Title: Data mining patented antibody sequences.

Authors: Konrad Krawczyk[1,*], Andrew Buchanan[2], Paolo Marcatili[3]

Affiliations:

[1] NaturalAntibody, Hamburg, Germany
[2] AstraZeneca, Cambridge, United Kingdom
[3] Technical University of Denmark , Lyngby, Denmark

* contact: konrad@naturalantibody.com

Abstract:
Patent literature should be a reflection of thirty years of engineering efforts in developing monoclonal antibody therapeutics. Such information is potentially valuable for rational antibody design. Patents however are not designed to convey scientific knowledge, but rather legal protection. It is unclear whether antibody information from patent documents, such as antibody sequences could be useful for the therapeutic antibody sphere in conveying engineering know-how rather than act as legal reference only. To assess the utility of patent data for therapeutic antibody engineering, we quantified the amount of antibody sequences in patents destined for medicinal purposes and how well they reflect the primary sequences of therapeutic antibodies in clinical use. We identified 16,526 patent families from major jurisdictions (e.g. USPTO and WIPO) that contained antibody sequences. These families held 245,109 unique antibody chains (135,397 heavy chains and 109,712 light chains) that we compiled in our Patented Antibody Database (PAD, http://naturalantibody.com/pad). We find that antibodies make up a non-trivial proportion of all patent amino acid sequence depositions (e.g. 10.95% of USPTO Full Text database). Our analysis of the 16,526 families demonstrates that the volume of patent documents with antibody sequences is growing with the majority of documents classified as containing antibodies for medicinal purposes. We further studied the 245,109 antibody chains from patent literature to reveal that they very well reflect the primary sequences of antibody therapeutics in

clinical use. This suggests that patent literature could serve as a reference of previous engineering efforts to improve rational antibody design.

Introduction

The binding versatility of antibodies has been used for medicinal purposes making them the most successful group of biotherapeutics[1]. Typical timelines involved in bringing these molecules to the market are slow, however more and more molecules are approved in the US and EU each year[1]. Successful exploitation of antibodies by either experimental[2,3] or computational techniques[4] relies on our ability to understand what makes a successful antibody-based therapeutic[5,6].

Therapeutic antibodies on the market and in late stage clinical trials have been previously studied by experimental[2,7] and computational[6] approaches to identify properties that make a successful biotherapeutic. Such studies[2] however only focused on 137 approved or post-phase-I antibodies (Clinical Stage Therapeutics, or CSTs), which is a small dataset in the light of the mutational space available to antibodies[8]. CSTs, are high-quality data-points that are end results of a long engineering process of selecting a molecule from a number of viable candidates. The single successful therapeutic molecule is therefore only partially representative of the engineering process. Full public disclosure of the efforts involved in developing a therapeutic antibody constituting intermediate sequences and selection decisions is not desirable because of the commercial value of such know-how, which needs to be legally protected.

Because of the need to protect the know-how involved in engineering therapeutic antibodies, relevant information needs to be disclosed in patent documents. Previous approaches to extract information on patent antibody landscape[9] or specific antibody formats[10] focused on keyword and patent classification searches. One can broadly discern between patents on antibody techniques (e.g. phage display, humanization) and novel antibody molecules. It is the patents on novel molecules that could be of particular engineering interest as these reflect the constructs that might find their way into the clinic. The disclosure of antibody sequence and target information[11] in such patents reveals to a certain extent the engineering choices as such molecules have been

subjected to myriad prior tests to be suitable candidates for expensive legal protection and further clinical trials.

The purpose of patent literature is not conveying scientific knowledge, but legal protection. In this work we assessed the utility of patent data for therapeutic antibody engineering efforts by establishing the extent to which antibodies from patents reflect therapeutics in clinical use. For this purpose, we identified patent documents that contained antibody sequences, to quantify how many of these were destined for medicinal purposes and how well they reflect advanced stage therapeutics.

Results.

Antibodies account for a non-trivial proportion of sequences deposited in patent documents.

We identified documents with antibody sequences by downloading data from four data sources: USPTO (http://uspto.gov), WIPO (http://wipo.int), DDBJ[12], and EBI[13]. Choice of the data sources was motivated by the availability of biological sequences and coverage of patent documents worldwide. Biological sequence information is not universally available in patent documents in all jurisdictions[14]. In certain cases, the data is not freely available, but rather accessible for a fee (e.g. European Patent Office). Primary access to biological sequences in machine-readable format is freely available from the USPTO and WIPO. USA is the largest pharmaceutical market[15], compelling pharmaceutical companies developing a novel antibody therapeutic to seek patent protection within the jurisdiction of USPTO. Similarly, it is common to seek protection under the auspices of WIPO PCT system in order to spread the coverage of the patent documents across many jurisdictions worldwide. Furthermore, data from certain major jurisdictions, such as EPO, JPO and KPO are available via third parties such as DDBJ[12] and EBI[13]. Therefore, we argue that datasets made available via USPTO, DDBJ, WIPO and EBI provide a reasonable coverage of the worldwide antibody sequence patents.

We extracted raw sequence data from USPTO, WIPO, DDBJ and EBI on Jan 30th 2020, with the particulars of parsing the heterogenous sources described in Methods. From each dataset we extracted raw, redundant amino acid and nucleic acids sequences. Sequences containing exclusively nucleotides were translated to amino acids using IgBlast[16] as described previously[17]. Raw amino acid sequences were analyzed using ANARCI[18] to identify antibody variable region chains (V_H, V_L, including scFvs). We report the number of raw sequences analyzed and the resulting identified antibodies in Table 1.

We find a higher proportion of sequences identified as antibodies in amino acid depositions which account for as many as 10.95% and 12.09% of USPTO-FT and DDBJ datasets respectively. In fact, large portion of sequences deposited in patents are very short; for instance in USPTO-FT only 1,811,694 (32.73%) amino acid sequences are longer than 50 amino acids, and antibodies make up 30.50% of these. This stands to show that antibodies make up a non-trivial volume of all the sequences deposited in patent documents.

Antibody sequence data in patents is however redundant to a large extent when one considers a unique sequence to be defined by its variable region. Combining all the non-redundant V_H and V_L sequences from our datasets we count 245,109 unique antibody domains (135,397 heavy chains and 109,712 light chains). This suggests that many antibody variable region sequences are listed as part of multiple patent documents. Not all of these sequences however are guaranteed to have been developed for medical applications, which can be determined by analyzing the text content of patent documents.

Table 1. Published biological sequences and proportion thereof identified as antibody chains. We extracted raw sequences from USPTO (divided between the full text, FT, and long listing repository PSIPS), DDBJ, WIPO and EBI. The total number of raw sequences is given in column Total Raw. Of these we show how many were identified by ANARCI as containing an antibody chain (column Ab-identified). In the column "% Total" we report the proportion of identified antibody sequences out of the total of raw sequences. Both Total Raw and Ab-identified columns report the redundant number of sequences so as to exemplify the volume of antibody depositions in patent sequences – we report the number of unique heavy (H) and light (L) chains in the parentheses in column "Ab-identified".

