シニア・リサーチフェロー

研究成果報告書

平成　28 年　4 月　28 日 提出

公益財団法人　がん研究振興財団
理　事　長　髙　山　昭　三　殿

報告者氏名：　足立　美保子　　印

研究課題：　がん全ゲノム・エピゲノムデータ解析パイプラインの開発と
（テーマ）　臨床を指向した新たな発がん分子機構解明への応用

研究期間：　自　平成　27 年　8 月　1 日

至　平成　28 年　3 月　31 日

研究指導者：氏名　　柴田　龍弘　　印

公益財団法人　がん研究振興財団

(1)シニア・リサーチフェロー期間中の研究について

１．要旨

研究の初年度にあたる27年度では、Genome Rearrangement (GR)検出プロセスの全自動パイプライン化と、BWA-mem移行に伴うバージョンアップを行い、また将来予定している公開頒布に備え、各ツールのパッケージ化を行った。胆道がん全ゲノム解読データを用いたGR検出では、9000箇所以上のGRを検出し、これらのGRの分布が遺伝子領域に強く偏っていることを見出した。遺伝子領域でのGRの高度な集積がみとめられる一方で、GRが生じた遺伝子での有意な発現量変化がみとめられない場合が全体の９割近くを占めていた。がんゲノム全解読データからウィルスゲノムとヒトゲノムとの接合点を検出するツールを開発し、肝がんWGSデータを用いてhepatitis B virusのゲノム挿入部位の同定を行った。今年度の研究成果を元に、GRにより移動するゲノム配列についてのアノテーションを付加するプロセスをパイプラインに追加する等の改良を行う。またGRで働いたと考えられるTransposon(TE)の同定を試みる他、既存のTE検出ツール（TraFiCなど）を実装し、がんゲノム内のTEを検出するパイプラインの構築も目指す。

略称：

GR: Genome Rearrangement

PE: Paired End

WGS: Whole Genome Sequencing

２．序

研究の背景並びに目的

この半世紀ほどの目覚しい分子生物学の発展により、多くのがんにおいて発がんにいたるメカニズムが解明され、分子標的薬など新たな作用機序からなる抗がん剤が多数開発されてきた。これにより個々の病態に対応した個別化治療が徐々に可能になり、かつてに比べ治療成績も向上している。しかしながら、ひとたびがんの脅威を脱したのちも、再発・転移のリスクは患者さんのその後の生活につきまとい続ける。抗がん剤の攻撃を逃れて僅かでも体内に残留したがん細胞は、巧みにその性質を変え、抗がん剤の標的から逃れて増殖を始めるかもしれないし、原発組織のくびきを逃れ、全く異なる器官で異なる性状のがん組織を形成しているかもしれない。

がんが完治困難である大きな理由の一つは、その多様な変化能にあるといえる。非常に早いペースで新しい性質をもつがん細胞が生じうるため、抗がん剤との鼬ごっことなることも珍しくない。さらに、がん組織では一般に免疫機能が抑制されるため、免疫系による異常な細胞の排除が進まず、新たな異常細胞の出現がさらに助長されていると考えられる。このような変幻自在で強靭ながん細胞の変化を捉え、その情報を元に再発・転移リスクを予測し、事前に対処することが出来れば、がん罹患者の精神面におけるQOLの大きな向上が見込まれる。

このような変化能の背景にあるのは、がんゲノムに見られる多様で多層的なゲノム異常の高度な蓄積とエピゲノム状態の変化であると考えられているが、つい最近まで個々のがんにおいて全ゲノムレベルでの異常を包括して解析することは困難であった。近年の高速シークエンス技術革新により、がんゲノム全体の解読が可能になり、理論的には包括的ゲノム異常が同定可能となったが、全ゲノム配列データの扱いの難しさやがん領域における情報解析専門家の不足から、この分野の研究が十分に進んでいるとはいいがたい状況である。

そこで、本研究計画では３年間の実施期間を設定し、効率的な情報解析手法の開発とその臨床応用を主眼としながら、以下の目的を達成することを目指す。

１・がんゲノムに生じる変異を全て検出対象とする自動解析パイプラインの構築

２・がんゲノム変異データとエピゲノムデータとの統合

３・がんゲノム異常に対応する免疫反応（Tcellレパトアプロファイル）の探索

研究計画の初年度にあたる27年度では、特に解明が進んでいない Genome Rearrangement変異（以下、GRと略）について注力し、検出パイプラインの構築と、更に得られたGRの２次解析を行った。

３．方法

3-1. GR検出パイプライン構築とパッケージ化

所属研究分野の十時泰ユニット長により開発されたGR検出パイプライン（Totoki et al. Nat Genet 2011, Totoki et al. Genome Res 2014）を、全自動化、高速化し、大規模化するWGS解析への実用性を高めた。このとき、並列計算による効率化を行うことで、計算時間の短縮・メモリの節約を実現した。また、公開・頒布の準備として、本パイプラインのパッケージ化も実施した。

本パイプラインが使用するマッピンングソフトはBWA-alnとBWA-memの両方に対応しているが、より精度のよい最新のBWA-mem版についてここでは述べる。またシークエンスデータはIlluminaのpaired readシークエンスデータを想定しているが、他のシークエンサーでも多少修正を加えれば対応可能だと考えている。

本パイプラインでは、入力として

1）rearrangementを検出するサンプル(腫瘍部サンプル)のpaired read（以下PEと略）マッピング結果ファイル（samフォーマット）

2）1)のコントロールとして使用するサンプル(非腫瘍部サンプル)のPEマッピング結果ファイル（samフォーマット）

3）PEリードデータ 1）の最大insert length

4）PEリードデータ 2）の最大insert length

を与えることにより、GR検出プロセスを開始する。GR検出パイプラインの主要なプロセスは以下の４ステップにより構成される。

【ステップ1. PEリードの分類】

GRが生じたゲノム領域に由来するPEリードは、リファレンスゲノムにimproperな形式（PEリード間の距離または方向が通常と異なる）でマッピングされると考えられる（図１）。本パイプラインにおいては、

1）PEリードの両エンドの配列決定方向の組合せ
（properな組合せでは、Forward-Reverseの対となる）

2）PEリードの両エンド間の距離

3）PEリードがマップされる染色体番号

を指標としてGR情報（improperなPEマッピング情報）を含むPEリードを入力データから抽出し、PEリード群を、Forward-Forward [:inversion type] / Reverse-Reverse[:inversion type] / Reverse-Forward[:tandem-repeat type] / Forward-Reverse[:deletion type]（インサートサイズが入力時に指定した閾値を越えるもの） /

