TOHOKU UNIVERSITY

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

東北メディカル・メガバンク機構
TOHOKU MEDICAL MEGABANK ORGANIZATION

World-wide Trends of Genome/Omics in Medicine

Clinical Implementation of Genome/Omics Medicine

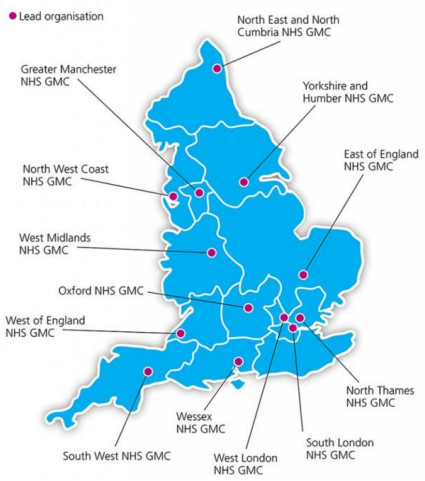
- Impact of Next Generation Sequencer (NGS)
- Clinical sequencing (CS) in hospitals in US
- the first trial: Medical College of Wisconsin (2010)
  - Followed by Baylor Medical College (2011) and spread
- President Obama Precision Medicine Initiative (2015)
- 1. Clinical sequencing of germline (innate) 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

Obama's PMI

Another Trends: Large-scale Genomic Cohort/Biobank

World-wide Spread of Genomic Biobank

- UK biobank
  - a large long-term biobank study in the United Kingdom (2006-2010, 62M£, 2011-16, 25M£)
  - 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. 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 Resourse 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.

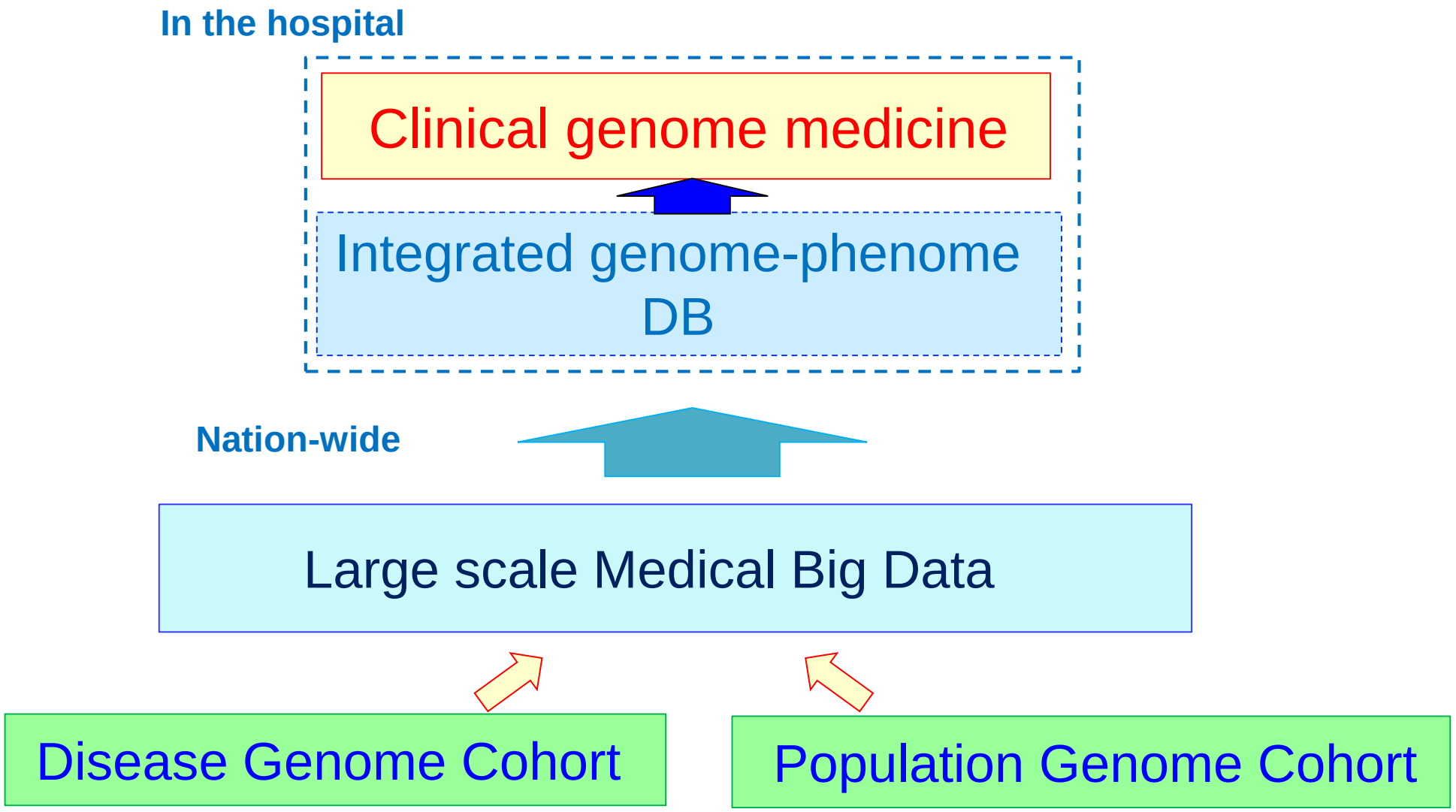


NHS Genome Medical Center

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)

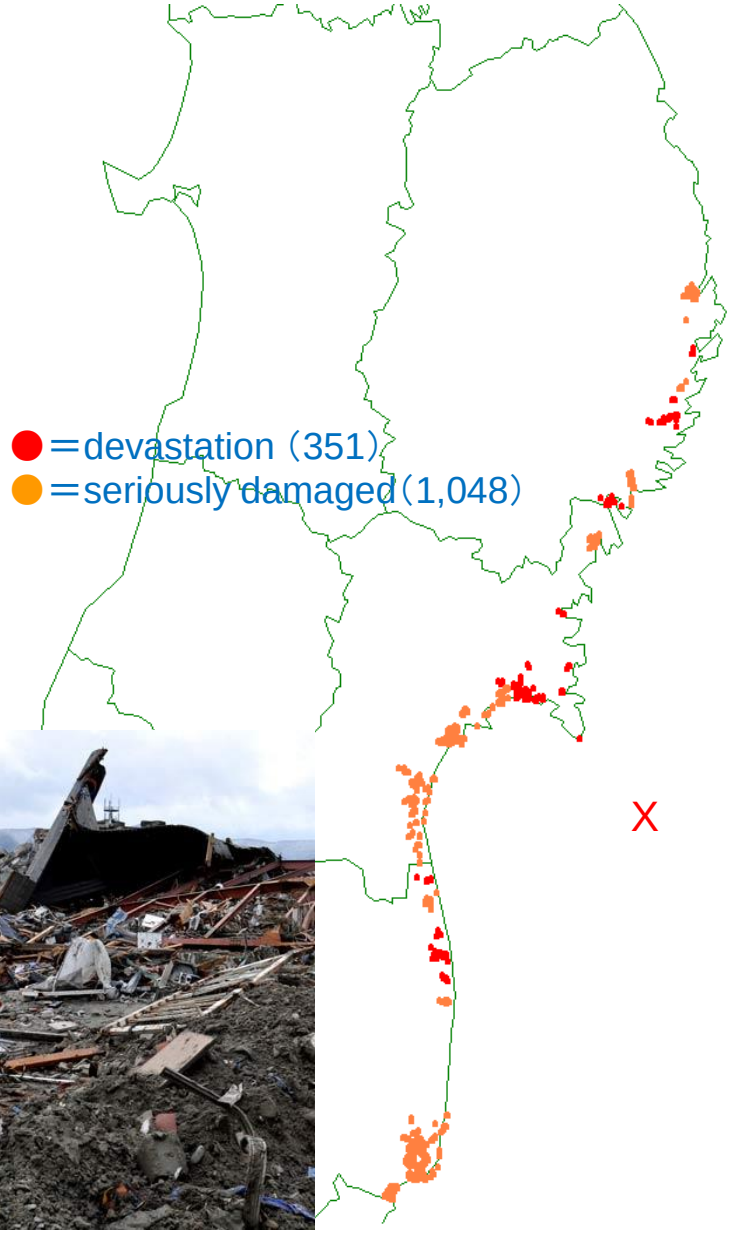
These Two Trends merge in near future and support the genome/omics medicine



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



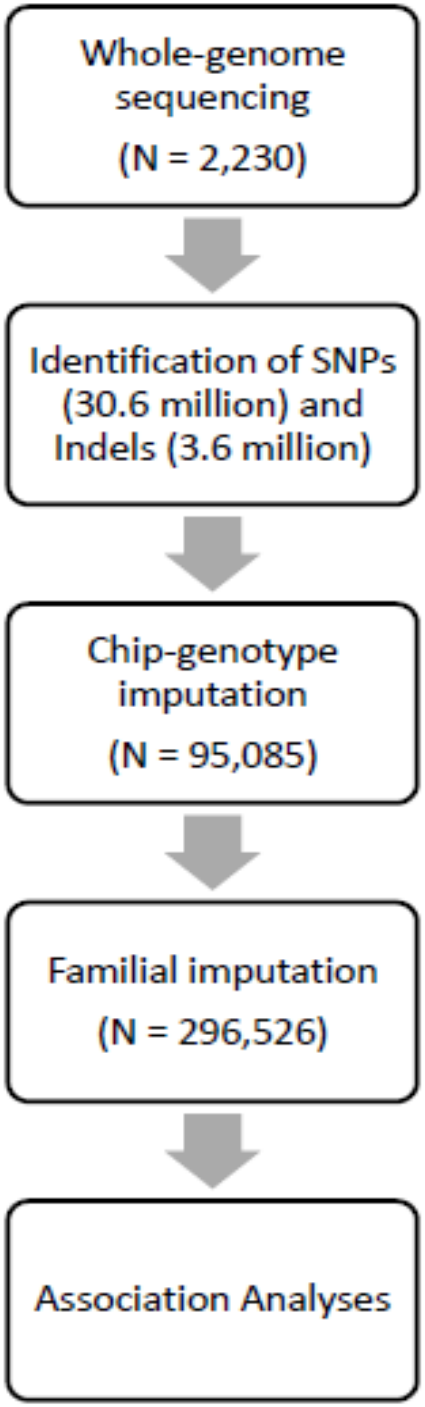
Two types of Cohort Study in ToMMo

- Residential Cohort
- Birth-Three generation cohort

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



Residential Cohort

1070 genomes

Delopement of Japonica array

This year, 200,000 genome including three generation cohort

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

Environmental factors
Whole genome sequence

Japanese genome structure
iJGVD / genome variation database

Japonica Array with
Genotype imputation

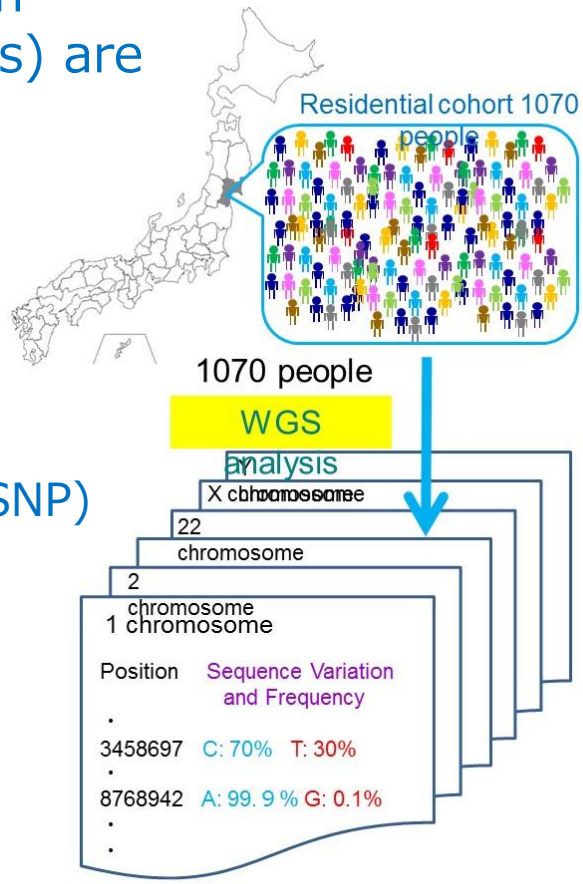
transmission disequilibrium test
IBD (identity by descent) mapping etc.

