

平成28年度シニア・リサーチフェロー

研究成果報告書

平成 29 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



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 28 年 4 月 1 日
至 平成 29 年 3 月 31 日

研究指導者：氏名 柴田 龍弘



公益財団法人 がん研究振興財団

要旨

全ゲノム配列解析(WGS)技術の進展とデータ生成の低廉化により、近年では十分な症例数のがん WGS データを使った解析が可能となった。これらの WGS 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、新たながん化機序が次々に明らかにされている。H28 年度では、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、全く新しい機序の報告が相次いだ。また、より直接的に細胞の性質を反映する遺伝子発現や発現制御等の情報についても全ゲノム規模のデータが得られるようになり、SRA や GEO 等の公共データベースに膨大な量の蓄積がなされている。これらのゲノムデータは、細胞のがん化に決定的な役割を果たすドライバー変異や、がんが悪性化するメカニズムの全貌を解明しうるポテンシャルを有しており、その情報を最大限活用するためには、これらのゲノムデータを統合した解析が必要であると考ええる。

そこで我々は、WGS データとエピゲノムデータを統合して扱う解析パイプラインの開発を研究計画の主題に置き、これを用いた大規模症例データの解析から新しい発がんの分子機序の解明を目指している。研究計画の 2 年目にあたる H28 年度では、初年度に構築した WGS データ解析パイプラインにより、胃がん 81 症例 WGS データの Genome Rearrangement(GR)変異を検出した。さらに、前述の 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追加実装した。

Insulator Neighborhood 異常検出パイプラインは、GR が集中する IN と、GR で破綻する IN に含まれる遺伝子の発現変化を調べるもので、これにより 13908 個の IN 構造、胆道がん WGS から検出した 9764 個の GR、胆道上皮細胞株から得た H3K27ac-ChIP デー

タ、及び胆道がん 73 症例の WGS データと/RNAseq データ対を統合した解析を行った。その結果、GR による IN 構造の破綻にともなって IN 内部遺伝子の発現が変化する、5 つの IN が抽出された。このうちの 1 つの IN では、内包する 3 遺伝子全てで発現量が有意に変化した。胆道上皮細胞の 3K27ac-ChIP データにより、この IN 領域の上流 10-20kb に強いエンハンサークラスターがあることが確認されており、これらのエンハンサーが IN 内部の遺伝子に作用したと考えられる。

3'UTR 解析についてはパイロット版を完成させ、胆道がん 182 症例の RNAseq データを使用して、3' UTR の分解シグナル配列の領域に生じた GR により発現量(細胞中での蓄積量)が増加した転写産物の検出を試みた。これはエキソンごとに発現量(RPKM 値)を求めた後、転写産物の発現量を近似する代表エキソンの高発現に反して、3'UTR の RPKM 値が極端に落ちる転写産物を抽出するもので、胆道がんデータの解析の結果 106 症例から 188 個の遺伝子(COSMIC 遺伝子 7 個を含む)に由来する異常 mRNA の高発現が抽出された。

胃がんについても胆道がんと同様に WGS データの症例に対応する RNAseq データの解析が現在進められており、データが揃い次第、上に挙げたゲノムデータの統合解析を行う予定である。H28 年度はその準備である GR 検出までをすませ、発生頻度やその hot spot、hot spot ごとに特徴的な GR イベントの偏りなどについて検討した。これらの傾向は昨年度検出した胆道がんの GR とは異なっており、細胞種によるエピゲノム状態の違いやがん種による変異選択(がんの進化)の違いなどが GR のパターンに反映したと考えられる。

目次

1. 序文

2. 胃がん81症例WGSゲノムデータからの Genome Rearrangement(GR) 変異の検出

2-1 序

2-2 方法

2-3 結果及び考察

3. Insulated neighborhood破綻による遺伝子の発現異常

2-1 序

2-2 方法

2-3 結果及び考察

4. 3'UTR 欠落による遺伝子の発現異常

2-1 序

2-2 方法

2-3 結果及び考察

5. 結び

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

略語:

WGS : Whole Genome Sequence

GR : Genome Rearrangement

BP : Break Point

IN : Insulator Neighborhood

1. 序文

全ゲノム解読(WGS)技術の登場により、がんゲノムにおける染色体構造異常を含めた様々なゲノム変化の蓄積が明らかにされてきた。さらに、ゲノム全領域に渡るメチル化状態の低下やヒストン修飾異常など、ゲノムワイドなエピゲノム状態の変化についても、解析技術の目覚ましい向上により多くの報告がなされている。大規模WGSデータ解析により、従来のコーディング領域を中心とした解析では発がんの機序が不明であったがんについてもドライバー変異/遺伝子が同定され、新たな治療標的が得られることが期待されている。2014年、Northcottらにより報告された髄芽腫のEnhancer hijackingによるGF1活性化(Northcott., Nature 2014)、2016年にHniszらが報告したinsulated neighborhood(IN)の破綻による急性リンパ芽球性白血病のTall1活性化(Hnisz et al., Science 2016)などは、WGSデータがあつて初めて解明された新規発がん機構の代表例である。

本研究計画の研究目的の一つとして、全ゲノム解読(WGS)データ自動解析パイプラインの構築を進めている。既存パイプラインを使わず、自家開発したパイプラインを持つことの強みは、上述のような新しいがん化機序が発見されたとき、速やかに解析パイプラインに組み込むことができる点にある。そこで、研究計画の2年目にあたる平成28年度においては、insulated neighborhood構造を仮定し、genome rearrangement (GR)変異がゲノム中の遺伝子発現制御機構を破綻させる可能性を、WGS・RNAseq・ChIP-seqデータを統合して評価するパイプラインを開発し、前年度に構築した自動解析パイプラインに追加実装した。さらに、昨年6月、片岡らにより報告された、GRによる遺伝子の3'UTR欠損が原因となるPD-L1の異常発現機構(Kataoka et al., Nature 2016)についても、早速この機構から生じる発がんを解析対象とすべく、解析パイプラインの開発・実装を行い、胆管がん81症例のWGSデータ、及び182症例のRNAseqデータを解析した。また、新たに生成された胃がん81症例のWGSデータを解析し、胃がんに特徴的なゲノム異常の傾向を見出した。

