



新生命科学分野開拓とスーパーコンピュータ「京」

2013/9/19 九州大学医学部百年講堂

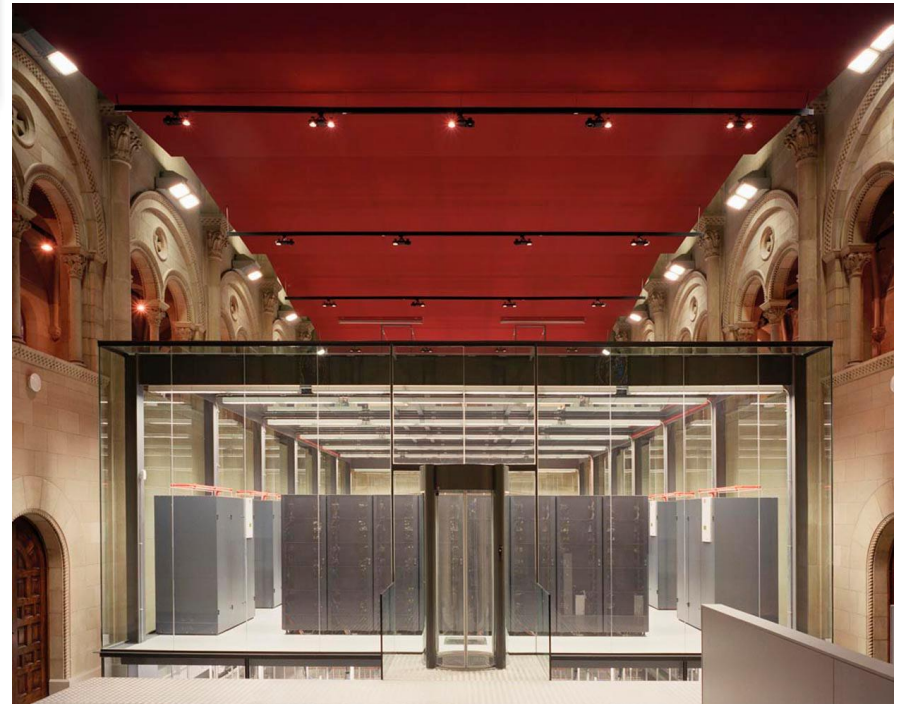
大規模エピゲノムプロジェクトと データ解析

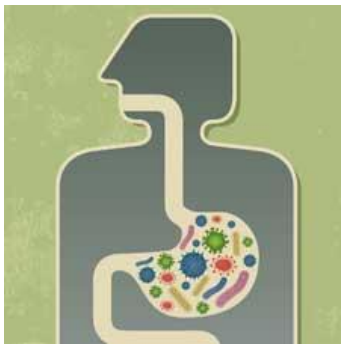
須山幹太

九州大学 生体防御医学研究所 情報生物学分野



BSC (Barcelona Supercomputing Center)





(Picture taken from *Sci. Am.*)

ゲノミクス：我々がもつもう1つ別のゲノムの話

Liping Zhao

Nature 465, 879–880 (17 June 2010) | doi:10.1038/465879a

Published online 16 June 2010

ヒトに棲み着いているすべての微生物の集合ゲノムの解読という、ヒトのマイクロバイオーム解析の基盤となる研究が終了した。この研究は、ヒトの健康と病態の両方を解明するうえで重要である。



nature International weekly journal of science

Search

▶▶ Take Nature Publishing Group's readership survey for the chance to win a MacBook Air.

Journal home > Archive > Article > Full Text

Journal content

- [Journal home](#)
- [Advance online publication](#)
- [Current issue](#)
- [Nature News](#)

Article

Nature 464, 59–65 (4 March 2010) | doi:10.1038/nature08821; Received 14 August 2009; Accepted 23 December 2009

A human gut microbial gene catalogue established by metagenomic sequencing

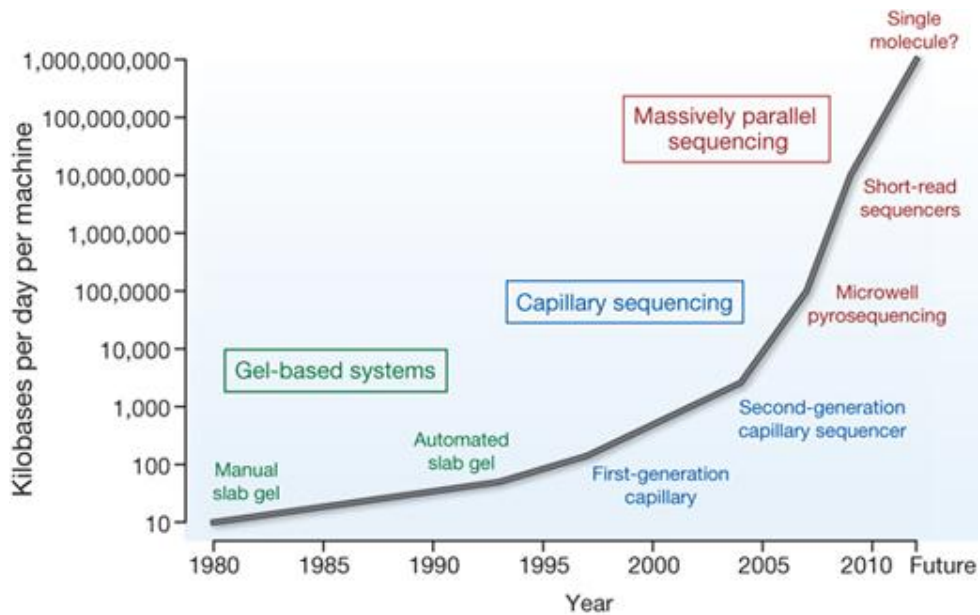
Qin *et al.*

腸内微生物がヒトの健康に与える影響を理解するには、その遺伝的潜在能力を評価することが極めて重要である。今回我々は、ヨーロッパ人124名の便検体から得た576.7ギガ塩基の配列に由来する重複しない微生物遺伝子330万個に関して、イルミナ社ゲノムシーケンスシステムを使ってメタゲノムの配列決定、組み立て、および特性解析を行った。ヒトの全遺伝子の約150倍に相当するこの遺伝子セットは、このコホートの一般的な（より頻度の多い）微生物遺伝子の圧倒的多数を含んでおり、一般的なヒトの腸内微生物遺伝子の大部分はこれに含まれると考えられる。この遺伝子群は、コホートの各個人にほぼ共通であった。その99%以上が細菌由来であったことから、コホート全体では一般的な細菌が1,000~1,150種存在しており、各個人には最低でもそのうちの160種が存在していて、それらもほぼ共通していることが示された。我々は、全個人およびほとんどの細菌にみられる機能の観点から、それぞれ最小限の腸内メタゲノムと最小限の腸内細菌ゲノムを定義し、記載する。

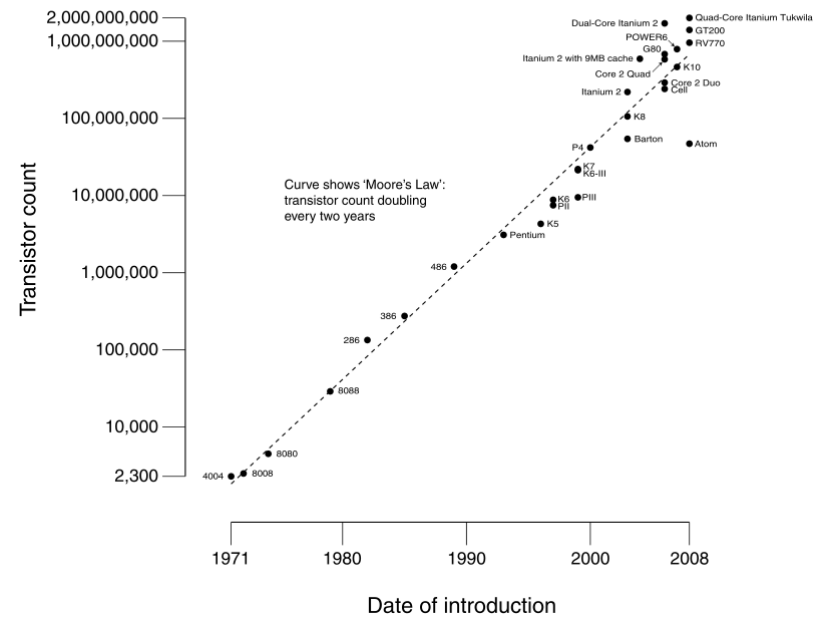
▶ Top

シーケンス技術の進歩

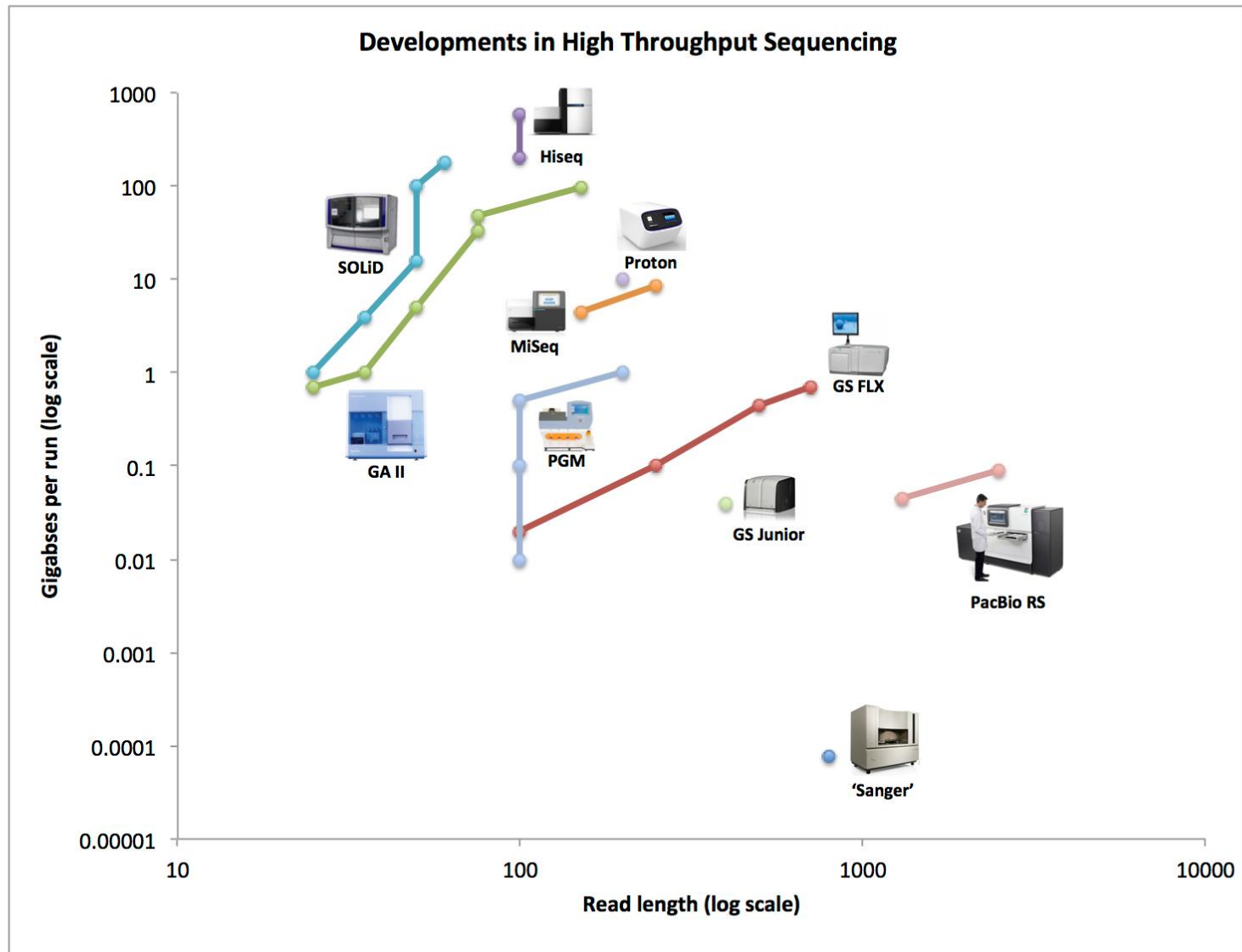
(+コンピュータの性能の進歩)



