離散的配列情報と高次元生命情報をつなぐ
統計と機械学習

Center for iPS Cell Research and Application, Kyoto University
Risa Karakida Kawaguchi

Kawaguchi Lab started since 05.01.2022!

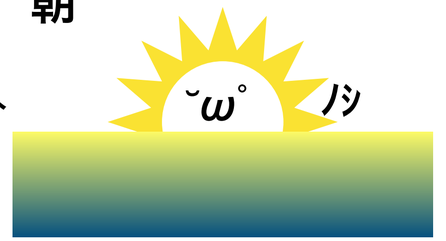
@ 定量生物学の会 チュートリアル（2022/12/07）

アウトライン：配列解析を中心とした統計と機械学習のお話

1. 生物学における配列解析から統計モデルへ (30min)

- RNA二次構造予測を例とした配列と機能を結びつける統計モデルの導入
- アルマジロの一細胞解析による非遺伝的変動検出

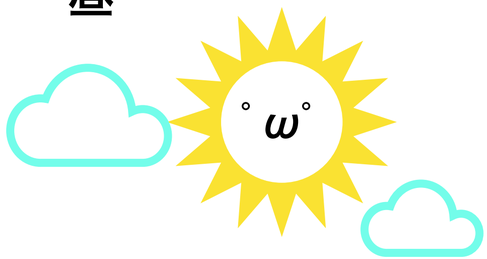
朝



2. ラジオゲノミクスにおける機械学習の実応用　(20min)

- MRI画像解析から学ぶヘテロな医療データの取り扱い
- 次元圧縮とファインチューニング

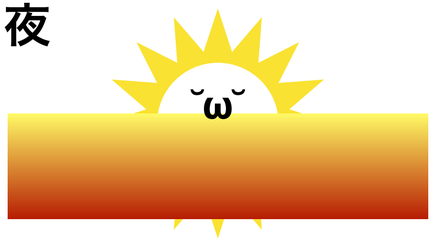
昼



3. メタ統合解析のススメ 機械学習・人工知能技術 実践編 (余った時間)

- 共発現ネットワーク解析
- 一細胞エピゲノムのメタ統合解析と深層学習によるモチーフ予測

夜



想定している前提知識

- 主に大学院の学生さんに向けた内容になります
- ゲノミクス・トランスクリプトームを中心とした生物学の基礎
- Pythonのコードがなんとなく読める・書ける
- いろいろなトピックをかいつまんで行きます
- 資料は英語と日本語混ざりです、ご容赦くださいませ

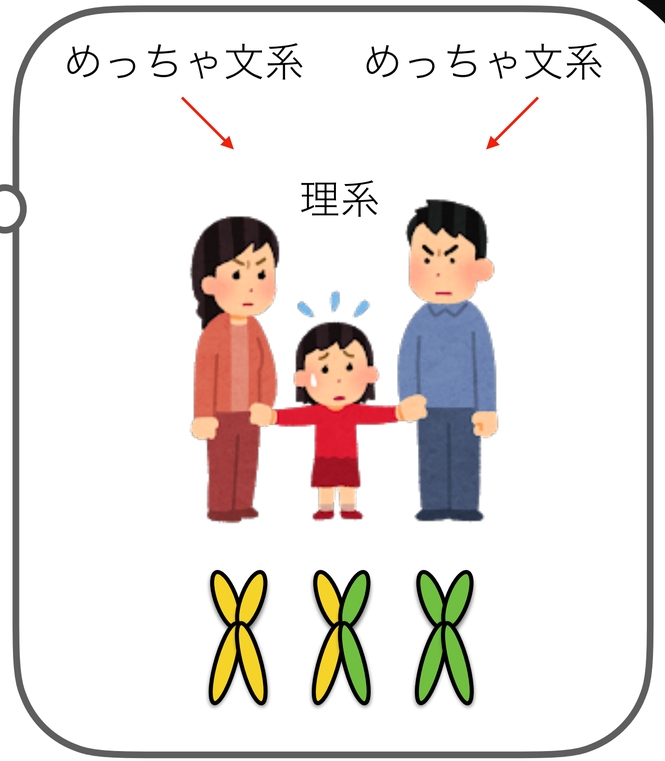
ゲノムへ興味を持った高校時代

生物の授業で…

- ゲノム配列は生物の設計図
- 染色体は父と母から受け継がれる

(父＋母) / 2 = 自分 … ?

（うっそだぁ）

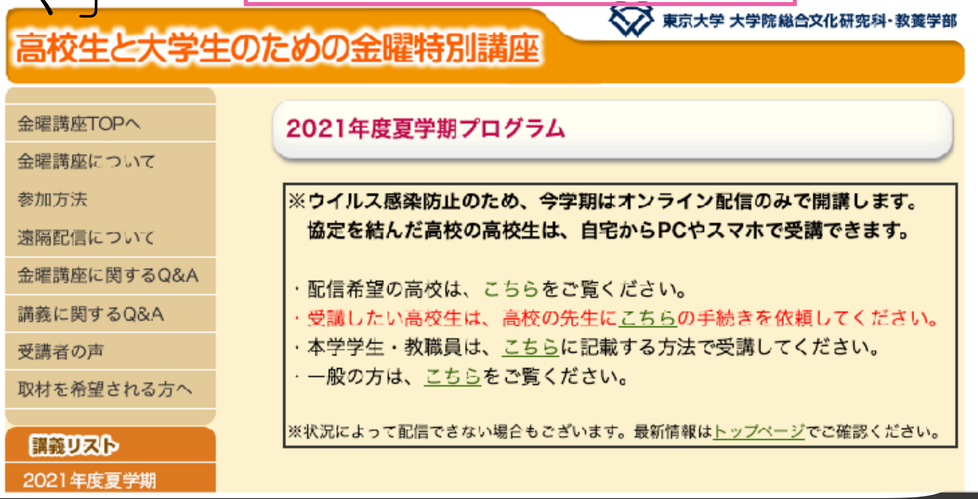


ゲノムは個体の性質をどこまで決定するのか？

文理が決められなかった高校時代

- 「**面白い**」と「**得意**」の不一致
  - 得意なもの：国語・英語・社会系・確率（！？）
  - 壊滅的なもの：物理・数学
- 担任の先生のアドバイス
  - 「毎日寝る前に文理どっちにするかノートに書く」
  - 「進振りがある東大にすれば？」
  - 金曜講座・大学院の一般講義・OC
  - 入学前に生物情報科学科への進学を決定

東大TVでオンライン化！！！



http://high-school.c.u-tokyo.ac.jp/lecture_time/2021s.html

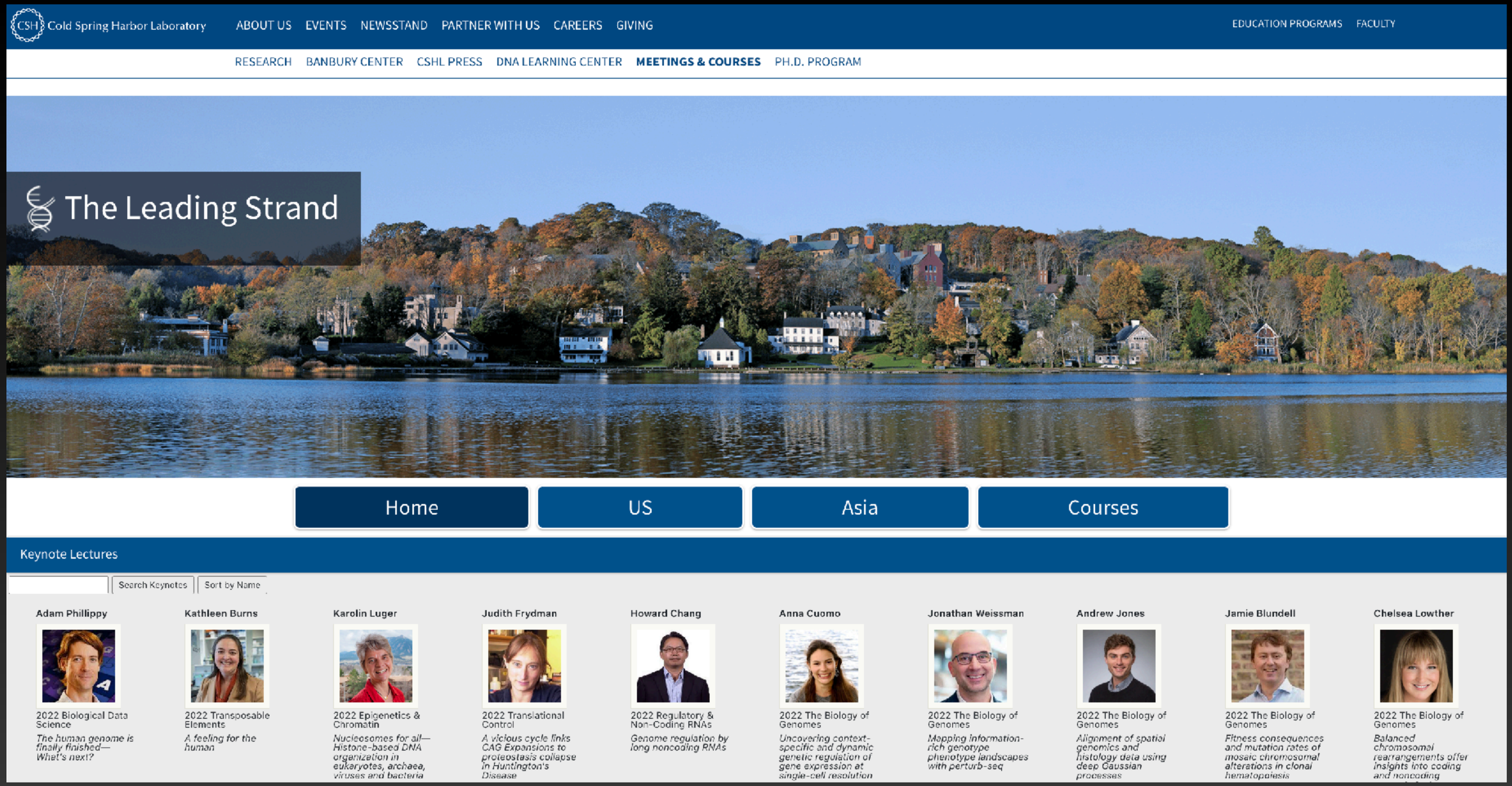
所在地

From Google Map

1890年に設立されたアメリカ合衆国ニューヨーク州にある**生物学・医学系研究所**

多くのノーベル賞受賞者を輩出。でも規模がこじんまりしていてアットホーム（）

- 世界中から研究者が集まるミーティング・コース
- 多様かつ豊富な海の幸
- 国際会議の最終日ディナーはいつもロブスター
- 突如出現する巨大カブトガニ



↑ 最新のBiological Data Scienceのキーノート

leadingstrand.cshl.edu

留学のきっかけとなった論文


Predictability of human differential gene expression

Megan Crow[a], Nathaniel Lim[b,c,d], Sara Ballouz[a], Paul Pavlidis[b,c], and Jesse Gillis[a,1]

Crow M, et al. PNAS, 2019.

- **RNA-seqをやる目的**
  - 1. 複数間のサンプル比較により発現変動遺伝子（DEG）を見つける
    - 例：健常 vs 疾患、通常 vs ストレス化
  - 2. 発現変動パスウェイを見つける
    - GO・KEGGなどのエンリッチメント解析
  - → 特定の条件下で遺伝子の役割を理解・創薬ターゲットに

→ どんなDEGでもよいのか？

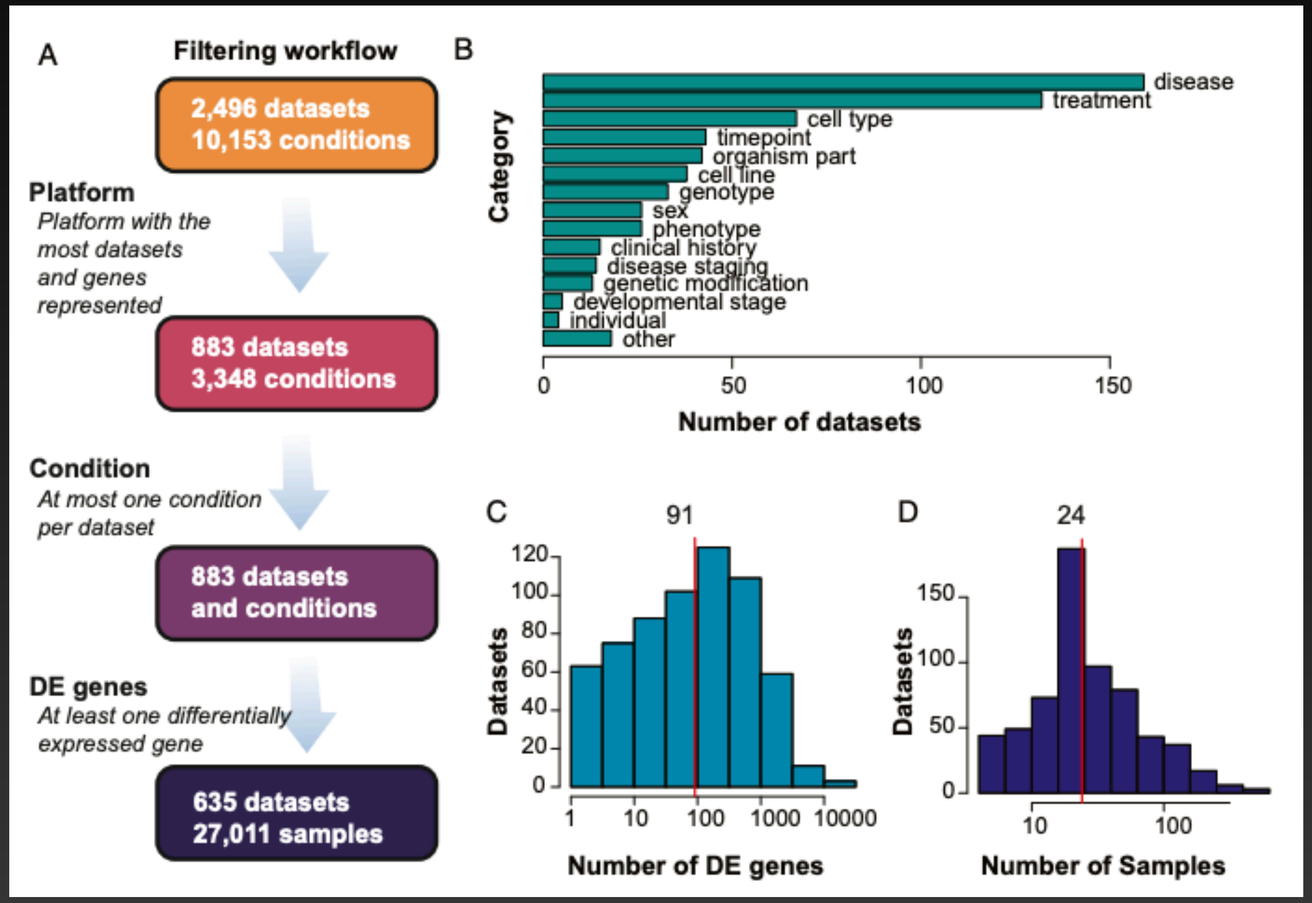
エンリッチメント解析
サーバーに投げられる
遺伝子の10%はキナーゼ
（Huang DW, et al. NAR, 2009）

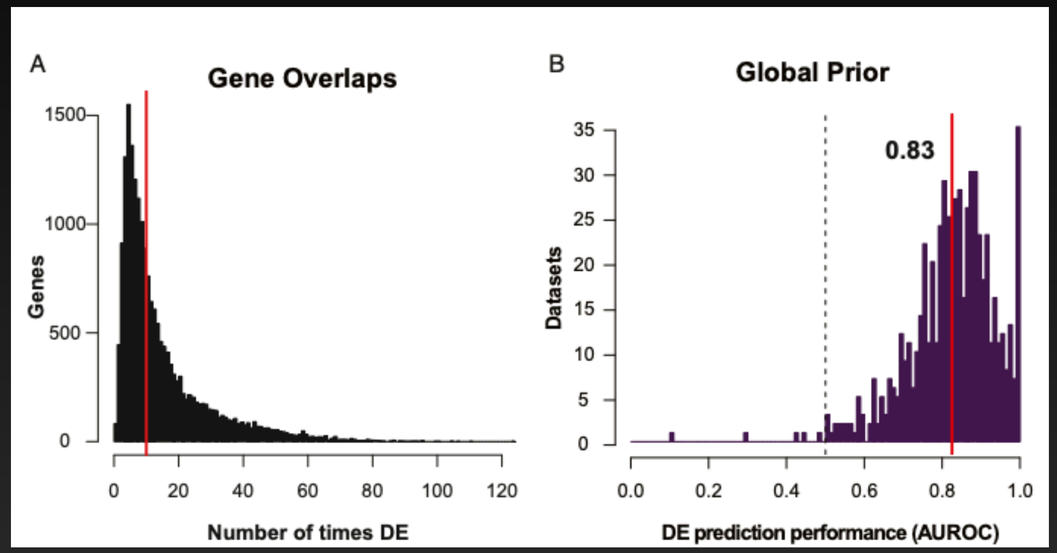
留学のきっかけとなった論文

Predictability of human differential gene expression

Megan Crow[a], Nathaniel Lim[b,c,d], Sara Ballouz[a], Paul Pavlidis[b,c], and Jesse Gillis[a,1]

Crow M, et al. PNAS, 2019.





- 大量の遺伝子発現データセットを収集
- DEの事前分布を作成
- 平均AUC=0.83をマーク

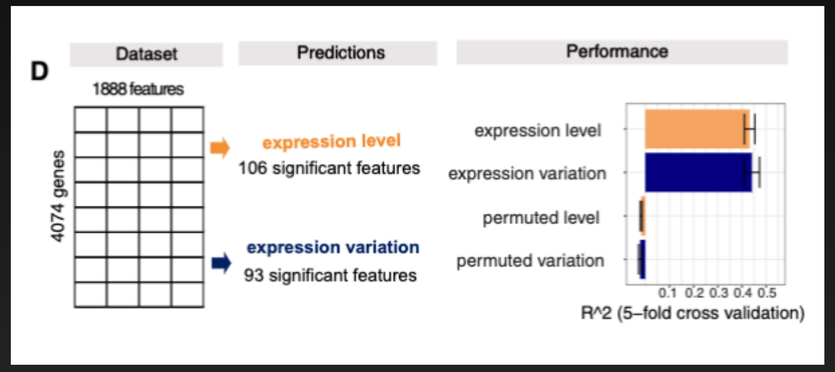
Article

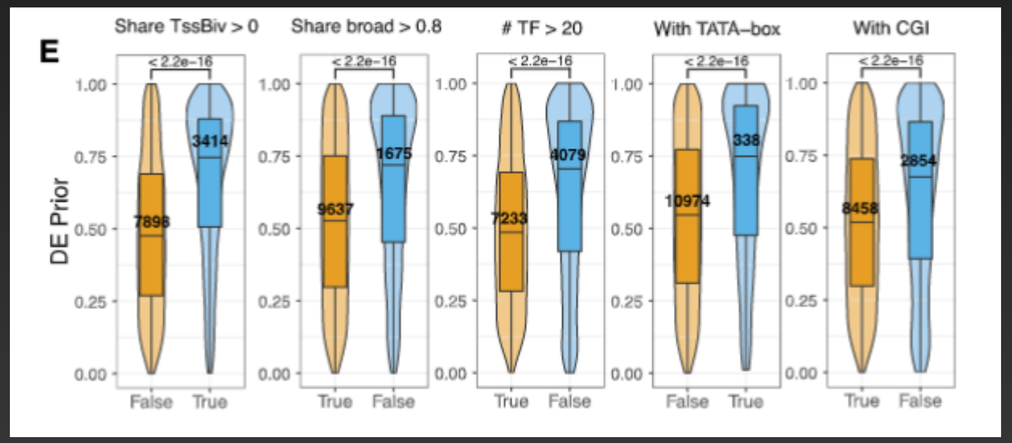
TRANSPARENT PROCESS OPEN ACCESS

molecular systems biology

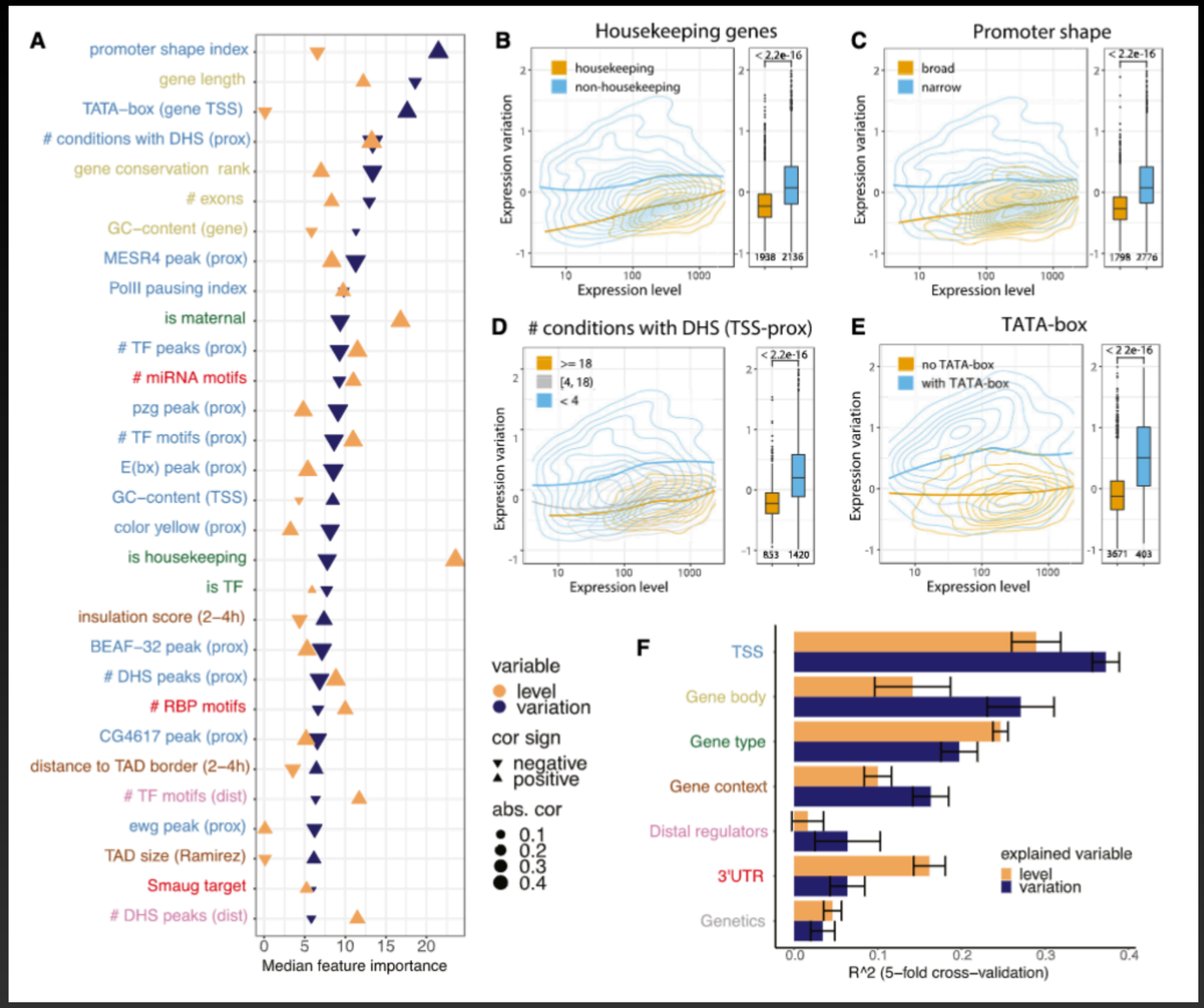
Predictive features of gene expression variation reveal mechanistic link with differential expression

Olga M Sigalova[1], Amirreza Shaeiri[2], Mattia Forneris[1], Eileen EM Furlong[1,*] & Judith B Zaugg[2,**]





