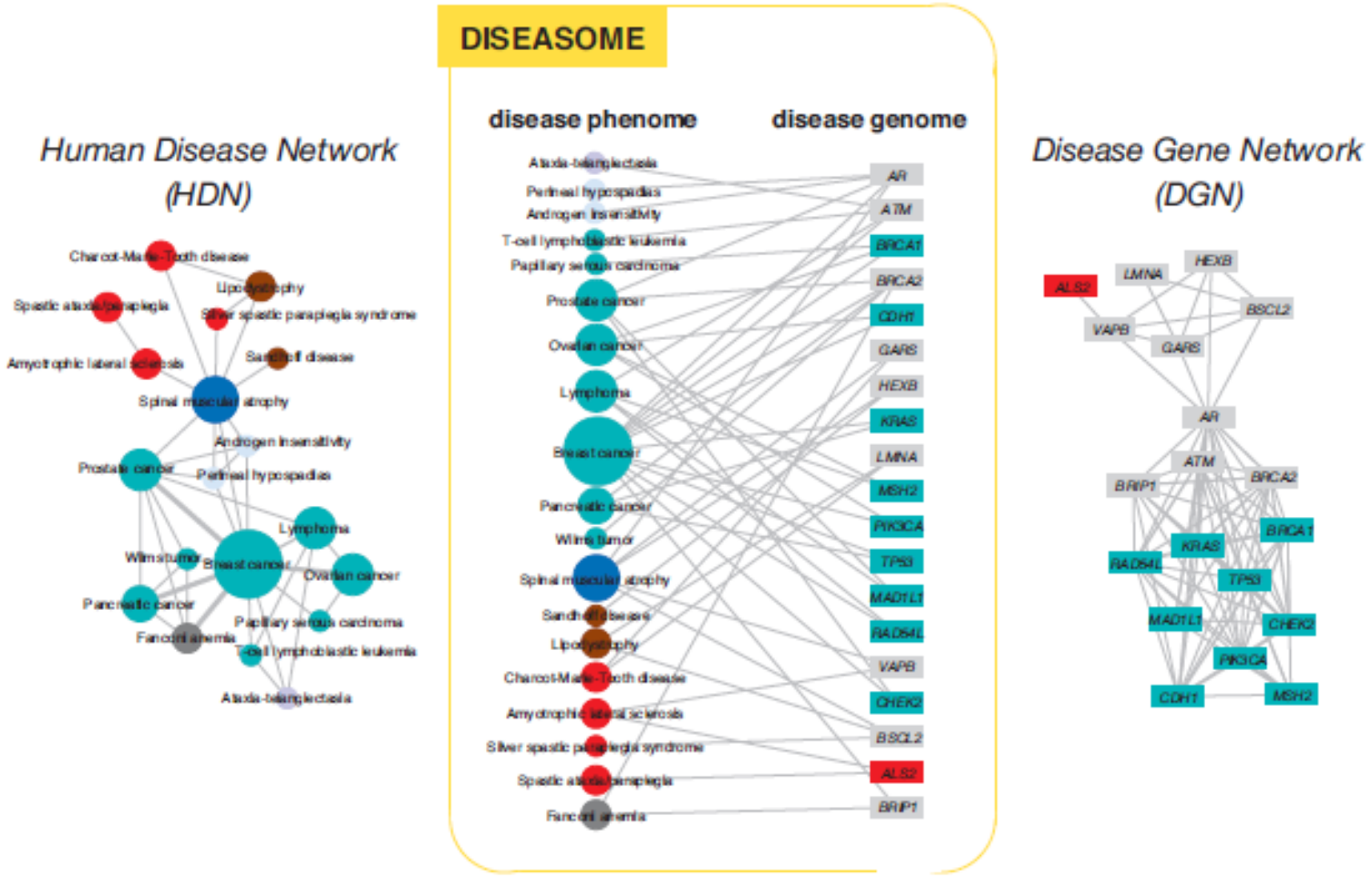
The first generation type
Diseasome and Disease Gene

- OMIM 1,284 diseases and 1,777 disease gene
- Human disease network（HDN)
  - 867 diseases connected to other disease
  - 516 diseases form a gigantic cluster
    - Hub colon cancer, breast cancer
    - Cancer connected through P53 ,PTEN to most strongly connected
    - Not influenced by organ or pathological phenotype
  - Overcome the convential phenotypic classification
- Disease Gene network（DGN)
  - 1377 genes connected other genes
  - 903 genes form a gigantic cluster
    - P53がハブ
- Comparison with random network
  - Size of gigantic cluster is significantly small
- Disease genes module
  - Expressed in same tissue

TMDU

Diseasome

(Goh,Barabasi et al.)