2. 胃がん 81 症例 WGS ゲノムデータからの Genome Rearrangement(GR)変異の検出

2-1 序

胃がんは本邦において罹患者数の多いがんのひとつである。基礎研究の進展や早期診断技術の向上によって5年生存率は2007年には74.4%に上昇した。その一方で、高齢化を反映し、胃がんの罹患者数は未だに増加傾向にある。

胃がんゲノムの特性として、点変異が高度に蓄積されることが報告されている(Lawrence et al., Nature 2013)。より規模の大きい変異であるGRにおいてもその傾向は維持されるのか、Transposal Element (TE)活性化やCommon Fragile Site (CFS)の観点からも検討を行った。

2-2 方法

本研究計画の初年度で開発した自家パイプラインにより81症例の胃がん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、(表2-1)。これは、paired read sequencingから得られたimpropered Paired End リード(paired read間の方向と距離が異常)により検出されたGRのゲノム上の領域と、そのイベント型(欠失/逆位/重複/転座/転移)について出力するものである。

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	
GC002	1.47	2.16	207	GC118	1.58	2.32	177	GC334	1.70	2.19	41	GC437	1.57	2.19	28
GC003	1.48	2.08	94	GC121	1.47	1.95	181	GC339	1.81	2.34	294	GC440	2.06	3.27	578
GC004	1.56	2.16	141	GC124	1.79	2.29	214	GC340	1.51	1.98	179	GC443	1.74	2.98	43
GC029	1.87	2.04	358	GC125	1.75	2.23	187	GC352	1.84	2.31	101	GC448	1.78	2.76	303
GC035	1.61	2.09	34	GC129	1.82	2.07	213	GC354	1.68	2.26	687	GC449	2.16	3.60	56
GC036	1.84	2.03	736	GC130	1.45	2.36	375	GC361	1.54	2.29	31	GC450	2.26	3.89	70
GC038	1.51	2.34	69	GC132	1.51	2.01	191	GC364	1.57	2.30	127	GC451	2.19	3.58	402
GC042	1.55	2.35	84	GC133	1.75	1.99	177	GC365	1.69	2.02	88	GC452	2.17	3.57	60
GC045	1.48	2.28	267	GC136	1.77	1.95	263	GC366	1.76	2.24	943	GC453	2.25	4.06	138
GC050	1.52	2.05	42	GC138	1.76	2.39	87	GC367	1.71	2.24	276	GC454	2.22	3.56	144
GC051	1.78	2.39	2	GC143	1.80	2.30	781	GC370	1.55	2.31	463	GC455	2.18	3.76	28
GC059	1.83	2.26	723	GC146	1.54	1.97	408	GC371	1.70	2.24	1423	GC458	2.15	3.68	53
GC061	1.49	2.27	71	GC152	1.55	2.11	114	GC377	1.46	2.28	53	GC459	2.47	4.25	165
GC062	1.49	2.31	478	GC158	1.77	2.09	221	GC379	1.72	2.21	76	GC461	2.09	3.34	984
GC077	1.63	2.34	341	GC162	1.59	2.00	72	GC383	1.71	2.19	157	GC463	2.02	3.27	242
GC096	1.59	2.26	771	GC164	1.75	1.99	395	GC385	1.59	2.36	52	GC467	1.53	2.07	81
GC101	1.88	2.04	1	GC202	1.54	2.14	139	GC396	1.66	2.25	219	GC469	2.15	3.12	86
GC104	1.62	1.97	222	GC301a	1.53	2.20	465	GC401	1.61	2.10	279	GC470	1.89	2.95	281
GC108	1.60	2.22	128	GC301b	1.53	2.05	93	GC404	1.64	2.04	66				
GC115	1.50	1.95	424	GC320	1.50	2.05	875	GC416	1.84	2.03	156				
GC116	1.50	2.33	353	GC331	1.45	1.97	96	GC417	1.69	2.26	134				

表 2-1. GR 検出に使用された胃がん WGS データ

2-3 結果及び考察

自家開発パイプラインを用いて、胃がん81症例のWGSseqデータからGR変異の検出を行った結果、20,857箇所のGRが検出された(図2-1)。これは前年度解析した胆道がんGRの2倍近い検出数であり、体細胞点変異と同様、胃がんゲノムの高度な変異蓄積を反映したと考えられる。症例ごとに見ると、最もGR頻度が高い症例(GC371)では1423のGRが検出された一方で、最もGR頻度が低い症例(GC101)では1個のGRが検出されるにとどまり、胆道がんと同様にGR変異頻度は症例ごとに大きく異なった。

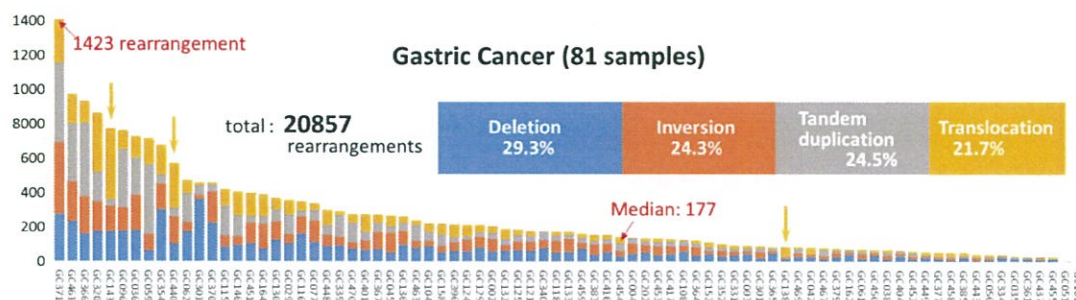


図 2-1. 胃がん WGS81 症例から検出された GR の概観

次にGRの4つのイベント型(欠失/重複/転座/転移)に注目すると、全症例での検出結果をマージした20,857GRにはイベント型に偏りは無いが、症例ごとに発生するGRイベントの頻度には偏りが見られ、GC143、GC440、GC138などの症例ではゲノムに生じたGRの4割以上がTransposition型で占められた(図2-1、黄矢印)。これらの症例においては、複数の領域で活性化Transposable Element(TE)が観察された。これらの症例以外でも、発生したGRの3割以上が転移で占められる症例が全症例数の65%(53/81)を占めた。これは胆道がんでの数値(33/88=37%)に比べても大きい数値であり、高い変異率と同様に、胃がんの特徴的な傾向といえる。

