TOHOKU
UNIVERSITY

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

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

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

accounting

Laboratory system

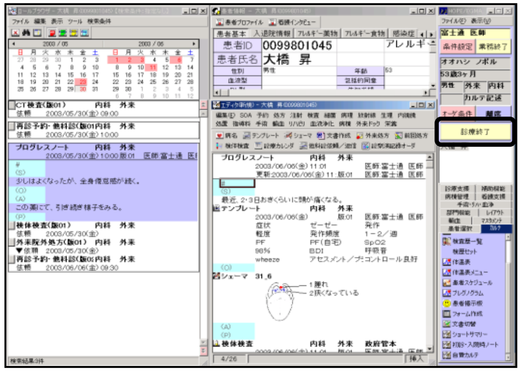
2nd generation: CPOE (Computerized Physician Order Entry): 1980s-

Order-entry/result reporting system of laboratory or radiological test, drug prescription

3rd generation: EHR/EMR : 2000s-

Electronic Health/Medical Record

DisplayScreen of EHR/EMR


患者ID 0099801045

患者氏名 大橋 昇

性別 男性 年齢 53

CT検査(版01) 内科 外来
依頼 2003/05/30(金) 09:00

再診予約・他科診(版01内科 外来
依頼 2003/05/30(金) 10:00

プログレスノート 内科 外来
2003/05/30(金) 10:00 版01 医師:富士通 医師

(S)
少しはよくなったが、全身倦怠感が続く。
(O)
(A)
この薬にて、引き続き様子をみる。
(P)

検体検査(版01) 内科 外来
依頼 2003/05/30(金)

外来院外処方(版01) 内科 外来
▼依頼 2003/05/30(金)

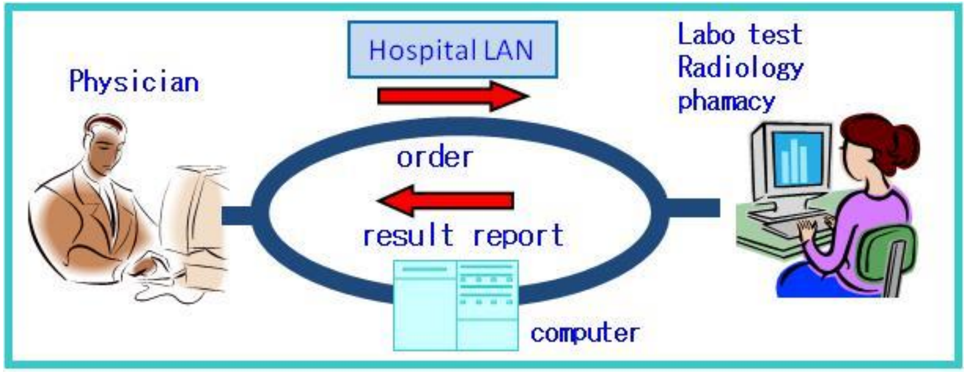
再診予約・他科診(版0 内科 外来
依頼 2003/06/06(金) 09:30

プログレスノート 内科 外来
2003/06/06(金) 11:01 医師:富士通 医師
更新 2003/06/06(金) 11: 版01 医師:富士通 医師

(S)
最近、2・3日おきぐらいに頭が痛くなる。

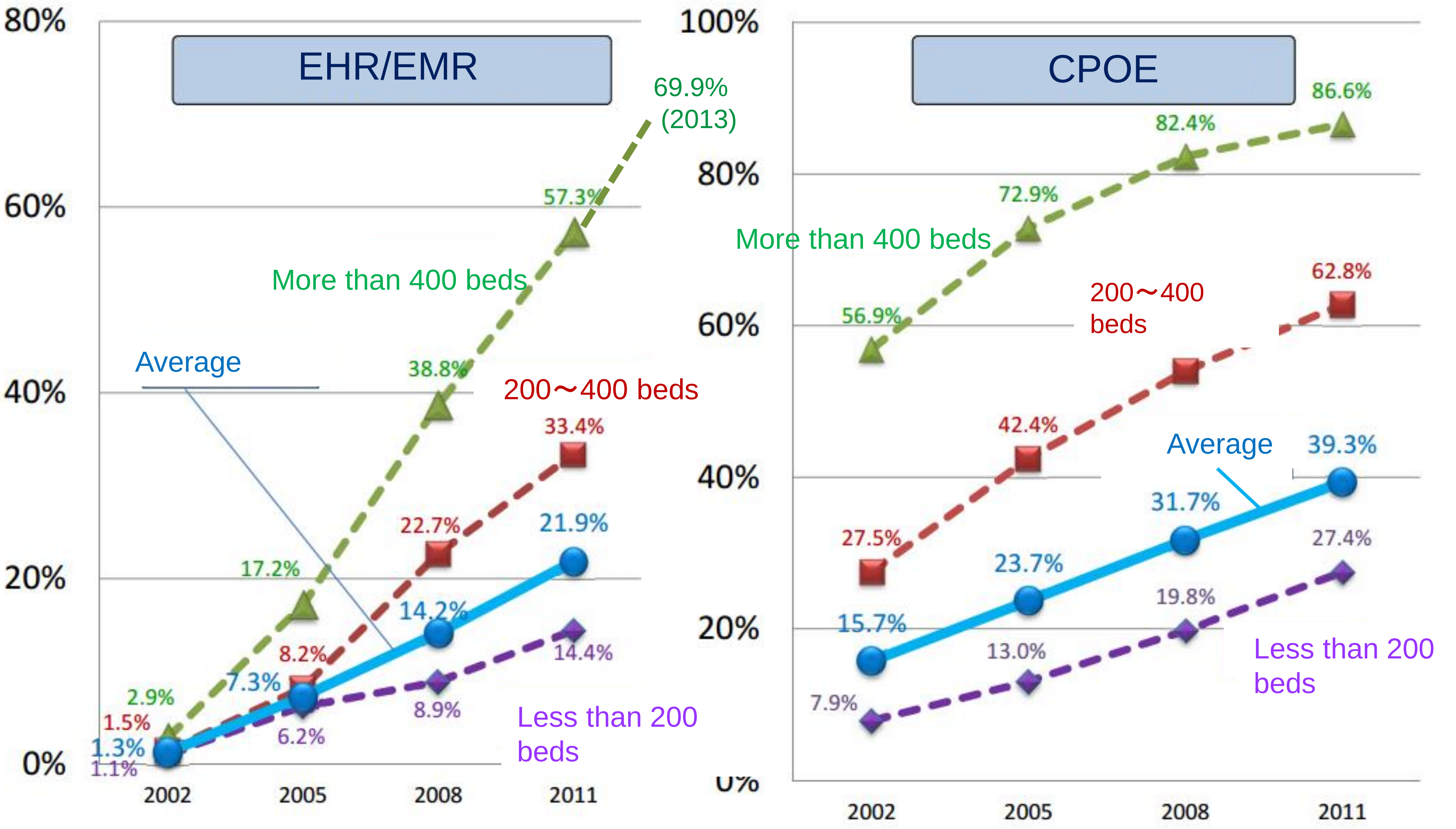
テンプレート 内科 外来
2003/06/06(金) 版01 医師:富士通 医師
症状 ゼーゼー 発作
軽度 発作頻度 1－2／週
PF PF(自宅) SpO2
98% BDI 呼吸音
wheeze アセスメント／プ: コントロール良好

(O)
シェーマ 31_6
1 腫れ
2 狭くなっている
(A)
(P)
検体検査 内科 外来 政府管本



Concept of CPOE

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
  - Established 2015.1, “Intermediate report”, 2015.7
  - 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

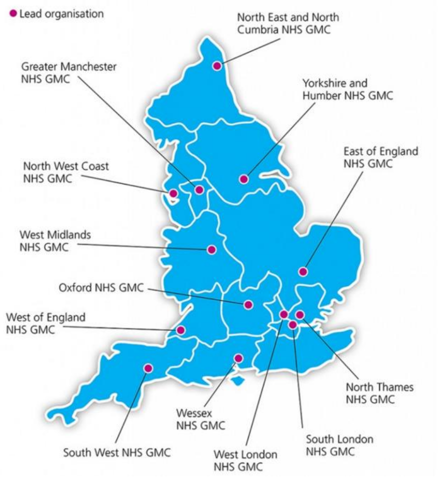
Obama's PMI

1. Clinical sequencing of germline (innate) genome
   - To find 'causative gene' of undiagnosed and inherited disease at POC (hospital)
   - 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
  - 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, 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 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.

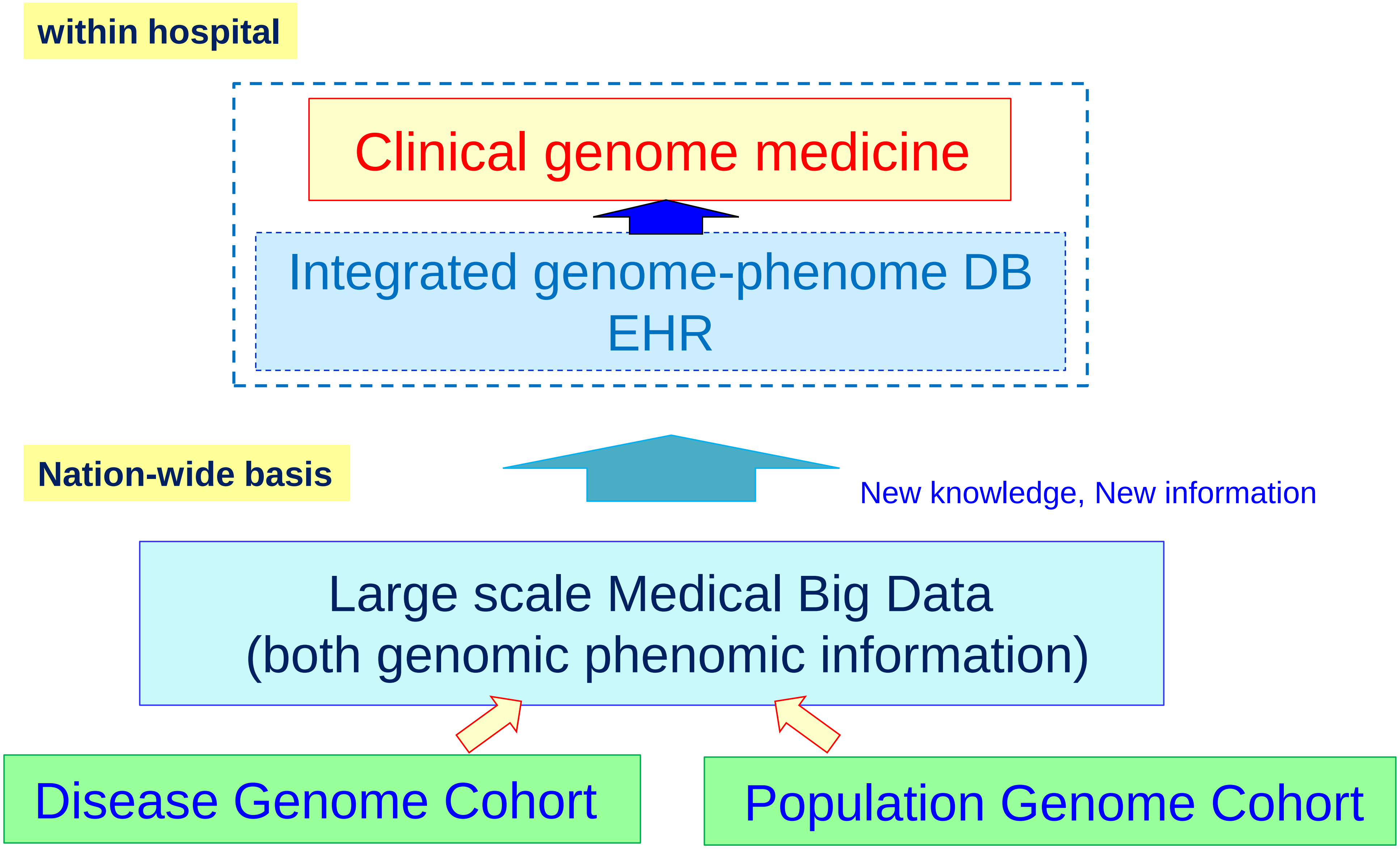


NHS Genome Medical Center
(Genomic England)

