【表紙】

平成２９年度シニア・リサーチフェロー

研究成果報告書

平成　30 年　4 月　28 日 提出

公益財団法人　がん研究振興財団
理　事　長　堀　田　知　光　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　29 年　4 月　1 日

至　　平成　30 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

要旨

本研究計画では、全ゲノムシークエンス(WGS)データとエピゲノムデータを統合して扱う自動解析パイプラインの開発を主題とし、さまざまながん種のWGS解析で実践を重ねながら、新たに報告された非コード領域におけるドライバー異常を検出する独自ツールの開発・実装を行ってきた。研究計画の最終年度にあたる平成29年度では、softclip mappingデータを起点とする新たなGenome Rearrangement (GR)検出ツールを開発し、先に構築したGR検出パイプラインに追加実装した。これを用いて、従来のpaired end (PE)リードによる検出を行った胆管がん88症例、胃がん81症例から再度GR検出を行った。その結果、100bp以下の小規模欠失を中心に、胆管がん88症例では6,251の、胃がん81症例では11,021のGRが新たに検出された。また、胃がんの中小規模GRからは、大規模GRとは分布の異なる、中小GRに固有のhotspotを見出すことができた。

GRで生じるbreakpointの殆どは非コード領域に生じており、GR変異が遺伝子発現に与えるインパクトを評価するためには、活性化enhancer/promoterの位置情報が不可欠となる。胆管上皮細胞においてはゲノム上の制御領域を網羅的に調べた研究の報告は未だになく、今回新たにWGS解析で使用した胆管がん症例から作成した胆管がん株化細胞、および健常人胆道上皮細胞それぞれ3サンプルを用いて、ChIP-seq によりH3K27ac及びH3K27me3の結合部位を決定した。健常細胞由来であるHBDEC2、HBDEC5の間で共通するヒストン修飾領域と、がん細胞由来であるCC3、CC6間で共通するヒストン修飾領域を求め、この2群の包含関係からがん細胞群に特異的な1249個のヒストン修飾領域、及び健常群に特異的な37個のヒストン修飾領域を同定することができた。ROSEアルゴリズムを使用してSuper enhancer領域を定義したところ、Super enhancer領域上に縦列重複型のGRが集中して分布するケースが観察された。

さらに、研究計画の主眼のひとつであるWGSデータとエピゲノムデータの統合解析の端緒として、WGS、RNAseq、抗H3K27ac-ChIP-seqデータを使用したドライバー変異GR、およびそのターゲット遺伝子を検索するパイプラインを構築した。これにより胆管がん73症例、胃がん61症例から検出されたGRと39,482遺伝子発現について解析を行った結果、GR挿入に伴う発現上昇が期待される遺伝子として胆管がんでは118（上流制御領域46、下流制御領域46、UTR5、CDS21）、胃がんでは211（上流制御領域93、下流制御領域85、UTR8、CDS25）の候補が見つかった。これらの候補遺伝子とその周辺に生じたGR、および活性エンハンサー分布をIGVブラウザ上で表示し、GRが編集した後のローカルゲノム構成を推定して、遺伝子発現を変化せしめた制御領域を同定する試みを始めている。

目次

略語：

WGS ： Whole Genome Sequence

GR ： Genome Rearrangement

BP ： Break Point

1. 序文

近年の次世代シーケンス技術をめぐる進化はめざましい。低コスト化と共に多様化が進み、特にエピゲノム分野においては染色体間距離を決定するChIA-PET(Fullwood et al., Nature 2009)や、転写調節分子のアクセシビリティを評価するATACseq(Buenrostro et al., Current Protocols in Molecular Biology 2015)など、次々に新たな次元の情報が生み出されている。これらの多層エピゲノムデータにより、従来では解析が困難であった遺伝子発現制御領域でのゲノム変異も解析の俎上にのせられ、がん研究に新たな局面を拓いている。2014年、Northcottらにより報告されたGenome rearrangementによるenhancer-遺伝子間距離の変化によりがん遺伝子（GF1）の活性化が引き起こされるEnhancer hijacking機構(Northcott., Nature 2014)、ゲノム上の染色体局所構造であるInsulator Neighborhoodの破綻ががん遺伝子の異常発現を引き起こす機構(Hnisz et al., Science 2016)や、3'UTR領域が欠失したPD-L1遺伝子の転写産物が、細胞による分解を免れて細胞内に蓄積することが発がんの契機となる機構(Kataoka et al., Nature 2016)などは、WGSデータとエピゲノムデータの統合により初めて解明された新規発がん機構の好例である。WGSデータであれ、RNAseqデータ等のエピゲノムデータであれ、単一の解析データの解析から結論を導く研究は既に古いスタイルとなりつつあり、2018年現在では、十分に大規模な症例群を対象とし、WGSを基底とし多元的な複数のエピゲノムデータを一元化して発がんに至る要因を捉えた研究が注目を集めている。

申請者が所属する研究グループにおいても、大規模症例群から得たエキソームデータ、及びRNAseqデータやメチル化データ等のエピゲノムデータを解析対象とした研究の蓄積があり、既に複数の成果が論文にまとめられている(Totoki et.al., Nature genetics 2014, Nakamura et.al., Nature genetics 2015, Hama et.al., Nature Communications 2018)。しかしながら、WGSデータ解析（特にGR検出）を効率よく進めるためのパイプラインは確立されておらず、研究の進捗を遅延させる要因の一つとなっている。既にBreakdancer(Chen et.al., Nature Mathods 2009)、Delly(Rausch et.al., Bioinfomatics 2012)、国内においてはGenomon(Chiba et.al., Bioinfomatics 2015)などの既成のGR解析ツールが複数公開されているが、このような外部の研究者により開発されたツールでは検出スコープが開発者の要求に合わせられており開発した時点で想定にない未知のゲノム変異を検出することは難しく、後追い研究になりがちである。また、WGS解析についてまわる巨大な計算機コストについても、解析完了までの計算時間、メモリ消費量、必要ディスク量のいずれの節約を優先させるかについてのユーザー側のリクエストを反映させうる余地が乏しく、これも解析効率を落とす要因となる。さらに、解析対象とするWGSデータの配列長、read depthなどに依存してGRの検出能（疑陽性/偽陰性の傾向）は異なるが、既成ツールでは解析中間ファイルが残されず、検出プロセスの詳細がブラックボックスになりがちであるため、このような変化に対応することは非常に困難である。

そこで我々は、WGSデータとエピゲノムデータとの統合をして扱う解析パイプラインの開発を本研究計画の主題とし、さまざまながん種のWGS解析で実践を重ねながら、新たに報告されたがん化機序の検出ツールの追加実装を行ってきた。平成27年度には現在の解析パイプラインの基盤となる、リードのクオリティチェックからGR検出までを自動化して行うパイプラインを構築し、これを乳がんWGS10症例、胆道がんゲノム88症例に適用した。また、肝がんゲノム解析でのニーズに応え、ゲノム内のHBVインサート部位の同定を行うツールを開発し、これにより肝臓がんWGS6症例の解析を行った。研究計画の二年目にあたる平成28年度にはInsulator Neighborhood(IN)破綻による遺伝子の発現異常の検出パイプライン、及び3'UTR領域が欠失した転写産物の異常蓄積を検出するパイプラインを開発し、WGS解析パイプラインに実装ののち、これを胆管がん88症例に適用してこれらの機序による発がんの可能性について検討した。研究計画の最終年度にあたる平成29年度では、GRの検出スコープを中小規模GRに拡大したオプションの追加実装を行った。さらに、WGSから検出したGRががん化ドライバーとして機能する可能性について、RNAseqデータ、ChIPデータを統合して評価を行うパイプラインを作成した。これをWGSとRNAseqデータの両方が揃う胆管がん73症例、胃がん61症例に適用して、がん遺伝子の発現亢進を促すGRとその対象遺伝子の組み合わせの絞込みを行った。

