トランスオミクス研究のための横断的生命科学データ解析基盤の構築

生体防御医学研究所
トランスクリプトミクス分野

大川恭行・前原一満

先駆的科学計算に関するフォーラム2019
(Forum on Advanced Scientific Computing 2019)
～先端的計算科学研究プロジェクト成果報告 ～

背景

次世代シーケエンサー：NGS

計算拠点にデータを転送するだけでもひと仕事

Kyushu Univ. total output (HiSeq1500/2000)

Output per Run	450 Gb (~300Gbyte)
Single Reads Passing Filter	4.5 billion reads

Hiseq 3000/4000

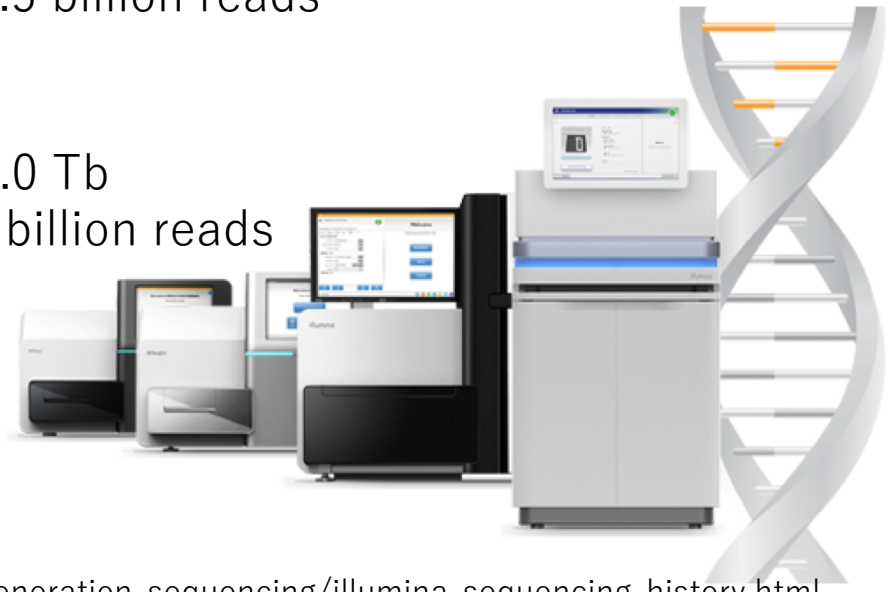
b:単位はbyteではなく、base（塩基）

Output per Run	400 Gb
Single Reads Passing Filter	2.1-2.5 billion reads

NovaSeq

Output per Run	2.0-6.0 Tb
Single Reads Passing Filter	2-20 billion reads

短いDNA配列の大規模並列読み取りマシン

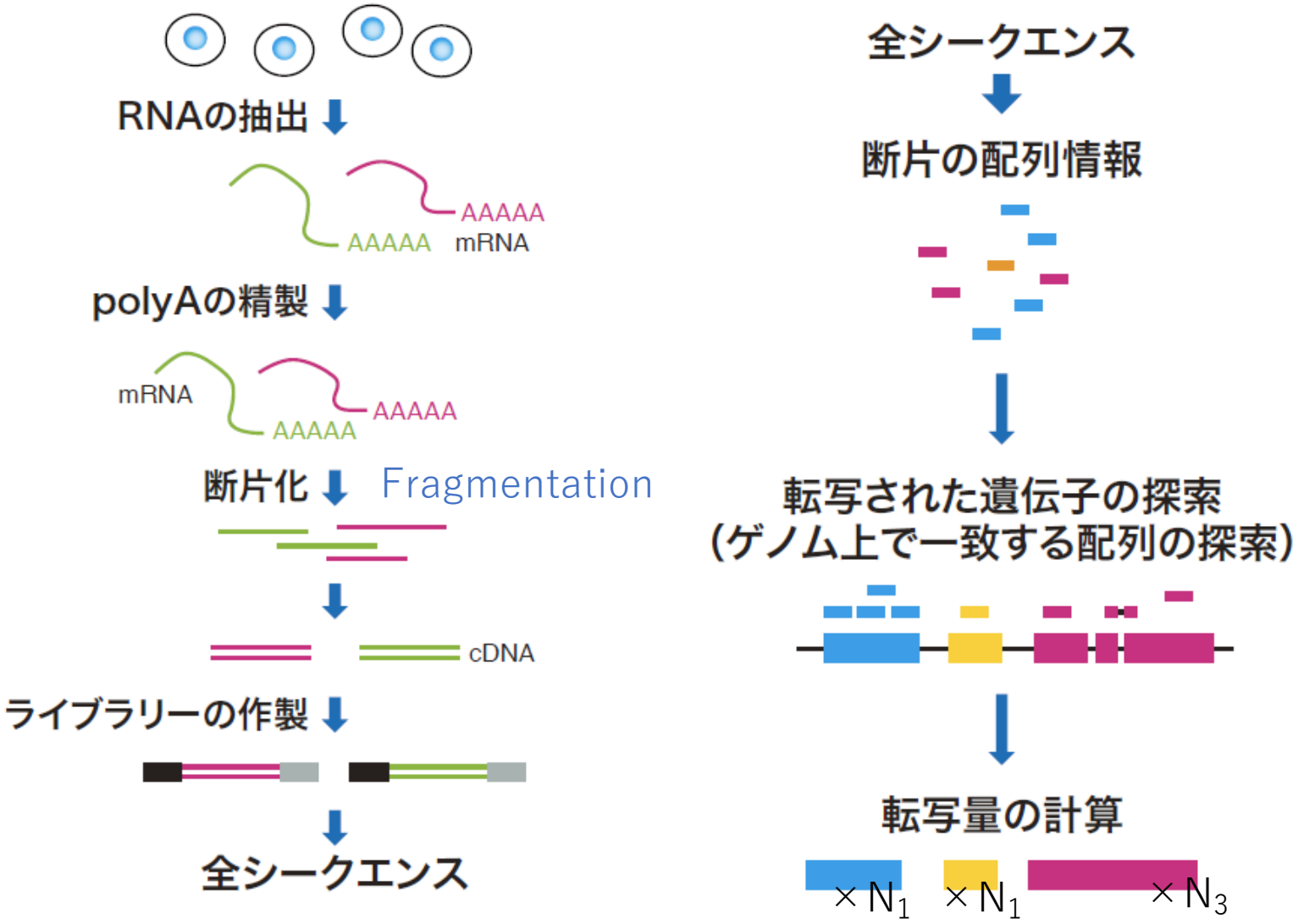


（遺伝子工学＋画像データ解析技術）

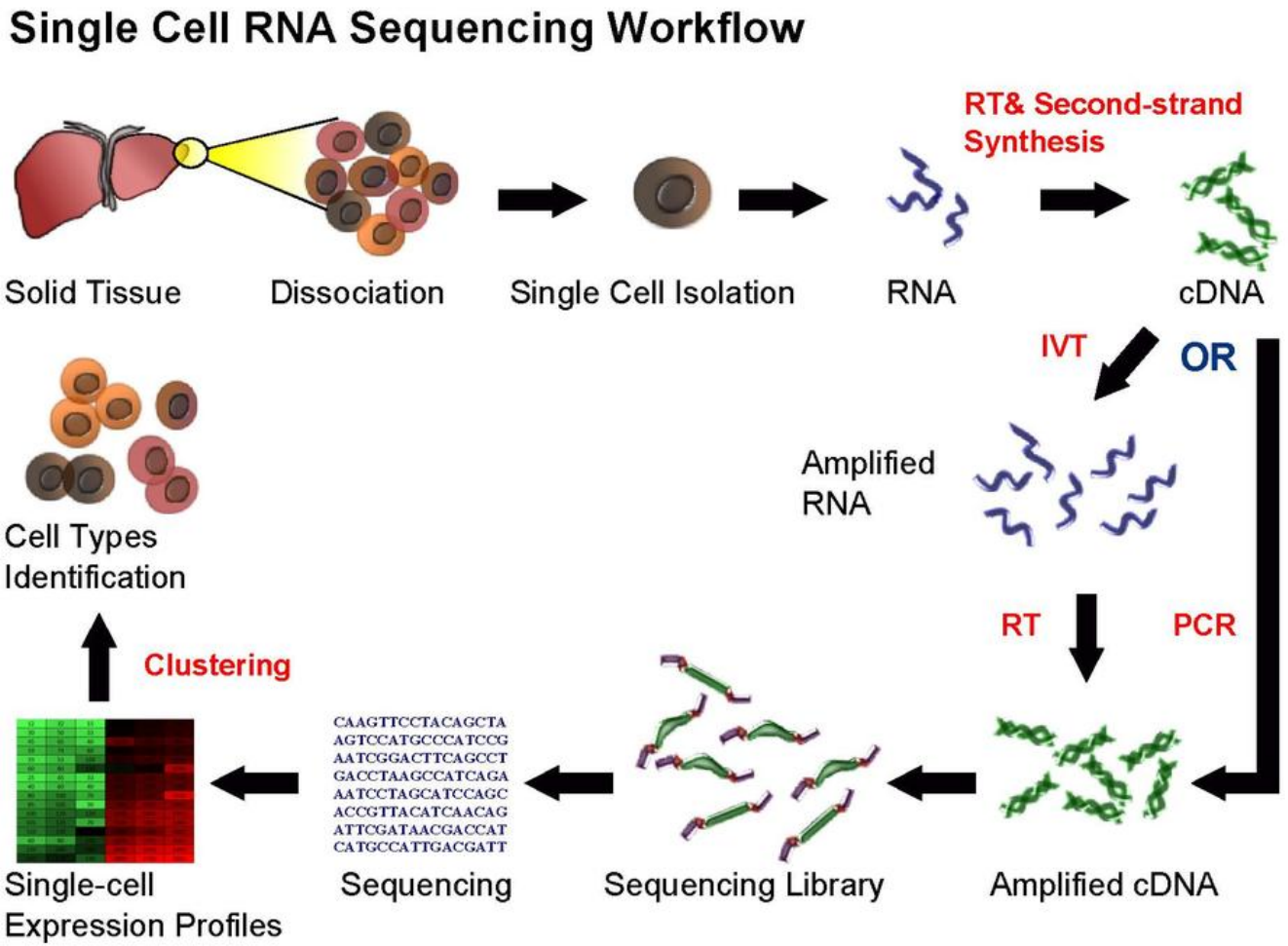
https://jp.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html

背景

例；RNA-seq （遺伝子発現の定量）



単一細胞解析技術の隆盛



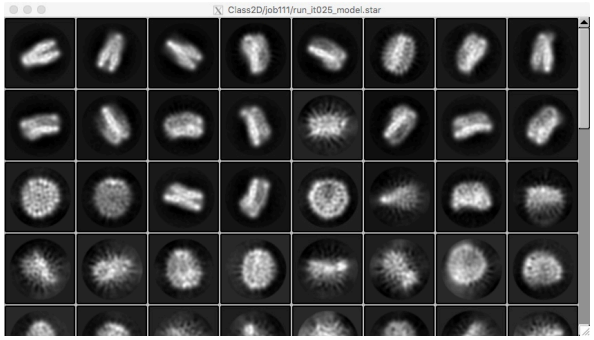
https://en.wikipedia.org/wiki/Single_cell_sequencing

背景

クライオ電子顕微鏡による単粒子解析

試料から得た数千〜数十万粒子分の低感度電顕画像

クラス分類・平均