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

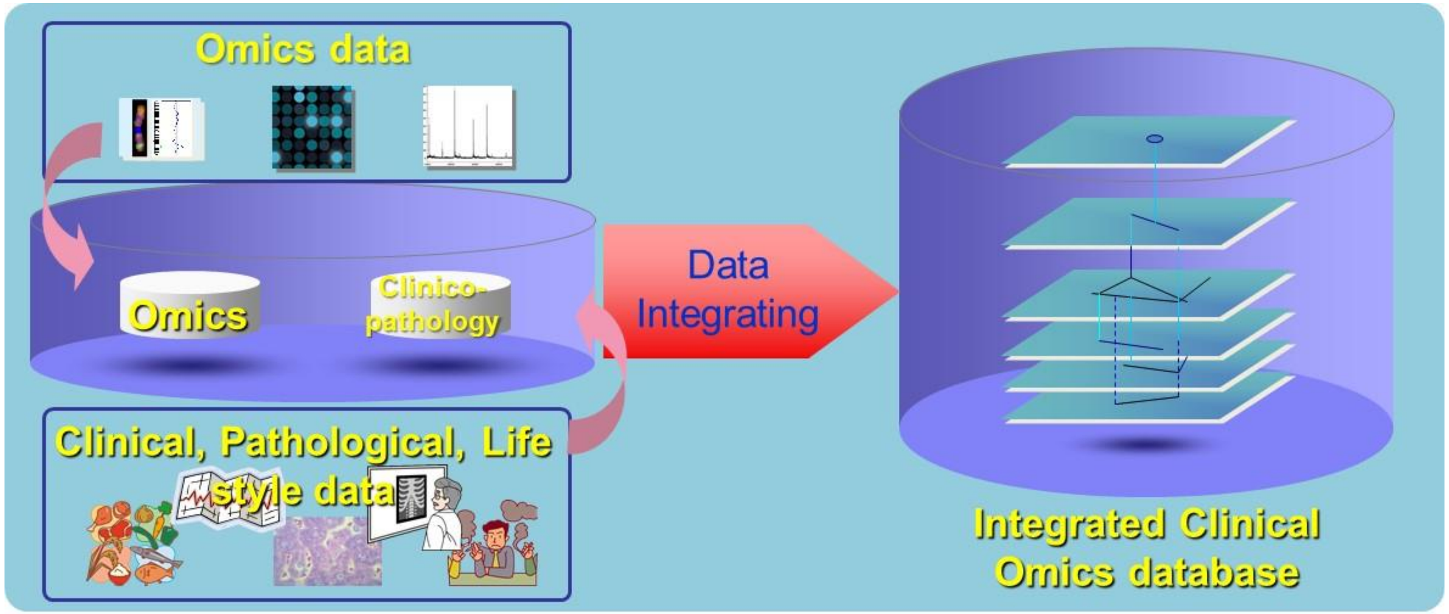
integrated Clinical Omics Database (iCOD)

Genome Medicine in Japan

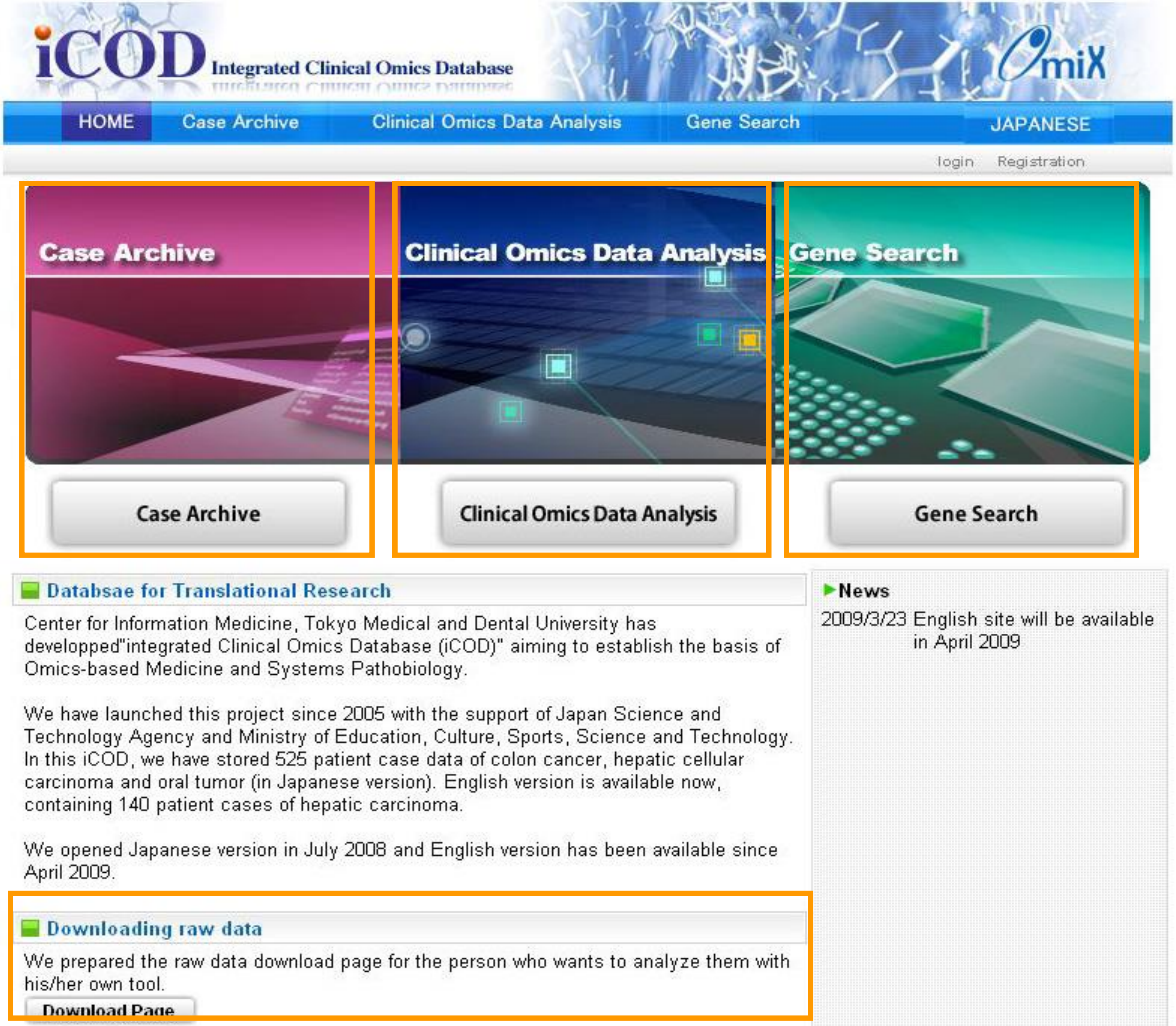
Integrated Clinical Omics Database (iCOD) Project of Japan (2005~)

- 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


Omics data
Omics
Clinico-pathology
Data Integrating
Clinical, Pathological, Life style data
Integrated Clinical Omics database

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

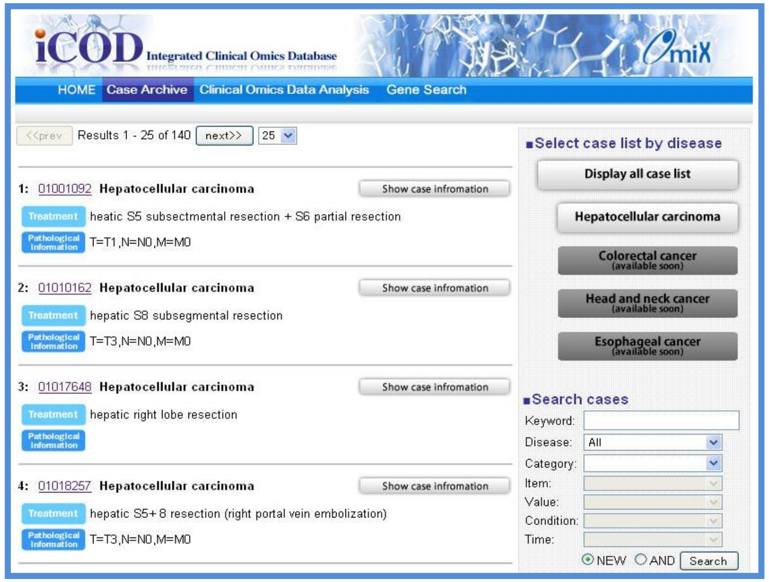
Downloading raw data

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

News

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

Case archive


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 partial resection
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

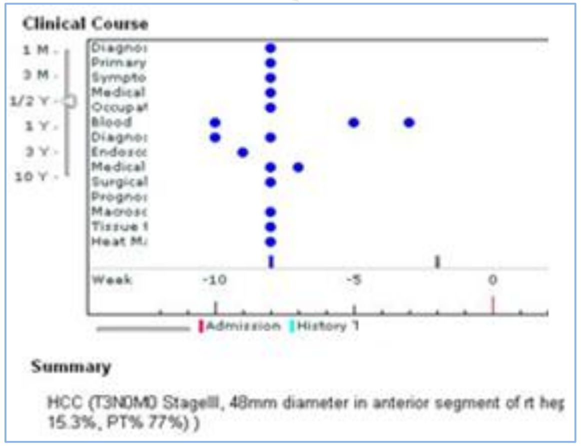
■Select case list by disease

- Display all case list
- Hepatocellular carcinoma
- Colorectal cancer (available soon)
- Head and neck cancer (available soon)
- Esophageal cancer (available soon)

