GQ Life Sciences

FREE VS PROFESSIONAL GRADE IP SEQUENCE SEARCHES

Content is King

If you're looking to protect your own sequences or want to make sure you're not infringing on anyone else's IP, then you need to review what's already out there. Anyone who has done an IP sequence search before can tell you that it's not an easy thing to get right. You want to make sure that you cover all the relevant data, that these data are searched in the right way, and that you have an efficient way to handle the results. You also don't want to accidentally disclose information that should stay confidential. Failure to bring these things together in the right way can impact search results, conclusions, and ultimately lead to flawed business decisions. Let's look at some of the common pitfalls in IP sequence searching and how to overcome them.

A free IP sequence search usually involves the following steps: search the GenBank patent divisions on the NCBI BLAST website, go through the alignments one by one, and look up related patent information on the web. Findings are tracked on a printout of the BLAST results or in a spreadsheet, which is very inefficient, and as explained below, certainly not comprehensive.

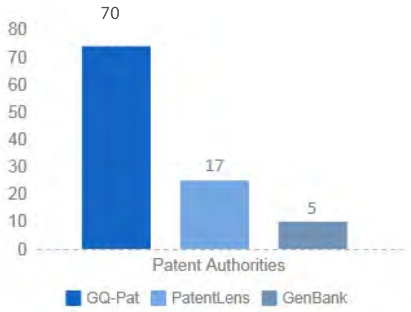
PatentLens searching is much friendlier, and much more information can be extracted, but its filtering, analysis and reporting capabilities still fall far short of those found in GenomeQuest® and its coverage, although more complete than GenBank, is still incomplete.

Search Everything You Can

Let's begin with an overview of the number of countries covered. Neither GenBank nor PatentLens approach GQ-Pat in global coverage.

Now let's look at these two free alternatives to GQ-Pat in a bit more detail:

Figure 1 - Patent Authorities Covered



GenBank Patent Division

The biggest challenge in IP sequence searching is finding a reliable, complete, and up-to-date source of patent information.

As of June, 2017 , the GenBank patent division (GBPAT) contains about 41 million sequences. As a comparison, GQ-Pat, our IP-related sequence database, contains over 370 million sequences. That is almost tenfold as many sequences!

The GenBank patent division mostly consists of documents filed in the US, Europe, and Japan. It is updated once every two months. It is also noteworthy that it contains only US granted patents – there are NO US PATENT APPLICATIONS IN GENBANK PATENT DIVISION[1,2] . Furthermore, it contains no Japanese patent protein sequences, only nucleotides.

