様式６

申請先学部　医　学部　採択番号　No.　1

２０１９年度大阪大学未来基金【住野勇財団】学部学生による自主研究奨励事業研究成果報告書

ふりがな 氏名	やまだ　こうき 山田　航暉	学部 学科	医学部医学科	学年	２年
ふりがな 共同 研究者氏名	いしかわ　かいと 石川　海斗	学部 学科	医学部医学科	学年	１年
	まつもと　やすなり 松本　康成		医学部医学科		１年
					年

アドバイザー教員 氏名	鈴木　顕	所属	大阪大学大学院医学系研究科遺伝統計学

研究課題名	RNAseq 解析パイプライン「ikra」の開発及び、「ikra」を用いた RNAseq メタ解析手法の確立
研究成果の概要	研究目的、研究計画、研究方法、研究経過、研究成果等について記述すること。必要に応じて用紙を追加してもよい。（先行する研究を引用する場合は、「阪大生のためのアカデミックライティング入門」に従い、盗作剽窃にならないように引用部分を明示し文末に参考文献リストをつけること。）

研究目的

今日、次世代シークエンサの登場により、２１世紀初頭に比べ格段に解析スピードが上がり、ゲノム解析技術が向上した。ヒトの全ゲノムをシーケンスするコストも年々減少しており、それに伴って得られたデータの量は増加している背景があり、大量に存在するデータを用いて解析していくことが必須である。

そして、次世代シークエンサを用いた解析として代表的なものに RNA-seq が挙げられる。従来のマイクロアレイでは困難であった解析が RNA-seq により行われるようになった。網羅的に RNA の発現量を解析していくことで、細胞集団を示すマーカー探索、細胞の分化に関わる遺伝子の同定などを行うことができ、また、新規転写物や新規スプライシングバリアントの探索など、現在様々な分野において RNA-seq が用いられている。

RNA-seq は、大量の配列データが得られる次世代シークエンサを利用して遺伝子発現解析における様々な目的に使用される手法で、遺伝子配列情報の取得、低発現遺伝子の検出、定量性などにおいて高いパフォーマンスを持っている。特に遺伝子発現変動解析においては、マイクロアレイに比べ非常に広いダイナミックレンジを持つため、低発現・高発現の両方の転写産物を高い精度で検出でき、遺伝子の発現変動を網羅的かつ定量的に解析できる。

次に RNA-seq 解析の流れを説明していく。まず、次世代シークエンサからサンプルごとのシーケンスしたデータを取得する。得られたデータのクオリティコントロールを行い、アダプター配列やクオリティの低い配列を除去する。そしてリファレンス配列としてトランスクリプトーム配列にリードを擬似的にマッピングし、リードがゲノム上のどの領域から得られたものかを知る。そして各領域に何本のリードがマッピングされたか、発現量を取得する。最後に、得られた発現量を用いて、統計的に解釈を施していき生物学的な意味を見いだしていく。一例として、RNA-seq 解析のためのインタラクティブな web アプリケーション iDEP(http://bioinformatics.sdstate.edu/idep/)があり、発現変動遺伝子(DEG)やパスウェイについてグラフィカルに調べることが可能である。

