No	質問	回答（記名なしの場合、飯田先生のご回答）
1	ASURATのmakeSignMatrix関数でエラーで止まってしまいます。メモリは十分ありますが、サイズの問題なのか止まってしまうので、サンプルコードなど提供されるとありがたいです。	ご質問ありがとうございます。各ステップの計算結果を確認しながら進めると良いと思います（やられていたら申し訳ありません）。記号の情報は metadata（sce）にあります。パイプラインは SingleCellExperiment オブジェクトをもとに進行しますので、よろしければ以下を参照ください。 Chapter 4 The SingleCellExperiment class | OSCA Introduction https://bioconductor.org/books/3.13/OSCA.intro/the-singlecellexperiment-class.html ASURATのサンプルコードは以下にございます。 ASURATBI | GitHub https://github.com/keita-iida/ASURATBI?tab=readme-ov-file
2	ASURATDBのインストールがうまくいかないです。	ご質問ありがとうございます。ASURATDBをインストールする際、いくつかのライブラリが同時にインストールされるはずです。以下の「Imports」に記載されたツールです。 ASURATDB | GitHub https://github.com/keita-iida/ASURATDB/blob/main/DESCRIPTION エラーの原因として、これらのインストールが上手くいっていないか、バージョンの衝突などが考えられます。ちなみに、Rのツール開発では、車輪の再発明は非推奨ですので、どうしてもこうした既存ツールを使わざるを得ず、バージョン衝突問題などの不具合が生じてしまいます（不便だなぁと、私も思っているところです）。
3	細胞のアノテーションについてです。どの程度に細かくクラスタリングされた状態でアノテーションをおこなうものでしょうか。	ご質問ありがとうございます。大切な点であると思います。データ解析を行う際は、研究目的も含め、データの中身を理解していることが前提となります。クラスタリング粒度を決定する決め手となるのは、その結果が生物学的に示唆的であるか、という点につきるかと思います。
4	細胞のクラスタリングを行っていく上で、様々な知識データベースを重ね合わせると確からしさは向上していく、あるいは既存の解釈でない情報が得られるようになるように拝見したのですが、その中の誤情報や不適切な情報を見分ける・排除するようにASURATはできているのでしょうか。	ご質問ありがとうございます。非常に重要な点であると思います。結論から言えば、ASURAT ver1 にはそのような機能は搭載しておりません。ユーザーが入力したキーワードを含まない（あるいは含む）記号のみを選択する仕組みは入れてありますが、ご質問の内容とは異なりますね？記号はクラスタリングへの寄与度に応じてランキングされて出てきますので、上位の記号から示唆的なものを選んでいるという状況です。ご指摘の点、なかなか難しい問題です。
5	#1の先生への質問です。これまでは形態学的な分類で細胞種は分類されてきたと思いますが、そういった情報との遺伝子発現を利用したクラスタリングとに相関関係があるかということも、形態による細胞腫分類のデータベースができれば可能なのでしょうか？	ご質問ありがとうございます。非常に面白いご質問ですね。細胞形態による細胞種分類と、遺伝子発現による分類の相関係数を直接的に計算するためには、細胞形態と発現の情報がシングルセルレベルで同時に得られている必要があるため、それはまだ難しいと思われます。ただし、全く不可能でもないと思います。例えば、平均的な細胞サイズが異なるいくつかのscRNA-seqデータを統合解析すれば、ある程度可能になります。私のこの回答は、データベースというのとは若干異なるかもしれません。
6	ミエロイド細胞を濃縮してシーケンスする場合もあると思いますが、そういったデータでも影響なく解析は可能でしょうか	ご質問ありがとうございます。ある程度均一な細胞集団で解析を行うということですね。やってみなければ分かりませんし、研究の目的にも依存するため何とも言えませんが、おもだった問題点は思いつきませんでした。