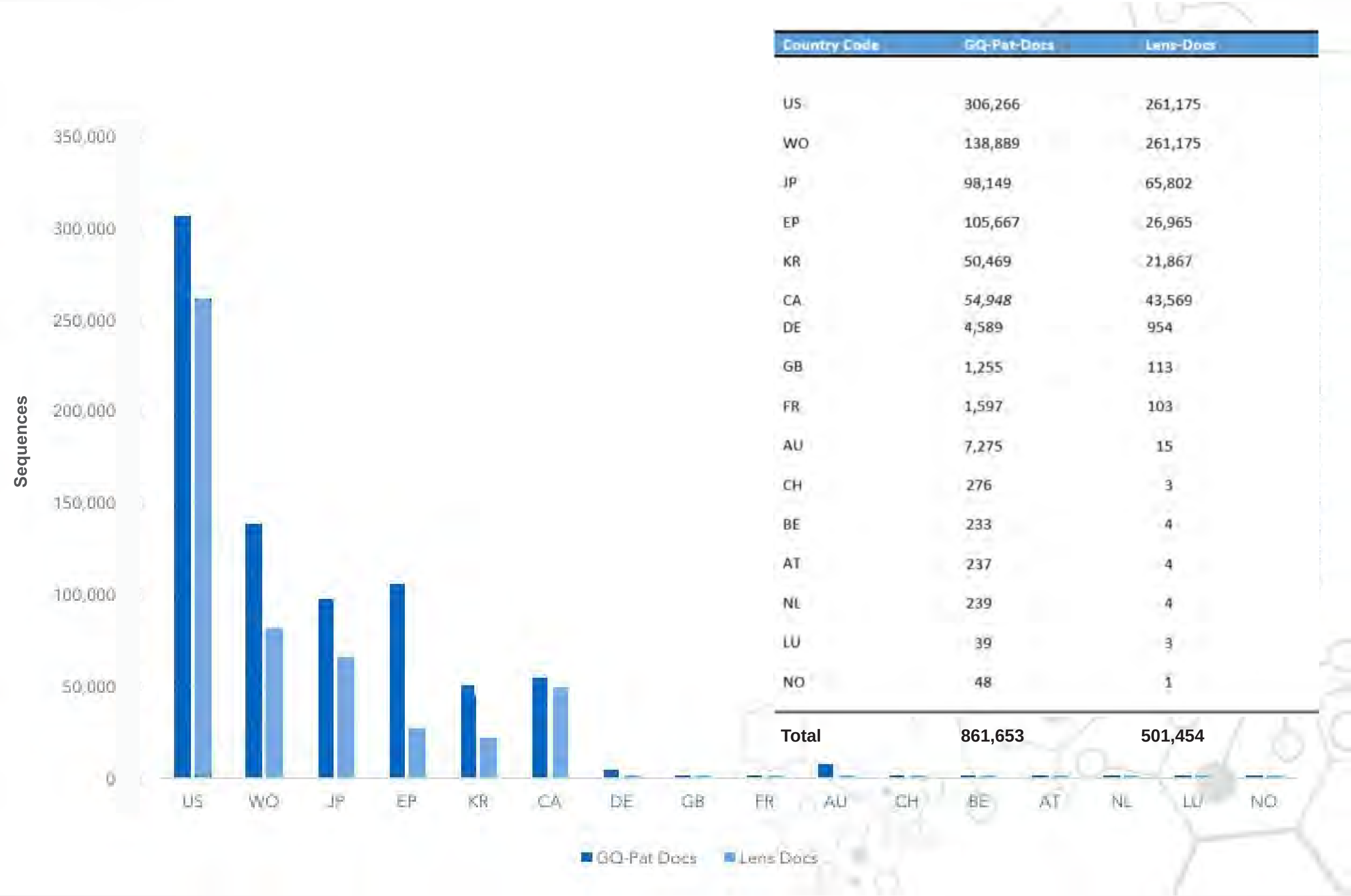
In contrast, our database (GQ-Pat) is continuously updated with data streams from patent offices all over the world, including the US, Europe, China, Brazil, India, and the WIPO/PCT offices, for a total of 70 different patent authorities! Of course, it contains US applications, and Japanese patent protein and DNA sequences. And GQ-Pat also includes the appropriate parts of databases like GenBank, EMBL, and DDBJ.

In comparison to GQ-Pat, the GenBank patent division is a very incomplete source of information. You should think twice before using it to answer business critical questions.

PatentLens

The Lens is very impressive sizewise, with slightly more than 269 million sequences[3], about 73% of the size of GQ-Pat. But—let's really compare. First, patent documents:

