2013年度 卒業論文

超並列シーケンサーデータを用いたRNA編集サイトの検出手法の開発とその情報学的解析

Development of detection method for RNA editing sites based on high-throughput sequencing data

慶應義塾大学 環境情報学部

石黒 宗

2013 年度 卒業論文

超並列シーケンサーデータを用いた RNA 編集サイトの検出手法の開発とその情報学的解析

Development of detection method for RNA editing sites based on high-throughput sequencing data

慶應義塾大学 環境情報学部

石黒 宗

要旨

RNA 編集とは、DNA が RNA へ転写される段階で起こる転写後修飾の一種である。修飾を受けた RNA はゲノムと異なる遺伝情報を持ち、タンパク質機能の多様化や、他の非翻訳 RNA との相互作用を介した遺伝子の発現制御へも関与する。ここ数年、超並列シーケンサーと呼ばれる高出力な塩基配列決定技術が普及し、ヒトやマウスなど高等真核生物に発現する RNA は、高頻度で編集を受けていることが明らかとなってきた。しかしながら、超並列シーケンサーデータを用いた RNA 編集サイトの検出手法は、ソフトウェアとしての実装が現在では一つしかなく、その機能は解析に十分であるとは言い難い。そこで本研究は、再現率と適合率と呼ばれる指標を導入することにより、既存の RNA 編集サイトの検出手法の比較を可能にし、高精度な検出手法と擬陽性を減少させるフィルタリング手法および実験デザインに関する議論を得た。その知見をもとに高精度かつ高速な RNA 編集サイトの検出ソフトウェア Ivy 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, Bioinformatics, High-throughput sequencing

2014 年 1 月 20 日

目 次

第1章　序論

1.1　はじめに

1953 年、James Watson と Francis Crick は、遺伝情報の実体は 4 種類のデオキリボ核酸が相補的に水素結合した二重らせん構造であること鮮やかに示した (Watson *et al.*, 1953)。5 年後の 1958 年、F. Click は"Ideas for protein synthesis" (図 1.1) において、遺伝情報は DNA-RNA-タンパクへと一方向性を持って伝達されるという着想を得ており、1970 年には分子生物学におけるセントラルドグマ (中心原理) を提唱した (Crick *et al.*, 1970)。1960 年代から 70 年代には、Marshall Nirenberg や Sydney Brenner など多くの研究者により、コドンの発見と遺伝暗号が解読されると、塩基配列がアミノ酸を表現する基本原則が明らかとなった。黎明期における分子生物学は、物理学者が多く参入した経緯も相まって、分子のことばで生命の持つ普遍的性質の理解を標榜する学問として誕生し、熾烈な競争の下に急速な進展を遂げた。

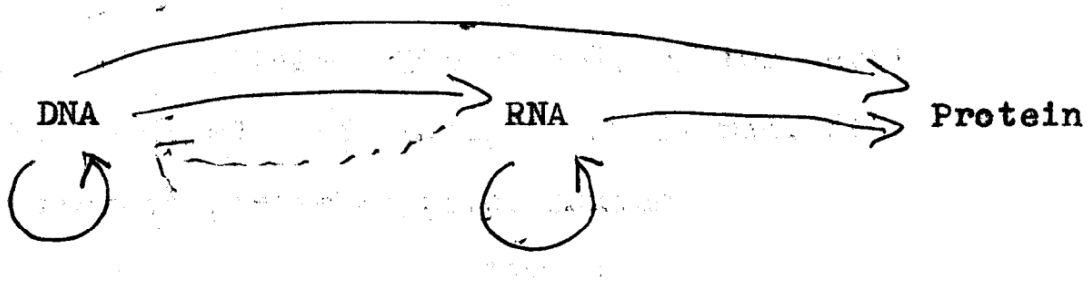

図 1.1: *"Ideas on Protein Synthesis"*に描かれたセントラルドグマのアイデア。矢印の方向は、遺伝情報の流れを表している。破線は RNA から DNA へ向いており逆転写酵素を予見させるが、クリックが 1958 年当時に想定していたかは定かでない。

分子生物学は生命の普遍性を明らかにする過程で、内在的に多様性を生み出す数々の機序をも明らかにしてきた。1970 年、Howard Temin および David Baltimore は一本鎖 RNA を鋳型として、相補的な DNA を合成する逆転写酵素を発見した (Baltimore, 1970; Temin and Mizutani, 1970)。この発見は、セントラルドグマから逸脱し、RNA から DNA へも遺伝情報は伝達しうることを示した画期的な発見であった。その後も、真核生物における遺伝子はスプライシング機構によってイントロンが切り出され複数のエクソンから構成されること (Chow *et al.*, 1977)、転写直後の新生 RNA は 3' 末端にポリアデニル化 (Edmonds and Abrams, 1960)、5' 末端にはキャップ構造が付加されて成熟すること (Wei *et al.*, 1975)、線虫からはコドンが異なるアミノ酸をコードする例外が発見されるなど (Hamashima *et al.*, 2012)、セントラルドグマの概念もまた時代とともに拡張されてきた。このことは、本質的には、生命が逸脱や例外を含む多様性を担保する存在であることに他

ならない。生命の持つ普遍的な性質の理解を目指した分子生物学において、多様性という個別的な生命現象をその対象することが逆説的には、生命の理解へと接続される重要なアプローチの一つだと私は考えている。

真核生物の多様で複雑なシステムは、セントラルドグマからの逸脱の他に塩基やタンパク質への修飾が生命活動に不可欠であることは、現在の分子生物学においては前提となっている。細胞の内外の情報伝達には、キナーゼによるタンパク質のカスケード的なリン酸化反応が用いられ、リン酸化された分子が核内に移行し転写因子の活性を調節することで遺伝子発現の状態と量が制御される (**?**)。ヒストンは、アセチル基やメチル基による化学修飾を受け、染色体構造の変化を伴った遺伝子発現の調節や細胞の分化状態が決定されるなど (**?**)、修飾の持つ生体内機序の制御例は枚挙にいとまがない。真核生物は、有限個の遺伝子にその種数を規定されながらも、転写および翻訳機構への修飾が多様性の生成を駆動する原動力であると言えるだろう。

RNA 編集 (RNA editing) は、転写後修飾 (Post-transcriptional modification) の一種である。RNA 編集は、DNA にコードされた遺伝情報が RNA ポリメラーゼにより転写された直後、RNA が別の塩基に修飾される現象を指す (Wulff and Nishikura, 2010)。RNA 編集を受けた転写物は、元来の遺伝情報とは異なった配列情報を持つこととなり、こういった意図的な修飾は、DNA から正確なコピーとしての RNA を合成するという転写機構からの逸脱と解釈することができる。興味深いことにヒトやマウス、ショウジョウバエといった高等真核生物においては、特定の遺伝子領域内へ RNA 編集を起こし、タンパク活性の変化や機能の調節に積極的に利用されているとの報告が蓄積している (Pullirsch and Jantsch, 2010)。また、ここ数年の超並列シーケンサー (High-throughput sequencing) によるトランスクリプトームの網羅的かつ定量的な計測は、非翻訳 RNA や遺伝子間領域などこれまで研究対象とされなかった転写領域における RNA 編集に光をあてており、非翻訳領域における RNA 編集の生物学的な意義にも、今日大きな注目が集まっている (Nishikura, 2006, 2010)。しかしながら、超並列シーケンサーを用いた情報学的な RNA 編集サイトの検出や解析は、ここ数年に立ち上がった新しい領域であることから、一度の実験で得られる膨大なシーケンスデータから精度よく編集サイトを同定する手法とその機能解析に、確立された手法は未だ登場していない。

本論文は、大きく 4 つの内容を取り扱う。第一章では、既往の研究により明らかにされてきた RNA 編集機構の多様な生理学的な機構を概観するほか、超並列シーケンサーを用いた編集サイトの検出手法の発展について概説する。第二章では、既存の RNA 編集検出手法について性能評価を行い、高精度な検出に寄与する検出手法とそのパラメータについて議論する。第三章では、精度よく高速に RNA 編集サイトを検出するソフトウェア・パッケージの開発を行った。第四章では、ヨコヅナクマムシのシーケンスデータから編集サイトを検出し、その情報学的解析を行った。本研究は、超並列シーケンサーから得られる大量のシーケンスデータを用いた RNA 編集サイトの簡便な検出を可能にし、真核生物における転写後修飾とその制御機能の更なる解明に貢献することを期待するものである。

1.2 真核生物における RNA 編集と ADAR の役割

1.2.1 転写後修飾機構としての RNA 編集

RNA 編集は転写物への一塩基修飾を指し、鞭毛虫のミトコンドリアから初めて発見された (Benne *et al.*, 1986)。真核生物では、ADAR (Adenosine deaminase acting on RNA) によるアデニン (A) からイノシン (I) へ修飾される A-to-I 編集、APOBEC (Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) によるシトシン (C) からウラシル (U) への C-to-U 編集がこれまで報告されている (Benne *et al.*, 1986)。APOBEC のターゲットとなる遺伝子は非常に少なく、ヒトでは APOB と NF1 遺伝子のみが知られている (Cappione *et al.*, 1997)。植物においては主にシロイヌナズナのミトコンドリアや色素体において、PPR (pentatricopeptide repeat) タンパクファミリーによる C-to-U 編集が知られる (Meng *et al.*, 2010; Shikanai, 2006)。ヒトやマウス、ショウジョウバエにおいては、高頻度に A-to-I への編集が知られている他、マウスにおいてはヒトに対して C-to-U 編集が高い頻度が見られることが明らかになっている (Danecek *et al.*, 2012; Gu *et al.*, 2012)。

ADAR の他に tRNA への A-to-I 編集を触媒するタンパクとして、Adenosine deaminase acting on tRNA (ADAT) と呼ばれるタンパクファミリーが同定されている (Maas *et al.*, 1999)。ADAT は tRNA のアンチコドンあるいはその近傍の塩基へ特異的に A-to-I 編集を起こす。ADAT はヒトから酵母まで真核生物において広く保存され、大腸菌においてもオーソログ遺伝子 tRNA adenosine deaminase A (TadA) が同定されている (Wolf *et al.*, 2002)。このことから、アデノシンからイノシンへの修飾をするタイプは、原核生物から真核生物まで広く存在し、ADAR と ADAT はともに deaminase ドメインを有することから、TadA から進化した可能性が示唆されている (Gerber and Keller, 2001)。

高等真核生物における A-to-I 編集は、アミノ酸置換を伴いタンパク質機能の多様化に寄与する他に、遺伝子間領域 (Intergenic region) や非翻訳領域 (Untranslated region, UTR)、Alu 領域といったレトロトランスポゾンにおける編集が高頻度で起きていることも明らかになってきた (Bazak *et al.*, 2013; Ramaswami *et al.*, 2012)。大規模な超並列シーケンサーデータを用いた情報学的解析は、こういった非翻訳 RNA への編集を高解像度かつ定量的な解析を可能にする強力なアプリケーションである。本章では、RNA 編集研究に関する最新の研究成果を含めて分野を概観すると同時に、大規模なシーケンシングデータを使った情報学的な A-to-I 編集サイトの検出手法についても最新の知見を含めて概説する。

1.2.2 ADAR による RNA 編集とその作用機序

ADAR は、二本鎖 RNA 結合タンパクの一種として知られ、二本鎖 RNA (Double-stranded RNA, dsRNA) と選択的に結合し、アデノシンからイノシンへの加水分解的な脱アミノ化反応 (Deaminase) を触媒する酵素である (Bass, 2002; Keegan *et al.*, 2004; Valente and Nishikura, 2005)。イノシンへと置換された修飾塩基は、転写機構においてグアノシンとして認識される。ADAR による修飾を

A-to-I 編集と呼び、修飾後のグアノシンに着目した場合は、A-to-G 編集とも表記するが本質的には同一である。

A-to-I 編集は、A から G への一塩基置換であるため、非同義置換によるアミノ酸の変異や終止コドンが置換され転写物が伸長するリードスルー、イントロン-エクソン境界への編集によるスプライスサイトの新生および欠失などの機能がこれまでに報告されている (Flomen *et al.*, 2004; Fukui and Itoh, 2010; Meng *et al.*, 2010)。図 1.2 に A-to-I 編集の脱アミノ化による塩基修飾反応の模式図を示す。

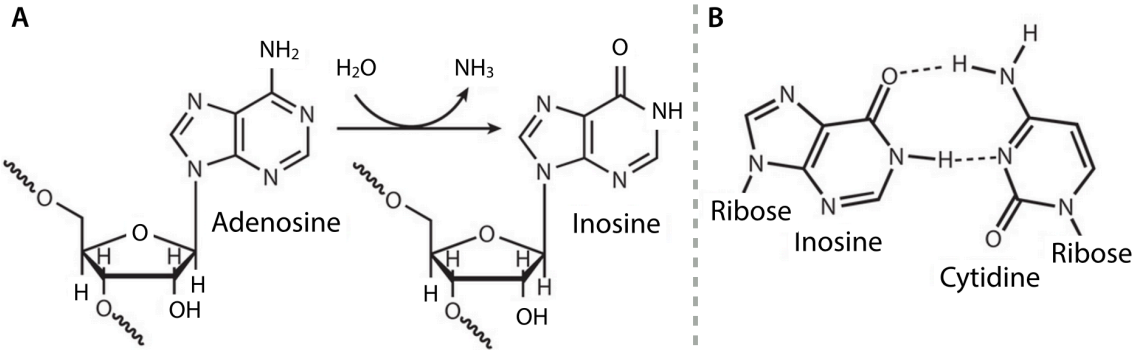

図 1.2: ADAR の生化学的な作用機序 (Nishikura (2010) より改変)

A: ADAR によるアデノシンからイノシンへの化学修飾が触媒される様子を示す。アデノシンは ADAR による脱アミノ化反応によってイノシンへの一塩基修飾を受ける。イノシンへの修飾は同時に標的となった二本鎖 RNA の二次構造を変化させる。**B**: 編集を受けたイノシンは、シチジンと塩基対を形成する。

図 1.3 に ADAR の持つ機能ドメイン構造を示す。ADAR はアデノシンからイノシンへの塩基修飾を触媒する deaminase ドメインと二本鎖 RNA に結合する dsRBD (Double-stranded RNA binding domain) の２つの機能ドメインをヒト、マウス、線虫、ショウジョウバエは共通して有する (Jin *et al.*, 2009)。dsRBD は 65 残基程度の長さの中に α-β-β-β-α という特徴的なドメイン構造を持ち、直接的に二本鎖 RNA と接触するため A-to-I 編集に必須の機能ドメインの一つである (Barraud and Allain, 2012; Cho *et al.*, 2003; Lai *et al.*, 1995)。ヒトの ADAR1 においては、Z-DNA-binding ドメインを 1 つから 2 つ有しているが、これは ADAR のターゲットとなる転写物、特に siRNA など短鎖の二本鎖 RNA との結合における親和性を高めるためだと考えられている (Qi *et al.*, 2012)。また、ヒトの ADAR3 のみに特徴的にアルギニンリッチな一本鎖 RNA 結合ドメインの R-domain を持つが、この生物学的な意義については明らかにされていない (Chen *et al.*, 2000)。ヒトの 2 つの ADAR1、L 型と S 型はスプライシングアイソフォームとして知られ、それぞれ異なるプロモーターから転写されることが分かっている (Kawakubo and Samuel, 2000)。*in vivo* では、ADAR1 および ADAR2 はホモダイマーの形成が脱アミノ化反応には必須であることが報告されている (Gallo *et al.*, 2003)。

ADAR1 および ADAR2 は、真核生物における昆虫やイカ、脊椎動物など進化的に広く分布し、高い配列保存性を示す。酵母や原生生物、植物においては発見されていない。ヒトにおいては、

これまでに ADAR1、ADAR2、ADAR3 の 3 種類が同定されている。このうち、ADAR1 および ADAR2 は酵素活性が確かめられているが、ADAR3 に関しては機能ドメインは保存されているものの、脱アミノ化活性については不明である (Melcher *et al.*, 1996; Nishikura, 2006)。また、ADAR3 は、脊椎動物の ADAR2 の遺伝子重複により生じたとも考えられている (Jin *et al.*, 2009)。

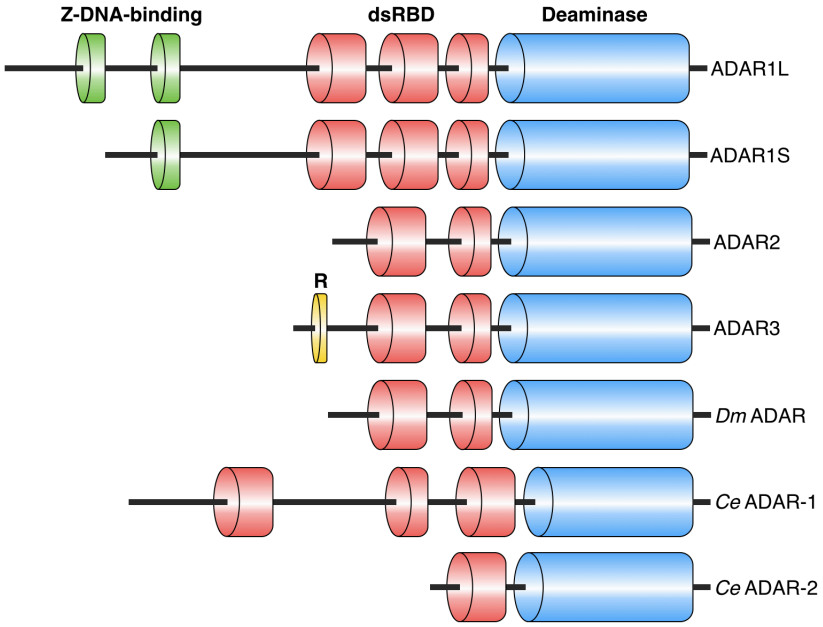

図 1.3: ADAR タンパクのドメイン構造

ヒト、線虫、ショウジョウバエにおける ADAR のスプライシングバリアントを含めたドメイン構造を示す。左から順に、Z-binding (緑)、Double stranded RNA binding (赤)、Deaminase (青)、Arginine-rich (黄) の通り、それぞれ機能ドメイン構造ごとに色分けした。上から ADAR1L、ADAR1S、ADAR2、ADAR3 はヒトから同定された ADAR のバリアント、*Dm* ADAR はショウジョウバエ、*Ce* ADAR-1、*Ce* ADAR-2 は線虫における ADAR である。

1.2.3 二本鎖 RNA の二次構造と修飾塩基の選択性

ADAR による A-to-I 編集は、イノシンへ修飾される位置に高い選択性のある場合と、選択性が低く RNA へランダムに置換する 2 つのタイプに分類される (Nishikura, 2010)。位置選択性の高い編集は、20 塩基程度の短い二本鎖 RNA を形成し ADAR のターゲットとなるのに対し、ランダムに編集される非選択的な編集サイトは、二本鎖 RNA が 100 塩基以上であることが多く、非選択的な編集は、最大でも 50%程度がイノシンへと修飾される傾向が見られることが報告されている。これまでの研究から、特定のアデノシンが選択的に編集するためには、二本鎖 RNA の形成する特異的に二次構造が重要であることが明らかとなっている (Lehmann and Bass, 1999)。

古くから A-to-I 編集の研究対象となっているグルタミン酸受容体 (Guanosine receptor-2, GluR2) やセロトニン受容体 (Serotonin receptor-2C, 5-HTR) は、編集の入る位置に高い特異性を示し、特定

のアデノシンのみが選択的にグアノシンへと置換される。このような位置特異的な A-to-I 編集においては、隣接するエクソン-イントロン境界における相補的な配列、ECS (Editing-site complamentary sequence) およびその二次構造の形成が不可欠であることが知られている (Higuchi *et al.*, 1993)。

GluR2 においては、編集サイト毎に ADAR1 または ADAR2 のどちらか一方に優先的に編集されることが知られており、位置毎における ADAR の選択性は、ADAR の持つ dsRBD の数とドメイン間の配列長の相違が ADAR と二本鎖 RNA との相互作用を変化させることに起因するとの報告がある (Nishikura, 2006)。ヒト B 細胞において、siRNA による *Adar1* および *Adar2* のノックダウンを別々に行った解析では、同定された A-to-I 編集のうち、20%程度が *Adar1* および *Adar2* の共通したターゲットとなっていることが報告されている (Wang *et al.*, 2013)。また、*Adar1* のみをサイレンシングした場合、A-to-I 編集サイトの総数は 1/3 程度に減少することから、ヒト B 細胞においては *Adar1* による A-to-I 編集が優勢的であると考えられる。図 1.4 に二本鎖 RNA に結合する ADAR とそのサイトを模式的に示した。

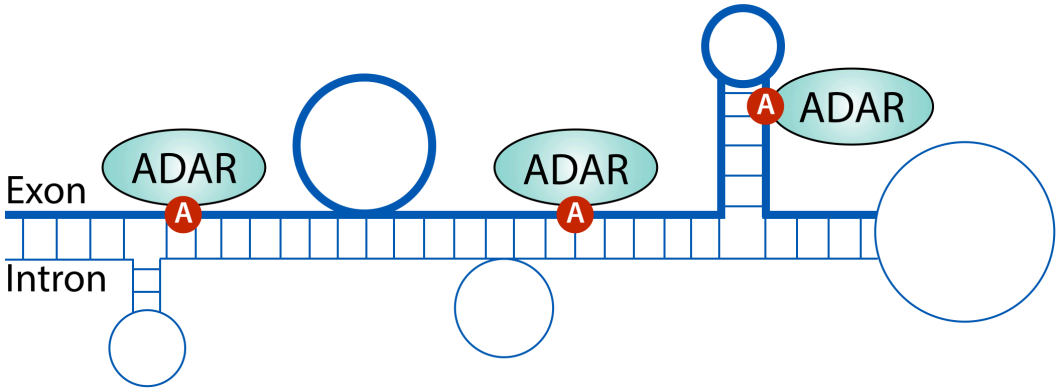

図 1.4: ADAR による A-to-I 編集の模式図

ADAR は二次構造を形成した二本鎖 RNA へ特異的に結合し、アデノシンからイノシンへの塩基置換を触媒する。多くの転写物は複数の編集サイトを持つことが知られる。エクソン領域 (太線) に隣接するイントロン領域 (細線) の間に二次構造が形成され、3 箇所に編集が起きている様子を表す。

