Integration of Genomic and Phenomic Information in Medicine
~Tohoku Medical Megabank Experience~

Special Adviser to the Executive Director
Tohoku Medical Megabank Organization, Tohoku University
Professor Emeritus
Tokyo Medical and Dental University

Hiroshi Tanaka
**Clinical Implementation of Genome/Omics Medicine**

– Impact of Next Generation Sequencer (NGS)
– Clinical sequencing (CS) in hospitals in US
  – the first trial: Medical College of Wisconsin (2010)
    • Followed by Baylor Medical College (2011) and spread
  – President Obama Precision Medicine Initiative (2015)
– 1. Clinical sequencing of germline (innate) genome
  • To find ‘causative gene’ of undiagnosed disease at POC (hospital)
– 2. Clinical sequencing of somatic genome of cancer tissue
  • Memorial Sloan Kettering CC, MD Anderson CC etc. (2012)
  • TCGA (2006～）、ICCG (2008～） revealed driver/passenger mutations
  • Identify the driver mutation and assign molecularly-targeted drug

**Another Trends: Large-scale Genomic Cohort/Biobank**
World-wide Spread of Genomic Biobank

- **UK biobank**
  - investigate the respective contributions of genetic predisposition and environmental exposure (nutrition, lifestyle, etc)
  - about 500,000 volunteers in the UK
  - Aged from 40 to 69. Initial enrollment took place over four years
  - the volunteers will be followed for at least 25 years thereafter

- **Genomics England**
  - four-year 100,000 Genomes Project, 2013-2017
  - Disease oriented genomic biobank
  - perform whole genome sequencing of 100,000 participants.
  - focusing on rare diseases, cancer, and infectious diseases

- **BBMRI (Biobanking and BioMolecule Resource Research Infra)**
  - More than 300 biobanks in Europe recruited to join BBMRI.
  - Harmonization and Standardization to pool biobank data

- **Many other biobanks**
  - Estonia, Singapore, Australia, Taiwan etc.
Information Basis for Genome Medicine Big Data

• Biobank
  – an organized collection of human biological material and associated information stored for research purposes
  – Genomic Biobank
    • repositories of human DNA and/or associated data, collected and maintained for biomedical research

• Types of Biobank
  – Disease-oriented (genomic) biobank
    • BioBank Japan (BBJ : 2003-) 200,000 patients,
    • World pioneer project for starting GWAS study to discover disease susceptibility gene
  – Population-based (genomic) biobank
    • UK biobank, etc. In Japan, Tohoku Medical Megabank (ToMMo: 2012-) 150,000 healthy people for at least 20 years

• Change of the functional role in genome erra
  – Former: transplantation, source of therapeutics (umbilical blood, stem cell etc.)
  – Present: information basis for genome/omics medicine, Medical Big Data

• Towards Individualized Medicine and Healthcare
  – Disease mechanism and etiology have a vast variety of (personalized) intrinsic subtypes
  – Big Data (many patient cases) are necessary to collect/find out as many individualized or stratified subtypes (stratified patterns)
These Two Trends merge in near future and support the genome/omics medicine.

In the hospital

Clinical genome medicine

Integrated genome-phenome DB

Nation-wide

Large scale Medical Big Data

Disease Genome Cohort

Population Genome Cohort
Tohoku Medical Megabank Project
The Great East Japan Earthquake Disaster

- 14:46, March 11, 2011
- Earthquake off the Pacific coast of Japan
- Magnitude 9.0
- Powerful tsunami waves reached heights of up to 40.5 m
- Most disastrous earthquake that has ever experienced in Japan after World War II
- The number of dead and missing persons are:
  - Miyagi Prefecture 10,817
  - Iwate Prefecture 5,815
  - Fukushima Prefecture 1,814
  - Total 18,550 (incl. other areas)
- Medical institutions: hospital, clinics
devastation 351
seriously damaged 1,048
Two types of Cohort Study in ToMMo

- Residential Cohort
- Birth-Three generation cohort

**Residential Cohort**

- 1070 genomes
- Development of Japonica array

This year, 200,000 genome including three generation cohort

Finally, 150,000 genome analysis: WGS and Japonica array

**deCODE Study**

Iceland deCODE Genetics

- Family-based Prospective Cohort
- 296 K participants (whole nation)
- DNA samples from 95 K (1/3)
- Family history available from 1650

Whole-genome sequencing (N = 2,230)

Identification of SNPs (30.6 million) and Indels (3.6 million)

Chip-genotype imputation (N = 95,085)

Familial imputation (N = 296,526)

Analysis for Gene-environment interactions
Deep whole genome sequencing
Japanese Healthy Population
Whole Genome Sequencing in Tohoku Medical Megabank Project

