新生命科学分野開拓とスーパーコンピュータ「京」

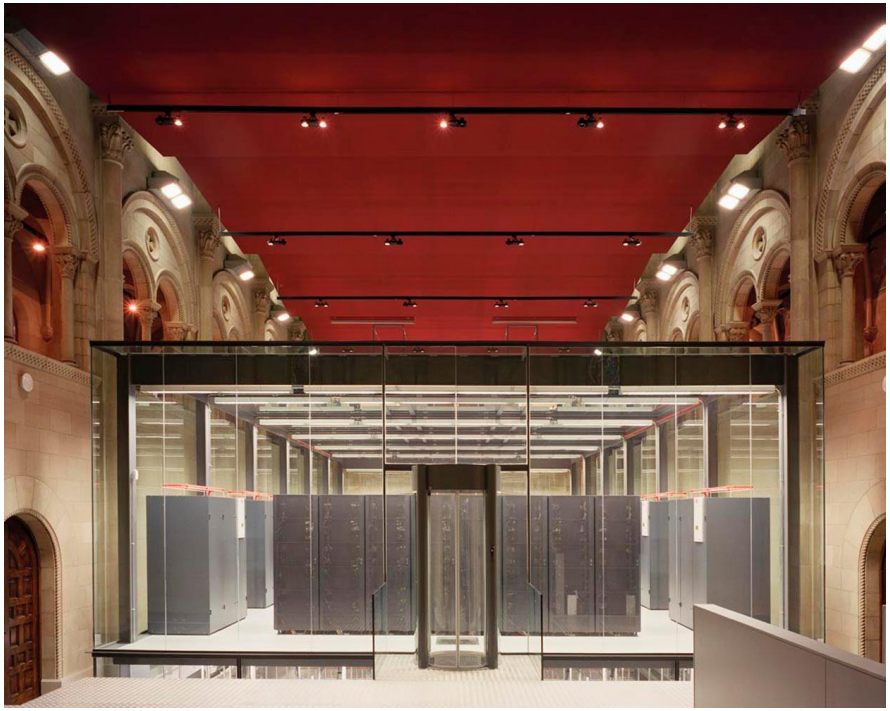
2013/9/19 九州大学医学部百年講堂

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

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

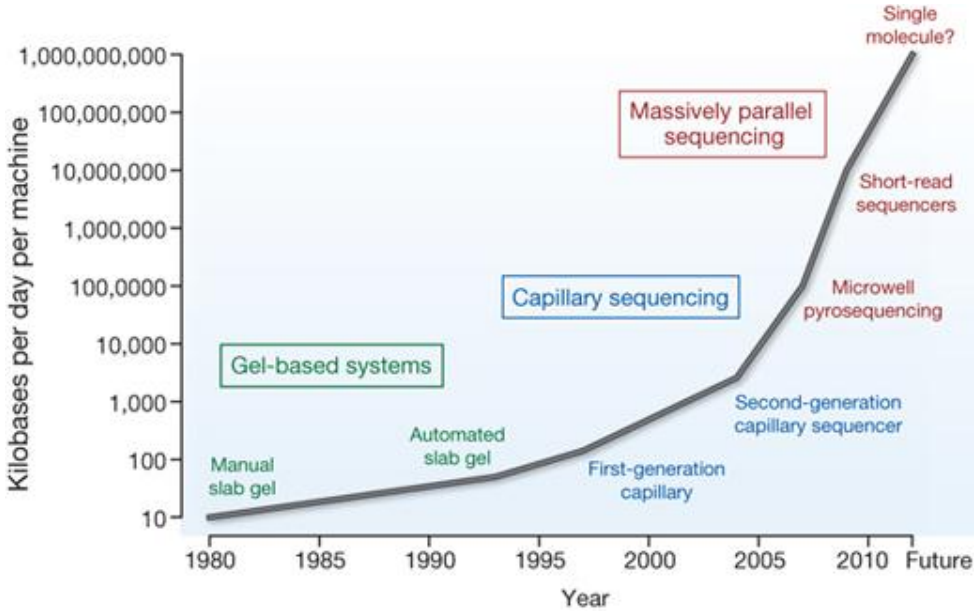
Qin *et al.*

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。

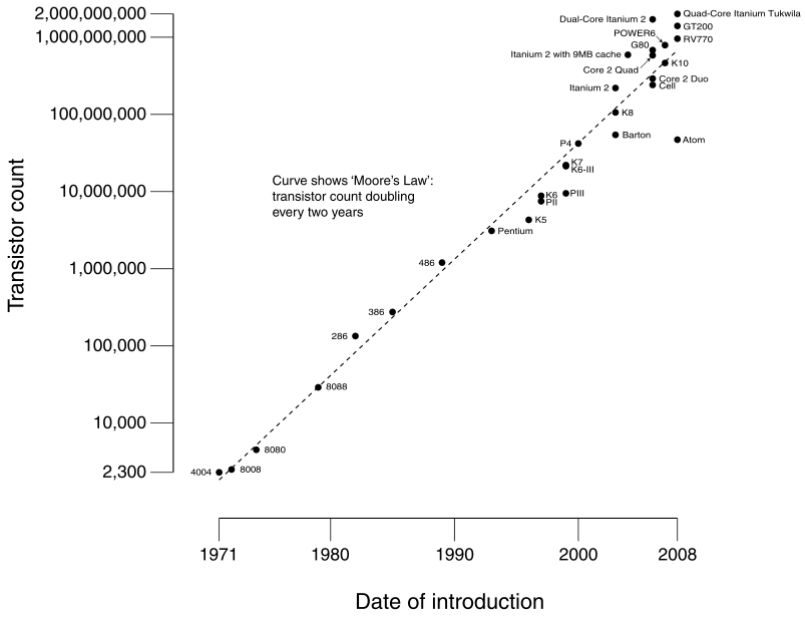
▲ Top

シーケンス技術の進歩

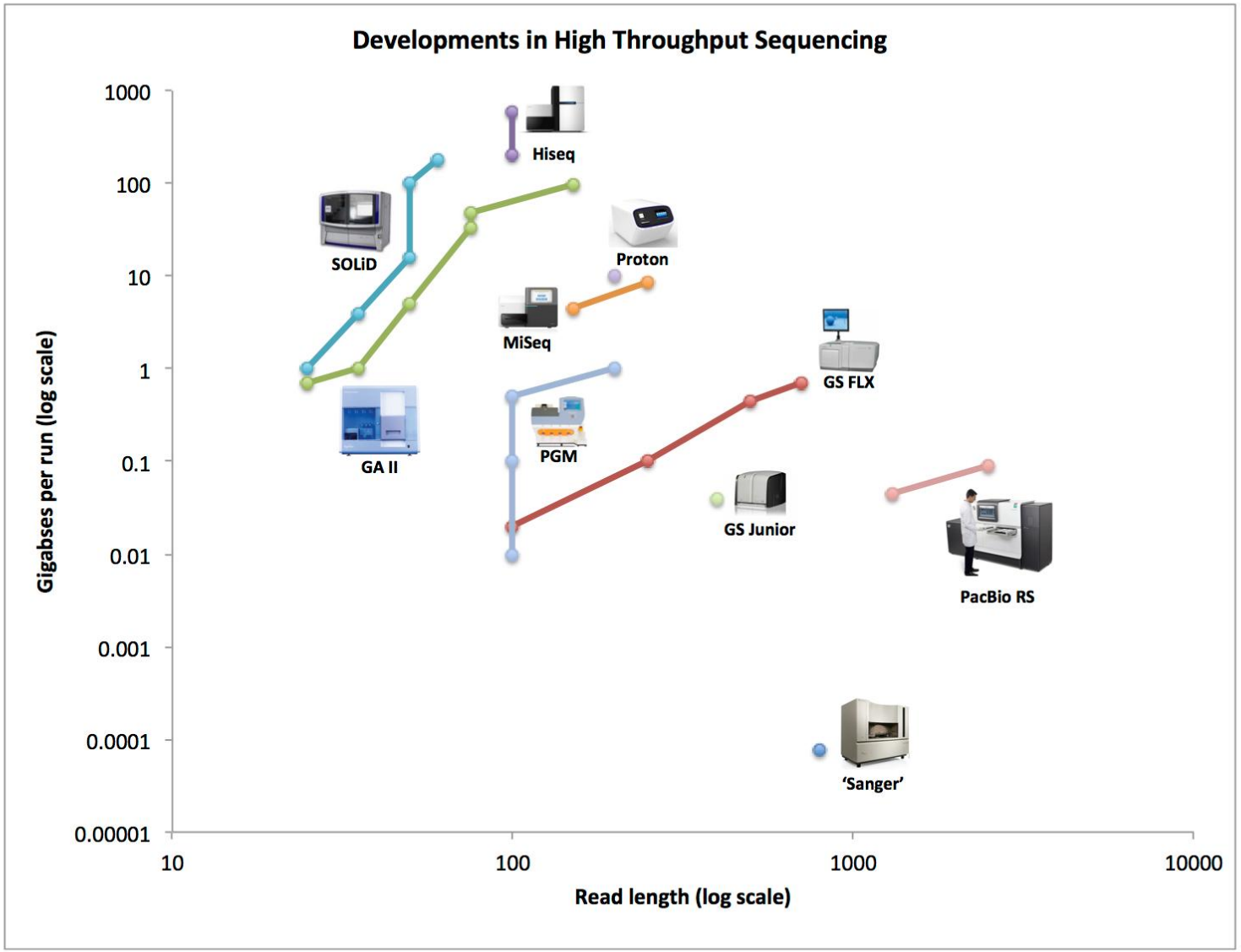
（＋コンピュータの性能の進歩）



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



http://dx.doi.org/10.6084/m9.figshare.100940

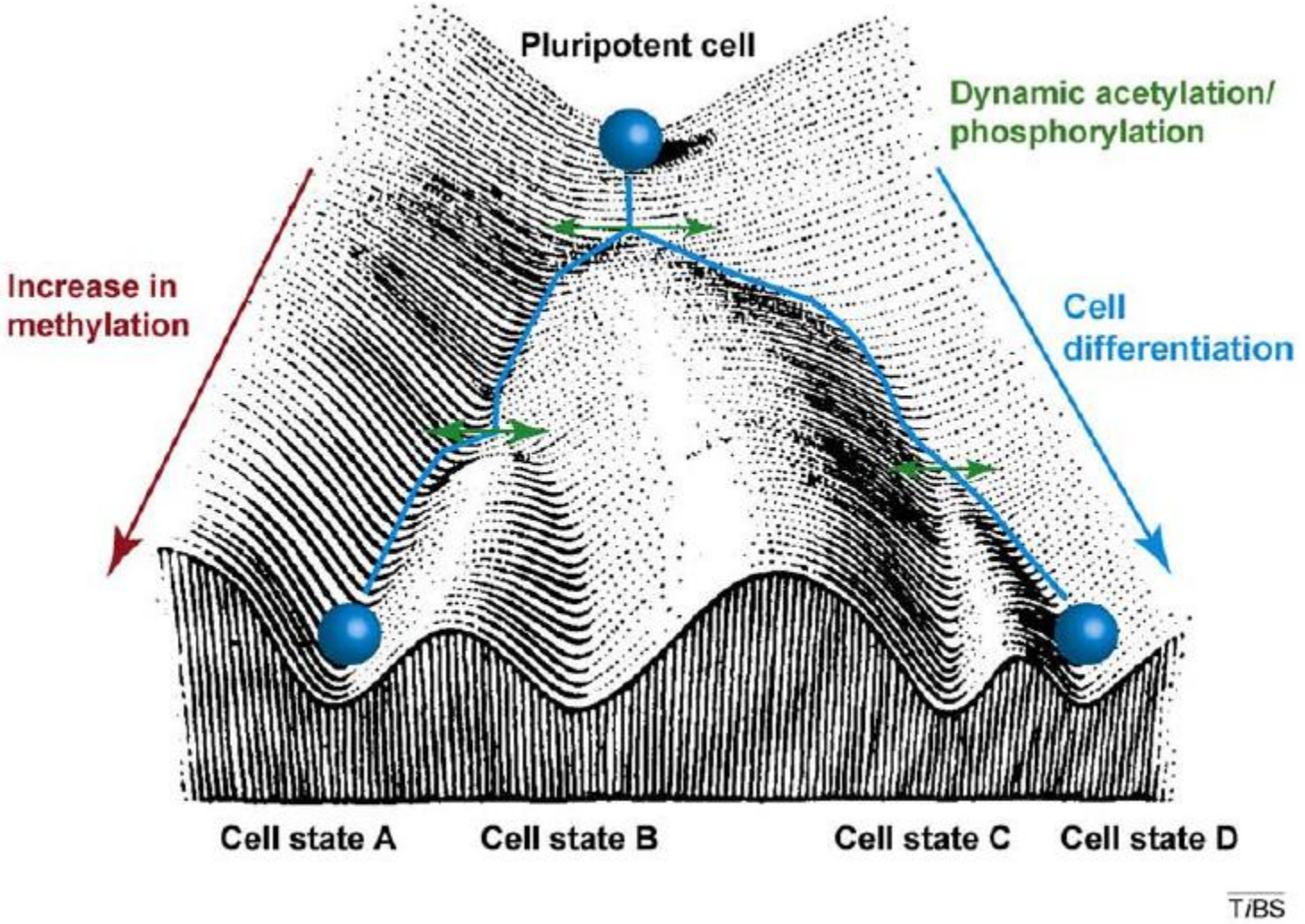
Illumina HiSeqのパフォーマンス

HiSeq System Performance Parameters

	HIGH OUTPUT RUN MODE*			RAPID RUN MODE*		
Read Length	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 × 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 × 50 bp Greater than 80% of bases above Q30 at 2 × 100 bp			Greater than 85% of bases above Q30 at 2 × 50 bp Greater than 80% of bases above Q30 at 2 × 100 bp Greater than 75% of bases above Q30 at 2 × 150 bp		