1.2.4 *Adar* 遺伝子の欠損による生理学的な影響

生体内における *Adar* の発現が生理学的に重要な役割を持つことは、ヒトやマウス、線虫などの変異株や遺伝子サイレンシングを用いた実験により明らかにされてきた。ショウジョウバエにおいては、*Adar1* の欠損により、協調的な運動行動の欠損や加齢依存的な神経変性が報告され (Palladino *et al.*, 2000)、線虫においては *Adar1* および *Adar2* のホモ変異体は、走化性が欠損することが報告されている (Tonkin *et al.*, 2002)。マウスにおける *Adar1* および *Adar2* はともに致死性を示す。*Adar1* の変異体では初期発生段階において赤血球新生における異常によるアポトーシスが原因で致死の表現型を示し (Hartner *et al.*, 2009)、*Adar2* の欠損は、生後 20 日以内にけいれん重積により死に至る (Higuchi *et al.*, 2000)。この原因は、グルタミン酸受容体における編集不全を原因とするこ

とが分かっている。ヒトにおいては、ADAR による編集不全に起因した疾患が報告されており、*Adar1* のホモ変異体は、遺伝性対側性色素異常症を引き起こす他、グルタミン酸受容体における不完全な編集は、筋萎縮性側索硬化症の原因となることが明らかとなっている (Miyamura *et al.*, 2003)。また、未編集のセトロニン受容体が精神疾患の原因となることも指摘されている (Slotkin and Nishikura, 2013)。

1.2.5 ADAR の発現と細胞内局在

ADAR は核内と細胞質のどちらにも局在することが知られる。ADAR1 については核内と細胞質に局在する ADAR1L および ADAR1S が知られるほか (Patterson and Samuel, 1995)、ADAR2 は核に局在することが知られる (Desterro *et al.*, 2003)。また、ADAR3 は脳特異的に発現していることが知られる (Melcher *et al.*, 1996)。核内で起こる A-to-I 編集は、RNA ポリメラーゼ II による転写と同時、すなわち Co-transcriptional に作用していることがショウジョウバエにおける核内の新生 RNA を用いたトランスクリプトーム解析から報告されている (Rodriguez *et al.*, 2012)。Co-transcriptinoal な A-to-I 編集は、スプライシングの効率に影響を与えているデータも示されており (Laurencikiene *et al.*, 2006)、ADAR1 および ADAR2 の細胞内局在とスプライシング機構との関係性については、今後より詳細が明らかにされるだろう。

ADAR2 は自己編集 (self-editing) と呼ばれる機構が知られており、発現している ADAR2 の mRNA に対して数カ所の A-to-I 編集を起こす (Gan *et al.*, 2006)。結果として、ADAR2-mRNA は 4 番イントロンと 5 番エクソンとの間に 47 塩基が挿入されたフレームシフト変異を引き起こし、新しいスプライスバリアントは編集活性を持たないことが報告されている (Hang *et al.*, 2008)。このことは、ADAR2 が自身の活性を A-to-I 編集により制御するという非常に興味深い現象である。

ヒトの 8 つのセルラインを用いて、核内と細胞質を分離した RNA-seq 解析からは、核内で起こる A-to-I 編集が細胞質で起こるものより数倍以上高頻度である傾向が見られることから、A-to-I 編集の多くは Co-transcriptional に起きていることを示唆していると考えられる。核内と細胞質ともに 70%がセルライン特異的な A-to-I 編集を占めることが報告されている (Chen, 2013)。

1.2.6 RNA 編集の進化的な起源と意義

A-to-I 編集による塩基修飾は、現存する生物が祖先種の遺伝子配列を獲得する復帰突然変異としての役割を持つという仮説が提唱されている (Chen, 2013; Pinto *et al.*, 2014)。Chen (2013) は、ヒト、チンパンジー、アカゲザルのゲノム配列から祖先種のゲノムを推定し、ヒトゲノムと推定した祖先種ゲノムの塩基配列を比較すると、ヒトの 90%以上の A-to-G 編集サイト (=グアニン) は祖先種においてもグアニンであることを示した。このことは、進化の過程でアデニンへの一塩基置換が起こり、再びその箇所を A-to-G 編集を起こすことにより、祖先配列を獲得している可能性が持たれている。このような A-to-G 編集による復帰突然変異のような現象を Chen (2013) は RNA memory として提唱している。しかしながら、現存しない祖先種のゲノム配列を検証することは

難しく、祖先配列との比較結果は見かけ上の現象なのか、RNA memory が何らかの制御を受けた結果であるのかは議論の余地があると考えられる。

1.3 A-to-I 編集の持つ多様な生理学的機能

1.3.1 タンパク機能の調節と多様化

真核生物における RNA 編集の重要な役割として古くから認識されてきたものは、翻訳領域内におけるアミノ酸配列の変化を伴った A-to-I 編集である。AMPA 型グルタミン酸受容体は、GluR2 サブユニットは他のサブユニットに対して膜上のアミノ酸がグルタミン (Q) から陽電荷を有するグリシン (R) へと変化しており、その結果として GluR2 サブユニットを有する AMPA 型グルタミン酸受容体のみ、カルシウムに対する非透過性を示す (Higuchi *et al.*, 1993)。このカルシウム透過性に関する制御は Q/R 調節とも呼ばれ、A-to-I 編集によるグルタミン (CAG, Q) からグリシン (CIG, R) への非同義置換によって達成されている。哺乳類の神経細胞においては、100%に近い割合で Q/R 調節が行われており、RNA 編集率の低下は細胞内へのカルシウムの透過率をさせ、結果として細胞死やヒトの場合は疾患の原因となることが報告されている (Slotkin and Nishikura, 2013)x。

タコの遅延整流性のカリウムイオンチャネル K_V は、14 箇所の編集が起こることが発見されている (Garrett and Rosenthal, 2012)。この編集を受けた $K_V1.1$ チャネルへの 4 箇所の A-to-I 編集は、イソロイシンからバリンへの非同義置換を引き起こし、ゲートの開閉を加速させることが報告された。寒冷性のタコは、編集を受けることにより、温帯性のタコよりも低い温度での活動を可能にしていることが示唆されている。

1.3.2 二本鎖 RNA の安定性への寄与

ADAR による A-to-I 編集は、二本鎖 RNA が形成する折りたたみ構造を局所的、また大域的に変化させる。RNA 分子における塩基対形成は、I:U および G:C のワトソン・クリック型塩基対で安定的に形成するが、A-to-I 編集により一般的なワトソン・クリック型塩基対ではなく、I:U がペアリングするゆらぎ塩基対 (Wobble base pair) を形成する (Barraud and Allain, 2012)。このゆらぎ塩基対は、二本鎖 RNA を熱力学的に不安定化させる。この逆に、A:C などのゆらき塩基対に対する A-to-I 編集は、二本鎖 RNA の安定化に寄与すると考えられる。このように、A-to-I 編集による塩基対形成の変化は、局所的または大域的に二本鎖 RNA の安定性を変化させると考えられている (Nishikura, 2010)。

1.3.3 スプライシング機構との関連性

真核生物のほぼ全てのエクソン-イントロン境界は、GU-AG 配列と呼ばれる強いコンセンサス配列が保存されており、スプライシング機構においてイントロンとエクソンの境界を決めている

(**?**)。3' スプライスサイトはドナーサイト (donor site)、5' スプライスサイトはアクセプターサイト (acceptor site) と呼ばれ、AU 配列または AG 配列への A-to-I 編集はそれぞれ新生のアクセプターサイト、ドナーサイトを形成する。また、AG 配列への A-to-I 編集は、3' スプライスサイトを欠失することがこれまでに報告されている (Valente and Nishikura, 2005)。また、このようなスプライスサイトの新生や欠失が二本鎖を形成した Alu 配列に起こる場合、Alu 配列の一部がエクソン化 (exonization) し、mRNA に新たなエクソンとして取り込まれる例がヒトの Nuclear prelamin A recognition factor (NARF) タンパクでは知られている (図 1.5) (Sorek *et al.*, 2004)。A-to-I 編集による Alu 配列のエクソン化は、新たな機能を持つタンパク質のバリアントを生み出すため、進化的に重要な役割を担っていると考えられている。

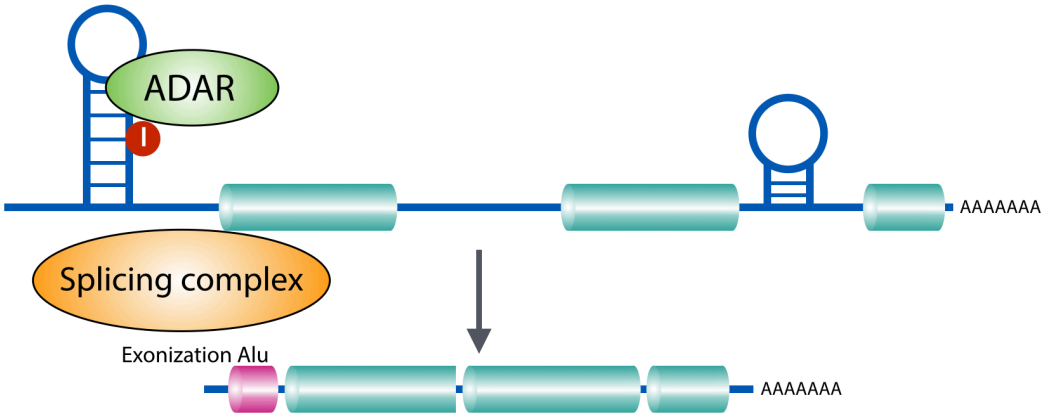

図 1.5: A-to-I 編集による Alu 領域のエクソン化 (Nishikura (2006) より改変)

イノシンは、スプライシング機構においてグアノシンとして認識される。二本鎖 RNA を形成した Alu 配列を標的とした A-to-I 編集は、GU 配列や AG 配列といったスプライスサイトを新生あるいは欠損させる。スプライスサイトの新生と欠損は、Alu 配列の構成性エクソン (Constitutive exon) へのそれぞれ取り込みと排除を起こす。

1.3.4 miRNA への編集と遺伝子発現制御

低分子 RNA (Small RNA) の一種として知られる miRNA は、複数の RNA 結合タンパクと相互作用し、転写された mRNA に対して分解を促進し、遺伝子の翻訳抑制を行う真核生物において重要な低分子 RNA の一つである (Krol *et al.*, 2010)。miRNA は、初めは数百から数千塩基程度の pri-miRNA (primary-miRNA) として転写され Drosha-DGCR8 複合体によりヘアピン構造の pre-miRNA として切り出され、細胞質へ輸送される。細胞質では、pre-miRNA は Dicer-TRBP 複合体による切断を受けた後に、21 塩基程度の二本鎖 RNA となり成熟する。miRNA は、RISC (RNA-induced silencing complex) へと取り込まれ、標的となった遺伝子の 3' 末端の非翻訳領域に結合し、mRNA の分解に作用する (Carthew and Sontheimer, 2009)。

pri-miRNA から miRNA へと成熟する過程では、二本鎖 RNA を形成するため Drosha や Dicer

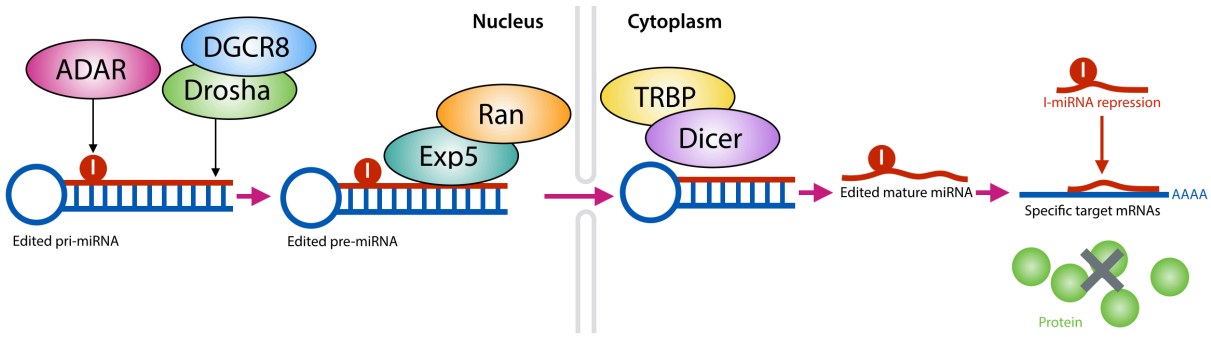

図 1.6: miRNA への A-to-I 編集による遺伝子発現調節 (Nishikura (2006) より改変)

miRNA が核から細胞質へ輸送され、成熟して機能する経路を示す。miRNA は、成熟過程において A-to-I 編集を受ける。Drosha-DGCR2 複合体や TRBP-Dicer 複合体における切断部位の近傍への A-to-I 編集は、miRNA 産生量の増加と減少に影響している。シード配列への編集は、標的となる遺伝子を変化させることが知られる。

による切断と同様に ADAR による A-to-I 編集もその標的となる。ADAR が二量体で機能する場合、ADAR による pri-miRNA への編集は大きく 3 タイプに分類することができる。一つ目は、pri-miRNA の Drosha による切断部位の近傍に編集が入ることで、pri-miRNA は切断されずにイノシン特異的に認識する Tudor-SN (Tudor Staphylococcal Nuclease) による分解を受ける。二つ目は、pri-miRNA が Dicer の切断部位の近傍に編集が入ることで、Exportin-5 (Exp5) を介して細胞質へ輸送されるが、Dicer への切断に抵抗性を示すため、成熟せず miRNA は機能しない。三つ目は、正常な miRNA として成熟するが、厳密にターゲットを識別するシード配列に編集が入ることにより、標的遺伝子が変化する場合である (Slotkin and Nishikura, 2013)。

最近では、ADAR と Dicer が複合体を形成し、miRNA 産生の活性化にも ADAR が関与していることが明らかとなってきた (Ota *et al.*, 2013)。ADAR-Dicer 複合体は、A-to-I 編集能を持たないことも明らかとなっており、miRNA の合成経路と ADAR の多様な関係性が明らかにされている。

1.3.5 siRNA 産生の制御と RNAi 経路とのクロストーク

miRNA の生合成過程における A-to-I 編集と同様に、Dicer のよる切断を受けて成熟する siRNA (small interfering RNA) もまた A-to-I 編集の影響を受けていることが明らかとなっている (Yang *et al.*, 2005)。siRNA は Dicer によって 21 塩基から 24 塩基程度の二本鎖 RNA へと切断され、miRNA と同様に RISC を形成する。RNAi 経路においては、siRNA と相補的な配列を有する遺伝子が転写抑制の標的となるため、Argonaute タンパクなどの複合体として RISC に取り込まれた siRNA は標的遺伝子を認識する重要な役割を持つ (Filipowicz, 2005)。

このように、RNAi 経路において Dicer と ADAR は共通して二本鎖 RNA を標的とするため、siRNA の Dicer の切断部位への A-to-I 編集は RNAi 経路へ拮抗的に作用する。線虫においては、ADAR による siRNA への編集が RNAi 経路を阻害し、遺伝子発現へ影響を及ぼすことが報告されている (Knight and Bass, 2002)。

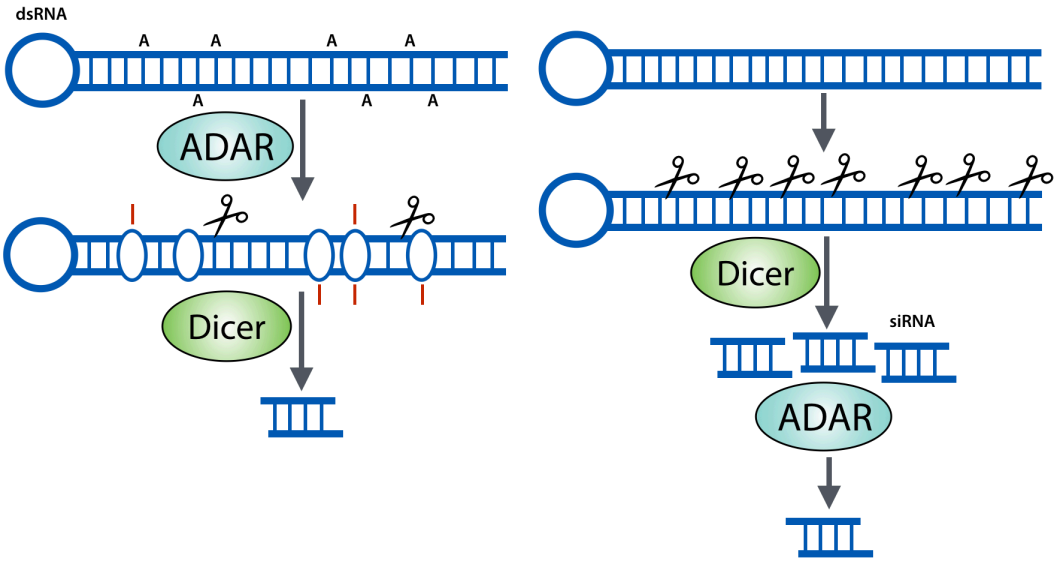

図 1.7: ADAR による siRNA の制御と RNAi 経路への影響

長鎖非翻訳 RNA は、Dicer による切断を受けて内在性 siRNA として機能する。現在考えられている A-to-I 編集による siRNA の制御を左右にそれぞれ示した。左は、A-to-I 編集を受けた長鎖ヘアピン二本鎖 RNA へのが Dicer への切断抵抗性を示すために、内在性 siRNA の産生量が減少する場合である。右は、既に生成された siRNA に対して、細胞質に局在する ADAR1 が強く結合することにより、siRNA の RISC へのローディングを阻害し、siRNA 濃度を減少させる場合である。

ウィルスなど外来性 RNA に由来する siRNA に加えて、マウスの卵母細胞を用いた実験では、SINE などレトロトランスポゾンが形成する長鎖ヘアピン (lhRNA, long-hairpin dsRNA) 二本鎖 RNA が Dicer の標的となって、内在性 siRNA (endogenous siRNA) が生成する経路が報告された (Watanabe *et al.*, 2008)。この内在性 siRNA は、トランスポゾン活性に対して抑制性に機能する siRNA である。*in vitro* の実験結果から、A-to-I 編集を受けた内在性 siRNA は Dicer による切断に対して抵抗性を示すことが分かった (Scadden and Smith, 2001)。A-to-I 編集によって二本鎖 RNA は、A:I から A:U へ塩基対形成が変化する。A:U 塩基対は、Dicer の切断に対して抵抗性を示すため、A-to-I 編集は内在性 siRNA の産生量を減少させる、或いは正常に機能しない内在性 siRNA の産生の制御に A-to-I 編集が関与する可能性を示唆している。加えて、この拮抗的な作用とは真逆に、RNAi 経路へ促進する機能も同時に ADAR1 は有していることが報告された。これは、ADAR と Dicer がホモ二量体の複合体を形成した場合である。

1.4 超並列シーケンサーによる RNA 編集サイトの検出とその手法開発

1.4.1 網羅的な RNA/DNA difference サイトの検出

ここ数年、超並列シーケンサーに代表される配列決定技術の躍進的な発展は、多サンプルのゲノム (DNA-seq) およびトランスクリプトーム (RNA-seq) のシーケンスを可能にしてきた。RNA

編集研究においては、超並列シーケンサーから得られる網羅的で定量性のある RNA-seq データや DNA-seq データを用いることにより、A-to-I 編集サイトを検出する情報学的手法が精力的に開発されてきた (Bahn *et al.*, 2012; Bazak *et al.*, 2013; Chen, 2013; Dillman *et al.*, 2013; Kleinman *et al.*, 2012; Lagarrigue *et al.*, 2013; Li *et al.*, 2011; Park *et al.*, 2012; Peng *et al.*, 2012; Pinto *et al.*, 2014; Ramaswami *et al.*, 2012, 2013; Rodriguez *et al.*, 2012; Sakurai *et al.*, 2014; St Laurent *et al.*, 2013; Zhu *et al.*, 2013)。超並列シーケンサーから測定されるトランスクリプトームやゲノムの配列情報は、数 GB から数百 GB の配列データを扱うことになるため、こういった大規模な編集サイトの検出は本質的に情報学的な解析が必須となっている。図 1.8 に解析の基本的なパイプラインを示した。

