

# シングルセル遺伝子解析 の高速化

エンドツーエンドの Scanpy パイプラインでインテル® Xeon®  
プロセッサが NVIDIA\* Tesla\* A100 を凌駕

Sanchit Misra インテル ラボ 上級科学的研究員  
Narendra Chaudhary インテル ラボ 科学的研究員

貢献者 : Padmanabhan Pillai インテル コーポレーション 上級科学的研究員、Bharat Kaul 同 Parallel Computing Lab ディレクター、Henry Gabb 同シニア主席エンジニア、Andrey Gorshkov 同 AI ソフトウェア・エンジニアリング・マネージャー、Pavel Yakovlev 同 AI フレームワーク・エンジニア、および Gurbinder Gill Katana Graph シニア・ソフトウェア・エンジニア

計測の分解能の向上は、さまざまな分野に革命をもたらしてきました。例えば、顕微鏡や望遠鏡の発明は科学に驚くほど大きなインパクトを与えました。シングルセル解析は、生物学における同様の革命の好例です。人間の身体は 40 兆個近い細胞からできています。歴史的に、これらの細胞は、ときには一度に数百万個という大きな単位で調べられてきたため、細胞間の違いをとらえることができませんでした。シングルセル解析は、細胞の個性を見るものです。新しいタイプの細胞を識別し、これらの細胞を区別するメカニズムを明らかにし、特定の病気や薬に細胞がどのように反応するかを示すことによって、細胞分化の謎が解明され始めています。この比較的新しい分野は、癌から Covid-19 関連の研究に至るまで、生物学的発見への計り知れない可能性をすでに示しています。

データ計測技術の進歩により、シングルセル・データの量は急速に増加しており、個々のデータセットのサイズも同様の速度で増加しています。このようなデータの解析には、通常、データ・サイエンス・パイプラインの実行が含まれます。パイプラインのステップは、パラメーターを変更しながら繰り返されることが多いため、ほぼリアルタイムで実行できる対話型のパイプラインが有効です。

## インテル® Xeon® プロセッサー 1 基で、130 万個のマウス細胞の ScRNA-seq 解析をわずか 7 分半で実現

細胞分化のさまざまな側面を研究するシングルセル解析には、多くの種類があります。シングルセル RNA-seq (scRNA-seq) 解析は、細胞間の遺伝子発現プロファイルの違いを研究します。この解析には、個々の細胞の遺伝子発現を測定する高度な技術であるシングルセル RNA シークエンシングが利用されます。

scRNA-seq 解析の典型的なワークフローは、各細胞の遺伝子発現量の行列から始まります (図 1)。データの前処理では、ノイズを除去し、データを正規化して、データセットの個々の細胞におけるすべてのヒト遺伝子の活性を取得します。このステップでは、データ収集から生じるアーティファクトを修正するため、マシンラーニングがしばしば使用されます。その後、次元縮小を行い、類似した遺伝子活性を持つ細胞をグループ化するためクラスタリングを行い、クラスターを可視化します。800,000 以上のダウンロード実績がある [Scanpy](#) (英語) は、この解析に最もよく使用されているツールキットの 1 つです。

130 万個のマウス脳細胞からなるデータセットでは、図 1 に示すパイプラインが通常、標準 Scanpy 実装 (ベースライン) を使用した Google Cloud Platform\* (GCP\*) 上の [単一 CPU インスタンス \(n1-highmem-64\)](#) で [5 時間近く](#) (英語) かかります。同じパイプラインについて、NVIDIA は、NVIDIA RAPIDS\* を使用した [単一の NVIDIA\\* A100 GPU 上で 686 秒](#) (英語) というエンドツーエンドの実行時間を報告しています。



図 1. 遺伝子活性行列から始まり、異なる細胞クラスターの可視化までのシングルセルの RNA シークエンシング・データの解析手順を示すパイプライン。

インテル ラボは、[インテル® oneAPI データ・アナリティクス・ライブラリー \(インテル® oneDAL\)](#) チームおよび [Katana Graph](#) (英語) と協力して、より優れた並列アルゴリズムを使用してパイプラインを高速化し、アーキテクチャーに合わせてパフォーマンスをチューニングしています。この取り組みはまだ進行中ですが、表 1 と図 2 に現在のパフォーマンスとクラウド利用コストを示します。この結果は、[Intel Investor Meeting 2022](#) (英語) で発表されました。GCP\* 上の同じ単一 CPU インスタンス (n1-highmem-64) で、パイプライン全体をわずか 626 秒で終了できるようになりました。第 3 世代インテル® Xeon® スケーラブル・プロセッサー (開発コード名 Ice Lake) が動作する新しい n2 インスタンス・タイプでは、パフォーマンスがさらに向上しています。また、パイプラインのメモリー要件を軽減したため、ハイメモリーの n2-highmem-64 インスタンスの代わりにハイ CPU の

n2-highcpu-64 インスタンスを使用できるようになりました。GCP\* 上の n2-highcpu-64 のシングル・インスタンスでは、パイプライン全体がわずか 459 秒 (7.65 分) で終了しています。これは、ベースラインの 5 時間と比較して約 40 倍高速です。NVIDIA\* A100 GPU のパフォーマンスと比べても約 1.5 倍高速です。