1レーンで30 Gbase

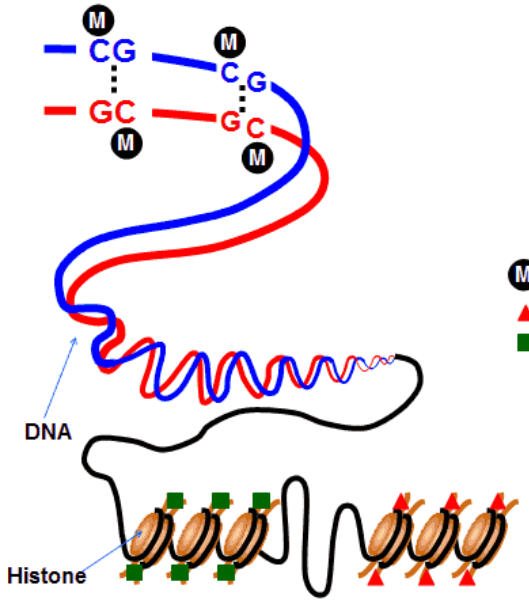
Waddiongton's Epigenetic Landscape

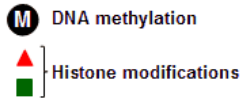


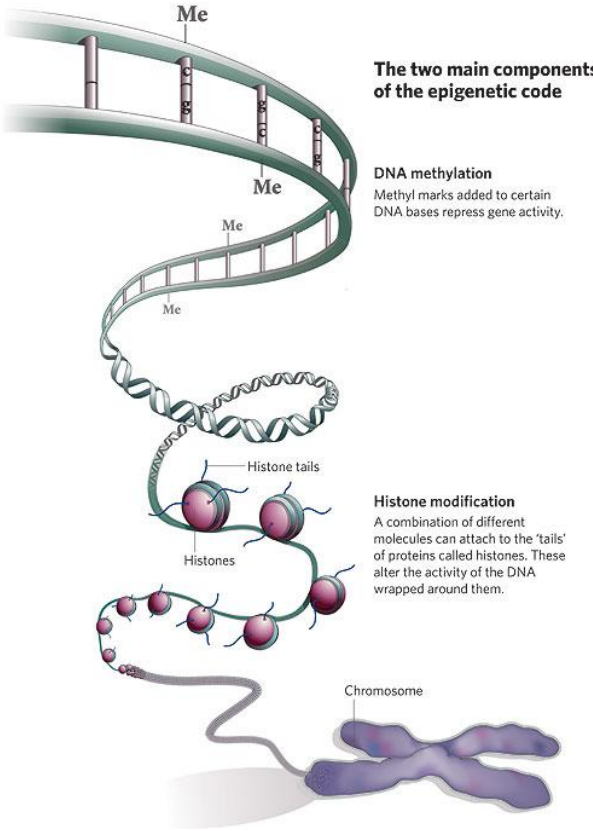
Barth and Imhof, 2010

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。

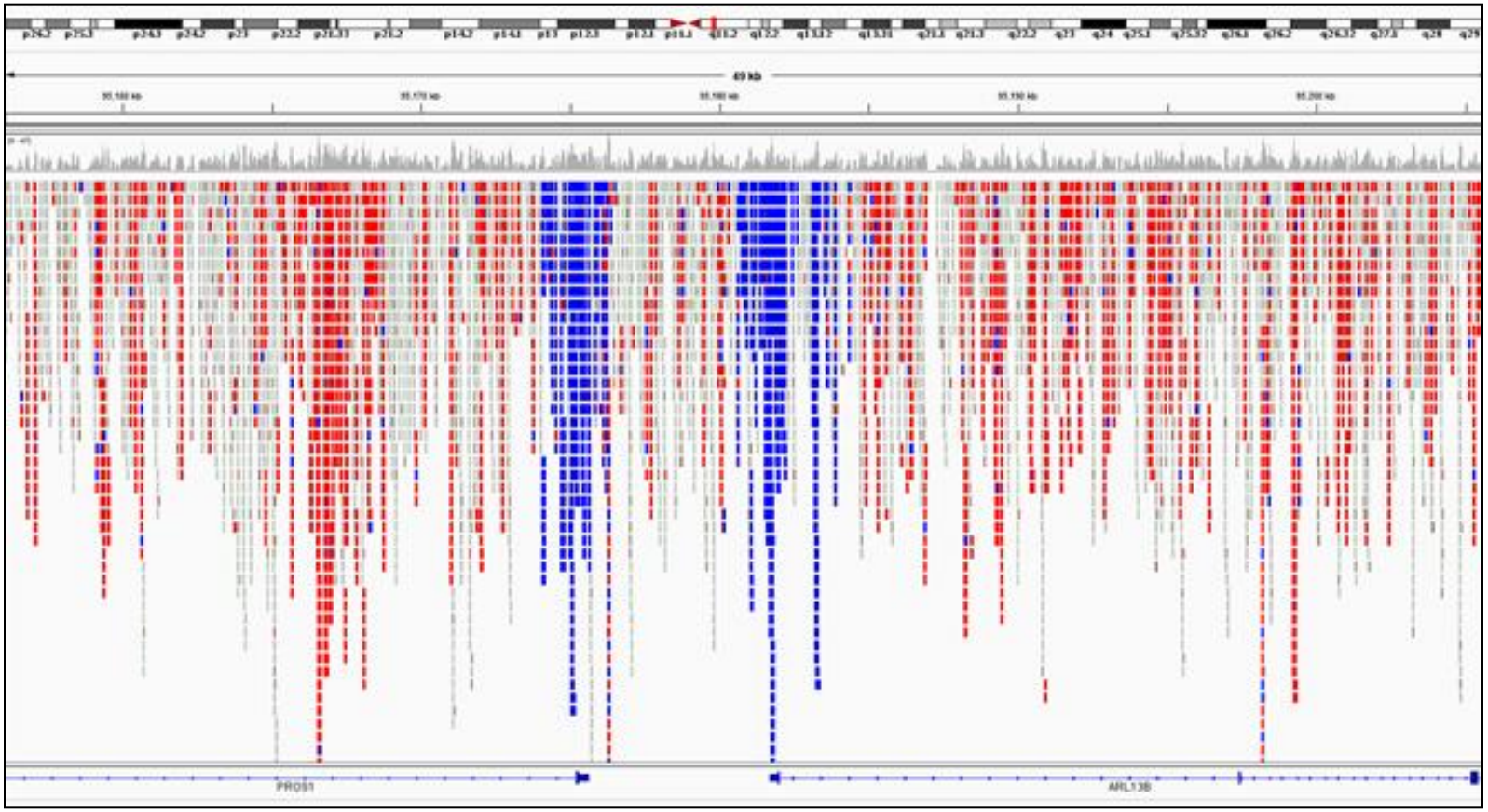
具体的には、たとえば、同一個体の各細胞は同じDNAを持っているが、エピジェネティックな違い（すなわちクロマチンの修飾の違い）が一因となって、組織特異的な遺伝子発現制御がなされているといえる。

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！

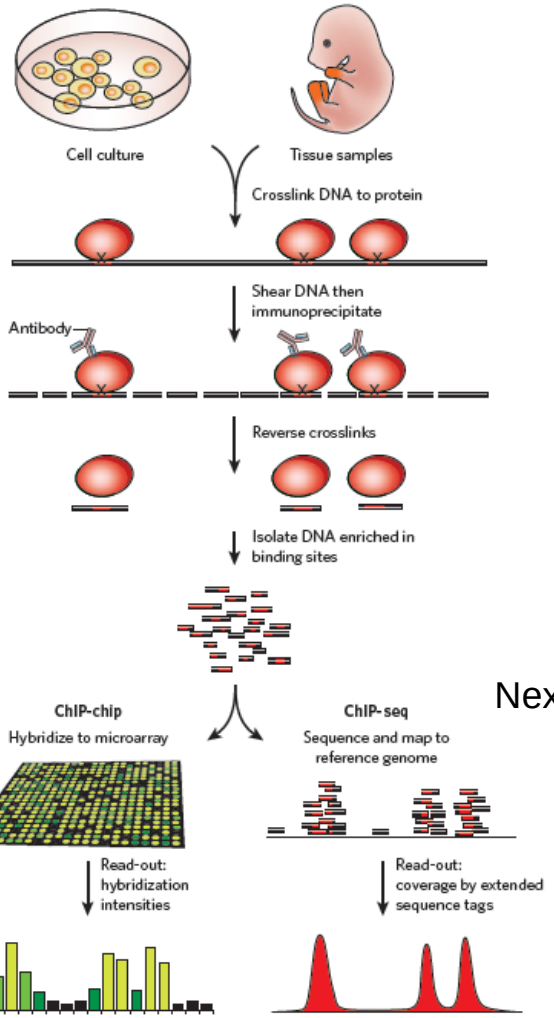
次世代シーケンサーの応用

- 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



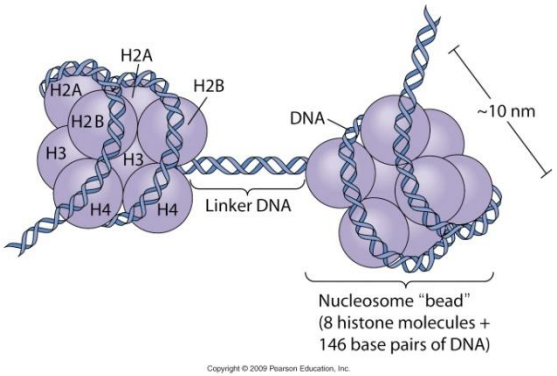
Microarray

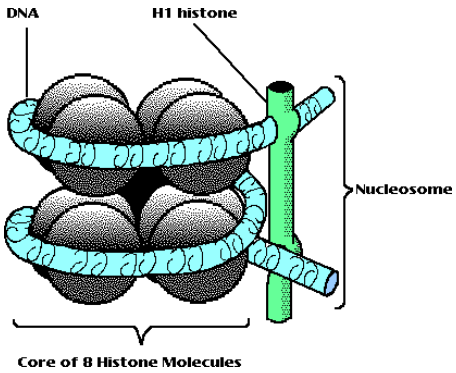
Next-generation sequencing technology

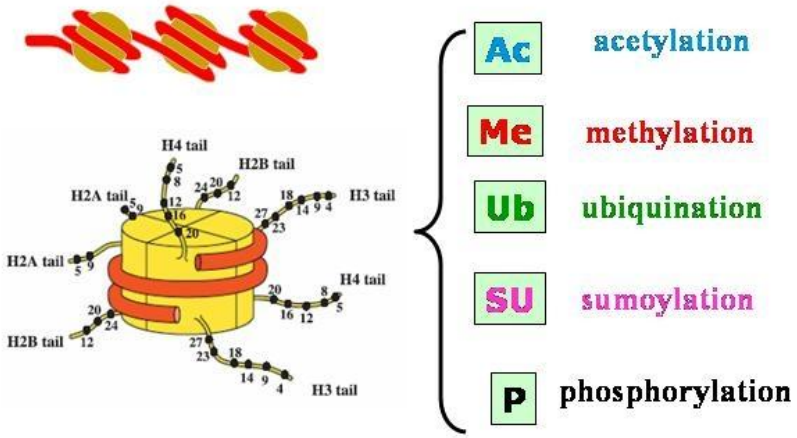
Visel *et al.* 2009

"*ChIP-chip*" and "*ChIP-seq*"

