卒論論文ダイジェスト

RNA 編集サイトの検出ソフトウェアの設計と実装

慶應義塾大学環境情報学部　石黒　宗

論文要旨

RNA 編集とは、転写物へ位置特異的に一塩基置換を引き起こす転写後修飾の一種として知られ、アデニン (A) からイノシン (I) への A-to-I 編集がヒトやマウス、ショウジョウバエから多数報告されている。この A-to-I 編集は ADAR と呼ばれる二本鎖 RNA 結合タンパク質によって触媒されることが知られており、翻訳の段階で置換されたイノシンはグアノシンとして認識されるため、編集を受けた転写物は翻訳の過程において、非同義置換によるスプライシングサイトの変化やタンパク質の高次構造の変化、miRNA への編集を介した遺伝子発現の抑制など転写調節に幅広く関与していることが報告されている。近年、RNA-seq データを用いたゲノムワイドな編集サイトの同定が多数の組織およびセルラインを用いて行われ、ヒトでは数万箇所の編集サイトが報告されている。RNA 編集サイトはゲノムと転写物の一塩基のミスマッチとして検出可能だが、シーケンシングやマッピングに起因した擬陽性を多く含むため、真の編集サイトと擬陽性を高精度に分離する検出手法がこれまで多く提案されている。しかしながら、解析に使用された手法の多くはソフトウェアとして公開されておらず、RNA-seq データを対象とした編集サイトの検出ソフトウェアは現時点で一つ存在するのみである。そこで本研究では、既存のソフトウェアよりも高速かつ低メモリで動作し、アラインメントデータへの統計的なフィルタリング手法、実験デザインを考慮した解析を可能にする RNA 編集サイトの検出パッケージの開発を行った。本パッケージは、既存のソフトウェアと比較して高速か低メモリで動作し、付属するベンチマーキングツールによって、検出した編集サイトの検出精度を定量的に評価することを可能にした。尚、本パッケージは、GPL の元、オープンソースのフリーウェアとして `https://github.com/soh-i/Ivy` においてソースコードを公開している。

1.1 研究背景

現在、RNA-seq データを対象とした RNA 編集サイトの検出ソフトウェアは、REDItools (Picardi and Pesole, 2013) の一つの実装に限られている。そのため、SNP や SNV を DNA-seq データから検出する変異解析用のソフトウェアとして開発された Samtools mpileup (Li *et al.*, 2009) や GATK (McKenna *et al.*, 2010)、SOAPsnp (Yu and Sun, 2013) を転用した研究例も複数ある (Chen and Bundschuh, 2012; Danecek *et al.*, 2012; Peng *et al.*, 2012; Sanjana *et al.*, 2012)。流用を可能にしているのは、RNA 編集サイトも SNP/SNV の検出も本質的にはショートリードのマッピング結果からゲノム配列との一塩基ミスマッチを検出することにほかならないからである。しかしながら、DNA-seq と RNA-seq のアラインメント結果を観察すると、一般に RNA-seq データは DNA-seq に対して数百倍の変異箇所が見られる。これは、RNA 分子の不安定性や複数のマッピングバイアスが影響しているからである。

こういった現状において、一つのソフトウェアで RNA 編集サイトの検出が完結した例はこれまでになく、実験で得られた RNA-seq データを参照ゲノム配列へ適切なパラメータでマッピングし、そのアラインメントについて数個から多い時には 20 以上のフィルタリングを通し、最終的に通過した箇所を RNA 編集サイトとしてリストするという方法が用いられる。変異解析のソフトウェアを用いた場合でも、下流解析では独自のフィルタリング過程をほぼ必ず設けており、擬陽性を減少させる工夫が行われている。そのため、必然的に情報解析のワークフローは複数のフィルタリングと条件分岐によって複雑化している。

