研究論文

がんゲノム変異の解析　-　COSMIC および ClinVar について†

菅野知里*，田島実歩*，荒木和浩**，中村建介*

Mutations in cancer cells – Analysis of COSMIC and ClinVar databases

Chisato Kanno, Miho Tajima, Kazuhiro Araki and Kensuke Nakamura*

We here describe the characteristics of genome mutations found in cancer cells.
Key words : cancer, bioinformatics, next generation sequencer, oncogene, oncomutation

1　はじめに

日本人の三人に一人はがんで亡くなると言われており，死因として最も多いものとなっている．[1] 近年の分子生物学や測定技術のめざましい進歩により，がんに対する早期発見と治療法に関する状況は格段に進歩したが，依然として我々にとっての大きな脅威であり続けている．がんはヒト自身の細胞が変化し，無制限増殖，転移浸潤，血管新生，免疫系やアポトーシスによる制御の回避，悪液質と呼ばれる恒常的な炎症状態，などの能力を獲得し，人間自身の細胞に由来しながら，通常のヒトの細胞とは異なる別の生物の様に振る舞うことから，悪性新生物(neoplasm)とも呼ばれている．ここで起きる「変化」にはエピゲノム的な修飾状態の変化も含まれるが，その主要な要素はゲノムの変異であり，がん細胞のゲノム配列を解析することで種類を特定し，個々のがん細胞の治療に最適な指針を定めることができると考えられている．2005 年ごろから普及した次世代シーケンサー(NGS)はそれまでのサンガー法に比べて飛躍的な解析速度の向上をもたらした．これにより，個々のがん細胞についてそのゲノム配列を解析することも可能になりつつある．それに伴い，典型的ながんに関連する遺伝子（oncogene）とその変異（oncomutation）に関する情報が蓄積されている．がん細胞における変異はいくつかの観点から分類できる．ひとつ目の観点は，その変異が親から遺伝的に引き継がれた変異であるか，個体の中で発生した変異であるかである．図１に示す様に前者は親から受け継いだ卵細胞がそもそも持つ変異であり，生殖細胞系列（germline）変異と呼ばれる．後者は，個体の成長過程で分裂する細胞に生じる変異であり，体細胞（somatic）変異と呼ばれる．がん細胞における変異を分類する上でのもうひとつの観点は，その変異が何らかの形でがんの進行を促進するもの（ドライバー変異）か，がん細胞によく見られるが，進行には直接関連しないと考えられるもの（パッセンジャー変異）かの区別である．

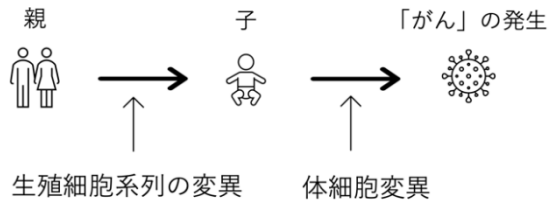

図 1.　遺伝性腫瘍における生殖細胞系列変異と体細胞変異

生殖細胞系列の変異は，その存在だけではがん化を決定づけるものではないが，後天的な変異の蓄積によって，がんになりやすいか否かの傾向を支配する．一方，我々がより関心を持っているのは，個体の成長過程において起きる体細胞変異のなかでも直接，がんの進行を促すと考えられるドライバー変異である．

がん細胞における変異を含む代表的なデータベースとして，ClinVar[2]と COSMIC [3]が挙げられる．ClinVar はがんに限らず様々な遺伝性疾患にみられる変異を含むデータベースであり，おもに「生殖細胞系列変異」に対して，個々の変異の疾患への影響の強さが経験に基づいて，病原性(pathogenic)あるいは良性(benign)に分類されている．一方，COSMIC はがん細胞にみられる「体細胞変異」を含む変異を網羅的に集めたものであり，どの様ながん細胞に見られたか，といった情報は含んでいるが，病原性についての記述は基本的には含まれない．

我々の研究室では一昨年より群馬県立がんセンターとの共同研究により，原発巣不明がんおよび希少がんのエキソームシーケンシング解析を行っており，これらの細胞でがん化に影響を与えるドライバー変異の特定を試みようとしている．個々の症例における変異の解析を進めるにあたり，既知のがん体細胞変異，特にドライバー変異の持つ特徴を把握しておくことが重要であると考え，