7	ASURATを利用する際に、必要なデータの前処理があれば教えてください。	ご質問ありがとうございます。ASURATは汎用ソフトウェアですので、チュートリアルには前処理も含め、全てできるように記載しております。データセットの性質に合わせ、以下の Vignettes をご覧いただければと思います。 ASURATBI | GitHub https://github.com/keita-iida/ASURATBI
8	非がん組織の解析に対して、がん細胞のデータベース情報があることで解析に影響は出ないでしょうか？	ご質問ありがとうございます。ASURATは知識データベースの入力を必要とするため、どのような場合であっても影響は受けます。膵がん腫瘍のデータ解析では、正常な細胞種別と、疾患のデータベースを結合して用いていました。記号はクラスタリングへの寄与度に応じてランキングされて出てきますので、上位の記号から示唆的なものを選んでいるという状況です。
9	scRNA-seqが自身の仮説の補強に使われがちな昨今、ソシュールの目指した言語学は、この問題の根本的な解決になり得るのでしょうか？	ご質問ありがとうございます。とても嬉しいです。ソシュールが記号学において目指したのは、我々の発話行為が言語体系によって束縛を受けるのだとすれば、意味の生産行為はどのように行われるのか？という根源的な問いであったと言われています（丸山圭三郎の解釈）。私はASURATの開発を通じて、記号学を生命の文脈に沿って完全に数学化することで、生命を記号の観点から理解してみたいと考えており、例えば、生命における意味の生産行為とは何だろうか？などの疑問に答えたいと考えています。ご質問の答えなっているか大変不安ですが。

No	質問	回答（記名なしの場合、飯田先生のご回答）
10	生データの良し悪しの評価はどのレベルで行えば良いでしょうか？	ご質問ありがとうございます。生データの良し悪しは、最初のデータ品質管理（QC）で分かる場合が多いです。典型的なのは、細胞数に比べて全リード数が少なすぎる場合や、ミトコンドリア関連遺伝子に割り当てられたリード数が異常に多い場合などで、後者は細胞が死につつある状態であると言われています。あとはデータ解析をやってみて、何も見つからない場合などにも、よく生データの品質が疑われますが、こればかりは結果論であることも多く、実はちゃんと解析したら示唆的な結果が得られたというような経験もあります。
11	がん細胞に囲まれた正常（？）細胞とありましたが、どういう生物学的意味があるのでしょうか？	ご質問ありがとうございます。正確には「がんに巻き込まれた正常細胞」でしょうか。おそらく、がんの浸潤により、正常な機能を失いつつある正常細胞のことを指していると考えられますが、具体的なことはまだ分かっていません。より詳細な推定を行うためには、細胞間相互作用などの分析が有効であろうと考えています。
12	ソシュールのロジックで生物学的記号論を試みると。結論が恣意的になりすぎる可能性はありますか？実在から乖離するリスクは？	ご質問ありがとうございます。はい、結局は人間が結論を出すため、ご指摘の可能性はあると思います。むしろ、そうなりすぎないように、事前にデータの内容や研究目的を深く理解し、生物学的に示唆的な結論を出す必要があります。ASURATが実行できるのは、人間が行う恣意的な生物学的解釈を、可能な限り自動化しているということになります。そのため、ある程度の恣意性を許しており、その自由度が“最適な”クラスタリングを得るために必要になります。ちなみに、ソシュール理論のすごいところは、恣意性を認めているという点にあります。客観性は科学の大原則と考えられていますが、本当にそうかな？と疑うことは面白い試みです。
13	非モデル生物に適用するためにはどのようなアプローチが考えられますか？	とても示唆的なご質問だと思います。ヒントは記号の作り方にあり、知識データベースを用いない記号の生成方法を考える必要があります。構想はあるのですが、未着手であり、今後の展開をお待ちいただければと思います。
14	最後のスライド，様々な記号の生成が可能とのことですが，ユーザの望む形が入るのではないでしょうか？	ご質問ありがとうございます。記号にはスコアがつくため、完全にユーザーが望む結論が得られるとは限りません。ただ、もちろんご指摘のようなバイアスが生じる可能性はあります。ASURATは人間が行う生物学的解釈という恣意的行為を、可能な限り自動化しているため、ある程度の恣意性を許しているわけです。しかし、最後は人間の決断に頼っているため、事前にデータの内容や研究目的を深く理解し、生物学的に示唆的な結論を出す必要があります。