Disease which have more than one gene share

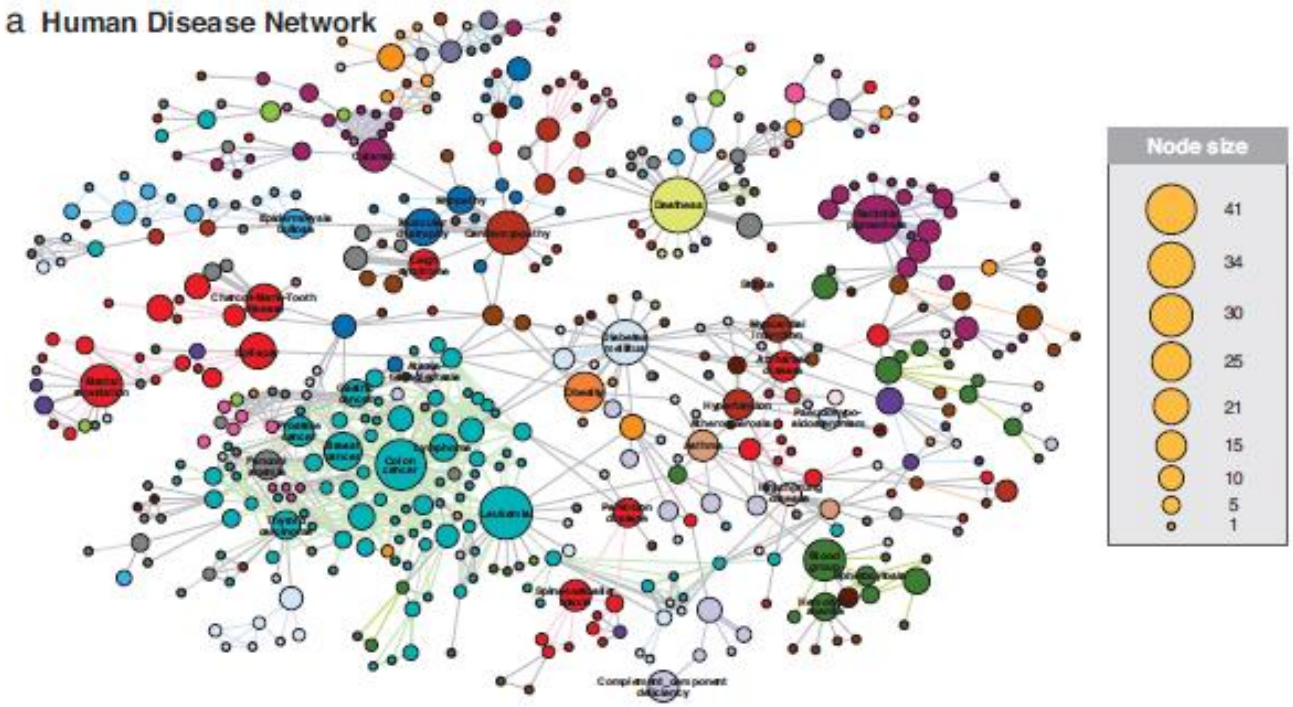
Disease gene which has shared

Kwang-Il Goh*, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi The human disease network PNAS2007