†　原稿受理 令和３年２月２６日　Received March　26，2021
*　生命情報学科（Department of Life Science and Informatics）
**　群馬県立がんセンター　腫瘍内科

ClinVar および COSMIC からこうした情報を抽出することを企図して様々な解析を進めており，本稿ではその一部を紹介する．

2　材料と方法

2・1　ClinVar

ClinVar データベースファイルは毎週更新されており今回は 2021 年 1 月時点での最新版である clinvar20210131.vcf をダウンロードして使用した．vcf フォーマットは 1 行に一種類の変異とその属性を記述したテキストファイルであり 78 万 9969 行からなる．このファイルから後述する「がん」とその関連疾患を記述するキーワードを含む行のみを抽出し，得られた 11 万 3375 件の変異に関する情報を解析の対象とした．

2・2　COSMIC

COSMIC では変異のタイプによりいくつかの利用可能なファイルが準備されている．今回はそれらのうち CosmicMutantExport.tsv というファイルをダウンロードした．2021 年 1 月の時点でこのファイルは 4722 万 2279 のエントリを含む．COSMIC の.tsv ファイルは一つのエントリ（行）に観測された一つの変異を含む．ClinVar とは異なり同一の変異（同じ位置で起きる同じ変異）が複数回（症例の数だけ）出現するため，同じ変異をまとめ情報を整理するプログラムを作成し，1190 万 4650 箇所の変異からなる cosmic3_org.out というファイルに変換し解析の対象とした．

2・3　解析

テキストファイルでダウンロードされた COSMIC, ClinVar データベースについて，C および python で記述したプログラムおよびスクリプトにより解析を行った．また，awk, grep, cut, sort, uniq 等の標準 Unix コマンドを適宜利用した．

3　結果

3・1　がん関連疾患を記述するキーワードについて

ClinVar には，がん以外の遺伝性疾患に関与する遺伝子変異の情報も多く含まれる．これらの中からがんに関連する疾患の遺伝子変異のみを抽出するために COSMIC に含まれるがんを記述する病名のエントリを整理し，キーワードの抽出を行った．その結果得られた 32 のキーワードを表 1 に示す．

表 1．がんおよびその関連疾患 を表す 31 のキーワード

adenoma	epithelioma	leiomyoma	neoplasm
blastoma	fibroma	leukemia	papilloma
cancer	fibromatosis	lipoma	pecoma
carcinoma	ganglioma	melanoma	sarcoma
chordoma	glioma	meningioma	schwannoma
cranioparingioma	haemangioma	mesothelioma	thyoma
cytoma	hyperplasia	myxoma	tumor
endothelioma	keratocyst	neoplasia	tumour

3・2　変異の種類と病原性について（ClinVar）

ClinVar データベースファイルより表 1 のキーワードを含むエントリ 11 万 3375 件を抽出し，それぞれの変異のタイプにより分類した結果を表 2 に示す．ClinVar に含まれるデータはほとんどが遺伝子内の変異であり，遺伝子外の変異もわずかに含まれているが表 2 では割愛した．表 2 で，網掛けをした 2 つのカラムのうち左側は，良性な場合と病原性の場合の比率を示したもので，コロン (:) の右側の数値が大きいほど病原性の比率が高い．また，網掛けをした右側のカラムは，病原性が不明な変異を含む，全体の変異数に対する，良性あるいは病原性が記述されているものの割合で，この数値が低ければ先ほどの比率による判定の信頼性が高くないと言える．