Translocation の５群に分類した。

【ステップ２．PEリードの選択】

各分類について、下記の基準で解析対象とするPEリードを限定し、ミスマッピングによる擬陽性GRの検出を未然に防いだ。

1)　リファレンスゲノムの複数箇所にマッピングされるPEリードの削除

2)　マッピングクオリティスコアが10未満のリードの削除

3)　ミスマッチが4個以上のリードの削除(リード長が100bpの場合)

4)　PCR duplicationによるリード(リファレンスゲノムの同じ場所にマッピングされるPEリード)の削除

【ステップ３．PEリードのクラスタリング】

前ステップで分類、選択されたPEリード群のそれぞれについて、下記の手順でクラスタリングを行った。

1)　PEの構成リードであるForwardリードとReverseリードに分けて、それぞれについて２リード間の距離が400bp以下であれば同一のクラスターとして、Forwadクラスター群とReverseクラスター群を作成する。

2)　Forwadクラスター群とReverseクラスター群間で対のPEリードだけを選び、対のPEリードだけを含むPairedクラスター(ペアのForwad, Reverseクラスター)を作成する。このペアのForwad、Reverseクラスター間領域に、GRで生じるBreakpointが存在すると考えられる。

3)　4個以上のPEリードを含むPairedクラスターをGRの候補とする。

【ステップ４．擬陽性GRのフィルタリング】

リファレンス配列は完全ではないので、リファレンス配列の不正確性からPEリードのミスアライメントが起こり、偽陽性のGRが検出される場合が多くある。同じ形式のミスアライメントはコントロールサンプル（非腫瘍部サンプル）でも起こるので、コントロールサンプルでも同様にGR検出を行い、コントロールサンプルでも検出されるGRを取り除くと、多くの偽陽性GRが取り除くことが可能であり、コントロールサンプル数が多いほど効果が大きい。この原理に従って、偽陽性GRを少なくするために、88症例分の胆道がん非腫瘍部サンプルのマッピング結果とオーバーラップするPEリードを含むPairedクラスターを偽陽性として取り除いた。これによって偽陽性GRが大幅に取り除かれたと思われる。

以上のプロセスを経て、GRは一対のPairedクラスターに挟まれた領域として出力される。

BWA-memへの対応に伴うバージョンアップ

2015年11月には、マッピングソフトをBWA-alnからBWA-memに変更したことに伴い、バージョンアップを行った。BWA-memはICGCにおける標準解析プロトコールで採用されたマッピングソフトでもあり、現在世界中で最も使用されている高精度マッピングソフトである。それまで使用されていたBWA-alnと比べ、soft-clipping readのマッピング（全長アライメントが出来ないリードのローカルアライメント）の性能が大きく改善した。このことにより、我々の開発したGR検出パイプラインにおいても、PEリードに挟まれた領域だけではなく、リードの途中にBreakpointが存在する（Breakpointそのものを配列決定した）リードもデータとして使用可能になった。また、ミスアライメントによりimproperly pairとされるPEが少なくなり、正確性が向上し、さらなる計算時間、要求メモリ量の節約も実現した。

3-2. エピゲノムデータ（遺伝子発現量）を利用したGR２次解析

ゲノムに生じたGRが遺伝子発現に与える影響を調べる目的で、以下の解析を行った。遺伝子発現のデータとして、RNAseqデータ（RPKM値）を使用した。これは、所属研究分野の濱奈津子研究員から提供いただいたもので、WGSデータを行った胆管がんサンプルから調整したRNAを解析したものである。今回使用した88サンプルのWGSデータのうち、73サンプルで対応するRNAseqデータが解析ずみであった。

RefFlat 常染色体＋性染色体上の39842個の全遺伝子を対象として、遺伝子及びその周辺領域に生じたBreak Pointを同定したのち、対象領域にBreak Point挿入があったサンプル群/それ以外のサンプル（=background）群の２群に分け、２群間の遺伝子発現量の差について、Welchのt検定により検討した。

3-3. Virus Integration 検出ツールの開発

WGSデータからVirus Integrationを検出するツールを開発した。このツールは、WGSデータからヒトゲノムに全長でマッチする配列がなく、ウィルスゲノムの配列と部分一致するリードを抽出し、これらのリードに含まれる情報をもとに、Virus Integrationが生じたヒトゲノム上の領域、及び挿入された配列を出力するものである。

以下、本ツールにおけるデータの流れについて記述する。

【ステップ１．Virus Integration 情報を含むリードの抽出】

はじめに、全WGSデータから ウィルス様配列を含むリードを抽出した。各サンプルのWGSデータは bowtie2 によってヒトゲノム（HG19）にマッピングされ、ここでヒトゲノムに完全一致する（リード全長がミスマッチを含まずにアライメントされた）リードを対象から除外した。

つぎに、hg19にウィルスゲノムの全長（ここではHBVゲノム：3215b p を使用）を追加したレファレンスゲノム上へのアライメントをblastnにより行い、リード長の25%以上の長さが、95%以上の一致率でHBVゲノム上にアライメントされたリードを以後の解析対象とした。

【ステップ２．対象リードのカテゴライズ】

解析対象としたPEリードを、ヒトゲノムと ウィルスゲノムとの接合箇所により２グループに分類した。グループのひとつはエンドリードの途中に接合箇所を含む、すなわち接合箇所そのものが配列決定されているパターンであり、もうひとつはPEリードに挟まれる領域（インサート）に接合箇所が存在するパターンである。前者については、各接合箇所について、

・リード長の90%がヒトゲノムまたはHBVゲノム上に95%以上の一致度でアライメント

・上のアライメントにおいて、ミスマッチ数・ギャップ数はそれぞれ２以下

・接合部位におけるオーバーラップ・またはギャップは10bp未満

の３条件全てを満たす対象リードが２本以上存在したとき、HB Virus Integration 接合の候補とした。

後者については、PEのorientation（シークエンスの方向）、およびinsert size（PEリード間の距離）によってFF/RR/RF/FR_n（InsertSize>550を満たすFR）の４カテゴリに分類され、カテゴリごとにPEクラスタリングを行った。リードカテゴリ、およびPEクラスタリング法の詳細については、方法項1のGR検出パイプラインに準ずるが、ウィルスゲノムに由来するリード同士のオーバーラップについては、HBVゲノムのサイズを考慮して±50bpまでを許容した。

