がんゲノムのデータベース化と共有

トーゴーの日シンポジウム 2012
ライフサイエンスデータベース統合の医学への応用を探る

International Cancer Genome Consortium

柴田　龍弘
がんゲノミクス研究分野
国立がん研究センター　研究所

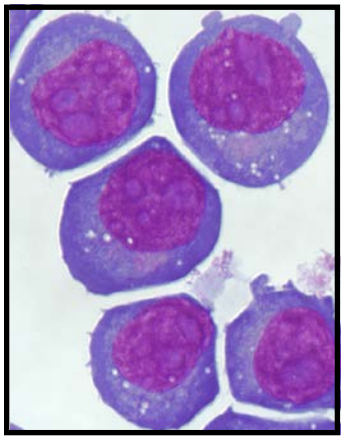
2012.10.5. @ Ginza

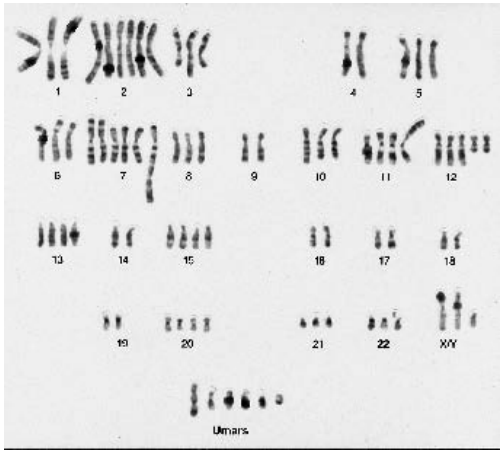
(C)2012柴田龍弘 (国立がん研究センター研究所)

Cancer

A Disease of the Genome

（正常のゲノムに異常が起こって発症する疾患）



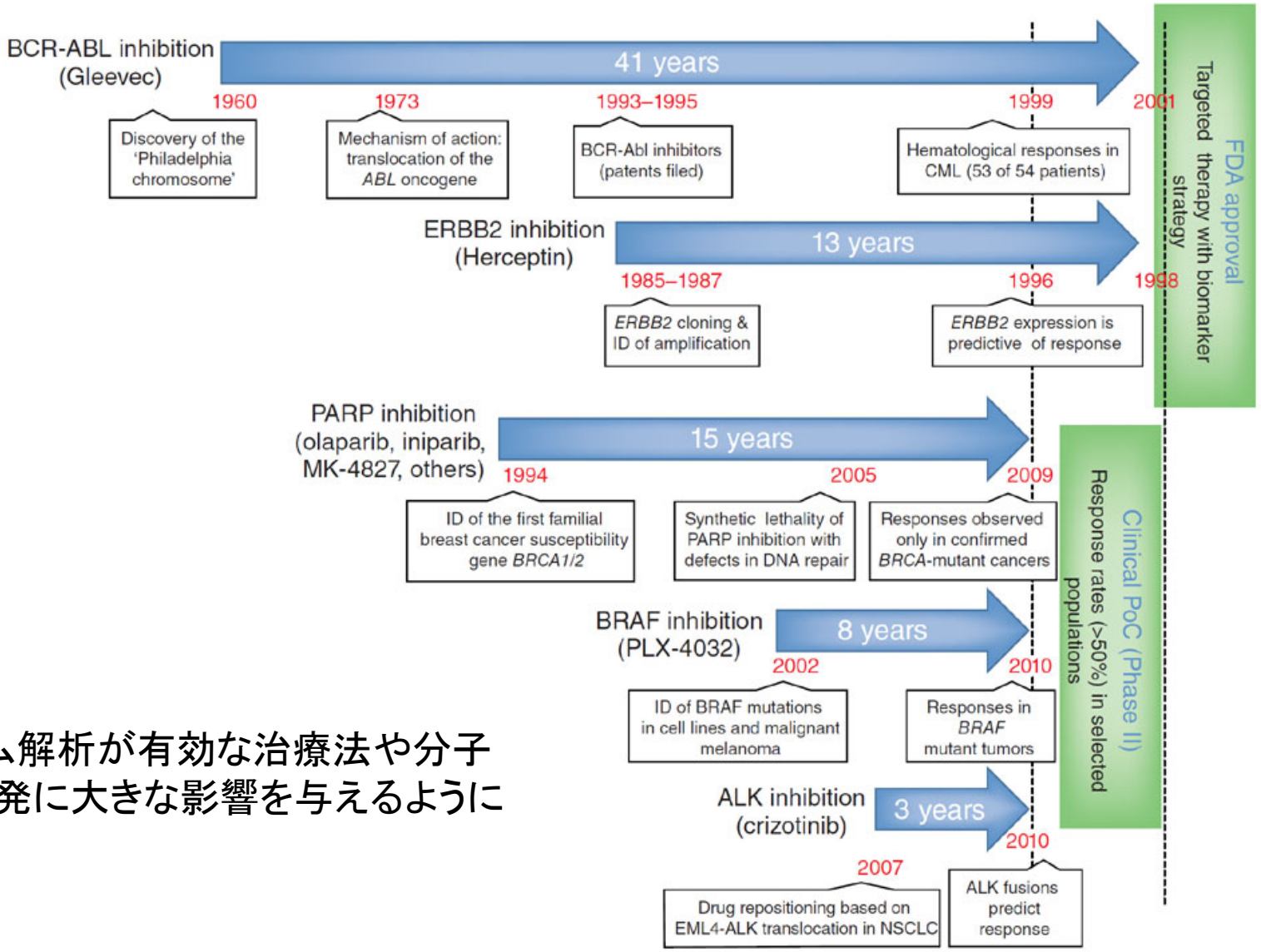


Challenge in Treating and Preventing Cancer:

- Every tumor is different
- Every cancer patient is different
- Every cancer type has unique ethnic and etiological backgrounds
- Every tumor contains heterogeneous clones

(C)2012柴田龍弘（国立がん研究センター研究所）

Cancer gene discovery and acceleration of molecular-target therapy

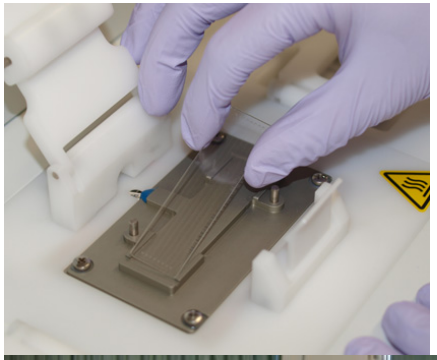


Chin et al, Nature Medicine, 2011

がんのゲノム解析が有効な治療法や分子診断法の開発に大きな影響を与えるようになった。

(C)2012柴田龍弘 (国立がん研究センター研究所)

次世代高速シークエンサーによるがんゲノム解析



次世代技術によって様々な解析が、より高精度に行なえるようになった。

- Genomic analysis: がん全ゲノム・全エクソーム解読による 突然変異・染色体構造異常・コピー数異常の全貌解明
- Transcriptome analysis: 全転写産物解読による融合遺伝子等の異常転写産物並びにスプライシング異常の同定、microRNAを含むsmall RNAの全貌解明
- Epigenetic analysis: ChIP-sequenceによるヒストン修飾やDNAメチレーションの全体像の描出

(C)2012柴田龍弘 (国立がん研究センター研究所)

Large cancer genome projects

ICGC

ICGC Goal: To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.