ClinVar の変異情報は，まずその起きる領域によって，アミノ酸配列を記述するコード領域（下半分）と，非コード領域（上半分）に分類される．非コード領域には，UTR, イントロン，翻訳されずに RNA として機能する領域等がある．これらの領域は一般に良性な変異が 4:1 から 10:1 のオーダーで大きく，がん化に影響する変異は比較的少ない．コード領域の前後に存在する UTR の変異はアミノ酸配列への翻訳を制御する機能があると考えられるが，それ以外の領域に比べて病原性が特に高いとは言えない．これら非コード領域の中で例外的に病原性の割合が高かった(40~80 倍)のが，スプライスドナーおよびアクセプターと呼ばれる領域であった（網掛けした上の 2 つの行）．これらはエキソンとの境界に近いイントロン領域にあって，スプライシングを制御する領域で，ここにある変異はスプライス異常を引き起こして，正常なタンパク質の発現を阻害するためであると思われる．コード領域の変異には，短い挿入欠失（INDEL）と一塩基置換(SNP)がある．表 2 ではインフレームインデルとフレームシフトの 2 つが INDEL によるもので，それ以外は基本的に SNP による変異である．インフレームインデルでは 3 の倍数の塩基が挿入・欠失することによりアミノ酸が数個増減するが，前後のアミノ酸配列は変化しないのに対し，フレームシフトでは 3 の倍数でない塩基が挿入・欠失するため下流のアミノ酸配列が大きく変化する．SNP 的変異のうち，非同義置換は一つのアミノ酸の種類が変わる局所的な変異であり，数個のアミノ酸が増減するインフレームインデルと影響が同程度と推定され，実際に同程度の病原性比率を持つ．また同義置換はアミノ酸の種類を変化させないため，非同義置換よりも病原性の比率が低い．同義置換により生じるアミノ酸配列は変異前と全く同じことから直感的には病原性を生じ得ない様に思えるが，RNA 干渉に対する感受性や，対応する tRNA の存在比率等により，病原性を生じるケースも多少存在すると考えられる．非同義置換については病原性・良性の区別がついていない変異が非常に多く，病原性の比率も極端に高いわけではないが，後述する COSMIC での数の多い変異でも見られる様に典型的なドライバー変異にも非同義置換が多く見られることから重要な変異パターンの一つである．

表２． ClinVar データの変異タイプと病原性

変異のタイプ		変異	良性 Benign	良性:病原性	病原性 Pathogenic	確定率 (%)	不明 Uncertain
非コード領域	非翻訳領域	3'-非翻訳領域(UTR)	532	11:1	50	36.7	1028
		5'-非翻訳領域(UTR)	285	4:1	73	37.1	617
		非コード転写領域	3077	13:1	253	86.6	631
	イントロン	イントロン内	5088	4:1	1165	76.7	2387
		スプライスドナー	26	1:39	1016	89.2	129
		スプライスアクセプター	14	1:79	1100	85.8	190
コード領域	アミノ酸配列不変	同義置換	14269	13:1	1087	95.7	705
	局所的アミノ酸配列変化	非同義置換	1643	1:4	6234	16.7	47866
		インフレームインデル	26	1:7	182	21.3	869
	アミノ酸配列の長さを変化させる	開始コドン変位	4	1:33	133	56.1	114
		終止コドンの消失	12	4:1	3	32.0	51
		終止コドンへの変異	21	1:226	4755	99.4	311
		フレームシフト	13	1:677	8804	99.4	604

表２の結果の中でも特に顕著なのは最下段の網掛けした２行の終止コドンへの変異とフレームシフトでありいずれも病原性の比率が 200 倍から 600 倍以上と極めて高い. また, 不明となっているエントリー数も少なく, 病原性の有無が確定している比率が高い．また，終止コドンへの変異（ナンセンス変異）はコード領域中のアミノ酸をコードしているコドンが終止コドンへと変化する変異で，フレームシフトと同様に，コードされているタンパク質が途中で切れてしまうため，病原性が極めて高いことが示されている．開始コドンの変異は翻訳の開始を妨げる可能性があるため，スプライスドナー・アクセプターの変異と同程度に病原性に関わる可能性が高い．一方，終止コドンの消失はアミノ酸配列の長さを変える変異の一つではあるが，影響としては通常のタンパク質の末尾に数十残基ほどが加わる程度で済むため病原性の比率が比較的低い．表２で網掛けをした特に病原性の確率が高い，スプライスドナーおよびアクセプタの変異，および終止コドンへのナンセンス変異とフレームシフトを含む 157 の遺伝子中，変異数の多い遺伝子 28 件をその変異数とともに表３に示す．これらはいずれもがん抑制遺伝子としてよく知られているものであり，これらの遺伝子が途中で切断されることにより機能が失われて病原性が生じていると考えられる．

表３． タンパク質を切断する変異により ClinVar で病原性となる遺伝子と ClinVar で該当する変異の数

遺伝子・変異数			
BRCA1・2762	PALB2・515	TP53・212	SDHA・84
BRCA2・2582	CHEK2・351	MEN1・211	SDHB・80
NF1・1426	DICER1・339	CDH1・204	NF2・79
ATM・787	PMS2・299	PTEN・200	FLCN・73
APC・755	BRIP1・297	NBN・137	MRE11・72
MSH2・741	RAD50・260	RAD51D・100	STK11・60
MSH6・713	RB1・228	RAD51C・92	FH・56