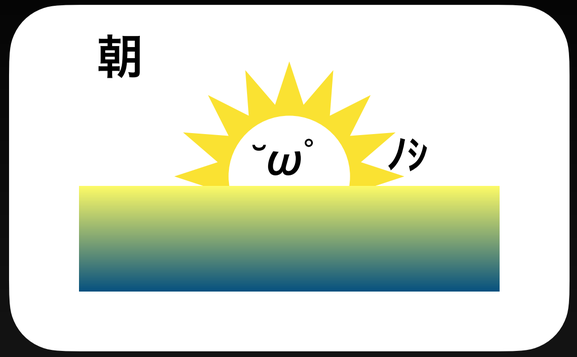
DE prior and other genes



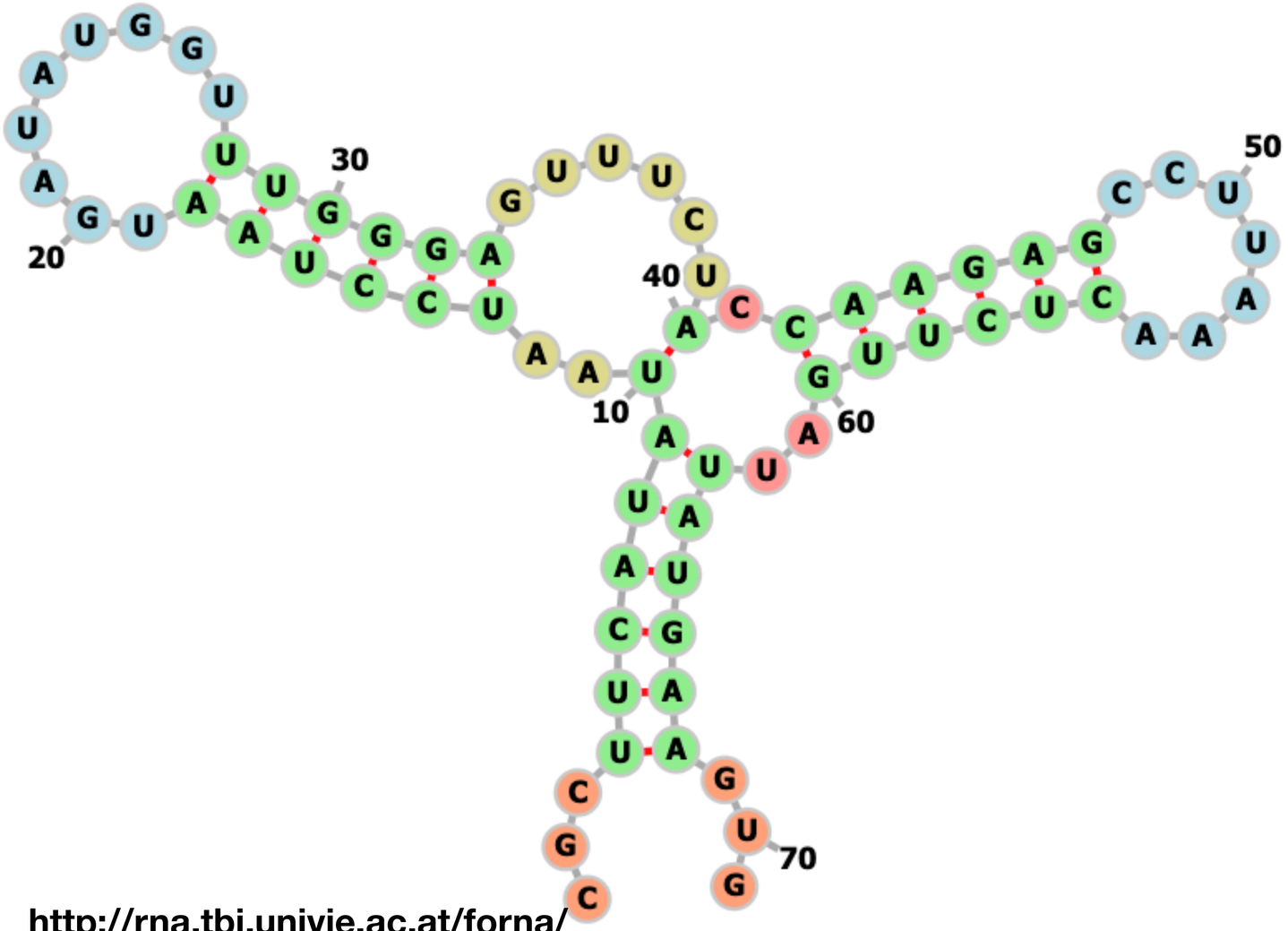
- 遺伝子の発現変動は多くの配列モチーフによって大部分が規定
- DEの事前分布にも特定の配列モチーフ

1. 配列解析から統計モデルへ

- ゲノム配列が遺伝子の様々な特性を決定する
  - そこに存在するバリエーション
- RNA二次構造予測
- ココノオビアルマジロの四つ子間の確率的変動の観測



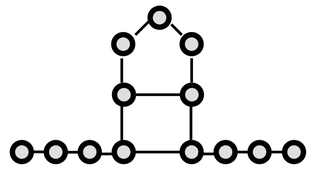
RNA二次構造とは？



http://rna.tbi.univie.ac.at/forna/

- RNAの相補的な塩基間に塩基対を形成
  - A-U, C-G + G-U
- スタッキングエネルギーがRNAの安定性に寄与 → **配列**に対する構造予測
- 他分子との結合のエネルギー障壁
  - RNA-RNA, RNA-DNA, RNA-RNA結合タンパク質

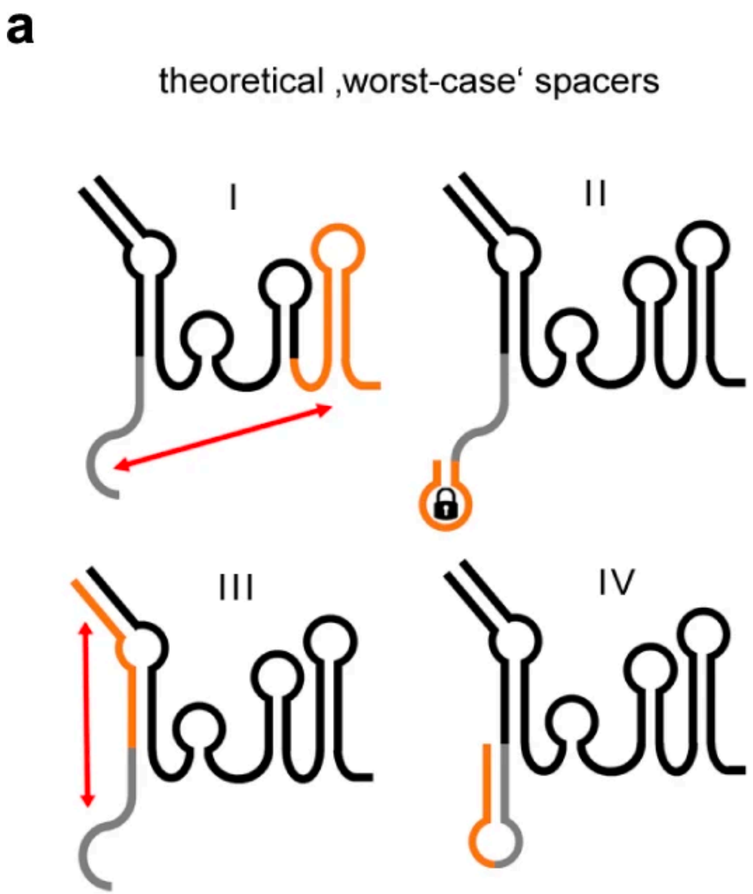
Example



...((...))...

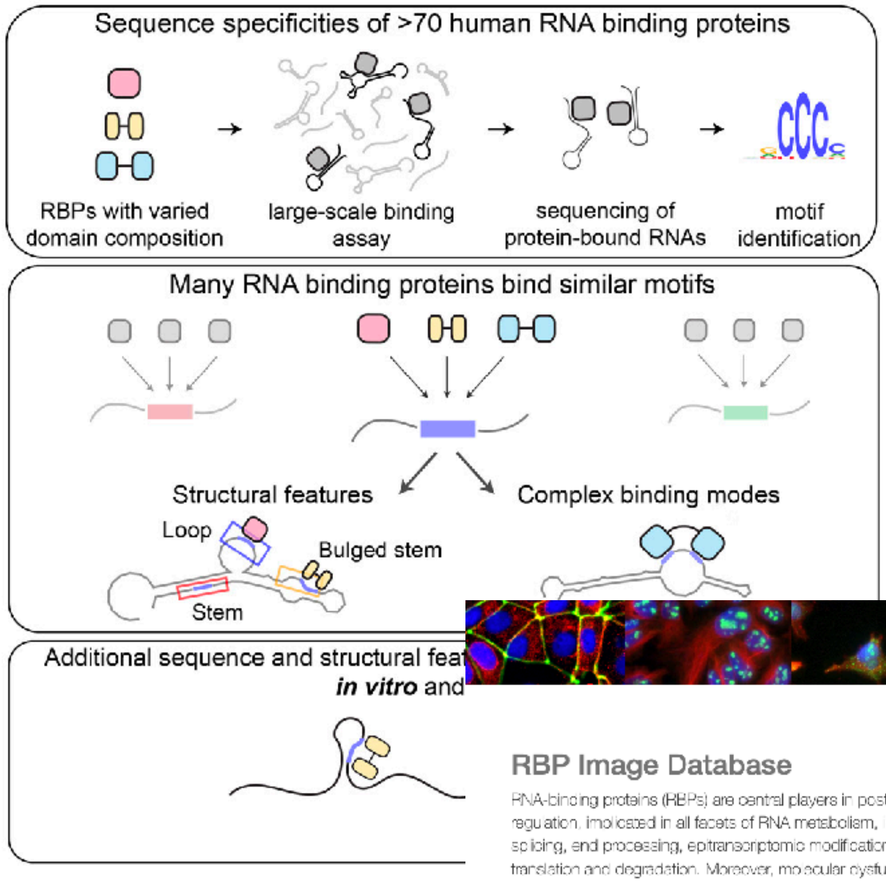
ハイスループット解析によるRNA構造の影響の解明

CRISPR gRNA 効率



Riesenberg S, et al. *Nature comm*, 2022.

RNA binding protein



Dominguez D, et al. *Molecular cell,* 2018

Van Nostrand EL, et al. *Nature* 2020. Bouvrette LPB, et al. *NAR*, 2022.

RNA二次構造予測モデルの歴史

塩基対があるほど安定 (Nussinov, 1978)

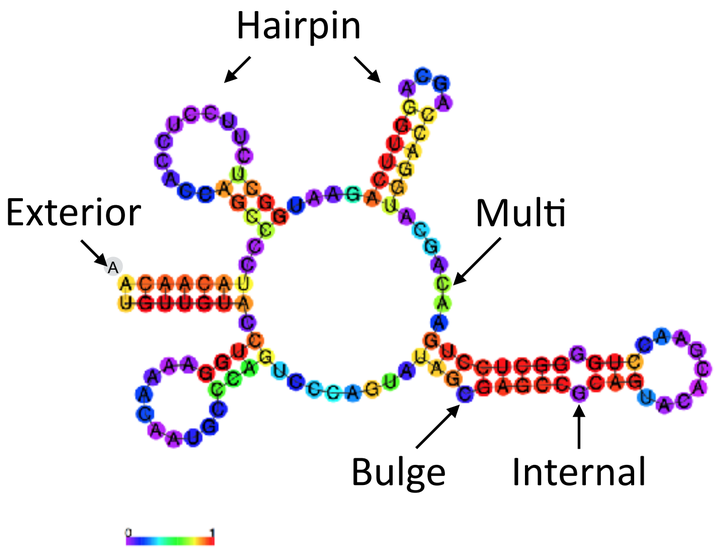
→ 安定な塩基対ほど安定 (Zuker, 1989)

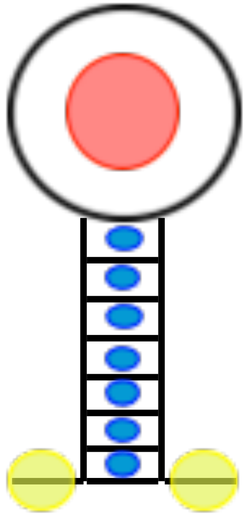
→ 部分構造から得られる熱力学的なモデル (Mathews, 1999)

→ 既知構造から推定されたパラメータ (Andronescu, 2010. Zakov, 2011.)

→ プロービングデータなど大規模データやディープラーニングの組み合わせ （SPOT-RNA 2019. MXfold2 2021.）

構造一つ一つに対し
自由エネルギー変化ΔGを計算





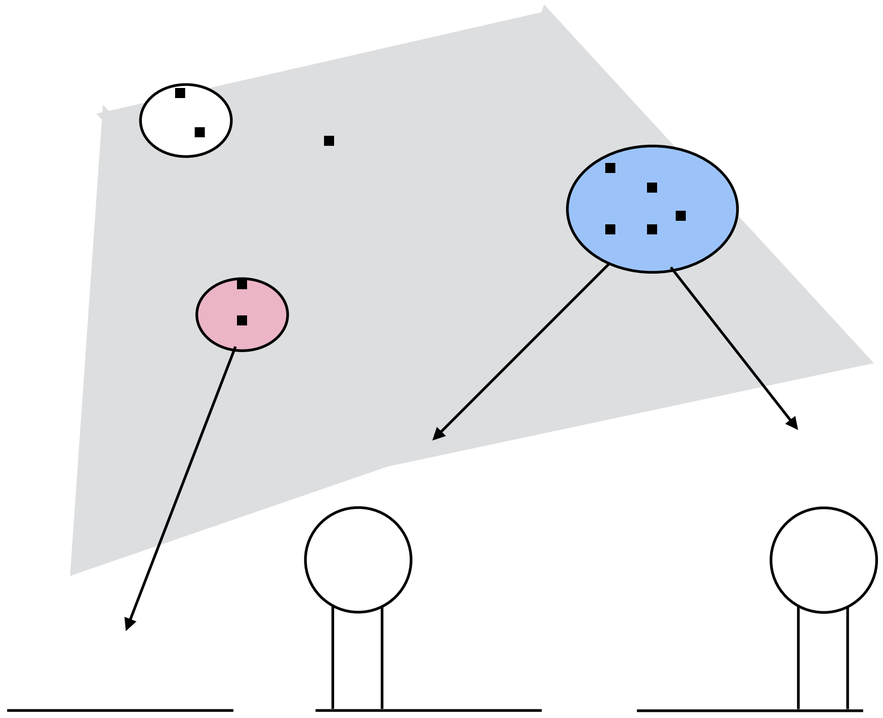
ΔG = Hairpin loop
+ Stacking loop
+ Dangling energy

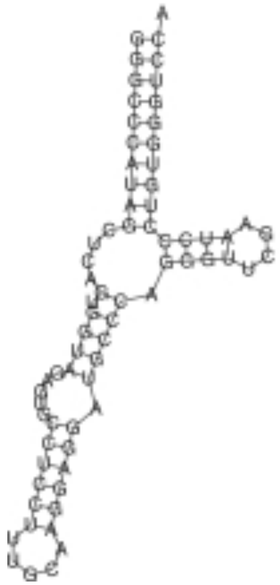
RNA二次構造予測の向上：最安定→期待精度最大

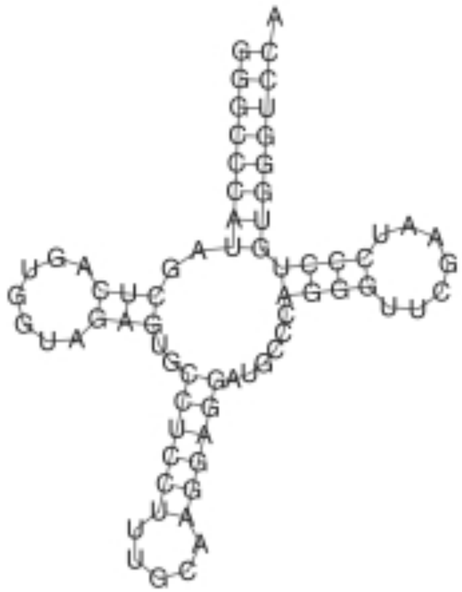
可能な構造集合全体

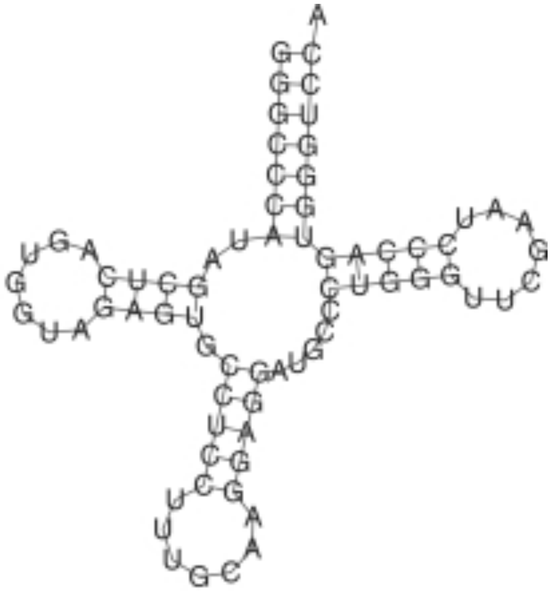
最適な安定構造の選択

安定構造の網羅的な評価









Sato K, et al. *NAR*, 2009.

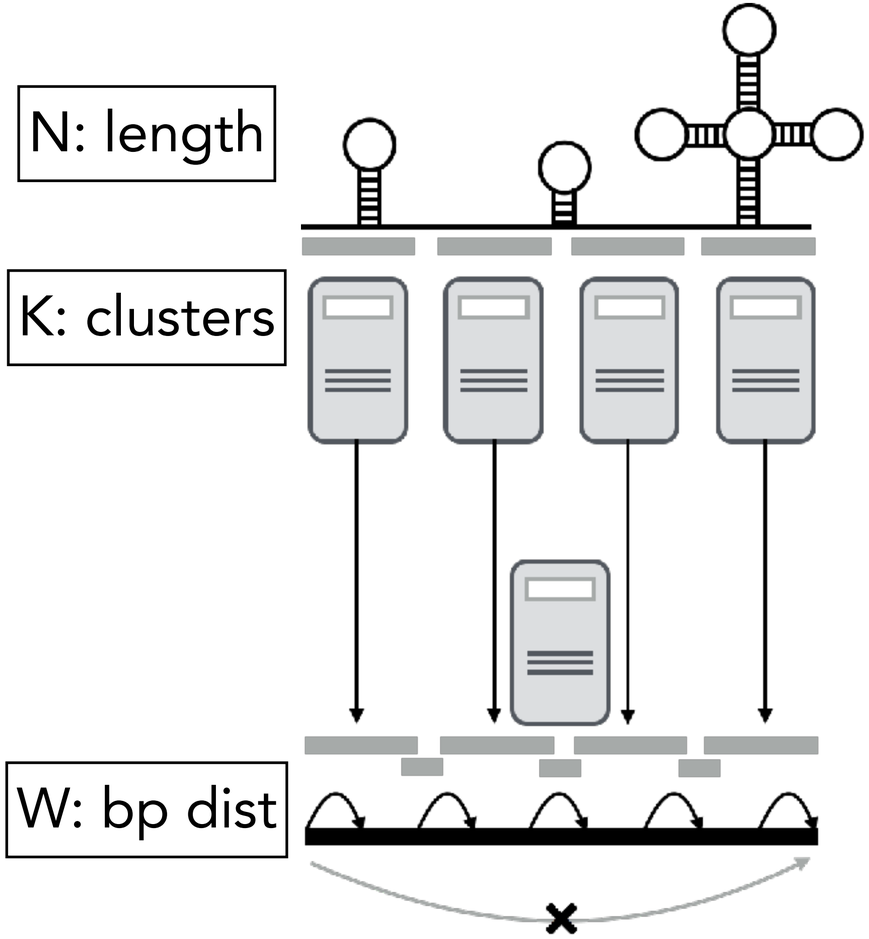
参考：100 ntの配列がとりうる構造 ≒ 10^{55}？

動的計画法 (DP) を適用することで効率的に

構造集合全体に対する期待値が計算できる

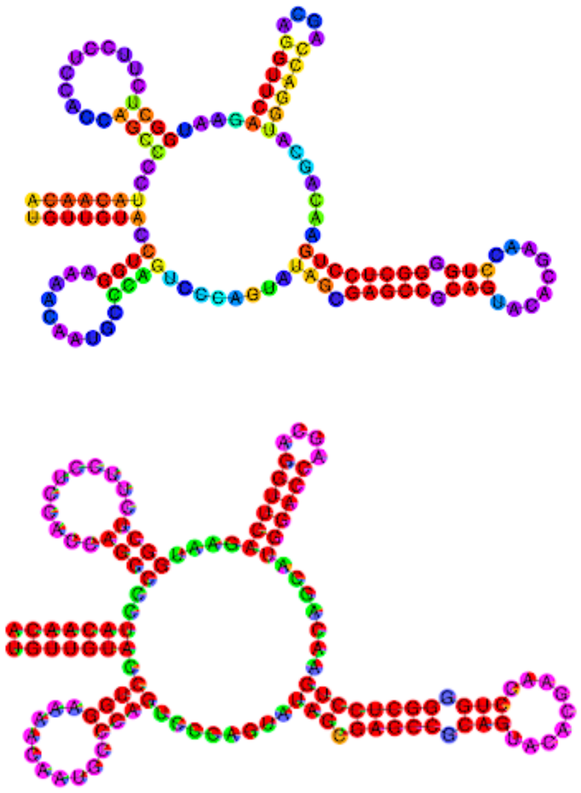
ParasoR

Kawaguchi R. et al. 2016. *BMC Bioinformatics*.
Download from https://github.com/carushi/ParasoR



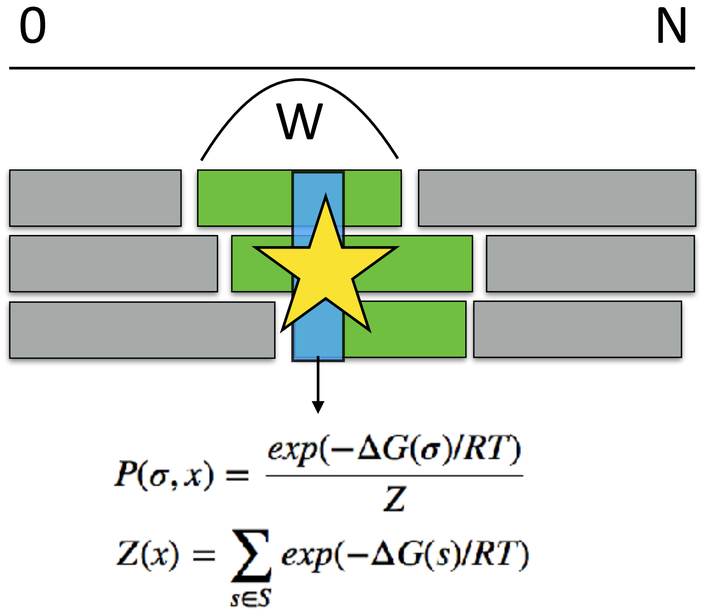
塩基対間距離の制限 (W) ＋並列計算

→ ヒトゲノムレベルの長さの構造予測



pre-mRNAの一部領域の

構造プロファイル可視化



変異による構造変化の

高速シミュレーション

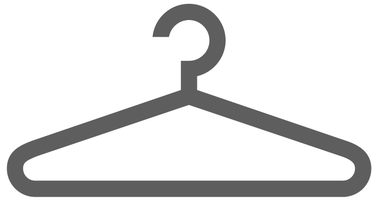
Kawaguchi RK and Kiryu H. *in press.*

in vivo構造は予測構造から大きく変化しうる

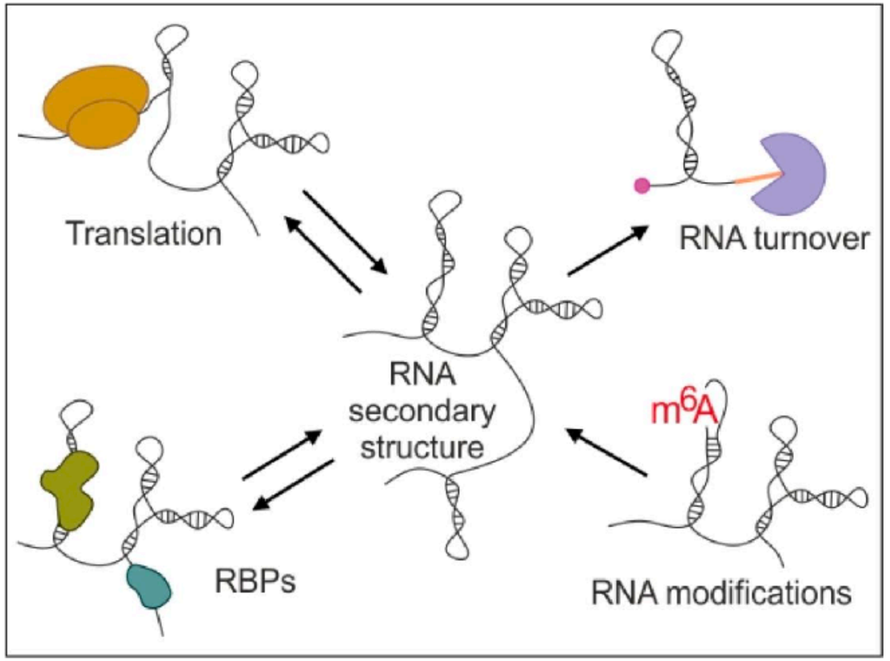
思ってたんと違う…

*in silico*予測

実際の構造



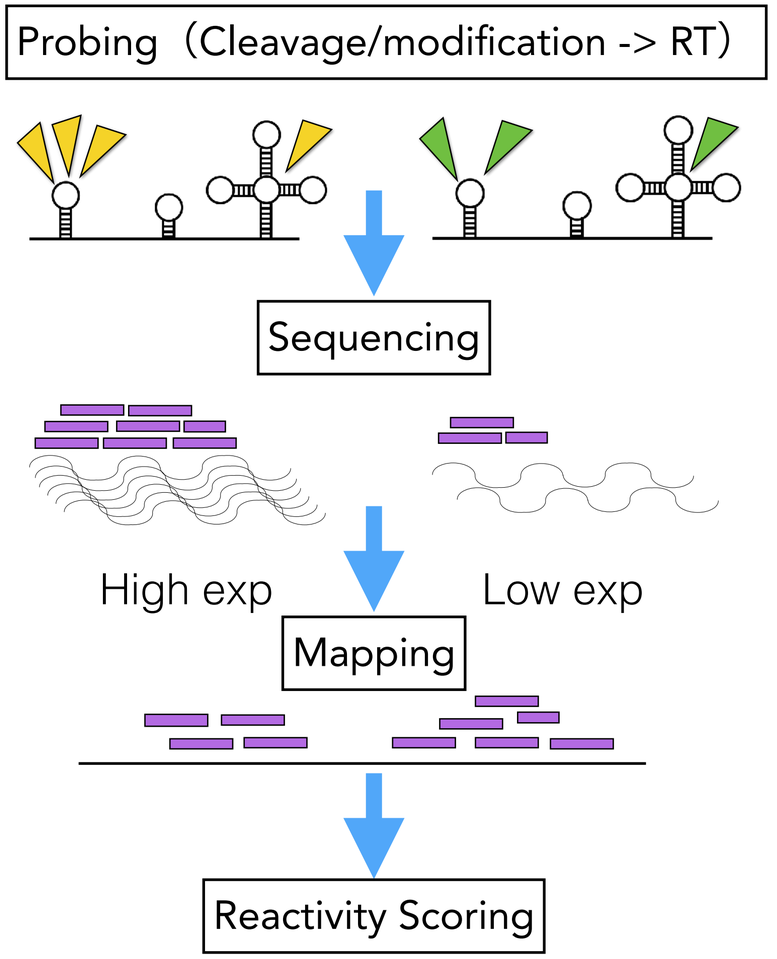
要因：リボソームなどの結合タンパクによるunwound、RNA修飾などなど…



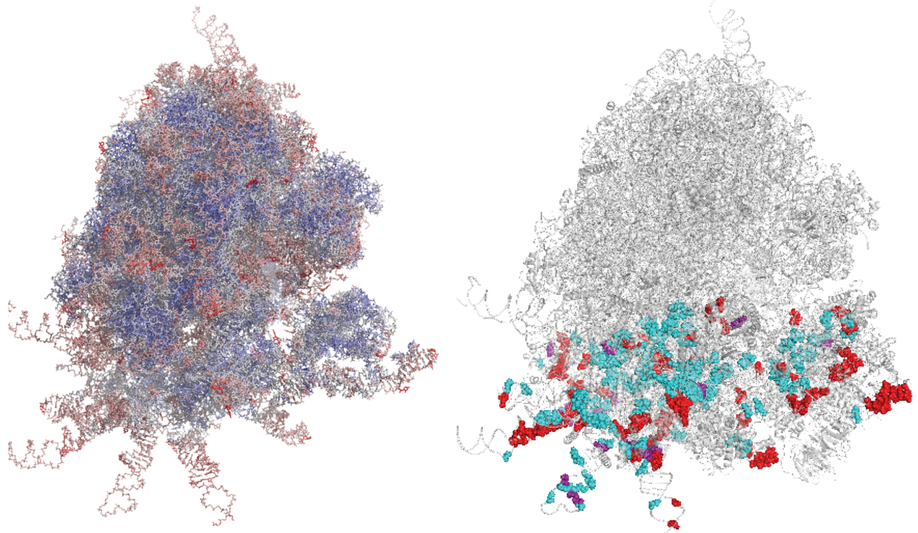
Andrzejewska A, et al. *Molecular Sciences*, 2020.

