OUKA
The University of Osaka
Institutional Knowledge Archive

Title	乳癌トランスクリプトームデータを用いた遺伝子発現と融合遺伝子の解析手法に関する研究
Author(s)	草田, 義昭
Citation	大阪大学, 2022, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/88158
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

乳癌トランスクリプトームデータを用いた
遺伝子発現と融合遺伝子の
解析手法に関する研究

提出先　　大阪大学大学院情報科学研究科

提出年月　2022年1月

草田　義昭

関連研究論文

1. 査読のある学術論文

   1-1 Yoshiaki Sota, Shigeto Seno, Hironori Shigeta, Naoki Osato, Masafumi Shimoda, Shinzaburo Noguchi and Hideo Matsuda. Improvement of detection performance of fusion genes from RNA-seq data by clustering short reads. Journal of Bioinformatics and Computational Biology, 17 (03), 2019, 1940008 (博士論文第 3 章)

2. 査読のある国際会議

   2-1 Yoshiaki Sota, Shigeto Seno, Yoichi Takenaka, Shinzaburo Noguchi, and Hideo Matsuda. Comparative analysis of transformation methods for gene expression profiles in breast cancer datasets. 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering, BIBE 2016, Tainan, Taiwan, November 1, 2016 (博士論文第 2 章)

内容梗概

遺伝子転写産物 (mRNA) 全てを扱うトランスクリプトームデータは、マイクロアレイや RNA シークエンス (RNA-Seq) などで一度に大量に計測する事が可能となり、発現解析や構造変異解析などに応用されている。また、それらの解析データは、日々公共のデータベース (Gene Expression Omnibus, GEO) に臨床データとともに登録されている。登録されたデータには、これまでのトランスクリプトーム解析が乳癌の治療成績に貢献したように新たな治療ターゲットであるバイオマーカーを見いだせる可能性がある。

本研究では、マイクロアレイや RNA-Seq の臨床応用やバイオマーカーの検出を困難にしている問題点にそれぞれ焦点を当てた。マイクロアレイの検討では、臨床と同じ逐次的な検体に対応できる前処理法を提案した。そして、提案手法の適用可能性を調べるためにデータベースの臨床因子とトランスクリプトームの発現値との一致率を向上させることを目標とした。RNA-Seq では、新たなバイオマーカーの一つである融合遺伝子の検出において前処理法を提案し、既存の手法による融合遺伝子の検出精度の向上を目標とした。本論文は以下の 4 章で構成されており、第 1 章では研究背景、目的、及び目標を述べた。

第 2 章では、マイクロアレイデータを臨床で用いるために課題となっている実験室や実験者などの相違に伴う誤差であるバッチ効果に対して、臨床的需要のある 1 検体単位でのバッチ補正法の開発を行った。バッチ効果は、新たなバイオマーカーを見出すための大きな障壁となっているだけでなく、見出したバイオマーカーや診断法を 1 検体単位でそのまま臨床診断として適用することを困難にしている。本研究では、シングルアレイによるバッチ補正法であるノンパラメトリック-Z-スケーリング (NPZ) 法を提案し、既存の手法に対して臨床因子との一致率の向上を目的とした。公共のデータベース GEO からエストロゲン受容体 (ER) とヒト上皮増殖因子受容体 2 (HER2) の免疫組織学的染色 (IHC) 法の結果を有する 2,817 症例のマイクロアレイの発現データを抽出した。続いて、マイクロアレイデータから背景補正及び対数変換のみを行ったもの (正規化なし) と既存の 3 つの正規化法 (MAS5、fRMA、及び RMX) の 4 通りのデータに対して、バッチ補正なしと 4 つのシングルアレイによるバッチ補正 (RANK、YuGene、Z、及び NPZ) とマルチアレイによるバッチ補正 (ComBat) の 6 通りの処理を加えて、4×6 で各々の ER と HER2 の IHC の結果と mRNA の発現の一致率を比較した。NPZ 法のバッチ補正を行うことで IHC と mRNA の発現の一致率は向上した。さらに ER 陽性リンパ節転移陰性乳癌の内分泌療法患者における mRNA による ER 陽性 HER2 陰性サブタイプとその他のサブタイ

プの再発曲線の解析においてバッチ補正なし (P=0.01) に対して NPZ 法 (P=1.14E-04) は予後をより明確に分けた。乳癌サブタイプによる治療効果の解析で最も化学療法感受性が高いとされる HER2 タイプの病理学的完全寛解率の比較において、バッチ補正なし (41.1%) に対して、NPZ 法 (44.6%) はより高い病理学的完全寛解率を有し治療効果を強調した。これらの結果から NPZ 法は腫瘍の特徴を維持した状態での逐次的なバッチ効果の補正に対して適用可能性を示した。

第 3 章では、乳癌細胞の RNA-Seq データから融合遺伝子の同定を行い、その検出精度の向上を目標とした。融合遺伝子は、遺伝子本来の機能を変化させうるため腫瘍形成の原因の一つであるとともに治療のターゲットであるバイオマーカーとして注目されている。しかしながら、RNA-Seq による融合遺伝子の同定の精度は十分ではない。その原因の一つが、RNA-Seq のリード長は比較的短いため、参照ゲノムの複数の場所にマッピングされることである。RNA-Seq では、発現量に応じて同一遺伝子から配列の一部が重複したリード（オーバーラッピングリード）が得られる。本研究では、オーバーラッピングリードの不一致部分を削除するためにリードの両端から数塩基をカットしたリード（シフテッドリード）を作成し、シフテッドリードの類似性を用いてリードをクラスタリングすることで、最終的に代表配列としてリードを伸長させるシフテッドショートリードクラスタリング (SSC) 法を提案した。そして SSC 法によって前処理されたリードを用いることで既存の融合遺伝子検出手法のリードのマッピング率が向上し、最終的に融合遺伝子の同定率が改善すると仮説を立てた。仮説を検証するために、4 つの細胞株 (BT-474、MCF-7、SKBR-3、及び T-47D) からの RNA-Seq データに SSC 法を適用し、非適用例と比較した。SSC 法にて 1 塩基だけシフトした場合にリードが平均 201 塩基から 217 塩基 (108%) に伸張し、参照ゲノムへのマッピング率が 89% から 94% へ改善した。既存の融合遺伝子検出手法である STAR-Fusion による融合遺伝子の同定率も 49% から 54% へと改善した。SSC 法は、マッピング率を向上させ、融合遺伝子同定率も改善させる有効な手法である可能性が示唆された。

第 4 章では、本研究で得られた成果を総括し、本研究の寄与と今後の展望について述べた。マイクロアレイの解析では、臨床的需要に即したデータの補正と生物学的特徴を維持したデータ統合の両立が求められており、本研究で提案した前処理法の NPZ 法は有力な一つの手法となりえることを示した。精度に課題のある RNA-Seq の融合遺伝子検出においても予め前処理としてリードを伸長させる SSC 法は、既存の手法の融合遺伝子検出の精度をさらに向上させその適用可能性を示した。日々蓄積されていくトランスクリプトー

ムデータに対して本研究で提案した手法は、いずれも前処理として働くため既存の手法だけでなく既存の手法と共通した入力形式で新たに開発される手法にもそのまま適応可能である。本研究の提案手法を適用することでこれらの手法の精度をより向上させる可能性があり、今後のさらなる精度の高いトランスクリプトーム解析に貢献することが期待される。

目次

図目次

表目次

第 1 章　緒論

1.1　乳癌における複合的な診断

乳癌は、年間 9 万人以上が罹患する (94,519 例 (男性 661 例、女性 93,858 例 全国がん登録罹患データ (罹患)[*1] より抜粋))、日本人女性で最も罹患率の高い疾患である。1960 年代以前の乳癌の診断および治療は、腫瘍の大きさや転移部位など臨床的因子を土台として行われていた。1960 年代にエストロゲン受容体の配列が同定され [1]、免疫組織化学染色法 (IHC) による診断が臨床で広く用いられるようになった。その結果、腫瘍自体に存在するタンパク質の違いによって異なる治療が行われるようになった。さらに 2000 年以降には、バイオインフォマティクスの進歩によって可能となった乳癌組織の網羅的な遺伝子発現による教師なしクラスター解析の結果、乳癌は大きく 4 つのサブタイプに再現性をもって自ずと分類されることが明らかとなったのである [2]。

トランスクリプトーム解析による乳癌サブタイプの知見は、乳癌は基本的に終末乳管小葉単位から発生し、その内側の内腔細胞 Luminal cell と外側の基底・筋上皮細胞 Basal cell、それらの中間に位置する Basoluminal を母体に発生したかどうかで臨床的特徴が異なっているとされていた [3, 4]IHC の知見を裏付ける結果であった。その臨床的特徴とは、以下の通りである。エストロゲン受容体やプロゲステロン受容体を発現する内腔細胞の特徴を多く有している Luminal サブタイプは、核異型が弱く予後良好だが、化学療法にあまり反応しない。その一方で、エストロゲン受容体やプロゲステロン受容体を発現しない基底細胞の特徴である Basal サブタイプは、核異型が強く予後不良だが、化学療法によく反応するという特徴をもっている [5]。乳癌サブタイプは、腫瘍の転写産物 (トランスクリプトーム) の発現を評価することで決定される。そのため分子診断が 2000 年代以降に臨床に応用され、臨床診断・IHC 診断と合わせて乳癌診療の基盤となっている。このように 1960 年代以前は、臨床診断による大きさなどの因子を用いた 1 つの乳癌として治療されていたが、IHC 法により、エストロゲン受容体陽性/陰性乳癌という 2 種類の乳癌として治療され、2000 年以降は、分子診断によって 4 種類の乳癌として治療されるに至っている。それらの分類に対応する治療法が開発されることによって乳癌の予後は確実に改善している [6]。しかし、現状で乳癌死をゼロにするまでには至っておらず、新たなバイオマーカー (病期の状態や治療効果の指標となる蛋白質や遺伝子などの生体内の物質) の

[*1] https://ganjoho.jp/reg_stat/statistics/data/dl/index.html#a14

探索が求められている。

1.2 腫瘍の微小環境

腫瘍は、腫瘍組織中の癌細胞がそれぞれ様々な要因によって変化すると考えられている。つまり、腫瘍が宿主に対して適応した結果として多様性が生まれ、腫瘍内の不均一性を形成していると考えられている。その様々な要因は、Hanahan 等によって”Hallmarks of Cancer”という概念として現在広く定着している。その要因は、(1) ゲノムの不安定化と変異 Genome instability and mutation、 (2) 無制限な複製による不死化 Enabling replicative immortality、(3) 増殖抑制の回避 Evading growth suppressors、(4) 細胞死 Cell death、(5) エネルギー代謝のリプログラミング Reprogramming energy metabolism、(6) 血管新生 Angiogenesis、(7) 免疫による攻撃からの逃避 Avoiding immune destruction、(8) 炎症の促進 Tumor-promoting inflammation、(9) 増殖シグナルの維持 Sustaining proliferative signaling、(10) 浸潤能および転移能の活性化 Activating of invasion and metastasis である。前述の要因が互いに密接に関わり複雑な腫瘍の微小環境を形成しているとされている [7]。これらの腫瘍の複雑性を解明し腫瘍の個性に応じて、個別に治療できる新たなターゲットを見出すために癌細胞の遺伝子や転写産物を解析するゲノム解析研究やトランスクリプトーム解析研究が行われている。

1.3 乳癌のバイオマーカー探索におけるトランスクリプトーム解析の意義

前述の乳癌サブタイプ分類をはじめとした分子診断は、広く受け入れられているが、日常診療として個別の患者に適応して検査できるものではない [8]。研究として乳癌患者の一部のサンプルのみが解析され乳癌サブタイプを主軸とした治療の方向性を提供するにとどまっている。したがって現状の医療では、IHC 診断の結果を組み合わせて、研究によって示された分子診断の”個別”を推定する形で”集団”として治療がなされている [9]。これらの研究で用いられたトランスクリプトームデータの多くは、公共のデータベースに既に登録されており、さらに今日多くの研究者が研究成果とともに、その元となった研究データを公共のデータベースに日々登録している。

公共データベースの一つである Gene Expression Omnibus (GEO) には、2016 年時点で 2889 の異なる生物に由来する 130 万を超えるサンプルを含む 54,640 の研究データが保存されていると報告されている [10]。さらにトランスクリプトームデータに加えて、臨床病理学的な因子や予後や治療効果も付随したデータセットも広く公開されている。さら

に、研究で示されている個別化された分子診断の臨床的な予後や治療効果に対する影響は、IHC 診断に比べて大きいためトランスクリプトームデータの重要性は非常に高くなっている [11, 12]。

乳癌の特徴を個別に捉え治療を行っていくためには、乳癌サンプルにおけるトランスクリプトームの解析が不可欠であると考えられる。その理由は、近年の全ゲノムシークエンス解析や全エキソーム解析による癌横断的な解析の結果、乳癌においては紫外線によるメラノーマやタバコによる非小細肺癌 [13, 14] など強烈な発癌物質があきらかではない。そのため発癌物質によって誘導される 1MB (100 万塩基) 当たりの体細胞遺伝子変異数 (Tumor Mutation Burden) は、乳癌では 1/MB 程度と癌の中では比較的少ないことが知られている [15]。さらに、肺癌の EGFR 遺伝子 [16] や大腸癌の K-ras[17] のようなドライバー癌遺伝子変異は同定されておらず、乳癌は遺伝子変異に関してはドライバー遺伝子変異が不在の腫瘍と考えられている。そのため、乳癌においては少数の遺伝子変異の評価よりも網羅的な遺伝子発現解析が重要視されており、前述のサブタイプ分類のように成果を上げてきた背景がある。さらにトランスクリプトーム解析で得られるバイオマーカーは遺伝子発現のみではない [18]。癌における異常な染色体の組み換えの後に発生し得る、遺伝子と遺伝子の異常な結合である融合遺伝子もまた、治療のバイオマーカーとなりえる [19, 20]。

前述のように、新たなバイオマーカーを探索することは乳癌において重要な意味を持つこと、そのためにトランスクリプトームの解析が重要であることを述べた。本研究ではトランスクリプトームデータの臨床応用における課題や貴重な臨床データが付随しているにも関わらずトランスクリプトームデータ解析を困難にしている問題点に焦点を当てる。

1.4　トランスクリプトームの解析手法と乳癌における臨床との関わり

トランスクリプトームとは、細胞内に存在する mRNA 全てを意味する。その解析方法には、代表的なものとして図 1.1 のようなものがある。

ノーザンブロット法は、抽出された mRNA を電気泳動によって分離した後にメンブレンに転写し、検査対象の遺伝子の mRNA 配列に標識された核酸プローブをハイブリダイゼーションさせて、標識された mRNA を検出する方法で、発現を定性的に評価する方法である。乳癌の診断や治療決定に不可欠な癌組織中の HER2 遺伝子の測定法として臨床使用されている。

リアルタイム RT-PCR は、ポリメラーゼ連鎖反応 (PCR) による遺伝子の逆転写した

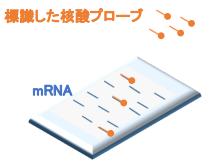
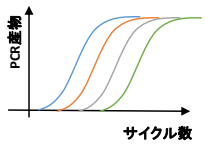
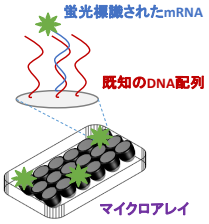
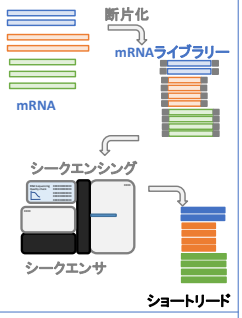
	ノーザンブロット法	リアルタイムRT-PCR	DNA マイクロアレイ	RNAシークエンス (RNA-Seq)
原理	電気泳動されメンブレンに転写されたmRNAに、標識した核酸プローブを用いて検出を行う 	ポリメラーゼ連鎖反応（PCR)によるmRNAの増幅を経時的に測定し、増幅率に基づいて定量を行う 	蛍光標識されたmRNAを既知のDNA配列が配置されたマイクロアレイに添加し、相補的に複合体を形成させ、発色させる。蛍光の配置と強度で発現を定量する 	次世代シークエンサを用いて断片化されたmRNAライブラリーの配列を解読し、発現を定量する 
データ特性	定性的	定量的	半定量的 （実験環境に依存）	定量的
解析規模	限定的 （数個~数十個）	限定的 （数個~数十個）	大規模 （既知の全mRNA）	網羅的 （含有される全mRNA）
乳癌での臨床応用例	抗HER2薬使用への HER2遺伝子発現診断	乳癌再発予測診断 Oncotype DX®	乳癌再発予測診断 MammaPrint® Curebest™	未開発

図 1.1　トランスクリプトームの解析手法

cDNA の標的配列の増幅を経時的に測定し、その増幅率を計測することで遺伝子の発現を定量的に測定する方法である。乳癌の再発予測診断 Oncotype DX™ では、リアルタイム RT-PCR 法を用いて 21 個の遺伝子発現を計測している。このようにノーザンブロット法やリアルタイム RT-PCR 法を用いた mRNA の測定は臨床応用されているものの、いずれの方法も測定できる遺伝子が多くても数十個に限定されるという制約がある。

近年、DNA マイクロアレイや RNA-Seq を用いた網羅的なトランスクリプトーム解析が行われるようになっている。DNA マイクロアレイは、蛍光標識された mRNA を既知のユニークな DNA 配列が配置されたマイクロアレイに添加し、ハイブリダイゼーションさせて発色させる。用いるマイクロアレイチップは、網羅的に遺伝子解析を行うため遺伝子毎に位置が決められており、その蛍光の配置と強度を一枚の画像として保存し、その画像を数値に変換して発現量を定量する。大阪大学で開発された 95 個の遺伝子発現を用いた再発予測モデル Curebest™ では、この技術が用いられている。

RNA-Seq は、次世代シークエンサー (NGS) を用いて断片化された RNA ライブラリー配列を実際に解読してそのリードの数を用いて発現量を定量する。RNA-Seq は現在のト

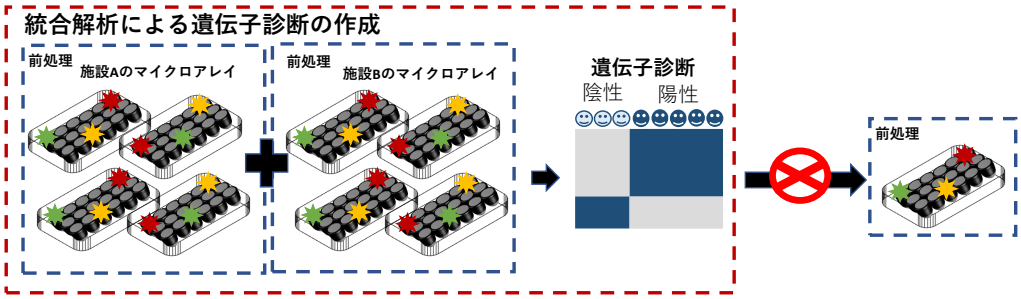
ランスクリプトーム解析の標準的なプラットホームとなりつつあるが、乳癌領域において検査としてこの手法を用いて臨床応用されたものは未だ開発されていない。トランスクリプトーム解析による分子診断である Oncotype DX™ は、乳癌診療ガイドラインでの保険適用内での使用が唯一明記され、Curebest™ なども保険適用外ではあるが臨床で使用可能である。これらの分子診断によって予測される再発リスクは、従来の臨床病期や IHC の所見より優先されて化学療法投与の是非が決定されるようになっているのが現状である [21, 22]。