CPU Transistor Counts 1971-2008 & Moore's Law



ハイスループット・シーケンシング技術の現況



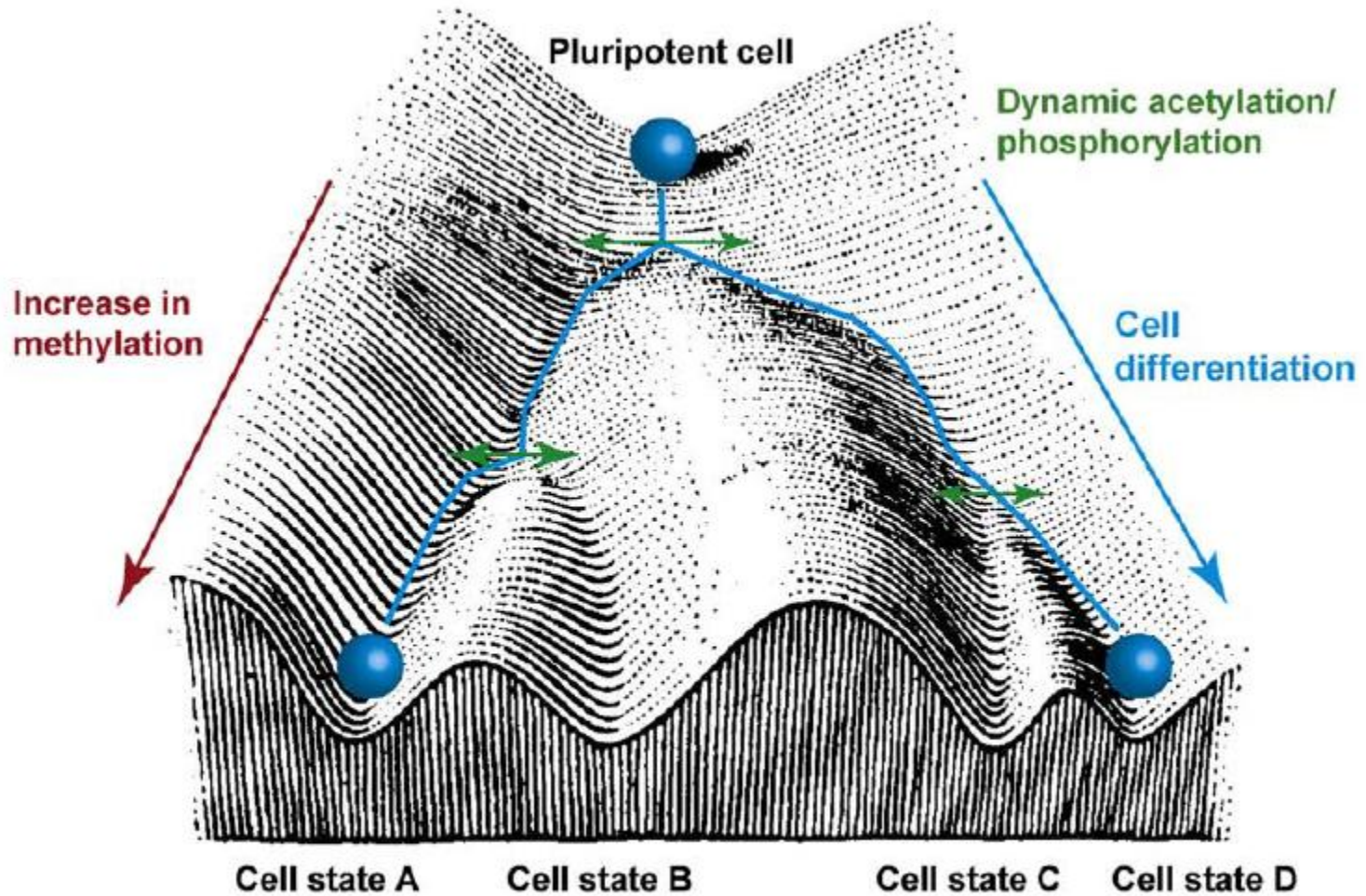
Illumina HiSeqのパフォーマンス

HiSeq System Performance Parameters

Read Length	HIGH OUTPUT RUN MODE*			RAPID RUN MODE*		
	Dual Flow Cell (HiSeq 2500 only)	Single Flow Cell (HiSeq 1500 or 2500)	Dual Flow Cell Run Time	Dual Flow Cell (HiSeq 2500 only)	Single Flow Cell (HiSeq 1500 or 2500)	Dual Flow Cell Run Time
1 x 36	95-105 Gb	47-52 Gb	2 days	18-22 Gb	9-11 Gb	7 hr
2 x 50	270-300 Gb	135-150 Gb	5.5 days	50-60 Gb	25-30 Gb	16 hr
2 x 100	540-600 Gb	270-300 Gb	11 days	100-120 Gb	50-60 Gb	27 hr
2 x 150	N/A	N/A	N/A	150-180 Gb	75-90 Gb	40 hr
Reads Passing Filter	Up to 3 billion single reads or 6 billion paired-end reads	Up to 1.5 billion single reads or 3 billion paired-end reads		Up to 600 million single reads or 1.2 billion paired-end reads	Up to 300 million single reads or 600 million paired-end reads	
Quality	Greater than 85% of bases above Q30 at 2 x 50 bp Greater than 80% of bases above Q30 at 2 x 100 bp			Greater than 85% of bases above Q30 at 2 x 50 bp Greater than 80% of bases above Q30 at 2 x 100 bp Greater than 75% of bases above Q30 at 2 x 150 bp		

1レーンで30 Gbase

Waddington's Epigenetic Landscape

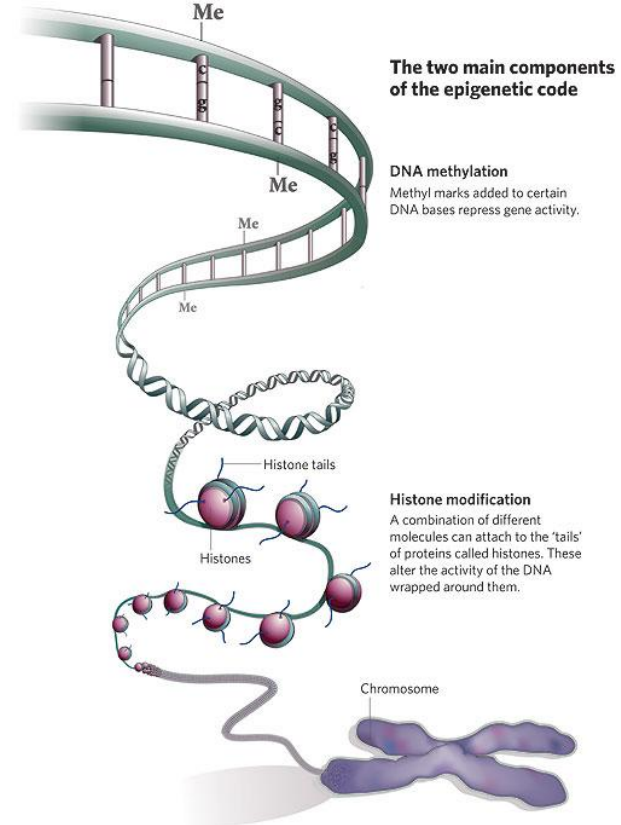
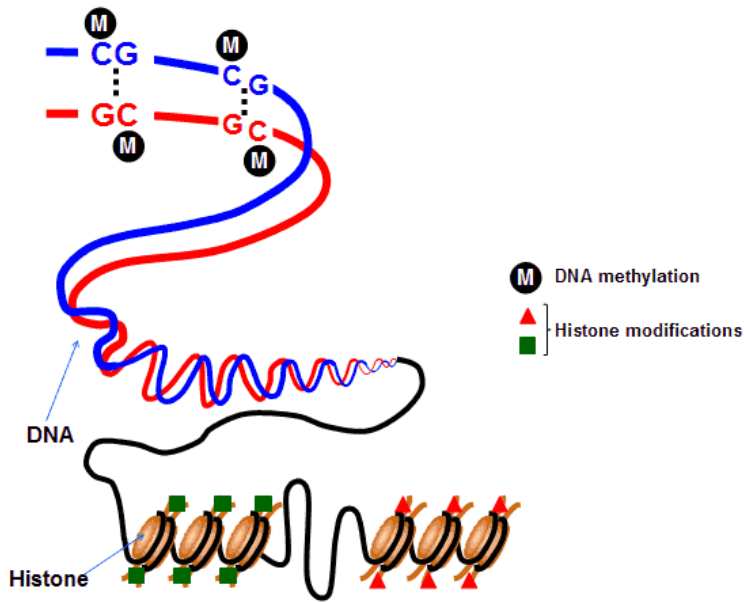


T/BS

Barth and Imhof, 2010

同じゲノムでの異なる組織に分化するのはなぜ？

DNAあるいはヒストンの修飾によるゲノムの制御



エピジェネティクスとは:

クロマチンへの後天的な修飾による遺伝子の発現制御を解析すること。

具体的には、たとえば、同一個体の各細胞は同じDNAを持っているが、エピジェネティックな違い（すなわちクロマチンの修飾の違い）が一因となって、組織特異的な遺伝子発現制御がなされているといえる。

次世代シーケンサーにより、ゲノムワイドなクロマチン修飾の解析が急速に進んでいる！

次世代シーケンサーの応用

- Re-sequencing

 - 1000 genomes project

 - バイサルファイト法によるDNAメチル化解析

- Protein-DNA interaction

 - ChIP-seq(ヒストン修飾)

 - ChIP-seq(転写因子)

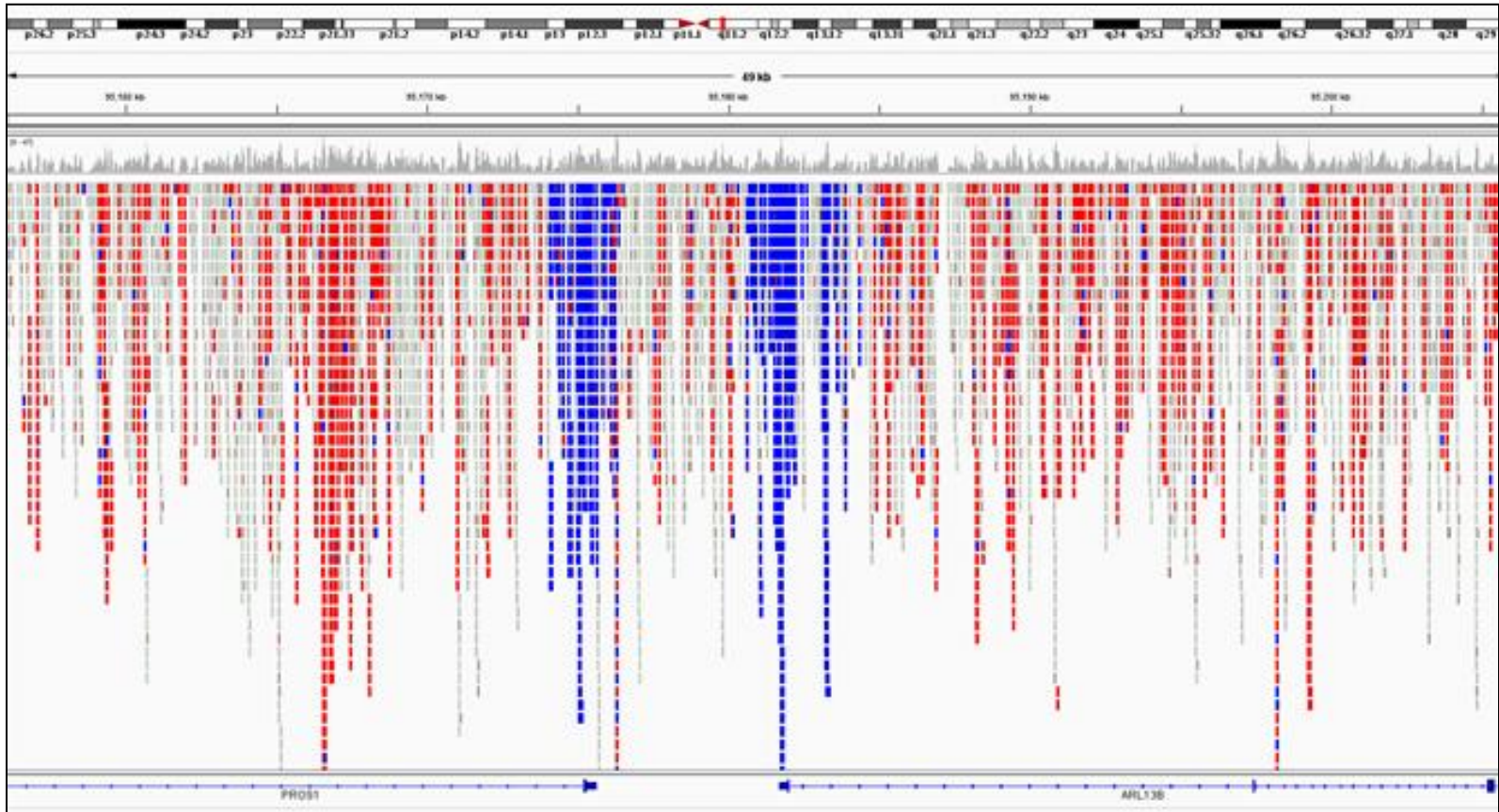
- Transcriptome and detection of RNA genes (ex. miRNA)

 - RNA-seq

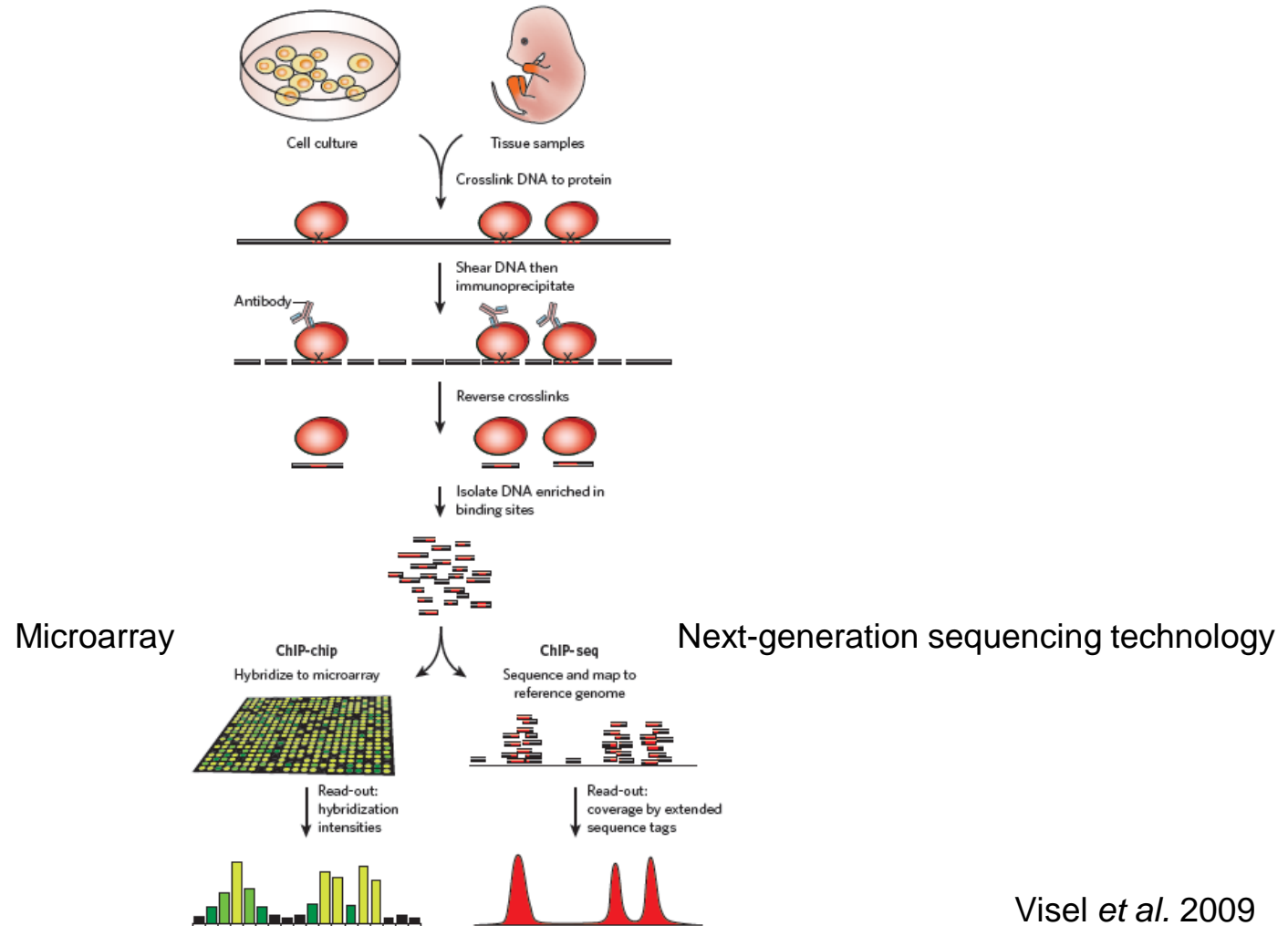
- Protein-RNA interaction

 - CLIP-seq (cross-linking immunoprecipitation followed by high-throughput sequencing)