ヒストンとヌクレオソーム構造



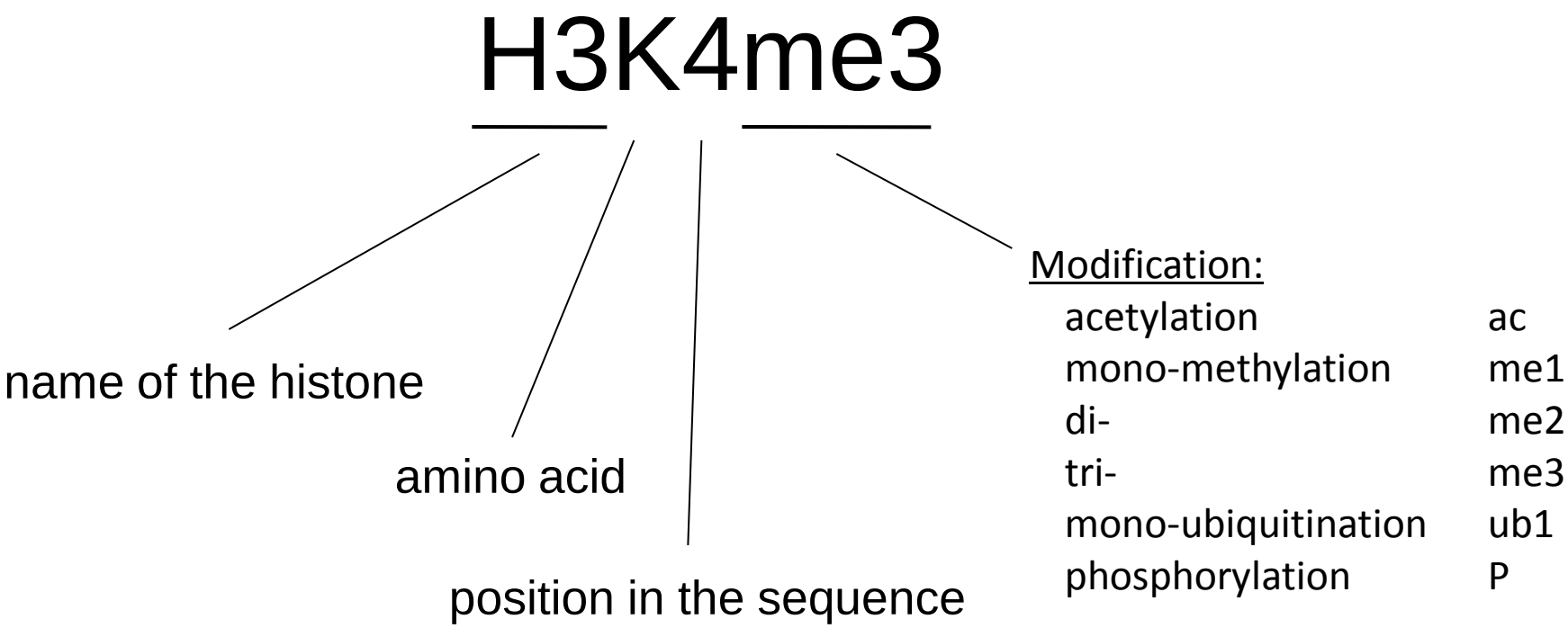




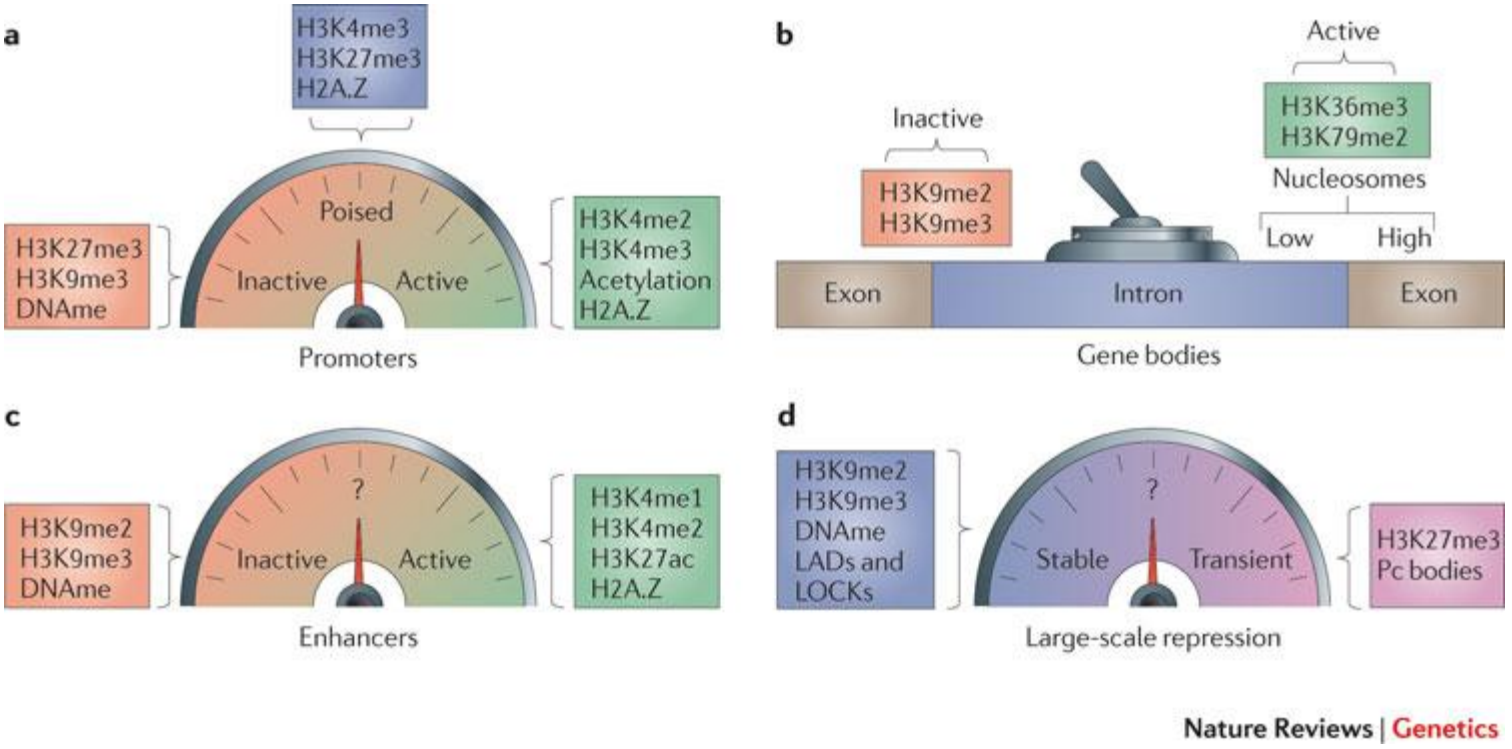
The figure illustrates nucleosome models and major posttranslational modifications which play essential roles in gene expression regulation and disease processes

- 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 ac
mono-methylation me1
di- me2
tri- me3
mono-ubiquitination ub1
phosphorylation P

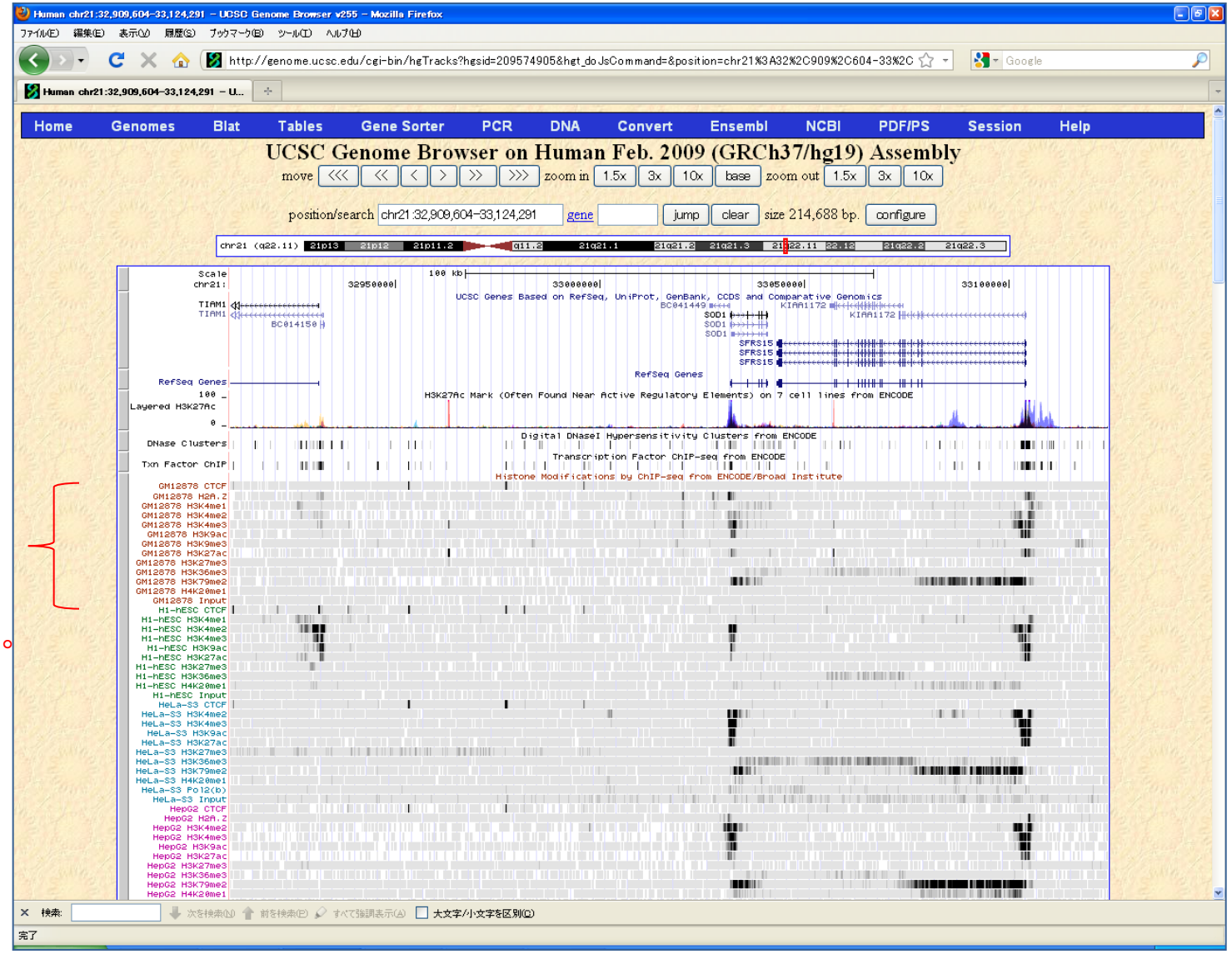
ヒストン修飾と生物学的機能の関連



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

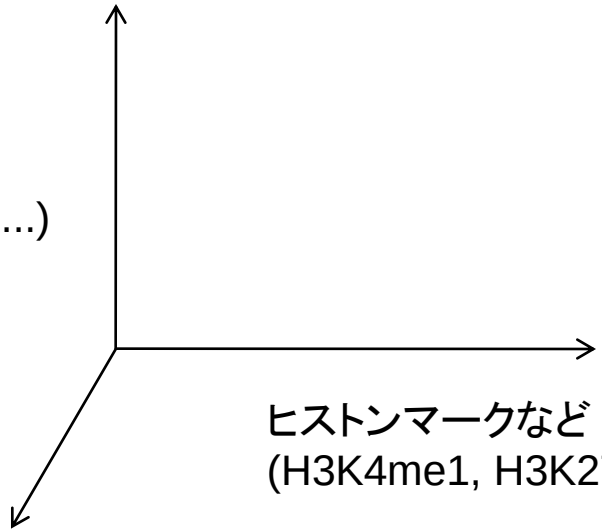
様々なcell-lineにおける様々なエピジェネティック・プロファイル

同一色のブロックが
各cell-line。
その中の各行が
様々なエピジェネ
ティック・プロファイル。



エピゲノムデータの多次元的な広がり

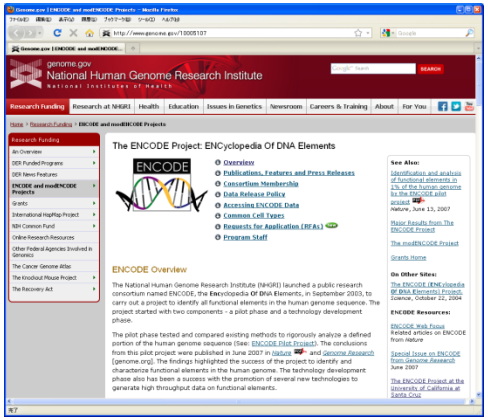
組織・発生段階・性差
(brain, liver, muscle, ...,
fetal/adult, male/female, ...)



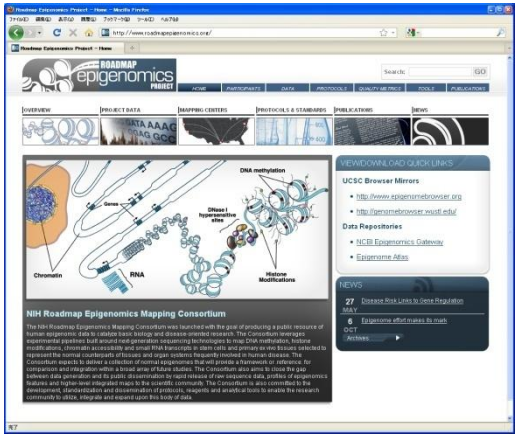
ヒストンマークなど
(H3K4me1, H3K27Ac, DNaseI, ...)

生物種

- ENCODE
  mouseENCODE,
  modENCODE (for drosophila, worm)



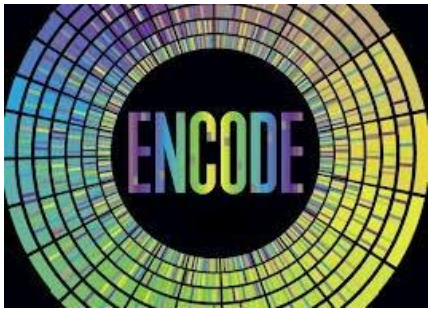
- Roadmap Epigenomics Project



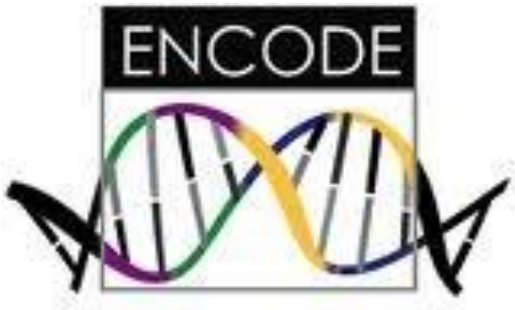
- International Human Epigenome Consortium (IHEC)

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.