また、上記【ステップ1】において検出されたウィルス配列様PEリードの中には、improperな様式でウィルスゲノムにマッピングされるものが存在する。このようなPEリードに、HBVゲノム内部で生じるGRを反映していると考えられるため、【ステップ２】の２カテゴリに分類し、virus integration と同様のクラスタリングを行ってGRを同定した。

４．結果並びに考察

4-1. GR解析パイプライン構築と胆管がんWGSデータからのGR検出

PEリードによって配列決定されたWGSデータの入力から、GRを検出する自動パイプラインを構築した。GR検出までの所要時間は、標準的なファイルサイズ（500GB, 750Mb PEread）で約9時間、最大要求メモリサイズは約5GBであった（方法項1-ステップ1で抽出されるimproper readの量・構成に依存して計算量は変化する）。本パイプラインは、所属研究室での胆管がんGR検出を担当したほか、所属研究所内の別研究室や東京大学医科学研究所との共同研究でも利用され、GR解析結果を提供した。

4-1-1. GR HotSpot

88対の胆管がんDNAと同一提供者の正常組織DNAを用いたWGSデータを使用してGR検出を行った。検出結果の概観を図2a, 図2bに示す。最も多くGRが検出されたBD244では634箇所のGRが検出される一方、最も検出数が少ないBD192では4箇所の検出にとどまるなど、サンプルによりGRの検出数は大きく異なった。一方で、ゲノム上のごく狭い領域で同タイプのGRが異なるサンプル間で共通して発生しているケースも観察された。そこで、88サンプルから検出された総計9764箇所のGR Break Point（GRによって生じるゲノム断裂点を近似した座標)をゲノム座標上にプロットし、GRが高頻度に生じるゲノム領域（=GR Hot Spot）について解析した結果を図2cに示す。最も顕著にGRが発生した領域は22番染色体の30M付近であり、ここでは1Mbほどの領域に25サンプル79個のGRに由来する84箇所のBreak pointの集積が見られた。

4-1-2. Transposable Element

22番染色体の30Mb近傍の遺伝子、および発生したGRのタイプを調べた。79箇所のGRのうち、59箇所まではTranslocationタイプのGRであり、図3aで示した一般的なGRタイプごとの発生率に対してTranslocationタイプに強く偏る傾向が見られた。Translocationタイプの GRにより、22番染色体30Mb周辺の領域が転移する先について、Circos Plotで表示したものを図2dに示した。小さな領域がゲノムのさまざまな部位に繰り返しTranslocationする様子から、この領域には活性化 Transposable Element が含まれると考えられる。図3cにおいて、Break Point頻度のピークが見られるほかの領域についても同様の観察をおこなった結果、Chr22以外にも活性化TEを含む可能性が高い領域を複数同定した。現在、これらのHotSpot領域に含まれるTransposable Element の同定、およびTransposable Elementが新たに挿入する先のゲノム領域で生じるエピジェネティックな変化について解析を進めている。

4-1-3. 遺伝子ごとに見たBreak Pointの集積

88例の胆管がんサンプルから検出されたGRは9764個であった。もしこれらのGRがランダムに発生するならば、ゲノム上におけるGR発生箇所の分布は、9764*2/30億塩基対から、６Mbにつき１回という比較的疎なものとなるはずである。しかしながら、図2cで示すHotSpotに相当するゲノム領域を詳しく調査した結果、遺伝子とその周辺領域で高度にBreak Pointが集積するケースが多く観察された。

そこで、RefSeqデータベース記載の全ての遺伝子について、このようなBreak Pointの集積の有無を網羅的に調査した。遺伝子とその周辺の領域を、"遺伝子本体（gene body）", "遺伝子近傍の上下流2000b(promoter領域)", "さらにその上下流100Kb"の３区分にわけ、各区分内に生じたBreak Pointの個数の集計をとった。refseqに記載のある全38942遺伝子のうち、上の３区分のいずれかに１以上のBreak Pointを含むものは8713遺伝子存在した。表2では、対象となった8713遺伝子のうち、高度にBreak Pointが発生した上位24遺伝子について示している。

これらの中には、FHIT、RAD51B、FGFR2のがん遺伝子のほか、FGFR2と融合遺伝子を形成するBICC1が含まれており、それぞれ33、25、19、12のBreak Pointが遺伝子とその近傍領域に生じていた（図3、左側グラフ）。図3aに示したFHIT遺伝子では、遺伝子全体に渡り、イントロン部位でBreak Pointが高頻度に生じていた。また、5'端上流に４つのBreak Pointが位置する一方で、3'端下流では対象的に250KB余りものBreak Pointの無い領域が続いていた。RAD51B遺伝子においては、遺伝子上流・下流ともBreak Pointは疎であり、遺伝子内部のイントロン領域にBreak Pointの集積が見られた。FGFRでは、最終イントロンに14のBreak Pointが集中しており、特に10番染色体123,241Mb-123,242Mbの1000bpの区間には６サンプルに由来する14のBreak Pointが存在していた。

胆管がんにおいては、FGFR2の最終エキソンが脱落して別遺伝子のエキソンが接続したFGFR融合遺伝子のトランスクリプトが既に報告されており（Nakamura.et.al., Nat Gen 2015）、本解析で得られた結果は、このような融合遺伝子がゲノム本体に生じたrearrangementにより生じることを明確に裏付けるものである。図3dにはFGFR2の融合相手の一つとされるBICC1でのBreak Point分布をしめした。FGFR2とBICC1の融合遺伝子に由来するトランスクリプトームには、BICC1の10aa程のC'端ドメインの一部がFGFRに結合したType1と、700aa超の長いドメインが結合するType2の2種が報告されているが、前者が3番イントロン・4番イントロン間に位置するBreak Pointに、後者が16番イントロン位置するBreak Pointに対応するものと考えられる。

4-1-4. 遺伝子領域に強く偏ったGR分布

9764箇所のGRのうち、少なくとも片方のBreak Pointが遺伝子内部で生じているものは6037箇所、遺伝子とその上下流100Kbまでの領域で生じたものは8560箇所であった。ゲノム全体のうち、遺伝子領域（機能性分子をコードする配列，イントロン含む）の割合はたかだか3割程度であることから、胆管がんゲノムで検出されたGRは遺伝子とその周辺領域に強く偏って存在するといえる。GRが生じる機構については未だ解明されない部分が多いが、明確な配列依存性は認められておらず、ゲノム中のランダムな部位で生じると考えられている。おそらく、GRの発生自体はランダムに起こるが、その後の選択淘汰によって生存に有利な影響を及ぼすGRが発生した細胞が生き残った結果、遺伝子領域へのGRの集積が進んだ結果と推測される。