Analysis for Gene-environment interactions

Deep whole genome sequencing Japanese Healthy Population

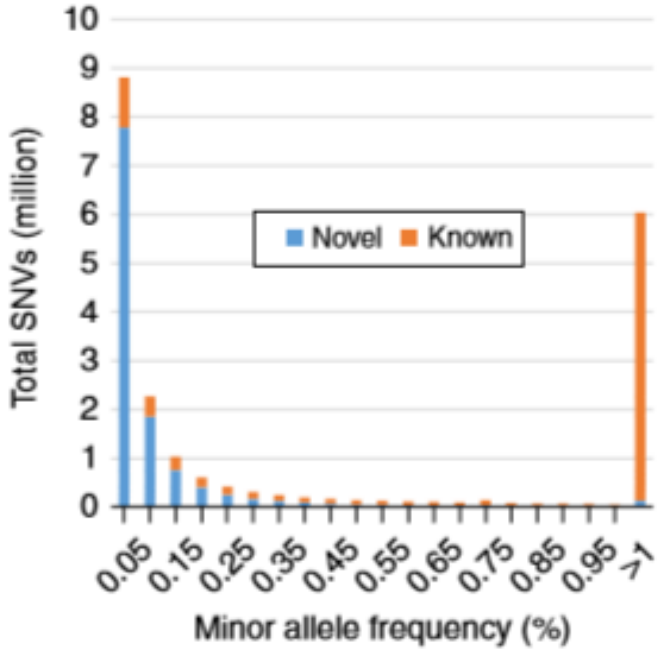
Whole Genome Sequencing in Tohoku Medical Megabank Project

- 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%



The numbers of novel and known SNVs in each MAF bin.

The novel SNV frequency begins to dominate for lower MAF

Summary of WGS of Japanese individuals and variant detection in autosomes.

Total samples	1,070
Total raw bases	100.4 trillion bases
Mean sequenced depth	32.4 ×

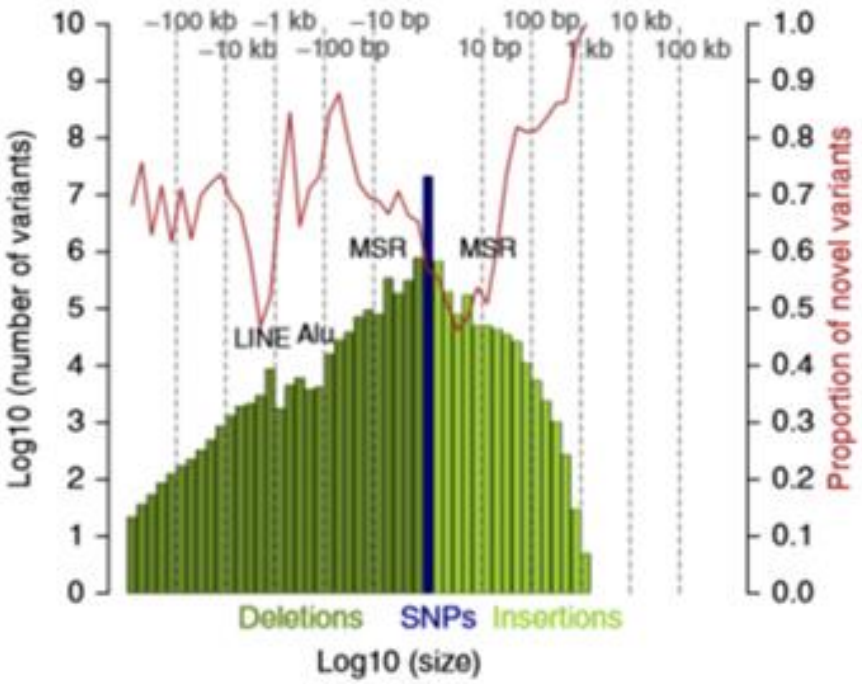
SNVs	High-confidence SNVs
Total	21,221,195
Number of known variants*	9,219,783
Number of novel variants*	12,001,412
Novelty rate	56.55%
Average number per sample	2,716,853
Average individual heterozygosity	1,532,773

Deletions	1 bp ≤ length < 100 bp	100 bp ≤ length
Number of sites overall	1,969,302	47,343
Number of novel variants†	1,429,636	—
Novelty rate	72.60%	—
Number of inframe/frameshift	3,112/4,454	—
Average number per sample	190,857	2,654

Insertions	1 bp ≤ length < 100 bp	100 bp ≤ length
Number of sites overall	1,384,230	9,354
Number of novel variants†	1,037,839	9,354
Novelty rate	74.98%	—
Number of inframe/frameshift	1,577/2,506	—
Average number per sample	159,359	45

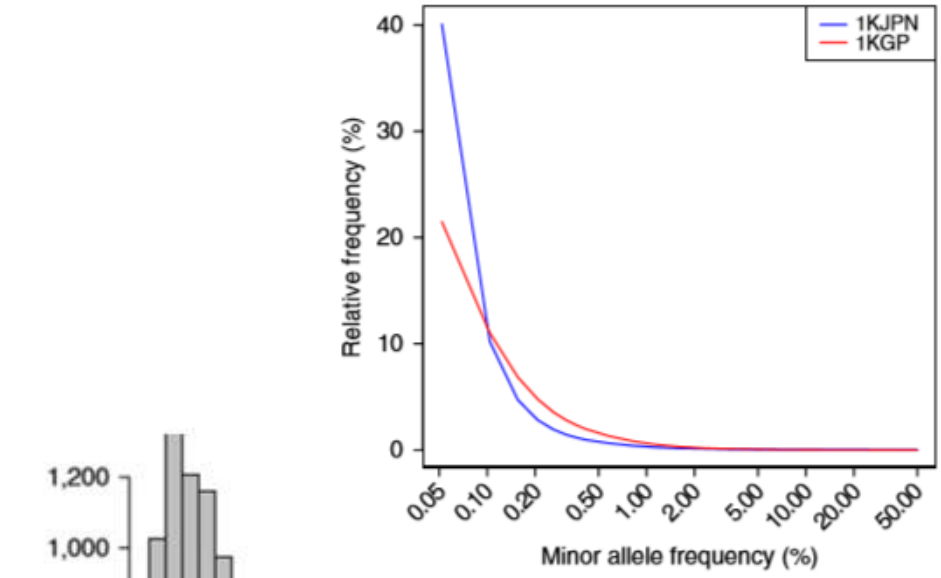
Copy number Variants	25,923

Statistics of Indel and SNV

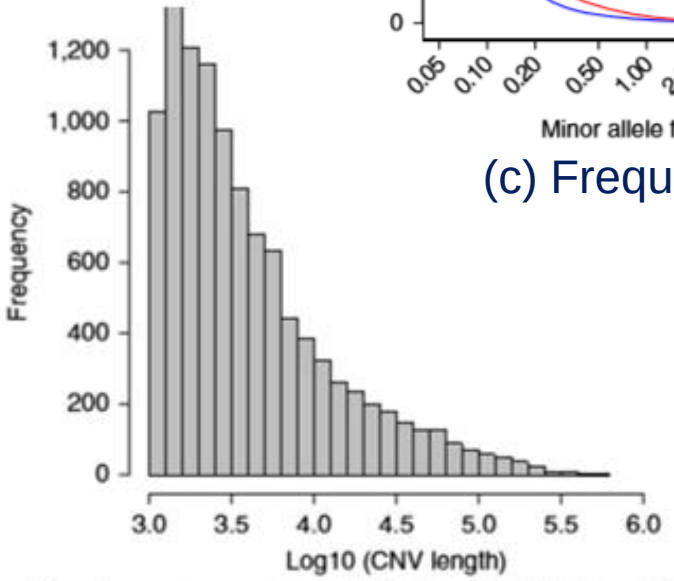


The size-frequency spectrum of SNVs, deletions and insertions discovered by high-coverage sequencing in 1KJPN. Novelty rates are shown by the red line. Peaks corresponding to long interspersed elements (LINE), Alu and microsatellite repeat (MSR) are shown.

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



(c) Frequency of SNV



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

(b) Size-frequency of CNV

Japonica Array

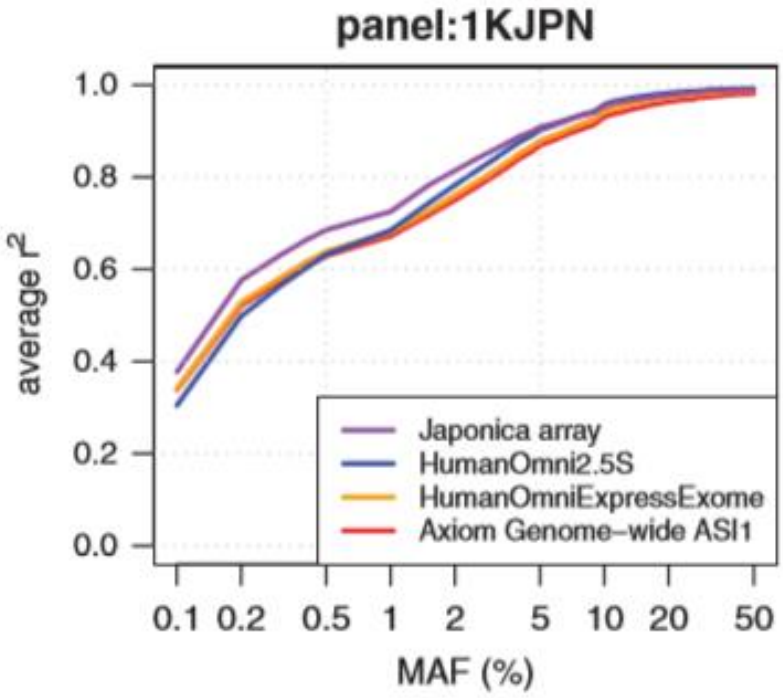
- Novel custom-made SNP array
  - based on the 1KJPN panel, for whole-genome imputation of Japanese individuals.
- The array contains 659, 253 SNPs
  - tag SNPs for imputation,
  - SNPs of Y chromosome and mitochondria,
  - SNPs related to previously reported genome-wide association studies and pharmacogenomics.
- 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