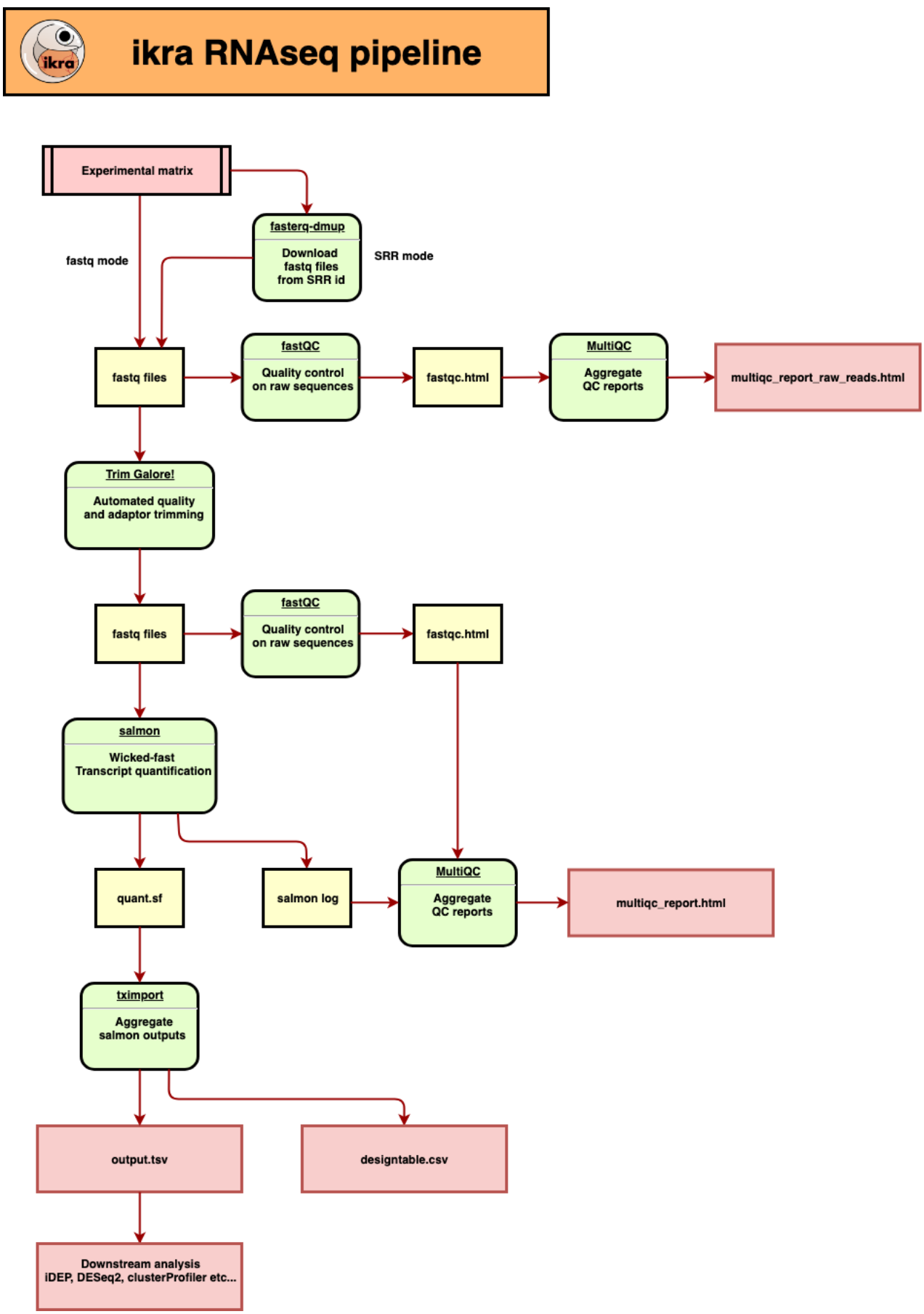

図１　パイプラインの流れ

以上のように RNA-seq 解析は現代の生命科学分野において非常に重要であるが、解析の手数の多さ、マニュアルで行うことによるサンプルの取り間違いなど様々な人為的なミスが考えられるため、RNA-seq 解析の一連の処理を自動で行う仕組みを作成することは有効であると考えられる。そこで、RNA-seq 解析の自動パイプラインである「ikra」(https://github.com/yyoshiaki/ikra, 図 2)の開発に今回の自主研究において取り組む。「ikra」は以上のデータ取得、クオリティコントロール、RNA 定量、可視化を全て自動で行えるものである。(図 1: パイプラインの流れ)

このように RNA-seq 解析をしていく上で非常に有用な「ikra」だが、使用していく中で改善すべき点がいくつか見つかった。実際の使用における問題点を解決すること、そして自分たちの考えたテーマで実際にパイプライン「ikra」を用いて解析し、RNA-seq 解析における「ikra」の有効性を確認するとともに新たな改善点を探ることを目的に我々は研究を行なっていった。

図 2　パイプライン「ikra」のロゴ

研究計画

RNA-seq 解析パイプラインツール ikra の開発と、データベース上の様々なデータを元にした ikra の動作確認・解析の２軸で研究を行なった。

開発・解析に際して、個人所有の PC では、メモリやストレージなどに限界がある。そのため、クラウド上にてリソースを使用することができる、AWS (Amazon Web Services) を利用して開発を行う。

