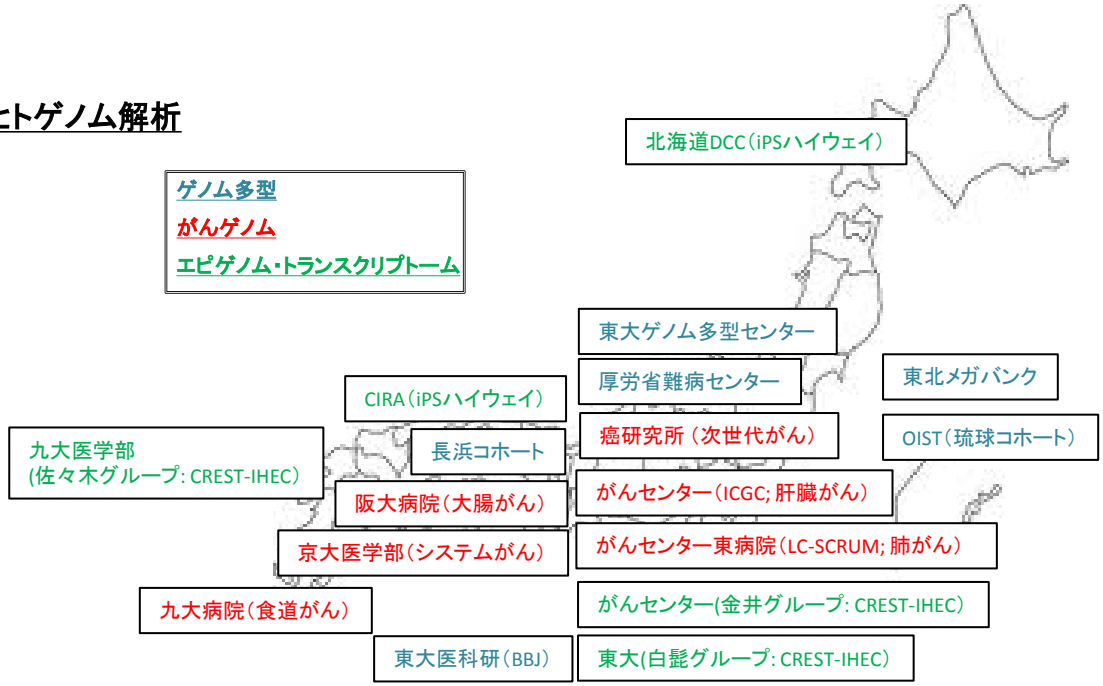
疾患ヒトゲノム変異の生物学的機能注釈を目指した多階層オミクスデータの統合

東京大学新領域創成科学研究科
菅野純夫

背景2

全国に展開するヒトゲノム解析



ゲノムデータは急速に蓄積している

ヒトオミクスデータ推定蓄積量

ゲノム多型(WGS/WES）: >2000人
がんゲノム(WGS/WES/Target Seq）: >1000症例
トランスクリプトーム(RNA Seq）: >1000例
エピゲノム(BS/ChIP Seq）: <100例

+培養細胞＋PDX＋モデル系: >5000例
+マウス等モデル生物: ???例

+個別研究者の蓄積するオミクス情報: ???例

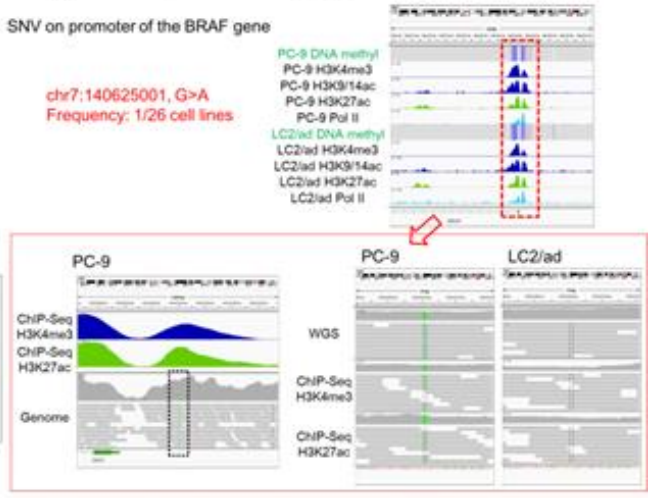
多くの生物種に応用可能

背景3

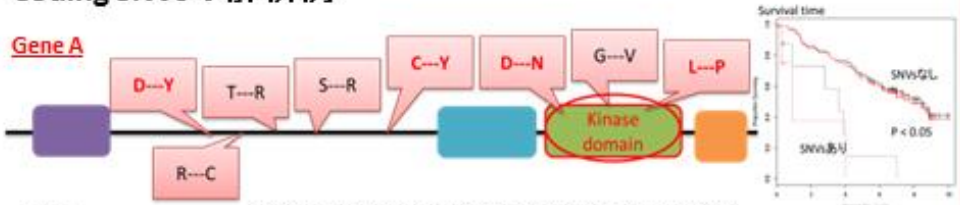
データ統合が目指すヒトゲノム臨床応用研究（がんへの応用を例に）

WGS/WES解析

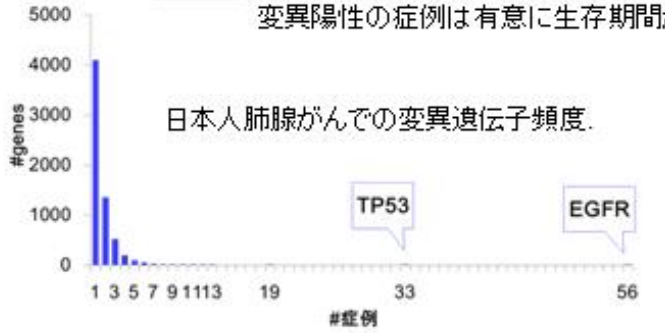
Regulatory SNVsの解析



Coding SNVsの解析例



変異陽性の症例は有意に生存期間が短い.