2. softclip型検出ツールの開発と既存パイプラインへの追加実装

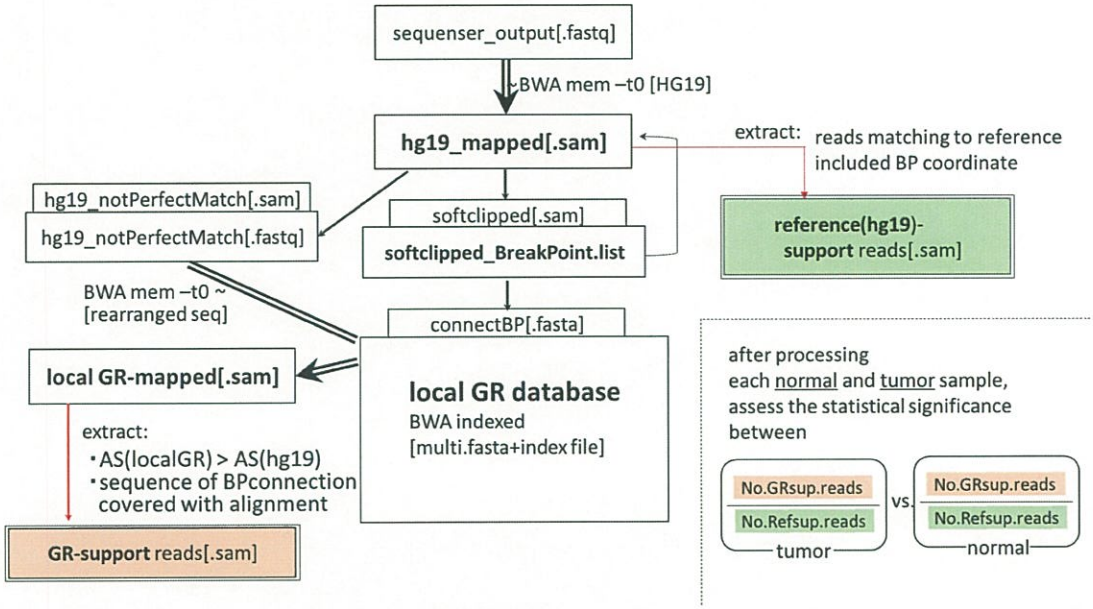
2-1 序

2013年に公開されたBWA-MEMマッピングアルゴリズムの登場に伴い、ショートリードの部分配列がマッピングされた部位の座標情報、すなわちsoftclip mappingデータが利用可能となった。従来法では1リードのマッピング先は原則1箇所であり、そこにアライメントできないリードの部分配列はummapped readとして処理されていたが、BWA-MEMでは全長アライメントが出来ないリードを部分配列に分け、それぞれにゲノム上のマッピング先を決定した結果がsupplementary alignmentとして出力される。この情報により、PEリードのインサート領域で生じたGRが対象となる、これまでのGR検出アルゴリズムで取り落とされていた小中規模（～1Kb）GRの検出が可能となった。GRの発生機序には染色体橋の形成と破壊によるchromothripsis説、修復酵素複合体の機能不全説など諸説あるが、サイズの異なるGRではそれぞれ異なる発生機序が関与している可能性が考えられる。H27年度に開発したGR検出パイプラインにおいては、PEリードのインサート領域に含まれるGRのみが検出対象であったが、softclipリードデータを利用したGR検出ツールの開発により、中小規模GRに特異的なhot spotが新たに見出されないか検討した。

H29年度ではsoftclip mappingデータを起点とする新たなGR検出ツールを開発し、先に構築したGR検出パイプラインに追加実装した。これを用いて、従来のPEリードによる検出を行った胆管がん88症例、胃がん81症例から再度GR検出を行い、softclipリードによるGR検出数の増分およびそのサイズ分布やhot spotについて検討を行った。

2-2 方法

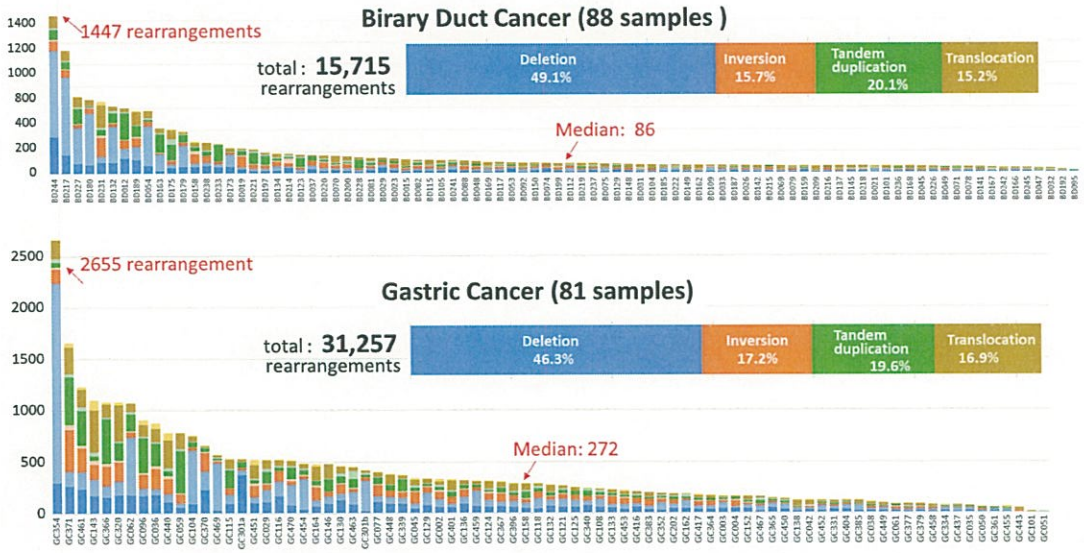
H29年度に新たに構築した、softclipリードを起点とするGR検出パイプラインのフローチャートを図1-1に示した。このパイプラインでは始めにシーケンサーから出力されたWGSリードを、t0オプションを指定したBWA-MEMによりリファレンスゲノム（hg19）にマッピングし、ここからsoftclip mappingを含むリード（以下、softclipリードと記述）のみを抽出する。softclipリードからPCR重複を除いたのち、配列中に含まれるbreakpoint座標情報をリスト化する。次に、リファレンスゲノムから切り出された1500bpの部分配列がリストに含まれるbreakpointで連結した、3000bpの再構成ゲノム配列群からなるデータベースを作成する。この再構成ゲノム配列上に、再度BWAによりWGSリードをマッピングし、対象となる3000bpの領域にマッピングされるリードをGR-supportリード、またはreference-supportリードのいずれかに分類する。すなわち、前者は再構成配列上に、後者は再構成される前のオリジナルのreference配列上に、全長がproperlyにアライメントされるリードを意味する。同一症例のがん細胞サンプル/正常細胞サンプルについてそれぞれ対象breakpointごとにGR-supportリード数、及びreference-supportリード数を求め、この４つの値によりFisher検定を行って、p値を算出した。



[図1-1] ： softclip mapping read からのGR検出フローチャート

2-3 結果及び考察

新たにsoftclip型リードの情報をもとにGR検出を行うパイプラインを開発し、H27年度に構築したPE型GR検出パイプラインに追加実装した。この従来型パイプラインは、既に胆管がん88症例、胃がん81症例のGR検出に適用されており、それぞれ9,464、20,236のGRが検出されている。これらの症例群を対象として改訂版パイプラインにより再度GR検出を行い、softclipリードの利用により新たに検出されるGRについて検討を行った。