15	細胞や遺伝子オントロジーを元に解析されているかともいますが、まだ明らかになっておらず、細胞や遺伝子オントロジーに登録されていないようなfunctionについてはどのように考えられていますでしょうか。	ご質問ありがとうございます。ASURAT ver1 では、ご指摘の内容は未解決であり、リミテーションでした。その解決策について、ヒントは記号の作り方にあり、知識データベースを用いない記号の生成方法を考える必要があります。構想はあるのですが、未着手であり、今後の展開をお待ちいただければと思います。
16	各記号の生物学的解釈は常に容易ですか？	大変鋭いご質問です。記号そのものの解釈は簡単ですが、問題はたくさんある記号の中から、どれを結論に用いるかという点にあります。この問題は、必ずしも容易ではありません。私の場合は、まず標準的な解析を行い、よい結論が得られなかったらASURATを用いるようにしています。結局は人間が結論を出すため、事前にデータの内容や研究目的を深く理解し、生物学的に示唆的な結論を出す必要があります。
17	どのツールを用いても、どこまで細かく分類するか（resolutionをどれくらいに設定するか）というのが難しいように感じています。ASURATでも同様のように見えますが、何か先生の中でどの程度細かくするかの基準の目安はありますか？	ご質問ありがとうございます。大切な点であると思います。データ解析を行う際は、研究目的も含め、データの中身を理解していることが前提となります。クラスタリング粒度を決定する決め手となるのは、その結果が生物学的に示唆的であるか、という点につきるかと思います。
18	ASURATを使用して生物の種間比較を行うことはできますか？	ご質問ありがとうございます。これはすごい質問です。実は、富山大学の先生がASURATを用いてこの問題に取り組んでおられ、非常に妥当なASURATの応用であると考えています。ありがとうございます。
19	これまでの、遺伝子の場合、遺伝子ごとに発現量の数値データが出てきたと思うのですが、ASURATにおいて記号を用いる場合、記号ごとにどのような数値が割振られるのでしょうか？	ご質問ありがとうございます。ASURAT ver1 では、ラフに述べると、記号に紐づく遺伝子の発現量の平均値をその記号のスコアとしています。より正確には、事前に発現量を中心化し（細胞間での平均値を0に揃え）、正に強く相関し合う遺伝子セット（ΩsまたはΩv）と、相関の弱い遺伝子セットΩwの重みつき平均を使って計算しています（←わかりづらいと思いますが、実際には簡単です。しかし、ここはある種のモデリング部分であり、まだ試行錯誤的であるというのが正直なところです）。
20	Monocyteのサブクラスの数の変化を把握するためには、各サブクラスのマーカーが把握できるだけのリード数を読めば良いと思いますが、例えば、疾患状態の変化に伴うMonocyteの各サブクラスの細胞の機能的な変化を把握するためには、さらにどの程度のリード数のシーケンスを読む必要がありますか。	ご質問ありがとうございます。リード数に関しては、通常の10xプロトコルで取得されたデータであれば概ね大丈夫だとは思いますが、細胞数や遺伝子数など、データ依存でもあるため、正確な数値を述べることは難しいです。私が解析したデータについては Figshare にもアップしているため、もし気になる場合はご参照ください。 Single-cell and spatial transcriptome datasets | Figshare https://doi.org/10.6084/m9.figshare.19200254.v3
21	膵管腺癌の空間オミクス解析は、どのようにして位置情報を保持しているのでしょうか？機械学習による画像認識でしょうか？	ご質問ありがとうございます。申し訳ありません、実験プロトコルの詳細について、私は詳しくないため、Visium のプロトコルをご参照ください。なお、原論文（Moncada et al., Nat Biotech, 2020）では Visium という用語は出てきませんが、今日で言うところの Visium 法に該当するようです。

No	質問	回答（記名なしの場合、飯田先生のご回答）