回転角等の推定を通してタンパク質の三次元像を再構成する

当プロジェクト参加グループ

トランスクリプトミクスG代表

大川恭行・教授・九州大学/生体防御医学研究所

構造生物学G代表

神田大輔・教授・九州大学/生体防御医学研究所

バイオインフォマティクスG代表

須山幹太・教授・九州大学/生体防御医学研究所

統合オミクスG代表

久保田浩行・教授・九州大学/生体防御医学研究所

フェノミクスG代表

林克彦・教授・九州大学/医学研究院

エピゲノミクスG代表

中島 欽一・教授・九州大学/医学研究院

大川G

論文成果（謝辞記載）

次世代シークエンスデータ解析に活用

Chromatin integration labelling method enables epigenomic profiling with lower input.

Harada A, Maehara K, Handa T, Arimura Y, Nogami J, Hayashi-Takanaka Y, Shirahige K, Kurumizaka H, Kimura H, Ohkawa Y.

Nature Cell Biology, 2019. (189 samples)

Histone H3.3 sub-variant H3mm7 is required for normal skeletal muscle regeneration

Harada A, Maehara K, Ono Y, Taguchi H, Yoshioka K, Kitajima Y, Xie Y, Sato Y, Iwasaki T, Nogami J, Okada S, Komatsu T, Semba Y, Takemoto T, Kimura H, Kurumizaka H, Ohkawa Y. ***Nature Communications,*** 2018.

(117 samples)

MNase, as a probe to study the sequence-dependent site exposures in the +1 nucleosomes of yeast.