4-2. エピゲノムデータとの統合解析の試み

GRはゲノムの数百bp～数千bp以上の広範な領域が挿入・欠失・逆位・転座される大規模な変異であるため、遺伝子内部や遺伝子周辺で生じたGRは、遺伝子の機能及び発現量に大きな影響を与えると考えられた。例えば1-2-3であげたFGFR2-BICC1は、GRが新たな機能をもつ遺伝子を生み出す好例である。また、遺伝子領域へのBreak Pointの集積は、GRが遺伝子機能を細胞の生存に有利な方向へ変化させた可能性を示唆するものである。そこで、GR検出に用いた胆管がん88サンプルのうち、RNAseqによる遺伝子発現解析を行った73サンプルを対象に、GRによるBreak Point挿入が遺伝子発現におよぼす影響を評価した。すなわち、遺伝子とその上下流100Kbまでの区間でBreak Pointが検出されたサンプルをBP群、それ以外のサンプルをbackground群と定義し、遺伝子ごとにBP群とbg群での発現量（RPKM値）分布の差を検定した。この検定は２つ以上のサンプルで対象領域からBreak Pointが検出された3894遺伝子すべてについて行った。遺伝子内部から検出されたBreak Point数の上位24位までの遺伝子について、BP群とbg群の平均値（分散）、平均値同士の比、WelchのT検定を行った結果を表2に示した。

予想に反して、RAD51Bを含む上位8個の遺伝子で遺伝子発現量に有意な変化はみられなかった。上位24遺伝子のうち、有意に発現量が変化したものは6遺伝子で、その中の4遺伝子までがBP群で発現量が上がっていた。この傾向は検定対象全体に共通しており、3894遺伝子のうちの約９割を占める3433遺伝子でBP群とbg群の間に有意な発現量変化はみられなかった。有意な変化があった遺伝子群では、BP群で発現量が上昇したものが269遺伝子、下

がったものが192遺伝子と、GRが発現上昇に働く遺伝子がやや多い結果となった。FGFR2、BICC1ではBP群で発現量は共に上昇していた（図3c、図3d）。

遺伝子領域に発生したGRが発現量に影響を与えないのならば、これらのGRはどのような機構を介して細胞の生存に有利に働いたのだろうか？一つの仮説として、コーディング領域そのものでなく、発現調節に関わるUTRやプロモーター・エンハンサー等の転写因子結合領域がGRにより別の遺伝子近傍に移動し、その遺伝子発現を調節した可能性が考えられる。このような例は既にいくつか報告があり、たとえば髄芽腫サブタイプⅢ、Ⅳではこれまでドライバー遺伝子が不明であったが、GRによりDDX遺伝子のエンハンサー領域とGFII遺伝子との距離が短縮し、GFIIの転写活性が顕著に上昇することでがん化が進むことが明らかにされた［Northcott, Nature2014］。胆管がんサンプルの中にもがん化に至る分子機構が不明なサンプルが残されているが、GRで再構築されたゲノム領域の機能を詳しく調べることで、新たなドライバー遺伝子を同定できる可能性も考えられる。

4-3. 肝がんWGSデータからのゲノム挿入（Virus genome Integration）の検出

Virus genome integrationとは細胞に感染したウィルスゲノムが宿主ゲノムに自らのゲノム断片を挿入させる現象であり、がんを引き起こすウィルスではMerkel cell polyoma virus（メルケル細胞がん），Human papillomavirus（乳頭腫），hepatitis B virus（肝臓がん)で、ヒトゲノム内への挿入が確認されている。そこで、がんゲノム全解読データからウィルス（外来）ゲノムとヒトゲノムとの接合点を検出するツールを開発し、肝がんサンプルのWGSデータを用いてhepatitis B virus(HBV)のゲノム挿入部位の同定を試みた。

ウィルスゲノムでは塩基の変異速度が大きいため、リードのアライメントに際してミスマッチを許容し、ローカルな（短い）アライメントに長じたツールが望ましい。そこで、本解析パイプラインではアライメントツールにbowtie2とblastnを併用し、bowtie2で入力の大半を占めるヒトゲノムに完全一致するPEリードを同定して解析対象から除外したのち、blastnで部分一致も含めたHBVゲノムとのアライメントを行った。エンドリードの部分的な類似配列をデータとして採用した結果、エンドリードの途中にヒトゲノムとウィルスゲノムの接合点が存在するケース（図4：紫線）でも検出が可能となった。

本パイプラインを表３で示す肝臓がん６サンプルに適用した。入力にはリード長50～125bpのPEリードによるWGSデータを使用した。ゲノム挿入検出パイプラインを適用した結果、HX14, HX19, HX25, HX28, HX33, HX35から、それぞれ2, 0, 13, 4, 3, 7の計29ゲノム挿入部位が検出された。HX19サンプルではHX33,HX35とほぼ同規模の入力データ（約80Gb）

を使用したにも関わらず、HBV類似配列を持つPEリードそのものがほぼ検出されなかった。

6つの肝がんWGSサンプルから得られた29箇所のゲノム挿入部位のうち、7箇所まではTERT上流にあたる15番染色体1.29Mb付近の狭い領域に集中していた（図4）。HBV陽性肝がんの発がん機序のひとつとして、HBVプロモーターのTERT上流部への挿入が引き起こすTERTの発現量の上昇が既に報告されており、本検出パイプラインにおいてもこれを裏付ける結果を得ることができた。

(2) シニア・リサーチフェロー期間中の研究成果を、今後の研究にどのように役立てたいと考えているか

今後の研究方針

研究の初年度にあたる27年度では、GR検出プロセスの全自動パイプライン化と、BWA-mem移行に伴うバージョンアップを行った。また、将来予定している公開頒布に備え、各ツールのパッケージ化を行った。胆道がん88症例全ゲノム解読データを用いたGR検出では、計9762箇所のGRを検出し、これらのGRの分布が遺伝子領域に強く偏っていることを見出した。遺伝子領域でのGRの高度な集積がみとめられる一方で、GRが生じた遺伝子での有意な発現量変化がみとめられないケースが全体の９割近くを占めていた。このような遺伝子では、コーディング領域そのものではなく、その上下流に位置するUTRや転写調節領域の移動先が重要な意味をもつのかもしれない。