22	ASURATを動かすにはどの程度のPCが必要でしょうか？	ご質問ありがとうございます。今回取り上げた敗血症のコホートデータは、約10万細胞のscRNA-seqデータでした。2022年時点での最高スペックのMacBook Pro（96 GBメモリ）を用いて、ギリギリ計算可能でした。とはいえ、10万細胞だとtSNEにも相当時間がかかります。 ちなみに、最も計算時間がかかるのは相関グラフの作成です。遺伝子が2万個あると膨大な計算時間がかかるので、私の場合は、極端に発現量の少ない遺伝子を除き、遺伝子数を1万個弱に絞り込んだうえでASURATを実行しています。一方、細胞数はあまり問題にならない感じがします。
23	ASURATはヒト以外にもマウスでも利用可能でしょうか？	ご質問ありがとうございます。気にされているのはデータベースを集めてくるところでしょうか。ヒトとマウスのデータを収録した「ASURATDB」を作成し、GitHubにて配布しています。 ASURATDB｜GitHub https://github.com/keita-iida/ASURATDB
24	ASURATは組織の種類に制限はございますでしょうか？	ご質問ありがとうございます。データベースを適切に選ぶことができれば、おそらく問題はないと思いますが、確証はありません。私は現在、心筋症の解析のために左心室から取得した空間トランスクリプトームのデータを扱っています。ただ、左心室にどのような細胞がいるか分からないことが問題です。心筋細胞のデータベースがあるらしく、これを利用できるならと思っているところです。
25	ASURATに用いた記号空間で得意分野・不得意分野などあるのでしょうか？例えば、免疫細胞の分類に優れるが間質細胞は難しいなど。また新しい免疫細胞の分類はよく出現しますが、実際それに対応するのか検証しているでしょうか？	ご質問ありがとうございます。後半部分、特に良いご質問だと思います。私自身、ASURATを用いた経験はまだ多くありませんので、明確な回答は難しいです。なんとなくですが、オルガノイドのデータとは相性が良い気が・・・やはりいい加減な回答は控えたいと思います。新しい免疫細胞の分類の検証は良いテーマですね。将来、またクラスタリングツールをつくる機会があれば、ツールの検証として考えたいと思います。ありがとうございます。
26	今後、さまざまな細胞分類パッケージが出現するとしたら、どのパッケージを信じればよいですか。複数使用して検証する必要があるのでしょうか。	ご質問ありがとうございます。調べたところ、391個のクラスタリングツールが発表されていました。 Tools検索結果｜scRNA-tools https://www.scrna-tools.org/tools?sort=pubs&cats=Clustering どのパッケージを使用するかは、データや研究目的にも依存するため、一概には言えませんが、汎用ツールの選択は難しい問題です。理想的には複数使用して検証するのが良いのでしょうが、個人的意見としては、経験上、Seurat が最も良いと感じています。
27	ASURATの網羅性について教えてください。またde novoの情報についてのアプローチはあるでしょうか？	ご質問ありがとうございます。「ASURATの網羅性」が何を指すのか、ちょっと汲み取れませんでした。De novo の情報についてのアプローチとしては、複数の記号を組み合わせることで文として新たなアノテーションをつけるということは考えられます。例えば、「〇〇細胞の中で、××生物過程があり、△△パスウェイが活性化している」。しかし、おそらくお聞きになられたいところはそうではなく、知識データベースにない生物機能はどうするのか？というご質問かもしれません。これについて、ヒントは記号の作り方にあり、知識データベースを用いない記号の生成方法を考える必要があります。構想はあるのですが、未着手であり、今後の展開をお待ちいただければと思います。
28	ASURATで作ることができなかった記号の組み合わせの解釈と取り扱いはどうすれば良いでしょうか？	ご質問ありがとうございます。現在は、ASURATで作ることができなかった記号は捨ててしまっています。しかし、「〇〇細胞の中で、××生物過程があり、△△パスウェイが活性化している」というような、記号の組み合わせを考えることは大事であり、単語ではなく文で細胞集団をクラスタリングするというアプローチは面白いと考えています。