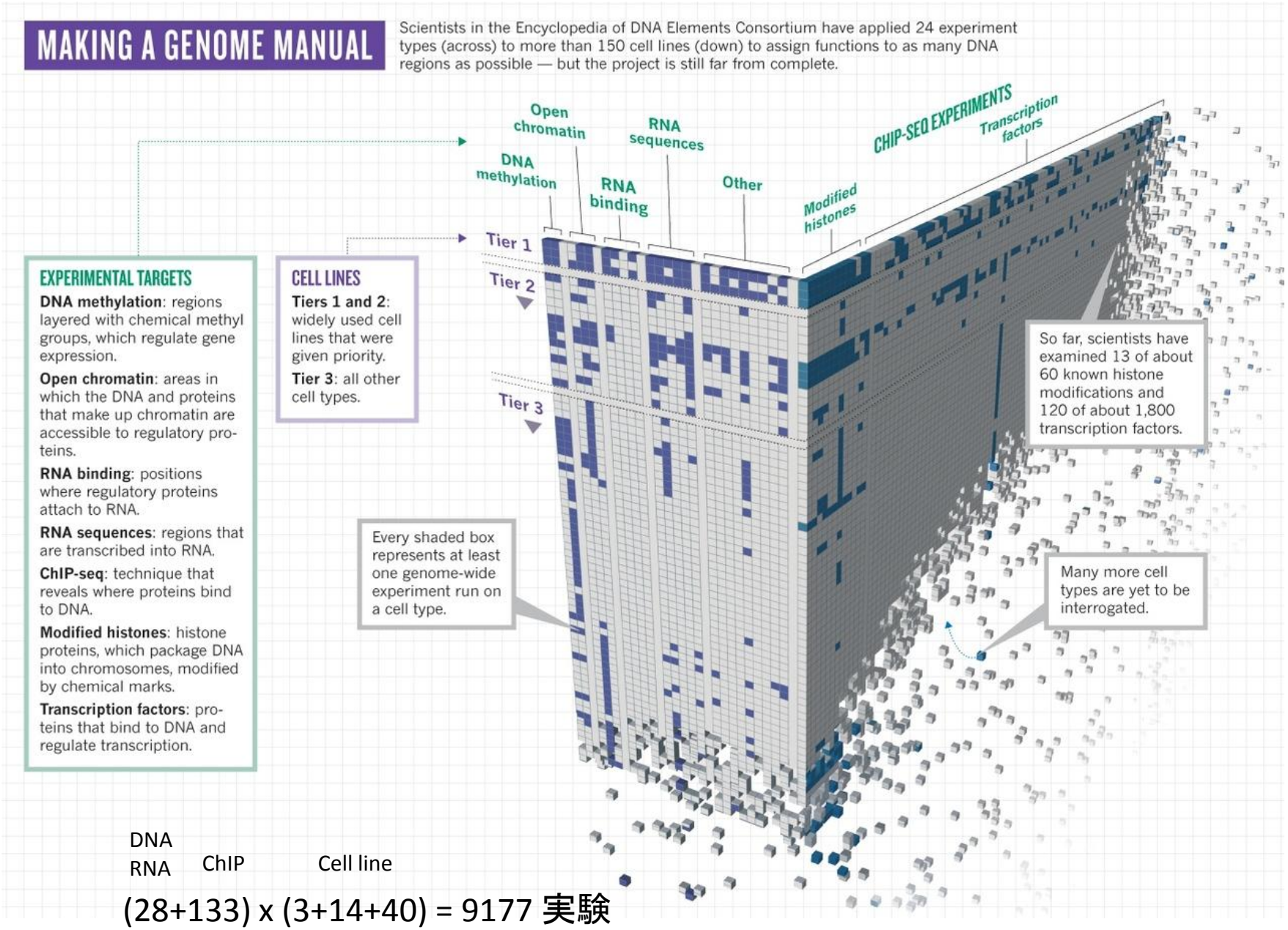
Pilot projectは2007に終了

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, ...



DNA
RNA ChIP Cell line

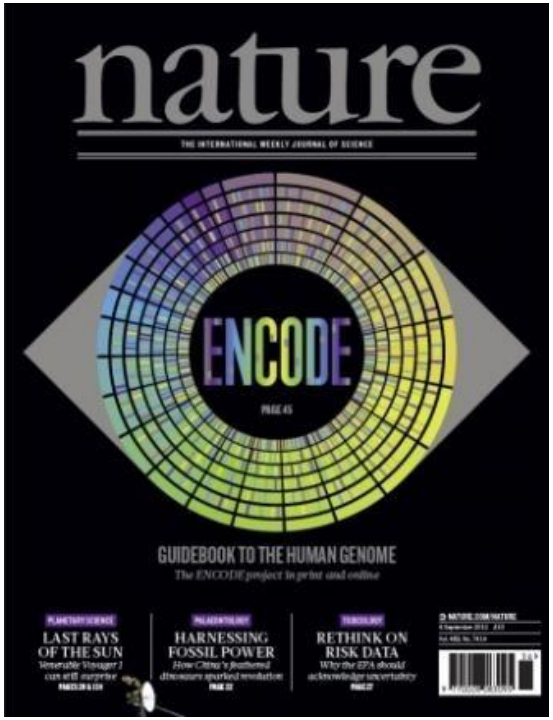
(28+133) x (3+14+40) = 9177 実験

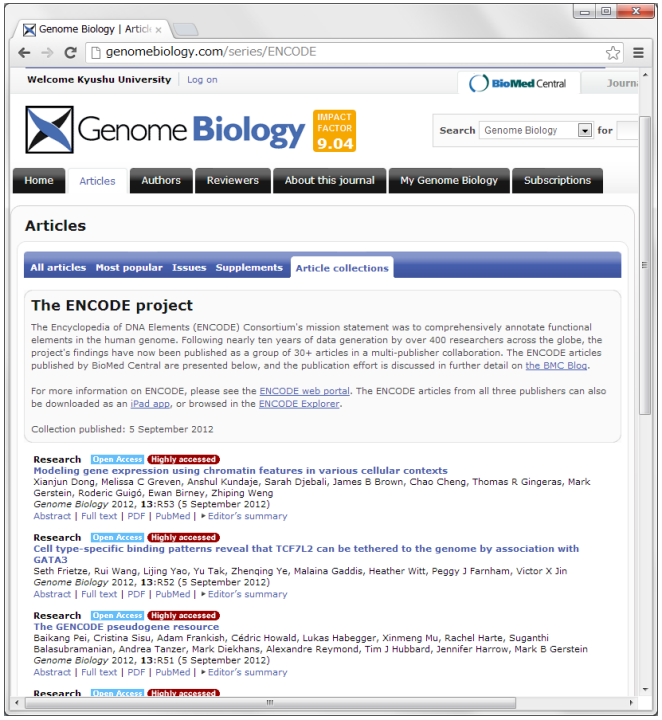
Maher, *Nature* 2012

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

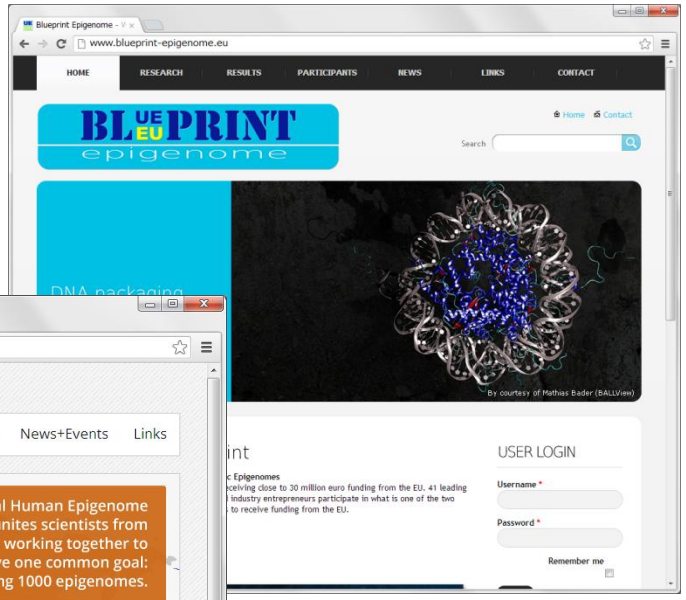




エピゲノム・プロジェクト

Roadmap Epigenomics

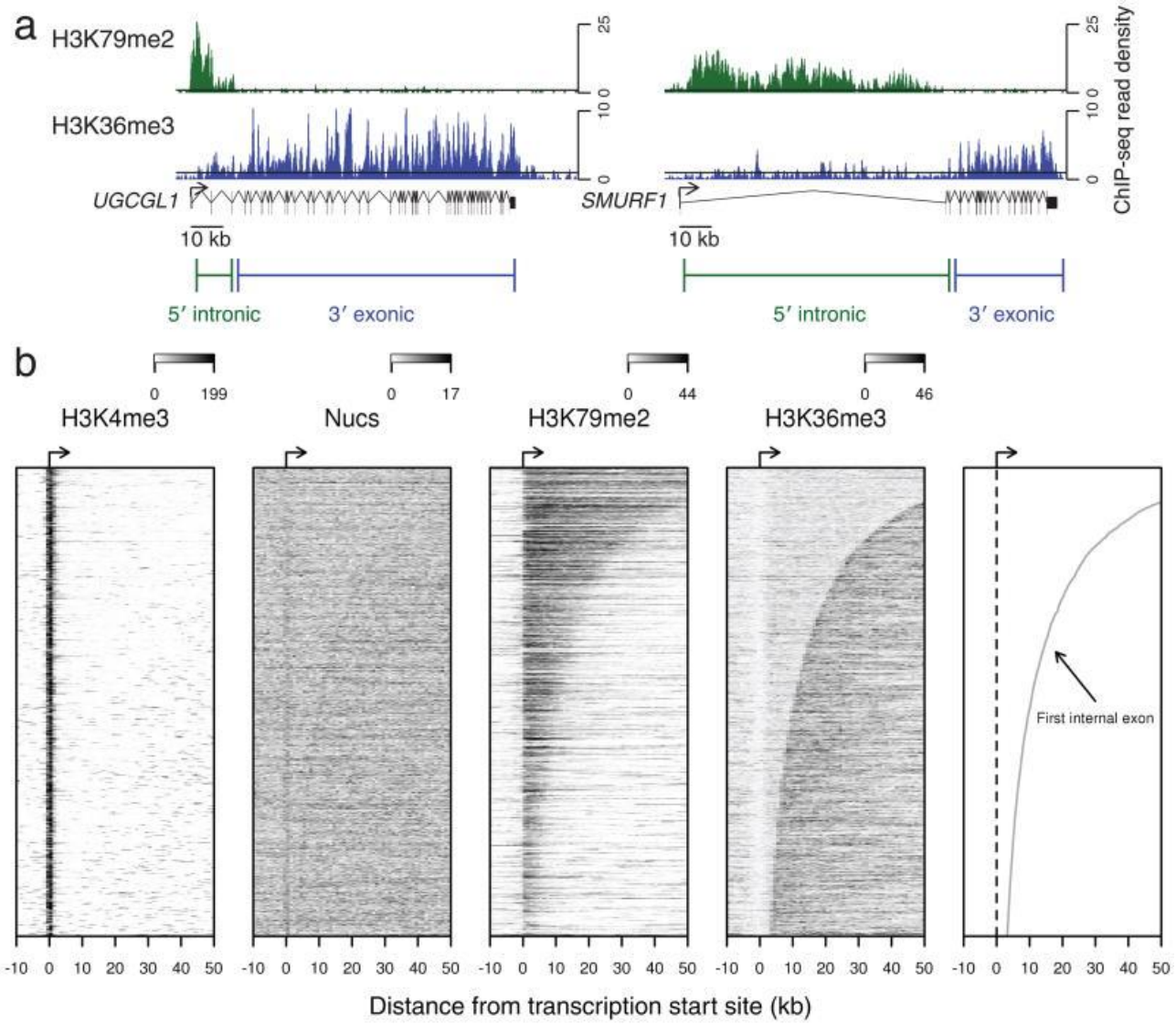
Blueprint Epigenomics



IHEC: International Human Epigenome Consortium

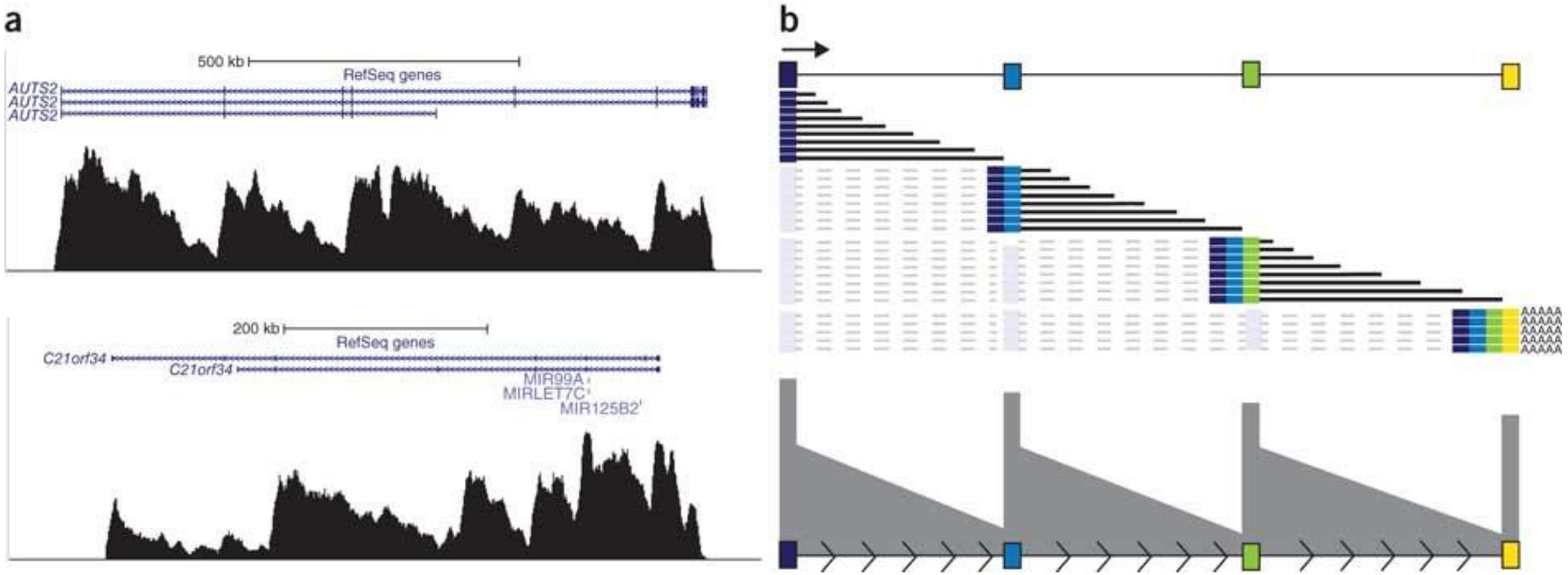
次世代シーケンサーが可能にした
新しいゲノミクス研究

イントロン/エクソン構造がゲノム中ですでにマークされている



Huff *et al. Nat. Struct. Mol. Biol*. 2010.