The Cancer Genome Atlas

Understanding genomics to improve cancer care

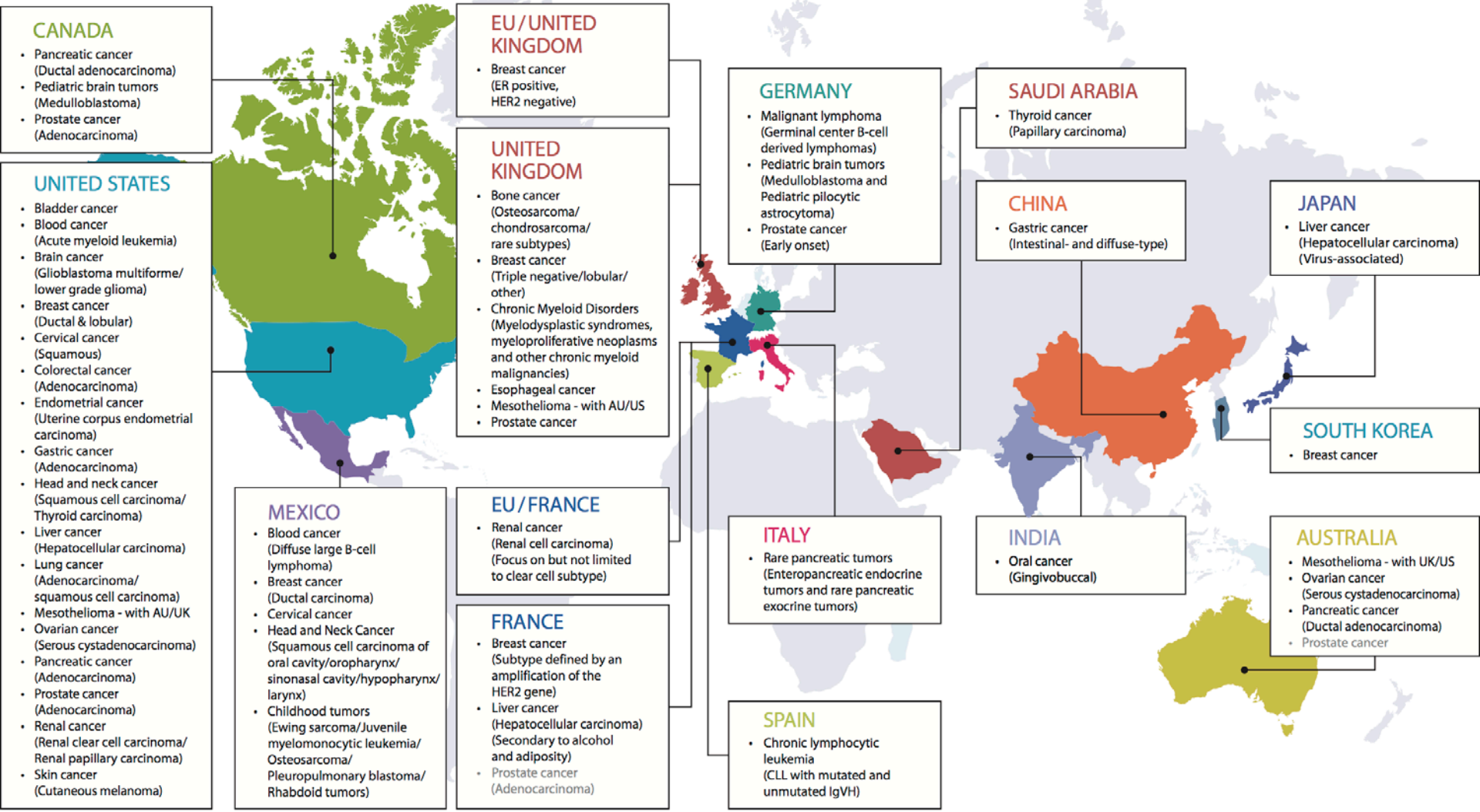
Following the success of The Cancer Genome Atlas (TCGA) Pilot Project, NIH announced in September 2009 that it is investing $275 million in TCGA over the next two years of this five-year program to chart the genomic changes involved in more than 20 types of cancer.

The most important contribution to science of these large-scale projects is the generation and transfer of resources, databases and technologies to the scientific community.

(C)2012柴田龍弘 (国立がん研究センター研究所)

Current status of ICGC

(C)2012柴田龍弘 (国立がん研究センター研究所)



Whole genome sequencing of HCC genome

nature genetics

High-resolution characterization of a hepatocellular carcinoma genome

Yasushi Totoki[1], Kenji Tatsuno[2], Shogo Yamamoto[2], Yasuhito Arai[1], Fumie Hosoda[1], Shumpei Ishikawa[3], Shuichi Tsutsumi[2], Kohtaro Sonoda[2], Hirohiko Totsuka[4], Takuya Shirakihara[1], Hiromi Sakamoto[4], Linghua Wang[2], Hidenori Ojima[5], Kazuaki Shimada[6], Tomoo Kosuge[6], Takuji Okusaka[7], Kazuto Kato[8], Jun Kusuda[9], Teruhiko Yoshida[4], Hiroyuki Aburatani[2] & Tatsuhiro Shibata[1]