実験的構造プロービング法によるin vitro/vivo二次構造解析

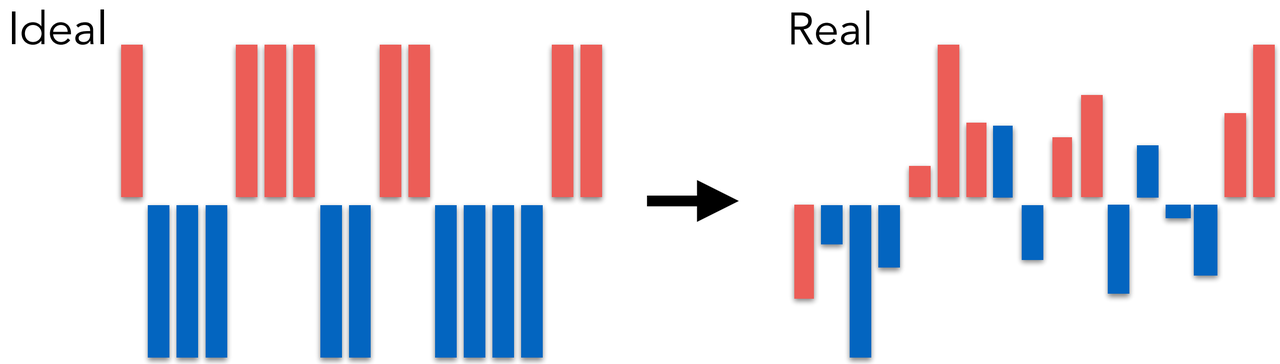
- ハイスループットな構造プロービング法（2010-）
  - 二次構造やアクセシビリティ依存的に修飾
  - シーケンシングによる領域の特定（RTの停止・切断）
- in vivo/vitro間の構造差の可視化



in vivo and in vitro



Kawaguchi R. et al. 2019. *BMC Bioinformatics*.



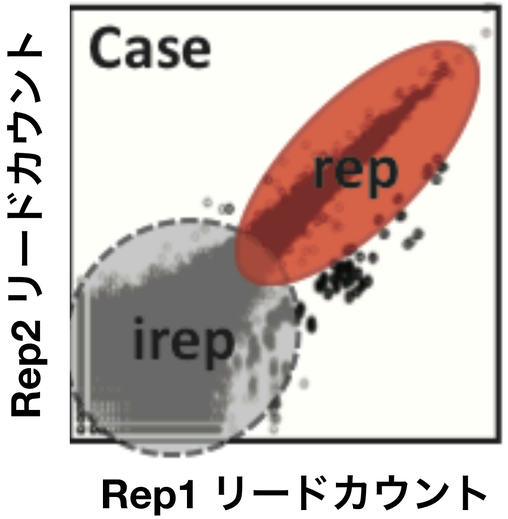
reactIDR

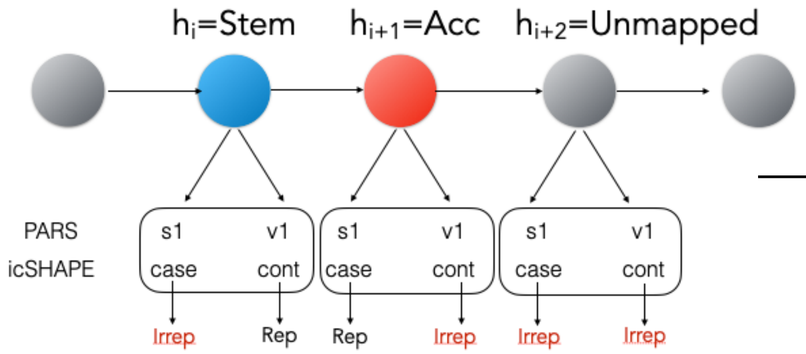
Kawaguchi R. et al. 2019. *BMC Bioinformatics*.
Download from https://github.com/carushi/reactIDR

Irreproducible Discovery Rate (IDR)　＋隠れマルコフモデル（HMM）

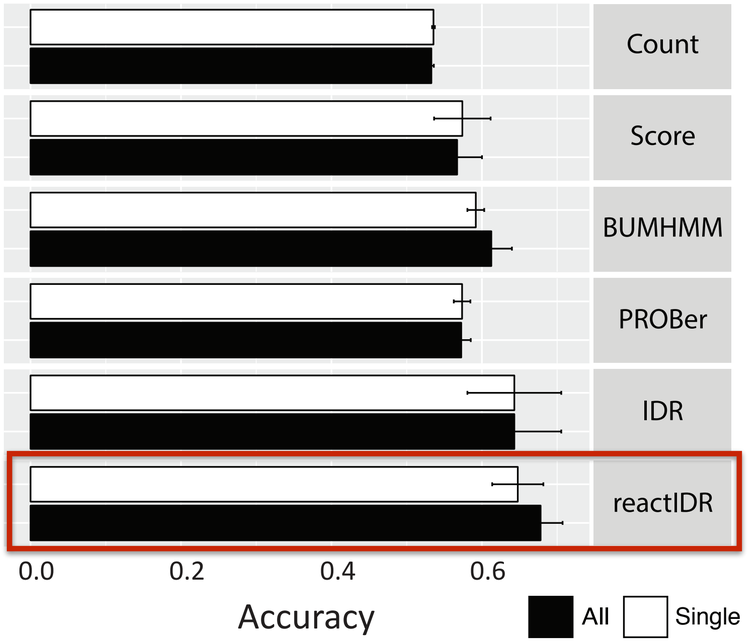
レプリケイトの間での非再現性確率

修飾が入る場所は連続しやすい



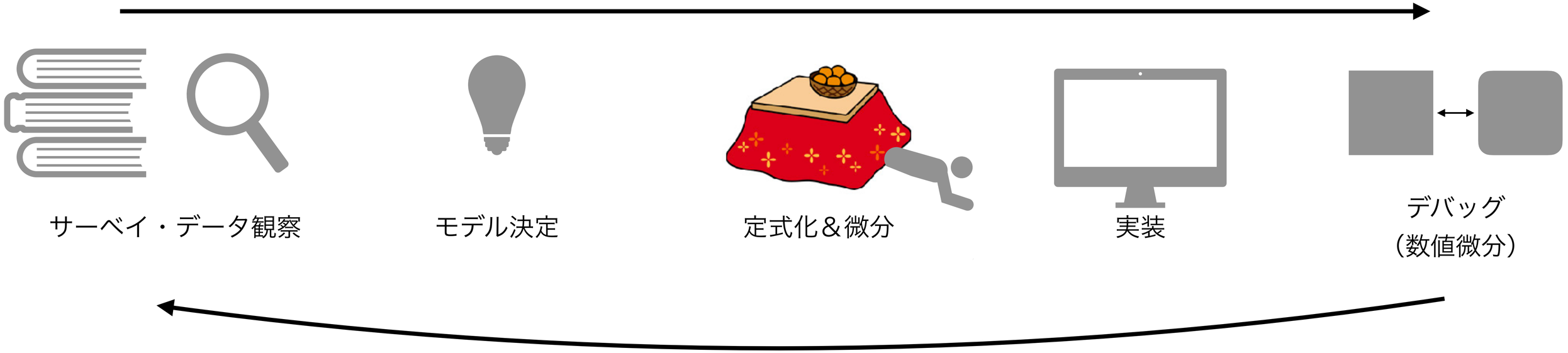


Best accuracy



- ChIP-seqで用いられるコピュラモデルによる順位相関のモデル化
- HMMにより連続的に修飾のある（ループ領域・ステム領域）場所を推定

統計手法開発をするときのルーティン



1. RNAの構造プロービングデータを解析→再現性問題
2. 再現性をモデル化する手法（IDR）を見つける
3. 検証データに当てはめてみる
4. プロービングデータの特性（連続性）を考慮する必要→隠れマルコフモデル（HMM）を追加

Model optimization: EM algorithm for reactIDR

```python
def lbfgs_qfunc(alpha, *args):
    global APPROX_GRAD
    index, sindex, theta, hmm = args[0], args[1], args[2], args[3]
    theta[index] = float(alpha)
    if APPROX_GRAD:
        update_amount = -hmm.q_function_grad_single_variable(sindex, index)
        return update_amount
    else:
        value = -hmm.q_function(sindex, theta)
        update_amount = -hmm.q_function_grad_single_variable(sindex, index)
        return value, np.array([update_amount])

def EM_LBFGS_step(self, sindex, init_theta, init_lhd, fix_mu, fix_sigma, eps, new_lhd = [0.0]):
        global APPROX_GRAD
        prev_theta, prev_lhd = init_theta, init_lhd
        min_vals, max_vals, max_change = bound_variables()
        for index, (min_val, max_val) in enumerate(zip(min_vals, max_vals)):
            if index == 0 and fix_mu: continue
            if index == 1 and fix_sigma: continue
            theta = prev_theta.copy()
            alpha = fmin_l_bfgs_b(lbfgs_qfunc, theta[index], approx_grad=APPROX_GRAD, args=(index, sindex, theta, self.HMM), bounds=[(min_vals[index], max_vals[index])], factr=eps)
            theta[index] = float(alpha[0][0])
            if len(new_lhd) < 3 or new_lhd[index] + eps >= prev_lhd:
                theta, changed_params = clip_model_params(theta)
                prev_theta = theta
        sys.stdout.flush()
        return prev_theta, max(new_lhd)

def EM_iteration_grad(self, iter_count, N, lhd, fix_mu, fix_sigma, alpha):
    break_flag = False
    thetas, pseudo_lhds = [], []
    for j in range(len(self.v)):
        prev_theta = self.get_IDR_params(j)
        theta, pseudo_lhd = self.EM_LBFGS_step(j, prev_theta, lhd, fix_mu=fix_mu, fix_sigma=fix_sigma, eps=1e7) # (extremely high precision), or 1e7 for moderate accuracy
        thetas.append(copy.deepcopy(theta))
        pseudo_lhds.append(pseudo_lhd)
    for j in range(len(self.v)):
        prev_theta = self.get_IDR_params(j)
        sum_param_change, mean_pseudo_val_change = self.check_value_change(iter_count, j, prev_theta, thetas[j], pseudo_lhds[j])
        if not (iter_count > N/2. and (sum_param_change < EPS and mean_pseudo_val_change < EPS)):
            pass
        else:
            break_flag = True
    return break_flag
```

Qの各パラメータによる一回微分

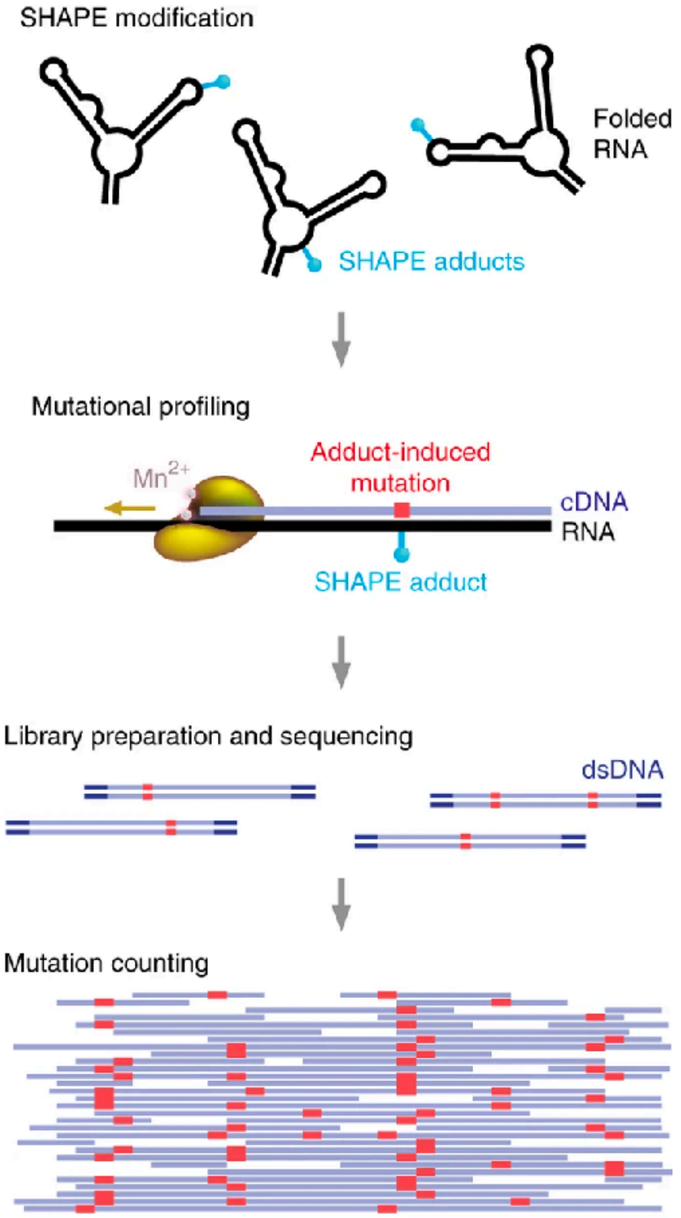
→ 各パラメータの微分計算はcythonで高速化

L-BFGSによる最適化

（一階微分・二階微分を受け取る）

EMイテレーション一回

切断・RTストップからMutational profiling（MaP）へ

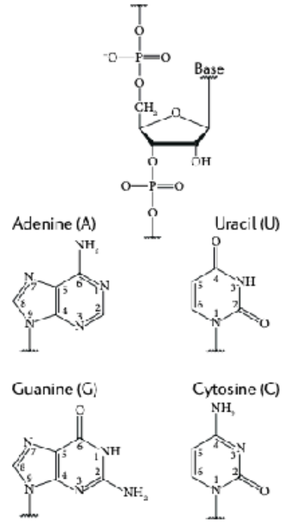


Siegfried NA, et al. *Nature Methods*, 2014

- ポリメラーゼの最適化による複数箇所の同時検出
- 修飾のための試薬選択

Strobel EJ, et al. *Nat Rev Genet*, 2018

	Probe	Primary modification sites	Half-life at 37°C	Refs
SHAPE	*N*-methylisatoic anhydride (NMIA)	2′ OH of all nts	260 s	30
	1-methyl-7-nitroisatoic anhydride (1M7)	2′ OH of all nts	14 s	76
	1-methyl-6-nitroisatoic anhydride (1M6)	2′ OH of all nts	31 s	77
	Benzoyl cyanide (BzCN)	2′ OH of all nts	0.25 s	78
	2-methylnicotinic acid imidazolide (NAI)	2′ OH of all nts	33 min	39
	2-methyl-3-furoic acid imidazolide (FAI)	2′ OH of all nts	73 min	39
	2-(azidomethyl)nicotinic acid imidazolide (NAI-N_3)	2′ OH of all nts	33 min	45
Base pairing	Dimethyl sulfate (DMS)	G N7, A N1 and C N3	Quenched	20, 70
	N-cyclohexyl-*N*′-(2-morpholinoethyl) carbodiimide metho-p-toluenesulfonate (CMCT)	G N1 and U N3	Quenched	61
	Kethoxal and other 1,2-dicarbonyl compounds	G N1 and C2-amine	Quenched	67
Solvent accessibility	Hydroxyl radical (•OH)	Backbone	Quenched	23, 26
	Nicotinoyl Azide (NAz)	G C8 and A C8	Solvent quenched, ps time scale	68



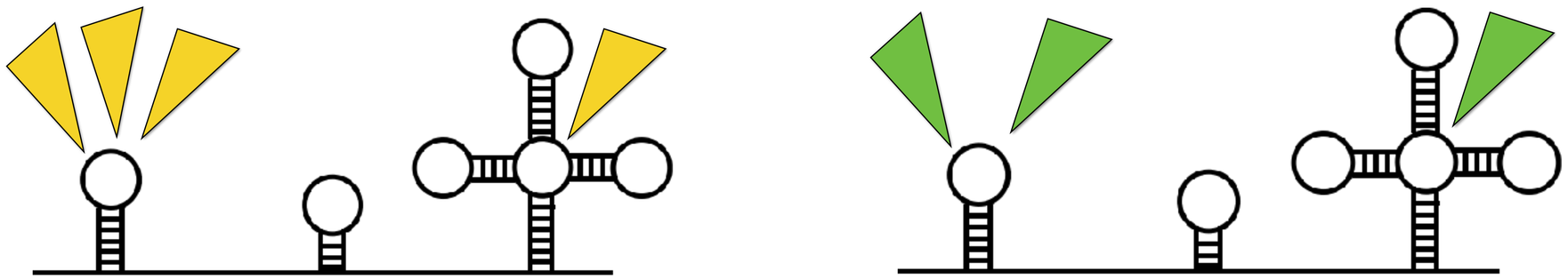
Carlson PD, et al. *Cell*. 2018

Probe	Structural feature probed	Molecular weight (g/mol)	Quenching half-life (37 °C)	Modifies	Used with in-cell probing?
1-methyl-7-nitroisatoic anhydride (1M7)	Nucleotide dynamics	222.2	14 s	2′ OH (all nts)	Y
Benzoyl cyanide (BzCN)	Nucleotide dynamics	131.1	0.25 s	2′ OH (all nts)	
2-methylnicotinic acid imidazolide (NAI)	Nucleotide dynamics	187.2	33 min	2′ OH (all nts)	Y
2-(azidomethyl)nicotinic acid imidazolide (NAI-N3)	Nucleotide dynamics	228.2	33 min	2′ OH (all nts)	Y
Dimethyl sulfate (DMS)	Base pairing context	126.1	User defined quenching	G, A, and C	Y
N-cyclohexyl-*N*′-(2-morpholinoethyl) carbodiimide metho-p-toluenesulfonate (CMCT)	Base pairing context	423.6	User defined quenching	G and U	
Kethoxal	Base pairing context	148.2	User defined quenching	G and C	
Hydroxyl radical	Solvent Accessibility	17.01	User defined quenching	Backbone	Y
Nicotinoyl Azide (NAz)	Solvent Accessibility	148.1	Solvent quenched, ps timescale	G and A	Y

ナノポア長鎖RNAのダイレクトシーケンシングによる構造プロービング

https://nanoporetech.com/

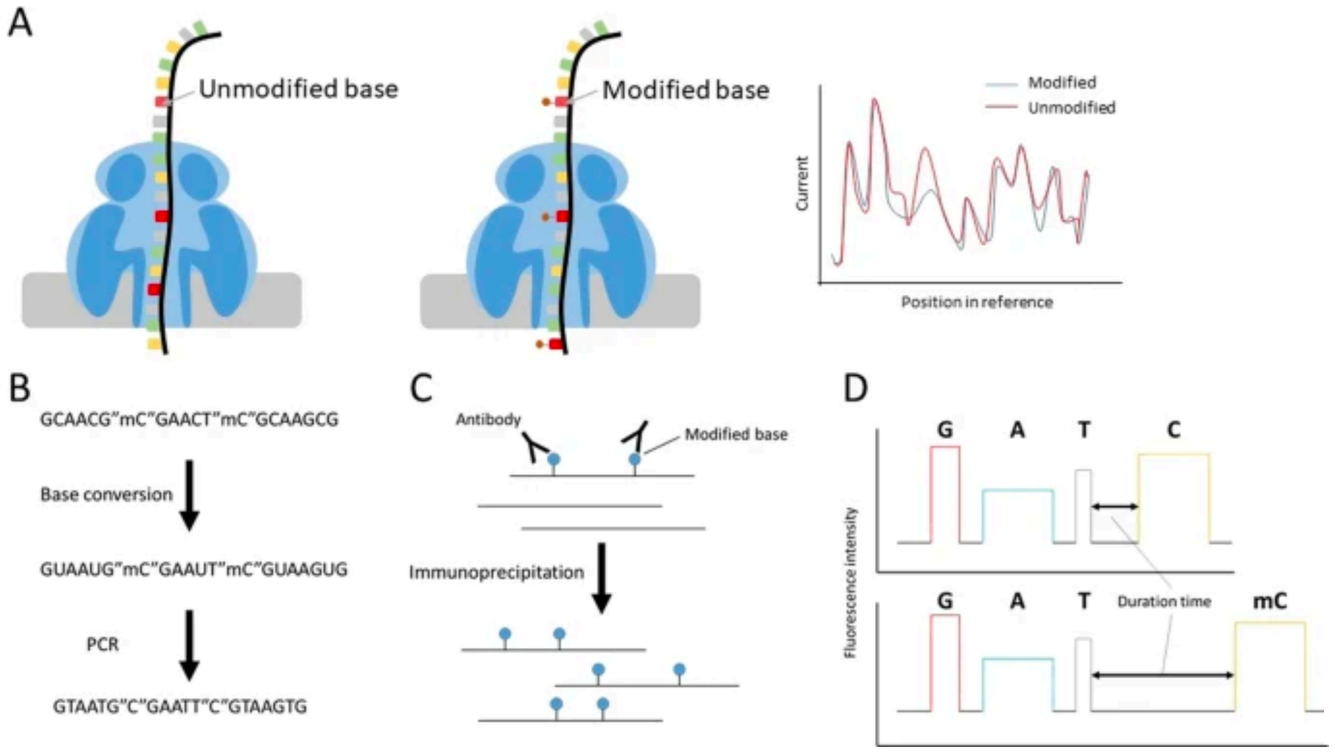
x



Review Article | Open Access | Published: 11 October 2019

Recent advances in the detection of base modifications using the Nanopore sequencer

Liu Xu & Masahide Seki



- RNAがポアを通過する際の電流・Duration timeの変化から修飾パターンを推定
- 周囲の配列依存性があるため多様な修飾にはまだ弱い

ARTICLES
https://doi.org/10.1038/s41587-020-0712-z
nature biotechnology
Check for updates

Determination of isoform-specific RNA structure with nanopore long reads

Jong Ghut Ashley Aw[1,9], Shaun W. Lim[1,9], Jia Xu Wang[1,9], Finnlay R. P. Lambert[1,2], Wen Ting Tan[1], Yang Shen[3], Yu Zhang[1], Pornchai Kaewsapsak[1], Chenhao Li[3], Sarah B. Ng[4], Leah A. Vardy[5], Meng How Tan[1,6], Niranjan Nagarajan[3,7] ✉ and Yue Wan[1,7,8] ✉

