

疾患ヒトゲノム変異の生物学的機能注釈を 目指した多階層オームクスデータの統合 (第2期)

菅野純夫

東京大学新領域創成科学研究科
メディカル情報生命専攻

○	菅野 純夫 	スガノ スミ オ	東京大学	大学院新領域創成 科学研究科	教授	KEROデータベース開発を 統括し、データ収載と各ブ ラウザ、関連解析ツールの 開発の統括及びRDF化の 実装と臨床情報データを含 むメタデータオントロジーの 整備を行う。
	鈴木 穰	スズキ ユタ カ	東京大学	大学院新領域創成 科学研究科	教授	多層オーミクスデータの整 備とデータベース化の統括
*	入江 拓磨	イリエ タク マ	東京大学	大学院新領域創成 科学研究科	特任助教	データ解析
*	横山 裕子	ヨコヤマ ヒ ロコ	東京大学	大学院新領域創成 科学研究科	学術支援職 員	データ入力補助
○	土原 一哉	ツチハラ カ ツヤ	国立がん研究センター	先端医療開発センター ゲノムトランスレーショナ ルリサーチ分野	分野長	がんゲノム多層オーミ クスDBの開発と運用
○	徳永 勝士	トクナガ カ ツシ	東京大学	大学院医学系研究 科	教授	ヒトゲノム多型・変異 データベースの拡充
	澤井 裕美	サワイ ヒロ ミ	東京大学	大学院医学系研究 科	助教	ヒトゲノム多型・変異 データの収集
○	井ノ上 逸朗	イノウエ イ ツロウ	国立遺伝学研究所	人類遺伝研究部門	教授	ヒトゲノム多型・変異 データの収集
*	早野 崇英	ハヤノ タカ ヒデ	国立遺伝学研究所	人類遺伝研究部門	特任研究員	ヒトゲノム多型・変異 データの収集
	杉本 竜太	スギモト リョウタ	国立遺伝学研究所	人類遺伝研究部門	技術補佐員	ヒトゲノム多型・変異 データの収集

ヒト応用研究を志向したオミクス情報の統合

統合1

転写開始点/トランスクリプトーム情報
(TSS/RNA Seq)

(発現量と転写開始点)

統合1

クロマチン情報 (ChIP Seq)

(ChrHMMパターンで示すヒストン修飾)

統合2

ヒトゲノム
変異情報の統合

統合DB (DBKERO)

(それぞれの検体での変異部位)

統合1

DNAメチル化情報 (BS Seq)

(BS Seqによる異常メチル化検出)

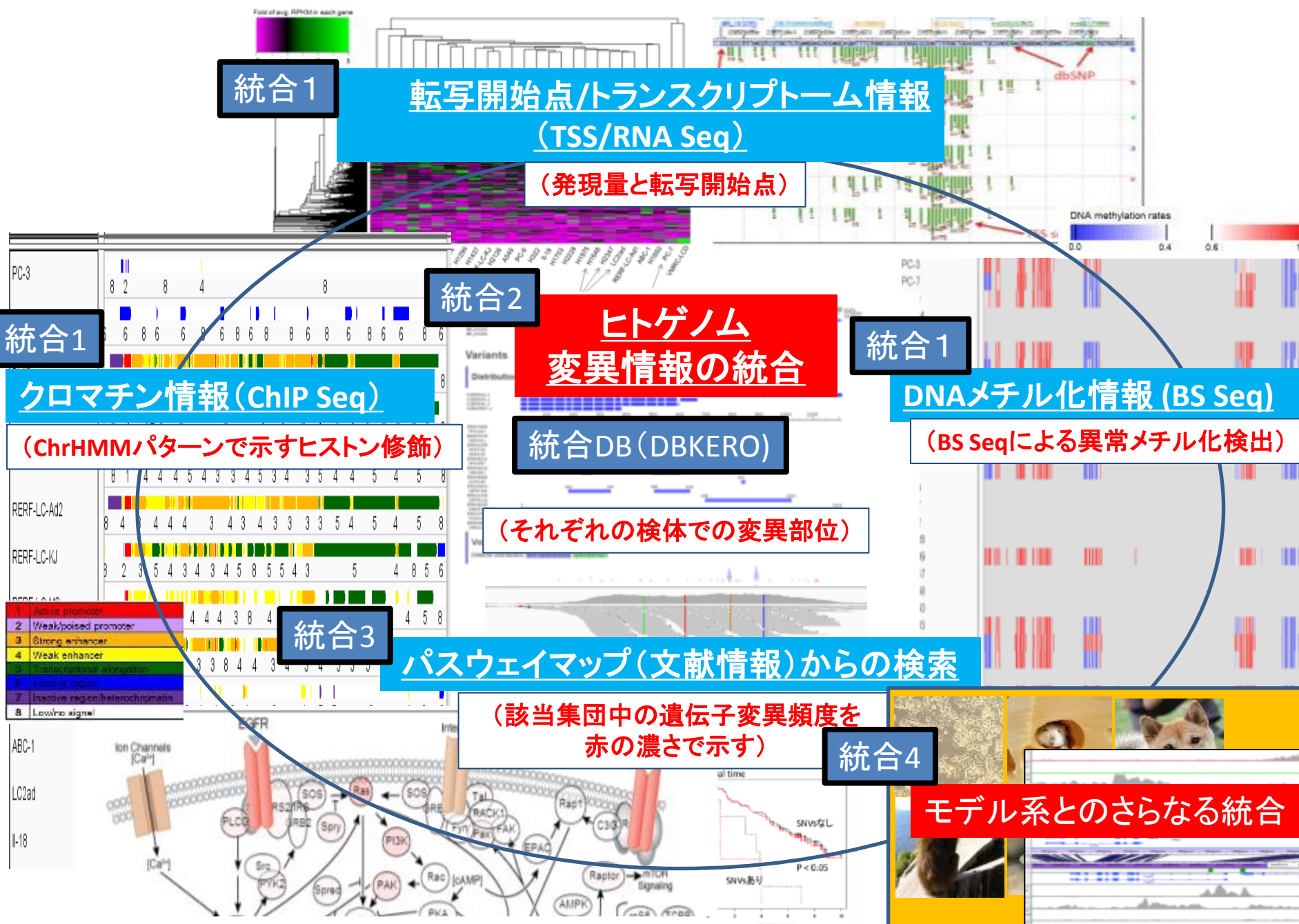
統合3

パスウェイマップ (文献情報) からの検索

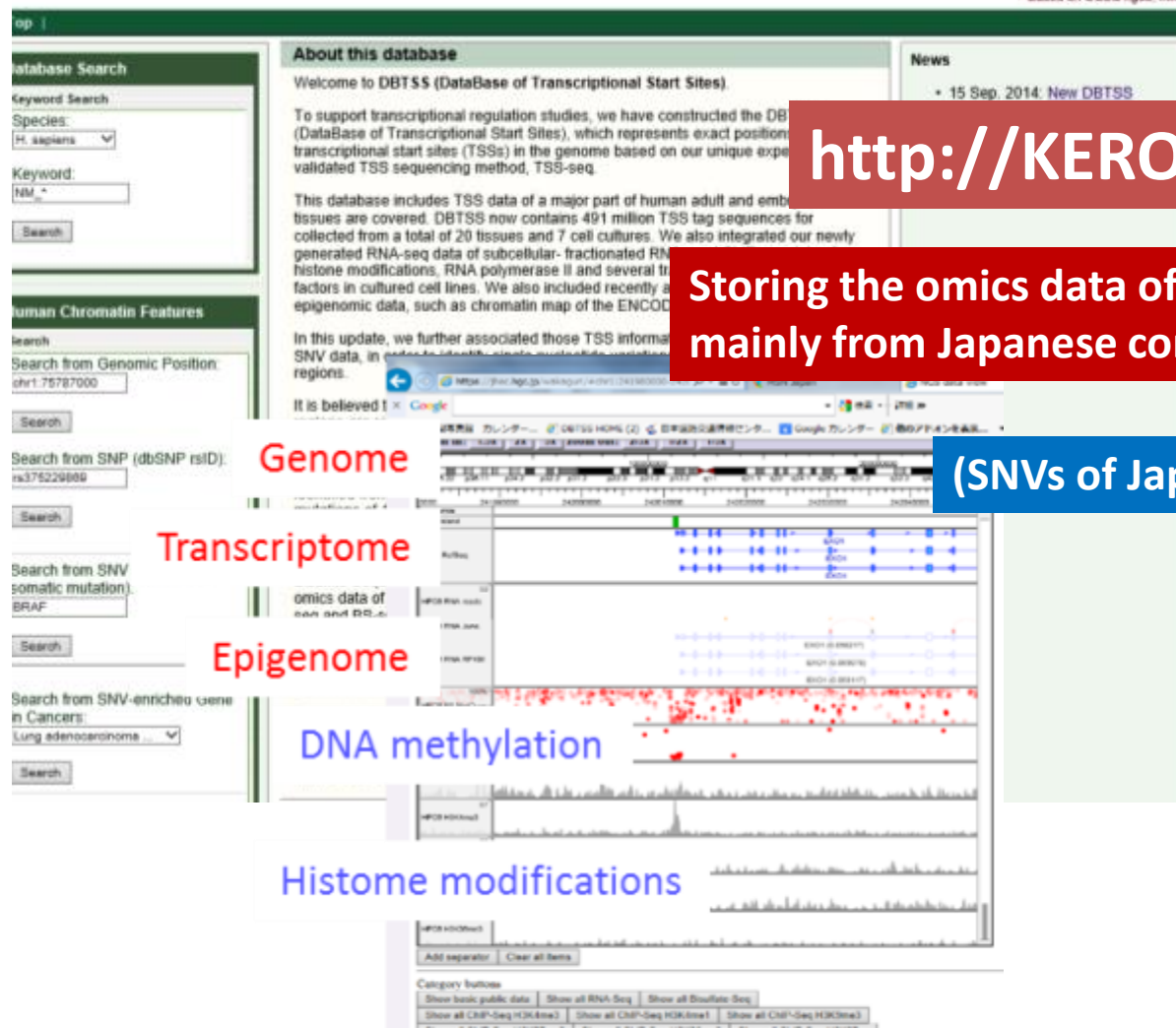
(該当集団中の遺伝子変異頻度を
赤の濃さで示す)

統合4

モデル系とのさらなる統合



DBKERO

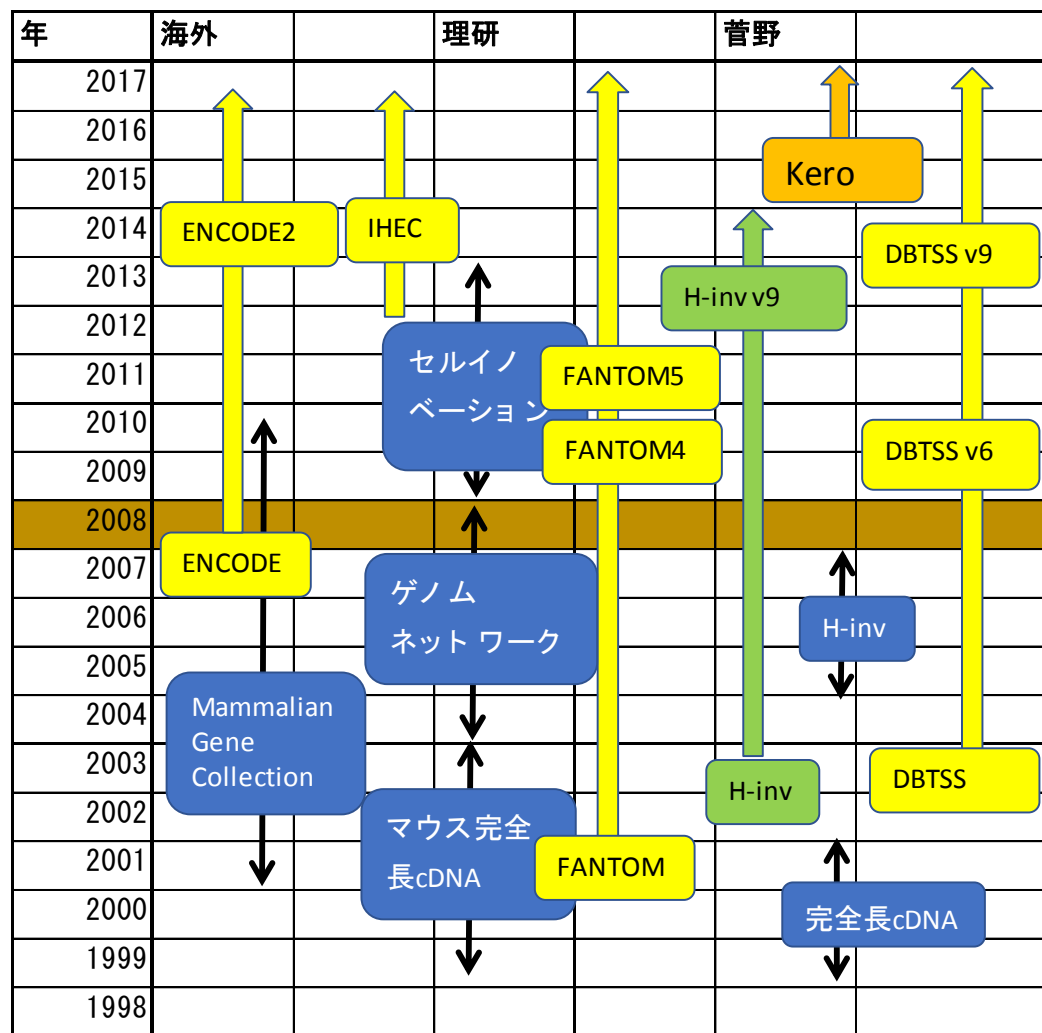
Release 9.0 Updated (September 15 2014)
Based on UCSC hg38, mm9<http://KERO.hgc.jp>Storing the omics data of a total of >5.000 points
mainly from Japanese community