・症例間で変異遺伝子が重複することは例外的な遺伝子を除いて、まれ

・Passenger変異<->Driver変異の区分が困難

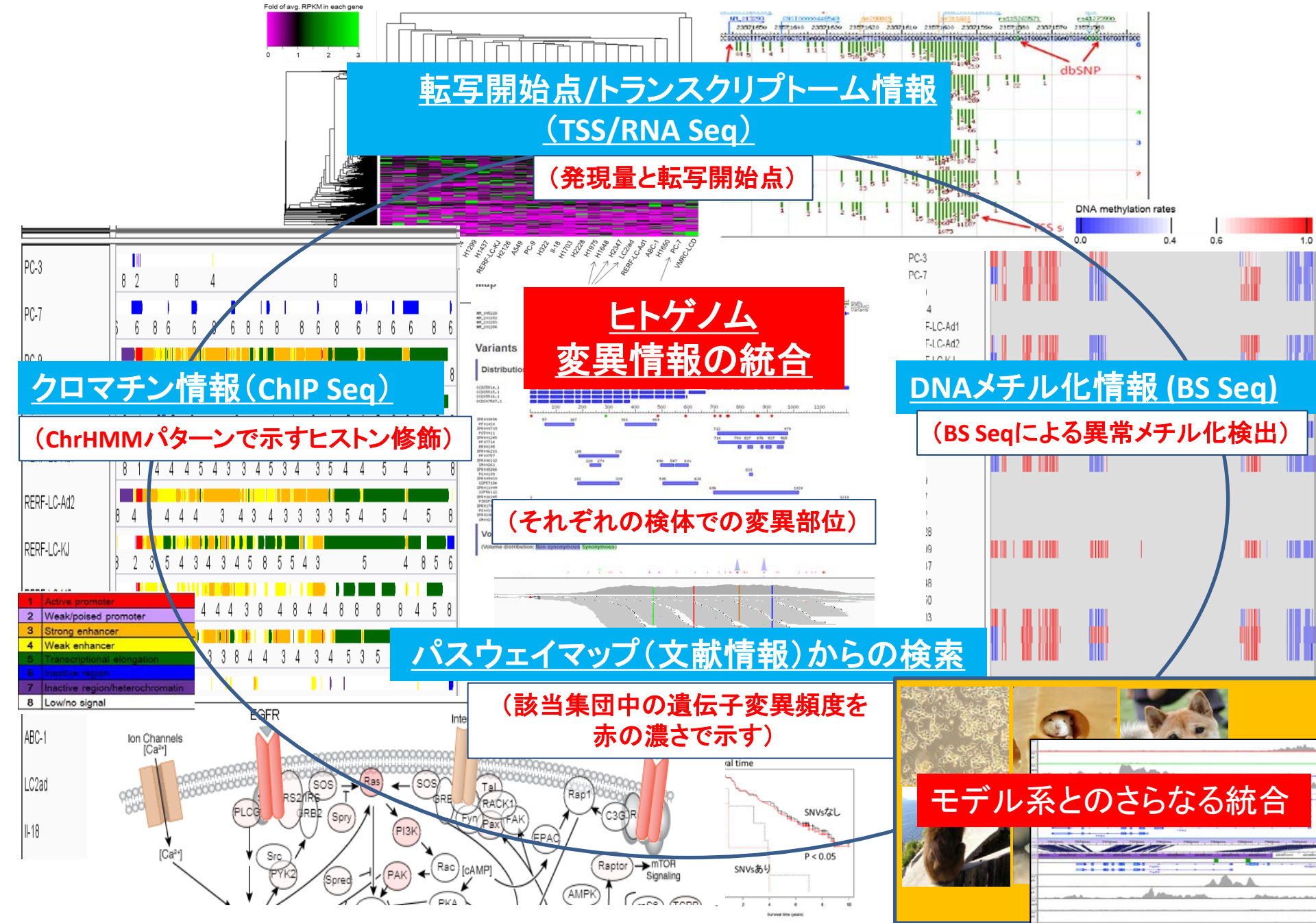
・Regulatory SNPについての情報が圧倒的に不足

×

ゲノム情報が創薬ゲノミクス・臨床応用へ直結しない

構想

ヒト応用研究を志向したオミクス情報の統合（EGFR遺伝子を例に）



データベースの公開

- DataBase of Transcriptional Start Sites -
DBTSS

Release 9.0 Updated (September 15 2014
Based on UCSC hg38, mm1

Top |

Database Search
Keyword Search
Species: H. sapiens
Keyword: NM_*
Search

Human Chromatin Features
Search
Search from Genomic Position: chr1:75787000
Search
Search from SNP (dbSN rs375229869
Search
Search fr somatic mutation): BRAF
Search
Search from SNV-enriched Gene in Cancers: Lung adenocarcinoma ...
Search

About this database

Welcome to **DBTSS (DataBase of Transcriptional Start Sites)**.

To support transcriptional regulation studies, we have constructed the DBTSS (DataBase of Transcriptional Start Sites), which represents exact positions of transcriptional start sites (TSSs) in the genome based on our unique experimentally validated TSS sequencing method, TSS-seq.

This database includes TSS data of a major part of human adult and embryonic tissues are covered. DBTSS now contains 491 million TSS tag sequences for collected from a total of 20 tissues and 7 cell cultures. We also integrated our newly generated RNA-seq data of subcellular- fractionated RNAs and ChIP-seq data of histone modifications, RNA polymerase II and several transcriptional regulatory factors in cultured cell lines. We also included recently accumulating external epigenomic data, such as chromatin map of the ENCODE project.

In this update, we further associated those TSS information with public and original SNV regio

It is b omic

News

- 15 Sep. 2014: New DBTSS opened.