超並列シーケンスデータを用いた RNA 編集サイトの検出には、現在二つの問題がある。一つ目は、高精度な検出のために解析が複雑化し、簡便かつ高速な解析が困難となっていることである。使用したソフトウェアや解析方法の詳細なパラメータに関しては、論文中では記述される。そのため、論文ごとに解析手法の記述には粒度の違いが見られ、完全な再現が困難な場合もある。こういった現状では、仮に先行研究ごとにシーケンスデータが公開されていたとしても、複雑な解析パイプラインを再現し、優れた手法を他のデータへ適用することや、追証実験を行い難いという問題を発生させる。二つ目の問題は、新規の検出手法によって編集サイトを検出した場合に、検出精度の検証方法がばらつき、手法やパラメータの影響についての比較検討が困難だということである。卒業論文の第 2 章では、検出手法の精度比較を主題とし、情報検索の分野で利用されてきた適合率や再現率の導入による解決方法の提案を試みたものであった。

本研究では、上記二つの問題を解決するため、超並列シーケンスデータを対象とした RNA 編集サイトの高速かつ高精度な検出に加え、精度検証を行うソフトウェア・パッケージ Ivy の開発を行った。Ivy はコマンドラインツールとして実装され、RNA 編集サイトを検出するためのツールと精度検証を行うためのベンチマークツールが付属されたパッケージである。Ivy は、GNU GPLv3 (GNU General Public License version 3) の元、オープンソースのフリーウェアとして、GitHub の `https://github.com/soh-i/Ivy` においてソースコードを公開している。

1.2 システムの設計

1.2.1 ivy の設計と実装

Ivy は Unix 環境で動作するコマンドラインツールとして Python v2.7.5 によって実装された。図 1.1 には、Ivy システムの設計の全体像を示した。Ivy は、オブジェクト指向プログラミングによ

る開発手法を取り入れており、適切なクラス設計によりユーザーとなる研究者からの追加機能の要望にも柔軟に対応できるような拡張性の高い実装を実現している。

ivy は、ユーザーから与えられた RNA-seq/DNA-seq のアラインメントファイルと参照ゲノム配列を解析のパラメータを引数として受け取り、動作する。基本的な動作として、受け取った引数から参照ゲノム配列の一塩基ごとにアラインメント結果を解析する。一塩基ごとのアラインメント情報の取得は、ストリーミングで処理され、各種のフィルタリング処理が行われる。設定されたフィルタリングを通過した最終的な候補サイトは、VCF ファイルへと書き出され、ivy による計算は終了する。edit_bench は、検出された RNA 編集サイトの精度検証を行うためのベンチマークツールとして開発された。精度検証には、再現率、適合率および F 値と呼ばれる指標を用いた。

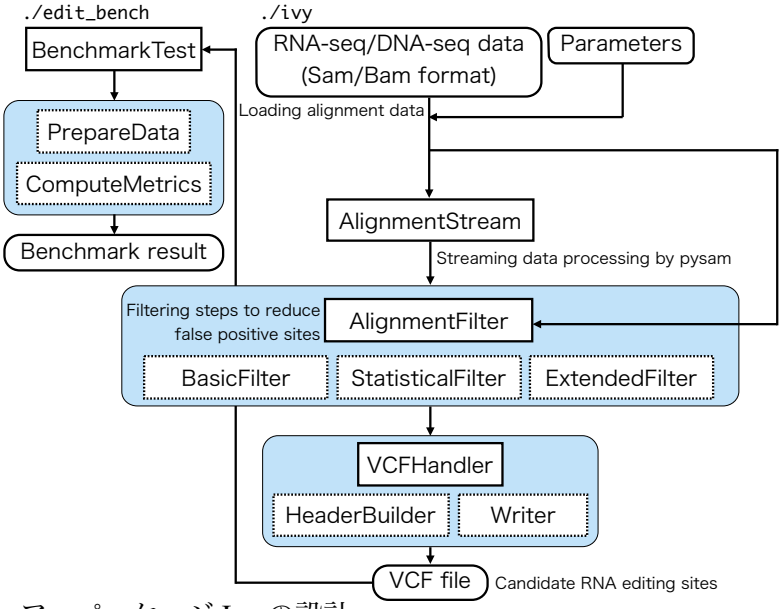

図 1.1: ソフトウェア・パッケージ Ivy の設計

設計された Ivy の全体像を示す。ここで示した全体像は、実装を簡略化して示している。矢印は、入力として受け取った RNA-seq/DNA-seq データと解析パラメータを受け取り、最終的に RNA 編集サイトが検出されるまでの流れを示す。