高速化とメモリー要件の軽減により、クラウド・コンピューティングのコストを大幅に軽減できました (表 1)。GCP\* 上の n2-highcpu-64 インスタンスのコストはわずか \$0.29 です。これは、ベースライン Scanpy を実行する n1-highmem-64 の約 1/66、NVIDIA\* A100 GPU の 1/2.4 です。

パイプライン・ステップ	CPU n1-highmem-64 64 vCPU (ベースライン Scanpy)	GPU a2-highgpu-1g Tesla* A100 40GB GPU (GPU により高速化された Scanpy)	CPU n1-highmem-64 64 vCPU (CPU により高速化された Scanpy)	CPU n2-highmem-64 64 vCPU (CPU により高速化された Scanpy)	CPU n2-highcpu-64 64 vCPU (CPU により高速化された Scanpy)
データの読み込みと前処理	1120	475	16.9	11.3	16.6
PCA	44	17.8	6.9	5.6	5.6
t-SNE	6509	37	216.2	175.6	172.8
K 平均法 (1 反復)	148	2	11.1	7.8	8.1
KNN	154	62	73.8	60.0	64.2
UMAP	2571	21	167.4	100.8	96.2
Louvain クラスタリング	1153	2.4	13.9	10.3	8.9
Leiden クラスタリング	6345	1.7	52.8	36.4	34.5
サブグループの再解析	255	17.9	23.9	20.8	19.2
Rest	39	49.2	42.7	33.6	32.8
エンドツーエンドの実行時間 (秒)	18338	686	625.7	462.1	458.8
オンデマンド料金 (US\$/ 時間) <sup>1</sup>	3.786	3.673	3.786	4.192	2.294
総コスト (US\$)	19.284	0.700	0.658	0.538	0.292

表 1. さまざまな GCP\* インスタンスにおける 130 万個のマウス脳細胞の scRNA-seq 解析の実行時間とクラウドコスト。最初の 2 つのカラムは単一 CPU インスタンス (n1-highmem-64) 上のベースライン Scanpy と単一 GPU インスタンス (a2-highgpu-1g) 上の GPU により高速化された Scanpy の公表されている (英語) 実行時間とクラウドコスト。最後の 3 つのカラムは、2 つの世代の CPU インスタンス・タイプの単一 CPU インスタンス (n1-highmem-64、n2-highmem-64、n2-highcpu-64) 上の CPU により高速化された Scanpy の測定された実行時間とクラウドコスト<sup>2</sup>。

<sup>1</sup>2022 年 5 月 15 日現在の <https://cloud.google.com/compute/vm-instance-pricing> の記載に基づく。