- Whole genome sequencing (WGS) of 1,070 healthy Japanese individuals was executed
  - by PCR-free sequencing
  - more than 30X coverage (average 32.4X).
- First results of WGS in healthy Japanese
- Single laboratory, single protocol and single measurement method
- Would be a basis for personalized medicine and prevention
- Very rare as well as novel single-nucleotide variants (SNVs) are identified
  - Totally 21.2 million SNV
  - 12 million novel SNV
- A reference panel of 1,070 Japanese individuals (1KJPN)
  - From the identified SNVs, we construct 1KJPN,
  - including some very-rare SNVs.
- Information of Genome Sequences
  - Information of statistical frequency of SNV (up to singleton SNP)
  - Genome sequences are open by controlled access
- From this panel, we designed custom-made SNP array for Japanese
  - Japonica array
  - 650 thousand SNV
Data Processing and variant discovery

• **Material**
  – 1344 candidates were selected from biobank
    • Considering **traceability** of participants’ information
    • Quality and abundance of DNA sample for SNP array and WGS
  – 1070 samples were selected by measured results by Omni2.5
    • By filtering out close relatives and outliers
  – Sequenced by Illumina Hiseq2500
    • Using **PCR-free protocol**

• **Variant discovery**
  – 21.2 million high confident SNV
  – 12 million novel SNVs
    • After several filtering procedure, high confident SNVs
    • Reference genome: GRCh37/hg19
    • False discovery rate <1.0%
### Summary of WGS of Japanese individuals and variant detection in autosomes.

<table>
<thead>
<tr>
<th>Total samples</th>
<th>1,070</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total raw bases</td>
<td>100.4 trillion bases</td>
</tr>
<tr>
<td>Mean sequenced depth</td>
<td>32.4 ×</td>
</tr>
</tbody>
</table>

#### SNVs

<table>
<thead>
<tr>
<th></th>
<th>High-confidence SNVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>21,221,195</td>
</tr>
<tr>
<td>Number of known variants*</td>
<td>9,219,783</td>
</tr>
<tr>
<td>Number of novel variants*</td>
<td>12,001,412</td>
</tr>
<tr>
<td>Novelty rate</td>
<td>56.55%</td>
</tr>
<tr>
<td>Average number per sample</td>
<td>2,716,853</td>
</tr>
<tr>
<td>Average individual heterozygosity</td>
<td>1,532,773</td>
</tr>
</tbody>
</table>

#### Deletions

<table>
<thead>
<tr>
<th></th>
<th>1 bp ≤ length &lt; 100 bp</th>
<th>100 bp ≤ length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sites overall</td>
<td>1,969,302</td>
<td>47,343</td>
</tr>
<tr>
<td>Number of novel variants†</td>
<td>1,429,636</td>
<td></td>
</tr>
<tr>
<td>Novelty rate</td>
<td>72.60%</td>
<td></td>
</tr>
<tr>
<td>Number of inframe/frameshift</td>
<td>3,112/4,454</td>
<td></td>
</tr>
<tr>
<td>Average number per sample</td>
<td>190,857</td>
<td>2,654</td>
</tr>
</tbody>
</table>

#### Insertions

<table>
<thead>
<tr>
<th></th>
<th>1 bp ≤ length &lt; 100 bp</th>
<th>100 bp ≤ length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sites overall</td>
<td>1,384,230</td>
<td>9,354</td>
</tr>
<tr>
<td>Number of novel variants†</td>
<td>1,037,839</td>
<td>9,354</td>
</tr>
<tr>
<td>Novelty rate</td>
<td>74.98%</td>
<td></td>
</tr>
<tr>
<td>Number of inframe/frameshift</td>
<td>1,577/2,506</td>
<td></td>
</tr>
<tr>
<td>Average number per sample</td>
<td>159,359</td>
<td>45</td>
</tr>
</tbody>
</table>
Statistics of Indel and SNV

(a) Size-frequency of Del, SNP, Ins

The size-frequency spectrum of SNVs, deletions and insertions discovered by high-coverage sequencing in 1KJPN. Novelty rates are shown by the red line. Peaks corresponding to long interspersed elements (LINE), Alu and microsatellite repeat (MSR) are shown.

(b) Size-frequency of CNV

Size-frequency spectrum of CNVs estimated from high-coverage sequencing data in the genic regions in 1KJPN.

(c) Frequency of SNV

Relative frequency (%) vs. Minor allele frequency (%)
**Japonica Array**

- **Novel custom-made SNP array**
  - based on the 1KJPN panel, for whole-genome imputation of Japanese individuals.

- **The array contains 659,253 SNPs**
  - tag SNPs for imputation,
  - SNPs of Y chromosome and mitochondria,
  - SNPs related to previously reported genome-wide association studies and pharmacogenomics.