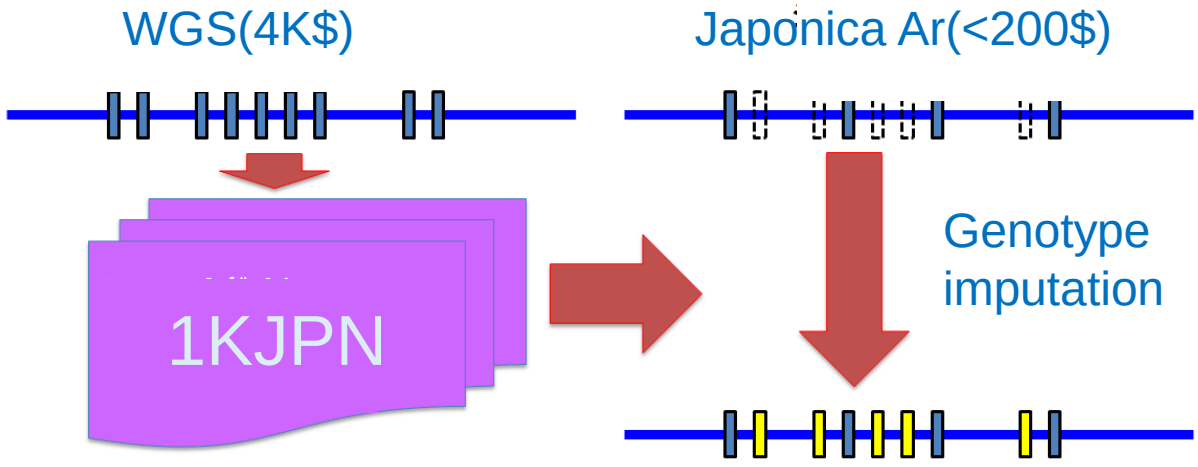
Category of SNPs on the Japonica array

Category	*Number of SNPs*[a]	*Array occupancy rate*
Tag SNPs (including X chromosome)	638 269	96.8%
Pharmacogenomics markers	2028	0.31%
Y chromosome	275	0.04%
Mitochondria	70	0.01%
NHGRI GWAS catalog	10 798	1.64%
HLA	3906	0.59%
Untaggable functional SNPs	3990	0.61%
Total	659 253	—

Abbreviations: GWAS, genome-wide association studies; SNP, single nucleotide polymorphism.
[a]Some SNPs are overlapped among categories.



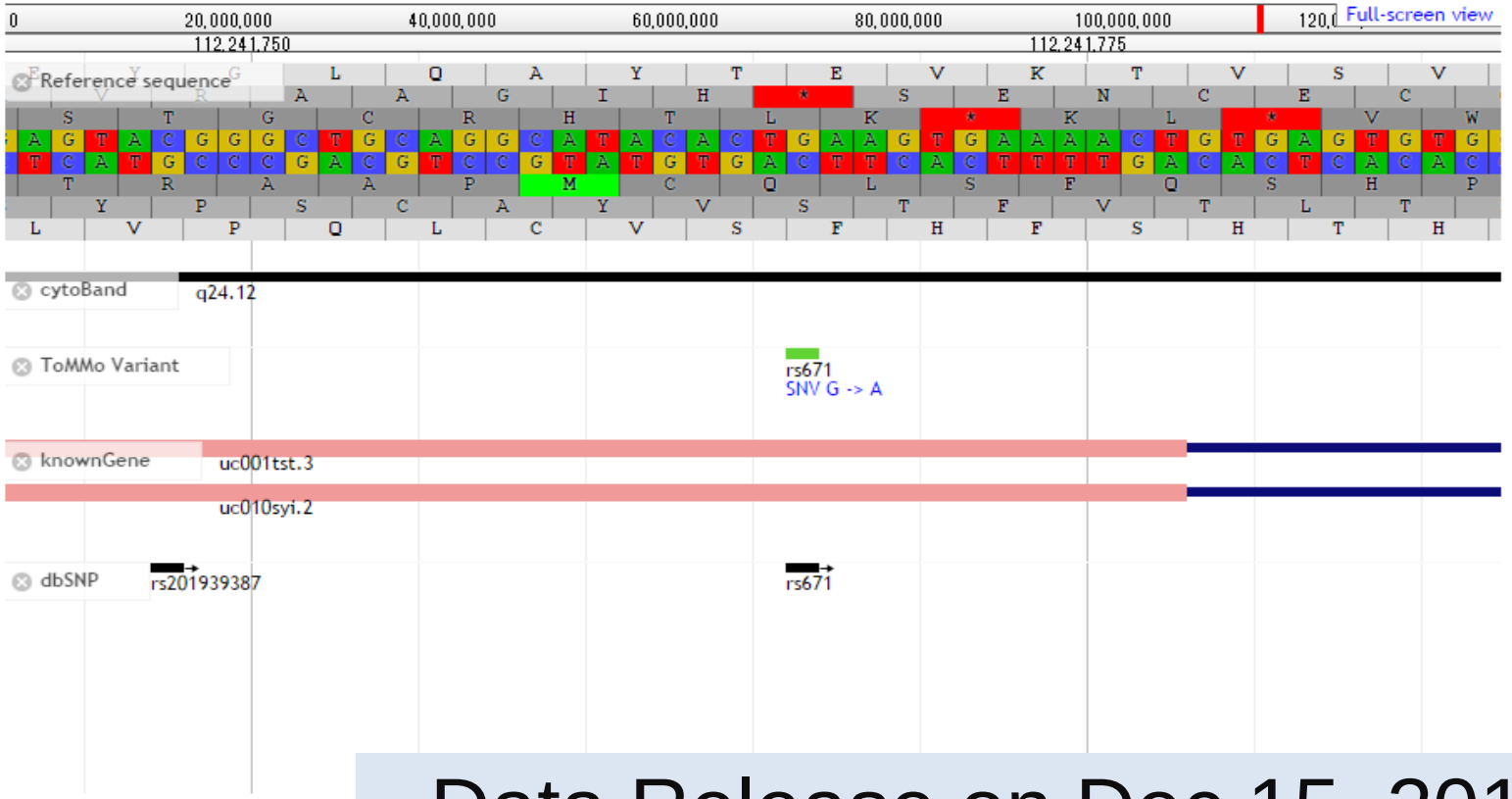
Japonica array (96sample)



INTEGRATIVE Japanese Genome Variation Database

The genome cohort study of Tohoku Medical Megabank Organization

- 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



iJGVD

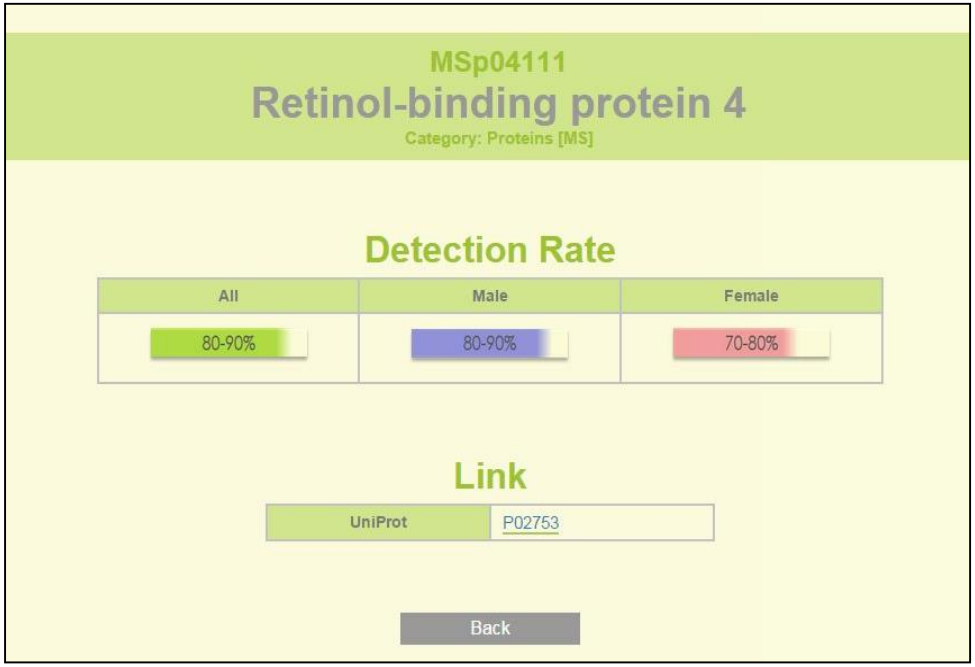
http://ijgvd.megabank.tohoku.ac.jp/

Data Release on Dec 15, 2015

Japanese Multi omics reference panel : jMorp

▲▼ToMMo Compound ID	▲▼Name	▲▼Category	▲▼Mean	▲▼SD	▲▼CV	▲▼N	▲▼Notes
TCC000019	Tetradecanedioic acid	Metabolites [MS]	3.16e+2	3.73e+2	1.18e+0	501	*provisional*
TCC000058	PG(15:0/0:0)	Metabolites [MS]	7.59e+1	1.01e+2	1.33e+0	489	
TCC000204	Biocytin	Metabolites [MS]	3.42e+1	6.25e+1	1.83e+0	499	
TCC000256	9,12,13-TriHOME	Metabolites [MS]	4.81e+1	9.58e+1	1.99e+0	496	*provisional*
TCC000776	L-leucyl-L-proline	Metabolites [MS]	4.59e+1	3.18e+1	6.92e-1	501	
TCC000800	LysoPE(18:1(11Z)/0:0)	Metabolites [MS]	1.41e+2	2.56e+2	1.81e+0	479	
TCC000821	(Z)-3-Oxo-2-(2-pentenyl)-1-cyclopenteneacetic acid	Metabolites [MS]	9.36e+2	9.21e+2	9.85e-1	501	

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 interacts with genetic system

Environment

Exposures, Nutrition, Lifestyle

Gene

Metabolic pathways, Signaling pathways

Phenotype (disease)

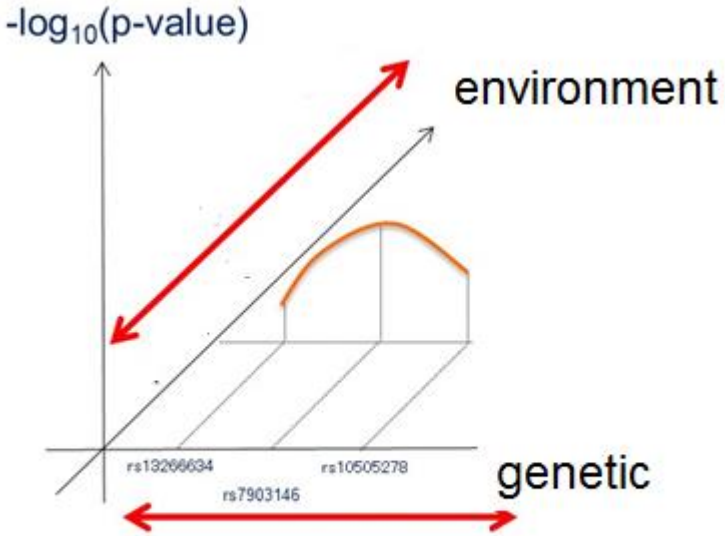
Precise Stratification

Personalized prevention

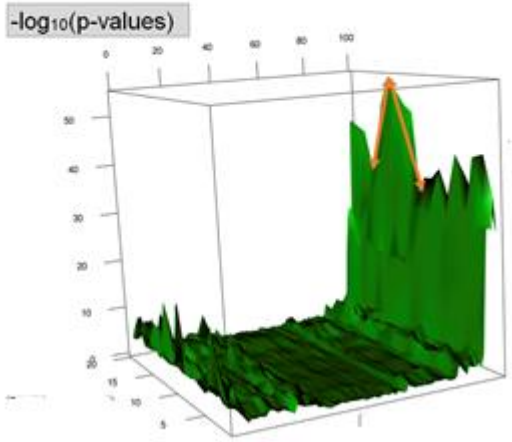
Idiosyncratic Effect of Combination of GxE factors