遺伝子プロモーターにみられる低メチル化領域

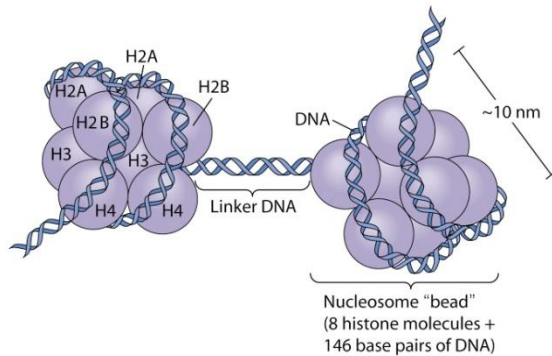


New methods to analyze protein-DNA interaction in genome-wide scale

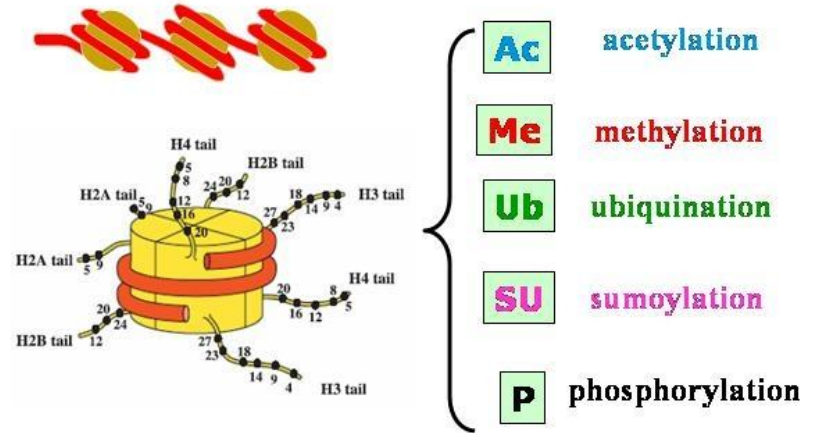
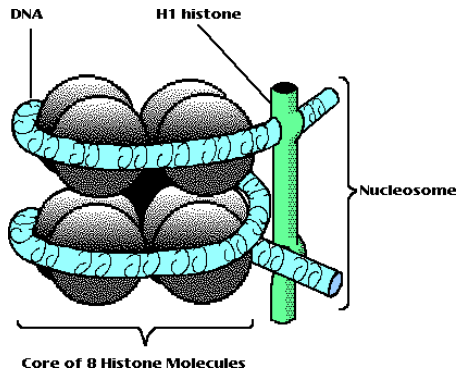


"ChIP-chip" and "ChIP-seq"

ヒストンとヌクレオソーム構造



Copyright © 2009 Pearson Education, Inc.



The figure illustrates nucleosome models and major posttranslational modifications which play essential roles in gene expression regulation and disease processes

- ・ヌクレオソームは、4つのコアヒストンタンパク質 (H2A, H2B, H3, H4) とリンカーヒストンH1で構成される。
- ・以前は、DNAをパッケージングする静的な足場として考えられてきた。
- ・最近になって、**ヒストンの様々な修飾には**、遺伝子の転写活性に直接影響していることをはじめ、クロマチン凝集やDNAへのアクセスのし易さを調節するといった**多彩な機能がある**ことがわかってきた。

ヒストン修飾の命名

H3K4me3

name of the histone

amino acid

position in the sequence

Modification:

acetylation

mono-methylation

di-

tri-

mono-ubiquitination

phosphorylation

ac

me1

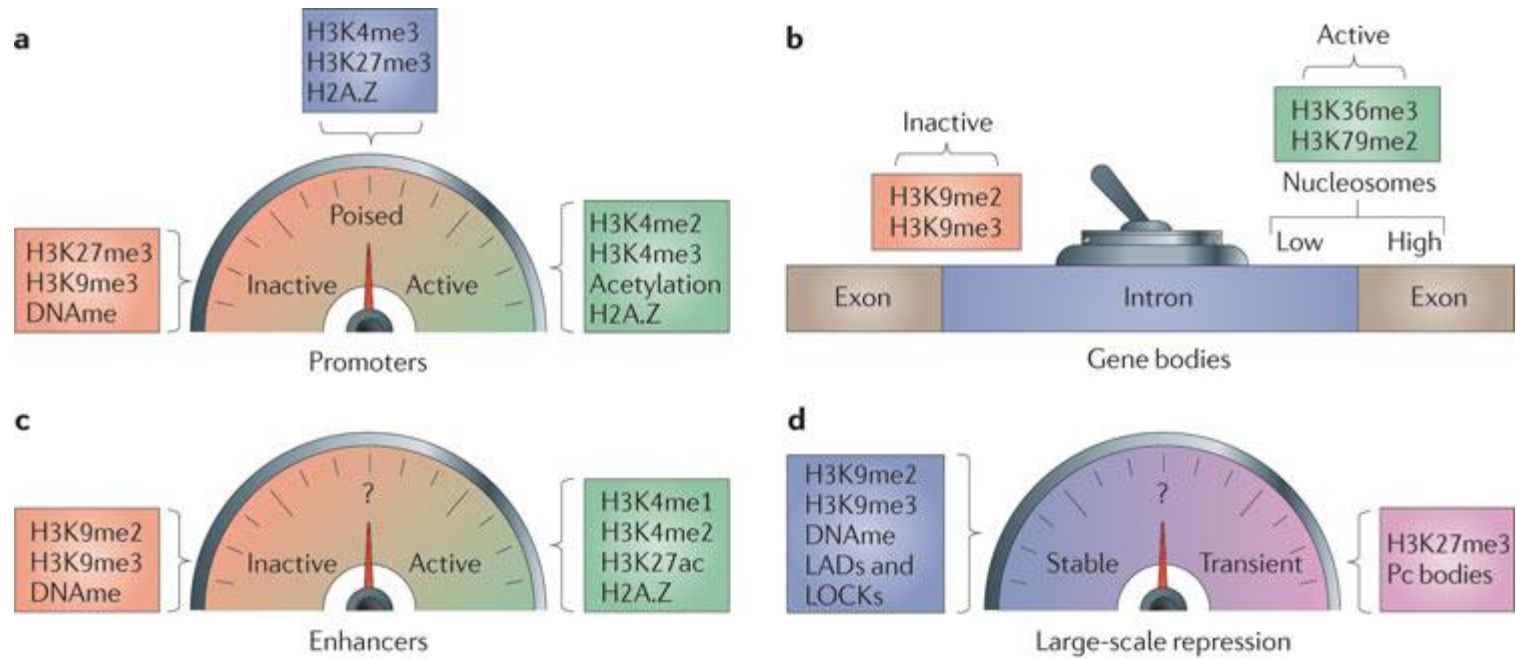
me2

me3

ub1

P

ヒストン修飾と生物学的機能の関連



Nature Reviews | **Genetics**

Zhou *et al. Nat. Rev. Genet.* 2011.

エピゲノムデータの多次元的な広がり

組織・発生段階・性差
(brain, liver, muscle, ...,
fetal/adult, male/female, ...)

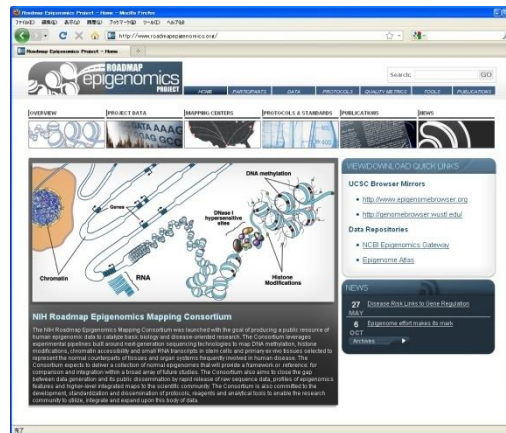
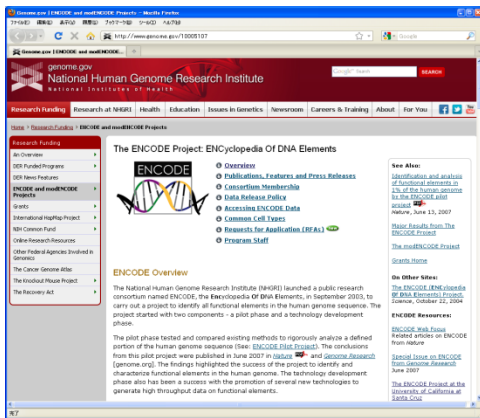
生物種

ヒストンマークなど
(H3K4me1, H3K27Ac, DNaseI, ...)

- ENCODE
mouseENCODE,
modENCODE (for drosophila, worm)

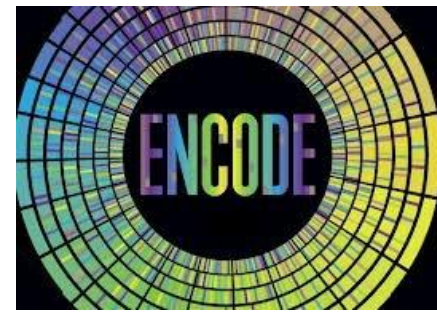
- Roadmap Epigenomics Project

- International Human Epigenome Consortium (IHEC)

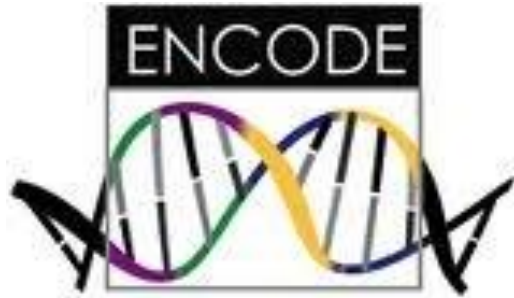


An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*



The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.



一言でいうと"regulome"が目標

Facts and figures:

Launched by NHGRI in Sept. 2003.

Pilot projectは2007に終了

Oct. 2007に4年で\$80Mの予算で'production phase'として継続決定

1000 Genomes Projectとも一部オーバーラップ

32 groups, >440 scientists, 24 standard types of experiment

120 transcription factors, ...

MAKING A GENOME MANUAL

Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.

EXPERIMENTAL TARGETS

DNA methylation: regions layered with chemical methyl groups, which regulate gene expression.

Open chromatin: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.

RNA binding: positions where regulatory proteins attach to RNA.

RNA sequences: regions that are transcribed into RNA.

ChIP-seq: technique that reveals where proteins bind to DNA.

Modified histones: histone proteins, which package DNA into chromosomes, modified by chemical marks.

Transcription factors: proteins that bind to DNA and regulate transcription.

CELL LINES

Tiers 1 and 2: widely used cell lines that were given priority.

Tier 3: all other cell types.

Every shaded box represents at least one genome-wide experiment run on a cell type.

So far, scientists have examined 13 of about 60 known histone modifications and 120 of about 1,800 transcription factors.

Many more cell types are yet to be interrogated.

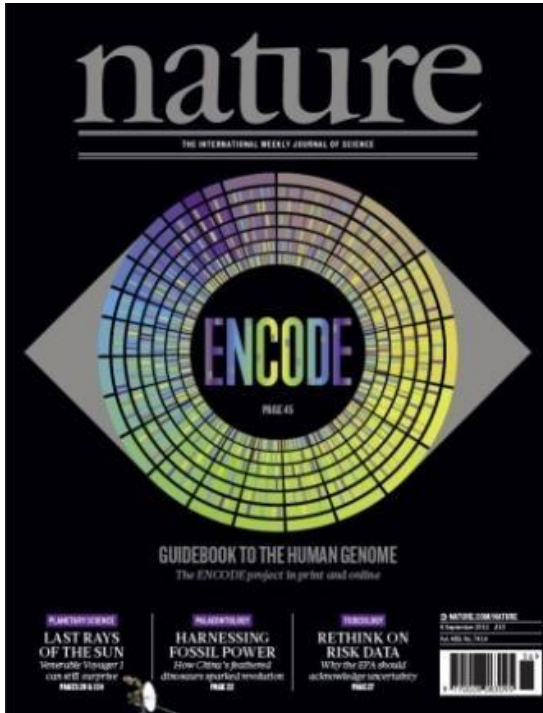
DNA
RNA ChIP Cell line

$$(28+133) \times (3+14+40) = 9177 \text{ 実験}$$

1実験10Gbyteとすると、基本データだけで100Tbyte

The results of the ENCODE project (30 papers)

September, 2012



A screenshot of the Genome Biology website, showing the ENCODE project articles. The browser address bar displays "genomebiology.com/series/ENCODE". The page features the Genome Biology logo with an impact factor of 9.04. The navigation menu includes "Home", "Articles", "Authors", "Reviewers", "About this journal", "My Genome Biology", and "Subscriptions". The "Articles" section is active, and the "The ENCODE project" article is highlighted. The article text describes the ENCODE project's mission and lists several research papers published in the September 2012 issue of *Genome Biology*.