TMDU

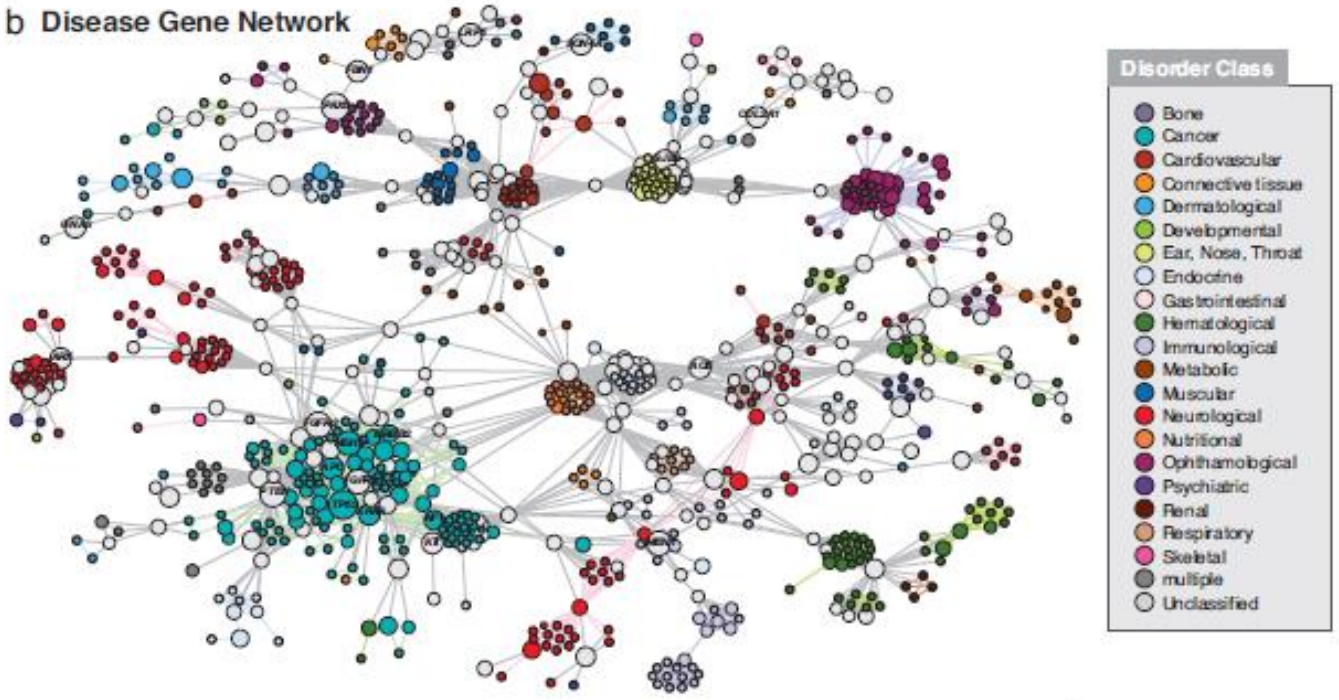
HDN

Node **size**
N causative genes
Link thickness
N shared genes among diseases



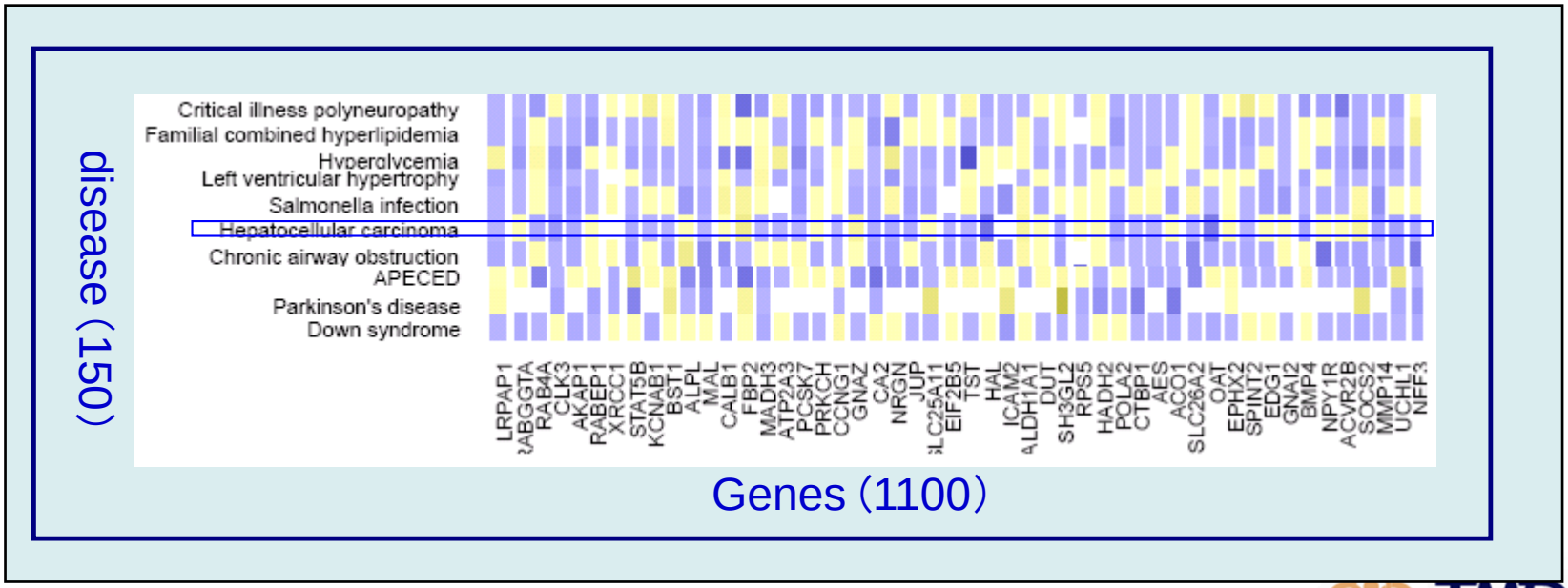
DGN

Node size
Nr cansative of genes for the disease に比例



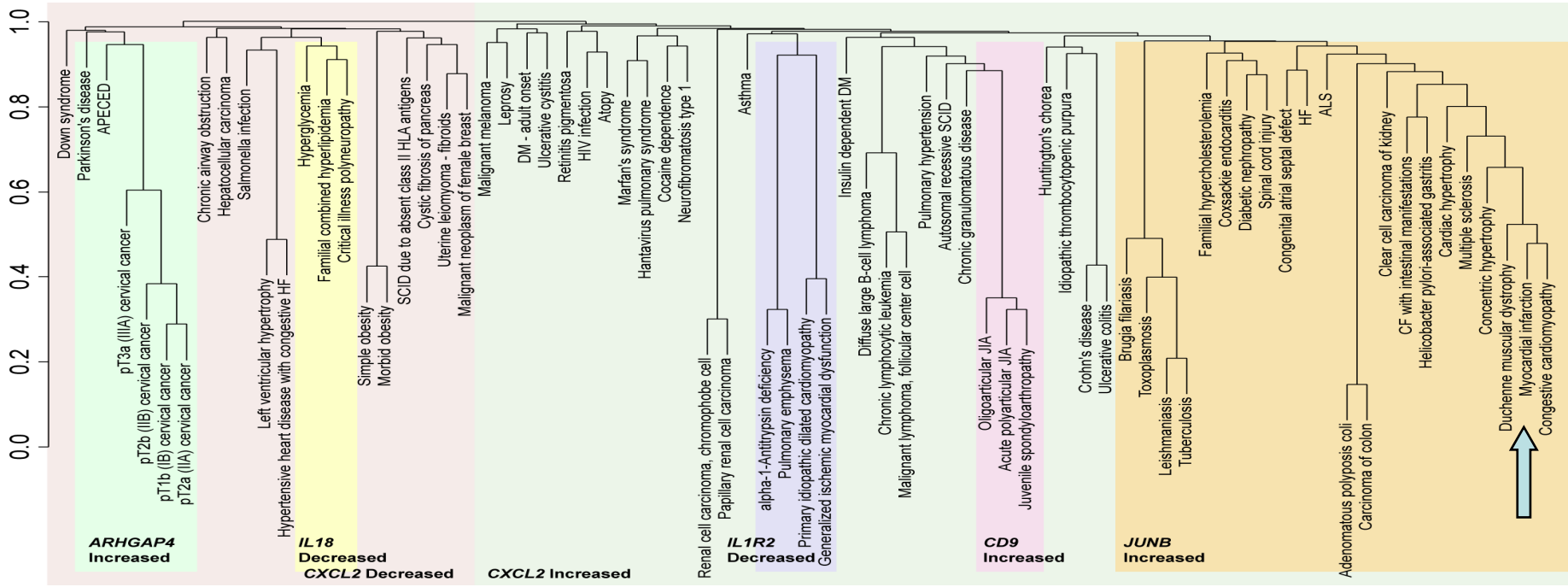
The second generation type **GENOMED** (A.Butte et al)

- Use of GEP DBのGEO (Gene Expression Ominibus)
  - 700000 samples
- To obtain average GEP for diseases



Gene-Expression Nosology of Medicine

- Cluster applied of GEP of diseases
  - Unexpected results not predicted byt conventional organ based classification
  - Cytokine, receptor oriented classification
- Myocardial infarction and Duchenne type dystrophy very close

