ロングリードRNA-seqを用いた
融合遺伝子検出手法に関する研究

提出先　　大阪大学大学院情報科学研究科

提出年月　2024年1月

増田　圭吾

学位取得に関わる発表論文

学術論文

Keigo Masuda, Yoshiaki Sota, and Hideo Matsuda. Detecting Fusion Genes in Long-Read Transcriptome Sequencing Data with FUGAREC. IPSJ Transactions on Bioinformatics. （第 2 章、採録決定）

国際会議

Keigo Masuda, Yoshiaki Sota, and Hideo Matsuda. A Novel Method for Fusion Gene Detection using Both End-Fragment Sequences of Long Reads. In 2022 9th International Conference on Biomedical and Bioinformatics Engineering (ICBBE 2022), pp.88–93, March 15, 2023 （第 2 章）

Keigo Masuda, Yoshiaki Sota, and Hideo Matsuda.Accurate Detection of Fusion Genes in Long-Read Transcriptome Datasets from Multiple Cancer Cell Lines. In 2024 14th International Conference on Bioscience, Biochemistry and Bioinformatics (ICBBB 2024), January 12-15, 2024 （第 3 章、採録決定）

内容梗概

融合遺伝子は 2 つの正常な遺伝子が融合した異常な遺伝子である。融合遺伝子は腫瘍形成において原因的な役割を果たし、ヒトのがんの罹患原因の約 20% を占めると言われている。また、融合遺伝子は医薬品の標的となると同時に薬剤の効果を予想するためにも利用され、融合遺伝子を正確に検出する手法は、臨床現場や医薬品開発現場で求められている。

2000 年半ば以降、がん細胞のゲノムや転写産物の配列を解析することで、融合遺伝子を検出する方法が主流である。近年では、第三世代シークエンサーが開発され、個々の転写産物の全長の配列を読み取るロングリード RNA-Seq が可能となった。ロングリード RNA-Seq を用いた融合遺伝子の検出は、読み取られた塩基配列（リード）と参照ゲノムの相同な塩基配列を整列（アライメント）させ、2 つの異なる遺伝子に部分的にアライメントされるリードを特定することで行われる。しかし、ロングリード RNA-Seq で得られるリードは 1 塩基あたりの読み取り精度が低いという欠点がある。この欠点のため、アライメントされない領域（ギャップ）がリードの融合点近傍に発生し、融合遺伝子および融合点が正しく検出できないという問題点がある。加えて、先行手法はゲノム上での距離が近い遺伝子が融合する可能性を考慮に入れておらず、ゲノム上の距離が近い融合遺伝子を検出できないという問題がある。

本研究では、新規の融合遺伝子検出手法を提案することで、先行手法の問題点を解決し、新規の融合遺伝子の発見につなげることを目的とした。提案手法は先行手法に対して、(1) 融合点のエキソン境界固定、(2) ギャップの再アライメント、(3) 融合点のクラスタリングの 3 つの処理を加えて拡張した手法である。(1) と (2) は融合点をエキソンの境界に合わせて補正することで、融合点を正確に特定する狙いがあり、(3) はギャップによって発生する融合点のずれを吸収することで、偽陽性を防ぐ狙いがある。まず、シミュレーションデータを用いて提案手法の融合遺伝子の検出性能を評価した。その結果、提案手法は先行手法よりも高い融合遺伝子の検出性能を有していることを示した。また、前述の (1) と (2) により検出できる融合遺伝子が増加すること、(3) により偽陽性が減少することを示した。

次に、培養がん細胞のシークエンスデータから融合遺伝子を検出するために提案手法の拡張を行なった。その後、複数の培養がん細胞を用いて提案手法の融合遺伝子の検出性能を評価した。その結果、提案手法は先行手法よりも高い融合遺伝子の検出性能を有してい

ること、ゲノム上の距離が近い融合遺伝子も検出できることを示した。これにより、新規の融合遺伝子の候補を提示できることを示した。

目次

図目次

表目次

第 1 章　序論

1.1　本論文の背景

1.1.1　遺伝子および融合遺伝子の背景

生物の遺伝情報は DNA(デオキシリボ核酸: deoxyribonucleic acid) に保存されており、この遺伝情報全体をゲノムと呼ぶ。ゲノムは A(Adenine)、T(Thymine)、G(Guanine)、C(Cytosine) という 4 種類の塩基の配列で保存されている。各遺伝子が細胞内で機能するためには、DNA から RNA(リボ核酸: ribonucleic acid)、RNA からタンパク質へと変換されていく必要がある。DNA から RNA への変換は転写と呼ばれ、転写の過程でスプライシングを受けると、ゲノム上の遺伝子をコードしない領域であるイントロンが除去され、遺伝子をコードするエキソンのみがつなぎ合わされる。エキソンが繋ぎ合わさってできた転写産物の全体像をトランスクリプトームと呼び、細胞が属する組織の違いや、疾患に応じて、できる転写産物やその転写量は異なる。ゲノムやトランスクリプトームを分析することで、様々なレベルで生物の特徴を明らかにすることができる。

融合遺伝子は 2 つの正常な遺伝子が融合した異常な遺伝子である。融合遺伝子は、ゲノムの組み換えに依存する融合遺伝子とゲノムの組み換えに依存しない融合遺伝子が存在し、ゲノムの組み換え非依存的な融合遺伝子は、シススプライシングやトランススプライシングが原因で発生する (図 1.1)。シススプライシング融合遺伝子は、ゲノム上で隣接している 2 つの遺伝子が融合してできた融合遺伝子である。一方、トランススプライシング融合遺伝子は、染色体上で 2 つの遺伝子がコードされているゲノム領域どうしの空間的な距離が近いときに生じる融合であり、ゲノム上での距離も近い可能性がある [1]。ゲノムの組み換えにより生じる融合遺伝子もそうであるが、特にシススプライシング融合遺伝子やトランススプライシング融合遺伝子は近年になって注目されるようになったため、融合遺伝子の持つ機能については、ほとんど解明されていないものが多い [1]。また、上述のスプライシングによる融合遺伝子はがん細胞だけでなく、正常な組織の細胞からも検出されているが、どのような機能を果たしているかはまだ明らかとなっていない [1]。

1.1.2　融合遺伝子とがんの関連

最初の融合遺伝子は、1980 年代初頭の慢性骨髄性白血病患者で発見された [2]。現在までに、乳がん [3]、前立腺がん [4]、膀胱がん [5]、大腸がん [6]、卵巣がん [7] 、肺がん [8]、

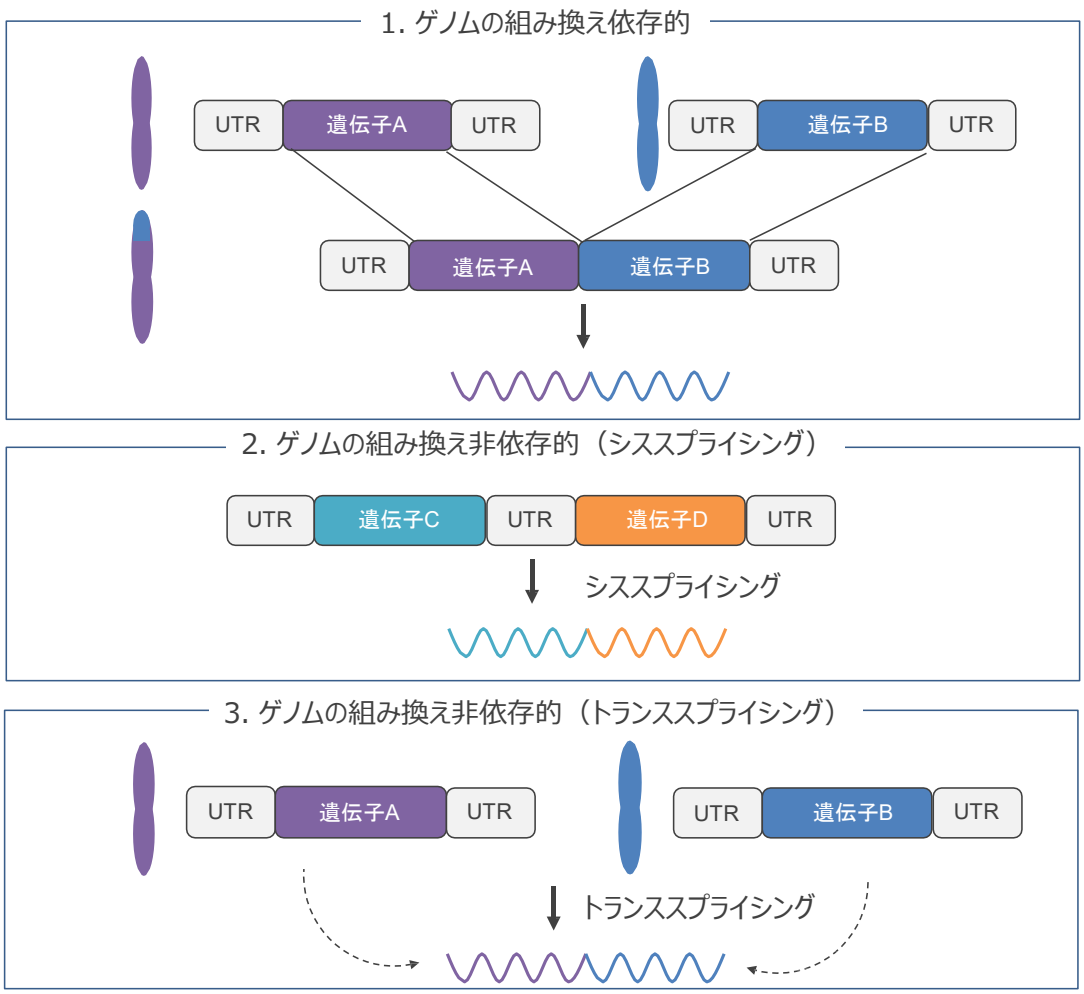

図 1.1　融合遺伝子が発生する模式図

滑膜肉腫 [9] など複数のがんの発症に融合遺伝子が関与していることが報告されている。また、融合遺伝子は腫瘍形成において原因的な役割を果たし、ヒトのがんの罹患原因の約 20% を占めているとも言われている [10]。融合遺伝子は、がん抑制因子やがん原遺伝子の発現を変化させたり、融合遺伝子がコードする融合タンパク質が正常なタンパク質の機能を改変することで、細胞のがん化を促進させたりすることが知られている。[11, 12, 13]。

いくつかの融合遺伝子については、分子標的薬が開発されており、臨床現場で効果を発揮している [14]。例えば、BCR 遺伝子と ABL1 遺伝子の融合遺伝子である BCR-ABL1 に由来する融合タンパク質を阻害するイマチニブメシレート [15] や、EML4 遺伝子と ALK 遺伝子の融合遺伝子である EML4-ALK に由来する融合タンパク質を阻害するクリゾチニブ [16] は、融合タンパク質の機能を阻害する薬剤として有名である。最近では、RET(rearranged during transfection）遺伝子が他の遺伝子と融合することで発症する、

表 1.1　一般的な融合遺伝子検出の手法

	RT-PCR 法	FISH 法	IHC 法	NGS 法
コスト	中	高	低	高
感度	高	中	中	高
特異度	高	高	低	高
未知の融合検出	不可	可	可	可
同時検査できる遺伝子数	1	1	1	複数

RET 融合遺伝子陽性非小細胞肺がんを対象とした分子標的薬であるセルペルカチニブが開発され [17] 、本邦では 2021 年 9 月に治療薬として適用を得ている。

また、融合遺伝子は医薬品の標的となるだけでなく、治療薬の効果を予測するためのバイオマーカーとしても利用されている [18]。例えば、薬物療法の治療方針を決めるため、肺がんでは ALK 融合遺伝子 [19]、慢性骨髄性白血病では BCR-ABL 融合遺伝子 [20] が利用されている。以上のような背景から、融合遺伝子は、がん治療に対して重要な情報を提供することができると考えられており、現在も活発に研究されている。

1.1.3　融合遺伝子の検出方法

がんの治療を成功させるためには、融合遺伝子を正しく検出することも重要である。代表的な融合遺伝子の検出方法として、RT-PCR（Reverse transcription polymerase chain reaction）法、蛍光 in situ ハイブリダイゼーション（Fluorescence in situ hybridization: FISH）法、免疫組織化学染色（Immunohistochemistory:IHC）法、次世代シークエンス (Next-Generation Sequencing: NGS) 法がある [21] (表 1.1)。4 つの手法はそれぞれ融合遺伝子の検出メカニズムが異なるため、感度（ある融合遺伝子 A が存在する状況下で融合遺伝子 A を検出できる能力)、特異度（ある融合遺伝子 A が存在しない状況下で融合遺伝子 A が存在しないと判断する能力)、未知の融合遺伝子検出の可否、同時調査できる遺伝子数に違いがある [22]。NGS 法は、NGS で読み取った塩基配列（リード）を分析して融合遺伝子を検出する方法である。NGS 法は、感度と特異度が高いだけでなく、1 度の実験で未知の融合遺伝子も含めて網羅的に検出可能である。そのため、現在は NGS 法を用いて融合遺伝子を検出することが主流になっている [23]。

NGS 法は、全ゲノムシークエンス（Whole Genome Sequencing :WGS) 解析、全エク

ソームシークエンス（Whole Exome Sequencing :WES) 解析、全トランスクリプトームシークエンス (RNA-Sequencing :RNA-Seq) 解析の 3 種類がある (表 1.2)。3 種類の手法はシークエンスする対象が違うことで、得られる情報量やコストが異なり、目的に応じて使い分けられる。全ゲノムを対象に大規模に融合遺伝子を検出したい場合は WGS、エキソン領域に絞って効率よく融合遺伝子を検出したい場合は WES、効率よく細胞で発現している融合遺伝子を検出したい場合は RNA-Seq が用いられる。

WGS は全ゲノムを対象に網羅的に解析するのに対し、WES はゲノム上のタンパク質をコードするエキソン領域のみを解析する手法である。WGS はタンパク質をコードしない領域であるイントロンも含めて融合遺伝子を検出できるという利点があるが [24]、大規模にシークエンスを行う必要があり、膨大な費用と計算コストが必要である [25]。WES は、エキソン領域のみを解析するため、低コストで融合遺伝子を検出できるという利点があるが、検出できる融合遺伝子に制限がある。また、WGS と WES の共通の問題点として、検出した融合遺伝子が機能しているかどうかを評価できない点が挙げられる。

RNA-Seq は、細胞中に存在する転写産物を網羅的に解析する手法である。RNA-Seq で検出した融合遺伝子は、実際に細胞内で発現していることが担保されている。そのため、ゲノムを対象とした手法より、がんに関与している可能性が高い融合遺伝子を検出することが可能である。このことから、現在では RNA-Seq を用いた融合遺伝子検出が広く行われている。

RNA-Seq による融合遺伝子の検出は、転写産物をシークエンスして得られた塩基配列（リード）と、参照ゲノムや参照トランスクリプトームを対応づけ、複数の遺伝子にまたがるリードを検出することで行われる。参照ゲノムとは、ゲノム解読プロジェクトなどで解読された大量のリードをアセンブルする（リード同士を繋げて対象の配列を復元する）ことで構築された標準的な参照配列である。参照トランスクリプトームは、参照ゲノムと同様に作成された細胞内の全転写産物の標準的な参照配列である。リードと参照ゲノムや参照トランスクリプトームを対応づけるためには、アライメントという複数の配列の中から共通部分を探し出し、お互いの配列を整列させる処理が必要である (図 1.2)。一般的に、参照ゲノムや参照トランスクリプトームのような参照配列（リファレンス）の方がアライメントしたいリード（クエリ）よりも塩基配列が長いため、完全に配列が一致することはない。リファレンスの塩基配列の中で、クエリの塩基配列が部分的に一致する配列（相同配列）が見つかり、部分的に整列されればアライメントされたとみなされる。

表 1.2　NGS を用いた融合遺伝子検出の特徴

	WGS	WES	RNA-Seq
解析対象	全ゲノム	全エキソン	全転写産物
コスト	大	小	小
融合遺伝子の転写確認	不可	不可	可

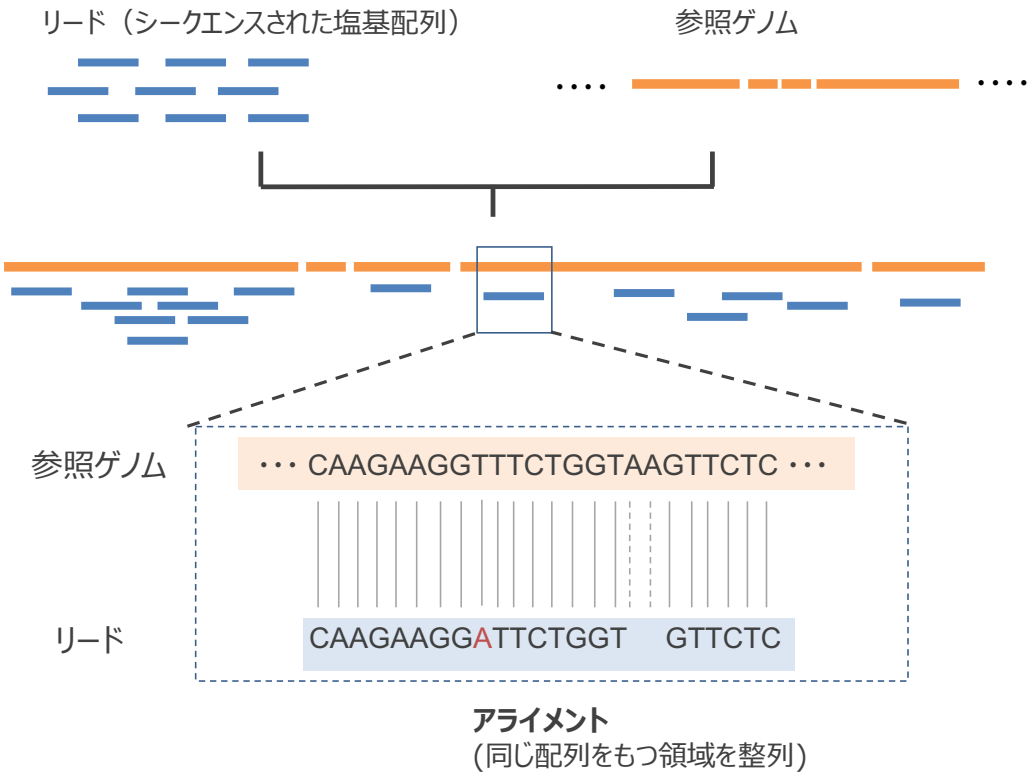

図 1.2　参照ゲノムへのリードのアライメントの模式図

1.1.4　先行手法の問題点

近年では、個々の転写産物の全長をシークエンスすること (ロングリード RNA-Seq) が可能となり、一般的なショートリード RNA-Seq では検出できない融合遺伝子の検出が期待されている。一方で、ロングリード RNA-Seq は 1 塩基あたりに約 7% の誤りを含むことが知られている [26]。そのため、ロングリード RNA-Seq から得られるリードを 1 塩基単位で正確に、参照ゲノムに対してアライメントすることは困難である。特に、2 つの異