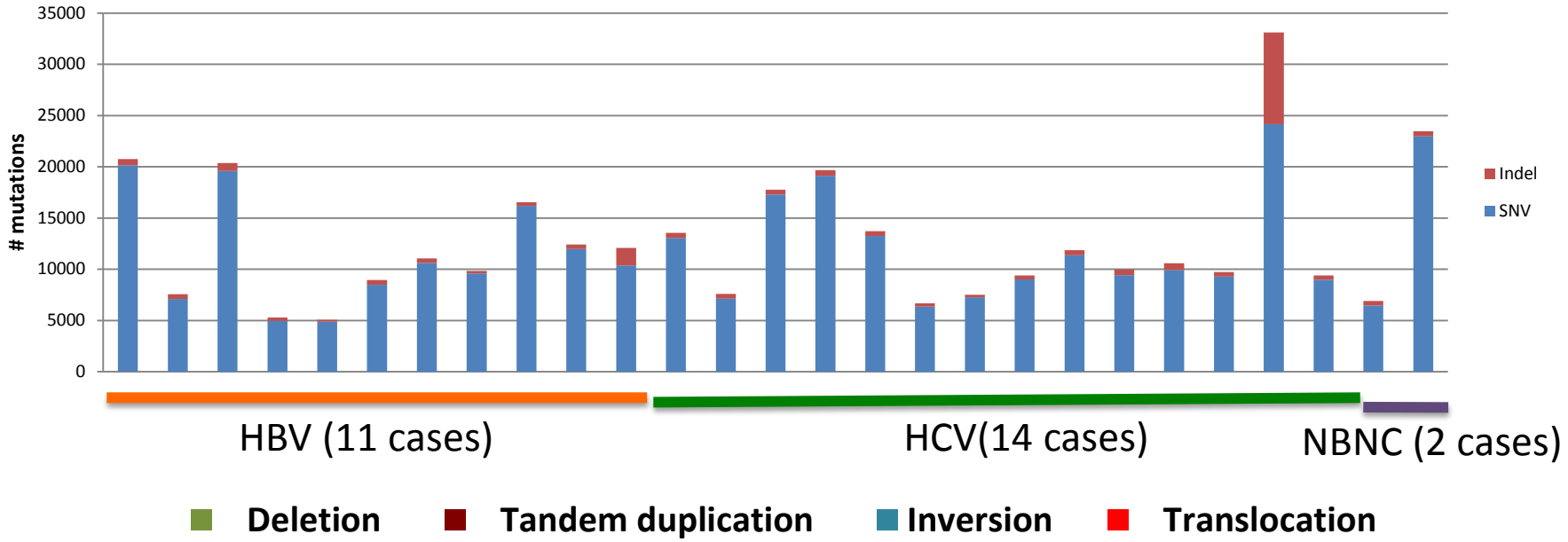
nature genetics

Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators

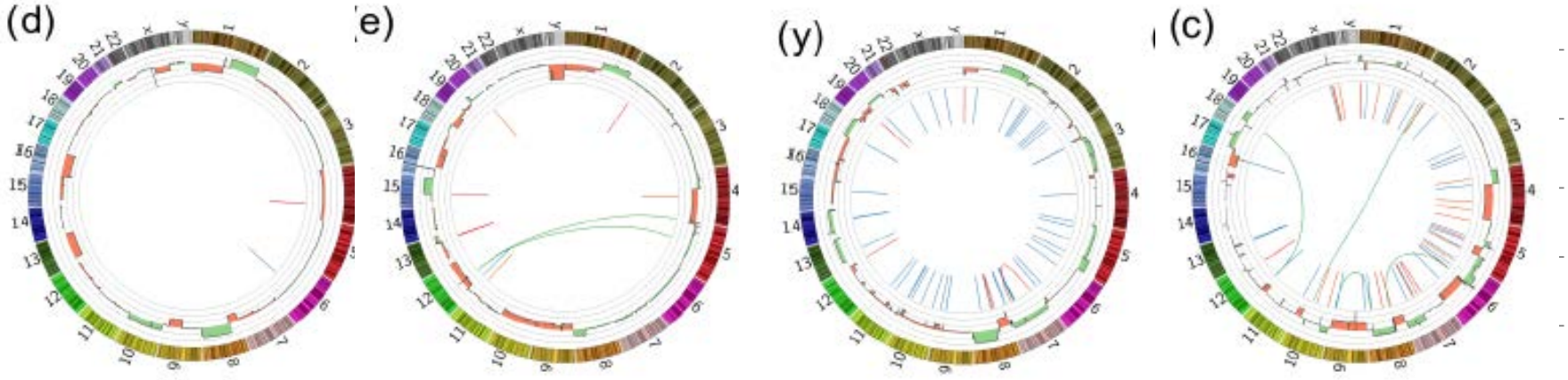
Akihiro Fujimoto[1,16], Yasushi Totoki[2,16], Tetsuo Abe[1], Keith A Boroevich[1], Fumie Hosoda[2], Ha Hai Nguyen[1], Masayuki Aoki[1], Naoya Hosono[1], Michiaki Kubo[1], Fuyuki Miya[1], Yasuhito Arai[2], Hiroyuki Takahashi[2], Takuya Shirakihara[2], Masao Nagasaki[3], Tetsuo Shibuya[3], Kaoru Nakano[1], Kumiko Watanabe-Makino[1], Hiroko Tanaka[3], Hiromi Nakamura[2], Jun Kusuda[4], Hidenori Ojima[5], Kazuaki Shimada[6], Takuji Okusaka[7], Masaki Ueno[8], Yoshinobu Shigekawa[8], Yoshiiku Kawakami[9], Koji Arihiro[10], Hideki Ohdan[11], Kunihito Gotoh[12], Osamu Ishikawa[12], Shun-ichi Ariizumi[13], Masakazu Yamamoto[13], Terumasa Yamada[12], Kazuaki Chayama[1,9], Tomoo Kosuge[6], Hiroki Yamaue[8], Naoyuki Kamatani[1], Satoru Miyano[3], Hitoshi Nakagama[5,14], Yusuke Nakamura[1,15], Tatsuhiko Tsunoda[1], Tatsuhiro Shibata[2] & Hidewaki Nakagawa[1]

(C)2012柴田龍弘 (国立がん研究センター研究所)

Whole genome sequence of 27 HCCs with distinct epidemiological backgrounds



Deletion **Tandem duplication** **Inversion** **Translocation**

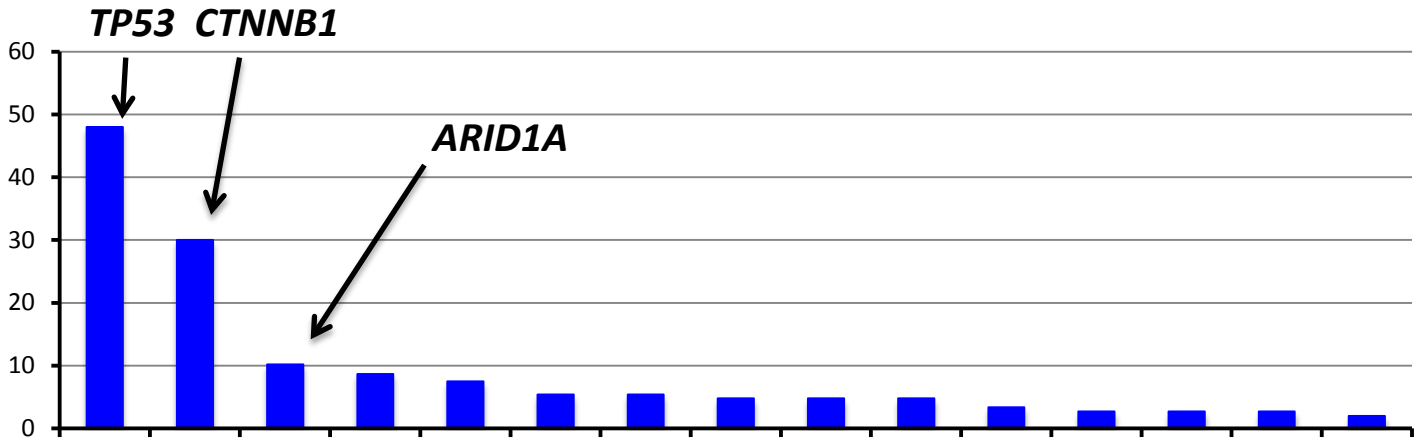


(C)2012柴田龍弘 (国立がん研究センター研究所)

Recurrently mutated genes in HCC

Compared to the background mutation rate, we identified 15 significantly mutated genes including *TP53* and *CTNNB1 (FDR<0.05)*.

We then determined the mutation frequency of these genes in 120 additional HCCs.



Pathway analysis revealed “chromatic regulator” as significant signature in this mutation set.