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

		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



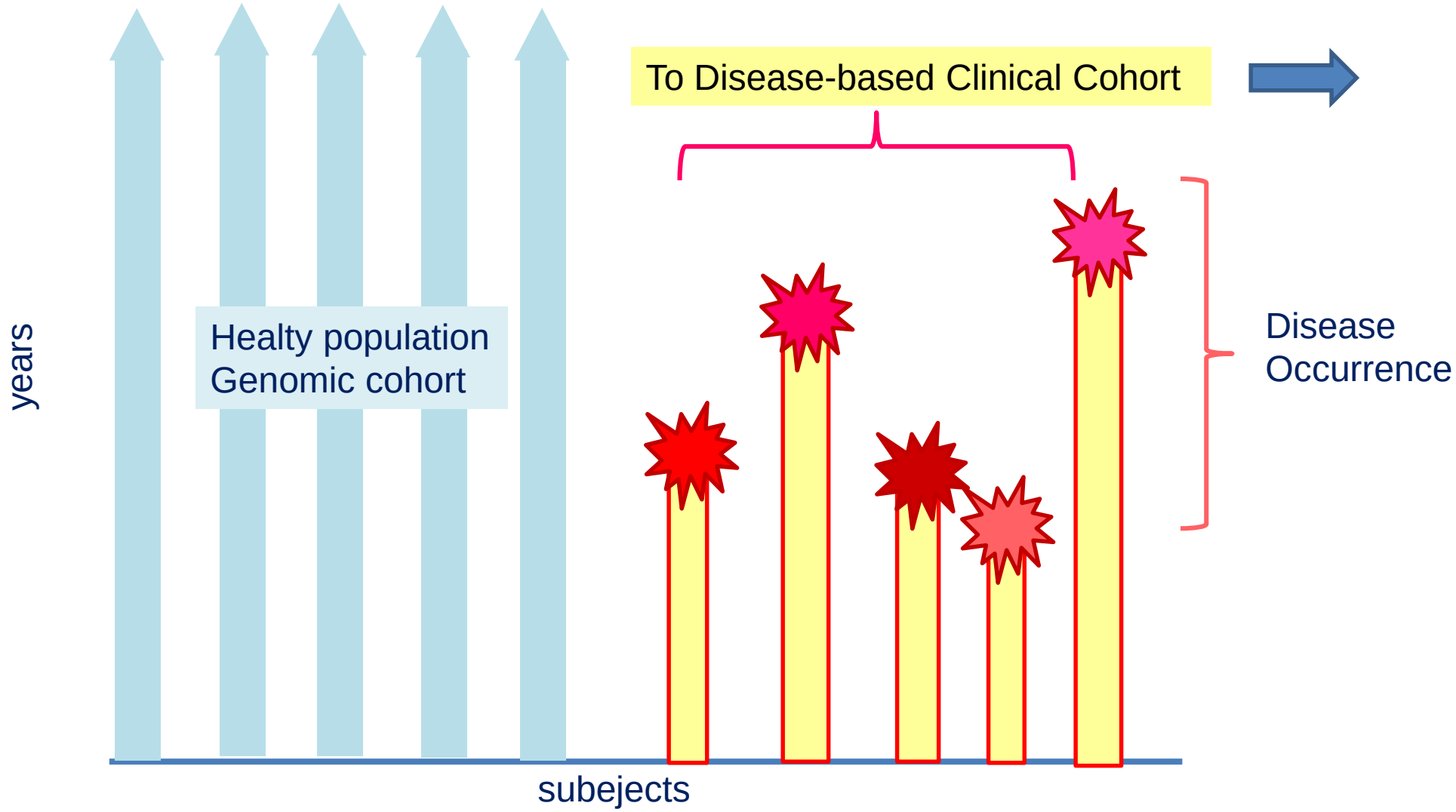
Simulation result



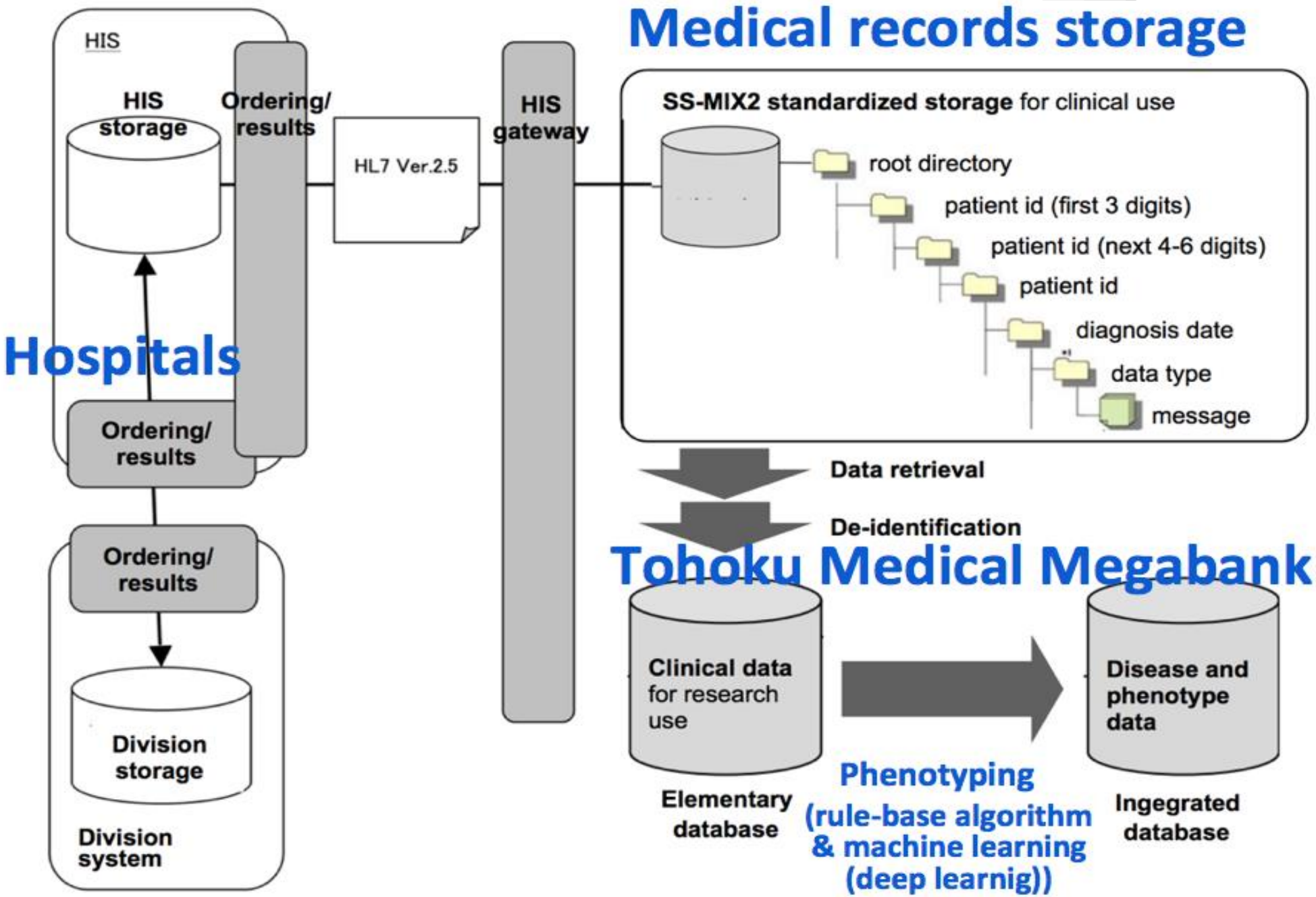
L. Le Marchand, JH. Hankin, LR. Wilkens, et alCombined Effects of Well-done Red Meat, Smoking, and Rapid N-Acetyltransferase 2 and CYP1A2 Phenotypes in Increasing Colorectal Cancer Risk, Cancer Epidemiol. Biomarkers Prev 2001;10:1259-1266

Towards Disease-oriented Biobank

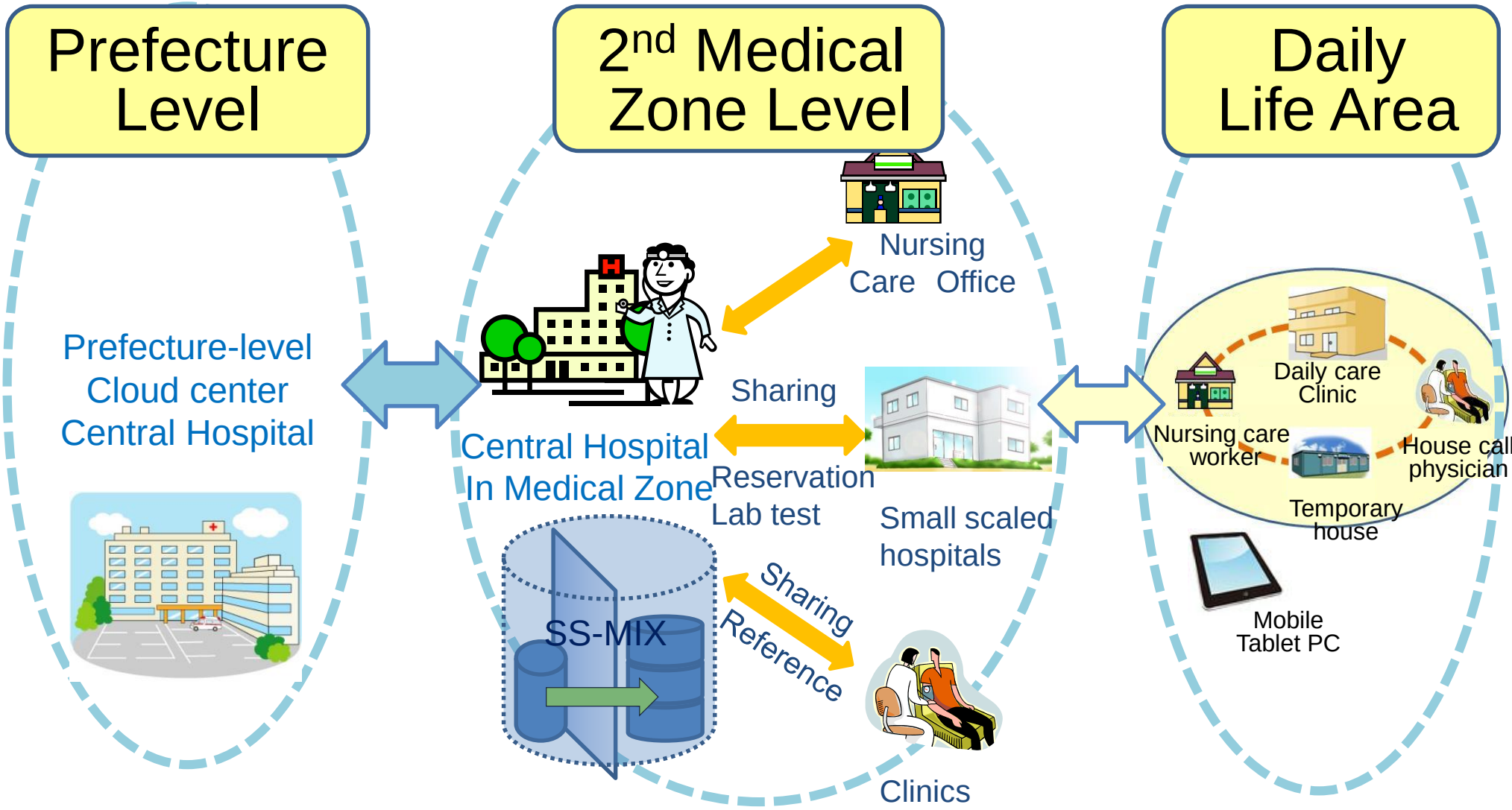
Population and Disease-based Mixed Genome Cohort