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

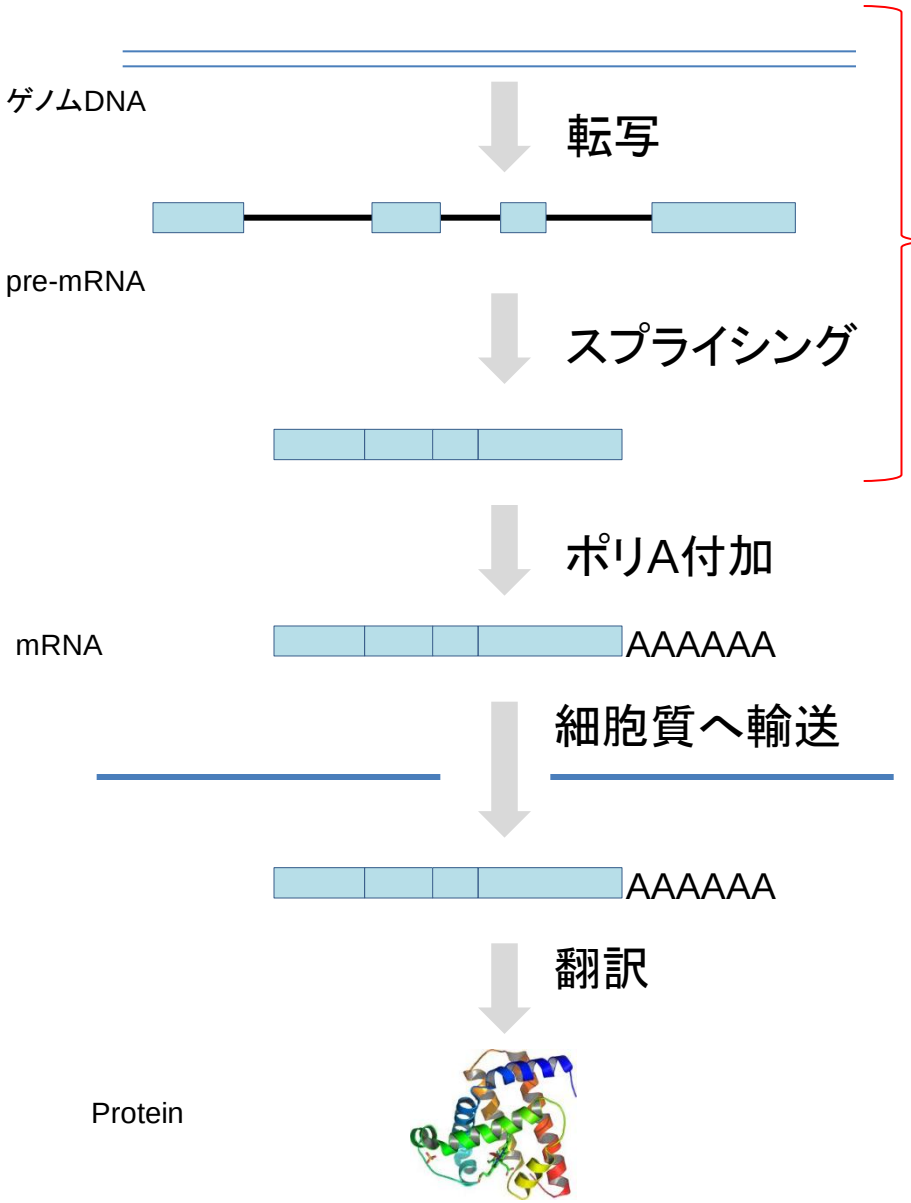


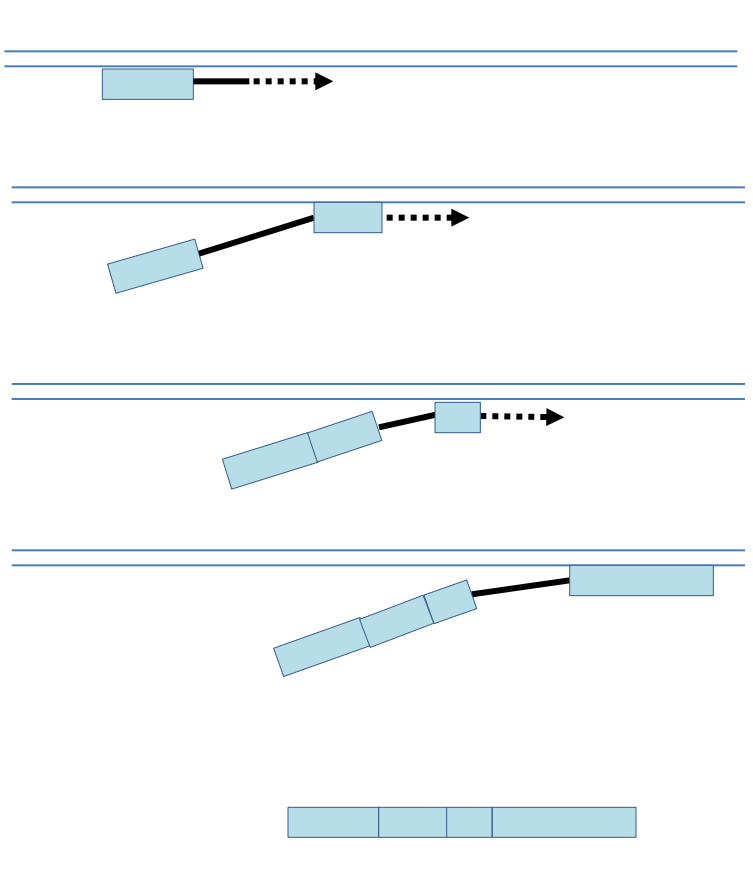
Ameur *et al. Nat. Struct. Mol. Biol*. 2011.

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

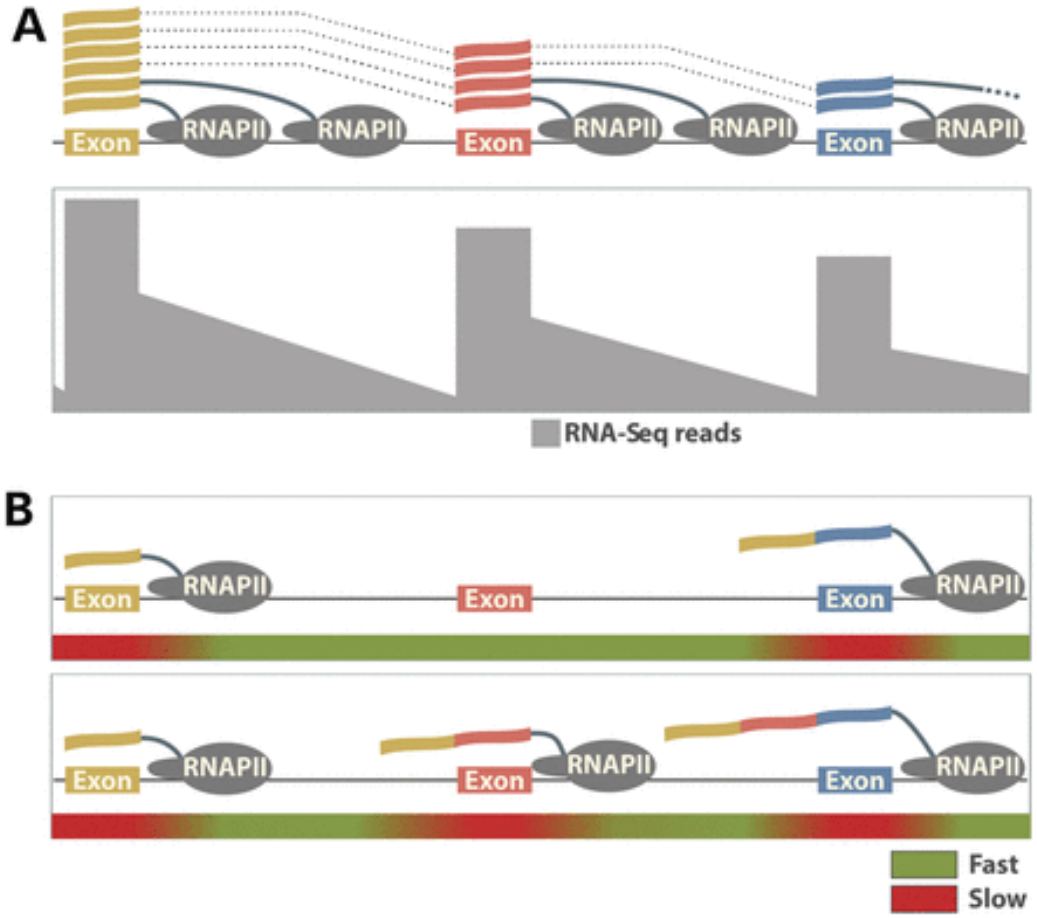
でも実際は・・・





co-transcriptional splicing

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



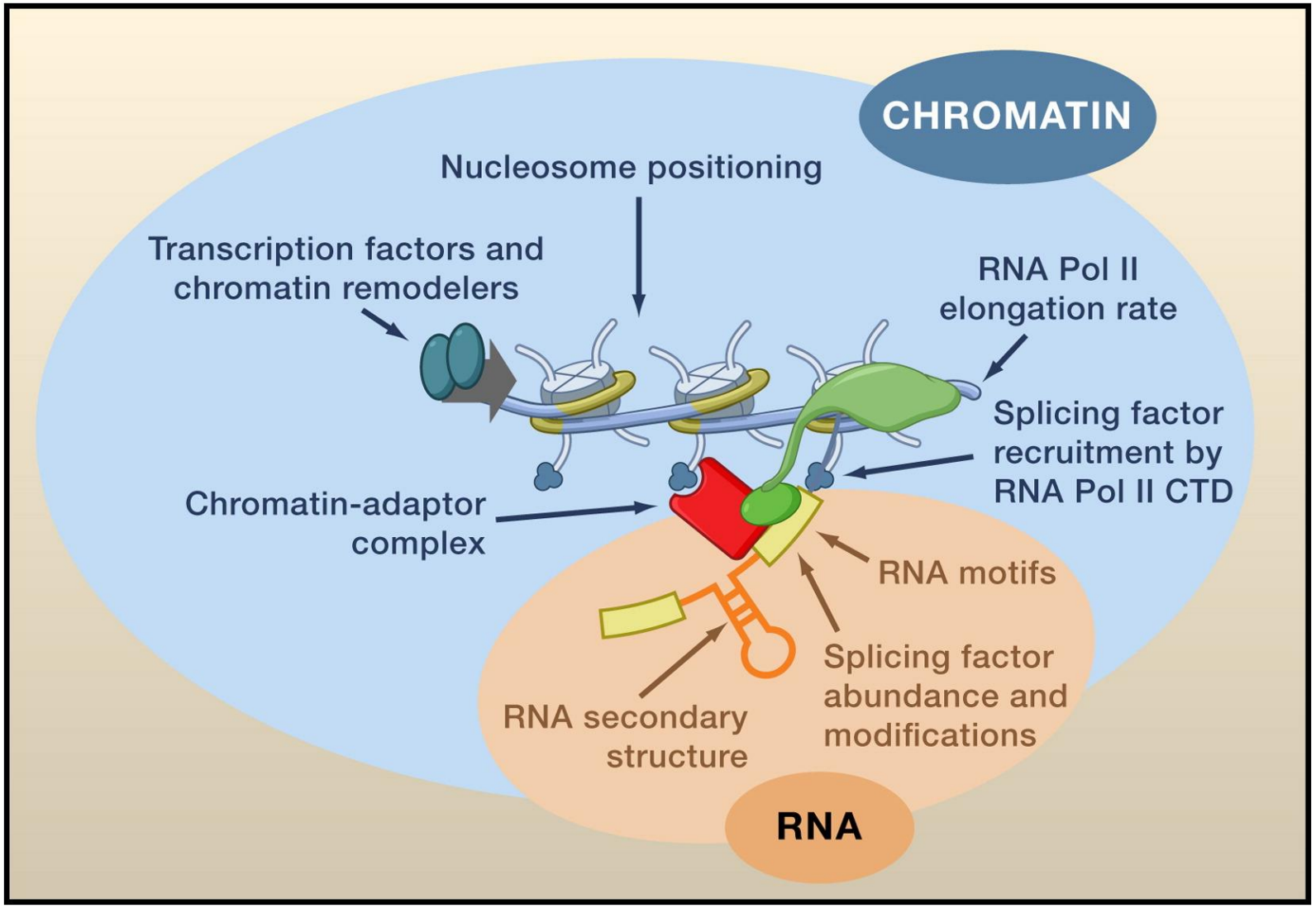
Brown *et al. Hum. Mol. Genet*. 2012.

