

ゲノム情報学

Jeffrey Fawcett

RIKEN iTHEMS (jeffrey.fawcett@riken.jp)

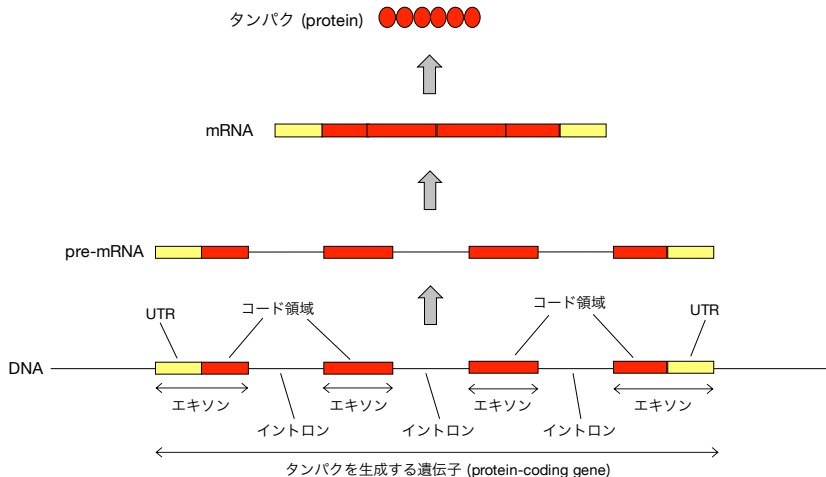
June 2018

ゲノム：遺伝情報の1セット

種	塩基数	遺伝子数
出芽酵母	0.121 億	6,000
ショウジョウバエ	1.75 億	17,000
フグ	3.9 億	28,000
イネ	4.5 億	40,000
マウス	27 億	23,000
ヒト	33 億	21,000
タマネギ	150 億	?
バッタ	650 億	?
肺魚	1300 億	?
キヌガサソウ	1500 億	?

- 塩基数も遺伝子数も「複雑さ」とあまりリンクしない

コード領域：ヒトゲノムの<2%



- それ以外は？ほとんどが機能のないジャンク DNA？

ゲノムを決定した後どうするの？

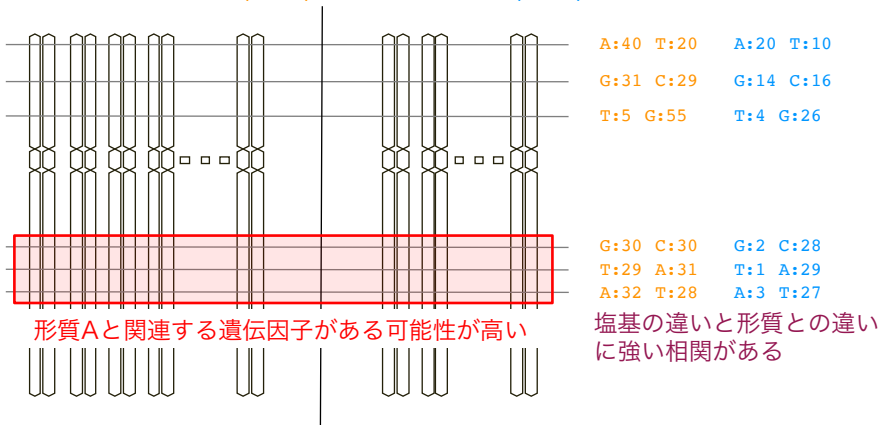
- 配列を決定しただけではあまり意味がない（ただのデータ）
- 遺伝・進化の理論に基づいた予測が重要
 - ▶ 遺伝子予測
 - ▶ 遺伝子予測だけではどの遺伝子（領域）がどの形質と関連しているか分からない
 - ▶ どうやって重要な形質の原因領域を予測するの？

今回の講義の内容

- どうやって重要な形質の原因領域を予測するの？
 - ▶ ゲノムワイド関連解析
 - ▶ 自然選択を受けた領域の探索
 - ▶ 組み換えと連鎖
- ゲノムの非コード領域
 - ▶ トランスポゾン
 - ▶ ENCODE Project
- 生物の進化と全ゲノム重複

ゲノムワイド関連解析 (GWAS)

形質Aを持たない個体(n=30) 形質Aを持つ個体(n=15)