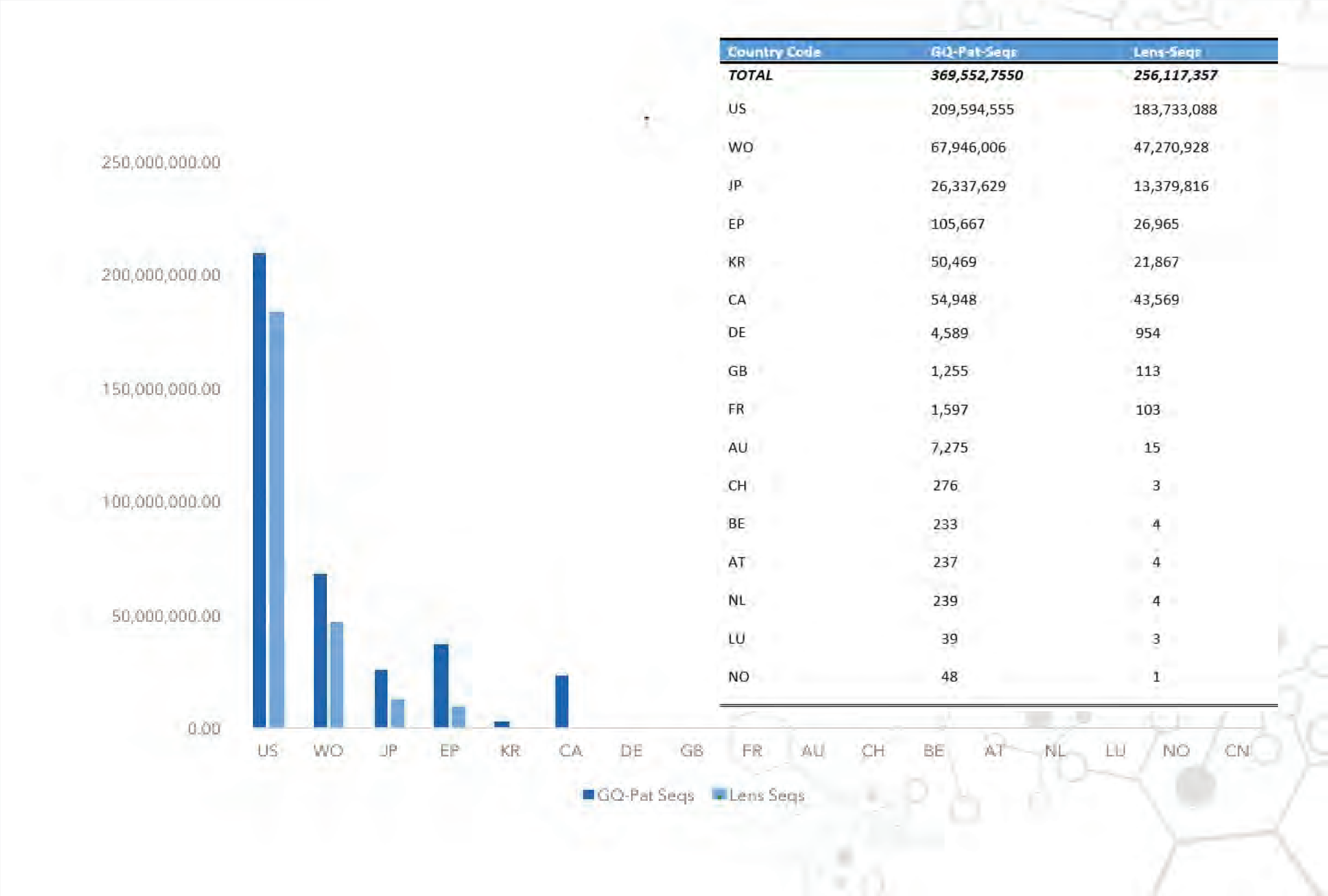
Figure 2 – Document Count Comparison
GQ-Pat to PatentLens



Country Code	GQ-Pat-Docs	Lens-Docs
US	306,266	261,175
WO	138,889	261,175
JP	98,149	65,802
EP	105,667	26,965
KR	50,469	21,867
CA	*54,948*	43,569
DE	4,589	954
GB	1,255	113
FR	1,597	103
AU	7,275	15
CH	276	3
BE	233	4
AT	237	4
NL	239	4
LU	39	3
NO	48	1
Total	**861,653**	**501,454**

Next, let’s compare the number of sequences in each of these databases:

