慶應義塾大学学術情報リポジトリ
Keio Associated Repository of Academic resouces

Title	RNA編集サイトの検出ソフトウェアの設計と実装
Sub Title	
Author	石黒, 宗(Ishiguro, So)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2013
Jtitle	生命と情報 No.20 (2013.) ,p.116- 123
JaLC DOI	
Abstract	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やsiRNA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きた。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精度かつ, 高速に動作するRNA編集サイトの検出ソフトウェアの開発を行った。開発したRNA編集サイトの検出ソフトウェアをグリア芽細胞腫由来のRNA-seqデータへ適用したところ, 既存のソフトウェアと比較して同等のメモリ効率で2倍程度高速に動作することが確かめられたほか, 全ての染色体において高い再現率を示す手法であることが明らかとなった。本研究は, 超並列シーケンサーデータを用いたRNA編集サイトの高精度かつ高速な検出手法の開発に貢献することが期待される。
Notes	慶應義塾大学湘南藤沢キャンパス先端生命科学研究会 2013年度学生論文集 卒業論文ダイジェスト
Genre	Technical Report
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO92001004-00000020-0116

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the KeiO Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

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

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やsiRNA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きた。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精度かつ、高速に動作するRNA編集サイトの検出ソフトウェアの開発を行った。開発した RNA 編集サイトの検出ソフトウェアをグリア芽細胞腫由来のRNA-seqデータへ適用したところ、既存のソフトウェアと比較して同等のメモリ効率で2倍程度高速に動作することが確かめられたほか、全ての染色体において高い再現率を示す手法であることが明らかとなった。本研究は、超並列シーケンサーデータを用いた RNA 編集サイトの高精度かつ高速な検出手法の開発に貢献することが期待される。

キーワード: RNA editing, High-throughput sequencing, Bioinformatics

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、編集サイトの検出には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 (YuandSun, 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)。このような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参照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。これらの多くは、ADARなど生物学的な事象を背景にした塩基修飾ではなく、RNA分子の不安定性や複数のマッピングバイアスなどを原因とした技術的なエラーに起因する。

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する。

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する。Ivyは、GNU GPLv3 (GNU General Public License version3)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。

1.2 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: Ivyの設計の概要.

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、VCFv4.1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*etal.*,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。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

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つである。GSMIDは、塩基配列データなどが公開されているNCBI GEOの登録IDを指す。

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

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。

Ivyは、 ivy -f hg19.fa- r U87MG_1_chr1.bam -G gene.gtf --one-based のように実行した。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、 REDItools-1.0.3/REDItoolsDenove.py -i U87MG_1_chr1.bam -f human_hg19.fa -l -e -E -d -p -u -Wのように実行した。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と実行した。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strandbias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。bcftoolsview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は低い適合率を持つことが示された。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を示した。

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、低い再現率を示すことが明らかとなった。

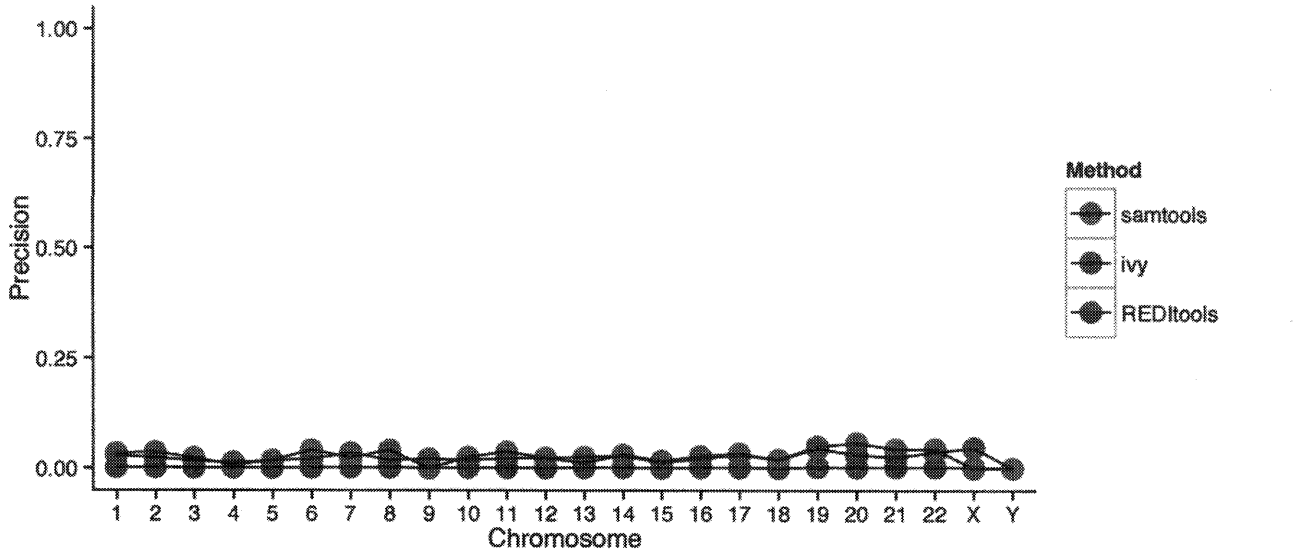

図1.2: 染色体ごとの適合率の比較結果

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。赤がSAMtools、青がivy、緑がSAMtoolsによる適合率をそれぞれ示す。