1.2.2 入出力の形式

Ivy の実行時に入力された BAM (Binary sequence alignment/map format) ファイルは、Pysam ライブラリを使用して、リファレンスゲノムへのアラインメント結果の取得に用いている。Pysam は、C 言語で書かれた BAM のパーサーライブラリ (Samtools C API) のラッパーであり、内部では直接 C 言語の API を呼び出しているため高速にアラインメント情報を取得可能であることから ivy に使用した。

ivy によって検出された A-to-I 編集サイトは、VCF v4.1 によって出力される。この VCF フォーマットは、SNP や SNV の検出といった変異解析に標準的に用いられているフォーマットを指し、1000 genomes project など国際プロジェクトでも採用されているデータ形式である。RNA 編集サイトも SNP も本質的にはゲノムのある座標における一塩基置換として表現可能であるから、検出した RNA 編集サイトも VCF 形式で出力することが望ましいと考えた。VCF を出力フォーマット

とする利点として、変異解析のために開発された他のミドルウェアを組み合わせた更なる解析が可能となる点である。SNP 解析では検出した SNP それぞれの遺伝子名やアミノ酸置換の有無などを Annovar (Wang *et al.*, 2010) とったソフトウェアを用いてアノテーションする場合が多い。ivy で出力された結果もまた VCF であるから、Annovar など他のツールと連携させた下流解析を容易に行うことができるという利点を持つ。REDItools は、独自のタブ区切りテキストを出力とする。

1.3 本手法の性能評価

1.3.1 性能評価に用いた RNA-seq データ

本研究によって開発された RNA 編集サイトの検出ソフトウェア ivy の性能評価を行った。評価軸は、ソフトウェアとしての性能と検出手法としての精度を複数の観点から評価した。

性能評価をするにあたり、RNA 編集サイトの検出を目的とした先行研究でシーケンスされた RNA-seq データの再解析を行った。ヒトを対象とした性能評価には、SRA (Sequence Read Archive, `www.ncbi.nlm.nih.gov/sra`) において公開されている Bahn *et al.* (2012) のシーケンスデータを用いた。Bahn *et al.* (2012) の手法は、高い精度を示した研究事例であると同時に、siRNA による *Adar* のノックダウン株を同時にシーケンスしているため、実装した*–adar_nul]* オプションの効果も検証できると考えた。加えて、アラインメントデータを同時に公開していることから、マッピング処理におけるデータの再現性の問題を回避することが出来ることも理由の一つである。以下に取得したデータの内訳を示す。

表 1.1: 検証に用いたヒトの RNA-seq サンプルの内訳

Sample	GSM ID	Cell line	Tissue	Replicate
Adar_control	GSM693747	U87MG	Glioblastoma	2
Adar_null	GSM693746	U87MG	Glioblastoma	2

Bahn *et al.* (2012) によってシーケンスされたヒトのグリア芽細胞腫由来のセルライン U87MG の RNA-seq (Adar_control) と siRNA によるノックダウン株の RNA-seq データ (Adar_null) の情報を示す。二種類のサンプルは、どちらも 2 回の実験を行った生物学レプリケートがあり、合計のサンプル数は 4 つとなっている。

ivy の実行には、参照ゲノム配列や遺伝子アノテーションを必要とする。これらのデータは、UCSC の提供する参照ゲノム配列や遺伝子のアノテーションを `ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com/Homo_sapiens/UCSC/hg19/Homo_sapiens_UCSC_hg19.tar.gz` より取得し、アノテーションは `genes.gtf`、参照ゲノム配列は `genome.fa` をそれぞれ用いることで解析を行った。