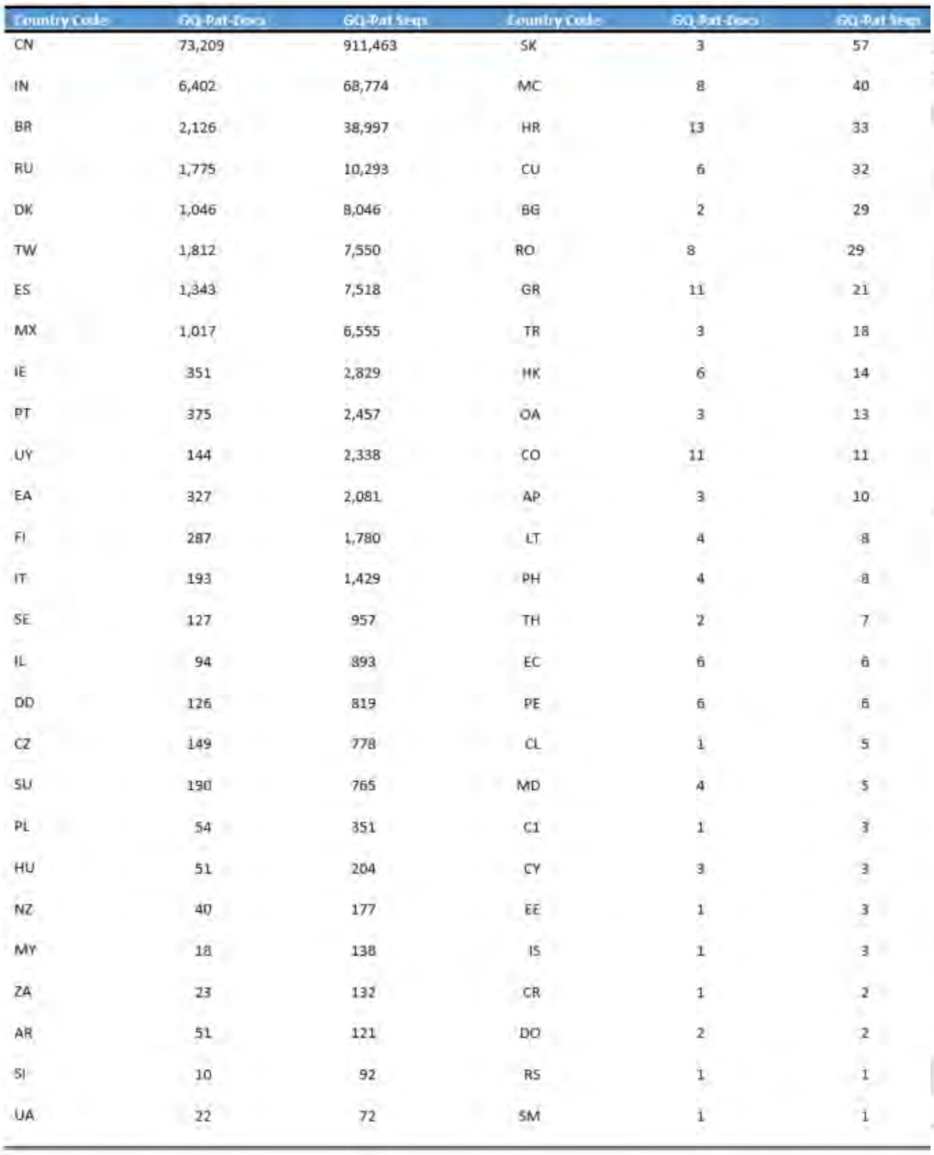
Figure 3 – Sequence Count Comparison
GQ-Pat to PatentLens



Country Code	GQ-Pat-Seqs	Lens-Seqs
TOTAL	***369,552,7550***	***256,117,357***
US	209,594,555	183,733,088
WO	67,946,006	47,270,928
JP	26,337,629	13,379,816
EP	105,667	26,965
KR	50,469	21,867
CA	54,948	43,569
DE	4,589	954
GB	1,255	113
FR	1,597	103
AU	7,275	15
CH	276	3
BE	233	4
AT	237	4
NL	239	4
LU	39	3
NO	48	1

And finally, here is a list of the patent authorities covered in GQ-Pat, but not in either Genbank-Pat or in Patent Lens, all 54 of them!