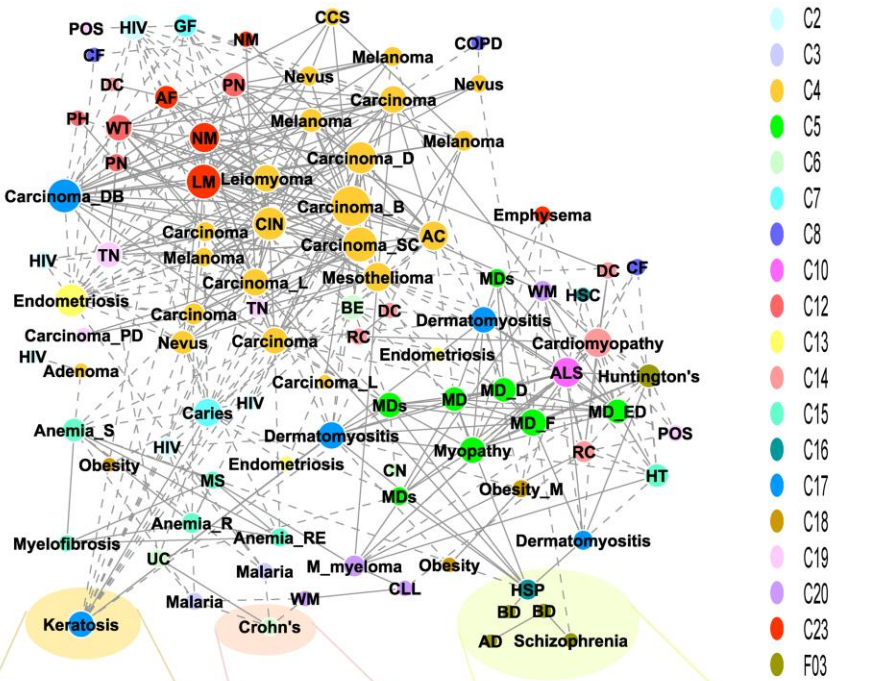


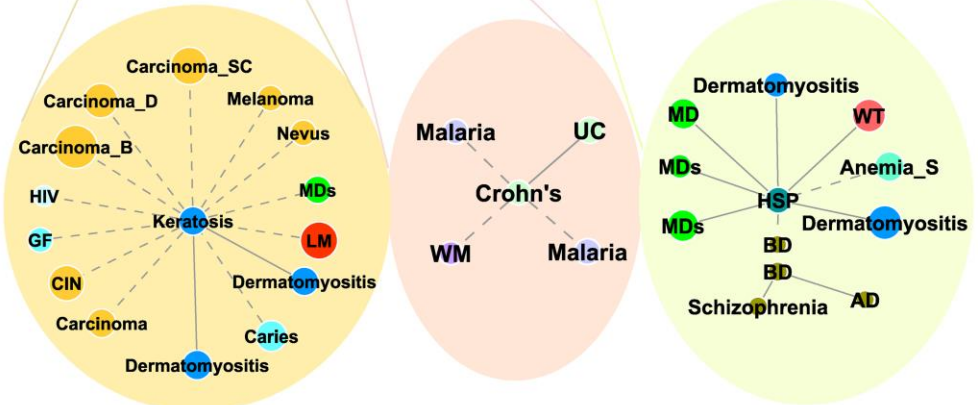
Transcriptional Profiling による疾患ネットワーク

Hu, Agarwal 遺伝子発現プロファイルとGSEA関連尺度によるリンク

疾患（disease-disease）645 nodes
疾患- 薬（disease-drug）5008 pairs

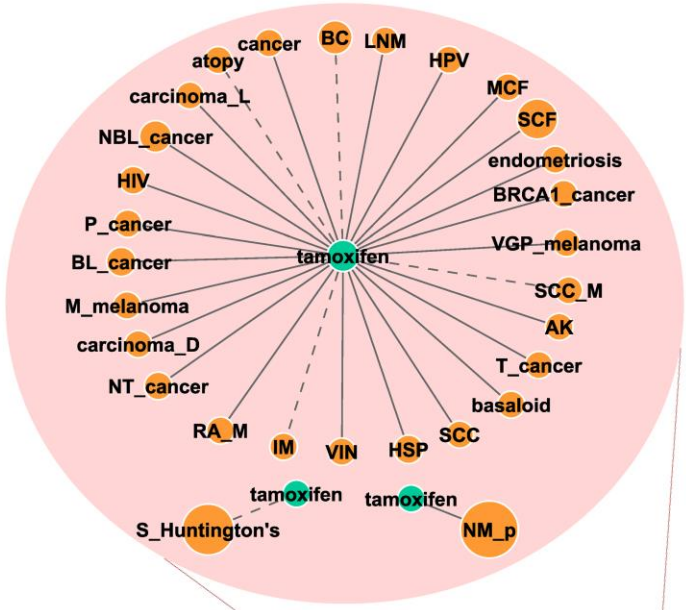
Solar keratosis 日光性角化症
⇒ cancer(squamous)
Crohn’s disease
⇒　マラリア
Hereditary Spastic Paraplegia
（遺伝性痙攣性対麻痺）
⇒bipolar双極性うつ病





カラーはMeSH
同一カテゴリー

Transcriptional Profiling



Disease-drug network

orange drug

green disease

Tamoxifen (breast cancer)