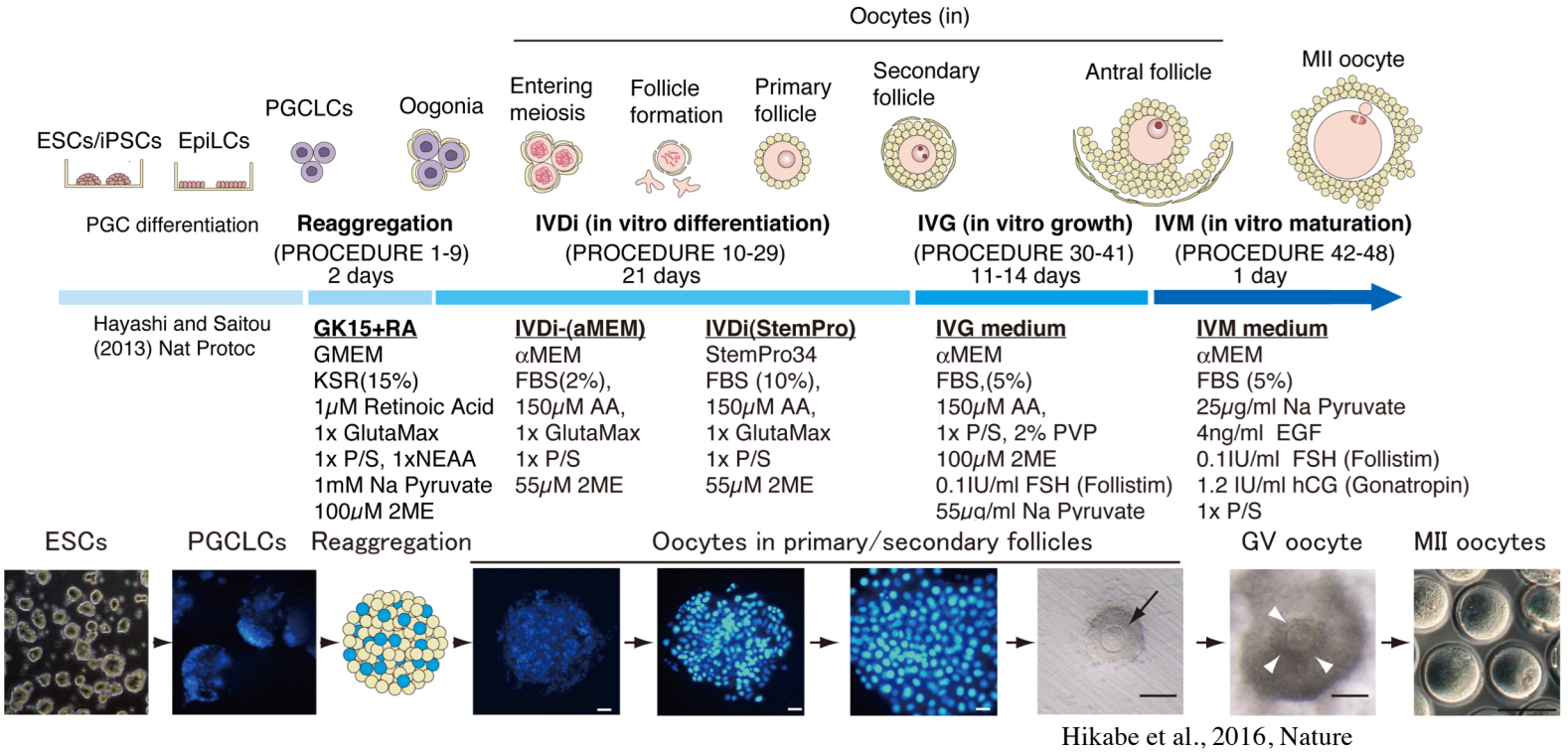
Luo D, Kato D, Nogami J, Ohkawa Y, Kurumizaka H, Kono H.

Nucleic Acids Res., 2018

林G

ES/iPS細胞から機能的な卵子への分化誘導系

本研究では、多能性幹細胞であるES/iPS細胞から誘導された卵子の性質を評価するために生体内とES細胞由来卵子のトランスクリプトームを比較した。

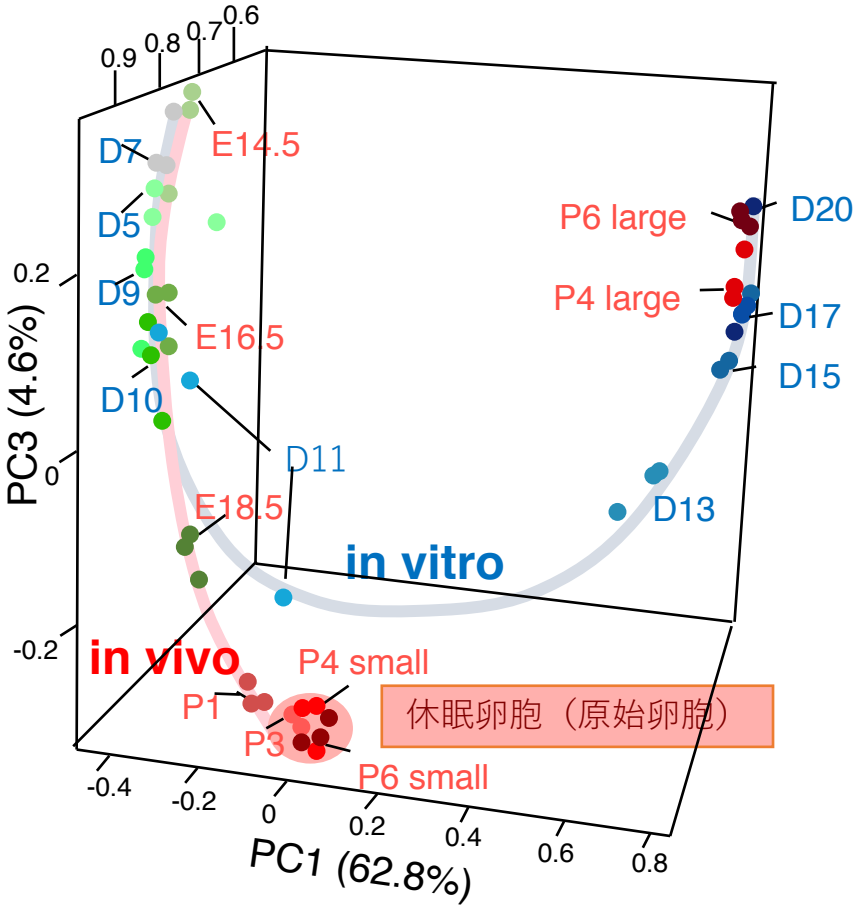


Hikabe et al., 2016, Nature

（過去の本プロジェクト成果）

林G

ES細胞由来卵子の性状解析



ES細胞由来の卵子と生体由来の卵子の発生過程における遺伝子発現比較を行ったところ、ES細胞からは休眠卵子である原始卵胞が形成されないことが明らかになった。

本解析において、従来350時間かかっていた、総計1,805,494,592塩基のゲノムへのアラインメント、種々の統計処理が、ITOの利用により25時間に短縮された。