(SNVs of Japanese population)

Database for multi-omics analysis

参考資料

Keroの位置づけ：機能ゲノムDBの系譜



Keroの2つの統合：

- 1) Keroは自前だけでなく他のグループのデータも統合。
- 2) ヒトを中心に疾患に関連した機能ゲノムデータを統合

- 自前のデータを使った機能ゲノムDB
- 大規模RNA解析プロジェクト
- 完全長RNAカタログ化DB

2008年にRNA-seqの論文が出てNGSを使った機能ゲノム解析の時代に入った。

公開データ概要

ヒトゲノム多型/変異

正常ゲノム/がんゲノム

日本人: 5272検体/567症例

外国人: 60706検体/9825症例



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

難病情報センター

次世代遺伝子解析装置を用いた難病の原因究明、治療法開発プロジェクト



大阪大学
OSAKA UNIVERSITY



九州大学
KYUSHU UNIVERSITY

ExAC Browser (Beta) | Exome Aggregation Consortium

DBKERO



Experiencing Oncology Research at Clinical Trial Center
早期・探索臨床研究センター



次世代がん研究シーズ戦略的育成プログラム
Project for Development of Innovative Research on Cancer Therapeutics

トランスクリプトーム

日本人正常トランスクリプトーム: 71検体

がん細胞トランスクリプトーム: 42データ点

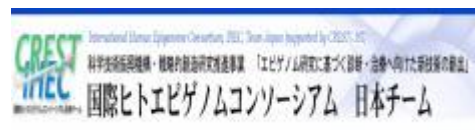
がん細胞TSSその他: 85データ点

統合

エピゲノム

正常エピゲノム: 550 データ点

がん細胞エピゲノム: 281データ点



国立がん研究センター

マウス等モデル生物種

エピゲノム: 337データ点

TSSその他: 36データ点

トランスクリプトーム: 33検体



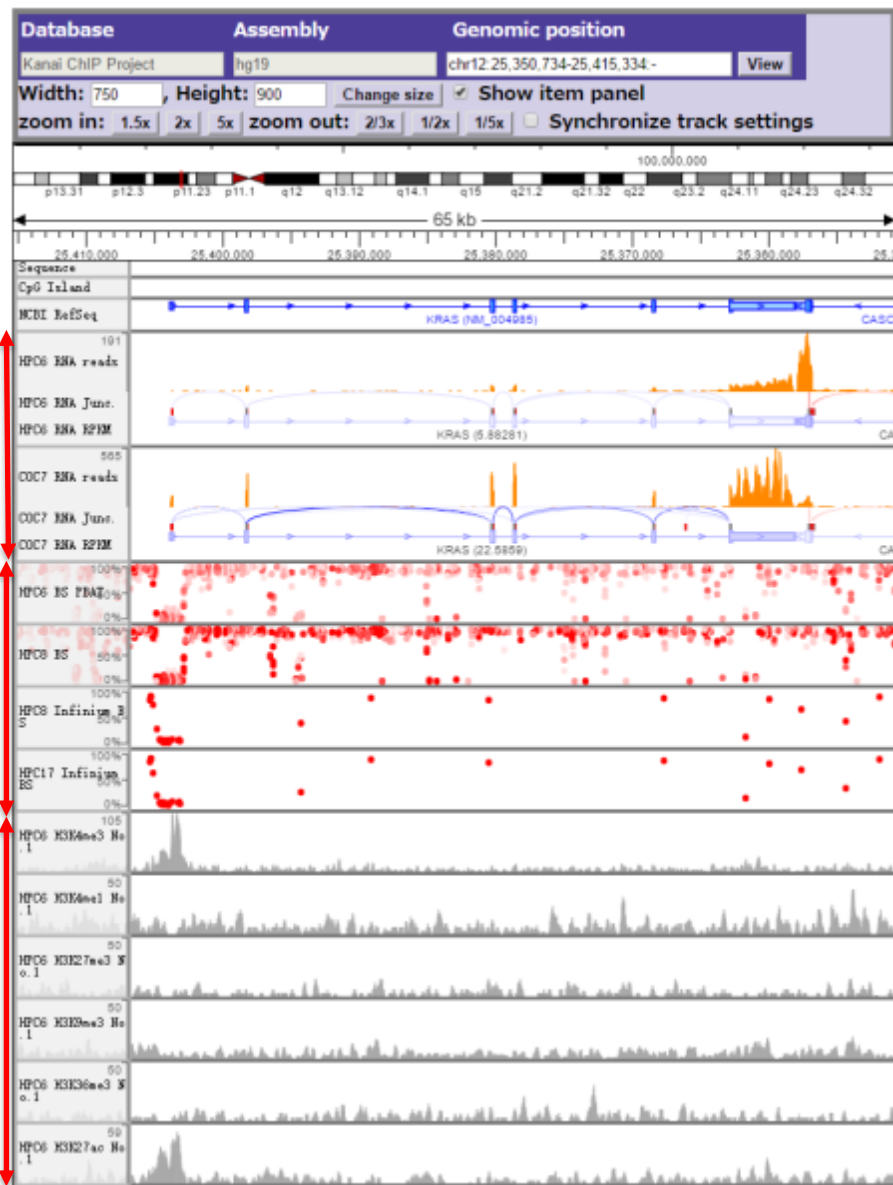
文部科学省科学研究費新学術領域研究
ゲノム支援



Japanese Genotype-phenotype Archive

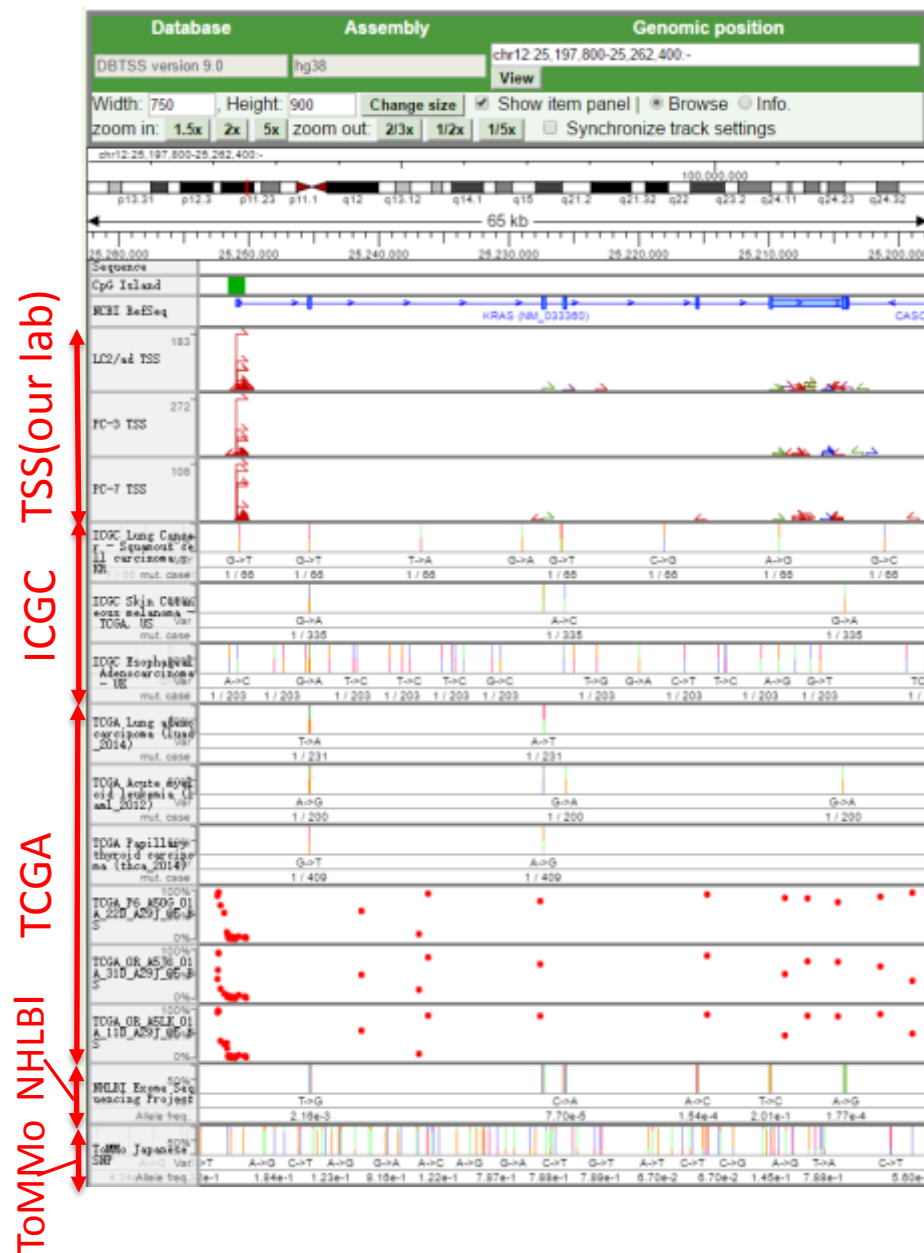
マルチオミックス

(RNA-seq, BS-seq, ChIP-seq, TSS-seq, SNV, etc...)