Source	Sequence Type	Total Raw	Ab-identified (unique Heavy (H), Light (L))	%Total
USPTO FT	Amino Acid	5,534,127	606,036 (H=52,388,L=38,922)	10.95
	Nucleotide	7,068,248	229,547 (H=21,169,L=17,009)	3.24
USPTO PSIPS	Amino Acid	25,527,942	470,317 (H=33,806,L=24,086)	1.84
	Nucleotide	176,840,912	376,567 (H=35,802,L=46,374)	0.21
DDBJ	Amino Acid	4,412,209	533,762 (H=61,999,L=46,015)	12.09
	Nucleotide	44,968,142	413,485 (H=35,290,L=28,502)	0.91
WIPO	Amino Acid	10,275,174	435,218 (H=67,533,L=49,699)	4.23
	Nucleotide	13,490,560	160,542 (H=35,747,L=27,275)	1.19
EBI	Amino Acid	10,368,431	713,620 (H=73,450,L=50,326)	6.88
	Nucleotide	12,349,772	38,366 (H=15,792,L=13,339)	0.31

Patent landscape of documents containing antibody sequences

We analyzed the text content of patents containing antibody sequences so as to establish what proportion of these list molecules for medicinal purposes. We connected all the redundant antibody sequences to their patent documents and identified a patent family for each. A patent family can be regarded as identifying documents with the same subject matter across several

138 jurisdictions. Altogether our 245,109 sequences are distributed among 16,526 patent families.
139 We extracted the metadata from the patent documents, such as titles, abstracts, inventors and
140 classifications. We used this information to determine the proportion of patents destined for
141 medicinal applications by analyzing their classifications, and whether the inventors and listed
142 targets resemble entities and molecules associated with development of monoclonal antibody
143 therapies.
144
145 **Most patent documents citing antibody sequences are destined for medicinal applications.**
146
147 We analyzed the patent classifications of the 16,526 patent families that indicate the purpose of
148 the invention described in each document. We extracted the Cooperative Patent Classification
149 (CPC, developed by USPTO and EPO, https://www.cooperativepatentclassification.org/)
150 designations from the documents as this was the most common listed scheme, covering 15,951
151 (96.52%) out of 16,526 families. Patent classifications according to CPC have a section, class,
152 subgroup, main group and a subgroup (e.g. classification C07K16/2866 has section C, class 07,
153 subclass K, main group 16 and subgroup 2866). We divided the 15,951 families according to
154 their CPC classifications excluding the subgroup (e.g. C07K16/2866 becomes C07K16) to reveal
155 the general categories the documents fall into and present results in Table 2.
156
157 Subgroup C07K16 that indicates immunoglobulins, is the most common classification, present in
158 13,790 (86.45%) of the 15,951 patent families. Families listing antibodies for medicinal purposes
159 (A61K39) account for 9,459 (59.30%) of the 15,951 families. Furthermore the more general
160 medicinal categorization A61K (preparations for medical, dental or toilet purposes) accounts for
161 11,398 (71.45%) of the 15,951 patent families. This indicates that the majority of documents
162 citing antibody sequences are developed for medicinal purposes, such as novel treatments or
163 diagnostics. This is well reflected by the organizations that submit such patent applications,
164 where 9 out of top 10 and 69 out of top 100 are pharmaceutical companies associated with
165 development of monoclonal antibody therapies for a range of targets and disease indications (see
166 Supplementary Section 1).
167
168

169 **Table 2.** Subclasses of the patent classifications. Most common subclasses associated with
170 patents including antibody sequences according to the Cooperative Patent Classification (CPC,
171 https://www.cooperativepatentclassification.org/). There were 15,951 patents containing
172 antibodies with CPC classification and the percentage of families in each class is expressed as a
173 proportion of this number.
174

CLASS	TOTAL FAMILIES (%)	DESCRIPTION
C07K16	13,790 (86.4)	Immunoglobulins [IGs], e.g. monoclonal or polyclonal antibodies (antibodies with enzymatic activity, e.g. abzymes
C07K2317	12,001 (75.2)	Immunoglobulins specific features
A61K39	9,459 (59.3)	Medicinal preparations containing antigens or antibodies
C07K2319	3,451 (21.6)	Fusion polypeptide
G01N33	3,105 (19.4)	Investigating or analysing materials by specific methods not covered by groups G01N1/00 - G01N31/00
C07K14	3,037 (19.0)	Peptides having more than 20 amino acids; Gastrins; Somatostatins; Melanotropins; Derivatives thereof
A61K47	2,392 (14.9)	Medicinal preparations characterised by the non-active ingredients used, e.g. carriers or inert additives; Targeting or modifying agents chemically bound to the active ingredient
A61K38	2,058 (12.9)	Medicinal preparations containing peptides
C12N15	1,972 (12.3)	Mutation or genetic engineering; DNA or RNA concerning genetic engineering, vectors, e.g. plasmids, or their isolation, preparation or purification; Use of hosts therefor
A61P35	1,900 (11.9)	Specific therapeutic activity of chemical compounds or medicinal preparations
A61K45	1,671 (10.4)	Medicinal preparations containing active ingredients not provided for in groups
A61K31	1,415 (8.8)	Medicinal preparations containing organic active ingredients
G01N2333	1,329 (8.3)	Assays involving biological materials from specific organisms or of a specific nature
C12N5	816 (5.1)	Undifferentiated human, animal or plant cells, e.g. cell lines; Tissues; Cultivation or maintenance thereof; Culture media therefor

175 **Targets of antibodies in patent documents correspond to known therapeutic targets.**
176

177 We checked to what extent antibody targets reported in patent literature reflect those of known
178 therapeutic antibodies. Each patent family in PAD was scanned for antibody target (see
179 Methods). Therapeutic antibodies in clinical use together with their associated targets were
180 compiled from the WHO lists of International Nonproprietary Names[19] (INNs, e.g. list 122[20])
181 IMGT [21], Antibody Society (http://www.antibodysociety.org) and Thera-SAbDab[22,23], resulting
182 in 563 unique INNs. We grouped the targets by number of patent families and therapeutic
183 antibodies they were associated with. We present the results for top 30 targets sorted by the
184 highest number of patent families in Table 3.

The number of patent families associated with a target appears to correspond to a larger number of therapeutic antibodies against the same target. Top 10 targets sorted by number of their patent families account for 114 (20.24%), top 30 account for 223 (39.60%) and top 100 account for 369 (65.54%) out of 563 therapeutics. Therefore, targets from patents listing antibody sequences provide a reasonable reflection of the targets of currently available therapeutic antibodies. In fact the greater number of patent families per target can be associated with an earlier date of the said target being mentioned in a patent document (Figure 1A). It does not mean however that the patent space for monoclonal antibodies is saturated as the number of new targets mentioned is increasing (Figure 1B). This suggests that studying patent documents including antibody sequences could provide an early indication of their targets and thus activity in the field of therapeutic antibodies.

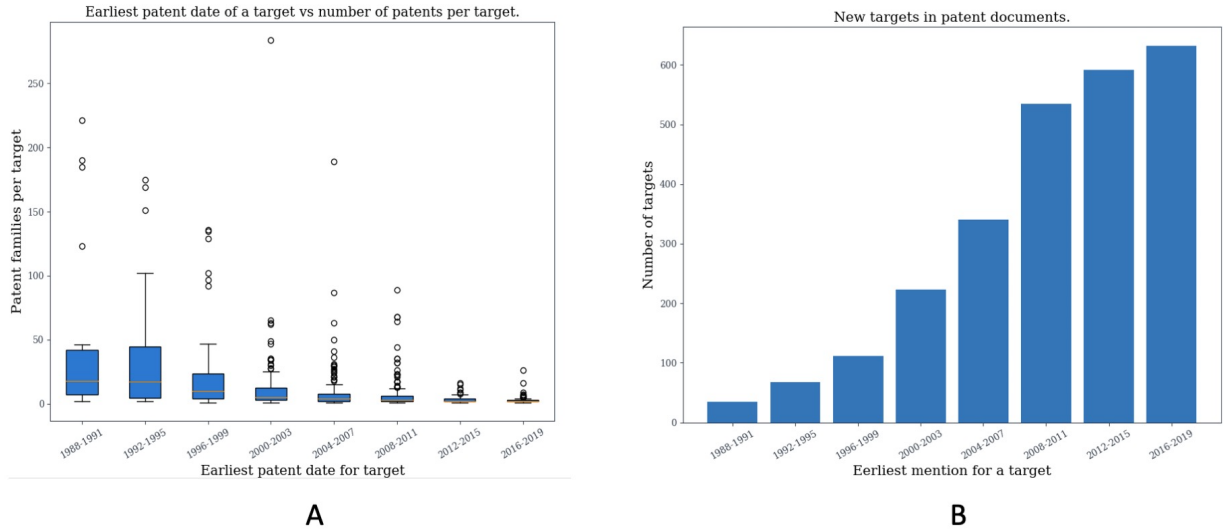


Figure 1. Target usage in patent documents reporting antibody sequences**.** A) Relationship between number of patent families per target and the earliest mention of the target in patent documents containing antibodies. For each target, we noted the earliest date among patent documents citing it and grouped these into 4-year intervals. Within each interval we noted the total number of patent families for a given target and plotted the aggregate for each time interval. B) For each 4-year interval, we plot the number of new target names that were first introduced in a patent document at that time.