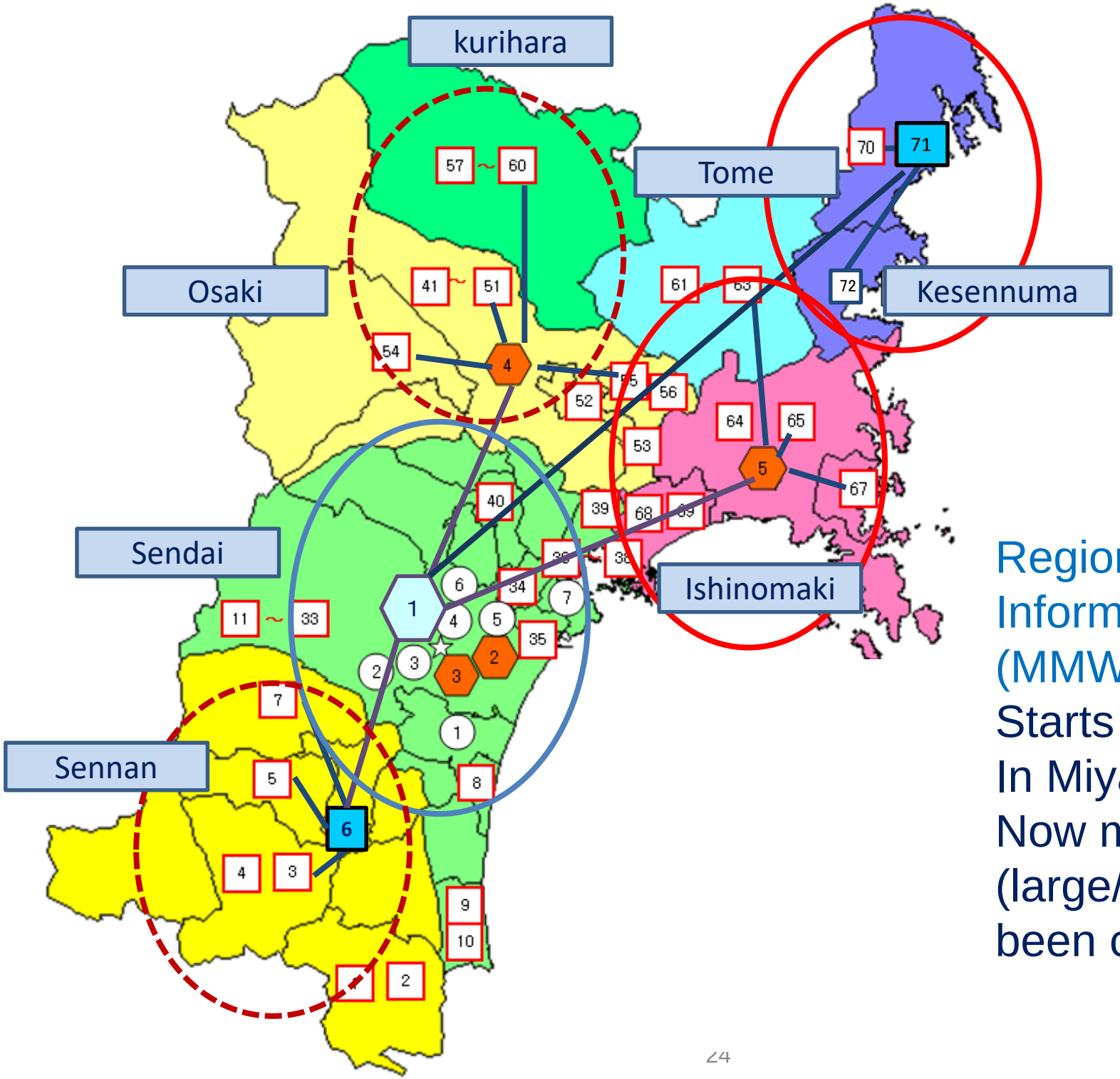


Clinical data collection and phenotyping



Disaster-tolerant Multi-hierarchical Regional Healthcare IT system





Regional Healthcare Information Network (MMWIN)

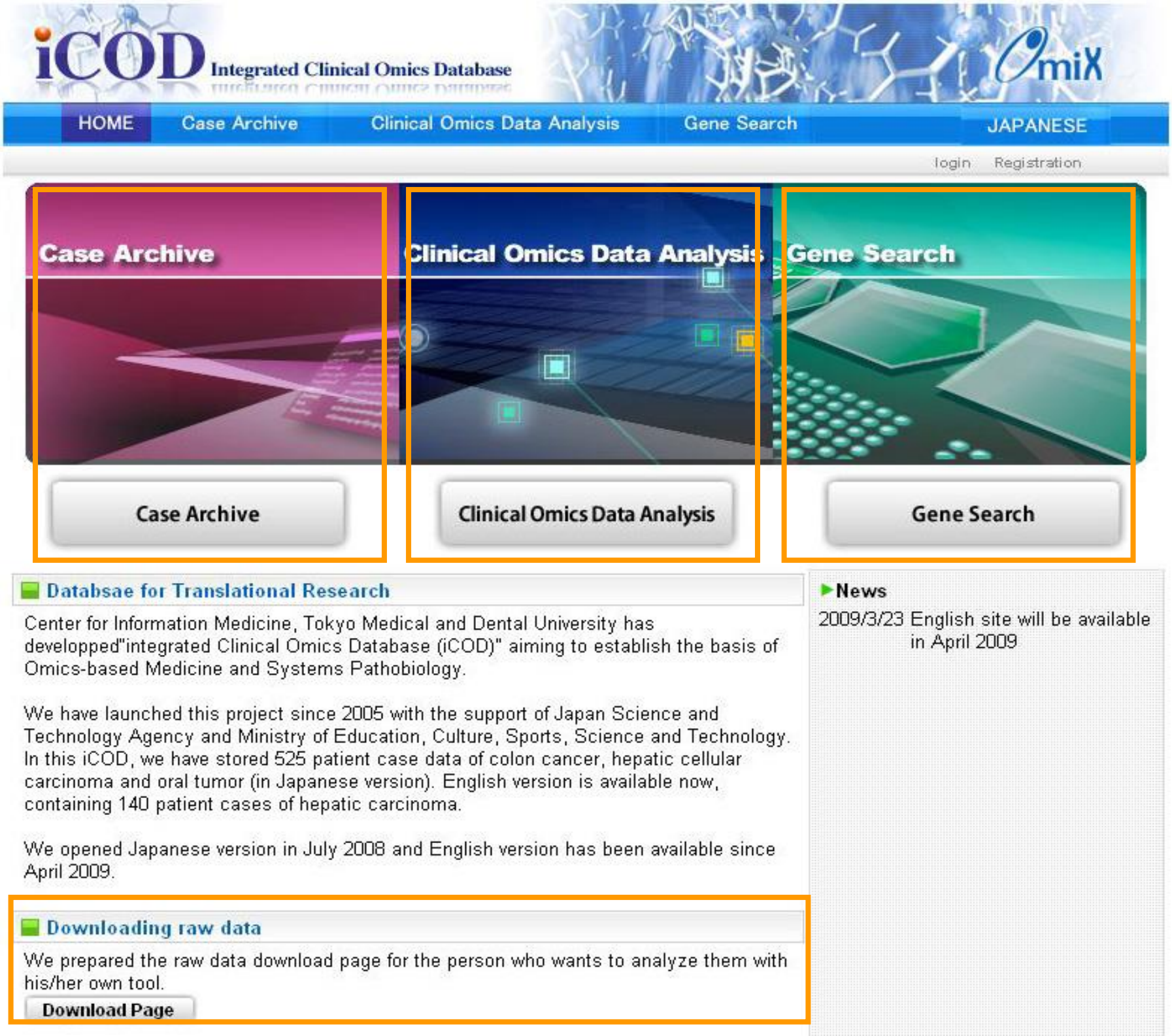
Starts August 10, 2012

In Miyagi Prefecture

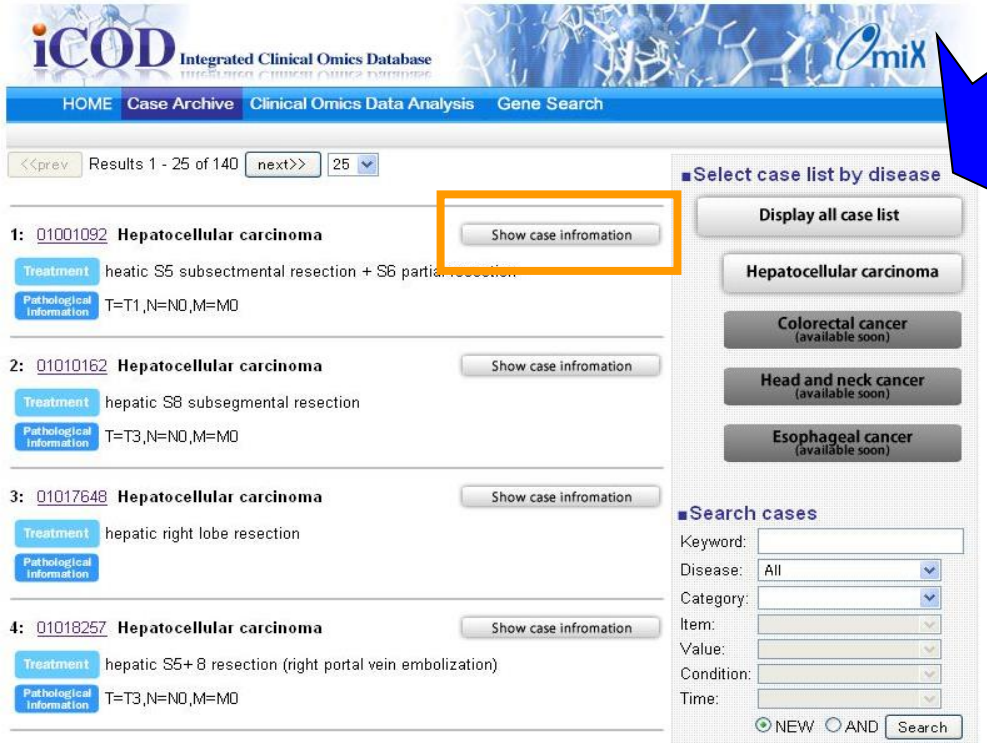
Now most of hospitals (large/medium size) have been connected