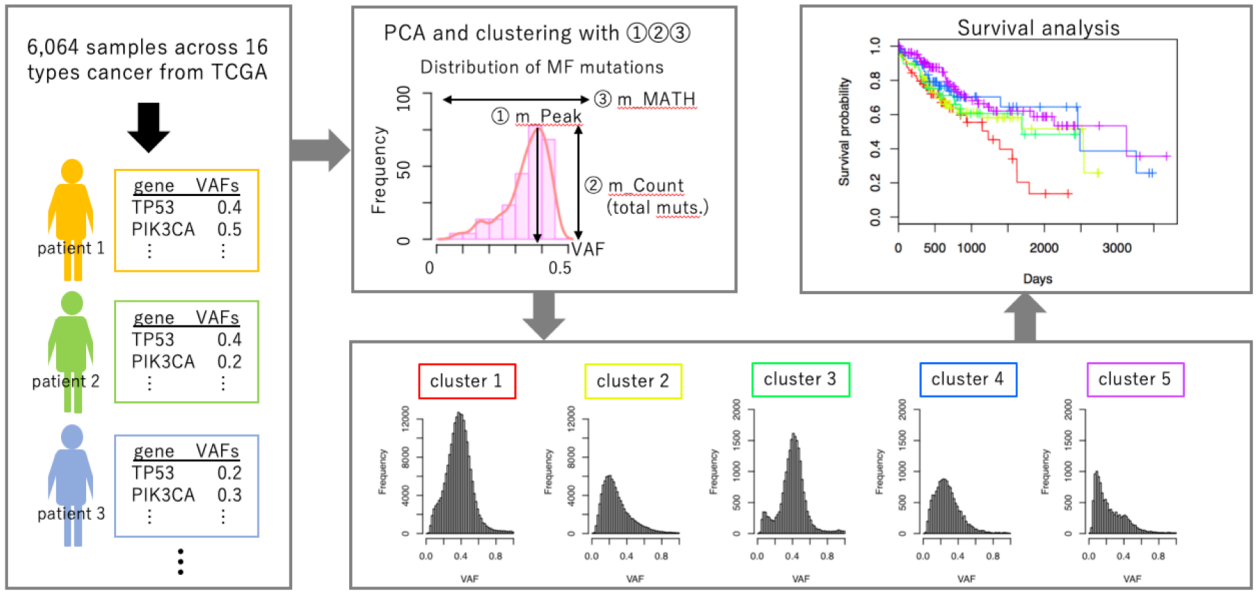
Shimamoto et al., in revision

須山G

腫瘍内不均一性を基にしたがんの予後解析

【研究目的】

- がんの治療の難題となっている腫瘍内不均一性とがん患者の予後との関連について解析する



【研究内容】

1）各サンプルの持つ変異データから、細胞集団中の変異頻度を計算
2）変異頻度のパターンから予後との関連を解析
3）予後の悪いサンプルにおける遺伝子発現パターンの解析（進行中）

須山G

【スパコン使用の意義】

- TCGAデータベースからがんのゲノム研究に関する大量のデータをダウンロードし、使用するにあたってのストレージ（現在使用している16種類のがん約6,000サンプル→約30TB）

- ダウンロードした大量のデータをマルチスレッドにより迅速に処理

（例：STARを用いたfastqファイルのマッピング1サンプルあたり：約1時間→約5分に）

【研究成果（謝辞記載）】

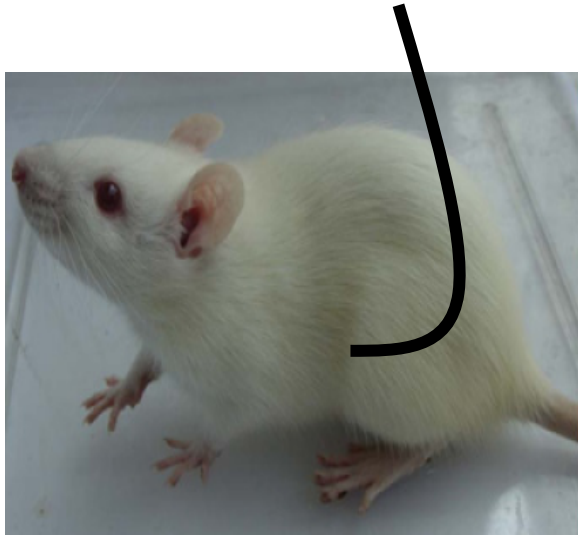
本プロジェクトのリソースを利用した研究成果は論文（2報）として報告した

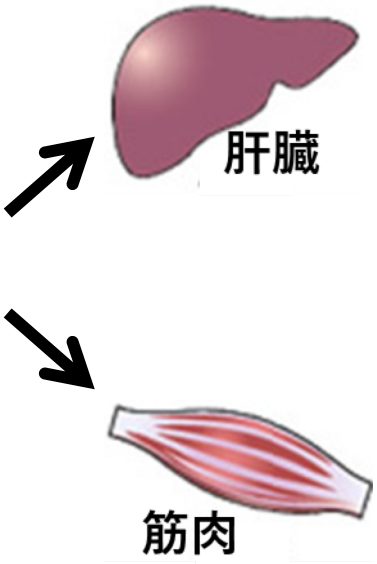
- Kikutake C, Yoshihara M, Sato T, Saito D, Suyama M. Intratumor heterogeneity of HMCN1 mutant alleles associated with poor prognosis in patients with breast cancer. Oncotarget. 2018; 9 (70): 33337–33347.
- Kikutake C, Yoshihara M, Sato T, Saito D, Suyama M. Pan-cancer analysis of intratumor heterogeneity associated with patient prognosis using multidimensional measures. Oncotarget. 2018; 9 (102): 37689–37699.

