Integration of Genomic and Phenomic Information in Medicine
～integrated Clinical Omics DB (iCOD) and Tohoku Medical Megabank (TMM)～

Special Adviser to the Executive Director
Tohoku Medical Megabank Organization, Tohoku University
Professor Emeritus
Tokyo Medical and Dental University
Hiroshi Tanaka
General situation of EHR and genome/omics medicine in Japan
History and Evolution of Medical ICT in Japan

Adoption of ICT in Healthcare was relatively early in Japan

For a long period (1970s-2000s), Medical ICT has been developed and primarily for administration and medical practice within the hospital.

1st generation: **Departmental system**: 1970s -
financing (accounting) system, departmental computerized system of clinical laboratory or pharmacy

2nd generation: **CPOE (Computerized Physician Order Entry)**: 1980s-
Order-entry/result reporting system of laboratory or radiological test, drug prescription

3rd generation: **EHR/EMR**: 2000s-
Electronic Health/Medical Record
Adoption rate of EHR/EMR in Japan

In opening a new clinic, 70-80% of them adopts EHR/EMR
Governmental Policies for realization of genomic medicine in Japan

• Headquarters for Healthcare Policy
  – Council for Promotion of Genome Medicine Realization
  – Propose the main direction for realization of genome medicine in Japan

• Ministry of Health, Labour and Welfare
  – Project for Practical Implementation of Genome Medicine
  – Headquarters for Promotion of Genome Medicine, 2015.9
  – Integration Project of Clinical Genomic DB (AMED)

• Japan Agency for Medical Research and Development (AMED)
  – Unified Research Funding Agency, 2015.4
  – “Initiative on Rare and Undiagnosed Diseases (IRUD)”, 2015.10
  – Working Group for Promotion of Genome Medicine, report 2016.2
  – Platform Project for promotion of genome medicine
  – Research foundation project for Three BioBanks
Practicing Genome Medicine in Japan

- **National Cancer Center**
  - Cancer Diagnosis by “NCC oncopanel”
  - SCRUM-JAPAN
    - Business-Academia Collaboration Cancer genome consortium
- **Shizuoka Cancer Center**
  - “HOPE” project
  - Identify the driver mutation for cancer and assign the most appropriate molecularly targeted anticancer drug
- **Kyoto University Hospital**
  - “Oncoprime” project
- In some of above clinical implementations, genomic information is integrated into EHR
Two Major Streams in the trends of Genomic Healthcare

- **Clinical Genome Medicine**
  - Clinical Implementation
- **Genomic Cohort / Biobank**
  - International Spread

Both need an integration of genome and phenomic (clinical and environmental) information
1. Clinical Implementation of Genome Medicine

• Impact of Next Generation Sequencer (NGS)
  – Clinical sequencing (CS) started to be used in hospitals in US
  – the first trial: Medical College of Wisconsin (2010)
    • Followed by Baylor Medical College (2011) and spread

• Clinical Implementation of Genome Medicine
  – Now, several tens hospitals in US, mostly three types
  1. Clinical sequencing of germline (innate) genome
    • To find ‘causative gene’ of undiagnosed and inherited disease
    • End the “Diagnostic Odyssey”, 25%～40% success
  2. Clinical sequencing of somatic genome of cancer tissue
    • Memorial Sloan Kettering CC, MD Anderson CC etc. (2012)
    • TCGA (2006〜)、ICCG (2008〜): driver/passenger mutations
    • Identify the driver mutation and assign appropriate molecularly-targeted drug
  3. Personalized medication
    • based on the polymorphism of drug metabolizing enzyme of patient

• President Obama: Precision Medicine Initiative (2015)
2. World-wide Spread of Genomic Cohort/Biobank

- **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
- **UK biobank**
  - investigate the respective contributions of
    - genetic predisposition and environmental exposure (nutrition, life style, etc)
    - about 500,000 volunteers in the UK, Aged from 40 to 69, followed for 25 y.
- **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.
Biobank as Information Basis for Genome Medicine

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

• Types of Biobank
  – Disease-oriented (genomic) biobank
    • BioBank Japan (BBJ : 2002-) 200,000 patients, World first GWAS study for disease susceptibility gene
  – Population-based (genomic) biobank
    • Tohoku Medical Megabank (TMM: 2012-) 150,000 healthy people for at least 20 years

• Towards Personalized Medicine and Healthcare
  – Disease mechanism and etiology have a vast variety of (personalized) intrinsic subtypes
  – Big Data (many patient cases) are necessary to collect/exhaust as many personalized subtypes
These Two Trends would merge and support the genome/omics medicine