■Search cases

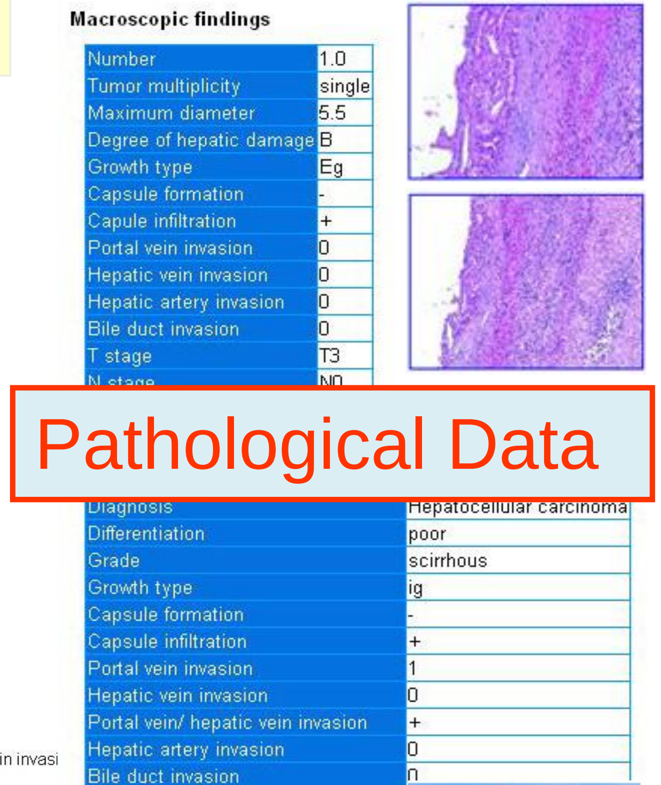
Keyword:
Disease: All
Category:
Item:
Value:
Condition:
Time:
● NEW ○ AND Search

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



Summary

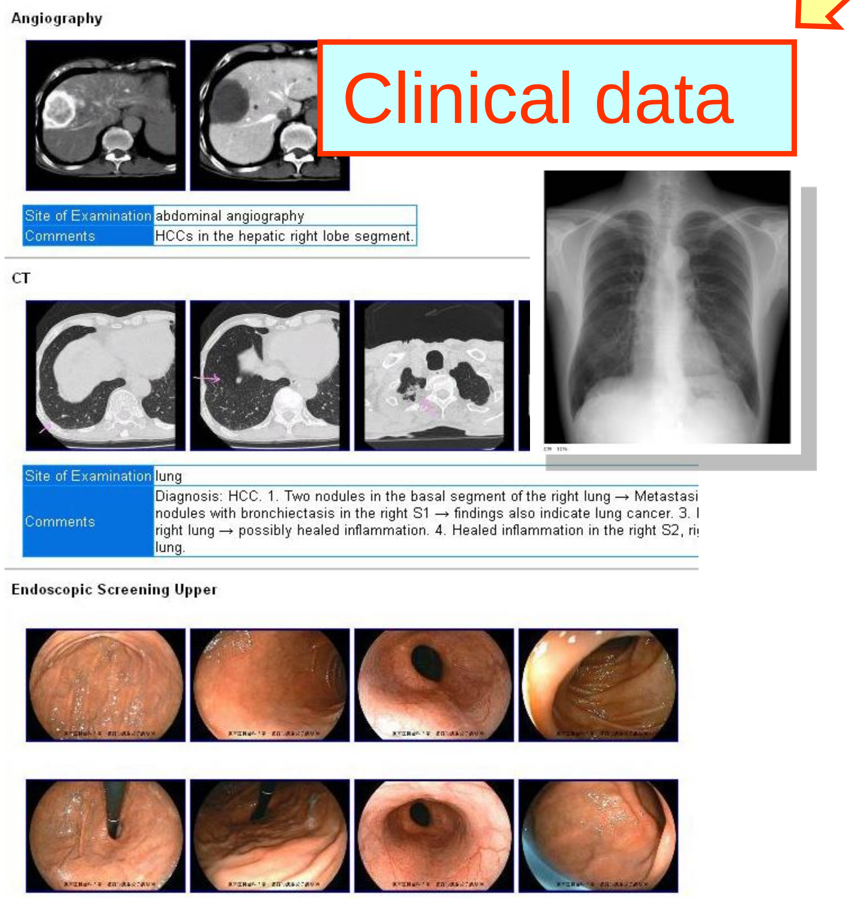
HCC (T3N0M0 StageII, 48mm diameter in anterior segment of rt hep 15.3%, PT% 77%))


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

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
SM	
pStage (TNM stage)	
Observation of non-cancerous region	

Pathological Data


Angiography

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

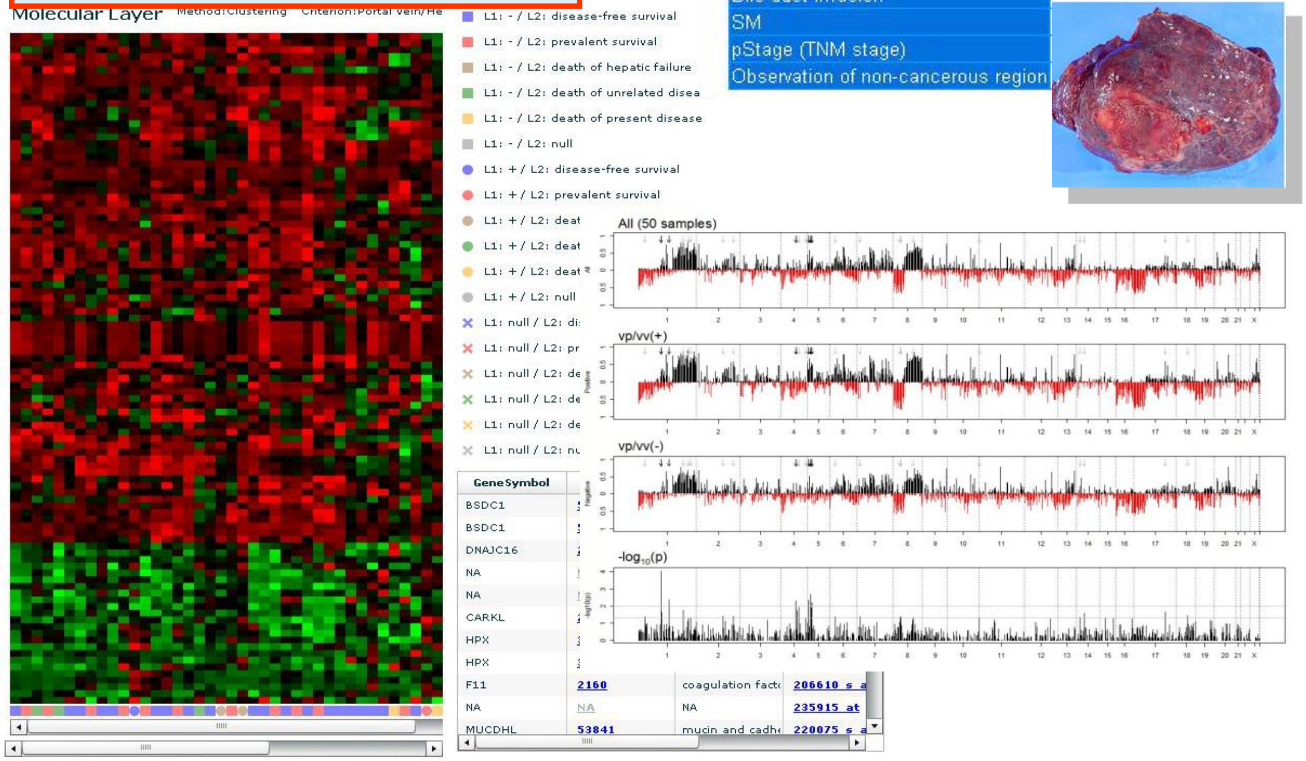
CT

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

Clinical data

Molecular Data



- L1: - / L2: disease-free survival
- L1: - / L2: prevalent survival
- L1: - / L2: death of hepatic failure
- L1: - / L2: death of unrelated disea
- L1: - / L2: death of present disease
- L1: - / L2: null
- L1: + / L2: disease-free survival
- L1: + / L2: prevalent survival
- L1: + / L2: deat
- L1: + / L2: deat
- L1: + / L2: deat
- L1: + / L2: null
- L1: null / L2: di
- L1: null / L2: pr
- L1: null / L2: de
- L1: null / L2: de
- L1: null / L2: de
- L1: null / L2: nu

Gene Symbol			
BSDC1			
BSDC1			
DNAJC16			
NA			
NA			
CARKL			
HPX			
HPX			
F11	2160	coagulation fact	206610_s_a
NA	NA	NA	235915_at
MUCDHL	53841	mucin and cadh	220075_s_a

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

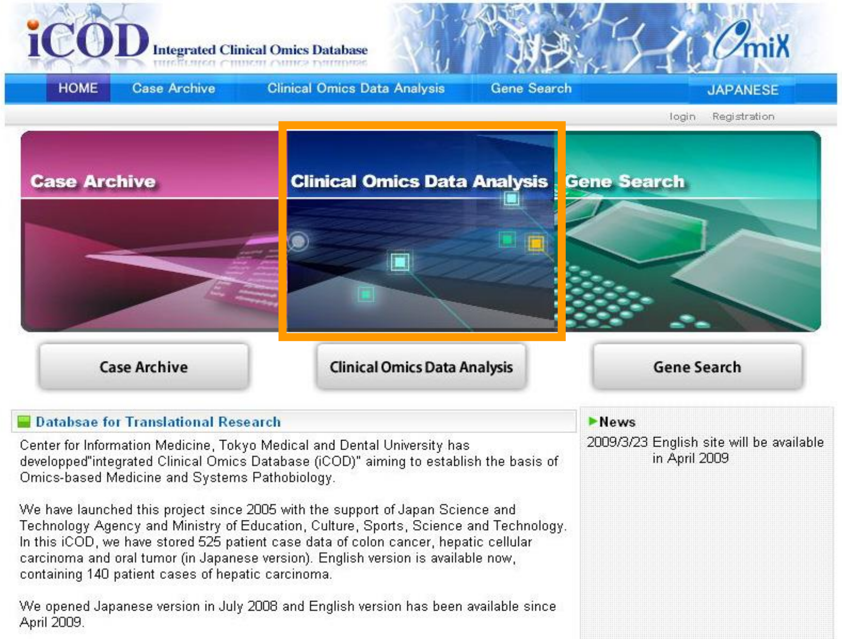
- Gene expression profile
- Copy number variation

(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**

Clinical Omics Data Analysis


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

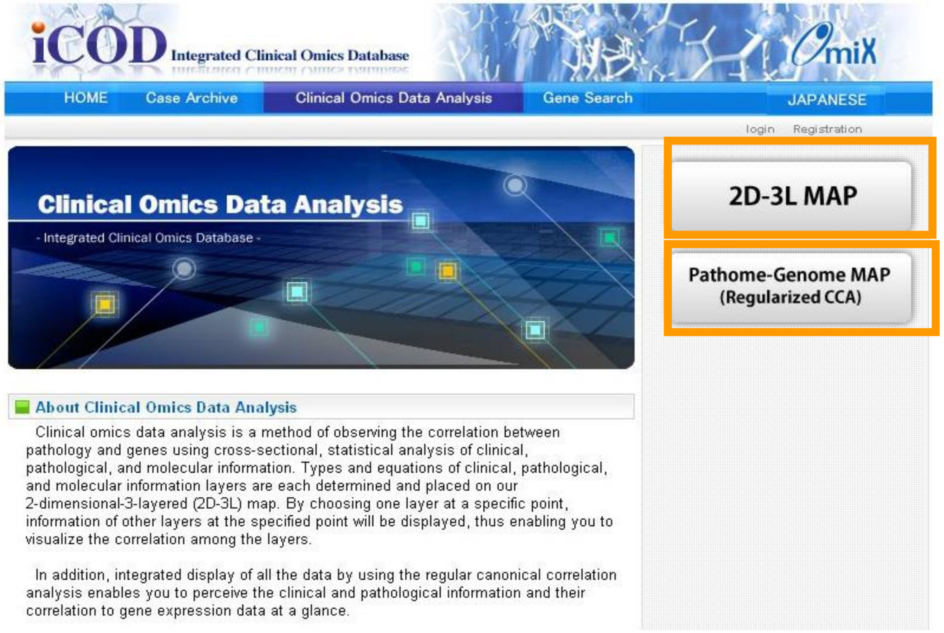
Center for Information Medicine, Tokyo Medical and Dental University has developped"integrated Clinical Omics Database (iCOD)" aiming to establish the basis of Omics-based Medicine and Systems Pathobiology.