Negative

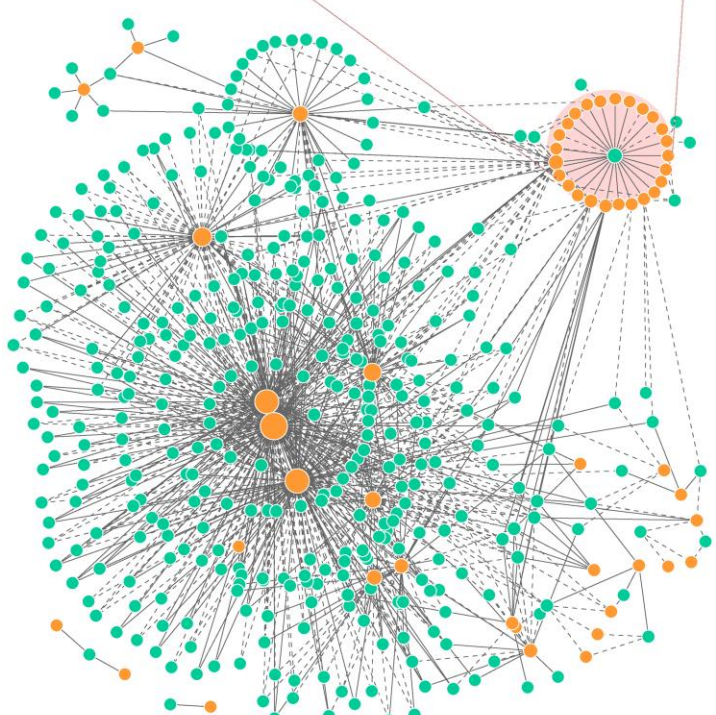
⇒ atopy

⇒mast cell increase、allergy supress

positive

Side efect

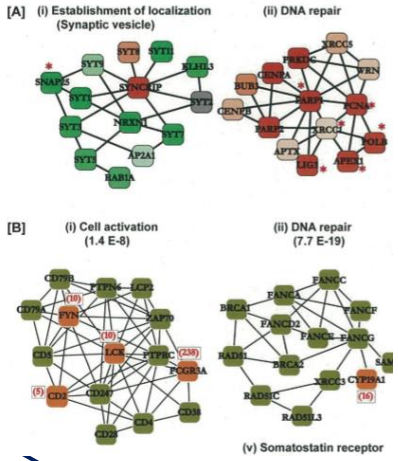
⇒ carcinogenic

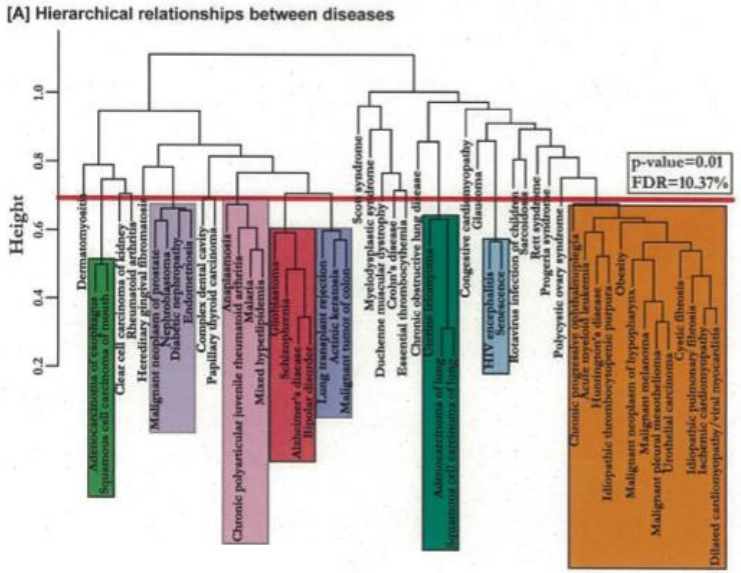


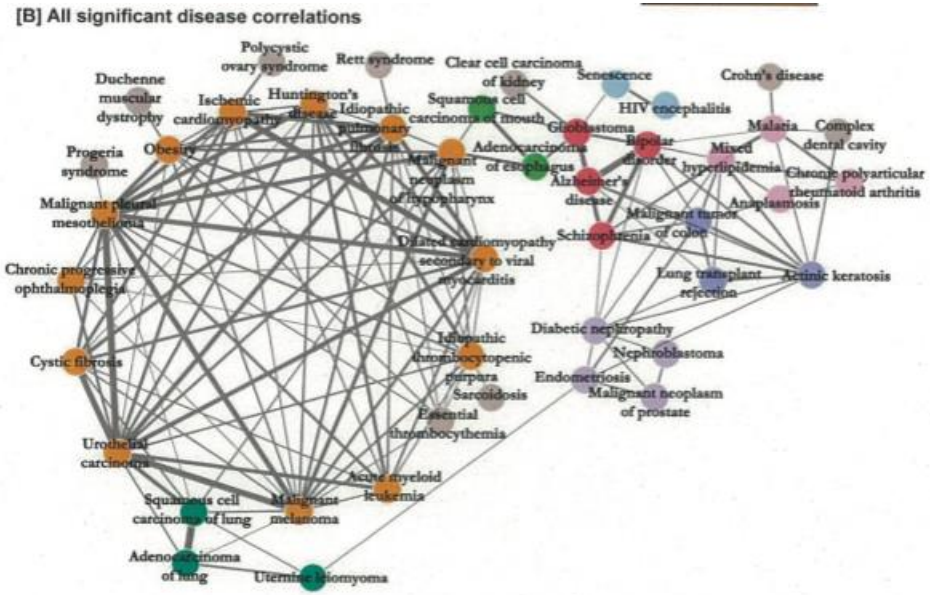
ご清聴ありがとうございました

Transcriptomeの変化をPPIに投影した疾患ネットワーク（Butte)