なる遺伝子にアライメントされる融合遺伝子由来のリードにおいて、アライメント先の遺伝子が切り替わる位置（融合点）の近傍にシークエンスエラーが入ると、その部分のアライメントができずにギャップが発生する（図 1.3）。このギャップが問題となり、先行手法ではギャップを含むリードにおいて、融合遺伝子と融合点を正しく特定することができないという問題がある。融合遺伝子の機能を解析するためには、翻訳により得られるタンパク質のアミノ酸配列を正しく予測できる必要があるが、先行手法はギャップのせいで融合点が正確に求められないため、融合遺伝子の機能解析の妨げとなっていた。

また、先行手法は、ゲノム上での距離が近い遺伝子が融合する可能性を考慮に入れておらず、融合する 2 つの遺伝子間の距離が近いと、融合遺伝子の検出候補から除外するという処理をしている。そのため、先行手法では、シススプライシング融合遺伝子やトランススプライシング融合遺伝子のようなゲノム上での距離が近い遺伝子の融合遺伝子は検出できないという問題がある。

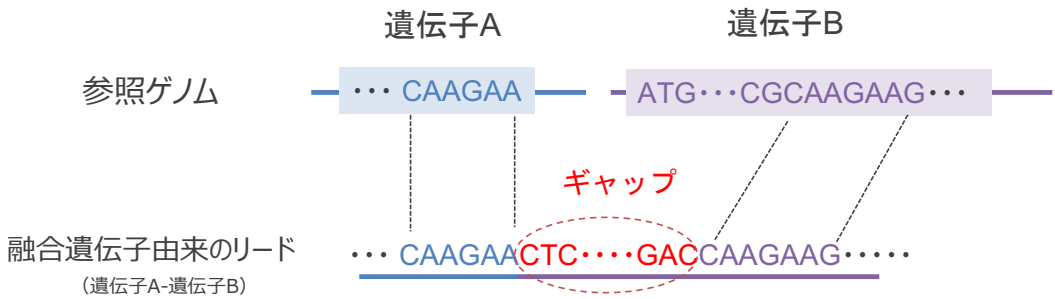

図 1.3　アライメント時に発生するギャップの模式図

1.2　本論文の目的

先行手法では、ギャップを持つリードから融合遺伝子と融合点を正しく検出できないという問題がある。加えて、シススプライシング融合遺伝子やトランススプライシング融合遺伝子のようなゲノム上での距離が近い遺伝子の融合遺伝子は検出できないという問題がある。本論文では、先行手法の問題を克服する手法を提案し、新規の融合遺伝子の発見につなげることを目的とする。また、融合遺伝子および融合点を正確に特定することで、がん細胞だけでなく、正常な組織の細胞で融合遺伝子がどのような役割を果たしているのかといった融合遺伝子の機能の解明に貢献することを目指す。

1.3　本論文の構成

本論文は 4 章から構成される。第 1 章では、本研究における背景と目的および本論文の全体構成について述べた。第 2 章では、先行手法の解決すべき問題を明示した上で、その問題を解決する新たな融合遺伝子検出アルゴリズムを提案する。次に、シミュレーションデータを用いて、提案手法の性能を先行手法と比較し、得られた結果をもとに考察を行う。第 3 章では、実際のがん細胞のシークエンスデータから融合遺伝子を検出するために提案手法の拡張を行なう。次に、複数の培養がん細胞のシークエンスデータを用いて、提案手法の性能を先行手法と比較し、得られた結果をもとに考察を行う。第 4 章では、本研究の結論と将来の展望について述べる。

第 2 章　融合遺伝子検出アルゴリズムの構築と評価

2.1　緒言

本章では、RNA-Seq を活用した融合遺伝子の検出手法を先行手法として紹介する。その後に先行手法の問題点を挙げ、その問題を解決するための新たな手法を提案する。シミュレーションデータを用いて、提案手法の性能を先行手法と比較し、得られた結果をもとに考察を行う。

2.2　融合遺伝子の検出と RNA-Seq

本節では、次世代シークエンス (Next-Generation Sequencing: NGS) を用いた融合遺伝子の検出方法について説明する。NGS を用いた融合遺伝子の中でも転写産物に着目した、RNA-Seq を用いた融合遺伝子の検出方法について説明し、融合遺伝子を検出する上で発生する問題について述べる。

2.2.1　ショートリード RNA-Seq を用いた融合遺伝子検出

ショートリード RNA-Seq は、DNA 配列を断片化して増幅後に配列を解読する。その特性上、ショートリード RNA-Seq で得られるリードは 150 塩基程度と短いが、得られる総リード数は多く、カバレッジ（参照ゲノムにアライメントした際、同じ領域にアライメントされるリード数）が大きいという特徴がある。これまでに、ショートリード RNA-Seq から得られるリードの特性を活かした、様々な融合遺伝子検出ツールが提案されている。採用されるアライメントツールや偽陽性の可能性が高い融合遺伝子をフィルタリングする基準などに違いはあるが、マッピングベースとアセンブルベースの 2 つに分けられる [27]。

マッピングベースのアプローチでは、ショートリードを参照ゲノムにマッピング（短い配列を参照ゲノムにアライメントする）し、異なる遺伝子にアライメントされるリードやキメラリード（リードの前半と後半で異なる染色体などにマッピングされるリード）を識別することで融合遺伝子を検出する。代表的なマッピングベースのツールとして、TopHat-Fusion[28]、STAR-Fusion[27]、Arriba[29] が有名である。マッピングベースのアプローチでは、アセンブルできない断片的なリードも活用し、融合を検出するため感度が高い傾向がある [27]。一方で、ショートリードのアライメントが困難なリピート領域

（ゲノム中の同じ配列が反復する領域）における融合遺伝子を特定できないという問題がある [30]。

アセンブルベースのアプローチでは、ショートリードをアセンブル（リードをつなぎ合わせて元の配列を再構築する）して転写産物を再構築する。その後、再構築した転写産物を参照ゲノムにアライメントさせ、キメラリードを検出することで、融合遺伝子を検出する (図 2.1)。代表的なアセンブルベースのツールとしては、JAFFA-Assembly[31] や TrinityFusion[27] がある。アセンブルベースの手法は、マッピングベースのアプローチでは困難であったリピート領域に対して、一定の効果が期待できる。その一方で、カバレッジが不足すると転写産物の再構築ができず、融合遺伝子が検出できなくなる欠点がある。アセンブルベースの手法は発現量が小さい融合遺伝子の検出に弱く、マッピングベースの手法と比較して感度が劣ることも報告されている [27]。最近では、JAFFA-Hybrid [31] のようなマッピングベースとアセンブルベースを組み合わせた手法も開発され、ショートリードの問題点を克服するべく現在も研究が進められている。

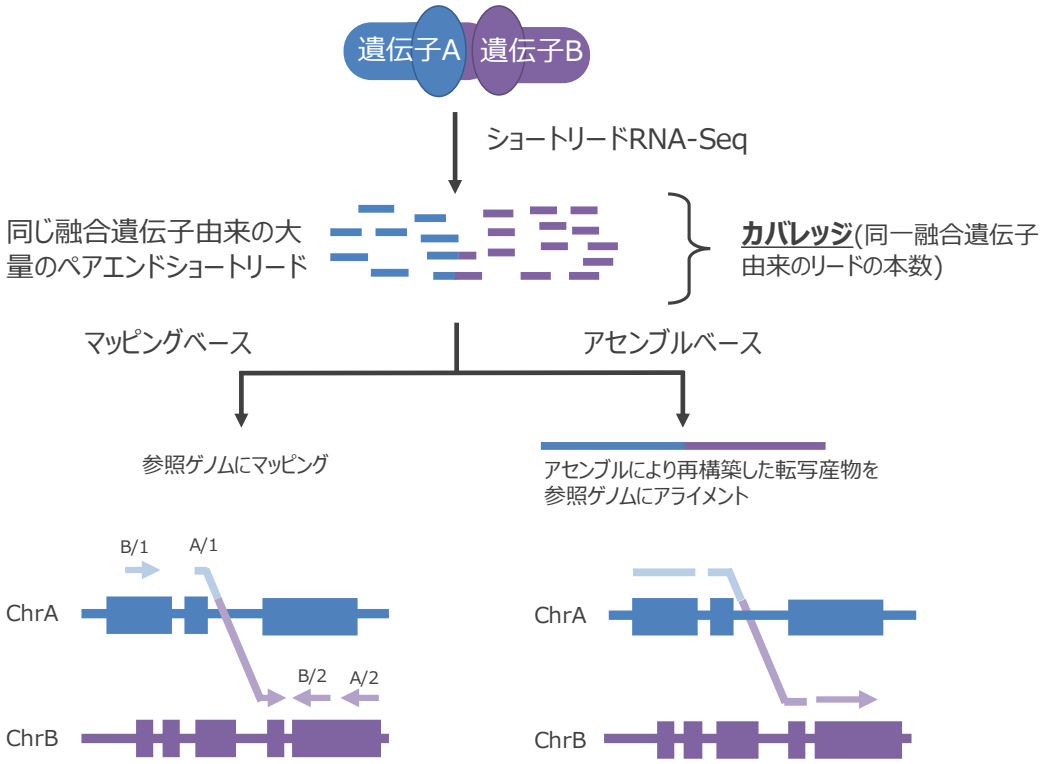

図 2.1　ショートリード RNA-Seq を用いた融合遺伝子検出の模式図

2.2.2 ロングリード RNA-Seq を用いた融合遺伝子検出

次世代シークエンサーの 150 塩基程度のショートリードしか得られないという欠点を補うように、第三世代シークエンス (Third Generation Sequencing、TGS) 技術が開発され、2017 年ころから急速に普及している。第三世代シークエンス技術を牽引するのが、Pacific Biosciences（PacBio）[32] と Oxford Nanopore Technologies (ONT)[33] である。

PacBio シークエンスは、検出感度が非常に高い蛍光色素を用いることで、試料を増幅することなく、1 分子をシークエンスする方法である。ONT シークエンスは、通電性があるポリマー製の膜に電圧をかけた状態で、膜に埋め込まれた膜タンパク質の微細孔（ナノポア）に DNA を流し、ヌクレオチド毎に変化する特徴的な電流の乱れを検知することで、塩基配列を決定するシークエンス方法である。PacBio と ONT でシークエンスの原理は異なるが、どちらも試料を増幅することなく、1 分子をリアルタイムにシークエンスすることが可能である。連続した 1 万塩基以上のロングリードが取得可能で、個々の転写産物は全長を 1 度にシークエンスすることが可能である。

ロングリード RNA-Seq は全長を 1 度にシークエンスできる一方で、ショートリード RNA-Seq と比較して 1 塩基あたりの読み取り精度（シークエンスアイデンティティ）が低いという問題がある (表 2.1)。PacBio シークエンスされたロングリードの 1 塩基あたりのエラー率は、11%〜15% [34] と報告されている。ONT シークエンスされたロングリードのエラー率は 1 塩基あたりのエラー率は 2018 年時点では 15% 程度 [35] であり、2021 年時点では 7% 程度 [26] だと報告されている。ロングリード RNA-Seq は、一度に得られる総リード数が少なく、カバレッジが小さい [30]。そのため、ロングリード RNA-Seq には、ショートリード RNA-Seq 用の融合遺伝子検出ツールを使用することができない。ロングリード RNA-Seq を用いて融合遺伝子を検出するためには、ロングリード RNA-Seq の特徴を考慮したアルゴリズムが必要である。

表 2.1　ショートリード RNA-Seq とロングリード RNA-Seq

	ショートリード RNA-Seq	ロングリード RNA-Seq
リード長	150 塩基前後	1 万塩基以上
シークエンスアイデンティティ	99.9% 以上	85%-90%
総リード数	大	小
カバレッジ	大	小

ロングリード RNA-Seq を用いた融合遺伝子検出では、ロングリードを参照ゲノムや参照トランスクリプトームにアライメントして、異なる遺伝子にアライメントされるキメラリードを検出する方法が一般的である (図 2.2)。現在までにロングリード RNA-Seq から融合遺伝子を検出するツールとして、LongGF [36]、 JAFFAL [37]、 FusionSeeker [38] が開発されてきた（表 2.2)。

表 2.2　ロングリード RNA-Seq を用いた融合遺伝子検出ツール

	LongGF	JAFFAL	FusionSeeker
論文公開	2020 年 12 月	2022 年 6 月	2023 年 4 月
アライメント対象	参照ゲノム	参照ゲノムと参照トランスクリプトーム	参照ゲノム
ギャップが大きいリード	除外	除外	対象
サポートリード数の下限	1 本以上	1 本以上	3 本以上

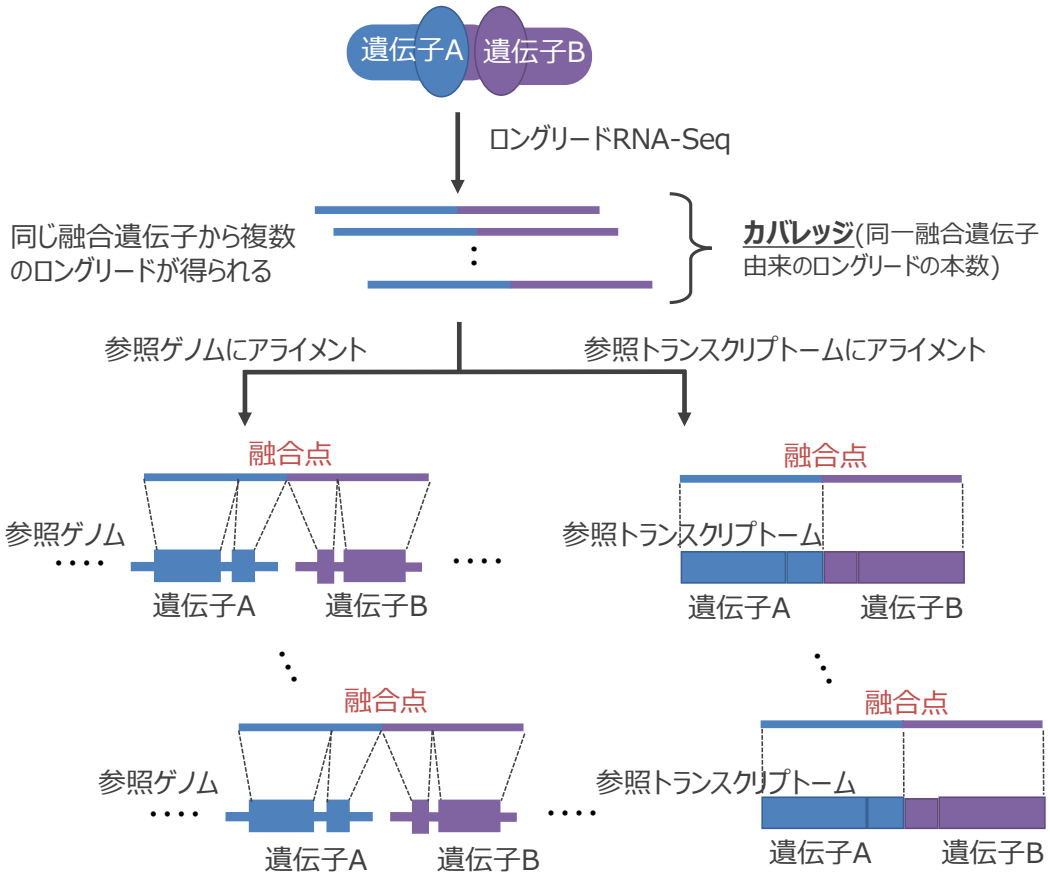

図 2.2　ロングリード RNA-Seq を用いた融合遺伝子検出の模式図

LongGF は、ロングリード RNA-Seq を用いた、融合遺伝子検出のための先駆的なツールである。ロングリードをゲノムにアライメントし、複数の遺伝子にアライメントされるリードを同定することで融合遺伝子を検出する。JAFFAL は、参照ゲノムと参照トランスクリプトームの両方にアライメントするなど、複数のフィルタリングを追加することで LongGF よりも低い偽陽性率で融合遺伝子を検出するツールである。FusionSeeker は、他の 2 手法とは異なり、遺伝子融合を検出するだけでなく、その融合転写産物を再構築までを行うツールである。ある融合遺伝子が複数のリードから検出される場合、そのリードが融合遺伝子を支持するという意味でサポートリードと呼ぶが、FusionSeeker では融合転写産物を再構築する都合上、サポートリードが 3 本以上を検出対象としている。そのため、FusionSeeker はサポートリードが 3 本未満の融合遺伝子は検出することができない。

次節では先行手法の問題点について説明する。

2.3 先行手法の問題点

前項で説明したように、ロングリード RNA-Seq を用いた融合遺伝子検出では、ロングリードを参照ゲノムや参照トランスクリプトームにアライメントして、異なる遺伝子にアライメントされるキメラリードを検出する方法が一般的である (図 2.2)。ただし、ロングリード RNA-Seq で得られるリードのシークエンスアイデンティティは低いため、ロングリードを参照ゲノムや参照トランスクリプトームに 1 塩基単位で正確にアライメントすることは困難である。その中でも、融合点は遺伝子の配列が切り替わる場所であり、シークエンスアイデンティティが低いことによる配列の不一致が起こると、アライメントへ与える影響が大きい。ゲノムにアライメントされない融合点近傍の領域（ギャップ）が発生した場合、先行手法では融合遺伝子と融合点が正しく特定できないといった問題が発生する (図 2.3)。融合遺伝子の機能を解析するためには、翻訳により得られるタンパク質のアミノ酸配列を正しく予測できる必要があるが、先行手法はギャップが原因で融合点が正確に求められないため、融合遺伝子の機能解析の妨げとなっている。

ギャップが融合遺伝子検出に与える問題に対応するため、LongGF や JAFFAL ではギャップが 15 塩基以上あるリードは対象外とする戦略を取り、FusionSeeker は融合転写産物を再構築する戦略をとっている。しかし、LongGF や JAFFAL ではギャップ長が大きい（15 塩基以上）リードからは融合遺伝子を検出できないという問題があり、FusionSeeker はサポートリード数が 3 本未満の低発現の融合遺伝子は検出できないという問題がある。

また、上記に加えて、第 1 章で説明したシススプライシング融合遺伝子やトランススプライシング融合遺伝子のようなゲノム上での距離が近い遺伝子の融合遺伝子は検出できないという問題がある。

まとめると、先行手法では次の 2 つの問題点がある。1 点目は、ギャップが原因で融合遺伝子および融合点を正確に検出することが困難であり、融合遺伝子の機能解析が十分に行えない点である。2 点目は、ゲノムの組み換えによる融合遺伝子の検出を前提としているため、ゲノム上での距離が近い遺伝子の融合遺伝子を検出することができない点である。