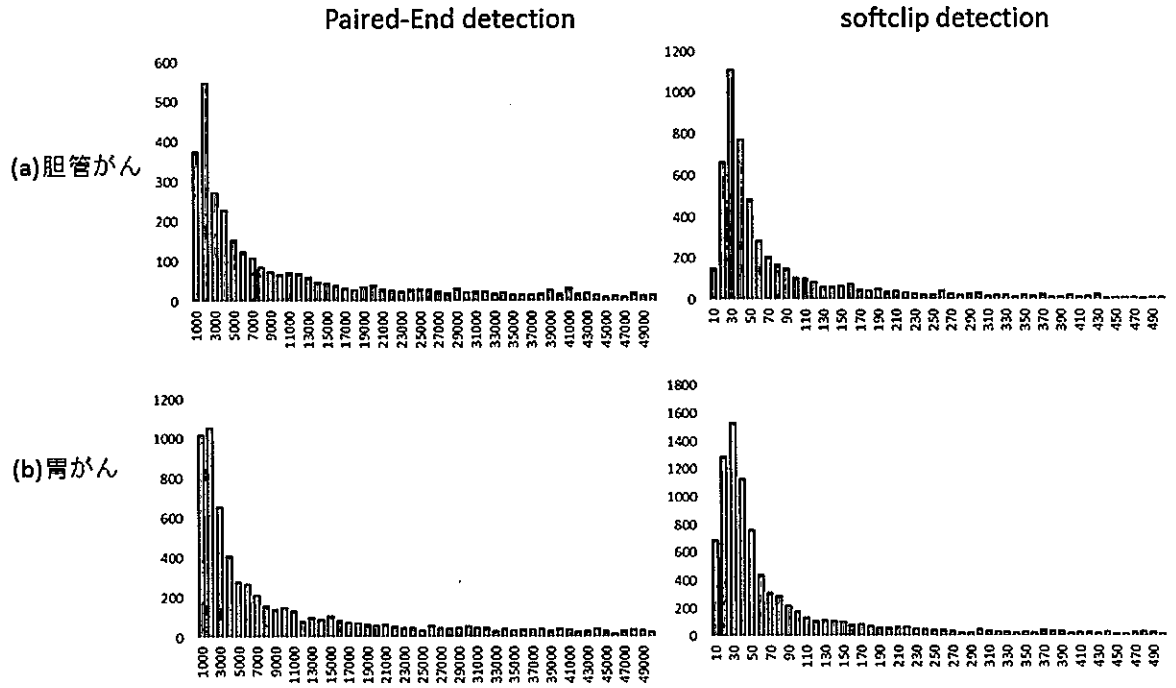
[図 1-2]：改訂後 GR 検出パイプラインによる検出数概観

帯グラフと棒グラフの色は対応しており、青色は欠失型 GR を、橙色は転位型 GR を、緑色は縦列重複型 GR を、茶色は転座型 GR をそれぞれ示す．棒グラフで淡色で示された部分が soft clip 型検出で新たに追加された GR を示す．

	Gastric Cancer		Bile duct cancer	
total No.new-GR	11021		6251	
deletion	8559	78%	4914	79%
inversion	583	5%	367	6%
tandem_duplication	1147	10%	789	13%
translocation	732	7%	181	3%

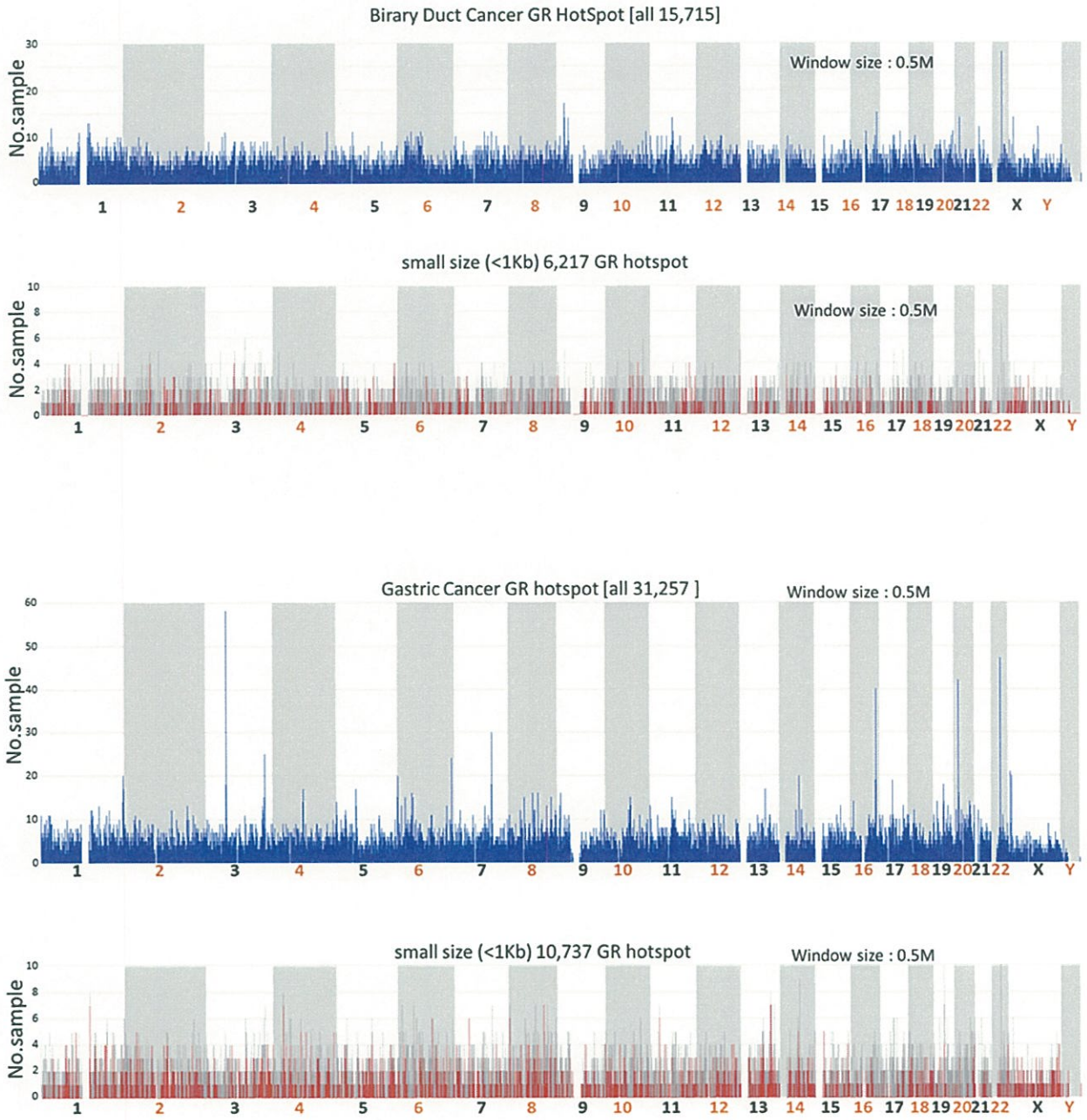
[表 1-1]：新たに検出された GR のイベント分布

図1-2に、胆管がん、胃がんそれぞれの改訂版パイプラインのGR検出結果概要を示した。胆管がん88症例からは新たに6,251のGRが加わった計15,715のGRが、胃がん81症例からは11,021のGRが加わった31,257のGRが検出された。図1-2の各症例の検出数を示した棒グラフ中、淡色で示された部分によりsoftclip型検出で新たに加わったGRを示している。増分の多くは水色で示される欠失タイプのGRであり、胆管がんでは79%（4,914/6,251）、胃がんでは78%（8,559/11,021）を占めていた。縦列重複タイプのGRがこれに続き（胆管がんで13%、胃がんで10%）、転位型、染色体転座型のGRも僅かながら新たに検出された（表1-1）。



[図 1-3]：従来型 GR/ 新しく検出された GR のサイズ分布

従来型のimproper-PEリードから検出されるGR群と、新たにsoftclipリード検出されたGR群について、translocationを除くGRでGRのサイズ（2つのbreakpointに挟まれる距離）の分布について図1-3にまとめた。左側にPE検出型GRのサイズについて、右側にsoftclip型GRのサイズについて棒グラフでそれぞ



[図 1-4]：従来型 GR/ 新しく検出された GR の hot spot

上段：胆道がん 88 症例、下段：胃がん 81 症例から検出された GR 発生の hotspot. 青線で示した棒グラフは全ての GR での発生頻度を、灰色線は小規模 GR に限定した発生頻度を、赤色線は（小規模 GR での発生数）－（中～大規模 GR 発生数）による頻度を示す.