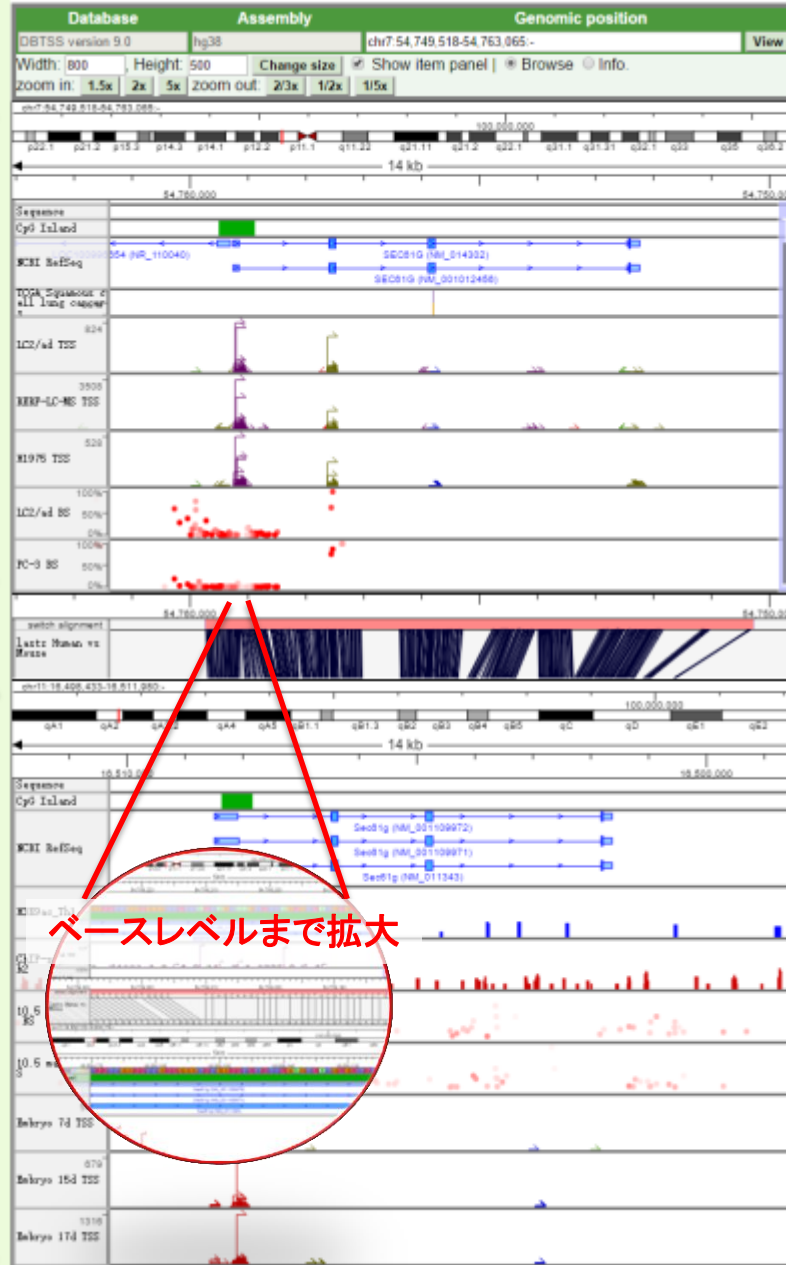


様々な機関のデータを統合

(研究室内データ、国内ToMMo等、国外ICGC/TCGA等)

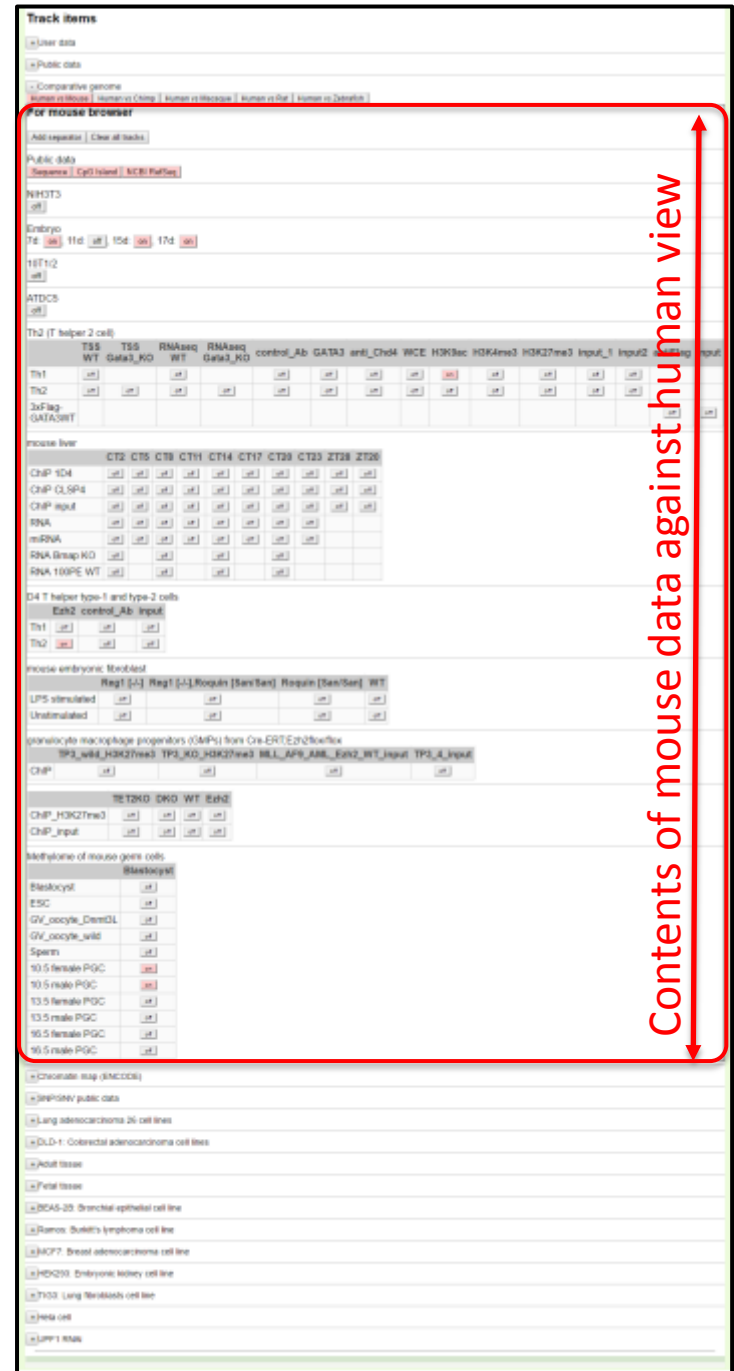


Human
mouse
alignment



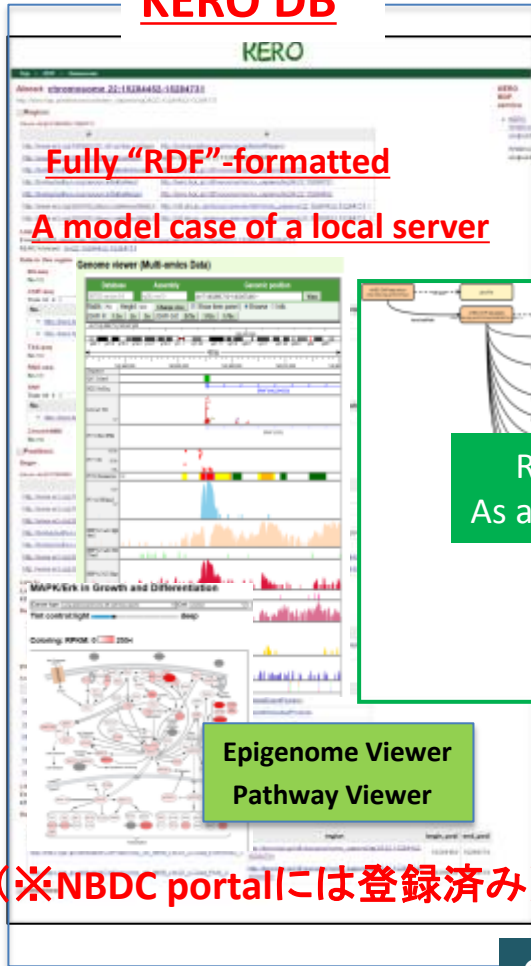
ベースレベルまで拡大

Contents of mouse data against human view



RDFを用いたデータ統合の試行

KERO DB



Fully "RDF"-formatted
A model case of a local server

Genome viewer (Multi-omics Data)

Epigenome Viewer
Pathway Viewer

(※NBDC portalには登録済み)

Integration of
the local data

Data retrieval

RDF/Sparkle Search
As a versatile "language"

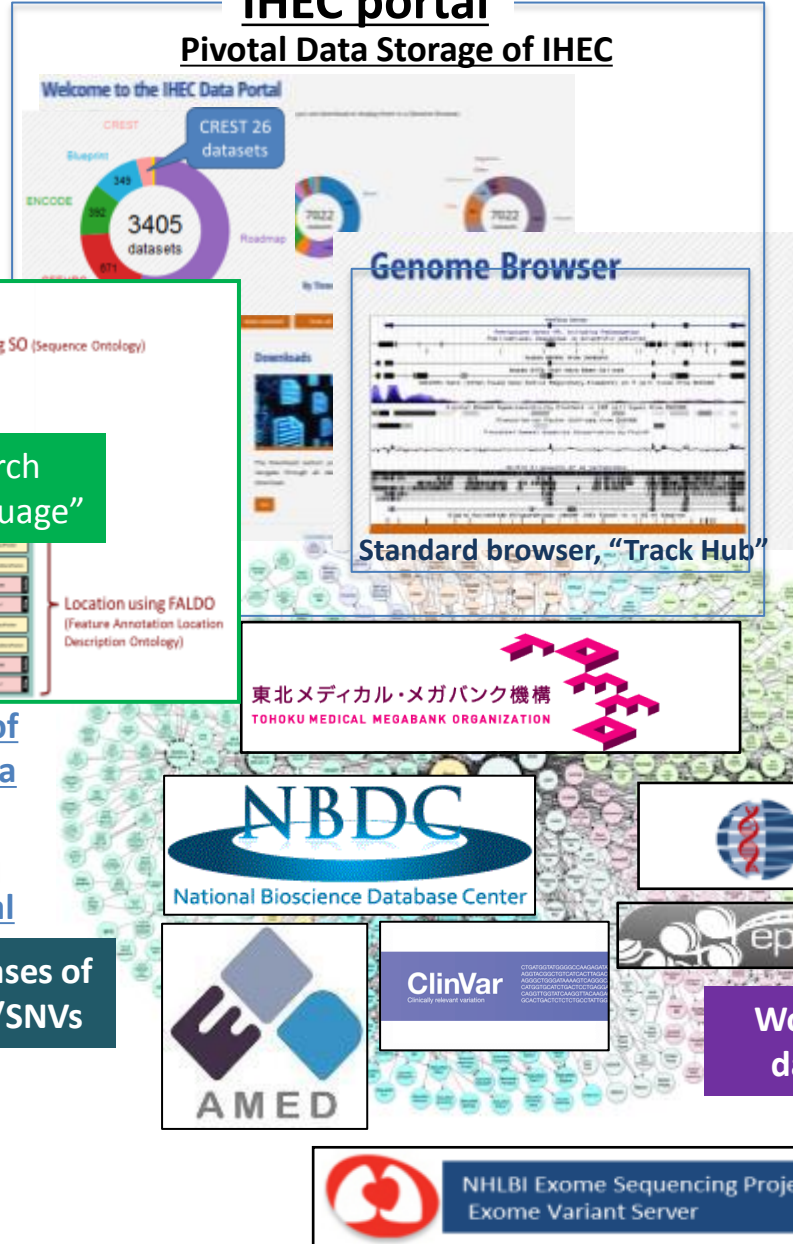
Integration of
the local data

Data retrieval

Genome Databases of
Japanese SNPs/SNVs

IHEC portal

Pivotal Data Storage of IHEC



Welcome to the IHEC Data Portal

CREST 26 datasets

3405 datasets

Genome Browser

Standard browser, "Track Hub"

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

NBDC
National Bioscience Database Center

ClinVar
Clinically relevant variation

International Cancer Genome Consortium

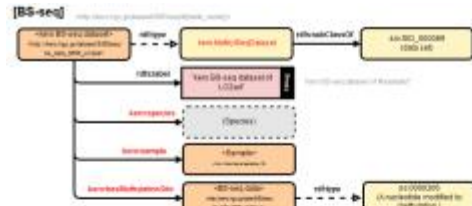
ROADMAP
epigenomics
PROJECT

World-wide
databases

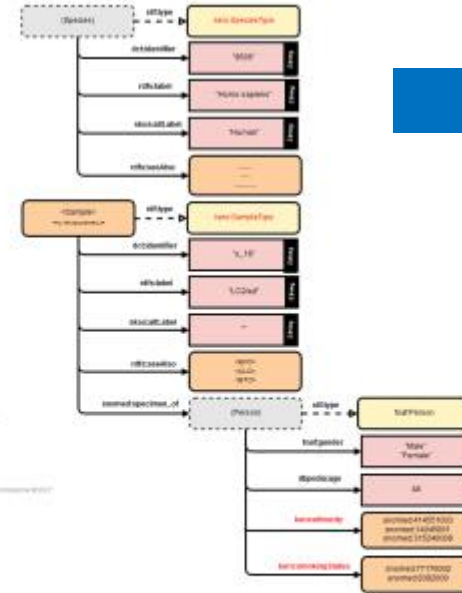
AMED

NHLBI Exome Sequencing Project (ESP)
Exome Variant Server

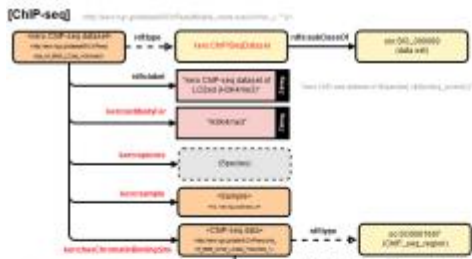
Schema (“Triple”) of RDF for multi-omics data