206 **Table 3. Top 30 targets in patent documents**. We extracted the targets of the antibodies in
207 patent documents and present top 30 ranked by the number of families where they were
208 mentioned. For each target, we show the number of patent families mentioning the target
209 (#Families), the number of therapeutics on the market/in the clinic against it (#Therapeutics) and
210 the cumulative number of therapeutics covered by the top targets (#Therapeutics cumulative).

RANK	TARGET	#FAMILIES	#THERAPEUTICS	#THERAPEUTICS (CUMULATIVE)
1	pd1	284	20	20
2	cd3	221	20	40
3	her1	190	17	57
4	pdl1	189	12	69
5	tnfa	185	6	75
6	her2	175	9	84
7	cd20	169	14	98
8	influenza	151	5	103
9	cmet	136	4	107
10	vegfa	135	7	114
11	amyloid beta	129	8	122
12	hiv	123	2	124
=13	il6	102	7	131
=13	cd40	102	9	140
=13	cd19	102	6	146
14	ctla4	97	6	152
=15	il17	92	8	160
=15	igf1r	92	6	166
16	pcsk9	89	8	174
17	her3	87	6	180
=18	rsv	73	5	185
=18	cd38	73	5	190
=19	tau	68	5	195
=19	lag3	68	7	202
20	ox40	65	6	208
21	bcma	64	1	209
=22	il23	63	5	214
=22	cd47	63	2	216
23	ang2	62	2	218
24	vegfr2	56	5	223

211

212 **There is a growing number of patent documents associated with antibody sequences.**
213
214 We analyzed the timestamps associated with patents in order to check whether there is a growing
215 trend in releasing documents with antibody sequences and what proportion thereof is made up of
216 molecules for medicinal indications. Each patent family lists several dates corresponding to the
217 activity associated with the patent. We noted the earliest and most recent dates for each patent
218 family to reflect the original submission dates and the most up-to-date activity respectively.
219
220 We plotted the earliest dates for each patent family in our dataset which indicates that the
221 number of patent documents containing antibody sequences is steadily rising (Figure 2). The
222 most recent dates associated with the same patent documents (Figure 2) shows a more acute rise
223 since 2016 which indicates strong activity within the earlier submitted patents. Since not all
224 patent families are explicitly destined for medicinal applications, we have plotted the
225 corresponding earliest and most recent dates for the 9,459 documents classified as medicinal
226 preparations containing antibodies (Supplementary Figure 1) which recapitulates the increasing
227 number of patent documents being released.
228
229 Increasing patent activity in documents listing antibody sequences for medicinal indications is in
230 line with the rising approval rates for antibody-based biologics[1,24]. Given that the patents are an
231 early sign of approvals to come, it suggests that we can expect more biologics in the clinics in
232 the foreseeable future. Since majority of such patent documents are indeed listing antibodies for
233 medicinal purposes, the broad characteristics of the molecules listed in patent documents could
234 provide an indication of the engineering choices in their design.

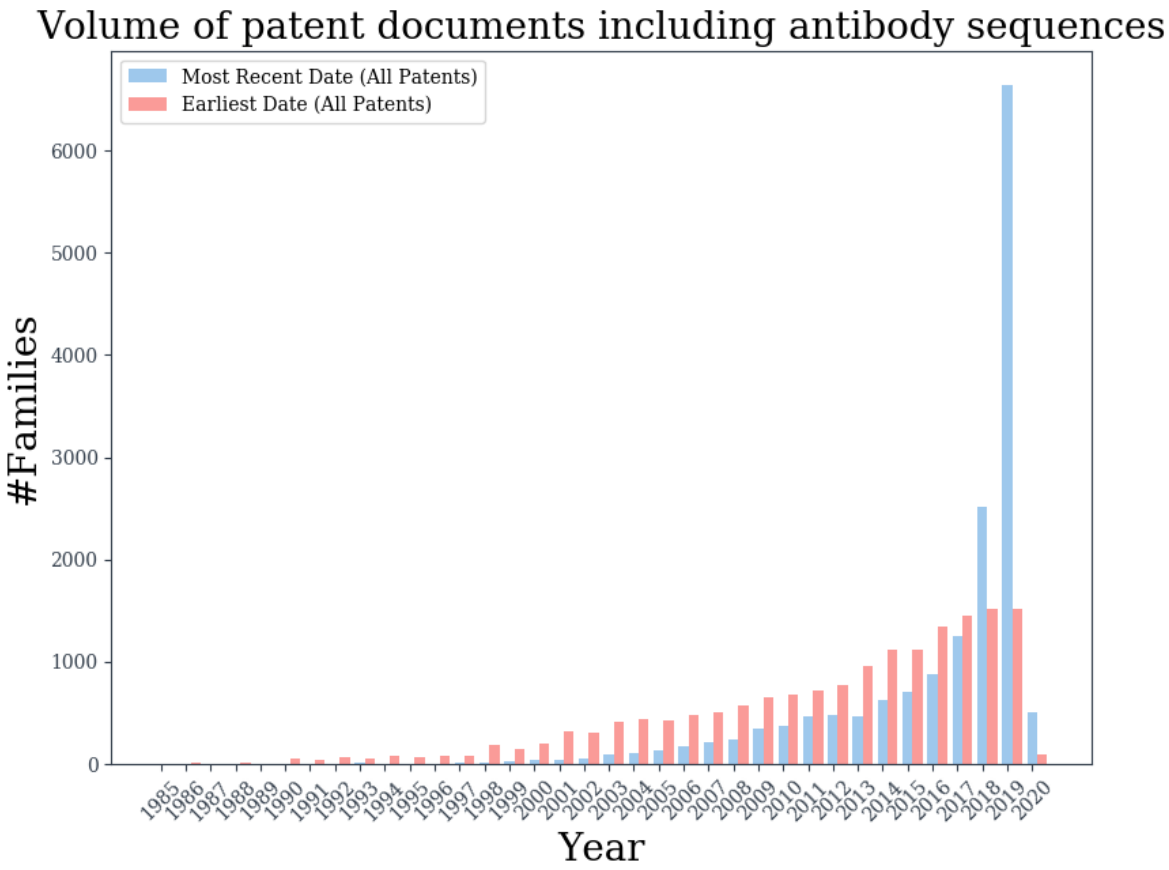


Figure 2. The volume of patent family documents listing antibody sequences per year. For each patent family we noted the earliest and most recent dates of any documents associated with it and the aggregate numbers of these are given by red and blue bars respectively. The apparent low activity in 2020 can be attributed to the fact that data contributed in 2020 only account for January that year.

The sequence landscape of patented antibodies.

Antibody sequences found in patent documents could reflect the broad decisions taken by engineers shaping these molecules before they arrive in the clinic. However, not all antibody sequences found in patent documents are destined for medicinal applications. For this reason we analyzed the broad sequence characteristics of antibodies from patent documents to establish to what extent they are a reflection of therapeutic antibodies in clinical use and vice-versa. We performed this analysis by looking at all of our antibodies from all patents (AllPatAb) and just the subset associated with documents classified as containing antibodies for medicinal

251 applications (MedPatAb). Altogether AllPatAb consisted of 135,397 heavy chains and 109,712
252 light chains whereas MedPatAb consisted of 93,067 heavy chains (68.73% of all heavy chains)
253 and 67,667 light chains (67.67% of all light chains).
254
255 **Most antibody sequences from patents align to human and mouse germline V region genes.**
256
257 We checked the patterns of organism-specific germline gene usage in antibody sequences
258 originating from patent documents. Since organism reporting is not consistent in patent
259 documents, we aligned the sequences in PAD to HMMs created from IMGT germline sequences
260 for fifteen organisms: human, mouse, alpaca, rhesus, rabbit, rat, pig, cow, macaque, zebrafish,
261 trout, salmon, dog, horse and chicken. For each organism and germline, we noted the total
262 number of patent antibody sequences aligning to a given germline as well as the number of
263 families they originated from.
264
265 We show the number of MedPatAb sequences that aligned to one of our fifteen organisms in
266 Table 4 with the corresponding distribution for AllPatAb sequences in supplementary Table 2.
267 Majority of the unique heavy sequences from patents for medicinal indications align to human
268 germlines (72.80% of unique sequences), followed by mouse (15.39% of unique sequences). The
269 same holds true for light chains with 67.72% of MedPatAb sequences aligning to human and
270 19.68% to mouse germlines. Antibodies aligning to either mouse or human germlines are most
271 frequently found within protein families. Human-aligned heavy and light chains can be identified
272 in 75.69% and 69.76% patent families respectively. Mouse-aligned heavy and light chains can be
273 found in 52.18% and 53.93% patent families respectively. This broad proportion is also reflected
274 in all the antibody sequences from patents (AllPatAb), indicating that the medicinal patent
275 classification does not skew the broad trend of majority of patented sequences aligning to human
276 and mouse germlines. The alignment to those two organisms reasonably reflects the human focus
277 of antibody development and the rodent antibodies that often server as a basis for humanized
278 therapeutics[25].
279
280

Germline V gene usage of antibodies from patent documents corresponds to a large extent with germline V gene usage of therapeutic antibodies.