1.5 トランスクリプトームの解析の課題と研究の背景

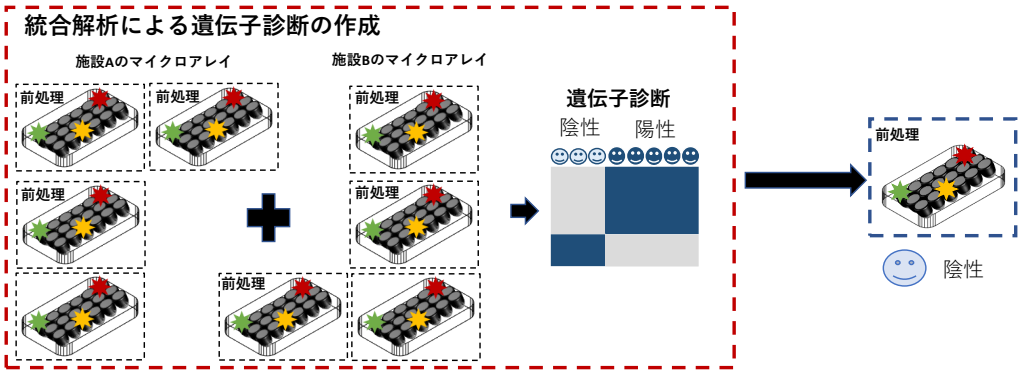
前項で述べたようにリアルタイム RT-PCR 法は、少数の遺伝子を定量化することに優れているが、癌の個性を評価する観点においては限界のあるプラットホームである。一方でマイクロアレイや RNA-Seq では一度に網羅的に遺伝子発現を評価することが可能である。しかしながら、それらには下記の問題点がある。

マイクロアレイは、各アレイの発現データを光信号として 1 枚の画像として記録し、それを後に発現量として定量している。その光信号を保存する環境の差異から測定されるデータに無視できないほどのバックグラウンドのばらつきを伴っている。これらのばらつきはバッチ効果と呼ばれており、その効果が大きいほどデータの取得環境が異なり、データの統合を困難にさせている。これまでの研究結果からバッチ効果は、測定する施設間、検査者、天候、気温、湿度、電圧や周波数等様々な因子によって大きく左右されることが報告されている [23]。マイクロアレイの研究データを統合解析する際に遺伝子発現変動よりもバッチ効果が大きくなり正しく解析できないことがある。そのためマイクロアレイによるトランスクリプトーム解析では、前処理を適用することでバッチ効果を取り除く (バッチ補正) 様々な方法が提案されている [23]。前述の前処理によって発現量は大きく変動するが、これまでに癌組織データを用いた前処理としての推奨された解析法は未だ示されていない。

さらにマイクロアレイを臨床応用する観点では、前処理の手法にも制約が加わる。臨床検査は、検体が逐次的に取得されるため、複数の検体を蓄積して補正するような方法を選択できない。図 1.2 のように蓄積した研究データを統合してバッチ補正を行い、バイオマーカーを探索したとしてもそれらの遺伝子診断を臨床サンプルにそのまま用いるためには予め逐次的に補正する手法を確立する必要がある。このような背景や制約の中で、前処理の違いにより実際に適切にバッチ効果が除去され、生物学的特性が適切に得ら



複数の施設のデータを施設単位で補正して統合して作成された遺伝子診断をそのまま1例に適用できない



複数の施設のデータを一例単位で補正して統合して作成された遺伝子診断は、そのまま適用できる

図 1.2　遺伝子診断作成におけるデータ統合の問題点

れているかどうかについても不明瞭なのが現状である。これまでに臨床応用されている Curebest™ は、バッチ効果の問題を一つの施設のみで解析することで実現している。さらに、マイクロアレイには、乳癌治療において欠かせないエストロゲン受容体などの発現値も取得可能であり、IHC 検査を置き換えることも期待される。そのため、この問題が解決されることで、マイクロアレイの臨床応用の適用可能性が広がる可能性がある。

一方で RNA-Seq によるトランスクリプトーム解析では、既知の遺伝子のみを解析対象としたマイクロアレイの限界を超えたグローバルな遺伝子発現パターンを解読するための強力なツールである。さらに、遺伝子の機能的異常である発現データにとどまらず遺伝子自体の異常も同時に提供するため現在のトランスクリプトーム解析の中核を担う手法となっている。RNA-Seq によって検出できる遺伝子の異常には、構造変異 (たとえば、遺伝子と遺伝子の結合である遺伝子融合)、遺伝子のコピー数多型 (CNV)、一塩基多型などの体細胞変異、一塩基多型 (SNP) などの生殖細胞変異およびウイルスの存在などがある。RNA-Seq によって検出できる遺伝子の機能的異常は、発現レベルの定量的推定と、転写物アイソフォーム、キメラ RNA、および RNA 編集部位の検出など様々である [18]。そ

れらは、全てバイオマーカーとなりえる。

しかしながら、RNA-Seq は、遺伝子自体の異常検出において、ゲノム解析に比べて体細胞変異や CNV 解析の精度が低いこと、融合遺伝子検出精度が低いことなどの問題点がある [18, 24, 25]。RNA-Seq の技術的な問題点の根本にあるのは、遺伝子から転写されてスプライシングされた mRNA をシークエンスするが、参照配列に対してスプライシングを加味してマッピングしなければならない点であると考えられる。スプライシングとは、遺伝子 DNA の塩基配列から転写された mRNA の前駆体のうち、翻訳されないイントロン部分を除いたたんぱく質合成に関わるエクソンを含む部分のみ連結する反応のことである。つまり、スプライシングのためリード全体をそのまま参照配列にマッピングさせることができず、部分部分で分割してマッピングすることになる。その際にシークエンスリードの長さが足りないことで、複数の遺伝子座にマッピングされ精度低下へとつながっていると考えられる。さらに、遺伝子の機能解析においても発現データで NGS の違いによるバッチ効果が問題になる [26] などの課題もある。そのため RNA-Seq を用いたトランスクリプトーム解析は、コストの面もあるが、上記のような技術的な課題があり、未だ臨床検査として用いられてはいないのが現状である。

そのような背景の中で、遺伝子自体の異常のひとつである融合遺伝子は、融合遺伝子が高頻度に確認される癌研究において近年精力的に用いられている [20, 27]。それは、融合遺伝子は、癌の原因だけでなく治療のターゲットとなりえることが慢性骨髄性白血病の治療から示されているためである [28, 29, 30]。さらに、融合遺伝子は癌細胞に特異的に認められるため、有望なバイオマーカーとなることが期待されている。

1.6 本研究の目的

乳癌においては、トランスクリプトーム解析が大きな役割を果たし、治療の効果や再発予後に直結する乳癌サブタイプ分類を見出すなど様々な乳癌の個性を明らかにしてきた。しかし、現在最も使用される乳癌サブタイプ分類はたった 50 個のみであり [31]、トランスクリプトームから得られる情報を正確に取得・評価することは、さらなる癌の個性を解明する可能性がある。さらにその結果として新たな治療ターゲットを見出し、個別に治療がなされることが期待される。

乳癌マイクロアレイデータには、これまでのトランスクリプトーム研究成果を築き上げた 1 万症例を超える得難い貴重な臨床データ (年齢・薬剤の反応・ホルモン受容体・HER2 遺伝子・再発の有無・日数) が付随して公共データベースに登録されておりデータの価

値を一層高めている。また、RNA-Seq 解析においても、TCGA (The Cancer Genome Atlas Program) を中心とした腫瘍横断的な解析の主要プラットフォームはショートリードシークエンスであり、乳癌だけでも 2000 例を超えるデータが臨床因子と共に登録されており利用可能である。そのため現時点において付随する乳癌の臨床データの最も多いマイクロアレイデータ [第 2 章] やショートリードペアエンド RNA-Seq データ [第 3 章] を研究の対象とした。

本研究の目的は、網羅的なトランスクリプトーム解析であるマイクロアレイ・RNA-Seq のいずれも臨床応用の障害となっている下記の問題点をそれぞれ改善することを目的とした。第 2 章では、逐次的に検体が加わる臨床に即したマイクロアレイのバッチ補正に対して焦点を当てた。目的は、異なるバッチから得られたマイクロアレイデータを乳癌の特徴を維持した状態で、逐次的に前処理する手法の提案とその適応可能性の評価することである。第 3 章では、RNA-Seq 解析の前述の課題の中で、癌特異性の高い融合遺伝子の検出に焦点を当てた。目的は、ショートリードペアエンド RNA-Seq データから融合遺伝子検出に対して前処理法をを用いることで既存の手法の融合遺伝子検出の精度を向上することである。

1.7　本論文の構成

本論文は、4 章構成である。第 1 章では、マイクロアレイデータの統合解析においてバッチ効果が問題になることや RNA-Seq の融合遺伝子解析の精度が低いという問題点を提起した。その解決策としてそれぞれ、第 2 章では、公共データベースにおける乳癌マイクロアレイデータにノンパラメトリック-Z-スケーリングによるバッチ補正を行うことで、IHC との一致率やサブタイプ分類精度が向上することを示した。第 3 章では、乳癌細胞株のショートリードペアエンド RNA-Seq データに対してショートリードクラスタリングによって読み取りリードを伸長することで融合遺伝子の検出精度を向上させる結果を示した。第 4 章では本研究のまとめと展望を示した（図 1.3）。

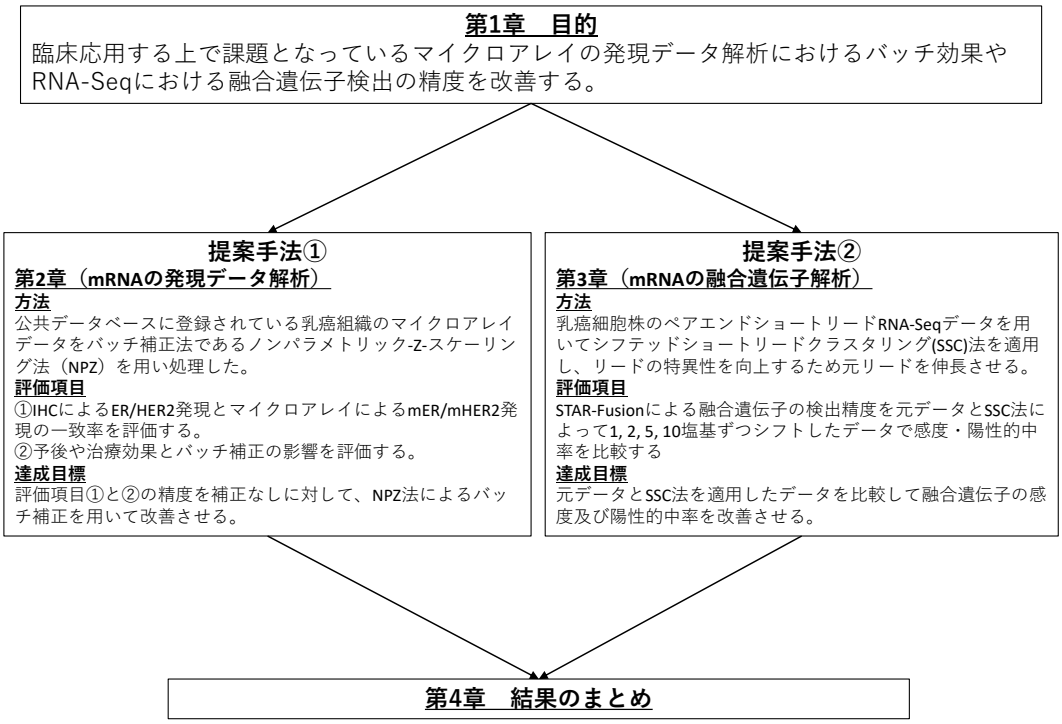


図 1.3　本論文の構成

第 2 章　マイクロアレイによる発現データのバッチ補正の検討

2.1　緒言

近年、マイクロアレイや RNA-Seq による遺伝子発現解析によって転写産物の情報を網羅的に取得することが可能になった。乳癌の治療方針決定のために欠かすことのできない IHC 法によるエストロゲン受容体 (ER) やヒト EGFR 関連物質受容体 2 (HER2) の発現が、リアルタイム-PCR 法やマイクロアレイ法などのトランスクリプトームの発現とよく相関することが示されており [32]、これまで非常に多くのトランスクリプトームを用いた乳癌の研究が行われてきた。日常臨床においても遺伝子発現解析を応用した多重遺伝子診断は、乳癌において急速に認知され、実際に使用されるようになっている。例えば遺伝子発現情報に基づいて判定される Oncotype DX™ (real-time PCR; Genomic Health, Redwood City, CA, USA) や MammaPrint®(マイクロアレイ; Agendia, Amsterdam,

The Netherlands)[33] は、欧米のガイドラインでその使用を強く推奨することが明記されている。日本でも 2021 年 12 月に Oncotype DX™ の保険収載が決定され、一部のサブタイプでは化学療法の投与の決定に使用されることが期待されている。さらに、顕微鏡による病理学的な細胞形態の違いをスケール化している組織学的核異型度よりも、マイクロアレイによる遺伝子発現に基づいた核異型度予測モデルである genomic grade index (GGI)[34] の方が、後方視的検討において予後とより相関することも示され、医療における分子診断の重要性がますます高まっている。

これらの遺伝子発現データは Gene Expression Omnibus (GEO)[*2]や ArrayExpress[*3] をはじめとした公共データベースに、研究で利用された実験データが日々登録され続けている。これらのデータは、各研究者単位の少数データから数百症例が含まれるものもあり様々ではあるが、それらを統合解析することで新たなバイオマーカーや各疾患における重要な遺伝子の働きを解明できる可能性がある。実際に筆者の作成した免疫関連遺伝子による乳癌化学療法感受性予測モデル (immune-related 23-gene signature: IRSN-23) は、自験データトレーニング 58 例とバリデーション 59 例で化学療法の感受性を高精度に予測することを示したが、さらに、公共データベースの 6 つの研究の合計 901 例を統合解析することでその再現性を確認することができた [35]。しかしながら、これらのデータの統合にはマイクロアレイの特性から後述する克服すべき課題がある。

2.1.1 マイクロアレイの前処理についての先行研究と課題

前章で触れたとおりマイクロアレイはチップ上の標識プローブの発光を画像として保存してそれを数値化している。撮影された画像はより中心が明るくなることなどの特性を有している。さらに、たとえ同じマイクロアレイチップを用いていたとしてもチップのわずかな傾きなどがノイズとして含まれている。その特性やノイズを事前に正規化する必要がある。そのためこれまでに様々な正規化法が報告されている [36]。

Robust multiarray average (RMA) 法 [37] は、マルチアレイ正規化と呼ばれる正規化の方法の一つである。マルチアレイ正規化は、実験検体すべてのマイクロアレイの発現量を統合して正規化を行う方法であるため、各検体間の共通ノイズに対して非常に有効な正規化法の一つである (表 2.1)。前述の筆者の開発した抗癌剤感受性 IRSN-23 遺伝子診断

[*2] http://www.ncbi.nlm.nih.gov/geo/
[*3] http://www.ebi.ac.uk/arrayexpress/

表 2.1　マイクロアレイの各種正規化方法

	RMA	MAS5	RMX	fRMA
バックグランド補正	Yes	Yes	Yes	Yes
ミスマッチプローブの情報の利用	No	Yes	Yes	No
プローブ毎の正規化	quantile	No	No	quantile(frozen)
プローブセット全体の正規化	No	Yes	Yes	No
要約	Median polish	Tukey' biweight	rmx polish	median polish
他のサンプルの情報の利用	Yes	No	No	No

の研究では、この手法を用いて正規化を行い研究データにおいて良好な結果を得た。しかしながら、新たに検体が加わった場合や検体を除いた場合に個々の推定発現値が変化するという問題点があり、同時に前処理するサンプル数によって発現値が異なる。その推定発現量によって計算された多重遺伝子診断の結果も異なるため、研究などでは施設毎や実験毎にまとめて正規化をする必要がある。IRSN-23 遺伝子診断は、研究において高い化学療法感受性予測を再現性を示していてもこの問題を根本的に解決できないため臨床応用には至っていない。特に逐次的に検体を解析する必要がある臨床検査においては、症例の蓄積を待って解析することができないため、この問題は非常に重要である (図 2.1)。さらに、第 1 章で述べた様にマイクロアレイを用いて作成された遺伝子診断をそのまま臨床に用いるためには、遺伝子診断を構築する際に既に臨床の逐次性を考慮した逐次的な前処理を行う必要がある (第 1 章の図 1.2)。

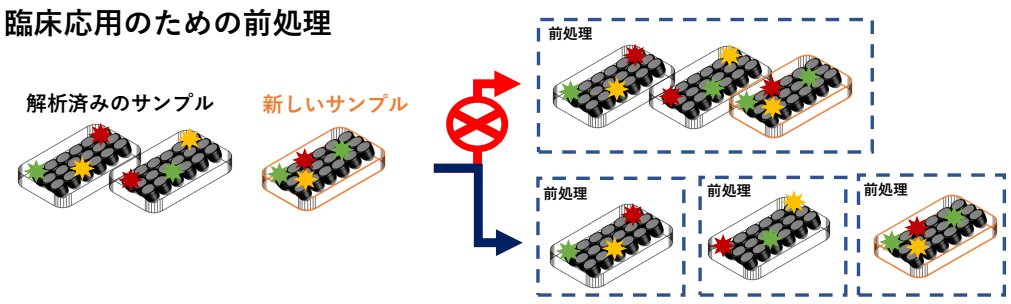


図 2.1　臨床応用のためのマイクロアレイの前処理

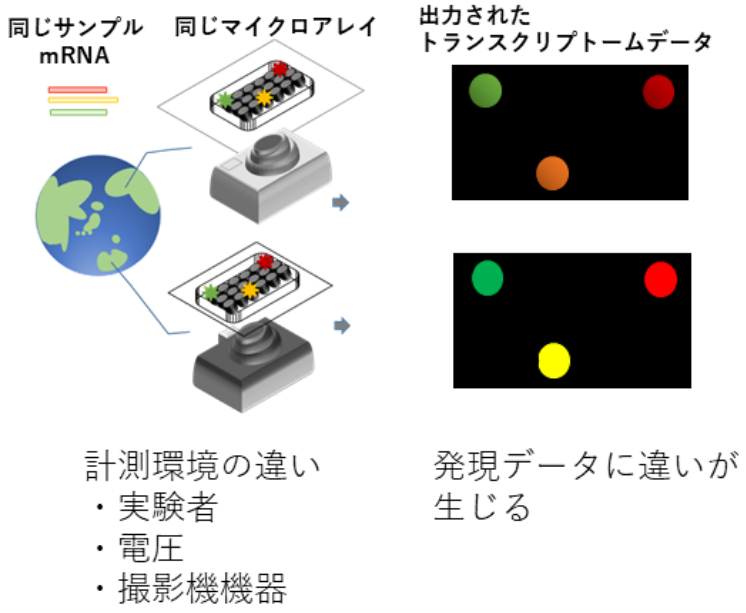


図 2.2　バッチ効果

一方、Microarray Suite 5.0 (MAS5)[38] や frozen robust multiarray analysis (fRMA)[39] や radius minimax (RMX)[40] に代表される正規化法は、シングルアレイ正規化であり、各々のサンプルごとに正規化を行うことができる。MAS5 は、Affymetrix のマイクロアレイに配置されている標的遺伝子配列 25 塩基 (パーフェクトマッチ) とその近傍に標的遺伝子とほぼ同じ配列であるが、意図的に 13 番目の塩基を 1 塩基を置換した配列 (ミスマッチ) の発現の差異をバックグランド補正として利用している。さらに、各遺伝子に対して複数の標的遺伝子配列がアレイ上に配置されているため、1 つの遺伝子の値として要約する必要がある。その際にアレイ全体を 4 × 4 個の区画に分けて各区画の最小値から 2% の発現の平均値をバックグラウンドとしてその標準偏差を算出する。そして各遺伝子毎に全区画の中心との距離を算出してその距離を重みとして平均化 (距離の大きさに応じて重みを付ける Tukey の Biweight 推定法を使用) して正規化する。RMX は、MAS5 の拡張版であり、要約の際に Tukey の Biweight 推定に代わって robust radius-minimax (rmx)estimator[41] を用いて行う手法であり、推定のばらつきが明らかに減少し、再現性が向上すると報告されている。fRMA は、RMA の改良版として報告され、正規化に必要なリファレンス分布とプローブ効果の推定値の情報を予め固定 (通常であれば、複数のマイロアレイに対してサンプル毎に発現強度順に並べ替えてそれらの平均値を補正値として値を置き換えるが、それを固定値で行う) して正規化を行う手