久保田G

生体内における肝臓と筋肉のシグナル伝達経路の違いをモデルを用いて定量的に明らかにする

インスリン刺激（門脈）





同じインスリンの標的臓器

⇒シグナル伝達経路の特性は同じ？異なる？

刺激濃度も異なれば応答も異なる

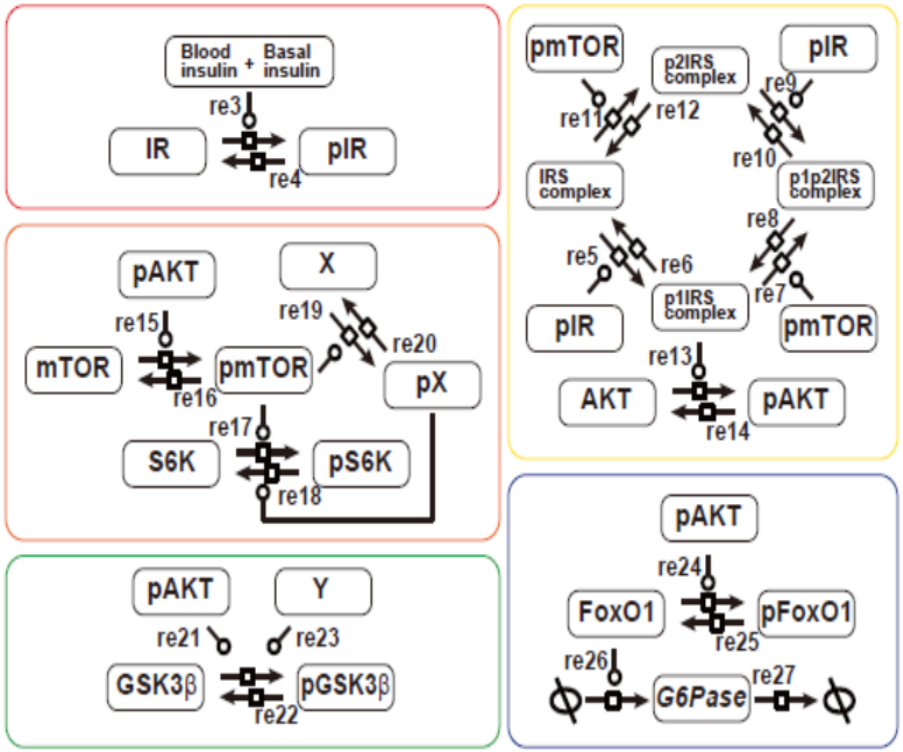
モデルでの定量的な比較

シミュレーションによる解析

久保田G

モデルの作成（パラメータの推定）には 大量の計算リソースを必要とする

シグナル伝達経路モデルの概要



モデルパラメータの推定
- 進化的アルゴリズムによる最適化
- 約50個のパラメータを推定
- 乱数のシードを変えて計算（200シード）
- 20時間/1計算

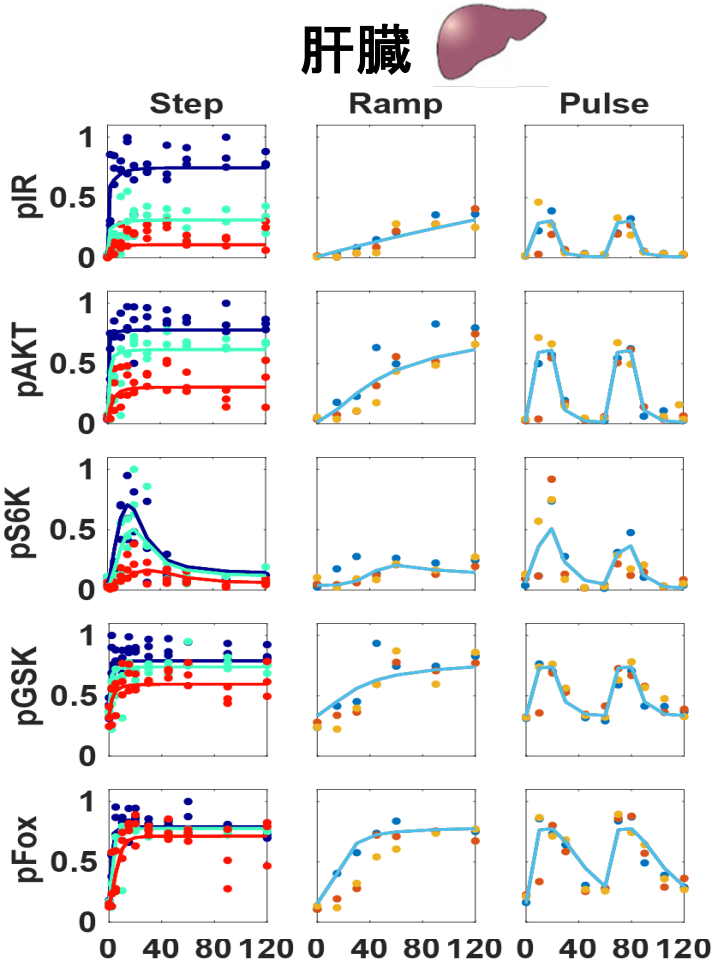
⇒ 20×200=4000時間の計算コスト

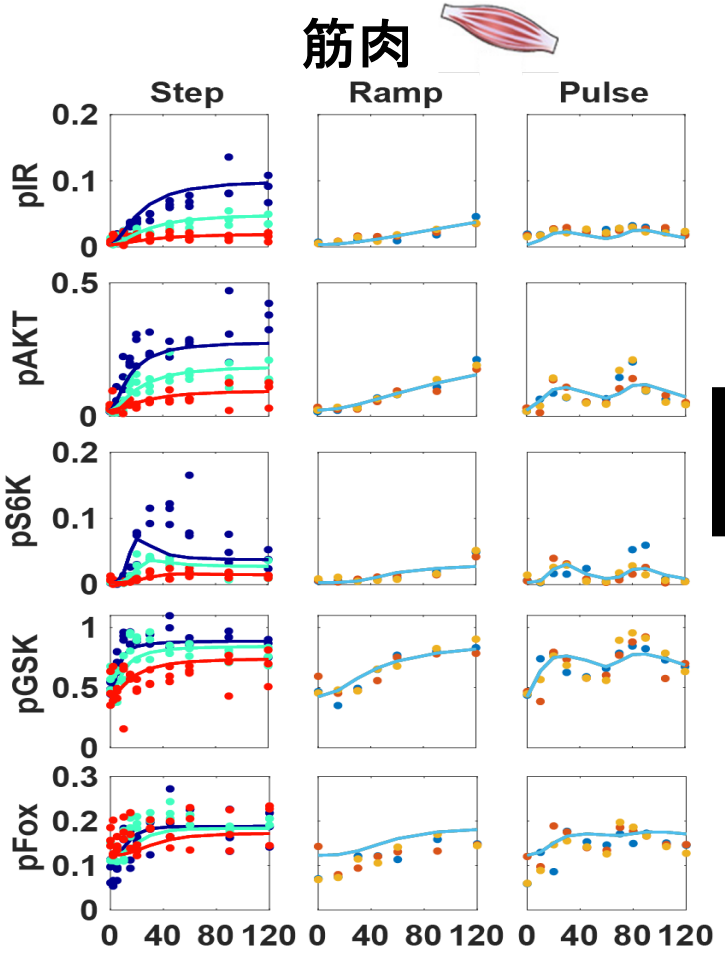
ITOで200コアの同時計算

⇒ 1日で1実験終了

久保田G

モデルの推定結果（暫定）

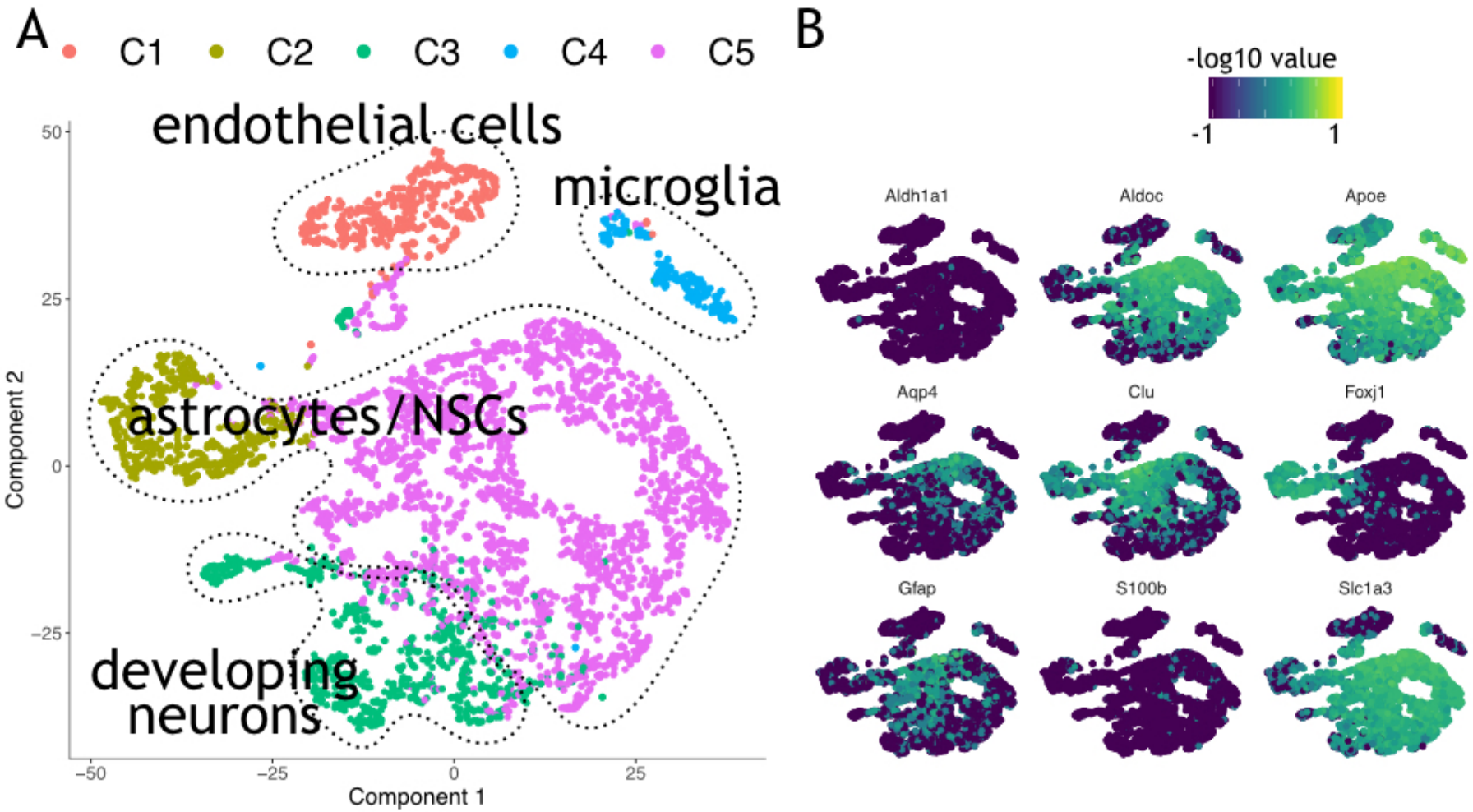




詳細な解析へ

中島G

脳におけるアストロサイトと神経幹細胞（NSCs）の不分離（従来法　左:Rtsne、右:マーカー発現）