れ分布を示した。胆管がん、胃がんともにPE検出型では1Kb-2Kb程度のGRが、新たに加わったsoftclip検出群では20-40bpサイズのGRが最頻値となっており、これまで取り逃していた小規模・中規模GRが新たに検出対象となったことが確認された。

これまでの解析から、胆管がん・胃がんの大規模GRには細胞種ごとに異なるエピゲノム状態が反映された固有のhot spotが存在することが明らかになっていた。そこで新たに検出された小中規模GRで、大規模GRの傾向を踏襲しない独自のhot spotが存在するかを検討した（図1-4）。全ゲノム領域をウィンドウサイズ0.5M（前後のオーバーラップ50Kb）として分割し、各ウィンドウの領域中でGRが生じた症例数を縦軸に、全ゲノム領域を横軸にとり、全てのGR（図1-4aの胆管がんでは15,715、図1-4bの胃がんでは31,257）の各ウィンドウにおける発生頻度を青色の棒グラフで示した。また、今回のパイプライン改訂により新たに検出された小中規模GR（胆管がん6251、胃がん11,021）のみを対象に同様のグラフを作成した。後者のグラフでは、各領域における小中規模GRの発生回数を灰色の棒グラフで、この数値から同ウィンドウにおける大規模GRの発生回数の差分をとったものを赤色の棒グラフで示している（正数のみ）。胆管がんではchr22:30Mb付近、chr10:113Mb付近に小中規模GRのピークが見られたが、この領域は大規模GRでも高頻度にGRが生じるhotSpotとなっており、小中規模GRに固有のピークは見出されなかった。一方、胃がんにおいては大規模GRのhotspotとオーバーラップしない、小中規模GRに固有のhotspotがchr1:145Mb付近、chr4:31Mb付近、chr6:48Mb付近、chr8:127Mb付近、chr14:55Mb付近に認められた。これらのhotspotを構成する小中規模GRは密に分布しており、ウィンドウサイズを10Kbに絞ってもウィンドウ内に4症例以上でGRが発生するhotspotが10箇所みとめられた。

3. 胆管がん抗H3K27ac/抗H3K27me3-ChiP 解析

3-1 序

遺伝子間領域を含む全ゲノムデータの決定が容易となった現在、遺伝子以外の領域のアノテーション情報の重要性が高まっている。既に完了したENCODEプロジェクトにより、主要な9種類のヒト細胞では全てのゲノム領域に機能エレメントによるアノテーションが付加されている。しかしながら、機能エレメントの分布は細胞種ごとに大きく異なるため、ENCODEで対象となった細胞以外のアノテーションに際しては参考程度にしかならず、また同一細胞種であっても、発生プロセスやがん化により制御領域の活性化状態はダイナミックに変化することが複数の論文から報告されている［Ooi.et.al., NatComm2016］。 GRで生じるbreakpointの殆どは遺伝子間領域に生じており、GR変異が細胞に与えるインパクトを評価するためには、活性化enhancer/promoterの位置情報が不可欠となる。胃がんではDengらにより、5サンプルのH3K4me1、H3K4me3、H3K27ac、H3K36me3の結合部位がChIP-seqで決定されており、

GEOから既に公開されているので(Deng et.al., Nature Communications 2014)、これを4章で述べるGRのアノテーションに使用した。しかし、胆管上皮細胞においてはゲノム上の制御領域を網羅的に調べた研究の報告がないため、今回新たにWGS解析で使用した胆管がん症例から作成した胆管がん株化細胞、および健常人胆道上皮細胞それぞれ3サンプルを用いて、ChIP-seq によりH3K27ac及びH3K27me3の結合部位を決定した。

得られたChIP-seqデータにMACSアルゴリズム(Zhang et.al., Genome Biology 2008)を適用してピークを同定し、サンプル間でのピークの包含関係を調べることから、正常細胞と胆管がん細胞でそれぞれ特異的に使用される調節領域の抽出を行った。また、ROSEアルゴリズム(Warren et.al., Cell 2013)を使用し、胆管がん細胞および胆管上皮細胞のsuper enhancerの同定を試みた。IGVブラウザによりenhencer/super enhancer 及びGRのゲノム上での分布を表示したところ、一部のsuper enhancer上で縦列重複GRが密集する特徴的な分布が観察された。

3-2 方法

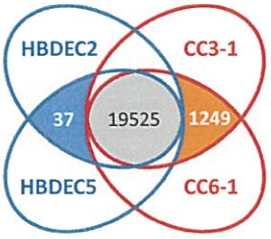
表2-1に調節領域同定に使用した６サンプルを示した。サンプルIDに"HBDEC"を含む３サンプルは、健常個体の胆管がん上皮細胞の初代継代培養に由来するものである（HBDEC2_3H10はHBDEC2細胞を株化させたものとなる）。"CC"をサンプルIDに持つ3サンプルは、胆道がん上皮細胞を株化させた細胞に由来するものである。各サンプルにつき、H3K27ac, H3K27me3を抗体に用いてChIP-seqを実施し、それぞれの抗体のゲノム上における結合領域を決定した。また、コントロールとして全細胞抽出液（Whole Cell Extract, WCE）でChIP-seqを行った。

これらのChIP-seqデータから、MACSアルゴリズムによりピークコールを行った。narrow peakであるH3K27acのピークコールにはMACS1.4をディフォルトパラメーターで、broad peakであるH3K27me3のピークコールにはnolamda オプション、及びnolambda オプションをつけたMACS2により実施した。また、macs1.4によって定義されたH3K27acピークと、その入力に使用したH3K27acリードデータ、及びコントロールリードデータを使用し、ROSEによるsuper enhancer領域同定を試みた。

3-3 結果と考察

GRパイプライン改訂により、胆管がん88症例のWGSデータから1万5千を超えるrearrangement部位が同定された。その多くは遺伝子間領域に生じており、これらのGRが細胞に及ぼす影響を評価するために、胆管がん細胞で発現調節を行う領域を同定したい。H3K27acは遺伝子発現を促す活性エンハンサー領域及び活性プロモーター領域のマーカーとして、H3K27me3は発現が抑制されたサイレントなゲノム領域のマーカーとして一般に使用されている(Menno et.al., PNAS 2010)。この2つのマーカー

