第２４回道南医学会医学研究奨励賞（医師部門）

がんゲノムプロファイリングに有用なミスセンス変異有害性予測ツール

函館五稜郭病院　がんゲノム医療センター　○池　田　　健

【要旨】

ミスセンス変異の有害性を予測するコンピュータ・ツールが多数公開されている。各ツールには、予測モデルの違い等によってそれぞれ特徴があり、予測精度も異なっている。がんゲノムプロファイリング検査がカバーする遺伝子を対象にした検討では、多種類のツールの結果を統合して予測する手法の精度が高かった。

【キーワード】：ミスセンス変異、がんゲノムプロファイリング検査、コンピュータ予測

【はじめに】

アミノ酸置換を伴うミスセンス変異の有害性は、患者と集団における変異頻度の比較、変異と患者の臨床的相関関係、遺伝子やタンパク質の構造に関する知見など、多くのエビデンスをもとに確定される[1)]。しかし、エビデンスの十分ではない変異や意見の分かれる変異も多く、例えば公開データベースClinVarではそれぞれVUS（Variants of Unknown Significance）、CIP（Conflicting interpretations of pathogenicity）とラベリングされる[2)]。ミスセンス変異の有害性を予測するためのコンピュータ・ツールが多数開発されていて[3)]、構築に用いたデータや予測モデルの違いによってそれらの予測精度には差があることが知られている[4)-11)]。

国内で薬事承認されているがんゲノムプロファイリング検査（がん遺伝子パネル検査）のうち、FoundationOne CDxでは309遺伝子のミスセンス変異を同定することが可能である[12)]。本報告では、がんゲノムプロファイリングでの使用を想定し、 dbNSFP（database for nonsynonymous SNP’s functional predictions）[13)]によって得られる上記309遺伝子のミスセンス変異に対するスコアをもとに、各ツールの予測精度を比較した。

【方法】

FoundationOne CDxがカバーする309遺伝子について、ClinVarに登録されているミスセンス変異を手動でダウンロードした（78,090変異、2021年3月25日～27日）。それらの変異を、ClinVarにおける臨床上の意義が明確であることを条件に5,858変異に絞った。各ミスセンス変異について、35のコンピュータ・ツールが評価するスコアを、Rパッケージmyvariant[14)]経由でdbNSFP（database for nonsynonymous SNP’s functional predictions、v4.0）より一括取得した。コンピュータ・ツールは以下の通り。VEST4[15)]、PROVEAN[16)]、DEOGEN2[17)]、SIFT4G[18)]、SIFT[19)]、FATHMM-XF[20)]、PolyPhen-2[21)]、MutationAssessor[22)]、MutationTaster[23)]、LRT[24)]、MPC[25)]、FATHMM-MKL[26)]、FATHMM[27)]、PrimateAI[28)]、LIST-S2[29)]、GERP++[30)]、bStatistic[31)]、SiPhy[32)]、REVEL[33)]、ClinPred[34)]、BayesDel[35)]、MutPred[36)]、MVP[37)]、M-CAP[38)]、Eigen[39)]、MetaLR[40)]、MetaSVM[40)]、DANN[41)]、GenoCanyon[42)]、phastCons（phastCons100way_vertebrate、phastCons30way_mammalian、phastCons17way_primate）[43)]、phyloP（phyloP100way_vertebrate、phyloP30way_mammalian、phyloP17way_primate）[44)]。

以上のツールをLiらに準じ3種類に分類した[6)]（表１）。タンパク質機能への影響から予測する方法、進化系統樹の配列保存性によって予測する方法および他のツールのスコアを統合して予測する方法（アンサンブル法）の3種類である。いくつかのコンピュータ・ツールで見られる問題点として、訓練データとテストデータの重複が挙げられている[45)]。今回の検討では、各ツールの構築に用いられたデータとの重複を避けるために、登録年が2019年、2020年あるいは2021年の比較的新しいデータのみを使用した。

予測精度は、ROC曲線（Receiver Operatorating Characteristic curve、受信者動作特性曲線）およびROC曲線下面積（AUC: Area under the curve）によって評価した。全ての統計学的手法はR言語によった。

【結果】

最もArea under the curve（AUC）の高いコンピュータ・ツールはClinPred（0.959）であり、BayesDel（0.953）、REVEL（0.942）、VEST4（0.940）、DEOGEN2（0.93

1）と続いた（表１、図１）。上位三つまでをアンサンブル法が占めた。最もAUCの低いツールはbStatisticで0.573であった。

コンピュータ・ツールのタイプ別では、タンパク質機能予測が平均0.841（標準偏差0.068；中央値0.860）、進化系統樹利用が平均0.717（標準偏差0.080；中央値0.704）、アンサンブル法が平均0.870（標準偏差0.097；中央値0.903）であった（図2）。タンパク質機能予測と進化系統樹利用、アンサンブル法と進化系統樹利用の平均値の差に統計学的有意差を認めた（それぞれp<0.001、p=0.001、マン・ホイットニーU検定）。タンパク質機能予測とアンサンブル法の間には統計学的有意差を認めなかった（p=0.2）。