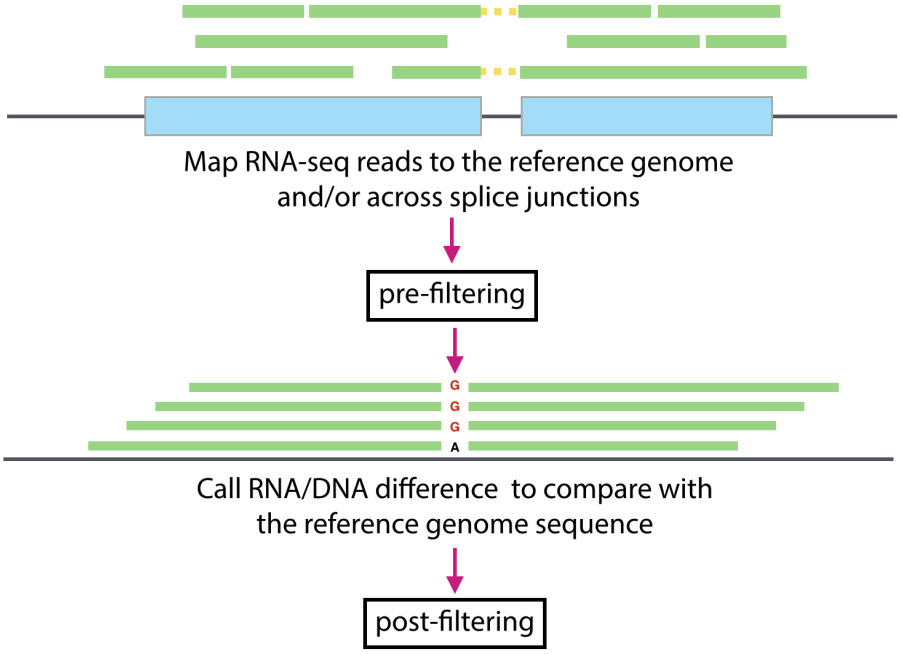

図 1.8: 超並列シーケンサーを用いた基本的な RNA 編集サイトの同定手法の解析パイプライン

超並列シーケンサーから得られた RNA-seq や DNA-seq データは、まずリファレンスとなるゲノム配列・遺伝子配列へマッピングを行う。次に、カバレッジやマッピングクオリティ、重複リード (PCR duplicated reads) の除外などを行う (pre-filtering 処理)。RNA 編集サイトは、ゲノム配列との一塩基ミスマッチとして検出することができるため、RNA とゲノム配列の比較を行い、編集サイトを検出する。その後に、SNP やリピート配列、スプライスサイト周辺の編集サイトを排除し、最終的な RNA 編集サイトの候補を得る (post-filtering 処理)。

2011 年に Li *et al.* (2011) の研究チームは、27 個体のヒトの RNA-seq および DNA-seq データから、大量の RDD サイト (RNA/DNA difference site) が検出されたという非常に興味深い研究成果を発表した (Li *et al.*, 2011)。RDD サイトは、4 × 3 の A-to-G 編集を含めた全ての塩基置換のパタンを指す。解析には、B 細胞 (immortalized B cell) が用いられ、Illlumina Genome Analyzer IIx による 50 塩基の 1.1 億リードをヒトゲノムにマッピングし、合計で 28,000 箇所以上の RDD サイトを検出した。10,000 箇所以上がエクソン領域における RDD サイトであった。ADAR によると考

えられる A-to-I 編集は 6,700 箇所、APOBEC による C-to-U 編集は 1,200 箇所程度を報告した。また、その他の A-to-C 編集や T-to-G 編集など修飾酵素が発見されていない RDD に関しては、未知のメカニズムが関与する可能性を示唆した。この結果は、同一個体から得られたゲノムとトランスクリプトームが高頻度で編集を受けている可能性を投げかけた意味において強いインパクトを示した。

ところがこの論文が掲載された後、解析結果に対する 4 本の追従論文が発表され、Li *et al.* (2011) によって検出された編集サイトの 95%以上は擬陽性の可能性が高いことが指摘された (Kleinman and Majewski, 2012; Lin *et al.*, 2012; Pickrell *et al.*, 2012; Schrider *et al.*, 2011)。論文では、主に RNA-seq リードを参照ゲノムへマッピングする際のバイアスについて検証が行われ、多くのエラーとその原因が報告された。こういった背景から、Li *et al.* (2011) の研究とその追証は、アラインメントデータへの適切なフィルタリングが編集サイトの正確な検出には極めて重要な過程であることを示し、複数のフィルタリング手法を組み合せたヒューリスティックな RNA 編集サイトの検出手法が開発される契機となった。現在では、検出される全体の RDD のうち、A-to-G 編集が 80%以上のデータを示すこと、報告された既知の RNA 編集サイトとの一致を示すことが検出手法の妥当性を示す主要な論調となっている。

1.4.2 RDD サイトに見られる多種のエラーとバイアス

Pickrell *et al.* (2012) は、Li *et al.* (2011) のシーケンシングデータを再解析した結果から 88%から 93%が参照ゲノムへのマッピングエラー、シーケンシングエラー、遺伝的な多型に起因した技術的エラー (technical artifacts) だと結論づけた。主要なエラーの原因は、ショートリードのマッピングにおける positional bias と strand bias の二つである。

Positional bias は、RDD サイトをカバーしたショートリードにおける位置の偏りである。*ReadLen* bp のショートリードにおける RDD サイトの位置 i は、$\{i, ReadLen - i\}$ である。i の分布は DNA と RNA の二つのデータに対して同一である、という帰無仮説を t 検定により統計的な検定を実施した。結果、8,000 箇所の RDD サイトは、ショートリードの両端 5bp に極端に集中するという positional bias が顕著に見られることを報告した。

Strand bias は、RDD サイトをカバーするショートリードの方向 (forward/reverse strand) に見られる偏りである。RDD サイトをカバーする RNA と DNA の 2 つのデータにリードの向きを加えた (DNA+/DNA-/RNA+/RNA-)4 つのクラスに分割し、それぞれ 4 つのクラスでリード数をカウントする。RDD をカバーするリードとそれ以外のリードにおける方向は、同一であるという帰無仮説に対して、フィッシャーの正確確率検定により偏りを検定した。結果、positional bias と同様に、RDD サイトにおいてはサポートするショートリードの向きが forward あるいは reverse に有意に偏ること結果が示された。

超並列シーケンサーから得られるショートリードの配列は、ゲノム中の反復配列へ適切にマッピングすることが難しく、本来とは異なったゲノム座標へのミスアラインメントの頻度が高くな

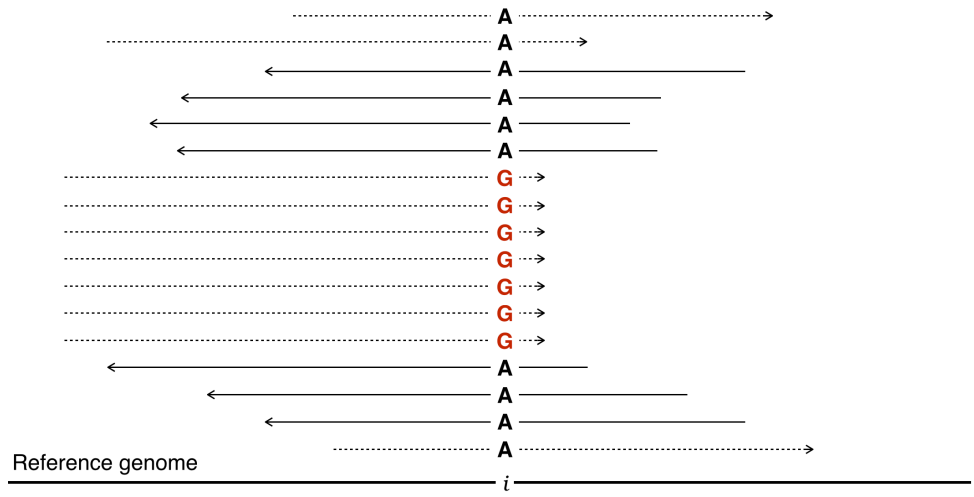

図 1.9: 検出された A-to-G 編集サイトに見られる positional bias および strand bias

実線は forward strand、破線は reverse strand のショートリードがそれぞれ参照ゲノム配列へマッピングされた結果を表す。黒文字で示した A に対して、ミスマッチ塩基は赤文字の G であることから、A-to-G 編集が検出されている。参照ゲノムとマッチしている A 塩基は、forward リードと reverse リードともにカバーされているが、ミスマッチの G 塩基は reverse strand のリード (破線) のみからカバーされており、strand bias が見られる。同時に G 塩基へのミスマッチは全てのリードが末端の同位置にあり、positional bias が見受けられる。

る (Schrider *et al.*, 2011)。誤った座標へのアラインメントを避けるため、BLAST (Basic local search tool) (Altschul *et al.*, 1990) などのローカルアラインメントを行うツールを使い、候補サイトをカバーするショートリードを参照ゲノム配列へ再度アラインメントする (re-alignment) 手法がとられている。リアラインメント結果から、ゲノムの複数箇所にマップされたリードを除くことで、候補サイトにに関する精度の高いマッピング結果が得られることが示されている。この反復配列へのマッピングに類似した問題として、spliced アラインメントを行った場合に、エクソンとエクソンの境界のスプライスサイト周辺にミスマッチが集中することが知られている。そのため、既存の遺伝子構造アノテーションを利用し、スプライスサイト周辺の 5 塩基程度は解析から除外するフィルタリングが行われる場合が多い。加えて、検出された編集サイトが、ADAR による転写後修飾ではなく、SNV や SNP などゲノム上の変異を反映している可能性が考えられる。そのため、ヒトでは、1000 Genomes Project など国際的なプロジェクトによって同定された SNP のサイトを除外するフィルタリングが行われる。

1.4.3 検出された編集サイトの実験的な検証方法

RNA-seq データを用いた RNA 編集サイトの網羅的な解析は、数千から数万のオーダーで候補を検出する。これらのサイトが真の編集サイトであるかを結論付けるため、現在ではサンガー法による実験的な検証が用いられる場合が多いが、全てのサイトをシーケンシングし直すことはスループットの問題から現実的ではないため、10 箇所から 100 箇所をランダムに抽出し、実験的に

検証する場合が殆どである。サンガーシーケンシングにおいても、対象となる編集サイトがゲノムの反復領域に由来する場合、PCR プライマーの特異性が保証されず、編集サイト周辺のみを増幅できない問題がある。これは、Peng *et al.* (2012) の研究においてサンガーシーケンシングによって検証されたサイトを追証した Piskol *et al.* (2013) によって指摘されている。追証実験から、反復配列に由来した編集サイトは、PCR バイアスによって正確な検証が行えないことを指摘している。

アミノ酸置換を引き起こす編集の場合には、質量分析器によって得られるペプチド配列のフラグメントから、アミノ酸置換の有無を確認する手法が用いられる (Li *et al.*, 2011)。この方法は、非同義置換を伴ったコード領域における編集サイトの検証に限られる。

1.4.4 A-to-I 編集サイトの検出に適切な実験デザイン

超並列シーケンサーから出力される RNA-seq データは、転写物の発現量やスプライシングを考慮した遺伝子構造を推定する用途においては、精度および再現性ともに高く、有用な手法として広く用いられている (Marioni *et al.*, 2008; Mortazavi *et al.*, 2008)。ところが RNA 編集の検出においては一塩基の解像度が求められるため、実験およびデータ解析で混入するノイズの影響を相対的に受けやすくなる。これまでに読み取りエラー (Base call error) の原因として、DNA と比較して RNA は分子として安定性を欠くこと、cDNA 合成における逆転写酵素の活性、PCR の配列ごとに見られる増幅バイアス、特定の塩基配列に高頻度の読み取りエラーが発生することなどが指摘されている (Aird *et al.*, 2011; Meacham *et al.*, 2011; Nakamura *et al.*, 2011)。このように、ノイズと ADAR に由来した A-to-I 編集サイトを正確に分離することが、高精度な検出に必要であり、適切に実験をデザインし、シーケンシングデータを取得する必要がある。既往の研究においても、精度の高い検出手法の開発を成功させているものは、情報学的解析のみならず、実験デザインにも優位性が見られる。

先行研究では、カバレッジが 10× 以上、75 塩基以上のリード長、paired-end シーケンシングが事実上の標準となっている。加えて Ramaswami *et al.* (2012) や Peng *et al.* (2012)、Zhu *et al.* (2013) らの解析では、転写物の向きを考慮した strand-specific RNA-seq が用いられており、センス鎖とアンチセンス鎖を分離した解析を行っている。転写物の方向性を考慮することは、塩基置換のパタンをより正確な分類を可能にする。Rodriguez *et al.* (2012) らによる解析では、生物学的レプリケートを 2 サンプル用意し、G 検定による統計的な検出手法が用いられている。また、Bahn *et al.* (2012) や Rodriguez *et al.* (2012) らのハエやショウジョウバエにおける解析では、*Adar* 変異体を同時にシーケンシングし、ADAR の発現時のみ見られた A-to-I 編集サイトの検出を行っている。ただし、マウスなど *Adar* の変異体が致死性を示す場合には適用が不可能である。また、対象生物種やセルラインのトランスクリプトームと同時にゲノム配列のシーケンスにより、編集サイトがゲノム上の SNV (Single nucleotide variant) などの変異を反映しているか否かを判断可能となるため、多くの研究で用いられている。

ここ数年、ADAR による A-to-I RNA 編集に関する研究は、目まぐるしい進展を遂げている。真

核生物において発見当初こそ、ランダムで生物学的意義については不明であった非翻訳領域におけるRNA編集であったが、今日では遺伝子の多様な発現制御、スプライシング機構へも影響しているというRNA編集の役割を大きく拡張し、新しい描象が解き明かされつつある。このような現象を貫いているのは、ADARおよびADARと相互作用するDicerやDroshaといった他の因子もまた二本鎖RNA結合タンパクということである。今後、ADARと二本鎖RNA結合タンパクの相互作用は、実験と情報解析の両側から更に研究されることで、ADARの新たな機能は次々と発見されることが予想される。二本鎖RNA結合タンパクとの相互作用を介したADARの理解は、真核生物におけるRNA編集の意義を解き明かす今後の大きな流れとなるであろう。

こういった背景から、大規模な超並列シーケンスデータからノイズと生物学的背景に由来する事象とを精度よく分離可能な実験デザインと情報学的手法を組み合わせた解析の重要性は、今後さらに高まることが予想される。

第2章　RNA編集サイトの検出手法の精度比較

2.1　研究背景

RNA-seqデータを用いたRNA編集サイトの検出は、これまでにヒト、マウス、ショウジョウバエを主な対象生物種として情報学的な手法が多数、開発されてきた。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されるが、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。そのため、ADAR由来のRNA編集サイトと擬陽性とを高精度に分離する検出手法がこれまでに考案されてきた (Lee *et al.*, 2013)。

ところが、論文ごとに解析対象となる生物種や組織、セルライン、シーケンシング手法などに相違が見られ、各々の手法の検出精度の評価・比較は困難な状況にある。また多くの研究はこれまでに同定された既知の編集サイトとの一致 (共通項) を確認しているに過ぎず、検出結果に含まれる擬陽性の影響を適切に評価するには至っていない。RNA編集サイトを高精度に検出する手法を開発するにあたっては、既知の結果との一致と同時に、検出結果に含まれるノイズとその割合を評価する必要があると考えられる。

バイオインフォマティクスの分野においては、タンパク質の立体構造予測コンテスト CASP (Critical Accesment of protein Structure Prediction) や BALIBASE (Benchmark alignment database) (Thompson *et al.*, 2005) に代表されるように、その分野で開発されたアルゴリズムや手法は、予測精度や計算時間といった複数の指標により性能評価 (Benchmarking test) が実施され、手法の標準化が行われてきた経緯を持つ。本研究もこういった文脈に位置づけることができる。

異なる検出手法についての精度評価は、生物種やサンプル毎に適したフィルタリング手法とそのパラメータを明らかにすることが目的である。序論で前述したとおり、RNA編集サイトの高精度な検出には、様々なフィルタリング手法を複合的に組み合わせることによって擬陽性の排除が行われており、手法ごとにゲノムへのマッピングにおける許容ミスマッチ数、最小のリードカバレッジなど、擬陽性の排除に使用されたフィルタリング手法は異なる。検出手法の性能評価により、高精度な検出手法が明らかとなると、他の手法との差を高精度な検出に寄与するパラメータの候補として抽出することが可能である。このような知見は、汎用性のある新たなRNA編集サイトの検出手法の開発に貢献できると考えられる。

本研究は、ヒト・マウス・ショウジョウバエのRNA-seqデータを用いた既存のRNA編集サイトの検出手法に関する性能評価を行い、高精度な検出に寄与するパラメータの探索を行ったものである。性能評価には再現率および適合率といった指標を導入し、各手法の検出精度について定量的な比較を行った。結果、生物種ごとに検出手法の精度を明らかにした。この結果を受け、シー

ケンシング手法などの実験デザインおよび検出手法が検出精度にどのように寄与しているのかについて議論する。

2.2 対象と手法

2.2.1 性能評価に用いた指標

これまでに開発されてきた RNA 編集サイトの検出手法を統一的な指標によって性能評価を行うため、適合率 (Precision)・再現率 (Recall)・F 値 (F-measure) と呼ばれる 3 つの指標を導入した。この 3 つの指標は情報工学において検索精度の測定において用いられており、情報検索の結果に対して検索ノイズと検索漏れという 2 つの側面からアルゴリズムの性能を測定するものである (Davis and Goadrich, 2006)。

以下にぞれぞれの指標とその計算方法について示す。この 3 つの指標は、算出される値が高いほど高精度であることを表す。以下で説明する際の検出サイトは、各手法により検出された全 A-to-I 編集サイトを指し、正解サイトは過去に報告された既知の全編集サイトを表す。

(1) 適合率 (Precision)

適合率は、式 2.1 に定義される。検出サイトと真の editing サイトの積集合に対する検出サイトの割合として計算される。適合率は検出サイトに正解が含まれる割合を意味しており、検出サイトの中に正解サイトとの一致が濃縮されることで高い適合率が達成される。

$$Precision = \frac{TP}{TP+FP} = \frac{CandidateSites \cap TrueEditingSites}{CandidateSites} \qquad \text{式 (2.1)}$$

(2) 再現率 (Recall)

再現率は、式 2.2 に定義され、正解セットに対して検出サイトが正解した割合を示す。ゲノム中における多くの箇所を edititing サイトとして検出することで原理的に再現率は 1 に近づくが、一方で擬陽性が増大するため適合率は 0 に近づく。このように、両者はトレードオフの関係にあることから、正確な RNA 編集サイトの検出には、再現率の向上よりも適合率を増加させる手法やパラメータを見出すことが重要である。

$$Recall = \frac{TP}{TP+FN} = \frac{CandidateSites \cap TrueEditingSites}{TrueEditingSites} \qquad \text{式 (2.2)}$$

(3) F 値 (F-measure)

F 値は、適合率および再現率の調和平均として式 2.3 のように定義される。F 値の導入により、再現率と適合率の 2 つの評価基準を一元的に取り扱うことが可能となる。

$$Fmeasure = 2 \times \left(\frac{precision \times recall}{precision + recall} \right) \quad \text{式 (2.3)}$$

2.2.2 正解セットの構築

検出精度を評価するためには、真の RNA 編集の集合を定義した正解セットの構築が必須となる。そこで、これまでの研究によって報告された A-to-I 編集サイトを収集した DARNED (Database of RNA editing site, `http://darned.ucc.ie/`) (Kiran and Baranov, 2010) と呼ばれるデータベースに登録されている全データをヒト・マウス・ショウジョウバエの 3 種ごとに取得し、正解セットを構築した。DARNED は、実験的な手法により同定された A-to-I 編集サイトの他に、情報学的な手法により同定されたサイトの双方を含み、文献と紐付けられた編集サイトに関するメタデータと共に公開している。表 3.1 に、生物種ごとに収集した正解セットの総数をまとめた。尚、精度評価の実施にあたり、取得した全てのサイトを正解セットとして扱い、同定手法の如何については区別しなかった。

表 2.1: DARNED より収集した A-to-I 編集サイトの内訳

Species	Reference genome version	Studies	Answer sites
H. sapiens	hg19	20	333,216
H. sapiens	hg18	22	259,705
M. musclus	mm10	4	8,341
M. musclus	mm9	4	8,352
D. melanogaster	dm3	3	1,969

DARNED により取得した 3 種それぞれの正解セットの情報示す。3 種それぞれに対応するゲノムのバージョン、正解セットの検出に関わる先行研究の数、収集した正解セットの総数をそれぞれ示す。

2.2.3 性能評価に用いた検出手法

RNA 編集サイトの検出精度に関する性能評価を実施するにあたり、(i) RNA-seq データを用いた検出手法であること、(ii) シーケンスデータが Sequence Read Archive (`http://www.ncbi.nlm.nih.gov/sra/`) において公開されていること、(iii) 同定された RNA 編集サイトの全リストが取得可能であることを評価の条件とし、ヒト・マウス・ショウジョウバエの 3 種類からそれぞれ評価する先行研究を選出した。表 2.2 にその一覧を示した。一覧では、便宜的に先行研究ごとに分類しているが、先行研究によっては異なる組織やセルラインに多数の解析を行っている場

合があるため、この場合に関してはサンプルごとに性能比較を行うようにした。これにより同一手法における組織やセルライン毎での精度の違いを明らかにすることができると考えられた。

表 2.2: 性能評価に用いた検出手法のリスト

Study	Samples (tissue/cell line)	Identified sites
H. sapiens		
Ramaswami, *et al.,* 2012	GM12878 (lymphoblastoid cell line)	74,166
Ramaswami, *et al.,* 2013	Lymphocyte cell line	230,402
Peng, *et al.,* 2012	Lymphoblastoid cell line in Han Chinese individual	22,696
Park, *et al.,* 2013	14 human cell lines (ENCODE project)	13,821
Zhu, *et al.,* 2013	16 human tissues, 2 cell lines (Illumina BodyMap 2.0)	2,246
Bahn, *et al.,* 2012	U87MG (glioblastoma cell line)	12,791
M. musculs		
Gu, *et al.,* 2012	White adipose, Femurs, liver	244
Dillman, *et al.,* 2013	Cerebral cortex, 4 embryonic mice	177
Lagarrigue, *et al.,* 2013	Liver, Adipose	363
D. melanogaster		
Rodriguez, *et al.,* 2012	Fly head	1,351
Graveley, *et al.,* 2011	Fly head	973
Ramaswami, *et al.,* 2013	Fly head	850

本研究において精度評価の対象となった先行研究の検出手法を生物種ごとに分類して示した。先行研究が用いたセルラインや組織と最終的な A-to-I 編集サイトの総数を表記した。異なるセルラインや組織の結果を統合的に解析している場合は、サンプル欄に複数を列挙した。

