論文に出る300遺伝子の選定について

生命科学論文において，ある特定の遺伝子について研究された報告は数多く存在する．本書にてリストアップした「論文に出る300遺伝子」は，遺伝子と論文の関係をまとめたデータであるgene2pubmedを2023年12月31日にダウンロードし，それをもとに論文数の多い順にランキングを計算した，その上位300個の遺伝子である．

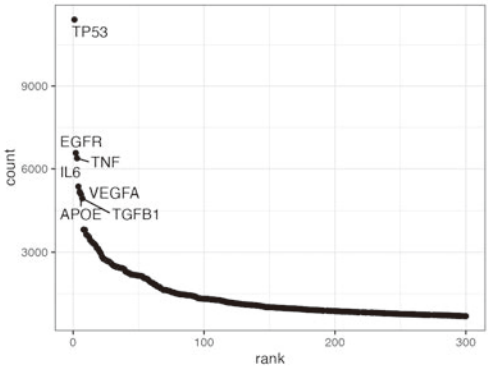

図1 「論文に出る300遺伝子」のgene2pubmed掲載論文数

図1に示したように最も論文数の多い遺伝子は*TP53*であり，その報告数には偏りがあることがわかる．つまり遺伝子にはよく知られているものとそうではないものが存在し，有名な遺伝子は全体のうちの一部を占めているにすぎない．gene2pubmedに記載されたヒトの161,647遺伝子（ノンコーディング遺伝子なども含まれているため数が多い）と，今回選抜した300遺伝子を対象に報告論文数の要約統計量を**表**に記す．

表　gene2pubmed に記載された論文数の要約統計量

	gene2pubmed収載ヒト遺伝子（161,647）	論文に出る300遺伝子（300）
最小値	1	685
第 1 四分位数	1	817
中央値	1	1,008
平均値	12.53	1,420
第 3 四分位数	1	1,550
最大値	11,408	11,408

多くの遺伝子はカウントされた論文数が1報である一方，「論文に出る300遺伝子」の論文数は最低でも685報は存在し，比較的よく研究された遺伝子であることが窺える．

ここで，gene2pubmed についてもう少し触れておきたい．gene2pubmed とは，NCBI（National Center for Biotechnology Information）が提供しているデータセットで，特定の GeneID（遺伝子）と tax_id（生物の分類）とそれに関連する PubMed_ID（論文）を関連付けた情報が含まれている．詳細は https://ftp.ncbi.nlm.nih.gov/gene/README もご確認いただきたい．gene2pubmed データは毎日更新されており，今回使用したデータセットには52,731,417行以上の情報が登録されていた．さらに human（tax_id：9606）に限定しても2,029,188行と大規模なデータとなっていた．**図2**にはそのごく一部分を示す．

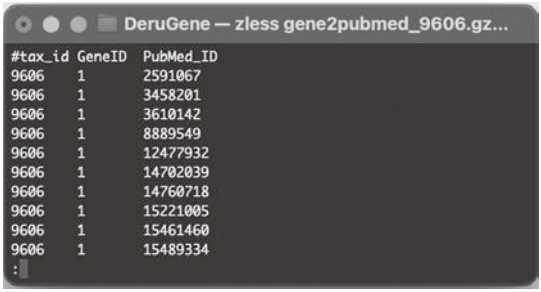

図2　gene2pubmedの一部分

2行目は，Human（Taxonomy id：9606）のとA1BG alpha-1-B glycoprotein（GeneID：1）の論文情報（Pubmed_ID：2591067）が記載されている．

gene2pubmedデータは，NCBIのFTPサイト（https://ftp.ncbi.nlm.nih.gov/gene/DATA/）からテキスト形式でダウンロードすることができる．RやPythonやシェルスクリプトなどのプログラミング言語に慣れている人であれば，さまざまな目的で活用することができる．例えば，特定の遺伝子IDを検索して，その遺伝子に関連するすべてのPubMed論文を抽出するなど，アイディアしだいで活用の幅は広がる．

プログラミングしない場合でも活用できる情報として，われわれが2021年の論文（https://doi.org/10.3390/biomedicines9050582）に使用したnumber of publications per human geneの表（https://doi.org/10.6084/m9.figshare.14140850.v1）を紹介したい（**図3**）．2021年の情報であることに留意は必要だが，このデータをダウンロードして活用いただけば，任意のヒト遺伝子の注目度合いを推し量ることができる．具体的な方法に関しては実験医学2021年12月号に掲載された「バイオDXで発見する見逃されていた低酸素応答性遺伝子（https://doi.org/10.18958/6949-00001-0000838-00）」も参照いただきたい．

	A	B	C	D	E
1	GeneID	Pubmedcnt	rnk	Symbol	description
2	1	31	9952.0	A1BG	alpha-1-B glycoprotein
3	2	272	921.0	A2M	alpha-2-macroglobulin
4	3	6	19233.0	A2MP1	alpha-2-macroglobulin pseudogene 1
5	9	245	1059.0	NAT1	N-acetyltransferase 1
6	10	704	228.0	NAT2	N-acetyltransferase 2
7	11	5	19942.0	NATP	N-acetyltransferase pseudogene
8	12	198	1444.0	SERPINA3	serpin family A member 3
9	13	28	10668.0	AADAC	arylacetamide deacetylase
10	14	43	7802.0	AAMP	angio associated migratory cell protein
11	15	42	7963.0	AANAT	aralkylamine N-acetyltransferase
12	16	72	4811.0	AARS1	alanyl-tRNA synthetase 1
13	17	2	24500.0	AAVS1	adeno-associated virus integration site 1
14	18	45	7492.0	ABAT	4-aminobutyrate aminotransferase
15	19	753	203.0	ABCA1	ATP binding cassette subfamily A member 1

図3　Number of publications per human gene
https://doi.org/10.6084/m9.figshare.14140850.v1より.

本書では「論文に出る300遺伝子」の概要がまとめられているので, 各遺伝子についてID検索することなくすばやく要点を把握できるが, gene2pubmedに含まれているIDについて詳細情報を入手したい場合は, 次の**図4〜6**のようにIDやURLを入力すると検索できるので, 必要に応じてお試しいただきたい.

図4　tax_id：9606
https://www.ncbi.nlm.nih.gov/taxonomy/?term = 9606

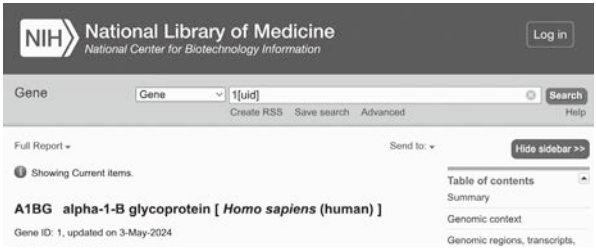

図5　GeneID：1

https://www.ncbi.nlm.nih.gov/gene/?term =1

図6　PubMed_ID：2591067

https://pubmed.ncbi.nlm.nih.gov/2591067/