

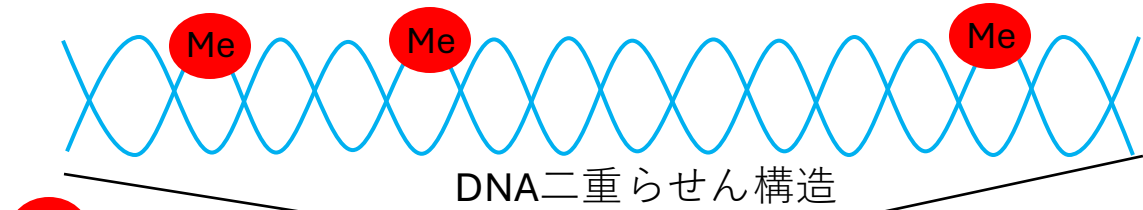
Genome Language Model (GLM)

への期待

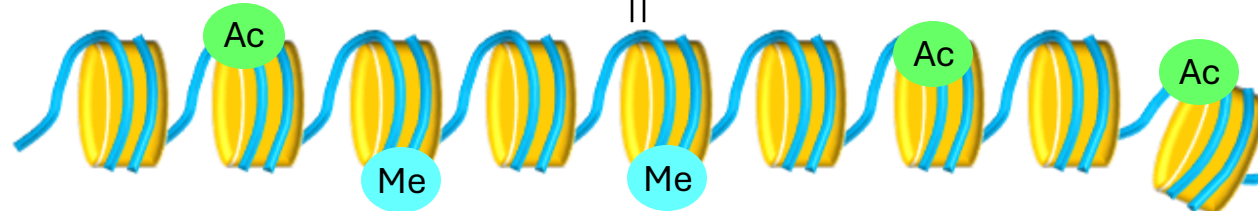
森下 真一

東京大学

DNAメチル化



Me DNA methylation

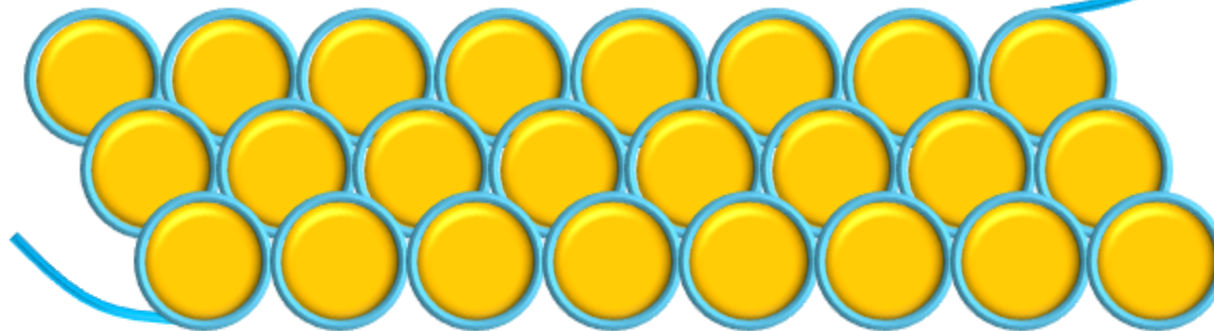


ヒストン修飾と
ヌクレオソーム構造

Ac Histone acetylation

Me Histone methylation

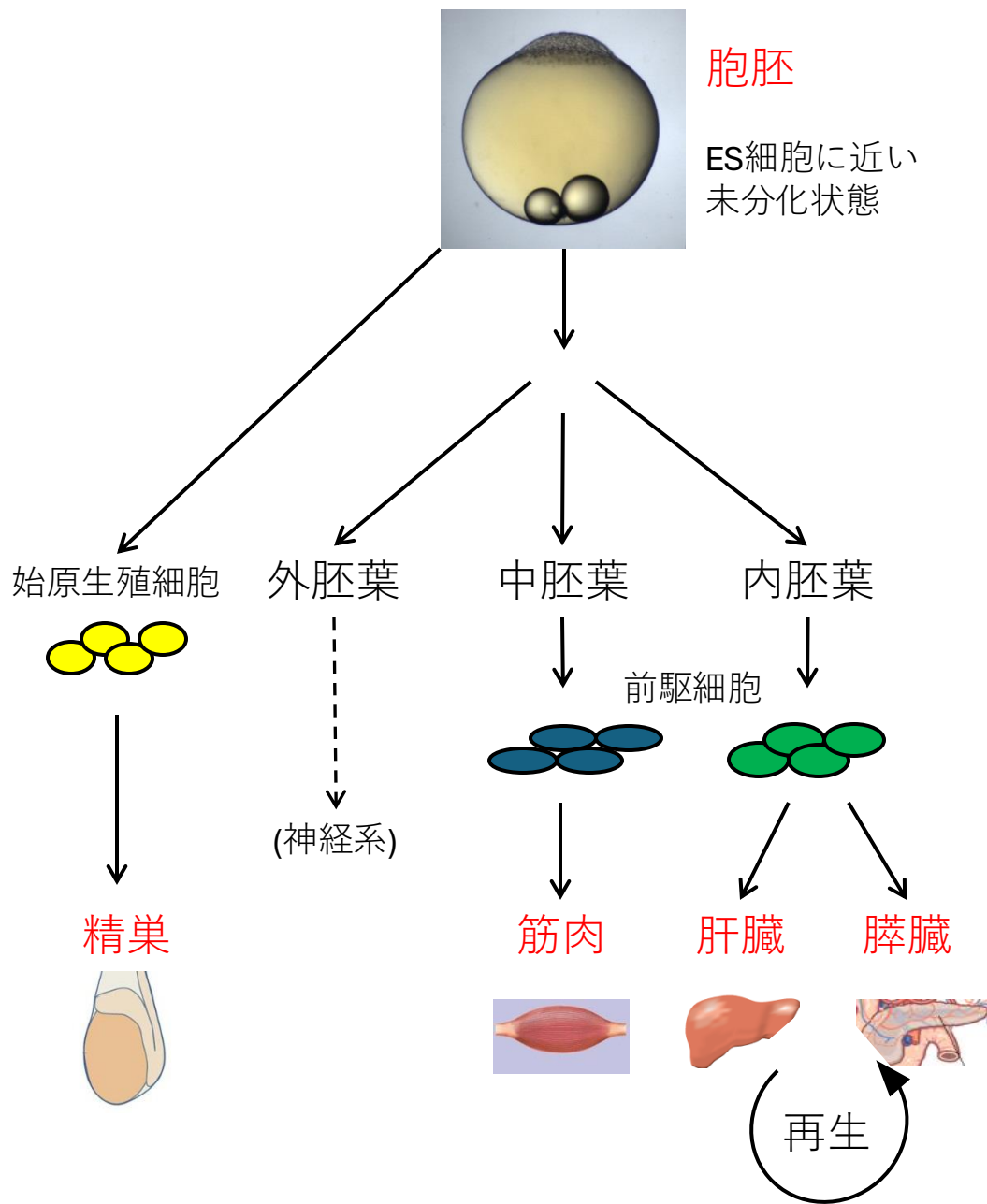
1972, Roger Kornberg



3次元クロマチン
構造の動態

2009, Erez Lieberman-Aiden





発生過程では多分化能が
獲得・維持される

細胞分化の過程で、
同一のDNAが細胞へ分配

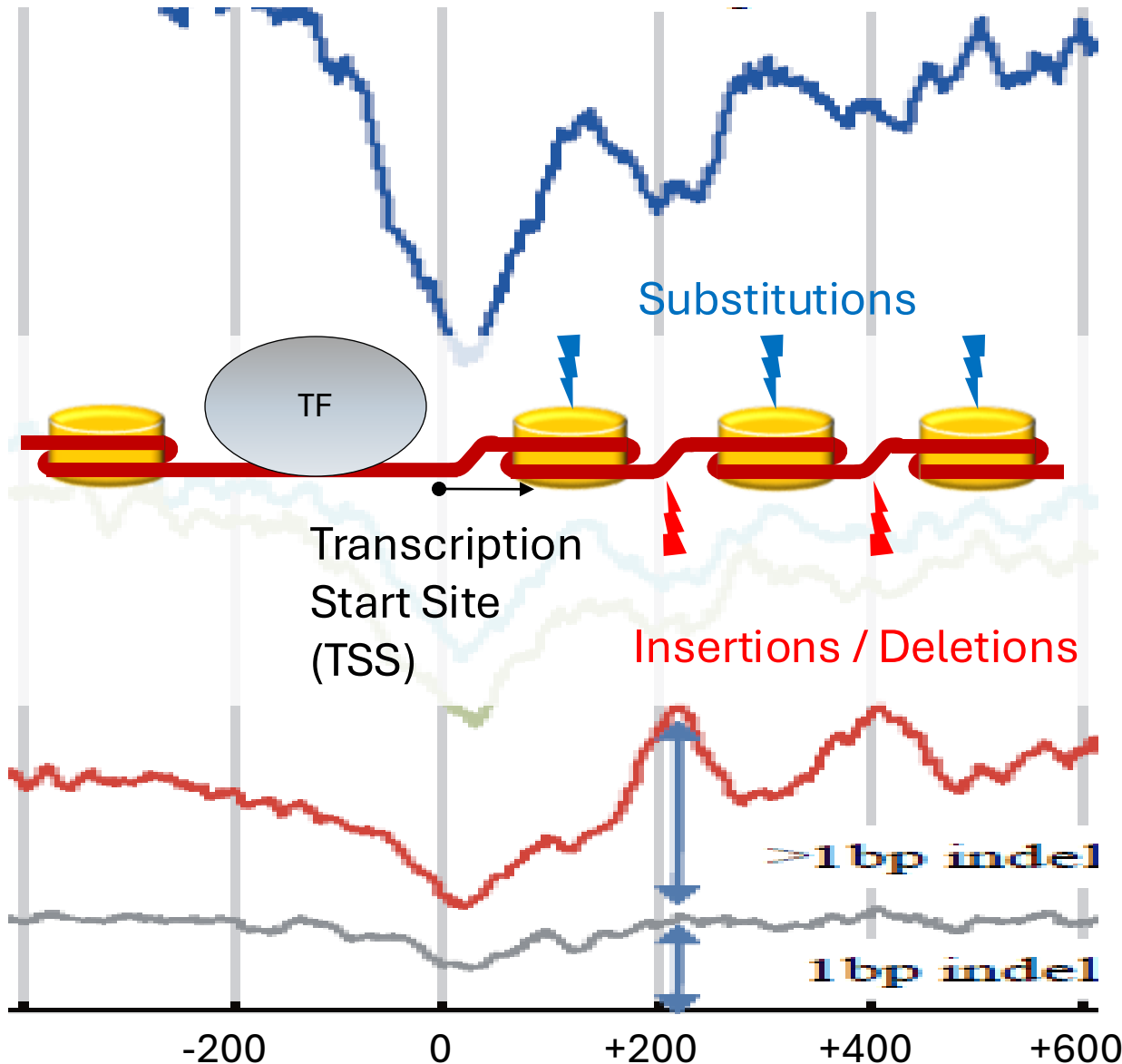
DNAの変化は僅か
(生殖細胞 $1/10^8$ 体細胞?)

変化しやすい指標:

- Epigenetic code
 - ✓ DNAメチル化
 - ✓ ヒストン修飾
 - ✓ ヌクレオソーム分布