2.3　検出性能の比較結果

2.3.1　生物種毎のごとの精度比較

(1)　ヒト

ヒトについては、6 つの先行研究について合計 27 サンプルを用いた性能評価を行った。再現率、適合率による各手法の検出精度を図 2.1A に示す。対象とした検出手法の全体としての精度の傾向は、どの手法も共通して 0.2 以下の低い再現率を示し、中でも最も高い再現率は、Alu 領域から検出されたサンプルを用いた Ramaswami *et al.* (2012) の 0.15 であった。適合率については、手法とサンプルによって大きく分散することが示された。適合率に関しては Ramaswami *et al.* (2012) および Bahn *et al.* (2012) が共に再現率 0.98 を示し、高精度な手法であることがことが示された (図 2.1A)。ENCODE プロジェクトにおいて、用いられた手法はどのセルラインに対しても共通して低い再現率と適合率を示す傾向が観察された。中でも、セルライン GM12892 における再現率が 0.13 と最も高く、適合率については、0.0005 程度とどの手法も極端に低かった。そのため、F 値はどのセルラインについても呼応して低い傾向が見られた。F 値は、再現率の最も高かった Ramaswami *et al.* (2012) が 0.15 であり、最も高い F 値を示す手法であることが分かった。

再現率と適合率を用いた性能評価を行ったところ、検出手法ごとに再現率、特に適合率が大きく分散する傾向が観察された。この原因を探るため、再現率および適合率が検出された編集サイト数との関係性についての解析を行った結果を図 2.1B に示す。その結果、高い再現率を示す手法は、検出した編集サイトの総数も 60,000 サイト以上と相対的に多い傾向が見られた。適合率については、二つのパタンが観察された。適合率が 0.3 以下のグループでは、適合率の上昇と検出された編集サイト数には、相関は見られず、10,000 箇所以上を検出した Ramaswami *et al.* (2013) の脳のサンプルに対して、1000 箇所程度を同定した HSMM、GM12891、GM12892 セルラインの方が高い再現率を示した。また、再現率が 0.5 以上については、この傾向が顕著に見られ、検出サイトの増加と精度の上昇の間に関係性は見られなかった。

(2)　マウス

マウスを対象とした検出手法の性能評価の結果を図 2.2 に示した。再現率は、用いた 3 つの手法は、最大でも Gu *et al.* (2012) の 0.03 が最大であり、極端に低い値を示した。適合率は、Gu *et al.* (2012) が最も高い 0.99 を示し、続いて Dillman *et al.* (2013) が 0.75 であることが示された (図 2.2A)。図 2.2B は、ヒトの解析と同様に検出精度と検出した編集サイト数の関係性をプロットした結果であるが、再現率は検出編集サイト数と 3 つ全ての手法において正の相関が見られた。対して、適合率は、Lagarrigue *et al.* (2013) の手法を除くと、正に相関する傾向が観察された。他の手法とサンプルの 3 つに関しては、検出数と適合率は強く相関することが示された。

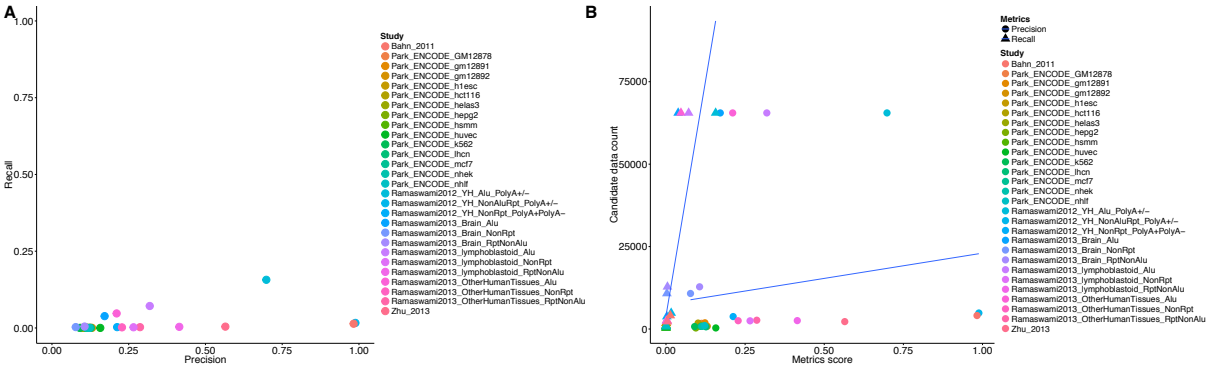

図 2.1: ヒトにおける検出手法の精度評価

A: 横軸に適合率、縦軸に再現率をそれぞれ示した。図中の色分けは、手法あるいはセルラインや組織ごとに変えて表した。ヒトを対象とした手法について、縦軸にそれぞれの手法によって検出された A-to-I 編集サイト数を示し、横軸に適合率および再現率を図示した。**B**: 検出精度の指標はそれぞれ黒丸は適合率、黒三角は再現率に対応する。また、青線は適合率および再現率と候補サイト数の単回帰直線を示した。

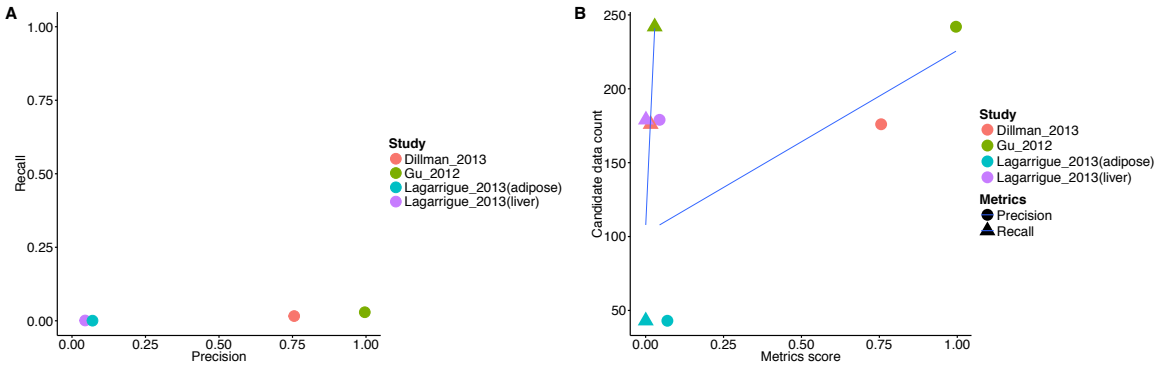

図 2.2: マウスにおける検出手法の精度評価

A: マウスにおける検出精度の結果を示す。**B**: 候補となった A-to-I 編集サイト数と検出精度との関係性を示す。青線は、候補サイト数と検出精度の回帰直線を示す。

(3) ショウジョウバエ

ショウジョウバエを対象とした検出手法の精度比較の結果を図 2.3A に示す。最大の適合率を示した手法は Rodriguez *et al.* (2012) の 0.49 であった。また、再現率に関しては、Graveley *et al.* (2011) (modENCODE プロジェクトの手法) が最大を示した。図 2.3B に示したように、候補となる編集サイトを多く検出した手法ほど、再現率と適合率が向上するという傾向が確認された。また、Ramaswami *et al.* (2013) において用いられた検出手法は、再現率ならびに適合率はどちらも他 2 つの手法に対して低い精度であることが示された。

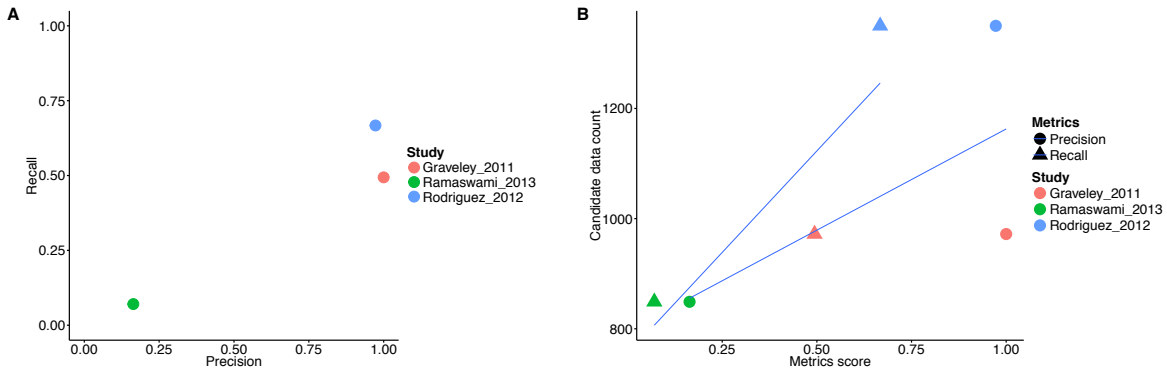

図 2.3: ショウジョウバエにおける検出手法の精度評価

A: ショウジョウバエにおける検出精度の結果を示す。**B**: 候補となった A-to-I 編集サイト数と検出精度との関係性を示す。青線は、候補サイト数と検出精度の回帰直線を示す。

2.3.2 検出手法の詳細とパラメータ

図 2.4 では、今回の精度検証に用いた全手法の F 値を示した。全体の傾向として、ショウジョウバエに用いられた手法の精度が高く、次はマウスに適用された手法が F 値の上位に位置することが示された。対して、ヒトを対象とした手法は全てにおいて、マウスおよびショウジョウバエよりも高い順位を示すものは見られなかった。

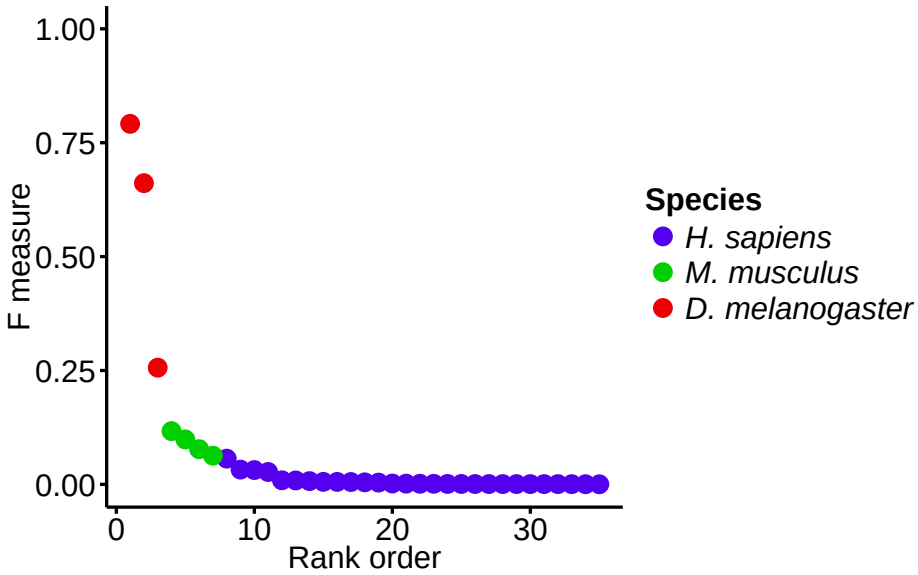

図 2.4: 生物種および検出手法ごとの F 値の順位

赤はヒト、青はマウス、緑はショウジョウバエを対象とした手法をそれぞれ示しており、各手法に関する F 値を昇順に並べて示した。

		Ramaswami, 2013	Ramaswami, 2012	Peng, 2012	Park, 2013	Zhu, 2013	Bahn, 2012	Gu, 2012	Dillman, 2013	Lagarrigue, 2013	Rodriguez, 2012	Graveley, 2011
Data	DNA-seq											
	RNA-seq		PolyA+	PolyA+/-	PolyA+	PolyA+				PolyA+		PolyA+
	Adar null strain											
	Replicates		2								2	
Aligner		BWA	BWA	SOAP2/ SOAPSplice	Bowtie	Bowtie	Bowtie/BLAT/ Tophat	Bowtie		mrsFAST/Tophat	TopHat	
Reference	Genome	hg19	hg19	hg18	hg19	hg19	hg18	mm9	mm9	mm9	dm3	dm3
	Transcripts		Gencode, RefSeq, Ensembl and UCSC		GENCODE							
pre-filter	Remove duplicate reads	Picard	Samtools rmdup		Samtools rmdup							
	Realign to indel											
	Recalibrate baq											
	Uniquely mapped reads											
	Caller	GATK	In house	SOAPsnp	In house	In house	In house	In house	samtools		In house	In house
post-filter	Base call quality	> 25	>= 25	>= 20				>= 20				
	Mapping quality	> 21	> 21									
	Coverage		> 10	5					>= 5	>= 10	>= 10	
	Alternative N bases		> 2	> 2	> 2	>= 2						
	Base modification									(DNA+RNA)/depth >= 90%		
	1 variant within N bp span			5						5		
	Discard mismatch in the Nbp	6	6				5					
	Highly similar region	BLAT	BLAT	BLAT					BLAT			
	Supporting mismatch			> 2 reads	> 10%	> 5%		>= 5%		>= 10%		>= 5 reads
	Expression level					5(FPKM)						
	Effective signal					> 95%						
	Strand bias											
	Extreme edited sites			100%								
	Ubiquitous events					> 3 tissues/cell lines			Shared by all samples	Shared by 5/6 samples	Shared by 5/6 samples	least 2 adult samples
	N bp of splice junction	5	4		5							6
	N bp repeat	> 6		> 1						>= 3		
	Homo-polymer		> 6bp									
Knowledge filter	SNP sites	dbSNP	dbSNP, 1000 genomes, EVS	dbSNP, 1000 genomes								
	ChIP-seq data				call CNV							
	CNV sites			CNVnator								
	Mask repeat seq.		RepeatMasker annotation							SciRoKo		
	hypervariable sites	abParts in UCSC										
	Bidirectional transcripts											

図 2.5: 性能評価に用いた検出手法において使用されたフィルタリング手法とそのパラメータ

赤はヒト、青はマウス、緑はショウジョウバエを対象とした手法をそれぞれ示す。濃くハイライトされた手法は、各生物種において最も精度の高い手法を示す。

2.4 議論

本研究は、情報検索の分野で用いられてきた適合率・再現率および F 値と呼ばれる精度評価の指標を導入し、RNA 編集サイトの検出手法の精度を定量化し、異なる複数の手法の直接的な比較を可能にした。先行研究においては、T-to-G や C-to-A など全塩基置換の個数を集計し、全体の分布において A-to-G への置換が濃縮されたピークとして見られること、または DARNED のデータを使い、既知の報告例との一致の割合を示すことにより、検出手法の妥当性を示してきた。従って検出手法ごとに異なったデータや方法で検出精度の妥当性を評価しており、検出手法を比較することは困難であった。

ヒトを対象にした手法の性能評価によると、どの手法も平均的に低い再現率を示す傾向が見られたが、これは DARNED から収集した正解セットが 330,000 箇所以上と検出された編集サイトを数倍上回る母数であったことが原因だと考えられる。この影響を受けて、ヒトの RNA-seq データに用いられた手法は総じて、低い再現率を示したと考えられる。ENCODE プロジェクトによる解析も低い再現率を示しているが、この原因も同様に検出された編集サイトの総数が正解セットに対して少ないことが原因だと考えられる。ENCODE プロジェクトにおいては、多くのセルラインを用いた解析が行われたが、最終的な検出サイトは 1,000 から 2,000 サイト程度であり、これは他の手法による検出された編集サイトの 1/10 程度である。この傾向を裏付ける結果として、再現率は検出サイト数の増加に対して比例的に上昇する、という傾向が見られている (図 2.1B)。再現率の計算上、候補数が増加すると再現率もまた高くなるからである。対して、適合率にはこのような線形的な関係が見られないのは、検出サイト数は増加する分、擬陽性の割合も同時に高くなるため、再現率とは異なる傾向が見られたのだと考えられる。

ヒトの RNA-seq データを用いた解析では、Ramaswami, *et el.,* 2012 が最も高い精度を持った手法であることが明らかとなった。この手法において精度の高い検出には、実験段階では、Strand-specific シーケンシングの他に *Adar* の変異体を同時にシーケンシングしている点である。また、データ解析においては、ショートリードをゲノムにかぎらず、遺伝子モデルへもマッピングしている点である。このようなリードのマッピング戦略は、スプライスサイト周辺に集中して見られるミスマッチの減少に寄与していると考えられる。また、他の手法に対して、Base call quality や Mapping quality のしきい値が厳格である他、ホモポリマー領域のフィルタリングを行っている。ホモポリマーは、シーケンサーによる配列決定に誤読が誘発されやすい繰り返し領域である。このような実験デザインおよびフィルタリング手法により、Ramaswami, *et al.,* 2012 は高精度化を達成していると考えられた。

マウスの RNA-seq データを用いた検出手法は、共通して再現率が極端に低い傾向が示された。この低い再現率の原因は、マウスを用いた検出手法は正解セットに対して、検出した編集サイトが少ない点が考えられる。10,000 サイトからなる正解セットに対して、既知のサイトと完全一致 (適合率 1.0) した場合にも、再現率は最大でも 0.1 にとどまるからである。適合率に関しては、少ない母数に正解セットとの一致が濃縮された結果、Gu *et al.,* 2012 や Dallman *et al.,* 2013 では高い

適合率を示した。

ショウジョウバエにおける検出手法の精度比較を行った結果、再現率に関しては非常に高い結果を示した。これは正解セットおよび検出手法のどちらもが、ショウジョウバエの頭部から抽出されたRNAであり、組織の同一性が高い再現率に寄与している可能性が考えられる。適合率について議論する。最もF値の高かったRodriguez *et al.,* 2012は、*Adar*の変異体を用いている。サンプル毎に2つの生物学的レプリケートを使っている。このような実験デザインは、高精度な検出には必要となると考えられる。

本研究は、超並列シーケンサーを用いたRNA編集サイトの検出手法の精度比較を初めて実施した結果、ヒト、マウス、ショウジョウバエにおける複数の検出手法の持つ検出精度を明らかにした。こういった方法は、新たなシーケンシングデータの情報解析にとどまらず、新規の検出手法の開発にも貢献できると考えられる。

今後、より正確な性能評価には、全ての手法において同一のRNA-seqデータを使って再解析する必要があると考えている。本解析では、入力となるRNA-seqデータと手法の双方が先行研究ごとに異なり、純粋に手法のみの比較が行えていない可能性が示唆される。そこで、解析対象となったRNA-seqデータと検出手法の影響を明確に分離し、RNA-seqデータを比較する全ての手法で同一にした上で、性能評価を行う必要があるだろう。そのためには、ENCODEプロジェクトなどの豊富なデータを用いて、純粋な手法の比較を行うことが望まれる。加えて、計算方法が正解セットのサイズに依存的であるため、多数のサイトを検出する手法は高い再現率を示しやすく、これは結果的にF値の上昇に直結する。検出サイト数に応じて正解セットをランダムサンプリングするなどの正解セットの適切な正規化も有効となると考えられる。

第3章　RNA編集サイトの検出ソフトウェアの開発

3.1　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 (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、論文中に記述されるため、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。本卒業論文の第 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 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。

3.2 システムの設計

3.2.1 動作環境

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 3.1 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、RNA 編集サイトを検出するための Ivy と、精度検証を行うための edit_bench の二つのコマンドラインツールが付属するパッケージである。

Ivy の開発は、Mac OSX 10.9 上の Python v.2.7.5 で行われており、Python3 およびそれ以下のバージョンでは動作試験は行っていない。Mac 以外の Unix 環境として、RHEL (Red Hat Enterprise Linux) 上での動作は確認している。Ivy は、複数の C 言語および Python によって書かれたライブラリに依存している。入力として受け取るアラインメントデータの処理には Pysam v0.7.7、VCF (Variant call format) (Danecek *et al.*, 2011) の処理には pyVCF v0.6.4、フィッシャーの正確確率検定には fisher v0.1.4 に依存する。edit_bech は、図の生成に matplotlib v1.4.1+に依存する。これらの依存関係は、インストール時に自動的に解決を試みる仕様だが、高速化のために一部のライブラリのインストールには、事前に gcc や clang など C コンパイラを用意しておく必要がある。インストール自体はコマンド一つで簡便に行えるようになっているほか、PyPI (Python package index) への登録を予定しており、インストール時にソースコードのダウンロードも不要となる予定である。

3.2.2 Ivy の設計と実装

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。以降では、詳細な Ivy のクラス設計と実装を概観してゆく。

AlignmentStream クラスは、入力されたバイナリ形式のアラインメントの情報が含まれるオブジェクトを一塩基ごとに読み出す中核的なクラスである。このように参照ゲノムの一塩基ごとに RNA-seq/DNA-seq のアラインメント結果へと変換する過程を pileup と呼び、pileup した pysam.AlignedRead オブジェクトは、ショートリードの配列、リードの向きやクオリティ、参照ゲノムへの座標などアラインメント結果に関する多くの属性 (Attributes) を持つ。AlignmentStream クラスはこれらの属性を参照しながら、委譲 (Deligation) された AlignmentFilter クラスや AlignmentReadStats クラスが、アラインメント結果のフィルタリングや統計的なバイアスのフィルタリングを行う実装を持つ。AlignmentStream での pileup 処理は、数百 MB から数十 GB のアラインメントファイルを対象とするため、ヒトゲノムの解析では一番染色体へ限定的に解析したとしてもオブジェクトのサイズはメモリ空間を圧迫しかねない。そのため、AlignmentStream.pileup_stream メソッドは、pileup した返り値をイテレータとする実装を持ち、省メモリでの解析を実現している。また、Ivy は Logger クラスを持ち、解析の各段階における累積時間、解析に使用された全オプション、発生した警告やエラーの出力などを制御している。Ivy が期待と異なった動作をした場合には、Logger クラスから出力された冗長なログが有用だと考えられる。デフォルトでログの出力レベルは最低限に抑えられており、必要に応じて有効にして使う実装とした。