また、バイオインフォマティクスの分野における最先端の知識を習得するため、ライフサイエンス統合データベースセンター(DBCLS)に訪問させていただき研修を行う。実際に生命科学に関する様々なデータベースを管理されておられる方々からの直接のフィードバックをいただき、開発に活かす。

さらに、ikra をオープンなパイプラインとして広く利用してもらうため、使用上の課題点について使用されている方からの意見を参考に、解決・バージョンアップしていく。

研究方法

個々人で RNA-seq を用いて研究を行なっている論文を選び、その論文データを再解析し、論文に書かれている結果と同じ結果が出ることを確認する。正しいデータを取得・解析するため、論文の内容の面でも吟味しながら、選び方に注意した。またそのように実際に ikra を実行する際に見受けられる、ikra の改善点をメンバーで共有し開発の工程に取り入れた。また、ikra の開発を行っているソフトウェア開発のプラットフォームである GitHub 上に寄せられる、ikra を実際に使用した方々からの意見もメンバー内で話し合った。こういった方法で、ikra の改善点を見つけ、その点を改善していくという形で開発を行った。

図 3　DBCLS

DBCLS
Database Center
for Life Science

DBCLS（図 3;DBCLS のロゴ）での研修では、バイオインフォマティクスの第一人者である、坊農秀雅先生をはじめとするバイオインフォマティクスの知識に秀でた先生方に幅広いご指導をいただいた。合宿では、実際に日々RNA-seq を用いた研究、解析をされている先生方から、ikra に関する意見を聞くことができた。ikra の開発に関する話や助言、技術の解説はもちろん、広くバイオインフォマティクスという技術に関する知識を教えていただいた。坊農秀雅先生は、バイオインフォマティクスに関する様々な書籍を執筆されている。研修では、それらの書籍を参考に RNA-seq 以外にも様々な解析手法について自らの手でプログラムを実行、様々な解析も行うことができた。書籍の執筆者に直接質問などをしながらコードを実行し解析をするという体験は、DBCLS での研修ならではであり、開発において有効となるスキルの向上ができた。（図 4;著書「生命科学者のためのデータ解析実践道場」）

DBCLS では、生命科学にまつわるさまざまなデータベースの作成、管理が行われている。バイオインフォマティクスで研究するにあたり、不可欠な知識であるデータベースの扱い方なども教わった。最先端の研究が行われている場所でのワークアウトによって、新しい発想や、バイオインフォマティクス、プログラミングの豊富な知識を得ることができた。

図 4

研究成果

研究目的でも記載の通り、RNA-seq の解析においては、「データの取得、クオリティコントロール、マッピング、可視化」と 1 つのサンプルあたりに対する処理が多く、さらに実際の現場では何サンプルも扱うことが多いため、これまでのマニュアルでの解析には手間がかかっていた。処理時間の多さ、処理を繰り返すことによるタイプミス、サンプルの取り間違いなど人為的なエラーも免れない。そういった観点から、今回のパイプライン「ikra」の開発に着手した次第だが、実際多くの方に使用してもらうことができ、よいフィードバックを得ることができた。Web 上でオープンでプログラムを配布する、オープンソースの仕組みを取り入れたことによる。（図 5）