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 available in April 2009


iCOD Integrated Clinical Omics Database
OmiX
HOME
Case Archive
Clinical Omics Data Analysis
Gene Search
JAPANESE
login
Registration
Clinical Omics Data Analysis
- Integrated Clinical Omics Database -
2D-3L MAP
Pathome-Genome MAP
(Regularized CCA)

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 (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)

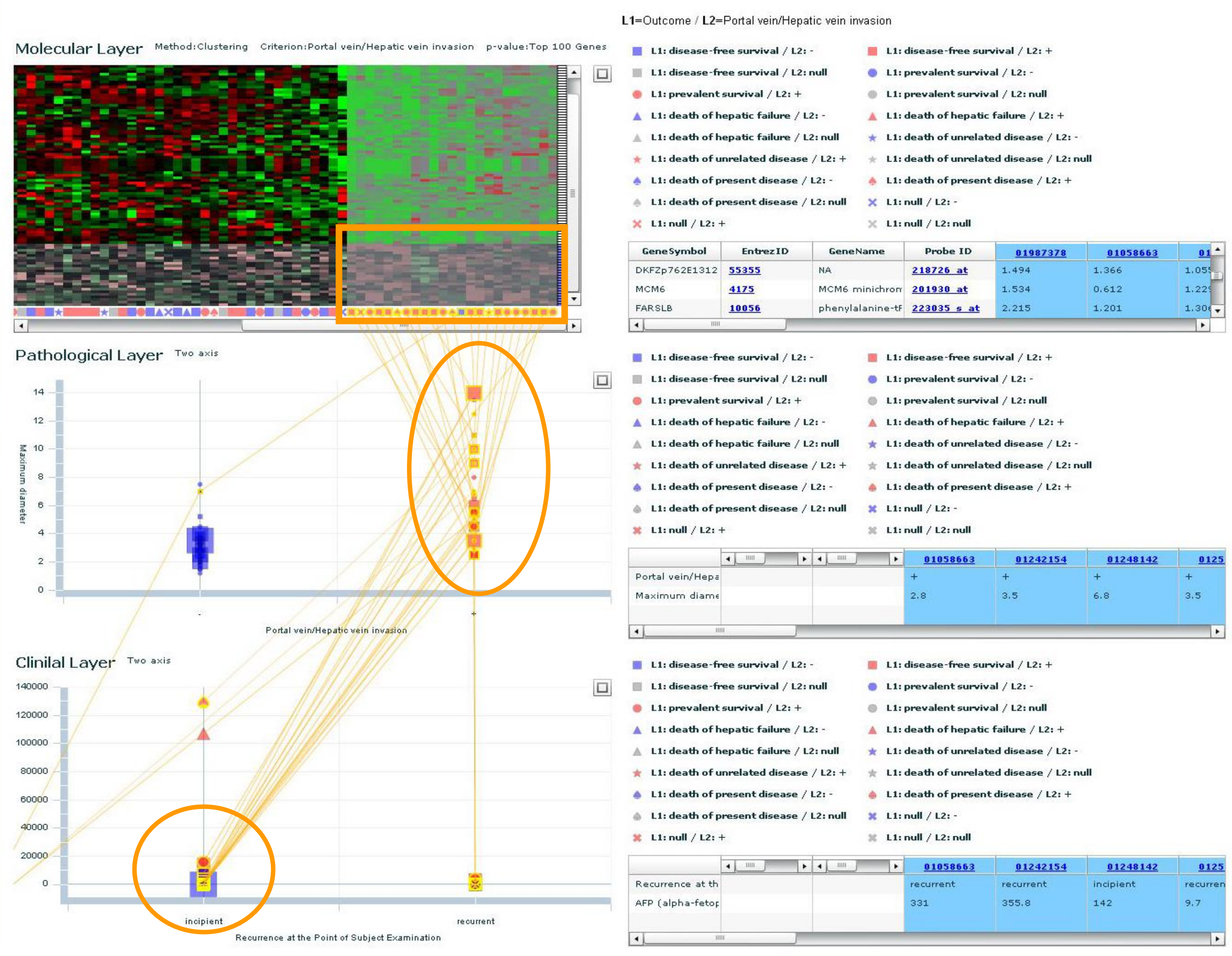
2 Dimensional – 3 Layered Map

Patient points in three 2D coordinates (molecular, pathological and clinical) are connected to show the corresponding relation between genome, pathological and clinical conditions.

Molecular Layer

Pathological Layer

Clinical Layer



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

Molecular Layer Method:Clustering Criterion:Portal vein/Hepatic vein invasion p-value:Top 100 Genes

- ■ L1: disease-free survival / L2: -
- ■ L1: disease-free survival / L2: +
- ■ L1: disease-free survival / L2: null
- ● L1: prevalent survival / L2: -
- ● L1: prevalent survival / L2: +
- ● L1: prevalent survival / L2: null
- ▲ L1: death of hepatic failure / L2: -
- ▲ L1: death of hepatic failure / L2: +
- ▲ L1: death of hepatic failure / L2: null
- ★ L1: death of unrelated disease / L2: -
- ★ L1: death of unrelated disease / L2: +
- ★ L1: death of unrelated disease / L2: null
- ♠ L1: death of present disease / L2: -
- ♠ L1: death of present disease / L2: +
- ♠ L1: death of present disease / L2: null
- ✖ L1: null / L2: -
- ✖ L1: null / L2: +
- ✖ L1: null / L2: null

GeneSymbol	EntrezID	GeneName	Probe ID	01987378	01058663	01
DKFZp762E1312	55355	NA	218726_at	1.494	1.366	1.05
MCM6	4175	MCM6 minichron	201930_at	1.534	0.612	1.22
FARSLB	10056	phenylalanine-tF	223035_s_at	2.215	1.201	1.30

Pathological Layer Two axis

- ■ L1: disease-free survival / L2: -
- ■ L1: disease-free survival / L2: +
- ■ L1: disease-free survival / L2: null
- ● L1: prevalent survival / L2: -
- ● L1: prevalent survival / L2: +
- ● L1: prevalent survival / L2: null
- ▲ L1: death of hepatic failure / L2: -
- ▲ L1: death of hepatic failure / L2: +
- ▲ L1: death of hepatic failure / L2: null
- ★ L1: death of unrelated disease / L2: -
- ★ L1: death of unrelated disease / L2: +
- ★ L1: death of unrelated disease / L2: null
- ♠ L1: death of present disease / L2: -
- ♠ L1: death of present disease / L2: +
- ♠ L1: death of present disease / L2: null
- ✖ L1: null / L2: -
- ✖ L1: null / L2: +
- ✖ L1: null / L2: null

			01058663	01242154	01248142	0125
Portal vein/Hepa			+	+	+	+
Maximum diame			2.8	3.5	6.8	3.5

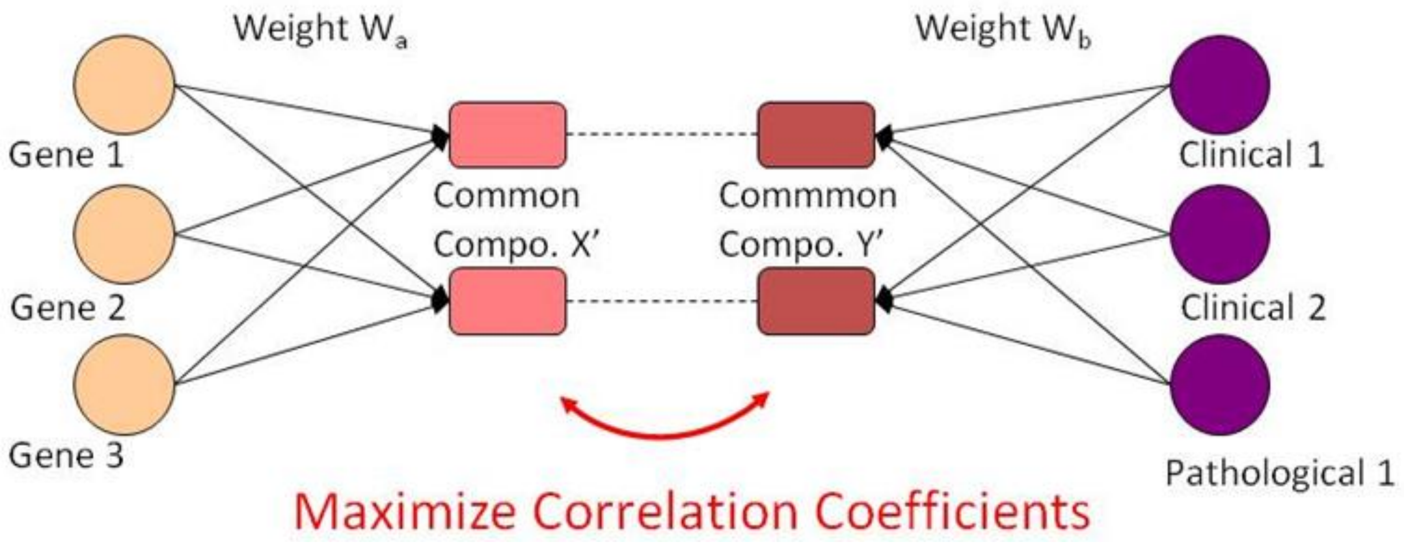
Clinilal Layer Two axis

- ■ L1: disease-free survival / L2: -
- ■ L1: disease-free survival / L2: +
- ■ L1: disease-free survival / L2: null
- ● L1: prevalent survival / L2: -
- ● L1: prevalent survival / L2: +
- ● L1: prevalent survival / L2: null
- ▲ L1: death of hepatic failure / L2: -
- ▲ L1: death of hepatic failure / L2: +
- ▲ L1: death of hepatic failure / L2: null
- ★ L1: death of unrelated disease / L2: -
- ★ L1: death of unrelated disease / L2: +
- ★ L1: death of unrelated disease / L2: null
- ♠ L1: death of present disease / L2: -
- ♠ L1: death of present disease / L2: +
- ♠ L1: death of present disease / L2: null
- ✖ L1: null / L2: -
- ✖ L1: null / L2: +
- ✖ L1: null / L2: null