The ENCODE project

The Encyclopedia of DNA Elements (ENCODE) Consortium's mission statement was to comprehensively annotate functional elements in the human genome. Following nearly ten years of data generation by over 400 researchers across the globe, the project's findings have now been published as a group of 30+ articles in a multi-publisher collaboration. The ENCODE articles published by BioMed Central are presented below, and the publication effort is discussed in further detail on the [BMC Blog](#).

For more information on ENCODE, please see the [ENCODE web portal](#). The ENCODE articles from all three publishers can also be downloaded as an [iPad app](#), or browsed in the [ENCODE Explorer](#).

Collection published: 5 September 2012

Research [Open Access](#) [Highly accessed](#)
Modeling gene expression using chromatin features in various cellular contexts
Xianjun Dong, Melissa C Greven, Anahil Kundaje, Sarah Djebali, James B Brown, Chao Cheng, Thomas R Gingeras, Mark Gerstein, Roderic Guigó, Evan Birney, Zhiping Wang
Genome Biology 2012, **13**:R53 (5 September 2012)
[Abstract](#) | [Full text](#) | [PDF](#) | [PubMed](#) | [Editor's summary](#)

Research [Open Access](#) [Highly accessed](#)
Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3
Seth Frieze, Rui Wang, Lijing Yao, Yu Tak, Zhenqing Ye, Malaina Gaddis, Heather Witt, Peggy J Farnham, Victor X Jin
Genome Biology 2012, **13**:R52 (5 September 2012)
[Abstract](#) | [Full text](#) | [PDF](#) | [PubMed](#) | [Editor's summary](#)

Research [Open Access](#) [Highly accessed](#)
The GENCODE pseudogene resource
Baikang Pei, Cristina Sisu, Adam Frankish, Cédric Howald, Lukas Habegger, Ximeng Mu, Rachel Harte, Suganthi Balasubramanian, Andrea Tanzer, Mark Diekhans, Alexandre Reymond, Tim J Hubbard, Jennifer Harrow, Mark B Gerstein
Genome Biology 2012, **13**:R51 (5 September 2012)
[Abstract](#) | [Full text](#) | [PDF](#) | [PubMed](#) | [Editor's summary](#)

エピゲノム・プロジェクト

Roadmap Epigenomics

The screenshot shows the Roadmap Epigenomics website with a navigation menu including HOME, PARTICIPANTS, BROWSE DATA, PROTOCOLS, COMPLETE EPIGENOMES, TOOLS, and PUBLICATIONS. A sub-menu is open for BROWSE DATA, showing OVERVIEW, PROJECT DATA, MAPPING CENTERS, PROTOCOLS & STANDARDS, PUBLICATIONS, and NEWS. A large diagram illustrates DNA methylation, DNase I hypersensitive sites, RNA, Chromatin, and Histone Modifications. Below the diagram is the NIH Roadmap Epigenomics Mapping Consortium logo and a brief description of the consortium's goals.

Blueprint Epigenomics

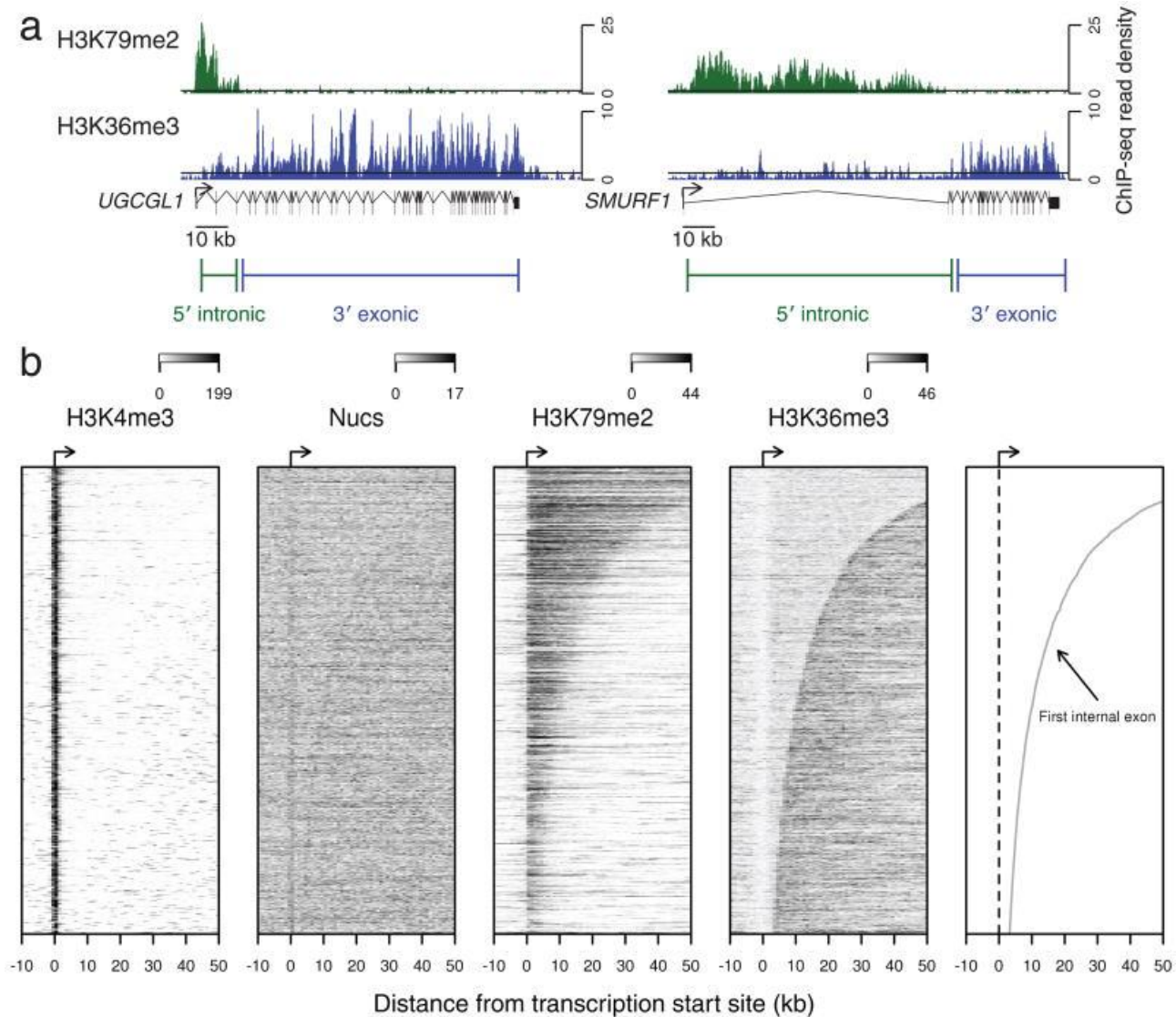
The screenshot shows the Blueprint Epigenomics website with a navigation menu including HOME, RESEARCH, RESULTS, PARTICIPANTS, NEWS, LINKS, and CONTACT. A large blue banner displays the 'BLUEPRINT epigenome' logo. Below the banner is a 3D visualization of DNA packaging. A 'USER LOGIN' section is visible on the right side of the page.

The screenshot shows the IHEC website with a navigation menu including About, Research, Outcomes, Epigenomics, News+Events, and Links. A large world map is displayed with a text box stating: "The International Human Epigenome Consortium (IHEC) unites scientists from all over the world working together to achieve one common goal: deciphering 1000 epigenomes." Below the map is a blue banner with links for Welcome to IHEC, Voices of IHEC, IHEC Protocols, and Annual Meeting. The main content area is divided into three sections: Research, Why Epigenomics?, and News+Events, each with a representative image.

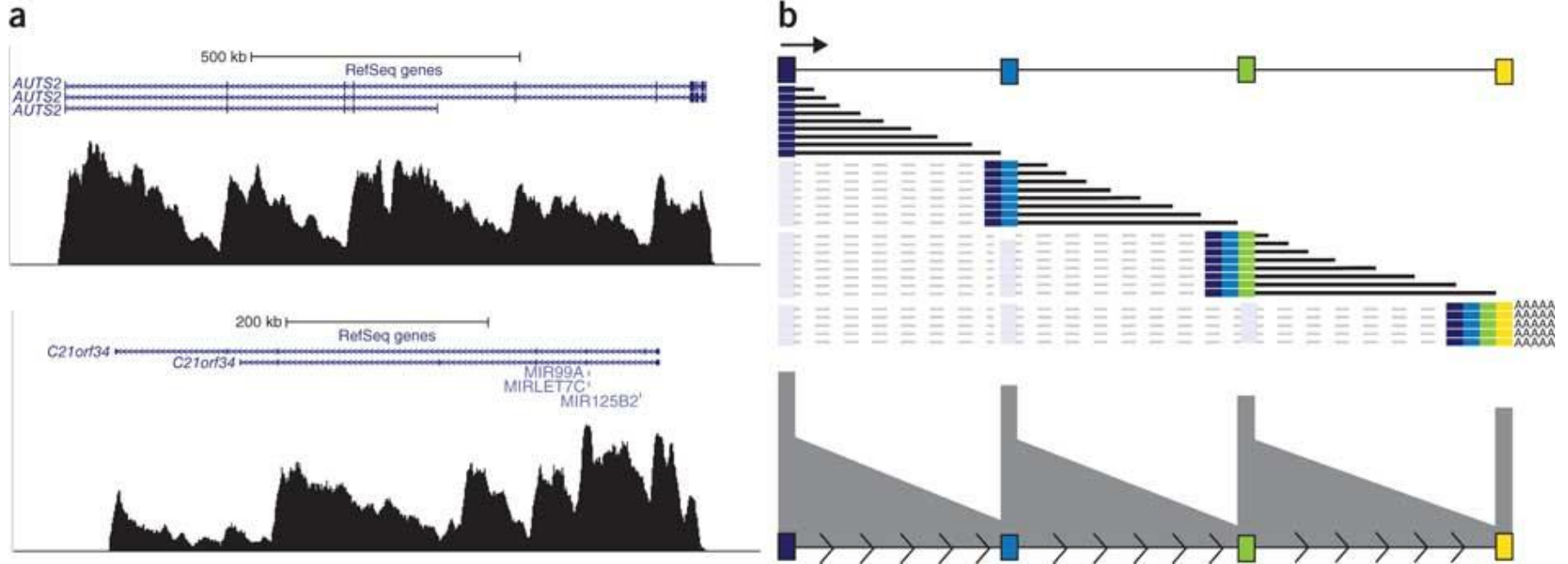
IHEC: International Human Epigenome Consortium

次世代シーケンサーが可能にした
新しいゲノミクス研究

イントロン/エクソン構造がゲノム中すでにマークされている



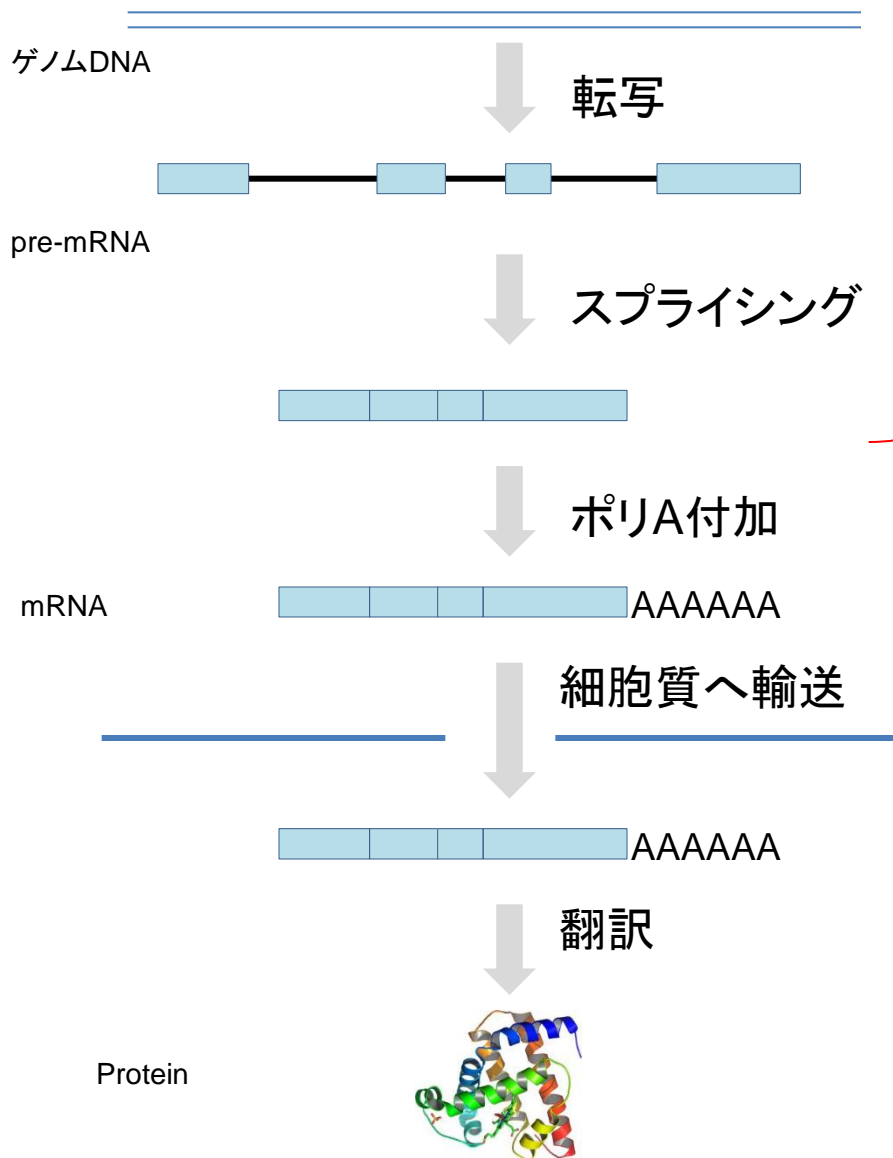
転写と共役したスプライシング (co-transcriptional splicing)