そこで次年度では、今年度の研究成果を元に、GRにより移動するゲノム配列についてのアノテーションを付加するプロセスをパイプラインに追加する。すなわち、① GRで移動が起こった配列に含まれる機能性配列（転写因子の結合部位など），② 移動先の領域近傍にある遺伝子、移動により相対距離が変わった遺伝子，についての情報を付加するようにツールの改良を行う。同時に、遺伝子の発現量を決定するGR以外の要素である、コピー数多形や塩基変異などのゲノム変異データ、メチル化状態やヒストン修飾などのエピゲノムデータを統合し、その上で遺伝子発現との関係についても解析を進める。

胆管がんWGSデータからはTransposable Element(TE)の活性化と考えられる、多くのtranslocationタイプのGRが検出されていた。これは、正常ゲノムで機能している内在性レトロウィルス・レトロトランスポゾンの転写抑制機構が、胆管がんゲノムにおいて破綻していることを示唆する。これらのTE様GRについては、転位された配列を推測し、GRで働いたと考えられるTEの同定をこころみるほか、既存のTE検出ツール（TraFiCなど）を実装し、適宜パラメーター・フィルターに調整を加えて、がんゲノムに含まれる全てのTEを検出対象とするパイプラインの構築も目指す。

３ヵ年計画の最終年度までには、ゲノムにおいてTEの活性化やGRが亢進された結果生じるトランスクリプトームの変化や、このような細胞に対する固有の免疫反応（Tcell レパトアプロファイル）の検討を予定しているため、次年度内には更にRNAseq解析パイプラインの実装やTcell レパトアプロファイル解析の準備にも着手する。

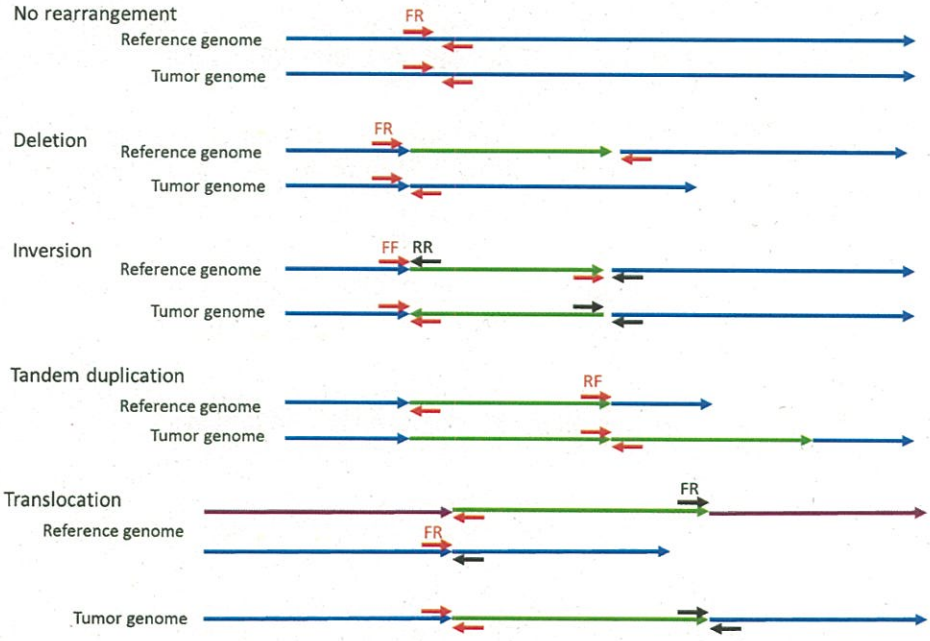
シニア・リサーチフェロー研究課題の目指すものと今後の発展

報告書の序に記述したとおり本研究計画の最終ゴールは、がん細胞の高度な変化能を捉え

ることである。ひとつのゴールは変化能の背景にあるがんゲノムでの高度なゲノム異常蓄積と、全長にわたるエピゲノム異常から、性質変化が生じる（細胞のがん化が進む）機序を解明することであり、ゴールのもうひとつは、活性化TEに由来するがん細胞固有のトランスクリプトーム変化や、がんの性質を反映する免疫細胞レパトアプロファイルの同定を通じて、診断・治療に役立つ新たなバイオマーカーを創出することである。

どちらのゴールも達成に至る道のりは非常に長いものであるが、本研究計画で構築する自動WGS解析パイプラインを介して研究交流を広げ、これらの目標実現に貢献を果たしたい。

a) 4つのGenome rearrangement タイプと、対応するPaired End read のマップ様式

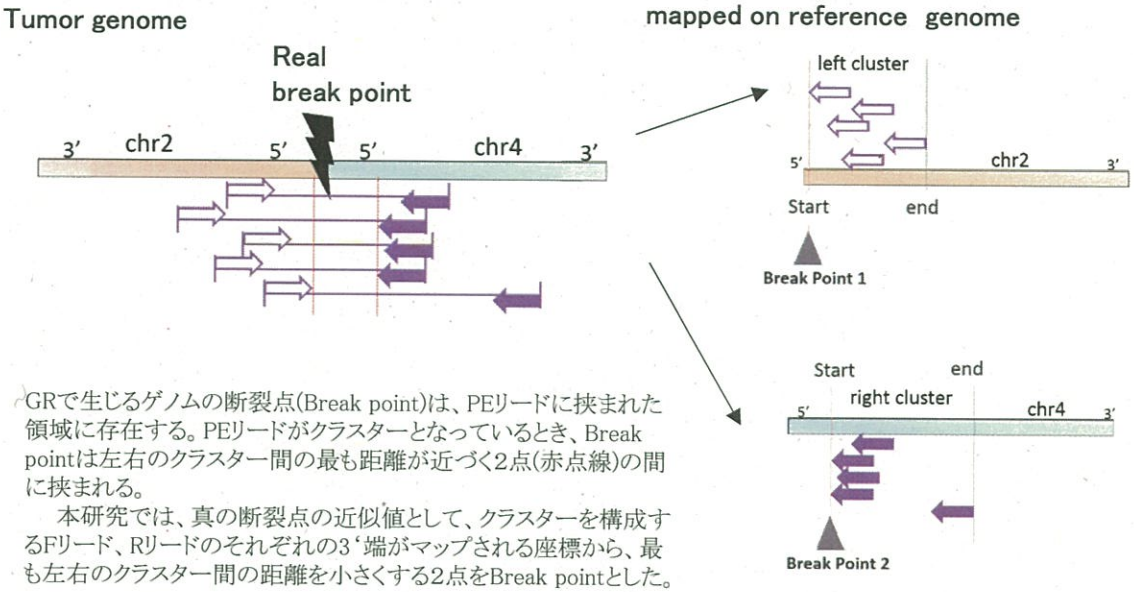


十時泰ユニットリーダーより許可を得て掲載

b) Genome rearrangement の出力例とその図解

chr2とchr4で生じた Reverse-Reverse translocation typeで例示：

PE_direction	ID	rearrangement type	No. read	left cluster					right cluster				
				chr	start	end	size	gene	chr	start	end	size	gene
RR	EX01	translocation	5	chr2	100,000,000	100,000,200	200	geneX	chr4	10,000,500	10,000,800	300	geneY