- Mar 2021.
- NAI-N3 modification

Cell Genomics
CellPress
OPEN ACCESS

Technology

Direct detection of RNA modifications and structure using single-molecule nanopore sequencing

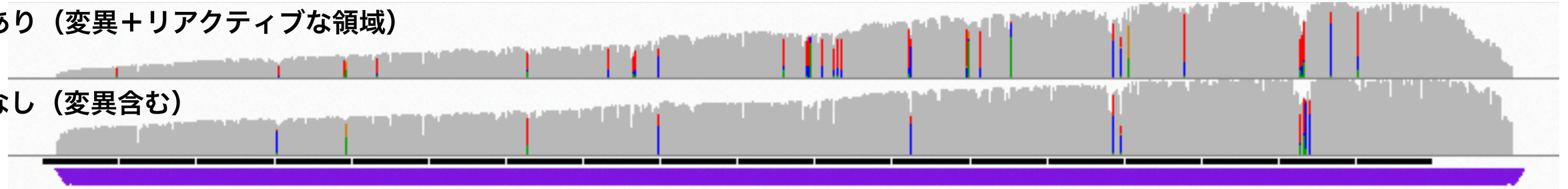
William Stephenson,[1,4,6,*] Roham Razaghi,[2] Steven Busan,[3] Kevin M. Weeks,[3] Winston Timp,[2] and Peter Smibert[1,5]
[1]Technology Innovation Lab, New York Genome Center, New York, NY, USA
[2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
[3]Department of Chemistry, University of North Carolina, Chapel Hill, NC, USA
[4]Present address: Genentech, South San Francisco, CA, USA
[5]Present address: 10x Genomics, Pleasanton, CA, USA
[6]Lead contact

- Feb. 2022.
- Aclm
- Data
  - Human pri-miRNA 17~92 (951nt)
  - S. Cerevisiae + E. Coli rRNA

ナノポアリードのアラインメントから修飾検出ワークフロー（実データ）

SHAPEあり（変異＋リアクティブな領域）

SHAPEなし（変異含む）

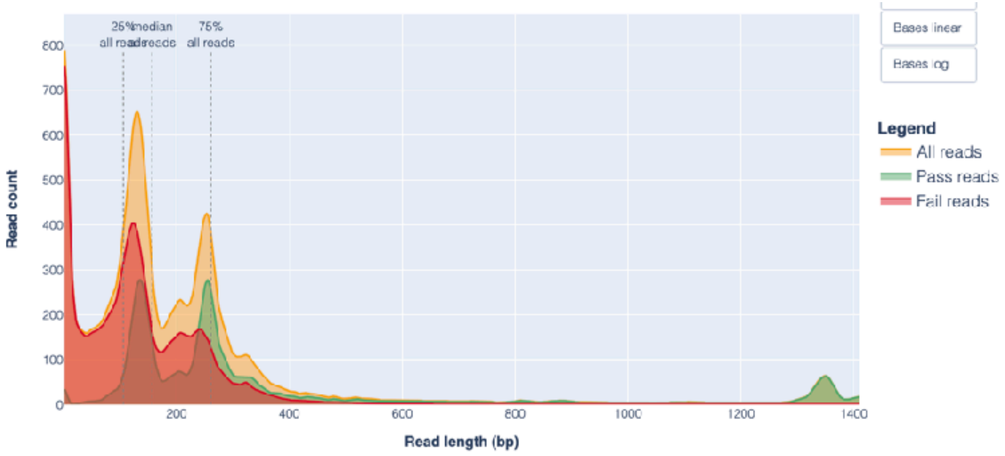


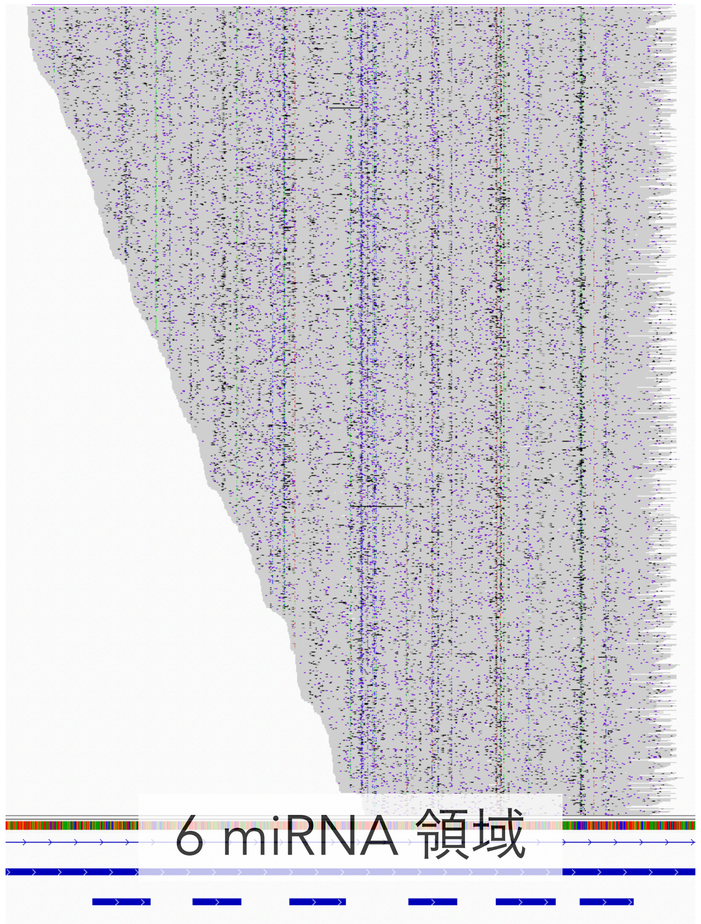
- エラーを許してアラインメント→再ベースコール
- SHAPEあり/なしの比較から統計的に修飾を検出（Tombo）

実験的にまだ多くの困難が伴う → 共同研究のお話待ってます！

複数Adductにより短い

Fail readsの極端な増加→





遺伝子配列では決まらないバリエーションもある

非遺伝要因

環境・外部刺激

初期発達時の
確率的なノイズ？

遺伝子配列では決まらないバリエーションもある

非遺伝要因

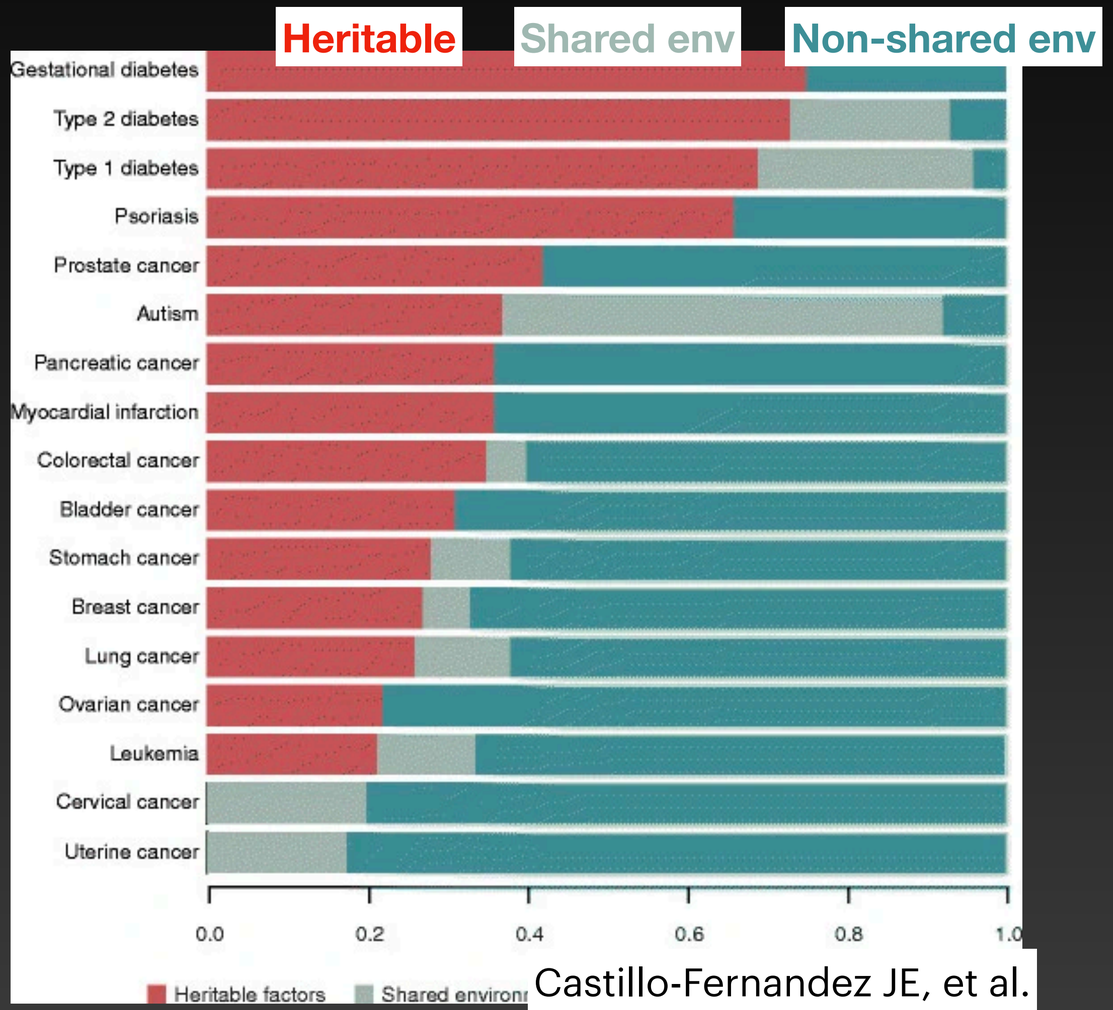
環境・外部刺激

初期発達時の

確率的なノイズ？

Human twin study
since 1875 (Francis Galton)

e.g., aging, cancer, autoimmune disease, psychiatric, and neurological traits



Castillo-Fernandez JE, et al.
Genome Medicine, 2014.

Phenotypic variation
- Genotype
- Environment
= ...?

https://www.nasa.gov/feature/nasa-twins-study-confirms-preliminary-findings

Phenotypic variation
- Genotype
- Environment
= ...?

RESEARCH ARTICLE

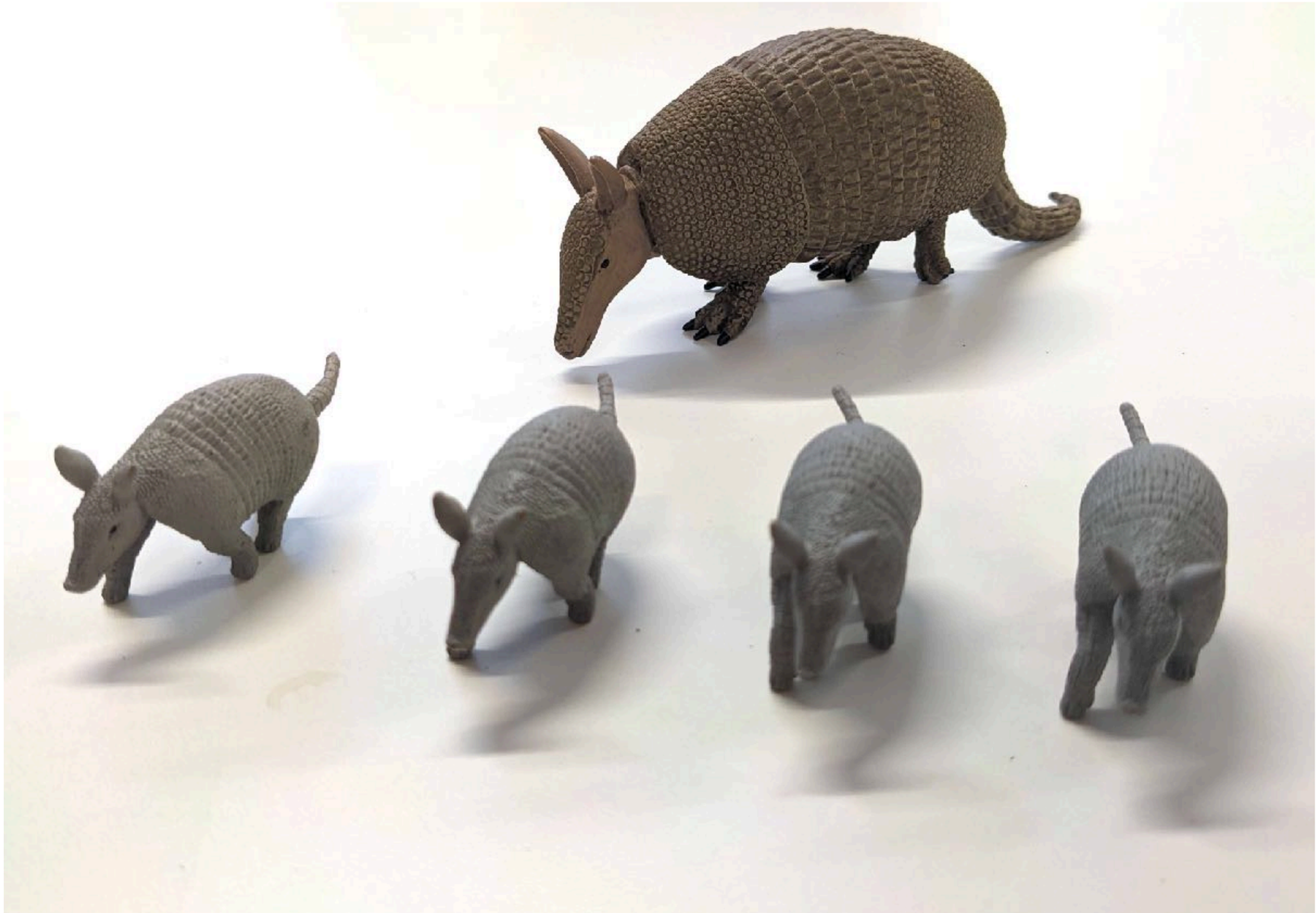
HUMAN PHYSIOLOGY

The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight

Francine E. Garrett-Bakelman[1,2*], Manjula Darshi[3*], Stefan J. Green[4*], Ruben C. Gur[5*], Ling Lin[6*], Brandon R. Macias[7*], Miles J. McKenna[8*], Cem Meydan[1,9*], Tejaswini Mishra[6*], Jad Nasrini[5*], Brian D. Piening[6*†], Lindsay F. Rizzardi[10*‡], Kumar Sharma[3*], Jamila H. Siamwala[11*§], Lynn Taylor[8*], Martha Hotz Vitaterna[12*], Maryam Afkarian[13], Ebrahim Afshinnekoo[1,9], Sara Ahadi[6], Aditya Ambati[6], Maneesh Arya[7], Daniela Bezdan[1,9], Colin M. Callahan[10], Songjie Chen[6], Augustine M. K. Choi[1], George E. Chlipala[4], Kévin Contrepois[6], Marisa Covington[14], Brian E. Crucian[14], Immaculata De Vivo[15], David F. Dinges[5], Douglas J. Ebert[7], Jason I. Feinberg[10], Jorge A. Gandara[1], Kerry A. George[7], John Goutsias[10], George S. Grills[1¶], Alan R. Hargens[11], Martina Heer[16#], Ryan P. Hillary[6], Andrew N. Hoofnagle[17], Vivian Y. H. Hook[11], Garrett Jenkinson[10**], Peng Jiang[12], Ali Keshavarzian[18], Steven S. Laurie[7], Brittany Lee-McMullen[6], Sarah B. Lumpkins[19], Matthew MacKay[1], Mark G. Maienschein-Cline[3], Ari M. Melnick[1], Tyler M. Moore[5], Kiichi Nakahira[1††], Hemal H. Patel[11], Robert Pietrzyk[7], Varsha Rao[6], Rintaro Saito[11‡‡], Denis N. Salins[6], Jan M. Schilling[11], Dorothy D. Sears[11], Caroline K. Sheridan[1], Michael B. Stenger[14], Rakel Tryggvadottir[10], Alexander E. Urban[6], Tomas Vaisar[17], Benjamin Van Espen[11], Jing Zhang[6], Michael G. Ziegler[11], Sara R. Zwart[20], John B. Charles[14§§], Craig E. Kundrot[21§§], Graham B. I. Scott[22§§], Susan M. Bailey[8§§], Mathias Basner[5§§], Andrew P. Feinberg[10§§], Stuart M. C. Lee[7§§], Christopher E. Mason[1,9,23,24§§], Emmanuel Mignot[6§§], Brinda K. Rana[11§§], Scott M. Smith[14§§], Michael P. Snyder[6§§], Fred W. Turek[19§§]

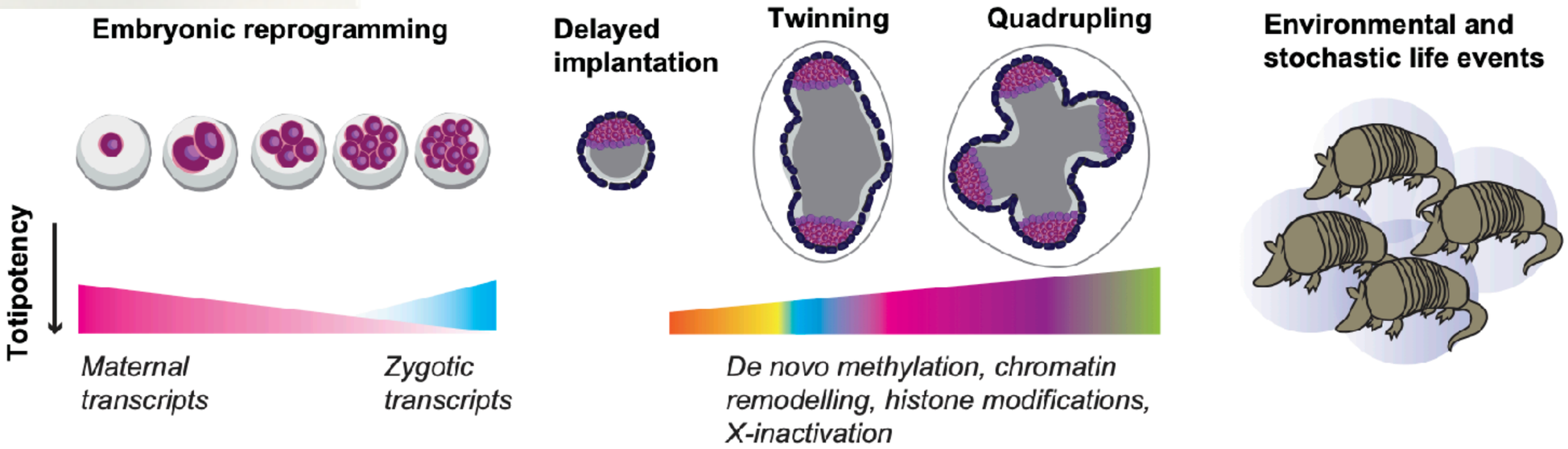
Concept: quantify unknown developmental stochasticity
under tight control of **genotype** and **environment**

Polyembryony - Genetically identical quadruplets of nine-banded armadillo



How to study armadillo quadruplets from birth

- Not yet possible to let armadillos mate in lab colony
- Pregnant females are caught from wild
  - Around 1 year delay of implantation
  - Trace the life history of quadruplets from birth



Same genotype still can produce (less) diverse variations

Non-heritable factors

~~Environmental~~

Stochastic “noise” at early development?

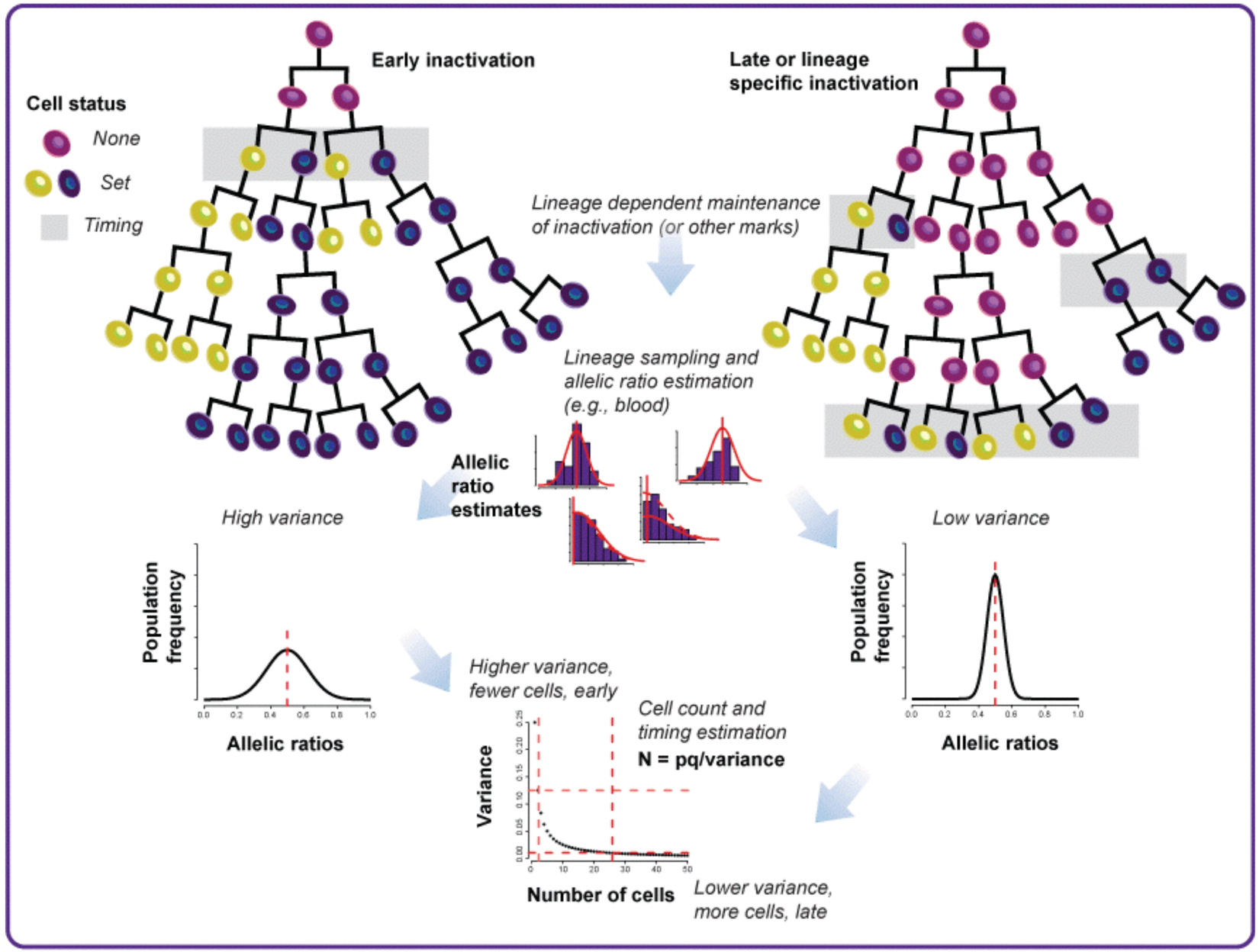
Under well-controlled environment (lab colony)

Who is who?