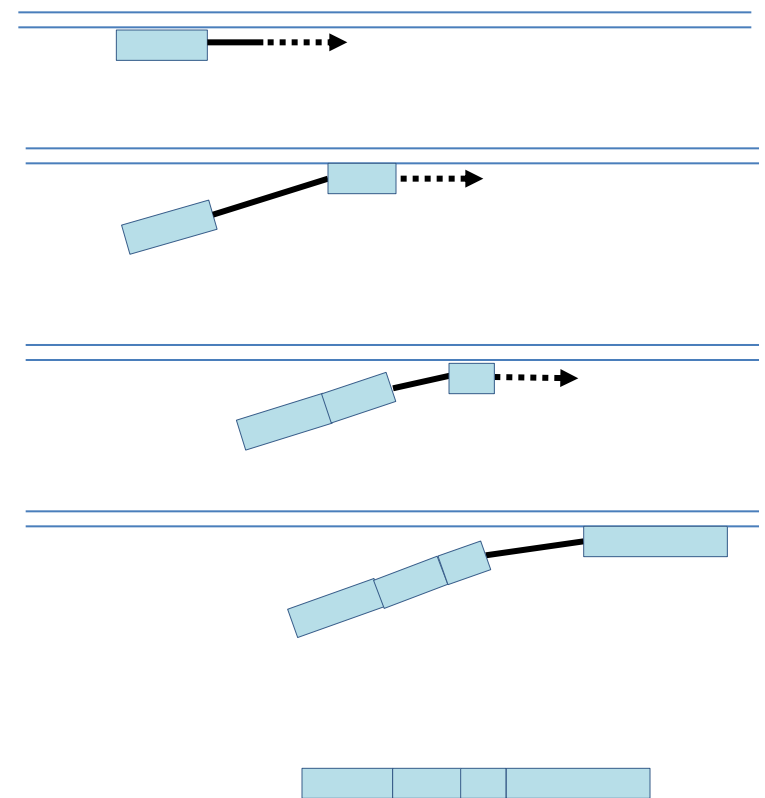
Ameur et al. Nat. Struct. Mol. Biol. 2011.

多くのイントロンは転写と共役してスプライスされる

よく教科書に見られる記述

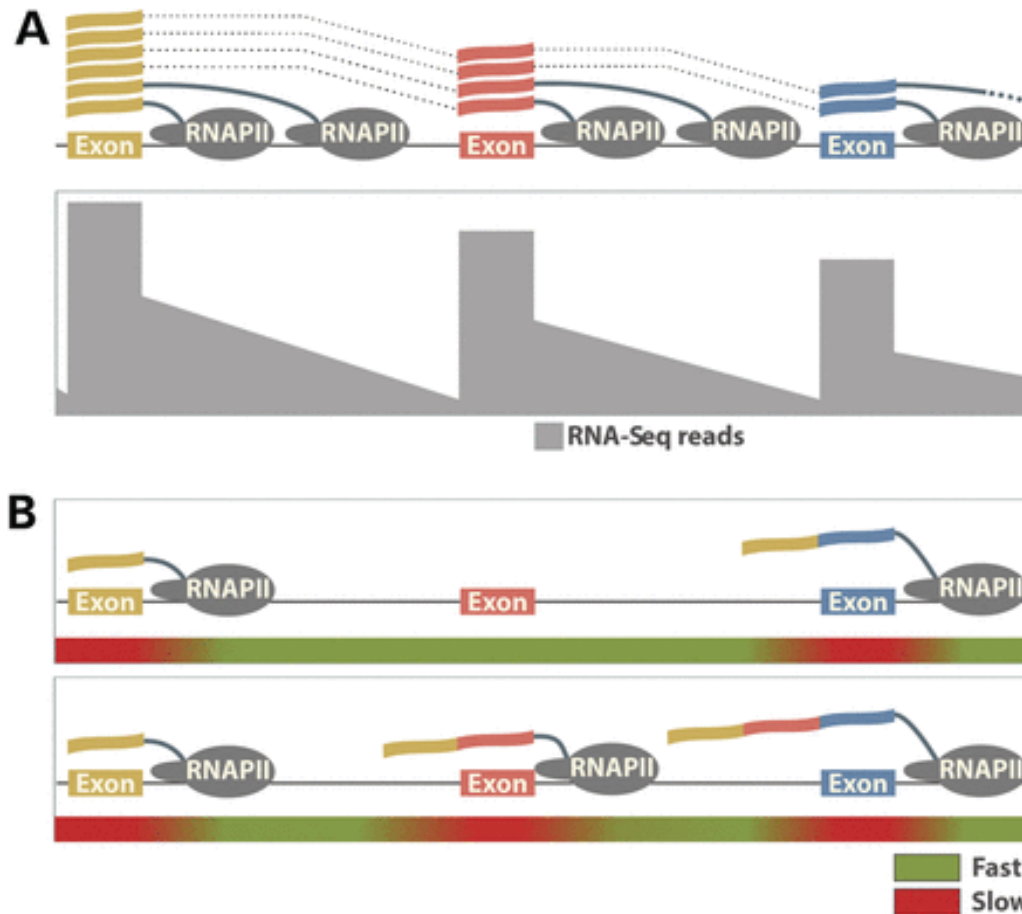


でも実際は...



co-transcriptional splicing

(A) 転写と共役したスプライシングと (B) 転写速度によるエクソンスキップの制御

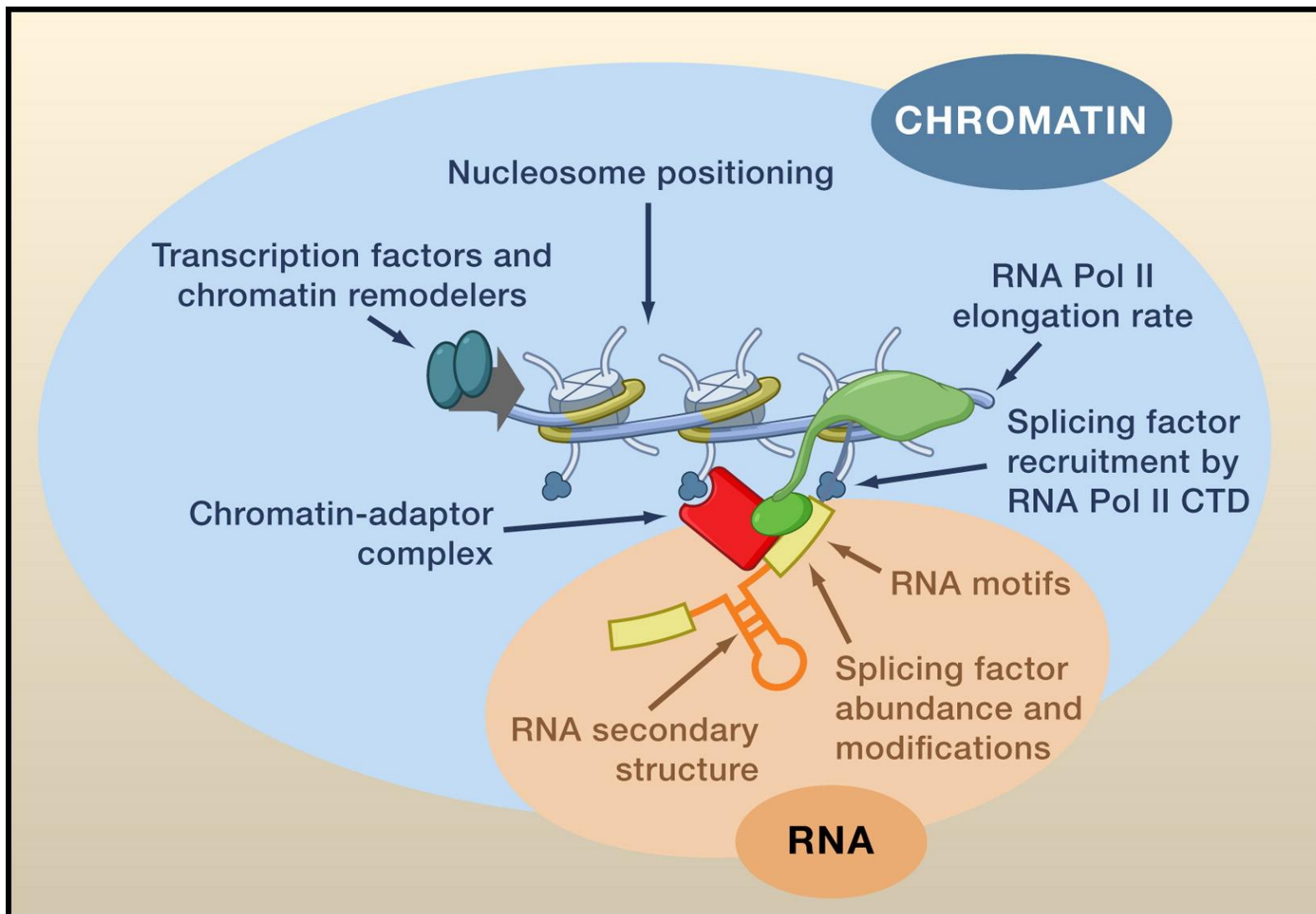


Brown et al. *Hum. Mol. Genet.* 2012.

- A. 多くの遺伝子で、転写が進むに従って順次スプライシングが起きていることがわかってきた。
- B. ある選択的スプライシングにおいては、RNase polymerase IIの転写速度の違いにより、エクソンのスキップと取り込みが制御されている (kinetic control of alternative splicing)。