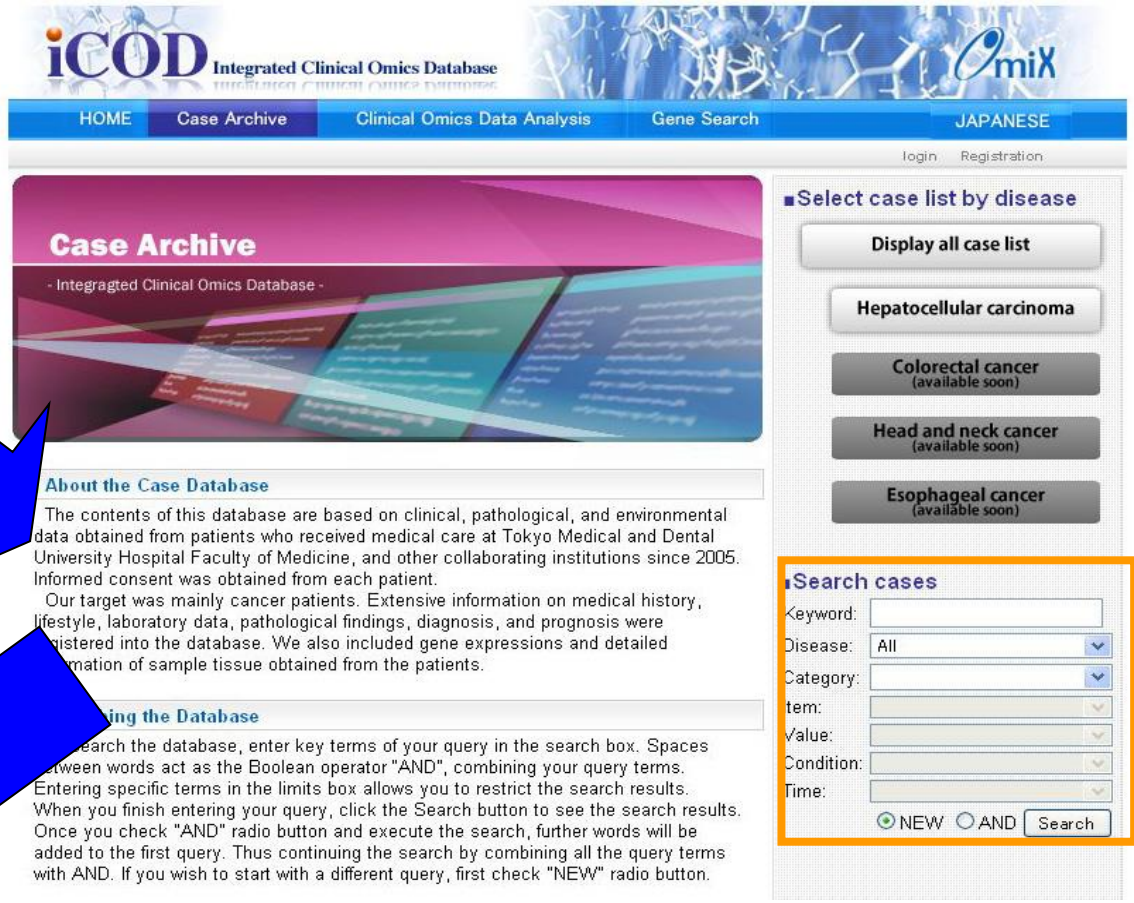
Clinical genome/omics database

- 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)

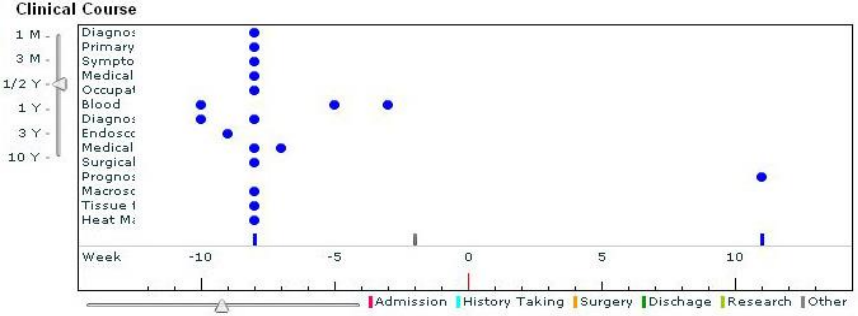

iCOD
Integrated Clinical Omics Database
OmiX
HOME
Case Archive
Clinical Omics Data Analysis
Gene Search
JAPANESE
login
Registration
Case Archive
Clinical Omics Data Analysis
Gene Search
Case Archive
Clinical Omics Data Analysis
Gene Search
Databsae for Translational Research
Center for Information Medicine, Tokyo Medical and Dental University has developped"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
News
2009/3/23 English site will be available in April 2009


iCOD Integrated Clinical Omics Database
OmiX
HOME
Case Archive
Clinical Omics Data Analysis
Gene Search
JAPANESE
login
Registration
Case Archive
Clinical Omics Data Analysis
Gene Search
Case Archive
Clinical Omics Data Analysis
Gene Search
Databsae for Translational Research
Center for Information Medicine, Tokyo Medical and Dental University has developped"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.
News
2009/3/23 English site will be ava
in April 2009


iCOD Integrated Clinical Omics Database
OmiX
HOME
Case Archive
Clinical Omics Data Analysis
Gene Search
<<prev
Results 1 - 25 of 140
next>>
25
1: 01001092 Hepatocellular carcinoma
Show case infromation
Treatment
heatic S5 subsectmental resection + S6 parti
Pathological Information
T=T1,N=N0,M=M0
2: 01010162 Hepatocellular carcinoma
Show case infromation
Treatment
hepatic S8 subsegmental resection
Pathological Information
T=T3,N=N0,M=M0
3: 01017648 Hepatocellular carcinoma
Show case infromation
Treatment
hepatic right lobe resection
Pathological Information
4: 01018257 Hepatocellular carcinoma
Show case infromation
Treatment
hepatic S5+8 resection (right portal vein embolization)
Pathological Information
T=T3,N=N0,M=M0
Select case list by disease
Display all case list
Hepatocellular carcinoma
Colorectal cancer
(available soon)
Head and neck cancer
(available soon)
Esophageal cancer
(available soon)
Search cases
Keyword:
Disease:
All
Category:
Item:
Value:
Condition:
Time:
NEW
AND
Search


iCOD Integrated Clinical Omics Database
OmiX
HOME
Case Archive
Clinical Omics Data Analysis
Gene Search
JAPANESE
login
Registration
Case Archive
- Integragted Clinical Omics Database -
About the Case Database
The contents of this 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
istered into the database. We also included gene expressions and detailed
mation of sample tissue obtained from the patients.
ing the Database
earch the database, enter key terms of your query in the search box. Spaces
tween 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.
Select case list by disease
Display all case list
Hepatocellular carcinoma
Colorectal cancer
(available soon)
Head and neck cancer
(available soon)
Esophageal cancer
(available soon)
Search cases
Keyword:
Disease:
All
Category:
tem:
Value:
Condition:
Time:
NEW
AND
Search

Clinical Course



01010162

Summary

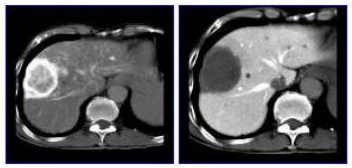
HCC (T3N0M0 StageIII, 48mm diameter in anterior segment of rt hepatic lobe, solitary), 70's male, LC(Child-Pugh A(5), Liver damage B (ICGR15 15.3%, PT% 77%))

Diagnosis

Hepatocellular carcinoma

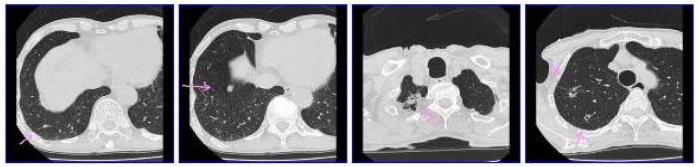
Clinical data

Angiography



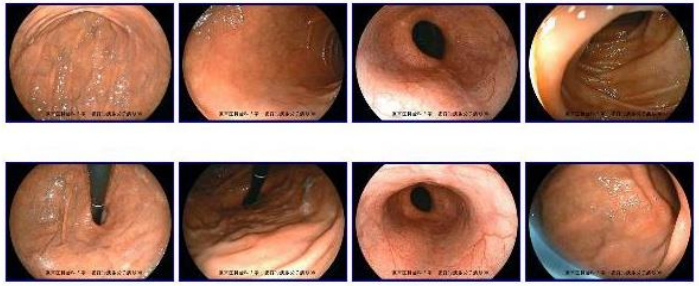
Site of Examination	abdominal angiography
Comments	HCCs in the hepatic right lobe segment.

CT



Site of Examination	lung
Comments	Diagnosis: HCC. 1. Two nodules in the basal segment of the right lung → Metastasi nodules with bronchiectasis in the right S1 → findings also indicate lung cancer. 3. I right lung → possibly healed inflammation. 4. Healed inflammation in the right S2, ri lung.

Endoscopic Screening Upper



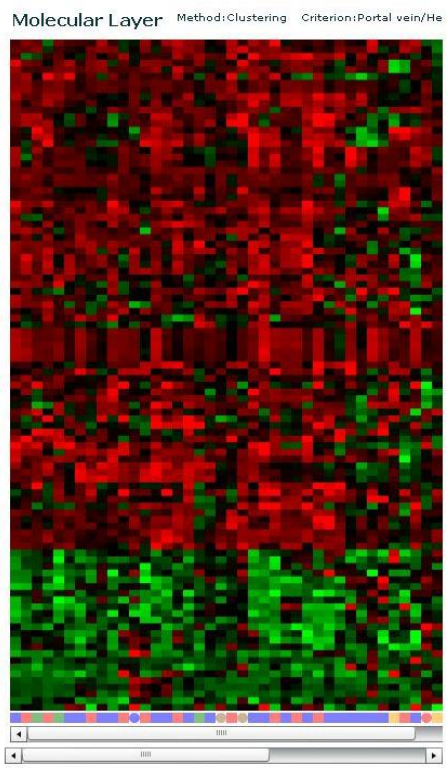
Macroscopic findings

Number	1.0
Tumor multiplicity	single
Maximum diameter	5.5
Degree of hepatic damage	B
Growth type	Eg
Capsule formation	-
Capule infiltration	+
Portal vein invasion	0
Hepatic vein invasion	0
Hepatic artery invasion	0
Bile duct invasion	0
T stage	T3
N stage	N0
M stage	M0

Pathological Data

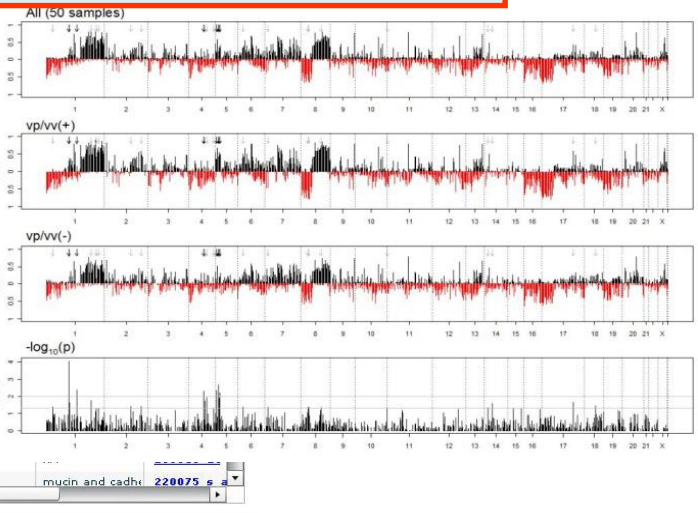
Diagnosis	Hepatocellular carcinoma
Differentiation	poor
Grade	scirrhous
Growth type	ig
Capsule formation	-
Capsule infiltration	+
Portal vein invasion	1
Hepatic vein invasion	0
Portal vein/ hepatic vein invasion	+
Hepatic artery invasion	0
Bile duct invasion	0
	-
stage)	III
non-cancerous region	NL