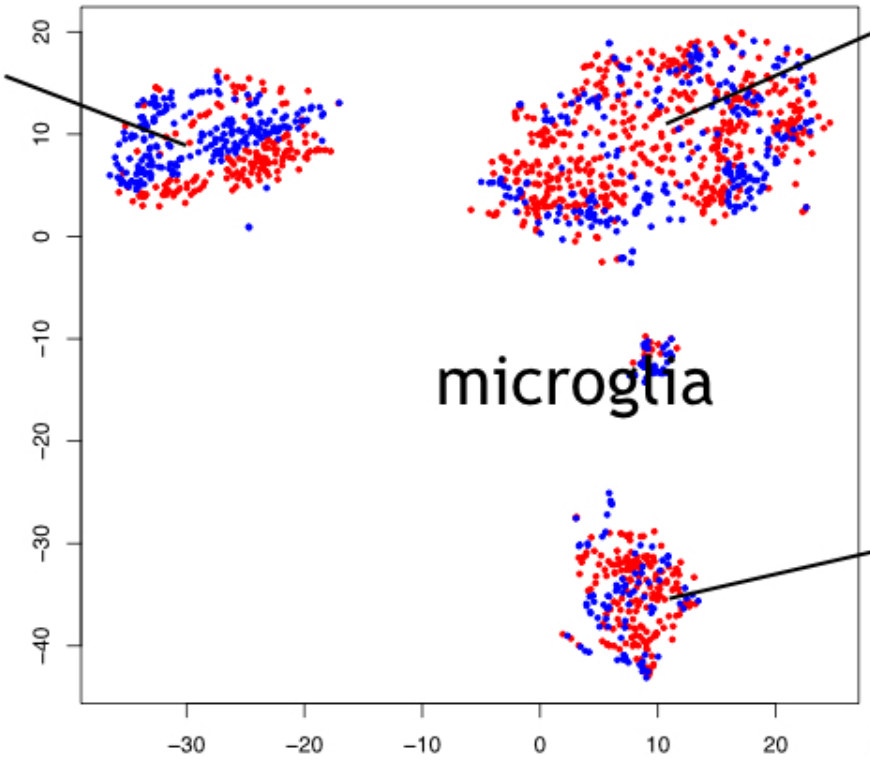


アストロサイトとNSCsは全く性質の異なる細胞であるにもかかわらず
遺伝子発現パターンが酷似しており、従来法では２群をうまく分離できていなかった

中島G

python-bhtsneを用いたクラスタリングによる遺伝子改変の影響を受ける細胞群の推定

total
WT: 861
KO: 592



野生型(WT) vs ノックアウト動物(KO)

分離能の大幅改善を突破口として、
大量データ処理からのPseudotime精細化も実現し、
病態発現における標的細胞群の同定に成功した

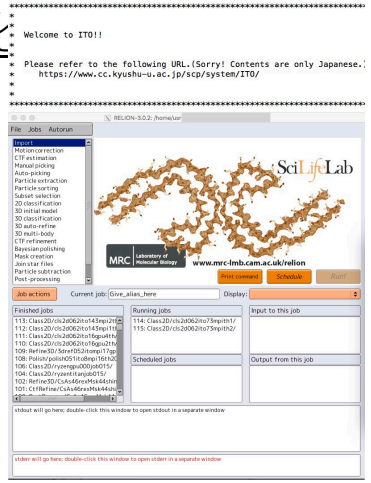
神田G

ITOを用いた並列計算による電顕単粒子解析の高速化

単粒子解析

- 数万～数百万の蛋白質等の電顕像から立体構造を算出。
- 単粒子解析の計算は並列化が非常に有効。
- 近年、検出器及び解析法の発展により結晶解析レベルの分解能に到達。
- 構造計算ソフト**Relion** が進展に大きく寄与、ベイズ統計を利用するため計算量が増大。
- 一般的に数百～千コアの計算機リソースが必要。 **Relion** 2 よりGPGPU 化。