GRで生じるゲノムの断裂点(Break point)は、PEリードに挟まれた領域に存在する。PEリードがクラスターとなっているとき、Break pointは左右のクラスター間の最も距離が近づく2点(赤点線)の間に挟まれる。

本研究では、真の断裂点の近似値として、クラスターを構成するFリード、Rリードのそれぞれの3‘端がマップされる座標から、最も左右のクラスター間の距離を小さくする2点をBreak pointとした。

図1. Genome rearrangement 検出

ID	No. Input read*		No.GR	ID	No. Input read*		No.GR	ID	No. Input read*		No.GR	ID	No. Input read*		No.GR
	Normal	Tumor			Normal	Tumor			Normal	Tumor			Normal	Tumor	
BD012	770	1,627	499	BD078	878	1,223	33	BD148	922	1,122	77	BD209	697	1,359	52
BD015	788	1,194	107	BD079	872	1,058	36	BD149	761	1,344	54	BD214	876	1,121	106
BD019	815	1,123	193	BD081n	612	1,576	99	BD150	661	1,283	66	BD215	849	1,207	71
BD021	1,103	1,737	24	BD082	1,215	1,644	92	BD158	745	1,154	221	BD216	913	1,228	49
BD023	848	1,138	112	BD088	715	1,338	112	BD159	646	1,257	45	BD217	667	1,331	409
BD026	699	1,354	45	BD092	899	1,172	56	BD162	914	1,133	56	BD218	674	1,345	36
BD029	855	1,178	111	BD095	824	1,128	10	BD163	695	1,417	257	BD219	693	1,375	64
BD031	742	2,257	46	BD101	688	1,193	39	BD166	1,057	1,191	20	BD220	964	2,319	119
BD032	658	1,180	17	BD104	922	2,016	53	BD167	999	1,616	21	BD221	892	1,422	179
BD033	1,201	1,988	33	BD105	716	1,418	85	BD168	715	1,483	43	BD222	668	1,390	39
BD037	816	1,127	135	BD109	832	1,128	32	BD169	1,111	2,275	58	BD226	824	1,121	33
BD045	646	1,302	34	BD112	797	1,132	49	BD173	957	1,220	114	BD227	899	1,150	355
BD047	776	1,180	22	BD115	1,019	1,520	86	BD175	959	1,204	314	BD228	701	1,354	121
BD048	770	1,456	99	BD117	658	1,360	65	BD179	867	1,203	184	BD231	897	1,119	544
BD049	868	1,141	27	BD123	873	1,463	128	BD180	932	1,139	190	BD233	984	1,101	238
BD053	686	1,352	84	BD129	1,026	1,541	66	BD185	897	1,202	38	BD236	902	1,180	43
BD054	784	1,569	197	BD132	1,125	1,440	261	BD187	874	1,112	52	BD237	970	1,120	56
BD069	924	1,120	42	BD134	1,107	1,982	120	BD189	788	1,133	425	BD238	884	1,639	238
BD070	672	1,336	124	BD137	849	1,597	41	BD192	798	1,430	4	BD241	970	1,557	80
BD071	987	1,208	24	BD141	1,104	1,548	27	BD197	855	1,184	141	BD242	835	1,470	22
BD074	800	1,169	85	BD142	1,037	1,192	44	BD199	872	1,196	70	BD244	921	1,168	644
BD075	746	1,420	41	BD145	675	1,382	42	BD200	871	1,214	114	BD245	914	1,332	25

＊単位:Mb

表1. Genome rearrangement検出に使用された胆管がんWGSデータ

gene	No. break point			mean(sigma)		change rate	p-value
	gene body	~ 2K	~ 100K	BP group	background		
geneA	76	0	1	1.3547 (2.295450)	1.3864 (1.591019)	0.98	0.56895
geneB	50	0	1	3.1741 (4.485309)	3.1930 (5.259642)	0.99	0.63190
FHIT	31	0	2	2.5195 (2.687073)	2.8742 (2.122159)	0.88	0.24745
geneC	26	0	4	1.7114 (1.102291)	1.8971 (2.508340)	0.90	0.60668
geneD	24	0	2	5.8257 (4.438788)	9.1859 (4.915032)	0.63	0.06404
RAD51B	23	0	2	1.1069 (0.641184)	1.5640 (1.052323)	0.71	0.39085
geneE	22	0	2	1.7884 (1.574089)	4.6494 (5.376861)	0.38	0.11466
geneF	20	0	1	2.4642 (4.158784)	2.7523 (5.311254)	0.90	0.80619
geneG	20	0	3	4.7830 (0.141746)	7.3477 (3.996129)	0.65	3.80E-04
geneH	19	0	2	1.2358 (1.885463)	0.7505 (0.869811)	1.65	0.92789
geneI	19	0	2	2.4469 (2.750667)	3.3446 (5.099510)	0.73	0.85885
FGFR2	16	1	2	44.8639 (44.292384)	15.5798 (18.368378)	2.88	3.83E-04
geneJ	16	0	6	2.3703 (1.214200)	3.2212 (2.422544)	0.74	0.61802
geneL	16	0	0	1.4836 (2.239029)	1.2551 (1.764526)	1.18	0.10687
geneM	15	0	2	25.8184 (12.269925)	44.9952 (49.744372)	0.57	0.43015
geneN	14	0	5	1.9544 (1.340876)	6.5597 (7.182900)	0.30	0.01705
geneO	14	1	3	4.1943 (2.426698)	10.0863 (7.081891)	0.42	0.01081
geneP	14	0	2	8.7584 (7.050465)	8.3068 (6.510959)	1.05	0.64243
geneQ	13	0	0	4.7667 (1.708904)	9.0854 (5.494085)	0.52	0.15701
geneR	13	0	3	12.5791 (5.068220)	13.8879 (12.655225)	0.91	0.76460
BICC1	12	0	0	59.6866 (27.589815)	19.5257 (27.041201)	3.06	1.93E-06
geneS	12	0	5	34.7181 (16.923796)	20.0511 (6.294769)	1.73	0.30198
geneT	12	0	1	6.4096 (0.307941)	2.7122 (2.271332)	2.36	0.00E+00
geneU	12	1	3	0.8588 (0.199370)	1.1791 (0.755223)	0.73	0.35763

