Medical Big Data and Knowledge Discovery

Tokyo Medical and Dental University
Dept. Bioinformatics, Medical Research Institute
Hiroshi Tanaka
Basis of Genome/Omics
Genome and Omics

1990 Human Genome Project
2003 HGP finished

Omics = Ome (Whole) + ics (study)

Cell

Transcript Translation

nucleus

DNA

mRNA

タンパク質

plasma

Nucleus

genotype

gene
genomics

transcriptome

transcriptomics

proteome

proteomics

metabolome

metabolomics

sequencer

DNA microarray

Mass Spect

TMDU
Genome of about 1000 organisms

Broad Inst.
How to use genome and omics in medicine

1st generation “Genomic Medicine” (1990～)

- Human genome ～0.5% different, mutation/polymerorphism, 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”

• 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
Disease Genes

Published Genome-WideAssociations through 12/2012
Published GWA at p≤5X10^-8 for 17 trait categories

NHGRI GWA Catalog
www.genome.gov/GWASTudies
www.ebi.ac.uk/fgpt/gwas/
Genome Omics Medicine and medical Big Data
The second genome revolution

Next generation sequencer
13 years ⇒ 1 day, 350 B dollar ⇒ 1000 dollar

Ilumina 2500
WGS (Whole genome sequencing)
3GB (1 person) × 30 = about 100 Gbps
1 person WGS 27 hours

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

Ion Torrent

1000 dollar NGS
Ilumina Hiseq X (10 set)

DNA Sequencing Cost: the National Human Genome Research Institute
Sequence data

(p53: 17chr p13.1)

NCBI Sequence Read Archive (SRA)
Rapid Spread of Genome/Omics medicine
Clinical Implementation of Genome/Omics Medicine

More than 20 hospitals have implemented Genome/Omics medicine

<table>
<thead>
<tr>
<th>Institution</th>
<th>Major Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC Wisconsin</td>
<td>Using whole genome sequencing to establish diagnosis in patients with currently undiagnosed genetic disorders</td>
</tr>
</tbody>
</table>
| 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  
                          • SLC01B1*5 genotyping and statin adherence  
                          • Effect of genetic risk info on anxiety and adherence in T2DM |
| 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 College
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
Genome omics medicine and Big Data

NGS, high-throughput technology

Clinical Implementation of genome sequencing, omics measure.

Accumulation of Genome, omics data

Integration Molecular & Medical Info.

Clinical phenotyping (EMERGE project)

Medical Big Data

Knowledge Discovery

Genome-omics knowledge
Omics measurement

Breast Cancer Intrinsic classification

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

<table>
<thead>
<tr>
<th>intrinsic classification</th>
<th>ER</th>
<th>PgR</th>
<th>HER2</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminal A</td>
<td>ER(+) and/or PgR(+)</td>
<td>(−)</td>
<td></td>
<td>良好</td>
</tr>
<tr>
<td>Luminal B</td>
<td>ER(+) and/or PgR(+)</td>
<td>(+)*</td>
<td></td>
<td>A型</td>
</tr>
<tr>
<td>HER2 enriched</td>
<td>(−)</td>
<td>(−)</td>
<td>(+)</td>
<td>不良</td>
</tr>
<tr>
<td>Basal-like</td>
<td>(−)</td>
<td>(−)</td>
<td>(−)</td>
<td>不良</td>
</tr>
</tbody>
</table>

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
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
Discovery in Drug Repositioning
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 Significantly Differential Expression (SDE)
  - CMAP: Connectivity Map
    - Broad Institute, 1309 chemicals,
    - 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
    - ArrayExpress EBI
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にanti-convulsant topiramate,
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 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 based 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（第二世代型）
  - 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（第三世代型）
  - 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 conventional 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
Diseasome
(Goh, Barabasi et al.)

Disease which have more than one gene share

Disease gene which has shared

Kwang-II Goh*, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi
The human disease network
PNAS 2007
HDN

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

DGN

Node size
Nr causative 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 by 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
Transcriptional Profiling

Disease-drug network
orange drug
green disease

Tamoxifen (breast cancer)
Negative
⇒ atopy
⇒ mast cell increase, allergy suppress

positive
Side effect
⇒ 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1module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

Cell

Nucleus
Genome Omics Medicine and medical Big Data

Clinical Implementation of Genome/Omics Medicine

Huge amount of Data

Advances in Data Science
AI, Knowledge Discovery

Big Data-driven Clinical Science