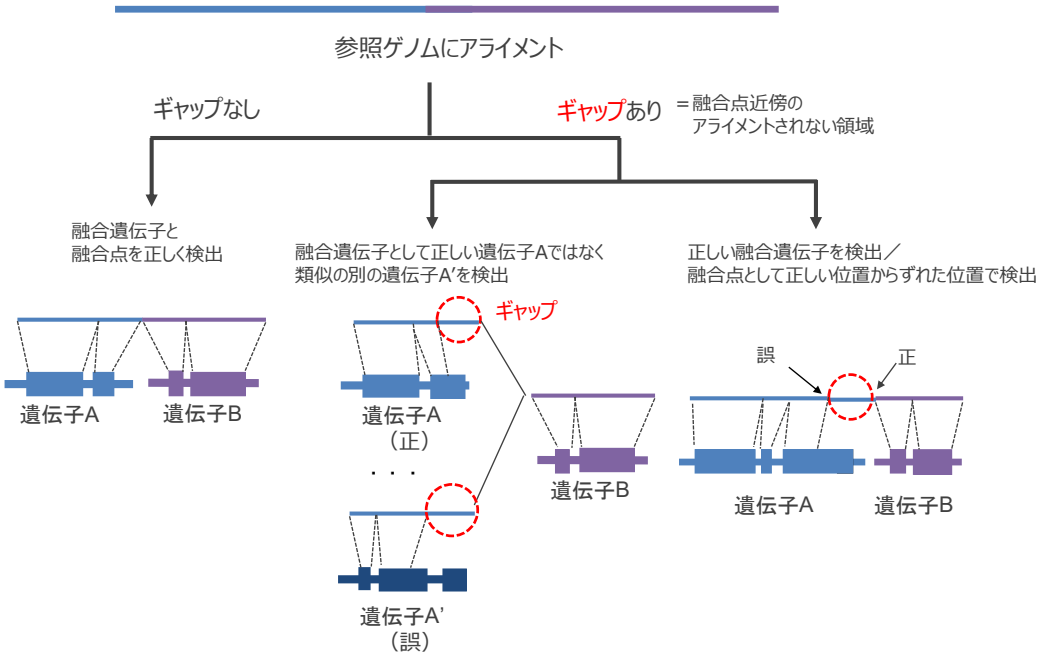

図 2.3　ギャップが融合遺伝子の検出に与える影響

2.4　提案手法

2.4.1　FLBEA（提案手法 1）

まずは、先行手法の問題点の 1 つ目である、ギャップが大きいリードから融合遺伝子を正しく検出ができないという問題を解決するために FLBEA (Full-Length and Both-End-End Alignment) を提案した [39]。問題の解決方針と手法のステップに分けて説明する。

2.4.1.1　問題の解決方針

ギャップが大きいリードは、正しい遺伝子の他に類似する別の遺伝子にアライメントされる場合がある（図 2.3）。この現象は融合点近傍にシークエンスエラーが多く含まれ、アライメントが不正確となっている可能性が高い。その一方、融合点から最も離れた両端の断片配列は正確にアライメントができる可能性がある。そこで FLBEA では、ロングリードの両端から断片配列を作成し、断片配列とロングリードの両方で一致する融合遺伝子のみを検出する戦略を取ることにした (図 2.4)。断片配列を検証的に使用することで融合遺伝子の偽陽性を防ぎつつ、ギャップが大きいリードからの融合遺伝子の検出を試みた。

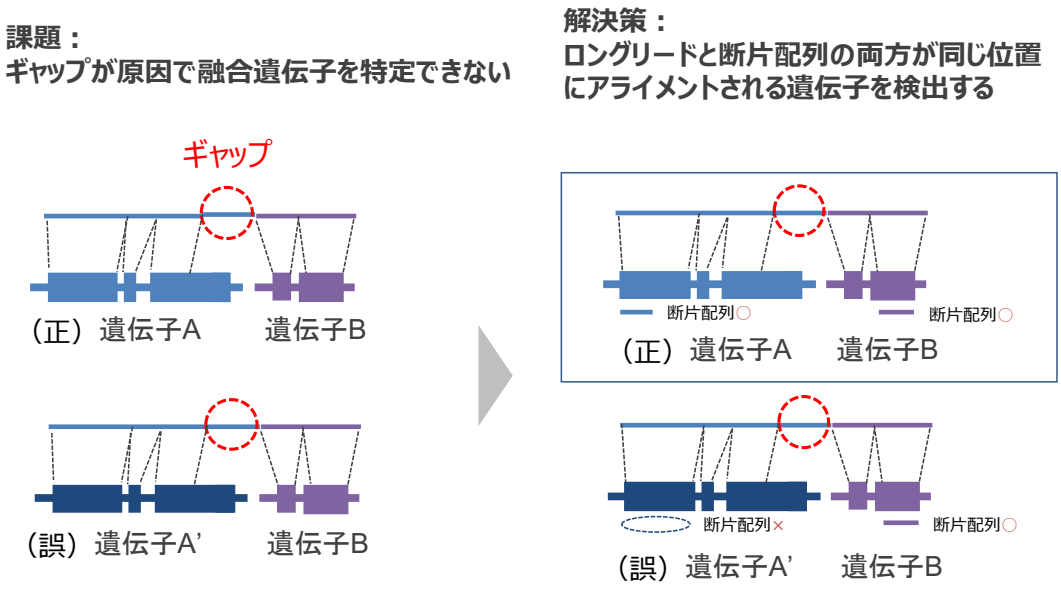

図 2.4　ギャップが原因で正しい融合遺伝子を検出できない問題に対する FLBEA の解決策

2.4.1.2　FLBEA のワークフロー

FLBEA における融合遺伝子検出のステップは次の通りである (図 2.5)。

ステップ 1:　ロングリードの両端から断片配列を生成する

ステップ 2:　断片配列を参照トランスクリプトームにアライメントする

ステップ 3:　2 つの異なる遺伝子にアライメントされるリードを抽出する

ステップ 4:　ロングリードと断片配列の両方を参照ゲノムにアライメントする

ステップ 5:　断片配列とロングリードがアライメントされる位置が一致するリードを抽出する

ステップ 6:　融合点でリードをクラスタリングする

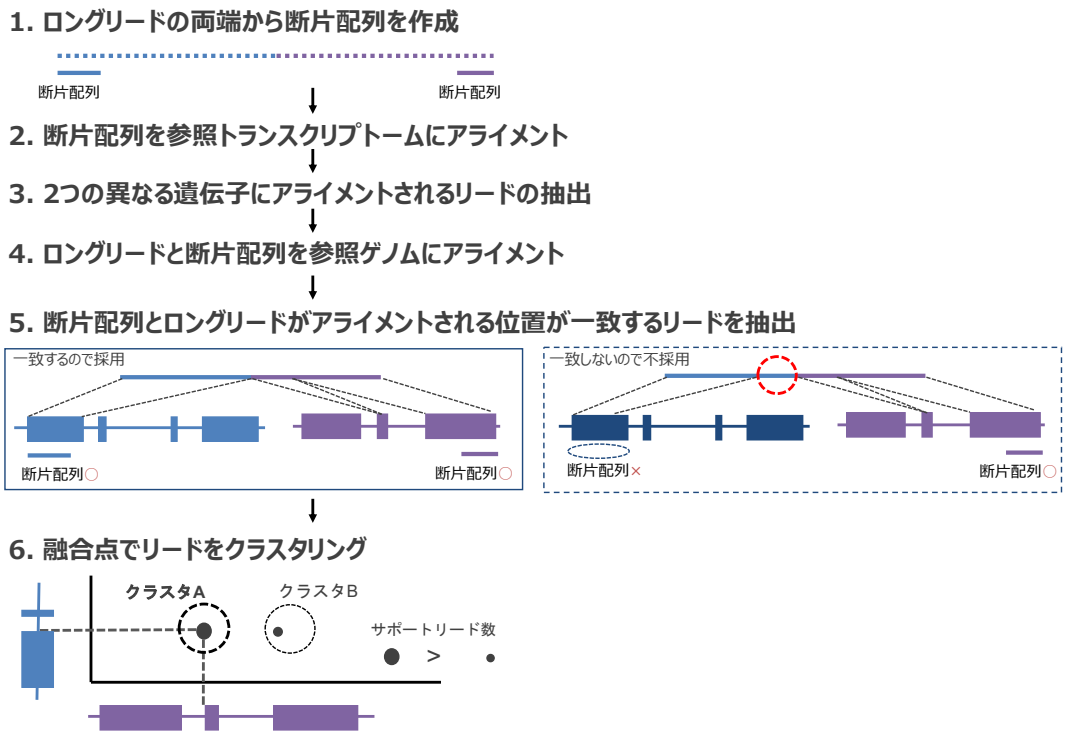

図 2.5　FLBEA の融合遺伝子検出の流れ

各ステップの内容は次の通りである。

ステップ 1： ロングリードの両端から断片配列を生成する。ロングリードの 5’ 側と 3’ 側から、それぞれ長さ 1 塩基の配列を切り出した断片配列を 5’ 断片配列と 3’ 断片配列とする。

ステップ 2： ロングリード用のアライメントツールである Minimap2[40] を用いてロングリードを参照トランスクリプトームにアライメントする。参照ゲノムにアライメントする前に、参照トランスクリプトームにアライメントすることで、正常な転写産物に由来するリードを除外することが可能である。また、Minimap2 は Nanopore と PacBio のシークエンスの際に発生するエラーの傾向を情報として内部に持っており、分析対象のロングリードが Nanopore か PacBio に由来するかをオプションで与えることで、より確からしいアライメント結果を得ることが可能である。

ステップ 3： ステップ 2 の参照トランスクリプトームへのアライメント結果をもとに、1 つの遺伝子にしかアライメントされない正常な転写産物由来のリードや 3 種類以上の遺伝子にアライメントされるリードを除外する。

ステップ 4： Minimap2 を用いてロングリードと 5’ 断片配列と 3’ 断片配列の 3 種類を参照ゲノムにアライメントする。

ステップ 5: ステップ 1 とステップ 3 で異なる遺伝子にアライメントされたロングリードは分析から除外する。

ステップ 6： ステップ 5 までを通過したリードを対象に、リードごとに融合点の座標を算出する。ロングリードの前半部分が参照ゲノムの染色体 c1 の座標 s1 から e1 に 5' 末端から 3' 末端にアライメントされ、ロングリードの後半部分が染色体 c2 の座標 s2 から e2 に 5' 末端から 3' 末端アライメントされた場合、染色体 c1 の e1 と染色体 c2 の s2 が融合点となる。一方、ロングリードの前半部分が 3' 末端から 5' 末端にアライメントされ、後半部分が 3' 末端から 5' 末端アライメントされた場合、染色体 c1 の s1 と染色体 c2 の e2 が融合点となる。

次に、算出した融合点をもとに N 塩基単位でクラスタリングを行う。各リードの融合点は各クラスターの代表の融合点とし、各クラスターの代表の融合点は、同じクラスターに属するリードの最頻の融合点として定義する。各クラスターの代表遺伝子名はクラスターに含まれるリードの過半数が支持する遺伝子名とし、過半数を占めるリードがない場合はそのクラスターおよびクラスター内のリードは分析から除外する。サポートリード数は「クラスターの総リード数」と「クラスターの代表遺伝子名と一致するリードが全体に占める割合」の積と定義する。最後に、融合点、サポートリード数を結果として出力する。以上がステップ 6 である。

2.4.1.3 FLBEA の問題点

FLBEA ではギャップが大きいリードからの融合遺伝子の検出を目指したが、2 つの問題点がある。1 点目は、ロングリードのシークエンスアイデンティティが低い場合、融合遺伝子の検出感度が低下することである。FLBEA は、ロングリードとロングリードの両端の断片配列の両方が参照ゲノムにアライメントされる遺伝子のみを検出することで、偽陽性を減らす手法である。断片配列は短いため、ロングリードよりもシークエンスアイデンティティの影響を強く受け、アライメントされるリード数がアイデンティティの低下に伴い急激に低下した。その結果、断片配列から検出できる融合遺伝子が減り、シークエンスアイデンティティが低い状況下では十分に機能しなかった。

2 点目は、依然としてギャップが大きいリードから融合点を正確に特定できないことである。FLBEA ではギャップ部分を補正するなどの措置は行なっていないため、融合点が不正確であるという問題は残っている。ギャップを補正するなどの直接的なアプローチを取り入れることで 2 点目の問題は改善の余地がある。

2.4.2 FUGAREC（提案手法 2）

前節の FLBEA の問題点に加え、先行手法の問題点の節で説明した次の 2 つの問題点を解決するため、新規の融合検出アルゴリズム FUGAREC (Fusion detection with gap realignment and clustering) を提案する。先行手法の問題点は、「ギャップが原因で融合遺伝子および融合点を正確に検出することが困難であり、融合遺伝子の機能解析が十分に行えない点」と「ゲノムの組み換えによる融合遺伝子の検出を前提としているため、ゲノム上での距離が近い遺伝子の融合遺伝子を検出できない点」である。

先行手法の問題の解決方針と手法のステップに分けて説明する。

2.4.2.1 先行手法の問題の解決方針

FUGAREC は先行手法に対して、(1) 融合点のエキソン境界固定、(2) ギャップの再アライメント、（3）融合点のクラスタリングの 3 つの処理を加えて拡張した手法である。（1）と（2）は融合点をエキソンの境界に合わせて補正することで、融合点を正確に特定する狙いがある。この処理は、遺伝子の融合はイントロンで起こることが多いことが知られているが [41]、融合点がエキソンの末端または開始点にある融合転写産物が作られることが多いことが背景にある。ギャップが原因でロングリードが複数の遺伝子にアライメントされ、正しい融合遺伝子の特定が困難な場合があるが、融合点をエキソン境界に合わせて補正することで、融合点が正確になり、正しい融合遺伝子の検出が可能となる (図 2.6)。

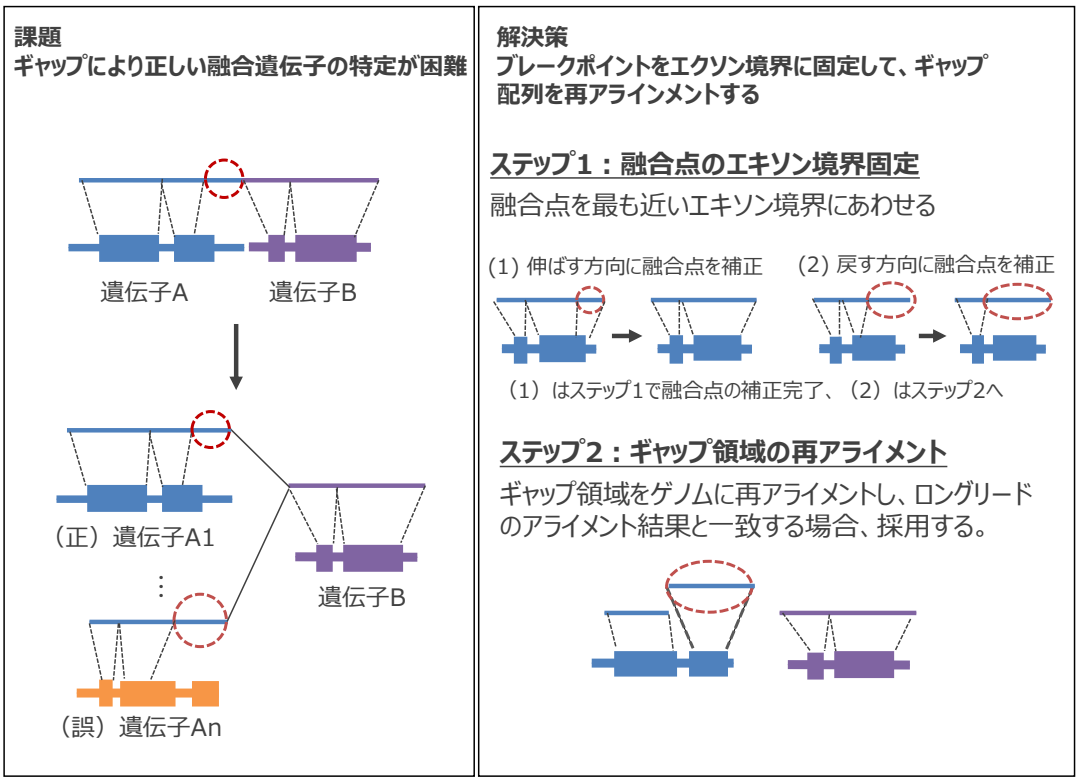

図 2.6　ギャップが原因で正しい融合遺伝子を検出できない問題に対する FUGAREC の解決策

(3) は融合点近傍のアライメントがずれて融合点が複数発生する場合、確からしい融合点を特定することで、偽陽性を防ぐ狙いがある。ギャップが原因で融合点近傍のアライメントがずれることがあるが、大部分は（1）と (2）で正しく補正される。しかし、本来とは違うエキソン境界に融合点が固定された場合やギャップの再アライメントに失敗した場合、1 つの融合遺伝子に対して複数の融合点が出現する場合がある。融合点をクラスタリングすることで、1 つの融合遺伝子に対し、最も確からしい融合点を特定することが可能となる (図 2.7)。