表2. 遺伝子とその近傍に生じたBreak Point と発現量の変化

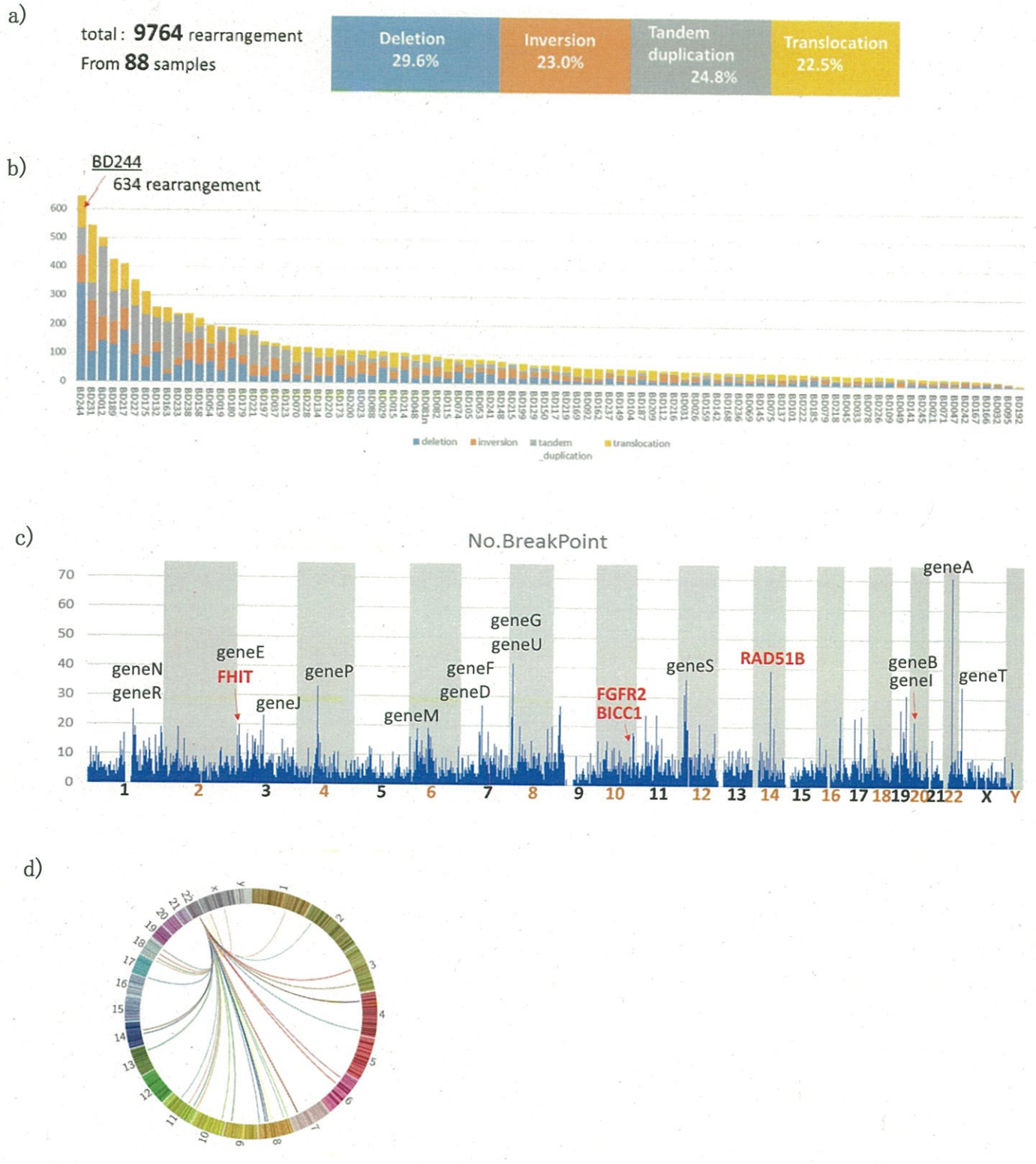


図2. 胆管がん88サンプルから検出された Genome rearrangement

88サンプルから得られた9764GRのタイプ別分布；b) サンプルごとの検出GR数とそのタイプ；c) ゲノム全長におけるBreak Pointの分布。ゲノム配列を1Mbのウィンドウ(オーバーラップ0.1M)に区切り、それぞれの区分において検出されたBreak Pointの数を縦軸にとった；d) c)において、特に顕著なピークが見られたchr22:30Mb付近で発生したTranslocationタイプのGRについて、転移先をCircosPlotで示した。CircosPlot内の線分の色は、検出されたサンプルによって塗りわけた。