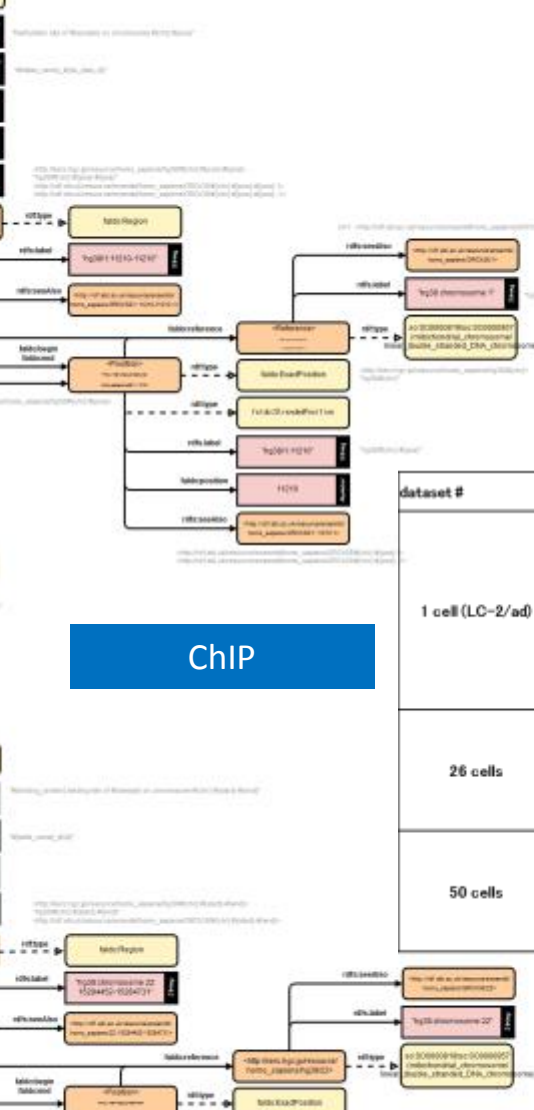
Genome



RNA/TSS



ChIP



ベンチマークテスト(ゲノム)

dataset #		WES	WGS (somatic)	WGS (somatic + germline)
1 cell (LC-2/ad)	データプロセス(データのimportにかかった時間)(秒)	15.769 13.250 12.443	47.723 54.478 50.984	3,128.971 3,341.555 3,916.970
	検索時間(3回平均)	0.054	0.121	0.338
	データ件数(RDBレコード数)	22,572	90,764	3,452,550
	タブ区切りファイル容量(未圧縮)	5M	21M	776M
	トリプル数 (subject-predicate-object)	406,344	1,633,802	62,145,950
	データ容量(推計)	40M	200M	9,356M
26 cells	データプロセス(データのimportにかかった時間)(秒)	356.088	2947.525	110,130.279
	検索時間(3回平均)	0.140	0.296	1.178
	データ件数(RDBレコード数)	549,563	3,235,943	85,440,166
	タブ区切りファイル容量(未圧縮)	117M	749M	19010M
	トリプル数 (subject-predicate-object)	7,129,104	55,113,680	1,500M
	データ容量(推計)	2,000M	12,293M	119,603M
50 cells	データプロセス(データのimportにかかった時間)(秒)	727.047	9694.896	220,000.000
	検索時間(3回平均)	0.083	0.189	ND (2-3秒?)
	データ件数(RDBレコード数)	1,054,554	6,190,528	163,994,212
	タブ区切りファイル容量(未圧縮)	224M	1431M	36475M
	トリプル数 (subject-predicate-object)	13,189,584	89,757,224	3,000M
	データ容量(推計)	5,000M	25,000M	230,000M

推計

Beacon検索

外部からのBEACON検索可能

Beacon Network Search Beacons

Search all beacons for allele

GRCh37 = 13:32936732 G > C Search

Response: All None Found Not Found Not Applicable

Organization: ASAPLab, LAC Biotech, Australian Genomics, Genomics Medical Ltd, etc.

NBDC Human Database

Search NBDC Human Database Beacon for Alternative Alleles [API help]

A NBDC Human Database Beacon will be a member of [Genomic Beacon Network](#).

GRCh37 = 12:112241786 A Search Example: ALDH2 Variant (GRCh37: 12:112241786 A)

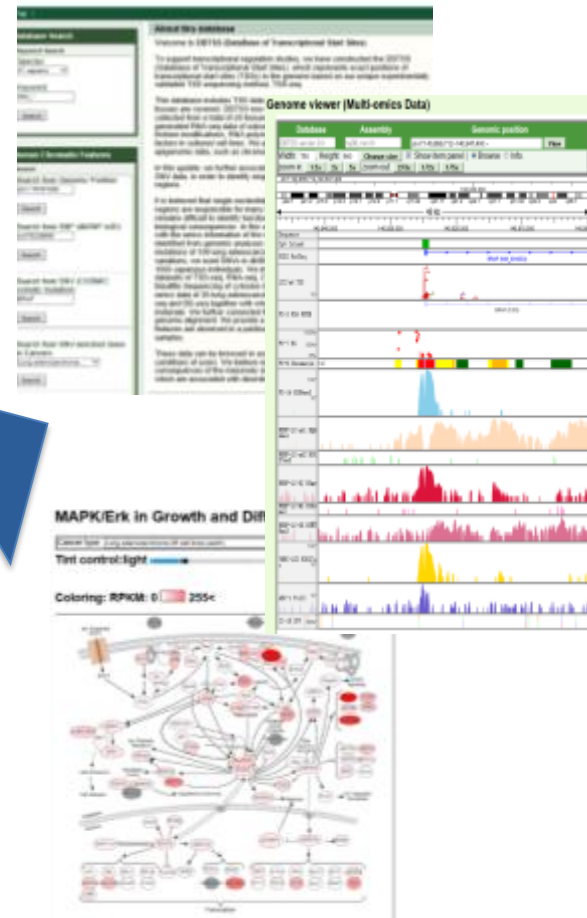
Beacon Query

- Reference: GRCh37
- Chromosome: 12
- Position: 112241786
- Allele: A

Response by Dataset

Dataset	Found	Link to Download Site	Error Messages
hum0013.v1.feq.v1	Found	http://humandb.biosciencedbc.jp/hum0013.v1	
hum0013.v3.T2008.1.v1	Not Found	http://humandb.biosciencedbc.jp/hum0013.v3	
hum0015.v1.feq.v1	Found	http://humandb.biosciencedbc.jp/hum0015.v1	
hum0029.v1.feq.v1	Found	http://humandb.biosciencedbc.jp/hum0029.v1	

Site Policies | Sitemap | Contact Us | Supported Resources



データベースの拡充

疾患ゲノム多層オーミクスデータベース拡充 (KERO; <http://kero.hgc>)

1. 外部参照データの更新 平成29～33年度
2. HGVDBとの統合(日本人ゲノム多型) 平成29～31年度;それ以降にさらなる融合
3. CREST-IHECデータの統合(日本人標準エピゲノム) 平成29～31年度
4. 先進ゲノム支援データ等(マウスを中心に)のモデル生物実験系データ
平成29年度～;随時更新
5. RDFによるデータ統合検索の運用の開始 平成29年度;実用化
6. 臨床ゲノムデータベース、バイオバンクデータベースとの連携 ～平成33年度

技術開発

7. 新規オーミクス解析技術(シングルセル、薬剤摂動)ブラウザの開発とデータ公開
平成30年度～

▪ **1) 菅野グループ (代表者グループ)**

※代表者グループのみの場合、その旨のみの記載で構いません。

- ・研究全体を統括し、それぞれの研究班の成果を KERODB として実装する。
- ・CREST-IHEC データに関して、最終確認を行い、公開を開始する。
- ・その他、RDF 化、新規計測技術データについて、データベースの設計に主導的な立場をはたす。

▪ **2) 土原グループ**

・「当該年度における研究開発の進め方」のうち、当該グループが担当する内容を記載してください。

- ・シングルセルデータのデータベース化の設計を行い、公開可能な形に加工できる一連のパイプラインを作成する。またプロトタイプ of データベースを作成する。薬剤摂動データについても同様の検討を開始する。検討について今年度到大枠を確定し、来年度以降の公開の基盤とする。
- ・RDF 特に臨床検体に関するメタデータについて、枠組みの整備を開始する。

▪ **3) 徳永グループ**

・「当該年度における研究開発の進め方」のうち、当該グループが担当する内容を記載してください。

- ・HGVDB の統合について研究を統括する。
- ・HGVDB データの移設を開始する。今年度、前掲したように全データの 1/3 について移設を完了する。
- ・HGVDB の統合にともない必要となる一連のプログラム、またブラウザを開発する。

▪ **4) 井ノ上グループ**

・「当該年度における研究開発の進め方」のうち、当該グループが担当する内容を記載してください。

- ・HLA 多型部分について、HGVDB の移設を行う。該当データセットをモデルケースに、GWAS 等の HGVDB データの円滑な移行と統合の枠組についての検証を行う。実施に際し、徳永 Gr と密接に連携して、移設の実務の多くを担当する。

Human Variation DB (旧徳永グループ)との統合

KERO DB

現在の相互リンク

ブラウザーの開発
データ管理の集約化
RDF検索の内部モデル

・表6 HGVDB（今後、統合予定）

名称	件数	概要
健常者データ	4 studies	ゲノムワイドSNP データ: 数千人規模
B型肝炎	3 studies	GWAS: 数千人規模
C型肝炎	3 studies	GWAS: 数千人規模
原発性胆汁性肝硬変	2 studies	GWAS: 数千人規模
自己免疫性肝炎	1 study	GWAS: 数百人規模
真性過眠症	1 study	GWAS: 数百人規模
ナルコレプシー	2 studies	GWAS: 千人規模
変形性関節症	1 study	GWAS: 数百人規模
パニック障害	1 study	GWAS: 千人規模
パーソナリティ、気分障害	1 study	GWAS: 数百人規模
1型糖尿病	1 study	GWAS: 数百人規模
2型糖尿病	1 study	GWAS: 数百人規模
不育症	1 study	GWAS: 数百人規模
インスリン自己免疫症候群	1 study	GWAS: 数百人規模

スチーブンスジョンソン症候群	1 study	GWAS: 数百人規模
自閉症	1 study	GWAS: 数百人規模
多系統萎縮症	2 studies	GWAS: 数百人規模
脳動脈瘤	1 study	GWAS: 数百人規模
緑内障	1 study	GWAS: 数百人規模
放射線被ばく	1 study	GWAS: 数百人規模
肺がん	1 study	GWAS: 数百人規模
地域集団	2 studies	GWAS: 数百人規模
HLA 配列データ 1	7 検体	HLA 領域ホモ接合セルライン
HLA 配列データ 2	10 検体	完全胞状奇胎
HLA 配列データ 3	1 検体	HLA 領域ホモ接合セルライン化前 DNA
膵臓上皮嚢腫のエクソーム解析	2 検体	エクソーム解析に伴う変異データ(pool)
四肢突発性筋痛症のエクソーム解析	2 検体	エクソーム解析に伴う変異データ(pool)
永久歯萌出不全のエクソーム解析	2 検体	エクソーム解析に伴う変異データ(pool)
家族性腎臓病のエクソーム解析	4 検体	エクソーム解析に伴う変異データ
IgA 腎症のエクソーム解析	52 検体	エクソーム解析に伴う変異データ (ゲノム支援)
慢性関節リウマチのエクソーム解析	59 検体	エクソーム解析に伴う変異データ (ゲノム支援)
大腸癌のエクソーム解析	50 検体	エクソーム解析に伴う変異データ (ゲノム支援)
永久歯欠如のエクソーム解析	48 検体	エクソーム解析に伴う変異データ (ゲノム支援)
MODY のエクソーム解析	59 検体	エクソーム解析に伴う変異データ (ゲノム支援)
膵臓上皮嚢腫のエクソーム解析	8 検体	エクソーム解析に伴う変異データ
家族性腎臓病のエクソーム解析	4 検体	エクソーム解析に伴う変異データ
コーデン病のエクソーム解析	5 検体	エクソーム解析に伴う変異データ
間質性腎炎のエクソーム解析	8 検体	エクソーム解析に伴う変異データ
ウォルフラム症候群のエクソーム解析	4 検体	エクソーム解析に伴う変異データ
卵巣癌のエクソーム解析	48 検体	エクソーム解析に伴う変異データ
日本人健常者	400 検体	エクソーム解析に伴う変異データ
多系統萎縮症のエクソーム	600 検体	エクソーム解析に伴う変異データ (ゲノム支援)
日本人ゲノム全ゲノム配列	10 検体	全ゲノム配列