29	スライドp22の説明において「他と強く相関する遺伝子セットを設定した」という説明があったかと思いますが、その設定はどのように行なっておられるのでしょうか。そこで主観が入ってしまうのか、客観的な数学的な指標があるのかご教授いただきたく思います。	ご質問ありがとうございます。まさにご指摘の通り、この部分で主観が入ります。しかし、実はここに自由度を持たせることで、シャープなクラスタリングが得られるまで記号の生成を試すことができるという大きなメリットが生まれています。ただ、そうは言っても客観的な数学的指標も大事であるというご意見には賛同します。そこで、ASURAT ver2 では、数学的に定義された分離指標をもとに、記号空間をつくることができるようになっています。ただ、こうしてつくった記号空間は、今度は解釈性が弱くなっており、やはり ASURAT ver1 の方が強力なのでは？となって現在苦心しているところです。

No	質問	回答（記名なしの場合、飯田先生のご回答）
30	今回ご紹介いただいたMonocyteのサブクラスを特定する程度の情報を得るためには、1細胞あたりに換算してどのくらいのリード数のシーケンスを読む必要をありますか？リード数が十分でない場合、どういった影響が想定されるでしょうか。	ご質問ありがとうございます。リード数に関しては、通常の10xプロトコルで取得されたデータであれば概ね大丈夫だとは思いますが、細胞数や遺伝子数など、データ依存でもあるため、正確な数値を述べることは難しいです。私が解析したデータについては Figshare にもアップしているため、もし気になる場合はご参照ください。 Single-cell and spatial transcriptome datasets | Figshare https://doi.org/10.6084/m9.figshare.19200254.v3 リード数が十分でない場合、全体的なRNAカウントが下がるため、遺伝子×細胞のリードカウントテーブルが0だらけになります。そうすると、例えば UMAP 上で遺伝子発現レベルを可視化した際に、それが発現差のある遺伝子なのかが極めてわかりづらい、というような状況が起こり得ます。
31	biological processというのはすべての細胞でみられる生物学的過程で使われる遺伝子ということでよろしいでしょうか	ご質問ありがとうございます。はい、その通りだと思います（私がご質問の意図を正しく理解していれば）。
32	scRNA-seqの各解析ステップにおいて、標準的なソフトウェアを教えて頂けると有難いです。またバージョン間でどの程度バラつきが出るのかご経験があれば伺いたいです	ご質問ありがとうございます。いわゆる「汎用ツール」というものを使えば、大体すべてのステップを網羅することができます。scRNA-seq データの汎用解析ツールとして最もよく使用されるのが Seurat です。使いやすく、解析結果も安定しています。以下URLもご参照ください。バージョン間での解析結果のばらつきですが、最も重篤なのが、前のバージョンで動いていた関数が、新しいバージョンでは動かなくなるといった現象です。他は、統計解析におけるp値が若干変化する、tSNEやUMAPによる低次元可視化の結果の見た目が変わる、といった程度かと思いますが、データ依存ということもあり、あまり明確な回答はできないかと思います。 tools検索結果 | scRNA-tools https://www.scrna-tools.org/tools?sort=pubs&cats=Clustering
33	scRNA-seqの検出細胞数を、そのまま生体内での細胞数として解釈しても良いでしょうか？	ご質問ありがとうございます。生体内での細胞数とscRNA-seqの検出細胞数は全く異なります。相関もないと思います。scRNA-seqにより検出される細胞数は、実験プロトコル、シークエンスプロトコル（ドロップレット方式か、ウェル方式かなど）によっても大きく変わります。もしかして、細胞数の割合のことでしょうか？末梢血のデータなどであれば、取得部位によらず均一な細胞割合が得られるとは思うのですが。