【考察】

発癌に重要な役割を担うゲノム変異をドライバー変異といい、発癌プロセスのなかでたまたま獲得された変異をパッセンジャー変異という[46]。その変異パターンは極めて多様性に富んでいて、例えば同じ大腸癌であっても個人差が非常に大きい[47]。最も確実にドライバー変異を同定する方法は生化学的解析やモデル動物、細胞株を用いた機能アッセイであろうが、実臨床では明らかに現実的ではない。遺伝子変異の生物学的意義を推測する補助的手段としてコンピュータ・ツールが利用されている[1)、3)]。

コンピュータ・ツールは手法によって3つに分類できる[6)]。タンパク質機能への影響から予測する方法、進化系統樹の配列保存性によって予測する方法および他のツールのスコアを統合して予測する方法である。三つ目の手法は便宜的にアンサンブル法と呼ばれることが多い。今回の検討では、タンパク質機能から予測する方法とアンサンブル法の予測精度が高かった。GhoshらはREVEL、MetaLR、Meta SVM、M-CAP、Eigenなどのアンサンブル法が高性能と報告している[5)]。さらに、精度の高いツールとしてREVELをあげる論文が目立つ[6)、8)、11)]。Borgesらは、高性能であったツールとしてClinPredとBayesDelを指摘し[10)]、今回の検討と同様である。Borgesらの対象とした疾患はがんではなく常染色体劣性遺伝疾患のムコ多糖症であった。体細胞変異が主体のがんと生殖細胞変異が主体の代謝性疾患という、臨床像の全く異なる疾患群においてもClinPredとBayesDelが高性能であったことは興味深い。タイプ別の予測精度の違いをもたらした要因については、さらなる検討が必要である。

コンピュータ・ツールは一般に特異性が低く、有害性の唯一のエビデンスとして用いることは避けるべきである[1)]。一方、良性が疑われる変異の場合は、コンピュータ・ツールの評価が良性を決定づけるデータとなり得るという[48)]。

人工知能の利用によってタンパク質三次元構造の予測精度は劇的に向上した[49)]。このような技術的進歩は、ミスセンス変異によるアミノ酸置換の機能的影響をコンピュータ上で正確にシミュレートできる可能性を示唆しており、ミスセンス変異有害性予測ツールのさらなる精度改善につながるものと期待される。

本論文に関する著者の利益相反；なし

【文献】

1) Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.
2) Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. Nucleic Acids Res. 2020;48(D1):D835-d44.
3) Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, Karchin R, et al. Computational approaches to identify functional genetic variants in cancer genomes. Nat Methods. 2013;10(8):723-9.
4) Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat. 2011;32(4):358-68.
5) Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. 2017;18(1):225.
6) Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res. 2018;46(15):7793-804.
7) Anderson D, Lassmann T. A phenotype centric benchmark of variant prioritisation tools. NPJ Genom Med. 2018;3:5.
8) Tian Y, Pesaran T, Chamberlin A, Fenwick RB, Li S, Gau CL, et al. REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. Sci Rep. 2019;9(1):12752.

9) Chen H, Li J, Wang Y, Ng PK, Tsang YH, Shaw KR, et al. Comprehensive assessment of computational algorithms in predicting cancer driver mutations. Genome Biol. 2020;21(1):43.
10) Borges P, Pasqualim G, Matte U. Which Is the Best In Silico Program for the Missense Variations in IDUA Gene? A Comparison of 33 Programs Plus a Conservation Score and Evaluation of 586 Missense Variants. Front Mol Biosci. 2021;8:752797.
11) Wilcox EH, Sarmady M, Wulf B, Wright MW, Rehm HL, Biesecker LG, et al. Evaluating the impact of in silico predictors on clinical variant classification. Genet Med. 2021.
12) FoundationOne CDx. [cited 2022 2/5]. Available from: https://www.foundationmedicine.com/test/foundationone-cdx
13) Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12(1):103.
14) Xin J, Mark A, Afrasiabi C, Tsueng G, Juchler M, Gopal N, et al. High-performance web services for querying gene and variant annotation. Genome Biol. 2016;17(1):91.
15) Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3:S3.
16) Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7(10):e46688.
17) Raimondi D, Tanyalcin I, Ferte J, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017;45(W1):W201-W6.
18) Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11(1):1-9.
19) Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001;11(5):863-74.
20) Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018;34(3):511-3.
21) Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.
22) Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. 2007;8(11):R232.
23) Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010;7(8):575-6.
24) Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19(9):1553-61.
25) Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, et al. Regional missense constraint improves variant deleteriousness prediction. bioRxiv. 2017.
26) Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31(10):1536-43.
27) Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57-65.
28) Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018;50(8):1161-70.
29) Malhis N, Jacobson M, Jones SJM, Gsponer J. LIST-S2: taxonomy based sorting of deleterious missense mutations across species. Nucleic Acids Res. 2020;48(W1):W154-w61.
30) Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput