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)。

より実際的な「転写」と「スプライシング」の関係

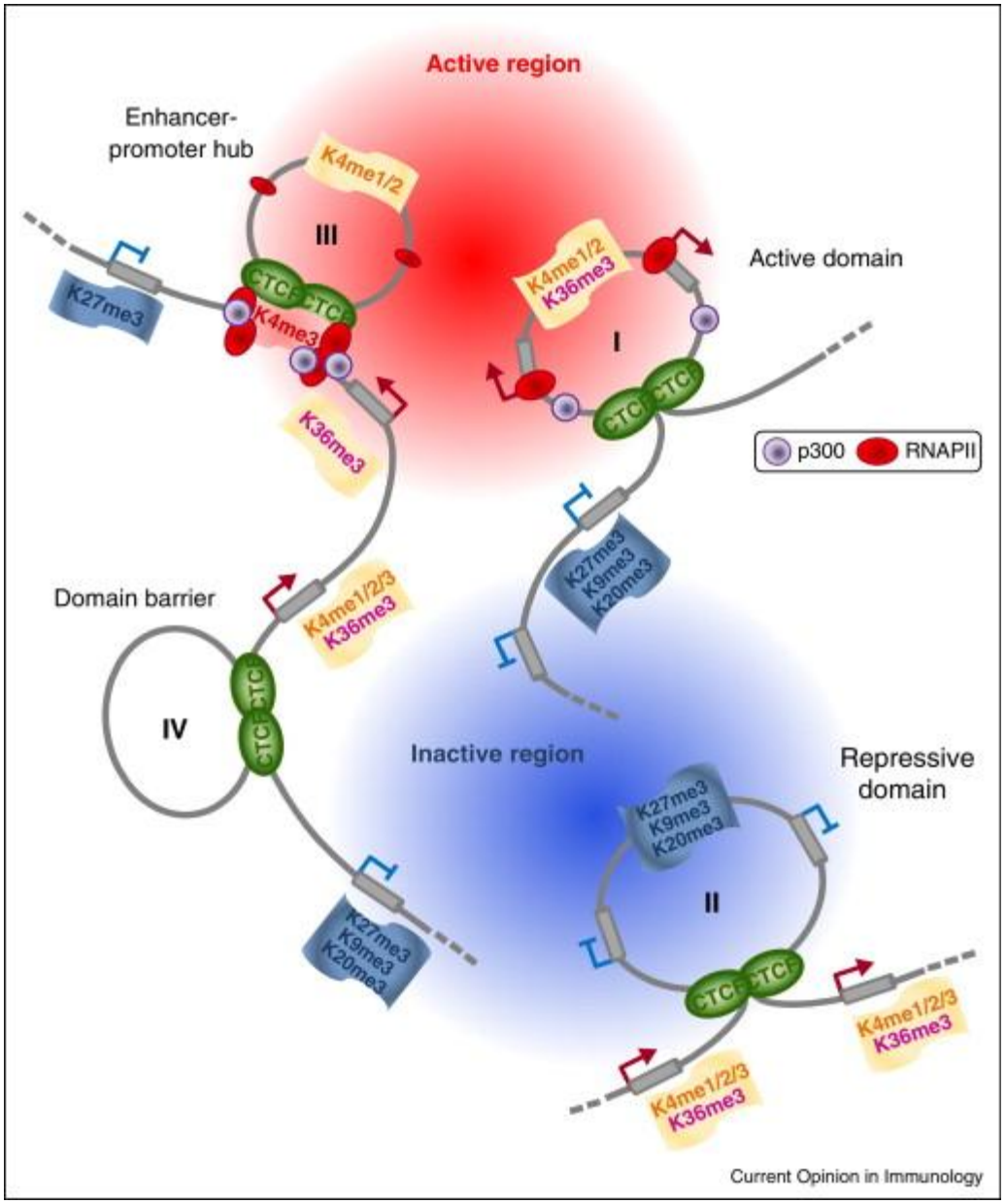
スプライシングにおける転写機構やクロマチン状態の関与



Luco *et al. Cell*, 2011

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。

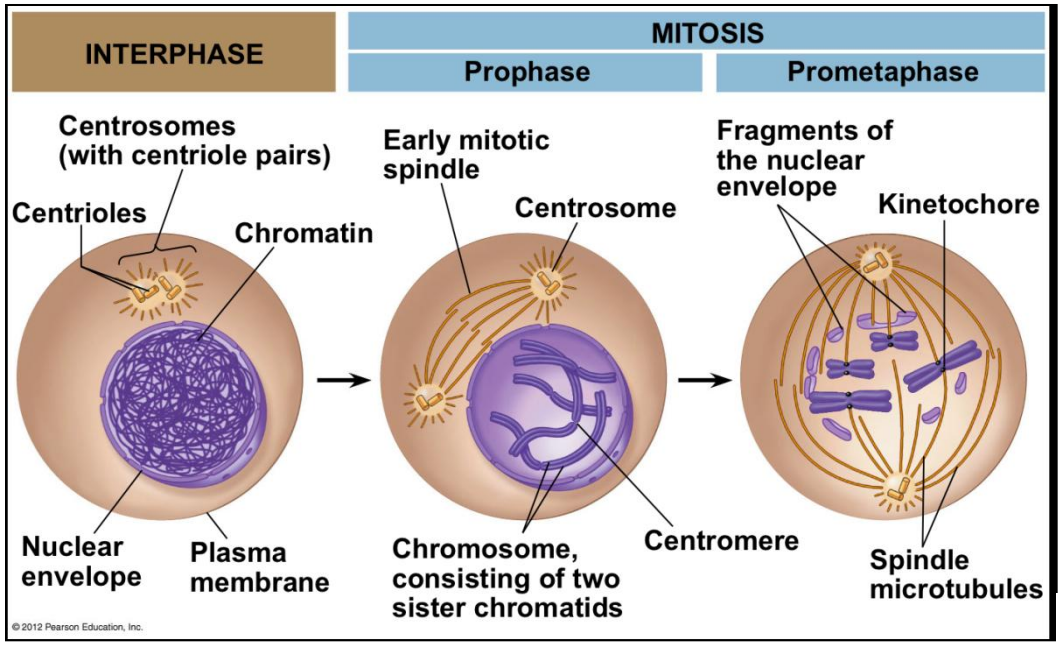
インシュレーター(insulator)によって規定されるゲノムの[活性/不活性]領域

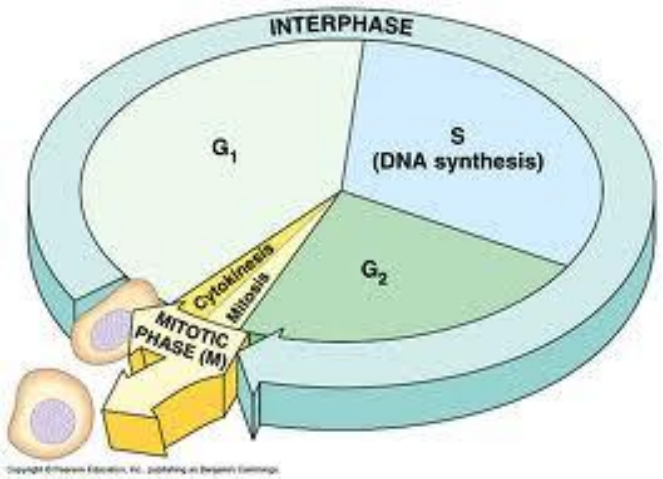


"Transcription factories"

CTCF： インシュレーターとして働くタンパク質

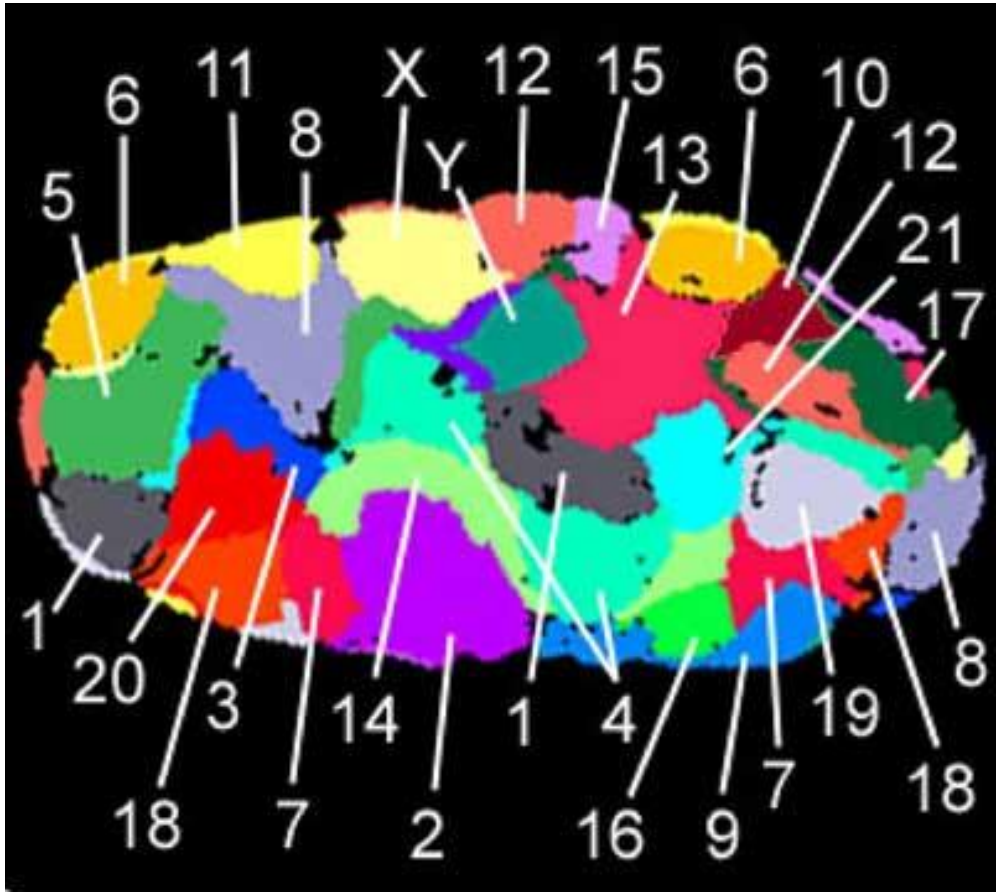
染色体は普段どんな形をしているか





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)

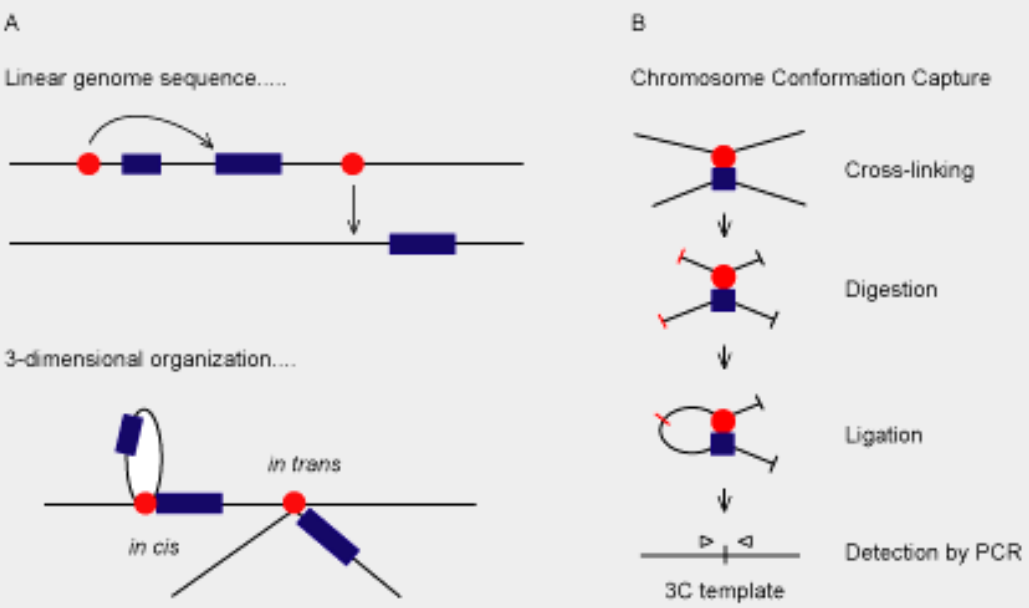
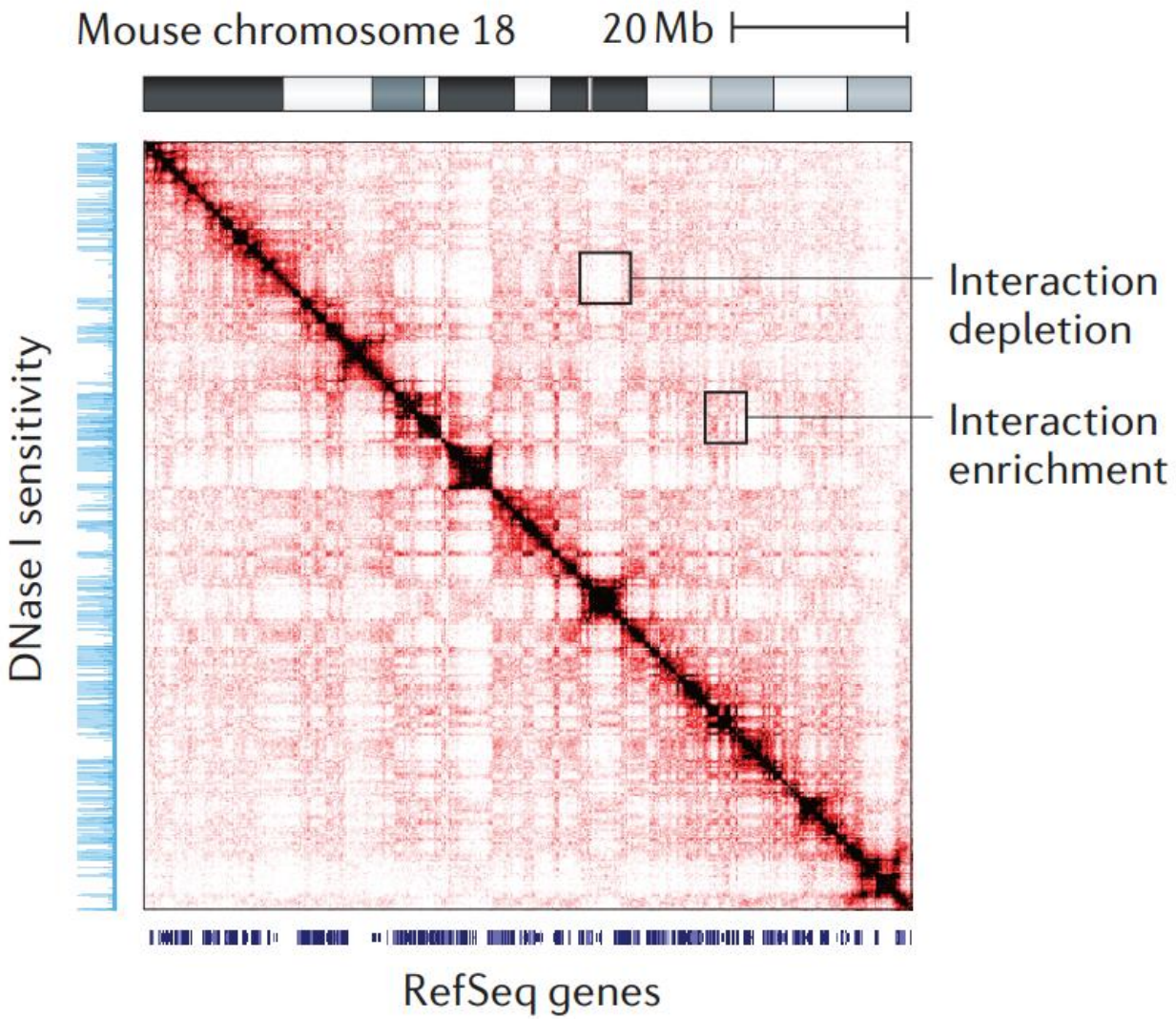


Figure 1. (a) Genes (blue rectangles) and regulatory elements (red circles) are linearly organized along chromosomes (top), but as a result of specific interactions between elements (indicated by arrows, both *in cis* and *in trans*) a complex three-dimensional network is formed inside the cell (bottom). (b) Schematic representation of the 3C assay. Chromatin is cross-linked, digested with a restriction enzyme and then ligated. Specific ligation products can be detected by PCR. [Figure from Dekker (2006), *Nature Methods*, 3(1): 17-21].