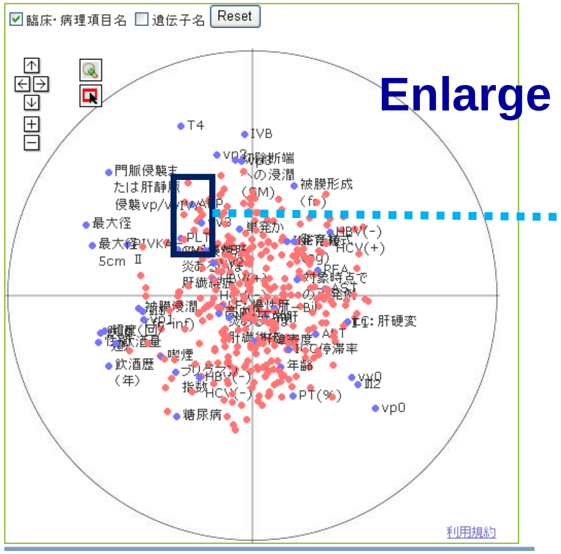
			01058663	01242154	01248142	0125
Recurrence at th			recurrent	recurrent	incipient	recurren
AFP (alpha-fetop			331	355.8	142	9.7

Pathome - Genome map

Canonical correlation analysis
Maximize the correlation coefficient
Between the linear combination of gene expression and clinic-pathological variables

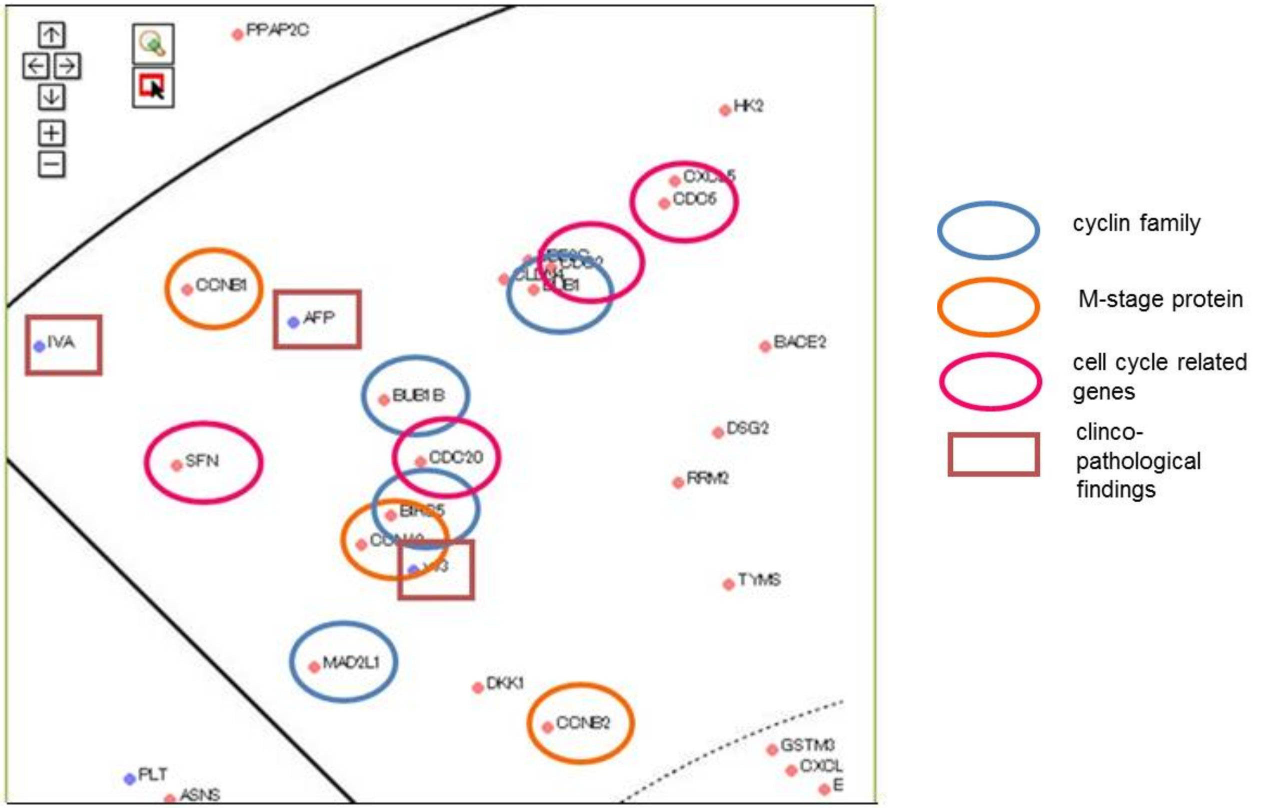


Pathome-Genome Map



☑臨床・病理項目名 □遺伝子名 Reset

Enlarge



cyclin family

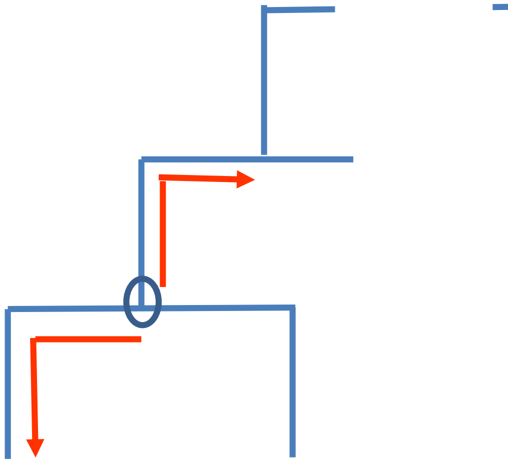
M-stage protein

cell cycle related genes

clinco-pathological findings

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

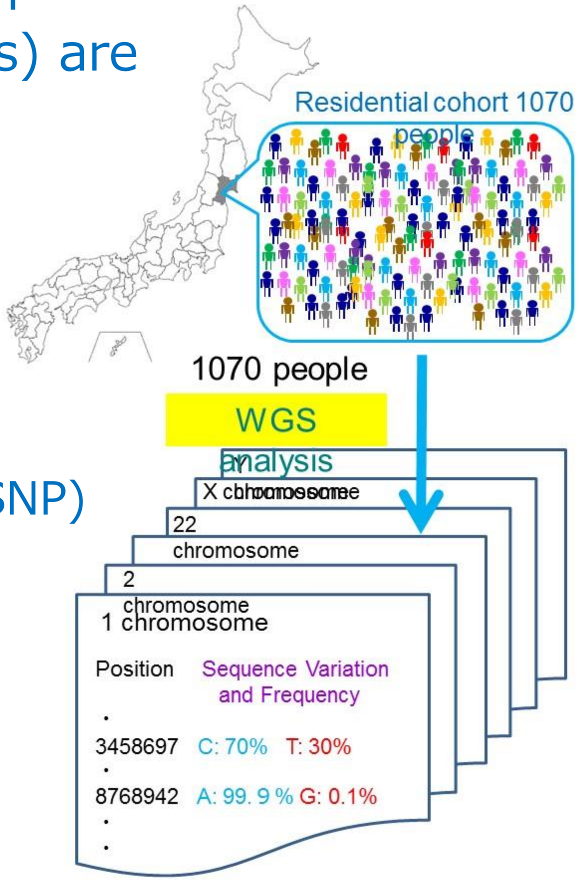


Concept ontology tree

First Results of TMM
Deep whole genome sequencing
Japanese Healthy Population

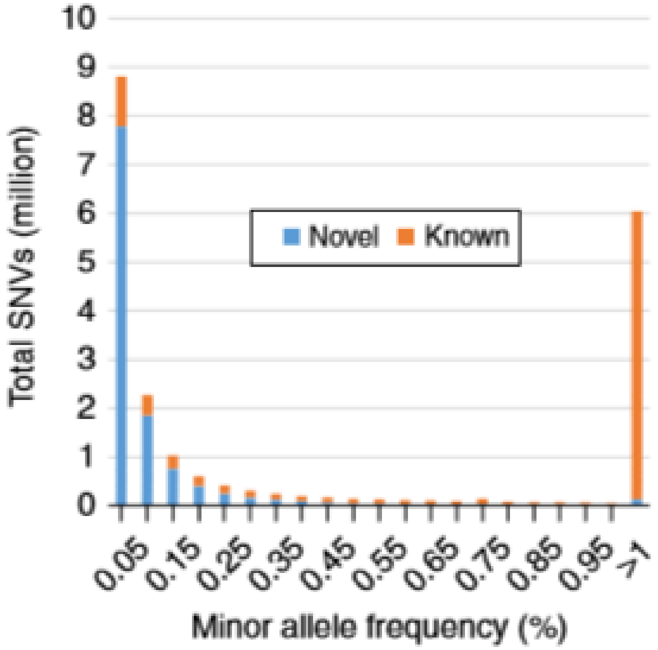
Whole Genome Sequencing in Tohoku Medical Megabank Project