ITOを用いた並列計算処理（CPU, GPU)により
単粒子解析の効率向上を試みた。

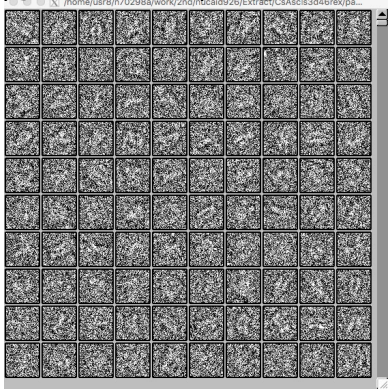


Relion 3.0
https://www3.mrc-lmb.cam.ac.uk/relion

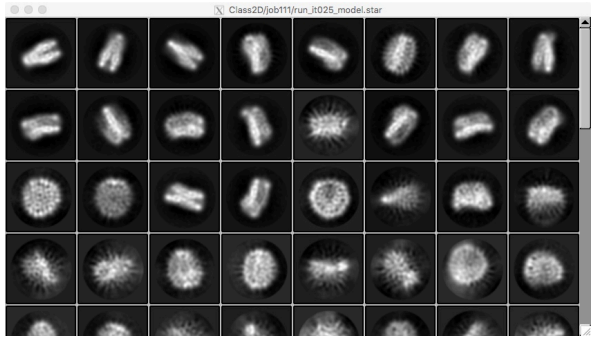
- ITOへのRelion 3.0 のインストール （GNUコンパイラー使用）
- 並列計算：MPIを使用（今回はOpenMPI 2.1.5) ; GPGPUはcuda 8.0を使用.

（評価法：以下の３プロセスについて行った）

1）Polish: 露光中の粒子の動きをトレース（第１step）、ブレ補正粒子像の算出（第２step）
2）Class2D: クラス平均像の算出（GPGPU）
3）ITOへの画像データ転送について



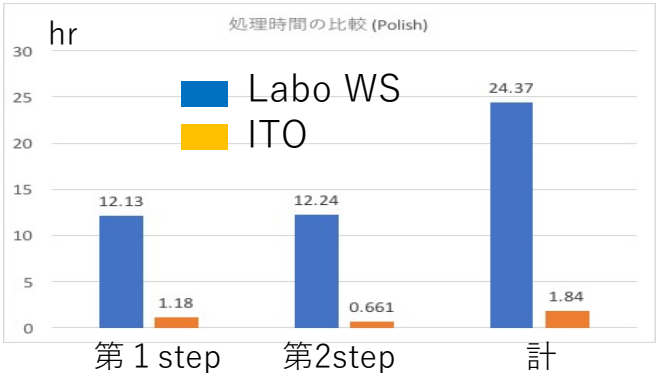
電子顕微鏡像（生画像データ）



ITOで算出したクラス平均像

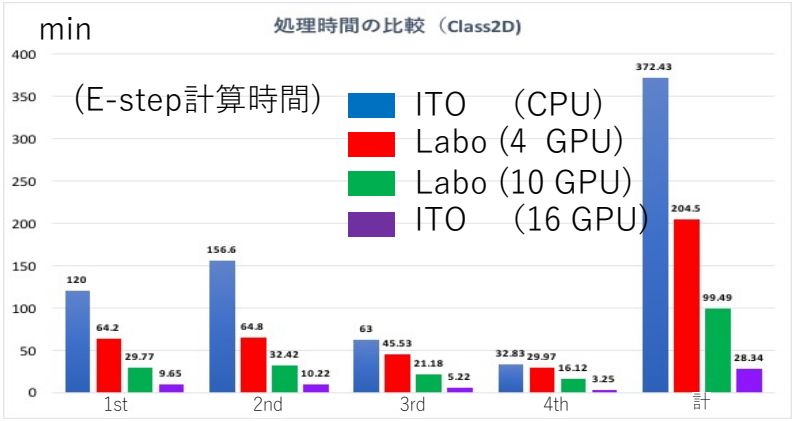
神田G

ITOを用いた並列計算による電顕単粒子解析の高速化



１）　CPU並列計算（MPI)

粒子像のブレ補正処理（Polish)は現在GPGPU化されていない。研究室のWS（Ryzen１６ core)と ITO (4 node：8MPI＊16 スレッド）を用いて計算、処理時間を比較した。
１０倍以上の高速化が見られた。
（研究室では８TB HDD（SATA3.0）を使用しており、ストレージのR/W速度の差も影響している可能性がある。）



2）　GPU並列計算（MPI & cuda)

187,968個の複合体粒子像　（200 x 200 画素）からクラス平均像（クラス数１００）を算出した。
並列計算が効果的なEMアルゴリズムの各E-stepの処理時間を比較。

ITO (CPU): 4 node-143 MPI で計算（CPUのみ）。
Labo (4 GPU)：GeForce 1080 x4 (通常のPCワークステーションで可能な構成)
Labo (１０ GPU)：GeForce 1080 x4 + 1080Ti x6 (研究室の１０-GPUワークステーション)
ITO (16 GPU): 4 node-16 GPUで計算（Tesla P100 x16)