Given that majority of antibodies from patents align to human germlines, we stratified these by the particular human V-region genes. In Table 4 and 5 we report the most common V-region genes medicinal patent sequences align to (corresponding numbers for all patents can be found in Supplementary Table 3). We compare the distribution of germline genes in patents to the germline usage in therapeutic antibodies to show to what extent patent submissions reflect current therapeutics.

The top heavy and light V region genes are identical among medicinal patented sequences, medicinal patents and therapeutics. The most used human heavy chain V-gene by sequence, family and therapeutic usage is IGHV3-23, accounting for 25.29% of all patented medicinal sequences, occurs in 15.56% of all medicinal families and accounts for 16.38% of therapeutics. The most frequently observed human light chain germline usage is IGKV1-39, accounting for 14.63% of all patented medicinal sequences, 12.84% of all medicinal patent families and 18.42% of therapeutic antibodies. Some of the most commonly observed genes might be the result of specific platform choices[26] that might attempt to recapitulate naturally observed frequencies[8] or focus on a small set of scaffolds[27]. The most frequently used germlines are broadly corresponding between patented sequences, medicinal patents and therapeutics, even though the ordering might not be the same. This indicates that the patent literature well reflects the choices of V-region genes of therapeutic antibodies in clinical use.

311 **Table 4.** Most common V-region gene species antibodies from patents aligned to. Antibodies
312 from patent documents destined for medicinal indications (MedPatAb) were aligned to fifteen
313 IMGT-derived[28] V region germlines from human, mouse, alpaca, rhesus, rabbit, rat, pig, cow,
314 macaque, zebrafish, trout, salmon, dog, horse and chicken. We noted the number of patent
315 sequences that aligned to the given species germline (#Unique Sequences) and the number of
316 patent families (#Patent Families) these originated from.
317

	PER SEQUENCE			PER FAMILY		
HEAVY CHAIN	**Organism**	**#Unique Sequences**	**Percentage**	**Organism**	**#Patent Families**	**Percentage**
	human	67754	72.80	human	7070	75.69
	mouse	14326	15.39	mouse	4874	52.18
	alpaca	7047	7.57	macaque	485	5.19
	rabbit	1313	1.41	horse	473	5.06
	macaque	1035	1.11	alpaca	403	4.31
	horse	799	0.85	rabbit	256	2.74
	chicken	417	0.44	chicken	46	0.49
	dog	291	0.31	dog	28	0.29
	rhesus	43	0.04	rhesus	23	0.24
	cow	30	0.03	cow	11	0.11
	pig	9	~0	pig	9	0.09
	rat	2	~0	rat	3	0.03
	salmon	1	~0	salmon	2	0.02
LIGHT CHAIN	**Organism**	**#Unique Sequences**	**Percentage**	**Organism**	**#Families**	**Percentage**
	human	45828	67.72	human	6312	69.76
	mouse	13320	19.68	mouse	4880	53.93
	rhesus	5333	7.88	rhesus	2238	24.73
	rabbit	1438	2.12	rat	361	3.98
	rat	778	1.14	rabbit	240	2.65
	chicken	505	0.74	chicken	46	0.5
	dog	220	0.32	cow	31	0.34
	cow	213	0.31	dog	21	0.23
	pig	17	0.02	pig	7	0.07
	horse	15	0.02	horse	7	0.07

318
319

Table 5. Top-20 most common human V-region genes antibodies from patents aligned to. For each patent antibody sequence for medicinal applications (MedPatAb) that aligned to human germline V-regions, we noted the IMGT V-region gene. We show the number of unique sequences that aligned to a given human V-region gene (Per Sequence) and number of patent families these originated from (Per Family). We also show the number of therapeutic antibody sequences in clinical use that align to the given V-region gene (Per Therapeutic).

	PER SEQUENCE			PER FAMILY			PER THERAPEUTIC		
HEAVY CHAIN	**Gene**	**#Sequences**	**Percentage**	**Gene**	**#Families**	**Percentage**	**Gene**	**#Sequences**	**Percentage**
	IGHV3-23	17140	25.29	IGHV3-23	2572	15.56	IGHV3-23	77	16.38
	IGHV1-2	6206	9.15	IGHV1-69	1369	8.28	IGHV1-69	39	8.29
	IGHV1-69	5334	7.87	IGHV3-30	1311	7.93	IGHV1-46	38	8.08
	IGHV3-30	4501	6.64	IGHV1-46	1136	6.87	IGHV3-33	26	5.53
	IGHV1-46	3840	5.66	IGHV1-2	1076	6.51	IGHV3-48	21	4.46
	IGHV3-33	2508	3.7	IGHV3-33	959	5.8	IGHV3-30	21	4.46
	IGHV1-18	2445	3.6	IGHV3-66	945	5.71	IGHV1-2	21	4.46
	IGHV1-3	1774	2.61	IGHV1-18	801	4.84	IGHV1-18	19	4.04
	IGHV3-66	1725	2.54	IGHV1-3	770	4.65	IGHV3-66	18	3.82
	IGHV5-51	1590	2.34	IGHV4-59	762	4.61	IGHV1-3	18	3.82
	IGHV4-59	1553	2.29	IGHV3-7	696	4.21	IGHV3-7	14	2.97
	IGHV3-48	1356	2	IGHV3-48	679	4.1	IGHV5-51	13	2.76
	IGHV4-4	1260	1.85	IGHV5-51	640	3.87	IGHV3-74	13	2.76
	IGHV7-4-1	1185	1.74	IGHV3-9	548	3.31	IGHV4-59	12	2.55
	IGHV3-7	1136	1.67	IGHV3-21	519	3.14	IGHV7-4-1	10	2.12
	IGHV3-21	1104	1.62	IGHV4-4	499	3.01	IGHV3-9	10	2.12
	IGHV3-9	1060	1.56	IGHV4-34	423	2.55	IGHV4-4	9	1.91
	IGHV3-15	1011	1.49	IGHV3-74	397	2.4	IGHV4-39	8	1.7
	IGHV3-11	917	1.35	IGHV3-11	392	2.37	IGHV4-34	8	1.7
	IGHV4-31	894	1.31	IGHV7-4-1	357	2.16	IGHV2-70	8	1.7
LIGHT CHAIN	**Gene**	**#Sequences**	**Percentage**	**Gene**	**#Families**	**Percentage**	**Gene**	**#Sequences**	**Percentage**
	IGKV1-39	6709	14.63	IGKV1-39	2123	12.84	IGKV1-39	70	18.42
	IGKV3-20	3882	8.47	IGKV3-11	1504	9.1	IGKV3-11	48	12.63
	IGLV1-51	2997	6.53	IGKV3-20	1335	8.07	IGKV3-20	35	9.21
	IGKV3-11	2753	6	IGKV4-1	1069	6.46	IGKV4-1	23	6.05
	IGKV4-1	2484	5.42	IGKV1-33	789	4.77	IGKV1-16	19	5
	IGKV3-15	1811	3.95	IGKV2-28	777	4.7	IGKV1-33	18	4.73
	IGKV1-5	1722	3.75	IGKV1-16	690	4.17	IGKV3-15	15	3.94