さらに、特にGRが集中して発生し、15症例以上でGRが観察される上位20のhotSpotにおいて、発生するGRイベントの頻度を調べたところ、領域によるGRイベントの頻度の差異が

Hot spot		Event type of Genome Rearrangements (GR)								total	note
chr(region)		deletion		inversion		tandem duplication		translocation		No.GR	
chr3	60M	246	84%	16	5%	25	9%	5	2%	292	←FHIT
chr14	58M	5	2%	16	6%	3	1%	224	90%	248	
chr19	30M	46	19%	129	52%	45	18%	28	11%	248	
chr22	29M	10	4%	26	11%	18	8%	174	76%	228	
chr20	14M	185	88%	5	2%	16	8%	4	2%	210	←FHIT ←WWOX
chr3	59M	154	78%	14	7%	28	14%	1	1%	197	
chr16	78M	173	89%	4	2%	17	9%	1	1%	195	
chr17	37M	20	14%	51	35%	42	29%	33	23%	146	
chr6	2M	77	74%	3	3%	8	8%	16	15%	104	
chr6	162M	83	89%	4	4%	4	4%	2	2%	93	
chr17	38M	9	10%	31	33%	28	30%	25	27%	93	
chr8	73M	7	8%	13	14%	5	5%	67	73%	92	
chr7	110M	76	84%	0	0%	7	8%	8	9%	91	
chr1	246M	57	67%	10	12%	11	13%	7	8%	85	
chr7	111M	57	70%	3	4%	21	26%	1	1%	82	
chr10	75M	13	16%	34	41%	21	26%	14	17%	82	
chr6	43M	15	19%	13	16%	37	46%	15	19%	80	
chrX	7M	59	76%	10	13%	7	9%	2	3%	78	
chrX	11M	12	16%	11	14%	5	6%	49	64%	77	
chr9	9M	53	72%	9	12%	2	3%	10	14%	74	

表 2-2. 上位 20 hotspot における GR イベントの頻度

明確に存在した(表2-2)。このうち、Transpositionが集中する上位4領域で、この領域がどこに転移されたかをcircos plotで可視化して観察したところ、限定された領域がゲノム全域に渡って転移されており、この領域内のTEの強い活性化が示唆された(図2-2上段)。胆管がんでは特徴的であったchr22の29Mb近傍の顕著なTE活性化は胃がんにおいても認められたが、その他のTE活性化領域は異なっており、がん種(細胞種)により異なるエピゲノム状態を反映していると推測される。

一方、欠失型が集中する2領域にはFHIT遺伝子、WWOX遺伝子がそれぞれ含まれており、これらの遺伝子領域内に多数のbreak point(BP)が認められたが、胆管がんゲノムのFGFR2-BICCで見られたBP分布のように、特定のイントロンへの集中は観察されなかった(図2-2下段)。これらの領域はCommon Fragile Site(CFS)として報告されていることから(Barlow et al., Cell 2013)、ゲノム不安定化に伴いCFSの脆弱性が顕著になった可能性が考えられる。