DBKEROへのCREST IHECデータの格納(一部完了)

Database Assembly Genomic position

DBTSS version 9.0 hg38, mm10 chr7:140,899,712-140,947,400- View

Width: 750 Height: 900 Change size Show item panel Browse Info

zoom in: 1.5x 2x 5x zoom out: 2/3x 1/2x 1/5x

chr7:140,899,712-140,947,400

Sequence

CpG Island

NCBI RefSeq

LCL/ad TSS

PC-3 RNA FPKM

PC-7 BS

PC-9 Chromatin

PC-14 H3K4me3

FERR-LC-ad1 H3K4me1

FERR-LC-ad2 H3K27me3

FERR-LC-RJ H3K4me3

FERR-LC-MS H3K36me3

FERR-LC-OK H3K9me3

VIMC-LCD H3K27ac

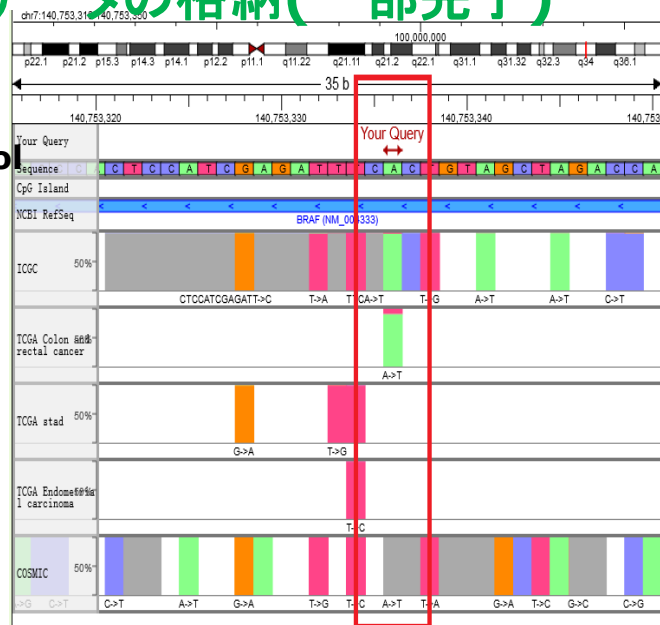
ABC-1 PolII

II-18 SNV

H3K4me3 H3K27me3 H3K9Ac H3K27Ac
H3K4me H3K9me3 H3K36me3 Pol II

ChIP-Seq

SNV



Select track items

Category buttons

Show basic public data

User data

Your Query

Public data

Sequence	CpG Island	NCBI RefSeq
----------	------------	-------------

Comparative genome

Human vs Mouse

Chromatin map

Chromatin Gm12878	Chromatin H1	Chromatin HMEC	Chromatin HSMM	Chromatin HepG2	Chromatin Huvec	Chromatin K562	Chromatin NHEK
-------------------	--------------	----------------	----------------	-----------------	-----------------	----------------	----------------

SNP/SNV public data

TCGA	TCGA Urothelial bladder carcinoma	TCGA Breast tumors	TCGA Colon and rectal cancer	TCGA Glioblastoma 2008	TCGA Glioblastoma 2013
	TCGA Acute myeloid leukemia	TCGA Squamous cell lung cancers	TCGA Ovarian carcinoma	TCGA stad	TCGA Endometrial carcinoma
ICGC	ICGC Acute Myeloid Leukemia - TARGET_US	ICGC Bladder Urothelial carcinoma - CN	ICGC Bladder Urothelial Cancer - TCGA_US	ICGC Bone	
	ICGC Breast Cancer - TCGA_US	ICGC Chronic Lymphocytic Leukemia - ES	ICGC Chronic Myeloid Disorders - UK	ICGC Colon Adenocarcinoma	
	ICGC Esophageal Adenocarcinoma - UK	ICGC Esophageal Cancer - CN	ICGC Gastric Cancer - CN	ICGC Brain Glioblastoma Multiforme - TCGA_US	
	ICGC Kidney Renal Papillary Cell Carcinoma - TCGA_US	ICGC Acute Myeloid Leukemia - KR	ICGC Brain Lower Grade Glioma - TCGA_US	ICGC	
	ICGC Liver Cancer - RIKEN_JP	ICGC Lung Cancer - Squamous cell carcinoma - KR	ICGC Lung Squamous Cell Carcinoma - TCGA_US	ICGC Mal	
	ICGC Oral Cancer - IN	ICGC Ovarian Cancer - AU	ICGC Ovarian Serous Cystadenocarcinoma - TCGA_US	ICGC Pancreatic Cancer - TCGA_US	
	ICGC Pancreatic Cancer Endocrine neoplasms - AU	ICGC Pediatric Brain Cancer - DE	ICGC Prostate Adenocarcinoma - CA	ICGC Prostate Aden	
	ICGC Rectum Adenocarcinoma - TCGA_US	ICGC Renal clear cell carcinoma - CN	ICGC Renal Cell Cancer - EU/FR	ICGC Skin Cutaneous mela	
	ICGC Thyroid Cancer - SA	ICGC Head and Neck Thyroid carcinoma - TCGA_US	ICGC Uterine Corpus Endometrial Carcinoma - TCGA_US		
	Others	26 lung adenocarcinoma cell lines	Lung adenocarcinoma 97 clinical samples	1000genome SNP	Clear-cell renal cell carcinoma
COSSMIC		Meyerson's lung adenocarcinoma	Myelodysplasia	ICGC	Nagahama Japanese SNP

Proposal; RDF as a framework to facilitate the data usage

DBTSS

Being RDF-formatted

IHEC portal

A model case of a local server

Pivotal Data Storage

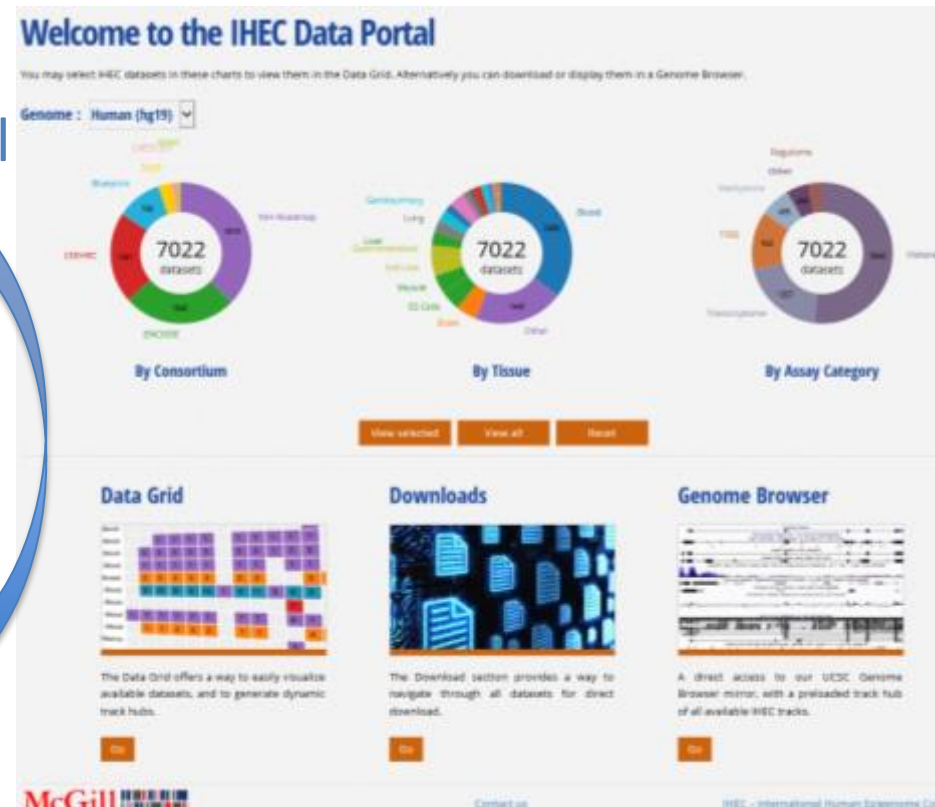
Bird-view data retrieval by “Track Hub”



Data retrieval

RDF/Sparkle
Search

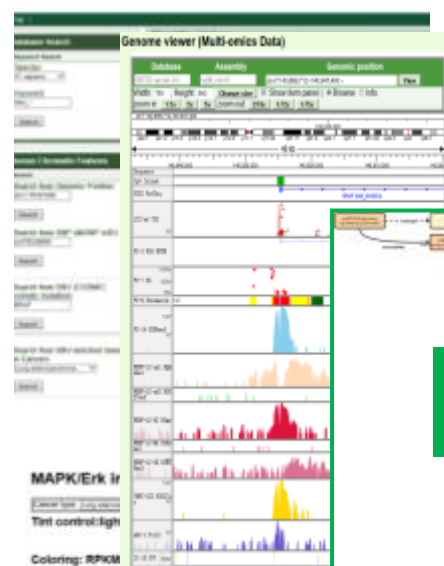
Integration of
the local data



RDF as a versatile tool for data integration

KERO DB

A model case of a local server



Integration of
the local data

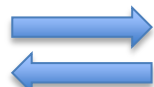


Data retrieval

RDF/Sparkle Search
As a versatile “language”

Epigenome Viewer
Pathway Viewer

Integration of
the local data



Data retrieval

Genome Databases of
Japanese SNPs/SNVs

IHEC portal

Pivotal Data Storage of IHEC



Genome Browser



Standard browser, “Track Hub”