Category	Term	Count	%	*P-value*	List Total	Pop Hits	Pop Total	Fold Enrichment	*q-value*
SP_PIR_KEYWORDS	phosphoprotein	162	55.1	6.92E-10	292	7263	19235	1.469292662	0.00000023
INTERPRO	IPR013032:EGF-like region, conserved site	19	6.5	8.02E-07	260	293	16659	4.154909425	0.00049
INTERPRO	IPR000742:EGF-like, type 3	15	5.1	2.14E-06	260	194	16659	4.954103886	0.0006
INTERPRO	IPR006210:EGF-like	15	5.1	3.25E-06	260	201	16659	4.781572905	0.0007
SP_PIR_KEYWORDS	egf-like domain	15	5.1	1.15E-05	292	230	19235	4.296083979	0.0019
UP_SEQ_FEATURE	domain:EGF-like 1	12	4.1	2.25E-06	292	120	19113	6.545547945	0.0039
SMART	SM00181:EGF	15	5.1	3.00E-05	175	201	9079	3.871641791	0.0043
SP_PIR_KEYWORDS	polymorphism	208	70.7	4.75E-05	292	11550	19235	1.18628951	0.0052
SP_PIR_KEYWORDS	calcium	28	9.5	9.42E-05	292	803	19235	2.2969515	0.0078
UP_SEQ_FEATURE	sequence variant	218	74.1	1.06E-05	292	11992	19113	1.189901144	0.0092
SP_PIR_KEYWORDS	**bromodomain**	6	2.0	2.89E-04	292	39	19235	10.13435195	0.019
SP_PIR_KEYWORDS	**chromatin regulator**	12	4.1	4.17E-04	292	213	19235	3.711171136	0.023
INTERPRO	IPR006209:EGF	10	3.4	1.66E-04	260	127	16659	5.045124167	0.025
SP_PIR_KEYWORDS	disease mutation	42	14.3	5.45E-04	292	1591	19235	1.738955426	0.026
INTERPRO	**IPR001487:Bromodomain**	6	2.0	3.69E-04	260	40	16659	9.610961538	0.044
SP_PIR_KEYWORDS	tumor suppressor	9	3.1	0.001157126	292	137	19235	4.327442256	0.047

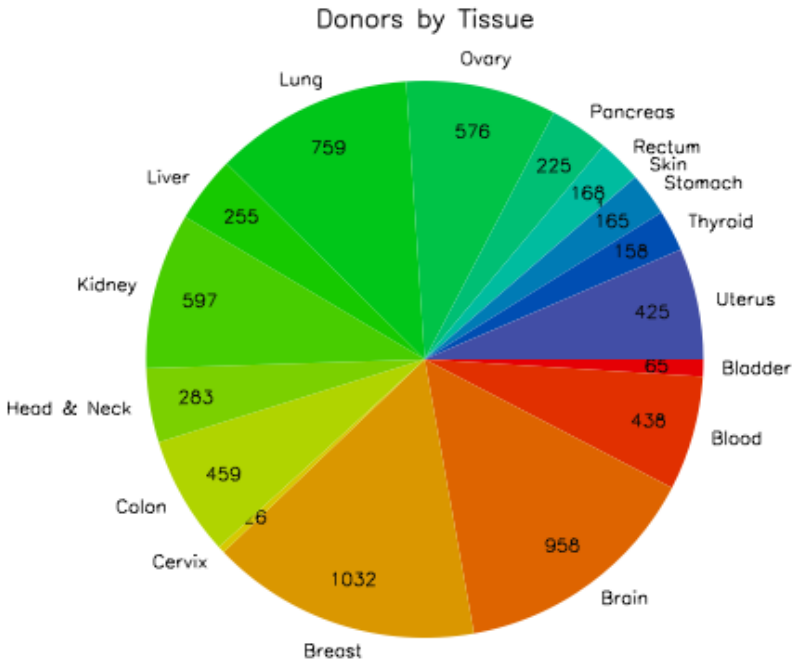
q-value was obtained by Benjamini and Hochberg’s FDR method

(C)2012柴田龍弘 (国立がん研究センター研究所)

ICGC HP Data portal

ICGC Dataset Version 9 (August 28th, 2012)

Cancer Projects: **36**



Total Donors: **6,590**

Gene Search

Go

Examples: TP53, ENSG00000133703, NM_000314

Database Search

Quick | **Advanced**

- Genes
- Samples
- Simple Mutations
- Copy Number Alterations
- Structural Rearrangements
- Gene Expression
- Methylation
- miRNA
- Exon Junction

Data Summaries

Genes | **Pathway**

- Affected Pathways - KEGG
- Affected Pathways - Reactome

(C)2012柴田龍弘 (国立がん研究センター研究所)

ICGC Data Releases Policy

ICGC Open Access Datasets	ICGC Controlled Access Datasets
➢ Cancer Pathology Histologic type or subtype Histologic nuclear grade ➢ Patient/Person Gender Age range ➢ Gene Expression (normalized) ➢ DNA methylation ➢ Genotype frequencies ➢ Computed Copy Number and Loss of Heterozygosity ➢ Newly discovered somatic variants	➢ Detailed Phenotype and Outcome Data Patient demography Risk factors Examination Surgery/Drugs/Radiation Sample/Slide Specific histological features Protocol Analyte/Aliquot ➢ Gene Expression (probe-level data) ➢ Raw genotype calls ➢ Gene-sample identifier links ➢ Genome sequence files

(C)2012柴田龍弘 (国立がん研究センター研究所)

ICGC Intellectual Property Policy（抜粋）

- All ICGC members agree not to make claims to possible IP derived from primary data (including somatic mutations) and to not pursue IP protections that would prevent or block access to or use of any element of ICGC data or conclusions drawn directly from those data.

ICGC Publication Policy（抜粋）

Investigators outside of the ICGC are free to use data generated by ICGC members, either en masse or specific subsets, but are asked to follow the guidelines developed at the Ft. Lauderdale meeting.

All data shall become free of a publication moratorium when either the data is published by the ICGC member project or one year after a specified quantity of data (e.g. genome dataset from 100 tumours per project) has been released via the ICGC database or other public databases. In all cases data shall be free of a publication moratorium two years after its initial release.

Data generatorの権利を担保しながら、できるだけ迅速に (prepublication dataを主体として) 広くデータが他の研究者にも活用できるようなルールを作る。

(C)2012柴田龍弘 (国立がん研究センター研究所)

がんゲノムデータベースの例

IARC TP53 Database

COSMIC (Catalog Of Somatic Mutations in Cancer)

ICGC (International Cancer Genome Consortium)

TCGA (The Cancer Genome Atlas)

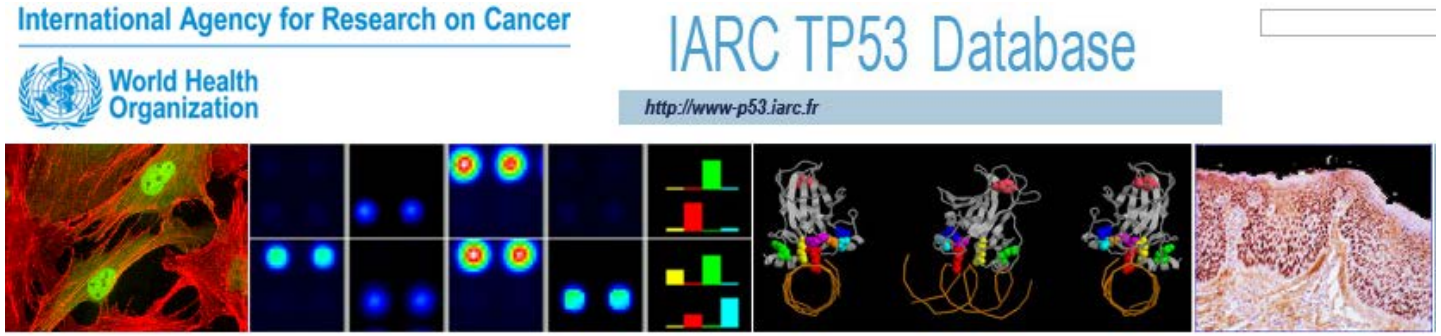
CCLE (Cancer Cell Line Encyclopedia)