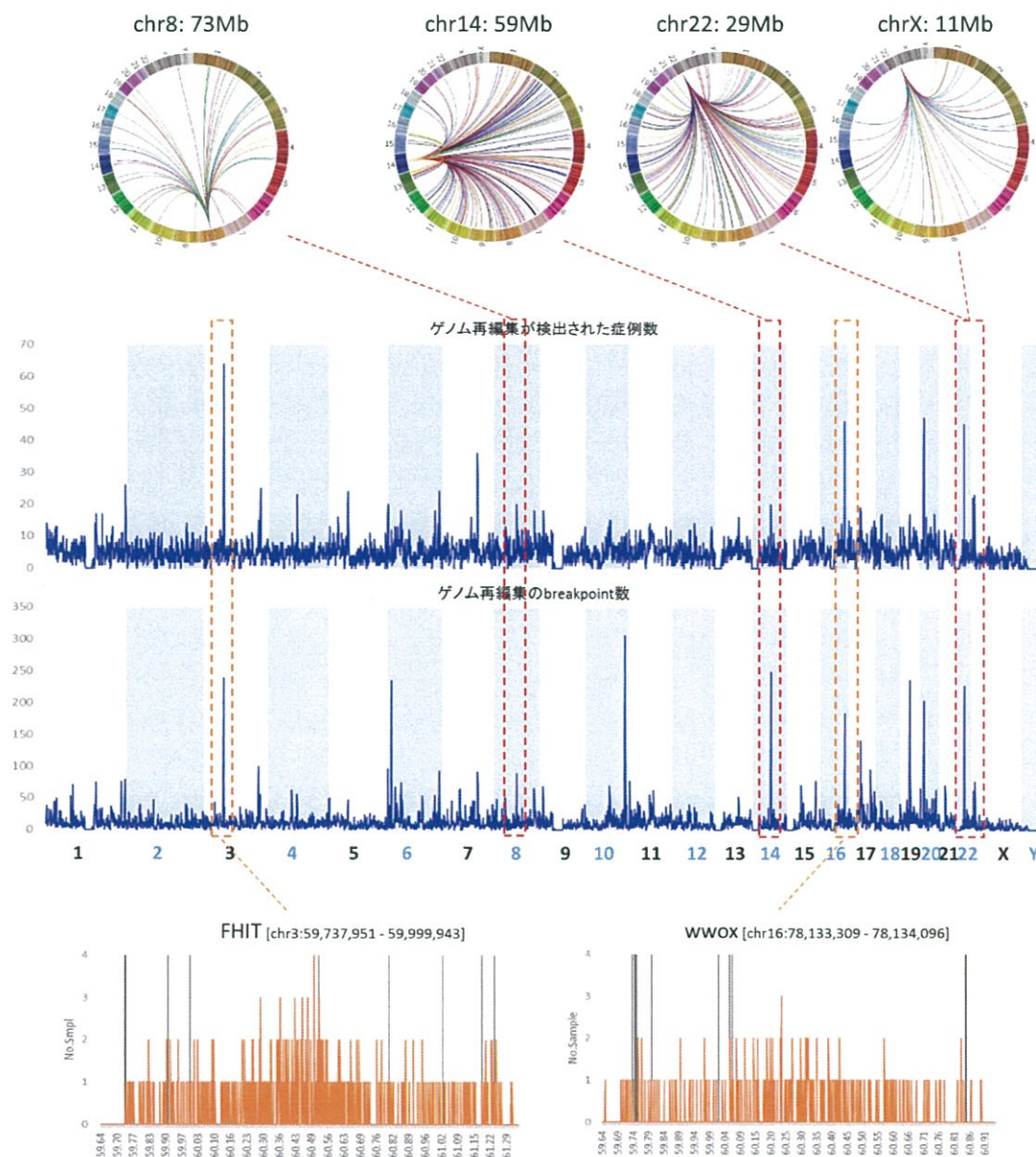


図 2-2. 胃がん 81 症例での GR hotSpot

下図で、横軸はゲノム座標、縦軸は Break Point(BP)の個数として、橙色の棒グラフで当該のゲノム領域から検出された BP を示した。灰色で塗りつぶされた領域は遺伝子のエキソン領域にあたる。

がん細胞においては、ほぼ例外なくゲノム不安定化が亢進することが知られているが、GR変異が生じる領域や好まれるGRイベントの傾向は、胆管がんゲノムと胃がんゲノムで大き

く異なった。これは、染色体局所構造(染色体フォールディングやヌクレオソーム構造等)や、CpGメチル化などの細胞に固有のエピゲノム状態を反映した結果であり、がんの進行により新たに生じる変異の予測に応用が可能と考えられる。

3. Insulated neighborhood破壊による遺伝子の発現異常

3-1 序

がん遺伝子そのものではなく、その制御領域に生じた変異によっても細胞のがん化は起こりうる。しかしながら、発現を制御する因子は制御領域のメチル化、ヒストン修飾、転写因子の結合、Insulator Neighborhoodに代表される染色体ループ構造、染色体間距離など多元的に存在する。さらにこれらのエピゲノム状態は細胞種・発達段階ごとに異なるため、対象とする細胞種ごとに実験的に決定する必要があり、制御領域上の変異評価を困難なものにしている。しかしながら、近年十分な症例数をそろえたWGSデータ解析により、GRが制御機構のかく乱を引き起こすメカニズムについても徐々に解明がなされるようになった。2016年、Hniszらにより報告された、ゲノムにおける局所的なループ構造の崩壊ががん遺伝子(Tal1)の異常発現を引き起こすケースはその好例である(Hnisz et al., Science 2016)。このループ構造はInsulated Neighborhood構造と呼ばれ、CTCF-コヒーシン蛋白質の結合によりゲノム上に100Kb~300Kb程度のサイズで形成されるもので、この構造により、ループ内部の遺伝子が外部の強力なエンハンサーの作用から隔離されると考えられている(図3-1右上)。Hniszらは、一部の急性リンパ芽球性白血病患者のゲノムにおいて、GR変異によりTal1遺伝子を含む大きさ約72kbのINループ構造が壊れていることを見出し、IN構造により隔離されていたSTIL遺伝子の制御因子がIN構造内部にあったTal1遺伝子の異常発現を促進することにより、がん化に至ることを示した。

このようなIN構造が形成される領域は、細胞種や発達段階によって異なるため、実験的

にCTCFとコヒーシンの共結合する領域を決定することが望ましい。しかしながら、コヒーシン結合領域の決定は汎用的に使えるコヒーシン抗体が開発されていないため難易度が高い。さらに、共結合部位の実験データが得られたとしても、隣りあった共結合部位を使用しない複雑なINループ構造(ループ内ループ構造など)の存在も報告されているため、ゲノム上のどの2点が結合点となってINループ領域が形成されるかは決定されない(Northcott., Nature 2014)。一方で、CTCF結合領域については公共データベース上において十分な量のデータが公開されている上、CTCFの結合モチーフについても研究が進んでいるため、信頼度の高い予測が可能である。

そこで、任意のIN領域リストとGR変異リスト、及び発現量データを入力に使用して、GR変異によるIN構造の破綻と発現量の変化の相関について解析を行うパイプラインを開発した。このパイプラインに公開データから得たconstitutive-IN領域リストを与え、WGSデータとRNAseqデータの両方が揃う胆管がん73症例の解析を行った。

3-2 方法

本パイプラインはGR領域のリストと任意のINのリストを入力とし、1)INごとのGR密度、2)IN破綻領域における遺伝子発現量の変化についての解析結果を出力する。INループ領域はINの開始/終了の2点のゲノム座標で定義されるものであれば任意に指定が可能であるが、ここではSRAから公開されている3種の血液系細胞(GM12878-リンパ芽球細胞, Jurkat細胞, K562-CML細胞)に共通する13908のIN領域を、細胞に共通して恒常的に使用される(=constitutiveな)IN領域とみなしてINリストに使用した。このリスト中のINループのサイズ分布は、最小のもので6577bp、最大のものでは3,302,723bpで、平均サイズは211,800bp(標準偏差204,202)であった。

はじめに、GR-BPをINループ領域上にプロットし、リスト中の各IN構造を破綻させる可能性を持つGR-BPについて計数した。ここで解析対象とした”IN構造を破綻させうる”GRとは