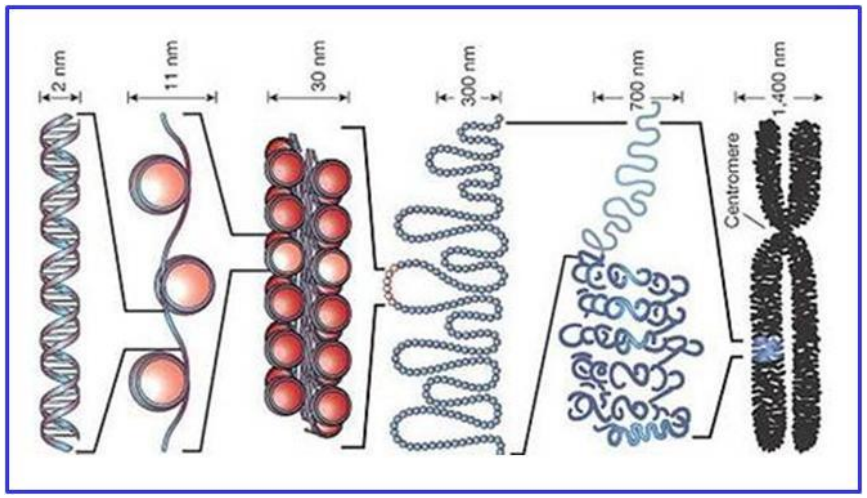
- 遺伝子発現プロファイルを直接使うのではなく
  4620Moduleに分解したPPIネットワークでの
  疾病での平均発現変化をつかう
  - 遺伝子発現プロファイルより疾病によって
    変化するmoduleを調べる　病気に対するPPIの応
    マラリアとクローン病
  - moduleの遺伝子の変化を平均して遺伝子の代わりに
    moduleの発現平均スコアを用いる
- 疾患の大半を占める＜共通疾患状態シグネチャア＞