- DNA 3次元折畳み構造
(クロマチン構造)

Chromatin associated genetic divergence

クロマチン構造は遺伝的多様性に深く関わる



メダカで最初に観察
Sasaki S et al. *Science*.
323(5912):401-4 (2009).

ヒトゲノムでも再確認
Tolstorukov, M.Y. et al.
Nature structural & molecular biology
18, 510-5 (2011).

JST より支援 (2006~2010)
課題名
マルチモーダル統合バイオDB

Joint Study with Andrew Fire, Cecilia Mello, Hiroyuki Takeda, et al.

高精度の個人ゲノムを読めば、
手間のかかる実験をせずに、
高精度のマルチモーダル情報を、
高速に推定できるか？

- プロモータ・エンハンサ活性
- ヌクレオソーム分布
- DNA修飾
- ヒストン修飾
- DNA折り畳み構造
- スプライシング異常

ゲノム解読の解像度とデータ量の変遷

ゲノム上のデータ密度（解像度）

● 技術の普及開始 ● UK BioBank 約50万人 時間軸 →



2002

2017

連鎖不平衡領域上の代表点 50万～100万点
マイクロアレー方式(別名 DNA チップ)



exon 上の配列 約 1 億塩基 Exome方式

2008

2022



全ゲノム配列 約30億塩基

Illumina short read

PacBio long read

Nanopore ultralong read

2010

2023

2011

2016

2027?

GLM の目標設定案

- 高精度ゲノムを解読して、どのような論文が書けるか検討
 - 実験研究者（医科学研究者やモデル生物学者）が、入力する高精度ゲノムを解読するのを手助け、計算したマルチモーダル情報を表示する快適なツールも必要か？ AlphaGenome も提供
- AlphaGenome 等のソフトウェア構成を解読して追いつく
 - 高速化や大規模化： AlphaGenome が扱えるゲノムサイズが高々100万塩基ぐらいなので、GPU計算機を十分に用意することが競争力を上げるには重要か？
- 訓練データ UK BioBank, HPRC, ENCODE, 4D Nucleosome 等
 - 異なるデータセットの補正、標準化
- 企業としては DeepMind, InstaDeep 等の海外企業が先行国内に GLM を設計開発できる人材を育成することが課題？