また、ゲノムの組み換えに非依存的な融合遺伝子も検出するため、先行手法とは異なり、融合点間の距離が近いものは除外するといった制限は設けないものとした。

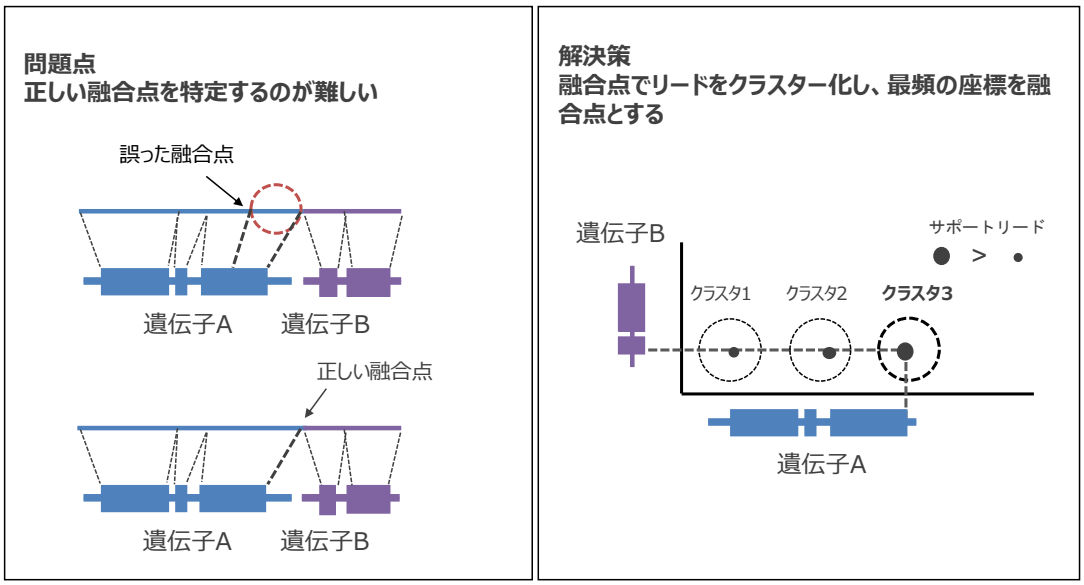

図 2.7　ギャップが原因で正しい融合点を検出できない問題に対する FUGAREC の解決策

2.4.2.2　FUGAREC のワークフロー

次に FUGAREC のワークフローを紹介する。FUGAREC は 6 つのステップから構成される (図 2.8)。

ステップ 1:　ロングリードを参照トランスクリプトームにアライメントする

ステップ 2:　2 つの異なる遺伝子にアライメントされるリードを抽出する

ステップ 3:　ロングリードを参照ゲノムにアライメントする

ステップ 4:　融合点をエキソン境界に固定する

ステップ 5:　ギャップ配列を参照ゲノムにアライメントする

ステップ 6:　融合点でリードをクラスタリングする

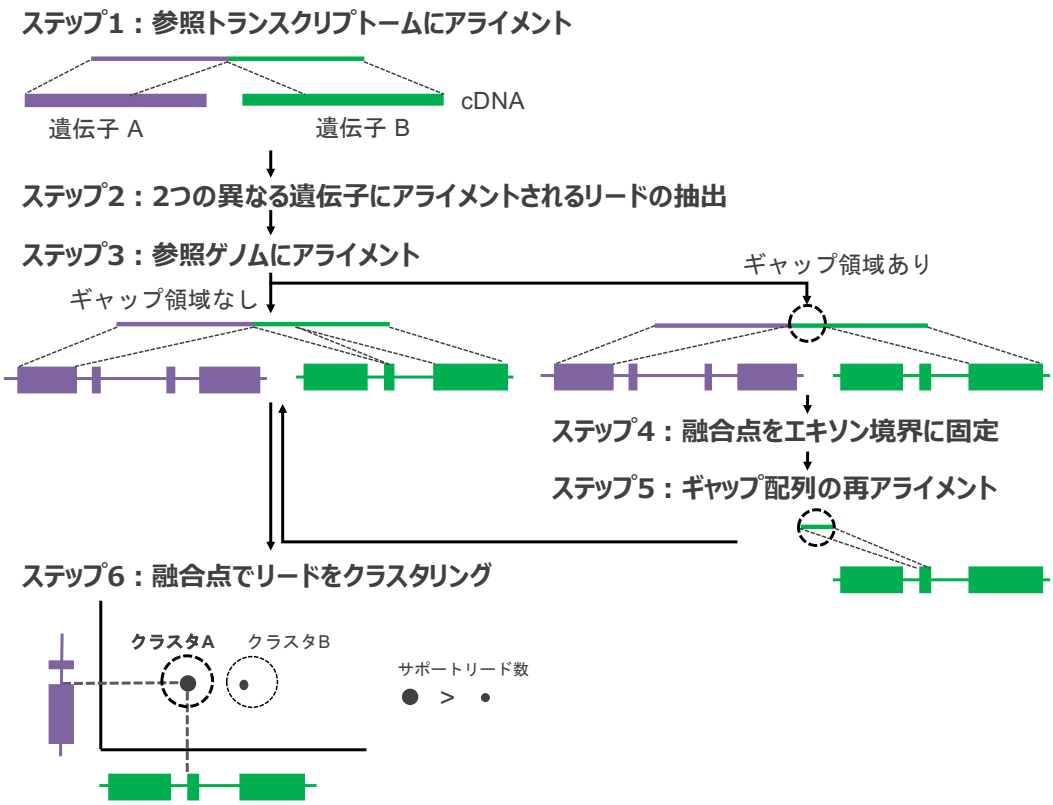

図 2.8　FUGAREC の融合遺伝子検出の流れ

各ステップの内容は次の通りである。

ステップ 1： Minimap2 を用いて参照トランスクリプトームにロングリードをアライメントする。

ステップ 2： 1 本の遺伝子にしかアライメントされない正常な転写産物由来のリードや 3 種類以上の遺伝子にアライメントされるリードを除外する。

ステップ 3： ロングリードを参照ゲノムにアライメントし、次の条件に当てはまるリードは分析から除外する。

・ステップ 1 とステップ 3 で異なる遺伝子にアライメントされたリード

・マッピングクオリティが q 以下のリード

・前半の遺伝子と後半の遺伝子にアライメントされたロングリードの配列が n 塩基以上オーバーラップするリード

次に、各リードについてギャップの塩基数を計算した。ロングリードの S1 から E1 が参照ゲノム中のある遺伝子にアライメントされ、ロングリードの S2 から E2 までが別の遺伝子にアライメントされている場合、S2 から E1 のアライメントされていない領域を

ギャップと定義する。ギャップが n 塩基以下のリードはステップ 6 へ、n 塩基より大きいリードはステップ 4 へ進む。

ステップ 4： ギャップが n 塩基より大きいリードは融合点が不正確である可能性が高いとみなし、融合点をエキソン境界に固定する。融合点をエキソン境界に固定する方法には、(A) アライメントを伸ばす方法に融合点を補正するパターンと、(B) アライメントを戻す方向に融合点を補正するパターンの 2 パターンがある。(A) のパターンでは、融合点からエキソン境界までの距離とギャップ長が一致するものを補正の対象とし、融合点からエキソン境界までの距離がギャップ長より n 塩基以上長い場合は解析から除外する。(B) のパターンでは、ギャップが増えることになるが、エキソン 1 個分が新しいギャップになる。新しいギャップは、ステップ 5 で参照ゲノムに再アラインメントする。

ステップ 5： 融合点をエキソン境界に固定した後、ギャップとして残った配列を BLAT(Blast Like Alignment Tool) [42] を用いて参照ゲノムに再アラインメントする。BLAT は短い配列を局所的にアライメントする際に使用されるアライメントツールである。ギャップ配列は非常に短い配列であるため、ギャップ配列の再アライメントにはロングリード用のアライメントツールである Minimap2 ではなく BLAT を採用した。ギャップ配列がアライメントされた遺伝子名とステップ 3 でロングリードがアライメントされた遺伝子名が一致するリードのみステップ 6 に進める。

ステップ 6： ステップ 5 までを通過したリードを対象に、リードごとに融合点の座標を算出する。ロングリードの前半部分が参照ゲノムの染色体 c1 の座標 s1 から e1 に 5' 末端から 3' 末端にアライメントされ、ロングリードの後半部分が染色体 c2 の座標 s2 から e2 に 5' 末端から 3' 末端アライメントされた場合、染色体 c1 の e1 と染色体 c2 の s2 が融合点となる。一方、ロングリードの前半部分が 3' 末端から 5' 末端にアライメントされ、後半部分が 3' 末端から 5' 末端アライメントされた場合、染色体 c1 の s1 と染色体 c2 の e2 が融合点となる。

次に、算出した融合点をもとに N 塩基単位でクラスタリングを行う。各リードの融合点は各クラスターの代表の融合点とし、各クラスターの代表の融合点は、同じクラスターに属するリードの最頻の融合点として定義する。各クラスターの代表遺伝子名は、クラスターに含まれるリードの過半数が支持する遺伝子名とし、過半数を占めるリードがない場合はそのクラスターおよびクラスター内のリードを分析から除外する。サポートリード数は「クラスターの総リード数」と「クラスターの代表遺伝子名と一致するリードが全体に占める割合」の積と定義する。最後に、融合点、サポートリード数を結果として出力する。

NR5A2 と TBC1D24 の融合遺伝子のリードをもとに FUGAREC の手順を示す (図 2.9)。

1. 参照トランスクリプトームにアライメント

クエリ情報				トランスクリプトーム情報	
リード	クエリ長	遺伝子1 アライメント終了位置	遺伝子2 アライメント開始位置	遺伝子1 遺伝子名	遺伝子2 遺伝子名
リード1	1666	476	605	NR5A2	TBC1D24
リード2	1710	584	630	NR5A2	TBC1D24
リード3	1675	556	599	NR5A2	TBC1D24

2. 2つの異なる遺伝子にアライメントされるリードの抽出

3.参照ゲノムにアライメント

クエリ情報					ゲノム情報	
リード	クエリ長	遺伝子1 アライメント終了位置	遺伝子2 アライメント開始位置	ギャップ長	遺伝子1 融合点	遺伝子2 融合点
リード1	1666	476	582	106	chr1:200080427	chr16:2547726
リード2	1710	590	607	17	chr1:200080341	chr16:2547726
リード3	1675	556	607	51	chr1:200080368	chr16:2547114

4. 融合点をエキソン境界に固定

クエリ情報					ゲノム情報			
リード	クエリ長	遺伝子1 アライメント終了位置	遺伝子2 アライメント開始位置	ギャップ長	遺伝子1 融合点	遺伝子2 融合点	遺伝子1 近いエキソン境界	遺伝子2 近いエキソン境界
リード1	1666	574	580	6	chr1:200080329	chr16:2547728	chr1:200080329	chr16:2547728
リード2	1710	602	605	3	chr1:200080329	chr16:2547728	chr1:200080329	chr16:2547728
リード3	1675	595	607	12	chr1:200080329	chr16:2547114	chr1:200080329	chr16:2547114

5. ギャップ配列の再アライメント

6. 融合点でリードをクラスタリング

クエリ情報					ゲノム情報				クラスター情報		
リード	クエリ長	遺伝子1 アライメント終了位置	遺伝子2 アライメント開始位置	ギャップ長	遺伝子1 融合点	遺伝子2 融合点	遺伝子1 近いエキソン境界	遺伝子2 近いエキソン境界	クラスターID	遺伝子1 融合点	遺伝子2 融合点
リード1	1666	574	580	6	chr1:200080329	chr16:2547728	chr1:200080329	chr16:2547728	A	chr1:200080329	chr16:2547728
リード2	1710	602	605	3	chr1:200080329	chr16:2547728	chr1:200080329	chr16:2547728	A	chr1:200080329	chr16:2547728
リード3	1675	595	607	12	chr1:200080329	chr16:2547114	chr1:200080329	chr16:2547114	A	chr1:200080329	chr16:2547728

クラスター A=chr1:200080—chr16:2547

図 2.9　FUGAREC の融合遺伝子検出の流れ（実例）

2.5　実験

2.5.1　データと実験条件

2.5.1.1　シミュレーションデータ

ロングリード RNA-Seq 用の融合遺伝子のシミュレーションデータを用いて融合遺伝子の検出性能を評価した。検証に用いた融合遺伝子のシミュレーションデータの概要を示す（表 2.3)。本シミュレーションデータは、2 つの遺伝子の塩基配列を結合させた後、ロングリードのシークエンスエラーを模すようにノイズが加えられたリードから構成されるデータである。2 つの遺伝子はランダムに選ばれ、それぞれの遺伝子のエキソン境界が融合点となるように 2 つの遺伝子の塩基配列を繋げて作成されている。加えられたロングリードのシークエンスノイズは Nanopore と PacBio の 2 種類があり、各モデルのデータ

セットには 2,500 種類の融合遺伝子が含まれる。また、2,500 種類の融合遺伝子は、5 段階のシークエンスアイデンティティ (75%、 80%、 85%、 90%、 95%) と 5 段階のカバレッジ (1、 2、 10、 50、 100) の組み合わせで、各グループに 100 種類の融合遺伝子が含まれるように調整されている。

本シミュレーションデータセットは JAFFAL[37] の論文で性能評価に使用されたデータセットである。本データセットを用いて評価することで、提案手法のシークエンスエラーに対する堅牢性およびカバレッジが稼げない発現量が低い融合遺伝子に対する検出感度の高さを評価することが可能である。なお、本データセットは融合遺伝子以外の正常な転写産物はリードは含まれていない。

融合遺伝子数は、最も少ないデータセットで 445 個、最も多いデータセットで 462 個であった。また、リード数は最も少ないデータセットで 17,479 本、最も多いデータセットで 17,889 本、平均リード長は最も少ないデータセットで 2,111 塩基、最も多いデータセットで 2,345 塩基であった。

表 2.3　シミュレーションデータセットの概要

データセット名	エラーモデル	アイデンティティ (%)	融合遺伝子の数	リード数	平均リード長
ONT 95%	Nanopore	95	459	17,745	2,312
ONT 90%	Nanopore	90	461	17,479	2,130
ONT 85%	Nanopore	85	462	17,531	2,114
ONT 80%	Nanopore	80	447	17,819	2,146
ONT 75%	Nanopore	75	452	17,889	2,150
Pac 95%	PacBio	95	445	17,495	2,345
Pac 90%	PacBio	90	454	17,500	2,128
Pac 85%	PacBio	85	450	17,566	2,111
Pac 80%	PacBio	80	455	17,779	2,150
Pac 75%	PacBio	75	460	17,500	2,199

2.5.1.2　リファレンスデータ

融合遺伝子の検出性能を評価するために使用するシミュレーションデータ [37] は

Gencode release ver19 [43] で作成されていた。そのため、参照ゲノムと参照トランスクリプトームは Gencode release ver19 [43] を用いた。

2.5.1.3 FLBEA の実行

シミュレーションデータに対し、提案手法 1 で説明した流れで FLBEA を実行した（2.4.1.2 節を参照）。各ステップで使用したパラメーターや条件は次の通りである。ステップ 1 の断片配列の長さ l は 200 塩基とした。ステップ 2 の Minimap は 2.17 のバージョンを使用し、シミュレーションデータが Nanopore の場合は-x map-ont、PacBio の場合は-x map-pb のオプションで実行した。ステップ 6 のクラスタリングの単位である塩基数 N は 1000 塩基とした。

2.5.1.4 FUGAREC の実行

シミュレーションデータに対し、提案手法 2 で説明した流れで FUGAREC を実行した（2.4.2.2 節を参照）。各ステップで使用したパラメーターや条件は次の通りである。ステップ 1 の Minimap は 2.17 のバージョンを使用し、シミュレーションデータが Nanopore の場合は-x map-ont、PacBio の場合は-x map-pb のオプションで実行した。ステップ 3、ステップ 4 の塩基数 n は 15 塩基、ステップ 3 のマッピングクオリティ q の閾値は 0 以上として実行した。ステップ 5 の BLAT は-out=blast8 -minScore=10 -stepSize=5 のオプションで実行した。また、ギャップ配列が複数の遺伝子にアライメントされる場合、E 値 0.05 以下かつマッピング率 0.5 以上の中で、E 値が最も低いアライメント結果を採用した。ステップ 6 のクラスタリングの単位である塩基数 N は 1000 塩基とした。

2.5.1.5 先行手法の実行

提案手法の融合遺伝子の検出性能を評価するための比較対象として JAFFAL と FusionSeeker を用いた。JAFFAL[37] の論文中および FusionSeeker[38] の論文中で、JAFFAL、FusionSeeker の融合遺伝子の検出性能がともに LongGF を上回ると報告されていたため、本研究では JAFFAL と FusionSeeker を比較対象とした。

JAFFAL は、設定ファイルがまとめられた JAFFAL.groovy を引数に与えて次の通りに実行した。

```
bpipe run JAFFAL.groovy.
```

FusionSeeker は、Nanopore のシミュレーションデータに対して、-datatype nanopore

のオプションを指定し、PacBio のシミュレーションデータに対して、-datatype isoseq のオプションを指定して実行した。

2.5.1.6 融合遺伝子の検出性能の評価方法

各ツールが提示する融合遺伝子の融合点と正解の融合点からの距離が 100 塩基以下であれば真陽性 (TP)、100 塩基より離れていた場合は偽陽性 (FP)、検出できなかった融合遺伝子は偽陰性 (FN) とした。また、以下の式に示すように、Precision、Recall、F1 スコアを算出した。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.2}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

Precision と Recall はトレードオフの関係にあり、閾値を調整することで片方だけを高めることが可能である。Precision と Recall の加重平均値で定義される F1 スコアは、モデルの総合的な評価指標として使用され、F1 スコアが高いモデルは性能が良いとされている。本研究では、モデルの総合的な評価指標として優れている F1 スコアを用いて、各手法の融合遺伝子検出性能を評価した。

2.5.2 実験結果

2.5.2.1 シミュレーションデータにおける融合遺伝子の検出性能

シミュレーションデータを用いて FUGAREC、FLBEA、JAFFAL、FusionSeeker の 4 手法の融合遺伝子の検出性能を比較した（図 2.10）（表 2.4）（表 2.5）。

1 本以上のリードでサポートされる融合遺伝を検出対象とした場合、FLBEA はシークエンスアイデンティティが 95% 以上の条件では JAFFAL よりも F1 スコアが高かったが、それ未満では JAFFAL には及ばなかった（表 2.4）。また、FLBEA はシークエンスアイデンティティが 80% 以下の条件では、他の手法と比較して圧倒的に F1 スコアが低かった。FUGAREC は全てのシークエンスアイデンティティで F1 スコアが最も高かった。

2 本以上のリードのサポートされた融合遺伝子を検出対象とした場合、ONT90% と Pac80% と Pac75% 以外のデータでは FUGAREC の F1 スコアが最も高かった（表

2.5)。ONT90% と Pac80% と Pac75% のデータについては JAFFAL が最も高かった。FUGAREC はサポートリードが 1 本以上、2 本以上どちらの場合においても F1 スコアは FusionSeeker よりも高かった。

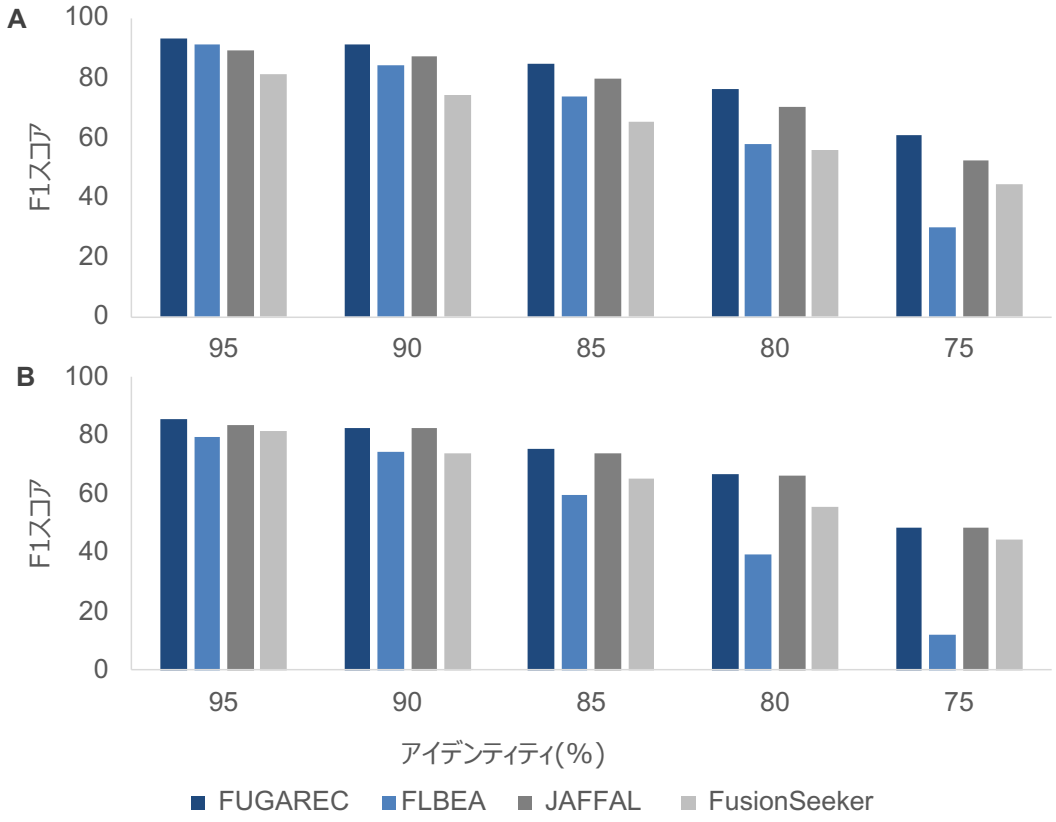

図 2.10　検出限界のサポートリード数が 1 本以上（A）と 2 本以上 (B) の F1 スコア

表 2.4　検出限界のサポートリード数を 1 とした場合の融合遺伝子の検出性能（F: FUGAREC、B: FLBEA、J: JAFFAL、S: FusionSeeker）