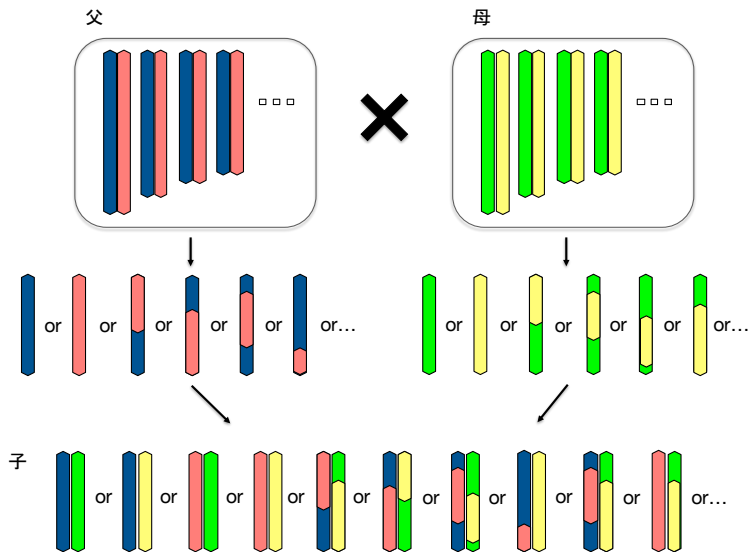


- SNP (Single Nucleotide Polymorphism): 一塩基多型

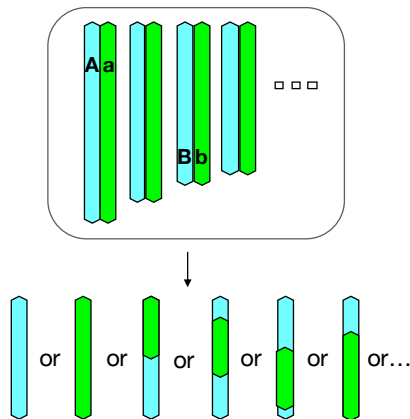
ゲノムワイド関連解析の課題

- サンプルの遺伝的な偏り
(例 形質 A を持つ個体に日本人が多い)
 - ▶ 全サンプル間の遺伝距離を考慮した線形混合モデル
- 多数の遺伝子が関与している形質が多い
(例 身長：数百～数千の SNP が少しずつ影響)
- 「相関」だけで原因かはわからない
- 領域は同定できるが原因の変異まで絞り込めない

組み換えと連鎖

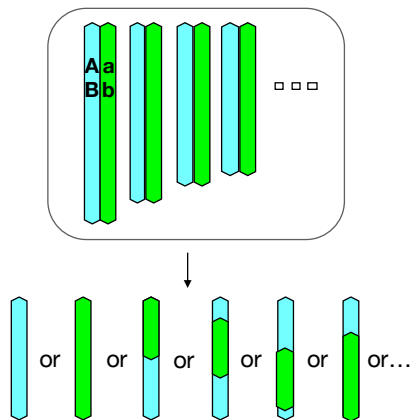


組み換えと連鎖



- A, a の頻度と B, b の頻度は独立 ($AB=Ab$, $ab=aB$)

組み換えと連鎖

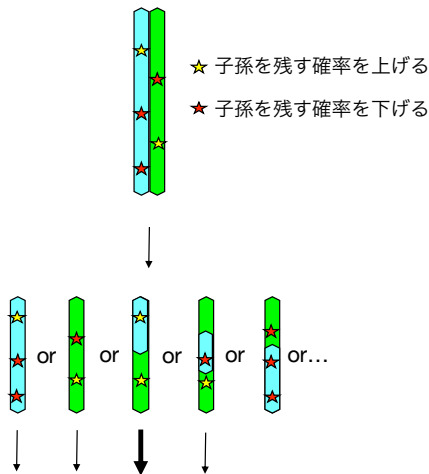


- A, a の頻度と B, b の頻度に強い相関 ($AB \gg Ab$, $ab \gg aB$)
- 重要な変異と一緒に近傍の重要でない変異も広まる

連鎖の影響

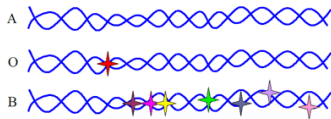
- **利点:** ゲノムを全部決定しなくても原因となる変異を含む領域を容易に同定できる
- **問題点:** 領域が広いことが多く、原因となる変異を同定するのが困難

進化における組み換えの重要性

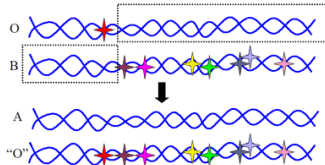


- 組み換えにより自然選択の効率が上がる

ABO 式血液型



There is only one difference between blood type A and O.
There are 7 between B and A.

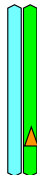


Swapping part of O with part of B can make an A blood type.
This is called recombination.

<http://genetics.thetech.org>

性染色体

● 常染色体から進化



▲ オス化を誘発
(温度などのシグナルで発現する場合も)

顕性 (1つで十分) の場合: XY型

潜性 (2つ必要) の場合: ZW型



▲ を含む領域の逆位

その領域で組み換えが抑制

性染色体の分化が始まる

自然選択（人為選択）の痕跡を見つける

- 重要な領域は変わたらまずい（変異が少ない）
- 子孫を残す確率を下げる変異は広まりにくい
- 子孫を残す確率を上げる変異は広まりやすい

遺伝暗号の冗長性

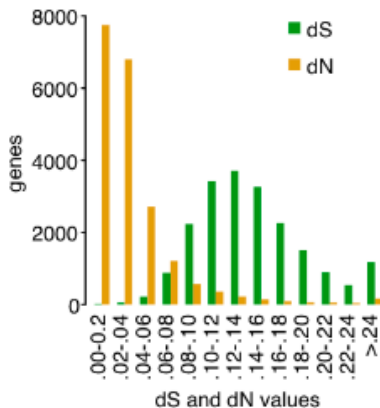
- 同義置換：アミノ酸を変えない変異
- 非同義置換：アミノ酸を変える変異

		Second Letter				
		T	C	A	G	
First Letter	T	TTT } Phe TTC } TTA } Leu TTG }	TCT } TCC } Ser TCA } TCG }	TAT } Tyr TAC } TAA Stop TAG Stop	TGT } Cys TGC } TGA Stop TGG Trp	T C A G
	C	CTT } CTC } Leu CTA } CTG }	CCT } CCC } Pro CCA } CCG }	CAT } His CAC } CAA Gln CAG }	CGT } CGC } Arg CGA } CGG }	T C A G
	A	ATT } ATC } Ile ATA } ATG Met	ACT } ACC } Thr ACA } ACG }	AAT } Asn AAC } AAA Lys AAG }	AGT } Ser AGC } AGA Arg AGG }	T C A G
	G	GTT } GTC } Val GTA } GTG }	GCT } GCC } Ala GCA } GCG }	GAT } Asp GAC } GAA Glu GAG }	GGT } GGC } Gly GGA } GGG }	T C A G