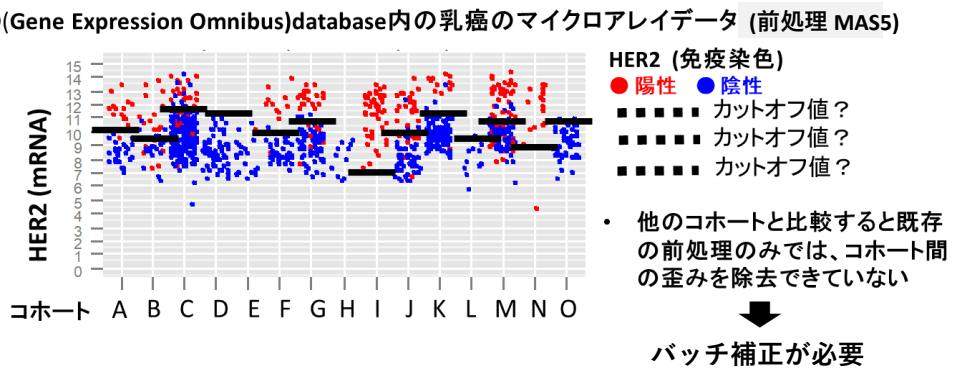


図 2.3　GEO データにおけるバッチ効果と HER2 遺伝子発現

法である。しかしながら、これらのシングルアレイ正規化は、マルチアレイ正規化よりも共通ノイズの補正は不十分であると報告されている [42]。

さらに、マイクロアレイには、第 1 章で述べたように”バッチ効果”によるデータのばらつきが存在する (図 2.2)。このバッチ効果は、多くの研究者によって GEO に登録されるデータを統合して解析する際に大きな障壁となっていると考えられる。図 2.3 では、GEO に登録されているコホート毎のマイクロアレイデータの各検体に対してそれぞれシングルアレイ正規化法の一つである MAS5 のみを行った HER2mRNA と HER2 の免疫染色との関係を示している。コホート単位では免疫染色の陽性・陰性は明確に分かれているが、各コホートをそのまま統合することが困難であることが示されている。

そのため、様々な付加的なバッチ補正 (RANK transformation 法 [43] や Z scale 法 [44]) がしばしば正規化後のデータに対して用いられている。最近ではサポートベクターマシーンに基づいた distance-weighted discrimination (DWD) 法 [45] や経験ベイズ法に基づいた ComBat [46] 法などのマルチアレイによるバッチ補正のバッチ効果に対する有効性が主に主成分分析 (PCA) を用いた解析などで報告されている。しかしながら、これらのバッチ補正もまた、複数の検体の情報を用いて補正を行っているために検体の増減に伴って陽性/陰性の値が後に変化しえるという問題点がある。そのため 1 例単位での同一の判別基準で再現性をもって判定できる値の算出が困難であり、臨床応用の需要に合致していない。

そのため Kim-Anh 等 [47] は、シングルアレイバッチ補正として累積割合値を使用した YuGene 法を提案し、バッチ補正を有効に行えることを報告した。しかしながら、YuGene 法は発現量総和から最大発現値より順に減算し累積割合値を算出するために、大きな外れ

値に影響を受ける可能性がある。この点においてRANK法は、単純に発現量の順番を数値に変換したものであるため、外れ値には影響を与えないが、発現データが一様分布へと変換されるなどの実際のデータ分布や密度と異なり大きく情報を失ってしまうという点に問題がある。

2.1.2 提案手法: ノンパラメトリック-Z-スケーリング

それらの問題点を解決するため、ノンパラメトリック-Z-スケーリング (NPZ) 法を提案した。手法の詳細は下記のとおりである。

$NPZ = (X_i - X_m)/NIQR$
X_i 各遺伝子の発現値, X_m 各遺伝子の発現値の中央値
$NIQR$ 正規分布の四分位範囲で正規化された各遺伝子の発現量の四分位範囲

このNPZ法は、各発現量を全発現量の中央値から減算した後に正規分布の四分位範囲で正規化された各遺伝子の発現量の四分位範囲で除法するというシンプルな手法である。この手法は、四分位範囲を用いているため外れ値の影響を受けにくい特徴があり、全発現量の中央値を用いて中央寄せしているため分布の形状に影響されないというメリットと同時にrank法とは異なり発現密度の情報を失わないという利点が想定される。NPZ法は、分析化学や食品化学の分野で使われているロバストZスコア法 [48] と類似した手法である。しかし、ロバストZスコア法は、データの中に存在する外れ値を検出するために用いられるのに対して、NPZ法はマイクロアレイを用いた遺伝子発現データのバッチ補正のための数値変換手法であり、データ処理の目的が異なる点が特徴である。つまり、このNPZ法の概念が外れ値の検出ではなく、マイクロアレイのデータに対してバッチ補正のために使われるのは本研究が初めてである。今回の検討では、臨床での場面に即して乳癌のマイクロアレイデータを1症例単位で正規化した後にバッチ補正を行うことで、IHCの結果や乳癌の予後や治療効果にどのような結果をもたらすかについて検討をおこなった。

2.2 方法

2.2.1 比較された前処理法

CEL[*4]ファイルからバックグランド補正 (R package “affy”中の expresso 関数における bg.correct = TRUE) 及び、2 を底とする対数変換のみをおこなった”正規化なし”に対して、各種 MAS5、RMX、および fRMA.RWA (fRMA-robust weighted average method) 等のシングルアレイ正規化を用いて行った (表 2.2)。

正規化後に下記のとおり、”バッチ補正なし”に対してシングルアレイによるバッチ補正 (Rank、YuGene、Z、 NPZ) とマルチアレイによるバッチ補正 (ComBat) を付加して比較検討した (表 2.2)。

表 2.2　正規化とバッチ補正

	方法	詳細
正規化	なし	シングルアレイベース
	MAS5	シングルアレイベース
	RMX	シングルアレイベース
	fRMA.RWA	シングルアレイベース
バッチ補正	なし	
	Rank	シングルアレイベース、分布仮定なし
	YuGene	シングルアレイベース、分布仮定なし
	Z	シングルアレイベース、ガウス分布
	NPZ	シングルアレイベース、分布仮定なし
	ComBat	マルチアレイベース、ガウス分布

fRMA.RWA: fRMA-robust weighted average method, Rank: rank transformation (R コアパッケージ), YuGene: semi-rank-based transformation (R YuGene パッケージ), NPZ: ノンパラメトリック-Z-スケーリング, Z: Z-score transformation (R コアパッケージ), ComBat: empirical Bayes method (R sva パッケージ)

[*4] Affymetrix Probe Results Format であり CEL 拡張子を使用する

2.2.2 使用したデータセット

患者は、公共のデータベース GEO から (1)〜(3) の基準を満たした 24 コホート、合計 2,817 症例が選択された (表 2.3)。 下記の基準は、以下の通りである。
(1)Affymetrix HG-U133 (GPL96) または、Affymetrix HG-U133 plus 2.0 (GPL570) のマイクロアレイを用いて得られた乳癌の発現データである。
(2)ER と HER2 の陽性/陰性の臨床情報が取得可能であり、現在の乳癌診療における国際的なカットオフであること [49, 50]。つまり、カットオフが ER: $\geq$ 10 fmol/mg cytosol protein または 陽性細胞が腫瘍の $\geq$ 10 % であり、 HER2 は、fluorescence in situ hybridization (FISH) 法で 2.0 以上もしくは IHC テストでスコアが 3+ であること。
(3) 症例が重複していないこと (全遺伝子発現が蓄積された症例と同じ発現パターンを示す症例は除外)

2.2.3 mRNA の ER と HER2 の閾値の決定方法

ER と HER の蛋白質とマイクロアレイの発現に関しては、既存の研究 [32] によってそれぞれ発現が一致することが知られているプローブ [ER; ESR1 (estrogen receptor 1, Affymetrix probe ID 205225_at), HER2; ERBB2 Erb-B2 receptor tyrosine kinase 2 (216836_s_at)] を選択した。IHC 染色と mRNA の発現の閾値を決定するため、まず ER/HER2 の陽性/陰性データに対して同一の乱数シードを用いたそれぞれ 1000 回のブートストラッピング抽出を行った。その 1000 回の抽出に対して、随時 ROC 曲線を用いて算出し、グラフの左上隅との距離が最小となる点をカットオフ値とした。最終的に 1000 個のカットオフの中央値を閾値と決定した。本論文では、マイクロアレイの ESR1、ERBB2 の発現を mER、mHER と表記し、最終的な閾値より大きいものを mER+、mHER2+、閾値以下を mER-、mHER2-とそれぞれ分類した。

2.2.4 各コホート毎の IHC の一致率の評価方法

検証用として用いる一つのデータセットを除いた残り全てのデータセットで、前述の通りブートストラップ法にて閾値を決定し、その閾値を用いて除かれていた 1 つの検証用データセットに適用 (Leave-one-out 交差検証 (LOOCV) 法) し、図 2.4 のように一致率を算出した。それをすべてのデータセットが 1 回ずつ評価されるまで繰り返した。最終的に、それぞれのコホートの IHC 染色と mRNA の発現の一致率の単純平均を比較し、

表 2.3　乳癌コホート (N = 2817) と臨床病理学的特徴

GEO			ER 発現			HER2 増幅		
データセット	プラットフォーム	N	陽性	陰性	不明	陽性	陰性	不明
GSE42822	GPL96	90(91) †	0	0	90	34	54	2
GSE33658	GPL570	11	11	0	0	0	0	11
GSE32646	GPL570	115	71	44	0	34	81	0
GSE32518	GPL96	50 (74) †	25	25	0	11	39	0
GSE29431	GPL570	54	0	0	54	25	25	4
GSE26971	GPL96	276 (277) †	0	0	276	0	276	0
GSE25066	GPL96	433 (508) †	261	172	0	4	418	11
GSE23593	GPL570	50	36	14	0	0	50	0
GSE23177	GPL570	116	116	0	0	0	116	0
GSE20271	GPL96	72 (178) †	32	39	1	22	47	3
GSE20194	GPL96	67 (278) †	33	34	0	36	31	0
GSE20181	GPL96	59	59	0	0	0	0	59
GSE18864	GPL570	84	0	0	84	18	64	2
GSE18728	GPL570	21	0	0	21	5	16	0
GSE17705	GPL96	298	298	0	0	0	0	298
GSE16446	GPL570	120	0	120	0	31	62	27
GSE16391	GPL570	55	55	0	0	3	42	10
GSE12093	GPL96	136	136	0	0	0	0	136
GSE10810	GPL570	31	19	12	0	0	0	31
GSE9195	GPL570	77	77	0	0	0	0	77
GSE6532	GPL96	206 (327) †	161	45	0	0	0	206
GSE6532	GPL570	87	87	0	0	0	0	87
GSE5460	GPL570	23 (127) †	12	11	0	23	0	0
GSE2034	GPL96	286	209	77	0	0	0	286

GEO : Gene Expression Omnibus　N :患者数　ER : エストロゲン受容体

HER2 : ヒト EGF 受容体 2　GPL96 : Affymetrix Human Genome U133A Array

GPL570 : Affymetrix Human Genome U133 Plus 2.0 Array　† : 除外前の患者数

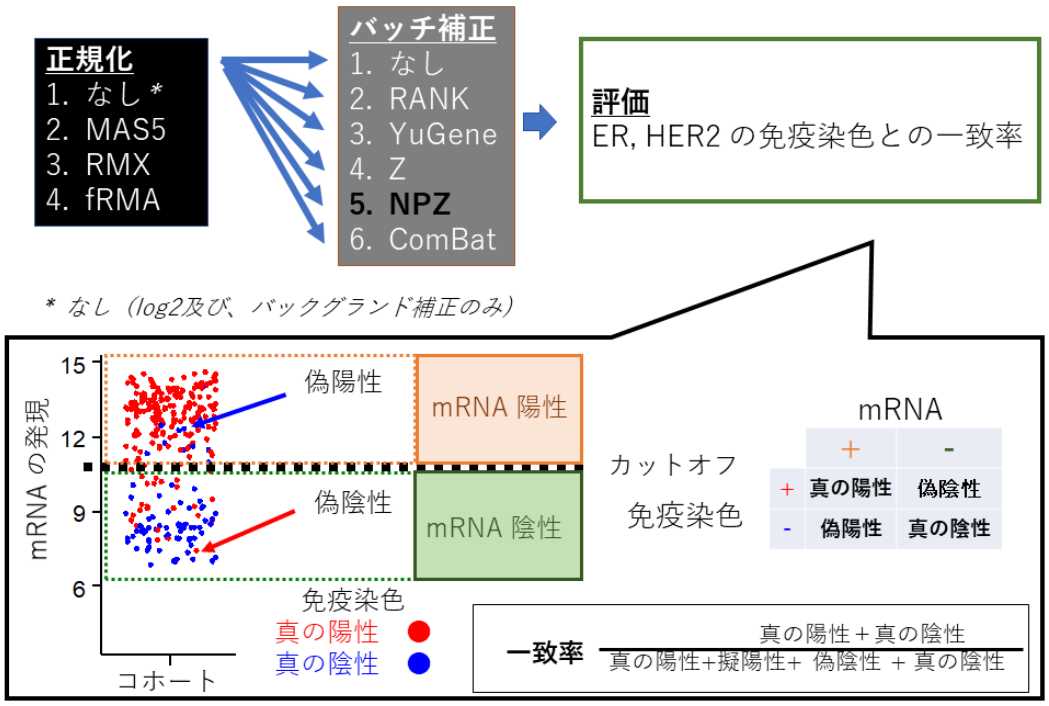


図 2.4　一致率の評価方法

バッチ補正による効果として評価した。

2.2.5　乳癌サブタイプにおける解析

まず、乳がんのサンプルを、ER と HER の IHC 染色の結果と組織学的核異型度の情報を用いて、4 つのサブタイプに分類した。分子サブタイプは、LumA (ER 陽性/HER2 陰性/組織学的核異型度 1 または 2)、LumB (ER 陽性/HER2 陰性/組織学的各異型度 3 もしくは、ER 陽性/HER2 陽性、全組織学的核異型度)、HER2 (ER 陰性、HER2 陽性、組織学的核異型度)、および Basal (ER 陰性、HER2 陰性、全組織学的核異型度) と定義した。続いて対応する発現データを、Parker[31] らによって記述された PAM50 という 50 個の遺伝子のマイクロアレイの発現に基づく乳癌サブタイプ分類に従って Luminal A(LumA)、Luminal B (LumB)、HER2-enriched (HER2)、Basal-like (Basal) に分類した。IHC と PAM50 で分類されたサブタイプの一致率は、各種正規化法と各種バッチ補正毎に計算した。さらに、前述の解析で予測されたサブタイプにおける抗がん剤の感受性を評価した。この解析は、治療効果の情報を持つ術前化学療法のサンプル 649 例を対象と

した。治療効果は、病理学的に腫瘍が完全に消失している pCR (病理学的完全奏効) と病理学的に腫瘍が残存している非 pCR と定義した。それぞれのサブタイプにおける pCR 率を、各種正規化法と各種バッチ補正毎に計算した。

2.2.6 Genomic grade index (GGI)

不特定多数の遺伝子におけるバッチ補正の影響を評価する目的で下記の遺伝子診断を用いてバッチ補正毎の予後との相関を評価した。Genomic grade index (GGI) は、Sotiriou[34] 等によって報告された 128 個のプローブ (97 遺伝子) を用いて細胞核の異型度を"low"と"high"に 2 分する多重遺伝子による核予測モデルである。この GGI は、各プローブの係数の重みを 1 とする等しい重みを用いており、主に細胞周期や増殖に関連する遺伝子の発現量で判定するモデルである。GGI は、乳癌の核異型度のグレードを高い一致率で予測し、尚且つ顕微鏡によるグレード分類よりも GGI が予後をより分けることが示されている [34]。今回の検討では、バッチ効果が残存すれば GGI による分類に影響を与えるため予後の分類精度が低下することが考えられる。また、GGI による high と low の比率が偏った場合に予後の比較に影響を与えるため high/low の比率が 1:1 となるように GGI の中央値で、High-GGI と Low-GGI に判別して予後との相関を比較した。

2.2.7 統計解析

2.2.7.1 主成分分析 (PCA)

主成分分析は、全発現量を用いて、第 1 主成分と第 2 主成分に対して行った。解析ソフトは、R コアパッケージを用いた。

2.2.7.2 生存分析

内分泌療法症例における無再発生存期間 (RFS) は、Kaplan-Meier 曲線を用いて算出し、検定は log-rank test における P 値を採用した。

2.3 結果

2.3.1 IHC とマイクロアレイ発現の一致率の比較

この検討では、GPL96 のプラットフォームから 1973 症例 (11 施設) と GPL570 のプラットフォームより 844 症例 (13 施設) の合計 2,817 例が選択された。まず、それぞれのプラットフォーム内でのバッチ補正の影響を調べた (表 2.4)。GPL96 内では、正規化

に加えてシングルアレイのバッチ補正 (RANK、YuGene、Z、NPZ の平均) を行うことで一致率の改善を認めた (mER; シングルアレイのバッチ補正 90.0% vs. バッチ補正なし 88.3%, mHER2; シングルアレイのバッチ補正 90.4% vs. バッチ補正なし 85.0%)。一方で、マルチアレイのバッチ補正では、一致率が低下する結果 (mER; 82.2%, mHER2; 82.3%) となった。同様な結果が、GPL570 においても認められた (シングルアレイのバッチ補正 (RANK、YuGene、Z、NPZ の平均): mER; 94.9%, mHER2; 79.2%, マルチアレイのバッチ補正: mER; 68.0%, mHER2; 77.7%)。シングルアレイのバッチ補正のなかでは、NPZ 法が全て平均以上であり各プラットホームで最も一致率が高い結果であった (GPL96 における mER; 90.1%, mHER2; 90.2%, GPL570 における mER; 95.0%, mHER2; 81.8%)。