Biol. 2010;6(12):e1001025.
31) McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet. 2009;5(5):e1000471.
32) Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009;25(12):i54-62.
33) Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016;99(4):877-85.
34) Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet. 2018;103(4):474-83.
35) Feng BJ. PERCH: A Unified Framework for Disease Gene Prioritization. Hum Mutat. 2017;38(3):243-51.
36) Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009;25(21):2744-50.
37) Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, et al. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun. 2021;12(1):510.
38) Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48(12):1581-6.
39) Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48(2):214-20.
40) Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24(8):2125-37.
41) Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2015;31(5):761-3.
42) Lu Q, Hu Y, Sun J, Cheng Y, Cheung KH, Zhao H. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci Rep. 2015;5:10576.
43) Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034-50.
44) Siepel AP, K. S.; Haussler, D. New methods for detecting lineage-specific selection. In: Apostolico AG, C.; Istrail, S.; Pevzner, P., A.; Waterman, M, editor. Proceedings of the 10th International Conference on Research in Computational Molecular Biology. Berlin: Springer; 2006. p. 190-205.
45) Grimm DG, Azencott CA, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015;36(5):513-23.
46) Haber DA, Settleman J. Cancer: drivers and passengers. Nature. 2007;446(7132):145-6.
47) Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007;318(5853):1108-13.
48) Bean LJH, Hegde MR. Clinical implications and considerations for evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Med. 2017;9(1):111.
49) Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.

表1　各コンピュータ・ツールのAUC

予測ツール	AUC	予測ツール	AUC	予測ツール	AUC
タンパク質機能予測		進化系統樹の利用		アンサンブル	
VEST4	0.940	PrimateAI	0.851	ClinPred	0.959
DEOGEN2	0.931	LIST-S2	0.810	BayesDel	0.953
SIFT4G	0.896	phyloP100way_vertebrate	0.809	REVEL	0.942
PROVEAN	0.889	SiPhy	0.725	MVP	0.923
SIFT	0.886	GERP++	0.709	MutPred	0.921
Polyphen2	0.868	phastCons17way_primate	0.704	M-CAP	0.903
FATHMM-XF	0.860	phastCons100way_vertebrate	0.694	Eigen	0.874
MutationAssessor	0.828	phastCons30way_mammalian	0.691	MetaSVM	0.856
LRT	0.803	phyloP30way_mammalian	0.669	MetaLR	0.846
MutationTaster	0.779	phyloP17way_primate	0.647	DANN	0.751
MPC	0.769	bStatistic	0.573	GenoCanyon	0.640
FATHMM-MKL	0.746				
FATHMM	0.740				
平均	0.841	平均	0.717	平均	0.870