Ivy は、ヒト、マウス、ショウジョウバエといった高等真核生物を主要な解析対象種とした編集サイトの検出を実行する。原理的には、pileup をするためのメソッドコール数は、ゲノムサイズに比例し、ゲノムサイズが大きな生物に対しては計算時間が増大することが予想される。そのため、複数のスレッドを使用した並列化が現実的な時間内での解析には必須だと考えた。Ivy は、python の標準ライブラリを使用した並列処理を提供する。検出の高速化のために使用可能なスレッド数に応じ、自動的に染色体数を均等に分割し、編集サイトを検出する。RNA 編集サイトは、染色体をオーバーラップして起こる現象ではないため、染色体を分割することによるアラインメントデータの損失などの問題は起こらないと考えた。

Ivy は、生物種、セルライン、シーケンシング手法、実験デザインなどの各種条件が異なるアラインメントデータが入力されることが想定される。こうったデータ生成プロセスに多様なバックグラウンドを持つデータに対応するため、可能な限りフィルタリングのオプションを提供し、多様なオプションとパラメータによる解析を可能とする実装を目指した。こういった複雑なオプションやパラメータを取る解析ツールは、簡便性と背反することが想定されるが、Ivy ではフィルタリングのオプションを複数のクラスに分類する実装を持つことで回避している。加えて、各クラスに分類されているオプションは、先行研究を参考にしたデフォルトの値が設定されているため、重複リードや INDEL リードの排除など基本的なオプションを改めて設定する必要はない。

基本的なフィルタリングクラス (BasicFilter) では、入力された RNA-seq/DNA-seq データに対して、リードカバレッジや塩基のクオリティ、アレル頻度、許容するミスマッチの種類などの基本的なフィルタリングを提供する。統計的なフィルタリングクラス (StatisticalFilter) は、positional bias や

strand bias などの統計的なアラインメントデータへのフィルタリング手法を提供する。SampleFilter クラスは、実験デザインに関するオプションを提供する。Strand-specific RNA-seq や *Adar* のノックダウンの RNA-seq などの入力に対応している。

表 3.1: Ivy に実装された入力データへのフィルタリング手法の一覧

Option name	Description	Default value
Basic filter class		
–min-mutation-frequency	Minimum allele frequency	0.1
–min-rna-coverage	Minimum RNA-seq read coverage	10
–min-dna-coverage	Minimum DNA-seq read coverage	20
–rm-duplicated-read	Remove duplicated mapped read	True
–rm-deletion-read	Remove deletion reads	True
–rm-insertion-read	Remove insertion reads	True
–min-rna-map	Minimum mapping quality of RNA-seq data	30
–min-dna-mapq	Minimum mapping quality of DNA-seq data	30
–min-rna-baq	Minimum phread-scaled base call quality in RNA-seq reads	28
–min-dna-baq	Minimum phread-scaled base call quality in DNA-seq reads	28
–num-allow-type	Number of allowing base modification type(s)	1
Statistical filter class		
–sig-level	Significance level for statistical testing	0.05
–base-call-bias	Considering base call bias	False
–strand-bias	Considering strand bias	False
–positional-bias	Considering positional bias	False
Sample filter class		
–strand	Use strand-specific RNA-seq data	False
–replicate	Considering biological replicates	False
–adar-null	Filter by Adar-null strain data	False

フィルタリング手法の一覧を示す。実行時に与えるオプション名とその説明、デフォルト値をそれぞれ示す。

3.2.3 edit_bench の実装

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、第 2 章で用いた再現率、適合率および F 値を用いた。各指標については、第 2 章を参照されたい。

edit_bench は大きく 4 つのクラスから実装され、DARNED から正解セットを HTTP 通信により取得し、適切な CSV ファイルにパースする Generator クラス、何らかの方法により検出された

RNA 編集サイトの VCF ファイルから、適切なデータ構造を生成と読み出しをする Reader クラス、精度検証のための各種指標を計算する Benchmark クラス、精度の検証結果を図によって可視化する Plot クラスを持つ。

Generator クラスでは、初回起動のみ引数に与えられた生物種名に対応した正解セットを DARNED より取得する必要から時間を要するが、2 回目以降はローカルに CSV ファイルとしてキャッシュするため高速に動作する。Reader クラスは、Ivy の出力である VCF を入力とし、データ構造に保存する他に、検出された編集サイトの個数などの情報を取得するメソッドを持つ。Benchmark クラスは、Reader クラスによって生成されたデータ構造を利用して、精度検証の指標を計算する。Plot クラスは、再現率と適合率に関する二次元プロットを生成するクラスであり、図の描画は補助的な機能ではなく、精度検証の結果から手法の妥当性を直感的に判断することを可能にしている。この描画機能は、複数回 Ivy を実行し、最も精度のよいパラメータの組み合わせを得る場合に使用されることを想定する。また、検証を行うサンプル数が多い場合、ラベルの色分けが離散的であると色が飽和する問題があるが、サンプル数に応じてカラーマップから連続的な階調を自動生成することで回避した。描画された結果は、PDF ファイルへと出力される。尚、図の生成には matplotlib パッケージに依存するが、他のクラスは Python の組み込みライブラリのみで動作する。

3.2.4 入出力の形式

超並列シーケンサーを用いた解析において、塩基配列データは Fastq フォーマットが標準的な形式となっており、Fastq データを参照ゲノム配列にマッピングすることによって、SNP の検出や遺伝子発現量の定量、ゲノムのアセンブリなどが行われ、RNA 編集サイトの検出も参照ゲノムへのマッピングが検出には必須である。このマッピングしたアラインメント結果を保持するデータ形式は、SAM/BAM 形式が事実上の標準フォーマットとして広く用いられている。SAM は参照ゲノムへマッピングされたショートリードの座標情報やクオリティ情報などを保持しており、座標にインデックスを貼り gzip で圧縮したバイナリ形式を BAM と呼ぶ。SAM と BAM は相互に変換が可能である。入力は、RNA-seq/DNA-seq リードをマッピングした結果から得られる BAM (Binary alignment/map format) 形式である。

入力された BAM ファイルは、Pysam (Python interface for the SAM/BAM sequence alignment and mapping format) 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。そのため、本ソフトウェア・パッケージでも pysam をメインのライブラリとして使用した。

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、

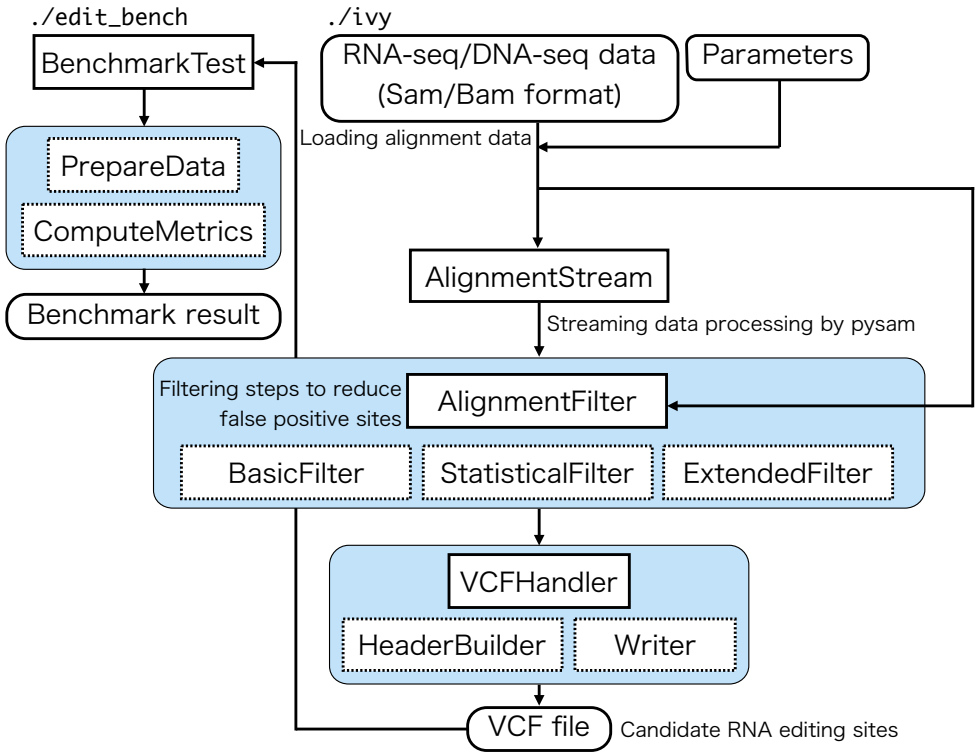

図 3.1: ソフトウェア・パッケージ Ivy の設計

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。

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。

3.2.5 ユーザーインターフェース

Ivy パッケージは、以下のコマンドで容易にインストールすることが可能となっている。Ivy のインターフェースは以下のようになっており、同梱されているサンプルデータを用いて解析を簡易的に試すことができるようになっている。以下に、ソースコードの取得とインストール方法を示した。

```
# 1. 最新版のソースコードの取得
```

```
$ curl https://github.com/soh-i/Ivy/archive/v.0.0.1-dev.tar.gz
$ tar zxvf Ivy-v.0.0.1-dev.tar.gz
$ cd Ivy-v.0.0.1-dev

# 2. パッケージのインストール
$ python setup.py install
```

(1) Ivy コマンドのインターフェース

Ivy のインストールが完了すると、Ivy と edit_bench コマンドが実行可能となるり、以下のようなコマンドで実際の解析を行うことが可能である。

```
$ ivy -f sample/hg19.fa -r sample/rna.bam -G gene.gtf -o sample_out.vcf
```

(2) edit_bench コマンドのインターフェース

edit_bench は、入力に検出した編集サイトが記述された VCF ファイルの他に、対応する生物種とそのゲノムバージョンを与える。以下、edit_bench コマンドのインターフェースと簡単な解析例を示す。

```
# 1. 基本的な使い方
$ edit_bench --vcf test.vcf --sp human_hg19 --plot --log bench.log

# 2. 正解セットを脳のサンプルに限定する
$ edit_bench --vcf test.vcf --sp human_hg19 --source brain --log bench.log
```

#1 は、基本的な使用方法について記した。`--vcf` が入力ファイル、`--sp` は生物種名とゲノムのバージョン、`--plot` は結果の可視化を意味している。#2 で示した例では、`--source` オプションを用いることで、この場合は脳から抽出された RNA に限定したサブセットを生成し、精度の検証を可能にしている。このサブセットの生成は、組織のほかにもセルラインを限定することが可能である。デフォルトでは、この機能は無効になっており、DARNED の全ての編集サイトを対象に、精度検証の指標は計算される。

3.3 本手法の性能評価

3.3.1 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それぞれ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、第二章で示したように、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 3.2: 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 つである。GSM ID 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。

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。

3.3.2 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。

(1) Ivy

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。

(2) REDItools

REDItools は最新のバージョン 1.0.3 を使用し、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、SNP およびスプライスサイトのアノテーション情報を除いたパラメータとなっている。これは Ivy ならびに SAMtools との解析に用いるパラメータを均一化するためである。

(3) SAMtools

SAMtools は、バージョン 1.0.19 を使用し、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。

3.4　性能評価の結果

3.4.1　計算機上でのパフォーマンス

比較に用いた 3 つのソフトウェアをそれぞれ実行し、解析に要したメモリの使用量および計算時間に関する計測を行った。

解析に要したメモリ容量を測定した結果を図 3.2 に示した。3 つのソフトウェアで比較を行ったところ、Ivy と REDItools は共に 20MB 程度の低メモリで動作することが示された。対して、SAMtools では 173MB のメモリを要求することが示された。

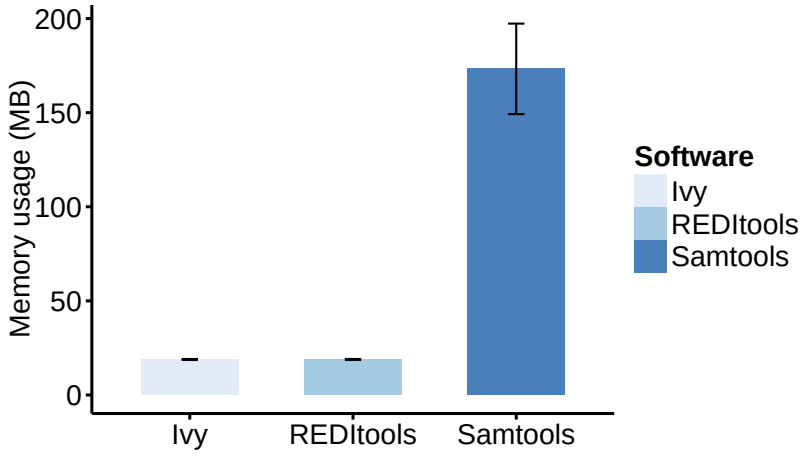

図 3.2: メモリの使用効率のベンチマーク

実行時におけるメモリ使用率 (MB) をそれぞれ Ivy、REDItools、SAMtools で比較した。エラーバーは、8 回のベンチマークを取った際の標準誤差を示す。

Ivy と REDItools、SAMtools について、表 3.2 のアラインメントデータを染色体ごとに分割した後に実行し、アラインメントデータのサイズに対する計算時間の結果を図 3.3 に示した。計測結果から、REDItools はゆらぎが見られるものの、基本的にデータサイズに対して線形に計算時間を要することが示された。REDItools との比較においては、Ivy はアラインメントデータサイズに依存せず、1.3 倍から 2 倍程度、高速に計算可能であることが示された。特に、75MB 以上のアラインメントデータに対しては、1.5 倍程度高速に動作しているため、絶対的な計算時間の短縮が見込まれた。

3.4.2　検出精度の検証

表 3.2 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 3.4 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、

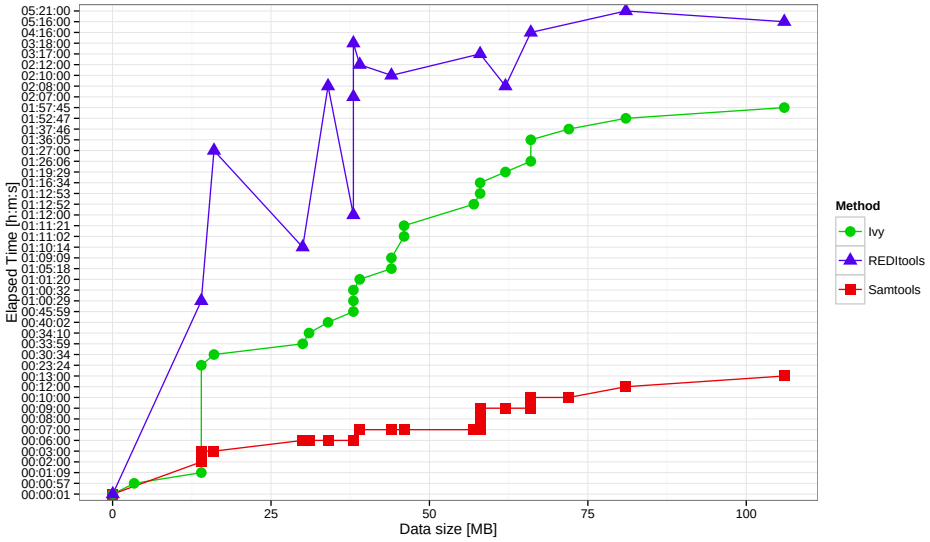

図 3.3: アラインメントデータのサイトと計算時間

サイズが異なるアラインメントファイルに対して、要した計算時間を縦軸にプロットした結果を示す。赤が SAMtools、緑が Ivy、青が REDItools それぞれ表す。データサイズが増加するにつれて、3 つのソフトウェアの計算時間は大きく異なる。

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。

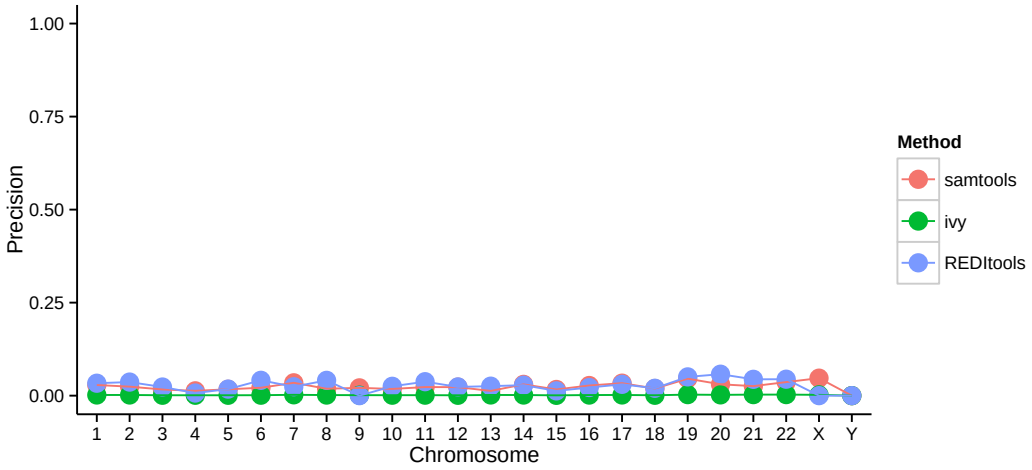

図 3.4: 染色体ごとの適合率の比較結果

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。尚、Y 染色体において検出された編集サイトは 0 である。

検出精度を再現率によって評価した結果を図 3.5 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。

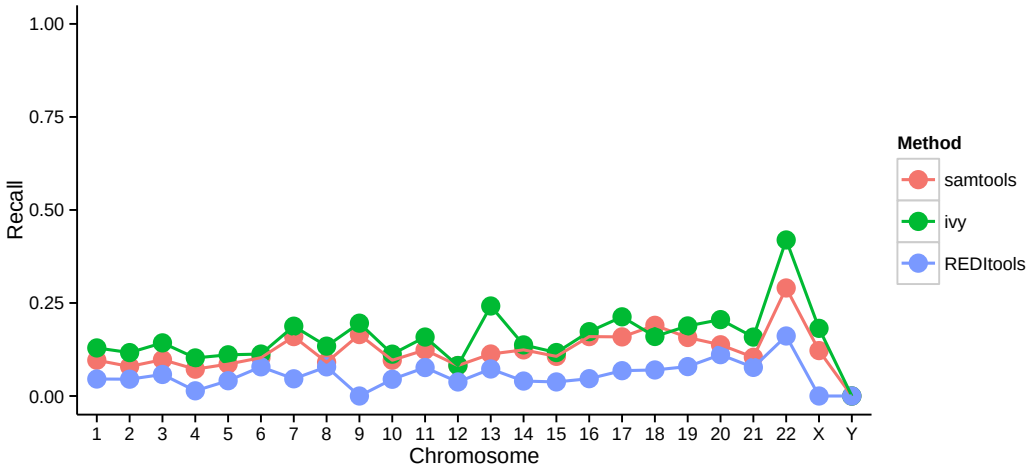

図 3.5: 染色体ごとの再現率の比較結果

染色体ごとの再現率を精度検証に用いた 3 つのソフトウェアで比較した結果を示す。尚、Y 染色体において検出された編集サイトは 0 である。

3.5 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 は、グリア芽細胞腫由来の RNA-seq データを用いて、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 (図 3.4 および図 3.5)。結果より、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すことが明らかとなった。

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。REDITools および SAMtools では検出されなかった一方、Ivy によって検出されたアンチセンス鎖における A-to-I 編集サイトを可視化した例を図 S.1 に示した。

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は必須だと考えており、次期バージョンのリリースで順次対応する予定で開発を進めている。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。こういった各種アラインメントデータへのフィルタリングの実装に加え、多様なセルラインや生物種の RNA-seq に対する検出精度を試験する必要性があると考えている。

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認識している。

第4章　クマムシにおけるRNA編集サイトの情報学的解析

4.1　研究背景

生命にとって生体内における水分子の損失は生存に深刻な影響を与える　(Cornette and Kikawada, 2011)。一方で、*Ramazzottius varieornatus* (和名: ヨコヅナクマムシ) を始めとする乾眠動物は、脱水により一時的に代謝の停止した乾眠状態となり、乾眠状態からは吸水によって生命活動を再開する (Tunnacliffe *et al.*, 2003)。このような可逆性を持つ乾眠状態は同時に、タンパク質の変性と凝集、核酸の損傷などを引き起こすため、多くの乾眠動物では遺伝子発現の様式を大規模に変化させ、保護物質としてのトレハロースやLEAタンパクを蓄積させている (Hengherr *et al.*, 2009; Welnicz *et al.*, 2011)。ところが、*R. varieornatus* は急速に乾眠状態へと移行するため、活動状態と乾眠状態では、遺伝子発現変動が僅かである特徴が明らかになってきた。このような急速に移行する *R. varieornatus* の乾眠は、定常的に発現している遺伝子群に加え、RNA編集を始めとする転写産物への修飾や、タンパク質のリン酸化修飾などによる制御が示唆されている。本研究では、RNA-seqデータを用いて *R. varieornatus* におけるRNA編集サイトのゲノムワイドな検出を行った。結果、*R. varieornatus* における全ての状態のトランスクリプトームから、他の真核生物と同様にADAR由来と考えられるA-to-I編集が優勢であることが示された。

本研究は、*R. varieornatus* におけるRNA編集と乾眠の分子機序との関係性について情報学的な解析を行ったものである。*R. varieornatus* より検出されたA-to-I編集サイトは、熱ショックタンパク質 (DnaJ) など分子シャペロンとして機能するタンパクのホモログから同定されており、*R. varieornatus* におけるRNA編集は、何らかの形で乾眠を支える機構への関与が考えられた。

4.2　対象と手法

4.2.1　解析データ

解析に用いたRNA-seqデータは、Illumina Genome Analyzer IIxによるリード長75bpのPaired-endシーケンシングされた *R. varieornatus* の活動状態 (Active)/乾眠状態 (Tun)/乾眠復帰後80分 (Recovery 80m)/復帰後240分 (Recovery 3h) の計4状態のデータを用いた。また、各状態におけるサンプル数はn=1である。ゲノムのシークエンスデータはIllumina GAIIxを用い、3つのレプリケーションでシーケンスされたデータを用いた。RNA-seqデータおよびDNA-seqデータは、