3･3　がん細胞で多く見られる変異について（COSMIC）

データベース COSMIC は，ClinVar よりデータ量が２桁も多く，がん細胞における体細胞変異について，網羅的に列挙したデータベースであるが，ClinVar とは異なり，個々の変異に対する病原性(pathogenic)あるいは良性(benign)といった記載はない．

表４には，各染色体の長さ，遺伝子数，ClinVar での病原性変異の数，COSMIC での変異箇所の数，および観測された変異の総数を比較した．ClinVar の病原性変異数は他の値との相関が薄く, COSMIC の変異箇所と変異数はそれぞれの染色体の長さよりも，遺伝子数，に弱い相関がある様に見受けられる．このことは COSMIC データの多くの部分が遺伝子領域周辺のゲノム配列を観測するエキソーム解析により得られていることを反映していると考えられる．また，ClinVar の病原性変異の分布との相関が薄いことは, COSMIC のデータにパッセンジャー的，あるいはがんとは無関係な変異も区別されずに多く含まれていることを示唆するものと考えられる．

表４．染色体ごとの ClinVar 病原性変異数と，COSMIC の変異箇所数(positions)，変異の総数(mutations)

染色体	長さ	遺伝子数	ClinVar pathogenic	COSMIC positions	COSMIC mutations
1	248956422	2093	768	897297	3059092
2	242193529	1350	2528	794598	2940044
3	198295559	1103	1285	680740	2668543
4	190214555	789	197	511672	2032199
5	181538259	958	1671	538073	1979304
6	170805979	1083	72	538707	1900936
7	159345973	1040	768	620875	2357900
8	145138636	715	300	493417	1753489
9	138394717	799	559	360434	1309300
10	133797422	760	745	459867	2090772
11	135086622	1396	2015	534123	1978338
12	133275309	1133	275	506167	2000362
13	114364328	349	4785	243107	688432
14	107043718	837	525	273700	1052925
15	101991189	658	167	293576	1022443
16	90338345	991	1566	327044	1282742
17	83257441	1284	6940	356918	1953998
18	80373285	319	115	204057	784978
19	58617616	1545	448	378976	1410164
20	64444167	555	11	231750	857532
21	46709983	262	9	99184	390370
22	50818468	505	625	145525	532202
X	156040895	868	44	339107	1246247
Y	57227415	47	0	4177	13979
MT	16569	13	46	0	0
Total	3088286401	21452	26464	9833091	37306291

この様に，病原性に関する明示的な情報を含まず，がん細胞に起きている様々な変異を含んでいる COSMIC データベースから，がんに関連する変異を特異的に選び出す方法のひとつとして，数多く起きている変異を探索することが考えられる. COSMIC に含まれる変異はすべてがん細胞で観測される変異であり，多くのがん細胞に共通して含まれる変異であれば，がんの形成に何らかの寄与をしている可能性がある，と考えることができる．特定の染色体の塩基位置で，多くの変異が観測される順に並べ替えを行った上位の 10 件について表 5 に示す．

表 5 . COSMIC で多くの変異が観測される塩基位置上位 10 件

観測数	染色体	塩基位置	遺伝子	アミノ酸変異	ClinVar 記載
120,728	7	140,753,336	BRAF	V600E	pathogenic
116,956	12	25,245,350	KRAS	G12D	pathogenic
53,235	3	41,224,633	CTNNB1	T41A	ConflictingInterpret.
42,849	9	5,073,770	JAK2	V617F	NA
37,107	17	7,675,008	TP53	R175H	pathogenic
36,738	3	41,224,646	CTNNB1	S45F	pathogenic
35,924	12	25,245,351	KRAS	G12C	pathogenic
28,557	17	7,673,802	TP53	R273H	pathogenic
27,683	17	7,673,220	TP53	R248Q	pathogenic
24,304	12	25,245,347	CTNNB1	G13D	drug_response

最も多い BRAF の V600E は細胞増殖シグナルの活性化を促す典型的なドライバー変異として知られており ClinVar にも pathogenic の記載がある．一方で，表 5 の 3，4 番目の変異は，それぞれ，解釈の不一致が見られる変異，および，ClinVar には記載のない変異(NA)，であるが，がん細胞で非常に多く観測されることから，病原性に関連する可能性が考えられる．ここに挙げた変異を含め観測数の多い変異の多くは 1 塩基変異の非同義置換であり，100 回以上観測されているものが 5,125 件，50 回以上観測されているものは 21,985 件あり，頻度が下がるにつれ ClinVar の記載も少なくなるが，ドライバー変異の候補として検討することができる．