- 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



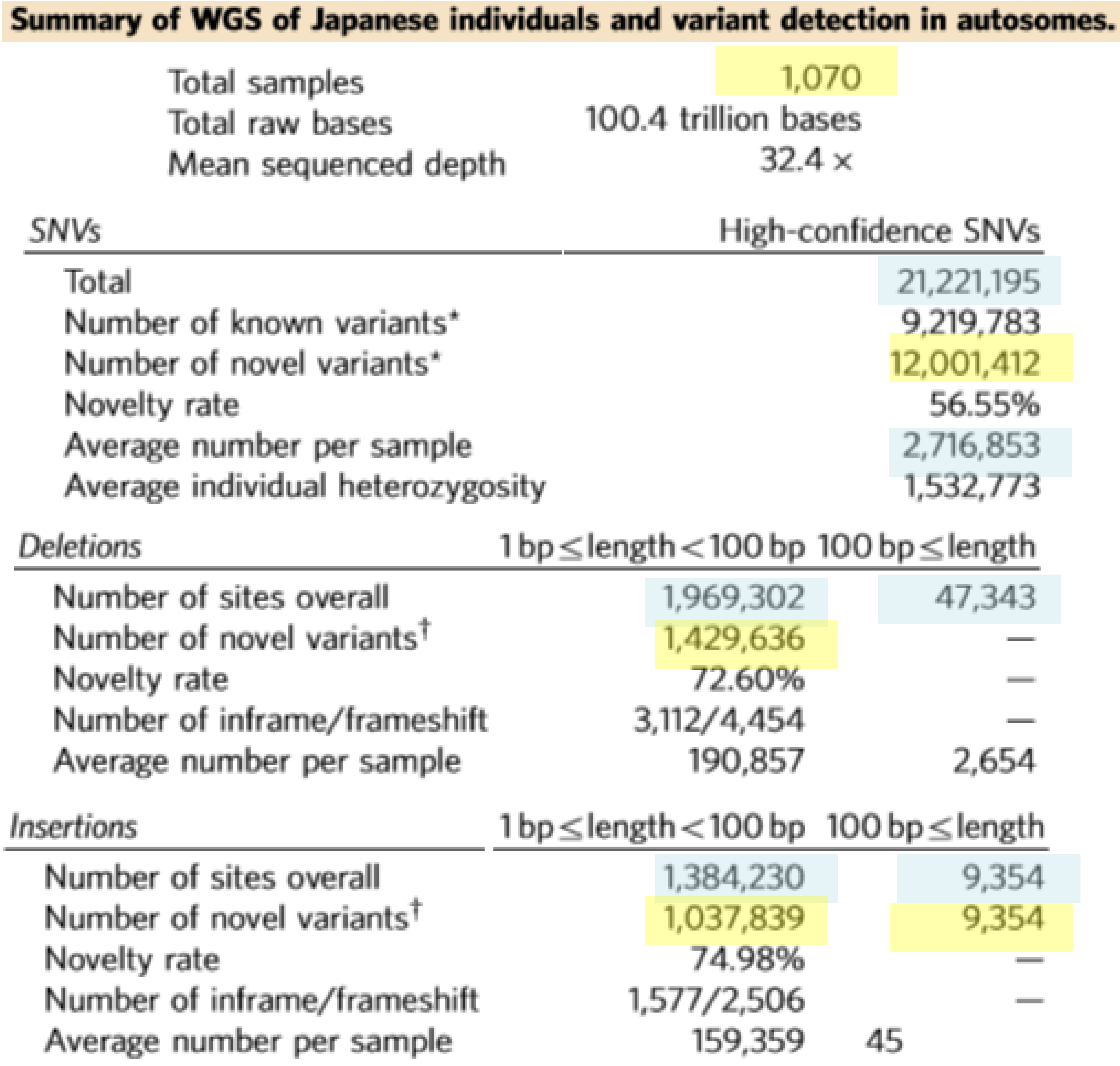
Data Processing and variant discovery

- 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%



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 ×

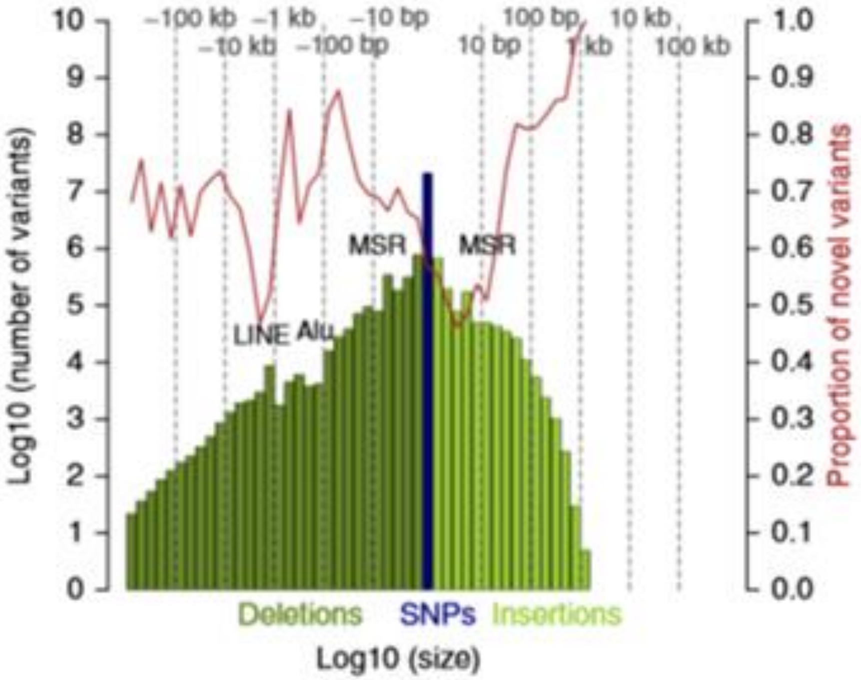
SNVs	High-confidence SNVs
Total	21,221,195
Number of known variants*	9,219,783
Number of novel variants*	12,001,412
Novelty rate	56.55%
Average number per sample	2,716,853
Average individual heterozygosity	1,532,773

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

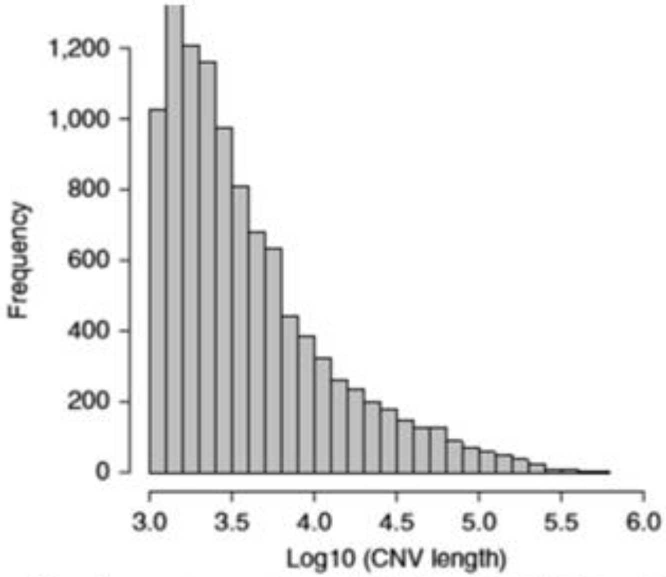
Copy number Variants 25,923

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



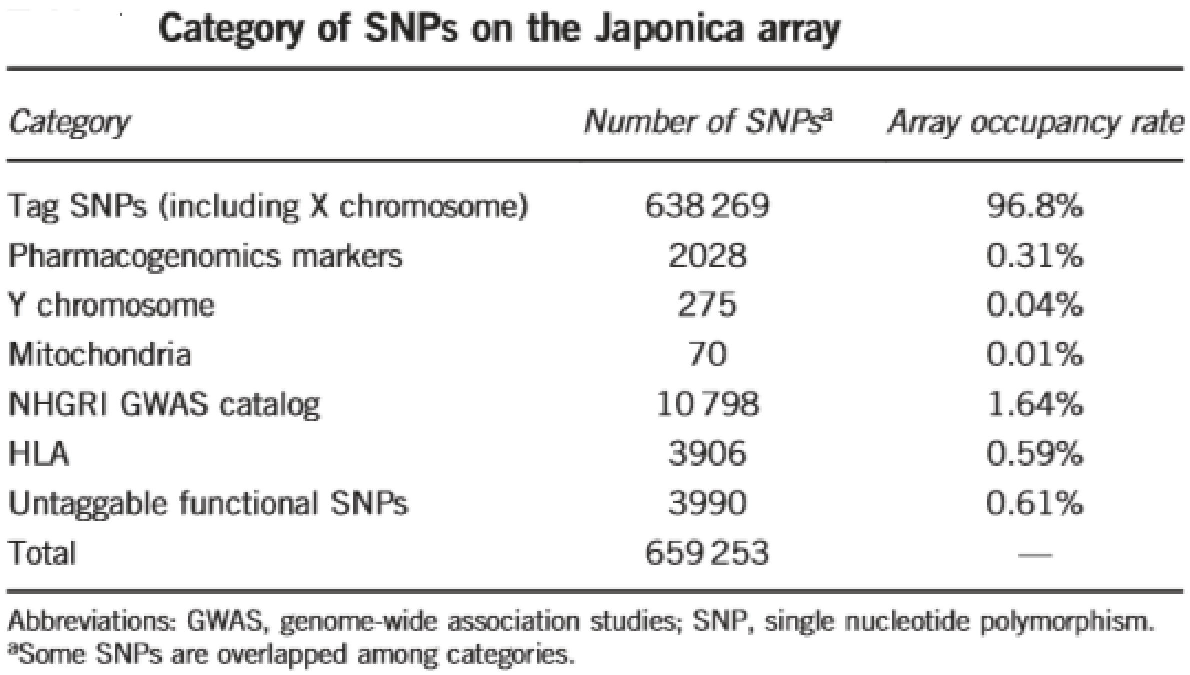
Size-frequency spectrum of CNVs estimated from high-coverage sequencing data in the genic regions in 1KJPN.

(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.

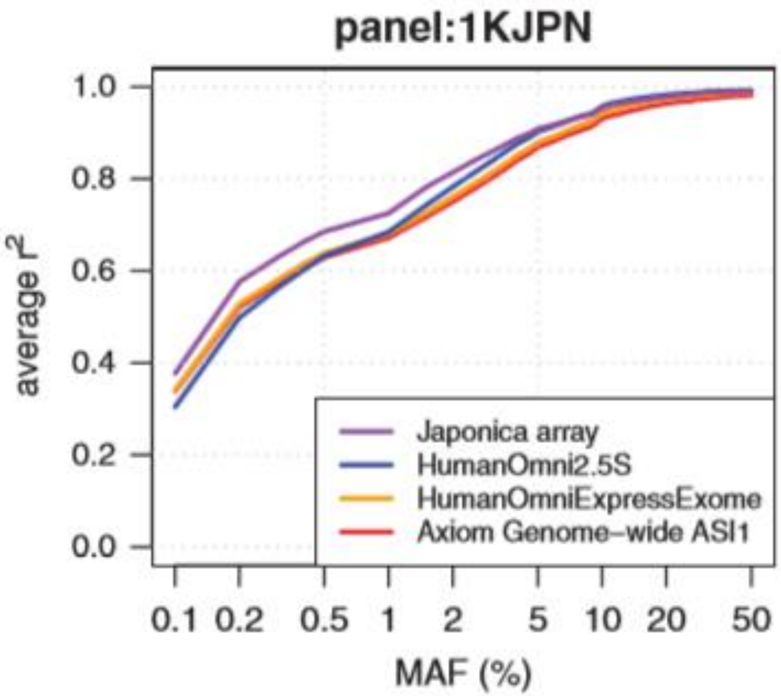
Japonica Array



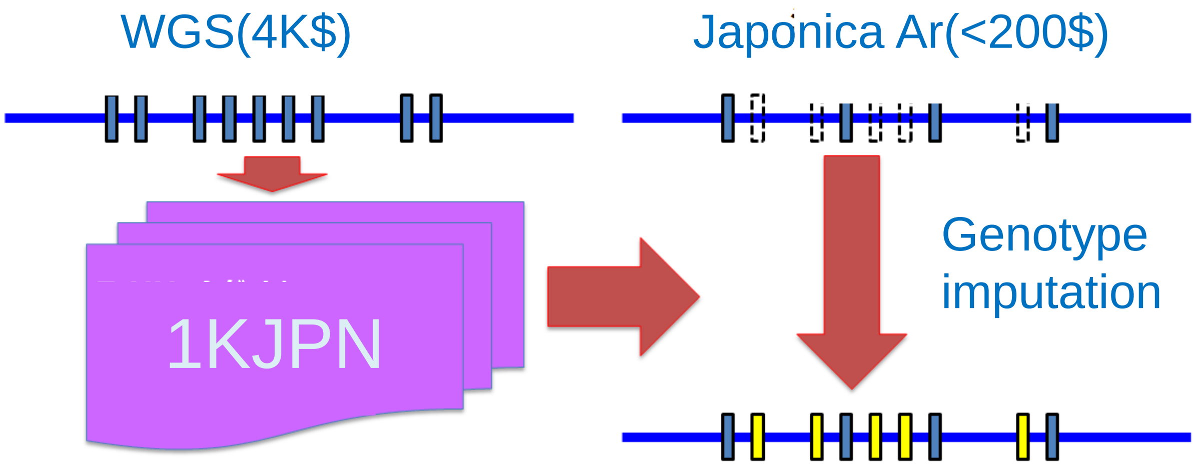
Category of SNPs on the Japonica array

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)