GemDBJ (Genome Medicine Data Base of Japan)

(C)2012柴田龍弘 (国立がん研究センター研究所)

がんゲノムデータベース

IARC　TP53　Database



IARC (International Agency for Research on Cancer)が運営するTP53変異に関する世界最大のデータベース。変異データに加えて、機能的な評価や細胞株・マウスモデルに関するデータも収納する。

The IARC TP53 Database compiles TP53 gene variations identified in human populations and tumor samples. Data are compiled from the peer-reviewed literature and from generalist databases.　The following datasets are available:

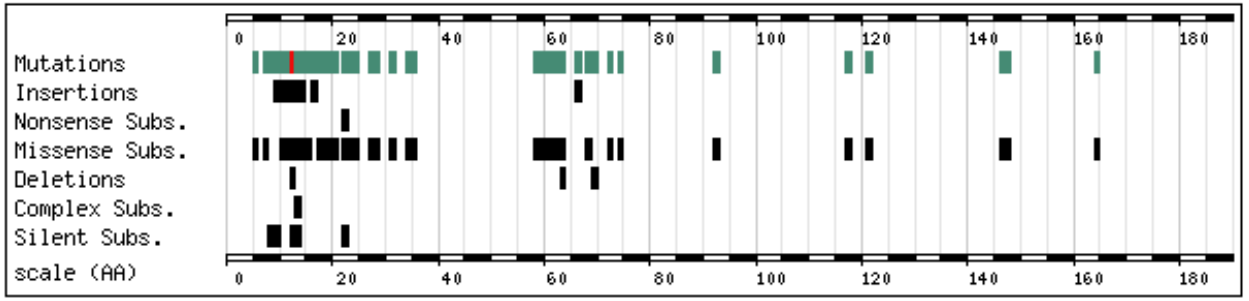
- TP53 **somatic mutations** in sporadic cancers
- TP53 **germline mutation** in familial cancers
- Common TP53 **polymorphisms** identified in human populations
- **Functional and structural properties** of p53 mutant proteins
- TP53 gene status in human **cell-lines**
- **Mouse-models** with engineered TP53

(C)2012柴田龍弘 (国立がん研究センター研究所)

がんゲノムデータベース

COSMIC (Catalog Of Somatic Mutations in Cancer)

英国サンガーセンターが運営する世界最大のがんにおける体細胞変異のデータベース。サンガーセンターの解析結果のみならず文献データや他のデータベース(ICGC, TCGA)の結果も収納している。

Gene Name	**KRAS** (HGNC Symbol) Synonyms: C-K-RAS, K-RAS2A, K-RAS2B, K-RAS4A, K-RAS4B, KI-RAS, KRAS1, KRAS2, NS3, RASK2,
Small Intragenic Mutation Summary	 Histogram - Click for a histogram of the full gene sequence and mutation details.

(C)2012柴田龍弘 (国立がん研究センター研究所)

The Cancer Genome Atlas

Understanding genomics to improve cancer care

Available Cancer Types	# Patients with	# Downloadable Tumor Samples	Date Last Updated

Data Browser

The output of this application reflects DCC data as of October 2010. Efforts are currently underway to update this site.

Help

Genes | Participants | Pathways

Search Criteria

"Use same search criteria on all tabs"

Disease Type

GBM – Glioblastoma multiforme

Genes

All Genes
Chromosome Region Add
Gene List

Clear

Participants

All Participants
Participants List

Clear

Copy Number - Genes

– Select to Add –

Genome_Wide_SNP_6 log2 ratio

<= -0.5 or >= 0.5
Frequency >= 20 % Avg. Across Patients

Copy Number - miRNAs

– Select to Add –

Gene Expression

– Select to Add –

miRNA Expression

– Select to Add –

DNA Methylation

– Select to Add –

HumanMethylation27

>= 0.5
Frequency >= 40 %

Validated Somatic Mutations

– Select to Add –

Any Non-Silent

Frequency >= 1 %

Correlations

– Select to Add –

Stomach adenocarcinoma [STAD]	255	162	09/23/12
Thyroid carcinoma [THCA]	412	357	09/28/12

(C)2012柴田龍弘 (国立がん研究センター研究所)

がんゲノムデータベース

CCLE (Cancer Cell Line Encyclopedia)

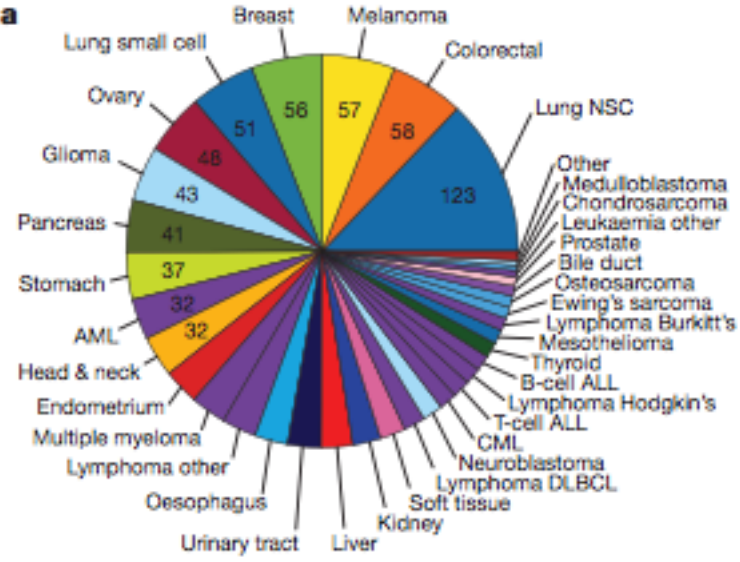
米国Broad研究所、ノバルティス生物医学研究所 (NBR)、ノバルティスゲノム研究所、が合同で進めているプロジェクトのデータを公開。