- **Better imputation performance**
  - for Japanese individuals than the existing commercially available SNP arrays
  - Common SNPs (MAF>5%), the genomic coverage of the Japonica array ($r^2>0.8$) was 96.9%
  - Coverage of low-frequency SNPs (0.5%<MAF≤5%) : 67.2%

- **High quality genotyping performance**
  - of the Japonica array using the 288 samples in 1KJPN;
  - Average call rate 99.7%
  - Average concordance rate 99.7% to the genotypes obtained from high-throughput sequencer.
### Japonica Array

**Category of SNPs on the Japonica array**

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of SNPs</th>
<th>Array occupancy rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag SNPs (including X chromosome)</td>
<td>638,269</td>
<td>96.8%</td>
</tr>
<tr>
<td>Pharmacogenomics markers</td>
<td>2,028</td>
<td>0.31%</td>
</tr>
<tr>
<td>Y chromosome</td>
<td>275</td>
<td>0.04%</td>
</tr>
<tr>
<td>Mitochondria</td>
<td>70</td>
<td>0.01%</td>
</tr>
<tr>
<td>NHGRI GWAS catalog</td>
<td>10,798</td>
<td>1.64%</td>
</tr>
<tr>
<td>HLA</td>
<td>3,906</td>
<td>0.59%</td>
</tr>
<tr>
<td>Untaggable functional SNPs</td>
<td>3,990</td>
<td>0.61%</td>
</tr>
<tr>
<td>Total</td>
<td>659,253</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: GWAS, genome-wide association studies; SNP, single nucleotide polymorphism.

Some SNPs are overlapped among categories.

---

**Panel: 1KJPN**

- **WGS(4K$)**
- **Japonica Ar(<200$)**

**Genotype imputation**

**Japonica array (96 sample)**
ToMMo integrated database enables to generate health-science big-data
Information in the integrated database will be open to research laboratories in Japan
ToMMo integrated data will be of important for new drug development for specific group of people

http://ijgvd.megabank.tohoku.ac.jp/

Data Release on Dec 15, 2015
Japanese Multi omics reference panel : jMorp

Metabolomics and Proteomics reference database from 500 cohort participants

https://jmorp.megabank.tohoku.ac.jp/
Integrated Database for genomic and environmental information
Gene-environment interactions causing common disease

Environment
Exposures, Nutrition, Lifestyle

Environment interacts with genetic system

Gene
Metabolic pathways, Signaling pathways

Phenotype (disease)

Precise Stratification
Personalized prevention
Idiosyncratic Effect of Combination of GxE factors

- Interaction of genomic and environmental factor
- Idiosyncratic Effect: colon cancer

<table>
<thead>
<tr>
<th></th>
<th>CYP1A2 Phenotype ≤ Median</th>
<th>CYP1A2 Phenotype &gt; Median</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Likes rare/medium meat</td>
<td>Likes well-done meat</td>
</tr>
<tr>
<td>Non-Smoker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAT2 Slow</td>
<td>1</td>
<td>1.9</td>
</tr>
<tr>
<td>NAT2 Rapid</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>Ever-Smoker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAT2 Slow</td>
<td>1</td>
<td>0.9</td>
</tr>
<tr>
<td>NAT2 Rapid</td>
<td>1.2</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Towards Disease-oriented Biobank
Population and Disease-based Mixed Genome Cohort

Healthy population Genomic cohort

To Disease-based Clinical Cohort

Disease Occurrence

years

subjects
Clinical data collection and phenotyping

Medical records storage

HIS storage

Ordering/results

HL7 Ver.2.5

HIS gateway

SS-MIX2 standardized storage for clinical use

root directory

patient id (first 3 digits)

patient id (next 4-6 digits)

patient id

diagnosis date

data type

message

Data retrieval

De-identification

Clinical data for research use

Elementary database

Ingegrated database

Disease and phenotype data

Phenotyping (rule-base algorithm & machine learning (deep learning))

Hospitals

Tohoku Medical Megabank
Disaster-tolerant Multi-hierarchical Regional Healthcare IT system

- Prefecture Level
  - Prefecture-level Cloud center
  - Central Hospital

- 2nd Medical Zone Level
  - Sharing
  - Central Hospital
  - Reservation
  - Lab test
  - Nursing Care Office
  - Small scaled hospitals
  - Clinics

- Daily Life Area
  - Nursing care worker
  - House call physician
  - Temporary house
  - Mobile Tablet PC

SS-MIX
Regional Healthcare Information Network (MMWIN)
Starts August 10, 2012
In Miyagi Prefecture
Now most of hospitals (large/medium size) have been connected
Clinical genome/omics database

- Before TMM project, we have been engaged in Government-commissioned project
- Integration of genome/omics information and clinic-pathological information project in various cancer
- iCOD (integrated Clinical Omics Data Base) project
  Mainly liver and colon cancer were dealt (about 1000 surgical cases)
- We measured Gene expression profiles and CNV were measured as genome/omics information.
- The relation between genome/omics and clinic-pathological information is analyzed
  (Shimokawa,Tanaka et al. PloS Genomics 2010)
Center for Information Medicine, Tokyo Medical and Dental University has developed "Integrated Clinical Omics Database (iCOD)" aiming to establish the basis of Omics-based Medicine and Systems Pathobiology.