genetics-notes.wikispaces.com

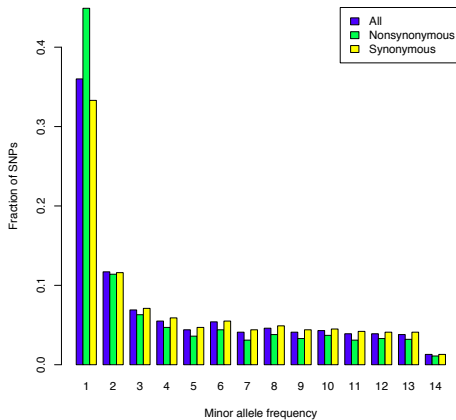
種間の比較

- *Arabidopsis thaliana* vs *A. lyrata*
- 遺伝子間の変異の頻度分布

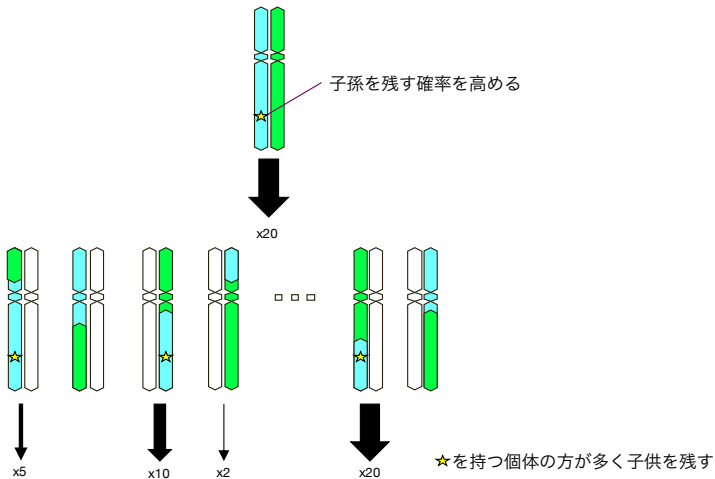


種内の比較

- 分裂酵母 30 個体

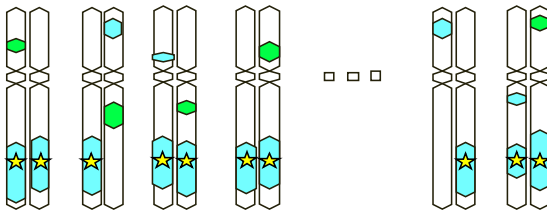


自然選択（人為選択）の痕跡を見つける



自然選択（人為選択）の痕跡を見つける

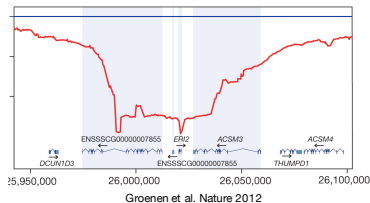
- 何世代も経つと組み換えが起こることで、
 - ▶ 重要な変異の周辺領域で多様度が低下する
 - ▶ 重要な変異を持つ塩基配列の頻度が上昇する



自然選択の痕跡の例

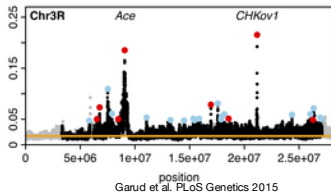
- 塩基多様度が著しく低下している領域

(例：ブタの家畜化の過程で人為選択を受けた領域)



- 特定の配列の頻度が非常に高い領域

(例：ショウジョウバエの環境への適応に関する自然選択を受けた領域)



自然選択（人為選択）の痕跡を見つける

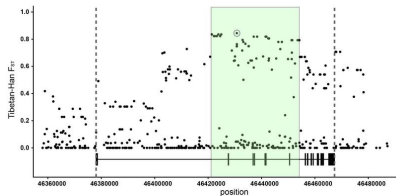
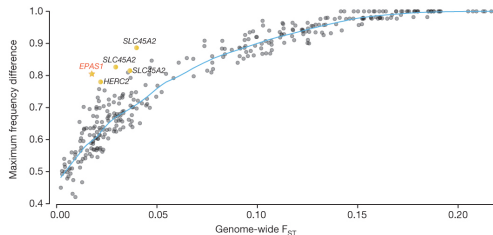


"Fumio Tajima and the Origin of Modern Population Genetics"
in *Genetics* (2016)

- 田島文夫（元東京大教授）
- DNA 配列間の平均塩基相違数（塩基多様度）と DNA 配列中の多型的な塩基部位数の差に基づいた統計量を提唱 (Tajima's D)
- 自然選択の有無や、過去に起きた個体数の増減を推測

チベット人の高地適応

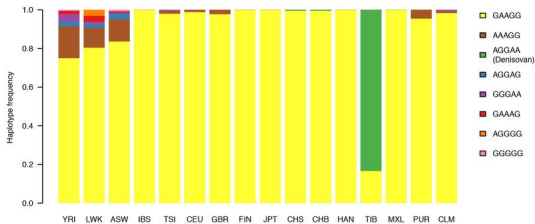
Genome-wide F_{ST} versus maximal allele frequency difference.