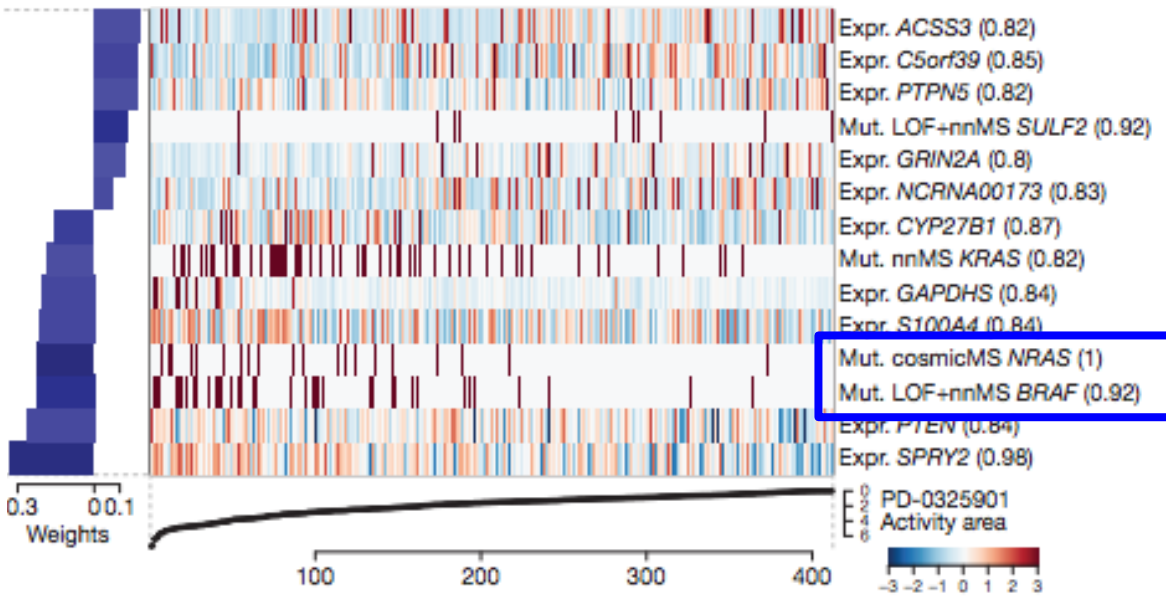
約1,000種類のがん細胞株に関する、コピー数 (SNP array), 遺伝子発現プロファイル、がん関連遺伝子変異、薬剤感受性データを公開している。

LETTER

doi:10.1038/nature11003

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





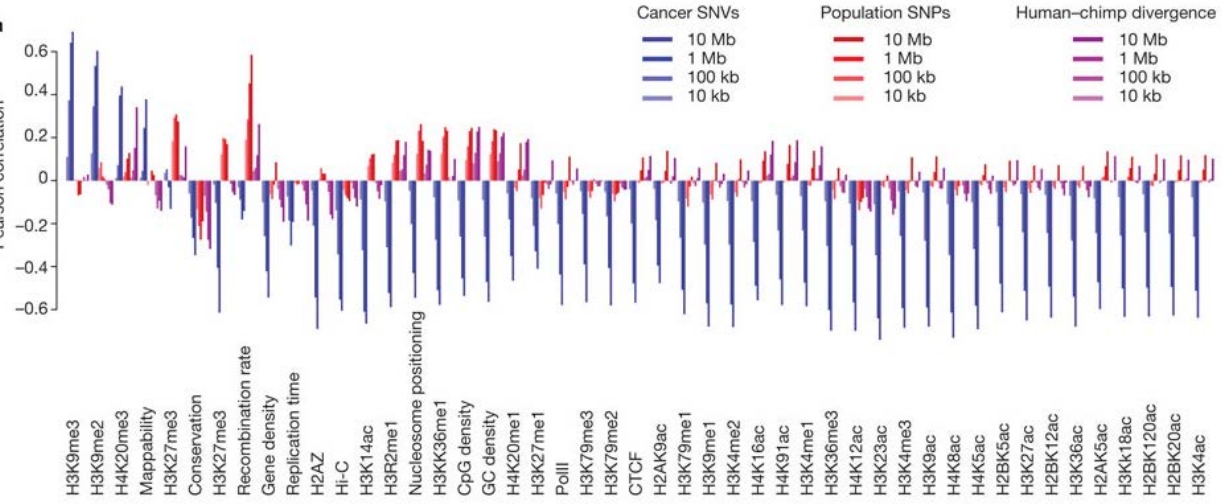
MEK阻害剤の奏功性と相関するゲノムマーカーの探索

(C)2012柴田龍弘 (国立がん研究センター研究所)

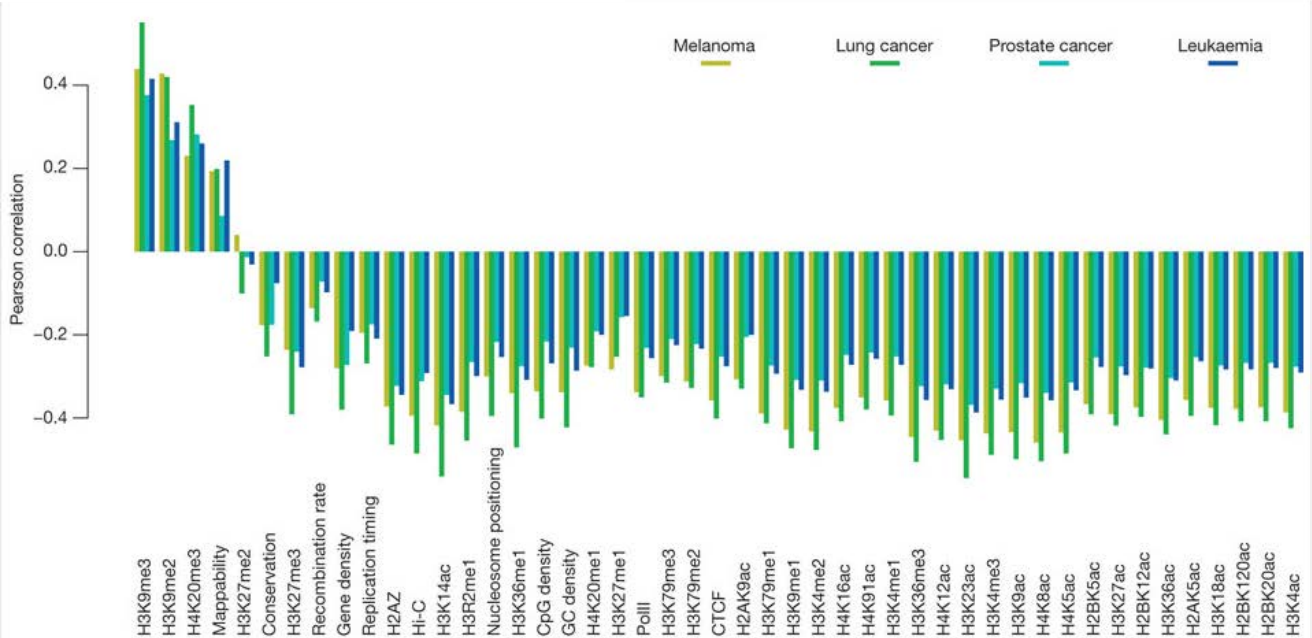
LETTER

doi:10.1038/nature11273

Chromatin organization is a major influence on regional mutation rates in human cancer cells