1.3.2 性能比較に用いたソフトウェア

ソフトウェアの検出精度や実行時間などに関する性能評価には、ivy v.0.0.1-dev の他に、REDItools v0.1.3 に同梱されている REDItoolDenovo.py および Samtools v.0.1.19 を用いた。REDItools は RNA-seq データを入力とした RNA 編集サイトの検出ソフトウェア、samtools は SNP や SNV を検出するためのソフトウェアである。samtools は厳密には RNA 編集サイトの検出を目的としたソフトウェアではないが、先行研究で用いられた例があるため比較対象として適当だと考えた。それぞれ 3 つのソフトウェアは、基本的にデフォルト値での実行を行った。以下に実行時の詳細を記した。

ivy は、以下のように実行した。

```
ivy -f hg19.fa -r U87MG_1_chr1.bam -G gene.gtf --one-based
```

実行時のオプションは、-r が RNA-seq のアラインメントデータ、-G は遺伝子のアノテーション、--one-based はゲノム座標の表現を 1-origin にするためである。

REDItools は、以下のように実行した。

```
REDItools-1.0.3/REDItoolsDenove.py -i U87MG_1_chr1.bam -f human_hg19.fa \
-l -e -E -d -p -u -W
```

実行時に用いた各種のフィルタリングパラメータは、-l で編集サイトのみを出力、-e で複数座標にマップされたリードの排除、-E で複数種の塩基置換が見られた箇所を排除、-d で PCR 重複したリードの排除、-p で適切なペアエンドリードのみを使用、-u ではマッピングクオリティの考慮、-W でホモポリマー領域のフィルターをそれぞれ意味する。このパラメータは、REDItools の論文 Picardi and Pesole (2013) において使用されているパラメータを参考にした。

Samtools は、以下のように実行した。samtools は mpielup とよばれるサブコマンドと bcftools の view と呼ばれるサブコマンドを組み合わせることで使用する。mpileup は、bam ファイルを pileup 形式に変換し、bcftools が変異箇所を検出する。

```
samtools mpileup -ugDSI -f human_hg.fa U87MG_1_chr1.bam | bcftools \
view -vcgIN
```

samtools mpileup はそれぞれ、-ugD は解析結果の出力に関するオプション、-S は strand bias の計算、-I は INDEL を検出しない、-f はリファレンスゲノムを意味する。bcftools view は、-v で変異箇所のみを出力、-cg により変異を検出、-I は INDEL のスキップ、-N は参照ゲノムが N の場合にスキップするオプションである。

1.4 検出精度の検証

表 1.1 における Adar_control の RNA-seq データに対して、Bahn *et al.* (2012) で報告されている 12,800 個の A-to-I 編集サイトについての再現性を比較することにより、検出精度を評価した。図 1.2 には、適合率による精度検証を行った結果を示す。Samtools と REDItools との比較において、Ivy は低い適合率が示された。また、samtools は、20 番から 22 番染色体などにおいては 3 つのソフトウェアの中でも比較的高い適合率が示された。

検出精度を再現率によって評価した結果を図 1.3 に示す。適合率を各染色体ごとに算出したところ、本研究によって開発した ivy は 18 番染色体を除いた全ての染色体において、他の二つのソフトウェアと比較して高い再現率を示した。ivy の次に高い再現率を示した手法は samtools であり、REDItools は全ての染色体を通して、低い再現率を示した。

転写物がセンス鎖とアンチセンス鎖のどちらから発現しているのかを考慮することは非常に重要である。サンプルに用いた RNA-seq データは、PolyA セレクションをした通常のライブラリ調整をしているため、転写物の方向が不明である。例えば、T-to-C ミスマッチが観察されたサイトがアンチセンス鎖から発現している場合、A-to-G 編集サイトとして検出する必要がある。ivy では既存の遺伝子モデルのアノテーション情報を利用することで、strand specific RNA-seq データでない入力の場合にも、適切なミスマッチパタンの分類を行うことが可能である。