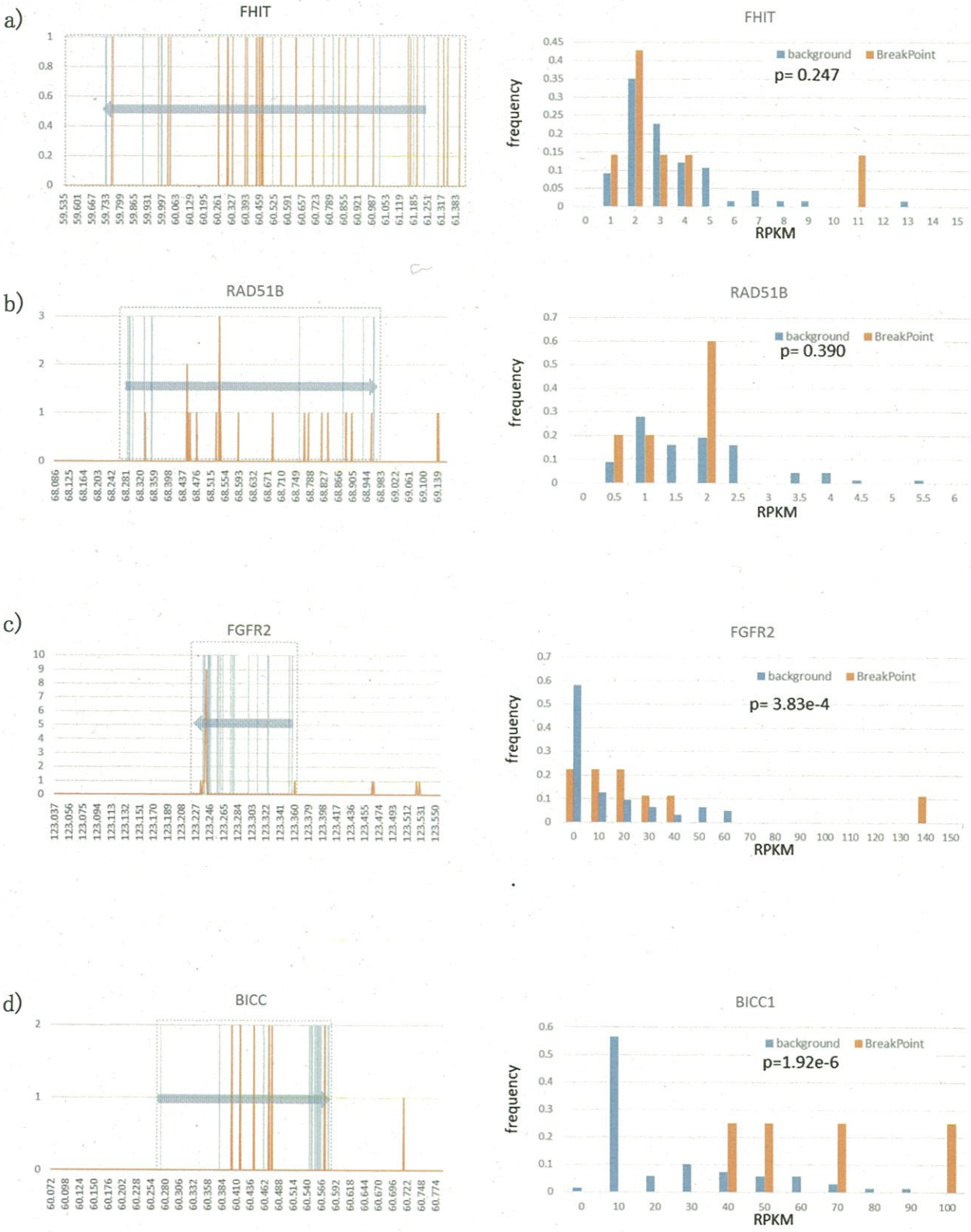


図3. Break Pointの高度な集積が見られた遺伝子とその発現量

右：遺伝子構造(エキソン：薄青色)とBreak Pointの頻発する領域。 左：右図、四角で囲まれた領域にBreakPointが生じたサンプル群(オレンジ)と、background群(青)の遺伝子発現量(RPKM)分布。縦軸はそれぞれの群における頻度を表す。

	read length	total read	mapped on HBV	No. virus integration	No.intra-rearrangement
HX14	50bp	693,288,506	3,408	2	1
HX19	101bp	782,002,422	2	0	0
HX25	100bp 125bp	2,030,033,190 54,028,184	10,918	13	14
HX28	100bp 125bp	1,615,928,114 236,255,690	3,522	4	12
HX33	100bp 125bp	745,748,049 47,519,371	753	3	2
HX35	100bp 125bp	750,911,466 44,009,959	808	7	1

表3. HB Virus integration 検出に使用された肝がんWGSデータ

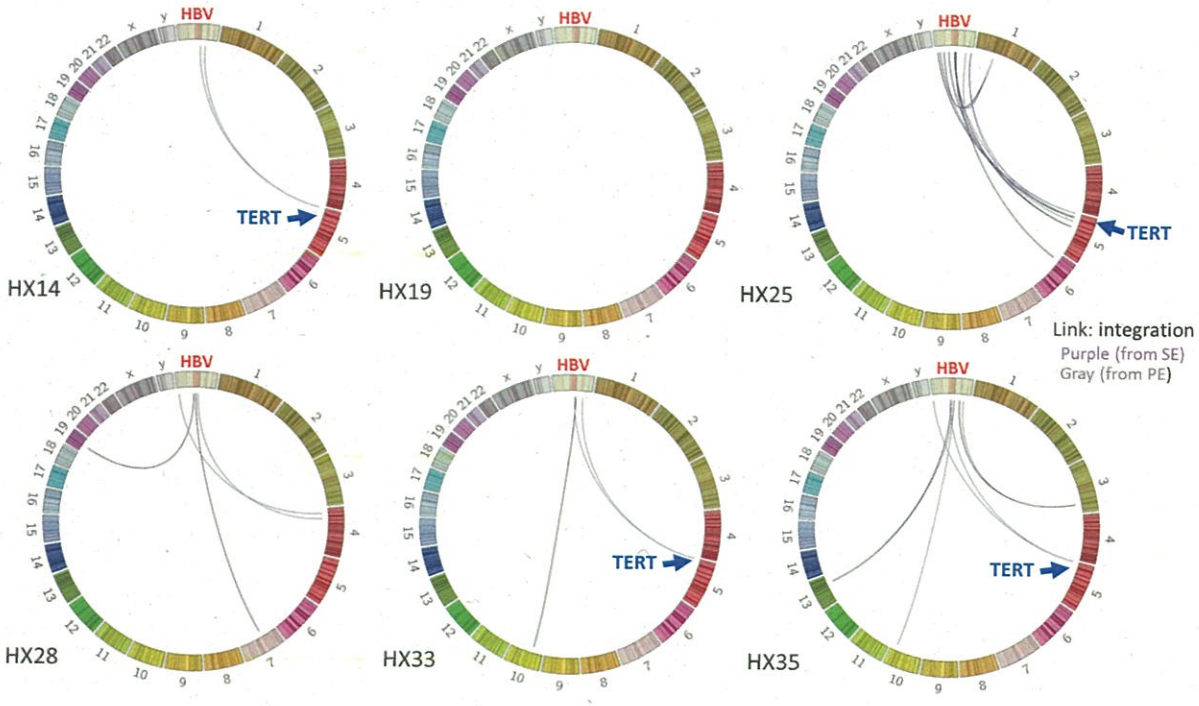


図4. 肝がん6サンプルから検出された Hepatitis B virusのヒトゲノムへの挿入

PEリードに挟まれた領域として検出された接合点を黒線で、エンドリード内部から検出された接合点を紫線で表示した。なお、両者が同一の接合箇所を出力した場合は、より正確な接合点の情報であるエンドリードの検出結果を残して重複を除いた。