WGS(4K$)

Japonica Ar(<200$)

Genotype imputation

Integrated Database for genomic and environmental information

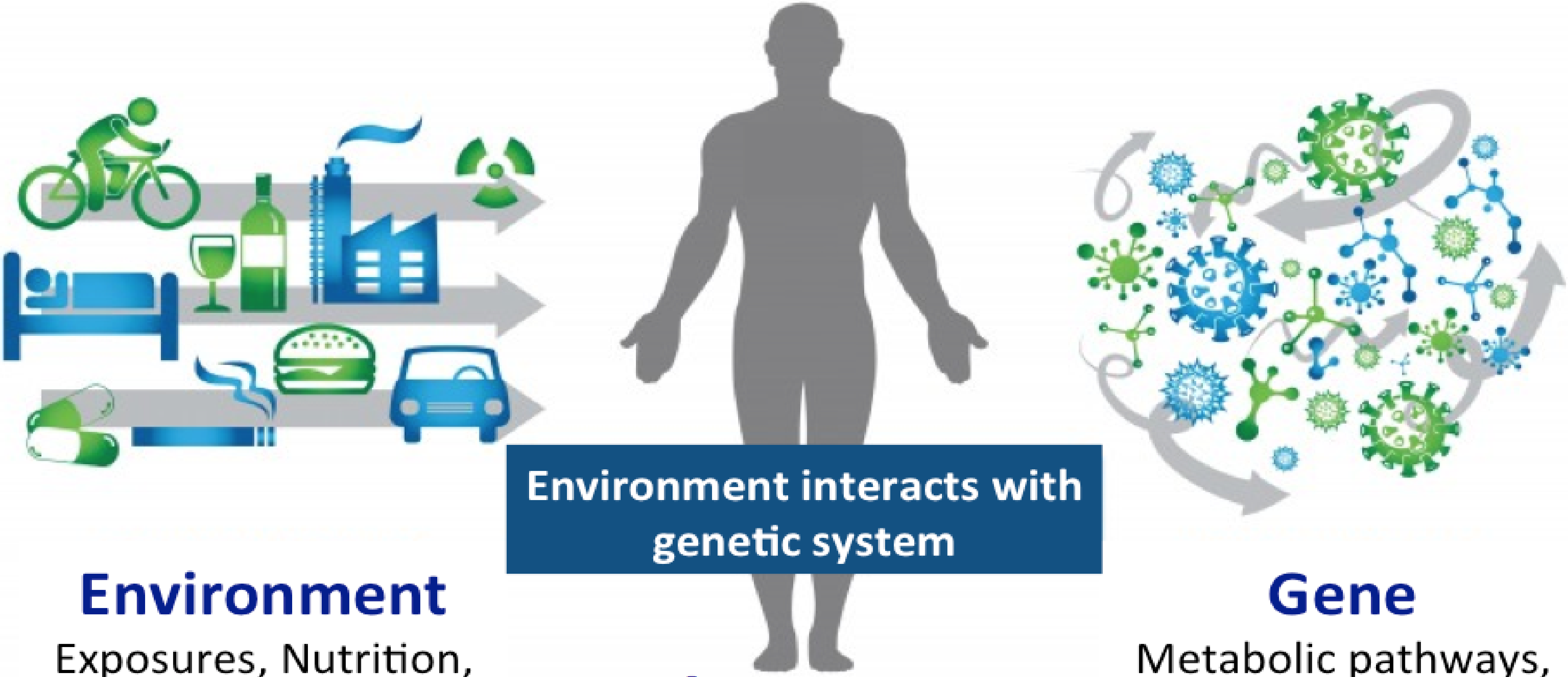
Towards the development of Information systems Tohoku Medical Megabank (TMM)

- 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 interacts with genetic system

Environment

Exposures, Nutrition, Lifestyle

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

		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

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

Each P value Estimation

Cochran-Mantel-Haenszel table

population		Disease (+)		Disease (-)	
		E (+)	E (−)	E (+)	E (−)
Gene1	0 (aa)	n_{00}	n_{01}	n_{00}	n_{01}
	1 (aA)	n_{10}	n_{11}	n_{10}	n_{11}
	2 (AA)	n_{20}	n_{21}	n_{20}	n_{21}

P value for G1x E1 → D

Gene set

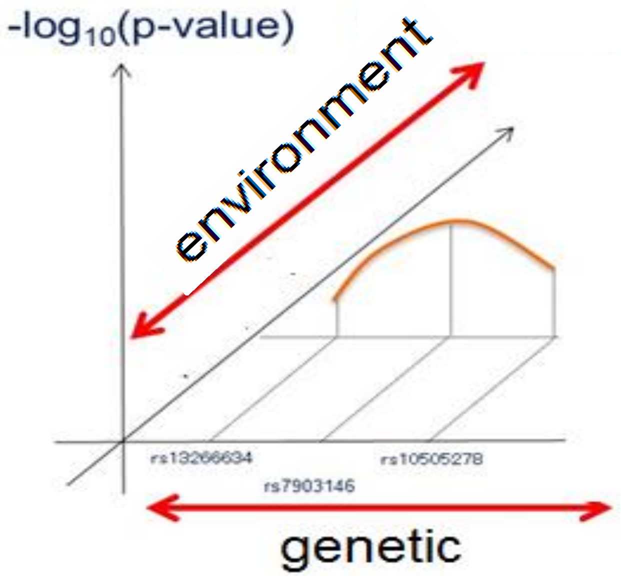
Environment factors

p	1	2	…	100
1	$7x10^{-14}$	$9x10^{-18}$	…	$3x10^{-22}$
2	$5x10^{-03}$	$2x10^{-04}$	···	$5x10^{-05}$
⋮	⋮	⋮	⋮	⋮
20	$3x10^{-17}$	$9x10^{-21}$	…	$4x10^{-22}$

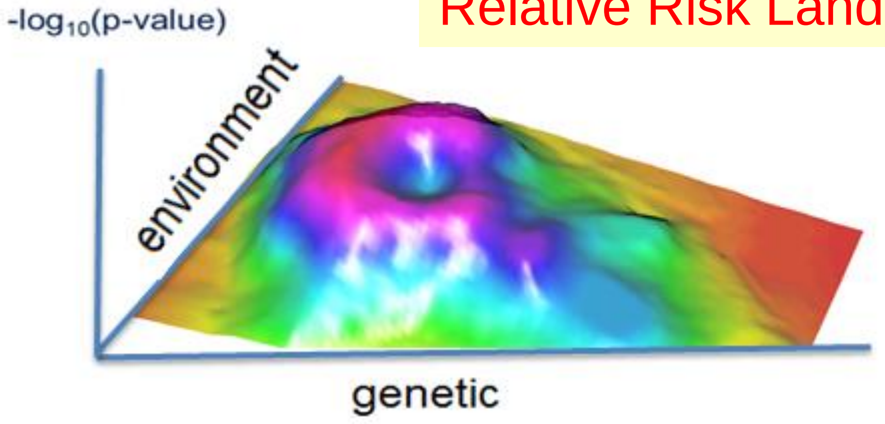
Gene allele X Environment = risk of Disease

Personalized prevention

Idiosyncratic Effect of Combination of GxE factors



Relative Risk Landscape



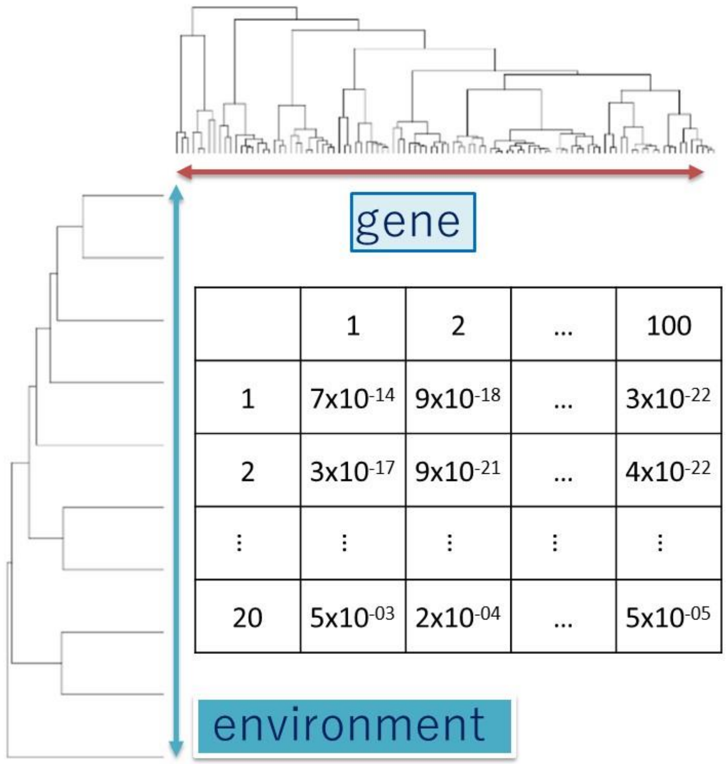
Each row of variables (genes, Environment factors) arer rearranged by hierarchical clustering

p value list

gene

Environment factors

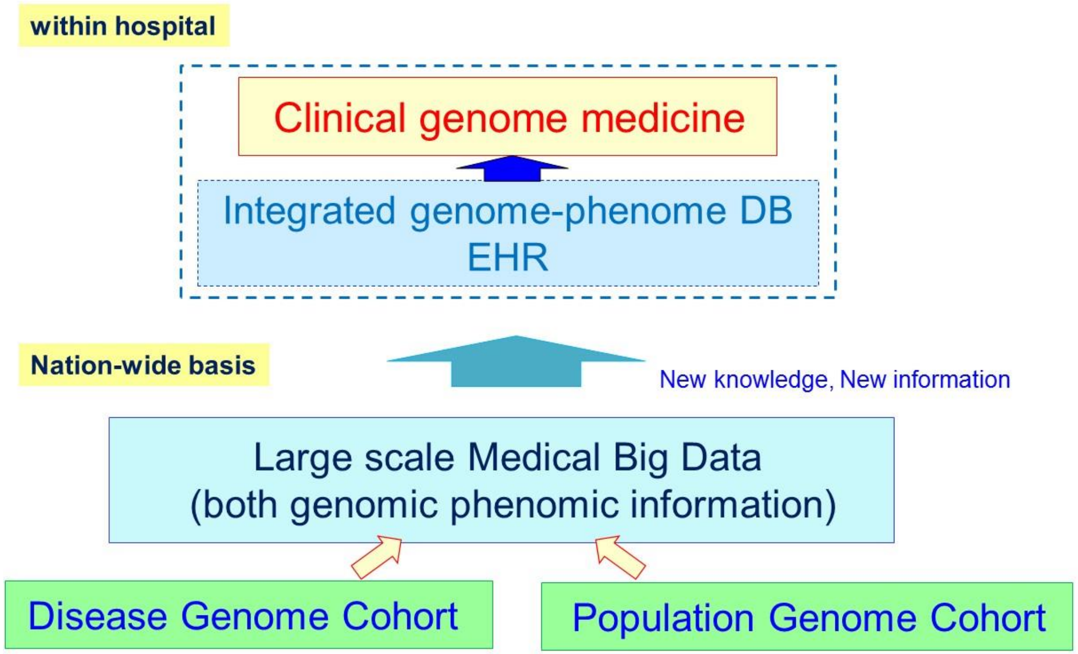
	1	2	…	100
1	$7x10^{-14}$	$9x10^{-18}$	…	$3x10^{-22}$
2	$5x10^{-03}$	$2x10^{-04}$	…	$5x10^{-05}$
⋮	⋮	⋮	⋮	⋮
20	$3x10^{-17}$	$9x10^{-21}$	…	$4x10^{-22}$