より実地的な「転写」と「スプライシング」の関係

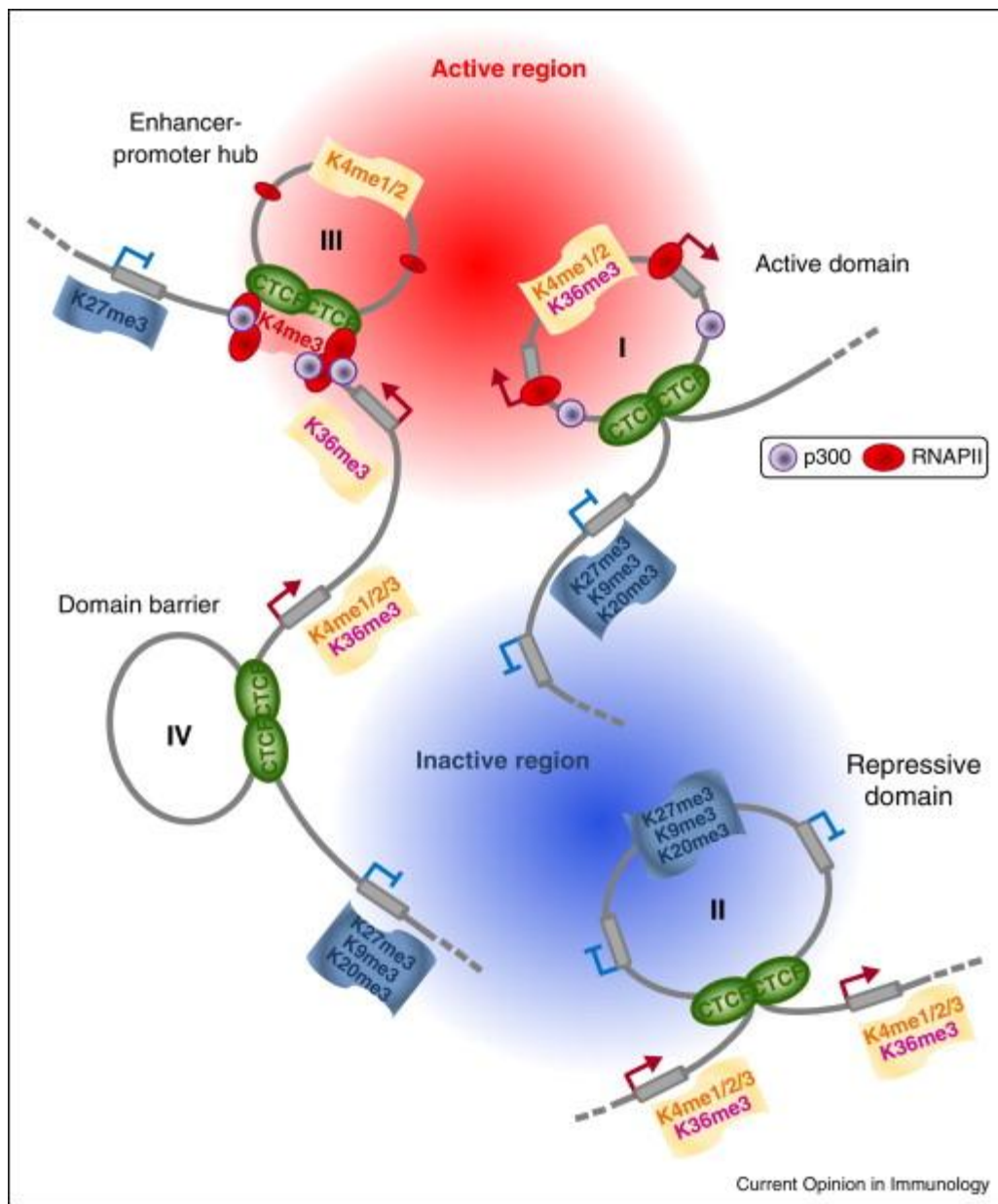
スプライシングにおける転写機構やクロマチン状態の関与



転写因子はプロモーターから遠く離れたところに結合することもある。



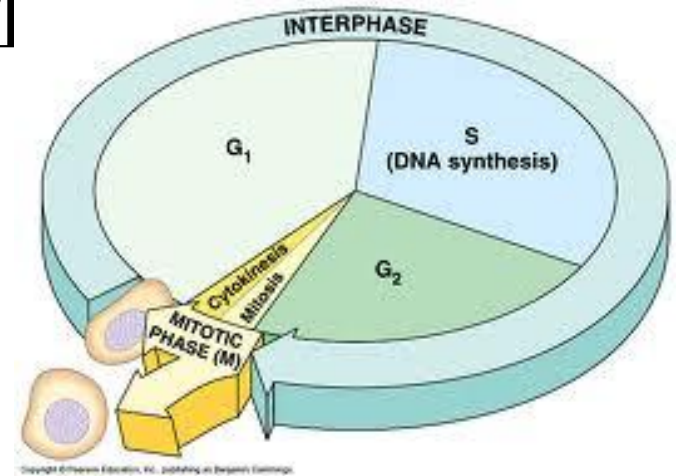
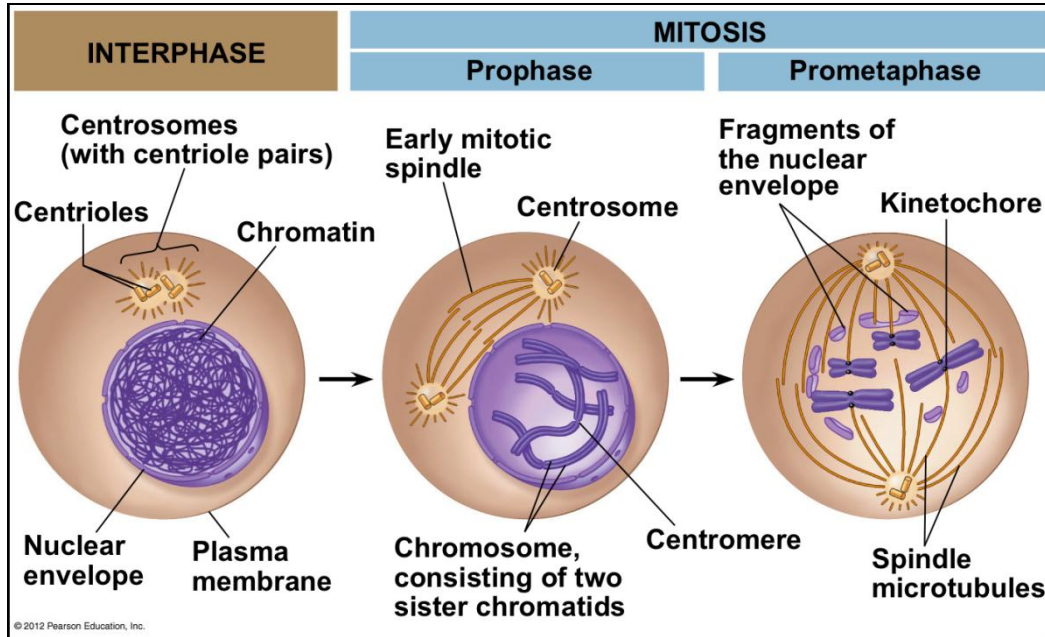
インシュレーター(insulator)によって規定されるゲノムの[活性/不活性]領域



"Transcription factories"

CTCF: インシュレーターとして働くタンパク質

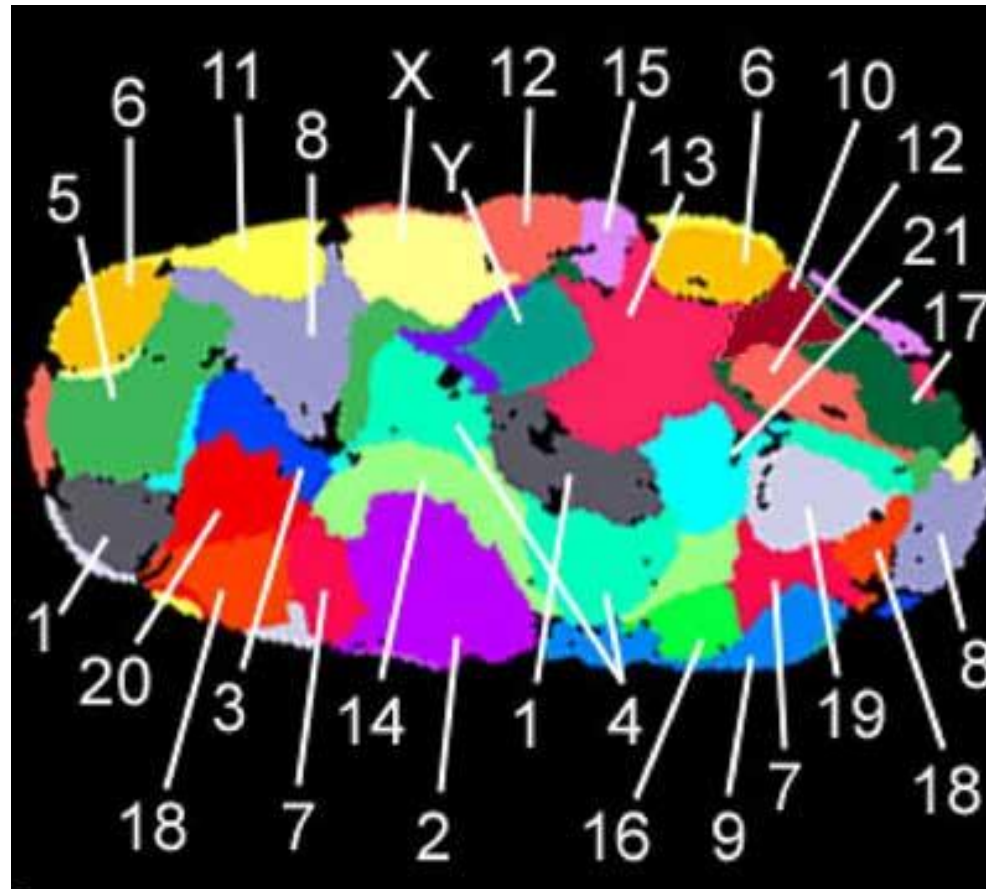
染色体は普段どんな形をしているか



間期(interphase)の染色体はランダムにほどけているわけではない

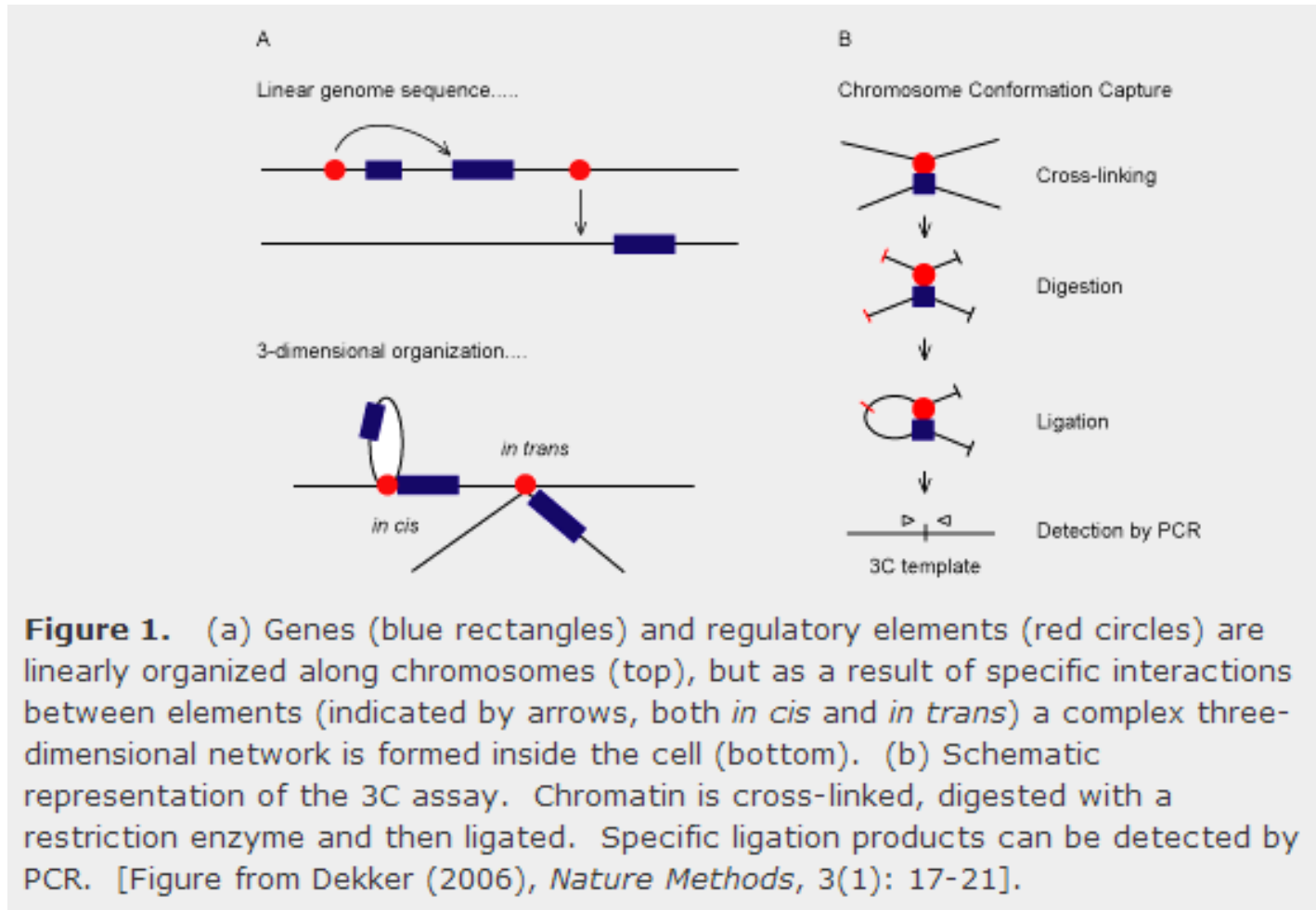
それぞれの染色体が決まった領域に広がっている

→ クロモソーム・テリトリー(chromosome territory; CT)



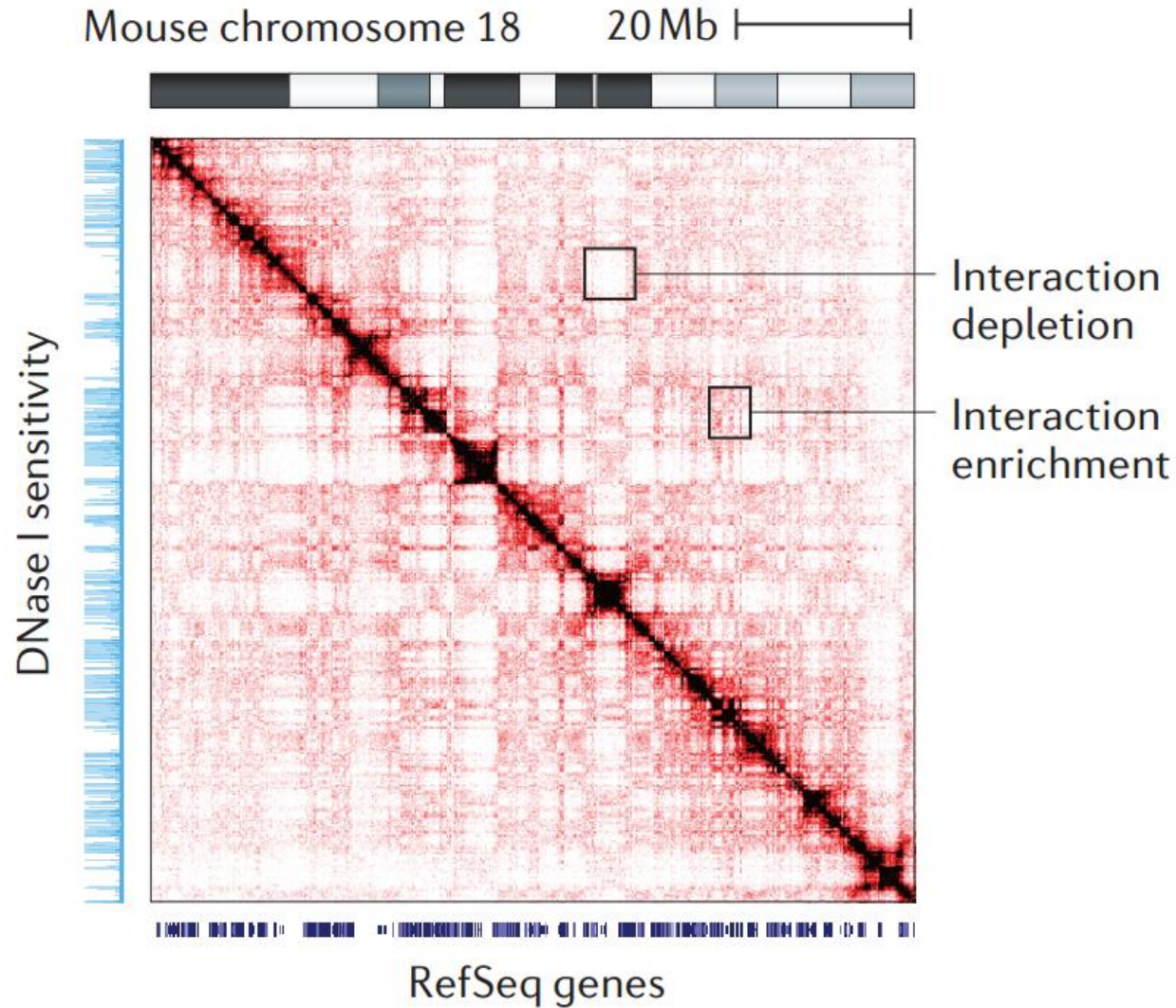
染色体間の空間的近さを測る方法

Chromosome Conformation Capture (3C)