	Precision (%)				Recall (%)				F1 (%)			
Dataset	F	B	J	S	F	B	J	S	F	B	J	S
ONT 95%	**99.5**	96.4	89.8	93.3	87.4	86.9	**88.7**	72.8	**93.0**	91.4	89.3	81.4
ONT 90%	**98.5**	85.7	92.9	82.5	**85.0**	83.1	82.0	67.5	**91.3**	84.4	87.1	74.2
ONT 85%	**97.7**	81.2	89.5	72.4	**74.5**	67.3	72.1	59.5	**84.5**	73.6	79.9	65.3
ONT 80%	**98.2**	72.6	94.3	67.8	**62.0**	48.1	55.7	47.7	**76.0**	57.9	70.0	56.0
ONT 75%	**100**	70.8	95.9	66.2	**43.8**	18.8	35.8	33.4	**60.9**	29.7	52.2	44.4
Pac 95%	**99.3**	93.2	88.4	90.9	**90.3**	89.2	89.2	74.2	**94.6**	91.2	88.8	81.7
Pac 90%	**97.3**	83.9	93.6	80.8	**86.6**	85.0	83.5	69.4	**91.6**	84.5	88.2	74.6
Pac 85%	**99.4**	77.1	90.2	71.7	73.1	66.0	**73.3**	60.9	**84.3**	71.1	80.9	65.9
Pac 80%	**99.3**	69.8	92.5	69.8	**59.1**	40.7	54.3	47.7	**74.1**	51.4	68.4	56.7
Pac 75%	**98.9**	66.1	93.8	63.3	**38.7**	17.0	33.0	27.4	**55.6**	27.0	48.9	38.2

表 2.5　検出限界のサポートリード数を 2 とした場合の融合遺伝子の検出性能（F: FUGAREC、B: FLBEA、J: JAFFAL、S: FusionSeeker）

	Precision (%)				Recall (%)				F1 (%)			
Dataset	F	B	J	S	F	B	J	S	F	B	J	S
ONT 95%	**100**	98.4	90.6	92.3	75.2	66.9	**77.8**	72.8	**85.8**	79.6	83.7	81.4
ONT 90%	**99.1**	92.9	94.9	82.5	70.7	62.3	**73.1**	67.5	82.5	74.5	**82.6**	74.2
ONT 85%	**99.3**	88.1	90.1	72.4	60.6	45.0	**62.8**	59.5	**75.3**	59.6	74.0	65.3
ONT 80%	**98.7**	80.6	95.0	67.8	50.3	26.0	**62.8**	47.7	**66.7**	39.3	66.6	56.0
ONT 75%	**100**	88.2	97.3	66.2	32.1	6.6	32.3	**33.4**	**48.6**	12.3	48.5	44.4
Pac 95%	**100**	96.5	89.1	90.9	76.0	67.6	**79.1**	74.2	**86.3**	79.5	83.8	81.7
Pac 90%	**98.8**	94.7	94.9	80.8	73.6	62.8	**74.4**	69.4	**84.3**	75.5	83.5	74.6
Pac 85%	**100**	85.2	91.0	71.7	61.8	43.6	**64.9**	60.9	**76.4**	57.6	75.7	65.9
Pac 80%	**99.1**	75.7	93.3	69.8	47.0	22.8	**49.2**	47.7	63.8	34.9	**64.5**	56.7
Pac 75%	**100**	71.0	94.1	63.3	27.0	4.8	**27.8**	27.4	42.5	9.0	**43.0**	38.2

2.5.2.2　カバレッジが低い融合遺伝子の検出性能

次に、カバレッジが低い（参照ゲノムにアライメントした際、同じ領域にアライメントされるリード数が少ない）融合遺伝子に対する検出感度を評価するため、カバレッジが 1 本、2 本、10 本、50 本、100 本の融合遺伝子のリードをそれぞれ取り出したデータに対して、各手法の検出感度を比較した（図 2.11）。FUGAREC、FLBEA、JAFFAL はカバレッジが 50 以上の融合遺伝子に対しては同程度の検出感度であった。一方、FUGAREC はカバレッジが 10 以下の融合遺伝子に対しては高い検出感度を示した。FusionSeeker はカバレッジ 2 以下の融合遺伝子は検出できなかった。また、どのカバレッジでも他の手法と比較して検出感度が低かった。

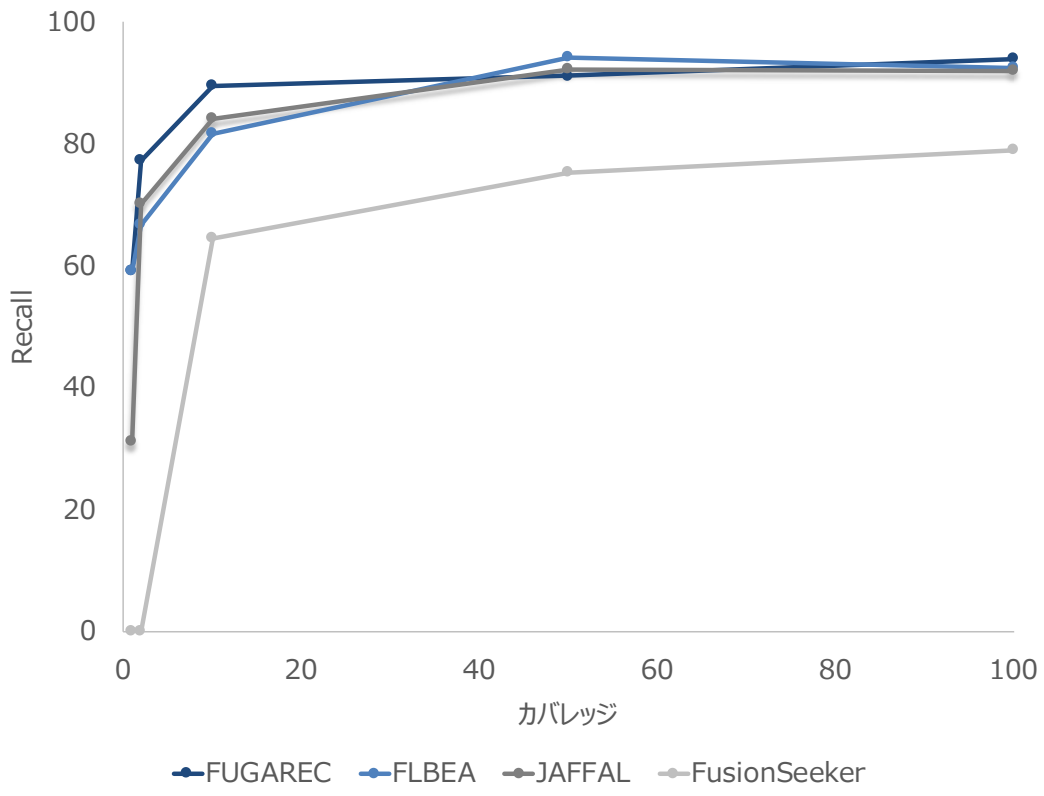

図 2.11　ONT 90% におけるカバレッジ別の融合遺伝子の検出感度

2.5.2.3　融合点のエキソン境界固定とギャップの再アラインメントの効果

融合点のエキソン境界固定とギャップの再アライメントによりギャップを補正することで、分析対象とするリード数がどの程度増えるかを検証するため、各ステップで除外されるリード数を調査した （図 2.12）。融合点のエキソン境界固定とギャップの再アライメ

ントがない場合、シークエンスアイデンティティが 95% のデータセットであっても、半数以上の融合遺伝子由来のリードが分析から除外された (図 2.12 (B))。一方、エキソン境界固定とギャップの再アライメントを行うことで、シークエンスアイデンティティ 95% のデータセットでは、70% 以上のリードを融合遺伝子由来のリードとして検出することができた (図 2.12(A))。融合点のエキソン境界固定とギャップの再アライメントにより、ギャップ長が原因でフィルタリングされるリード数および最終工程でフィルタリングされるリード数が減少した。結果として、検出できる融合遺伝子由来のリードの数は、シークエンスアイデンティティが 95%、90%、85%、80%、75% の場合、それぞれ 1.7 倍、2.6 倍、4.8 倍、7.8 倍、15 倍に増加した。融合点のエキソン境界固定およびギャップの再アライメントは、融合遺伝子の検出感度を向上させるために重要なステップであることが示唆される。

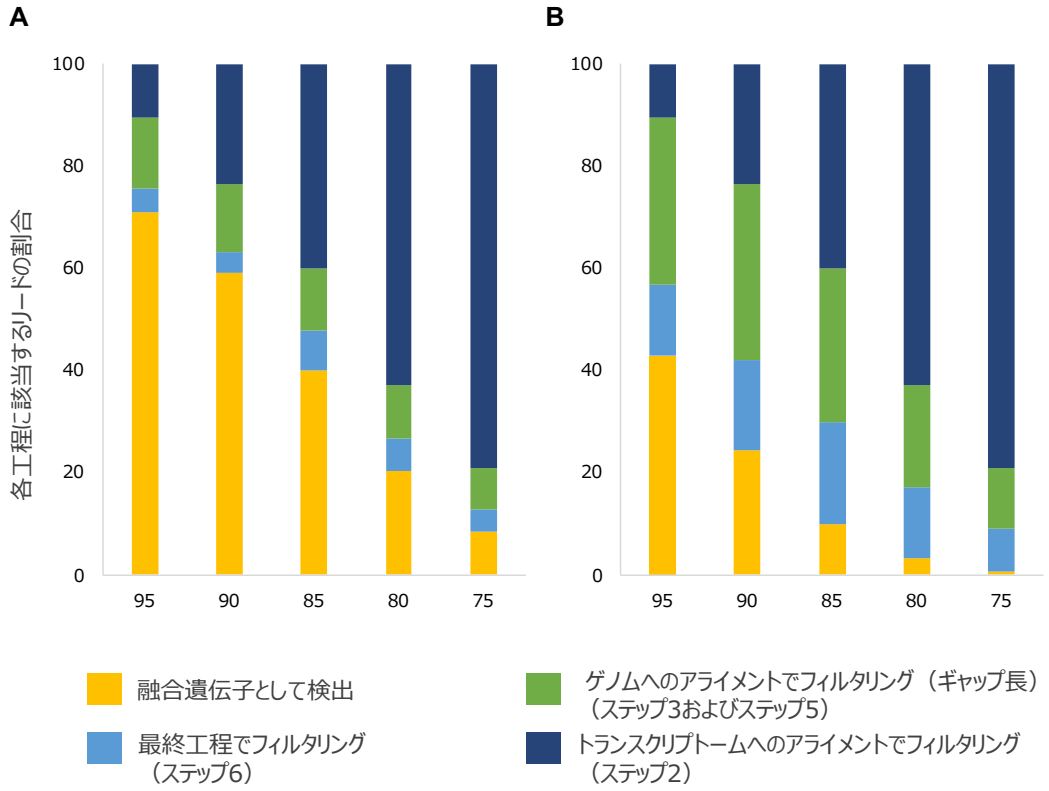

図 2.12　FUGAREC の各ステップで除外されるリードの割合

2.5.2.4　融合点のクラスタリングの効果

FUGAREC の融合遺伝子検出における融合点のクラスタリングの寄与を評価するた

め、クラスタリングの有無で F1 スコアを比較した（図 2.13）。融合点のクラスタリングの効果は、Recall と Precision でトレードオフの関係にあったが、ONT 75% のデータセットを除いて、全てのデータセットで F1 スコアが向上した。ONT 75% のデータセットでは、クラスタリング前に偽陽性がなかったため、F1 スコアが改善しなかった。融合点のクラスタリングは偽陽性を減らすことで、融合遺伝子検出の性能向上に寄与することが示唆された。

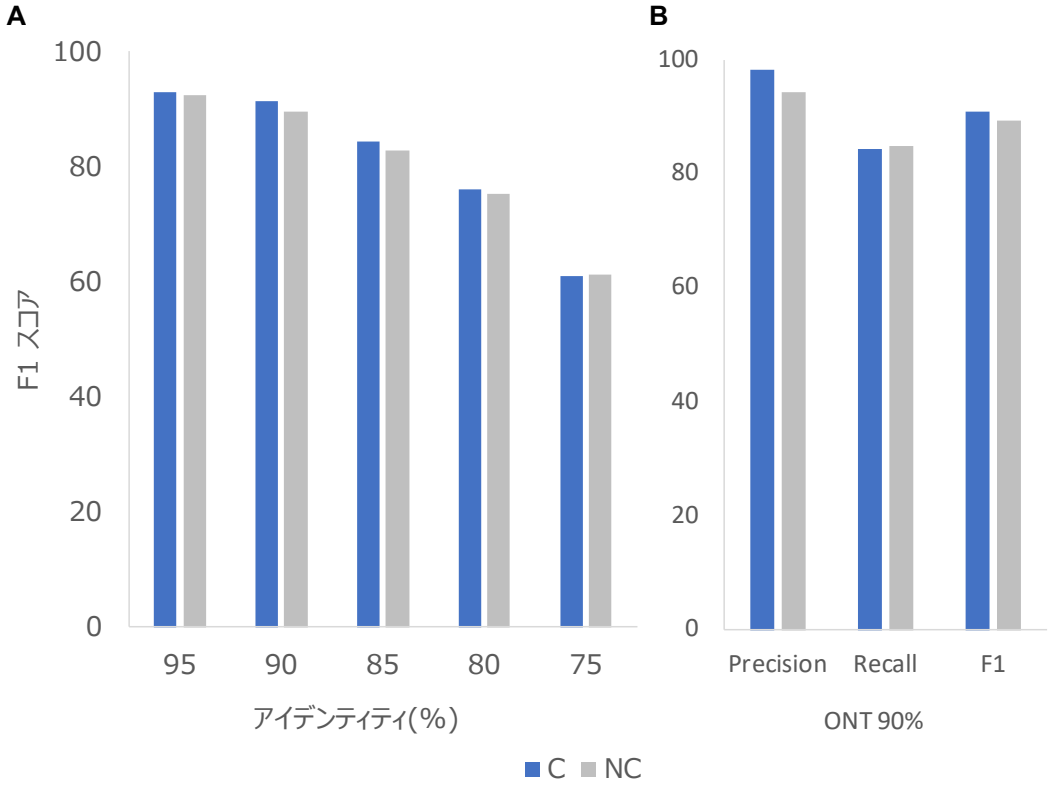

図 2.13　融合点のクラスタリングの融合遺伝子検出性能への寄与

2.5.3　考察

本章では、シミュレーションデータセットを用いて、先行手法や FLBEA と比較することで、FUGAERC の融合遺伝子の検出性能を評価した。FUGAREC は、どのシークエンスアイデンティティでも最も F1 スコアが高かった。シークエンスアイデンティティが低い条件で、融合遺伝子の検出性能が極端に低下する FLBEA の問題を克服したといえる。FLBEA はシークエンスアイデンティティが低い条件では、断片配列が参照ゲノムにアライメントされず、検出できる融合遺伝子の数が減ったことで極端に検出精度が低下し

ていた。それに対して FUGAREC では、融合点をエキソン境界に補正できるものは再アライメントを行わずに融合点を補正した。再アライメントを行う場合も、短い配列のアライメントに有利な BLAT を使用することで、参照ゲノムにアライメントできるリード数を増加させた。この 2 点が大きく寄与することで、低いシークエンスアイデンティティでも高い融合遺伝子の検出性能を保つことができたと考える。

また、JAFFAL と比較しても FUGAREC の F1 スコアは高い傾向があった。FUGAREC が先行手法よりも高い F1 スコアを取得できたのは、検出感度の向上と偽陽性の削減をバランスよく行うことができたからである。融合遺伝子の検出感度の向上には、(1) 融合点のエキソン境界固定、(2) ギャップの再アライメントが寄与していた。(1) と (2) は融合点をエキソン境界に補正することで、ギャップが原因でフィルタリングされる融合遺伝子由来のリード数を減少させた。その結果、FUGAREC では JAFFAL では検出できなかった複数の融合遺伝子を検出することができた（2.5.2.3 節を参照）。一例として JAFFAL が誤って検出した融合遺伝子 PSAP–PRDM10 のリードを示す (図 2.14 (A))。FUGAREC では、NMNAT1 の融合点をエキソン境界に固定することでギャップを 5 塩基短くでき（補正前：chr1:10035828、補正後：chr1:10035833）、PLA2G1B の融合点をエキソン境界に固定することでギャップを 31 塩基短くできた（補正前：chr12:120763792、補正後：chr12:120763823）。その結果、36 塩基あったギャップがなくなり、FUGAREC では PSAP–PRDM10 を検出することができた。

また、(1) 融合点のエキソン境界固定、(2) ギャップの再アライメントは、融合点を正確にすることで、偽陽性を防ぐことにも寄与していた。一例として、JAFFAL が誤って検出した融合遺伝子 C20orf194–RRH のリードを示す（図 2.14 (B))。ロングリードの前半部分の融合点を最も近いエキソンに補正するには、融合点を 907 塩基分移動させる必要があり、ロングリードの後半部分の融合点を最も近いエキソンに補正するには、融合点を 12 塩基分移動させる必要がある。合計 919 塩基分だけ融合点を移動させる必要があるが、ロングリードのギャップは 95 塩基しかないため、塩基数が一致しない。アライメントされている遺伝子および融合点が不正確である可能性が高く、偽陽性として除外すべきと判断して FUGAREC では該当リードを分析から除外した。その結果、FUGAREC では JAFFAL と同じ偽陽性が発生することはなかった。以上の結果から、(1)、(2) のステップは感度の向上および偽陽性の削減の両方に寄与しており、極めて重要なステップであるといえる。

一方、シミュレーションデータセットには正常な転写産物に由来するリードは含まれ

ず、融合遺伝子に由来するリードしか含まれていない。融合遺伝子に由来する異常なリードしかない状況であれば、異常なリードを積極的に判定する手法が有利になりやすい。FUGAREC は F1 スコアが非常に高かったが、データのバイアスが含まれる結果であることに注意したい。次の章では正常な転写産物のリードと融合遺伝子に由来する異常なリードの両方を含むがん細胞株のシークエンスデータで提案手法の有効性を評価する。