FastQC v0.10.0 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) を用いてリードの品質に問題がないことを確認した。RNA 編集サイトの検出には、各状態における RNA-seq データを用いた。DNA-seq リードは、ゲノムに見られる SNP をリストし、 編集サイトがトランスクリプトーム由来であることを保証するデータとして利用した。*R. varieornatus* のアノテーションは、最新の YOKOZUNA0703 を解析に使用した。

4.2.2 リードのマッピング

DNA-seq および RNA-seq のマッピングは、現在までに非常に多くのソフトウェアが開発されており、本研究でも様々なソフトウェアによるマッピング手法の検討を行った。結果、DNA-seq および RNA-seq リード共に TopHat v.2.0.2 (Trapnell *et al.*, 2009) が最適だと判断した。以下に計算に用いた詳細なパラメータについて述べる。

RNA-seq リードのマッピングにおいて、BWA や Bowtie といったアライナーは、ゲノムへのマッピングを前提として設計されていることから、転写物のリードに含まれるスプライスサイト周辺のリードは、ジャンクション領域への適切なマッピングが難しいことが指摘されている (Trapnell *et al.*, 2012)。これをふまえ RNA-seq リードにおいては、*R. varieornatus* ドラフトゲノムへ、TopHat v2.0.5 を用いた Splice junction mapping を実行した。計算に際してのオションについては、–read-realign-edit-dist 0 –max-multihits 1 –micro-exon-search –coverage-search –read-mismatches 2 –read-gap-length 0 –read-edit-dist 2 –b2-very-sensitive をそれぞれ設定した。リード内にはギャップを許さず、許容するミスマッチ数は 2 とした。また、micro-exon-search により短い領域のエクソンも検出できるようにした。加えて、ゲノム中にユニークにマップされたリードのみを用いた。マッピング後の処理として、INDEL (Insertion and Deletion) を含んだリードは、正確なマッピングが保証できないと考え、該当するリードは削除した。INDEL についての情報は、マッピングによって得られる BAM ファイルに記述されている。DNA-seq は、Illumina GAIIx を用いて 75bp の paired-end シークエンスされた 3 つのレプリケートをそれぞれ *R. varieornatus* ドラフトゲノムへマッピングした。パラメータは RNA-seq と同様に設定した。また、INDEL リードの削除も RNA-seq リードと同様に行った。

4.2.3 統計的フィルタリングを用いた RNA 編集サイトの検出

(1) RNA 編集サイトの検出

RNA 編集サイトの検出には、次世代シーケンサーを用いた変異解析で用いられる SAMools/bcftools (Li, 2011) を使用し、フィルタリングにはカスタムスクリプトを使用した。SAMtools/bcftools は、DNA-seq だけでなく RNA-seq へ適用し、RNA 編集サイトの同定に用いた研究も存在することから (Danecek *et al.*, 2012)、SNP 解析用のソフトフェアを使用することは妥当と考えた。本研究で

は、DNA-seq からも SNP の検出を行うことで、編集サイトと SNP の位置が一致した場合は除外するようにした。

(2) バイアスとその統計的フィルタリング

SAMtools/bcftools によって抽出したゲノム配列とのミスマッチポジションに対して、リードカバレッジが 20 未満のポジションは解析から除外し、ゲノム配列が N である場合はも同様に除外するようにした。このフィルタリングを通過した変異サイトについて、strand bias および base call bias の統計的フィルタリングを適用した。

(3) ADAR ホモログの予測

R. varieornatus において、ADAR ファミリーのホモログが存在しかつ、それぞれの 4 状態における転写物の発現量を確認することは重要である。4 状態の RNA-seq データから Cufflinks (Trapnell *et al.*, 2012) を用いて遺伝子発現量を定量した。加えて、ADAR が機能を有するためには、二本鎖 RNA を特異的に認識する DSRB (Double-strand RNA binding) ドメインや、ターゲットサイトとなるアデニンからグアニンへの置換を触媒する Deaminase ドメインが必須である。このことから *R. varieornatus* における ADAR ホモログ配列における、機能ドメインの予測を行った。予測には ADAR ファミリーのドメインから生成された隠れマルコフモデルを Pfam (Punta *et al.*, 2012) より取得し (PF00035.20/PF02137.13)、HMMER v.3.0 (Eddy, 2011) による予測を行った。このとき、E-value の閾値を 1e-20 とした。

4.2.4 検出手法の精度検証

(1) 適合率および再現率による精度検証

本研究で用いた RNA 編集サイトの同定手法を他の生物種に適用することにより、本手法の予測性能および再現性を評価した。評価には、第二章で説明した再現率および適合率を用いた。

(2) 正解となるデータセット

Rodriguez *et al.* (2012) による *D. melanogaster* (yellow white strain) の頭部から抽出された ZT14 (3 から 4 日の成熟個体) ステージにおける 1.3 億塩基の RNA-seq リード (Illumina GAII single-end) を公共データベースにより取得し解析を行った (GEO Accession No. GSE37232)。RNA-seq データのマッピングは論文に記載されたソフトウェアとパラメータを使用した。TopHat v.1.3 を使用し、-m 1 -F 0 –micro-exon-search –no-closure-search -G 20120309UCSC.gene.gtf –solexa-quals -I 50000 を実行した。遺伝子情報のアノテーションは、UCSC (University of California, Santa Cruz) の提供する 20120309UCSC.gene.gtf を用いた。遺伝子のアノテーションは、Illumina iGenome (`ftp://`

igenome:G3nom3s4u@ussd-ftp.illumina.com/Drosophila_melanogaster/UCSC/dm3/Drosophila_melanogaster_UCSC_dm3.tar.gz) から取得した。*D. melanogaster* における RNA 編集サイトの正解セットは、modENCODE プロジェクトによって同定された 1657 箇所 (Graveley *et al.*, 2011)、Rodriguez *et al.* (2012) で同定された 1969 箇所に共通していた 312 箇所を用いた。この 312 箇所を正解セットとして、G 検定、統計的なフィルタリングを行う前後の 3 つの予測セットを用意し、適合率と再現率による予測性能の評価を行った。

4.3 結果

4.3.1 RNA 編集サイトの検出

RNA-seq データを Tophat によるマッピングし、統計的なフィルタリングを行なっていない全ミスマッチパタンを図 4.1 に示す。4 状態のサンプルから検出されたミスマッチの総数は、54,809 箇所であった。各サンプルについて詳細を見ていくと、同定されたミスマッチパタンは、全てのサンプルにおいて T-to-C ミスマッチが 3000 サイト程度であり、最頻出のパタンであった。次いで A-to-C ミスマッチは 1000 サイト程度であり、A-to-G ミスマッチが優勢的に見られたサンプルはなかった。各サンプルにおけるミスマッチパタンの割合はよく類似しており、特定のサンプルのみで多く観察されるミスマッチパタンは見られなかった。このように、統計的フィルタリングを行う前の結果においては、ADAR による A-to-G 変異以外のミスマッチパタンが全サンプルにおいて例外なく見られる傾向であることが示された。このマッピング結果である図 4.1 に対して、各ポジションにおける最小カバレッジを 20、先述した 3 つのバイアスを統計的にフィルタリングすることにより、図 4.2 に示した結果を得た。図 4.2 が示すように、T-to-G ミスマッチや A-to-C ミスマッチは低く抑えられており、統計的フィルタリングによって 4 状態のトランスクリプトーム全てから合計 89 個の A-to-I 編集が検出され、最頻出のパタンであった。特に、T-to-G ミスマッチは、フィルタリングによって最も減少しているパタンであった。

4.3.2 変異サイトの特徴

(1) A-to-I 編集の割合

RNA 編集の見られた箇所はそのすべての転写物が ADAR による編集を受けるのではなく、*D. melanogaster* においては、10%〜20%にピークが見られることが明らかになっている。*R. varieornatus* における RNA 編集の割合を観察したところ、図 4.3 に示したように、10%付近にピークが観察される傾向が見られた。

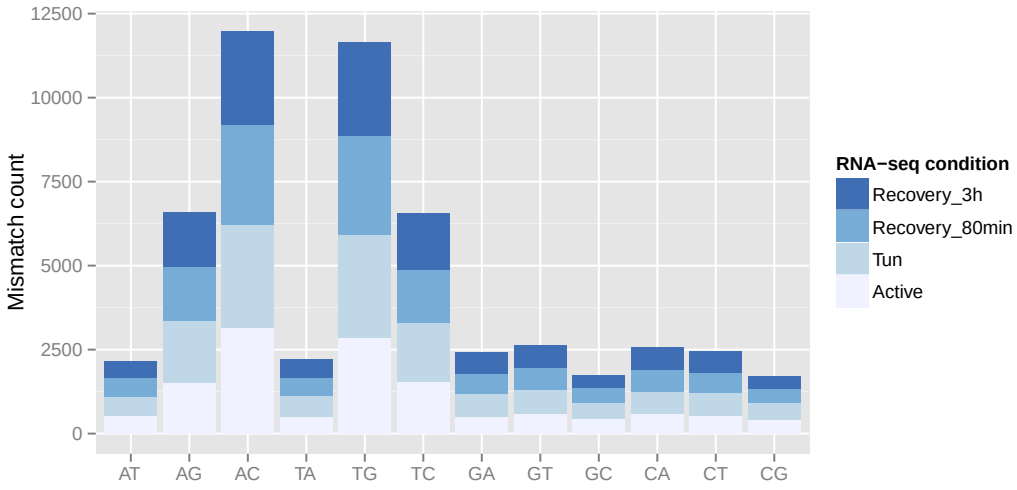

図 4.1: ゲノム配列へ RNA-seq データをマッピングした結果から直接的に得られた全ミスマッチパタン

Y 軸は、塩基置換の全 12 タイプのミスマッチを表しており、X 軸はそれに対応するミスマッチの観測された頻度を示す。用いた RNA-seq の 4 状態をそれぞれ色分けして表した。

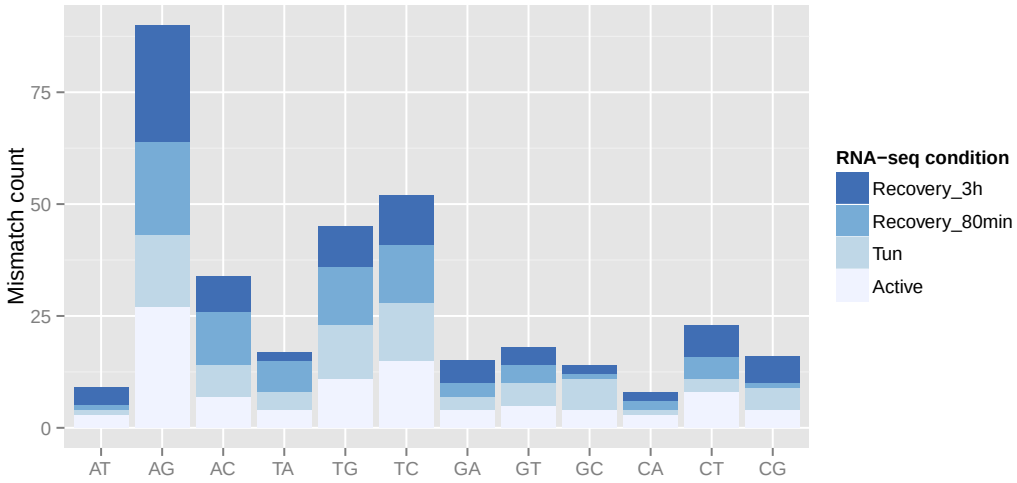

図 4.2: 統計的フィルタリングの適用後に観察された全ミスマッチパタン

統計的フィルタリングによって、A-to-G ミスマッチが最頻出であることが示された。図 4.1 で示した結果の 98.6%はフィルタリングされているため、頻度のスケールが異なることに留意。

(2) 遺伝子構造における分類

予測された A-to-I 編集サイトがゲノム構造中のどのような場所に特徴的に見られるのかを観察するため、A-to-I 編集サイトとコントロールにおいて、エクソン/イントロン/UTRs による分類を行った結果を図 4.4 に示す。各状態のサンプルとコントロールにおいて、遺伝子構造における分類

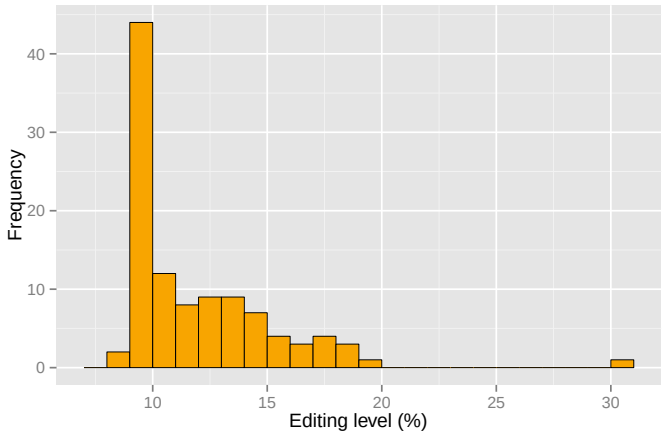

図 4.3: A-to-I 編集における編集率のヒストグラム

同定された A-to-I 編集サイトにおいて、編集されていた塩基の割合を編集 level として算出し、ヒストグラムで示した。編集 level は、各ポジションごとにおける、A および G 塩基の数をカウントし、$A_{count}/(A_{count}+G_{count})\times 100$ により算出した。

の比率に関連性が見られるかを評価するため、χ 二乗検定 ($p > 0.05$) による独立性の検定を行った。結果、両者の間にはエクソン/イントロン/UTRs の比率に有意差は見られないことが分かり ($p = 0.87$)、編集サイトがエクソンが多く位置し、イントロンや UTRs に少ないといった特徴は、編集サイトに特徴的ではないことが示された。

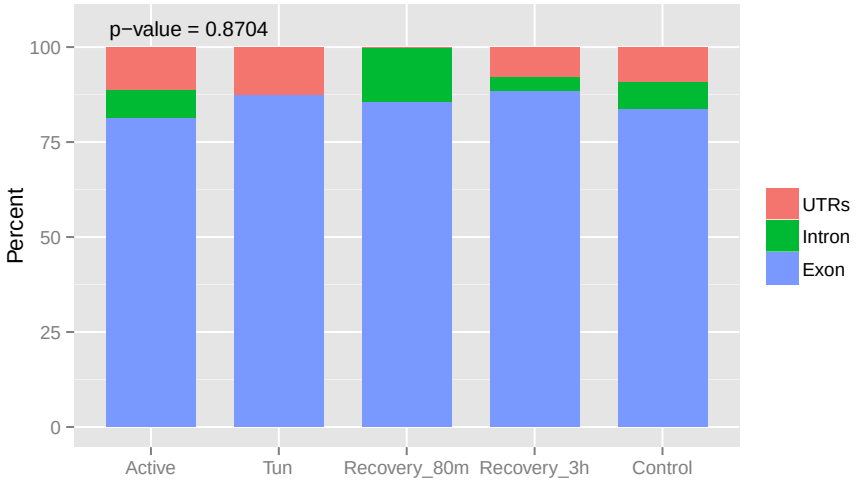

図 4.4: 遺伝子構造による分類

候補となる RNA 編集サイトが位置するエクソン、イントロン、UTRs における割合を示す。Y 軸に、それぞれ 4 状態のサンプルにおける割合を示した。A-to-G 以外の 11 タイプのミスマッチが見られたサイトをコントロールに用いた。

(3) RNA 編集サイトのスプライスサイトからの距離

ADAR による RNA 編集は、A から G への置換を引き起こすため、真核生物におけるエクソン-イントロン境界である Donnor/Acceptor サイトの AG-TG ルールを変化させる可能性がある。Donnor/Acceptor サイトが編集を受けることで、Intron retention や逆に新生スプライスサイトによって短いエクソンが転写される可能性が指摘されている。検出された A-to-I 編集サイトにおいて、近傍のスプライスサイトとの相対距離を求めた結果を図 4.5 に示す。スプライスサイトからの相対距離が編集サイト特有に見られる傾向かを検証するため、A-to-I 編集サイトと他の 11 のミスマッチパタンのサイトをコントロールとし、両者において統計的な有意差が見られるかを t 検定 ($p > 0.05$) によって検定した。t 検定に先立ち、コロモゴロフ・スミノフ検定による正規性ならびに等分散性の検定を行った。結果、A-to-I 編集サイトとコントロールサイトの間には有意差があることが示され ($p = 3.865e - 08$)、コントロールサイトに対して、A-to-I 編集サイトでは 5' 末端側へピークがシフトしている傾向が示された。ただ、これらの A-to-I 編集サイトから、スプライスサイトを変えるような RNA 編集サイトは同定されなかった。

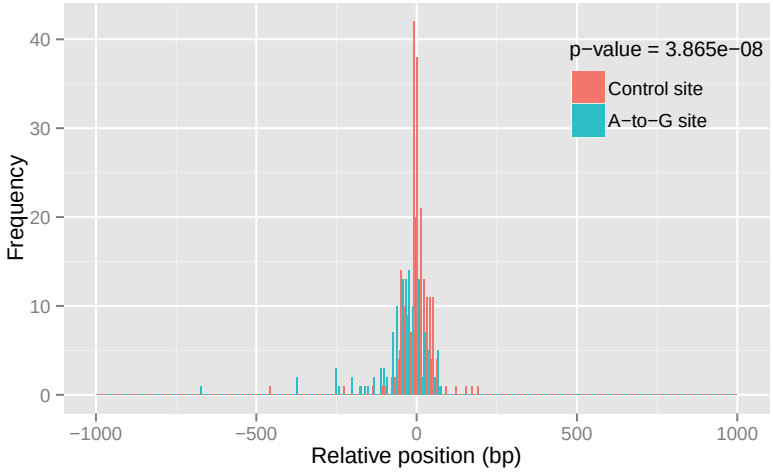

図 4.5: 検出された編集サイトのスプライスサイトからの相対的距離

X 軸は RNA 編集サイトの近傍に位置するスプライスサイトからの距離 (bp) を示している。相対距離の算出は、検出された RNA 編集サイトに最も近い Acceptor サイトまたは Donnor サイトからの相対距離を計算した。赤ラベルが A-to-I 編集以外のコントロールサイト、緑ラベルは A-to-I 編集サイトにそれぞれ対応する。

4.3.3 同定手法の精度検証結果

本研究による RNA 編集の同定手法の妥当性および予測性能を定量的に把握するため、適合率および再現率による同定手法の評価を行った。その結果を図 4.6 に示す。本手法を *D. melanogaster* yw strain ZT14 サンプルの RNA-seq データへ適用したところ、適合率は 3.7%、再現率は 18.7%で

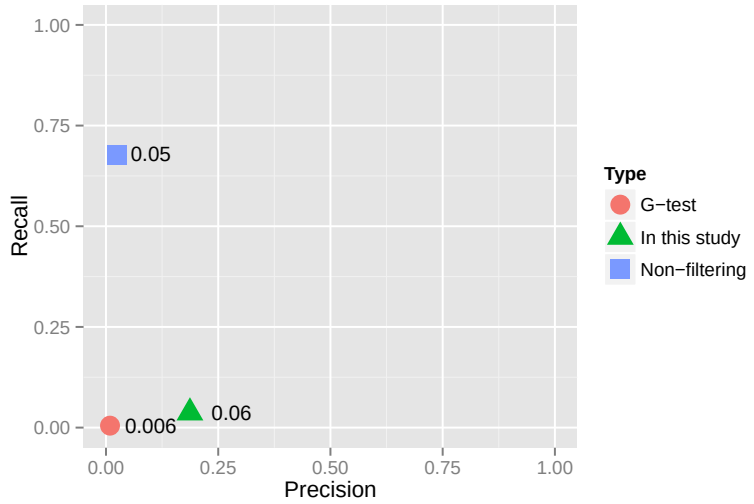

図 4.6: 適合率および再現率を用いた精度検証の結果

3 つの異なった予測手法における適合率と再現率を示す。赤は G 検定を用いた先学期の手法、緑は本研究による手法、青は本研究による統計的フィルタリングを行う前の結果にそれぞれ対応する。また、各点の横に算出した F 値を示した。

あることが示された。同様に、統計的フィルタリングを適用しなかった場合の結果は、適合率は 2.5%、再現率は 67.7%であった。G 検定を用いた結果は、適合率は 0.9%、再現率は 0.5%と最も低い予測性能を示していた。本研究による統計的フィルタリングを適用することにより、適合率は 12%向上し、再現率は 49%低下することが示された。本手法は、G 検定による検出と比較すると、予測性能は向上していることが示された。

4.3.4 同定された ADAR のホモログ

(1) ADAR ホモログの発現量

R. varieornatus における ADAR ホモログについて、既知の真核生物の ADAR のアミノ酸配列を用いた BLASTP による類似性検索から、ADAR1 と類似性の高い遺伝子が *R. varieornatus* においても見出された。ADAR には複数のファミリーが知られているが、*M. musculus* の ADAR1 との類似性が高いことが示された。加えて、RNA-seq を用いた発現量解析から、*R. varieornatus* における ADAR ホモログはトランスクリプトームの各状態において定常的に発現していることが確かめられた。また、各サンプルにおいては特異的な発現パタンは観察されなかった (図 4.7)。

(2) ADAR ホモログおいて予測された機能ドメイン

ADAR は、既知の報告によると、3 種類の機能ドメインから構成されていることが明らかになっていることから、HMMER によって Deaminase domain と Double-strand RNA binding (DSRB) ドメイン、Z-DNA binding domain を探索した。*R. varieornatus* のホモログから、Deaminase ドメイン