IGKV1-12	1627	3.55	IGKV1-5	669	4.04	IGKV1-12	12	3.15
IGKV1-33	1532	3.34	IGKV1-12	669	4.04	IGKV1-5	11	2.89
IGLV3-19	1479	3.22	IGKV3-15	633	3.83	IGLV1-40	10	2.63
IGKV2-28	1427	3.11	IGLV2-14	561	3.39	IGKV2-30	9	2.36
IGLV1-47	1377	3	IGKV1-27	558	3.37	IGKV2-29	9	2.36
IGLV1-44	1367	2.98	IGLV3-1	454	2.74	IGKV1-13	9	2.36
IGLV2-14	1310	2.85	IGLV1-44	427	2.58	IGLV3-21	8	2.1
IGLV3-1	1264	2.75	IGKV2-30	418	2.52	IGKV2-28	8	2.1
IGKV1-17	1123	2.45	IGLV3-21	412	2.49	IGKV1-27	7	1.84
IGLV1-40	1089	2.37	IGLV1-47	409	2.47	IGKV1-17	7	1.84
IGKV1-16	1076	2.34	IGLV1-40	407	2.46	IGLV1-47	6	1.57
IGLV3-21	1007	2.19	IGLV3-19	371	2.24	IGKV1-NL1	6	1.57
IGKV2-30	812	1.77	IGKV1-17	370	2.23	IGLV3-19	5	1.31

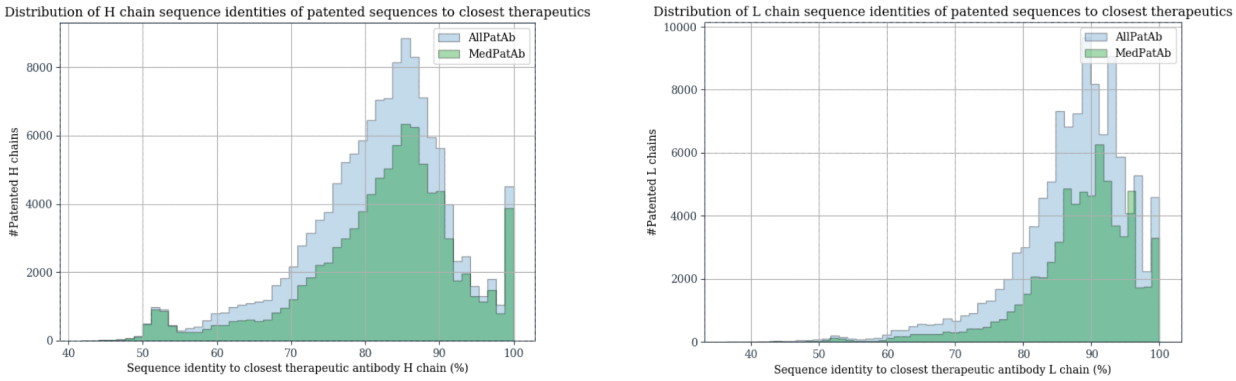
Antibodies from patent documents well reflect therapeutic antibody sequences, with the exception of CDR-H3 lengths.

The germline gene distribution of antibody sequences from patents appears to reflect the germline gene distribution of therapeutic sequences, though such comparison is not fit to indicate the actual sequence discrepancies between the two datasets. We checked to what extent patented sequences are a reflection of therapeutics by pairwise sequence comparisons between the two datasets.

For each of the 563 therapeutics we checked if we can find a perfect length-matched hit in PAD. For 546 (96.98%) out of 563 therapeutics we found a perfect length-matched hit in PAD. For the remaining 17 therapeutics without perfect matches, we found that the PAD version used for this study (Jan 2020) was out of date or there existed only high sequence identity matches as compared to Lens.org but not perfect ones (Supplementary Table 4).

For each antibody sequence from a patent, we noted the highest IMGT sequence identity to any therapeutic and present the results stratified by AllPatAb and MedPatAb sequences in Figure 3. Large proportion of PAD sequences align with high sequence identity to one of the 563 therapeutics. Total of 21,772 (16.08%) of heavy chain AllPatAb sequences and 17,378 (18,67%)

347 of heavy chain MedPatAb sequences have matches of 90% sequence identity or better to a
348 therapeutic sequence. Total of 44,919 (40,94%) of light chain AllPatAb sequences and 31,241
349 (46,16%) of light chain MedPatAb sequences have matches of 90% sequence identity or better to
350 a therapeutic sequence. Altogether this illustrates that many sequences in patent documents well
351 reflect the therapeutic antibody sequences currently in the clinical use. However there is also a
352 large number of sequences with matches below 90% sequence identity to either heavy or light
353 therapeutic heavy chain. This could reflect sequences that are only currently in development or
354 never found their way to the clinic as a result of failure, abandonment or otherwise.
355



356 A B

357 **Figure 3.** Closest matches of antibody sequences from patents to therapeutic antibodies. For
358 each sequence in AllPatAb and AllPatMed we noted the closest IMGT sequence identity to a
359 therapeutic antibody. A) Distribution of heavy chain sequence identities to closest therapeutic
360 heavy chain. B) Distribution of light chain sequence identities to closest therapeutic light chain.
361

362 Perfect matches between full variable region PAD sequences and therapeutics implicitly
363 indicates good correspondence in the CDR region. Arguably, the most diverse and thus the most
364 engineered portion of an antibody is its heavy chain CDR3 region, CDR-H3[29,30]. The length of
365 CDR-H3 has been previously shown to be a good estimator of overall developability of an
366 antibody, with therapeutic antibodies having shorter CDR-H3[6]. We contrasted the CDR-H3
367 lengths found in PAD to those in therapeutic, structural and natural human antibodies. We
368 extracted CDR-H3s from antibody structures found in the Protein Data Bank[31] that are regularly
369 collected by the Structural Antibody Database[22] (SAbDab). The natural human antibodies were

370 sourced from a deep Next Generation Sequencing (NGS) study by Briney et al.[8] downloaded
371 from the Observed Antibody Space database[17]. We found a total of 58,383 unique CDR-H3s in
372 all PAD sequences (AllPatAb), 37,247 unique CDR-H3s in antibodies from medicinal patents
373 (MedPatAb), 422 unique CDR-H3s in therapeutics, 2021 unique CDR-H3s in structures and
374 73,217,582 unique CDR-H3s from natural human antibodies. We plotted the distribution of
375 lengths for each of these datasets in Figure 4.
376
377 The distribution of CDR-H3 lengths from patent sequences does not appear to be different
378 between AllPatAb and MedPatAb sequences. Therapeutic CDR-H3s have the shortest median
379 lengths, followed by structures, patents and natural human antibodies. The shorter lengths in
380 structures might be reflective of large number of artificial/therapeutic antibodies that can be
381 found in SAbDab[23]. Lengths of CDR-H3s from patent sequences appear to be mid-range
382 between therapeutic and natural antibodies. This suggests that patent antibody sequences might
383 reflect certain amount of engineering of these molecules as they do not follow the natural
384 distribution, normally favoring longer lengths. Nevertheless patent antibody CDR-H3 do not
385 recapitulate the therapeutic CDR-H3 length distribution. Since vast majority of therapeutic CDRs
386 can be found in sequences from patent documents, the discrepancy with the therapeutic length
387 distribution can suggest certain engineering choices faced by those molecules not revealed in this
388 study.
389