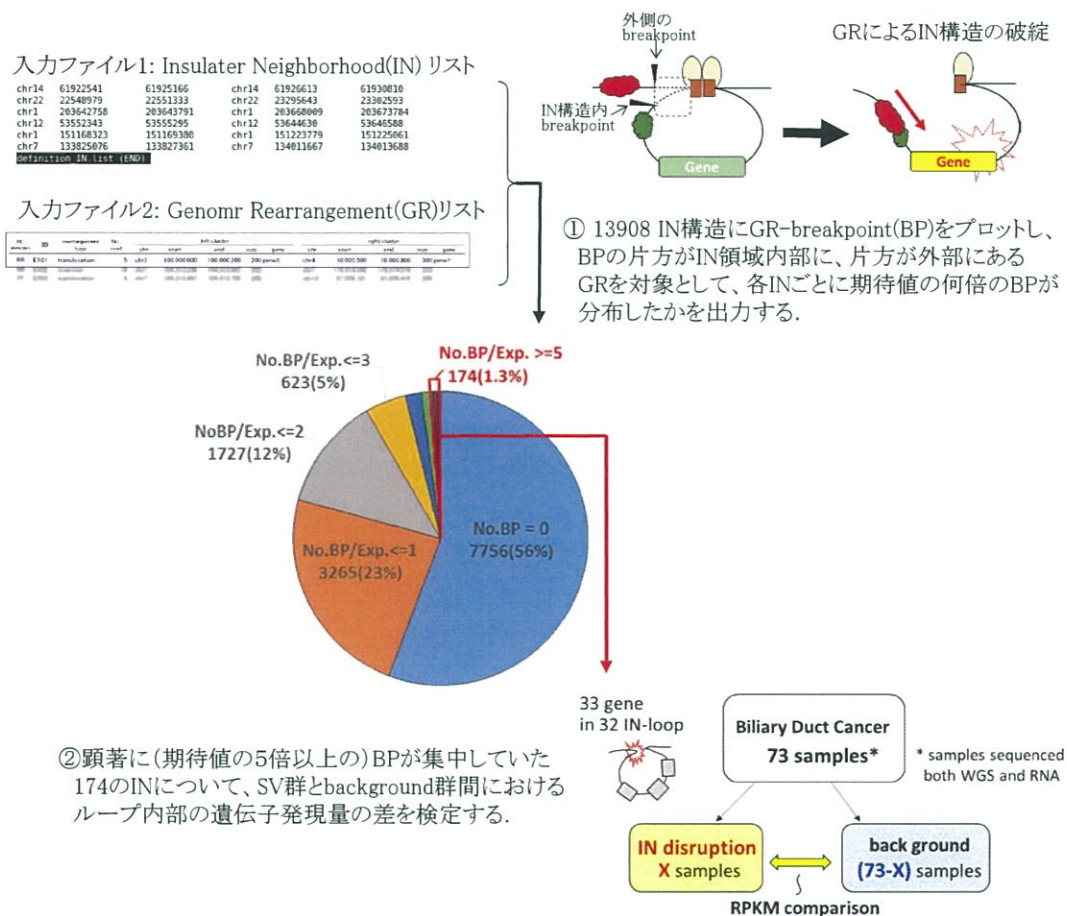


図 3-1. IN 構造を破綻させるゲノム異常解析の流れ

片側のBPがIN領域内に、もう一方が外部にあるGRをさす。対象となった7006GR(14012 BP)について、各INループ構造ごとに、IN領域の大きさから期待されるBPの出現数と、実際に観察されたBP数の比をとり、特にGR変異が集中したINループ領域を選定した。ここで、IN中のBPの期待値は、以下の式によって求めたものを使用した。

期待値=(全症例のbreakpoint数の合計 / ゲノムの解析対象領域[※])×Insulated neighborhoodのサイズ(bp)

※ 全ゲノム領域から、telomere, centromere, heterochromatin, sequence gap等を除いた数値

以上の操作により、特にGRの集積が認められたIN領域に含まれる遺伝子については、INの破綻と発現量の変化に相関がみられるかを検討した。すなわち、全症例を、IN結合部の前後にGRが生じた(=INが破綻)症例群と、INが維持されるback ground群に分け、二群間の遺伝子発現量の差をstudent-T testにより検定した。

また、今回解析対象とした胆管がんゲノムにおける活性化エンハンサー領域の推定のため、胆道上皮細胞株でH3K27ac抗体のChIPseqを行った(図3-2下段)。得られたChIPseqデータを、MACS2を用いて解析した結果、計69027のH3K27acピーク(d=211)が検出された。

3-3 結果と考察

WGSデータとRNAseqデータの揃う胆管がん73症例を使って、3種の血液系細胞に共通するIN構造をconstitutve-IN構造と仮定し、GRによるIN構造の破綻ががん化を促す可能性について検討した。BPの片側がINループの内側に、もう一端がループの外部にあるGRを対象とした結果、9764のGRの72%にあたる7006のGRが解析対象となった。GR変異が期待値を超えて集積したINループは、全体の21%にあたる2887個見出された。

IN ID	gene	IN-disrupted group(IN)			background group(b.g.)			p [ks-test]	RPKM ratio (IN/b.g.)
		N	ave.RPKM	σ	N	ave.RPKM	σ		
IN259	geneA	4	155.6	31.3	69	118.6	56.4	0.037	1.31
	geneB	4	55.2	17.9	69	40.6	23.8	0.236	1.36
IN348	geneC	4	0.4	0.3	69	0.3	1.0	0.574	1.12
	geneD	4	0.8	1.0	69	0.8	1.8	0.875	1.00
	geneE	4	2.8	1.1	69	1.7	1.2	0.037	1.62
IN353	geneF	4	171.2	99.4	69	96.9	87.7	0.118	1.77
	geneG	4	190.6	108.5	69	132.3	99.7	0.093	1.44
	geneH	4	4.4	3.2	69	4.1	4.4	0.778	1.09
	geneI	4	20.8	5.1	69	12.2	6.4	0.017	1.70
	geneJ	4	133.0	26.2	69	88.3	55.2	0.014	1.51
IN354	geneF	3	195.6	106.0	70	96.9	87.1	0.173	2.02
	geneG	3	191.2	132.9	70	133.1	99.2	0.340	1.44
	geneK	3	137.6	30.1	70	88.8	55.0	0.050	1.55
IN13619	geneL	3	106.8	52.1	70	431.3	292.1	0.009	0.25
	geneM	3	25.1	13.6	70	5.5	4.4	0.003	4.52
	geneN	3	16.0	6.8	70	4.6	3.4	0.001	3.46

表 3. IN 結合点を壊す GR により、内部の遺伝子発現が変化した IN

特に強くGRの集積が認められたIN構造、すなわち期待値の5倍以上のBPが3症例以上から観察された32個のINを対象に、GRによるIN構造の破綻と発現量の変化の相関を調べた(図3-1)。その結果、5個のIN構造中から、IN構造の破綻により発現量が変化する8遺伝子を見出した(表3)。このうちの7遺伝子で発現が上昇しており、その上位3遺伝子はいずれも他のがん種でがんとの関連が報告されているものであった。特に、ある染色体領域のINループにおいては、ループ中に含まれる遺伝子の全てが、有意に発現変化していた(図3-2a)。