2.3.2 乳癌サブタイプの予測

先の検討では、主に少数のプローブのみに注目して検討したため、バッチ補正がより多くの遺伝子にどのような形で作用するかについて検討した。PAM50 は、前述の通り 50 個の遺伝子を用いて、患者を 4 つの内在性乳癌サブタイプに分類するために使用されている。臨床現場では、マイクロアレイアッセイの代わりに IHC 染色を用いて、乳癌を同じ 4 つのサブグループに分類している。まず、ER、HER2、組織学的核異型度の情報を用いて、4 つのサブタイプ (LumA 235 名、LumB 200 名、HER2 210 名、Basal 391 名) にそれぞれ分類した。サンプルは、それぞれ各種の正規化とバッチ補正を経て、その発現データで PAM50 に分類された。すべての手法における平均精度は，LumA，LumB，HER2，Basal の各サブタイプにおいて，それぞれ 65.4%，59.1%，38.7%，73.2% であった (表 2.5)。全体として、今回の解析では HER2 グループの予測は困難であったが、シングルアレイによるバッチ補正のうち NPZ は、バッチ補正を行わなかった場合と比較してサブタイプ予測の精度を向上させた (41.0% vs. 38.3%)。一方、RANK 法は、各サブタイプで平均的な精度を下回った (LumA；63.4 %、LumB；58.0 %、HER2；36.9 %、Basal；72.5 %)。

2.3.3 治療効果に対する影響

バッチ補正が生物学的特徴を失うことなく適切にバッチ効果を取り除くことができるかどうかについて検討を行った。この問題を評価するため、臨床的な結果である (i) 化学療法感受性 (ii) 内分泌治療法感受性に着目してバッチ補正の効果を評価した。

表 2.4　mRNA と IHC 染色と ER/HER2 の一致率

		GPL96 (N = 1973)		GPL570 (N = 844)	
		ER/mER	HER2/mHER2	ER/mER	HER2/mHER2
正規化	バッチ補正	一致率 (%)	一致率 (%)	一致率 (%)	一致率 (%)
なし	なし	86.5	80.7	94.8	76.3
MAS5	なし	89.6	90.6	94.9	79.6
RMX	なし	86.9	82.3	94.8	76.3
fRMA.RWA	なし	90.2	86.3	94.8	76.3
(バッチ補正なしの平均)		(88.3)	(85.0)	(94.8)	(77.1)
なし	RANK	89.6	90.7	94.8	76.8
MAS5	RANK	90.1	89.9	95.0	82.2
RMX	RANK	90.0	89.3	94.8	76.8
fRMA.RWA	RANK	90.3	90.8	94.8	76.8
(RANK の平均)		(90.0)	(90.2)	(94.9)	(78.2)
なし	YuGene	88.6	91.2	94.8	77.0
MAS5	YuGene	90.1	89.3	95.2	82.0
RMX	YuGene	89.6	89.9	94.8	77.0
fRMA.RWA	YuGene	90.5	90.8	94.8	77.0
(YuGene の平均)		(89.7)	(90.3)	(94.9)	(78.3)
なし	Z	90.0	90.5	94.8	77.6
MAS5	Z	89.9	91.4	94.9	81.0
RMX	Z	90.4	90.4	94.8	77.6
fRMA.RWA	Z	90.0	91.3	94.8	77.6
(Z の平均)		(90.1)	(90.9)	(94.8)	(78.5)
なし	NPZ	89.7	88.8	95.0	81.6
MAS5	NPZ	90.0	92.2	94.8	82.3
RMX	NPZ	90.5	89.7	95.0	81.6
fRMA.RWA	NPZ	90.1	91.0	95.0	81.6
(NPZ の平均)		(90.1)	(90.4)	(95.0)	(81.8)
(シングルアレイのバッチ補正の平均)		(90.0)	(90.4)	(94.9)	(79.0)
なし	ComBat	81.9	80.3	68.2	78.1
MAS5	ComBat	82.4	84.2	67.2	76.4
RMX	ComBat	80.5	80.7	68.2	78.1
fRMA.RWA	ComBat	83.9	84.1	68.2	78.1
(マルチアレイバッチ補正 ComBat の平均)		(82.2)	(82.3)	(68.0)	(77.7)

GPL96: Affymetrix Human Genome U133A アレイ, GPL570: Affymetrix Human Genome U133 Plus 2.0 アレイ, ER: エストロゲン受容体, mER: マイクロアレイの estrogen receptor 1 (205225_at) を用いて判定された ER, HER2: ヒト上皮増殖因子受容体 2, mHER2: マイクロアレイの Erb-B2 receptor tyrosine kinase 2 (216836_s_at) を用いて評価された HER2.

表 2.5　IHC サブタイプ予測の精度

正規化	バッチ補正	LumA (n=235) 精度 (%)	LumB (n=200) 精度 (%)	HER2 (n=210) 精度 (%)	Basal (n=391) 精度 (%)
MAS5	なし	65.1	59.5	39.0	74.7
RMX	なし	70.1	62.9	38.5	70.2
fRMA.RWA	なし	66.7	60.6	37.6	72.5
(バッチ補正なしの平均)		(67.3)	(61.0)	(38.4)	(72.5)
MAS5	RANK	61.7	54.9	37.9	73.5
RMX	RANK	63.8	59.6	36.4	72.5
fRMA.RWA	RANK	64.8	59.6	36.4	72.5
(RANK の平均)		(63.4)	(58.0)	(36.9)	(72.8)
MAS5	YuGene	62.8	56.2	36.3	73.4
RMX	YuGene	64.3	53.9	36.4	71.4
fRMA.RWA	YuGene	66.0	59.8	37.4	71.4
(YuGene の平均)		(64.4)	(56.6)	(36.7)	(72.1)
MAS5	Z	64.9	59.9	40.7	75.4
RMX	Z	67.7	60.0	41.0	73.3
fRMA.RWA	Z	66.0	60.5	39.5	73.2
(Z の平均)		(66.2)	(60.1)	(40.4)	(74.0)
MAS5	NPZ	64.6	59.7	40.6	75.9
RMX	NPZ	67.4	59.9	43.4	74.6
fRMA.RWA	NPZ	66.0	60.1	38.9	73.0
(NPZ の平均)		(66.0)	(59.9)	(41.0)	(74.5)
(シングルアレイのバッチ補正の平均)		(65.0)	(58.7)	(38.7)	(73.3)

2.3.3.1 化学療法感受性とバッチ補正の影響

先の PAM50 のサブタイプ分類の予測を用いて、術前化学療法を行った 649 例の pCR 率について検討を行った。一般に、HER2 と Basal は術前化学療法の感受性が高いことが知られている (高い pCR 率)[9]。そのため、HER2 や Basal での感受性の結果が改善されていれば、バッチ効果を適切に除去できていることが想定されるため本解析を行った。すべての方法における平均 pCR 率は、LumA、LumB、HER2、Basal の各サブタイプで、それぞれ 3.0 %、11.6 %、43.1 %、33.1 %であった (表 2.6)。今回の解析では、NPZ はバッチ補正なしと比較して、HER2 と Basal の特性を強調していた HER2；44.6 %対 41.6 %、Basal；33.6 %対 32.5 %)。

2.3.3.2 内分泌治療法感受性とバッチ補正の影響

ER 陽性乳癌は、内分泌療法のターゲットであり、その治療により ER 陰性乳癌とは異なり予後が改善することが知られている [51]。また、ER 陽性乳癌の中で HER2 陽性乳癌はホルモン療法に抵抗性を示すことが知られている。

そのため、術後補助療法として内分泌療法のみを行った症例 (N = 814) を対象として、先のサブタイプの解析と同様に mRNA の情報で ER 陽性/陰性と HER2 陽性/陰性を判別し、LumA (マイクロアレイの mER と mHER2 で判別した ER+HER2-サブタイプ) とそれ以外のサブタイプに分類した。その 2 群の無再発生存期間 (RFS) と再発の有無についての経時的なデータを Kaplan-Meier 曲線と log-rank 検定で P 値を算出した。その再発曲線と P 値をバッチ補正ごとに比較した。この検討を行うことでバッチ効果が生物学的差異を縮小させて、バッチ効果を取り除くことで遺伝子診断による分類の差がより明確になるのではないかと想定し検討を行った。

図 2.5 は内分泌療法患者における再発曲線と mER/mHER2 遺伝子発現によるサブタイプ分類との関係を示しているが、MAS5 後のシングルアレイのバッチ補正を行うことで (RANK, P=3.86E-4; YuGene, P=0.001; Z, P=2.56E-4; NPZ, P = 1.08E-4, log-rank test) の RFS をバッチ補正なし (P = 0.011, log-rank test) やマルチアレイのバッチ補正 (P = 0.186, log-rank test) に比べて改善する事が示された。同様な結果が、fRMA や RMX の場合にも観測された。

これまで、ER や HER2 などの極限られた特定の遺伝子に着目して、バッチ補正と予後との関係を確認したが、さらにより多くの遺伝子を対象として検討を行った。対象とし

表 2.6　乳癌のサブタイプにおける治療効果

正規化	バッチ補正	LumA 症例 (n)	LumA pCR (%)	LumB 症例 (n)	LumB pCR (%)	HER2 症例 (n)	HER2 pCR (%)	Basal 症例 (n)	Basal pCR (%)
MAS5	なし	149	3.4	120	11.7	144	42.4	236	34.3
RMX	なし	121	2.5	106	12.3	147	41.5	275	30.5
fRMA.RWA	なし	148	2.0	98	14.3	144	41.0	259	32.8
(バッチ補正なしの平均)			(2.6)		(12.7)		(41.6)		(32.5)
MAS5	RANK	150	4.0	111	9.9	151	43.7	237	32.9
RMX	RANK	139	4.3	117	9.4	155	41.9	238	33.2
fRMA.RWA	RANK	155	2.6	94	10.6	149	43.0	251	33.1
(RANK の平均)			(3.6)		(9.9)		(42.9)		(33.1)
MAS5	YuGene	146	4.1	106	9.4	156	42.9	241	32.4
RMX	YuGene	140	3.6	113	9.7	155	41.9	241	33.2
fRMA.RWA	YuGene	149	2.7	96	10.4	144	43.8	260	32.3
(YuGene の平均)			(3.5)		(9.8)		(42.8)		(32.6)
MAS5	Z	151	3.3	123	10.6	142	44.4	233	34.3
RMX	Z	148	2.7	121	13.2	135	43.7	245	33.5
fRMA.RWA	Z	155	1.9	99	15.2	140	42.1	255	32.9
(Z の平均)			(2.6)		(12.8)		(43.4)		(33.5)
MAS5	NPZ	156	3.2	121	10.7	142	45.1	230	34.3
RMX	NPZ	141	2.8	141	12.1	132	46.2	235	33.6
fRMA.RWA	NPZ	155	1.9	98	14.3	141	42.6	255	32.9
(NPZ の平均)			(2.6)		(12.2)		(44.6)		(33.6)
(シングルアレイのバッチ補正の平均)			(3.1)		(11.3)		(43.4)		(33.2)

て用いたのは、GGI と呼ばれる乳癌で細胞核の異型度を判定する多重遺伝子診断である。この多重遺伝子診断で、High-GGI と判定された場合予後不良であることが報告されている。今回の検討では、バッチ効果を取り除くため元の発現データに対して各種バッチ補正を行っており、複雑な学習モデルでは、各プローブの重みが原著論文のパラメーターと異なることが想定される。一方で、今回用いた GGI は、各プローブの重みが元々等しく設計されており、その影響は少ないと考え本検討において採用した。検討の結果、MAS5 後

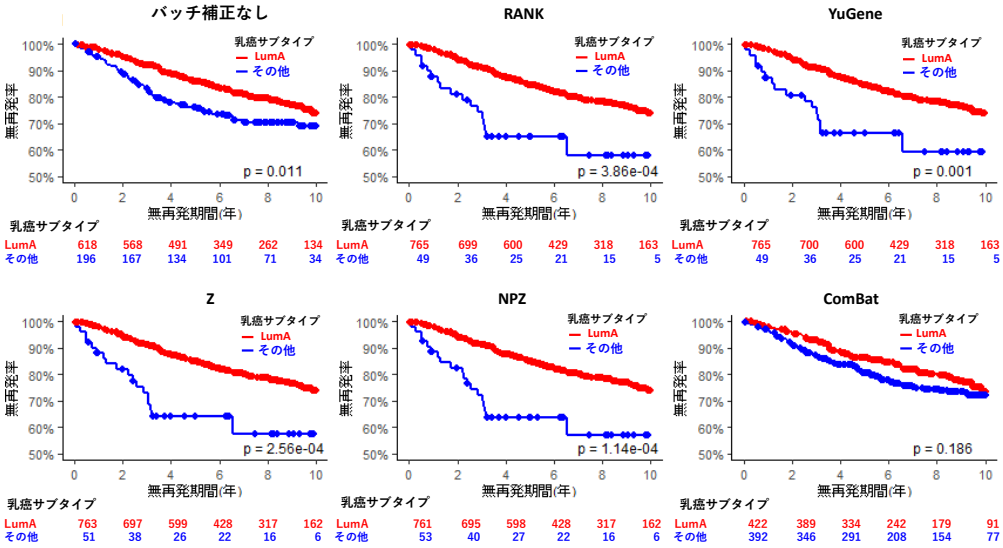


図 2.5　エストロゲン受容体陽性リンパ節転移陰性乳癌の内分泌療法患者における mER/HER2 遺伝子発現によるサブタイプ分類と MAS5 後の各種バッチ補正の再発曲線に与える影響

の NPZ (P = 4.57E-4, log-rank test) は、バッチ補正なし (P = 0.007, log-rank test) やマルチアレイバッチ補正 (P = 0.003, log-rank test) に比べて、10 年時の再発率は二つのグループの差をより明確にさせる傾向が見られた (図 2.6)。

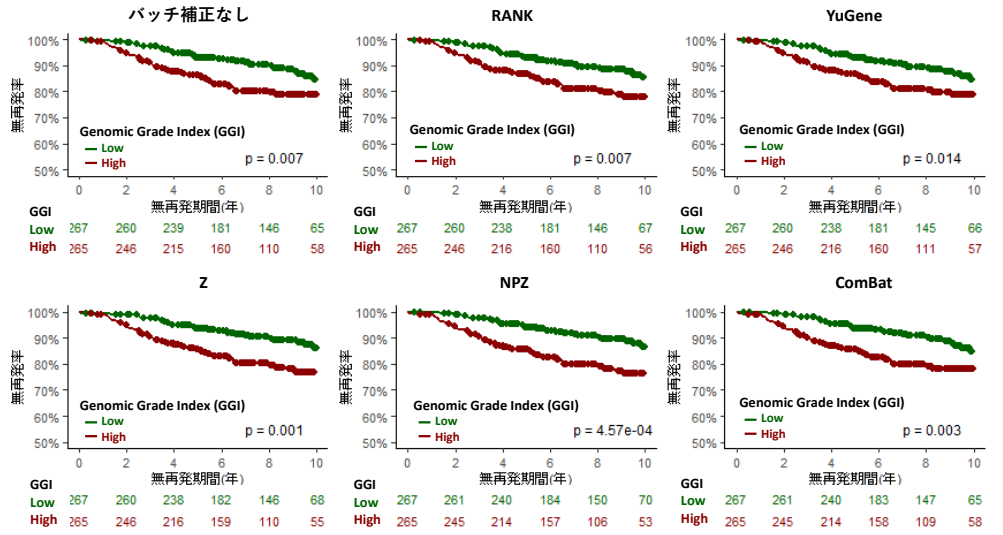


図 2.6　エストロゲン受容体陽性リンパ節転移陰性乳癌の内分泌療法患者における再発曲線と GGI

2.4 考察

マイクロアレイを用いた研究において前処理は非常に重要であり、これまでに様々な正規化が提案されてきた。しかしながら、バッチ効果をどのように取り除けるかによって生物学的意味が異なってくるが、いくつかの正規化にはバッチ補正も伴っているため正規化後にバッチ補正を行うべきかどうかについても明らかではない。さらにバッチ補正を行った結果を数千を超えるサンプルで実際に IHC 等でタンパク質の発現とどのように変化するかについてと比較検討した研究は、筆者が調べる限りにおいては報告されていない。

今回、IHC とマイクロアレイ発現の一致率の解析において、バッチ補正を行わない場合に正規化の手法間で IHC との一致率が大きくばらつく結果となった。正規化の手法間の結果の違いを比較すると特に ER 発現のように高発現と低発現に比較的明確に分布する図 2.10 下図と異なり、図 2.3 のように連続値として分布する HER2 発現の場合において、RMX は IHC との一致率を落としていた。その意味で本解析では、ER 発現に比べて HER2 発現がより敏感にバッチ補正の影響を表している可能性がある。RMX は、MAS5 と共通部分が多いが、低〜中発現の分散が異なること [40] が示されており、HER2 解析でバッチの影響をより強く受けているものと考えられる。fRMA は、発現値をソフトウェア内に固定された発現分布に対して quantile に置換しており複数のサンプルを用いなければならなかった RMA の問題点を克服しており、比較的良い結果が示されていた。MAS5 は、前述の通りミスマッチプローブとパーフェクトマッチプローブの差分と距離による重みづけという比較的シンプルな手法であることが良い結果につながったと考えられる。

各正規化に対してバッチ補正を行った結果は、シングルアレイの正規化を行うことで、バッチ補正なしに比べて IHC との一致率や乳癌の再発率や治療効果とより相関するという結果であった。しかしながら、バッチ補正の手法において正規化間での差が少なくなっていた。そのため、本解析においては最適な正規化と最適なバッチ補正については、言及するまでには至らなかったが、発現分布や発現値を喪失する RANK 法よりも NPZ や Z のように発現値の情報を維持した解析が全体として良い結果であった。しかしながら、臨床における検査の互換性や代替法と認められる一致率の基準は 95% とされている [52]。ER における IHC とマイクロアレイの一致率は、NPZ 法で改善したとしても GPL96 で 90.1%、GPL570 で 95.0%、HER2 も GPL96 で 90.4%、GPL570 で 81.8% と不十分であった。そのためマイクロアレイを IHC 検査の代替法として、すぐに臨床応用できるレベルにまでは補正できなかった。特に ER に比べて HER2 の一致率が低かった原因の一