AUC, area under the curve

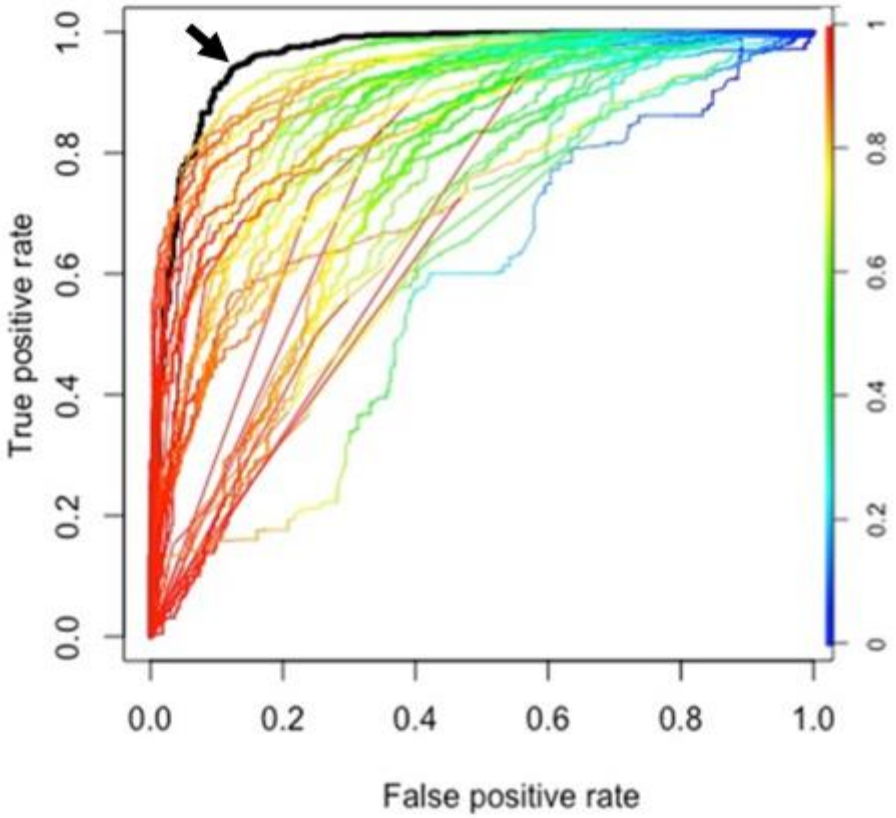

図１　各コンピュータ・ツールのROC曲線　矢印はClinpred

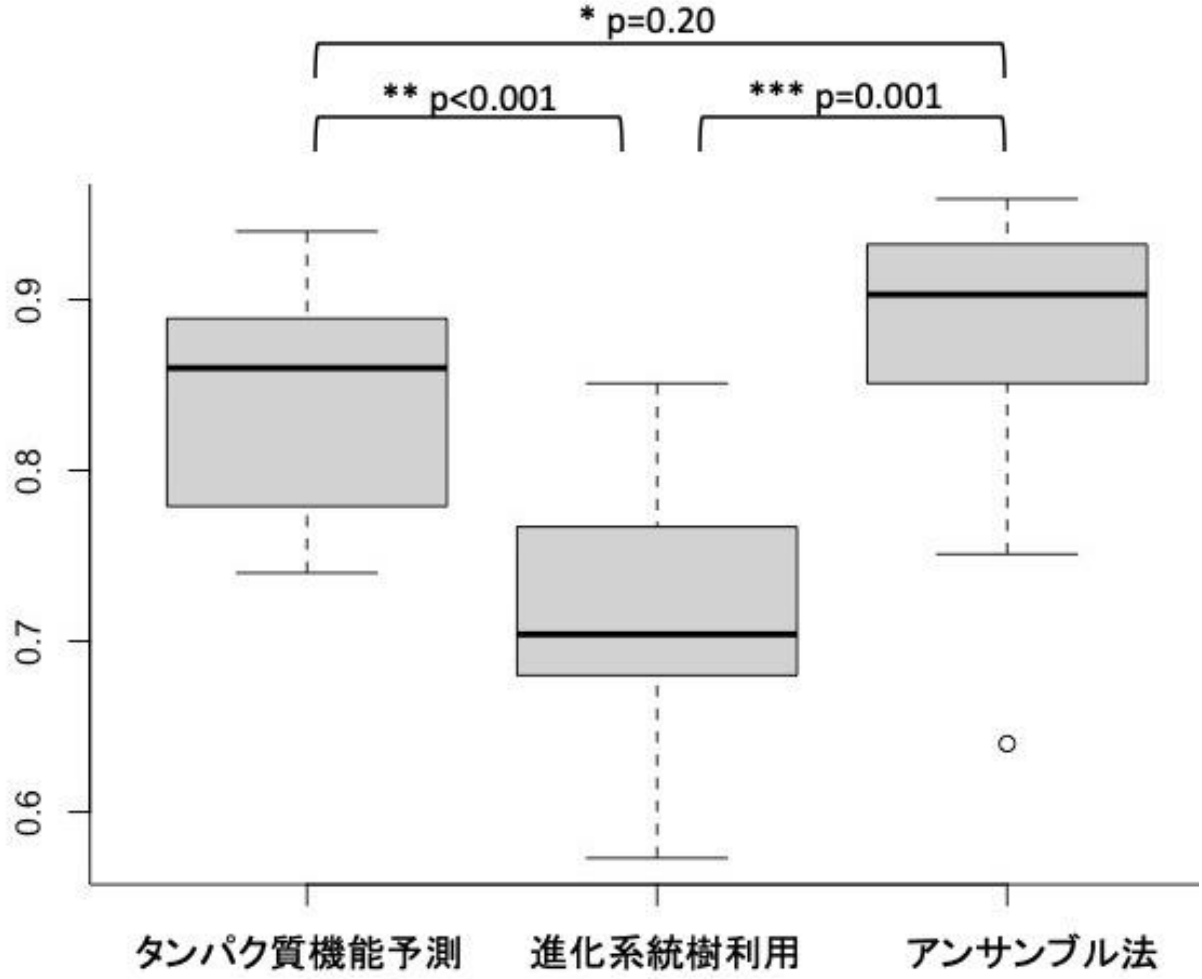

図２　コンピュータ・ツールのタイプ別のAUC