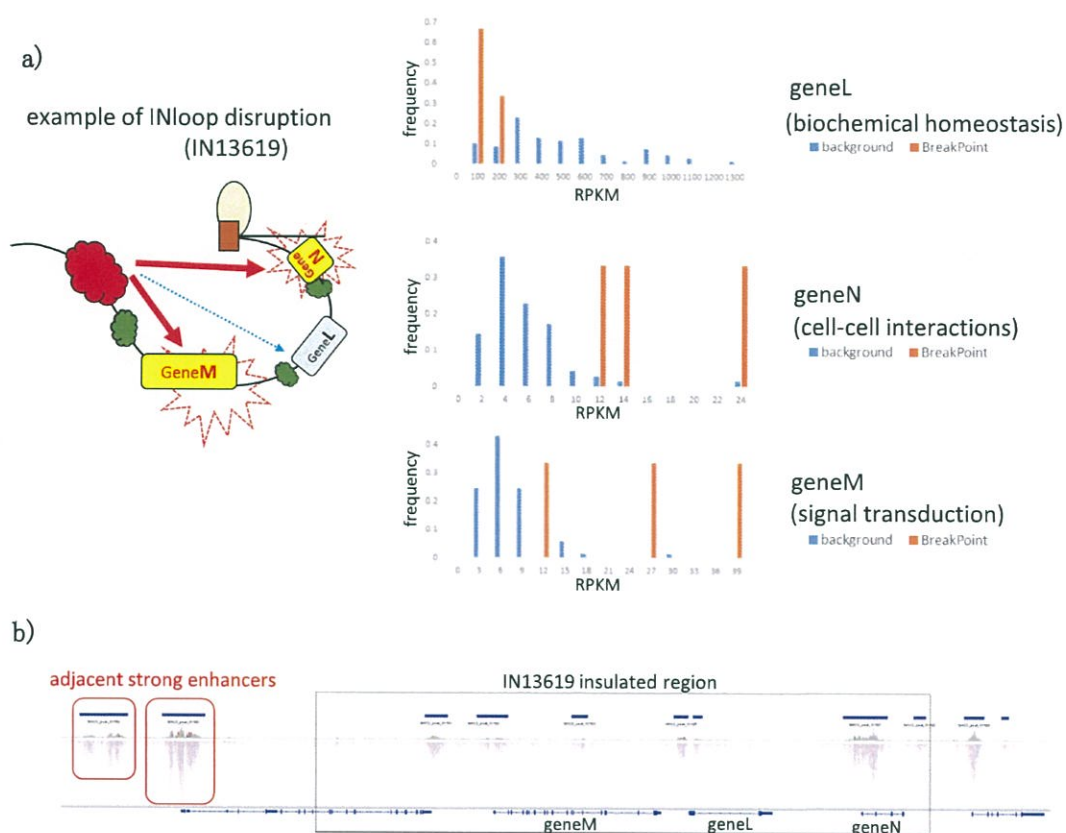


図 3-2. 破綻が予測される IN 構造の例

(a)GR により IN 構造が破綻したサンプル群(オレンジ)と、background 群(青)の遺伝子発現量(RPKM)分布を示した。縦軸はそれぞれの群における頻度を表す。(b)胆道上皮細胞における H3K27ac-ChIP の結果を示した。上段の青い bar は MACS2 により検出された H3K27ac ピークの領域を、中段はピークの高さ(リードのマッピング分布)を表す。

胆管上皮細胞からH3K27ac抗体を使ったChIPseq解析により、この領域の活性化エンハンサーを確認したところ、IN結合点の10~20Kb上流に、大きなH3K27acピークのクラスターがみつめられた(図3-2b)。INループ構造の破綻により、これらの強エンハンサーが遺伝子M、遺伝子Nに作用した結果、発現が上昇したと予測される。INループ構造は外部エンハンサーからの隔離機構なので、内部の遺伝子はそれぞれ別の制御機構下にあってもかまわない。ループ内に含まれる遺伝子の中で、一つだけ発現が下がった遺伝子Lは、他の2つの遺伝子に比べ桁違いに高い発現をしており、この遺伝子だけ別の制御機構下にあったと考えられる(図3-2a)。

胆道上皮細胞でのcohesin結合領域のデータは現在調整中であるが、このパイプラインにより、他の細胞種でのINデータや、潜在的なループ結合点であるCTCF結合部位同士の組み合わせなどを入力に使ったシミュレーションが容易となった。INリストとRNAseqデータ、WGSデータ、及びChIPデータを統合することで、制御機構を破綻させ、発がんドライバーとなりうるGR変異を絞り込むことが可能になると考えられる。

4. 3'UTR欠落による遺伝子の発現異常

4-1 序

2016年、片岡らにより解析された成人T細胞白血病リンパ腫の研究から、3'UTR配列中に含まれる分解シグナルの欠損により、細胞内にPD-L1転写産物が高度に蓄積し、これが発がんの契機となる機構が見出された(Kataoka et al., Nature 2016)。真核生物のmRNAの寿命は3'UTR配列上のシグナル配列等により調節されており、アミノ酸鎖への翻訳後直ちに分解をうけるmRNAがある一方、「急に」「大量に」必要になるような分子をコードするmRNAではそのまま細胞質中にとどまり、翻訳に複数回使用される。片岡らは3'UTRが欠損した症例の対応するWGSデータを解析し、PD-L1の第4エクソンの直後に生じたGRにより、3'UTRを含む3'側mRNAが欠損することを確認した。

PD-L1で見られたような、分解機構の破綻によるmRNAの蓄積が契機となる発がんは、他のがん関連遺伝子でも生じている可能性がある。そこで、検討対象をPD-L1遺伝子だけではなく、refseqに登録された遺伝子転写産物43766個に拡大し、3' UTRの分解シグナル配列に生じた変異により発現量(細胞中での蓄積量)が増加した転写産物を検出するパイプラインを構築した。

パイロット版にあたる本バージョンでは、3'UTR欠損mRNAの異常蓄積検出に特化し、片岡らがPD-L1異常転写産物の検出に使用した条件を参照しつつ、より汎用性がある検出条件の検討を行った。構築した異常mRNA検出パイプラインを使い、胆管がん細胞mRNA182症例の解析を行った結果、106症例の188遺伝子について、3'UTRが欠損したmRNAの異常蓄積が認められた。