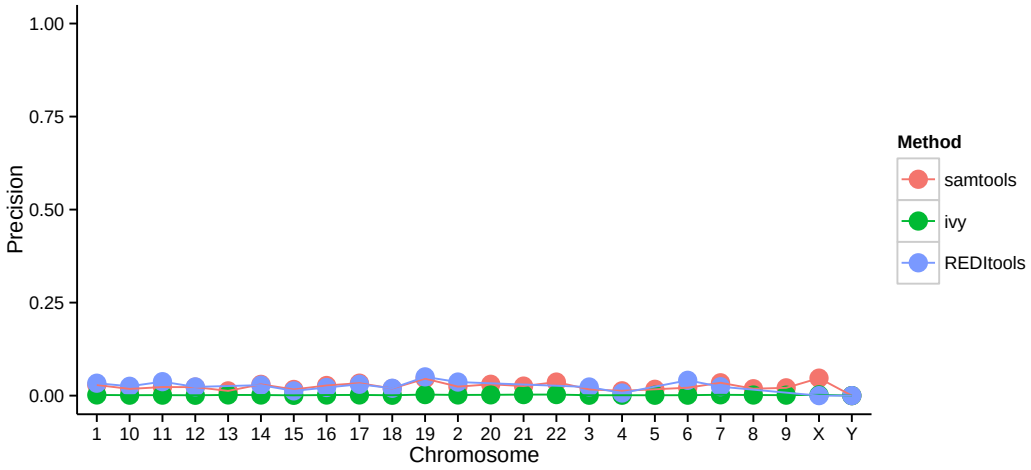

図 1.2: 染色体ごとの適合率

縦軸に適合率、横軸に染色体をそれぞれの手法ごとに示す。Y 染色体においては、精度比較に用いた 3 つのソフトウェア全てにおいて検出された編集サイトは 0 だった。

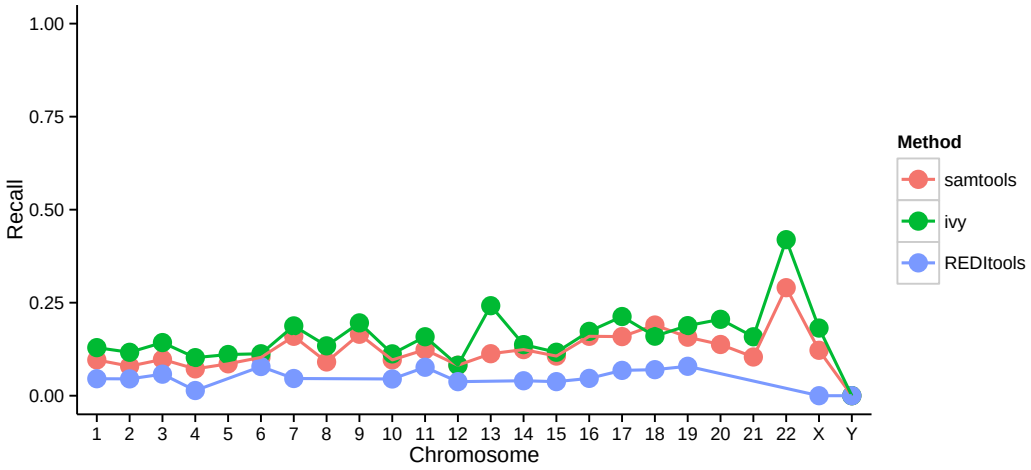

図 1.3: 染色体ごとの再現率

縦軸に再現率、横軸に染色体をそれぞれ比較した 3 つの手法ごとに色分けして示した。

1.5 議論

本研究は、RNA-seq データを用いた高精度かつ高速な RNA 編集サイトの検出手法の開発を目的としたソフトウェア・パッケージ Ivy の設計と実装を行い、オープンソースのフリーウェアとして公開した。