Ballouz S†, Kawaguchi RK†, et al. (in revision)

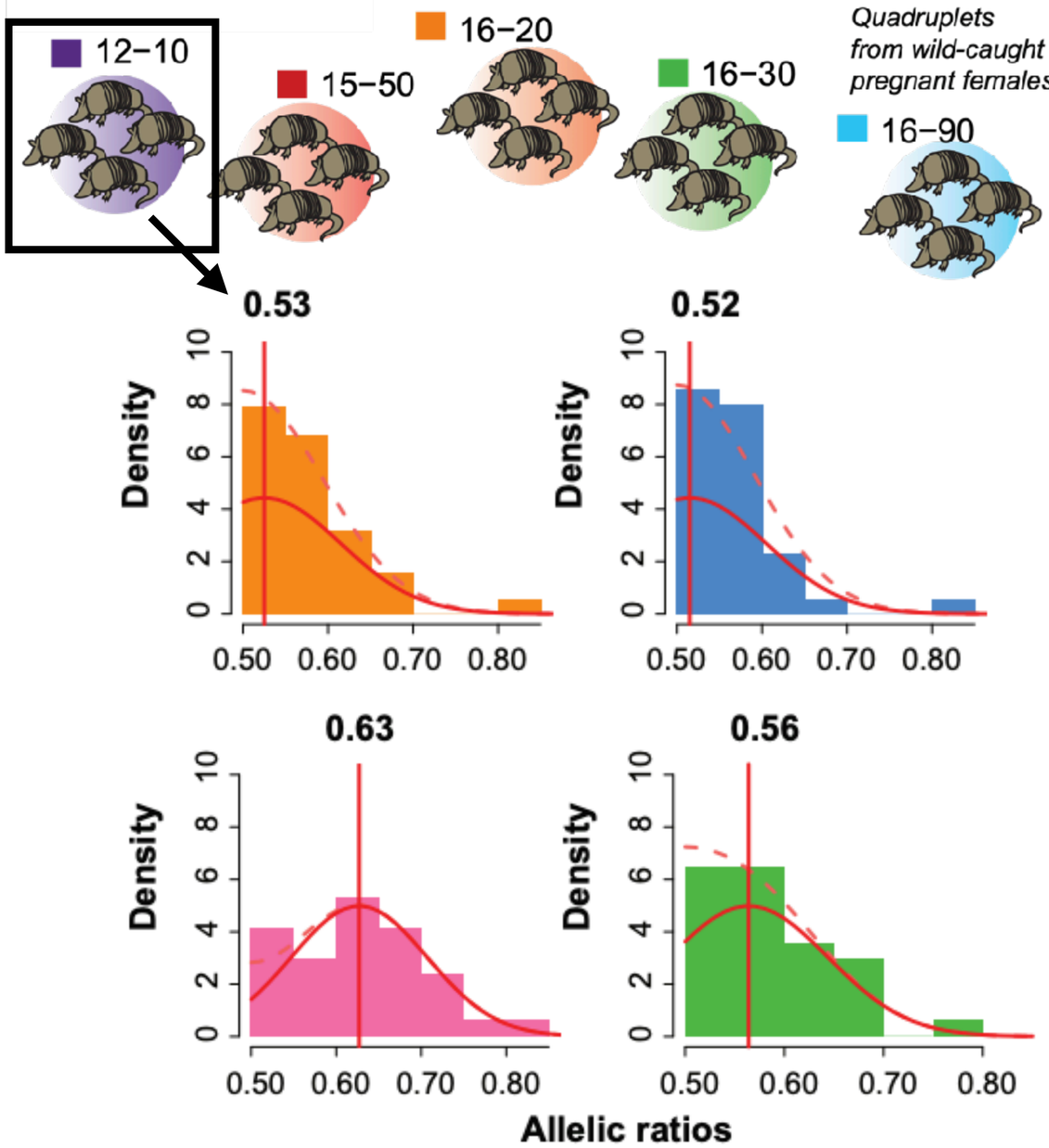
Concept: quantify canalized developmental stochasticity using **armadillo quadruplets** under tight control of **genotype** and **environment**

ランダムなX染色体の不活化が生み出すバリエーション

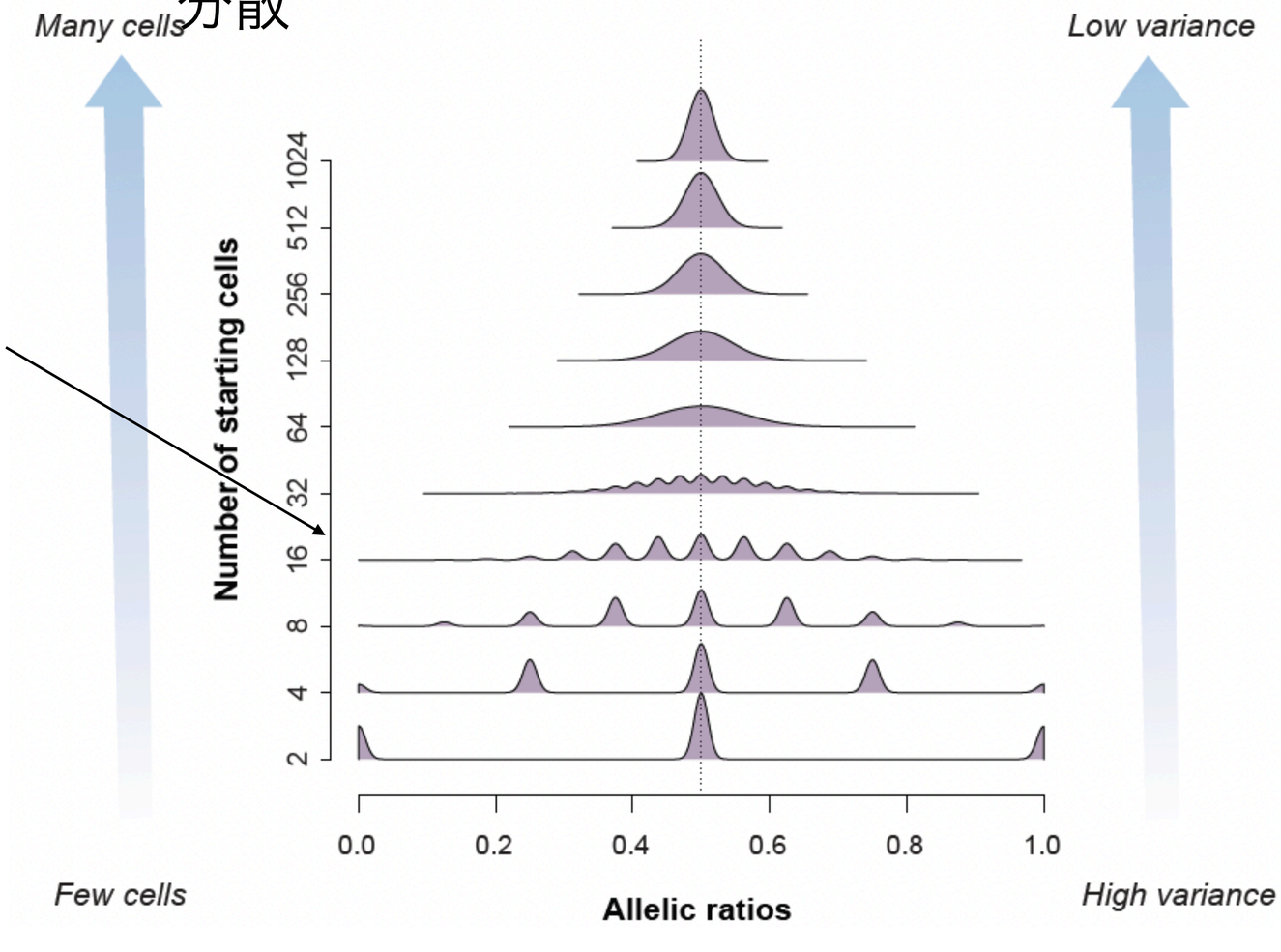


- 着床前後でX染色体のうちの一本が確率0.5で不活化
  - 初期：分散 大（細胞数 少）
  - 後期：分散 小
- X染色体上の遺伝子の**アレル特異的発現**
  - 偏り度合いを推定

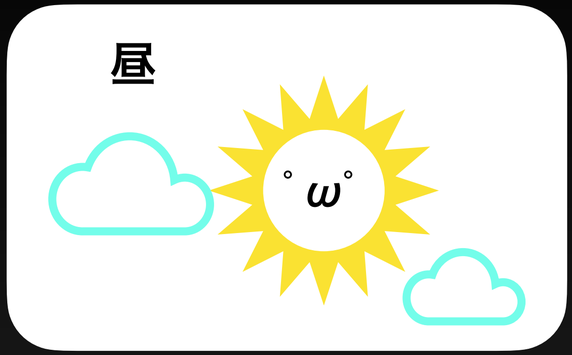
X染色体の不活化の分散から有効細胞数の推定



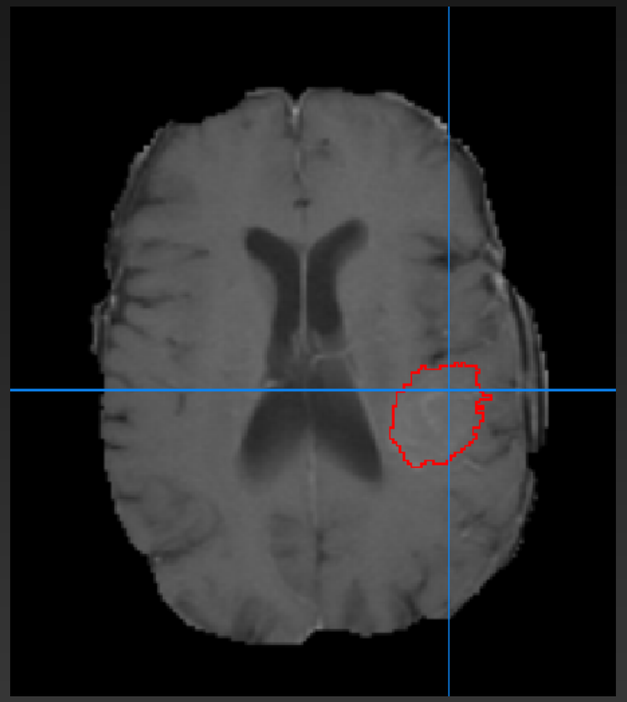
- 実際に四つ子間での偏りを観測
- 約25細胞で不活化が起きた場合に最も近い分散



2. ラジオゲノミクスにおける機械学習の実応用



- MRI画像解析から学ぶヘテロなデータの取り扱い
- 社会実装に沿った機械学習ワークフローデザイン
- 検証データにおける十分な多様性の確保

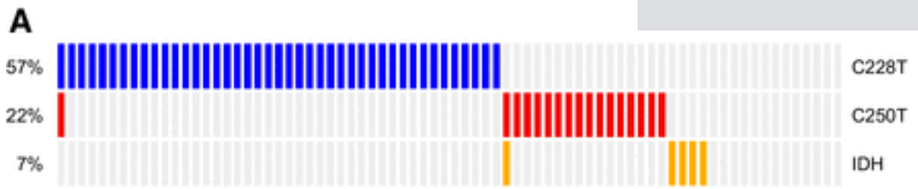


神経膠腫のラジオゲノミクス

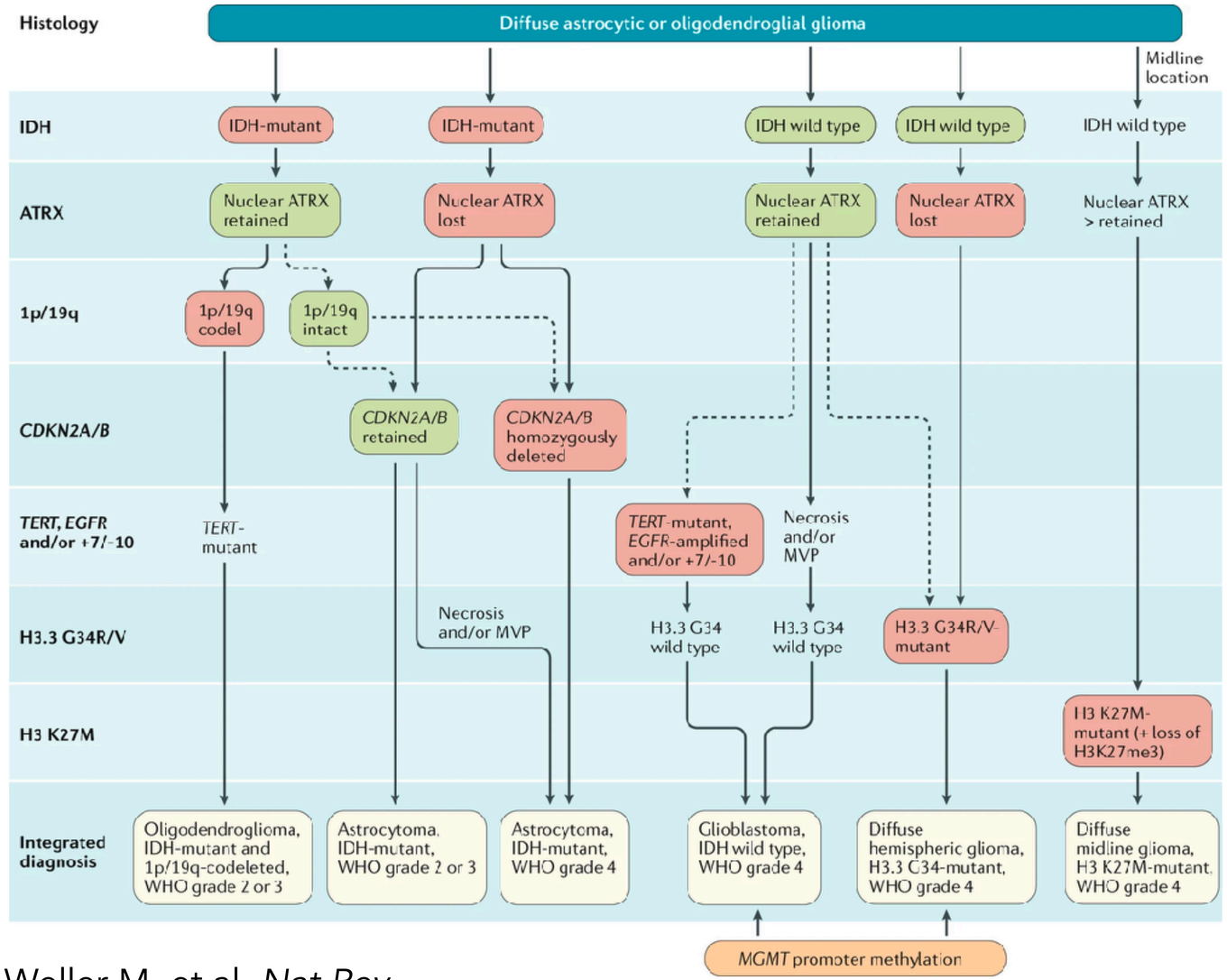
- グリオーマ（神経膠腫）ー 脳原発腫瘍
- 1-4のグレードに分類
  - グレード1-3: Low grade glioma （LrGG）
  - グレード4: グリオブラストーマ（GBM）

    5年後生存率 6.8%
- European Association of Neuro-Oncology (EANO) ガイドライン
  - 遺伝子情報とMRIの画像情報による診断

- GBM: IDH mutation 7%
- LrGG: IDH mutation 80%



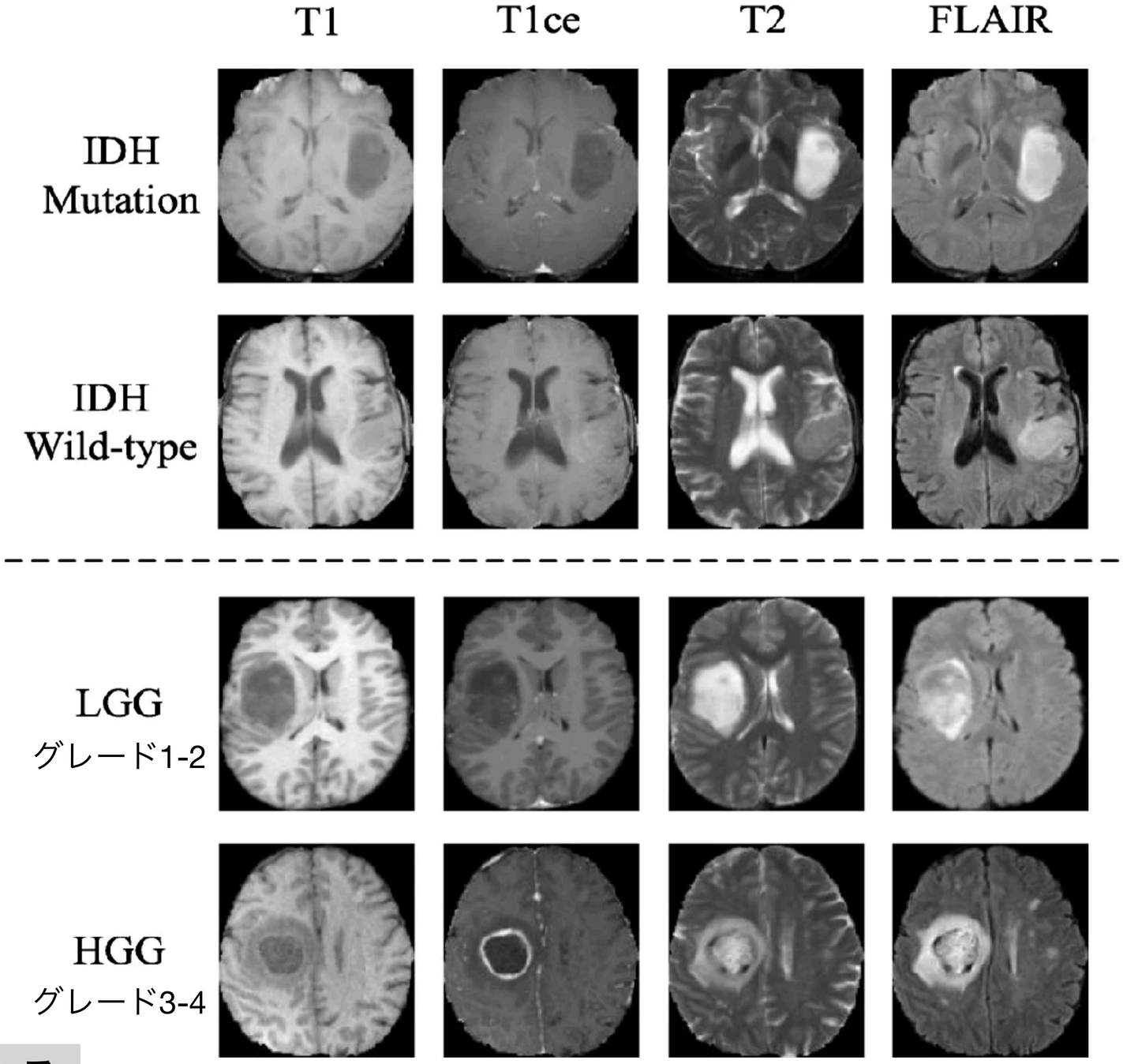
Eva Schulze Heuling, et al. *Molecular Cancer Research*, 2017.



Weller M, et al. *Nat Rev Clin Onco*, 2020.

MRI画像特徴

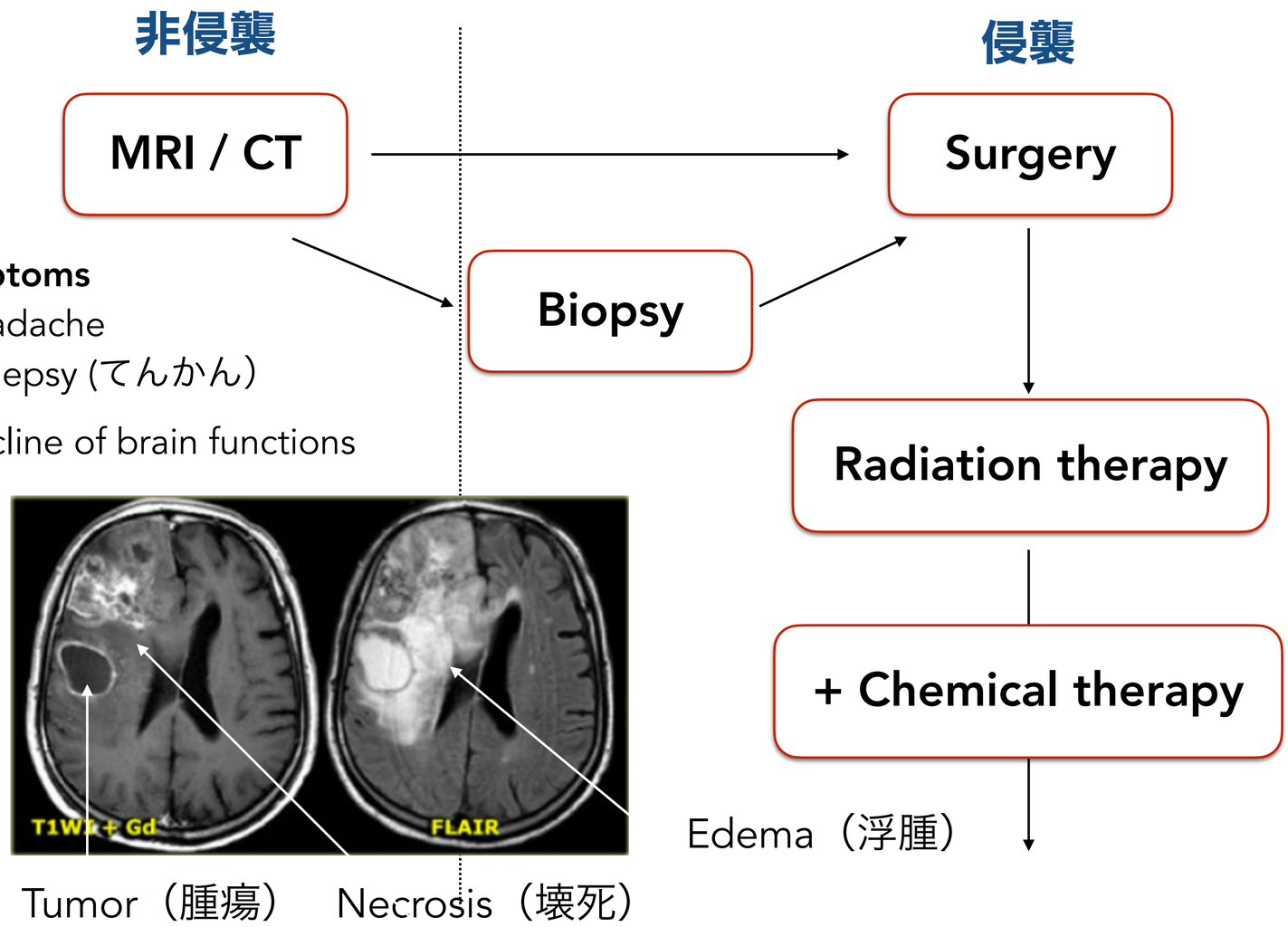
- T1緩和/T2緩和
  - 磁場のかけ方によりプロトンの量に応じて明るい場所が変化
  - T1：脂肪（Gdが腫瘍に溜まる）
  - T2 水
  - FLAIR 水抑制
- **LrGG**：T2/FLAIRで浮腫
- **HGG**：T1-contrast enhancedで輪郭強調・ネクローシス



腫瘍の特性と関連することが期待されている

Ge C, et al. *BMC Medical Imaging*, 2020.

グリオーマの診断・治療ワークフローとラジオゲノミクスの可能性



https://radiologyassistant.nl/neuroradiology/brain-tumor/systematic-approach

- 生検による遺伝子検査には時間がかかる
- 腫瘍内での非均一性も問題
- より非侵襲的な手法で悪性度や予後関連のバイオマーカーの予測ができないか？

Article

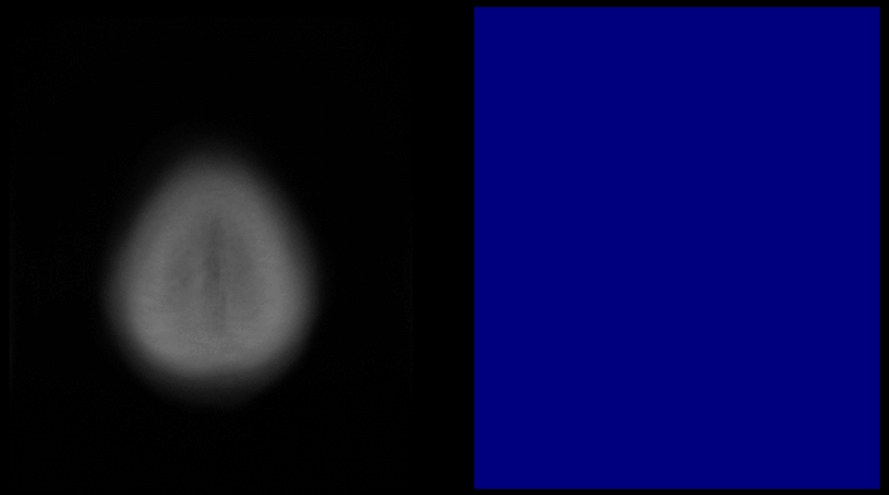
Assessing Versatile Machine Learning Models for Glioma Radiogenomic Studies across Hospitals

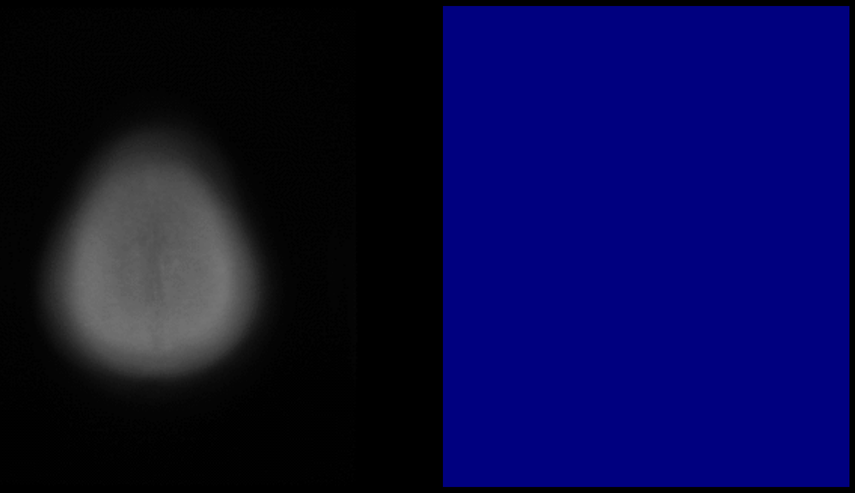
Risa K. Kawaguchi [1,2,3,†], Masamichi Takahashi [4,*,†], Mototaka Miyake [5], Manabu Kinoshita [6], Satoshi Takahashi [2,7], Koichi Ichimura [8], Ryuji Hamamoto [2,7], Yoshitaka Narita [4] and Jun Sese [2,9]

Kawaguchi RK, et al. *Cancers* (2021)

Average profile of tumor regions (Gd T1WI imaging)

Red: ROI = Tumor region



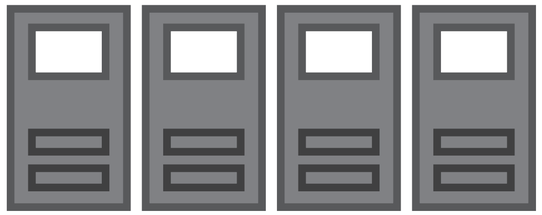


High　　MGMT methylation (>16%)　　Low

Can we predict genetic mutation and malignancy from brain MRI?