Sample ID	Cell type	Antibody	# Reads	% of >= Q30 Bases (PF)	Mean Quality Score (PF)	% hg19 Align (PF)	# peaks (MACS)	# Ac peaks not overlap me3	# peaks (ROSE)
HBDEC2	primary (normal)	K27ac	21,740,607	95.95	37.43	88.58	69499	36404	1049
		k27me3	23,527,400	98.08	38.39	84.90	21722		
		WCE	12,532,998	91.63	36.66	70.94			
HBDEC2_3H10	cell line (normal)	K27ac	58,087,464	96.92	37.85	89.66	36462		
		k27me3	15,784,458	98.36	38.50	82.91	127989		
		WCE	65,884,481	97.39	38.05	80.67			
HBDEC5	primary (normal)	K27ac	51,628,659	97.00	37.88	89.51	69021	68103	1714
		k27me3	28,677,785	98.04	38.38	85.37	31845		
		WCE	62,466,223	97.36	38.04	81.59			
CC3-1	cell line (tumor)	K27ac	36,667,990	96.05	37.46	88.79	37202	36714	1188
		k27me3	13,527,114	94.33	36.94	78.61	325311		
		WCE	15,622,891	93.54	36.77	73.74			
CC4_1		K27ac	22,916,651	98.29	38.49	88.76	28997		
		k27me3	27,917,057	98.39	38.51	85.44	65425		
		WCE	10,871,128	89.74	36.11	67.36			
CC6-1		K27ac	25,340,314	95.92	37.39	85.39	32746	32732	755
		k27me3	28,283,394	96.25	37.54	83.45	64070		
		WCE	13,429,538	92.97	36.61	72.45			



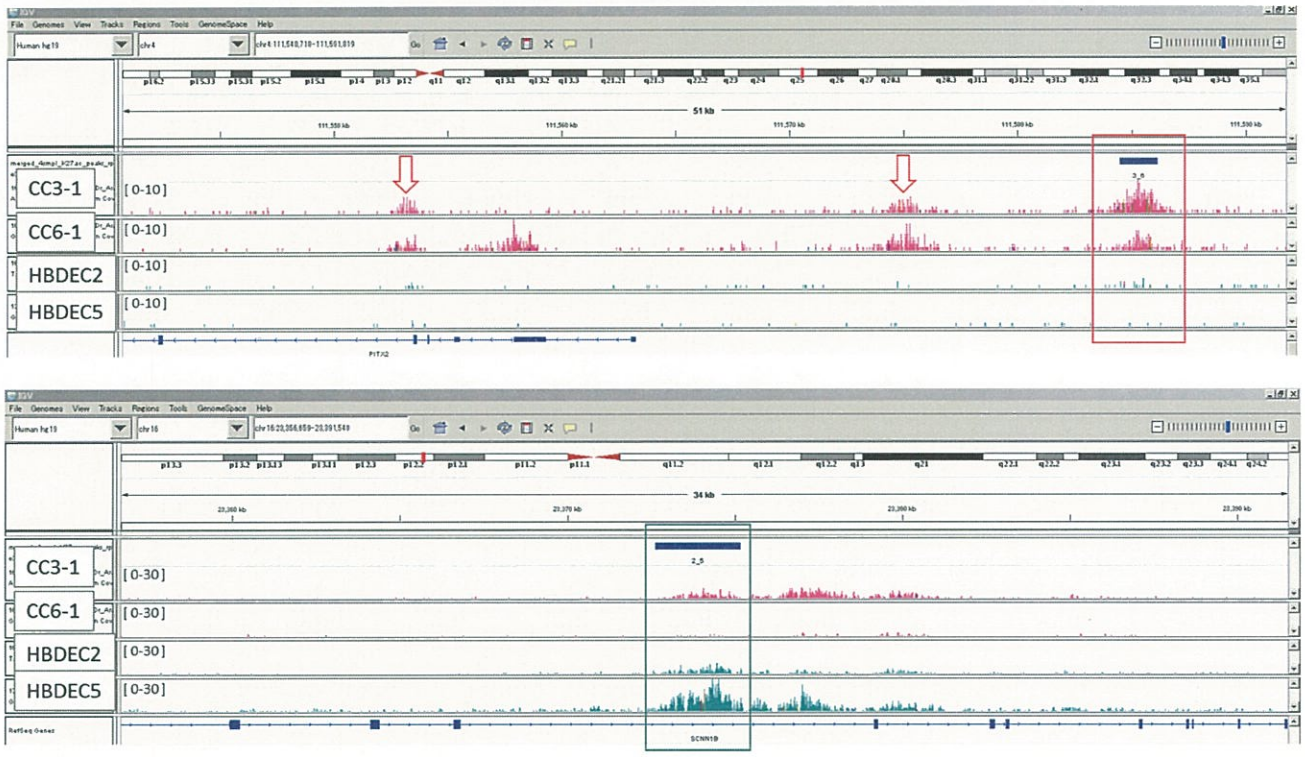
[表 2-1]：胆管上皮細胞の活性化 Enhancer 領域の同定に使用した細胞

を抗体とし、健常人由来胆管上皮細胞、および胆管がん細胞から樹立した株細胞のChIP-seqを行った。

6つのサンプルそれぞれから得られたリード数、及びMACSによりコールされたピーク数を表2-1に示した。確実に転写活性化能を有するエンハンサー領域を定義するため、H3K27me3ピークが観察された領域とオーバーラップする領域のH3K27acピークを除いたものをActive peakとし（表2-1, #Ac peaks not overlap me3）、今後の包含関係推定および4章で記述するGRアノテーションに使用した。なお、CC4-1のH3K27ac及びHBDEC2-3H10のH3K27me3では十分なChIP-seq出力が確保できず、偽陰性/疑陽性が多く含まれることが予想されたため、CC4-1およびHBDEC2-3H10は評価対象から除外した。

健常細胞由来であるHBDEC2、HBDEC5からはそれぞれ36,404個、68,103個の、がん細胞由来株細胞であるCC3、CC6からはそれぞれ36,714個、32,732個のActive peakが同定された。MACSアルゴリズムでは2相性のピークの高さ、及びピーク間の距離から抗体結合部位のピークコールを行うため、ピーク数はライブラリ調整環境（抗体結合以外のゲノムDNAの分解酵素の働き）に依存し、再現性の高い結果を得ることは難しい。そこで、健常細胞由来であるHBDEC2、HBDEC5の間で共通するActive peakと、がん細胞由来であるCC3、CC6間で共通するActive peakを求め、この2群の包含関係からそれぞれの群に特異的なActive peakを抽出することをこころみた。その結果、がん細胞群に特異的な1249個のActive peak、及び健常群に特異的な37個のActive peakが見出された（表2-1右）。

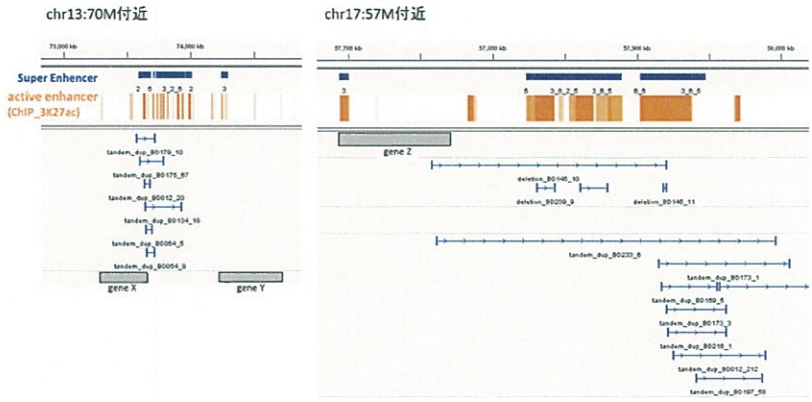
図2-1(a)として示したがん細胞群に固有のActive peakにおいては、健常群ではほとんどリードが張り付いていない領域に、新しくH3K27ac結合性リードのピークが生じていた。また、MACSがピークとして検出した領域以外にも、CC3、CC6の両方でH3K27ac結合性リードが密集してマッピングされる領域が複数あり、これらも細胞のがん化に伴い新たに生じた活性Enhancer領域として機能する可能性がある。