Within hospital

Clinical genome medicine

Integrated genome-phenome DB
EHR

Nation-wide basis

New knowledge, New information

Large scale Medical Big Data
(both genomic phenomic information)

Disease Genome Cohort

Population Genome Cohort
Integration of clinical genome/omics into EHR

integrated Clinical Omics Database (iCOD)
Genome Medicine in Japan
Integrated Clinical Omics Database (iCOD) Project of Japan (2005~)

- Integrated DB of genome/omics and EHR (clinical, life style,..)
  - Information basis for realization of genomic EHR.
- Government-commissioned collaborative project
  - Tokyo Medical & Dental University (TMD)
  - Riken
  - Nat. Inst. of Adv. Industrial Science and technology (AIST)
  - National Cancer Center(NCC)
- Totally 10 million $ for first 5 years, 2005-2010 (about 1000 cancer cases)
  Started Earlier than “Emerge project” in US
- But for Japanese situation of GM, iCOD project was too early

Shimikawa K, Tanaka H. et. al.
iCOD: an integrated clinical omics database based on the systems-pathology view of disease
BMC genetics (2010)
Databases 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 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.

Downloading raw data

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

Download Page
Clinical data

Molecular Data

Comprehensive list of the patient data on time-line from admission

Pathological Data

Case archive

Macroscopic findings

<table>
<thead>
<tr>
<th>Number</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor multiplicity</td>
<td>Single</td>
</tr>
<tr>
<td>Maximum diameter</td>
<td>5.5 cm</td>
</tr>
<tr>
<td>Degree of hepatic damage</td>
<td>Grade E2</td>
</tr>
<tr>
<td>Growth type</td>
<td>Cephalic cancer, proliferation</td>
</tr>
<tr>
<td>Capsule formation</td>
<td>+</td>
</tr>
<tr>
<td>Capsule infiltration</td>
<td>+</td>
</tr>
<tr>
<td>Portal vein invasion</td>
<td>0</td>
</tr>
<tr>
<td>Hepatic vein invasion</td>
<td>0</td>
</tr>
<tr>
<td>Hepatic artery invasion</td>
<td>0</td>
</tr>
<tr>
<td>bile duct invasion</td>
<td>0</td>
</tr>
<tr>
<td>T stage</td>
<td>T1</td>
</tr>
<tr>
<td>N stage</td>
<td>N0</td>
</tr>
<tr>
<td>M stage</td>
<td>M0</td>
</tr>
</tbody>
</table>

Pathological Data

Diagnosis: Hepatocellular carcinoma

Differential diagnosis: Benign hepatic tumour

Grade: Low

Growth type: Cephalic cancer, proliferation

Capsule formation: +

Capsule infiltration: +

Portal vein invasion: 0

Hepatic vein invasion: 0

Hepatic artery invasion: 0

Bile duct invasion: 0

Stage: T1, N0, M0

Histological examination

Histological examination of the liver shows immunostaining for the target markers:

- GOSCL
- ADCC
- RAS
- NAC
- CA
- AR

Genes:

- FLI1
- MMML
- DDIT4
- MACF1
- UBE3A
- CTNMBL1
- NRAS
- NOTCH1
- NRAS
- CTNMBL1

Conclusion:

The histological examination of the liver shows the presence of the target markers, indicating a diagnosis of Hepatocellular carcinoma with a low-grade malignancy.
Graphical presentation of relation between Genome/Omics and Clinic-pathological (EHR) data

- **iCOD**: comprehensive DB specially for cancer (colon, liver) patient data
- **Relation** between genome/omics and clinico-pathological phenotype is presented
  1. **Molecular data** of cancer surgical tissue
     - Gene expression profile
     - Copy number variation
  2. **Clinico-Pathological phenotype**
     - Lab test result, medical image (CT, MRI,..), drug history
     - Tumor size, stage, invasion
     - Clinical outcome, recurrence, metastasis
- **Not correlation network** among molecular and clinic-pathological findings, but
- **Two special graphical relation presentation**
2 Dimensional – 3 Layered (2D-3L) map
- Connect three different layers
  - Molecular, Pathological, Clinical Layer
- Axes of each 2D map
  - principal component (PCA) of the layer or user defined

Pathome - Genome map
- Canonical correlation analysis between \( G \) and \( P \)
- Both items are mapped into same plane
- The distance represents the relatedness between clinic-pathological phenotype (P) and genes activity (G)
Patient points in three 2D coordinates (molecular, pathological, and clinical) are connected to show the corresponding relation between genome, pathological, and clinical conditions.
Canonical correlation analysis
Maximize the correlation coefficient
Between the linear combination of gene expression and clinic-pathological variables