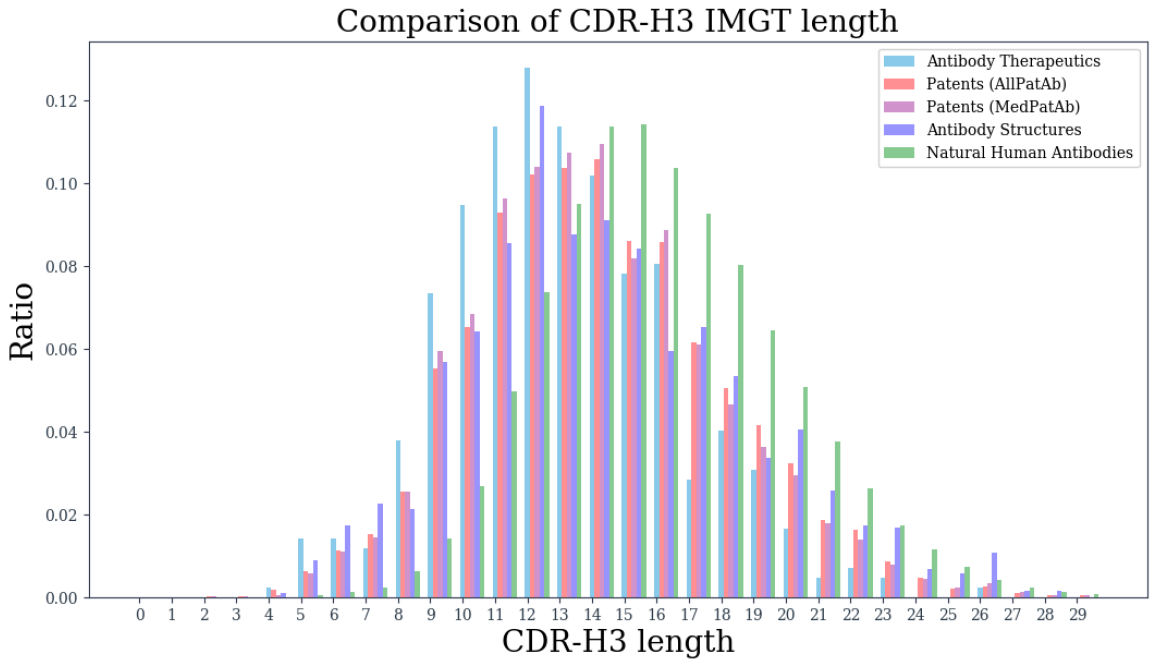


Figure 4. Distribution of CDR-H3 lengths. We plotted the distribution of CDR-H3 lengths from therapeutic antibodies (Antibody Therapeutics), antibodies from patents in PAD (Patents, stratified between AllPatAb and MedPatAb), structures of antibodies from the Protein Data Bank (Antibody Structures) and natural human antibodies from a deep Next Generation Sequencing study (Natural Human Antibodies).

Patent landscape of single domain antibodies.

Our earlier results revealed that majority of antibodies from patents align well to human or mouse germline V region genes, which recapitulates the widespread use of 'traditional' antibody format containing both heavy and light chains. The third most commonly identified organism was alpaca (Table 4), which suggests the single domain antibody (sdAb) format. The single domain antibodies are found naturally in camelids (camels, lamas, alpacas) and because of the lack of light chain are believed to have more favorable biophysical properties than antibodies, without detriment to their antigen recognition ability[32,33]. They have been commercialized as therapeutics by Ablynx under the protected name Nanobody® with first single domain antibody drug, Caplacizumab, recently approved[34]. Allowing the first sdAb drug in clinical use holds the promise of more molecules in this format in the near future[35], which can be reflected by patents.

409

410 We identified the total number of patent families in PAD having sdAbs to quantify the possible
411 number of molecules in this format in development, providing an orthogonal view to currently
412 known therapeutic candidates[35]. Patent families were classified as containing sdAbs if they were
413 classified as C07K2317/569 (Single domain, e.g. dAb, sdAb, VHH, VNAR or nanobody®) or
414 C07K2317/22 (from camelids, e.g. camel, llama or dromedary) or if they contained sequences
415 aligning to alpaca sdAb germlines. Using the classification method we identified 845 families
416 and using the alpaca germline method we found 867 families. There was an overlap between the
417 two, resulting in total of 1,176 families identified as containing sdAbs or 7.11% of all of our
418 16,526 families in PAD. Of the 1,176 families 586 (49.82%) were classified as containing
419 antibodies for therapeutic purposes.

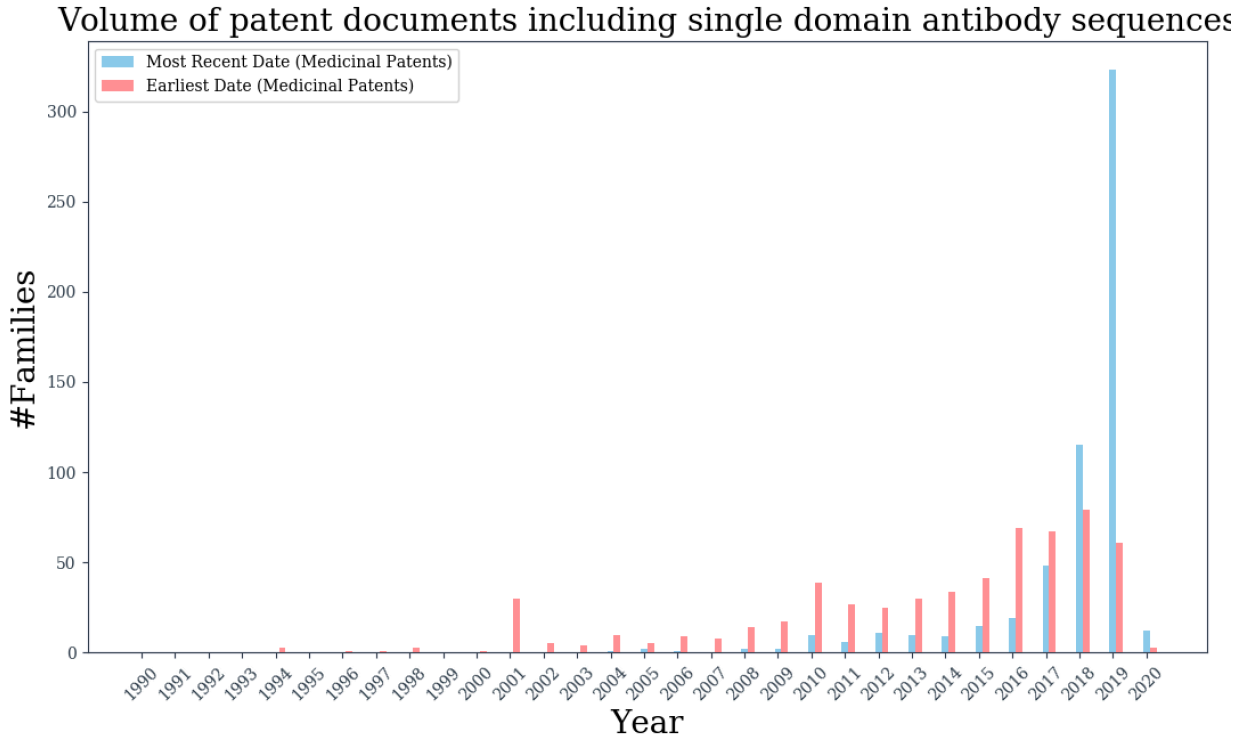
420

421 The top 30 organizations sorted by the number of families containing sdAb sequences
422 (Supplementary Table 5) well reflect the companies developing biotherapeutics in this format[35].
423 The list however contains more organizations than those currently reported as developing sdAb
424 therapies, suggesting that the field might be more nuanced, notwithstanding wide use of sdAbs
425 for imaging and diagnostic purposes[36]. From the list of known sdAb therapeutics, our list does
426 not contain AdAlta and Ossianix that report shark single domain antibodies, sequences of which
427 we do not identify. In fact, not all sdAbs that we identified follow the natural camelid format, as
428 there exist sequences of single domain human antibodies (e.g. US2011097339).