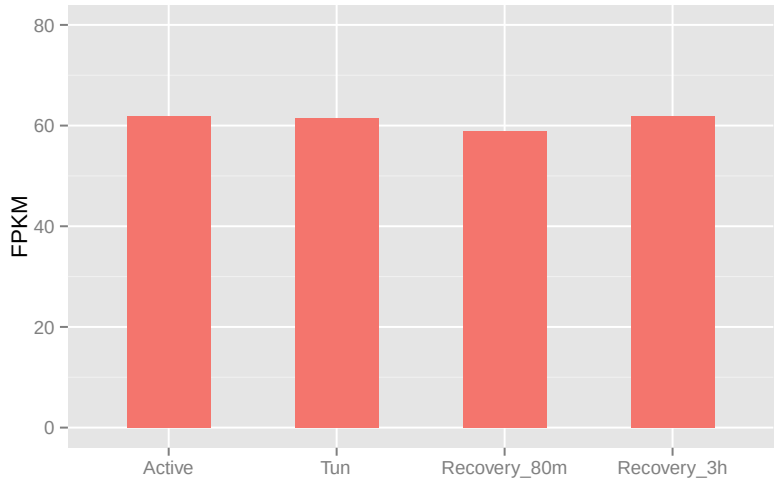

図 4.7: 各状態における ADAR ホモログの発現量

RNA-seq から定量した ADAR ホモログの発現量を示す。発現量の単位は FPKM (Fragment Per Kilobase per Million) であり、RNA-seq のリードカウントを遺伝子の長さで正規化した値である。

および DSRB ドメインがホモログ内に含まれていることが示された。対して、Z-DNA binding ドメインは類似性に有意差は見られなかった。

4.3.5 分子シャペロンに見られた RNA 編集サイト

DnaJ homolog subfamily C member 16 (DNAJC16) のエクソン内において RNA editing が検出されることが分かった。DnaJ タンパクは、Heat shock protein (HSP) ファミリーに属する分子シャペロンである。このタンパクは、熱ショックストレスによって会合や変性を受けたタンパクのリフォールディング機能を有しており、*R. varieornatus* を含めた乾眠動物では乾眠応答における主要なタンパクであることがこれまでに多く報告されている (Gusev *et al.*, 2011; Tyson *et al.*, 2012)。DnaJ への RNA 編集は、4 状態全てにおいて共通して起きていることも明らかになった。

4.4 議論

本研究は、*R. varieornatus* の 4 状態のトランスクリプトームデータからゲノムワイドな RNA 編集サイトの検出を行った。4 状態のトランスクリプトーム全てから、複数の転写物において RNA 編集による塩基置換が起きていることが示唆された。12 種類全てのミスマッチパタンは、図 4.1 で示したように、T-to-G や A-to-C ミスマッチが優勢的であったが、これらのミスマッチパタンは、Illumina のシーケンスエラーコレクションとして既に報告されている結果とも良く一致していたことから (Meacham *et al.*, 2011; Nakamura *et al.*, 2011)、擬陽性である可能性が高いと考えらた。この RNA-seq のマッピング結果へ、複数の統計的フィルタリングを適用することにより、図 4.2

で示すような、ADAR によると考えられる A-to-G ミスマッチが優勢を占めることが示された。この結果は、フィルタリング無しの結果で優勢的だった T-to-C ミスマッチや A-to-C ミスマッチは、フィルタリングによって低く抑えられることを示しており、本研究でマッピングしたデータへ適用したストランドバイアス、ベースコールバイアスによるフィルタリングによる擬陽性の減少が有効であることを示している。このように、初期のマッピング結果に対してフィルタリングを適用することによって初めて A-to-I 編集パタンが優勢を占める結果となる傾向は、これまでの報告とも一致するものだった (Alon *et al.*, 2012)。加えて、A-to-I 編集がフィルタリングによって優勢を占めるという結果は、ADAR による RNA 編集が確認されている *D. melanogaster* や *H. sapiens* など *R.varieornatus* 以外の真核生物からの報告とも一致するものであり、*R.varieornatus* においても ADAR による RNA 編集が主であることを示唆しているものと考えられる。

R. varieornatus における RNA 編集機構の存在を示唆する根拠は、ADAR ホモログにおける機能ドメインの探索からも示された。*R. varieornatus* の ADAR ホモログは、ADAR の機能ドメインとして重要な Deaminase ドメインならびに DSRB ドメインが有意に保存されていることが明らかとなり、相同性検索により見つかったホモログは、ADAR としての機能を有している可能性が高いことが示唆された。また、Z-DNA binding ドメインについては *R. varieornatus* は類似配列がなかったことも明らかとなった。このドメインは、*H. sapiens* の ADAR ファミリーから同定されたドメインであり、*Caenorhabditis elegans* や *D. melanogaster* からは発見されていないことを考慮すると、必須ドメインではないと推測された。加えて、*R. varieornatus* において ADAR ホモログが同定されたことは、ADAR ファミリー及び真核生物における RNA 編集の進化を研究する上での重要な示唆を提供するものと考えられる。

同定された RNA 編集サイトにおけるスプライスサイトからの相対距離を観察した結果 (図 4.5) からは、A-to-I 編集サイトは、コントロールサイトに対して 5' 末端側へピークがシフトしていることが示された ($p = 3.865e^{-08}$)。この特徴は、*R. varieornatus* における RNA 編集の特徴だと考えられた。また、ADAR がイントロン-エクソン境界において二本鎖 RNA となって二次構造を形成するとの報告 (Keegan *et al.*, 2005) からも、スプライスサイトの近傍に位置する編集サイトに関しては、イントロンを介した二次構造の形成のしやすさに関与している可能性が考えられた。同定された A-to-I 編集サイトを遺伝子構造によって分類した結果からは、検出された RNA 編集サイトとコントロールサイト間には統計的な有意差は見られず ($p = 0.87$)、編集サイトにエクソンが豊富に存在しているなどの特徴はなかった。この結果は、活動状態ならびに復帰後 3 時間のサンプル、コントロールにおけるイントロン/エクソン/UTRs の比率が類似していることから、RNA-seq に含まれる転写量を反映しているものと考えられた。このように、編集サイトが高頻度に観察される遺伝子構造上の場所と転写量には相関関係があると思われた。また、*H. sapiens* における A-to-I 編集の報告と類似しているものの (Ramaswami *et al.*, 2012)、*M. musculus* においては 3'UTR に編集サイトが偏在している報告 (Gu *et al.*, 2012) があることから、セルラインや生物種に因る部分が大きいと考えられた。本研究によって明らかにされた *R. varieornatus* における RNA 編集サイトのグローバルな特徴は、真核生物の RNA 編集に関する既知の報告と類似する点が多いことが明

らかになってきた。この結果は、マッピングデータへの統計的なフィルタリングの有効性ならびに *R.varieornatus* における RNA 編集も真核生物に近い制御を受けている可能性を示唆するものであった。

予測精度の検証結果は、図 4.6 に示したように、再現率は 3.7%、適合率は 18.7%であり、統計的フィルタリングの前後では、フィルタリングによって適合率が 12%向上していることが示された。フィルタリングを適用しない場合は非常に多くの擬陽性を含んだ編集サイトを予測するが、そのようなサイトが除外されることによって、適合率は向上したものと考えられる。対して、フィルタリングによって再現率が 49%低下することも明らかになった。RNA 編集サイトの検出は、多数の編集サイトの候補を出力するのではなく、母数は少なくとも確実な編集サイトを検出することが重要である。加えて再現率と適合率はトレードオフ関係にあるため、再現率が低下していることに関しては、本手法は問題がないと考えられる。また、僅かではあるが、フィルタリングによって F 値が増加することも示された。一方で、適合率は先学期の結果から向上はしているものの、依然として 20%程度の既知の A-to-I 編集サイトの予測にとどまっていることから、予測精度の向上は今後の課題である。

検出された RNA 編集は、熱ショックタンパクである DnaJ タンパクのエクソン中に起こるサイトが含まれていることが明らかになった。DnaJ ファミリーは、乾眠応答の際にタンパクの保護やリフォールディングを行うことがこれまでに確認されており、*Milnesium tardigradum* (オニクマムシ) からも報告がある (Frank *et al.*, 2011)。*R. varieornatus* においても、DnaJ タンパクは乾眠応答の際のタンパク保護などに関わっていると考えられ、DnaJ への編集は、その活性の制御などの可能性が考えられた。

第5章　結論

本論文は、超並列シーケンサーと呼ばれる高出力の塩基配列決定技術によって得られた大規模な塩基配列データから、RNA 編集サイトの検出ソフトウェアの開発ならびに、情報学的な解析に関する内容を扱った。各章での議論を簡単にまとめた後、超並列シーケンサーデータを用いることによって解くべき RNA 編集に関する問題を議論する。

二章では、既存の RNA 編集サイトの検出手法を統一的な指標を導入することにより、異なる検出手法の比較を可能にした。異なる検出手法の比較から、検出されたサイト周辺のリアラインメントや反復配列の除外など情報学的なフィルタリング手法に加えて、Strand specific RNA-seq や *Adar* のノックダウン株のシーケンシグなど実験デザインの重要性が示唆された。この成果は、より精度の高い編集サイトの検出に貢献することが期待された。

三章では、多くの研究では提案手法の実装がない状況において、超並列シーケンスデータを用いた RNA 編集サイトの検出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 編集サイトの検出ツールおよび検出精度を評価することのできるベンチマークツールが同梱されたソフトウェア・パッケージである。公共データベースにおいて公開されている RNA-seq データを用いて他のソフトウェアとの性能比較を行ったところ、Ivy は既存の RNA 編集サイトの検出ソフトウェアと比較して、同等のメモリ効率ながら 2 倍程度高速に動作することが示された他、高い再現率を示すことが明らかとなった。

四章では、ヨコヅナクマムシと呼ばれる乾眠動物の RNA-seq データを用いて RNA 編集サイトの検出を行った。非モデル生物であるため、完全にアセンブルされたゲノムがないといった解析の難しさがあったが、ADAR を発現量の定量および機能ドメインレベルで機能推定を行った結果、ヨコヅナクマムシにおいても A-to-I 編集は起こりうる可能性を示すことができた。検出された RNA 編集サイトの中には、熱ショックからのタンパクの保護に関わる因子への編集が観察されたことから、今後は実験的な検証と共に、RNA 編集と乾眠の関係性についての解析が進められることが期待された。

今日、超並列シーケンサーは年々高性能化しており、一度のランで得られる総リード数は、ムーアの法則を凌駕している。今後、RNA-seq データのエラー率は減少すると同時により高いカバレッジのデータを取得可能となることは容易に想像できる。このことは、RNA 編集サイトにおける編集の頻度を定量的に議論することを可能にすると考えられる。

RNA-seq データから十分な精度で編集率を推定することが可能になると、*Adar* の発現量と編集率の関係性が明らかとなるだろう。一部の遺伝子に関しては *Adar* の発現量と編集率の間には一貫

した相関性が見られないことが報告されており (Dominissini *et al.*, 2011)、単純に細胞内における ADAR の発現量と二本鎖 RNA との結合確率の上昇が編集率の増加へ寄与しないことは興味深く、ADAR への他の制御機構が機能している可能性がある。他にも、Alu などレトロトランスポゾンにおける RNA 編集の位置と頻度がどの程度制御されているのかは、内在性 siRNA の産生にも関与することから興味深い。レトロトランスポゾン領域におけるランダムに起きているように見える A-to-I 編集サイトに関しても、ADAR の発現量 (≒細胞内濃度) と編集率の関係を明らかにすることで、背後に何らかの制御を見出すことが可能だろう。

超並列シーケンサーから得られる RNA 編集サイトに関する定量的なデータは、ADAR の新たな機能と制御機序に関する問題を解く手がかりとなるだろう。本論文において議論してきた高精度かつ高速な RNA 編集サイトの検出ソフトウェアの開発が、RNA 編集の生物学的意味と ADAR を通した転写の調節機序の更なる解明へ貢献できることを期待したい。

謝辞

冨田研究室に配属されてからのおよそ 3 年半、研究室での会話や進捗のミーティング、学会発表など様々な場面で、厳しく有用な指摘があり、多くの勇気づけられる助言があった。本研究を行うにあたり、関わって頂いた多くの人に心からお礼を申し上げたい。

慶應義塾大学 政策・メディア研究科 荒川和晴講師には、学部 1 年秋学期より一貫して研究に関する指導をして頂きました。基本的に楽をして解析を済ませようとする怠惰な僕に対し、真っ当な解析方法を提示される荒川さんの指摘は、研究の質を少しずつ上げてくれるものでした。加えて、毎日の研究をいかに進めていくかというマネージメントなど、研究の方法論についても多くの気付きをもたらしたかけがえの無い経験でした。またスランプに陥った時も、なんとか次の研究につながるアドバイスを頂きました。深く感謝致します。

高校二年次に、SBP (Super bioscience program) に参加したきっかけが冨田勝教授との出会いでした。冨田さんとの出会いがなければ、SFC へ進学していなかったと思います。間違いなく人生における重要な出会いでした。所属した学部 2 年生までは悲しい程にプログラムが書けなくてこの先研究ができるのか心配なくらいでしたが、続けていくうち次第に自分の得意と言えるスキルへ変化していった感触があります。当時、卒業論文でバイオインフォマティクスにおけるソフトウェア開発を行うことは全く想像できませんした。こうった研究領域への関心が変化していくことに寛容な雰囲気がある冨田研で研究できたことは幸運でした。非常に恵まれた研究の機会と環境を与えて頂いた冨田さんにお礼申し上げます。同時に、秘書の見上さん、水上さん、平本さんには大学院の出願や学会発表、鶴岡での春・夏プロジェクトにおいて、大変にお世話となりました。感謝致します。

基本的に研究は三歩コマが進んだら二歩、悪いと四歩くらい後退するものでありましたが、それでもなんとかやってこられたのは同じような境遇の同期が周りにいたからだと思います。研究会の同期には公私（多くは不可分だが）問わず助けられました。土岐珠未氏や真流玄武氏とは正月を返上して昼夜問わず研究したし、今井淳之介氏にも世話になった。鶴岡に行ってしまった梅田栄美氏や臼井優希氏にも大きな感謝を表したいです。早期卒業した川崎翠氏にも遥かロンドンに向かって大きな感謝を送りたい。こういった多彩な顔ぶれの同期がいなかったら、研究生活はもっとつまらないものだったと想像します。本当にありがとう。

所属する G-language グループの皆様には本当に感謝しています。吉田勇太氏、そして特にグループの 5 人の後輩たちは、いつも後輩らしくないでかい態度で接してくれましたから、僕にとっても同期が増えたようで、なかなかよい関係を築くことができました。後輩たちとの分け隔てない関係は、楽しく気の休まるものでした。大下和希氏には、卒業後も付き合いがあるほど本当に

よくしていただきました。実装や解析に詰まった数えきれない回数、多くの時間を割いて一緒に解決して頂き、本当に感謝しています。

野崎慎氏には、研究に関する様々なアドバイスを頂きました。RNA 編集の研究は泥沼やでと最初に脅されましたが（実際そうでしたが）、解析に関するアドバイスや議論は常に本質的で有益でした。また、香川県に行くにあたり、はしごすべき讃岐うどんの名店を丁寧に教えて頂きました。どの店も最高の讃岐うどんでした。新土優樹氏からは研究に対する姿勢や研究の進め方などを影ながら多くを参考にさせて頂きました。玉木聡志氏とは技術的な話をするのが僕はすごく楽しかったです。いつでも楽観的のようにみえる玉木さんの生き方には何度も救われるようでした。松井求氏には、鶴岡に行った時などに RNA 編集に関する議論などができ、解析に関する幾つかの重要な指摘をして頂いたほか、根気強く研究に打ち込む姿には強い感銘を受けました。

本卒業研究は、SFC の卒業プロジェクトの履修者を対象とした株式会社 GREE の副社長 山岸広太郎氏の寄付による第一回山岸学生支援プロジェクトに採択して頂き、研究開発および学会発表が可能となりました。この場をお借りして感謝を申し上げます。山岸学生支援プロジェクトでは、専門が全く異なる方へ自身の研究と意義を伝えることの困難さに直面することができた他に、採択された他の学生とその指導教官の先生と交流する機会にも恵まれた大変に貴重な機会でした。

2013 年夏に開催された E-Cell sprint 2013 では、理化学研究所 生命システム研究センターの高橋恒一博士、海津一成博士を始めとした細胞シミュレーションを専門とした方たちと夜通し議論し、コードを書くことができた経験は、僕にとって象徴的な出来事となりました。これまで塩基配列など実データありきの研究以外を経験してこなかった僕にとって、生命現象すらも高度に抽象化されたプログラミングによって表現できるのだという驚きは今も鮮明に記憶しています。shafi さんには個人的にグルメ情報や進路について多くの助言を頂いた他、一杯のコーヒーのために丹沢まで 2 時間かけて名水を汲みに行ったりもしました。E-Cell sprint には学外からの学生も参加し、その後も仲良くできているのは本当に貴重です。このような機会を提供してくださった内藤泰宏准教授にも感謝致します。内藤さんには『細胞の物理生物学』の輪読会でお世話になった他、何気ない会話の中に大きな知の体系が見え隠れする瞬間があり、その度に強い感銘を受けました。

幸いにも数回の学会発表の機会に恵まれ、その度に学外の先生や学生の方とのよき出会いがありました。二度にわたって参加させて頂いた NGS 現場の会や分子生物学学会では、ポスター発表や懇親会において学外の多くの学生の方と知り合うことができました。彼らからは、同年代として多くの刺激や進学に関するアドバイスをもらい、研究の励みとなりました。また、山口大学 鈴木治夫博士には、学会などでお会いする度に優しい助言を頂きました。深謝致します。

三木研研究室 修士課程の小澤みゆき氏および武藤研究室 博士課程の中島博敬氏の両氏には、研究分野が離れているにも関わらず、非常に仲良くしていただき、研究を遂行するにあたって励みとなっていたことをここで表明させて頂きます。

最後になりましたが、大学生活を通して一貫して好き勝手にやらせてくれた両親と家族へ深く感謝し、卒業論文の締めとさせて頂きます。

参考文献

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*, **12**(2), R18.

Alon, S., Mor, E., Vigneault, F., Church, G. M., Locatelli, F., Galeano, F., Gallo, A., Shomron, N., and Eisenberg, E. (2012). Systematic identification of edited microRNAs in the human brain. *Genome Res.*, **22**(8), 1533–1540.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**(3), 403–410.

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.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, **226**(5252), 1209–11.

Barraud, P. and Allain, F. H.-T. (2012). Adar proteins: double-stranded rna and z-dna binding domains. *Curr Top Microbiol Immunol*, **353**, 35–60.

Bass, B. L. (2002). Rna editing by adenosine deaminases that act on rna. *Annu Rev Biochem*, **71**, 817–46.

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., Isaacs, F. J., Rechavi, G., Li, J. B., Eisenberg, E., and Levanon, E. Y. (2013). A-to-i rna editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res*.

Benne, R., Van den Burg, J., Brakenhoff, J. P., Sloof, P., Van Boom, J. H., and Tromp, M. C. (1986). Major transcript of the frameshifted coxii gene from trypanosome mitochondria contains four nucleotides that are not encoded in the dna. *Cell*, **46**(6), 819–26.

Cappione, A. J., French, B. L., and Skuse, G. R. (1997). A potential role for nf1 mrna editing in the pathogenesis of nf1 tumors. *Am J Hum Genet*, **60**(2), 305–12.

Carthew, R. W. and Sontheimer, E. J. (2009). Origins and mechanisms of mirnas and sirnas. *Cell*, **136**(4), 642–55.

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

Chen, C. X., Cho, D. S., Wang, Q., Lai, F., Carter, K. C., and Nishikura, K. (2000). A third member of the rna-specific adenosine deaminase gene family, adar3, contains both single- and double-stranded rna binding domains. *RNA*, **6**(5), 755–67.

Chen, L. (2013). Characterization and comparison of human nuclear and cytosolic editomes. *Proc Natl Acad Sci U S A*, **110**(29), E2741–7.

Cho, D.-S. C., Yang, W., Lee, J. T., Shiekhattar, R., Murray, J. M., and Nishikura, K. (2003). Requirement of dimerization for rna editing activity of adenosine deaminases acting on rna. *J Biol Chem*, **278**(19), 17093–102.

Chow, L. T., Gelinas, R. E., Broker, T. R., and Roberts, R. J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, **12**(1), 1–8.

Cornette, R. and Kikawada, T. (2011). The induction of anhydrobiosis in the sleeping chironomid: current status of our knowledge. *IUBMB Life*, **63**(6), 419–429.

Crick, F. *et al.* (1970). Central dogma of molecular biology. *Nature*, **227**(5258), 561–563.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and vcftools. *Bioinformatics*, **27**(15), 2156–8.

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.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

Desterro, J. M. P., Keegan, L. P., Lafarga, M., Berciano, M. T., O'Connell, M., and Carmo-Fonseca, M. (2003). Dynamic

association of rna-editing enzymes with the nucleolus. *J Cell Sci*, **116**(Pt 9), 1805–18.

Dillman, A. A., Hauser, D. N., Gibbs, J. R., Nalls, M. A., McCoy, M. K., Rudenko, I. N., Galter, D., and Cookson, M. R. (2013). mrna expression, splicing and editing in the embryonic and adult mouse cerebral cortex. *Nat Neurosci*, **16**(4), 499–506.

Dominissini, D., Moshitch-Moshkovitz, S., Amariglio, N., and Rechavi, G. (2011). Adenosine-to-inosine rna editing meets cancer. *Carcinogenesis*, **32**(11), 1569–77.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**(10), e1002195.

Edmonds, M. and Abrams, R. (1960). Polynucleotide biosynthesis: formation of a sequence of adenylate units from adenosine triphosphate by an enzyme from thymus nuclei. *J Biol Chem*, **235**, 1142–9.

Filipowicz, W. (2005). Rnai: the nuts and bolts of the risc machine. *Cell*, **122**(1), 17–20.

Flomen, R., Knight, J., Sham, P., Kerwin, R., and Makoff, A. (2004). Evidence that rna editing modulates splice site selection in the 5-ht2c receptor gene. *Nucleic Acids Res*, **32**(7), 2113–22.