L1=Portal vein/Hepatic vein invasion / **L2=**Outcome



L1: - / L2: disease-free survival
L1: - / L2: prevalent survival
L1: - / L2: death of hepatic failure
L1: - / L2: death of unrelated disease
L1: - / L2: death of present disease
L1: - / L2: null
L1: + / L2: dis
L1: + / L2: pre
L1: + / L2: dea
L1: + / L2: dea
L1: + / L2: death of pre
L1: + / L2: null
L1: null / L2: disease-fr
L1: null / L2: prevalent
L1: null / L2: death of h
L1: null / L2: death of
L1: null / L2: death of p
L1: null / L2: null

Molecular Data



GeneSymbol	Entre
BSDC1	55108
BSDC1	55108
DNAJC16	23341
NA	NA
NA	NA
CARKL	23729
HPX	3263
HPX	3263
F11	2160
NA	NA
MUCDHL	53841
mucin and cadh	220075 s a

Clinical Omics Data Analysis

- 2 Dimentional – 3 Layerd 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

Databsae for Translational Research

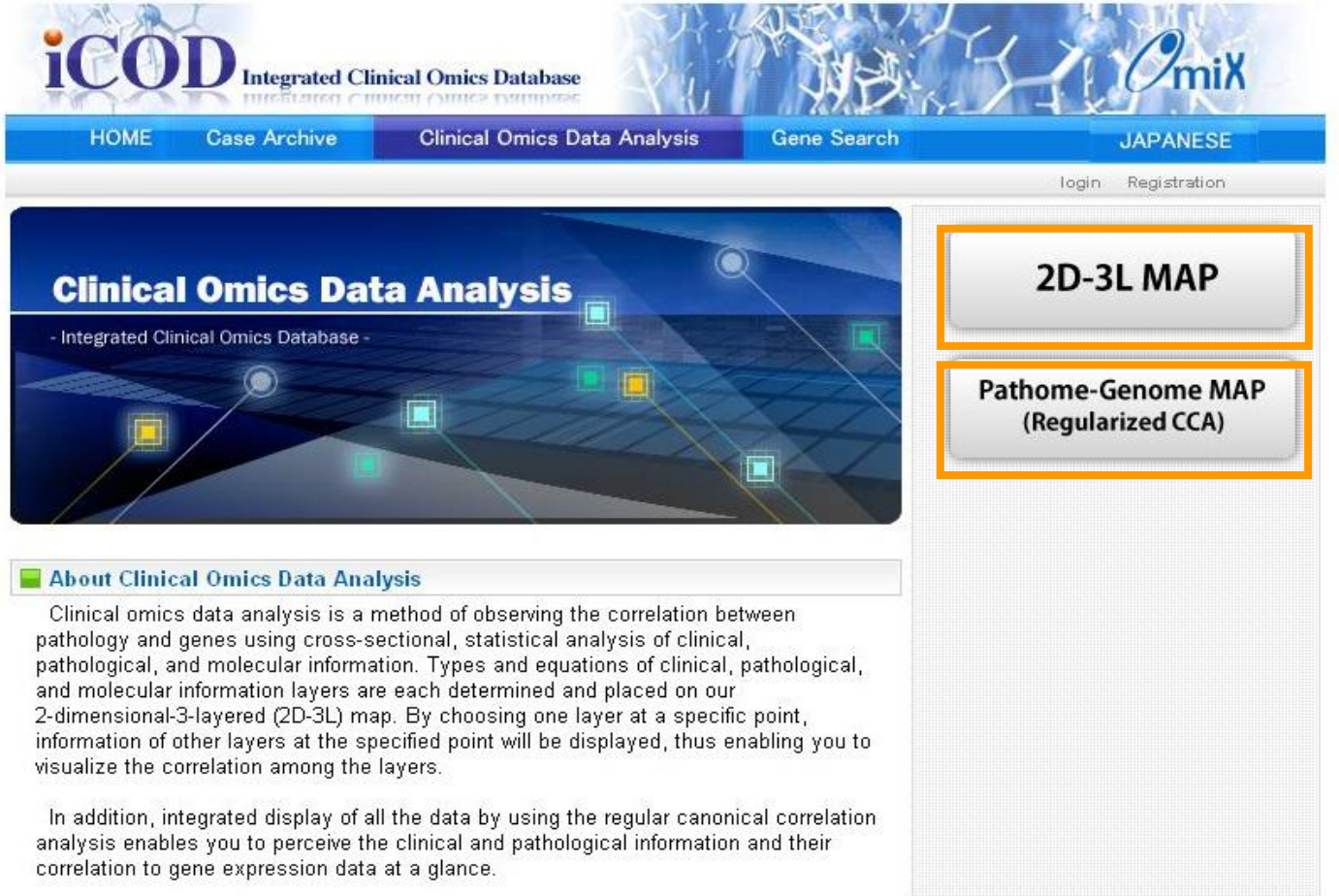
Center for Information Medicine, Tokyo Medical and Dental University has developped"integrated Clinical Omics Database (iCOD)" aiming to establish the basis of Omics-based Medicine and Systems Pathobiology.

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.

News

2009/3/23 English site will be available in April 2009



About 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.

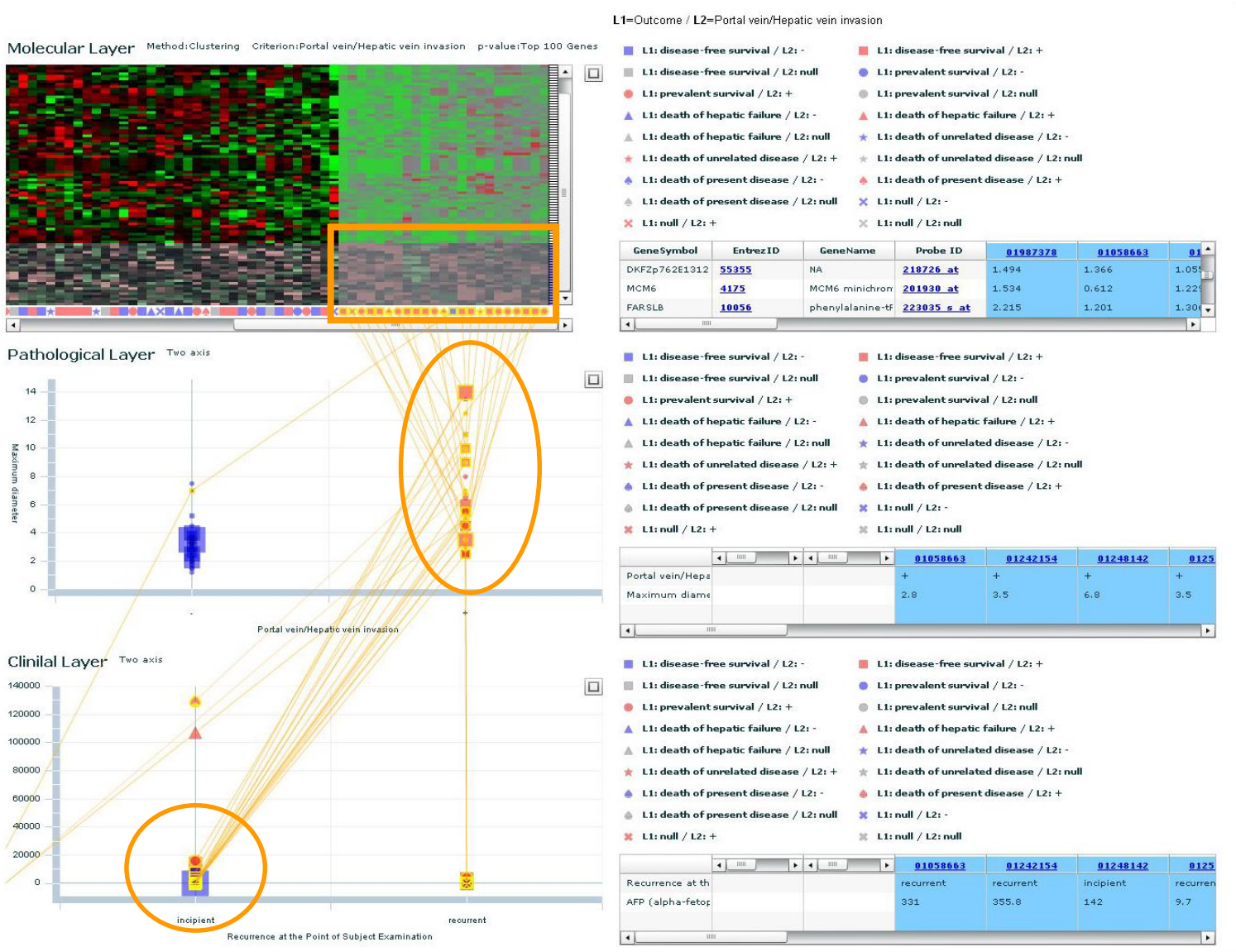
In addition, integrated display of all the data by using the regular canonical correlation analysis enables you to perceive the clinical and pathological information and their correlation to gene expression data at a glance.