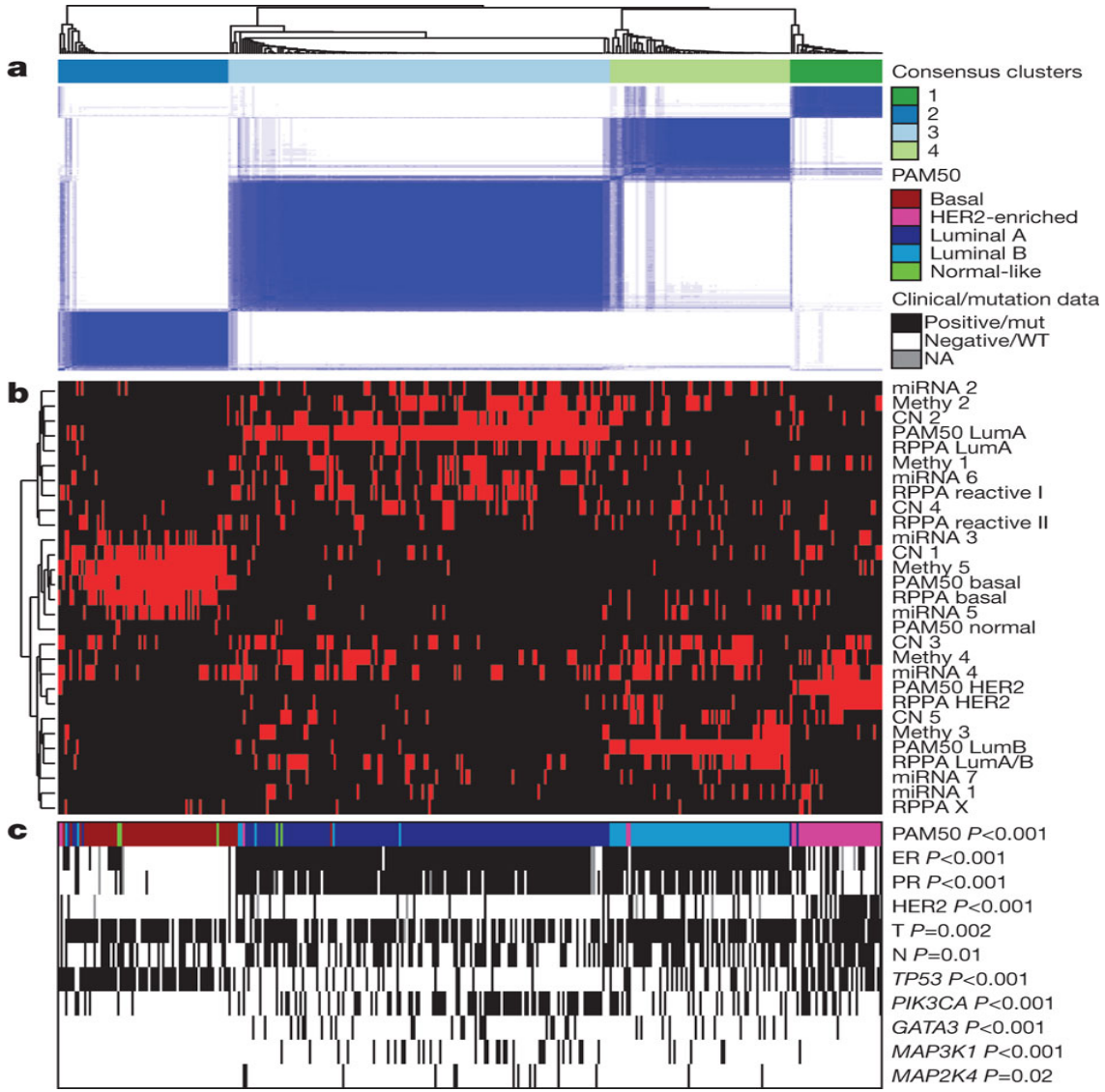


がんゲノム全解読データを用いて、体細胞変異率とクロマチンの状態との関連について検討した。



(C)2012柴田龍弘 (国立がん研究センター研究所)

データの活用 Integration of cancer genome data

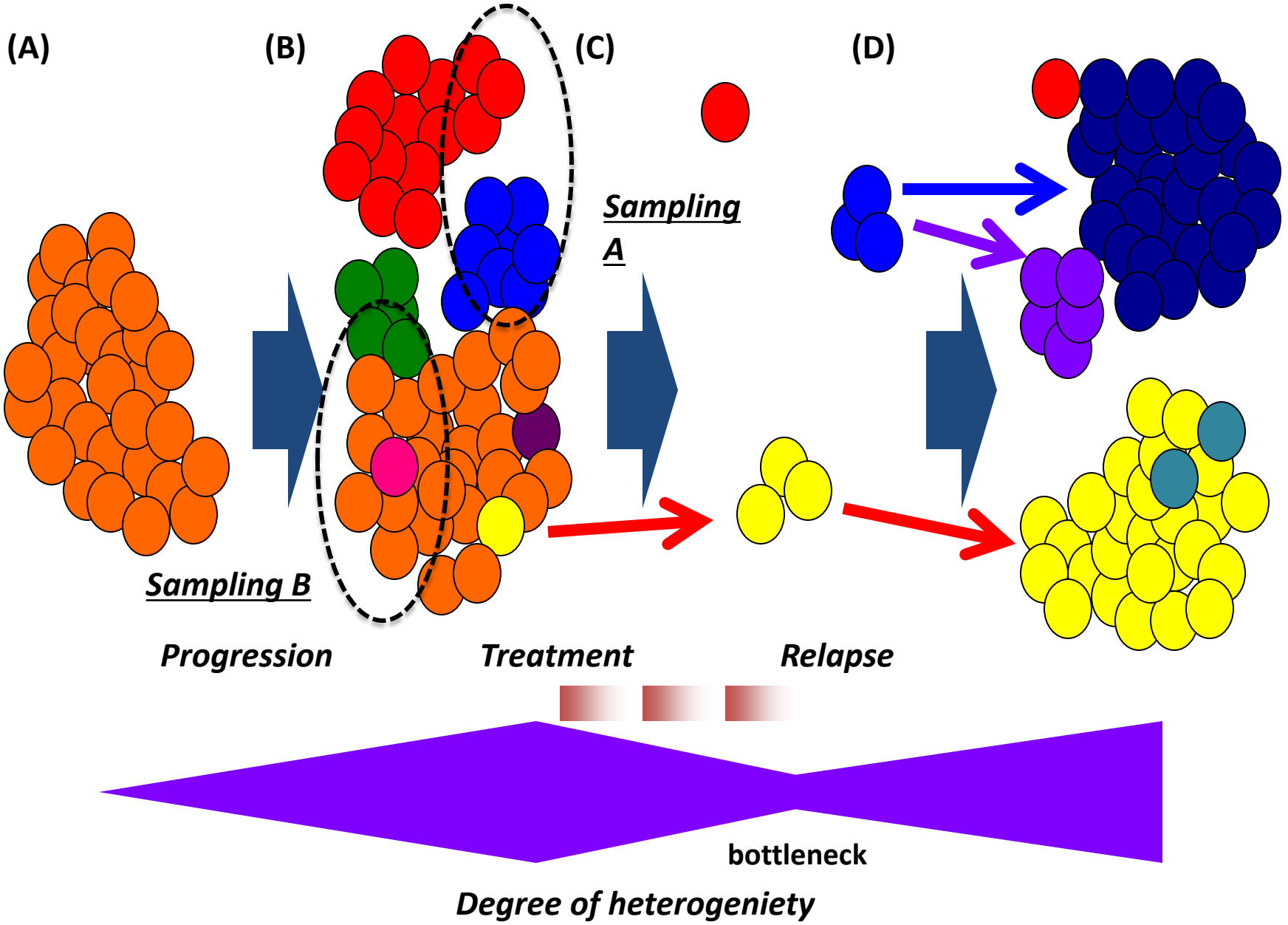


Coordinated analysis of breast cancer subtypes defined from five different genomic/proteomic platforms.