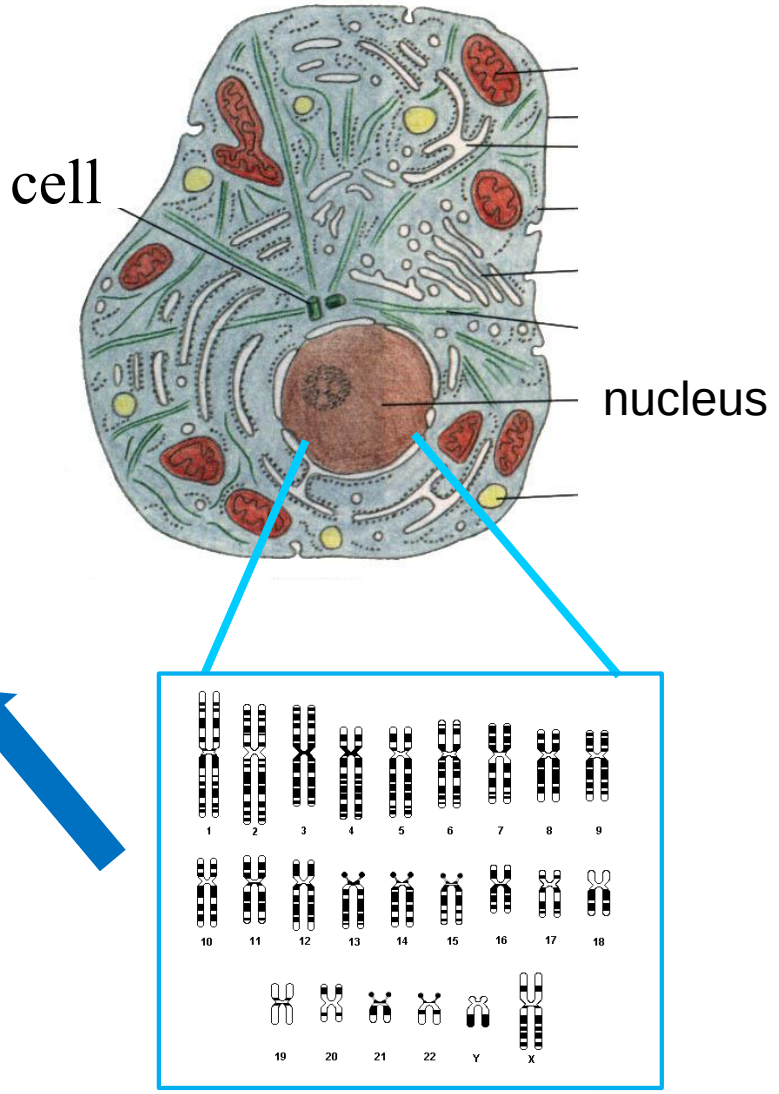




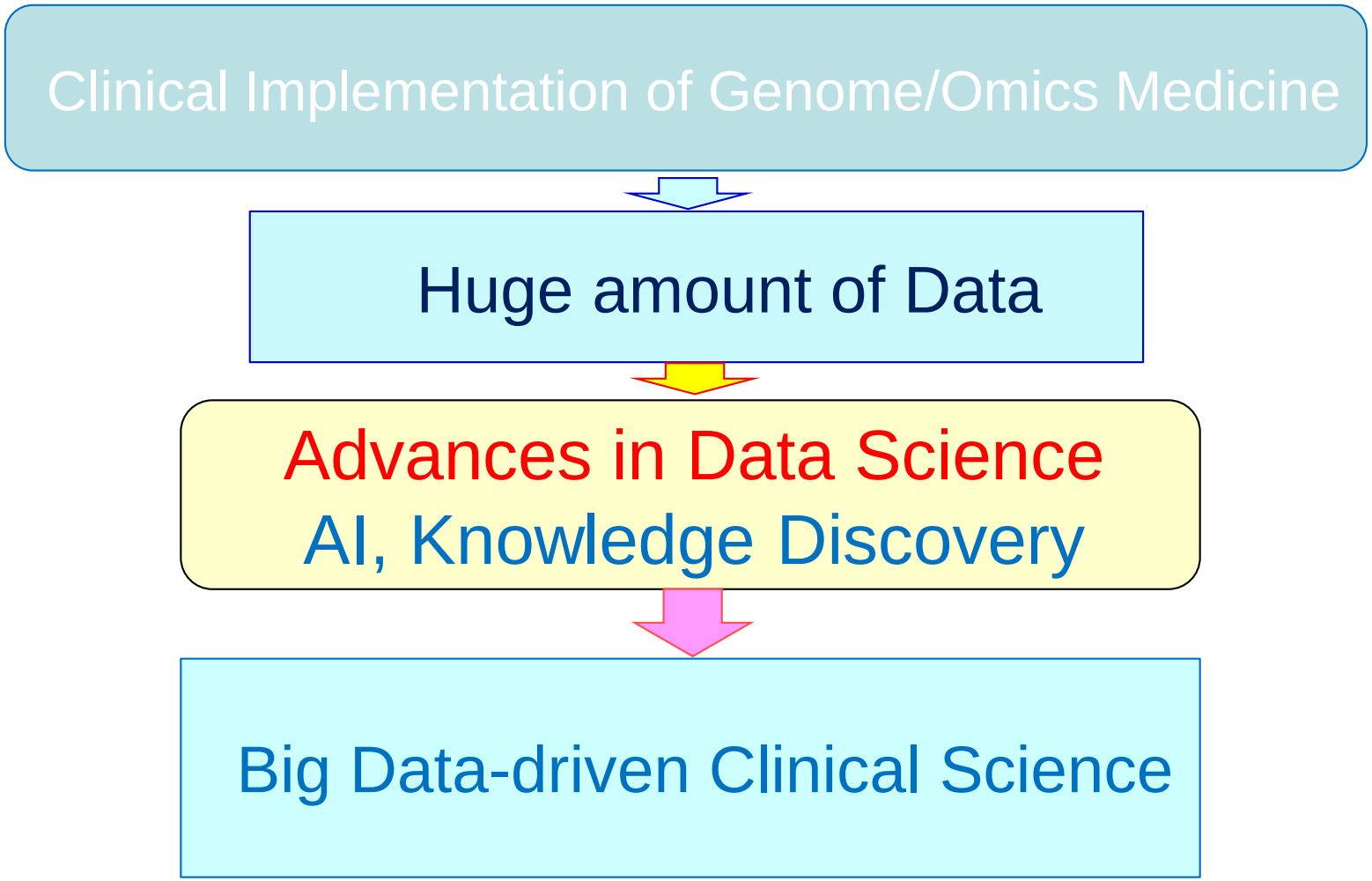


Genome Omics





Genome Omics Medicine and medical Big Data



TMDU