パフォーマンスについて。しかしながら、解析に用いられる計算機には数 GB から数百 GB 程度のメモリが搭載される場合が多く、ここで見られたメモリ使用量の相違は誤差程度だと考えられる。

一つ目は、高精度な検出のために解析が複雑化し、簡便かつ高速な解析が困難となっているこ

とである。本研究により開発された ivy は、既存の REDItools よりも精度よく既知の RNA 編集サイトを検出することが示された。

二つ目の問題は、新規の検出手法によって編集サイトを検出した場合に、検出精度の検証方法がばらつき、手法やパラメータの影響についての比較検討が困難だということである。本研究は、この問題に対して、edit_bench と呼ばれる検出精度の検証を可能にするコマンドラインツールを実装することによって解決を試みた。第 2 章では、検出手法の精度比較を主題とし、情報検索の分野で利用されてきた適合率や再現率の導入による解決方法の提案を試みたものであった。

将来的には、生物種、セルラインとその実験条件の組み合わせにおいて、最も再現率および適合率の高い解析パラメータを網羅的に探索し、*a priori* な情報をデータベースとして公開することを視野に入れている。これにより、ivy を用いる研究者は、対象となるサンプルや実験条件に適した ivyzzz のパラメータやフィルタリング手法をある程度の目星をつけることができるため、有用性が高いと考えられる。

現在、ivy の並列化の実装は、Python の multiprocessing モジュールを利用し、染色体ごとの並列処理に対応している。しかしながら、染色体やコンティグには総塩基長に数倍以上の差があり、現在の実装では長い染色体も一つ以上のスレッドを使用できない。将来的には、各スレッドが解析する塩基長を均一化することで、より効率的な計算が可能になると予想される。加えて、主要なクラスを Cython を介した C のコードに書き換えることで、計算時間の短縮化を検討している。

Ivy の開発は、現在はベータ版 (v.0.0.1-beta) のリリースにとどまっており、開発が続行されているプロジェクトである。これまでに議論したようなアラインメントデータへのフィルタリング手法の更なる実装に加えて、多様な RNA-seq データに対して安定した性能を示すことが大きな課題として残されていると言えるだろう。

謝辞

本研究を遂行するにあたり、慶應義塾大学政策・メディア研究科 荒川和晴講師には、開発と実装に関する多くの助言を頂いた。また、所属する G-language グループのメンバーは、進捗ミーティングを通して多くの問題を指摘してもらった。同大学環境情報学部 冨田勝教授には計算資源など恵まれた研究環境を提供して頂いたことを深謝する。

参考文献

Bahn, J. H., Lee, J.-H., Li, G., Greer, C., Peng, G., and Xiao, X. (2012). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res*, **22**(1), 142–50.

Chen, C. and Bundschuh, R. (2012). Systematic investigation of insertional and deletional RNA-DNA differences in the human transcriptome. *BMC Genomics*, **13**, 616.

Danecek, P., Nellåker, C., McIntyre, R. E., Buendia-Buendia, J. E., Bumpstead, S., Ponting, C. P., Flint, J., Durbin, R., Keane, T. M., and Adams, D. J. (2012). High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biol*, **13**(4), 26.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–9.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, **20**(9), 1297–303.

Peng, Z., Cheng, Y., Tan, B. C.-M., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X., Guo, J., Dong, Z., Liang, Y., Bao, L., and Wang, J. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nat Biotechnol*, **30**(3), 253–60.

Picardi, E. and Pesole, G. (2013). REDItools: high-throughput RNA editing detection made easy. *Bioinformatics*, **29**(14), 1813–4.

Sanjana, N. E., Levanon, E. Y., Hueske, E. A., Ambrose, J. M., and Li, J. B. (2012). Activity-dependent A-to-I RNA editing in rat cortical neurons. *Genetics*, **192**(1), 281–7.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, **38**(16), e164.

Yu, X. and Sun, S. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics*, **14**, 274.