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

NBDC
National Bioscience Database Center

AMED

ClinVar
Clinically relevant variation

International
Cancer Genome
Consortium

ROADMAP
epigenomics
PROJECT

World-wide
databases

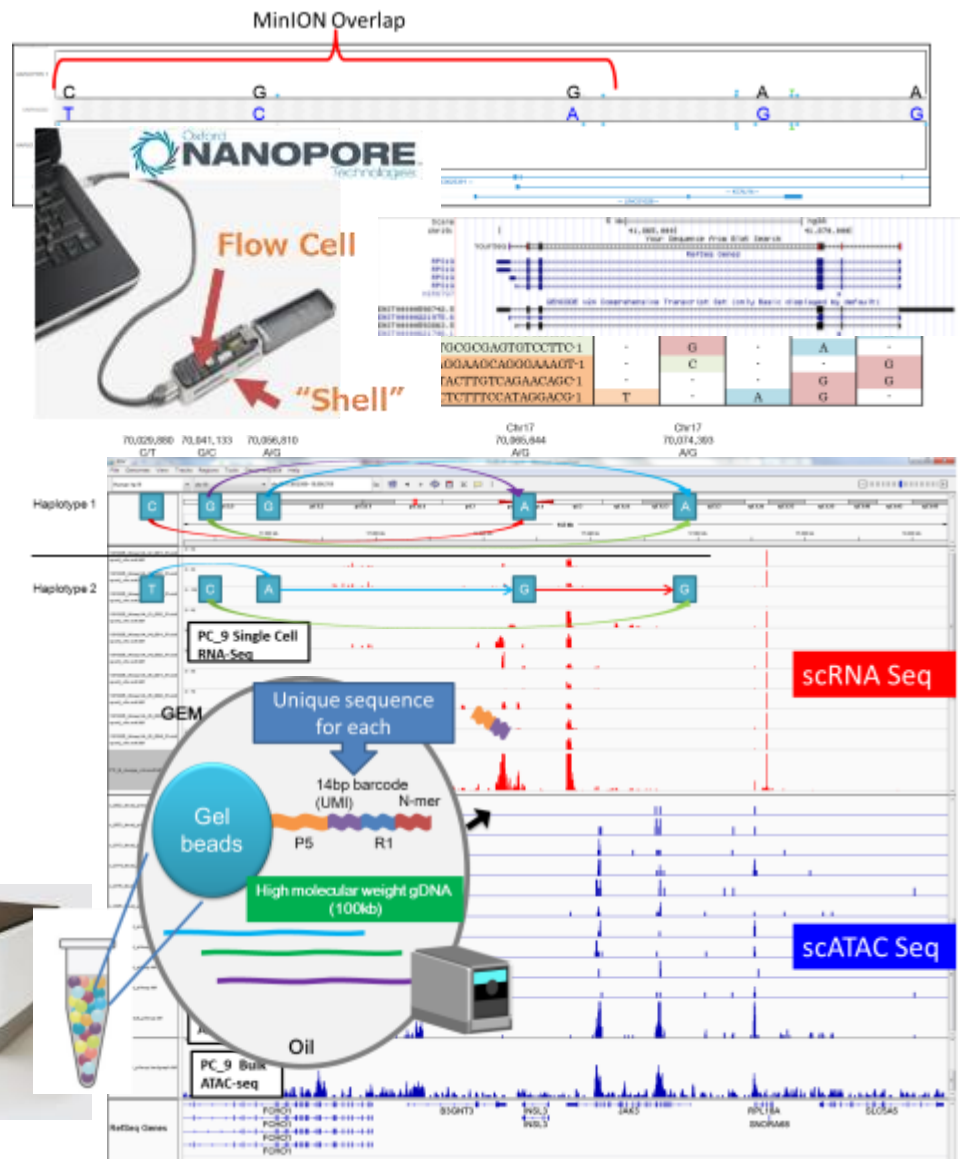
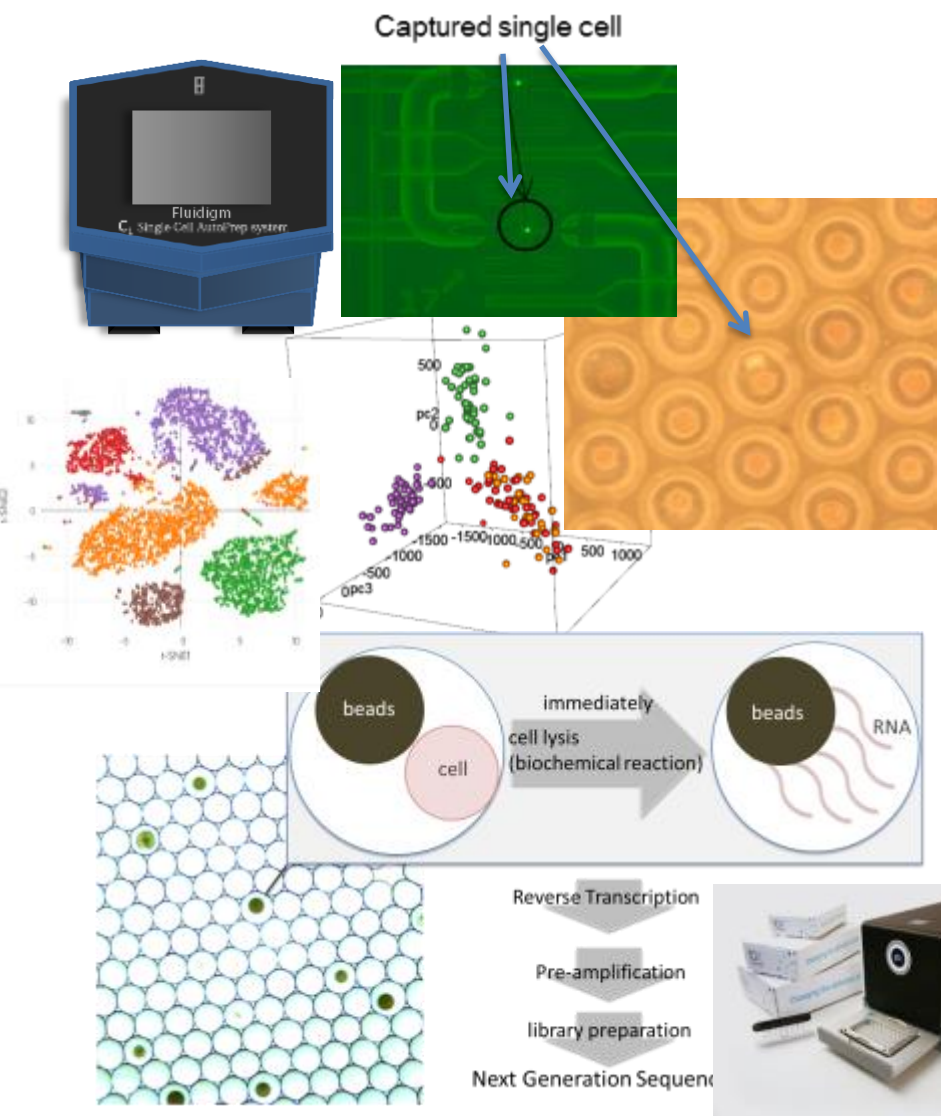


NHLBI Exome Sequencing Project (ESP)
Exome Variant Server

新規オーミクス解析技術データの収載(データ参照サイト機能)

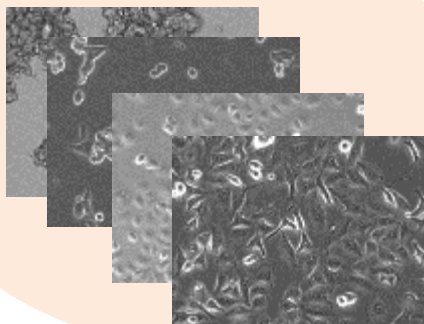
シングルセル解析 (C1, Chromium, Nx1)

ロングリード解析 (GemCode, MinION)



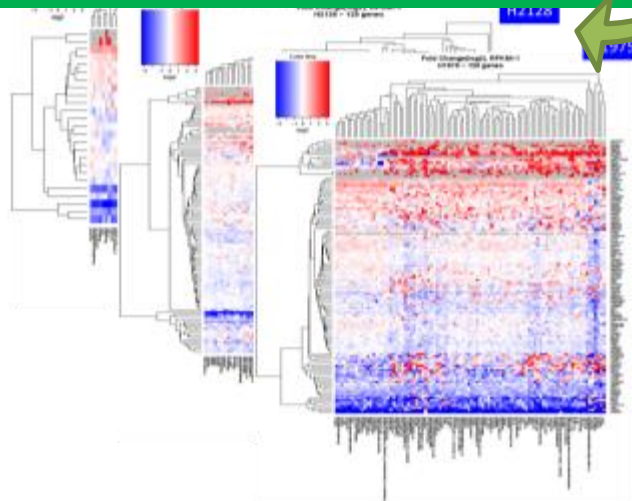
大規模薬剤摂動データ

26 lung adenocarcinoma



(National Cancer Center)

RNA/ATAC Seq x 96 compounds x 23 cell lines
 ≈ 5000 data points prepared



~Lung adenocarcinoma cell lines as a model system~

High-throughput drug screening system
 (Univ. of Tokyo)

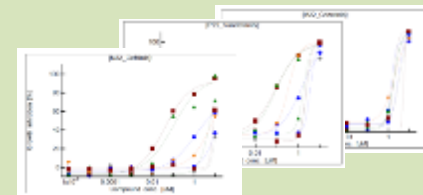
Sponsored by Fujifilm



HORNET-NX (Wako)

320 compounds

- ✓ Molecular-targeted drug
- ✓ Compounds for epigenomic changes etc...



High-throughput
 NGS sequencing
 (Univ. of Tokyo)

Illumina HiSeq2500



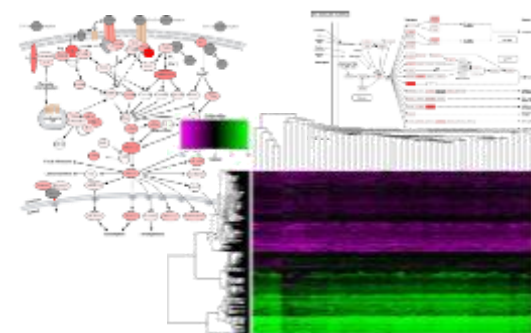
Assays for small
 amount and large
 numbers of samples

- RNA-Seq
- RRBS
- ATAC-Seq

Library prep
 • 96-well-based
 • automatics

Multi-omics sequencing data
 To KERO DB

- Epigenome (DNAmet/chromatin)
- Transcriptome



Constructing a catalogue of genome-wide multi-omics changes
 → Identification of molecular landscapes under drug perturbation

細胞株薬剤摂動外部参照データ

Cancer Program Resource Gateway

Login

Cancer Cell Line Encyclopedia (CCLE)



The Cancer Cell Line Encyclopedia (CCLE) project is an effort to conduct a detailed genetic characterization of a large panel of human cancer cell lines. The CCLE provides public access analysis and visualization of DNA copy number, mRNA expression, mutation data and more, for 1000 cancer cell lines.

[View All by Name](#)
[View Data Resources](#)
[View Tool Resources](#)

CCLE: BROAD cancer cell line encyclopedia

=479 cell lines;
24 anticancer drugs (2012時点)

➡ =~1000 cell lines;

Genome+ IC50+ transcriptome

LETTER

doi:10.1038/nature11003

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity

Jordi Barretina^{1,2,3,*}, Giordano Caponigro^{4,*}, Nicolas Stransky^{4,*}, Kavitha Venkatesan^{4,*}, Adam A. Margolin^{1,†}, Sungjoon Kim⁵, Christopher J. Wilson⁴, Joseph Lehar⁴, Gregory V. Kryukov¹, Dmitriy Sonkin⁴, Anupama Reddy⁴, Manway Liu⁴, Lauren Murray³, Michael F. Berger^{4,†}, John E. Monahan⁴, Paula Morais¹, Jodi Meltzer⁴, Adam Korejwa¹, Judit Jané-Valbuena^{1,2}, Felipa A. Mapa⁴, Joseph Thibault⁴, Eva Bric-Furlong⁴, Pichai Raman⁴, Aaron Shipway⁵, Ingo H. Engels⁵, Jill Cheng⁶, Guoying K. Yu⁶, Jianjun Yu⁶, Peter Aspesi Jr⁴, Melanie de Silva⁴, Kalpana Jagtap⁴, Michael D. Jones⁴, Li Wang⁴, Charles Hutton³, Emanuele Palescandolo³, Supriya Gupta⁴, Scott Mahan¹, Carrie Sougnez⁴, Robert C. Onofrio¹, Ted Liefeld⁴, Laura MacConaill³, Wendy Winckler⁴, Michael Reich¹, Nanxin Li⁵, Jill P. Mesirov³, Stacey B. Gabriel¹, Gad Getz¹, Kristin Ardlie³, Vivien Chan⁶, Vic E. Myer⁴, Barbara L. Weber⁴, Jeff Porter⁴, Markus Warmuth⁴, Peter Finan⁴, Jennifer L. Harris⁵, Matthew Meyerson^{1,2,3}, Todd R. Golub^{1,3,7,8}, Michael P. Morrissey^{4,*}, William R. Sellers^{4,*}, Robert Schlegel^{4,*} & Levi A. Garraway^{1,2,3,*}