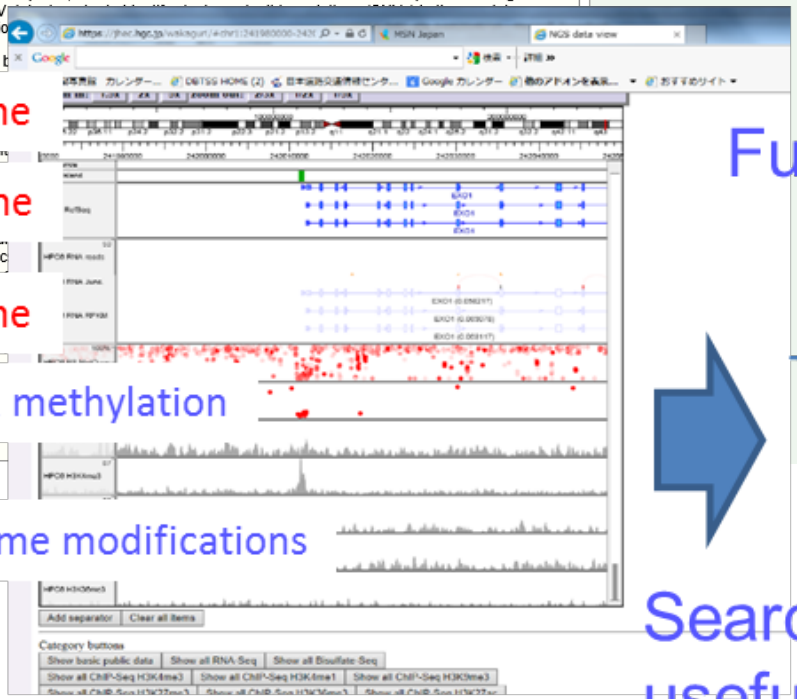
Genome

Transcriptome

Epigenome

DNA methylation

Histome modifications



Database for multi-omics analysis

Further integration with

- Image data
- Clinical information

Search for novel/more useful biomarkers at “bulk” level

Data contents

Number of dataset

Data contents

	Human	Mouse	Malaria	Chyzon	Rat	Chimpanzee	Macaque
TSS-seq	73	7	1	1	1	1	2
RNA-seq	42	0	0	0	0	0	0
ChIP-seq	255	0	0	0	0	0	0
RIP-seq	12	0	0	0	0	0	0
BS-seq	26	0	0	0	0	0	0
ChromHMM	36	0	0	0	0	0	0
SNV	49	0	0	0	0	0	0

Clinical samples

Data contents

Data provider	Cancer type	Number of samples	Reference
National Cancer Center Hospital East and University of Tokyo	Lung adenocarcinoma	97	PLoS One. 2013 Sep 12;8(9):e73484.
National Cancer Center Hospital East	Small cell lung cancers	57	J Thorac Oncol. 2014 Sep;9(9):1324-31.
ICGC	43 of ICGC DCC Project Codes	6,590	https://dcc.icgc.org/
Dr. Meyerson's Lab.	Lung adenocarcinoma	183	Cell. 2012 Sep 14;150(6):1107-20.
Dr. Ogawa's Lab.	Myelodysplasia	29	Nature. 2011 Sep 11;478(7367):64-9
	Clear-cell renal cell carcinoma	106	Nat Genet. 2013 Aug;45(8):860-7
TCGA	Gastric adenocarcinoma	295	Nature. Published online 23 July 2014
	Urothelial bladder carcinoma	131	Nature. 507 (7492):315-22.
	Glioblastoma	291	Cell. 155 (2):462-477.
	Clear cell renal cell carcinoma	446	Nature. 499 (7456):43-49.
	Endometrial carcinoma	373	Nature. 497 (7447):67-73.
	Acute myeloid leukemia	200	NEJM. 368:2059-2074.
	Breast tumors	507	Nature. 490 (7418):61-70.
	Squamous cell lung cancers	178	Nature. 489 (7417):519:525.
	Colon and rectal cancer	224	Nature. 487 (7407):330-337.
	Ovarian carcinoma	316	Nature. 474 (7353):609-615.
	Glioblastoma	91	Nature. 455 (7216):1061-1068.
HGVD	Normal (Japanese)	1,208	URL: http://www.genome.med.kyoto-u.ac.jp/SnpDB
Total		11,322	

＋ファーマコゲノミクス（＞2000）＋　東大多型センターとのリンク

A

- DataBase of Transcriptional Start Sites -

DBTSS

Release 9.0 Updated (September 15 2014)
Based on UCSC hg38, mm10

Top |

Database Search

Keyword Search
Species: H. sapiens
Keyword: NM_*
Search

Lift over:
hg38
chr1:99,950,000-100,050,000
Search

Human Chromatin Features

Search
Search from Genomic Position:
chr1:75,787,000
Search

Search from SNP (dbSNP rsID):
rs375229869
Search

Search from SNV (COSMIC: somatic mutation):
BRAF
Search

Search from SNV-enriched Gene in Cancers:
Lung adenocarcinoma ...
Search

SNV Summary in Cancers:
NM_*
Search

Pathway Map

Species: H. sapiens
Search

About this database

Welcome to **DBTSS (DataBase of Transcriptional Start Sites)**

To support transcriptional regulation studies, we have constructed the DBTSS (DataBase of Transcriptional Start Sites), which represents exact positions of transcriptional start sites (TSSs) in the genome based on our unique experimentally validated TSS sequencing method, TSS-seq.

This database includes TSS data of a major part of human adult and embryonic tissues are covered. DBTSS now contains 491 million TSS tag sequences for collected from a total of 20 tissues and 7 cell cultures. We also integrated our newly generated RNA-seq data of subcellular- fractionated RNAs and ChIP-seq data of histone modifications, RNA polymerase II and several transcriptional regulatory factors in cultured cell lines. We also included recently accumulating external epigenomic data, such as chromatin map of the ENCODE project.

In this update, we further associated those TSS information with public and original SNV data, in order to identify single nucleotide variations (SNVs) in the regulatory regions.

It is believed that single nucleotide variations (SNVs) in the transcriptional regulatory regions are responsible for many human diseases, including cancers. However, it remains difficult to identify functionally relevant SNVs from those having no explicit biological consequences. In this version of DBTSS, we attempt to associate SNVs with the omics information of the surrounding regions. We used SNVs which we identified from genomic analyses of various types of cancers, including somatic mutations of 100 lung adenocarcinoma and lung small cell carcinoma. For germline variations, we used SNVs in dbSNP as well as our unique dataset of variations in 1000 Japanese individuals. We integrated those SNV information with our original datasets of TSS-seq, RNA-seq, ChIP-seq of representative histone modifications and Bisulfite Sequencing of cytosine methylations of DNA. Particular, we present multi-omics data of 26 lung adenocarcinoma cells line for which TSS-seq, RNA-seq, ChIP-seq and BS-seq together with whole genome sequencing are collected from the same materials. We further connected the multi-omics data of model organisms by genome-genome alignment. We provide a unique data resource to investigate what genomic features are observed in a particular genomic coordinates in a wide variety of samples.

These data can be browsed in our new viewer which also supports versatile search conditions of users. We believe new DBTSS is helpful to understand biological consequences of the massively identified TSSs and identify human genetic valuations which are associated with disordered transcriptional regulations.

News

- 15 Sep. 2014: New DBTSS opened.