DC Koboldt *et al. Nature* **000**, 1-10 (2012) doi:10.1038/nature11412

(C)2012柴田龍弘 (国立がん研究センター研究所)

問題点2　Cancer genome complexity (dynamic heterogeniety)

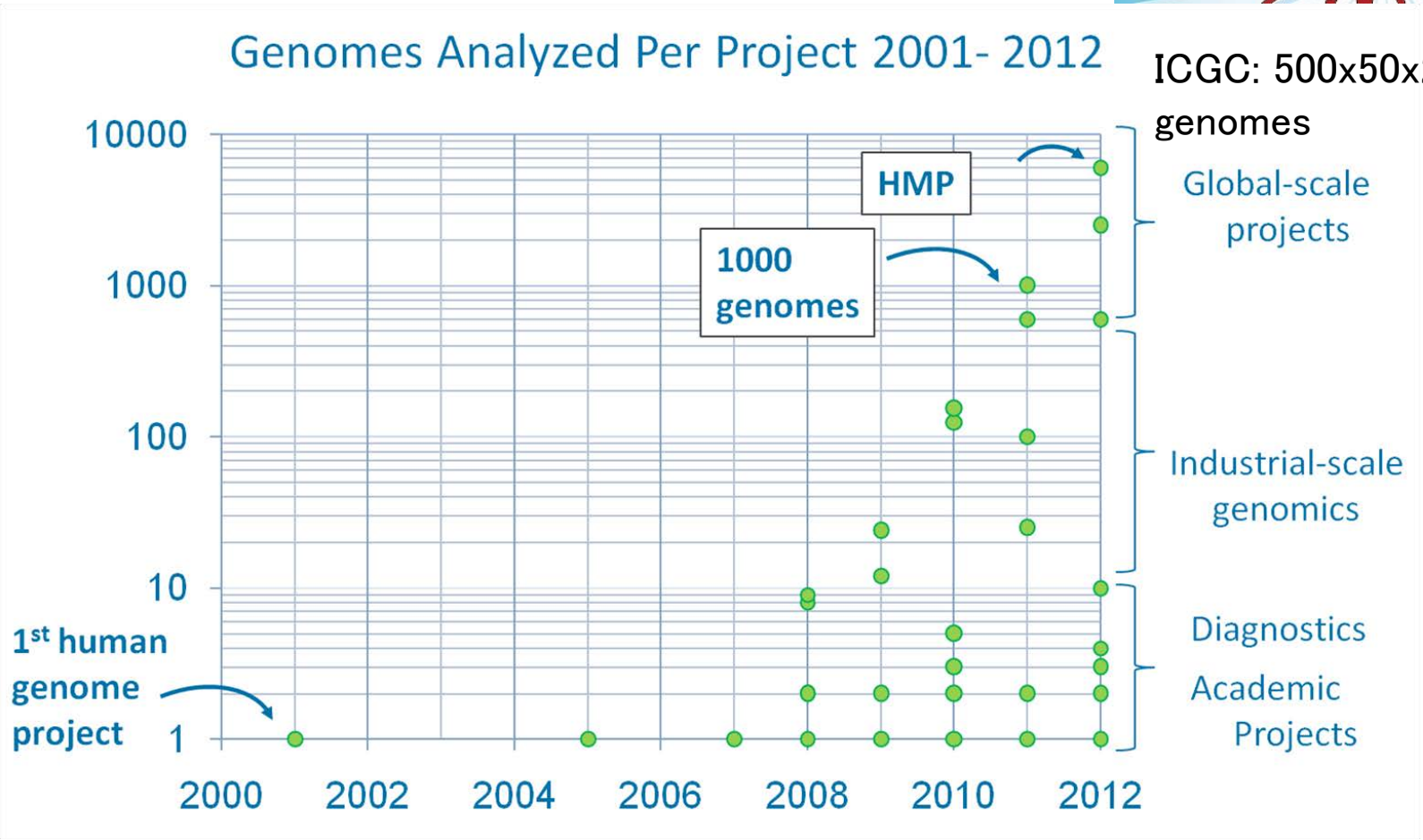


(C)2012柴田龍弘 (国立がん研究センター研究所)

問題点３　データ量の爆発

■ 安全性を担保した大量データの保存、管理
■ 複雑で大量の統合データの解析法、より大規模な計算機資源

NSF/NIH started BIG DATA program.



ICGC: 500x50x2=50,000 genomes

(C)2012柴田龍弘 (国立がん研究センター研究所)

Conclusions

- 高速シークエンス技術の革新により、多様ながんゲノムデータが大量に産出されるようになった。こうしたBig data自体、あるいはそれらをうまく統合することによって、がん研究やがんの治療・診断は大きく進むことが期待されている。また大型がんゲノムプロジェクトでは、当初からデータを如何に迅速に、使いやすく、一定の水準を担保した形で公開するのかについて、議論が進められている。

- がんゲノムデータベースの充実によって、Data 産出者とも協調しながら臨床情報も含めた大型データを活用する新たな情報解析研究分野の発展が今後期待される。同時に、臨床の現場においてはがんゲノムデータベースが個別化医療の実現に向けた強力な推進力となりつつある。

- がんゲノムデータベースの統合と共有化を考えていく上で、高速シークエンス技術やデータ解析技術の精度、対象となるがんの複雑性、加速的に増加するデータ量の保存、統合解析の方法論、等といった課題がまだ残されている。

(C)2012柴田龍弘 (国立がん研究センター研究所)

Acknowledgements

Cancer patients who agreed to use their samples for our research

NCC Research Institute

Yasushi Totoki, Fumie Hosoda, Yasuhito Arai, Takuya Shirakihara, Tomoko Urushidate, Shouko Ohashi, Tatsuhiro Shibata

Takashi Kohno, Hitoshi Ichikawa

Hiromi Sakamoto, Teruhiko Yoshida

Hitoshi Nakagama

NCC Central Hospital

Hidenori Ojima, Kazuaki Shimada, Takuji Okusaka, Tomoo Kosuge

Koji Tsuta, Shunichi Watanabe

Research Center for Advanced Science and Technology, U. of Tokyo

Hiroyuki Aburatani, Kenji Tatsuno, Shogo Yamamoto

Kyoto University

Kazuto Kato

National Institute of Biomedical Innovation

Jun Kusuda, Yoshihiko Sano, Sachiko Suematsu, Hideo Eno

RIKEN CGM

Hidewaki Nakagawa, Akihiro Fujimoto, Tatsuhiko Tsunoda, Kaoru Nakano, Kumiko Makino, Tetsuo Abe, Keith Boroevich, Michiaki Kubo, Noayuki Kamatani

Human Genome Center, Institute of Medical Science, U. of Tokyo

Masao Nagasaki, Tetsuro Shibuya, Yoko Tanaka, Rui Yamaguchi, Atsuji Niida, Satoru Miyano, Yusuke Nakamura

Wakayama Medical University

Masaki Ueno, Yuki Yamaue

Osaka Medical Center for Cancer & Cardiovascular Diseases

Akimasa Yamada, Osamu Ishikawa

Tokyo Women's Medical University

Syunichi Ariizumi, Masakazu Yamamoto

Hiroshima University

Yoshiiku Kawakami, Kouji Arihira Kazuaki Chayama

(C)2012柴田龍弘 (国立がん研究センター研究所)