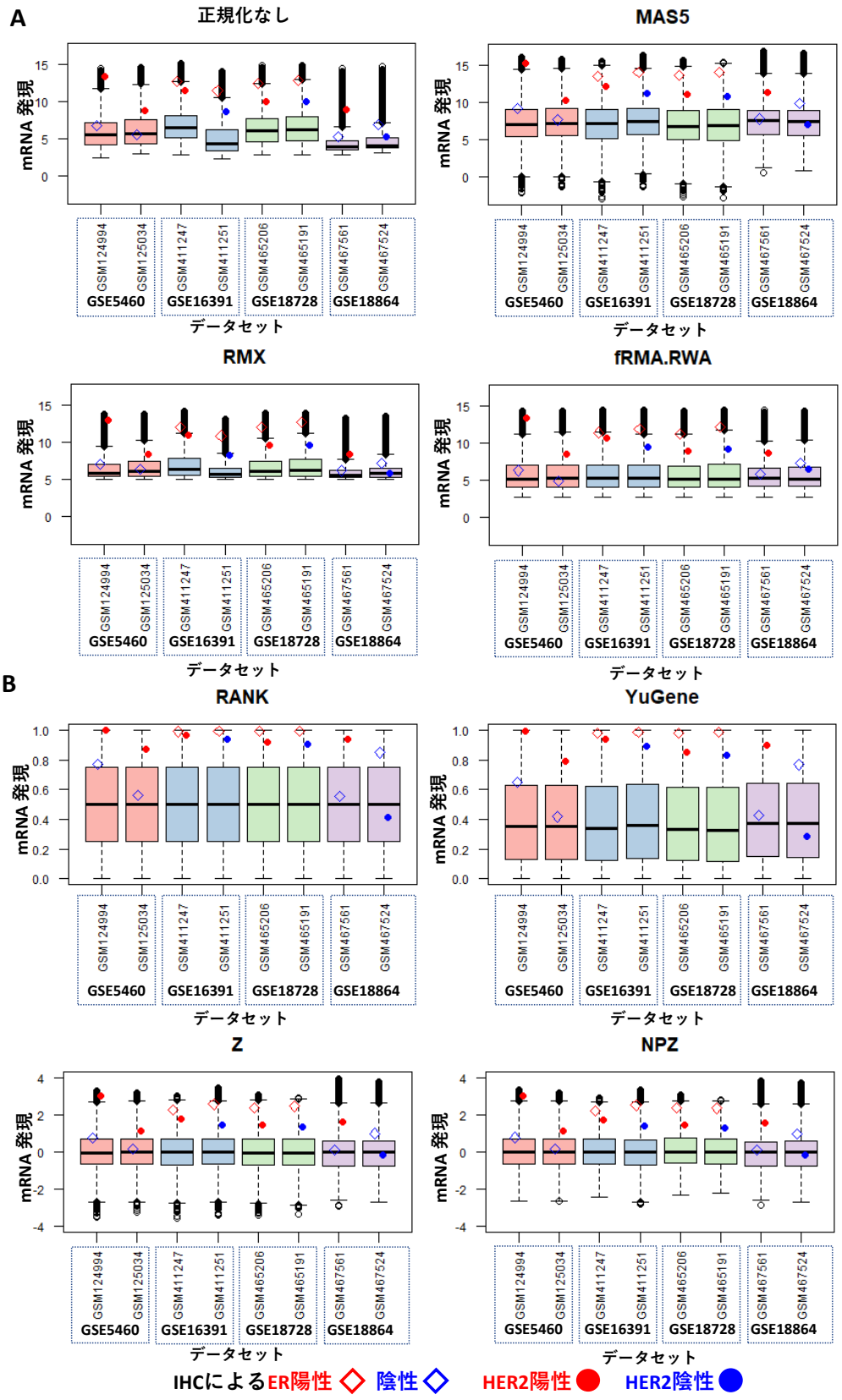


図 2.7　各種正規化 (A) と MAS5 後の各種シングルアレイのバッチ補正 (下) のデータの特徴

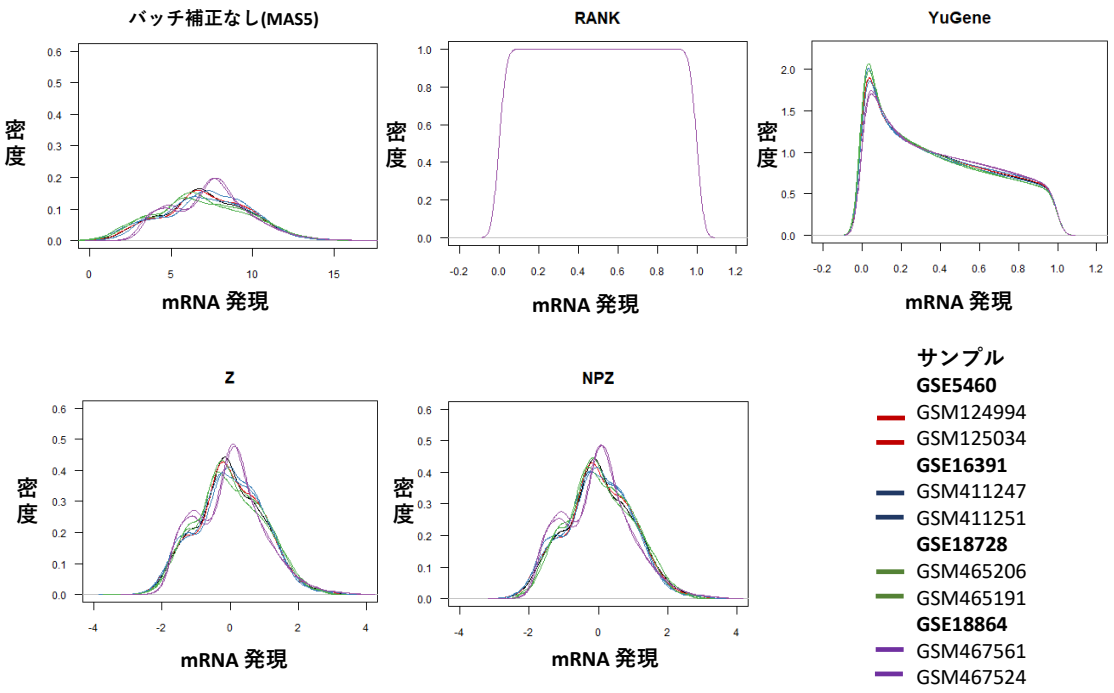


図 2.8　MAS5 後のデータにおける各種シングルアレイのバッチ補正の密度プロット

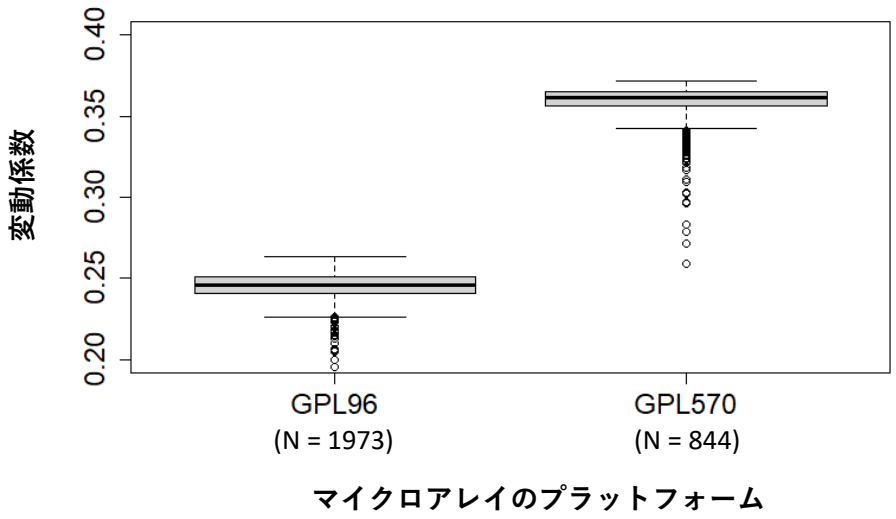


図 2.9　マイクロアレイの全発現量の変動係数とプラットフォームの違い

つとして、ER はタンパク質の発現を定量しているのに対して、HER2 はタンパク質の発現または、DNA のコピー数で陽性と陰性を区別して臨床で用いられているためその違いを生じた可能性が考えられた。

正規化の効果とバッチ補正の効果の違いについて比較して考察した。図 2.7A では、GPL570 の GSE5460、GSE16391、GSE18728、及び GSE18864 から 2 例ずつ合計 8 例での正規化の有無について全遺伝子発現を箱ひげ図を用いて比較したものである。正規化なしに比べて正規化を行うとコホート内では互いに比較可能となっているが、バッチ効果の影響が残存していることがわかる。図 2.7B では、各種バッチ補正の違いを比較しているが、Z に比べて NPZ は低発現での外れ値が少なく、さらに中央値で減算しているため、コホート間でのばらつきが少なくなっているが、図 2.8 のように密度分布も元の発現量の密度分布の形状を維持している。今回、NPZ が Z に比べて表 2.4 の一致率を改善したのは、GPL96 に比べて GPL570 のプラットフォームであった点および、MAS5 よりもバッチ補正なし、RMX、fRMA.RWA であった点から下記のことが考えられる。図 2.9 の示すように GPL96 は GPL570 に比べて各サンプル毎の全遺伝子の変動係数 (全遺伝子の標準偏差/全遺伝子の平均) を箱ひげ図にしたものであるが、よりばらつきが大きいことが示されている。さらに、図 2.7A で示すように正規化なし、RMX、及び fRMA.RWA は高発現に偏った外れ値を持っている。このようにばらつきが多く偏ったデータに対しては、IQR や中央値を用いる NPZ 法が、分散や平均値を用いる Z 法に比べてより安定して効果を示した可能性が考えられる。

通常、マイクロアレイのバッチ補正効果は、PCA グラフによって評価されることが多く、これまでマルチアレイによるバッチ補正が PCA グラフにおいてバッチ効果を改善することが報告されている [45]。PCA グラフにおける改善が効果的にバッチ効果を取り除けているかどうか？生物学的特徴を保てているかどうか？は実際には評価困難である。今回の検討を開始した当初は、これまでの報告通りマルチアレイによるバッチ補正が最も良い結果となると考えており、シングルアレイによるバッチ補正がいかにそれに近づけるかということを目標としていた。驚くべきことに、マルチアレイのバッチ補正は、ER/HER2 の IHC 染色との一致率や乳癌の再発や治療効果における検討においても最も悪い結果となった。既存の報告のように PCA を行った結果 (図 2.10)、PCA グラフは、マルチアレイによるバッチ補正によって一見個々のバッチ効果が適切に除去されているかのように見える。しかしながら、今回の結果からバッチ補正によって明らかに生物学的特徴も補正されている可能性が示唆された。その原因は、研究コホートによっては、ER 陽

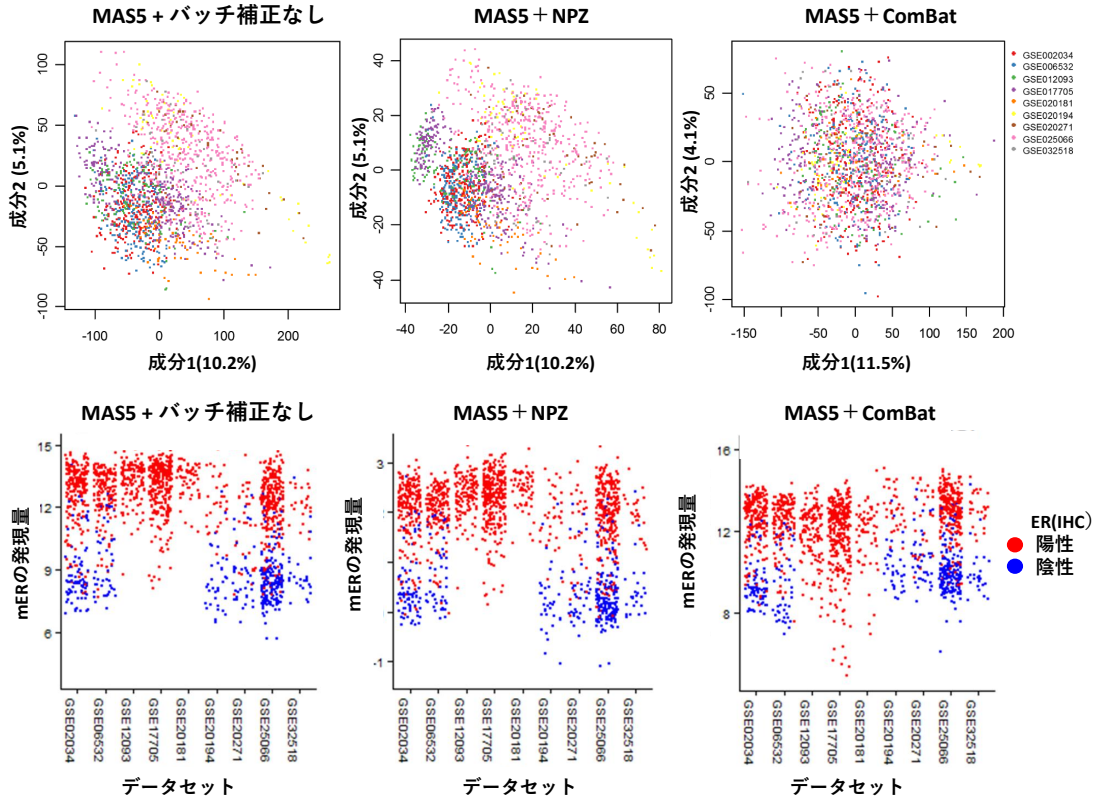


図 2.10　バッチ補正と主成分分析 (上) と mER の発現分布 (下)

性/陰性のみを対象としたものも含まれるためマルチアレイによるバッチ補正が適切に働かなかったものと考える。臨床では、検査前に ER 陽性や陰性を明確に予測して区別することはできない。さらに、1 症例ずつ結果を得る必要がある臨床検査においては、マルチアレイによるバッチ補正は非常に不便であり、このような設定では効果を発揮できないことが示された。そのような設定においても今回の結果は、正規化に NPZ 法等のシングルアレイによるバッチ補正を加えることで、効果的にバッチ効果を除去できることを示している。

2.5　結言

公共データベースに登録された乳癌のマイクロアレイデータセットを用いてm RNA と対応する蛋白質の発現の一致率を評価することで、データセット間のバッチ補正がどのように関与しているかについて検討を行った。シングルアレイのバッチ補正を追加することで生物学的特徴を失うこと無しにバッチ効果を取り除け、臨床のような逐次的にサンプル

が追加される場面においては非常に有用である可能性が示唆された。さらに、適切なバッチ補正により生物学的特徴がより強調され、治療効果や予後の予測を改善しうる可能性も確認できた。今回の提案手法である NPZ 法をはじめとしたシングルアレイによるバッチ補正が、マイクロアレイを臨床で用いた際の施設間格差を是正することが期待される。さらに、臨床応用を目指した多重遺伝子診断を構築する際に公共データベースに登録されたデータを統合する必要があるが、この NPZ 法等のシングルアレイのバッチ補正を用いることで、作成したモデルを逐次的に取得される検体にそのままに適用でき、早期臨床応用が促せる可能性が期待される。

第 3 章　ショートリードクラスタリングによる乳癌細胞株での融合遺伝子検出

3.1　緒言

融合遺伝子は、染色体の転座や挿入や逆位などに染色体が組みかえられることによって遺伝子と遺伝子が連結された新たな遺伝子であり、その遺伝子が転写・翻訳されて新たなタンパク質 (融合たんぱく質) を生成するものである。近年、臨床において融合遺伝子は、診断及び治療の両面で注目を集めている [28]。慢性骨髄性白血病では、癌細胞内で染色体の第 9 番染色体の一部と第 22 番染色体の一部が切断されて相互に入れ代わる相互転座がおこっている [53]。この慢性骨髄性白血病での異常な染色体はフィラデルフィア染色体とよばれており、第 9 番染色体にある ABL1 遺伝子が第 22 番染色体にある BCR 遺伝子の下流に転座し、BCR/ABL1 融合遺伝子が作られる [54, 19]。この融合遺伝子が作る異常な BCR/ABL1 蛋白は、恒常的に活性化されたチロシンキナーゼであり、細胞の不死化を引き起こすため白血病細胞を増殖させる [30]。このチロシン・キナーゼの働きを選択的に阻害する目的で第 1 世代のチロシンキナーゼ阻害薬イマチニブメシル酸塩が創薬され、慢性骨髄性白血病の特効薬として確立している [29]。異なる癌においても融合遺伝子は着目されている。染色体の転座によって生成された EML-ALK 融合遺伝子は、肺小細胞癌患者の約 5% に認められ [55]、さらに EML-ALK 融合遺伝子が転写され、産生された異常蛋白質 EML-ALK チロシンキナーゼは、発癌に関与することが報告された [55]。前述のように融合遺伝子を同定することは、癌の詳細なメカニズムや性質を解明し治療に寄与してきた。最近では、融合遺伝子を対象として創薬された薬の対象を見つける目的だけではなく、腫瘍マーカーとしての用途でも用いられるようになっており、より精度の高い融合遺伝子の検出が求められるようになっている [56]。

かつては、融合遺伝子の同定は、分子解析を用いており、検出される領域も、一度に同定される数も限定されていた。シークエンス技術の向上に伴い短時間かつ低コストでヒトのゲノム配列の情報を取得でき応用することが可能になり、同定される融合遺伝子の数も指数関数的に増加している [57]。シークエンスによって融合遺伝子を同定するには、当初は Whole genome sequencing (WGS) を用いて行われていたが、より安価であることに加えて、転写されてより機能に影響を与えると考えられる mRNA を用いた RNA-Seq がよく用いられるようになってきた。

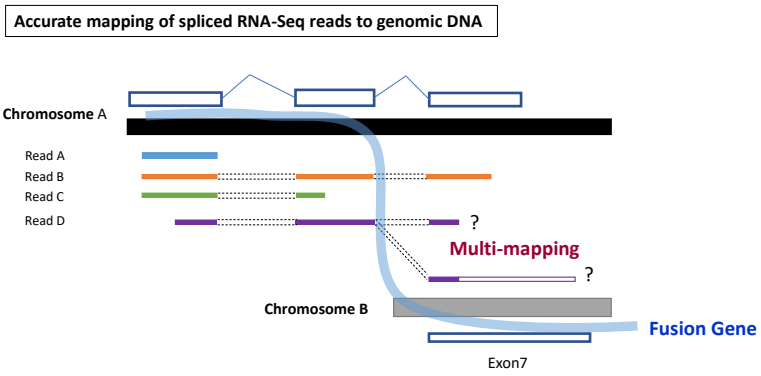


図 3.1　マルチマッピング

3.1.1　RNA-Seq の融合遺伝子検出の問題点

mRNA を用いる RNA-Seq では、WGS と異なりスプライスジャンクションも考慮してリードを参照配列にアライメントを行わなければならない。言い換えれば、配列そのものを参照配列に連続してアライメントすることができず、リードをエキソンとイントロンを区別して断続的にアライメントする必要がある。そのため、RNA-Seq を用いて融合遺伝子を同定するための様々な方法が現在までに報告されているが、融合遺伝子を同定する感度が低く、偽陽性が多い傾向がある [58] 。

RNA-Seq による低い融合遺伝子検出精度の原因の一つに参照ゲノムに対する参照配列の複数の loci にマップし得るマルチマッピング (multiple-mapped reads) がある (図 3.1)。一般的な RNA-Seq のリード長は、約 50〜100 塩基であるが、融合遺伝子のブレークポイントに適切にマッピングするためは、長さが不十分である。特に一塩基多型 (SNP) や繰り返し配列が存在する部分では、困難になることが想定される。結果としてマルチマッピングが起こりえる。リードがマルチマッピングすることは、融合遺伝子の同定に大きな影響を与える。それらは、図 3.1 のように融合遺伝子の候補となり得るからである。

融合遺伝子の候補を減らす手法として、近年、STAR アライナーを改良して融合遺伝子を検出するソフトウェアである STAR-Fusion が Hass 等によって報告された [59]。STAR-Fusion の主な改善点は、融合遺伝子のブレークポイントを予測するアルゴリズムの改善と融合遺伝子の候補のフィルタリングである。このソフトウェアは、従来提案されていた手法に比べて高速かつ高精度で融合遺伝子を同定することが可能であるが、図 3.1 のようなケースのマルチマッピングを根本から改善しているわけではない。

このリードが短いという点に対しては、シークエンスデバイスの技術革新で克服できないかという動きがある。従来のシークエンスではリードの長さは、約 50〜100 塩基が一般的であったが、近年、Oxford Nanopore や Pacbio は既にリード長が数百〜数キロの塩基のプラットフォームを提供している [60]。これらのデバイスの技術革新によってリード長が増加し、リードの特異性が向上し、multiple-mapped reads を改善する一つの手段となり得ると思われる。しかしこれまでの多くのシークエンスデータが従来のリード長で蓄積されており、また様々なアルゴリズムがそのリード長に対応するように開発されてきた。そのため、多くのソフトウェアは明らかに異なるリード長である数キロのリード長を想定されておらず現状では、そのまま用いて融合遺伝子検出に適用することができないという問題点がある。