References

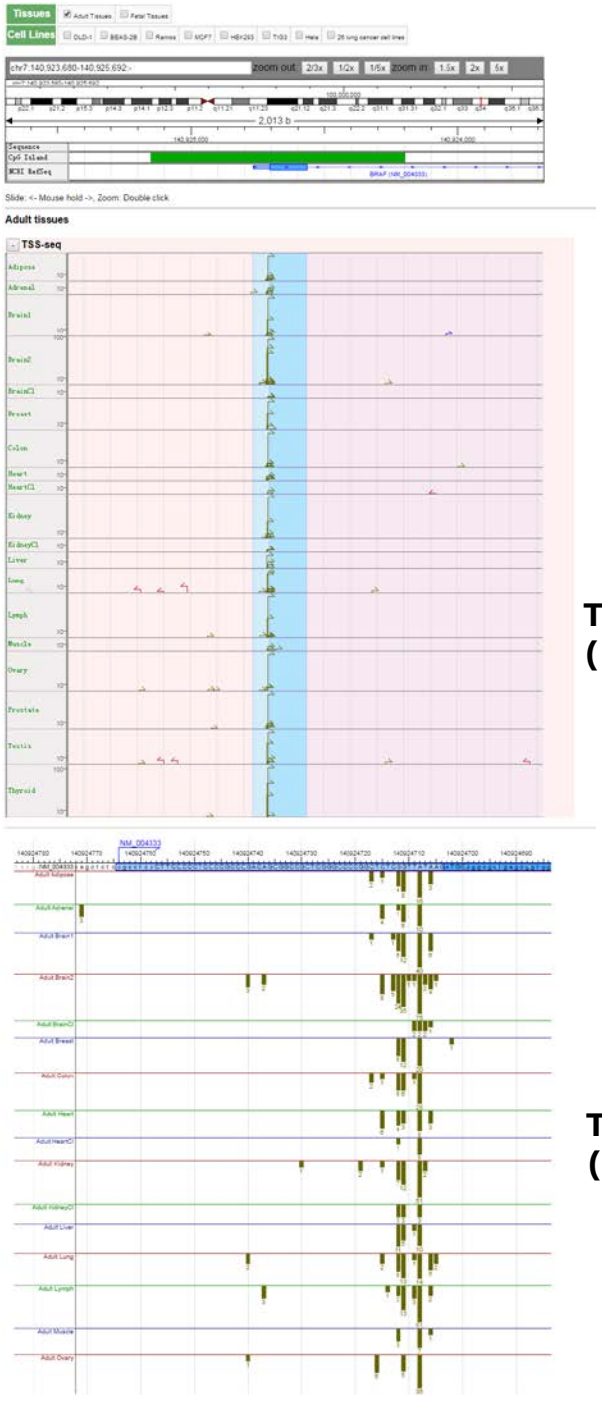
Suzuki A, Mimaki S, Yamane Y, Kawase A, Matsushima K, Suzuki M, Goto K, Sugano S, Esumi H, Suzuki Y, Tsuchihara K. Identification and characterization of cancer mutations in Japanese lung adenocarcinoma without sequencing of normal tissue counterparts. *PLoS One*. 2013 Sep 12;8(9).

Yamashita R, Sugano S, Suzuki Y, Nakai K. DBTSS: DataBase of Transcriptional Start Sites progress report in 2012. *Nucleic Acids Res*. 2012 Jan;40(Database issue):D150-4.

Yamashita R, Sathira NP, Kanai A, Tanimoto K, Arauchi T, Tanaka Y, Hashimoto S, Sugano S, Nakai K, Suzuki Y. (2011) Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res*. 2011 Mar 3.

Tsuchihara K, Suzuki Y, Wakaguri H, Irie T, Tanimoto K, Hashimoto S, Matsushima K, Mizushima-Sugano J, Yamashita R, Nakai K, Bentley D, Esumi H and Sugano S. (2009) Massive transcriptoional start site analysis of human genes in hypoxia cells. *Nucleic Acids Res*. 2009 Feb 22.

B



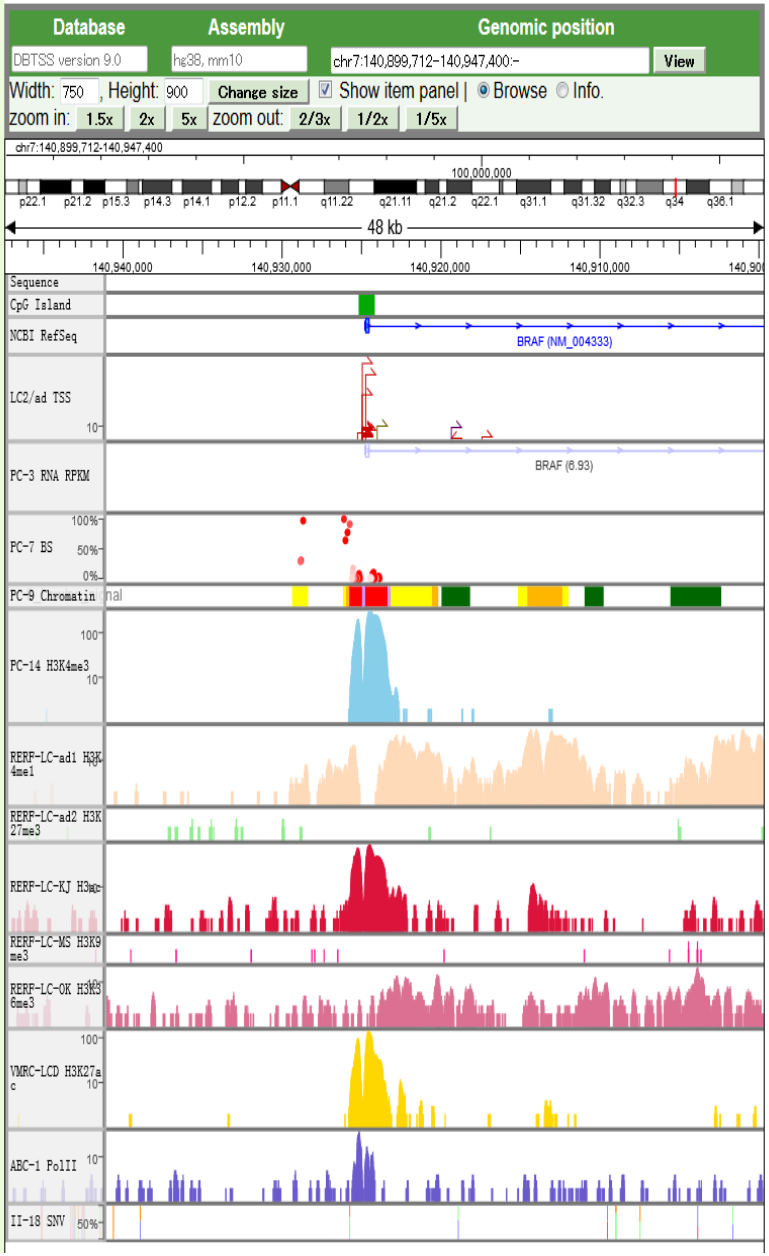
TSS Overview (Upper panel)