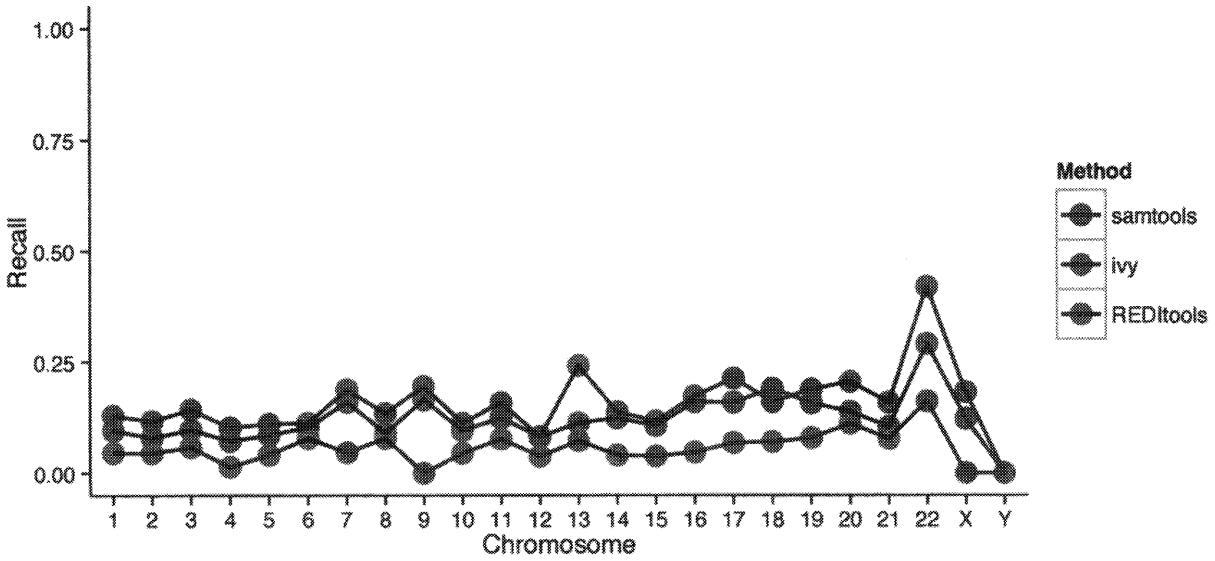

図1.3: 染色体ごとの再現率の比較結果

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。Ivyは、RNA編集サイトの検出と検出結果から精度検証を行うことのできるソフトウェア・パッケージである。

開発したivyを他のRNA編集サイトおよび変異解析のソフトウェアとの精度比較を行った結果 (図1.2および図1.3)、ivyは適合率が低い一方で3つのソフトウェアの中で最も高い再現率を示した。適合率が低かった原因については、RNA編集サイトとして検出した箇所がivyの場合は他のソフトウェアと比較して数倍程度多かったことが挙げられる。適合率は、検出した全サイトに正解が含まれる割合として計算される。このため、検出数が高くなるにつれて適合率は低くなる傾向にある。対して、ivyは一部の染色体を除いて高い再現率を示した。高い再現率はすなわちBahn *et al.*(2012)による結果を最もよく再現した手法であることを意味している。この高い再現率を示した原因として、ivyは他の2つのソフトウェアに対して、遺伝子のアノテーションを利用した転写物の方向性を考慮した計算を可能にした点が挙げられる。ADARによるA-to-G編集は、センス鎖の場合はA-to-G変異であるが、アンチセンス鎖の転写物に入った場合にはT-to-C変異として検出される。本研究で、精度検証に用いた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、転写物の方向は不明である。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を可能にしたことが、本手法の高い再現率に貢献していると考えられた。

開発したソフトウェア・パッケージには、精度検証を行うツールedit_benchが同梱されている。edit_benchは、新規にRNA編集サイトを検出した場合に、検出精度を比較可能な指標で評価することを目的として開発された。このツールは、ivyや他の研究によって同定されたRNA編集サイトの検出精度を簡便に測定するできることから、異なる検出手法やパラメータの統一的な比較を可能にしたと考えられる。

本研究により開発されたRNA編集サイトの検出ソフトウェアivyは、今後より再現率および適合率を向上させるための実装が求められる。特に、既存のソフトウェアよりも適合率が低いことは課題である。適合率を向上させるためには、現在は未実装であるスプライスサイト周辺のフィルタリングや、dbSNPなどのデータを用いたゲノムの変異箇所のフィルタリング、BLASTやBLATを用いた編集サイト周辺のリアラインメントが必要だと考えている。これらのフィルタリングはより厳格なフィルタリングを可能するため、検出サイトは減少することが予想されるが、同時に適合率が上昇することが期待される。

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、現在の実装では染色体は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な実装に変更する予定である。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-dev)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再現率および適合率を示すことが今後の開発に残された重要な課題だと考える。

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。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.