E Huerta-Sánchez *et al. Nature* (2014)

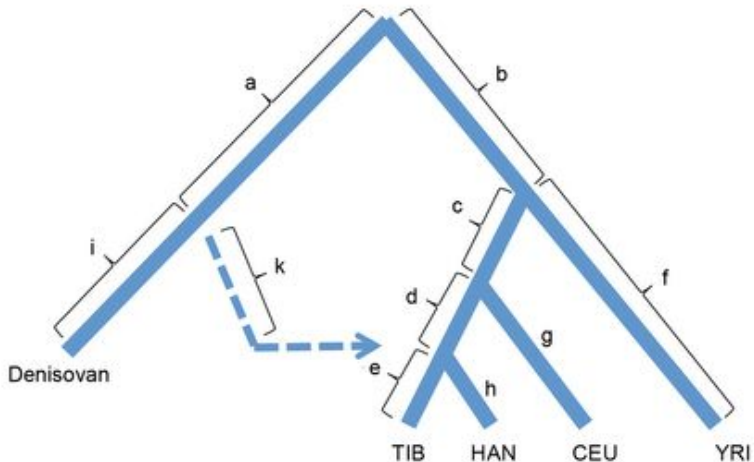
チベット人の高地適応

Haplotype pattern in a region defined by SNPs that are at high frequency in Tibetans and at low frequency in Han Chinese.



E Huerta-Sánchez *et al. Nature* (2014)

チベット人の高地適応



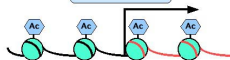
E Huerta-Sánchez *et al.* *Nature* (2014)

DNA やヒストンの修飾

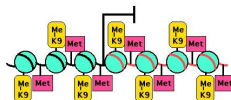
遺伝子発現のエピジェネティックな制御

DNA のメチル化とヒストンの化学修飾が
重要な役割をはたしている

発現している遺伝子



活性化
↓
不活性化



抑制されている遺伝子

Ac ヒストンの
アセチル化

Me K9 ヒストンH3
リジン9のメチル化

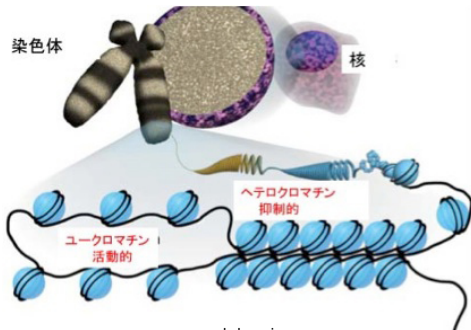
Met DNA のメチル化

www.biomol.med.saga-u.ac.jp

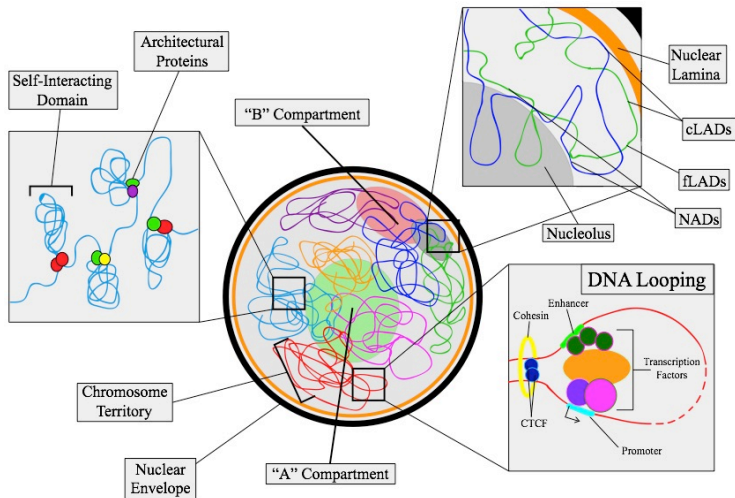
- エピジェネティクス (Epigenetics): 角谷徹仁 (東大教授)

クロマチンの構造

- ユークロマチン (euchromatin) vs ヘテロクロマチン (heterochromatin)

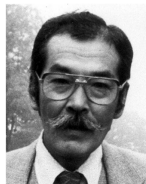


染色体の核内 (3D) 構造



en.wikipedia.org

非コード領域にはどんな情報があるの？



Susumu Ohno

Susumu Ohno: *"So much 'Junk' DNA in our Genome"*
in *Brookhaven Symposium on Biology* (1972)

Junk DNA の論拠

- ゲノムサイズが種によって全然違う
- “Genetic load”
 - ▶ 遺伝子あたりの有害突然変異率は 10^{-5} /世代
 - ▶ 遺伝子が仮に 10 万あるとヒトの有害突然変異率は 1/世代
 - ▶ これは多すぎるので考えづらい
 - ▶ ヒトの遺伝子数は約 3 万と予測
(Muller 1967, Crow and Kimura 1970)
 - ▶ ヒトゲノムにおける機能配列は多くて 6%
(Kimura and Ohta 1971)

McClintock: ノーベル賞 (1983)

- 斑入りトウモロコシ:DNA トランスポゾンの出入りが原因

Dr. Barbara McClintock



© American Philosophical Society Library—Barbara McClintock Papers/National Library of Medicine

Botanist, Geneticist, Nobel Prize Winner

First inspired by an undergraduate genetics course at Cornell in 1921, Dr. McClintock went on to earn her doctorate in Botany in 1927. Her career was focused on genetics thereafter. Dr. McClintock is best known for her research on gene regulation and genetic transposition.