TSS Details (Lower panel)

Figure 1

C

Genome viewer (Multi-omics Data)



Viewer control

Gene model

TSS-Seq

RNA-Seq

BS-Seq

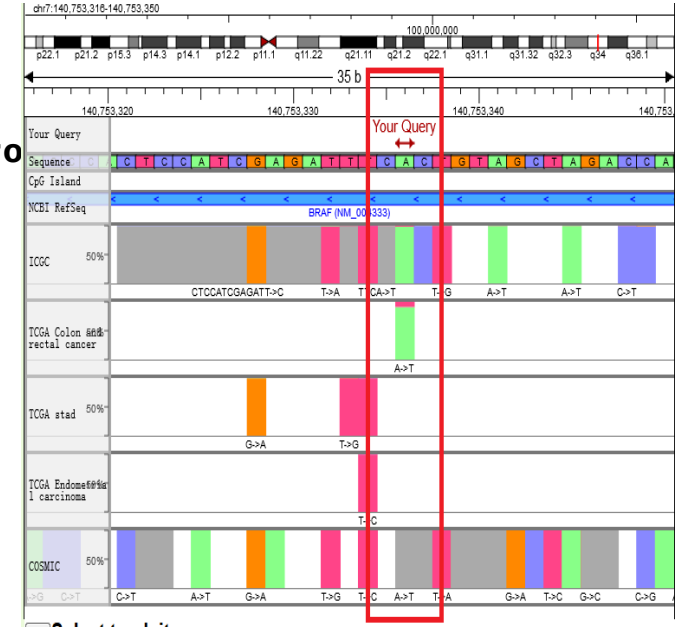
ChromHMM

ChIP-Seq

H3K4me3 H3K27me3 H3K9Ac H3K27Ac H3K4me H3K9me3 H3K36me3 Pol II

SNV

D



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 - TGCA, US | ICGC Breast Cancer - TCGA, US | ICGC Chronic Lymphocyclic 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, | 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 - AU | ICG | 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 melar | 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 | dbSNP build 137 | COSMIC | Meyerson's lung adenocarcinoma | Myelodysplasia | ICGC | Nagahama Japanese SNP

Figure 1

検索例1

SNV on promoter of BRAF

chr7:140625001, G>A

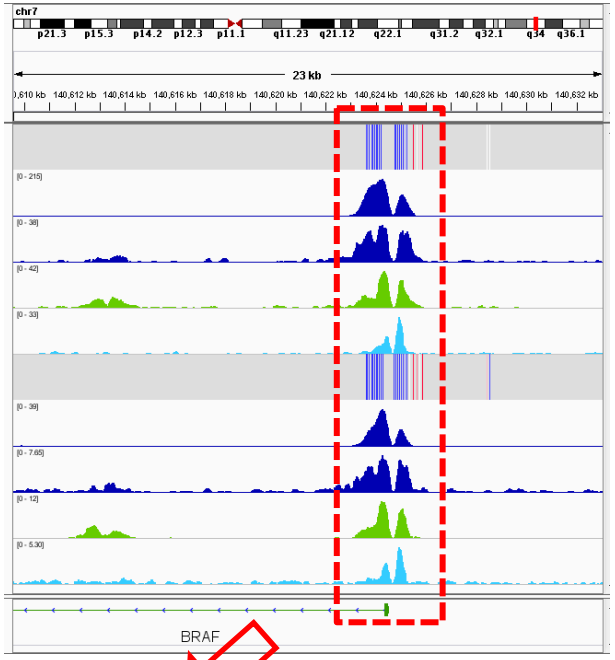
Frequency: 1/26 cell lines

＝疾患ゲノムのその座標で“何が起きているのか”を網羅的に検索

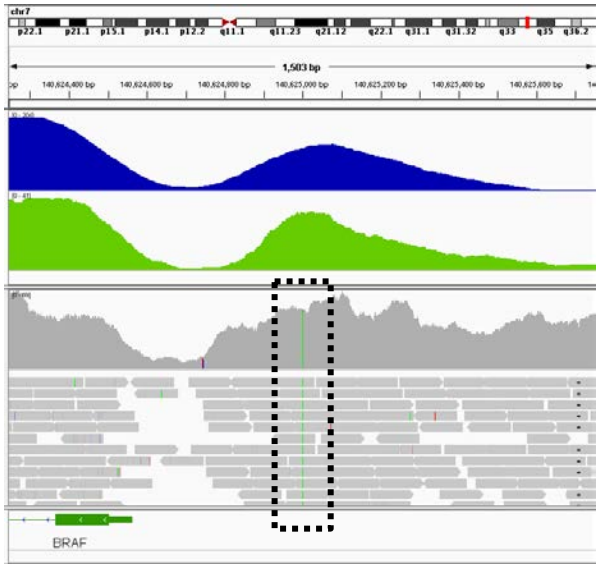
このゲノム変異はエピゲノム、トランスクリプトームに変化を与えない。

中立変異の可能性が高い？

PC-9 DNA methyl
PC-9 H3K4me3
PC-9 H3K9/14ac
PC-9 H3K27ac
PC-9 Pol II
LC2/ad DNA methyl
LC2/ad H3K4me3
LC2/ad H3K9/14ac
LC2/ad H3K27ac
LC2/ad Pol II