COSMIC Cell Line Catalogue

[Home](#) ▼

[Resources](#) ▼

[Curation](#) ▼

[Tools](#) ▼

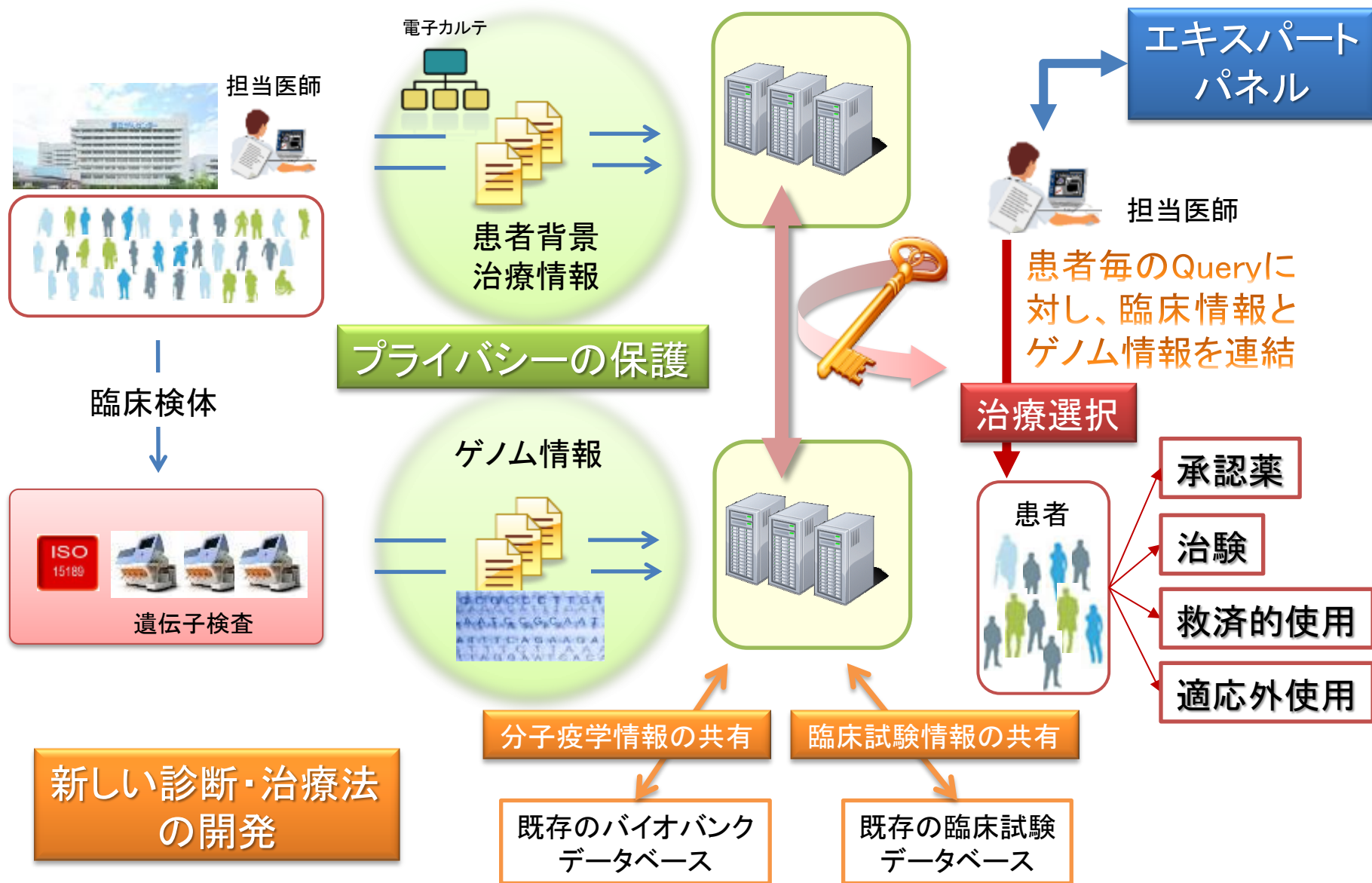
[Data](#)

=~1000 cell lines;

Cell Lines Project v80, released 13-FEB-17

技術開発

がん“最適化創薬”の実現に向けた 臨床情報・ゲノム情報管理システムの構築



JSPS “Platform for Advanced Genome Science (PAGS)”

<http://www.genome-sci.jp/>

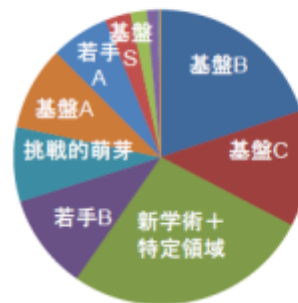
文部科学省科学研究費助成事業「新学術領域研究『学術研究支援基盤形成』」
先進ゲノム解析研究推進プラットフォーム

先進ゲノム支援

ホーム 先進ゲノム支援とは 公募要項 支援申請の流れ 支援申請 公募支援に関するFAQ お問い合わせ

新学術領域研究「先進ゲノム解析システム」
提供することにより様々な
この過程でゲノム解析シ
ひいては生命科学のすそ
ます。この支援にふさわ

②支援採択課題の元科研費

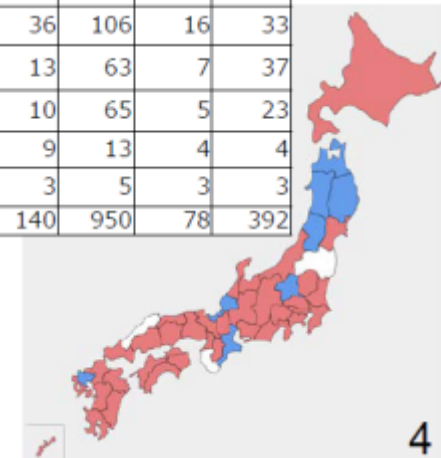


全期間の応募・採択数

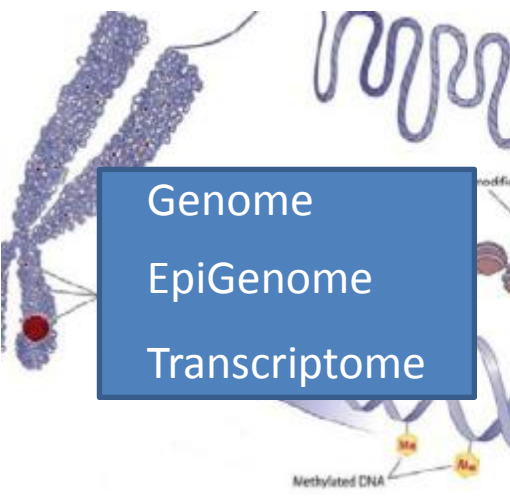
年度	H22	H23	H24	H25	H26	H27
応募数	166	237	200	172	161	211
採択数	69	62	81	88	91	74

③支援依頼者所在機関 (H26まで)

	応募		採択	
	機関	件数	機関	件数
国立大学	53	662	38	282
公立大学	16	36	5	10
私立大学	36	106	16	33
国立研究所	13	63	7	37
独法研究所	10	65	5	23
公立研究所	9	13	4	4
財団研究所	3	5	3	3
合計	140	950	78	392



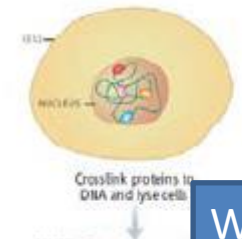
Providing latest analytical platforms for every layer of omics studies



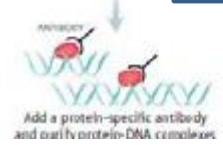
Sequencers
-HiSeq/MiSeq
-PacBio



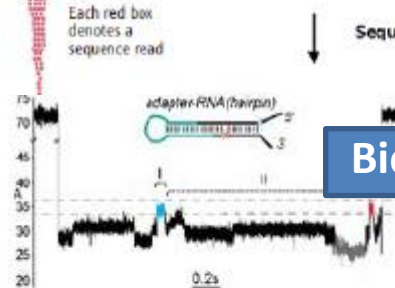
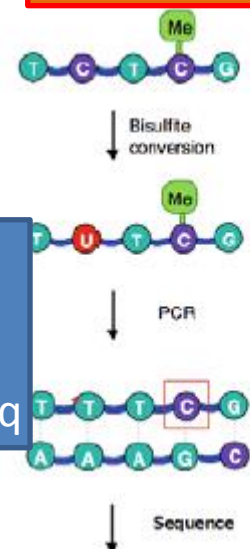
Analytics Platforms
-Single Cell
-USB Sequencer
-Long Read



Reverse crosslinks and isolate DNA



WES/WGS
ChIP Seq/BS Seq
RNA Seq/CLIP Seq



Bioninformatics



・表3 IHECデータ

金井チーム	器官数 (3)	肝臓(8 検体)、腸管(12 検体)、胃(14 検体)	エピゲノムデータセット数 (267)
白髭チーム	器官数 (1)	血管内皮(43 検体)	エピゲノムデータセット数 (207)
佐々木チーム	器官数 (1)	子宮内膜(17 検体)	エピゲノムデータセット数 (147)

・表4 先進ゲノム支援

ヒト	データセット数 (100)	トランスクリプトーム/エピゲノムデータセット数 (1,000)
マウス等モデル生物	データセット数 (200)	トランスクリプトーム/エピゲノムデータセット数 (2,000)

・表5 シングルセル、薬剤摂動多層オームクスデータ

シングルセルデータ肺腺癌培養細胞(26 細胞種)	臨床検体由来がん浸潤白血球(50 検体):臨床胃がん(50 検体)	
薬剤摂動多層オームクスデータ肺腺癌培養細胞 (26 細胞株)、胃がん培養細胞株 (20 細胞株)	トランスクリプトーム摂動データ (26 x 1,000 データポイント)	エピゲノム摂動データ (26 x 1,000 データポイント)

研究開発テーマ	タスク	小タスク	担当グループ	開始点	終了点	依存タスク
DBKERO拡張	CREST-IHECデータ 取載	Data QC DB設計 データ公開	菅野G	2017・4（PRG 開始）	2018・3	Data QC DB調整
	HGVDDBデータ取載	プロトタイプGWAS	菅野G、徳永G、井上 G	2017・4	2018・3	DBデザイン ブラウザ開発
		プロトタイプHLA	菅野G、徳永G、井上 G	2017・4	2018・3	DBデザイン ブラウザ開発
		DB完成	菅野G、徳永G、井上 G	2018・4	2021・3	プロトタイプGWAS/HLA
		データ公開	菅野G、徳永G、井上 G	2021・3	2022・3	DB完成
新規計測技術DB作 成	シングルセルData QC、DB設計	プロトタイプSc	菅野G、土原G	2017・4	2018・3	
	薬剤摂動Data QC、 DB設計	プロトタイプPb	菅野G、土原G	2017・4	2018・3	
	データベース公開		菅野G、土原G	2018・4	2020・3	Data QC、DB完成
RDF化	RNA Seq	運用開始	菅野G、	2017・4	2018・3	最終動作確認
	ChIP Seq	基礎検討	菅野G、	2017・4	2018・3	
		運用開始	菅野G	2018・4	2019・3	基礎検討の終了
	ゲノム多型	適用範囲検討	菅野G、土原G	2017・4	2018・3	
		部分的運用開始	菅野G、土原G	2018・4	2019・3	基礎検討の終了
		全面運用条件検討	菅野G、土原G	2018・4	2021・3	部分的運用の成功
		全面運用開始	菅野G、土原G	2021・4	2022・3	全面運用条件検討の成功

スケジュール

			29年度		30年度		31年度		32年度		33年度	
			上期	下期	上期	下期	上期	下期	上期	下期	上期	下期
疾患ゲノム多層 オーミクスデータ ベース拡充	番号	開発項目										
	1	外部参照データの更新										
	2	HGVDBとの統合 (日本人ゲノム多型)										
	3	CREST-IHECデータの統合 (日本人標準エピゲノム)										
	4	先進ゲノム支援データ等(マウスを中心 に)のモデル生物実験系データ										
	5	RDFによるデータ統合検索の 運用の開始										
技術開発	6	臨床ゲノムデータベース、 バイオバンクデータベースとの連携										
	7	新規オーミクス解析技術(シングルセル、 薬剤摂動)ブラウザの開発とデータ公開										

 菅野班

 土原班

 徳永班

 公開開始時期

年度	物品費		(3)旅費	(4)人件費・謝金	(5)その他	合計
	(1)設備備品費	(2)消耗品費				
2017年度	5,490	700	210	11,800	6,800	25,000
2018年度	5,500	690	210	11,800	6,800	25,000
2019年度	5,500	690	210	11,800	6,800	25,000
2020年度	5,500	690	210	11,800	6,800	25,000
2021年度	5,500	690	210	11,800	6,800	25,000
合計	27,490	3,460	1,050	59,000	34,000	125,000

No.	研究機関名	研究代表者名または 主たる共同研究者名	物品費		(3)旅費	(4)人件費・謝金	(5)その他	合計
			(1)設備備品費	(2)消耗品費				
01	東京大学	菅野 純夫	5,000	200	0	8,000	1,800	15,000
02	国立がん研究センター	土原 一哉	0	0	0	0	5,000	5,000
03	国立大学法人 東京大学	徳永 勝士	490	500	10	0	0	1,000
04	国立遺伝学研究所	井ノ上 逸朗	0	0	200	3,800	0	4,000
		合 計	5,490	700	210	11,800	6,800	25,000