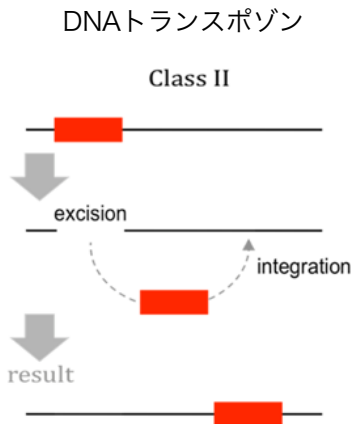
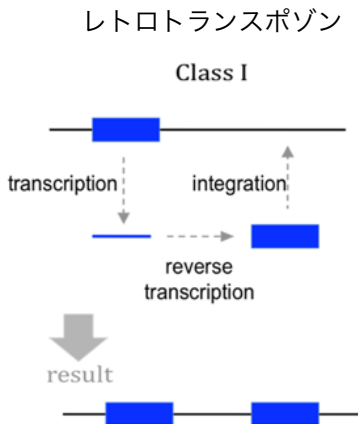
Because it was conceptually difficult and challenged accepted ideas at the time, her research was not immediately accepted by her colleagues. In the late 60s and early 70s, nearly 20 years after Dr. McClintock first published a paper on the subject, other researchers discovered the transposition process in bacteria and yeast; however, Dr. McClintock is generally credited with first discovering transposition. She was awarded the Nobel Prize in Physiology or Medicine for this discovery, the only woman to win the prize unshared in that category.



en.wikipedia.org

トランスポゾン

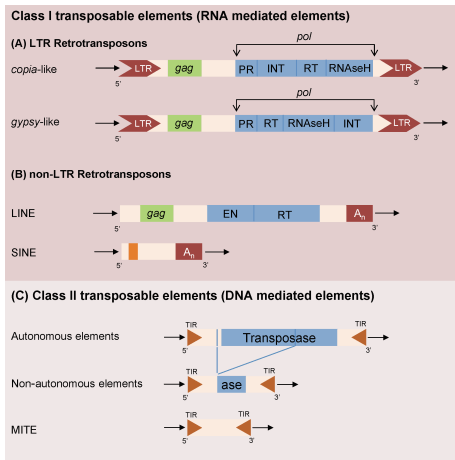
- ヒトゲノムの約 60%



<http://cubocube.com>

トランスポゾン

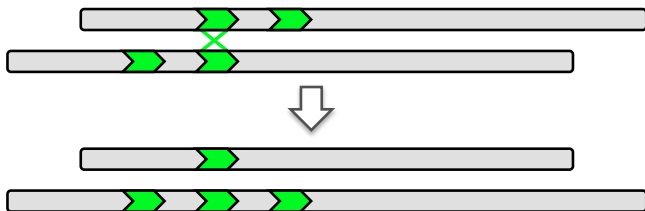
- 転移を可能にするタンパクを生成（そのための配列を持つ）
- 非常に多様



<http://forest.sisef.org>

トランスポゾンの悪影響

- タンパクを生成する領域へ挿入し、タンパクの機能を阻害
- 近傍遺伝子の発現に影響
- 間違った組み換えによる欠失、重複、再編成を誘発



トランスポゾン:進化の原動力

- 遺伝子発現の制御・調節の進化に大きく寄与
- トランスポゾンから新しい重要な遺伝子が生まれることも
 - ▶ 胎盤の遺伝子



トランスポゾンの制御

- メチル化・ヘテロクロマチン化により転写を抑制
- RNAi (RNA interference): Fire and Mello, Nobel Prize (2006)
 - ▶ 二本鎖 RNA を認識し、遺伝子の発現を阻害
 - ▶ ウイルスやトランスポゾンに対する防御機構

遺伝子を「いじる」実験手法

- RNAi: 遺伝子のノックダウン
 - ▶ 機能を阻害したい遺伝子に対応する二本鎖 RNA を注入
- CRISPR/Cas9: ゲノム編集
 - ▶ ゲノムの任意の箇所を切断し、任意の新しい配列を導入

非コード領域の機能の解明へ

- タンパクを生成する配列はヒトゲノムの $<2\%$
- それ以外（非コード領域）は何をしてるの？

Ultraconserved element

Ultraconserved Elements in the Human Genome

Gill Bejerano,^{1*} Michael Pheasant,³ Igor Makunin,³
Stuart Stephen,³ W. James Kent,¹ John S. Mattick,³
David Haussler^{2*}

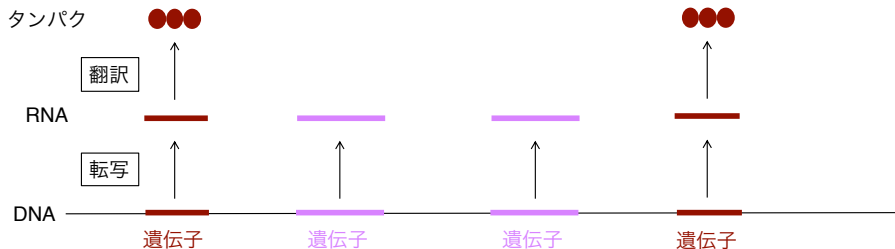
Science (2004)

- コード配列より遥かに保存されている非コード領域が多数
- ヒトとマウスで >200 塩基で全く同じ
- ニワトリ、フグとも非常に似ている
- 転写の調節・制御に関与 !?
- 植物では遠い種間で異常に保存されている配列はない

非コード領域の機能の解明へ

- 非コード領域にも重要な領域が多数存在
- 「変わったらまずい」パターンを示す領域はせいぜい 10%
 - ▶ まだ未解明なものが多い
- 残りは？ 方法論の問題？ ジャンク !?