34	クラスター間の遺伝子発現量を比較する際にデータ処理で気を付ける点はありますか？また、異なるデータセット間で発現量の比較をすることは可能でしょうか？	ご質問ありがとうございます。クラスター間の遺伝子発現量を比較する際にデータ処理で気をつけている点は、とにかくp値やFC（fold change）などの統計量のみに頼らず、必ず可視化を行うということです。私の経験談ですが、p値が非常に低い遺伝子について可視化を行ったところ、期待と全く異なっていたことがあります。それで、よく調べてみたところ、前の方のデータ処理で、誤って発現量データにマイナスの値を混入させていることが分かり、肝を冷やしました（マイナスの値が入ると、FCなどは通常計算できなくなります）。異なるデータセット間で発現量の比較を行うためには、データの品質がデータ間でだいたい均一である必要があります。Multiplex と呼ばれる実験プロトコルでは、異なるサンプルを混ぜてシークエンスするため、データ間での品質の違いがかなり軽減されます。
35	公共DBで公開されているscRNA-seqのデータの信頼性担保の仕方も教えてほしいです	ご質問ありがとうございます。私は、公共DBで公開されているscRNA-seqデータを用いるときには、時間がない場合を除き、かならず原論文を全部読み、内容を理解してからデータ解析に取り掛かります。論文の記述と図を読み、説得力があるかを吟味することで、その信頼性を確認しています。さらに、データ解析を実施すれば、論文の主張がどの程度正しいのかということも大体わかるかと思います。
36	scRNAseqで分類することは可能だと思うのですが、逆に、全ての細胞種でばらつき少なく、定常的に安定発現している遺伝子群についてのデータを抽出することはできるのでしょうか？	（飯田） ご質問ありがとうございます。クラスタ間で発現差の統計解析を行い、p値が大きく、FC（fold change）が小さい遺伝子を選択すればよいかと思います（Volcano plot等で）。ただ、ご指摘の内容について私は実施したことがないため、こちらに誤解があるかもしれません。 （粕川） 統計（検定）で発現の変化したものを取り出すことはできるのですが、逆に変化していないものを取り出すのは困難ではないかと思います（もし調査不足でそういう方法がありましたら申し訳ありません）。なので統計的手法ではなく、Fold change で見分ける方法や、分散や変動件数で評価する方法などが考えられるかと思います。qPCR用のreference geneを探索する解析手法もあり、そういったものも使えるかもしれません。以前私のいたグループで、発現量の頻度分布を単峰性の分布と多峰性の分布にフィッティングさせて、どちらのほうがよりよくフィッティングするかモデル選択的な方法を使って判断したこともあります。

No	質問	回答（記名なしの場合、飯田先生のご回答）
37	scRNA-seqではmRNAの3’末端側のみを読み取っていると思いますが、transcript単位での特定は難しいですか？	（飯田） ご質問ありがとうございます。全長RNAを読む実験技術について、私は明るくないため、実験の方に聞かれるのがよいと思います。 （粕川） 現状よく使われる 10x Genomic 社のプロトコルでは全長を決めることはできないのですが、SMART-seq2のように transcript 全体から断片を取り出してシーケンスする方法や、最近は long-readで読む方法も開発されてきていますので、scRNA-seqでも全長が分かるようにはなってきていると思います。ただ一般的に全長を読むことと発現プロファイルにおける発現量の定量性はトレードオフの関係になりますので、現状解析をするときは全長を決めるのか、発現プロファイルを取得するのかのどちらか一方に絞ることになるかと思います。
38	シングルセル解析で必要となるサンプル数の目安はありますか？	ご質問ありがとうございます。「サンプル数」とは、scRNA-seqのデータセットの数という理解で正しいでしょうか？それは研究の目的にもよるため、明確な回答はできませんが、通常は何がしかのコントロールサンプルを入れます。例えば、時系列データであれば day 0 のサンプル、介入実験であれば対照群のサンプル、などです。「サンプル数」が「細胞数」を指している場合は、最低でもデータQC後に2000細胞はあってほしいです。遺伝子数 p に比べて細胞数 n が少なすぎると、p >> n 問題という統計学上の困難が発生するためです。
39	プログラム細胞死を引き起こす細胞のシングルセル解析はどうやってやればいいですか？	ご質問ありがとうございます。こちらは少し曖昧であり、スタディデザインについて、捉え方も何通りか考えられるため、回答は控えたいと思います。でも、テーマはちょっと面白そうですね。