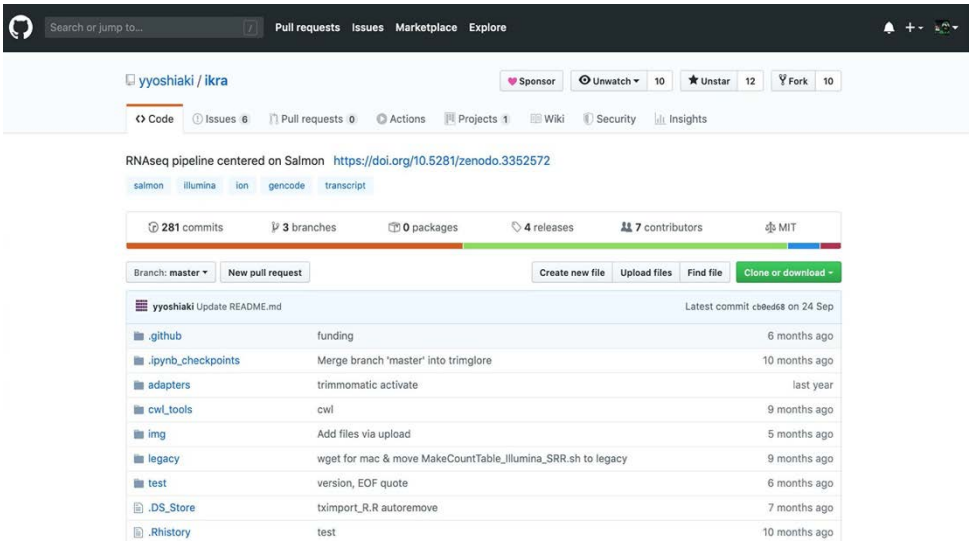

図 5　開発上の GitHub の様子

また、使用上における細部の不具合・改善点について、以下の点について改善することができた。

1.　世界中の研究者に閲覧・活用してもらうため、ドキュメント（説明書）を英語に対応させた。

2.　Zenodo という研究データレポジトリに登録することによって、DOI を取得することができた。これにより、パイプラインが使用された場合はその論文に引用されることができる。

3.　Travis CI というプログラムの動作性を自動でチェックさせることに対応させた。常に細かいエラーと戦うパイプライン開発においては、バグの修正後、正常にプログラム全体が動作するか確認することは必須であるが、その確認を自動化することによって開発上の利便性が高まった。

4.　パイプラインのオプションの拡充をすることができた。

さらに、パイプライン「ikra」を活用したものとして、RNA-seq のメタ解析にも取り組んだ。メタ解析においては、特に多くのサンプルデータを用いるため、デスクトップやラップトップの PC 以上の計算リソースが必要であったため、AWS を使用した。
今後ますます増えていくであろう大量の生命科学データを扱う際に、AWS 等のクラウドサービスを活用することは有用であると分かった。

また、今後の開発における展望として、ikra のパイプラインの出力結果をより上手く活用するために、IsoformSwitchAnalyzeR に対応させるなど、オプションを拡張させていきたい。
IsoformSwitchAnalyzeR(http://bioconductor.org/packages/release/bioc/vignettes/IsoformSwitchAnalyzeR/inst/doc/IsoformSwitchAnalyzeR.html)とは、RNA-seq データから予測される機能的結果を伴うアイソフォームスイッチの統計的識別と視覚化を可能にするツールであり、iDEP などのツールとともに活用させていきたい。
まとめ
今回の自主研究において、RNA-seq 解析パイプライン「ikra」の開発及び、「ikra」を用いた RNA-seq メタ解析手法について探ることができた。
RNA-seq の自動解析パイプラインは多くのサンプルを扱う際有用であり、また多様なデータを用いたメタ解析においても活用できることがわかった。
生命科学のデータ解析においては、日々日進月歩で技術の更新が行われており、それに伴ってこのパイプライン「ikra」もバージョンアップさせていくことが不可欠である。自主研究終了後も、バイオインフォマティクスの分野に貢献できるよう、パイプラインのさらなる開発を行っていきたい。

参考文献

1.　Hiraoka, Yu, Yamada, Kohki, Kawasaki, Yusuke, Hirose, Haruka, Matsumoto, Yasunari, Ishikawa, Kaito, & Yasumizu, Yoshiaki. (2019, July 27). ikra : RNAseq pipeline centered on Salmon. (Version v1.2.1). Zenodo. http://doi.org/10.5281/zenodo.3352573

2.　Mortazavi, A., Williams, B., McCue, K. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621-628 (2008) doi:10.1038/nmeth.1226

3. Ge, S.X., Son, E.W. & Yao, R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics 19, 534 (2018) doi:10.1186/s12859-018-2486-6

4. Kristoffer Vitting-Seerup, Albin Sandelin, IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences, Bioinformatics, Volume 35, Issue 21, 1 November 2019, Pages 4469-4471, https://doi.org/10.1093/bioinformatics/btz247