Non-coding RNA

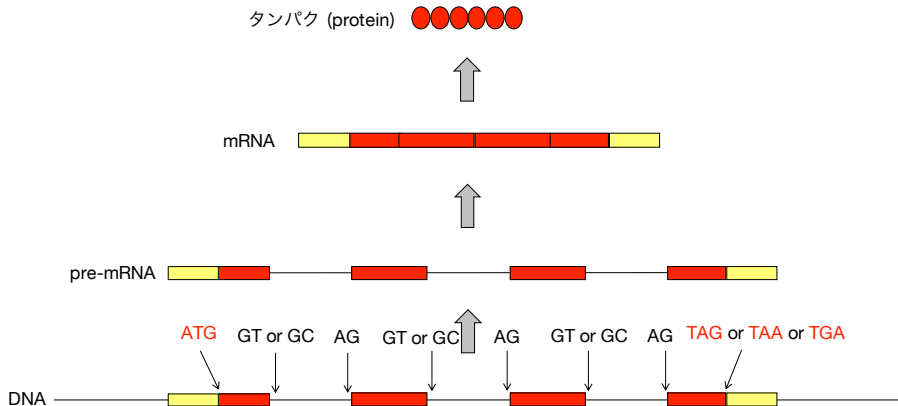


- 重要な機能を持つ RNA 遺伝子が多数存在
- タンパクを生成する遺伝子の発現の調節・制御に関与!?

Non-coding RNA

- ヒトゲノムの大部分が転写されている
- 「転写されている」＝「機能を持つ、重要」ではない
 - ▶ 例) pseudogene も転写される場合あり

Pseudogene (偽遺伝子)



- コード配列の途中に終始コドンが出来、本来のタンパクを生成できない

Non-coding RNA

- 重要な Non-coding RNA の割合は？機能は？
- はっきりしたルールがなく、まだ解明が必要
 - ▶ 大半が保存されていない
(変わったらまずいパターンが見られない)
 - ▶ 例) 転写されることが重要で配列はどうでもいい
 - ▶ 例) 最初の数十塩基だけが重要で他は必要ない

ENCODE project (2012 に発表)

- 非コード領域にはどんな情報があるのか？
- ヒトゲノムで機能を持つ配列を実験的に全部見つけよう！
- 実験データを大量に生成、重要なマイルストーン、だが、、、
- 「ヒトゲノムの 80% が機能を持つ！」
- 「ジャンク DNA はもはや死語！」

批判：「機能を持つ」とは？

- 狭義：変わったらまずい、その形質の存在意義
 - ▶ 例) 心臓の機能：血を供給
- 広義：その形質の属性、二次的な効果・影響
 - ▶ 例) 心臓の機能：体重に 300g 上乗せする。音を出す。
- ENCODE は広義の「機能」を使っている !?

批判：「機能を持つ」とは？

- ENCODE の定義：以下のいずれかに該当するもの
 - ▶ 転写されている
 - ▶ ヒストンの修飾がある
 - ▶ open chromatin 領域にある
 - ▶ 転写因子に結合する
 - ▶ メチル基の修飾がある
- 必要条件であっても十分条件ではない

ENCODE の問題

- 政治的なアピール ...
- 進化・理論に対する理解のなさ
- ヒト中心の研究

「進化の過程で無駄なものは除かれるはず」!?

- 完全な間違い !! 一あくまでも確率論
- A, a の集団内での頻度をそれぞれ p, q とすると、中立な場合、 A, a の固定確率はそれぞれ p, q
- 個体数 N の集団で生じた新しい変異が固定する確率は $1/(2N)$ (個体数が大きく影響する)
- 自然選択がある場合: 選択係数 s と N などに依存
- s が負の場合も 50-100 年前から理論研究が行われている
- イメージで語られる事象に対して定量的な議論を行えるのが理論の重要さ・面白さ !?

トランスポゾンの場合

- 個体数
- トランスポゾンのコピー数
- トランスポゾンの転移活性
- トランスポゾンの欠失確率
- トランスポゾンの新しい挿入に対する自然選択の強さ
- トランスポゾンのコピー数に対する自然選択の強さ
- 30-40 年前から理論研究が行われている

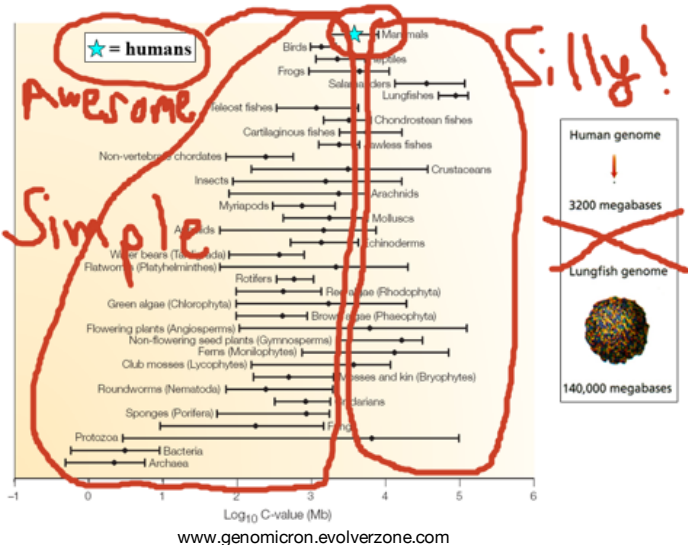
トランスポゾンの進化の例

- Au SINE:植物界に広く分布
- 持っている種も持っていない種もある
- 遠く離れた種同士で配列が非常に似ている
- 高頻度の水平伝播 or ゲノムにとって重要な機能?
- 約 70 種の植物のゲノムデータを比較解析
- 大昔から存在、トランスポゾンにとって重要な配列
(配列を維持したトランスポゾンだけが増幅し続けられた)