4　現状と今後の展望

今回，全エキソーム解析により得られた変異データからがん化に関連する変異を探索していくことを念頭に置いて, 既知がん遺伝子変異データベースの解析を行った. NCBI の SRA 等で公開されているデータや, すでに実施したがんゲノムのエキソームシーケンシングによる予備的なデータ解析からは，一つの検体について数十万件の SNP, INDEL を含む微小変異が検出されるが，その中から ClinVar のがんに関係する pathogenic 変異との単純な照合による一致だけを見た場合，1 検体あたり数件から多くて十数件程度であり，成熟したがん細胞では平均して 70 箇所程度あると言われるがんの生成に寄与する変異が充分に特定できているとは言えない．そうした中で，実際のがん細胞で観測される変異の中からがん化に寄与する変異を見出していくための手法として，本稿では 2 つの可能性を検討した．

ひとつ目は表 2 と表 3 で示された様に，がん抑制遺伝子におけるフレームシフトやナンセンス変異，およびスプライスドナー・アクセプター領域の変異がこれらの遺伝子の機能の阻害によりがん化を促進する可能性が非常に高いことである．ClinVar の解析からは，がん抑制遺伝子のフレームシフトおよびナンセンス変異は，遺伝子の下流，特に終端部に近い位置では病変性が低くなるケースも見られ，150 種程度のがん抑制遺伝子について終端付近以外で，上記の様な機能阻害が起き得る様な変異が観測されれば ClinVar に記載がなくても病原性に寄与している可能性が高い，と推定することができる．

ふたつ目は，表 5 で示した様に COSMIC での観測回数が多い変異はがん化を促進する変異であるケースが多いことである．先のがん抑制遺伝子の阻害が，やや間接的ながん化の促進であったのに対して，この場合には増殖シグナルの活性化などの，より積極的なドライバー変異が含まれる．通常の遺伝子多形との区別にも注意する必要があるが，がん細胞で特異的に多く見られる変異であれば未知のものであっても，詳細に解析していく変異の候補に挙げていくことができる．

がんの特徴のひとつとして，ゲノム修復の阻害などにより変異を蓄積しやすくする仕組みがある．上記の例でも特定の塩基位置で極めて多くの変異が観測されることについては何らかのメカニズムの存在も推定される．表 4 からは ClinVar データからがん化を誘導する遺伝子の分布に偏りがあることがわかる以外には，特に染色体ごとの変異の発生頻度の偏りは見られなかったが，こうした，がん化による変異の誘導についても考察したい．

本稿では，がん細胞のゲノム配列解析により得られる SNP・INDEL 等の微小変異について COSMIC・ClinVar で得られる情報と組合せて解析することで，どの様なアプローチが可能か，について考察した．実際のがん細胞の変異解析では，これらの微小変異の他に，より長い欠失や挿入，転座を含む大規模な組み替え，複数の遺伝子を含む広い領域にわたる欠損や増幅による CNV (Copy Number Variation) も観測できる．COSMIC にはこれらのデータについても記載があり比較検討が可能である. 我々の研究室では近年，植物オルガネラを対象として，ショートリードシーケンサーデータを利用した組み替え解析を手掛けてきており，がん細胞ゲノムの実際のデータの解析においても本技術を活用していきたいと考えている．ヒトゲノムは繰り返し配列が非常に多く，また，がんゲノム解析が難しい理由の一つとしてしばしば挙げられるゲノム不均一性（がん組織が正常細胞や多様ながん細胞の混合物であること）の問題もあり，様々な困難が予想されるが，新規ながん化メカニズムの特定を目指して取り組んで行きたい．

参考文献

1) 厚生労働省政策レポート「がん対策について」. https://www.mhlw.go.jp/seisaku/24.html

2) M.J. Landrum, *et al.*, ClinVar: improvements to accessing data, *Nucleic Acids Research*, **48**, D835-D844 https://doi.org/10.1093/nar/gkz972 (2020).

3) J.G. Tate *et al.*, COSMIC: the catalogue of somatic mutations in cancer, *Nucleic Acids Research*, 47. D941-D947 https://doi.org/10.1093/nar/gky1015 (2019).