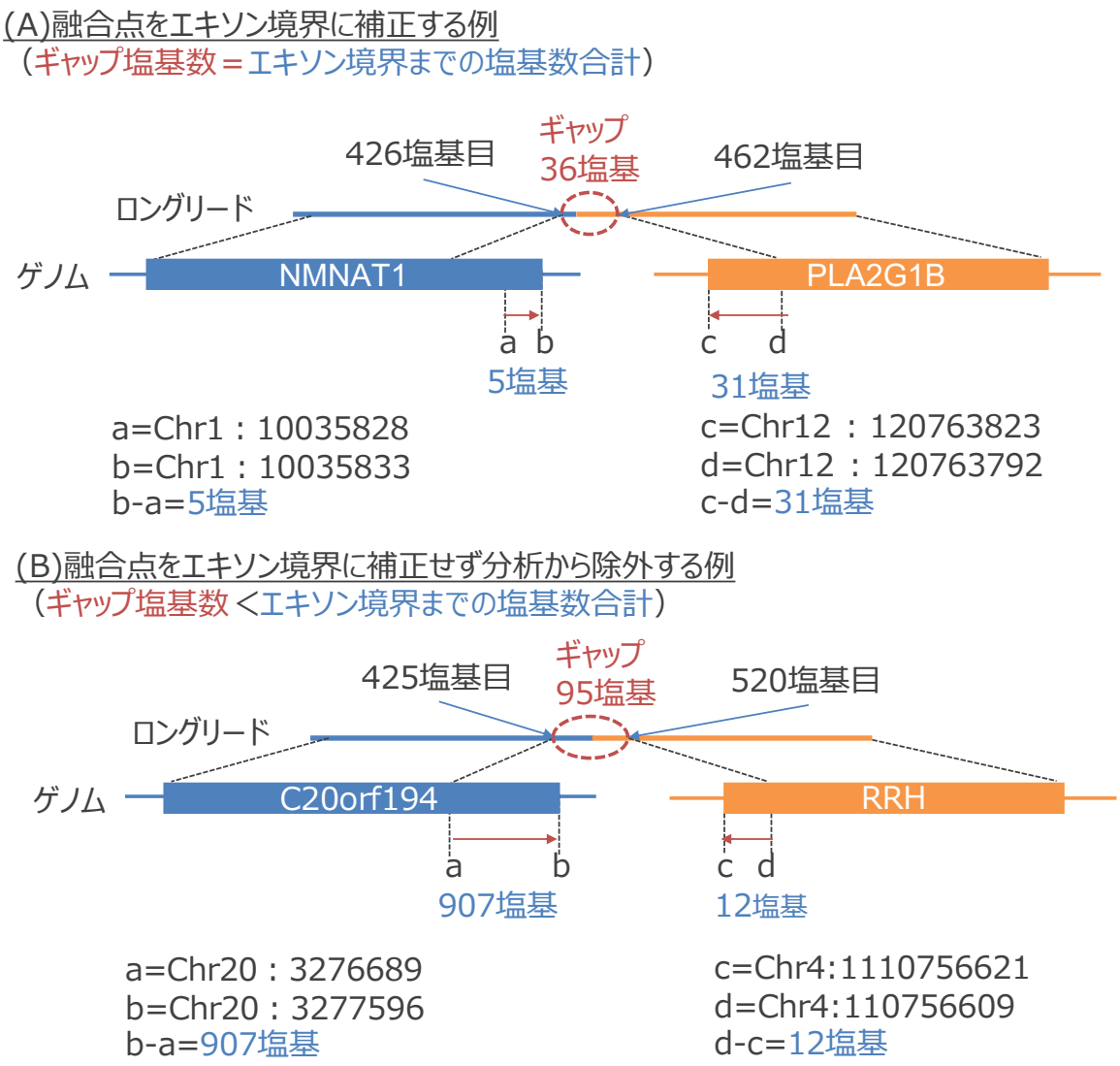

図 2.14　エキソン境界固定の実例

2.6 結言

本章では、ロングリード RNA-Seq を用いた既知の融合遺伝子検出ツールである JAFFAL、FusionSeeker の問題点を提示した上で、その問題点を解消するための手法として FUGAREC を提案した。先行手法の問題点として、ギャップが原因で融合遺伝子および融合点を正確に検出できず、融合遺伝子の機能解析が十分に行えない点、ゲノム上での距離が近い遺伝子の融合遺伝子を検出できない点を挙げた。

FUGAREC では、先行手法に対して、融合点のエキソン境界固定、ギャップの再アライメント、融合点のクラスタリングの 3 つの処理を追加して拡張した。また、ゲノム上での距離が近い遺伝子の融合遺伝子も検出するため、融合点間の距離の制約は設けないようにすることで先行手法の問題点の解決を試みた。

その結果、先行研究では検出できない融合遺伝子の検出や、融合点の正確な補正が可能となったことを示した。また、シミュレーションデータで FUGAREC が最も高い検出性能を有していることを示した。

第 3 章　培養がん細胞株での有効性の検証

3.1　緒言

第 2 章では、先行手法の問題点を解決するための手法として FUGAREC を提案し、シミュレーションデータで高い検出性能を有していることを実証した。本章では、実際の培養がん細胞のシークエンスデータを扱う。培養がん細胞のシークエンスデータは第 2 章で扱ってきたシミュレーションデータとは異なり、正常な転写産物が大半を占める。そのため、正常な転写産物のリードの中から融合遺伝子に由来する数少ないリードを特定する必要がある。また、シミュレーションデータとは異なり、培養がん細胞のシークエンスデータ中に含まれる融合遺伝子の数量および正解の融合点の座標が不明である。このような背景から、融合遺伝子の検出手法および評価方法を新たに設計する必要がある。

本章では、第 2 章で提案した手法をベースに、がん細胞株から融合遺伝子を検出できるように手法を拡張する。次に、複数の培養がん細胞のシークエンスデータを用いて提案手法の性能を先行手法と比較し、得られた結果をもとに考察を行う。

3.2　培養がん細胞株で手法の有効性を評価する意義

第 2 章で扱ったシミュレーションデータはロングリードのシークエンスエラーを考慮して作成されているものの、実際のがん細胞からシークエンスしたデータとは性質が異なるデータである。シミュレーションデータは、融合遺伝子のリードだけで構成され、正常な転写産物由来のリードが全く含まれていない。一方、実際の培養がん細胞のシークエンスデータは正常な転写産物由来のリードが大半を占める。第 2 章では融合遺伝子に由来する異常なリードを検出する能力を示すことができたが、正常なリードを正常と判断する能力は示すことができていない。また、がんの診断マーカーでの活用を想定した場合、実際のがん細胞からシークエンスしたデータから、偽陽性を少なく保ったまま異常な融合遺伝子由来のリードを検出できることが重要である。そのため、本章では培養がん細胞のシークエンスデータを用いて、提案手法の有効性を検証する。

3.3　検出手法

3.3.1　問題点

前述の通り、シミュレーションデータには正常な転写産物由来のリードが含まれていない。そのため、シミュレーションデータでは正常な転写産物を融合遺伝子と誤って検出することへの対策は不要であった。一方、培養がん細胞のシークエンスデータは正常な転写産物由来のリードが大半を占めるため、上記の対策が必要不可欠である。実際、シークエンスエラーが原因で、正常な転写産物が偶然 2 つの遺伝子に部分的にアライメントされる事象が発生していた。それらのリードを全て融合遺伝子と判定すると偽陽性が大幅に増加し、検出性能が低下するという問題が発生した。

3.3.2　検出手法の変更点

シークエンスエラーは偶然に起こるため、正常な転写産物由来の 2 本以上のリードが、誤って同じ融合遺伝子から転写されたリードと判定される可能性は極めて低い。そのため、サポートリードが 1 本しかない融合遺伝子は検出対象外とし、サポートリードが 2 本以上の融合遺伝子を検出対象とした。

加えて、培養がん細胞のシークエンスデータでは正確な融合点が不明であり、遺伝子名で評価する必要があるため、融合遺伝子名とサポートリード数を出力する仕様に変更した。

上記 2 点の変更は、先行手法も同様に行い、同一条件で比較した。

3.4　実験

3.4.1　データと実験条件

3.4.1.1　培養がん細胞

融合遺伝子の検出性能を評価するために、MCF7 と SKBR3 の 2 種類の乳がん細胞株、4 つのデータセットを用いた（表 3.1）。MCF7-ONT-1 のデータは、大阪大学大学院医学系研究科乳腺・内分泌外科により、次の手順でシークエンスされた。本シークエンスデータは、Sequence Read Archive (SRA) のアクセッション：SRX22168673 から入手可能である。

ステップ 1:　アンプリコン作成

ステップ 2:　ライブラリー作成

ステップ 3:　シークエンス

ステップ 1 では、SMARTer PCR cDNA Synthesis Kit を用いてアンプリコンを作成した。試料増幅のためのプライマーは、ポリ A 型のプライマーである SMARTer II A Oligonucleotide が用いられた。ステップ 2 では、Ligation Sequencing Kit (SQK-LSK109) と Native Barcoding Expansion 1-12 (EXP-NBD104) を用いてアンプリコンからライブラリーを作成した。ステップ 3 では、MinION シークエンサーを用いてシークエンスを行い、塩基配列を決定した。フローセルには Flow Cell R9.4.1（FLO-MIN106D）を用いた。

残りの 3 つのデータセットは、FusionSeeker[38] の論文中で融合遺伝子の検出性能の評価に用いられているデータセットである。次の通りに公開データより入手した。MCF7-ONT-2 は https://github.com/GoekeLab/sg-nex-data/より入手した。MCF7-Pac のデータセットは SRA よりアクセッション：SRP055913 で入手し、SKBR3-Pac のデータセットは SRA よりアクセッション：SRP150606 で入手した。

各データセットのリード数および平均リード長の情報を示す（表 3.1）。リード数は最も少ないデータセットでは 2,389,856 本、最も多いデータセットでは 10,285,910 本と 4 倍以上の差があった。また、平均リード長は最も少ないデータセットでは 691 塩基、最も多いデータセットでは 3,552 本と 5 倍以上の差があった。

表 3.1　培養がん細胞株データセットの概要

データセット名	細胞株	シークエンサー	リード数	平均リード長
MCF7-ONT-1	MCF7	Nanopore	4,181,740	1,281
MCF7-ONT-2	MCF7	Nanopore	10,285,910	691
MCF7-Pac	MCF7	PacBio Iso-Seq	2,389,856	1,741
SKBR3-Pac	SKBR3	PacBio Iso-Seq	3,070,545	3,552

3.4.1.2　既知の融合遺伝子

MCF7 と SKBR3 の既知融合遺伝子リストは、FusionSeeker[38] の論文の付録から取得した。MCF7 は 34 個、SKBR3 は 28 個を既知の融合遺伝子として使用した。

3.4.1.3　先行方法との比較方法

提案手法の融合遺伝子の検出性能を評価するため、比較対象として JAFFAL と Fusion-Seeker を用いた。JAFFAL および FusionSeeker の実行方法やオプションは、第 2 章で記載した手順で実行した （2.5.1.5 節を参照)。

先行手法との比較では、2 本以上のリードでサポートされる融合遺伝子を検出対象とし、検出した融合遺伝子が既知の融合遺伝子リストにあれば真陽性 (TP)、既知の融合遺伝子リストになければ偽陽性 (FP)、既知の融合遺伝子リストの中で検出できなかった融合遺伝子は偽陰性 (FN) とした。既知の融合遺伝子リストとの照合の際、シノニム（同じ遺伝子を意味するが Gencode のバージョンの違いによって別名で登録されている遺伝子）は同一の遺伝子として扱った。つまり、遺伝子 A のシノニムである遺伝子 a と遺伝子 B の融合を検出したケースにおいて、既知の融合遺伝子リストに遺伝子 A と遺伝子 B が登録されている場合は正解と判定した。本論文中では、VPS35L と C16orf62 をシノニムとして扱い、C16orf62-IOCK が検出できていれば正解と判定した。

各手法の Precision、Recall、F1 スコアを算出し、Precision と Recall の加重平均である F1 スコアの比較により、総合的な融合遺伝子の検出性能を評価した。

3.4.2　実験結果

3.4.2.1　がん細胞株シークエンスデータにおける融合遺伝子の検出性能

複数のがん細胞株のシークエンスデータを用いて、融合遺伝子検出の性能を評価した (表 3.2、表 3.3)。全てのデータセットにおいて、 F1 スコアは FUGAREC、JAFFAL、FusionSeeker の順で高かった。また、FUGAREC の F1 スコアは、 MCF7-Pac のデータセットを除いて 50% を超えていた。一方で、MCF7-Pac のデータセットでは F1 スコアは極端に低かったが、他の手法も同様の傾向であった。

表 3.2　各細胞株における真陽性と偽陽性の数

	TP			FP			FN		
Dataset	F	J	S	F	J	S	F	J	S
MCF7-ONT-1	**22**	**22**	19	**26**	50	93	**12**	**12**	15
MCF7-ONT-2	**21**	20	17	**19**	**19**	52	**13**	14	17
MCF7-Pac	23	**25**	19	**121**	149	177	11	**9**	15
SKBR3-Pac	12	12	**14**	**3**	21	40	16	16	**14**

表 3.3　各細胞株における融合遺伝子の検出性能

	Precision (%)			Recall (%)			F1 (%)		
Dataset	F	J	S	F	J	S	F	J	S
MCF7-ONT-1	**45.8**	30.6	17.0	**64.7**	**64.7**	55.9	**53.7**	41.5	26.0
MCF7-ONT-2	**52.5**	51.3	24.6	**61.8**	58.8	50.0	**56.8**	54.8	33.0
MCF7-Pac	**16.0**	14.4	9.7	67.6	**73.5**	55.9	**25.8**	24.0	16.5
SKBR3-Pac	**80.0**	36.4	25.9	42.9	42.9	**50.0**	**55.8**	39.3	34.1

3.4.2.2　検出した融合遺伝子の重複

各手法で検出した融合遺伝子の重複を調査した（図 3.1）。まずは手法間で共通して検出した融合遺伝子について集計した。3 手法で共通して検出した融合遺伝子は、MCF7-ONT-1、MCF7-ONT-2、MCF7-Pac、SKBR3-Pac のデータセットで、それぞれ 28、22、32、10 だけ存在し、そのうち 19、16、17、10 が正解であった。手法間の共通性に関して、FUGAREC と JAFFAL は検出した融合遺伝子の重複が多い傾向があり、FusionSeeker は他の 2 つの手法との重複が少ない傾向があった。

次に、各手法が唯一検出した融合遺伝子を集計した。FUGAREC は MCF7-ONT-1、MCF7-ONT-2、MCF7-Pac、SKBR3-Pac のデータセットで、それぞれ 4、9、67、0 の融合遺伝子を 3 手法の中で唯一検出し、1、2、0、0 が正解であった。JAFFAL は MCF7-ONT-1、MCF7-ONT-2、MCF7-Pac のデータセットで、それぞれ 4、9、67、15 の融合遺伝子を 3 手法の中で唯一検出し、1、1、1、0 が正解であった。FusionSeeker は MCF7-ONT-1、MCF7-ONT-2、MCF7-Pac のデータセットでそれぞれ 79、44、123、43 の融合遺伝子を

3 手法の中で唯一検出し、0、0、1、4 が正解であった。

次に、複数のツールで検出した融合遺伝子のうち、既知の融合遺伝子リストにない融合遺伝子の候補を集計した。3 種類のツールで検出したが、既知の融合遺伝子リストに存在しなかった融合遺伝子の数は、MCF7-ONT-1、MCF7-ONT-2、MCF7-Pac、SKBR3-Pac のデータセットで、それぞれ 9、6、15、0 の 30 個で重複を除くと 26 種類であった。FUGAREC と JAFFAL で検出したが、既知の融合遺伝子リストに存在しなかった融合遺伝子の数は、MCF7-ONT-1、MCF7-ONT-2、MCF7-Pac、SKBR3-Pac のデータセットで、それぞれ 10、5、25、3 の 43 個で 43 種類であった。FUGAREC と FusionSeeker で検出したが、既知の融合遺伝子リストに存在しなかった融合遺伝子の数は、MCF7-ONT-1、MCF7-ONT-2、MCF7-Pac、SKBR3-Pac のデータセットで、それぞれ 4、1、14、0 の 19 個で重複を除くと 18 種類であった。JAFFAL と FusionSeeker で検出したが、既知の融合遺伝子リストに存在しなかった融合遺伝子の数は、MCF7-ONT-1、MCF7-ONT-2、MCF7-Pac、SKBR3-Pac のデータセットで、それぞれ 1、2、26、2 の 31 個で 31 種類であった。