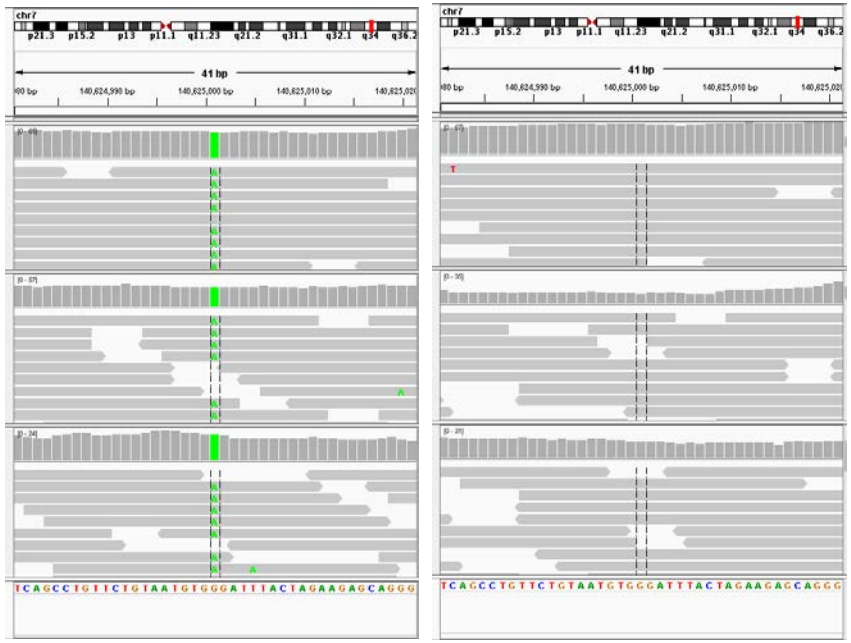


PC-9



PC-9

LC2/ad



A

TSS Viewer: BRAF

Description	Mutation frequency		
	Upstream distal: -50k to -1	Upstream proximal: -1k to -1	Gene body
Clear-cell renal cell carcinoma by Dr. Ogawa's Lab.	0/106	0/106	0/106
ICGC: Acute Myeloid Leukemia - KR	1/78	1/78	1/78
ICGC: Acute Myeloid Leukemia - TARGET, US	0/2	0/2	0/2
ICGC: Bladder Urothelial Cancer - TGCA, US	1/128	1/128	1/128
ICGC: Bladder Urothelial carcinoma - CN	1/103	1/103	1/103
ICGC: Bone Cancer - UK	0/66	0/66	0/66
ICGC: Brain Glioblastoma Multiforme - TCGA, US	5/268	5/268	5/268
ICGC: Brain Lower Grade Gliona - TCGA, US	1/268	1/268	1/268
ICGC: Breast Cancer - TCGA, US	1/943	1/943	1/943
ICGC: Breast Triple Negative/Lobular Cancer - UK	1/117	1/117	1/117
ICGC: Chronic Lymphocyclic Leukemia - ES	1/109	1/109	1/109
ICGC: Chronic Myeloid Disorders - UK	0/129	0/129	0/129
ICGC: Colon Adenocarcinoma - TCGA, US	25/216	25/216	25/216
ICGC: Early Onset Prostate Cancer - DE	1/11	1/11	1/11
ICGC: Esophageal Adenocarcinoma - UK	1/16	1/16	1/16
ICGC: Esophageal Cancer - CN	0/88	0/88	0/88
ICGC: Gastric Adenocarcinoma - TCGA, US	3/289	3/289	3/289
ICGC: Gastric Cancer - CN	0/9	0/9	0/9
ICGC: Head and Neck Thyroid Carcinoma - TCGA, US	232/393	232/393	232/393
ICGC: Kidney Renal Clear Cell Carcinoma -	1/404	1/404	1/404

Multi-omics data in cell lines

Cell	RNA-Seq (RPKM)	H3K4me3 (proximal)	H3K4me1 (distal)	H3K27ac (distal)	PolII (proximal)	H3K36me (gene body)	H3K27me3 (gene body)	H3K27me3 (distal)
A427	6.8	137.4	5.6	15.1	0	2	0.6	0.5
A549	3.4	34.1	0	6.7	16.8	1.6	0.3	0.4
ABC1	3.5	32.4	0	5.2	0	1.8	0.4	0.6
H1299	2.8	176.5	262.9	71.9	850.3	2.3	0.5	0.8
H1437	3	40.8	0	0	0	1.7	0.2	0.2
H1648	7.5	43.4	0	14.5	6.1	2.9	0.4	0.6
H1650	5.8	66.1	0	25.3	33	2.7	0.5	0.6