ITOの使用により、研究室のGPUワークステーションの3～４倍の高速化。
（GeForceとTesla P100の性能差も寄与）

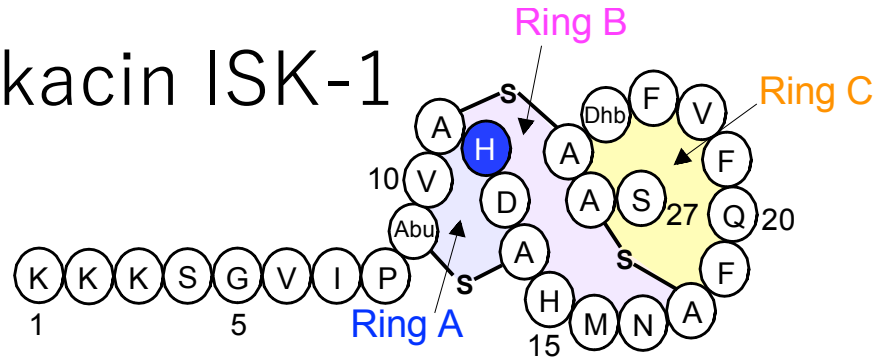
３）生医研ーITOへのデータ転送

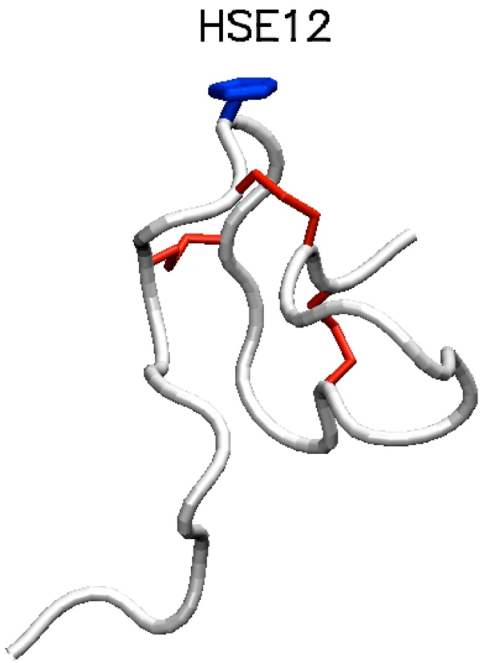
７月の大阪大学での測定データは７TBであった。FileZillaを使用し、ITOへの電子顕微鏡画像データ転送にかかる時間を計測した。80～110MB/secの速度で転送されることが確認できた。通常10 TB近いデータは3.5 インチHDD（SATA 3.0)の使用が主流であり、Localでのコピーが100 MB/sのオーダーであることを考慮すると、ITOへのデータ転送が解析の律速にはならないと考えられる。

結論：ITOの使用により、単粒子解析の大幅な解析速度の向上が見込める。
課題：ジョブ投入後の待ち時間、転休止期間。

秒オーダーの構造変化の分子動力学計算

nukacin ISK-1





モノスルフィド結合

Target MD (GENESIS)

restraints	:RMSD
N step	:5000
Time step	:2 fs
FF	:CHARMM

REUS (GENESIS)

restraints	:90 dihedral angles
Temperature	:326K
N replica	:64
N step	:50,000
Time step	:2 fs
Period	:5000

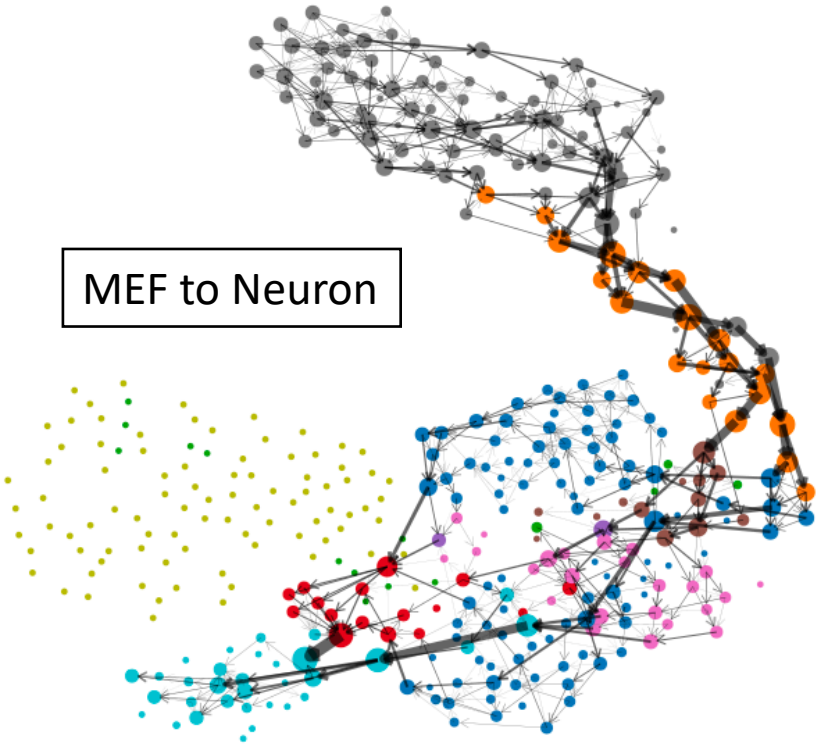
J Am. Chem. Soc., submitted

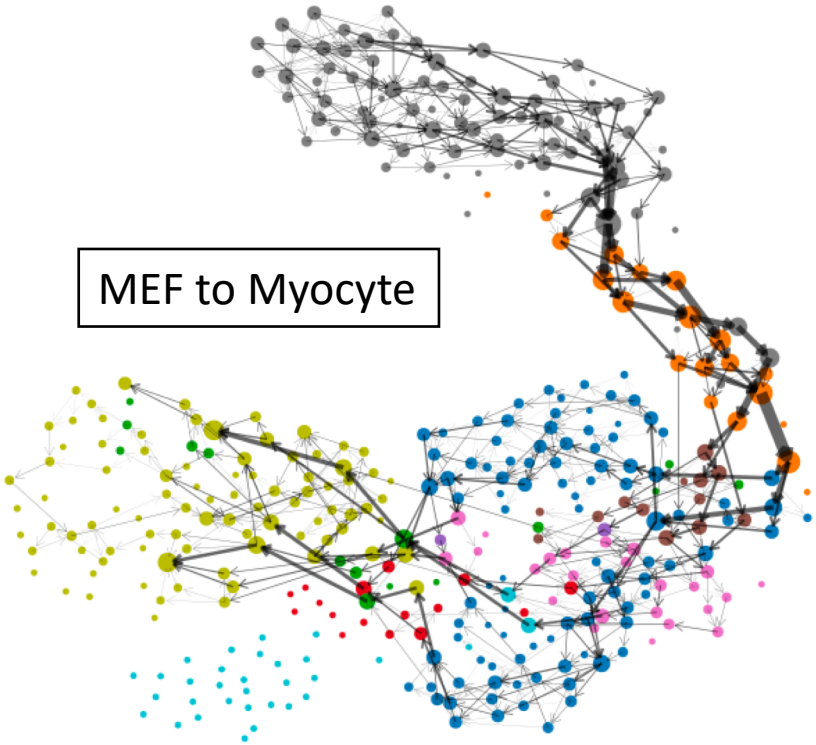
大川G

大規模な生物データを定性的に理解するための可視化・ダイナミクス抽出技術

ホッジ分解の数理的フレームワークを応用し、一細胞・一分子データの関係性（推移・因果）をスパースな有向グラフ上の**流れ**として簡潔に表現できる。

Incl. 距離行列構築、クリーク列挙





Maehara & Ohkawa, Biorxiv (preprint), 2019

まとめと今後の展開

- 昨今の大規模生命科学データ(NGS, Cryo-EM)の迅速な処理のためのプラットフォームとして利用→論文成果
- 生命科学分野の大規模データ解析を中心に、生医研・医学研究院グループで連携・利用することで、データ転送・計算処理に関わる問題意識を共有できた（ITO活用の裾野を広げられた）
- プロジェクト後半は（良い傾向として）リソース利用時間のグループ間衝突が起きはじめたため、次年度以降は各自独立のプロジェクトとしてより発展していってもらえれば。
- データサイエンス分野における新しい展開