達成目標1

(1)～第3年次末

達成目標	タスク
DBKEROの拡張	・ CREST-IHECの公開 ・ HGVDBデータの統合と試験的運用 ・ 先進ゲノム支援データの取載開始（100データセット）
新規ゲノム計測技術データの加工と試験的公開	・ シングルセルデータの部分的公開（10データセット） ・ 薬剤振動データの公開（100データセット）
RDF化の実装	・ RNA Seqデータの公開 ・ ChIP Seqデータ統合への予備的検討の終了 ・ ゲノムデータに関する基礎検討

達成目標2

(2) ～第5年次末

達成目標	タスク
DBKEROの拡張	・ 国際IHECデータの収載 ・ HGVDBデータの統合の完了 ・ 先進ゲノム支援データの公開（300データセット）
新規ゲノム計測技術データの公開	・ シングルセルデータの公開（合計100データセット） ・ 薬剤摂動データの公開（1000データセット） ・
RDF化の運用開始	・ RNA/ChIP Seqの運用開始 ・ ゲノムデータの部分的運用 ・ 全ゲノムデータ運用の予備的検討の修了 ・ メタデータのRDF化の予備的検討の修了

データベースの拡充

疾患ゲノム多層オーミクスデータベース拡充 (KERO; <http://kero.hgc>)

1. 外部参照データの更新
2. HGVDBとの統合(日本人ゲノム多型)
3. CREST-IHECデータの統合(日本人標準エピゲノム)
4. 先進ゲノム支援データ等(マウスを中心に)のモデル生物実験系データ
5. RDFによるデータ統合検索の運用の開始
6. 臨床ゲノムデータベース、バイオバンクデータベースとの連携

技術開発

7. 新規オーミクス解析技術(シングルセル、薬剤摂動)ブラウザの開発とデータ公開

研究開発の進め方

オントロジーの整備

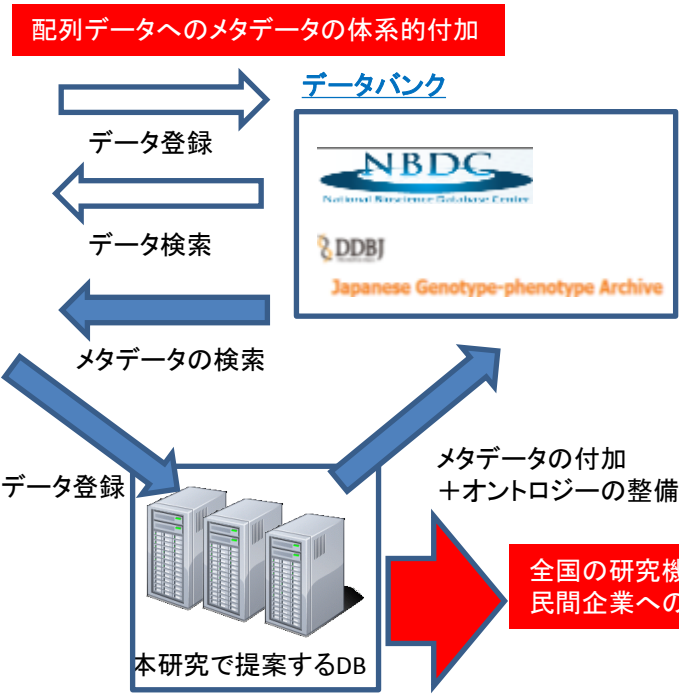


申請者らが自身によるメタデータの付加、オントロジーの整備が可能

将来展望



データバンク機関の機能補完



全国の研究機関
民間企業へのデータ発信

データベース・検索システムの構築

申請者らが運用する既存のデータベース・検索システムの転用

基本的な枠組みは構築済み！データは加工済み！！

現有するデータは初年度内での整備・公開を目指す



検索例とブラザの詳細は
配布資料参照



配布資料1

最終年度の高度検索システムの開発へと初年度から注力

高度検索システムの実装

配布資料2

エピゲノム・トランスクリプトームデータのパターンからの検索
パスウェイへの変異濃縮パターンからの検索

機微情報に留意した動的検索システムとの統合

毎日変化する臨床データ
ある程度静的なオミクスデータ

臨床-ゲノム統合データベース(開発中)
本データベースとの統合

多疾患への応用

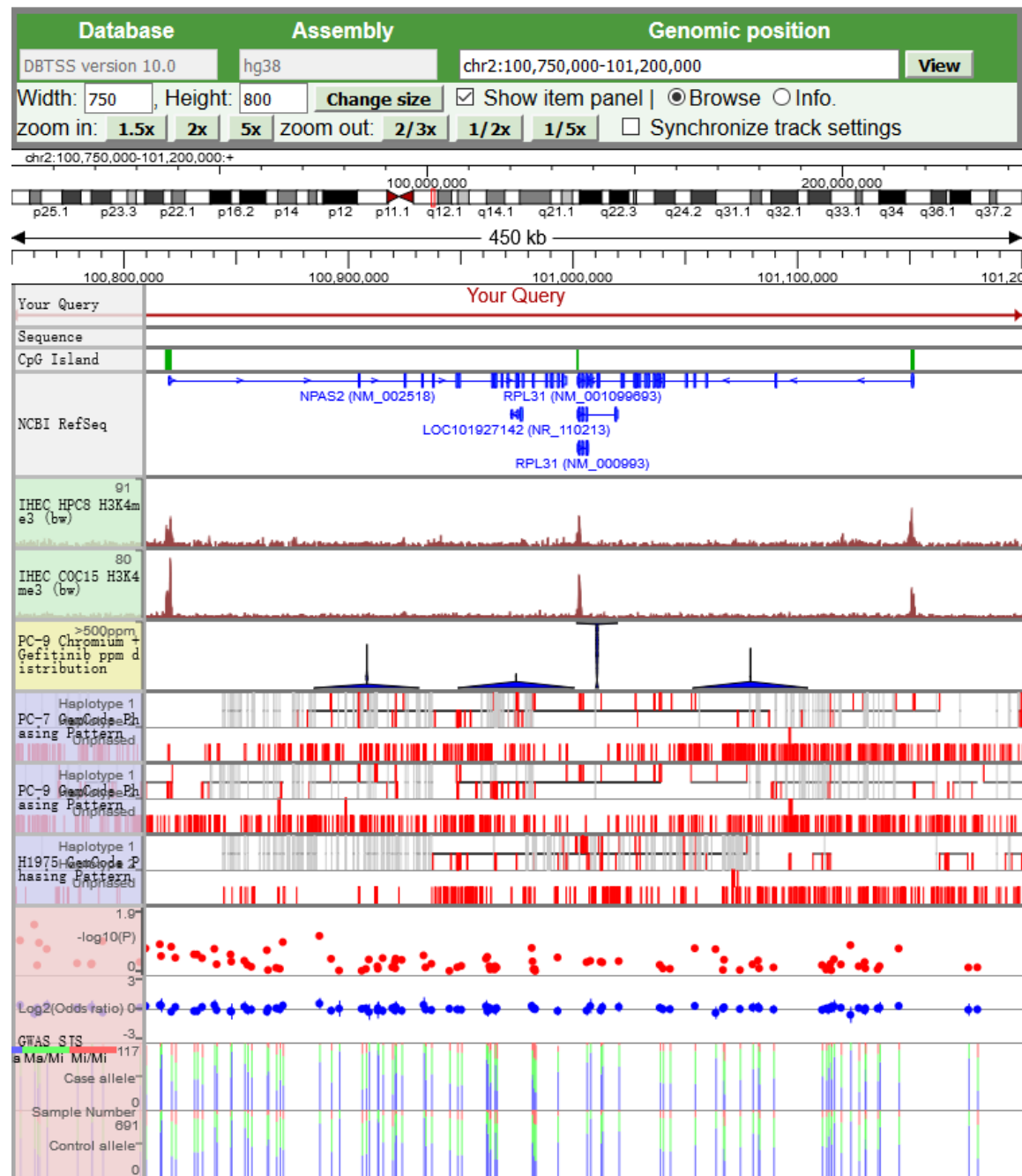
精神神経疾患
糖尿病、高血圧、
アレルギー疾患

臨床検体試料の
採取・検証実験が困難
モデル系のオミクスデータ統合の有用性

(人材の育成)

バイオインフォマティクス人材育成の対象の多くがnon-MD/大学院生
培養細胞系、モデル動物系を有効活用した情報学的基盤
の構築は、臨床検体に関する種々の倫理問題を回避

該当分野の人材育成も加速する。



IHEC (緑)

10x Chromium (黄)
(Single cell expression distribution)

10x GemCode (青)
(Phasing using GemCode)

GWAS (赤)
(作成中: p値、オッズ比、
case/control allele frequency)

Track item details:

+ IHEC (Japan)

+ Single cell (lung cancer)

+ 10x GemCode (lung cancer)

+ GWAS

+ User data

+ Public data

+ Comparative genome

トグルになっていて[+]をクリックすると、トラックを選択できるボタンが現れます。

Track item details

IHEC (Japan)

Type	Sample		Expression								DNA methylation			Infinium chip	H3K4me1	
			RNA-seq								BS-seq					
			Read depth				Exon/intron junction									
			Show all HPC	Show all COC	Show all JKU	Show all JKU	Show all HPC	Show all COC	Show all JKU	Show all JKU	Show all HPC	Show all COC	Show all JKU		Show all HPC	Show all COC
Hide all HPC	Hide all COC	Hide all JKU	Hide all JKU	Hide all HPC	Hide all COC	Hide all JKU	Hide all JKU	Hide all HPC	Hide all COC	Hide all JKU	Hide all HPC	Hide all COC	Hide all JKU	Hide all HPC	Hide all COC	
Liver	Liver, hepatocytes (HPC6)	Show all Hide all	IHEC HPC6 RNA-seq (bw)				IHEC HPC6 RNA-seq junction				IHEC HPC6 BS			1	IHEC HPC6 (1) H3K4me1 (bw) IHEC HPC6 (2) H3K4me1 (bw) IHEC HPC6 (3) H3K4me1 (bw)	
	Liver, hepatocytes, hepatitis C virus-positive chronic hepatitis (HPC8)	Show all Hide all	IHEC HPC8 RNA-seq (bw)				IHEC HPC8 RNA-seq junction				IHEC HPC8 BS			1	IHEC HPC8 H3K4me1 (bw)	
	Liver, hepatocytes, hepatitis B virus-positive chronic hepatitis (HPC17)	Show all Hide all	IHEC HPC17 RNA-seq (bw)				IHEC HPC17 RNA-seq junction				IHEC HPC17 BS			1	IHEC HPC17 H3K4me1 (bw)	
	Liver, hepatocytes (HPC20)	Show all Hide all	IHEC HPC20 RNA-seq (bw)				IHEC HPC20 RNA-seq junction				IHEC HPC20 BS			1	IHEC HPC20 (1) H3K4me1 (bw) IHEC HPC20 (2) H3K4me1 (bw) IHEC HPC20 (3) H3K4me1 (bw)	
	Liver, hepatocytes (HPC25)	Show all Hide all	IHEC HPC25 RNA-seq (bw)				IHEC HPC25 RNA-seq junction				IHEC HPC25 BS			1	IHEC HPC25 H3K4me1 (bw)	
	Liver, hepatocytes (HPC27)	Show all Hide all	IHEC HPC27 RNA-seq (bw)				IHEC HPC27 RNA-seq junction				IHEC HPC27 BS			1	IHEC HPC27 H3K4me1 (bw)	
	Liver, hepatocytes (HPC28)	Show all Hide all	IHEC HPC28 RNA-seq (bw)				IHEC HPC28 RNA-seq junction				IHEC HPC28 BS			1	IHEC HPC28 H3K4me1 (bw)	
	Liver, hepatocytes (HPC35)	Show all Hide all	IHEC HPC35 RNA-seq (bw)				IHEC HPC35 RNA-seq junction				IHEC HPC35 BS			1	IHEC HPC35 (1) H3K4me1 (bw) IHEC HPC35 (2) H3K4me1 (bw)	
	human colon absorptive epithelial cells (COC4)	Show all Hide all	IHEC COC4 RNA-seq (bw)				IHEC COC4 RNA-seq junction				IHEC COC4 BS				IHEC COC4 H3K4me1 (bw)	
human colon absorptive epithelial cells (COC5)	Show all Hide all	IHEC COC5 RNA-seq (bw)				IHEC COC5 RNA-seq junction				IHEC COC5 BS				IHEC COC5 H3K4me1 (bw)		