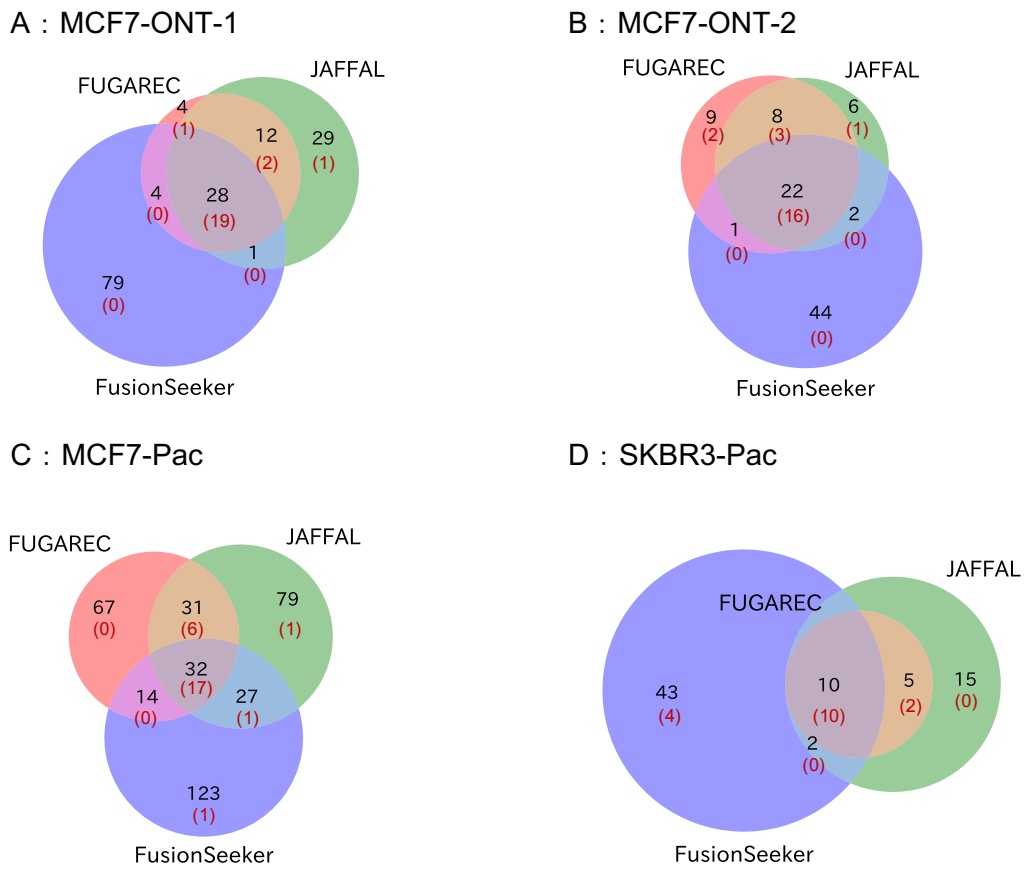

図 3.1　各手法で検出した融合遺伝子数の重複（黒字：検出した融合遺伝子数、赤字：正解の融合遺伝子数）

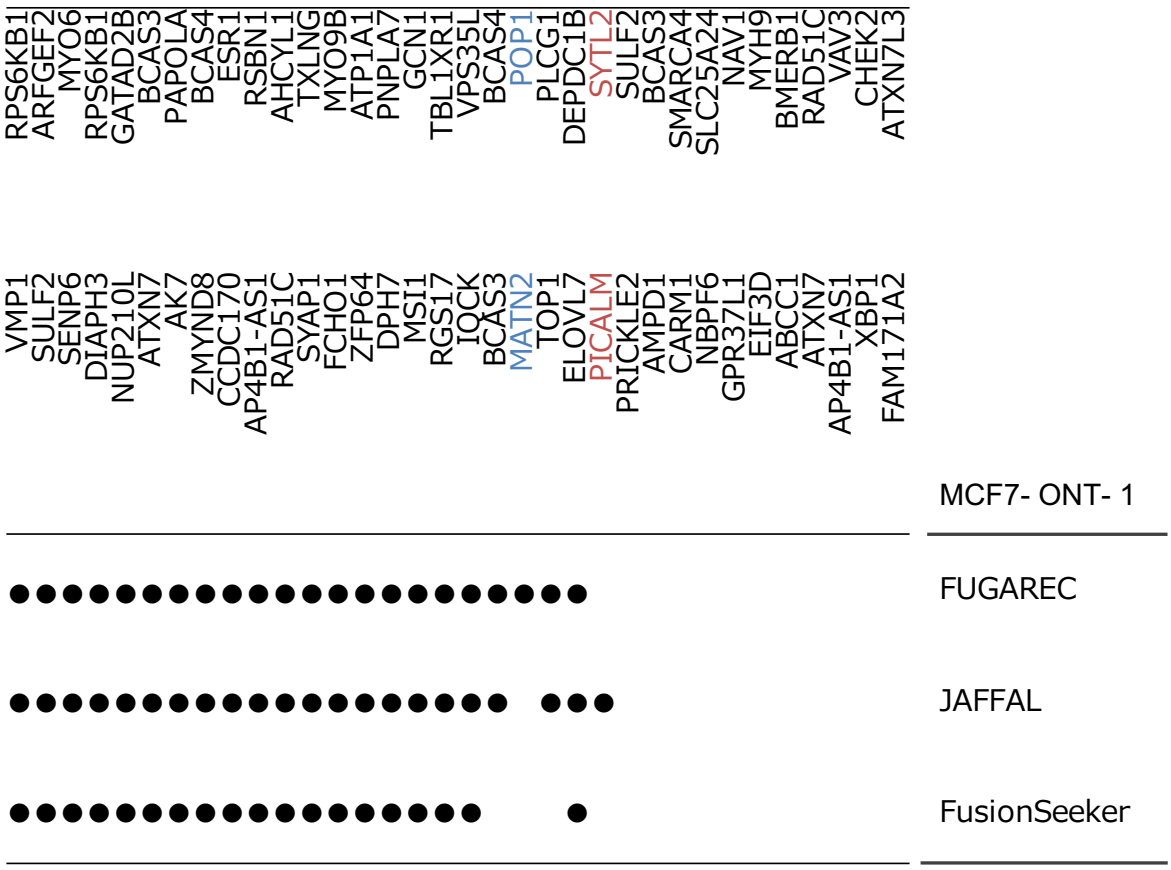

図 3.2　MCF7-ONT-1 の星取表

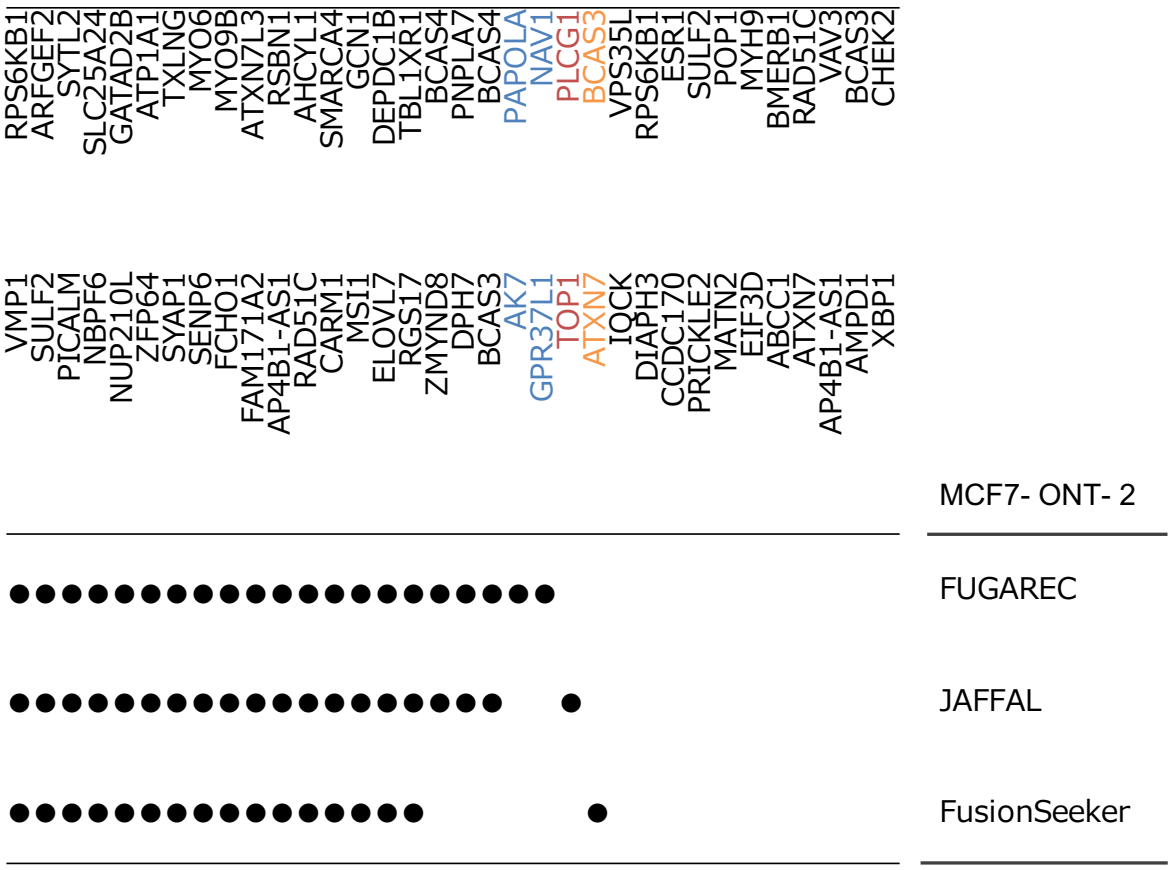

図 3.3　MCF7-ONT-2 の星取表

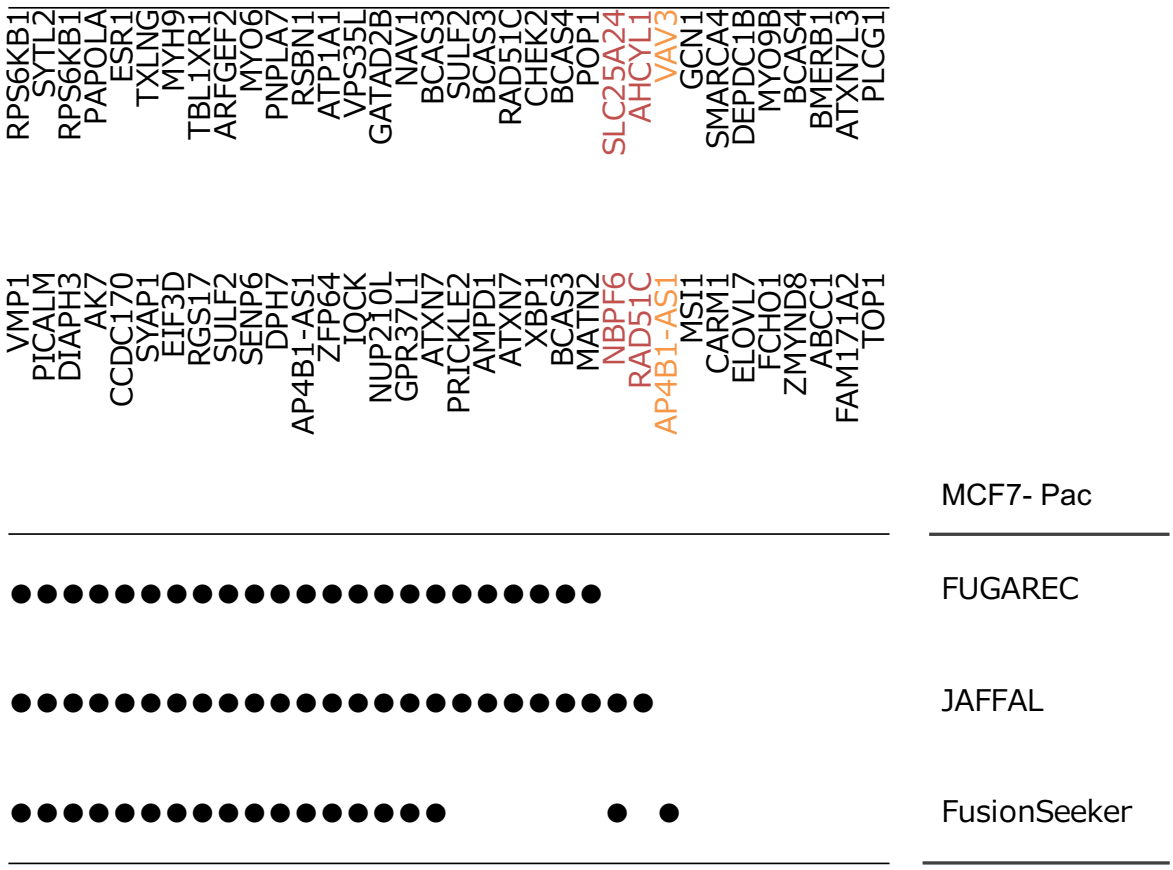

図 3.4　MCF7-Pac の星取表

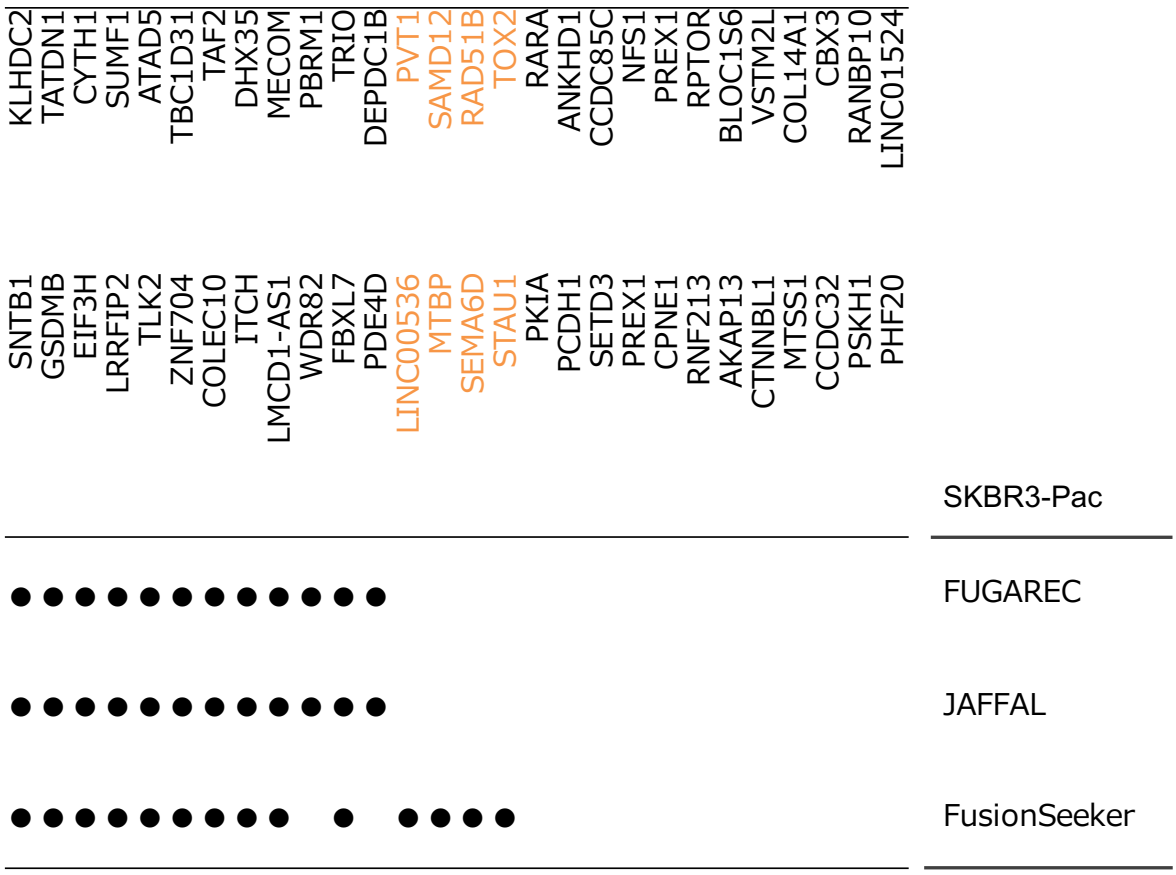

図 3.5　SKBR3-Pac の星取表

3.4.2.3　FUGAREC で唯一検出できた融合遺伝子

FUGAREC で唯一検出できた融合遺伝子は 3 種類あり、NAV1-GPR37L1、PAPOLA-AK7、POP1–MATN2 であった（図 3.2 、図 3.3)。NAV1-GPR37L1 と PAPOLA-AK7 は、アライメントの際にギャップが発生し、JAFFAL ではフィルタリングされ検出できなかった（表 3.4)。NAV1-GPR37L1 と PAPOLA-AK7 の元々のギャップ長はそれぞれで 56 と 32 であったが、エキソン境界固定とギャップの再アライメントでギャップを補正することでギャップ長は 3 まで減少し、FUGAREC では検出することができた（表 3.5)。POP1–MATN2 については、2 つの融合点間の距離が 200,000 塩基位内であり、ゲノムの組み換えの検出を想定する JAFFAL ではフィルタリングされて検出できなかった (表 3.6)。一方で、FUGAREC では融合点間の距離に制限は設けていないため検出することができた。

表 3.4　ギャップ長が原因で JAFFAL で検出できなかった融合遺伝子の実例

	5' 側ロングリード		3' 側ロングリード		
融合遺伝子	開始座標	終了座標	開始座標	終了座標	ギャップ長
NAV1–GPR37L1	59	**142**	**198**	1383	**56**
PAPOLA–AK7	25	**128**	**160**	997	**32**

表 3.5　ギャップを補正して FUGAREC で検出できた融合遺伝子の実例

	5' 側ロングリード		3' 側ロングリード		
融合遺伝子	開始座標	終了座標	開始座標	終了座標	ギャップ長
NAV1–GPR37L1	59	**163**	**166**	1383	**3**
PAPOLA–AK7	25	**134**	**137**	997	**3**

表 3.6　融合点間の距離が原因で JAFFAL で検出できなかった実例

	5' 側の遺伝子の融合点		3' 側の遺伝子の融合点		
融合遺伝子	染色体	座標	染色体	座標	融合点間の距離
POP1–MATN2	Chr8	99,042,689	Chr8	99,129,618	86,929

3.4.2.4　データセット間で共通して検出した融合遺伝子

複数のデータセット間で共通して検出した融合遺伝子は真に存在する可能性が高い。そこで FUGAREC で 4 種類のデータセットのうち 2 種類以上で共通して検出した融合遺伝子を調査した。2 種類以上のデータセットで検出され、既知の融合遺伝子リストにも存在していた融合遺伝子は 23 種類であった。一方、既知の融合遺伝子リストに存在しない融合遺伝子で 3 つのデータセットで検出された融合遺伝子は 1 種類、2 つのデータセットで検出された 9 種類であった（表 3.7）。

表 3.7　データセット間で共通して検出した既知の融合遺伝子リストにない融合遺伝子

融合遺伝子	MCF7-ONT-1	MCF7-ONT-2	MCF7-Pac	SKBR3-Pac
AC099850.1–VMP1	1	1	1	0
AC005808.3–NCOA3	1	1	0	0
ATP9A–RP11-347D21.2	1	1	0	0
BCAS3–PARD6B	0	1	1	0
CDKN2B-AS1–MTAP	1	1	0	0
EIF4E2–GIGYF2	0	1	1	0
GLCCI1–RPA3-AS1	1	1	0	0
GSN–SMIM22	1	1	0	0
RP11-446N19.1–RP11-96H19.1	1	0	1	0
RP11-977B10.2–SLC16A7	1	1	0	0

3.4.3　考察

本章では、複数の培養がん細胞株のシークエンスデータを用いて FUGAREC の融合遺伝子の検出性能を評価した。FUGAREC は全てのデータセットで F1 スコアが最も高く、他の手法よりも高い検出能力を有していることを示した。FUGAREC の F1 スコアは MCF7-Pac のデータセットを除いて 50% を超えていた。一方で、MCF7-Pac のデータセットにおける F1 スコアは極端に低かったが、他の手法も同様の傾向であった。これは MCF7-Pac のデータが PacBio RS という初期に開発されたシークエンサーでシークエンスされたデータであることが原因と考えられる。他のデータセットは比較的新しく、MCF7-Pac だけが他のデータセットとは性質が異なっていたと考えられる。

FUGAREC の F1 スコアが高かった理由には、他の手法と比較して偽陽性が少ないことが挙げられる。JAFFAL が誤って検出した融合遺伝子の 6 割以上は、融合点をエキソン境界に補正するステップで、偽陽性と判断して除外できていた。これは第 2 章の考察で実例を挙げて説明した原理と同じである（図 3.6）。この結果は、シミュレーションデータだけでなく、がん細胞株のシークエンスデータにおいても、融合点をエキソン境界に補正することが、偽陽性を削減するのに重要なステップであることを支持している。

一方、FUGAREC で検出できず、JAFFAL で検出できた融合遺伝子も存在した。その融合遺伝子は 4 種類で、SYTL2–PICALM、SLC25A24–NBPF6、PLCG1–TOP1、AHCYL1–RAD51C であった（図 3.2、図 3.3、図 3.4）。前半 3 つの融合遺伝子はトラン

スクリプトームへのアライメントの段階でフィルタリングされたことが原因で、残りの融合遺伝子はサポートリード 2 本以上の条件を満たさなかったことが原因で検出することができなかった。SYTL2–PICALM、SLC25A24–NBPF6、PLCG1–TOP1 はそれぞれ該当の遺伝子以外で ENPP1、MB、RP1-1J6.2 にアライメントされていた。FUGAREC は偽陽性を減らすため、2 つの遺伝子に部分的にアライメントされるリードのみから融合遺伝子を検出し、3 つの遺伝子に部分的にアライメントされるリードは検出対象外としている。そのため、FUGAREC では SYTL2–PICALM、SLC25A24–NBPF6、PLCG1–TOP1 の融合遺伝子を検出することができなかった。3 種類の遺伝子に部分的にアライメントするリードから、真の融合リードのみを検出することは、FUGAREC の今後の課題である。