B

ErbB / HER Signaling

Cancer type: Lung adenocarcinoma 26 cell lines (each) | Cell: LC2/ad | Unit: RF

Tint control:light — deep

Coloring: RPKM: 0 — 255<

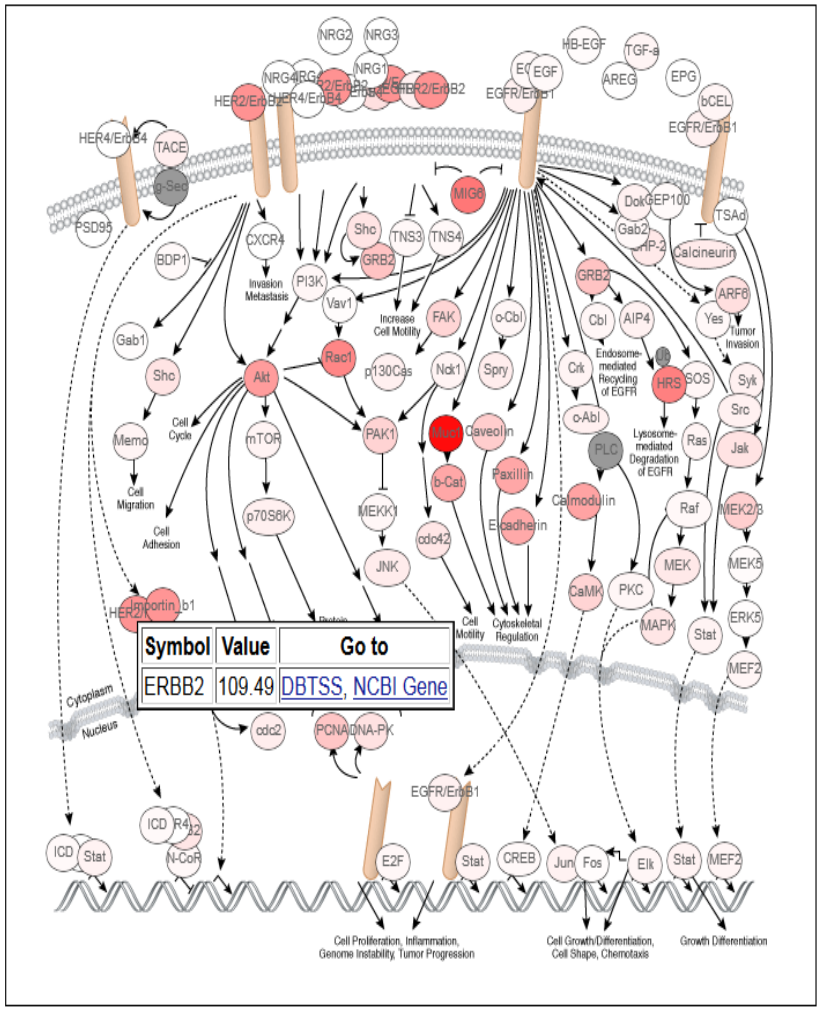


Figure 2

将来展望

データバンク機関の機能補完

データ生産者

病院

ゲノムセンター

個別研究者

提案者自身

配列データへのメタデータの体系的付加

データバンク

データ登録

データ検索

メタデータの検索

NBDC National Bioscience Database Center

DDBJ Japanese Genotype-phenotype Archive

データ登録

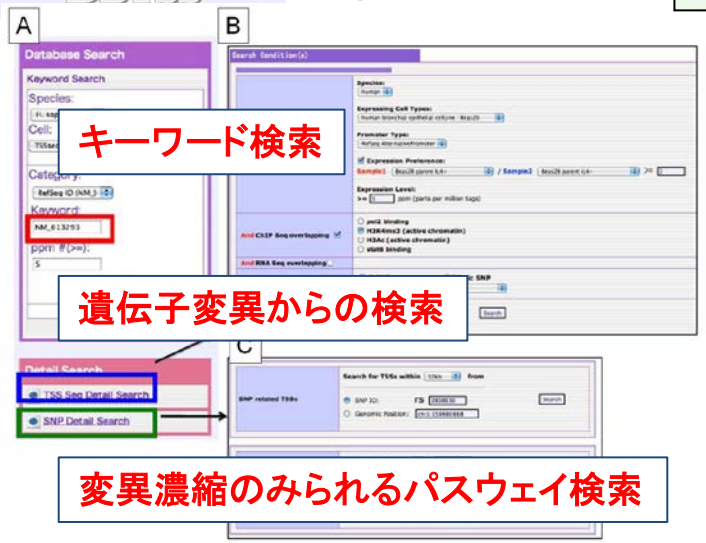
本研究で提案するDB

メタデータの付加
＋オントロジーの整備

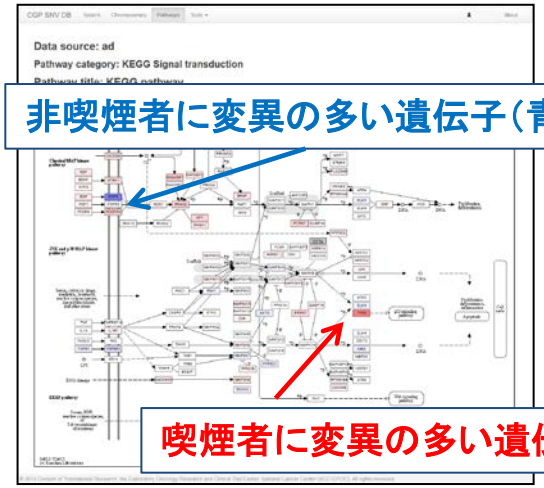
全国の研究機関
民間企業へのデータ発信

検索例（部分的には作成済み）

検索（テキスト検索）（公開DB）

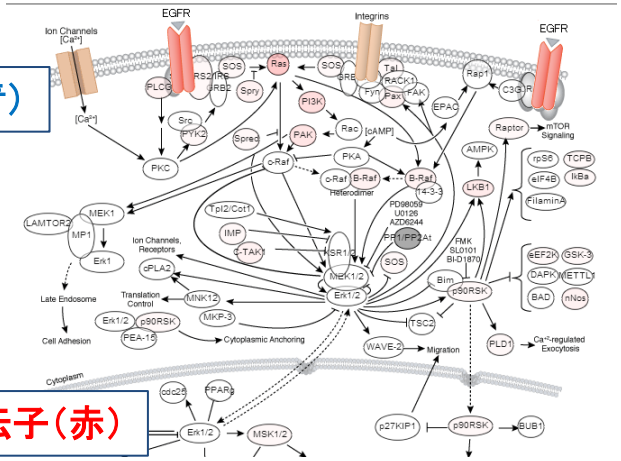