[図 2-1]：がん細胞および健常細胞 で特異的にあらわれるピークの例

一方で、図2-1(b)に示した健常群固有Active peakとされた領域には少数ながらもH3K27ac結合性リードがマッピングしている様子が観察された。がん細胞群に固有のActive peak数に比べ、健常群に固有のActive peak数が極端に小さいことからも、がん細胞では既存のエンハンサー活性は基本保持したまま、活性化領域を追加することで増殖に有利となる遺伝子の発現を調整していると推測される。

HBDEC2、HBDEC5、CC3、CC6の4サンプルについては、ROSEによるsuper enhencer領域の同定を行い、これと前項で定義したActive peak、及び2章で検出した15,715箇所のGRをIGVブラウザ上に表示させ、その分布を観察した。super enhancerが占める領域（図2-2,青い帯で示された領域）に、縦列重複型のGRが集中して分布するケースが数例観察された。ROSEアルゴリズムによるsuper enhancerの定義の一つ



[図 2-2]：super enhancer 領域で観察された縦列重複 GR の密集

は高密度なenhanecerピーク分布であり、集中して分布する活性化enhancer領域に効率よく転写酵素複合体が結合することにより強い転写活性が生じるとされている。がん細胞でsuper enhencer領域に生じた縦列重複は、高密度enhanecer領域を増幅させることによりさらにエンハンサー活性を強めさせ、がん化に有利な遺伝子の発現増幅に働いている可能性があるため、現在これらのGRが生じた症例で特に発現が上がっている遺伝子について検討を進めている。

4. WGS(SV)-RNAseq-抗H3K27acChIPデータを使用した統合解析パイプライン

4-1 序

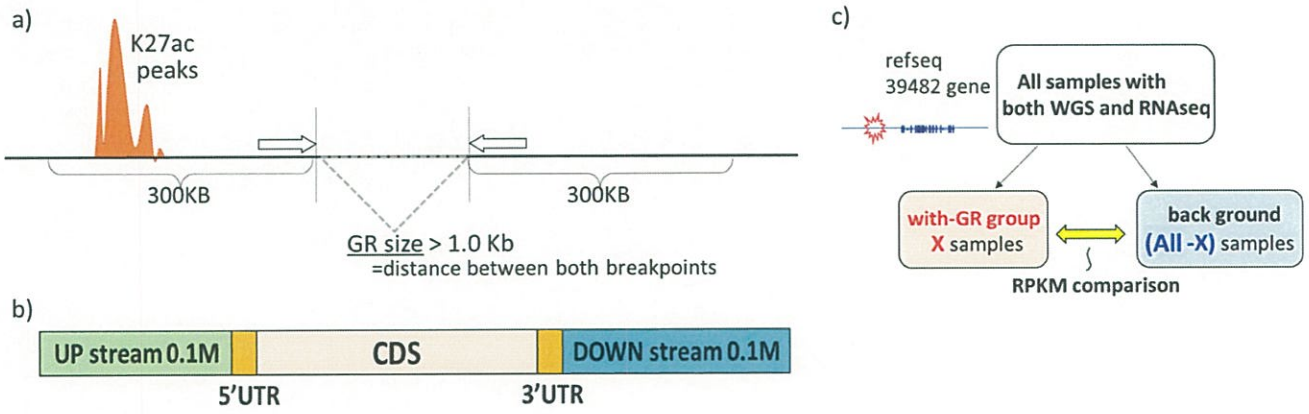
がん細胞で大量に発生するGRは遺伝子の発現、ひいては細胞の表現型にどのような影響を与えるのだろうか？これまでに遺伝子間領域に生じたGRが発がんドライバーになるケースは髄芽腫細胞(Northcott., Nature 2014)、急性リンパ芽球性白血病(Hnisz et al., Science 2016)など複数報告されており、他のがん種においても類似の機構ががん化に働いていることは十分に考えられる。大規模症例でWGSデータが生成され、対応するエピゲノムデータも出揃うようになった昨今、情報解析に求められるのはゲノムに生じた変異（因）とそれにより変化したエピゲノム情報（果）を統合的に解析し、複数の症例の結果を全ゲノムスケールで簡便に俯瞰できるシステムであると考える。

本研究では、その端緒としてWGSデータから検出したGR、H3K27ac及びH3K27me3のChIP-seqデータ、RNAseqにより得られた39,482の遺伝子の発現量データを使用し、GRの発生がきっかけとなって発現変動が起こる遺伝子について、ncRNAを含むゲノム上の全トランスクリプトを対象に解析を行うパイプラインを構築した。

4-2 方法

本パイプラインではbreakpointの両端が定義されたGRリスト、解析対象細胞の活性化エンハンサーマーカーを抗体としたChIPseqデータ、GR検出に使用したものと同一症例のがん細胞から抽出したRNAseqデータを使用する。同一症例からWGSデータ、RNAseqデータの両方が得られている症例は胆管がんで73症例、胃がんで61症例あり、これを母集団として以降の解析を行った。はじめに、個別のGRがゲノムに与えるインパクトについて近傍の活性エンハンサーの有無、及びGRの規模からアノテーションを行った。遺伝子近傍300Kb以内に活性Enhancer領域が存在し、GRの両端breakpoint間の距離が1Kb以上であるGRを"エンハンサーと遺伝子間の距離関係を変化させ、遺伝子発現を変動させうるGRをとみなし、このようなGRのみを解析に採用した。

次のステップでは、遺伝子ごとに母集団の症例を対象遺伝子近傍にGR-breakpointが生じたGR群と、



[図 3-1]：がん化のドライバーとなる GR 変異及びその対象遺伝子の推定

それ以外のbackground群（以下、bg群）にわけ、対象遺伝子の発現量（RPKM値）にGR群とbg群で有意な差が見られるかをt検定/コルモゴロフ・スミルノフ（KS）検定/wilcoxen検定により評価した。GR群に該当する症例が１しかないケースでは、RPKM値の母集団中でのZスコアを使用した。遺伝子近傍の領域は、遺伝子上流0.1M以内/遺伝子下流0.1M以内/5'UTR及び3'UTR/coding region(CDS)の4領域に分類し、領域ごとに評価を行った。p値が0.05を下回る、またはZスコアの絶対値が2を上回る遺伝子を、GRが発現変動を引き起こした遺伝子候補として出力し、これらの遺伝子のRPKM値のビーンズプロット（図3-2）及びIGVブラウザによるGRイベントの発生箇所の観察（図3-3）により、候補の絞込みを行った。

4-3 結果

がん細胞ではゲノムの恒常性が大きく損なわれ、核形異常から一塩基変異にいたるまであらゆるスケールでの変異が蓄積していることが知られている。GRもゲノム不安定性を反映の一つとしてがんゲノムに頻発するが、WGSデータが無ければ検出できない情報であることに加え、その殆どが遺伝子間領域で生じるため、これまで細胞への影響について十分に解明されていなかった。しかしながら、WGSデータの増加と遺伝子間領域の機能アノテーションの進展に伴い、2012年以降、発現制御領域に生じたＧＲが直接的にドライバー異常として働くケースが複数報告されるようになった。

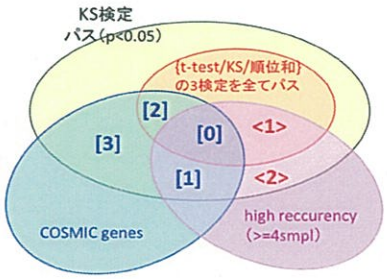
このような流れをうけ、がん化のドライバー異常となるGR、およびそのターゲット遺伝子を効率的に検索するパイプラインを構築し、これにより胆管がん73症例、胃がん61症例から検出されたGRと39,482遺伝子発現について解析を行った。　はじめに、検出されたGRの中から遺伝子発現を変化させるポテンシャルをもつ候補を絞り込んだ。ここでは、GR規模が十分に大きく（GRの2つのBP間距離が1Kbより大）、近傍300Kb以内に活性Enhancer/promoterが位置するものを採用した結果、胆管がんで13,021個（総検出量の83%）、胃がんで18,481個（同59%）が以降の解析の対象となった（図3-1a）。