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。

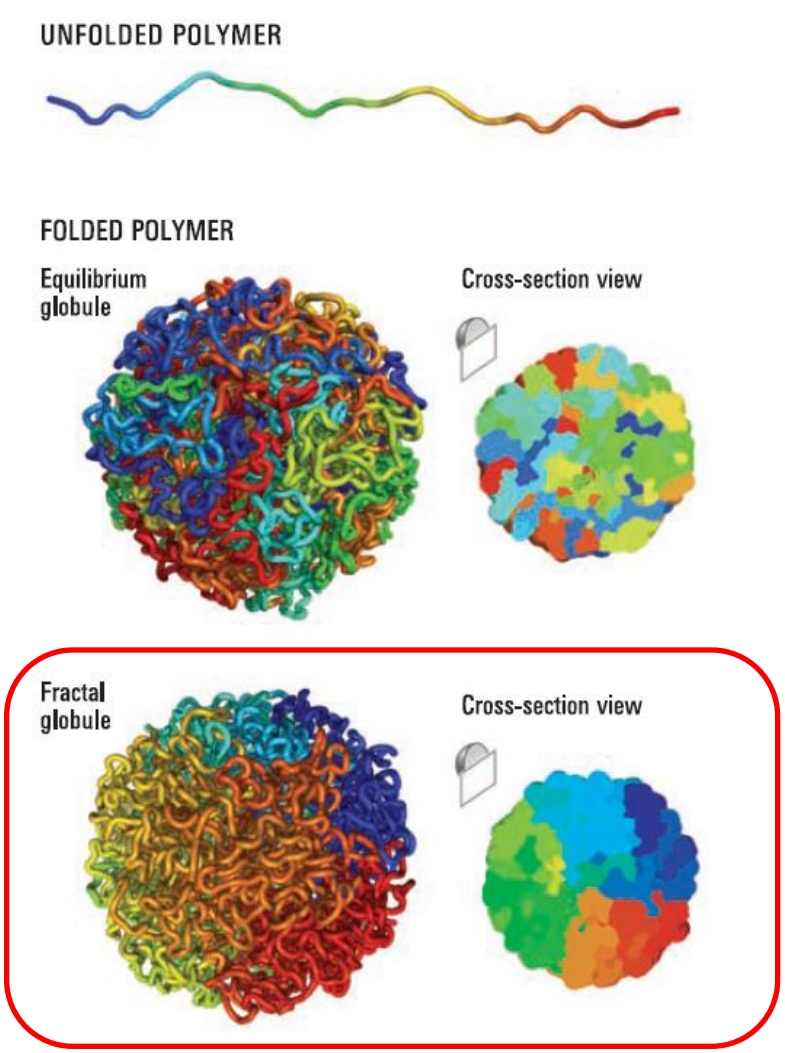
Hi-C法による染色体のinteraction map

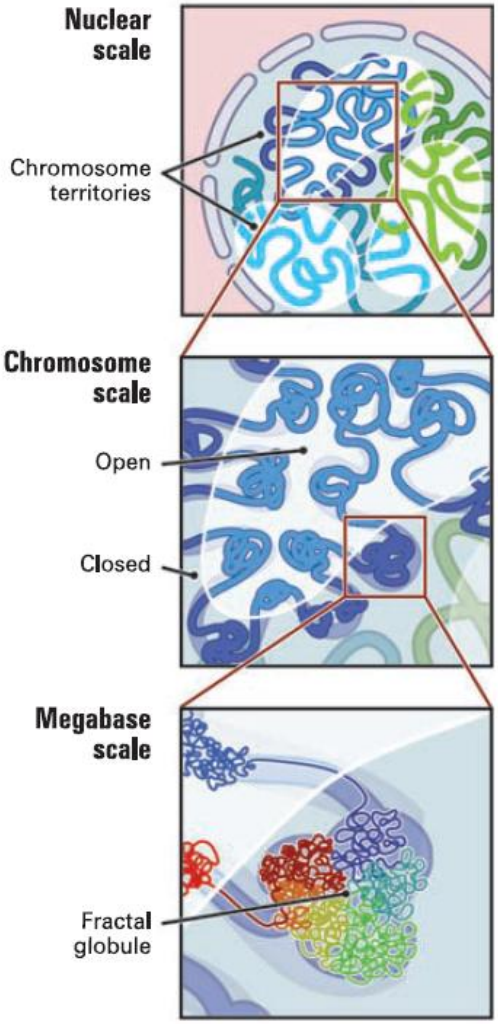


Dekker *et al., Nat. Rev. Genet*., 2013.

Hi-Cのデータから得られた染色体の立体構造モデル

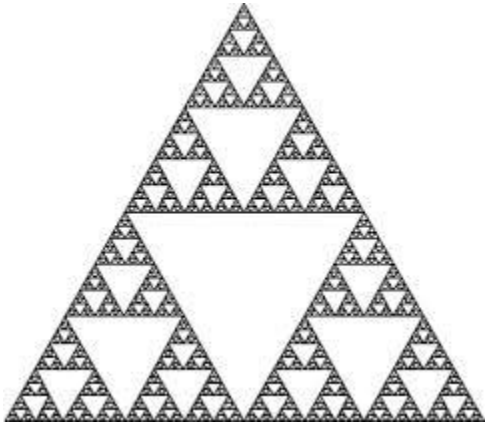
"Fractal globule"

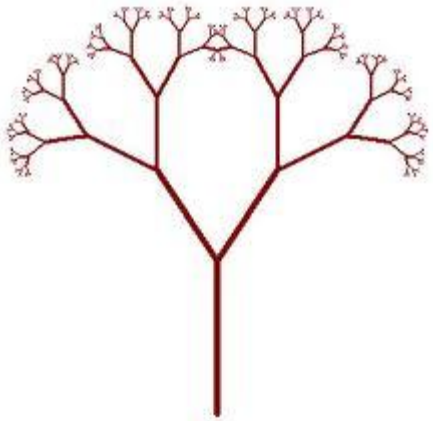




Lieberman-Aiden *et al.* (2011) *Science*

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.

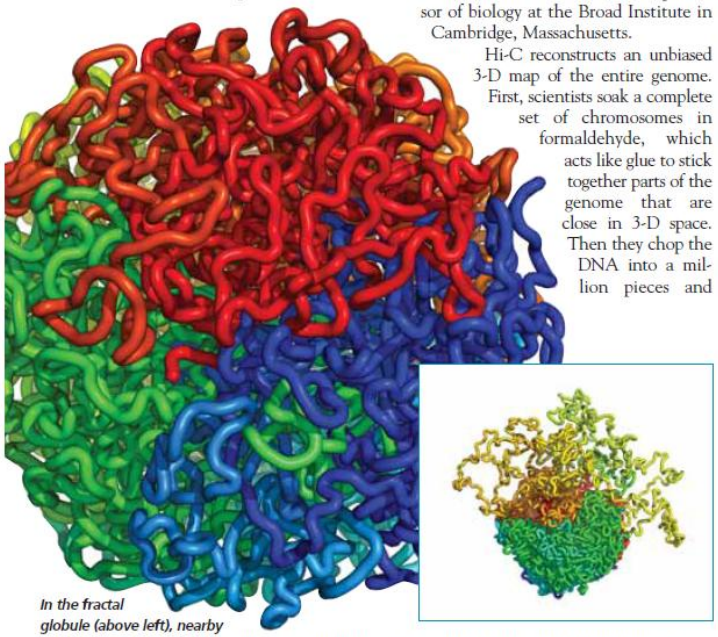
The team found that active, gene-rich and inactive, gene-poor sections cluster in separate parts of the nucleus. The active chromatin segments are like easily accessible papers spread out across a desk, whereas the inactive portions are densely packed, like folders in a file cabinet.

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



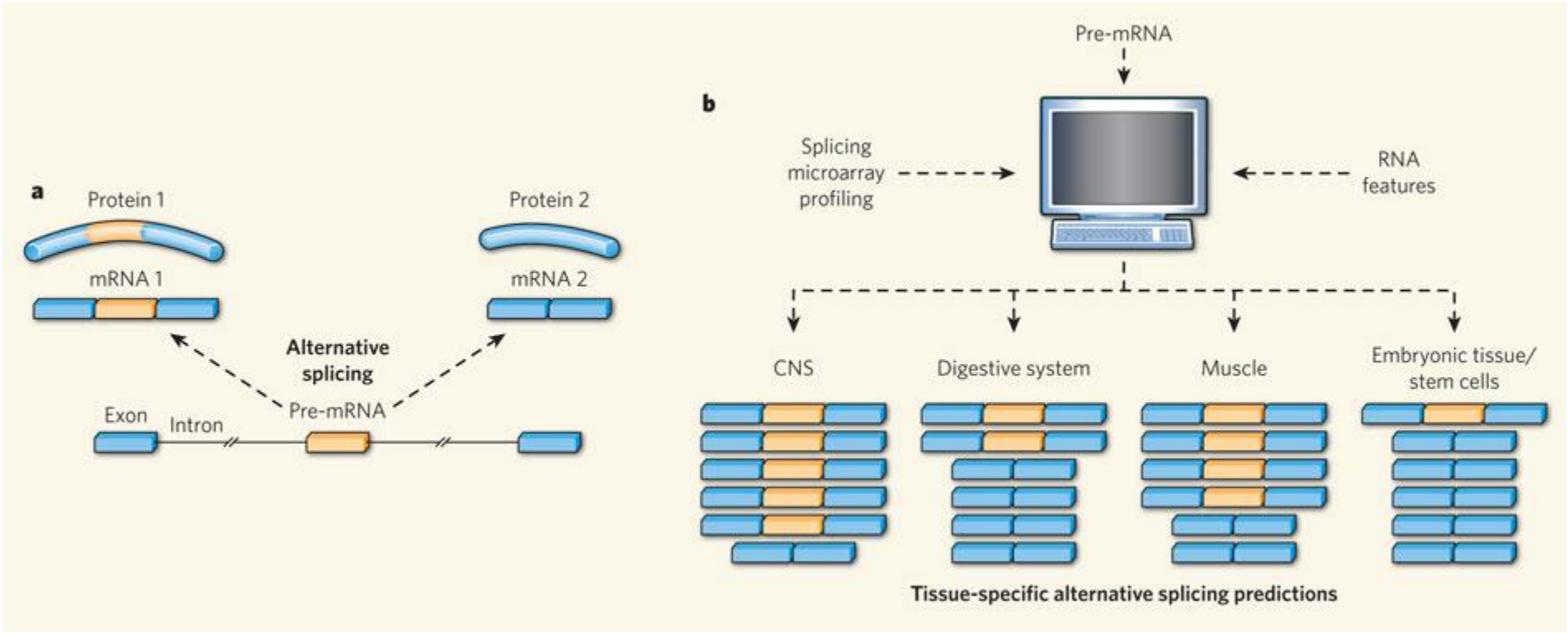
In the fractal globule (above left), nearby regions on a chain of DNA—indicated using similar colors—are packed into nearby regions in 3D space. The accessible DNA chain unravels easily (above right) because the globule lacks knots. Images courtesy of Leonid A. Mirny and Maxim Imakaev, reprinted from 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.