検索（クリッカブルマップ）（非公開DB）



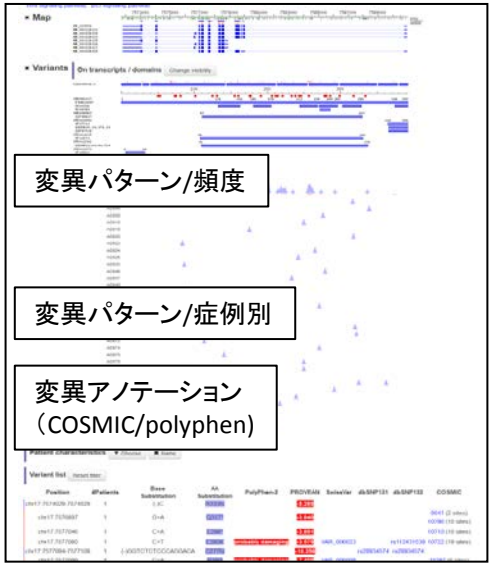
KEGGからの自動生成

（該当集団中の遺伝子変異頻度を赤の濃さで示す）

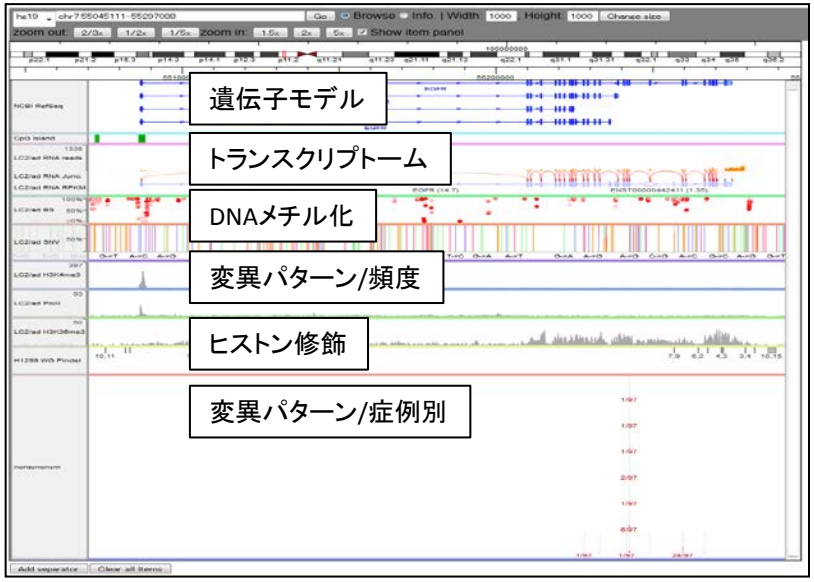


文献（ウェブ）からのマニュアル描画

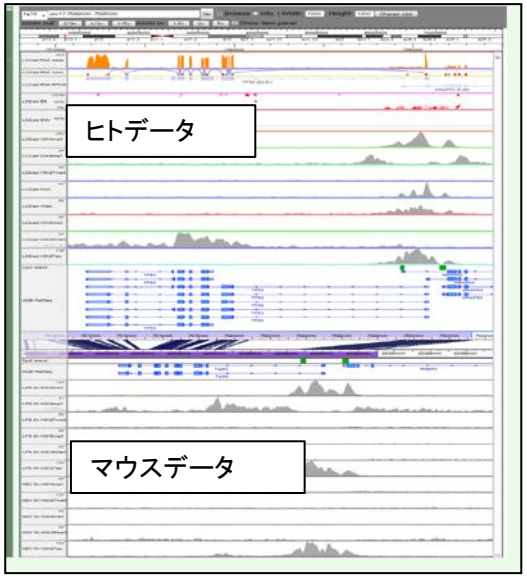
結果表示（変異情報）



結果表示（ゲノムブラウザ）



結果表示（比較ゲノム）



ヒト疾患ゲノム統合DB (DBTSSの拡張):

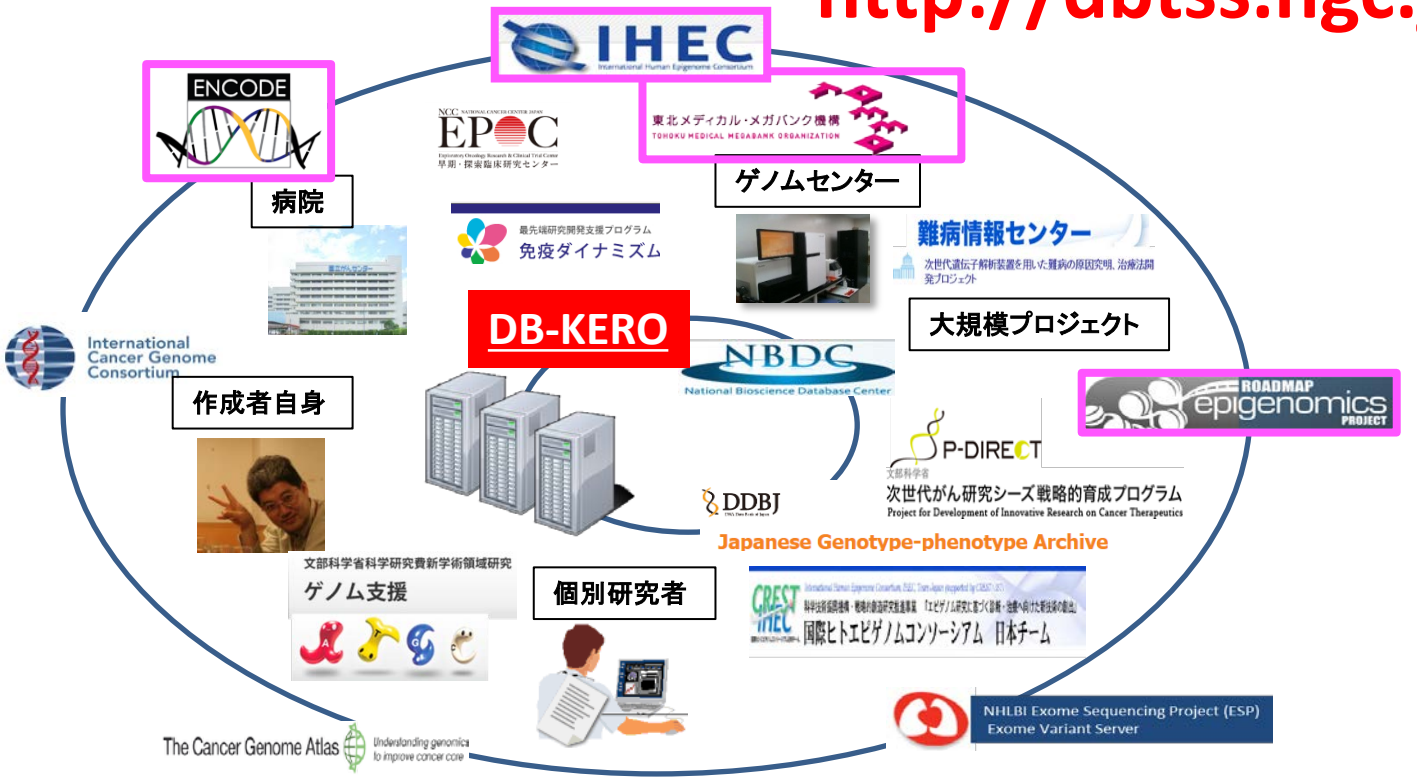
KERO(Kashiwa Encyclopedia of Regulatory Omics)

ヒトゲノム・エピゲノム・トランスクリプトームデータの統合

ヒト疾患ゲノム変異への機能的注釈

パターン検索システムの開発と実装

http://dbtss.hgc.jp/



オミクスデータ統合が加速するヒトゲノム臨床応用研究

＝疾患ゲノムのその座標で“何が起きているのか”を網羅的に検索

メンバー

菅野（東大）：研究の統括/医科研スパコンの運用

若栗、鈴木（東大）：データの加工とデータベース設計

土原（がんセンター）：ヒト疾患応用研究を志向した検索システムの実装

河野（DBCLS）：データベースの構築、オントロジーの整備