429

430 We checked the total number of sequences in PAD that could be identified as sdAbs. The 1,176
431 patent families that we identified as containing sdAbs hold a total of 48,849 unique heavy chain
432 sequences. Not all of such sequences are sdAbs as the patent document might have included
433 traditional antibodies as well. Therefore we calculated the number of sequences that were
434 identified as alpaca sdAb germlines and sequences found in one of the 1,176 families but
435 containing only heavy chains. We found a total of 12,914 unique sequences aligning to sdAb
436 alpaca germlines and 13,368 unique sequences found in 1,176 sdAb families containing heavy
437 chains only. There was an overlap between the two sequence sets and combining them resulted
438 in a total of 15,792 possible sdAb sequences, which makes up 11,66% of all the 135,397 heavy
439 chain sequences in PAD. Of the 15,792 possible sdAb sequences, 8,342 (52.82%) were found in

440 patent documents classified as containing antibodies for medicinal purposes. Therefore, single
441 domain antibody sequences appear to make up a non-trivial proportion of antibody sequences
442 found in patents which could be indicative of upcoming sdAb clinical trials and approvals.
443
444 In order to provide an indication of the possible activity to come in the field of single domain
445 antibodies, we plotted the earliest and most recent dates associated with any of the 586 sdAb
446 patent families classified as having antibodies for medicinal applications (Figure 5). There
447 appears to be a steady increase in the number of patent documents including sdAb sequences for
448 medicinal purposes (same holds true for all 1,176 patent families containing sdAb sequences,
449 Supplementary Figure 2). Given the steady rise in the number of patents containing sdAbs and
450 recent approval of Caplacizumab, one might expect more molecules in this format in clinical use
451 in the future.
452



453
454 **Figure 5**. Patents including single domain antibody sequences over time. For each of the 586
455 patent families in PAD identified as having sdAbs and classified as containing antibodies for
456 medicinal purposes, we noted the earliest and most recent dates, given as red and blue bars
457 respectively.

458 **Discussion.**

459

460 Successful exploitation of antibodies as therapeutics relies on ever deeper understanding of the
461 biology of these molecules. Many features of therapeutic antibodies can be found in naturally
462 sourced sequences[5], however effective biotherapeutic requires bespoke engineering for clinical
463 safety and developability[2]. We proposed that such biotherapeutic engineering knowledge could
464 be reflected in patent documents containing antibody sequences.

465

466 Our analysis of patents containing antibody sequences revealed that majority of such documents
467 are explicitly developed as containing antibodies for medicinal purposes. Vast majority of
468 therapeutic antibody sequences can be found in patent documents. Further to that, many
469 sequences from patents are within close sequence identity of therapeutic antibodies that are
470 approved or undergoing clinical trials. This suggests that thousands of antibodies from patents
471 could provide a reflection of engineering choices that were made during development therapeutic
472 molecules. Such data could offer an integrated collection of insights into the features that were
473 designed into antibodies to make them successful therapeutics.

474

475 This information could be readily exploited by computational methods[4]. It was previously
476 demonstrated that only 137 Clinical Stage Therapeutic (CST) antibodies can provide insights
477 into developability of these molecules[6]. As demonstrated by our analysis, there is an order of
478 magnitude more patented sequences that are close sequence matches to such CSTs. These could
479 indicate different variants and possible features of biotherapeutics, creating a more wholesome
480 picture of what makes a successful biotherapeutic.

481

482 Employing antibody sequences from patents however is not without its caveats. Unlike academic
483 literature, patent documents are not designed to convey knowledge but rather offer legal
484 protection. This might result in wide claims on sequence identities to proposed antibody variants
485 that could obfuscate the resulting therapeutic sequence. As we demonstrated, antibodies in
486 patents provide a good reflection of the therapeutics either approved or in clinical use. This
487 would suggest that even though claims could be quite wide on sequence space, many of them
488 appear to fall within the sequence identity orbit of currently available therapeutics. Therefore,

489 certain antibody sequences from patents could broadly reflect the engineering choices in the
490 design of these molecules.
491
492 The already large amount of antibodies from patent documents will most likely keep rising, as
493 we demonstrated by the growth in the number of such documents in the recent years. In fact
494 studying such patents could provide an early indication of approvals to come[37]. This might be
495 specifically true in the sphere of single domain antibodies. There is just one such approved
496 therapeutic on the market[34] and ten in clinical trials[35] (in 2019). We find a great number of sdAb
497 patents suggesting that the field might further develop in the near future, providing an alternative
498 to traditional monoclonal antibody therapy.
499
500 The ongoing increase of patents containing antibodies for medicinal indications will keep
501 contributing to an already ample body of knowledge of antibody engineering. This data could be
502 used to offer insights into the engineering choices in designing these molecules, accelerating
503 delivery of biotherapeutics to the clinic.
504
505 **Methods**
506
507 **Identifying antibody sequences in patent documents.**
508
509 Raw biological sequence data associated with patent documents was downloaded from four
510 freely available accessible services: the United States Patent and Trademark Office (USPTO,
511 https://www.uspto.gov/), the DNA Data Bank of Japan (DDBJ)[12], European Bioinformatics
512 Institute (EBI)[13] and World Intellectual Property Organization (WIPO, https://www.wipo.int/).
513 The USPTO data were divided between the full text submissions (https://bulkdata.uspto.gov/)
514 and lengthy sequence listings (http://seqdata.uspto.gov/). Using a custom Python script, the
515 USPTO full text submissions were scanned for nucleotide or amino acid sequences and listings
516 containing these, whereas USPTO PSIPS contained sequence listings only. Using a custom
517 Python script the WIPO FTP documents (ftp://tp.wipo.int/pub/published_pct_sequences) were
518 scanned for nucleotide and amino acid sequences. In both cases of USPTO and WIPO,
519 differences in sequence listing formats from different time periods was accounted for by

520 developing a custom Python parser for each case, transferring all the raw sequences and their
521 associated patent numbers into FASTA format. Data from DDBJ and EBI are available through
522 their ftp services (ftp://ftp.ddbj.nig.ac.jp/ddbj_database and ftp://ftp.ebi.ac.uk/pub/databases
523 respectively) and were readily available in FASTA format.
524
525 The nucleotide entries were scanned for antibody sequence by using IGBLAST[16] as described
526 previously[17], and their amino acid translations were noted. The raw amino acid sequences were
527 scanned for presence of antibodies using ANARCI[18]. We only kept those amino acid sequences
528 where all three CDR regions and all four framework regions could be identified and that
529 contained only 20 canonical amino acids. This resulted in a dataset of IMGT-numbered amino
530 acid sequences, associated with their patent numbers.
531
532 **Patent metadata acquisition and antibody target identification.**
533
534 Different patent numbers can point to the same document, submitted across several jurisdictions,
535 termed 'patent family'. For each patent number associated with a sequence we identified the
536 patent family by using the Open Patent Services API v. 3.2 (developers.epo.org/ops-v3-2). Using
537 the Open Patent Services API, we downloaded the metadata associated with each family which
538 included: family identifier, title, description, patent numbers with associated dates and
539 applicants.
540
541 The patent metadata was used for antibody target identification. Even though there exist certain
542 CPC classifications indicating what the antibody should bind to, we noted that they were not
543 universally present. Therefore we performed manual target annotation, supported by Named
544 Entity Recognition (NER). We applied the GENIA NER[38] parser to the titles and abstracts of
545 patent families. As with scientific publications titles and abstracts can be expected to reflect the
546 most important content of the document[39], in particular pertaining to the binding mode of the
547 reported antibody. The resulting annotations accelerated the manual process of annotating each
548 of our patent families with possible targets.
549
550

551 **Web Service.**

552

553 We make the data accessible for academic non-commercial use via web service accessible at
554 http://naturalantibody.com/pad. Users can search for antibody sequences by pasting the amino
555 acids of the variable domains. The input sequence is IMGT-numbered. The sequences in PAD
556 are IMGT-aligned to the input sequence and the top 50 best sequence identity matches are
557 displayed.