さらに、遺伝子とその周辺の領域を、期待されるGRの影響によって４つに分類した（図3-1b）。遺伝子発現制御に異常を生じさせる効果が期待される領域として①遺伝子上流0.1Mbおよび②下流0.1Mbを、転写・翻訳機構を乱す領域として③UTR領域を、④融合遺伝子を生じさせうる領域としてCDS(イントロンも含む)を想定し、refseq gene とUCSC known geneから重複を除いて調整した39,482の各遺伝子について、4分類ごとにGR挿入群/background群を定義し、2群間のRPKM値分布の有意差検定を行った（図3-1c）。検定はt検定/コルモゴロフ・スミルノフ（KS）検定/wilcoxen検定の３つの検定法により行い、KS検定でp<0.05または|Zscore|>2.0（GR群に属する症例数が1のとき）を満たすものを候補として出力した（表3-2，出力例）。

以上の操作を行った結果、GR挿入に伴う発現上昇が期待される遺伝子として胆管がんでは118（上流制御領域46、下流制御領域46、UTR5、CDS21）、胃がんでは211（上流制御領域93、下流制御領域85、UTR8、CDS25）の候補が抽出された（表3-1）。一方で、GR挿入により発現の減少が期待される遺伝子候補は、胆管がんでは28（上流制御領域9、下流制御領域11、UTR2、CDS6）、　胃がんでは47（上流制御領域21、下流制御領域13、UTR1、CDS12）と、発現上昇が見込まれる遺伝子数の1/4程度にとどまった。また、上記で得たＧＲ−遺伝子変化の組み合わせが細胞のがん化に及ぼすインパクトを、以下の3つのパラメーターによりクラス分けした（表3-1右図）。

がん化に及ぼすインパクトを評価する３つのパラメーター

	重要度高	重要度低
判別有意性	3検定の全てでp<0.05	KS検定でp<0.05
GR群の構成数	4症例以上	1～3症例
GRの対象遺伝子	COSMIC census gene	other



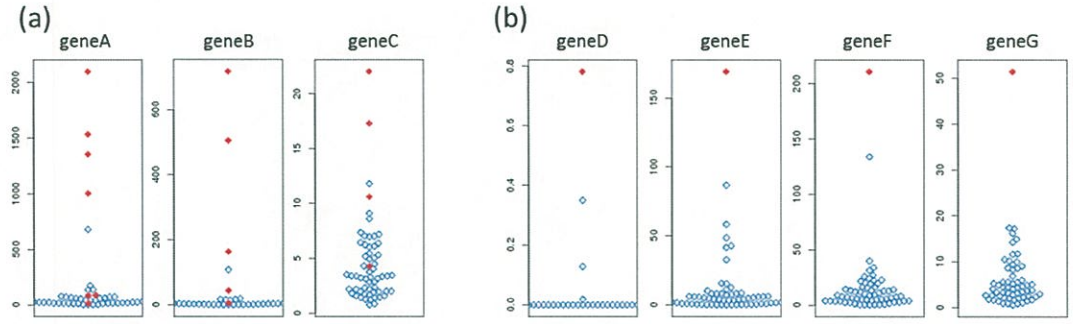
		Birary Duct Cancer				Gastric Cancer			
		UPs	DWs	UTRs	CDS	UPs	DWs	UTRs	CDS
exp. Increase	[0] validate_High + recurrency_High + cosmic	0	1	0	0	1	3	0	0
	[1] validate_pass + recurrency_High + cosmic	0	2	0	1	1	1	0	2
	[2] validate_High + cosmic	3	2	0	2	4	3	0	2
	[3] validate_pass + cosmic	31	30	5	13	25	20	6	10
	<1> validate_High + recurrency_High	7	6	0	4	46	41	2	4
	<2> validate_pass + recurrency_High	5	5	0	1	16	17	0	7
	total	**46**	**46**	**5**	**21**	**93**	**85**	**8**	**25**
exp. Decrease	[0] validate_High + recurrency_High + cosmic	1	0	0	1	1	0	0	2
	[1] validate_pass + recurrency_High + cosmic	0	1	0	0	0	0	0	1
	[2] validate_High + cosmic	1	2	2	1	3	4	1	0
	[3] validate_pass + cosmic	1	0	0	2	2	0	0	4
	<1> validate_High + recurrency_High	0	3	0	2	8	3	0	1
	<2> validate_pass + recurrency_High	6	5	0	0	7	6	0	4
	total	**9**	**11**	**2**	**6**	**21**	**13**	**1**	**12**

[表 3-1]：GR Break point の挿入により発現量が変化したと考えられる遺伝子

表3-1［0］はがん化に強く関連することが期待されるクラスで、他の症例群に比べGR挿入群で明確に発現が変化し、症例母集団中で高頻度に観察され、GRにより発現が変化する対象遺伝子ががん関連遺伝子であるものがここに分類される。このようなGR-遺伝子の組み合わせが、胆管がんでは3（発現上昇で1，減少で2）、胃がんでは7（発現上昇4、減少で各3）見出された。図3-2aに、胃がん下流制御領域で［0］クラスに分類された3つのGR群とその対象遺伝子について、ビーンズプロットでGR群（赤）とbackground群（青）での対象遺伝子のRPKM値分布を表現したものを示した。いずれのケースにおいても、圧倒的に高い発現量が観察された症例はGR群で占められており、GRが発現変動の原因となったと考えられる。一方、赤い点で示されるGR群症例の中にも特に高い発現量を示さないものも存在した。その原因として、本解析においては、GRイベントの種類（欠失/転位/縦列重複/転座）に関わらず、遺伝子周辺の対象領域にGR breakpointの挿入が見られたものを一律GR群としたことが考えられる。すなわち、同じカテゴリの領域に生じたGRであっても、生じたGRイベントによって対象遺伝子周辺の再構成後のローカルゲノム構成（遺伝子とその周辺の制御領域の配置）は一様ではなく、遺伝子発現に与える影響は各々異なると推測される。すなわち、GRイベントの種類によってenhancer配列を遺伝子近傍に導く場合と導かない場合があるので、さらにenhancer配列の移動の有無を見る必要があると考えられる。

category	gene	acc.	No.smpl BP	mean.RPKM BP group	mean.RPKM background	fold change	Zscore	p-value t-test(log)	p-value KS	p-value wilcoxen	
[0]	geneA		7	884.57	52.528	16.8		1.41E-02	1.87E-04	1.48E-03	(a)
	geneB		5	287.587	8.404	34.2		2.22E-02	2.15E-03	2.01E-03	
	geneC		4	13.552	3.856	3.5		3.57E-02	1.55E-02	7.04E-03	
[1]	gene,accessionID は論文作成中データであるため非公開		6	102.885	36.507	2.8		6.58E-02	4.22E-02	2.67E-02	
[2]			2	1344.038	7.433	180.8		1.09E-02	1.09E-03	1.78E-02	
			2	14.937	4.608	3.2		2.20E-07	6.56E-03	2.73E-02	
			2	28.186	16.836	1.7		9.87E-09	2.30E-02	4.50E-02	
			2	32.443	16.055	2		1.07E-01	3.93E-02	3.70E-02	
[3]	geneD		1	0.781	0.008	93.2	6.93				(b)
	geneE		1	169.021	8.926	18.9	6.09				
	geneF		1	210.604	12.664	16.6	6.24				
	geneG		1	51.376	5.402	9.5	6.22				
			1	0.247	0.039	6.3	3.78				
			1	168.558	37.71	4.5	3.76				