国立がん研究センターと産総研の共同研究

社会実装を想定した実験デザイン

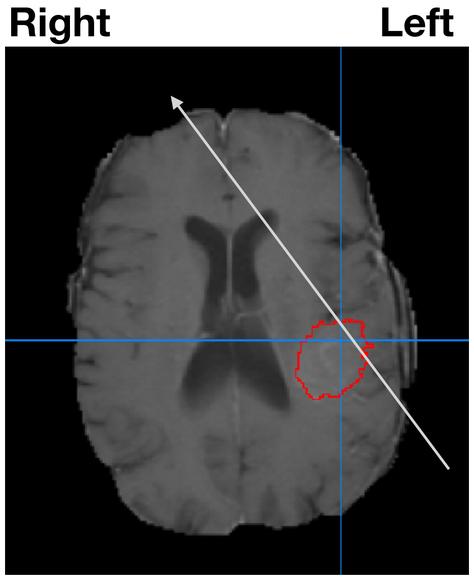


？？？

https://github.com/carushi/PABLO

- 公開データセットで予測器を学習
- 各病院で予測を行う
  - 病院ごとの撮影方法の違い、個人情報保護

解析するための画像変換で一苦労



- DICOMフォーマット
  - 生のアウトプット、患者情報などを含む
  - 1スライス1画像
- NIfTI（Neuroimaging Informatics Technology Initiative）フォーマット
  - qform （スキャナー基準）vs sform
  - 原点の位置
  - x: Right-to-Left, y: Posterior-to-Anterior, z:Inferior-to-Superior
  - RAS +/-
- ビューワーで確認しながら変換を進める
  - vinci, mango, FSLなど

Erratum for the Report “Neural mechanisms for lexical processing in dogs” by A. Andics, A. Gábor, M. Gácsi, T. Faragó, D. Szabó, Á. Miklósi

SCIENCE · 7 Apr 2017 · Vol 356, Issue 6333 · DOI: 10.1126/science.aan3276

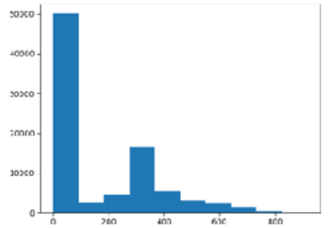
1,745 5

In the Report “Neural mechanisms for lexical processing in dogs,” the directions left and right were inadvertently switched in reporting the results from dogs’ brains. This was caused by an error in interpreting the coordinates of MRI images, specifically in the process of accounting for the different body positions of humans and dogs in the MRI scanner. This error does not affect the main conclusions of the paper. The HTML and PDF versions have been corrected.

医師が無意識に捉えている腫瘍の特徴量とは何か？

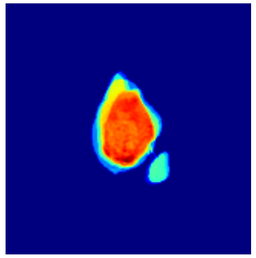
- 画像情報 (GD, T1, T2, FLAIR)

輝度の基本特徴量
15



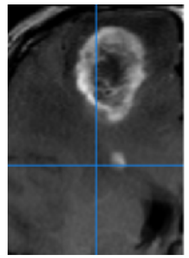
- min, max, mean, …

Pyradiomics
van Griethysen, J.K.M., et al. 2017
960

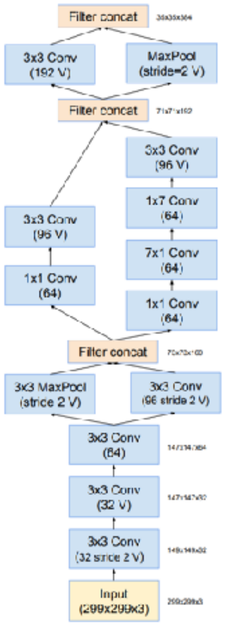


- GLCM-based texture
- First order features

Resnet-inception v2
Szegedy C, et al . 2016
3072



- trained for Imagenet



- 患者情報

解剖学的な腫瘍の場所情報
30

- Fraction of each anatomical tissue type in tumor regions

カルテ情報
3

- Age
- Sex
- KPS (Karnofsky Performance Score)
  - used to evaluate patient physical performances
  - <= 100%

=16221

Sklearnでいろいろな分類器を試す

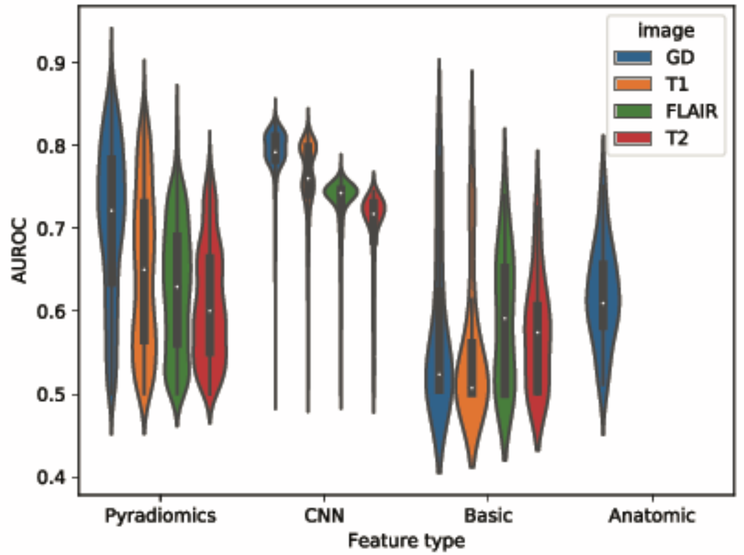
- 分類手法を呼び出す
  - clf = LinearDiscriminantAnalysis()
- 様々な手法がラップされているのでやることは基本的に同じ
  - fitted = clf.fit(x[train,:], y[train])
  - y_score = fitted.predict_proba(x[test,:])
  - y_pred = fitted.predict(x[test,:])
- 精度評価のための指標
  - ['roc_auc', 'precision', 'recall', ‘accuracy']
  - roc_curve(y[test], y_score)
- 交差検証
  - cv_results = cross_validate(x, y, cv=sample, scoring=metrics)
  - 交差検証のROCを書くときはちょっと面倒

特徴量の解釈の難しさ

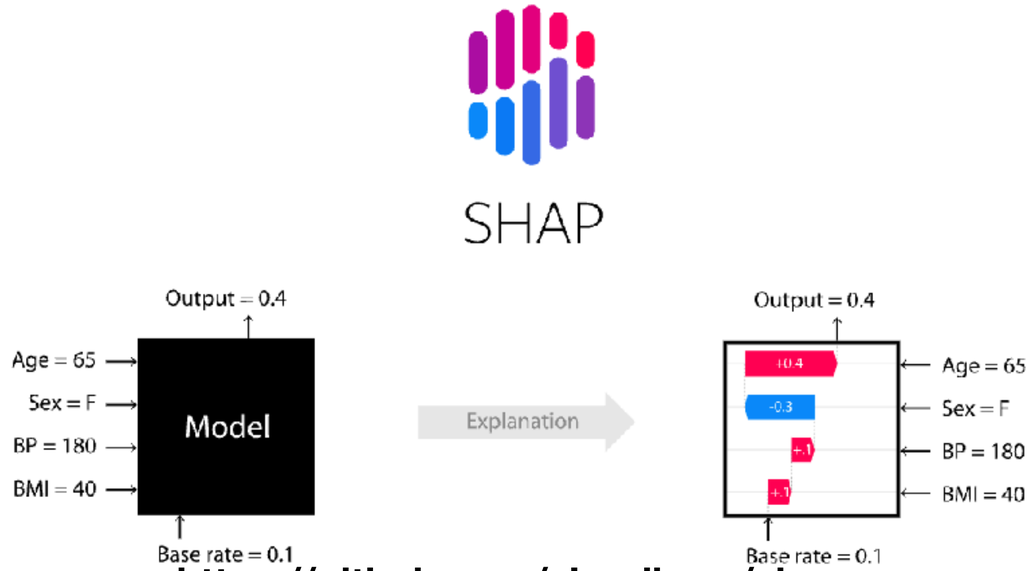
- 線形モデル ＜ 単純な非線形モデル ＜ 深層学習

参考：https://christophm.github.io/interpretable-ml-book/index.html

異なる精度を持つ特徴量群



シャープレイ値のようなスコア



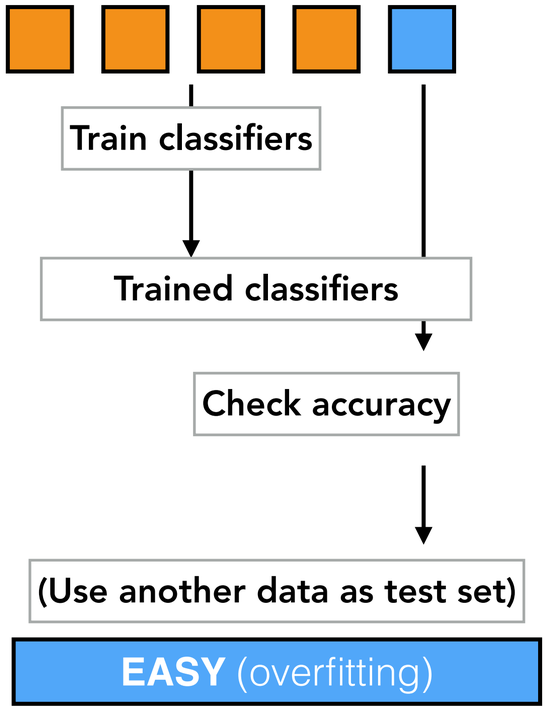
https://github.com/slundberg/shap

アテンション（属性マップ）

Bruce NDB and Tsotsos JK.
Journal of Vision, 2009.

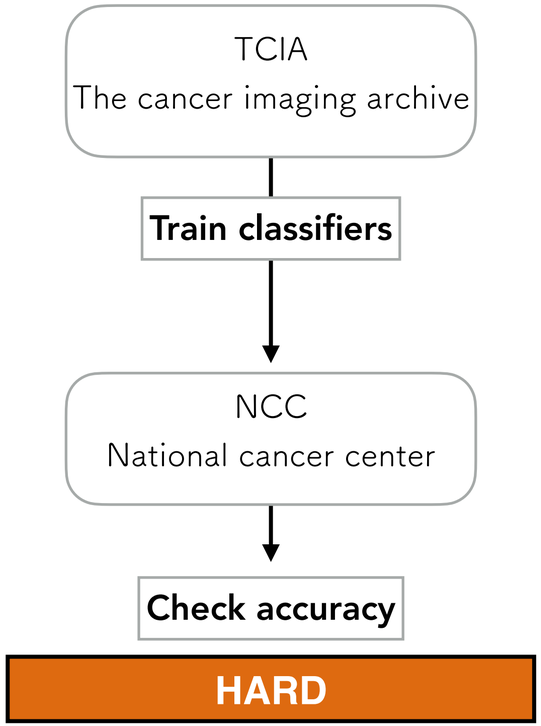
交差検証とデータセット間比較の精度比

1. 交差検証

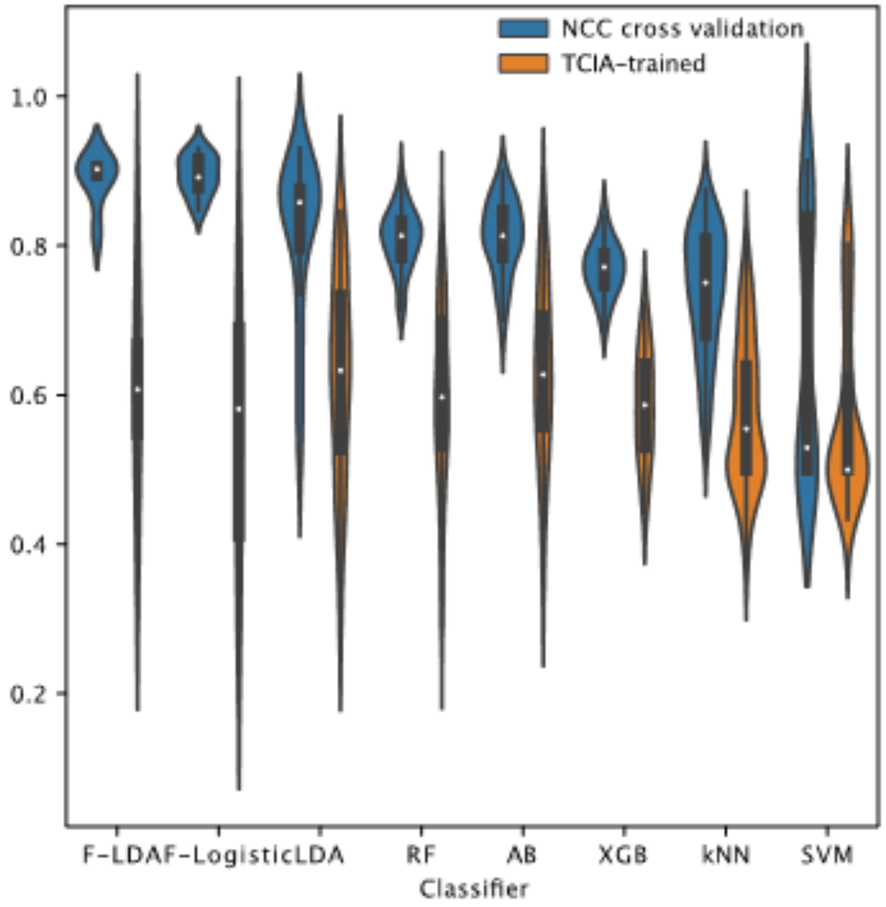


2. 広義の転移学習

= Practical application

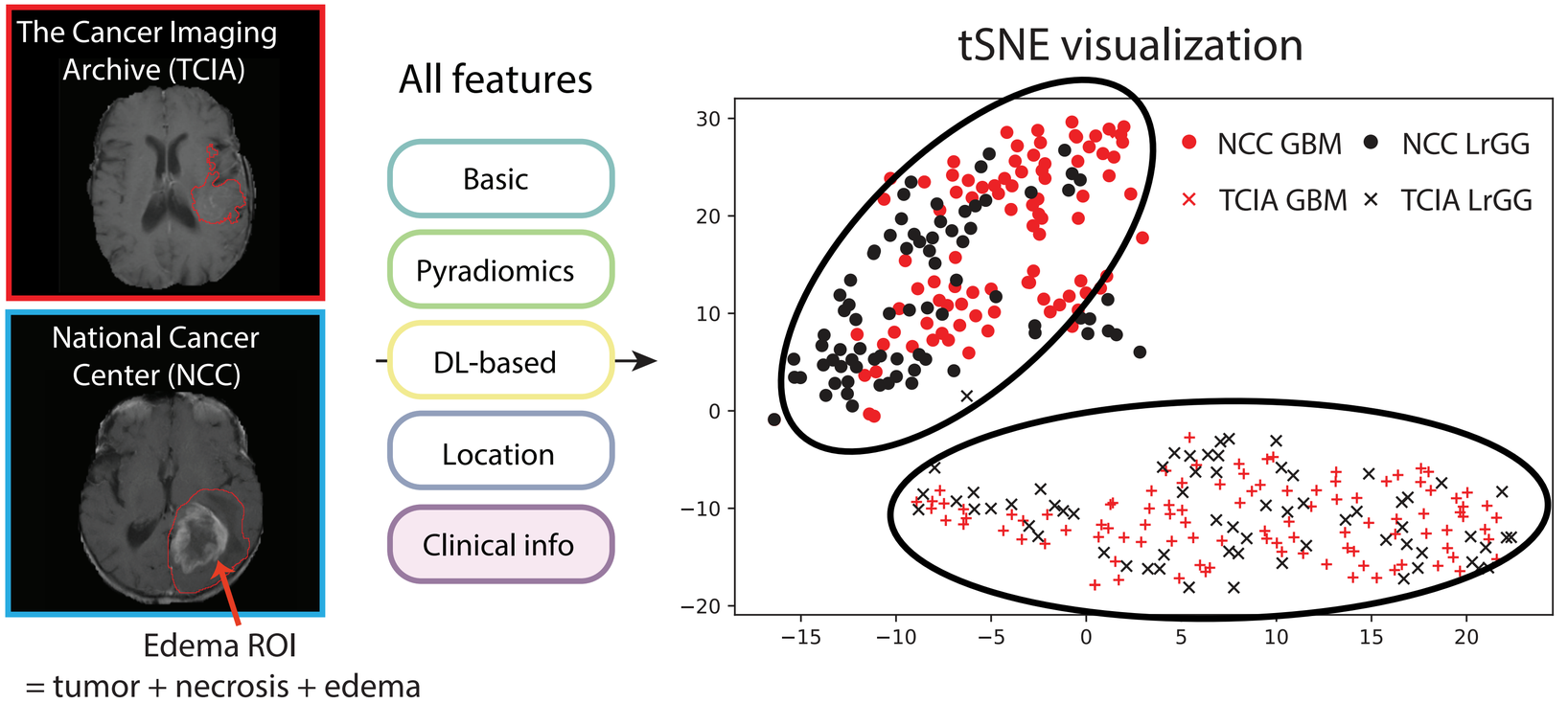


GBM予測のAUROC



- より訓練データに適応した予測器は精度が低下している

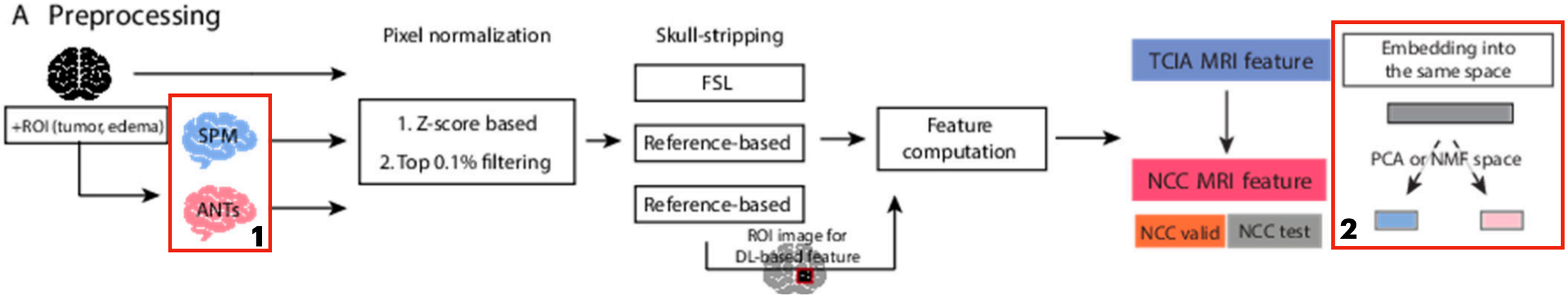
異なるデータセット間の系統的変動



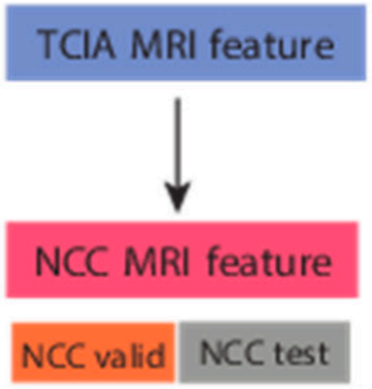
- <u>２つのデータセット間での精度を比較</u>
  - National Cancer Center (NCC)
    - 90 GBM and 76 LrGG
  - The Cancer Imaging Archive (TCIA)
    - 102 GBM and 65 LrGG

- <u>特徴量の系統的変動</u>
  - TCIAはすでに標準化・Skull-strippingされた脳空間での画像
  - 解像度なども違う
  - →ファインチューニングが必要