We have launched this project since 2005 with the support of Japan Science and Technology Agency and Ministry of Education, Culture, Sports, Science and Technology. In this iCOD, we have stored 525 patient case data of colon cancer, hepatic cellular carcinoma and oral tumor (in Japanese version). English version is available now, containing 140 patient cases of hepatic carcinoma.

We opened Japanese version in July 2008 and English version has been available since April 2009.

We prepared the raw data download page for the person who wants to analyze them with his/her own tool.
Database for Translational Research

Center for Information Medicine, Tokyo Medical and Dental University has developed Integrated Clinical Omics Database (iCOD) aiming to establish the basis of Omics-based Medicine and Systems Pathobiology.

We have launched this project since 2005 with the support of Japan Science and Technology Agency and Ministry of Education, Culture, Sports, Science and Technology. In this iCOD, we have stored 926 patient case data of colorectal cancer, hepatic cell carcinoma and oral tumor (in Japanese version). English version is available now, containing 140 patient cases of hepatic carcinoma.

We opened Japanese version in July 2000 and English version has been available since April 2003.

Case Archive

Integrated Clinical Omics Database

About the Case Database

The contents of this case database are based on clinical, pathological, and environmental data obtained from patients who received medical care at Tokyo Medical and Dental University Hospital Faculty of Medicine, and other collaborating institutions since 2005. Informed consent was obtained from each patient.

Our target was mainly cancer patients. Extensive information on medical history, lifestyle, laboratory data, pathological findings, diagnosis, and prognosis were entered into the database. We also included gene expressions and detailed location of sample tissue obtained from the patients.

Searching the Database

Search the database, enter key terms of your query in the search box. Spaces between words act as the Boolean operator "AND", combining your query terms. Entering specific terms in the limits box allows you to restrict the search results. When you finish entering your query, click the Search button to see the search results. Once you check "AND" radio button and execute the search, further words will be added to the first query. Thus continuing the search by combining all the query terms with AND. If you wish to start with a different query, first check "NEW" radio button.
Clinical data

Pathological Data

Molecular Data
Clinical Omics Data Analysis

• 2 Dimensional – 3 Layered map
  – Connect three different layers
    • (Molecular, Pathological, Clinical Layer)
  – Principal component or user defined axis

• Pathome - Genome map
  – Canonical Correlation analysis
  – Connect Pathological information and genes
Clinical Omics Data Analysis

Clinical omics data analysis is a method of observing the correlation between pathology and genes using cross-sectional, statistical analysis of clinical, pathological, and molecular information. Types and equations of clinical, pathological, and molecular information layers are each determined and placed on our 2-dimensional-3-layered (2D-3L) map. By choosing one layer at a specific point, information of other layers at the specified point will be displayed, thus enabling you to visualize the correlation among the layers.

In addition, integrated display of all the data by using the regular canonical correlation analysis enables you to perceive the clinical and pathological information and their correlation to gene expression data at a glance.
2 Dimensional – 3 Layered Map

Molecular Layer

Pathological Layer

Clinical Layer
Pathome - Genome map

Display both Pathological states and Genes

Clinical data

Gene
Canonical Correlation Analysis

Maximize Correlation Coefficients

Gene expression $X$

Clinico-pathological $Y$
Human Phenotype Ontology (HPO) is a controlled vocabulary used to describe phenotypic abnormalities seen in human disease.

- Growth abnormalities
- Abnormality of the nervous system
- Abnormality of the abdomen
- Abnormality of the muscles

11,000 classes
116,000 annotations
7,000 rare diseases

Marfan Syndrome [MIM:154700]
Arachnodactyly [HP:0001166]
Ectopia lentis [HP:0001083]
Phenotyping using HPO

Clinical summaries can be used for phenotyping by natural language processing using HPO.

HPO-Japanese is under development, and the first draft will be released until March 2016.
Conclusion

- Large scale genomic cohort/biobank
- Information basis for genome/omics medicine and healthcare
- Source for Clinical and healthcare Big Data
- Personalized medicine and personalized
- Collect as many personalized (stratified) patterns or disease subtype