4-2 方法

大規模な症例数のRNA sequencing (100bp paired read)から得られたmRNAデータを入力とし、有意に発現が高く、3'UTRが欠損した転写産物を出力するパイプラインを構築した。解析対象とする転写産物はrefGeneに記載のある全50129転写産物から、常染色体または性染色体上にのる、3個以上のエキソンで構成される遺伝子に由来する43766個の転写産物とした。

はじめに、mRNAをゲノム全長(hg19)にマップし、ここからPCR重複を除外したのち、intersectBedを使用して482,765エキソン(50129転写産物)上にマップされたリード数を計数した。これとhg19ゲノム全長にマップされたリードの総数からエキソンごとのRPKM値を算出した。これをもとに、各転写産物の発現量を近似する代表エキソンを、それぞれの転写産物で定義した。代表エキソンは、母集団中での平均RPKM値と、発現率(発現のある症例の割合)の積が最も高いエキソンとした。

この代表エキソンと3'UTRのRPKM値により、「3'UTRの発現量は低い」が代表エキソン

の発現量が上がっている」転写産物(遺伝子)と症例を抽出するために、以下の条件で絞り込みを行った。

(母データ: 182症例の全転写産物 7,904,079(43429*182))

- 0a) 代表エキソンのRPKM値が0.1を超える
- 0b) 線形回帰直線の傾き α が $0.75 < \alpha < 2.5$
- 1) 代表エキソンのRPKMのz-scoreが2.0を超える
- 2) RPKM比(3'UTR/代表エキソン)のz-scoreが-1.0[-0.5/-2.0]を下回る
- 3) RPKM比(代表エキソン/3'UTR) >3.5 かつ、代表エキソンのRPKM値 >2.0
:[Kataoka et al., Nature 2016]より引用

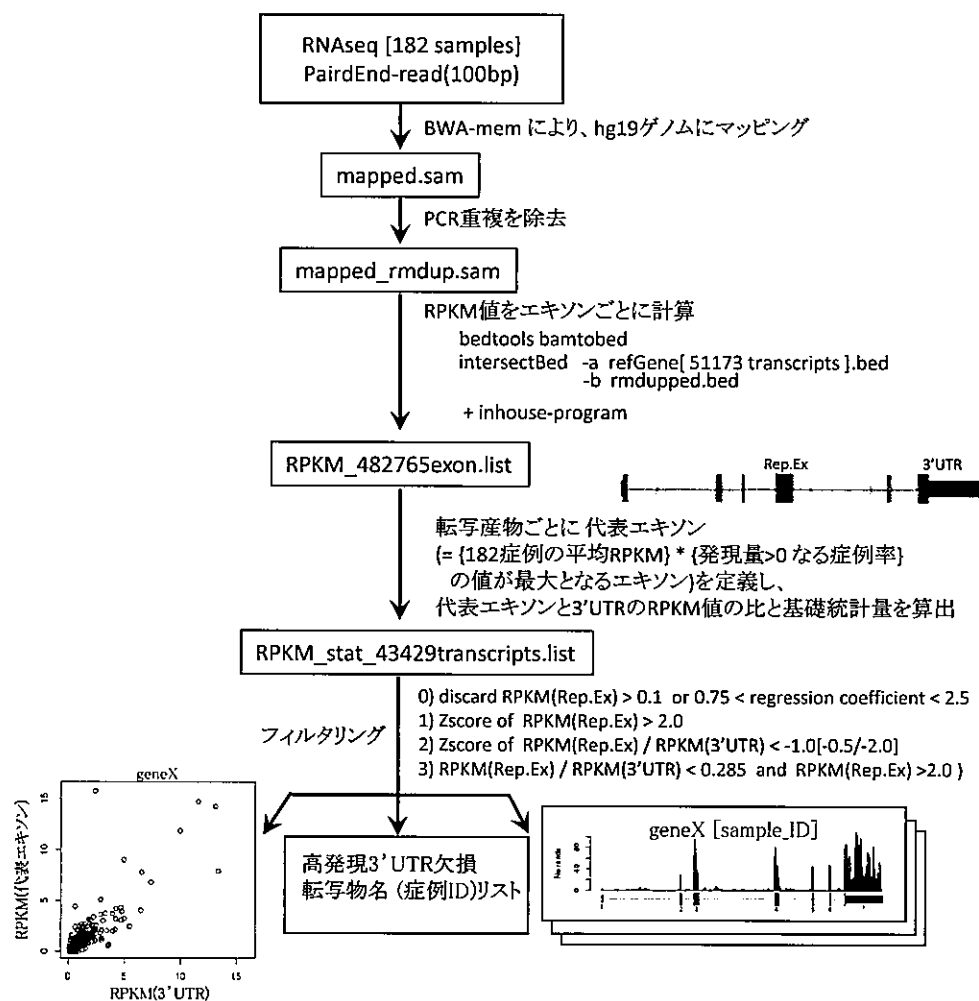


図 4-1. 3'UTR 欠落による異常発現遺伝子の検出の流れ

0a)及び0b)は信頼性の低いデータを排除するための条件であり、前者ではエクソンあたりにマップされるリード数が2未満である転写産物を、後者ではdegragateされるまでのスパンが短く、母集団の殆どの3'UTRが分解されている転写産物を除外している。以上の条件を満たすtranscriptを3'UTRが欠落して高発現するmRNAの候補として、症例ID及び母集団におけるz-scoreをつけたリストとして出力した。

また、パイプラインにはデータ可視化を支援する2つの出力オプションを実装した。1つは代表エクソンと3'UTRのRPKM散布図であり、遺伝子名(またはtranscriptID)を指定すると、母集団全ての対象transcriptについて横軸に3'UTRのRPKM値を、縦軸に代表エクソンのRPKM値をとった散布図を出力するものである(図4-2)。もう1つはエクソンごとのread-depthを可視化するもので、症例IDと遺伝子名(transcriptID)を指定すると、対象転写産物上にマップされたリード数を棒グラフとして出力する。これによりゲノム上のBPの候補の位置(GRが生じた場所)を推測することができる。