<sup>2</sup>2022 年 5 月 25 日現在のインテル社内でのテスト結果。

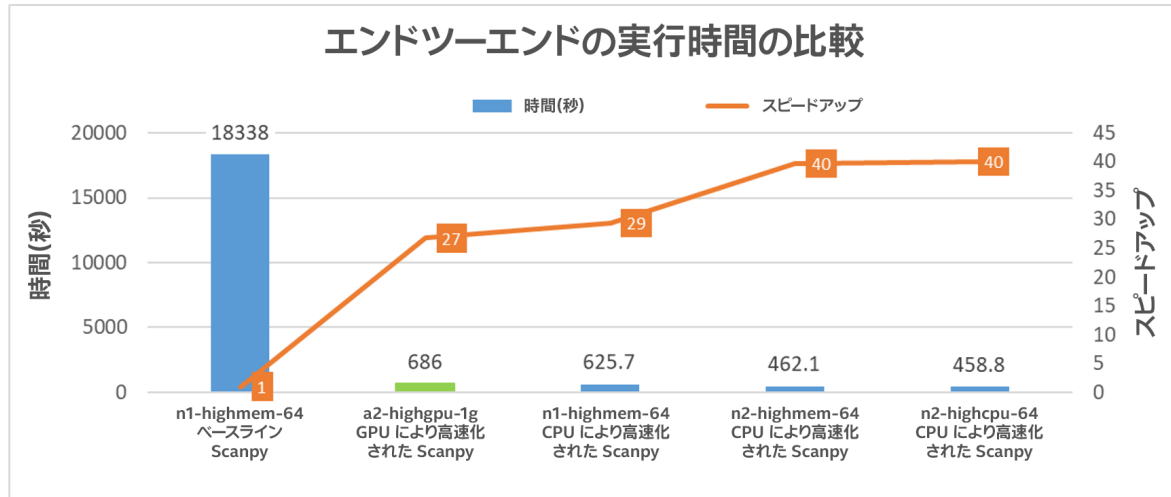


図 2. さまざまな GCP\* インスタンスにおける 130 万個のマウス脳細胞の scRNA-seq 解析の実行時間とスピードアップ。次のデータを使用：(1) 単一 CPU インスタンス (n1-highmem-64) 上のベースライン Scanpy と単一 GPU インスタンス (a2-highgpu-1g) 上の GPU により高速化された Scanpy の公表されている（英語）実行時間、(2) 2つの世代の CPU インスタンス・タイプの単一 CPU インスタンス (n1-highmem-64、n2-highmem-64、n2-highcpu-64) 上の CPU により高速化された Scanpy の測定された実行時間<sup>3</sup>。折れ線グラフは、n1-highmem-64 インスタンス上のベースライン Scanpy と比較したスピードアップを示しています。

## データ・サイエンス・パイプラインの高速化方法

以下は、このパイプラインのパフォーマンスを向上するため行ったステップの要約です。

- データの事前処理を効率化するため、ウォーム・ファイル・キャッシュを適用し、JIT コンパイラー Numba を使用してマルチスレッド化しました。これにより、ベースラインの事前処理パフォーマンスが 70 倍以上向上しました。
- また、K 平均法クラスタリング、KNN (K 近傍法)、PCA (主成分分析) の効率的な実装を含む [scikit-learn\\* 向けインテル® エクステンション](#) (英語) を使用しました。
- Scanpy は元々 scikit-learn\* の tSNE (t-distributed Stochastic Neighbor Embedding) 実装を使用していましたが、インテル® Xeon® プロセッサでは非効率的でした。tSNE の効率的な実装を構築することで、40 倍近い高速化を達成しました。
  - Barnes-Hut アルゴリズムの共有メモリー並列実装
  - Morton コードを用いた 4 分木の構築、ソート、要約の各ステップの並列化
- さらに、UMAP (Uniform Manifold Approximation and Projection) を最適化しました。
  - Python\* ソースコードの C++ への変換
  - 疑似乱数ジェネレーターの高効率なインテル® AVX-512/ インテル® AVX2 ベースの実装を作成
  - 固有値計算ステップにインテル® oneAPI マス・カーネル・ライブラリー (インテル® oneMKL) を使用
- Katana Graph との協力により提供された Louvain 法および Leiden 法コミュニティ検出アルゴリズムの高効率な実装をパイプラインに統合

<sup>3</sup>2022 年 5 月 25 日現在のインテル社内でのテスト結果。

これらの取り組みにより、大規模データセットの解析時間が大幅に短縮し、研究者はインテル® Xeon® プロセッサ上では 40 倍、NVIDIA\* A100 GPU 上と比べても 1.5 倍高速に作業を完了できるようになりました。

## まとめ

シングルセル解析は多くの分野に適用できます。腫瘍学、微生物学、神経学、生殖学、免疫学、消化器系や泌尿器系など、多くの分野で応用されています。作業時間の短縮により、さまざまな細胞をより深く理解できるようになり、大きな集団的利益をもたらす医学的進歩への道が開けることが期待されます。私たちは、scRNA-seq 解析パイプラインをさらに洗練させるため、tSNE、UMAP、Leiden ステップの改良に注力しています。

## システム構成

**GCP n1-highmem-64** : 1 インスタンス GCP n1-highmem-64 : 64 vCPU (開発コード名 Skylake)、合計メモリー 416GB、BIOS : Google\*、ucode : 0x1、Ubuntu\* 20.04、5.13.0-1024-gcp

**GCP n2-highmem-64** : 1 インスタンス GCP n2-highmem-64 : 64 vCPU (開発コード名 Ice Lake)、合計メモリー 512GB、BIOS : Google\*、ucode : 0x1、Ubuntu\* 20.04、5.13.0-1024-gcp

**GCP n2-highcpu-64** : 1 インスタンス GCP n2-highcpu-64 : 64 vCPU (開発コード名 Ice Lake)、合計メモリー 64GB、BIOS : Google\*、ucode : 0x1、Ubuntu\* 20.04、5.13.0-1024-gcp