558

559 **Disclosure Statement.**

560

561 The authors declare no conflict of interest

562

563 **References**

564

565 1. Kaplon H, Muralidharan M, Schneider Z, Reichert JM. Antibodies to watch in 2020.
566 MAbs 2020; doi: 10.1080/19420862.2019.1703531
567 2. Jain T, Sun T, Durand S, Hall A, Houston NR, Nett JH, Sharkey B, Bobrowicz B, Caffry
568 I, Yu Y, et al. Biophysical properties of the clinical-stage antibody landscape. Proc Natl
569 Acad Sci U S A 2017; doi: 10.1073/pnas.1616408114
570 3. Mason DM, Friedensohn S, Weber CR, Jordi C, Wagner B, Meng S, Reddy ST. Deep
571 learning enables therapeutic antibody optimization in mammalian cells. bioRxiv 2019;
572 doi: 10.1101/617860
573 4. Norman RA, Ambrosetti F, Bonvin AMJJ, Colwell LJ, Kelm S, Kumar S, Krawczyk K.
574 Computational approaches to therapeutic antibody design: established methods and
575 emerging trends. Brief Bioinform 2019; doi: 10.1093/bib/bbz095
576 5. Krawczyk K, Raybould MIJ, Kovaltsuk A, Deane CM. Looking for therapeutic antibodies
577 in next-generation sequencing repositories. MAbs 2019; doi:
578 10.1080/19420862.2019.1633884
579 6. Raybould MIJ, Marks C, Krawczyk K, Taddese B, Nowak J, Lewis AP, Bujotzek A, Shi
580 J, Deane CM. Five computational developability guidelines for therapeutic antibody
581 profiling. Proc Natl Acad Sci U S A 2019; doi: 10.1073/pnas.1810576116

582 7. Koenig P, Lee C V, Walters BT, Janakiraman V, Stinson J, Patapoff TW, Fuh G, Wilson
583 IA. Mutational landscape of antibody variable domains reveals a switch modulating the
584 interdomain conformational dynamics and antigen binding. Proc Natl Acad Sci U S A .
585 2017 doi: 10.1073/pnas.1613231114
586 8. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity
587 in the baseline human antibody repertoire. Nature 2019; doi: 10.1038/s41586-019-0879-y
588 9. Petering J, McManamny P, Honeyman J. Antibody therapeutics - the evolving patent
589 landscape. N. Biotechnol. 2011; doi: 10.1016/j.nbt.2011.03.023
590 10. Dumet C, Pottier J, Gouilleux-Gruart V, Watier H. Insights into the IgG heavy chain
591 engineering patent landscape as applied to IgG4 antibody development. MAbs 2019; doi:
592 10.1080/19420862.2019.1664365
593 11. Cole P. Patentability of genes: A European union perspective. Cold Spring Harb Perspect
594 Med 2015; doi: 10.1101/cshperspect.a020891
595 12. Tateno Y. DNA Data Bank of Japan (DDBJ) for genome scale research in life science.
596 Nucleic Acids Res 2002; doi: 10.1093/nar/gks1195
597 13. Li W, Mcwilliam H, de la Torre AR, Grodowski A, Benediktovich I, Goujon M, Nauche
598 S, Lopez R. Non-redundant patent sequence databases with value-added annotations at
599 two levels. Nucleic Acids Res 2009; doi: 10.1093/nar/gkp960
600 14. Jefferson OA, Köllhofer D, Ehrich TH, Jefferson RA. Transparency tools in gene
601 patenting for informing policy and practice. Nat Biotechnol 2013; doi: 10.1038/nbt.2755
602 15. Shibata S, Uemura R, Suzuki T. Comparative Analysis Between the Top-Selling Drugs in
603 the Japanese Pharmaceutical Market and Those in the United States, the United Kingdom,
604 France, and Germany. Ther Innov Regul Sci 2016; doi: 10.1177/2168479015604182
605 16. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain
606 sequence analysis tool. Nucleic Acids Res 2013; doi: 10.1093/nar/gkt382
607 17. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. Observed Antibody
608 Space: A Resource for Data Mining Next-Generation Sequencing of Antibody
609 Repertoires. J Immunol 2018; doi: 10.4049/jimmunol.1800708
610 18. Dunbar J, Deane CM. ANARCI: Antigen receptor numbering and receptor classification.
611 Bioinformatics 2015; doi: 10.1093/bioinformatics/btv552
612 19. Jones TD, Carter PJ, Plückthun A, Vásquez M, Holgate RGE, Hötzel I, Popplewell AG,

613 Parren PWHI, Enzelberger M, Rademaker HJ, et al. The INNs and outs of antibody
614 nonproprietary names. MAbs 2016; doi: 10.1080/19420862.2015.1114320
615 20. International nonproprietary names for pharmaceutical substances (INN): proposed INN:
616 list 122. WHO Drug Inf 2019; Available From
617 https://www.who.int/medicines/publications/druginformation/innlists/PL122.pdf
618 21. Poiron C, Wu Y, Ginestoux C, Ehrenmann F, Duroux P, Lefranc M-P. IMGT/mAb-DB:
619 the basis of IMGT data of therapeutic monoclonal antibodies. Bull Cancer 2010;
620 Available from http://www.imgt.org/mAb-DB/
621 22. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM.
622 SAbDab: the structural antibody database. Nucleic Acids Res 2013; doi:
623 10.1093/nar/gkt1043
624 23. Raybould MIJ, Marks C, Lewis AP, Shi J, Bujotzek A, Taddese B, Deane CM. Thera-
625 SAbDab: the Therapeutic Structural Antibody Database. Nucleic Acids Res 2019; doi:
626 10.1093/nar/gkz827
627 24. Walsh G. Biopharmaceutical benchmarks 2010. Nat Biotechnol 2010; doi:
628 10.1038/nbt0910-917
629 25. Almagro JC, Fransson J. Humanization of antibodies. Front Biosci 2008; Available from
630 https://www.bioscience.org/2008/v13/af/2786/fulltext.htm
631 26. Almagro JC, Pedraza-Escalona M, Arrieta HI, Pérez-Tapia SM. Phage Display Libraries
632 for Antibody Therapeutic Discovery and Development. Antibodies 2019; doi:
633 10.3390/antib8030044
634 27. Lee C V., Liang WC, Dennis MS, Eigenbrot C, Sidhu SS, Fuh G. High-affinity human
635 antibodies from phage-displayed synthetic Fab libraries with a single framework scaffold.
636 J Mol Biol 2004; doi: 10.1016/j.jmb.2004.05.051
637 28. Lefranc M paule, Giudicelli V, Regnier L, Duroux P. IMGT, a system and an ontology
638 that bridge biological and computational spheres in bioinformatics. Brief Bioinform 2008;
639 doi: 10.1093/bib/bbn014
640 29. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody
641 specificities. Immunity 2000; 13:37–45. doi: 10.1016/s1074-7613(00)00006-6
642 30. Knappik A, Ge L, Honegger A, Pack P, Fischer M, Wellnhofer G, Hoess A, Wölle J,
643 Plückthun A, Virnekäs B. Fully synthetic human combinatorial antibody libraries

(HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. J Mol Biol 2000; doi: 10.1006/jmbi.1999.3444

31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000; doi: 10.1093/nar/28.1.235

32. Muyldermans S, Cambillau C, Wyns L. Recognition of antigens by single-domain antibody fragments: The superfluous luxury of paired domains. Trends Biochem. Sci.2001; doi: 10.1016/s0968-0004(01)01790-x

33. Bannas P, Hambach J, Koch-Nolte F. Nanobodies and nanobody-based human heavy chain antibodies as antitumor therapeutics. Front. Immunol.2017; doi: 10.3389/fimmu.2017.01603

34. Duggan S. Caplacizumab: First Global Approval. Drugs 2018; doi: 10.1007/s40265-018-0989-0

35. Morrison C. Nanobody approval gives domain antibodies a boost. Nat. Rev. Drug Discov.2019; doi: 10.1038/d41573-019-00104-w

36. De Meyer T, Muyldermans S, Depicker A. Nanobody-based products as research and diagnostic tools. Trends Biotechnol 2014; doi: 10.1016/j.tibtech.2014.03.001

37. Pereira CG, Lavoie JR, Garces E, Basso F, Dabić M, Porto GS, Daim T. Forecasting of emerging therapeutic monoclonal antibodies patents based on a decision model. Technol Forecast Soc Change 2019; doi: 10.1016/j.techfore.2018.11.002

38. Tsuruoka Y, Tsujii J. Bidirectional inference with the easiest-first strategy for tagging sequence data. In: HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. 2005; Available from https://www.aclweb.org/anthology/H05-1059.pdf

39. Volanakis A, Krawczyk K. SciRide Finder: A citation-based paradigm in biomedical literature search. Sci Rep 2018; doi: 10.1038/s41598-018-24571-0