AHCYL1–RAD51C については、サポートリードが 2 本という条件を満たせずに検出することができなかった。FUGAREC では、同じ遺伝子の融合であっても融合点がエキソン 1 個分ずれていれば別の融合とみなし、サポートリードを 1 として扱っている。AHCYL1 は chr1:110546752-110547073 と chr1:110527307-110527794 にエキソンを保有しており、FUGAREC では融合点をそれぞれのエキソン境界に補正した。そのため、AHCYL1 と RAD51C の 2 つの遺伝子に部分的にアライメントされるリードは 2 本存在したが、FUGAREC ではサポートリードを 1 本と判定して検出することができなかった。ゲノム上の近い距離に複数のエキソンを持つような遺伝子においては、今回のように誤ったエキソンの境界に融合点を補正してしまい、検出できない可能性がある。この問題に対処することは FUGAREC の今後の課題である。

本研究では、各手法が検出した融合遺伝子に対し、既知の融合遺伝子リストに存在すれば正解、存在しなければ不正解として評価した。しかし、既知の融合遺伝子リストが全ての融合遺伝子を網羅できているとは限らない。そのため、各手法から検出されて不正解と判定された融合遺伝子の中に真の融合遺伝子が含まれている可能性がある。特に、1 つのデータセットから 2 種類以上の手法で検出した融合遺伝子や、複数のデータセット間で共通して検出した融合遺伝子は真である可能性がある。

3 種類の手法で検出したが、既知の融合遺伝子リストに存在しなかった融合遺伝子は 26 種類であった（3.4.2.2 節を参照）。これらは、既知の融合遺伝子リストが捕捉できていない可能性が高く、PCR で存在の有無を確認すべきである。そうすることで、新規の融合遺伝子を発見できる可能性がある。また、2 種類の手法で共通する融合遺伝子も次点での新規の融合遺伝子候補となる。既知の融合遺伝子リストに存在しなかった融合遺伝子で、FUGAREC と JAFFAL でのみ共通して検出した融合遺伝子は 43 種類、FUGAREC

と FusionSeeker でのみ共通は 18 種類、JAFFAL と FusionSeeker でのみ共通は 31 種類であった（3.4.2.2 節を参照）。JAFFAL と FusionSeeker だけを用いた場合と比較して、FUGAREC を使用することで、新規の融合遺伝子の候補を約 2 倍に増やすことができた。これは FUGAREC の成果の 1 つと考える。

また、FUGAREC で複数のデータセット間で共通して検出したが、既知の融合遺伝子リストに存在しなかった融合遺伝子は合計 10 種類存在した（3.4.2.4 節を参照）。いずれも大規模な融合遺伝子のデータベースである Mitelman Database[44] に掲載されておらず、新規の融合遺伝子の可能性がある。PCR での検証は今後の研究課題としたい。

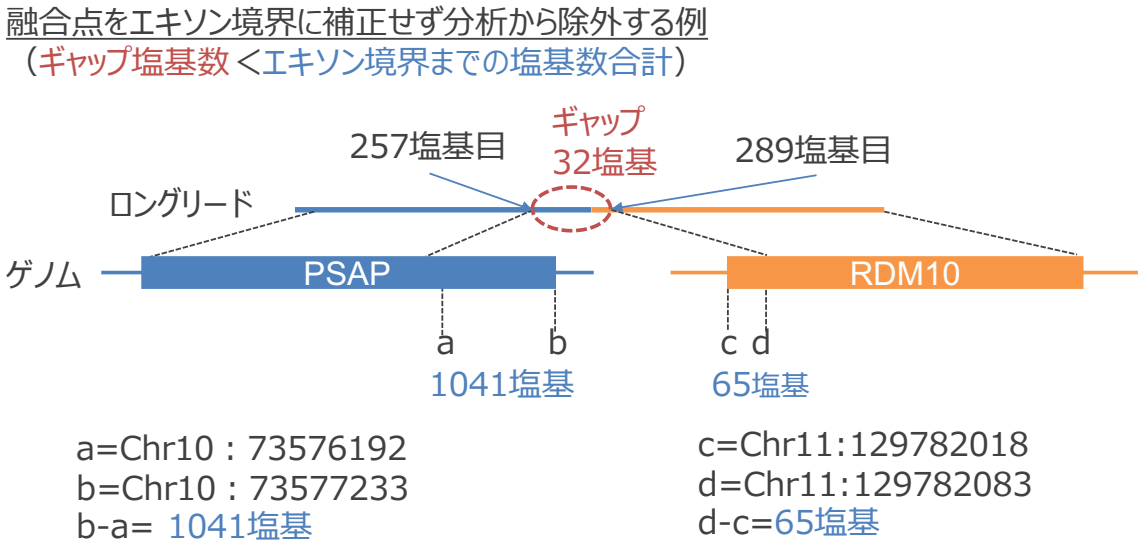

図 3.6　融合点のエキソン境界固定による偽陽性リードのフィルタリング実例

3.5　結言

本章では、複数の培養がん細胞のシークエンスデータを用いて、FUGAREC と先行手法の融合遺伝子の検出性能を比較した。その結果、FUGAREC の融合遺伝子の検出性能が最も高いことを示した。また、FUGAREC を用いることで、先行手法では注目できない新規の融合遺伝子の候補を提示することができた。加えて、先行手法では検出できなかったゲノム上の近い距離の融合遺伝子を検出することができた。一方、3 種類以上の遺伝子に部分的にアライメントされるリードから真の融合遺伝子を検出すること、ゲノム上の近い距離に複数のエキソンを持つような遺伝子に対しても正しい融合点を特定することは、今後の課題である。

第 4 章　結論

本研究は、ギャップが原因で融合遺伝子および融合点を正確に検出できず、融合遺伝子の機能解析が十分に行えない問題、ゲノム上での距離が近い遺伝子の融合遺伝子を検出できない問題に着目した。先行手法の問題を解決する手法を提案し、新規の融合遺伝子の発見につなげること、融合遺伝子の機能解明に貢献することを目的とした。

提案手法は、先行手法に対して、融合点のエキソン境界固定、ギャップの再アライメント、融合点のクラスタリングの 3 つの処理を追加して拡張した。また、ゲノム上での距離が近い遺伝子の融合遺伝子も検出するため、融合点間の距離の制約は設けないようにすることで先行手法の問題点の解決を試みた。

第 2 章では、シミュレーションデータを用いて、提案手法の融合遺伝子の検出性能を先行手法と比較し、提案手法が先行手法よりも高い融合遺伝子の検出性能を有していることを示した。また、融合点の正確な補正が可能となったことを示した。融合遺伝子の機能を解析するためには、融合遺伝子のアミノ酸配列を正しく予測できる必要があるが、先行手法ではギャップのせいで融合点が正確に求められず、融合遺伝子の機能解析の妨げとなっていた。提案手法を用いることで、融合遺伝子の機能解明を効率的に行うことができる。研究目的の 1 つである、融合遺伝子の機能解明への貢献の可能性は示せたと考える。

第 3 章では、第 2 章で提案した手法をベースに、がん細胞株から融合遺伝子を検出できるように手法を拡張した。複数の培養がん細胞株を用いて、先行手法と融合遺伝子の検出性能を比較し、先行手法よりも提案手法が高い融合遺伝子の検出性能を有していることを示した。また、提案手法を用いることで、先行手法では注目できない新規の融合遺伝子の候補を提示できること、先行手法では検出できなかったゲノム上の近い距離の融合遺伝子を検出できることを示した。これらの融合遺伝子は存在有無を PCR で確認することで、新たな融合遺伝子を発見できる可能性がある。実験的な検証は今後の研究課題であるが、研究目的の 1 つである、新規の融合遺伝子を発見できる可能性は示せたと考える。

これまで、融合遺伝子はがんの病態に伴って、ゲノムの組み換えの結果起こると考えられてきたが、ゲノムの組み換え非依存的に発生する融合遺伝子が存在し、がんの細胞だけでなく、正常な細胞にも存在することが明らかとなってきた。一方、そのような融合遺伝子が正常な細胞でどのような役割を果たしているかは十分に研究されていない。がん細胞における融合遺伝子の正確な検出だけでなく、融合遺伝子が正常な細胞でどのような役割を果たしているかの解明に提案手法が活用されることを期待する。

参考文献

[1] R. Dorney, B. P. Dhungel, J. E. Rasko, L. Hebbard, and U. Schmitz. Recent advances in cancer fusion transcript detection. *Briefings in Bioinformatics*, Vol. 24, No. 1, p. bbac519, 2023.

[2] J. D. Rowley. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature*, Vol. 243, No. 5405, pp. 290–293, 1973.

[3] C. Tognon, S. R. Knezevich, D. Huntsman, C. D. Roskelley, N. Melnyk, J. A. Mathers, L. Becker, F. Carneiro, N. MacPherson, D. Horsman, et al. Expression of the ETV6-NTRK3 gene fusion as a primary event in human secretory breast carcinoma. *Cancer Cell*, Vol. 2, No. 5, pp. 367–376, 2002.

[4] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X.-W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, Vol. 310, No. 5748, pp. 644–648, 2005.

[5] S. V. Williams, C. D. Hurst, and M. A. Knowles. Oncogenic FGFR3 gene fusions in bladder cancer. *Human Molecular Genetics*, Vol. 22, No. 4, pp. 795–803, 2013.

[6] S. Seshagiri, E. W. Stawiski, S. Durinck, Z. Modrusan, E. E. Storm, C. B. Conboy, S. Chaudhuri, Y. Guan, V. Janakiraman, B. S. Jaiswal, et al. Recurrent R-spondin fusions in colon cancer. *Nature*, Vol. 488, No. 7413, pp. 660–664, 2012.

[7] J. Salzman, R. J. Marinelli, P. L. Wang, A. E. Green, J. S. Nielsen, B. H. Nelson, C. W. Drescher, and P. O. Brown. ESRRA-C11orf20 is a Recurrent Gene Fusion in Serous Ovarian Carcinoma. *PLoS Biology*, Vol. 9, No. 9, p. e1001156, 2011.

[8] M. Soda, Y. L. Choi, M. Enomoto, S. Takada, Y. Yamashita, S. Ishikawa, S.-i. Fujiwara, H. Watanabe, K. Kurashina, H. Hatanaka, et al. Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. *Nature*, Vol. 448, No. 7153, pp. 561–566, 2007.

[9] C. Turc-Carel, I. Philip, M. Berger, T. Philip, and G. Lenoir. Chromosomal translocation (11; 22) in cell lines of Ewing's sarcoma. *Comptes Rendus des Seances de L'academie des sciences. Serie III, Sciences de la vie*, Vol. 296, No. 23,

pp. 1101–1103, 1983.

[10] F. Mitelman, B. Johansson, and F. Mertens. The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, Vol. 7, No. 4, pp. 233–245, 2007.

[11] J. N. Honeyman, E. P. Simon, N. Robine, R. Chiaroni-Clarke, D. G. Darcy, I. I. P. Lim, C. E. Gleason, J. M. Murphy, B. R. Rosenberg, L. Teegan, et al. Detection of a recurrent DNAJB1-PRKACA chimeric transcript in fibrolamellar hepatocellular carcinoma. *Science*, Vol. 343, No. 6174, pp. 1010–1014, 2014.

[12] R. Berthold, I. Isfort, C. Erkut, L. Heinst, I. Grünewald, E. Wardelmann, T. Kindler, P. Åman, T. G. Grünewald, F. Cidre-Aranaz, et al. Fusion protein-driven IGF-IR/PI3K/AKT signals deregulate Hippo pathway promoting oncogenic cooperation of YAP1 and FUS-DDIT3 in myxoid liposarcoma. *Oncogenesis*, Vol. 11, No. 1, p. 20, 2022.

[13] S. Kuravi, R. W. Baker, M. U. Mushtaq, I. Saadi, T. L. Lin, C. J. Vivian, A. Valluripalli, S. Abhyankar, S. Ganguly, W. Cui, et al. Functional characterization of NPM1–TYK2 fusion oncogene. *NPJ Precision Oncology*, Vol. 6, No. 1, p. 3, 2022.

[14] B. J. Druker, M. Talpaz, D. J. Resta, B. Peng, E. Buchdunger, J. M. Ford, N. B. Lydon, H. Kantarjian, R. Capdeville, S. Ohno-Jones, et al. Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia. *New England Journal of Medicine*, Vol. 344, No. 14, pp. 1031–1037, 2001.

[15] B. J. Druker. Imatinib as a paradigm of targeted therapies. *Advances in Cancer Research*, Vol. 91, No. 1, pp. 1–30, 2004.

[16] A. T. Shaw, B. Y. Yeap, B. J. Solomon, G. J. Riely, J. Gainor, J. A. Engelman, G. I. Shapiro, D. B. Costa, S.-H. I. Ou, M. Butaney, et al. Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis. *The Lancet Oncology*, Vol. 12, No. 11, pp. 1004–1012, 2011.

[17] L. J. Wirth, E. Sherman, B. Robinson, B. Solomon, H. Kang, J. Lorch, F. Worden, M. Brose, J. Patel, S. Leboulleux, et al. Efficacy of Selpercatinib in RET-

Altered Thyroid Cancers. *New England Journal of Medicine*, Vol. 383, No. 9, pp. 825–835, 2020.

[18] B. C. Parker and W. Zhang. Fusion genes in solid tumors: an emerging target for cancer diagnosis and treatment. *Chinese Journal of Cancer*, Vol. 32, No. 11, p. 594, 2013.

[19] M. Vincent, M. Kuruvilla, N. Leighl, and S. Kamel-Reid. Biomarkers that currently affect clinical practice: EGFR, ALK, MET, KRAS. *Current Oncology*, Vol. 19, No. s1, pp. 33–44, 2012.

[20] D. T. Yeung and T. P. Hughes. Therapeutic targeting of BCR-ABL: prognostic markers of response and resistance mechanism in chronic myeloid leukaemia. *Critical Reviews™ in Oncogenesis*, Vol. 17, No. 1, 2012.

[21] A. M. Chinnaiyan and N. Palanisamy. Chromosomal Aberrations in Solid Tumors. *Progress in Molecular Biology and Translational Science*, Vol. 95, pp. 55–94, 2010.

[22] 日本肺癌学会. 肺癌患者における ros1 融合遺伝子検査の手引き, 2017.

[23] Q. Wang, J. Xia, P. Jia, W. Pao, and Z. Zhao. Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. *Briefings in Bioinformatics*, Vol. 14, No. 4, pp. 506–519, 2013.

[24] H. Nakagawa and M. Fujita. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Science*, Vol. 109, No. 3, pp. 513–522, 2018.

[25] A. Sboner, L. Habegger, D. Pflueger, S. Terry, D. Z. Chen, J. S. Rozowsky, A. K. Tewari, N. Kitabayashi, B. J. Moss, M. S. Chee, et al. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biology*, Vol. 11, pp. 1–19, 2010.

[26] K. Sahlin and P. Medvedev. Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. *Nature Communications*, Vol. 12, No. 1, p. 2, 2021.

[27] B. J. Haas, A. Dobin, B. Li, N. Stransky, N. Pochet, and A. Regev. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biology*, Vol. 20, No. 1, pp. 1–16,

2019.

[28] D. Kim and S. L. Salzberg. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology*, Vol. 12, No. 8, pp. 1–15, 2011.

[29] S. Uhrig, J. Ellermann, T. Walther, P. Burkhardt, M. Fröhlich, B. Hutter, U. H. Toprak, O. Neumann, A. Stenzinger, C. Scholl, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Research*, Vol. 31, No. 3, pp. 448–460, 2021.

[30] R. Dorney, B. P. Dhungel, J. E. J. Rasko, L. Hebbard, and U. Schmitz. Recent advances in cancer fusion transcript detection. *Briefings in Bioinformatics*, Vol. 24, No. 1, 12 2022. bbac519.

[31] N. M. Davidson, I. J. Majewski, and A. Oshlack. JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Medicine*, Vol. 7, No. 1, pp. 1–12, 2015.

[32] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, Vol. 323, No. 5910, pp. 133–138, 2009.

[33] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, Vol. 4, No. 4, pp. 265–270, 2009.

[34] A. Rhoads and K. F. Au. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, Vol. 13, No. 5, pp. 278–289, 2015.

[35] F. J. Rang, W. P. Kloosterman, and J. de Ridder. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, Vol. 19, No. 1, pp. 1–11, 2018.

[36] Q. Liu, Y. Hu, A. Stucky, L. Fang, J. F. Zhong, and K. Wang. LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. *BMC Genomics*, Vol. 21, No. 11, pp. 1–12, 2020.

[37] N. M. Davidson, Y. Chen, T. Sadras, G. L. Ryland, P. Blombery, P. G. Ekert, J. Göke, and A. Oshlack. JAFFAL: detecting fusion genes with long-read transcriptome sequencing. *Genome Biology*, Vol. 23, No. 1, pp. 1–20, 2022.

[38] Y. Chen, Y. Wang, W. Chen, Z. Tan, Y. Song, H. G. S. V. Consortium, H. Chen, and Z. Chong. Gene Fusion Detection and Characterization in Long-read Cancer Transcriptome Sequencing Data with FusionSeeker. *Cancer Research*, Vol. 83, No. 1, pp. 28–33, 2023.

[39] K. Masuda, Y. Sota, and H. Matsuda. A Novel Method for Fusion Gene Detection using Both End-Fragment Sequences of Long Reads. In *Proceedings of the 2022 9th International Conference on Biomedical and Bioinformatics Engineering (ICBBE '21)*, pp. 88–92. ACM, 2022.

[40] H. Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, Vol. 34, No. 18, pp. 3094–3100, 2018.

[41] W. Li, R. Wan, L. Guo, G. Chang, D. Jiang, L. Meng, and J. Ying. Reliability analysis of exonic-breakpoint fusions identified by DNA sequencing for predicting the efficacy of targeted therapy in non-small cell lung cancer. *BMC Medicine*, Vol. 20, No. 1, p. 160, 2022.

[42] W. J. Kent. BLAT—the BLAST-like alignment tool. *Genome Research*, Vol. 12, No. 4, pp. 656–664, 2002.

[43] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Research*, Vol. 22, No. 9, pp. 1760–1774, 2012.

[44] Mitelman F, Johansson B, and Mertens F(Eds.). Mitelman database of chromosome aberrations and gene fusions in cancer (2024). *https://mitelmandatabase.isb-cgc.org.*

謝辞

本研究を進めるにあたり、多くの方にお力添えをいただきました。

指導教員として、研究の進め方から研究の細部の部分まで丁寧に見ていただき、終始暖かいご指導をいただいた大阪大学大学院情報科学研究科バイオ情報工学専攻 松田 秀雄教授に厚く御礼を申し上げます。

学位審査委員を務めていただき、研究のまとめ方について貴重なご意見をいただいた大阪大学大学院情報科学研究科バイオ情報工学専攻 松田 史生 教授、小蔵 正輝 准教授、 瀬尾 茂人 准教授、 大阪大学大学院医学系研究科外科学講座（乳腺・内分泌外科学） 下田 雅史 准教授に心から感謝いたします。

様々な局面において、的確な助言をいただいた大阪大学大学院医学系研究科外科学講座（乳腺・内分泌外科学） 草田 義昭 助教に心より御礼申し上げます。

大阪大学大学院情報科学研究科バイオ情報工学専攻 繁田 浩功 助教には研究環境の構築で多大なる援助をしていただきました。心から感謝いたします。