3.1.2　提案手法:ショートリードクラスタライング法

前述のようにリードが短いことによるマルチマッピングを減少させるためにリード長をなんらかの方法で伸長させることができないかと考えた。リードを少しだけ伸長することができれば、現状のソフトウェアを利用可能である。さらに、リードの特異性が上がるためマルチマッピングが減り、融合遺伝子検出の精度が向上するのではないかと考えた。その際に着眼したのがオーバーラッピングリードである (図 3.2)。

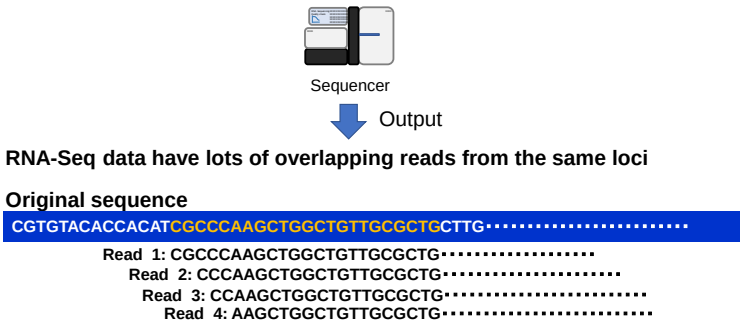


図 3.2　オーバーラッピングリード

シークエンスデータには沢山の重複リードや一部のリード配列を完全に共有するオーバーラッピングリードが含まれている。特に、高発現の遺伝子では非常に多くのリードが参照配列にマッピングされている。同じ遺伝子由来のオーバーラッピングリードを正確に同定することが出来れば、その情報を元にリードを伸張することが出来るのではないかと仮説を立てた。リードの伸張のために鈴木等 [61] が提案したショートリードクラスタリ

ング法を修正して用いることとした。この方法は、リードの mRNA の 5' 末端より数塩基 (左シフテッドリード)、mRNA の 3' 末端から数塩基削除した (右シフテッドリード)2 種類のタイプのリードを作成する。続いて清水等 [62] が開発した SlideSort ソフトウェアを適用して左シフテッドリードと右シフテッドリードの類似性を評価し、その塩基配列の共通部分を有向非巡回グラフ (Directed Acyclic Graph; DAG) に随時変換する。最終的に類似性によって連結された DAG の情報を用いて代表配列を決定する。この手法は、本来リードのエラーや SNP を検出することを目的に作成されたものである。このエラー修復用のショートリードクラスタリング法を新たに RNA-Seq データに適用するために改変した。この手法によって代表リードを検出して元リードを置き換えることでリードを伸長させ融合遺伝子検出精度の改善を試みた。

3.2 実験材料と方法

RNA-Seq データを品質管理した後に、後述する提案手法を用いた場合と用いない場合とで融合遺伝子の検出精度にどのような影響を与えるかについて、リード長の異なる RNA-Seq データに適用して検出精度をそれぞれ評価した。本実験で用いた融合遺伝子の真の正解に関しては、後述する通りデータベースに登録されたものを真の正解とした場合と RT-PCR 法での結果のみを正解とした場合との 2 つの方法で其々評価した。

3.2.1 使用したデータセット

大阪大学 (OU) データセット

4 つのヒト乳癌細胞株 (BT-474、MCF-7、SKBR-3、及び T-47D) の RNA-Seq データは、大阪大学乳腺内分泌外科から提供され本検討で使用した。そのデータの詳細は下記の通りである (表 3.1)。

表 3.1 OU データセットのシークエンス結果

サンプル	リード数	リード長	平均クオリティ	リードタイプ
BT-474	23,815,977	101 bp	36.6	ペアエンド
MCF-7	23,301,615	101 bp	36.6	ペアエンド
SKBR-3	22,318,738	101 bp	36.6	ペアエンド
T-47D	22,269,402	101 bp	36.6	ペアエンド

シークエンスリードアーカイブ (SRA) データセット

リード長が異なるデータセットでの融合遺伝子の検出精度を調べるために、SRA からの 2 つの RNA-Seq データセット (https://www.ncbi.nlm.nih.gov/sra) を取得した (表 3.2)。

表 3.2　SRA データセットのシークエンス結果

サンプル	リード数	リード長	リードタイプ	SRA ID
MCF-7 50bp	8,412,431	50 bp	ペアエンド	SRX025827
MCF-7 80bp	27,681,356	80 bp	ペアエンド	SRR5808854

3.2.2　RNA-Seq リードの品質管理

Bolger ら [63] が開発した Illumina NGS データ用のトリミングツール Trimmomatic-0.36 を用いて，低品質のリードや塩基のトリミングを行い，アダプター配列を除去した (解析コード: trim_galore –quality 20 –phred33 –stringency 3 –gzip –length 50 –paired [リード 1 データ] [リード 2 データ])。最終的に、このソフトウェアで処理されたリードは、後に比較される元のデータとした。上記のソフトウェアでは、SSC 法後のリードも同様に処理した。

3.2.3　提案手法: シフトテッドショートリードクラスタリング (SSC) 法

SSC 法は、次の 5 つのステップで構成されている。**ステップ 1: 重複したリードの除去** RNA-Seq データには、非常に多数の重複リードが含まれており、リードがユニークとなるように除去する。

ステップ 2: シフテッドリードの作成

各リードのシフト幅を w、最大のシフト幅を W と定義し、図 3.3 のように 0 から W まで w 塩基ずつリードを刈り込む。最初のステップでリードの先頭もしくは末尾から w 塩基分をカットしたリードを左シフテッドリード、右シフテッドリードとそれぞれ定義する。その結果、新たに W 種類の長さのリードが作成される。

ステップ 3: 類似リードの検索及びクラスタリング

左シフテッドリードと右シフテッドリードの類似性を比較した際のミスマッチの距離を d と定義する。ステップ 3 では、それぞれシフト幅 w 刈り込まれた左シフテッドリードと

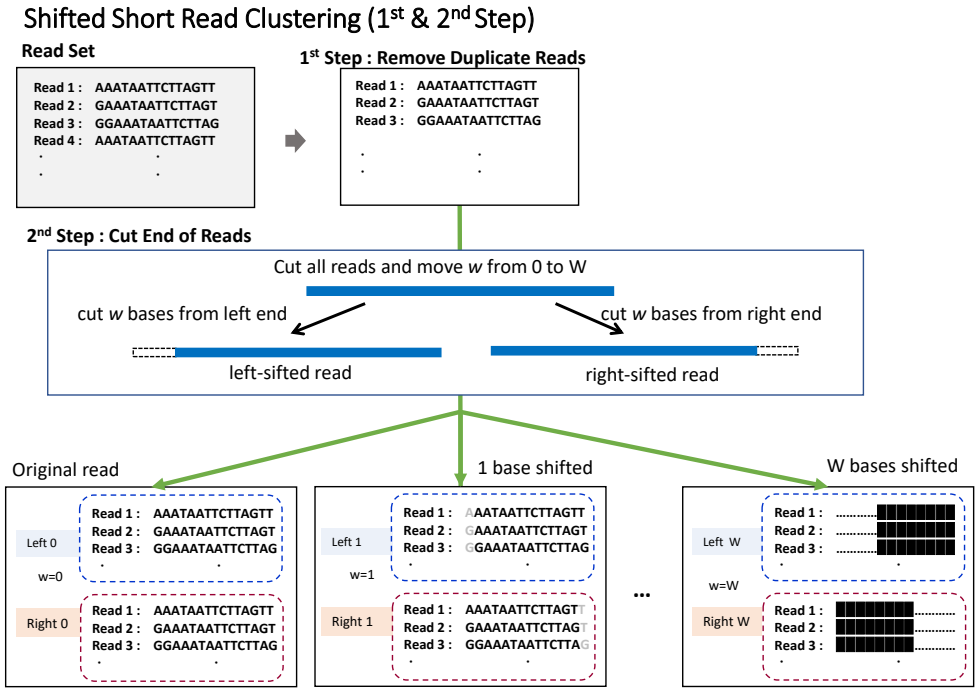


図 3.3　SSC 法 ステップ 1〜2

右シフテッドリードを距離 d で類似性を比較してクラスタリングする。本研究では、特異性を高めるために $d = 0$ とした。 また、リードの類似性を判断するために、類似したペアをリードの本数の増加に対して線形時間で見つけることができる SlideSort を用いて、シフテッドリードのクラスタリングを行う (図 3.4)。

ステップ 4: 類似リードの連結

生成されたリードクラスターは、1 つずつ DAG 形式に変換される。ここで、ノード、エッジ、ラベルはそれぞれ、ユニークな部分配列、リードの連結、観測された部分配列の数を示している。読み込みにミスマッチがある場合、グラフは分岐し DAG が伸長される。代表配列の決定には、分岐点で最も観測されたノードで選択するという多数決を採用する。代表配列の追跡及び決定は貪欲法にて決定され、非代表配列は追跡しない (図 3.5, 図 3.6)。

ステップ 5: リードの再構成

最終的に代表配列の DAG からのリードへ変換される。このステップで代表配列の DAG を構成する全てのリードを代表配列に置き換える (図 3.6)。

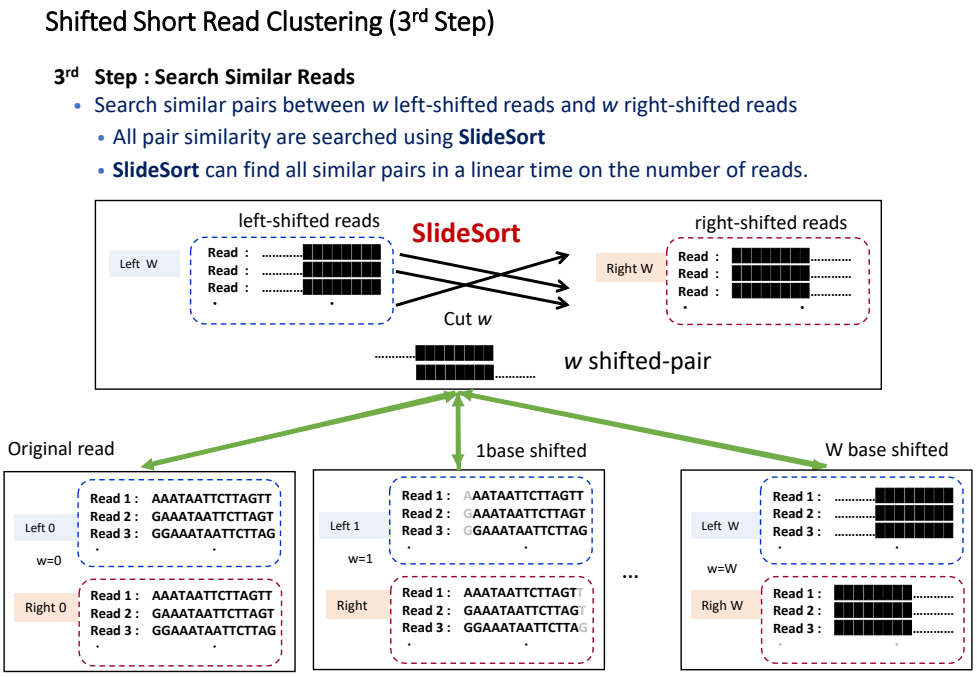


図 3.4　SSC 法 ステップ 3

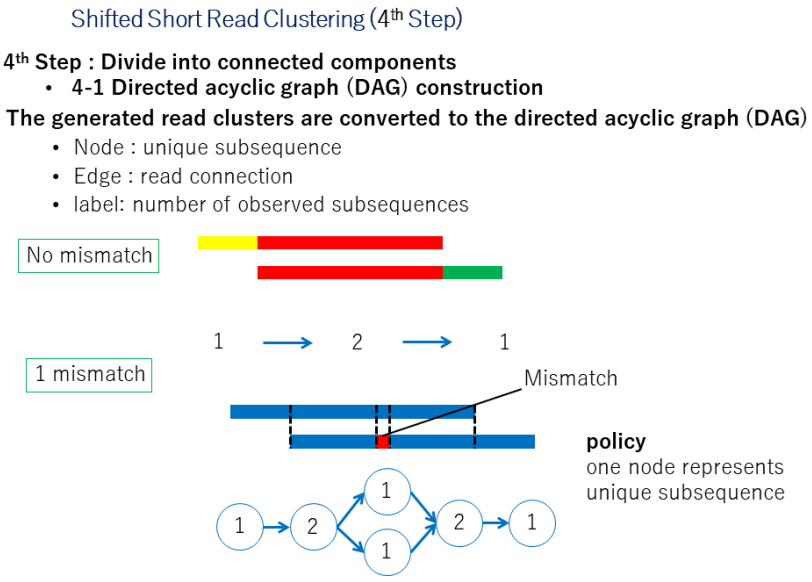


図 3.5　SSC 法 ステップ 4-1

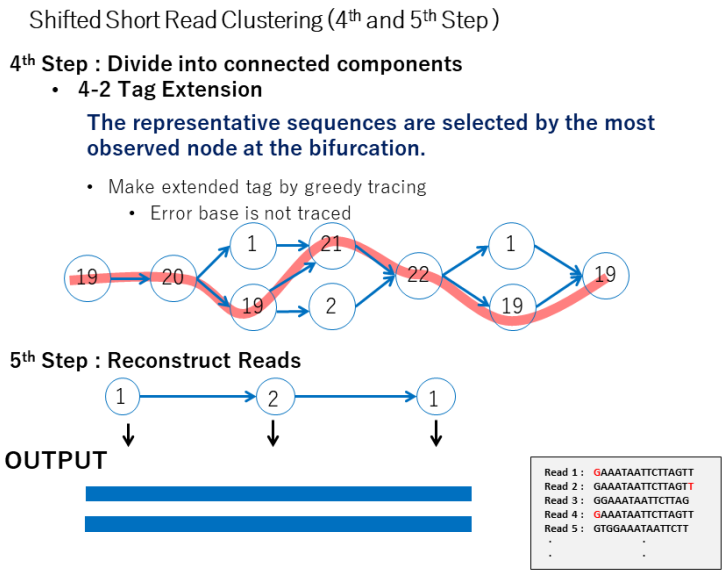


図 3.6　SSC 法 ステップ 4-2〜5

3.2.4　融合遺伝子の検出

融合遺伝子の検出には、ヒトゲノム GRCh38 をリファレンスゲノムとして STAR-2.6.0a と STAR-Fusion-v1.4.0 を用いた。

3.2.5　融合遺伝子の検出精度の評価方法

融合遺伝子の検出は、以下の定義で評価した。各評価について、真陽性 (TP)、偽陽性 (FP)、偽陰性 (FN) を算出した。感度 (または再現率、真陽性率) と陽性予測値 (PPV；または精度) は、以下の式に従って計算した。

$$\mathbf{Sensitivity = TP\ /\ (TP + FN)}$$

$$\mathbf{PPV = TP\ /\ (TP + FP)}$$

STAR-Fusion で同定された全ての融合遺伝子候補は、Cancer Cell Line Encyclopedia (CCLE; http://www.broadinstitute.org/ccle) のデータベースで検索し、本研究では、同一細胞株名で CCLE データベースに登録されている融合遺伝子を真の陽性とした (評価方法 1)。CCLE に登録されている融合遺伝子は、各種細胞株に対して、deep RNA-seq、WES、WGS、Reduced Representation Bisulfite Sequencing による DNA メチル化解析、マイクロ RNA 発現プロファイル、ならびにヒストン H3 修飾プロファイルのいずれかによって同定されている。つまり主に NGS で解析された融合遺伝子である。さらに、

全ての融合遺伝子候補は、PubMed の論文検索で、RT-PCR により検証された融合遺伝子のみを真の陽性とした解析でも評価された (評価方法 2)。この解析を行った理由は、偽陽性を多く含む NGS に比べて RT-PCR による融合遺伝子検出は、これまで広く良く行われており、数個の遺伝子のみが検出可能であるという制限があるが、非常に精度が高いことが知られている [20] ためである。

3.3 結果

3.3.1 OU データセットでのマッピング結果

OU データセットでは、4 つの細胞株の RNA-Seq の生データを SSC 法に適用した。この手法の性質上、シフト幅を増加させることでより多くのオーバーラッピングリードと同一クラスターに分類され、その結果としてリードが伸長される。そのため本研究では、SSC 法の有効性を検証するために、SSC 法のシフト幅を 1、2、5、10 塩基に変更した。シフト幅を 1、2、5、10 塩基に変更することで、オリジナルデータの 201 塩基が平均して 217、234、282、317 塩基に伸長された (表 3.3)。続いて、元データ (アダプター除去と QC コントロールのみ) と SSC 法で処理したデータをそれぞれ STAR-Fusion に適用した。STAR-Fusion では、元データのユニークマップリードの割合は、平均で 89 % (BT-474: 88 %、MCF-7: 88 %、SKBR-3: 92 %、T-47D: 87 %) となった。一方でリードを 1 塩基ずらした場合、平均で 94 %と最も高い (BT-474；93 %、MCF-7；94 %、SKBR-3；95 %、T-47D；92 %) 結果となった。しかし、SSC 法では、シフト幅が 2 塩基以上になると、単一遺伝子座に合致したリードの割合が減少した。その一方で複数の遺伝子座に合致したリードは、1 塩基ずらすと元のデータの 9.7 %から 4.6 %に減少し、10 塩基ずらすと元のデータの約 4 分の 1 以下に減少した。

3.3.2 OU データセットにおける STAR-Fusion の融合遺伝子の結果 (評価 1)

CCLE データベースに登録されている 145 個の融合遺伝子 (BT-474; n = 46, MCF-7; n = 55, SKBR-3; n = 30, T-47D; n = 14) を真の融合遺伝子と定義して融合遺伝子の検出精度について比較検討した。元データにおける融合遺伝子の感度は、平均で 47% (BT-474；48%, MCF-7；49%, SKBR-3；50%, T-47D；43%) であった (表 3.4)。

SSC 法で塩基を 1 つずらすと、真の融合遺伝子の数は 70 から 79 に増え、融合遺伝子の感度は平均で 54% に上昇した (BT-474；57%、MCF-7；53%、SKBR-3；57%、T-47D；50%)。5 塩基以上シフトさせるとシフト幅によって融合遺伝子検出の感度は元データよ

表 3.3　OU データセット (BT-474、MCF-7、SKBR-3、及び T-47D) における STAR-aligner によるアライメント結果

	元データ	シフトテッドショートリードクラスタリング			
		1 塩基シフト	2 塩基シフト	5 塩基シフト	10 塩基シフト
平均入力リード長	201	217±29	234±57	282±128	317±148
単一の遺伝子座にマッピング:					
リードの数	19518515	20824794	19648767	17913912	16756658
リードの割合 (%)	89%	94%	86%	80%	75%
平均のリード長	200	213	214	227	245
複数の遺伝子座にマッピング:					
リードの数	2127215	1019722	929272	711900	443976
リードの割合 (%)	9.7%	4.6%	4.1%	3.2%	2.0%

りは良いものの 10 塩基シフトにて 52% まで低下した。また、PPV についても検討した結果、元データの PPV は平均で 39% であった。SSC 法で 1 塩基シフトした場合、融合遺伝子候補は 176 個から 208 個に増加し、PPV は 37% であった。しかし、さらに SSC 法で 2 塩基以上シフトさせると、融合遺伝子候補の数が減り、PPV が改善された。最終的に、SSC 法で 10 塩基をシフトした場合、PPV は 51% となった。

表 3.4　OU データセットにおける STAR-Fusion の融合遺伝子の結果 (評価 1)

	元データ	シフトテッドショートリードクラスタリング			
		1 塩基シフト	2 塩基シフト	5 塩基シフト	10 塩基シフト
融合遺伝子候補の数	176	208	183	164	152
検出された真の融合遺伝子の数*	70	79	79	78	76
感度*	47%	54%	54%	53%	52%
陽性的中率 (PPV)*	39%	37%	42%	47%	51%