[表 3-2]：胃がんで遺伝子下流 0.1Mb に GR が生じた遺伝子で発現が上昇した遺伝子の出力例

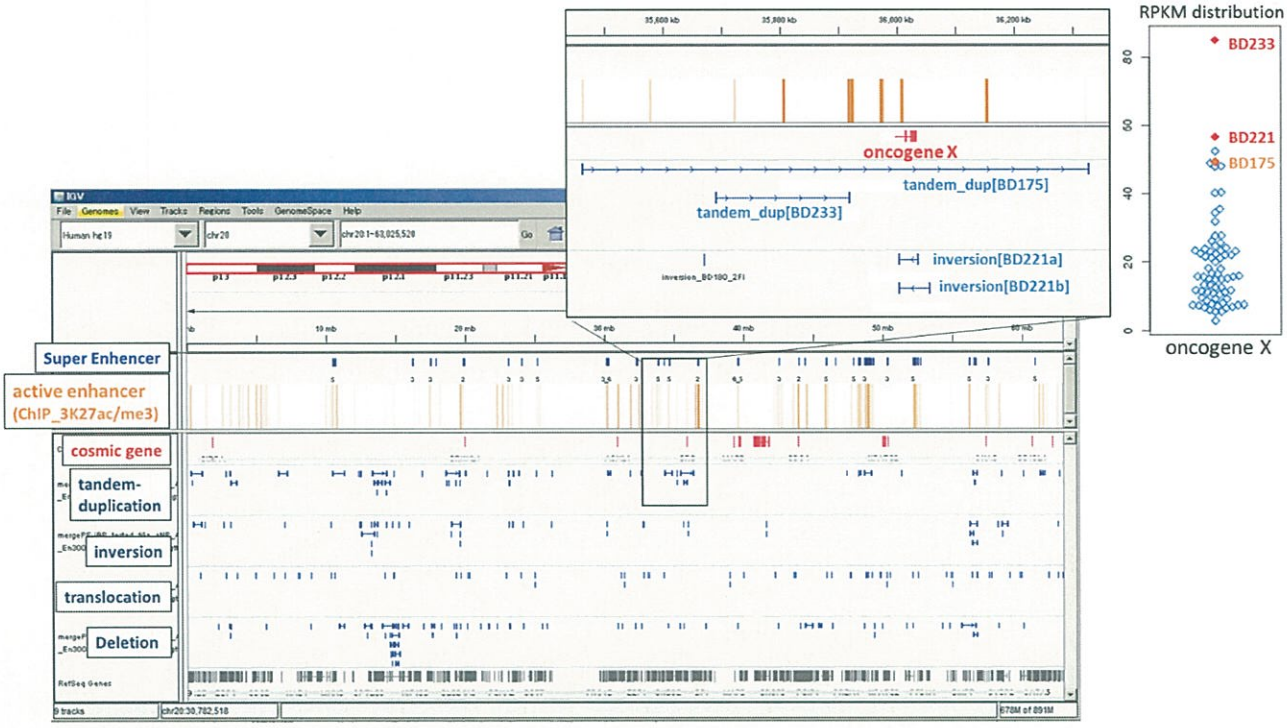


[図 3-2] :対象遺伝子における fpkm 値分布

赤い点が GR 群を、青い点が background 群をあらわす．(a)クラス[0]に分類された遺伝子における 61 症例の fpkm 分布．(b)同じく、クラス[3]に分類された遺伝子の fpkm 分布の例．

そこで、GRが遺伝子発現に影響を与える機序について検討するため、ローカルなゲノム構成についてIGVブラウザによる観察を行った例を図3-3に示した。この遺伝子は、胆管がんで上流遺伝子にGRが、起こり発現が上昇した遺伝子のクラス［3］として出力されたものの一つで、縦列重複型GRが入った症例と転位型GR症例によりGR群が定義されていた。しかし、ブラウザ上で近傍のGR分布を確かめるとこれらの他に遺伝子上流0.5Mbと下流0.3Mbをbreakpointとする縦列重複型GRが生じており、この症例もGR無し群に比べ十分高い発現量を示していた。今回構築したパイプラインにおいては遺伝子周辺の対象領域を0.1Mbとし、発生したGRイベントは考慮せずに使用したが、今後はIGVで遺伝子と周辺GRの分布を観察しつつ、対象領域の拡大やGRイベント型の限定等パラメーター調整を行うことが必要と考えられる。

一方で、全症例中での発生数は１症例にとどまるものの、GR症例で発現量が顕著に上昇するケースが複数のがん遺伝子でみとめられた（図3-2b）。これらはがん種において高頻度に発生する変異ではないものの、ゲノム不安定化によるGRの頻発そのものが、がん化のリスク向上の一因となることを示唆している。COSMICに登録されたがん遺伝子以外からも、GRにより発現量が変動する遺伝子候補が胆管がんで44（発現量増28，減16）、胃がんでは162（発現量増133、減29）出力された（表3-1<1>，<2>）。これらの遺伝子群の中には、未知のがん遺伝子や、がん化シグナルキャスケードの構成分子が含まれる可能性が考えられるため、IGVによるローカルゲノム構成の観察と並行して、遺伝子の機能や相互作用遺伝子の発現量について検討を行うことを予定している。



[図 3-3]：がん遺伝子 X / GR / enhancer の IGV ブラウザによる表示例

5. 結び

(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

本研究計画は、「全ゲノムデータと複数種のエピゲノム異常データを統合した自動解析パイプラインの構築と、これを用いた大規模症例の効率的な解析による新規発がん機序の解明」を目標に掲げ、平成27年度から平成29年度までの3年間実施された。この研究期間中、自動解析パイプラインを構築し、これにより5つのがん種（乳がん/胃がん/肝臓がん/小児脳腫瘍/胆管がん）計278症例のGR検出を完了し、それぞれ論文発表済みまたは準備中である。このほかに現在112症例が解析中であり、その後も複数種のがんで本パイプラインを使用した解析が予定されている。これらの中には公的データベースから入手した平均depth>100に及ぶ大規模症例のWGSの解析も含まれており、パイプライン構築時に想定した平均depth30～40のデータとは桁違いの計算機負荷が必要となったが、自家開発パイプラインならではでのアルゴリズムの透明性、パラメーター調整可能域の広さを利用して、効率よく解析を実施することができた。

エピゲノムデータについては、研究期間内にRNAseq、H3K27ac/H3K27me3-ChIP-seqの2種類を統合解析に取り込むことができた。これにより、がん化のドライバーとなりうるGRとその対象遺伝子についての組み合わせとして、胆管がんで118個、胃がんで211個の候補が得られた。一方で、平成28年度に追加実装したInsulator Neighborhood(IN)破綻による遺伝子の発現異常の検出パイプラインについては、予定していた抗cohesin ChIP-seqのデータが取れなかったため適用は先送りとなった。また、研究計画の中で予定していたTransposable Element活性化症例でのトランスクリプトーム解析については、WGSで高いTE活性が検出された症例でRNAseqデータが生成できなかったため、これもデータが出揃うまで保留中となっている。

本研究期間の3年間をふりかえっても、内外のプロジェクトにより生成された大規模症例WGSデータやエピゲノムが次々と公的データベースに蓄積され、これらのプロジェクトから各種がん発生・進展に関する新たな知見が報告されてきた。今後は自家パイプラインとしての柔軟性を活かしてこれらの既知公開データや知見を積極的に解析にとりこみ、十分な量のデータ解析結果をもとに新たながん治療のターゲットを提案することを目標とする。同時に、パイプライン内部ツールの計算機コスト最適化を進め、大規模WGS国際プロジェクトにおいても解析スピードで遅れをとることが無いよう解析高速化を目指す。