前処理・次元圧縮



1. 脳の標準化 **2. 輝度値の正規化** **3. 骨を除く** **4. 特徴量を計算** **5. 次元圧縮（PCA/NMF）**



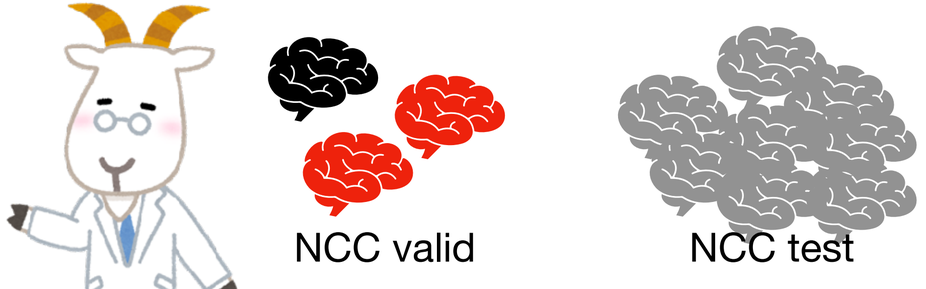
データセットを3分割

1. TCIA – 訓練データ
2. NCC valid - モデル選択用
3. NCC test - テストデータ

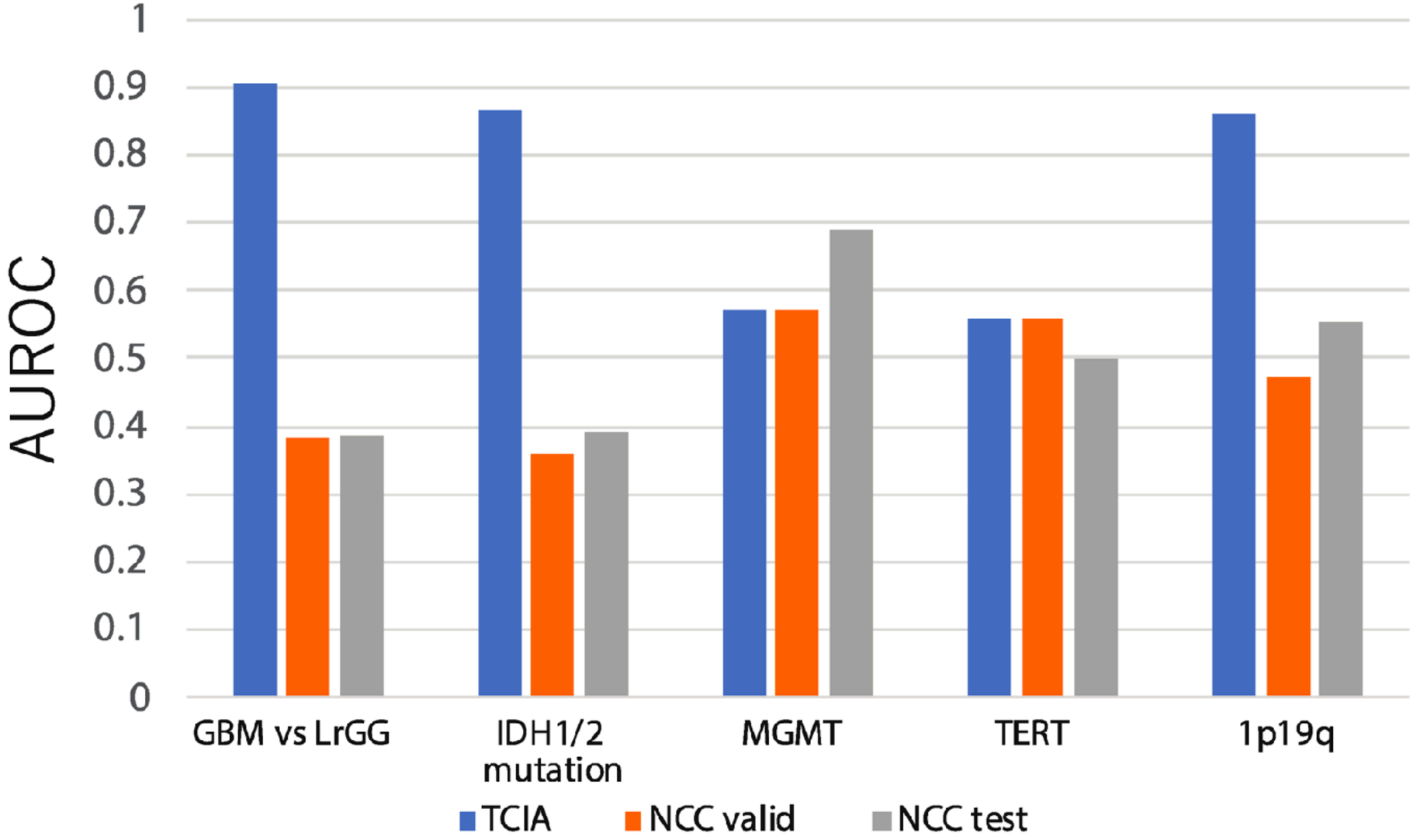
患者情報を統合する前に圧縮

TCIA/NCCそれぞれで同じ圧縮

- NCC valid/testはデータセット全体の中での方向性の情報は保持



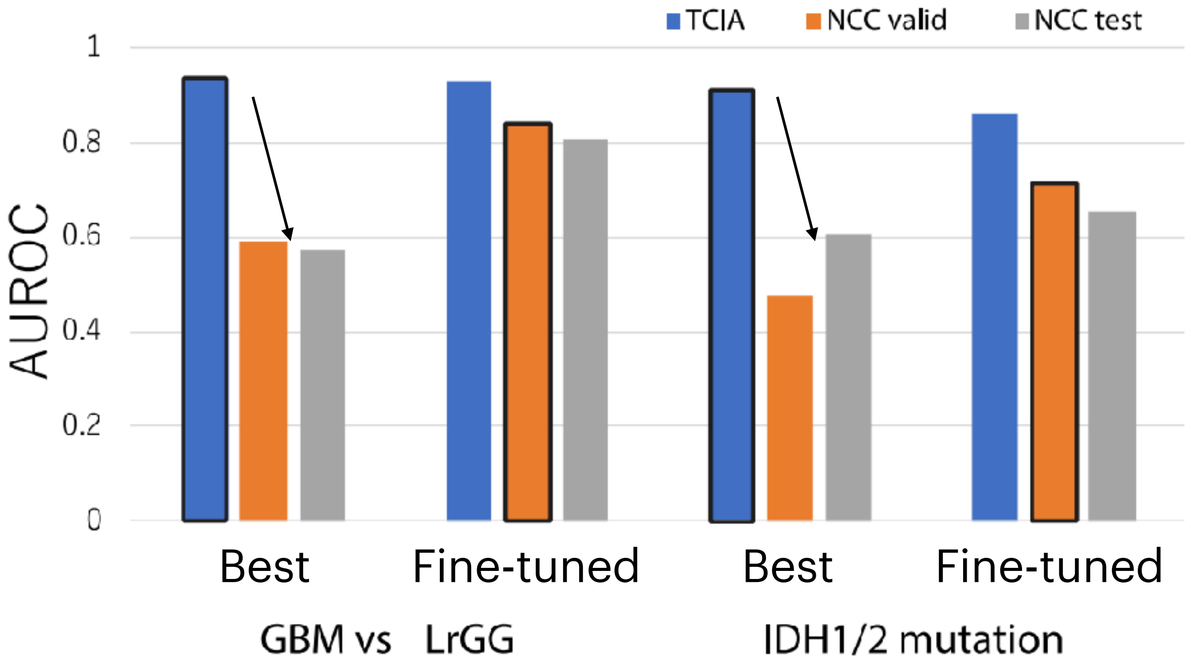
MRIによるバイオマーカー情報の予測精度



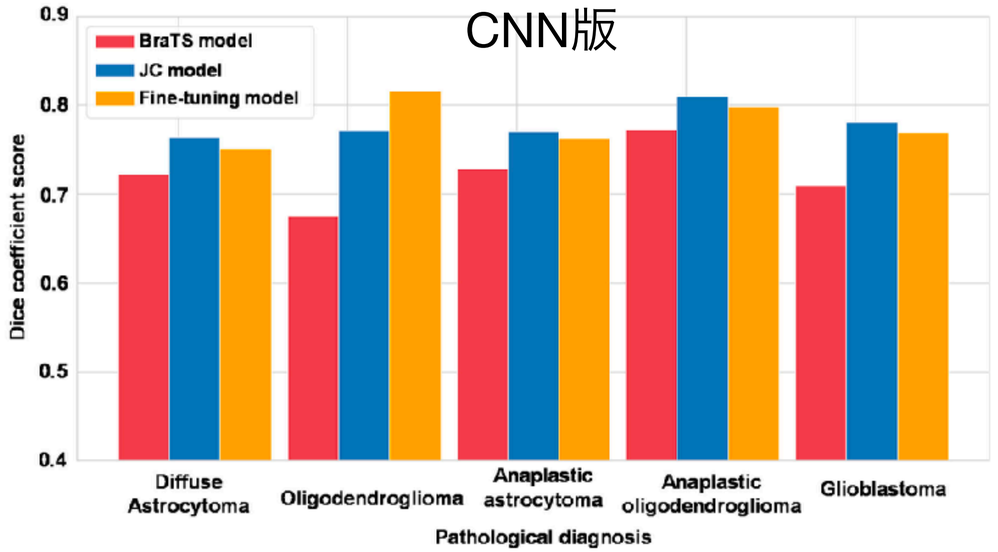
- **TCIAの交差検証**で最も精度の高かった手法を選択
- グレード情報・IDH変異・chr1p19qの共欠損でランダムでない予測精度
- NCCのデータセットでは精度の大幅な低下

ローカルな病院で一部データを利用したファインチューニング

- 分類器自体はTCIAだけで学習
  - Best: TCIAの交差検証（青）でモデル選択
  - Fine-tuned: NCC valid（オレンジ）でモデル選択
- ファインチューニングで精度の大幅な低下を防ぐことができる可能性



CNN版

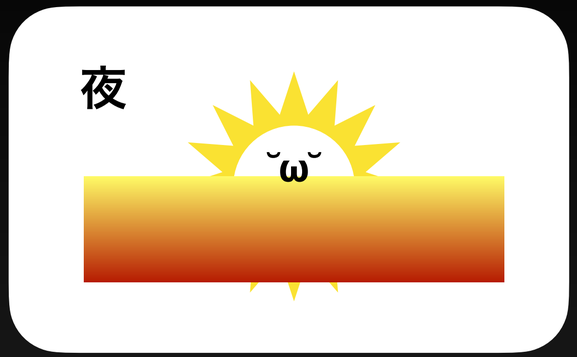


Takahashi S, et al. *Cancers* (2021)

大規模な画像データの学習を行う際に地味に注意すべきこと

- 最終的に計算された特徴量が元の画像から妥当な値であるかを確かめる
  - 画像と数値の検証はぱっと見では難しく何らかの基準が必要
  - 画像変換が途中で落ちたりデータ転送のエラーで中身が空になることも
- データセットの並びに特別な意味がある場合
  - 予測検証の際にシャッフルする
- それぞれの予測は独立したジョブとして走らせる
  - ランダムシードの共有などにより書き換えた際に再現性がとれなくなることも
- AUCなのかAccuracyなのかPrecisionなのかRecallなのか（正負の偏り）
  - 異なるデータセット間ではPrecisionは低かったがAUCは十分高い予測ができた

3. メタ統合解析のススメ 機械学習・人工知能技術 実践編



- 共発現ネットワーク解析
- 一細胞エピゲノムのメタ統合解析と深層学習による
  モチーフ予測
- Github: https://github.com/carushi/cb_lab/tree/main/code_collection/qb_221207
- Access to Jupyter notebooks via Google colab
  https://colab.research.google.com/github/carushi/cb_lab/

1. 遺伝子共発現ネットワークによるDEGクラスタリング

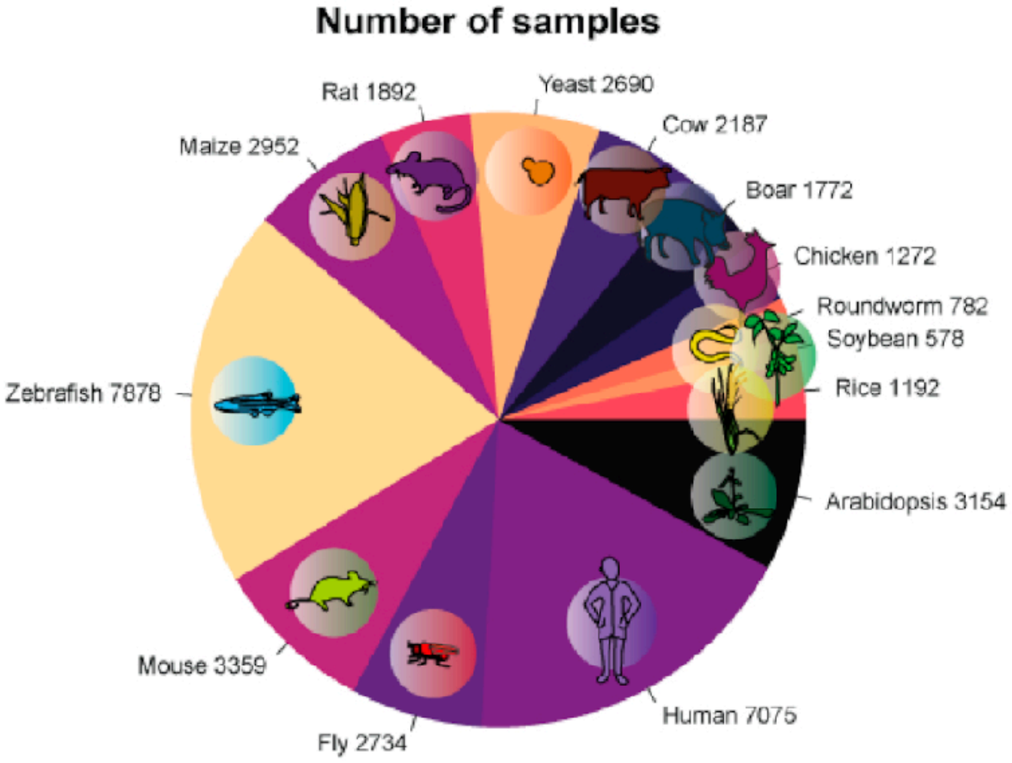
W566–W571 Nucleic Acids Research, 2020, Vol. 48, Web Server issue *Published online 11 May 2020*
doi: 10.1093/nar/gkaa348

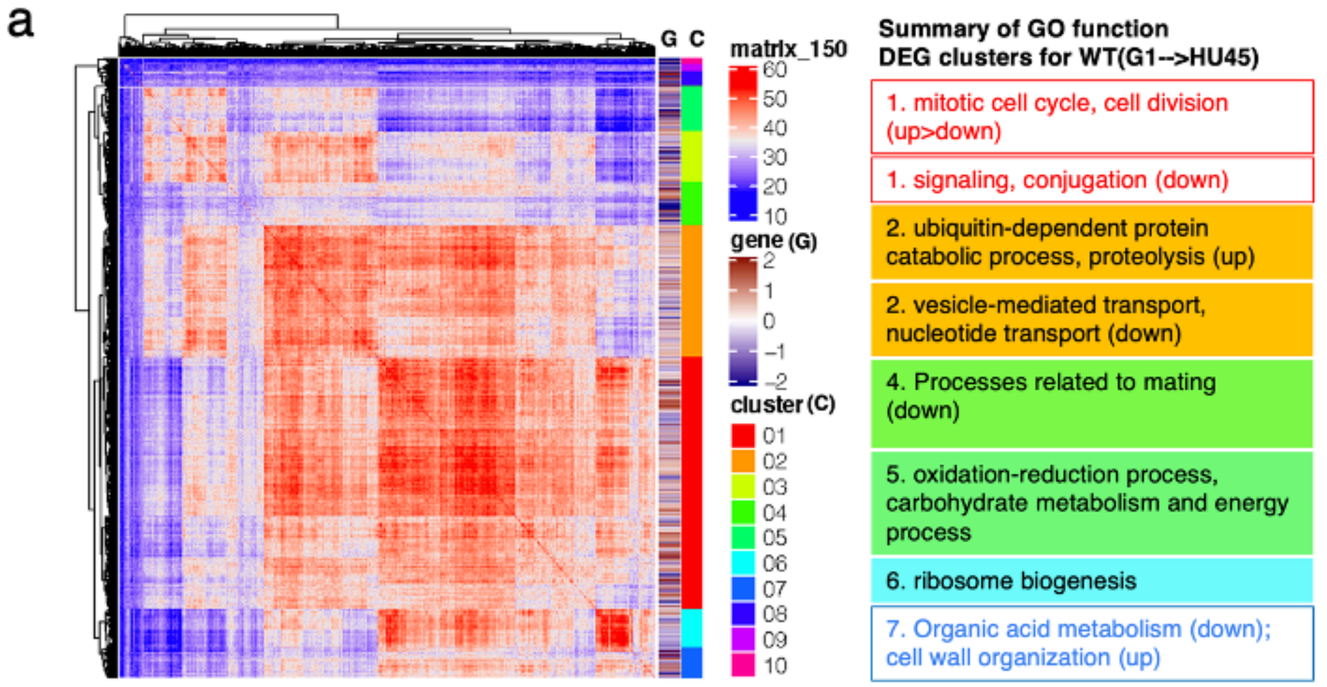
CoCoCoNet: conserved and comparative co-expression across a diverse set of species

John Lee†, Manthan Shah†, Sara Ballouz, Megan Crow and Jesse Gillis*

Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, 500 Sunnyside Blvd., Woodbury, NY 11797, USA

Received March 12, 2020; Revised April 21, 2020; Editorial Decision April 24, 2020; Accepted April 24, 2020



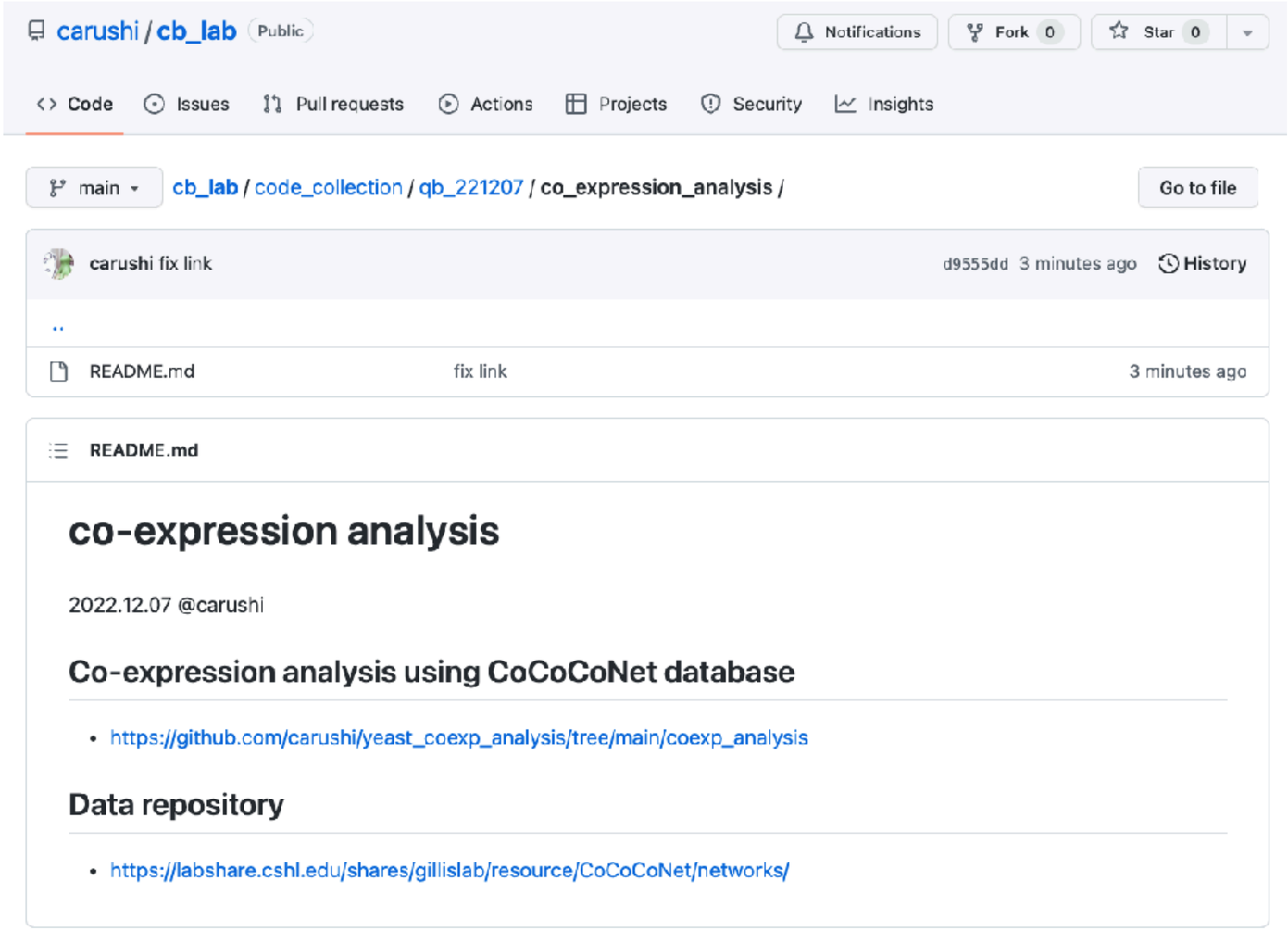


Shue YJ, et al. Accepted by *eLife.*

- 共通して変動する遺伝子群
  - 類似の機能・共通する制御因子
  - 機能解析の手がかりになる
  - 欠損値の推測にも

サンプルコード

- https://github.com/carushi/cb_lab/tree/main/code_collection/qb_221207/co_expression_analysis
- Rスクリプト
- DE-seq2により得られたDEGの共発現クラスター解析
- Dynamic tree cutによりクラスターに分類しヒートマップ化



2. 深層学習による細胞種特異的エピゲノム状態の予測

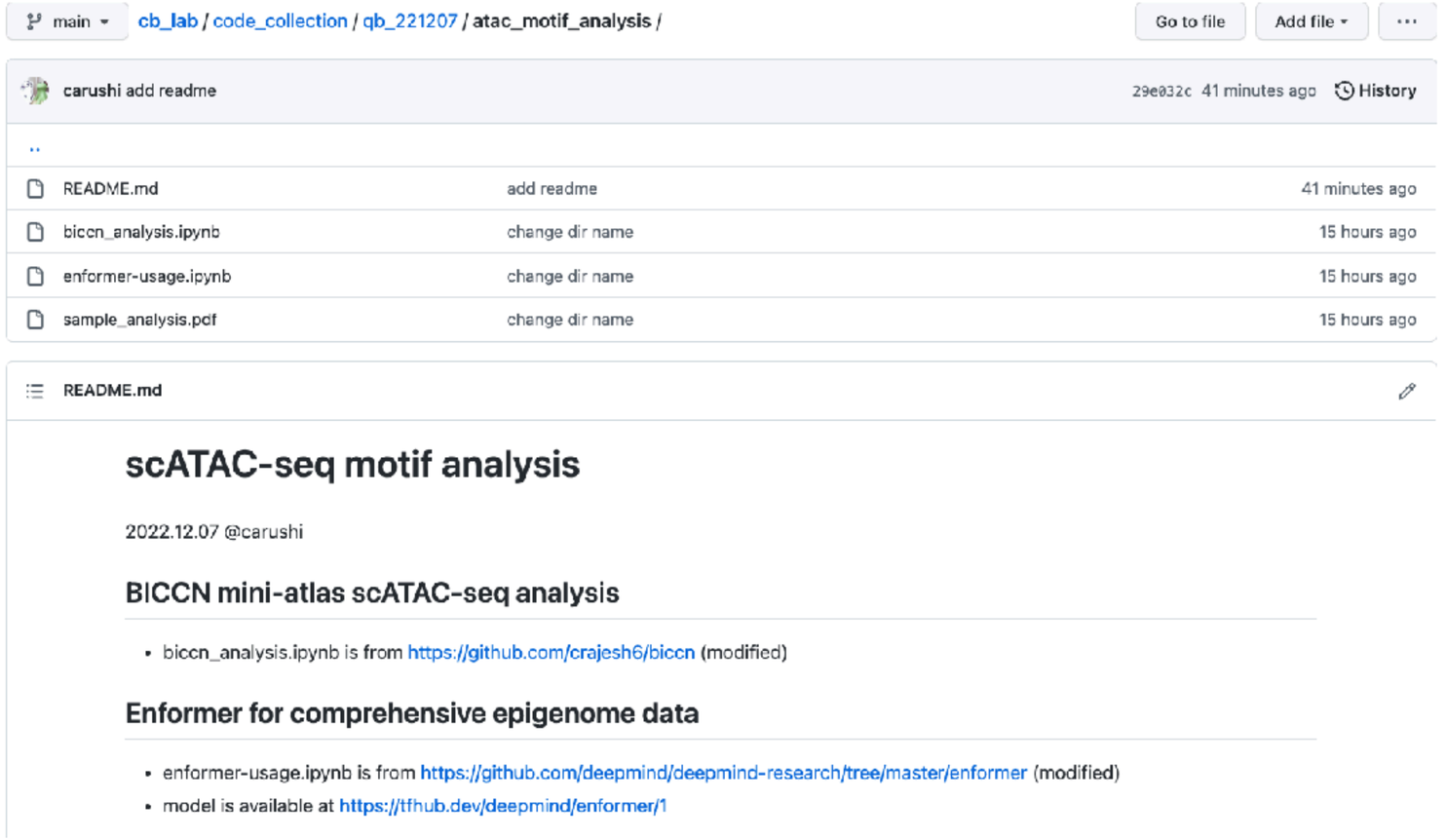
- github.com/carushi/cb_lab/qb_221207/README.md

  biccn-analysis.ipynb　　enformer-usage.ipynb

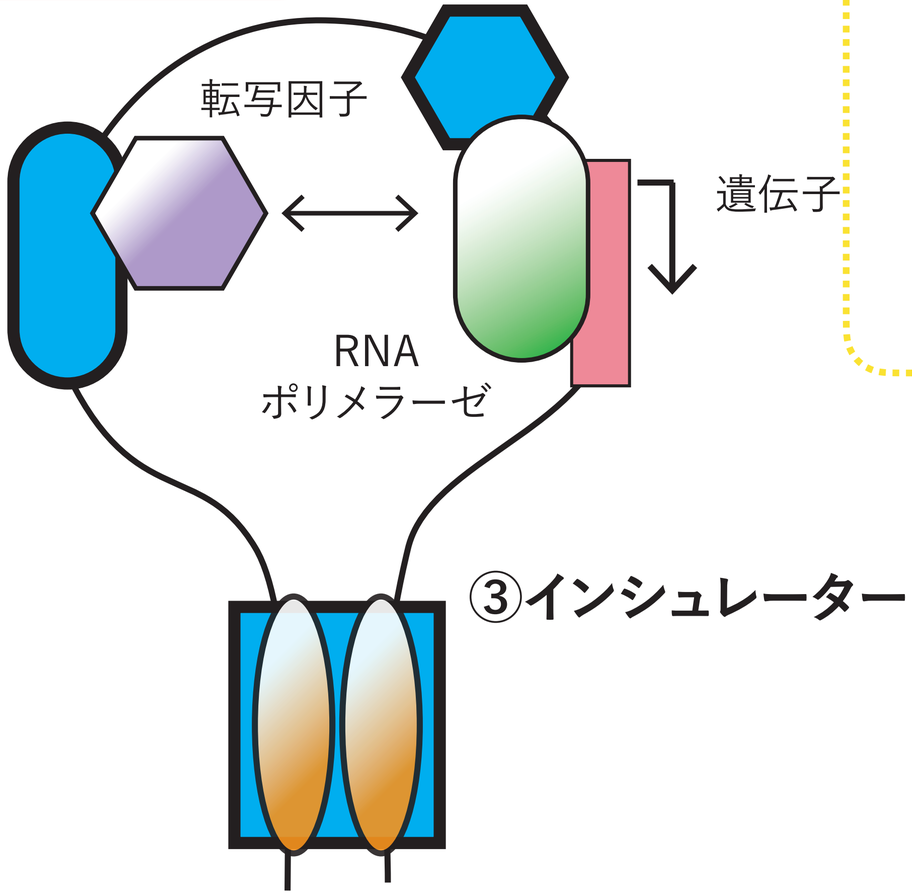
  - BICCN mini atlasのscATAC-seqデータを使った細胞種特異的モチーフ予測モデルの訓練から解析まで
  - Enformer（Basenji2）による条件依存的エピゲノム状態予測

サンプルコード

- Jupytor notebook＋Python
- https://colab.research.google.com/github/carushi/cb_lab/
- Dropboxからデータをダウンロード→モデルファイルをダウンロード→パッケージをインストールで手元でも動く