	1	2	…	100
1	$7x10^{-14}$	$9x10^{-18}$	…	$3x10^{-22}$
2	$3x10^{-17}$	$9x10^{-21}$	…	$4x10^{-22}$
⋮	⋮	⋮	⋮	⋮
20	$5x10^{-03}$	$2x10^{-04}$	…	$5x10^{-05}$

Summary

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


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

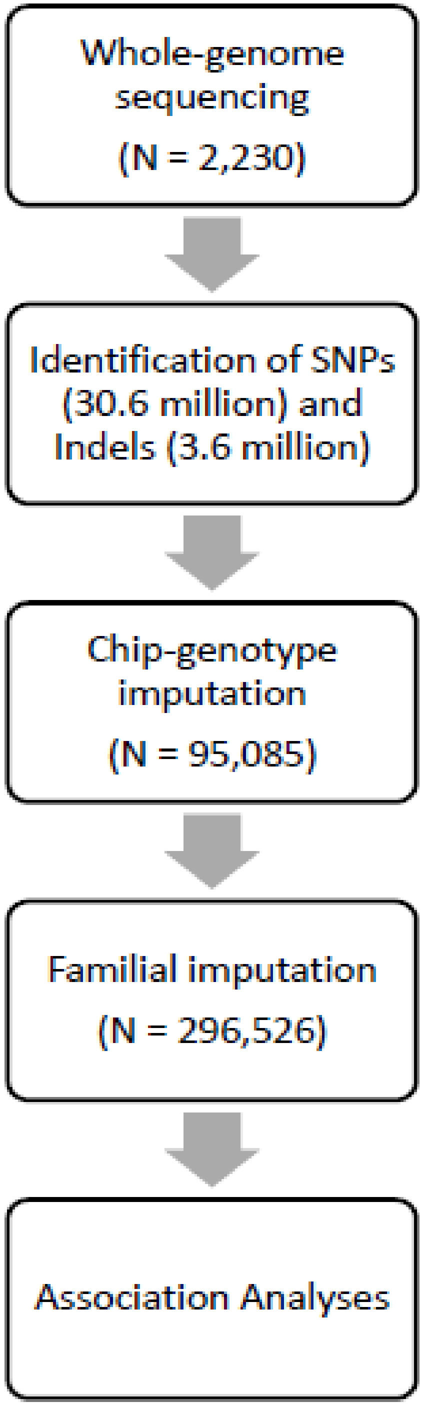
Two types of Cohort Study in ToMMo

- 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



Whole-genome sequencing

(N = 2,230)

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

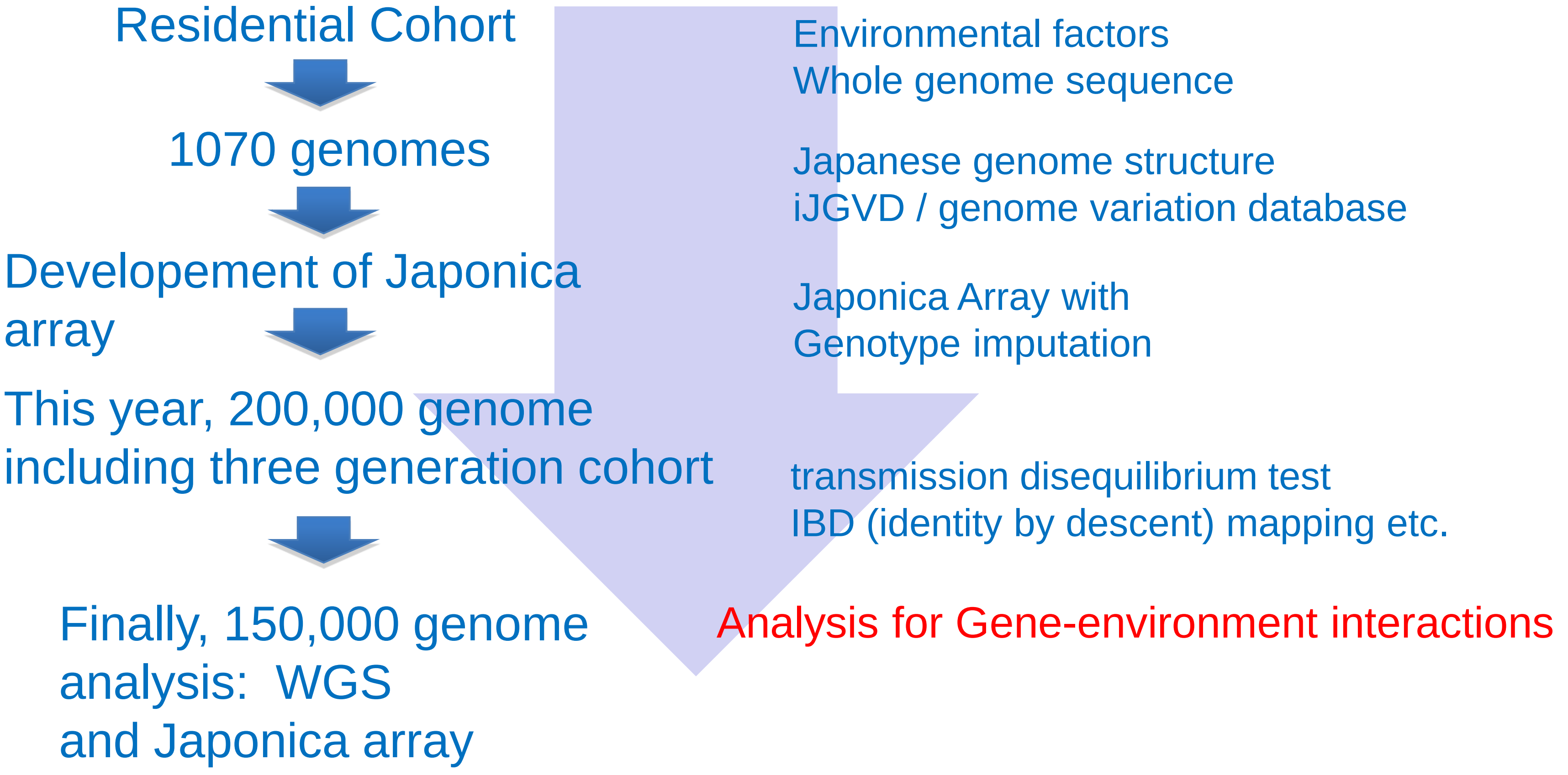
Chip-genotype imputation

(N = 95,085)

Familial imputation

(N = 296,526)

Association Analyses



Residential Cohort

1070 genomes

Developement 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

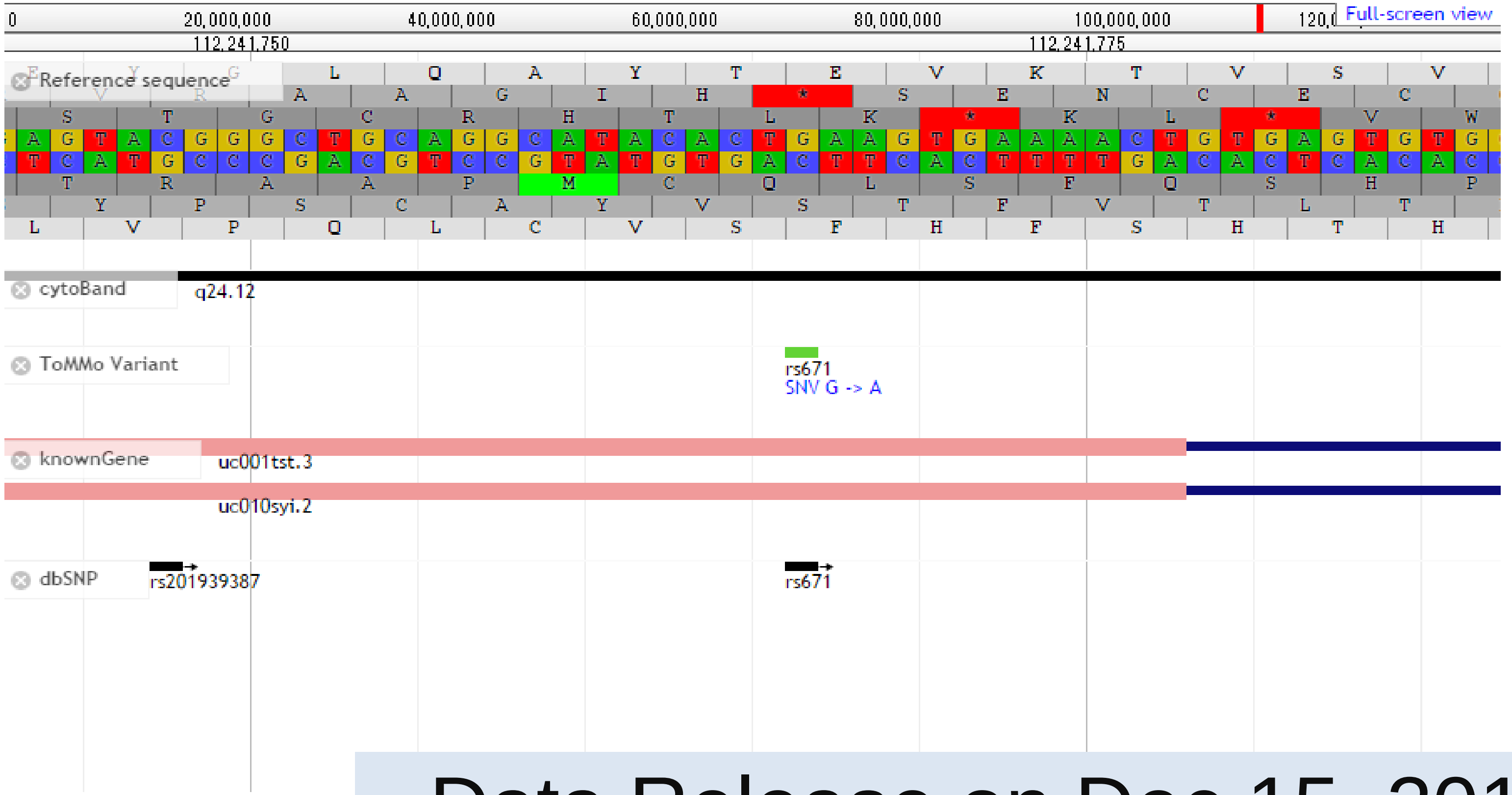
transmission disequilibrium test
IBD (identity by descent) mapping etc.

Analysis for Gene-environment interactions

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