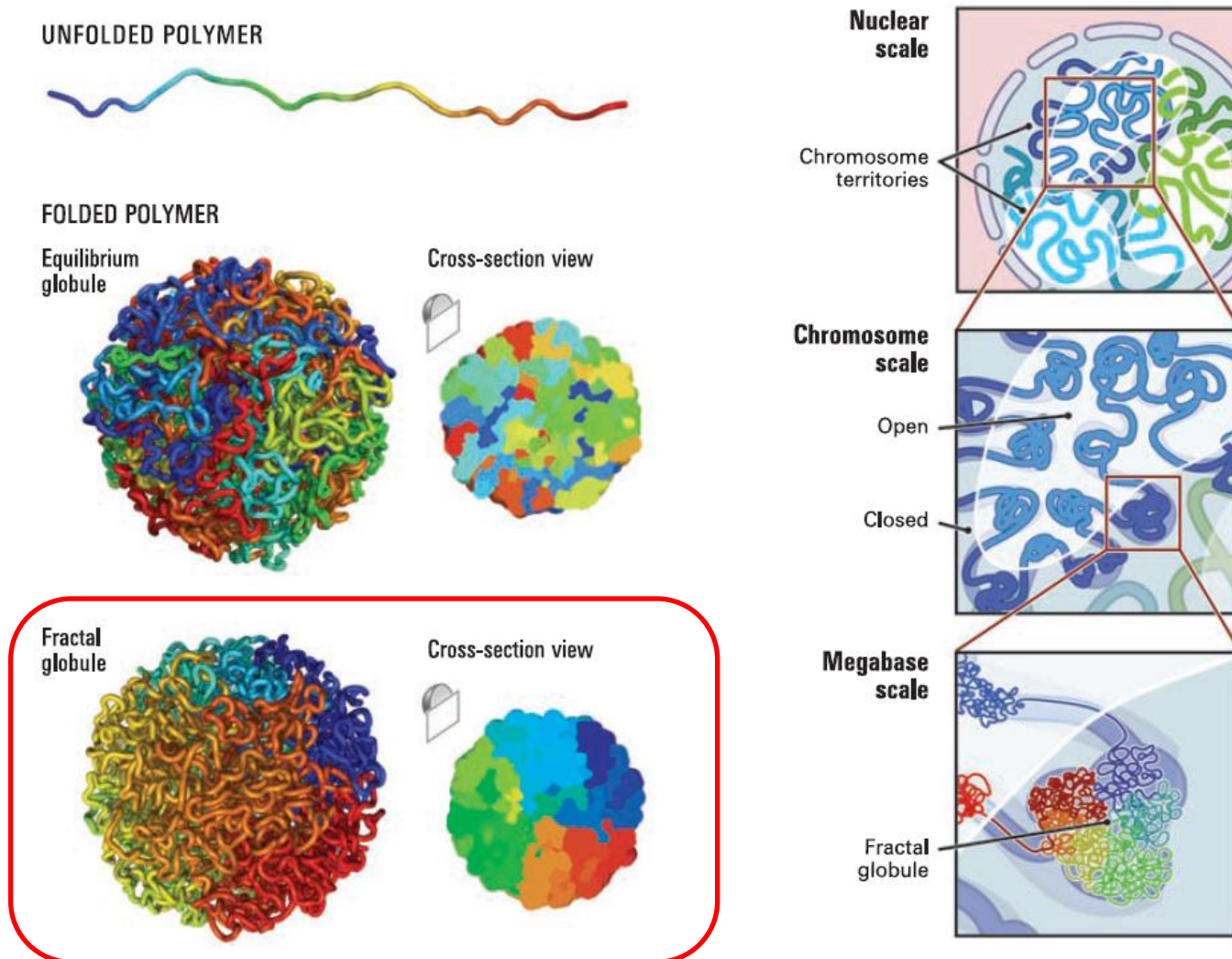
次世代シーケンサーを使えば、ゲノムワイドに空間的な近さを測ることができる。

Hi-C法による染色体のinteraction map

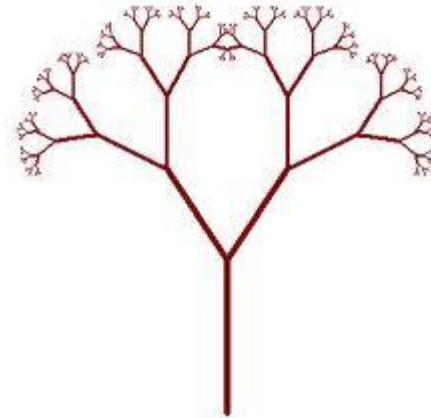
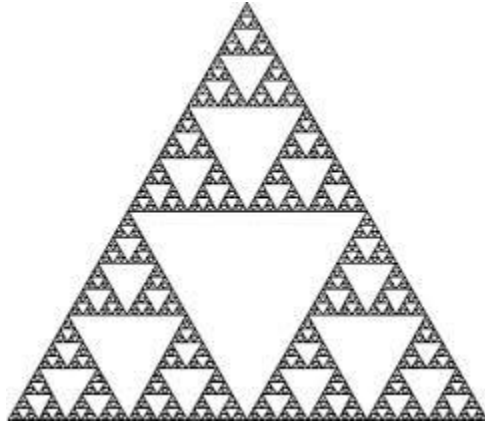


Hi-Cのデータから得られた染色体の立体構造モデル

"Fractal globule"



フラクタル図形の例



自己相似形

NewsBytes

New Technology Reveals the Genome's 3D Shape

Try taking a human hair as long as Manhattan and cramming it—unsnarled—inside a marble. This is the challenge faced by a 2-meter-long strand of DNA as it folds into its compact array of 23 chromosomes within a cell's nucleus. Previously, scientists only theorized about how DNA squeezes inside a nucleus without becoming a hopelessly tangled mass. Now a new technique called Hi-C reveals that DNA packs knot-free into its chromosomal patterns by assuming a rare geometric shape observed in snowflakes, crystals and broccoli.

"We've developed

a powerful new technique to look at chromosomes at an unprecedented resolution," says **Job Dekker, PhD**, cell biologist at the University of Massachusetts and coauthor of the study in the October 9, 2009 issue of *Science*. "What we found constitutes a breakthrough in our understanding of chromosome folding."

At the small scale, DNA wraps around proteins called histones and assumes its classical double-helix shape. At the large scale, chromosomes cluster in discrete sections within the nucleus called "territories." "Between the scale of chromosome territories and the scale of histones, effectively nothing has been known about the structure of the genome," says first author **Erez Lieberman-Aiden**, a graduate student in the lab of **Eric Lander, PhD**, professor of biology at the Broad Institute in Cambridge, Massachusetts.

Hi-C reconstructs an unbiased 3-D map of the entire genome.

First, scientists soak a complete set of chromosomes in formaldehyde, which acts like glue to stick together parts of the genome that are close in 3-D space. Then they chop the DNA into a million pieces and

perform massive parallel sequencing on the interacting fragments. Mapping software compares the sequences of attached fragments with a human genome reference sequence; based on the results, the scientists compute which parts of the folded DNA physically interact with each other.

The team found that active, gene-rich and inactive, gene-poor sections cluster in separate parts of the nucleus. The active chromatin segments are like easily accessible papers spread out across a desk, whereas the inactive portions are densely packed, like folders in a file cabinet.

Simulations revealed that DNA assembles into dense fractal globules—structures that look alike at different levels of magnification, such as the intricate geometrical form of a crystal. Genes are easily accessible, but when they're not in use, the structure spontaneously collapses into a tight, knot-free bundle.

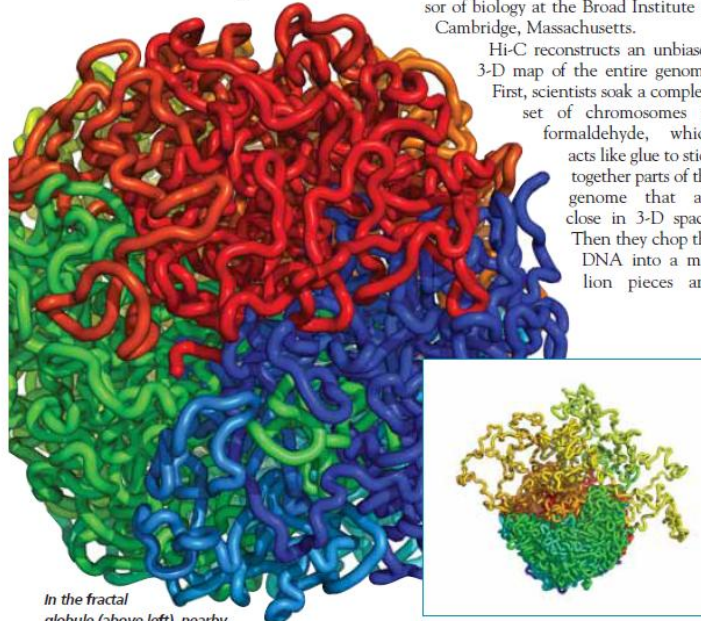
"This is the first spatial map of the genome," says **Tom Misteli, PhD**, cell biologist at the National Cancer Institute in Bethesda, Maryland. "It's a technical breakthrough that opens the doors to doing all sorts of interesting experiments."

Future experiments will investigate how the 3-D shape of DNA morphs depending on the activity of genes and disease states, like cancer. As genome sequencing becomes cheaper, Dekker says, it should be possible to obtain higher spatial resolution and even to reconstruct the shapes of individual genes.

—By **Janelle Weaver, PhD**

How DNA Goes A'Courtin'

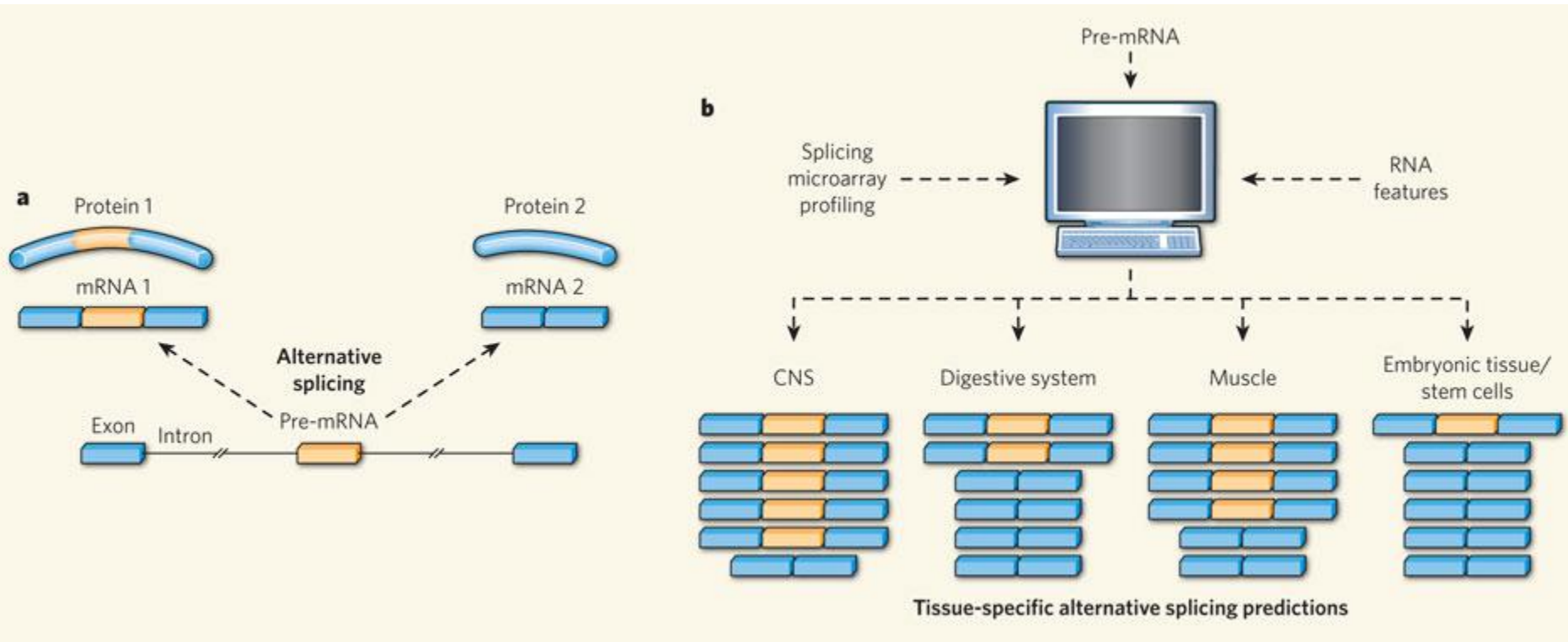
Until now, scientists have known little about how complementary single strands of DNA court one another before binding to form the classical double helix. But now, molecular dynamics simulations have identified that the binding—or hybridization—mechanism depends largely on the sequence of the DNA: Ordered sequences will meet and then slither lengthwise to find the correct match; but sequences that are random will connect at key sites then rapid-



In the fractal globule (above left), nearby regions on a chain of DNA—indicated using similar colors—are packed into nearby regions in 3D space. The accessible DNA chain unravels easily (above right) because the globule lacks knots. Images courtesy of Leonid A. Mirny and Maxim Imakaev, reprinted from Lieberman-Aiden, E., et al., *Comprehensive Mapping of Long-Range Interactions Reveal Folding Principles of the Human Genome*, *Science*, 326(5950): 289-293 (2009), with permission from AAAS.