2 Dimensional – 3 Layered Map

Molecular Layer

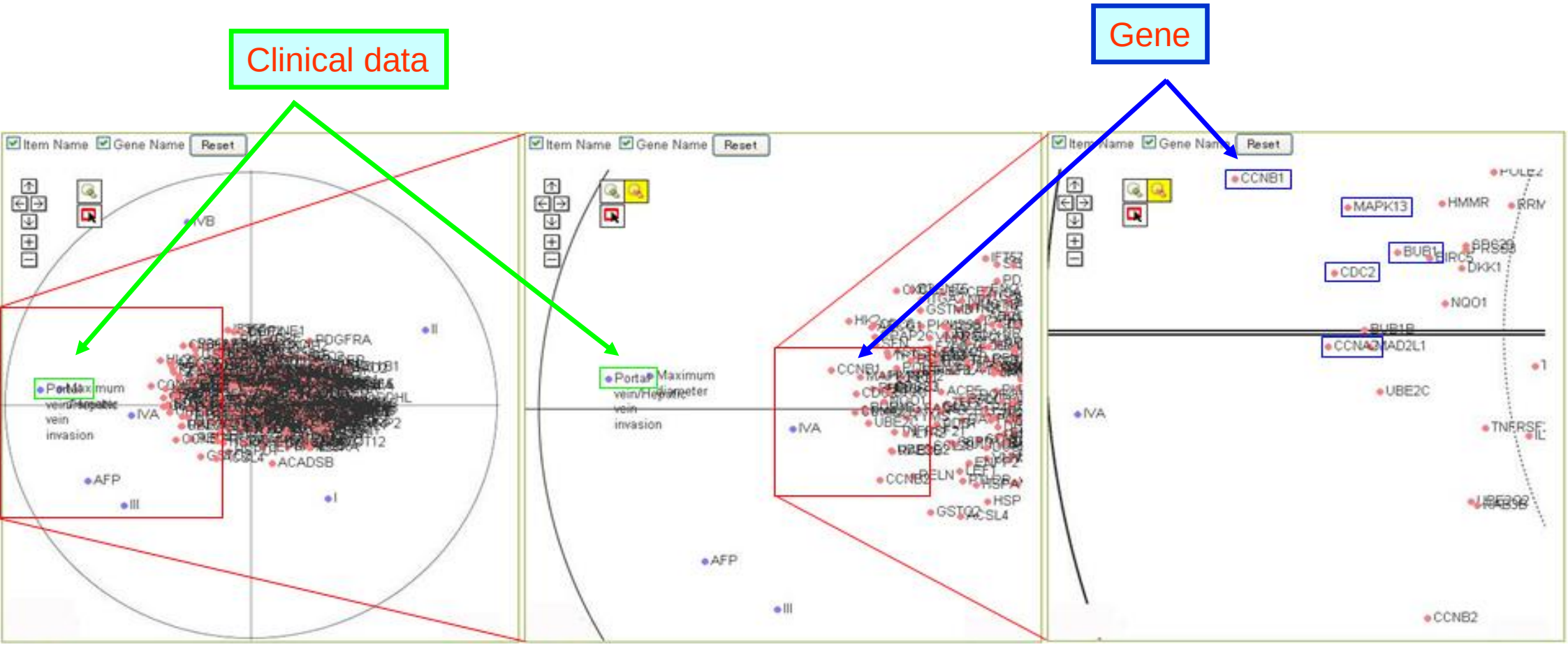
Pathological Layer

Clinical Layer

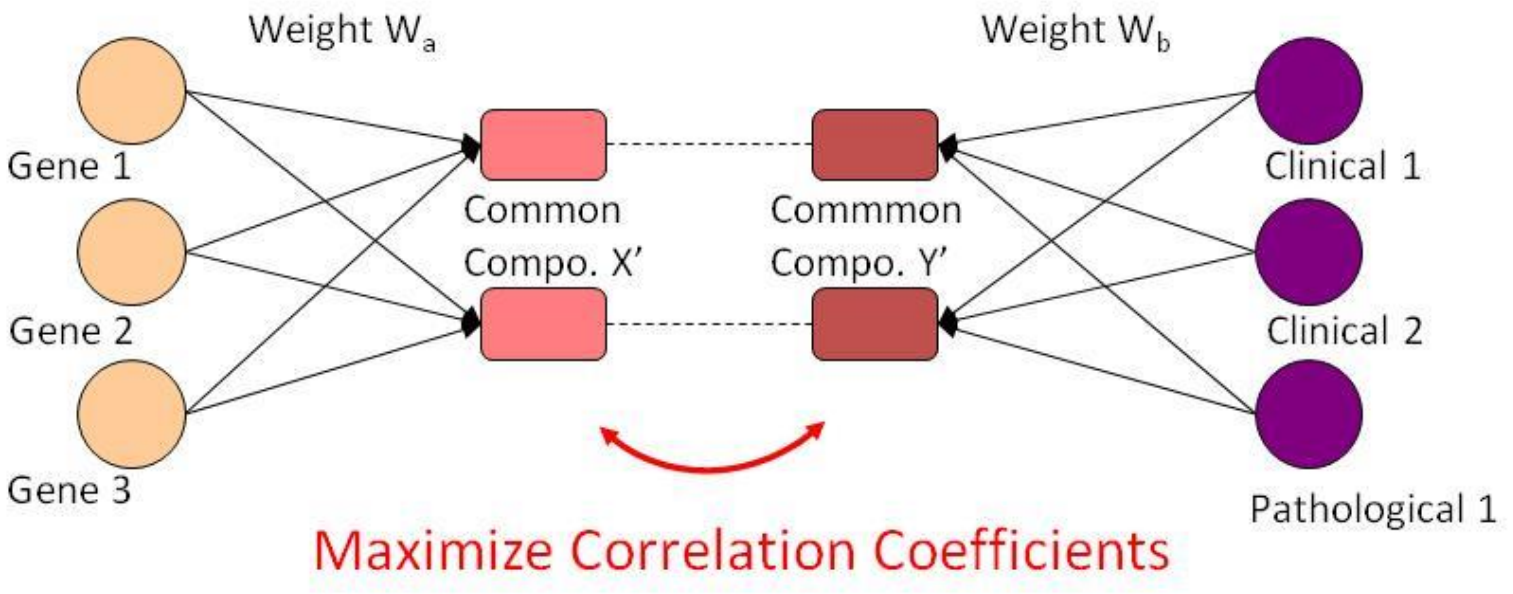


Pathome - Genome map

Display both Pathological states and Genes



Canonical Correlation Analysis

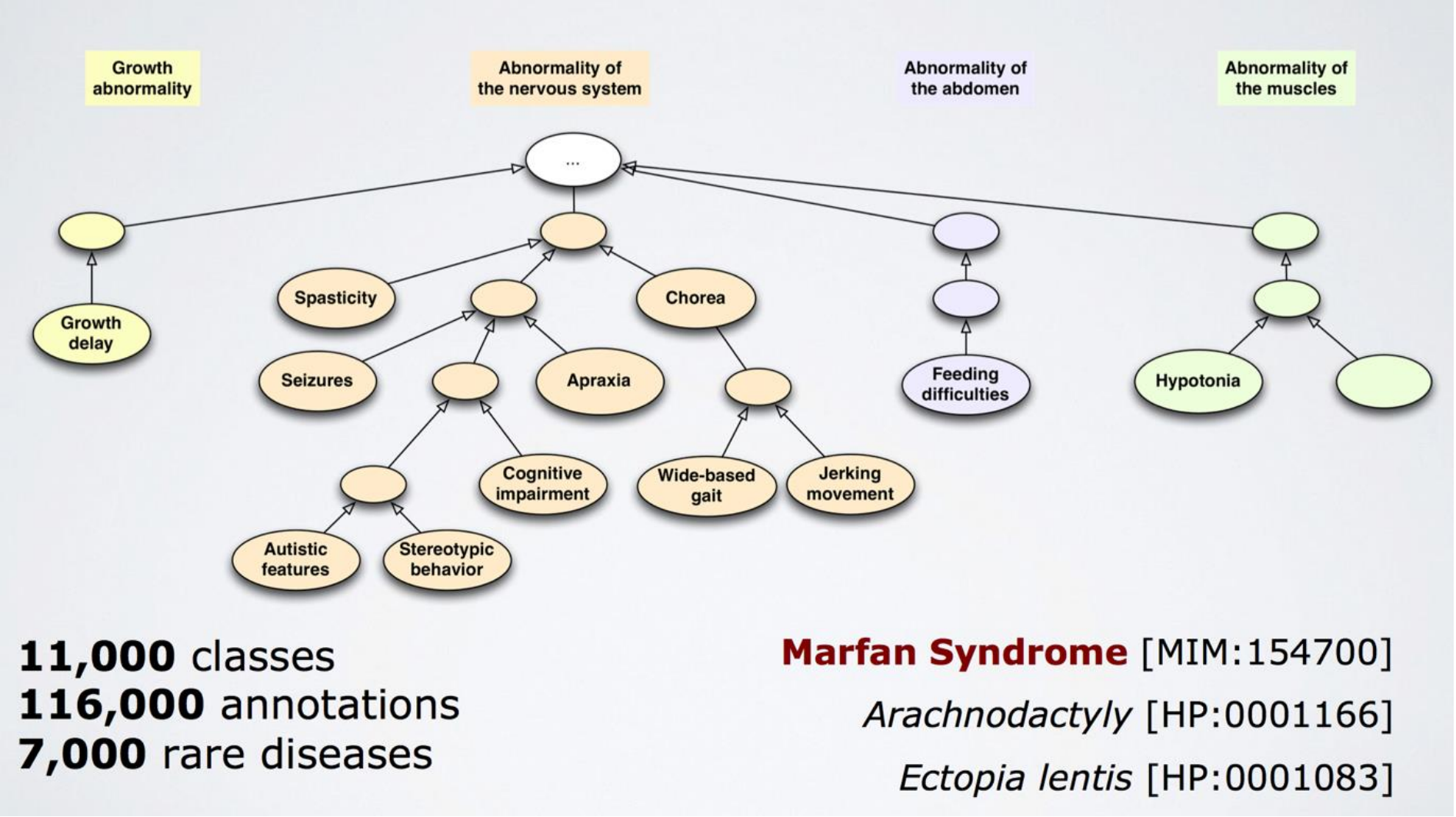


Gene expression X

Clinico-pathological Y

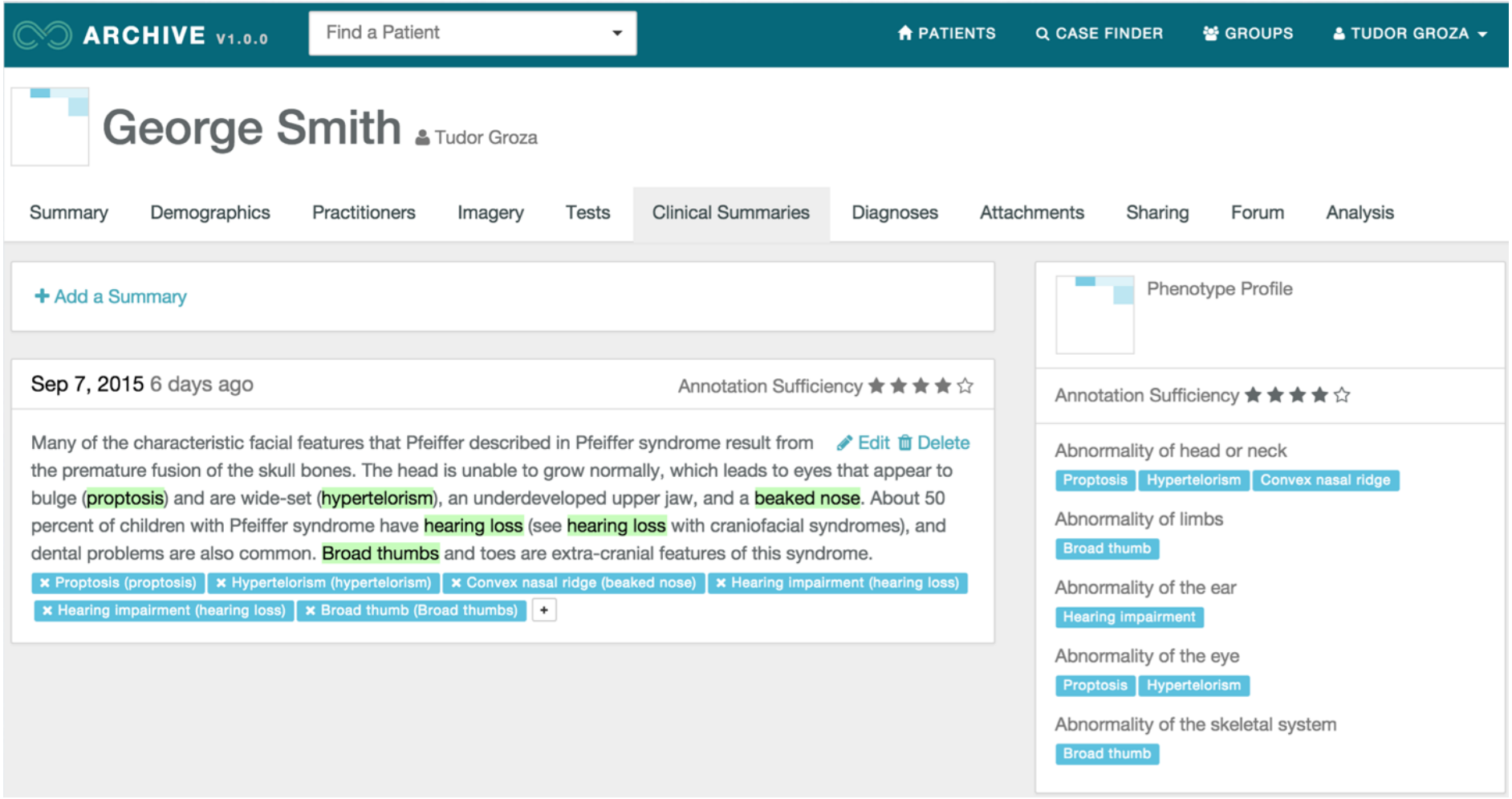
Human Phenotype Ontology (HPO)

Human Phenotype Ontology (HPO) is a controlled vocabulary used to describe phenotypic abnormalities seen in human disease



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