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-

スプライシングコードの解明

Breaking the second genetic code

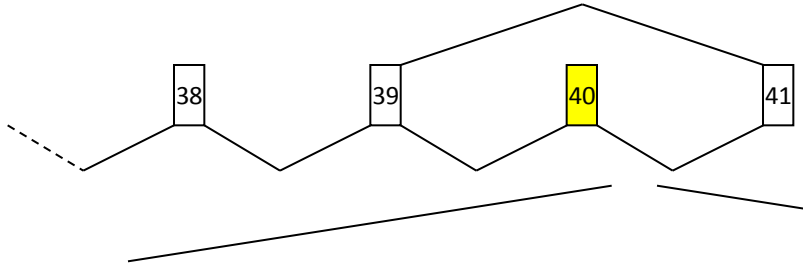


Ramón Tejedor and Valcárcel, *Nature,* 2010

(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)

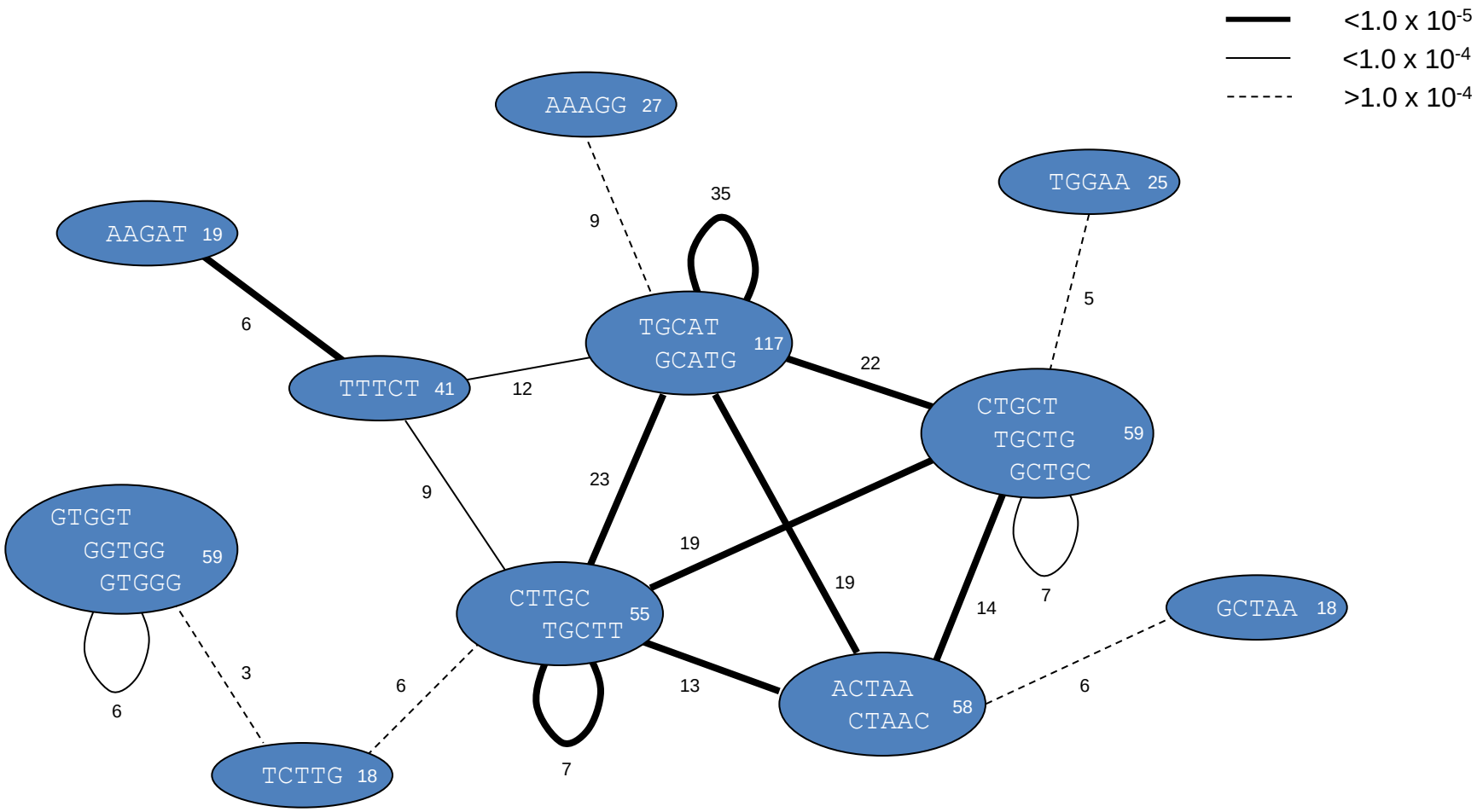



```
human       chr14             -   35087341  ACATTTCAGAAATTGTCACTAAATTTTTTCC--AGTATTA--TACTGACTAACCC-AGGTCTGCATGAAACACTAACA-T
chimpanzee  chr15             -   34251985  ACATTTCAGAAATTGTCACTAATTTTTTTCC--AGTATTA--TACTGACTAACCC-AGGTCTGCATGAAACACTAACA-T
macaque     chr7              -   98506834  ACATTTCAGAAATTGTCACTAAATTTTTTCC--AGTATTA--TACTGACTAACCC-AGGTCTGCATGAAACACTAACA-T
rat         chr6              -   75858765  ACATTTCAAAAATTATCACTAAATTTTTTCCCCAGAATTG--TACTAACTAACCC-AGGTCTGCATGAAACACTAACA-T
mouse       chr12             -   56530084  ACATTTCAAAAATTATCACTAAATTGTTCCCCGAGAACTG--TGCTAACTAACCC-AGGTCTGCATGAAACACTAACC-C
rabbit      scaffold_178393   -      17945  ACATGGCAGAAATTGTCACTACATTTTTTCC--AGATTTA--TACTAACCAACCC-AGGTCTGCATGAAACACTAACA-T
cow         chr21             -   30280498  ACATTTCAGAAATTGTCACTAAATTTCTTCC--AGAATTC--TACTTACTAACCT-AGGTCTGCATGAAACACTAACATT
dog         chr8              -   17234636  ACATTTCAGAAATTGTCACTAAATTTCTTCC--GGAATTA--TACTTACTAACCC-AGGTCTGCATGAAACACTAACACT
armadillo   scaffold_3577     -      24577  ACATTTCAGAAATTGTC-CTAAATT-CTTCC--AAAATTG--TTCTTACTAACAC-AGGTCTGCATGAAACACTAACACT
tenrec      scaffold_299940   +       3914  ATATTTCAGAAATTGTCACTAAATTTTTTTC---CAGTTA--TACTTACTAACCC-AGGTCTGCATGAAACACTAACA-C
opossum     chr1              -  286030148  ACATTTCAGAAGTTTTTACTAAATTTTTTCC--AAAGTTAGTTTTTTACTAACCCCAGGTCTGCATGAAA-ACTAACA-C
                                             * **  ** ** ** *  ***  **  *  *       *   *  * ** ***   ************** ******
```

2つのシス因子が一緒にあらわれる（共起）

Suyama *et al. Nucleic Acids Res.*, 2010

A network of co-occurring motifs

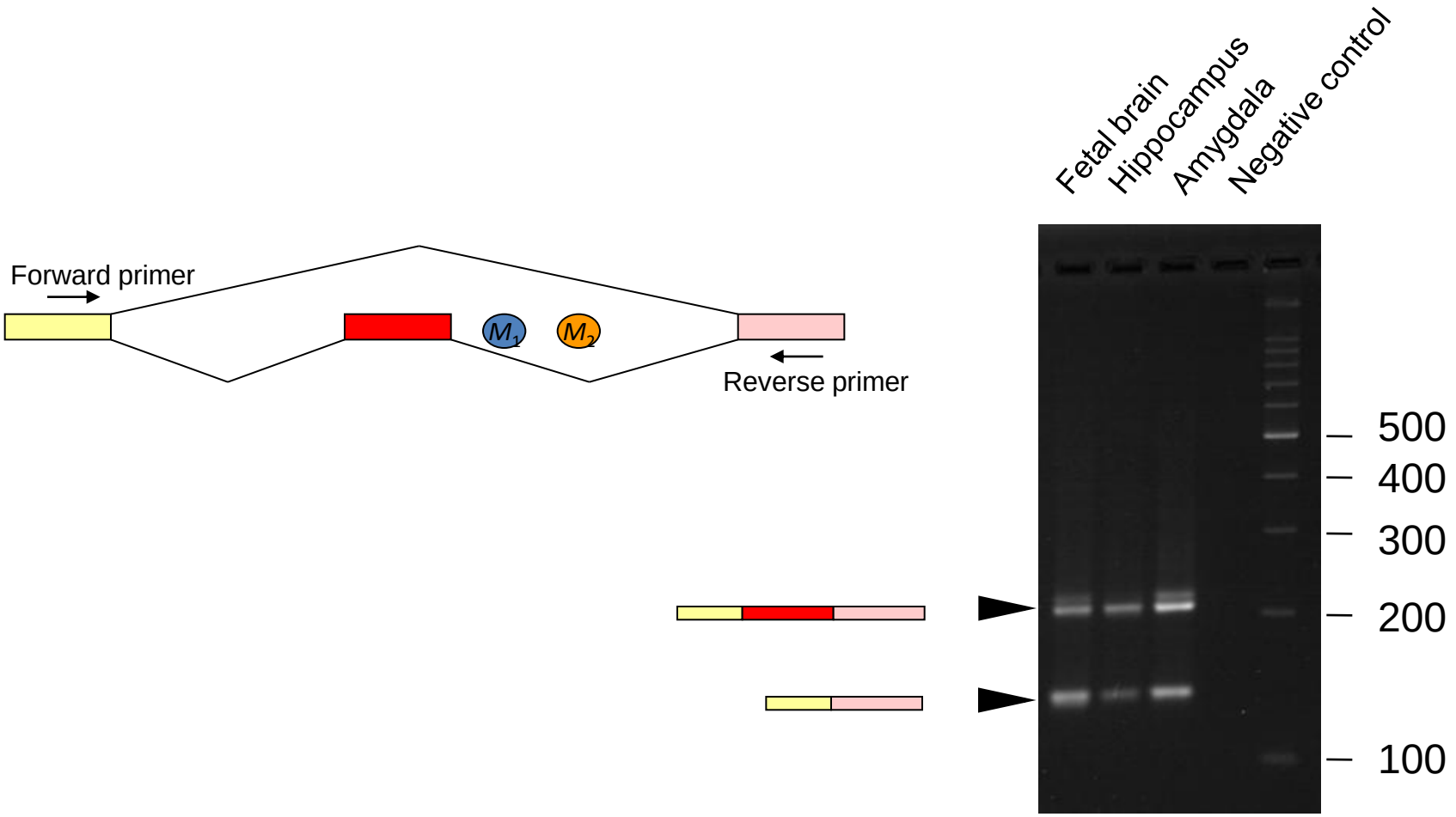


シス因子の組み合わせにより多様性を生み出す。

Suyama *et al. Nucleic Acids Res.*, 2010

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.

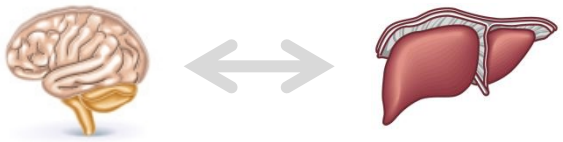


We randomly selected 10 predictions, and confirmed the skipping in 3 cases.

Suyama *et al. Nucleic Acids Res.*, 2010

ゲノミクス解析の今後

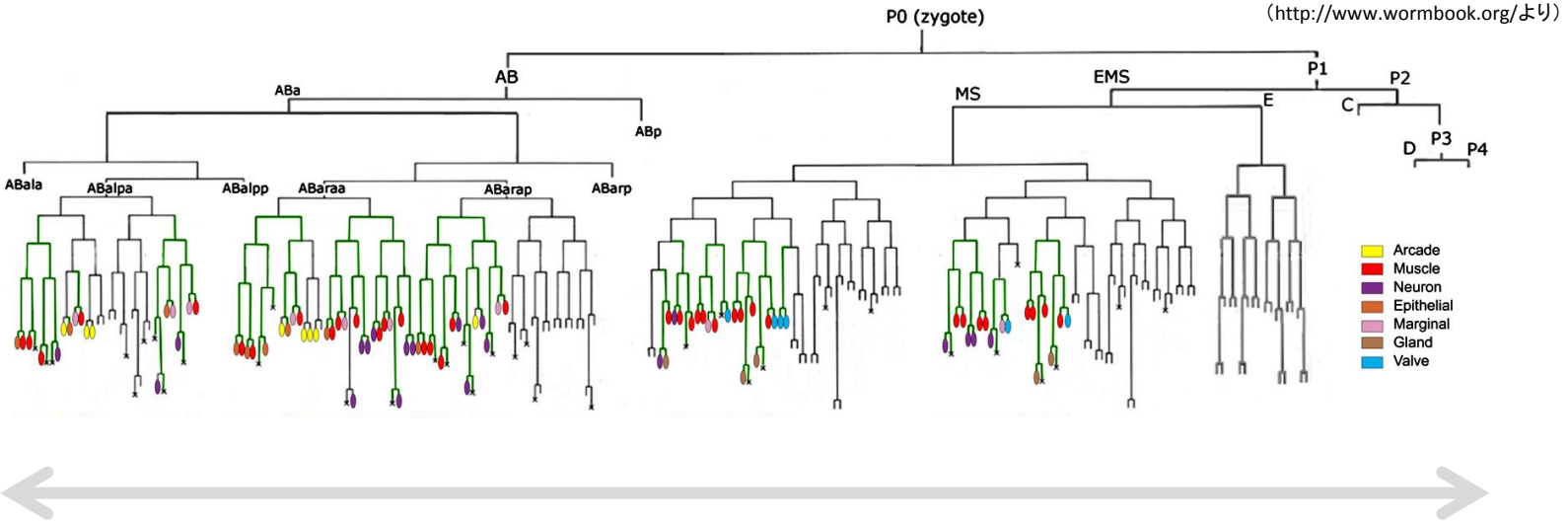
これまで:



2〜3の組織での発現比較

今後:

線虫の細胞系譜
（http://www.wormbook.org/より）



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

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

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

EMBL

Peer Bork

Eoghan Harrington

Bork Group Members

Universität Heidelberg

Magnus von Knebel Doeberitz

Svetlana Vinokourova

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」