スプライシングコードの解明

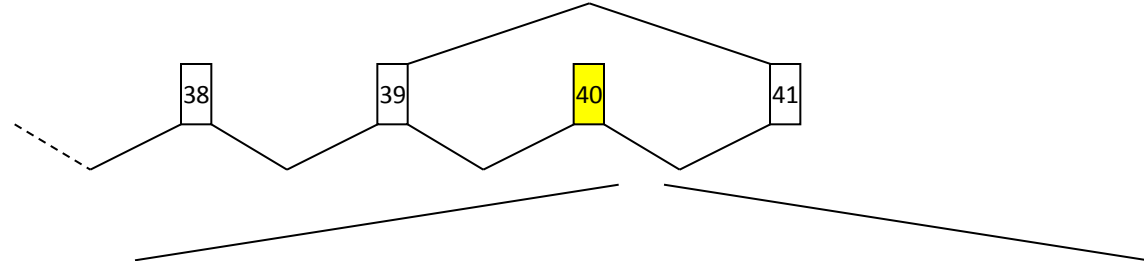
Breaking the second genetic code



Ramón Tejedor and Valcárcel, *Nature*, 2010
(Comment on Barash *et al.*, *Nature*, 2010)

スプライシングのシス因子の探索

GTPase activating Rap/RanGAP domain-like 1 protein (GARNL1)

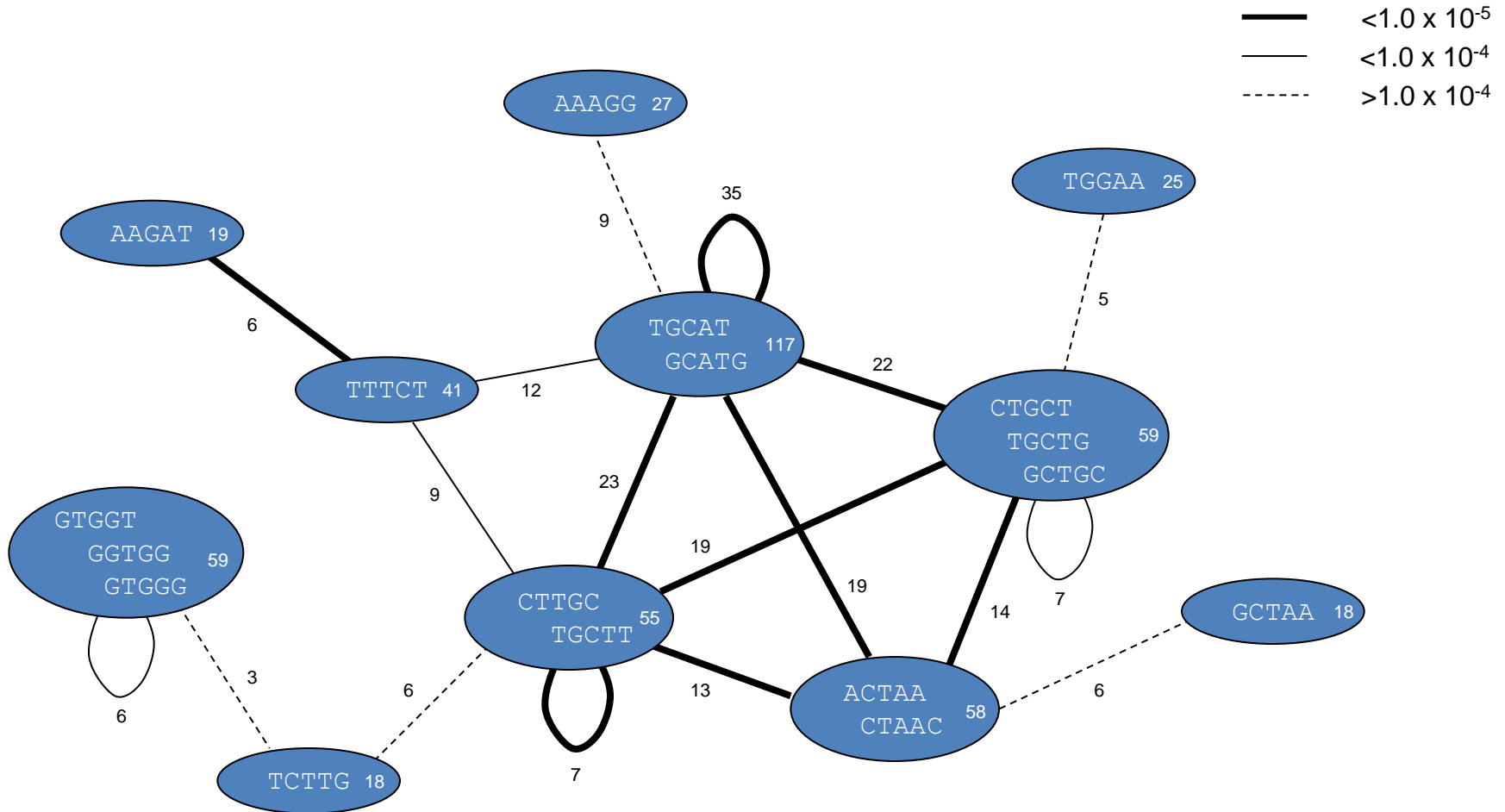


human	chr14	-	35087341	ACATTTTCAGAAATTGTCACTAAATTTTTTCC--AGTATTA--TACTGACTAAGCC-AGGTCTGCATGAAACACTAACA-T
chimpanzee	chr15	-	34251985	ACATTTTCAGAAATTGTCACTAAATTTTTTCC--AGTATTA--TACTGACTAAGCC-AGGTCTGCATGAAACACTAACA-T
macaque	chr7	-	98506834	ACATTTTCAGAAATTGTCACTAAATTTTTTCC--AGTATTA--TACTGACTAAGCC-AGGTCTGCATGAAACACTAACA-T
rat	chr6	-	75858765	ACATTTCAAAAATTATCACTAAATTTTTTCCCAGAATTG--TACTAACTAAGCC-AGGTCTGCATGAAACACTAACA-T
mouse	chr12	-	56530084	ACATTTCAAAAATTATCACTAAATTTGTTCCCGAGAAGT--TGCTAACTAAGCC-AGGTCTGCATGAAACACTAACC-C
rabbit	scaffold_178393	-	17945	ACATGGCAGAAATTGTCACTACATTTTTTCC--AGATTTA--TACTAACCAAGCC-AGGTCTGCATGAAACACTAACA-T
cow	chr21	-	30280498	ACATTTTCAGAAATTGTCACTAAATTTCTTCC--AGAATTC--TACTTACTAAGCC-AGGTCTGCATGAAACACTAACA-T
dog	chr8	-	17234636	ACATTTTCAGAAATTGTCACTAAATTTCTTCC--GGAATTA--TACTTACTAAGCC-AGGTCTGCATGAAACACTAACA-T
armadillo	scaffold_3577	-	24577	ACATTTTCAGAAATTGTC--CTAAAT--CTTCC--AAAATTG--TTCTTACTAACAC-AGGTCTGCATGAAACACTAACA-T
tenrec	scaffold_299940	+	3914	ATATTTTCAGAAATTGTCACTAAATTTTTTCC---CAGTTA--TACTTACTAAGCC-AGGTCTGCATGAAACACTAACA-C
opossum	chr1	-	286030148	ACATTTTCAGAAAGTTTTTACTAAATTTTTTCC--AAAGTTAGTTTTTACTAAGCCAGGTCTGCATGAAA-ACTAACA-C



2つのシス因子が一緒にあらわれる(共起)

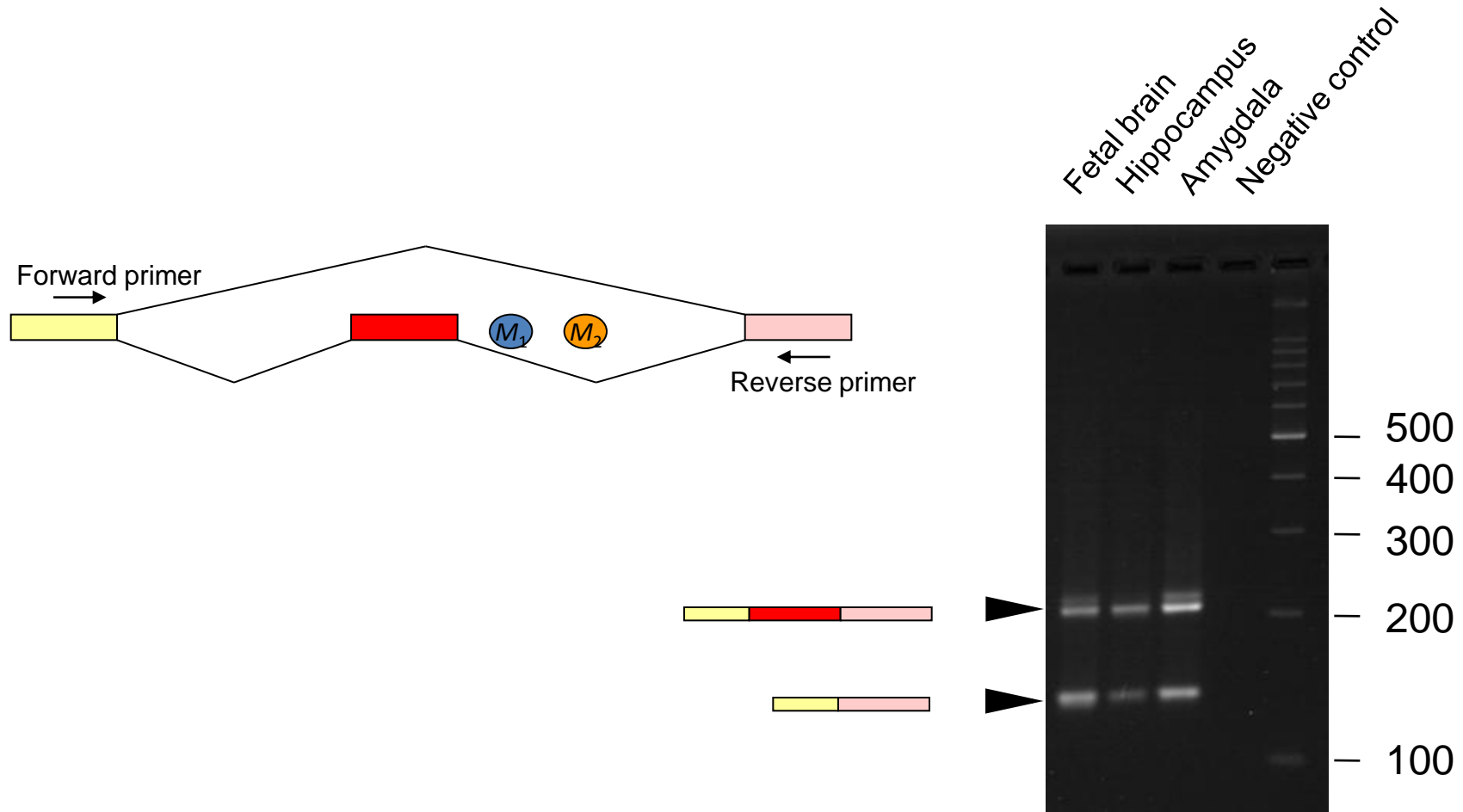
A network of co-occurring motifs



シス因子の組み合わせにより多様性を生み出す。

シス因子の共起から、未知のエクソンスキップを予測し、実験で検証

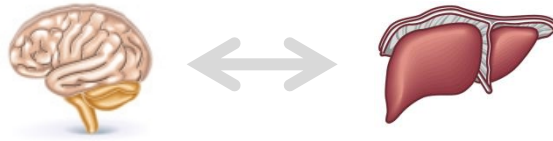
The 3rd exon of the ENST00000256858 transcript.



We randomly selected 10 predictions, and confirmed the skipping in 3 cases.

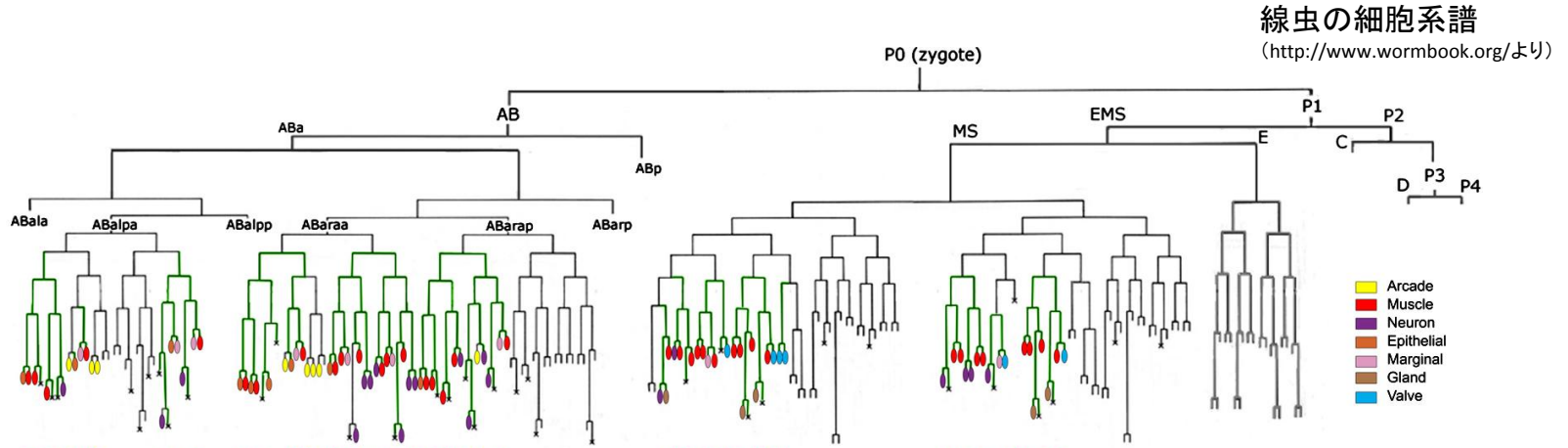
ゲノクス解析の今後

これまで:



2~3の組織での発現比較

今後:



全ての組織での発現比較

(ヒトの場合約200の細胞種)

データ蓄積の増加、その解析に要する計算量の飛躍的増加！

謝辞

九州大学 生体防御医学研究所

情報生物学分野

佐藤哲也

吉原美奈子

EMBL

Peer Bork

Eoghan Harrington

Bork Group Members

九州大学 医学研究院

先端医療医学部門 エピジェネティクス分野

大川恭行

Universität Heidelberg

Magnus von Knebel Doeberitz

Svetlana Vinokourova

かずさDNA研究所

理研免疫アレルギー科学研究センター

小原收

文科省 新学術領域「性差構築の分子基盤」

基盤研究C

特定領域研究「ゲノム」

JST CREST「エピゲノム研究に基づく診断・治療へ

向けた新技術の創出」