Frank, F., Daniela, B., Marcus, F., Ralph, O. S., and Thomas, D. (2011). Bioinformatics identifies tardigrade molecular adaptations including the DNA-j family and first steps towards dynamical modelling. *J Zool Syst Evol Res*, **49**, 120–126.

Fukui, T. and Itoh, M. (2010). Rna editing in p transposable element read-through transcripts in drosophila melanogaster. *Genetica*, **138**(11-12), 1119–26.

Gallo, A., Keegan, L. P., Ring, G. M., and O'Connell, M. A. (2003). An adar that edits transcripts encoding ion channel subunits functions as a dimer. *EMBO J*, **22**(13), 3421–30.

Gan, Z., Zhao, L., Yang, L., Huang, P., Zhao, F., Li, W., and Liu, Y. (2006). Rna editing by adar2 is metabolically regulated in pancreatic islets and beta-cells. *J Biol Chem*, **281**(44), 33386–94.

Garrett, S. and Rosenthal, J. J. C. (2012). Rna editing underlies temperature adaptation in k+ channels from polar octopuses. *Science*, **335**(6070), 848–51.

Gerber, A. P. and Keller, W. (2001). Rna editing by base deamination: more enzymes, more targets, new mysteries. *Trends Biochem Sci*, **26**(6), 376–84.

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., Brown, J. B., Cherbas, L., Davis, C. A., Dobin, A., Li, R., Lin, W., Malone, J. H., Mattiuzzo, N. R., Miller, D., Sturgill, D., Tuch, B. B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R. E., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J. E., Wan, K. H., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P. J., Brenner, S. E., Brent, M. R., Cherbas, P., Gingeras, T. R., Hoskins, R. A., Kaufman, T. C., Oliver, B., and Celniker, S. E. (2011). The developmental transcriptome of Drosophila melanogaster. *Nature*, **471**(7339), 473–479.

Gu, T., Buaas, F. W., Simons, A. K., Ackert-Bicknell, C. L., Braun, R. E., and Hibbs, M. A. (2012). Canonical a-to-i and c-to-u rna editing is enriched at 3'utrs and microrna target sites in multiple mouse tissues. *PLoS One*, **7**(3), e33720.

Gusev, O., Cornette, R., Kikawada, T., and Okuda, T. (2011). Expression of heat shock protein-coding genes associated with anhydrobiosis in an African chironomid Polypedilum vanderplanki. *Cell Stress Chaperones*, **16**(1), 81–90.

Hamashima, K., Fujishima, K., Masuda, T., Sugahara, J., Tomita, M., and Kanai, A. (2012). Nematode-specific tRNAs that decode an alternative genetic code for leucine. *Nucleic Acids Res*, **40**(8), 3653–62.

Hang, P. N. T., Tohda, M., and Matsumoto, K. (2008). Developmental changes in expression and self-editing of adenosine deaminase type 2 pre-mrna and mrna in rat brain and cultured cortical neurons. *Neurosci Res*, **61**(4), 398–403.

Hartner, J. C., Walkley, C. R., Lu, J., and Orkin, S. H. (2009). Adar1 is essential for the maintenance of hematopoiesis and suppression of interferon signaling. *Nat Immunol*, **10**(1), 109–15.

Hengherr, S., Worland, M. R., Reuner, A., Brummer, F., and Schill, R. O. (2009). High-temperature tolerance in anhydrobiotic tardigrades is limited by glass transition. *Physiol. Biochem. Zool.*, **82**(6), 749–755.

Higuchi, M., Single, F. N., Köhler, M., Sommer, B., Sprengel, R., and Seeburg, P. H. (1993). RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. *Cell*, **75**(7), 1361–70.

Higuchi, M., Maas, S., Single, F. N., Hartner, J., Rozov, A., Burnashev, N., Feldmeyer, D., Sprengel, R., and Seeburg, P. H. (2000). Point mutation in an ampa receptor gene rescues lethality in mice deficient in the rna-editing enzyme adar2. *Nature*, **406**(6791), 78–81.

Jin, Y., Zhang, W., and Li, Q. (2009). Origins and evolution of adar-mediated rna editing. *IUBMB Life*, **61**(6), 572–8.

Kawakubo, K. and Samuel, C. E. (2000). Human rna-specific adenosine deaminase (adar1) gene specifies transcripts that initiate from a constitutively active alternative promoter. *Gene*, **258**(1-2), 165–72.

Keegan, L. P., Leroy, A., Sproul, D., and O'Connell, M. A. (2004). Adenosine deaminases acting on rna (adars): Rna-editing enzymes. *Genome Biol*, **5**(2), 209.

Keegan, L. P., Brindle, J., Gallo, A., Leroy, A., Reenan, R. A., and O'Connell, M. A. (2005). Tuning of RNA editing by ADAR is required in Drosophila. *EMBO J.*, **24**(12), 2183–2193.

Kiran, A. and Baranov, P. V. (2010). Darned: a database of rna editing in humans. *Bioinformatics*, **26**(14), 1772–6.

Kleinman, C. L. and Majewski, J. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, **335**(6074), 1302.

Kleinman, C. L., Adoue, V., and Majewski, J. (2012). Rna editing of protein sequences: a rare event in human transcriptomes. *RNA*, **18**(9), 1586–96.

Knight, S. W. and Bass, B. L. (2002). The role of rna editing by adars in rnai. *Mol Cell*, **10**(4), 809–17.

Krol, J., Loedige, I., and Filipowicz, W. (2010). The widespread regulation of microrna biogenesis, function and decay. *Nat Rev Genet*, **11**(9), 597–610.

Lagarrigue, S., Hormozdiari, F., Martin, L. J., Lecerf, F., Hasin, Y., Rau, C., Hagopian, R., Xiao, Y., Yan, J., Drake, T. A., Ghazalpour, A., Eskin, E., and Lusis, A. J. (2013). Limited rna editing in exons of mouse liver and adipose. *Genetics*, **193**(4), 1107–15.

Lai, F., Drakas, R., and Nishikura, K. (1995). Mutagenic analysis of double-stranded rna adenosine deaminase, a candidate enzyme for rna editing of glutamate-gated ion channel transcripts. *J Biol Chem*, **270**(29), 17098–105.

Laurencikiene, J., Källman, A. M., Fong, N., Bentley, D. L., and Ohman, M. (2006). Rna editing and alternative splicing: the importance of co-transcriptional coordination. *EMBO Rep*, **7**(3), 303–7.

Lee, J.-H., Ang, J. K., and Xiao, X. (2013). Analysis and design of rna sequencing experiments for identifying rna editing and other single-nucleotide variants. *RNA*, **19**(6), 725–32.

Lehmann, K. A. and Bass, B. L. (1999). The importance of internal loops within RNA substrates of ADAR1. *J Mol Biol*, **291**(1), 1–13.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**(21), 2987–2993.

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.

Li, M., Wang, I. X., Li, Y., Bruzel, A., Richards, A. L., Toung, J. M., and Cheung, V. G. (2011). Widespread RNA and DNA sequence differences in the human transcriptome. *Science*, **333**(6038), 53–58.

Lin, W., Piskol, R., Tan, M. H., and Li, J. B. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, **335**(6074), 1302.

Maas, S., Gerber, A. P., and Rich, A. (1999). Identification and characterization of a human trna-specific adenosine deaminase related to the adar family of pre-mrna editing enzymes. *Proc Natl Acad Sci U S A*, **96**(16), 8895–900.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, **18**(9), 1509–17.

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.

Meacham, F., Boffelli, D., Dhahbi, J., Martin, D. I., Singer, M., and Pachter, L. (2011). Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, **12**, 451.

Melcher, T., Maas, S., Herb, A., Sprengel, R., Higuchi, M., and Seeburg, P. H. (1996). Red2, a brain-specific member of the rna-specific adenosine deaminase family. *J Biol Chem*, **271**(50), 31795–8.

Meng, Y., Chen, D., Jin, Y., Mao, C., Wu, P., and Chen, M. (2010). Rna editing of nuclear transcripts in arabidopsis thaliana. *BMC Genomics*, **11 Suppl 4**, S12.

Miyamura, Y., Suzuki, T., Kono, M., Inagaki, K., Ito, S., Suzuki, N., and Tomita, Y. (2003). Mutations of the rna-specific adenosine deaminase gene (dsrad) are involved in dyschromatosis symmetrica hereditaria. *Am J Hum Genet*, **73**(3), 693–9.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**(7), 621–8.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M. C., Hirai, A., Takahashi, H., Altaf-Ul-Amin, M., Ogasawara, N., and Kanaya, S. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res*, **39**(13), e90.

Nishikura, K. (2006). Editor meets silencer: crosstalk between rna editing and rna interference. *Nat Rev Mol Cell Biol*, **7**(12), 919–31.

Nishikura, K. (2010). Functions and regulation of rna editing by adar deaminases. *Annu Rev Biochem*, **79**, 321–49.

Ota, H., Sakurai, M., Gupta, R., Valente, L., Wulff, B.-E., Ariyoshi, K., Iizasa, H., Davuluri, R. V., and Nishikura, K. (2013). Adar1 forms a complex with dicer to promote microrna processing and rna-induced gene silencing. *Cell*, **153**(3), 575–89.

Palladino, M. J., Keegan, L. P., O'Connell, M. A., and Reenan, R. A. (2000). A-to-i pre-mrna editing in drosophila is primarily involved in adult nervous system function and integrity. *Cell*, **102**(4), 437–49.

Park, E., Williams, B., Wold, B. J., and Mortazavi, A. (2012). Rna editing in the human encode rna-seq data. *Genome Res*, **22**(9), 1626–33.

Patterson, J. B. and Samuel, C. E. (1995). Expression and regulation by interferon of a double-stranded-rna-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. *Mol Cell Biol*, **15**(10), 5376–88.

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.

Pickrell, J. K., Gilad, Y., and Pritchard, J. K. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, **335**(6074), 1302.

Pinto, Y., Cohen, H. Y., and Levanon, E. Y. (2014). Mammalian conserved adar targets comprise only a small fragment of the human editosome. *Genome Biol*, **15**(1), R5.

Piskol, R., Peng, Z., Wang, J., and Li, J. B. (2013). Lack of evidence for existence of noncanonical rna editing. *Nat Biotechnol*, **31**(1), 19–20.

Pullirsch, D. and Jantsch, M. F. (2010). Proteome diversification by adenosine to inosine rna editing. *RNA Biol*, **7**(2), 205–12.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Res.*, **40**(Database issue), 290–301.

Qi, N., Zhang, L., Qiu, Y., Wang, Z., Si, J., Liu, Y., Xiang, X., Xie, J., Qin, C.-F., Zhou, X., and Hu, Y. (2012). Targeting of dicer-2 and rna by a viral rna silencing suppressor in drosophila cells. *J Virol*, **86**(10), 5763–73.

Ramaswami, G., Lin, W., Piskol, R., Tan, M. H., Davis, C., and Li, J. B. (2012). Accurate identification of human alu and non-alu rna editing sites. *Nat Methods*, **9**(6), 579–81.

Ramaswami, G., Zhang, R., Piskol, R., Keegan, L. P., Deng, P., O'Connell, M. A., and Li, J. B. (2013). Identifying rna editing sites using rna sequencing data alone. *Nat Methods*, **10**(2), 128–32.

Rodriguez, J., Menet, J. S., and Rosbash, M. (2012). Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila. *Mol Cell*, **47**(1), 27–37.

Sakurai, M., Ueda, H., Yano, T., Okada, S., Terajima, H., Mitsuyama, T., Toyoda, A., Fujiyama, A., Kawabata, H., and Suzuki, T. (2014). A biochemical landscape of a-to-i rna editing in the human brain transcriptome. *Genome Res*.

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.

Scadden, A. D. and Smith, C. W. (2001). Rnai is antagonized by a–>i hyper-editing. *EMBO Rep*, **2**(12), 1107–11.

Schrider, D. R., Gout, J. F., and Hahn, M. W. (2011). Very few RNA and DNA sequence differences in the human transcriptome. *PLoS ONE*, **6**(10), e25842.

Shikanai, T. (2006). Rna editing in plant organelles: machinery, physiological function and evolution. *Cell Mol Life Sci*, **63**(6), 698–708.

Slotkin, W. and Nishikura, K. (2013). Adenosine-to-inosine rna editing and human disease. *Genome Med*, **5**(11), 105.

Sorek, R., Lev-Maor, G., Reznik, M., Dagan, T., Belinky, F., Graur, D., and Ast, G. (2004). Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. *Mol Cell*, **14**(2), 221–31.

St Laurent, G., Tackett, M. R., Nechkin, S., Shtokalo, D., Antonets, D., Savva, Y. A., Maloney, R., Kapranov, P., Lawrence, C. E., and Reenan, R. A. (2013). Genome-wide analysis of a-to-i rna editing by single-molecule sequencing in drosophila. *Nat Struct Mol Biol*, **20**(11), 1333–9.

Temin, H. M. and Mizutani, S. (1970). Rna-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, **226**(5252), 1211–3.

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). Balibase 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**(1), 127–36.

Tonkin, L. A., Saccomanno, L., Morse, D. P., Brodigan, T., Krause, M., and Bass, B. L. (2002). Rna editing by adars is important for normal behavior in caenorhabditis elegans. *EMBO J*, **21**(22), 6025–35.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**(9), 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, **7**(3), 562–578.

Tunnacliffe, A., Lapinski, J., and van Leeuwenhoek, A. (2003). Resurrecting Van Leeuwenhoek's rotifers: a reappraisal of the role of disaccharides in anhydrobiosis. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **358**(1438), 1755–1771.

Tyson, T., O'Mahony Zamora, G., Wong, S., Skelton, M., Daly, B., Jones, J. T., Mulvihill, E. D., Elsworth, B., Phillips, M., Blaxter, M., and Burnell, A. M. (2012). A molecular analysis of desiccation tolerance mechanisms in the anhydrobiotic nematode Panagrolaimus superbus using expressed sequenced tags. *BMC Res Notes*, **5**, 68.

Valente, L. and Nishikura, K. (2005). Adar gene family and a-to-i rna editing: diverse roles in posttranscriptional gene regulation. *Prog Nucleic Acid Res Mol Biol*, **79**, 299–338.

Wang, I. X., So, E., Devlin, J. L., Zhao, Y., Wu, M., and Cheung, V. G. (2013). Adar regulates rna editing, transcript stability, and gene expression. *Cell Rep*, **5**(3), 849–60.

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.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., Surani, M. A., Sakaki, Y., and Sasaki, H. (2008). Endogenous sirnas from naturally formed dsrnas regulate transcripts in mouse oocytes. *Nature*, **453**(7194), 539–43.

Watson, J. D., Crick, F. H., *et al.* (1953). Molecular structure of nucleic acids. *Nature*, **171**(4356), 737–738.

Wei, C. M., Gershowitz, A., and Moss, B. (1975). Methylated nucleotides block 5' terminus of hela cell messenger RNA. *Cell*, **4**(4), 379–86.

Welnicz, W., Grohme, M. A., Kaczmarek, L., Schill, R. O., and Frohme, M. (2011). Anhydrobiosis in tardigrades–the last decade. *J. Insect Physiol.*, **57**(5), 577–583.

Wolf, J., Gerber, A. P., and Keller, W. (2002). tada, an essential trna-specific adenosine deaminase from escherichia coli. *EMBO J*, **21**(14), 3841–51.

Wulff, B.-E. and Nishikura, K. (2010). Substitutional a-to-i rna editing. *Wiley Interdiscip Rev RNA*, **1**(1), 90–101.

Yang, W., Wang, Q., Howell, K. L., Lee, J. T., Cho, D.-S. C., Murray, J. M., and Nishikura, K. (2005). Adar1 rna deaminase limits short interfering rna efficacy in mammalian cells. *J Biol Chem*, **280**(5), 3946–53.

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

Zhu, S., Xiang, J.-F., Chen, T., Chen, L.-L., and Yang, L. (2013). Prediction of constitutive a-to-i editing sites from human transcriptomes in the absence of genomic sequences. *BMC Genomics*, **14**(1), 206.

研究業績

ポスター発表

- Soh Ishiguro, Kazuharu Arakawa, Masaru Tomita. **"Benchmarking test for the detection methods of RNA editing sites based on RNA-seq data"**, MBSJ2013 (the 36th Annual Meeting of the Molecular Biology Society of Japan), Kobe, Japan (Dec. 3-6, 2013)
- Soh Ishiguro, Kazuharu Arakawa, Masaru Tomita. **"Benchmarking test for the detection methods of RNA editing sites based on RNA-seq data"**, NGS Field the 3rd Meeting, Kobe, Japan (Sep. 4-5, 2013)
- Yuki Yoshida, Soh Ishiguro, Kazuharu Arakawa, Takekazu Kunieda, Hirokazu Kuwabara, Daiki Horikawa, Atsushi Toyota, Toshiaki Katayama, Fujiyama Akisao, Masaru Tomita. **"RNA-Seq データ用いたヨコヅナクマムシにおける細胞防御・修復関連遺伝子の同定"**, NGS Field the 3rd Meeting, Kobe, Japan (Sep. 4-5, 2013)
- Soh Ishiguro, Kazuharu Arakawa, Takekazu Kunieda, Hirokazu Kuwahara, Daiki D. Horikawa, Atsushi Toyoda, Toshiaki Katayama, Asao Fujiyama, Masaru Tomita. (2012) **"Identification of RNA editing sites in tardigrade transcriptome from RNA-Seq data"**, MBSJ2012 (the 35th Annual Meeting of the Molecular Biology Society of Japan), Fukuoka, Japan (Dec. 2-7, 2012)
- Soh Ishiguro, Kazuharu Arakawa, Takekazu Kunieda, Hirokazu Kuwahara, Daiki D. Horikawa, Atsushi Toyoda, Toshiaki Katayama, Asao Fujiyama, Masaru Tomita. (2012) **"Identification of RNA editing sites in tardigrade transcriptome from RNA-Seq data"**, CBI/JSBi/Omix2012 (Joint Conference on Informatics in Biology, Medicine and Pharmacology), Tokyo, Japan (Oct. 14-17)
- Nobuaki Kono, Kazuharu Arakawa, Kazuki Oshita, Gembu Maryu, Soh Ishiguro, Taiyo Miyashita, Hidetoshi Itaya, Yuki Yoshida, Masaru Tomita. (2012) **"Web Application for Pathway Visualization with Instinctive Interface for Metagenome data"**, CBI/JSBi/Omix2012 (Joint Conference on Informatics in Biology, Medicine and Pharmacology), Tokyo, Japan (Oct. 14-17)
- 石黒宗, 荒川和晴, 冨田勝. (2012) **"次世代シーケンサーを用いたヨコヅナクマムシにおける RNA editing の解析"**, 第 2 回 NGS 現場の会, 大阪, 日本 (2012 年 9 月 4 日-6 日)

口頭発表

- 石黒宗, 荒川和晴, 冨田勝. (2013) **"ヨコヅナクマムシの乾眠メカニズムの研究"**, 慶應義塾大学 SFC Open Research Forum 2013, 東京, 日本 (2013 年 11 月 22-23 日)

付録

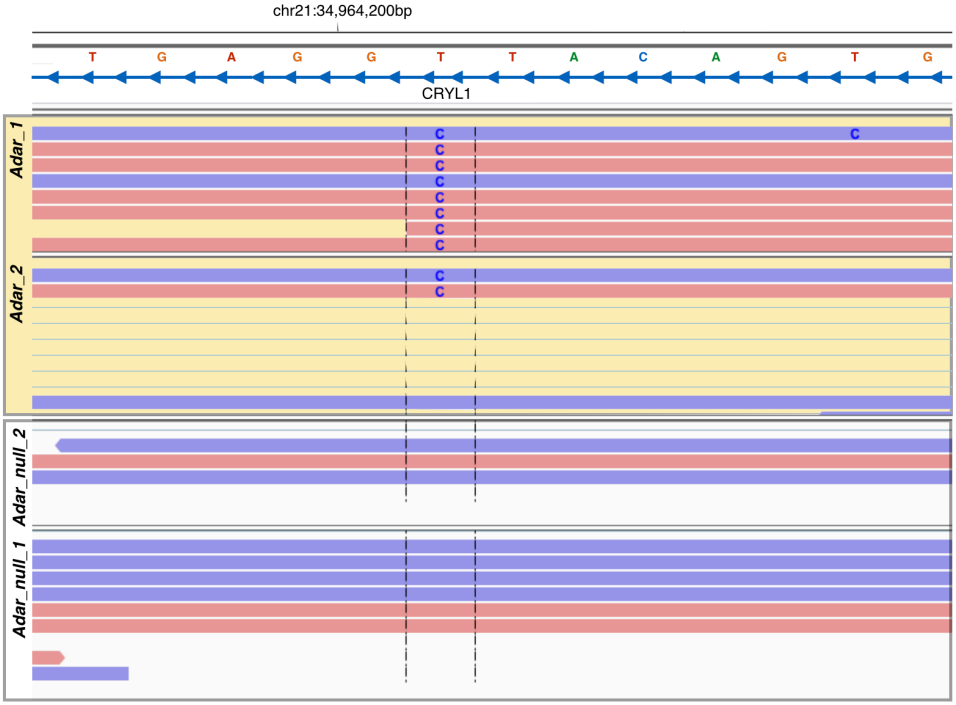

図 S.1: アンチセンス鎖における A-to-I 編集サイト

ivy によって検出された A-to-I 編集サイトを可視化した結果を示す。黄色くハイライトされたトラックは、*Adar* が発現しているサンプルを示し、白いトラックは siRNA によるノックダウン株の結果である。この CRYL1 遺伝子はアンチセンス鎖から発現しているため、A-to-I 編集サイトは、その逆鎖である T-to-C ミスマッチとして検出される。転写物の方向性を考慮しない場合、こういったサイトは A-to-I 編集サイトとして検出することは難しい。