*Cancer Cell Line Encyclopedia に登録されている 145 個の融合遺伝子を真の融合遺伝子と定義

表 3.5　OU データセットにおける STAR-Fusion の融合遺伝子の検出結果 (評価 2: RT-PCR によって確認されている融合遺伝子のみを対象)

					シフトテッドショートリードクラスタリング					
	5'	5'	3'	3'		1 塩基	2 塩基	5 塩基	10 塩基	
細胞株	遺伝子	染色体	遺伝子	染色体	元データ	シフト	シフト	シフト	シフト	参照
BT-474	ACACA	17	STAC2	17	○	○	○	○	○	[64]
BT-474	RPS6KB1	17	SNF8	17	○	○	○	○	○	[64]
BT-474	VAPB	20	IKZF3	17	○	○	○	○	○	[64]
BT-474	ZMYND8	20	CEP250	20	○	○	○	○	○	[64]
BT-474	RAB22A	20	MYO9B	19	○	○	○	○	○	[64]
BT-474	SKA2	17	MYO19	17	○	○	○	○	○	[64]
BT-474	STARD3	17	DOK5	20	○	○	○	○	○	[64]
BT-474	GLB1	3	CMTM7	3	○	○	○	○	○	[64]
BT-474	THRA	17	AC090627.1	17	○	○	○	○	○	[65]
BT-474	TOB1	17	SYNRG	17	○	○	○	○	○	[65]
BT-474	AHCTF1	1	NAAA	4	○	○	○	○	○	[65]
BT-474	MED1	17	STXBP4	17	○	○	○	○	○	[65]
BT-474	MED1	17	ACSF2	17	○	○	○	○	○	[65]
BT-474	STX16	20	RAE1	20	○	○	○	○	○	[65]
BT-474	MED13	17	BCAS3	17	○	○	○	○	○	[65]
BT-474	TRPC4AP	20	MRPL45	17	○	○	○	○	○	[65]
BT-474	PIP4K2B	17	RAD51C	17	○	○	○	○		[65]
BT-474	USP32	17	MED1	17		○	○			[65]
BT-474	DIDO1	20	TTI1	20		○	○	○	○	[64]
BT-474	LAMP1	13	MCF2L	13		○	○	○	○	[64]
BT-474	CPNE1	20	PI3	20						[64]
MCF-7	BCAS4	20	BCAS3	17	○	○	○	○	○	[64]
MCF-7	ARFGEF2	20	SULF2	20	○	○	○	○	○	[64]
MCF-7	RPS6KB1	17	VMP1	17	○	○	○	○	○	[64]
MCF-7	AC099850.1	17	VMP1	17	○	○	○	○	○	[65]
MCF-7	SMARCA4	19	CARM1	19	○	○	○	○		[65]
SKBR-3	TATDN1	8	GSDMB	17	○	○	○	○	○	[64]
SKBR-3	CSE1L	20	AL035685.1	20	○	○	○	○	○	[64]
SKBR-3	RARA	17	PKIA	8	○	○	○	○	○	[64]
SKBR-3	ANKHD1	5	PCDH1	5	○	○	○	○	○	[64]
SKBR-3	CCDC85C	14	SETD3	14	○	○	○	○	○	[64]
SKBR-3	SUMF1	3	LRRFIP2	3	○	○	○	○	○	[64]
SKBR-3	WDR67	8	ZNF704	8						[64]
SKBR-3	CYTH1	17	EIF3H	8	○	○	○	○	○	[64]
SKBR-3	DHX35	20	ITCH	20		○	○	○	○	[64]
SKBR-3	NFS1	20	PREX1	20						[64]

3.3.3　OU データセットにおける STAR-Fusion の融合遺伝子検出の結果 (評価 2)

先の検討で用いていた CCLE に登録されている融合遺伝子は、前述の通り NGS で検出された融合遺伝子が多く、実際に RT-PCR で検証された融合遺伝子は少ない。そのため CCLE に登録されている融合遺伝子全てが真の陽性とは言えないため、2000 年 1 月

以降に報告された MCF-7、BT-474、SKBR-3、T-47D の融合遺伝子に関するすべての論文を検索し (2018 年 8 月に PubMed を検索した時点)、それらの論文中で RT-PCR で確認された融合遺伝子を真の融合遺伝子としたところ、3 つの細胞株 (BT-474、MCF-7、SKBR-3) で合計 36 個の融合遺伝子が同定された。

元データにおける融合遺伝子の感度は 81% (BT-474；81%、MCF-7；100%、SKBR-3；70%) であった。SSC 法で 1 塩基シフトした場合、融合遺伝子の感度は 92% (BT-474；95%、MCF-7；100%、SKBR-3；80%) となり、元データよりも高い感度を示した。一方で、元データの PPV (20%) は、SSC 法で 1 塩基シフトしたデータ (18%) よりも高い結果となった (表 3.5)。評価 1 と同様にシフト幅を増加させると検出感度は徐々に低下し、10 塩基シフトした場合に 83% となった。PPV は、元データでは、16.4% に対して 1 塩基シフトした場合には 15.8% と若干低下したが、10 塩基シフトした場合に 19.7% まで上昇した。

表 3.6　SRA データセットにおけるアライメント結果と STAR-Fusion による融合遺伝子検出の結果

	MCF-7 データセット ペアエンド 50 塩基		MCF-7 データセット ペアエンド 80 塩基	
	元データ	1 塩基シフト	元データ	1 塩基シフト
平均入力リード長	99	116	159	170
単一の遺伝子座にマッピング:				
リード数	5838628	5880470	25546925	25348026
% 割合	75%	73%	92%	94%
マップされた平均リード長	98	104	159	169
複数の遺伝子座にマッピング:				
リード数	1730123	1079805	1843755	1296910
% 割合	22.1%	13.3%	6.7%	4.8%
融合遺伝子検出:				
融合遺伝子の候補の数	22	20	56	76
検出された真の融合遺伝子の数*	13	12	22	26
感度*	24%	22%	40%	47%
陽性的中率 (PPV)*	59%	60%	39%	34%

*Cancer Cell Line Encyclopedia に登録されている融合遺伝子を真の融合遺伝子として定義

3.3.4 SRA データセットにおけるマッピングと融合遺伝子検出の結果

SRA データセットの異なるリード長 (50bp、80bp) に対して、SSC 法の有用性を検証した。本研究では、SSC 法によって 1 塩基だけシフトしたリードを元データと比較した。MCF-7 の 50bp のデータセットでは、STAR にユニークにマッピングされたリードの割合は、SSC 法 (73%) よりも元データ (75%) の方が高かった。感度と PPV はほとんど改善されなかった。MCF-7 80bp データセットでは、OU データセットと同様に、オリジナルデータと比較して、 ユニークにマップされたリードの割合が 92% から 94% に増加し、感度も 40% から 47% に増加した (表 3.6)。

3.4 考察

融合遺伝子を検出するアルゴリズムはこれまでに様々なものが提案されているが、検索した限りでは入力されたリードをあらかじめ伸長して精度を高める方法は報告されていない。リードを伸長するためには、ゲノムへのマッピングが考えられるが、融合遺伝子のブレイクポイントからのリードはゲノムにマッピングされないため、その効果は限定的であると思われる。そこで、本研究では参照ゲノムを使わずにリードを伸長するために、オーバーラップするリードに着目した。同一遺伝子からのオーバーラップしたリードは部分的に異なるため、多数のリードから真のオーバーラップされたリードのペアを見つけることは困難である。そこで、リードを切断してシフトしたリードを作成する方法を採用した。この方法では、シフトされたリードを作成する際に、シフト幅に応じて一時的にデータ量が増えてしまうという問題がある。また、同時に増加したリードの類似性を素早く判断する必要がある。これらの問題を解決するために、SSC 法に SlideSort を組み込むことにした。SSC の SlideSort は、与えられた編集距離内にあるリードの類似性を、リードの数に比例した (線形) 計算時間で素早く判定するという重要な役割を果たしている。

今回の研究では、リードの長さが異なる場合 (80bp、101bp) でもリード伸長法の有効性が示されたが、これには 2 つの利点があると考えられる。1 つは、リードの伸長により、類似した配列に対する特異性が高まることである。実際にリードを伸長させることでマルチマッピングリードが減少した。もう 1 つは、トリミングの改善であると考えられる。実際に OU データセットの BT-474 細胞株において、元データからトリミングにて排除されるリードは、116,088 本であったが、1 塩基、2 塩基、5 塩基、10 塩基シフトするごとにトリミングされるリード数は、680,850 本、898,969 本、1,038,967 本、1,074,370 本に増

加した。これは、リードの伸長により、塩基数が少なすぎて区別がつかないアダプター配列やポリ (A) テールを除去可能にすることができた結果である。しかし、今回の研究では、リードが長くなるにつれて、ユニークにマッピングされるリードの比率や融合遺伝子の検出感度が悪くなることも明らかになった。この問題には、下記の原因が考えられる。シフト幅を大きくすると、本質的に望ましくないリードを結合する機会が増えていたその結果としてマッピングされるリードがシフトするごとに減少した。SRA データセットの解析では、元データでは、リード全体の中で複数の遺伝子座にマッピングされたリードが 20 %以上認めた。このような状況では、SSC 法では望ましくない結合が起こり、結果が悪くなる傾向が示されたものと考えられる。2 つ目の理由は、既存のソフトウェアが短いリードを想定して作られていることである。そのため、ショートリードの想定を超えた長いリードには不向きだと考えられる。マッピングされなかった BAM ファイルの結果から、参照配列と完全に一致しながらも、通常よりも長く伸びすぎたリードは、STAR にマップされなかった。その理由の一つは今回の STAR-Fusion のマッピングソフトである STAR の設定は、ペアリード間の距離などによってマッピングの精度が変化する。そのため、提案手法との精度を比較するためデフォルトの設定で使用したが、ペアエンド用のマッピングソフトが、過度にクロスオーバーしてマッピングすることを想定していないからである。

癌のメカニズムや特徴を解明するためには、融合遺伝子の検出感度が重要である。また、融合遺伝子が最近の治療法の標的となっていることから、薬剤候補を探索する上で、感度の重要性はますます高まっている。一方で、融合遺伝子を検査や腫瘍マーカーとして利用する場合には、PPV も非常に重要である。少ないシフトされたリードを用いるこの方法は、融合遺伝子の検出感度を向上させる可能性がある。また、より多くのシフト幅を用いる方法は、PPV の向上につながるという結果であった。本研究によって検出の評価 1 における感度は 80 塩基、102 塩基のデータ共に約 1.1 倍に増加し、約 50 %まで向上することができたが、癌の個性を評価する上では、不十分である。評価 2 を基準として解析した場合の感度は、約 81% から 92% まで増加して臨床応用における目標とされる 95% に迫っている。その一方で、PPV は、評価 1 で約 50%、評価 2 では 20% とまだまだ十分な結果とは言えないかもしれない。しかしながら、腫瘍マーカーとして融合遺伝子を用いる際には、RT-PCR の対象配列を作成するための融合遺伝子の候補にマッピングするリードの数で上位数十個を対象とすることが想定される。そのような観点では、OU データセットの評価 1 の BT-474 の解析では、元データで融合遺伝子にマッピングされたリー

ドの数上位 20 位中 12/21 個 (60%) が検出されており、シフト幅を 1、2、5、10 塩基に増加させると 14/21 個 (70%)、15/21 個 (71.4%)、17/25 個 (68%) と 15/21 個 (71.4%) が検出されており目的には比較的合致した精度であると考えられる。今回の研究では、どのシフト幅が最適なのかは言及できないが、目的に応じて元のリードよりもリードを伸長した方が良い結果が得られることが示唆された。

3.5 結言

本研究では、リードの伸長によって複数の遺伝子座にマッピングされるリードが減少することを実証した。さらに、シフトしたリード幅に応じて、融合遺伝子の検出感度や PPV が向上することを示した。この方法は、前処理として働くため既存の手法がそのまま使えるという点で非常に有用であると考えられる。しかしながらリードの過度な伸長にともなう本手法の欠点や 50 塩基のデータセットのような短いショートリードでの改善効果がない点など本手法の限界点や今後の改善点も明らかとなった。今回、この手法は、RNA-Seq データの元リードを伸長させることで融合遺伝子の検出精度を向上させたが、リードを伸長させることができるのは RNA-Seq に限ったことではない。本研究では、高発現の遺伝子で (リード数の多い遺伝子で) は正確な伸長を認めたが、リード数を予め規定したエクソーム解析などの DNA のデータではどのように作用するか不明瞭である。今後は、様々なデータセット (RNA-Seq だけでなく、エクソームや全ゲノム配列など) やリード長での検証が必要である。

第 4 章　結論

4.1　研究結果のまとめ

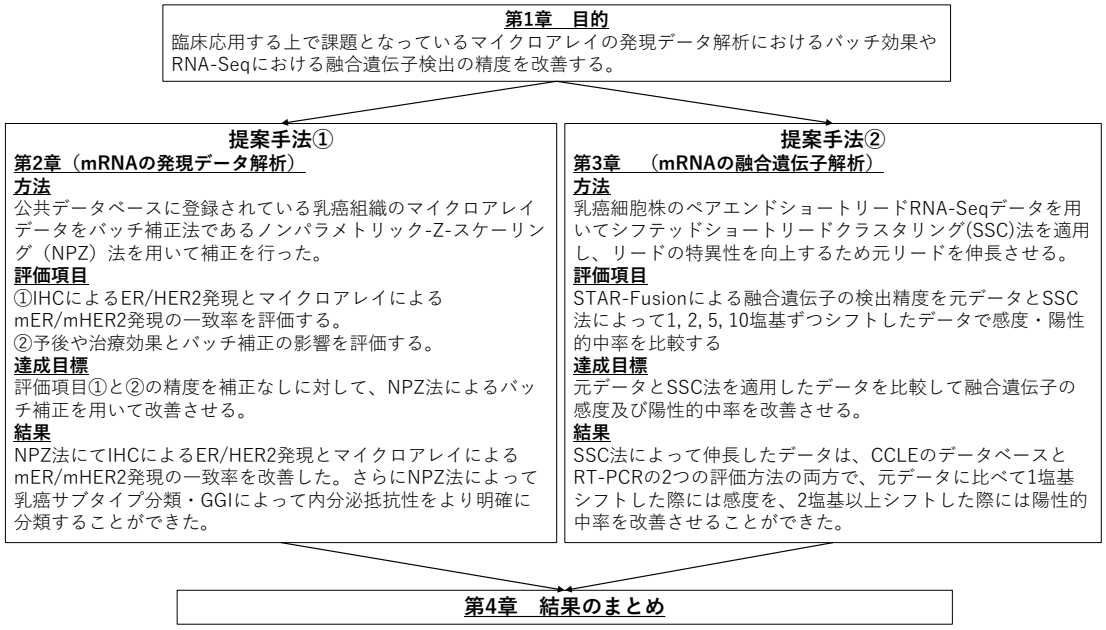


図 4.1　本論文の結果のまとめ

本論文では網羅的なトランスクリプトームの解析の課題についてマイクロアレイおよび RNA-Seq にそれぞれ焦点を当てて研究を行ってきた（図 4.1）。マイクロアレイの解析は、網羅的な発現量を一度に取得可能にする。しかし、その反面で、マイクロアレイの特性上異なる施設から得られたデータを統合して解析する際に”バッチ効果”によって困難にされている。また、RNA-Seq によってトランスクリプトームの解析は、発現定量にとどまらず融合遺伝子などの構造変異を捉えることが近年のシークエンス技術の革新によって可能となったが、その検出精度は不十分である。臨床応用のための 2 つの課題に対して、第 2 章ではマイクロアレイ、第 3 章では RNA-Seq データに対して前処理を加えることで問題解決を目指した。

第 2 章では、複数の施設から得られた乳癌のマイクロアレイのデータに対して、バッチ補正の有用性について検討した。臨床に使用される場面を想定し、サンプルごとに補正が完結するシングルアレイによるバッチ補正法である NPZ 法を提案した。各種正規化法に対して、バッチ補正の有無で臨床検査として確立している ER や HER2 の発現との一致

率がどのように変化するかについて検討した結果、NPZ 法をはじめとしたシングルアレイによるバッチ補正が、補正なしやマルチアレイによるバッチ補正よりも ER や HER2 の発現と高い一致率を認めた。この結果によって分類された乳癌のサブタイプによって内分泌療法の効果をバッチ補正なしに比べて大きな統計学的有意差をもって選別できることも同時に示した。また、これまでに報告されている遺伝子シグネチャーである GGI や PAM50 といったモデルに対しても適用した結果、それらの精度も改善することが示された。本検討においてこれまでのバッチ補正の標準的な手法であったマルチアレイによるバッチ補正が、ひとくくりに乳癌とされても各サブタイプによって遺伝子発現が大きく異なる癌腫の場合には、かえってその特性を失わせていることも確認されデータ統合の難しさが示された。各種解析において NPZ 法によるバッチ補正は、安定して効果を示しておりマイクロアレイの発現解析への適用可能性が示唆された。

第 3 章では、乳癌の細胞株を用いて RNA-Seq による融合遺伝子の検出について検討した。融合遺伝子を検出する手法は各種報告されているが、検出精度の問題点としてショートリードのリード長に一因があると仮説を立て、そのリード長を伸長することで検出感度や陽性的中率を改善できるかについて比較した。そのため同一遺伝子由来のリードを検出して伸長させるシフテッドショートリードクラスタリング法を提案して適用した結果、ペアエンドリード長が 80 塩基以上のデータにおいては、元データを伸長させることで STAR-Fusion の遺伝子の検出感度が改善されることが示された。また、リードを伸長させることで陽性的中率も改善する傾向にあったが、感度も若干低下しておりこの手法の限界点や問題点も確認された。シフテッドショートリードクラスタリングは、目的や用途に応じて使用を工夫することで既存のデータでは見逃されている融合遺伝子を検出できる可能性があり融合遺伝子検出に有用である可能性が示唆された。

4.2 乳癌研究における本研究の寄与と今後の展望

腫瘍を治療・克服するためには、まず適切に診断することが必要であり、肉眼的に評価する臨床診断から顕微鏡を用いた病理診断へ、そしてトランスクリプトームやゲノムを解析する分子診断が加わり、それらによって得られた結果による治療の個別化が求められている。つまり、乳癌の中には、腫瘍細胞や正常の乳管上皮細胞、免疫細胞、間質細胞が入り混じって形成された複雑な微小環境を少数の一部の遺伝子のみで説明し、治療することは困難であると考えられる。そのため先に述べた複雑な微小環境を捉えるためには網羅的なトランスクリプトームの解析が必要である。

ヒト遺伝子は約 2 万 2000 個 [66] 存在すると報告されているが、それらの全トランスクリプトームを解析する場合において縦軸を遺伝子数、横軸をサンプル数とした場合に、100 例解析したとしても縦長なアンバランスなデータとなるため統計解析上問題となり、正しい結果を得られない可能性がある。それを克服するためには同一癌腫を統合してサンプル数を増やすことが解決策の一つとして考えられるが、同一なプラットフォームを用いたとしてもバッチ効果があり困難である。また今日ではマイクロアレイもメーカによって様々なプラットフォームが提供されているが、搭載されている遺伝子数や配置されている場所も大きく異なりそれらを全て統合して解析することはほぼ不可能である。その一方で RNA-Seq は、実際に転写産物のリードを計測して発現量を算出してるため上記の問題解決に期待されている。しかしながら、これまで公共データに登録された乳癌研究における網羅的なトランスクリプトーム解析は、マイクロアレイデータが主力であり、それらにはヒトに実際に投与された抗癌剤感受性の結果や 10 年を超えた予後の結果など貴重な臨床データも付随している。