ENCODE の問題

- 政治的なアピール ...
- 進化・理論に対する理解のなさ
- ヒト中心の研究

生物界全体を見ることが大事



“Onion test”

Junk DNA が重要な普遍的な機能を持つと主張したいならまず次の問いに答えてみよう

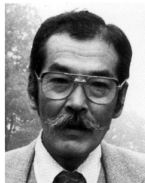
種	塩基数
ヒト	33 億
<i>Allium altyncoliucum</i>	70 億
<i>Allium cepa</i> (タマネギ)	150 億
<i>Allium ursinum</i>	300 億

- なぜタマネギがそれをヒトの 5 倍も持つ必要があるのか？
- なぜタマネギのある近縁種では半分でもよくて、別の近縁種では倍必要なのか？

どうやって新しい遺伝子が生まれるのか

- ゲノムの非コード領域のランダムな配列から
- トランスポゾンから
- 遺伝子の重複

Evolution by Gene Duplication (1970)

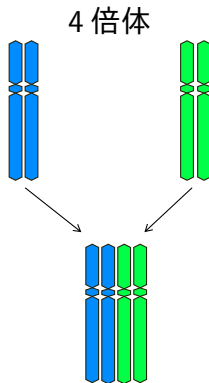
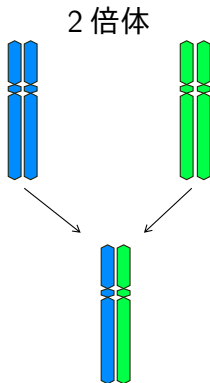


Susumu Ohno



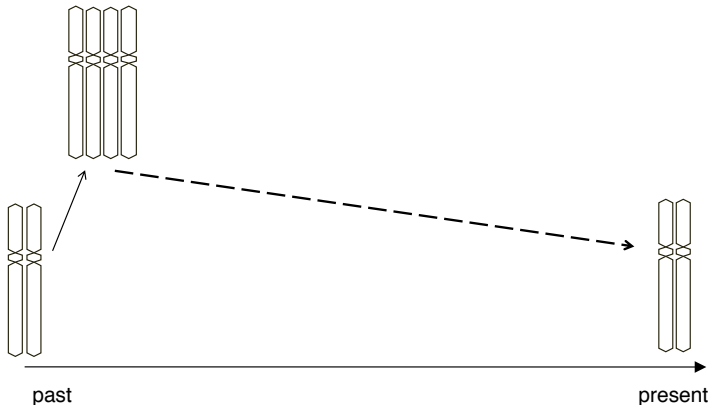
- 発現量が倍になる
- 片方が機能を維持していればもう片方は変わっても大丈夫

全ゲノム重複



全ゲノム重複

- 現在 2 倍体の種も昔 4 倍体だった？



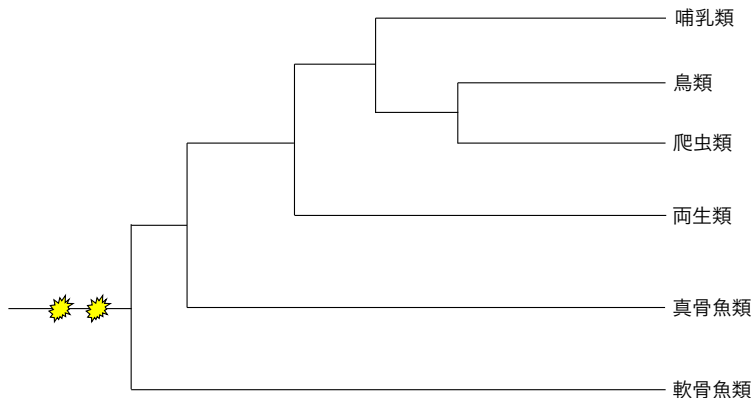
複雑さの問題

機械 A が機械 B を作る $\rightarrow$ 機械 A の複雑さ $>$ 機械 B の複雑さ?

- 答 1: 設計図のエラーが進化を可能にする
- 答 2: 機械 A が機械 B を 2 つ作ることがある

2R 仮説 (Ohno)

- 多くの脊椎動物の共通祖先で2回の全ゲノム重複
- 複数の脊椎動物のゲノム解読により（ほぼ）証明
- 脊椎動物の進化、複雑性の獲得に大きく寄与



生物の進化と全ゲノム重複



生命大躍進
第1集 そして"目"が生まれた



No.512
DNAの大事件！ 生命進化の謎

2015年7月26日 放送

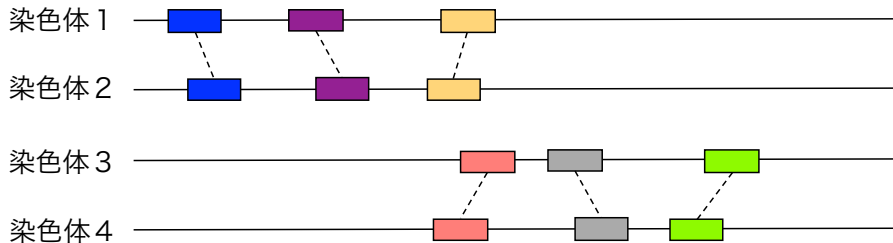
キーワード
全ゲノム重複

科学
技術
生命
科学

シロイヌナズナ *Arabidopsis thaliana*

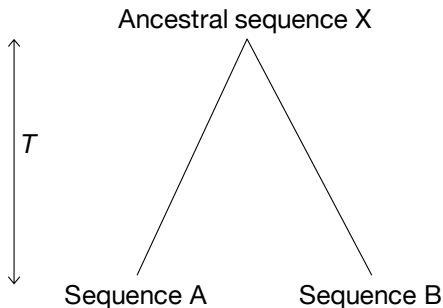
- モデル植物、2000 年全ゲノム解読
- 2 倍体、ゲノムが小さい (1.35 億塩基)
- 似たような遺伝子がいっぱい (重複遺伝子)
- いつ重複したの？