Pathome - Genome map

Pathome-Genome Map

Enlarge
Latter stage of the iCOD project

• “Integrating DB in life science” national project budget
• Development of Ontology system for Medical Concept
  – To obtain interoperability of concept or terminology with other life-science DB
  – When exact match between the concept or terminology in other DB is not found
  – generalization (upward) or specialization (downward) inference is executed along the ontology system to find interchangeable concept or terminology
• Theoretical sound but not so feasible
  – Took too much time to find the best much concept at that time
First Results of TMM
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></th>
<th>High-confidence SNVs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total samples</td>
<td>1,070</td>
</tr>
<tr>
<td>Total raw bases</td>
<td>100.4 trillion bases</td>
</tr>
<tr>
<td>Mean sequenced depth</td>
<td>32.4 x</td>
</tr>
</tbody>
</table>

### SNVs

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></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>

**Copy number Variants**

|                                | 25,923                 |
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.
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 SNPsa</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.  
*aSome SNPs are overlapped among categories.*

---

**Genotype imputation**

- **WGS(4K$)**
- **Japonica Ar(<200$)**
Integrated Database for genomic and environmental information
Towards the development of Information systems
Tohoku Medical Megabank (TMM)

• iCOD team (prof. Tanaka’s Lab, TMDU) was asked to collaborate with development of the information system of TMM
  – Appreciating iCOD development
  – Several members moved to TMM in 2012
  – But, TMM is biobank of healthy population
  – Integrating information with genome/omics is different, from clinical to environmental data

• TMM Systems for our division to develop
  (1) Information manage system for genomic cohort study
  (2) Integrated database of genomic and environmental information
Gene-environment interactions causing common disease

Environment
Exposures, Nutrition, Lifestyle

Environment interacts with genetic system

Phenotype (disease)

Gene
Metabolic pathways, Signaling pathways

Precise Stratification
Personalized Prevention
New Method for GxE relative risk estimation

• Interaction of genomic and environmental factor
  – Not additive, not multiple
  – Combination specific
• As first step to estimate GxE effect on relative risk of disease occurrence
• Comprehensive listing of GxE contingency tables

<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>

Each P value Estimation

Cochran-Mantel-Haenszel table

<table>
<thead>
<tr>
<th>population</th>
<th>Disease (+)</th>
<th>Disease (-)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E (+)</td>
<td>E (-)</td>
</tr>
<tr>
<td>0 (aa)</td>
<td>n₀₀</td>
<td>n₀₁</td>
</tr>
<tr>
<td>1 (aA)</td>
<td>n₁₀</td>
<td>n₁₁</td>
</tr>
<tr>
<td>2 (AA)</td>
<td>n₂₀</td>
<td>n₂₁</td>
</tr>
</tbody>
</table>

P value for G1x E1  ➔  D

Gene set

<table>
<thead>
<tr>
<th>p</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7x10⁻¹⁴</td>
<td>9x10⁻¹⁸</td>
<td>...</td>
<td>3x10⁻²²</td>
</tr>
<tr>
<td>2</td>
<td>5x10⁻⁰³</td>
<td>2x10⁻⁰⁴</td>
<td>...</td>
<td>5x10⁻⁰⁵</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>20</td>
<td>3x10⁻¹⁷</td>
<td>9x10⁻²¹</td>
<td>...</td>
<td>4x10⁻²²</td>
</tr>
</tbody>
</table>

Gene allele X Environment = risk of Disease
Personalized prevention
Idiosyncratic Effect of Combination of GxE factors

Each row of variables (genes, Environment factors) are rearranged by hierarchical clustering

<table>
<thead>
<tr>
<th>Environment factors</th>
<th>gene</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>7x10^{-14}</td>
<td>9x10^{-18}</td>
<td>...</td>
<td>3x10^{-22}</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>5x10^{-03}</td>
<td>2x10^{-04}</td>
<td>...</td>
<td>5x10^{-05}</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>3x10^{-17}</td>
<td>9x10^{-21}</td>
<td>...</td>
<td>4x10^{-22}</td>
</tr>
</tbody>
</table>

Relative Risk Landscape
Summary

- Two trends of genomic healthcare
  (1) Genome/omics clinical medicine in hospital
  (2) Large scale genomic cohort/biobank
- These two trends pursue same goal: Personalized and precise healthcare and equally indispensable.
- For both, integration of genome/omics information and phenomic information (clinical, environmental) is key importance.
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

Japonica Array with
- Genotype imputation
- Transmission disequilibrium test
- IBD (identity by descent) mapping etc.

Environmental factors
- Whole genome sequence
- Japanese genome structure
- iJGVD / genome variation database

Analysis for Gene-environment interactions

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)

Association Analyses
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