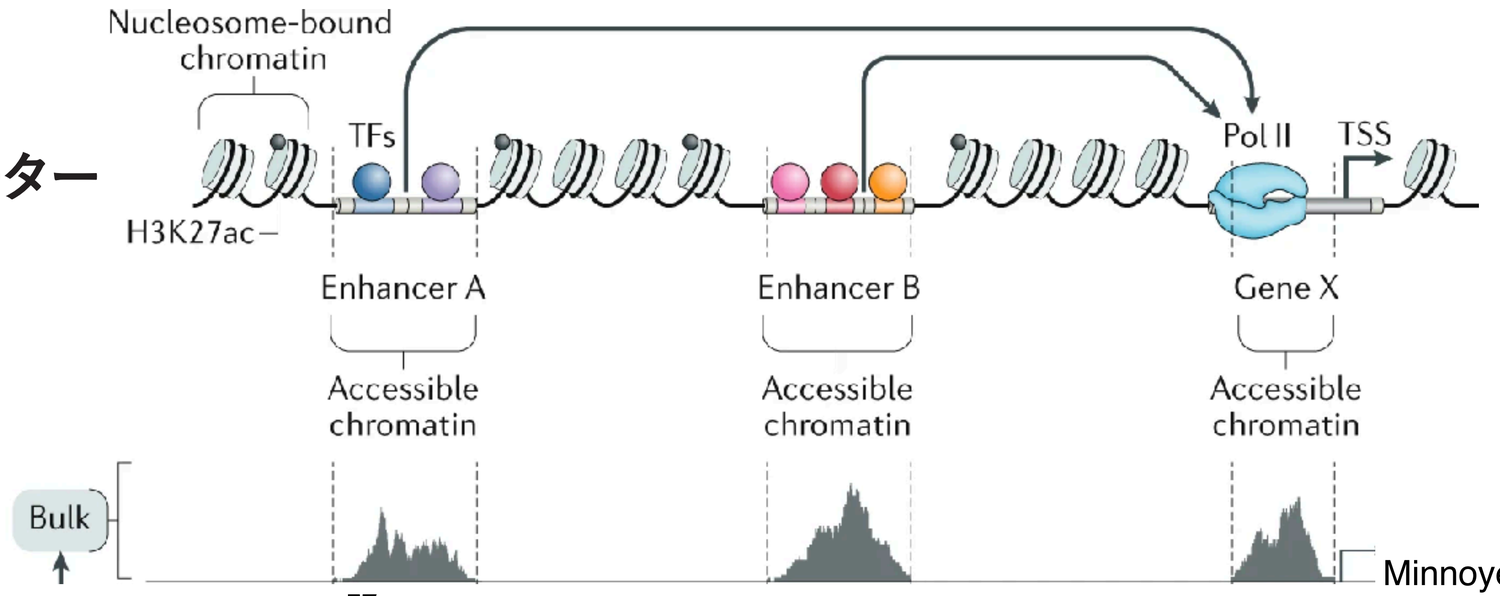


背景：エピゲノム解析による細胞種特異的エンハンサーの特定



エンハンサー

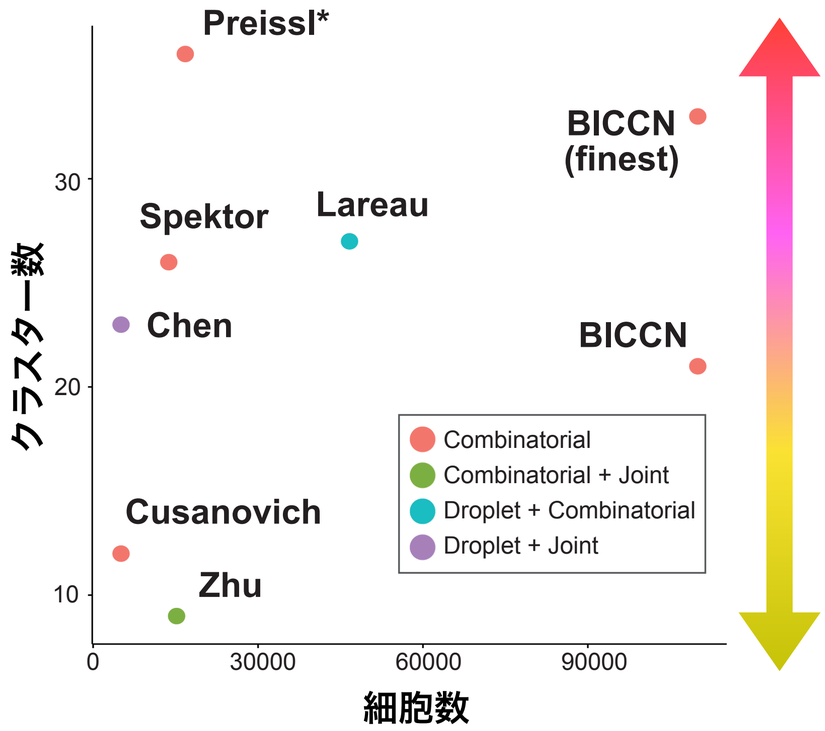
- シス制御因子の一つ
- **転写因子**の結合の足場として下流の遺伝子を制御
  - 組織・細胞種特異的な発現制御
- **一細胞ATAC-seq**などの手法により推定可能

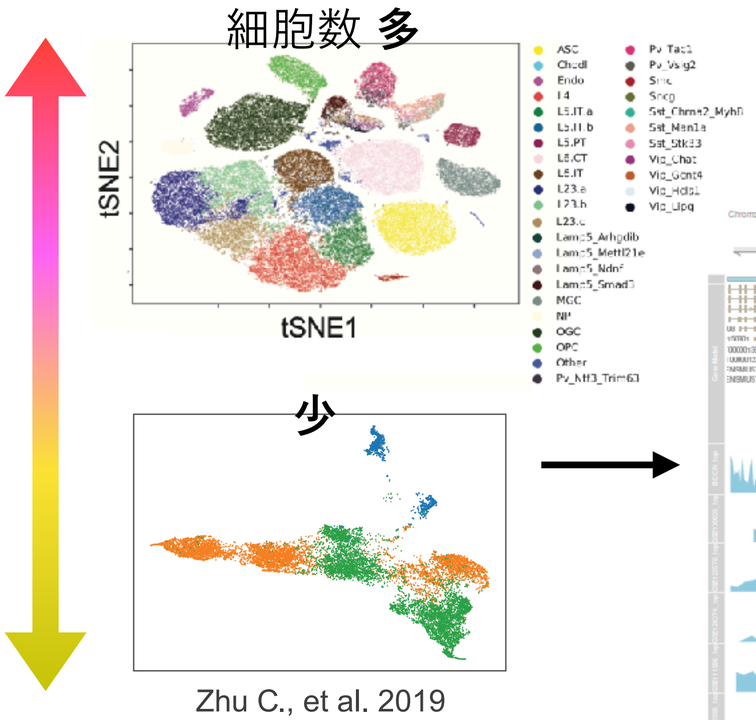


Minnoye L., et al. *Nat Rev*, 2021.

一細胞エピゲノム解析のメタ統合解析による不均一性の克服

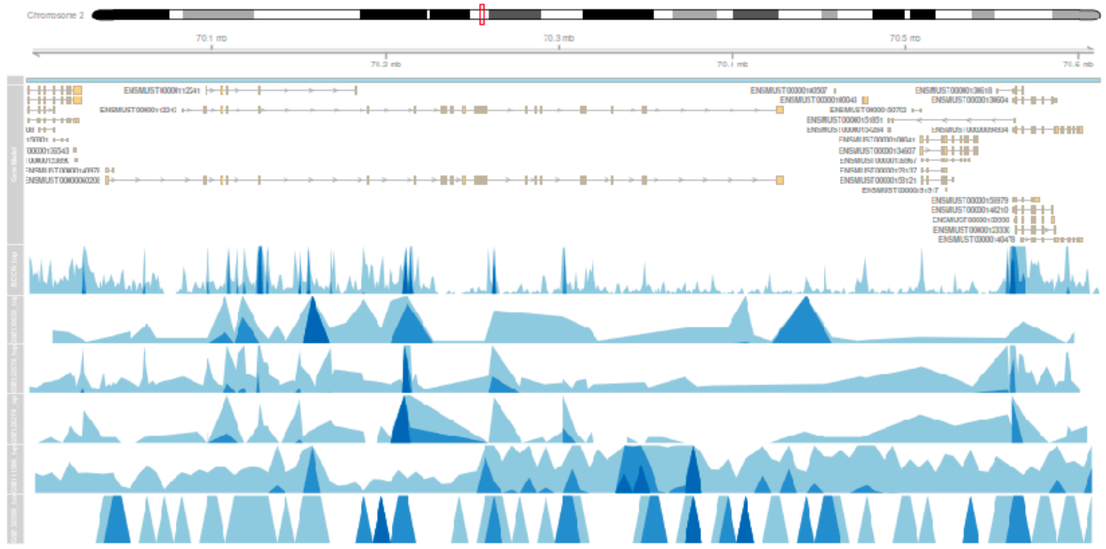
マウス脳一細胞ATAC-seqデータ収集





Zhu C., et al. 2019

Meta scATACサーバー上で公開



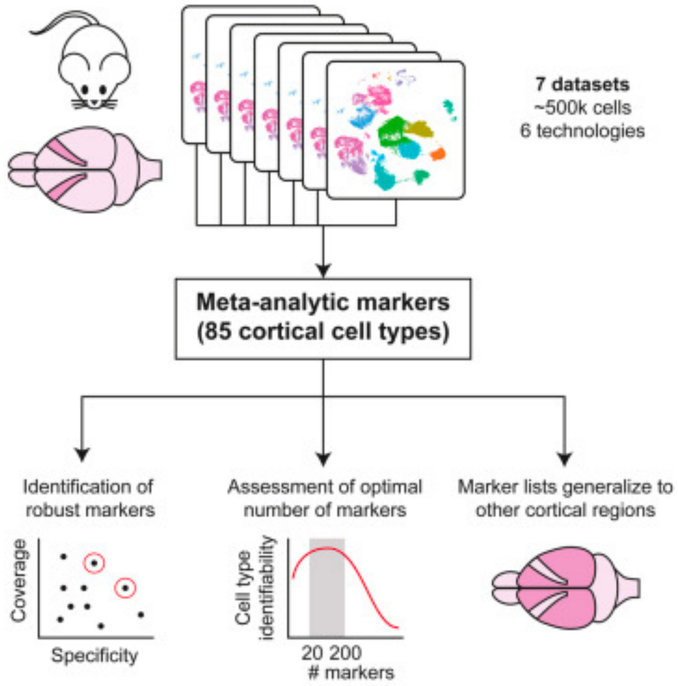
Shinyサーバー

https://gillisweb.cshl.edu/Meta_scATAC/

Kawaguchi RK, et al. Accepted by *Briefings in Bioinformatics*

メタ統合解析によるロバストな細胞種特異的バイオマーカーの構築

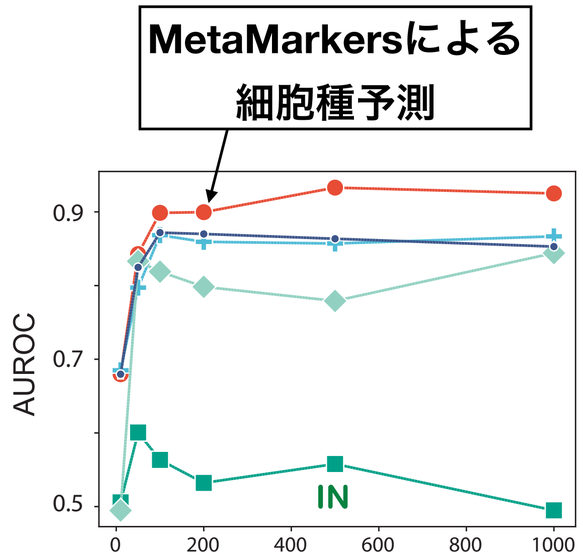
MetaMarkers: 一細胞RNA-seqの欠損を克服するマーカー遺伝子構築

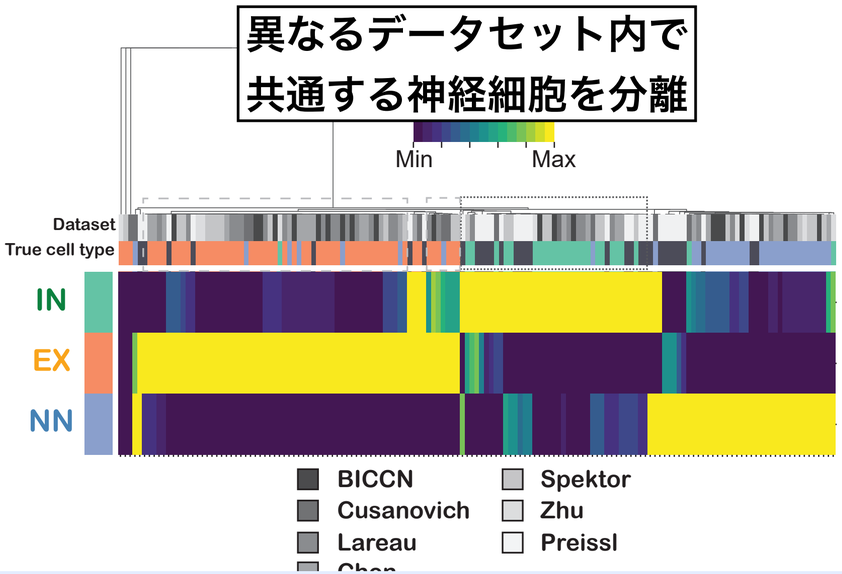


Fischer and Gillis. *iScience*, 2021

~**200**遺伝子の細胞種マーカーセット

Catactor: 非均一な一細胞ATAC-seqでもメタ解析的マーカー遺伝子の有効性を示す





メタ解析により再現性のあるエピゲノム解析が可能

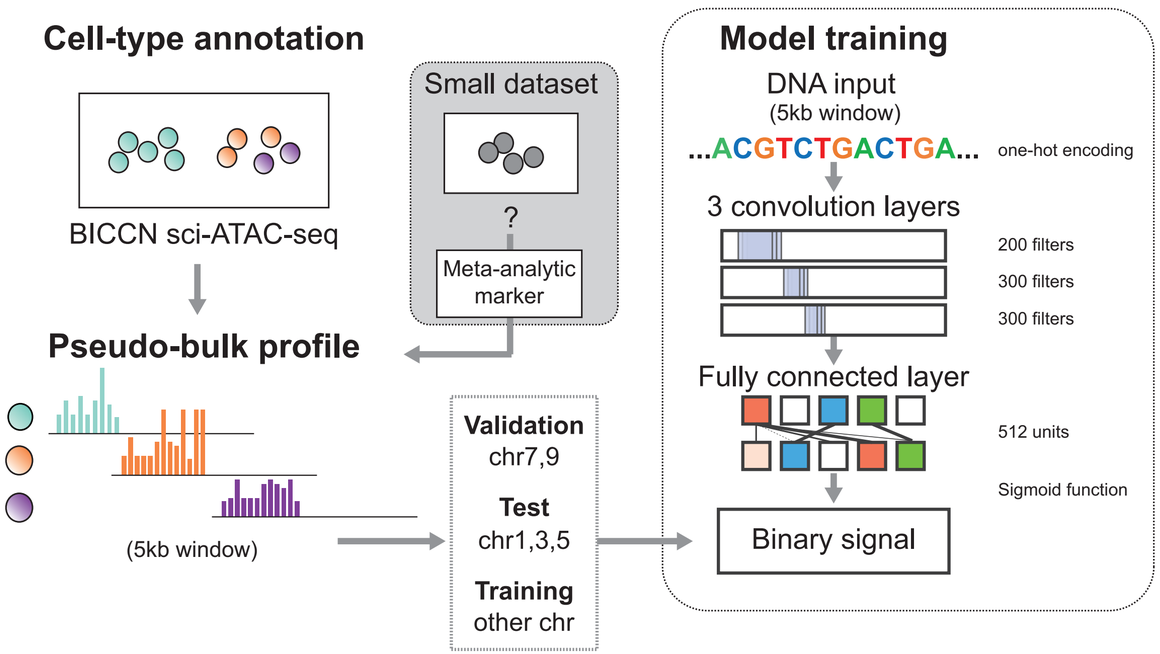
メタ統合解析×深層学習による細胞種特異的エンハンサー解析

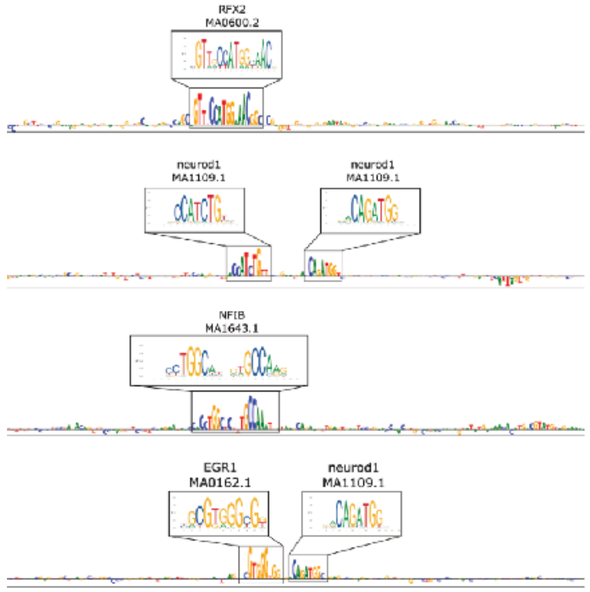
メタ解析による
細胞種アノテーション

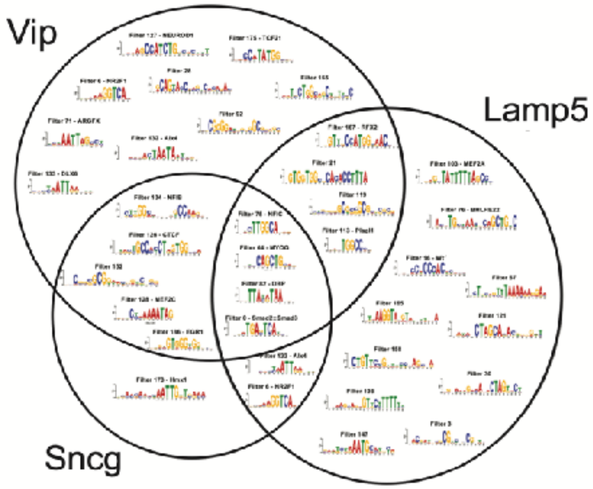
CNNによる
エピゲノム予測

属性マップによる
モチーフ領域検出

モチーフエンリッチ
メント解析

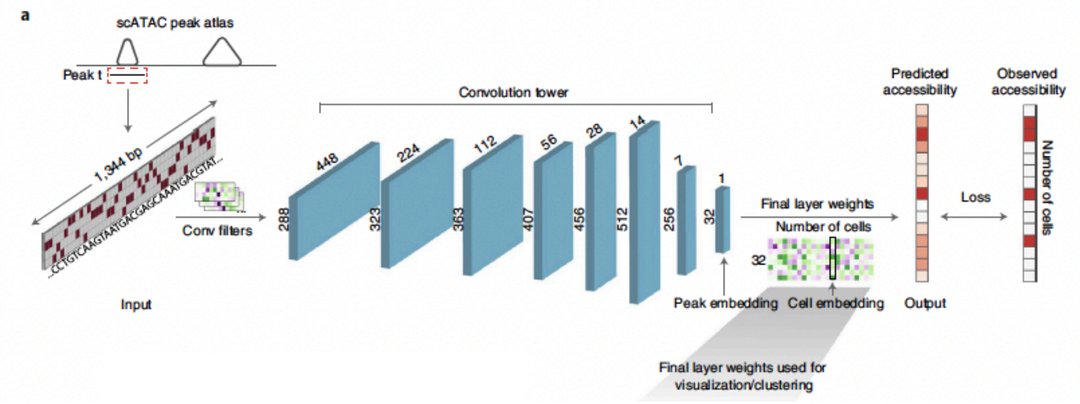




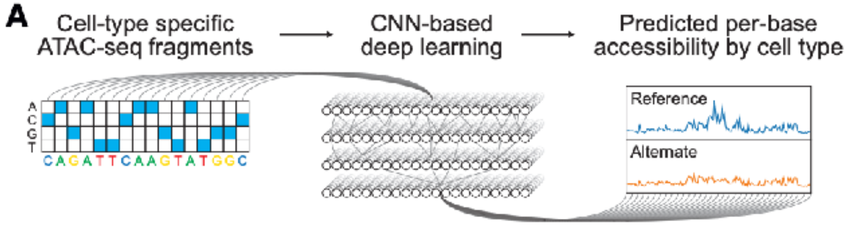


- メタ解析的マーカー遺伝子によって共通する細胞種を予測
- 細胞集団を集めて深層学習（CNN）を適用
- 属性マップ（attribution map）から細胞種特異的なモチーフ推定

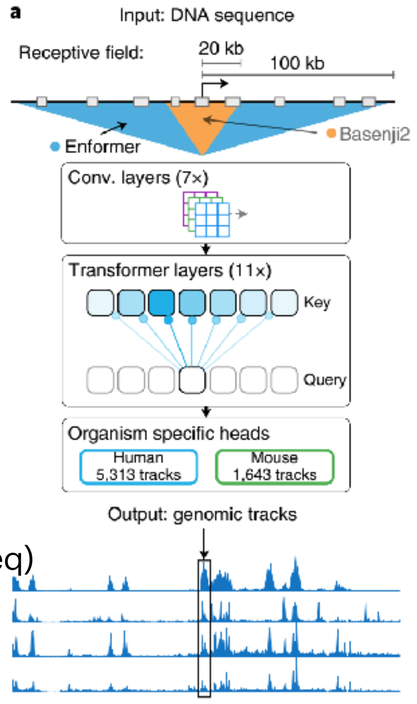
補足資料:先行研究で一細胞解析に利用されている深層学習モデル



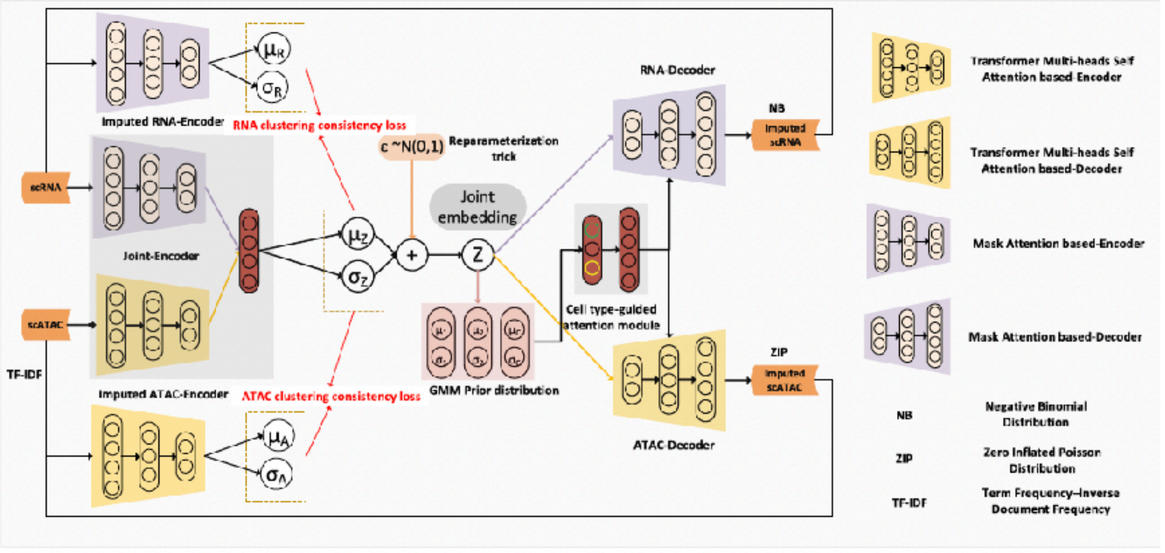
Yuan H and Kelly D. *Nature methods*, 2022 (scBasset: scATAC-seq data)



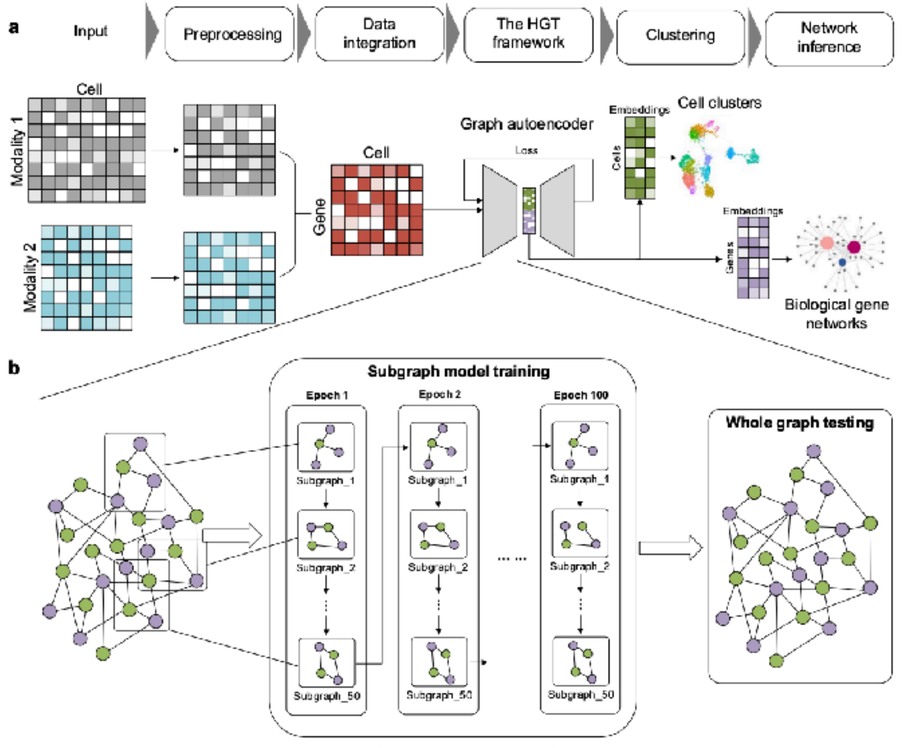
Wang SK, et al. Cell Genomics, 2022. (Bpnet trained for scATAC-seq)



Avsec Ž, et al. Nature methods, 2021. (Enformer)



Li G, et al. *Genome Biology*, 2022. (scMVP: scRNA-seq + scATAC-seq)



Ma A, et al. bioRxiv, 2021. (DeepMAPS)