本研究で提案した NPZ 法は、同一プラットフォームにおけるバッチ効果を様々な解析において安定して除去することが示されており現在までに蓄積されたデータを有効活用するための重要なツールとなりえる可能性が考えられる。遺伝子発現解析の新しい標準プラットフォームとなっている RNA-Seq においても過剰に分散したカウントデータであり”バッチ効果”が認識され、それらに対してマイクロアレイで開発されたバッチ補正法である ComBat が有効にバッチを補正できると近年報告されている [26]。NPZ 法は、外れ値に強い特性もあり RNA-Seq の補正に関しても今後貢献できると期待される。

遺伝子の解析目的も多岐にわたり、腫瘍の”ゲノムの不安定性と変異”の解明から原発巣で検出した融合遺伝子を血液中の循環 DNA から検出して再発の有無や治療効果に応用するといった”より個別化されたマーカーの検索”まで及んでいる。融合遺伝子の検出は、ロングリードを用いた RNA-Seq によって今後加速していくことが期待される。実際に今回提案したシフテッドショートリードクラスタリング法でのみ検出した融合遺伝子を同じ mRNA を用いたロングリードシークエンスにて検出しその再現性を確認している (資料未記載)。これらの技術革新は、今回の研究課題であったリードの短さを根本から解決する手法であるが、リードのシークエンス精度の低さから偽陽性も多く定まった標準的な解析ツールが確立していない。さらに TCGA をはじめとした多くの臨床データを有した膨大な RNA-Seq データは、ショートリードシークエンスが中心でありショートリードにおける精度の改善は不可欠である。しかしながら今回の検討で示された通り、現在のショー

トリードを用いた融合遺伝子検出ツールでの標準的な感度および PPV は約 50 %に満たない結果であり、腫瘍の生物学的意義や共通の変化を正確に捉えるに足るためにはまだまだノイズの多い不十分な結果である。本研究で提案したリードを伸長させるシフテッドショートリードクラスタリング法は、自験例のデータの感度を 5%、PPV を 10% 改善したが、それでもようやく 50% に達した程度である。また、トランスクリプトームにおける融合遺伝子検出は比較的新しい分野であり、本研究において用いた CCLE をはじめとした細胞株の融合遺伝子のデータベースにおいても真の融合遺伝子かどうかについて確証が得られない状態である。そのため、本研究では RT-PCR によって検証されている融合遺伝子に絞っても改善効果を検証する必要性があった。シフテッドショートリードクラスタリング法は、入力リードを伸長させて元リードを置き換えるため既存のソフトウェアをそのまま用いることを可能にするため、今後新たな融合遺伝子検出ソフトが開発された際にも、そのまま適用可能であるという利便性がありさらなる精度の向上が期待される。

癌研究におけるトランスクリプトーム解析は、前述の Hallmark を解明・把握するために第 1 章で述べた様に遺伝子自体の異常の解析と遺伝子の機能の異常の解析に今後改善が望まれる課題がある。本研究ではそれらの課題の一つであるマイクロアレイによるバッチ補正と融合遺伝子解析に既存の取り組んだにすぎないが、腫瘍本来の個性を正確に評価すると同時に多様性を把握・理解するためには一定のデータの集積の両方が重要である。本研究で提案した手法は、いずれも前処理法であり、既存の手法にそのまま適応可能であり、既存の手法の精度をより向上させることができることを示した。臨床での融合遺伝子の有用性が認識され、マイクロアレイの使用が臨床で一般化しつつあり、将来的に RNA-Seq の臨床応用が期待される今日において、過去のデータと現在とをつなぐ手法として、本研究が寄与し貢献することを今後期待する。

参考文献

[1] E. V. Jensen and E. R. DeSombre. Estrogen-receptor interaction: Estrogenic hormones effect transformation of specific receptor proteins to a biochemically functional form. *Science*, Vol. 182, No. 4108, pp. 126–134, 1973.

[2] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, et al. Molecular portraits of human breast tumours. *Nature*, Vol. 406, No. 6797, pp. 747–752, 2000.

[3] J. Taylor-Papadimitriou, M. Stampfer, J. Bartek, A. Lewis, M. Boshell, E. Lane, and I. Leigh. Keratin expression in human mammary epithelial cells cultured from normal and malignant tissue: relation to in vivo phenotypes and influence of medium. *Journal of Cell Science*, Vol. 94, No. 3, pp. 403–413, 1989.

[4] L. Ronnov-Jessen, O. W. Petersen, and M. J. Bissell. Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction. *Physiological Reviews*, Vol. 76, No. 1, pp. 69–125, 1996.

[5] L. A. Carey, E. C. Dees, L. Sawyer, L. Gatti, D. T. Moore, F. Collichio, D. W. Ollila, C. I. Sartor, M. L. Graham, and C. M. Perou. The triple negative paradox: primary tumor chemosensitivity of breast cancer subtypes. *Clinical Cancer Research*, Vol. 13, No. 8, pp. 2329–2334, 2007.

[6] A. Yoshimura, H. Ito, Y. Nishino, M. Hattori, T. Matsuda, I. Miyashiro, T. Nakayama, H. Iwata, K. Matsuo, H. Tanaka, et al. Recent improvement in the long-term survival of breast cancer patients by age and stage in japan. *Journal of Epidemiology*, Vol. 28, No. 10, pp. 420–427, 2018.

[7] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, Vol. 144, No. 5, pp. 646–674, 2011.

[8] 日本乳癌学会. 総論 2, 乳癌 intrinsic subtype と免疫組織化学的方法を用いた代替定義について」. 乳癌診療ガイドライン 2　疫学・診断編, 第 3 版, p. 225–226, 2015.

[9] G. Curigliano, H. J. Burstein, E. P. Winer, M. Gnant, P. Dubsky, S. Loibl, M. Colleoni, M. M. Regan, M. Piccart-Gebhart, H.-J. Senn, et al. De-escalating and escalating treatments for early-stage breast cancer: the st. gallen international expert consensus conference on the primary therapy of early breast cancer

2017. *Annals of Oncology*, Vol. 28, No. 8, pp. 1700–1712, 2017.

[10] E. Clough and T. Barrett. The gene expression omnibus database. In *Statistical Genomics*, pp. 93–110. Springer, 2016.

[11] A. Prat, E. Pineda, B. Adamo, P. Galván, A. Fernández, L. Gaba, M. Díez, M. Viladot, A. Arance, and M. Muñoz. Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast*, Vol. 24, pp. S26–S35, 2015.

[12] K. Oshima, Y. Naoi, K. Kishi, Y. Nakamura, T. Iwamoto, K. Shimazu, T. Nakayama, S. J. Kim, Y. Baba, Y. Tamaki, et al. Gene expression signature of tp53 but not its mutation status predicts response to sequential paclitaxel and 5-fu/epirubicin/cyclophosphamide in human breast cancer. *Cancer Letters*, Vol. 307, No. 2, pp. 149–157, 2011.

[13] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, Vol. 500, No. 7463, pp. 415–421, 2013.

[14] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, Vol. 499, No. 7457, pp. 214–218, 2013.

[15] Z. R. Chalmers, C. F. Connelly, D. Fabrizio, L. Gay, S. M. Ali, R. Ennis, A. Schrock, B. Campbell, A. Shlien, J. Chmielecki, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Medicine*, Vol. 9, No. 1, pp. 1–14, 2017.

[16] G. d. Cunha Santos, F. A. Shepherd, and M. S. Tsao. Egfr mutations and lung cancer. *Annual Review of Pathology: Mechanisms of Disease*, Vol. 6, pp. 49–69, 2011.

[17] A. Conlin, G. Smith, F. A. Carey, C. R. Wolf, and R. J. Steele. The prognostic significance of k-ras, p53, and apc mutations in colorectal carcinoma. *Gut*, Vol. 54, No. 9, pp. 1283–1286, 2005.

[18] M. Cieślik and A. M. Chinnaiyan. Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics*, Vol. 19, No. 2, pp. 93–109, 2018.

[19] M. P. Powers. The ever-changing world of gene fusions in cancer: a secondary gene fusion and progression, 2019.

[20] E. E. Heyer, I. W. Deveson, D. Wooi, C. I. Selinger, R. J. Lyons, V. M. Hayes, Sandra AO' Toole, M. L. Ballinger, D. Gill, D. M. Thomas ほか. Diagnosis of fusion genes using targeted rna sequencing. *Nature Communications*, Vol. 10, No. 1, pp. 1–12, 2019.

[21] J. A. Pardo, B. Fan, A. Mele, S. Serres, M. G. Valero, I. Emhoff, A. Alapati, and T. A. James. The role of oncotype dx® recurrence score in predicting axillary response after neoadjuvant chemotherapy in breast cancer. *Annals of Surgical Oncology*, Vol. 28, No. 3, pp. 1320–1325, 2021.

[22] E. Schaafsma, B. Zhang, M. Schaafsma, C.-Y. Tong, L. Zhang, and C. Cheng. Impact of oncotype dx testing on er+ breast cancer treatment and survival in the first decade of use. *Breast Cancer Research*, Vol. 23, No. 1, pp. 1–11, 2021.

[23] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, Vol. 8, No. 1, pp. 118–127, 2007.

[24] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, Vol. 21, No. 1, pp. 1–35, 2020.

[25] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, Vol. 40, No. 12, pp. 1413–1415, 2008.

[26] Y. Zhang, G. Parmigiani, and W. E. Johnson. Combat-seq: batch effect adjustment for rna-seq count data. *NAR genomics and bioinformatics*, Vol. 2, No. 3, p. lqaa078, 2020.

[27] Z. Huang, D. T. Jones, Y. Wu, P. Lichter, and M. Zapatka. conffuse: high-confidence fusion gene detection across tumor entities. *Frontiers in Genetics*, Vol. 8, p. 137, 2017.

[28] E. Olsson, C. Winter, A. George, et al. Serial monitoring of circulating tumor DNA in patients with primary breast cancer for detection of occult metastatic

disease. *EMBO Molecular Medicine*, Vol. 7, No. 8, pp. 1034–1047, 2015.

[29] C. Gambacorti-Passerini, L. Antolini, F.-X. Mahon, F. Guilhot, M. Deininger, C. Fava, A. Nagler, C. M. Della Casa, E. Morra, E. Abruzzese, et al. Multicenter independent assessment of outcomes in chronic myeloid leukemia patients treated with imatinib. *Journal of the National Cancer Institute*, Vol. 103, No. 7, pp. 553–561, 2011.

[30] R. Hehlmann, A. Hochhaus, M. Baccarani, et al. Chronic myeloid leukaemia. *The Lancet*, Vol. 370, No. 9584, pp. 342–350, 2007.

[31] J. S. Parker, M. Mullins, M. C. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, Z. Hu, J. F. Quackenbush, I. J. Stijleman, J. Palazzo, J. S. Marron, A. B. Nobel, E. Mardis, T. O. Nielsen, M. J. Ellis, C. M. Perou, and P. S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, Vol. 27, No. 8, pp. 1160–1167, Mar 2009.

[32] P. Roepman, H. M. Horlings, O. Krijgsman, M. Kok, J. M. B.-d. Mesquita, R. Bender, S. C. Linn, A. M. Glas, and M. J. v. d. Vijver. Microarray-based determination of estrogen receptor, progesterone receptor, and her2 receptor status in breast cancer. *Clin Cancer Res*, Vol. 15, No. 22, pp. 7003–11, 2009.

[33] L. J. v. t. Veer, H. Dai, M. J. v. d. Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. v. d. Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, Vol. 415, No. 6871, pp. 530–6, 2002.

[34] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. V. d. Vijver, J. Bergh, M. Piccart, and M. Delorenzi. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst*, Vol. 98, No. 4, pp. 262–72, 2006.

[35] Y. Sota, Y. Naoi, R. Tsunashima, N. Kagara, K. Shimazu, N. Maruyama, A. Shimomura, M. Shimoda, K. Kishi, Y. Baba, et al. Construction of novel immune-related signature for prediction of pathological complete response to neoadjuvant

chemotherapy in human breast cancer. *Annals of Oncology*, Vol. 25, No. 1, pp. 100–106, 2014.

[36] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, Vol. 7, No. 1, pp. 55–65.

[37] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, Vol. 4, No. 2, pp. 249–64, 2003.

[38] E. Hubbell, W. M. Liu, and R. Mei. Robust estimators for expression analysis. *Bioinformatics*, Vol. 18, No. 12, pp. 1585–92, 2002.

[39] M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, Vol. 11, No. 2, pp. 242–53, 2010.

[40] M. Kohl and H. P. Deigner. Preprocessing of gene expression data by optimally robust estimators. *BMC Bioinformatics*, Vol. 11, p. 583, 2010.

[41] H. Rieder, M. Kohl, and P. Ruckdeschel. The cost of not knowing the radius. *Statistical Methods and Applications*, Vol. 17, No. 1, pp. 13–40, 2008.

[42] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, Vol. 19, No. 2, pp. 185–93, 2003.

[43] C. H. Koh and L. Wong. Embracing noise to improve cross-batch prediction accuracy. *BMC Syst Biol*, Vol. 6 Suppl 2, p. S3, 2012.

[44] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker. Analysis of microarray data using z score transformation. *J Mol Diagn*, Vol. 5, No. 2, pp. 73–81, 2003.

[45] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou, and J. S. Marron. Adjustment of systematic microarray data biases. *Bioinformatics*, Vol. 20, No. 1, pp. 105–14, 2004.

[46] W. E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, Vol. 8, No. 1, pp. 118–27, 2007.

[47] K. A. Le Cao, F. Rohart, L. McHugh, O. Korn, and C. A. Wells. Yugene: a

simple approach to scale gene expression data derived from different platforms for integrated analyses. *Genomics*, Vol. 103, No. 4, pp. 239–51, 2014.

[48] 保母敏行, 飯田芳男, 石橋耀一, 岡本研作, 川瀬晃, 中村利廣, 中村洋, 平井昭司, 松田りえ子, 山崎慎一ほか. 日本分析化学会における標準物質の開発. 分析化学, Vol. 57, No. 6, pp. 363–392, 2008.

[49] K. H. Allison, M. E. H. Hammond, M. Dowsett, S. E. McKernin, L. A. Carey, P. L. Fitzgibbons, D. F. Hayes, S. R. Lakhani, M. Chavez-MacGregor, J. Perlmutter, et al. Estrogen and progesterone receptor testing in breast cancer: Asco/cap guideline update. 2020.

[50] A. C. Wolff, M. E. H. Hammond, K. H. Allison, B. E. Harvey, P. B. Mangu, J. M. Bartlett, M. Bilous, I. O. Ellis, P. Fitzgibbons, W. Hanna, et al. Human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline focused update. *Archives of Pathology & Laboratory Medicine*, Vol. 142, No. 11, pp. 1364–1382, 2018.

[51] A. Howell, J. Cuzick, M. Baum, A. Buzdar, M. Dowsett, J. F. Forbes, G. Hoctin-Boes, J. Houghton, G. Y. Locker, J. S. Tobias, and A. T. Group. Results of the atac (arimidex, tamoxifen, alone or in combination) trial after completion of 5 years' adjuvant treatment for breast cancer. *Lancet*, Vol. 365, No. 9453, pp. 60–2, 2005.

[52] M. E. Vergara-Lluri, N. A. Moatamed, E. Hong, and S. K. Apple. High concordance between herceptest immunohistochemistry and erbb2 fluorescence in situ hybridization before and after implementation of american society of clinical oncology/college of american pathology 2007 guidelines. *Modern Pathology*, Vol. 25, No. 10, pp. 1326–1332, 2012.

[53] P. C. Nowell, et al. Discovery of the philadelphia chromosome: a personal perspective. *The Journal of Clinical Investigation*, Vol. 117, No. 8, pp. 2033–2035, 2007.

[54] S. Faderl, M. Talpaz, Z. Estrov, and H. M. Kantarjian. Chronic myelogenous leukemia: biology and therapy. *Annals of Internal Medicine*, Vol. 131, No. 3, pp. 207–219, 1999.

[55] 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, p. 561, 2007.

[56] M. Eguchi-Ishimae, M. Tezuka, T. Kokeguchi, K. Nagai, K. Moritani, S. Yonezawa, H. Tauchi, K. Tokuda, Y. Ishida, E. Ishii, et al. Early detection of the pax3-foxo1 fusion gene in circulating tumor-derived dna in a case of alveolar rhabdomyosarcoma. *Genes, Chromosomes and Cancer*, Vol. 58, No. 8, pp. 521–529, 2019.

[57] F. Mertens, B. Johansson, T. Fioretos, and F. Mitelman. The emerging complexity of gene fusions in cancer. *Nature Reviews Cancer*, Vol. 15, No. 6, pp. 371–381, 2015.

[58] S. Kumar, A. D. Vo, F. Qin, and H. Li. Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data. *Scientific Reports*, Vol. 6, pp. 1–10, 2016.

[59] B. Haas, A. Dobin, N. Stransky, et al. STAR-Fusion: fast and accurate fusion transcript detection from RNA-Seq. *BioRxiv*, 2017. doi:10.1101/120295.

[60] M. Miyamoto, D. Motooka, K. Gotoh, et al. Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics*, Vol. 15, No. 699, pp. 1–8, 2014.

[61] K. Suzuki, D. Ueta, S. Seno, et al. A Method of Sequence Analysis for High-throughput Sequencer Data Based on Shifted Short Read Clustering. In *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)*, pp. 380–386, 2013.

[62] K. Shimizu and K. Tsuda. Slidesort: all pairs similarity search for short reads. *Bioinformatics*, Vol. 27, No. 4, pp. 464–470, 2010.

[63] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, Vol. 30, No. 15, pp. 2114–2120, 2014.

[64] H. Edgren, A. Murumagi, S. Kangaspeska, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biology*, Vol. 12, No. 1, pp. 1–13, 2011.

[65] S. Kangaspeska, S. Hultsch, H. Edgren, D. Nicorici, A. Murumägi, and O. Kallioniemi. Reanalysis of RNA-sequencing data reveals several additional fusion genes with multiple isoforms. *PloS One*, Vol. 7, No. 10, 2012.

[66] F. Collins, E. Lander, J. Rogers, R. Waterston, and I. Conso. Finishing the euchromatic sequence of the human genome. *Nature*, Vol. 431, No. 7011, pp. 931–945, 2004.

謝辞

本研究を遂行し学位論文を作成するにあたり、格別なるご指導とご高配を賜りました大阪大学大学院情報科学研究科 松田秀雄教授に謹んで深謝の意を表します。

本論文の審査において、貴重なご指導とご助言を賜りました大阪大学大学院情報科学研究科 松田史生教授、戸谷吉博准教授、小蔵正輝准教授に謹んで深謝の意を表します。

本研究の進捗や解析手法・論文作成などについてご指導およびご助言を賜りました大阪大学大学院情報科学研究科 瀬尾茂人准教授に謹んで深謝の意を表します。

本研究を進めるにあたり、融合遺伝子解析の解析コードのプログラミングについてご助言を頂いた大阪大学大学院情報科学研究科 繁田浩功助教に深く感謝申し上げます。

本研究を進めるにあたり、乳癌細胞株のデータの提供を賜りました大阪大学大学院乳腺内分泌外科 野口眞三郎教授 (現兵庫県立西宮病院院長)、下田雅史講師に深く感謝申し上げます。