似たような遺伝子がいっぱい



- 似たような領域がいっぱい = 領域ごと重複？
- 似ている度合いから重複時期を推定できる

似ている度合いと分かれてからの時間が比例



- $K = 2Tr$

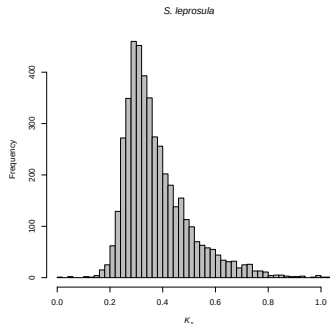
K : 塩基あたりの違い

T : 時間

r : 塩基置換率

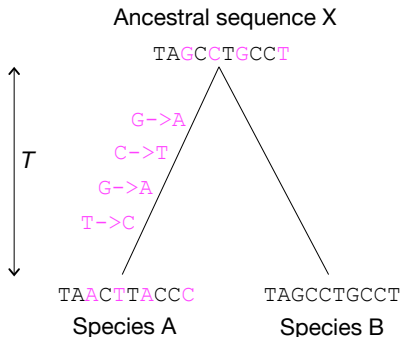
- ゲノムデータから K が求まるので、そこから $T = K/2r$

いつ重複が起きたかを推定



- 各重複ペアの同義置換を計算
- ほとんどが同程度の「似ている度合い」
- 全ゲノム重複？

塩基置換率 r が種によって大きく異なる



- $K_{XA} \neq K_{XB}$
- どの種を使うかで T が変わ
りうる

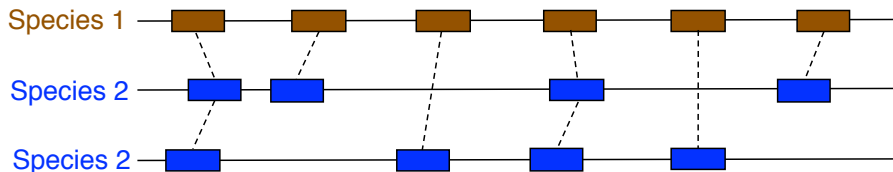
- 系統樹の研究者には知られた問題
- ほとんどの生物学者は十分認識しておらず

塩基置換率 r が種によって大きく異なる

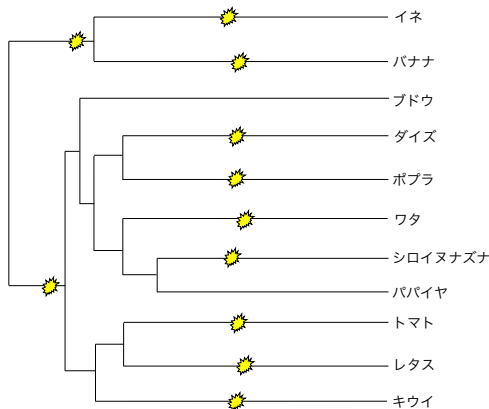
- 例) ポプラの全ゲノム重複
 - ▶ シロイヌナズナで推定された r を用いた場合 : 800-1300 万年前
 - ▶ 近縁種とのゲノム比較の結果 : >6000 万年前
- 草本性の植物 $\gg$ 木本性の植物

もう1つの手法

- 種間で遺伝子の並びを比較 (1:1 か 1:2 か)
- 全ゲノム重複の相対的な時期を推定

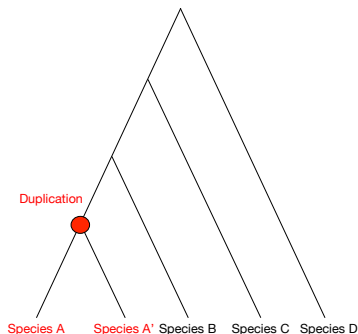


生物の進化と全ゲノム重複



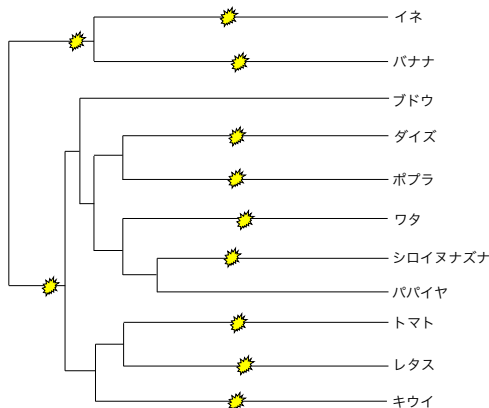
- 単子葉類の共通祖先で、双子葉類の共通祖先で
- 多くの種が独立に全ゲノム重複を経験している

全ゲノム重複の年代推定



- 系統樹の分野の手法を拝借
- 各枝の塩基置換率が異なりうるモデルを使用
- 化石に基づいた年代も事前情報として使用

生物の進化と全ゲノム重複



- ~6500 万年前の恐竜等が絶滅した時期の前後に多くの被子植物で全ゲノム重複
- 生物の絶滅への耐性や複雑性に大きく寄与？