4-3 結果

成人T細胞白血病リンパ腫で報告されたPD-L1のように、3'UTRの欠損(=分解シグナルの機能不全)により細胞内に異常蓄積する転写産物を検出する目的で、パイロット版となるパイプラインを構築し、これにより胆管がん182症例のmRNAデータの解析を行った。本解析ではRefSeqデータベースに登録された3つ以上の構成エクソンを持つ43429転写産物を対象として、胆管がん182症例のmRNAseqデータから、3'UTRが欠落して発現量が上がる転写産物の候補を抽出した。

182症例の全7,904,079転写産物(43429転写産物×182症例)から、十分量のリードがマップされ、不安定な(3'端の分解が進行した)転写産物を除外した2,850,512の転写産物に対し、以下のフィルターにより候補転写産物を調整した。すなわち、(1)182症例の母集団において有意に発現が高く、(2)3'UTR/代表エクソンのRPKM比が抑えられ、(3)片岡らが異

No. transcripts			
original data	7904079		
filter_0	2850512		
filter_1	133947		
filter_2	44427	7447	3638
filter_3	1817	438	19
cosmic	123	11	2
	Z < -0.5	Z < -1.0	Z < -2.0
filter_2: Zscore threshold			

filter_0: discard RPKM(Rep.Ex) > 0.1 or 0.75 < regression coefficient < 2.5

filter_1: Zscore of RPKM(Rep.Ex) > 2.0

filter_2: Zscore of RPKM(Rep.Ex) / RPKM(3'UTR) < -1.0[-0.5/-2.0]

filter_3: RPKM(Rep.Ex) / RPKM(3'UTR) < 0.285 and RPKM(Rep.Ex) > 2.0 }

表 4. データスケールの推移

常構造PD-L1の検出基準に使用した「代表エクソン/3'UTRのRPKM比が3.5を超え、かつ代表エクソンのRPKM値が2以上となる」条件で絞り込みを行った。その結果、438の転写産物(188遺伝子、106症例)が、3'UTRの欠失を伴う高発現転写産物の候補として出力された(表4)。これらの候補転写産物には、COSMICに記載のあるがん関連遺伝子が11転写産物(7遺伝子、5症例)含まれていた。

図4-2に、パイプラインにより出力された転写産物の一例を示した。ここで例にあげた transcriptXは、7個のエクソンから構成される遺伝子の転写産物である。この転写産物の例では、ほとんどの症例で対象転写産物の代表エクソンと3'UTRの発現量比は 0.84 ± 0.5 の範囲におさまるが(図4-2, 散布図)、図中赤線で囲まれた領域にプロットされた2症例では、代表エクソンのRPKM値は高いにも関わらず、3'UTRのRPKM値は低く抑えられている。エクソンごとのリード深度を示した棒グラフでも、通常のRPKM比を持つ3症例(図4-2右下)では、グラフ右端の3'UTRに相当する領域にリードのピークが明確に観察されることに対し、図4-2左上に示した2症例では、第2エクソンに300本以上のリードがマップされるにも関わらず、第3エクソン以降でほとんどリードのピークが見られなくなった。この2症例では、transcriptXの

3'UTRの分解シグナルが何らかの変異により損なわれ、分解速度が落ちた結果、細胞内異常蓄積が生じた可能性がある。来年度、これらの解析結果を深く考察して、必要があれば条件検討を行う予定である。

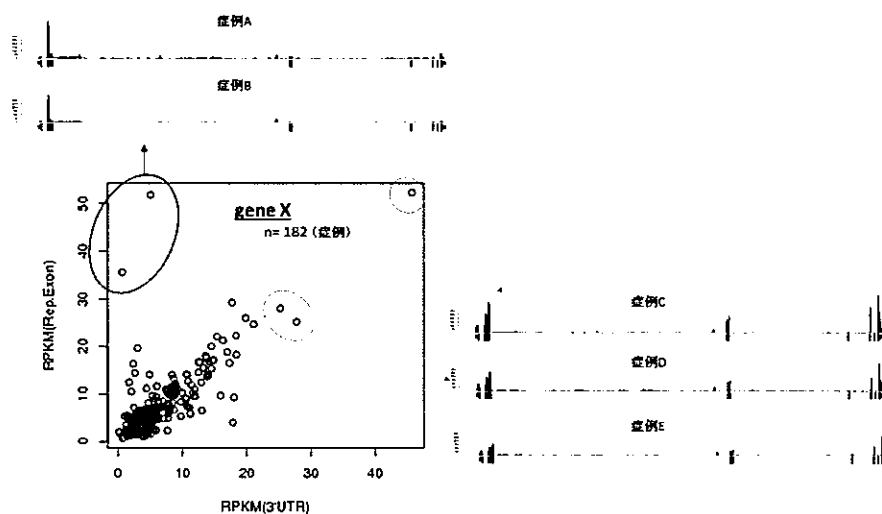


図 4-2. 異常発現遺伝子(転写産物)の抽出例

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

研究計画の 2 年目にあたる 28 年度では、新たに 2 つの発がん機序(Insulated neighborhood による発現制御機構の破綻、及び 3'UTR にコードされた分解シグナルの機能不全)を解析対象とするパイプラインを開発、実装し、これにより胆道がん症例の解析を行った。前者では胆管がん 73 症例の WGS データ及び RNAseq データ、胆管上皮細胞株の ChIP-seq(H3K27ac)データ、及び公開データである ChIP-seq(CTCF-cohesin)を統合した解析結果が出力されており、本研究計画の主眼であった「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」の実現に大きく近づいた。

今後は、現在準備が進められている胃がん WGS データと同一の症例から調整した RNAseq データの到着を待って、胃がんゲノムについても Enhancer Hijacking/Insulator Neighborhood 機構を想定した解析や、3'UTR 欠落転写産物の異常発現について解析を行う予定である。胃がんのエピゲノム情報については公共データベースに豊富な蓄積があり、これらのデータを積極的に利用して、ドライバー変異候補の絞り込みを行いたい。絞り込みのものは、データ解析から得られたドライバー変異候補を実験的に検証し、新たながん化機序の解明を目指すとともに、その検証結果をデータ解析にフィードバックすることで、解析パイプラインの精度・信頼性を向上させることを考えている。