Figure 4 – GQ-Pat Additional Country Coverage



Country Code	GQ-Pat-Docs	GQ-Pat Seqs	Country Code	GQ-Pat-Docs	GQ-Pat Seqs
CN	73,209	911,463	SK	3	57
IN	6,402	68,774	MC	8	40
BR	2,126	38,997	HR	13	33
RU	1,775	10,293	CU	6	32
DK	1,046	8,046	BG	2	29
TW	1,812	7,550	RO	8	29
ES	1,343	7,518	GR	11	21
MX	1,017	6,555	TR	3	18
IE	351	2,829	HK	6	14
PT	375	2,457	OA	3	13
UY	144	2,338	CO	11	11
EA	327	2,081	AP	3	10
FI	287	1,780	LT	4	8
IT	193	1,429	PH	4	8
SE	127	957	TH	2	7
IL	94	893	EC	6	6
DD	126	819	PE	6	6
CZ	149	778	CL	1	5
SU	190	765	MD	4	5
PL	54	351	C1	1	3
HU	51	204	CY	3	3
NZ	40	177	EE	1	3
MY	18	138	IS	1	3
ZA	23	132	CR	1	2
AR	51	121	DO	2	2
SI	10	92	RS	1	1
UA	22	72	SM	1	1

Data Quality

Many of the self-submission databases are known for high error rates. Sequence data is error-prone, there's no getting around that! Data derived from the same sequence listing filed with a patent authority is pretty straightforward; however, there are many, many more sequences that are NOT in the sequence listings, which are either directly submitted to Genbank, or which are hidden in the text of the patent and have to be manually extracted. Manual curation, as we call it, is expensive, but it's essential to obtain sequences not otherwise available – which are just as valid as prior art or legitimately claimed as those sequences in listings. And these sequences won't be found in Lens or GenBank Patent.

There's also a known issue with SEQ ID NO misnumbering from sequences originating at the JPO[4]. This problem was discovered here at GQ Life Sciences in 2013, and we've put measures in place to correct the SEQ ID Nos in GQ-Pat. Neither GenBank nor PatentLens has made this problem or its fix evident; if you are unaware of it, you will be reporting mismatched sequences and SEQ ID Numbers.

Keep Things Confidential

A discussion of using paid vs free databases wouldn't be complete without mention of confidentiality. At GQ Life Sciences we understand the confidential nature of IP searches. All submitted data is handled and stored on a secure private network. The communication between your browser and our servers is fully encrypted, and your user account is protected by a password. This means that you are the only one able to see your data, unless you explicitly share it with someone else in your organization. The same cannot be said for a public service like NCBI BLAST, where none of the communication is encrypted, your queries are logged, and anyone with the right URL can see your data.

Conclusions

Searching through sequence-related IP is a precise task. To do it right requires a comprehensive and updated database covering applications and patents from as many sources as possible. It requires using the right algorithms and parameters for the job. It also requires an efficient way to go from a list of search results to precise answers to the questions being asked. Any shortcut, incomplete database, undetected data anomaly, or incomplete solution is very likely to influence the outcome of a search and can easily shift the conclusions and your company's IP strategy. And after all, what is a searcher's biggest fear? Missing a key hit! And if a sequence isn't in the database you search, then it will be almost impossible to find – similar to using a magnifying glass to find a single lactobacillus in an entire barn!

Over the last 10 years GenomeQuest has been used by almost all of the biggest pharmaceutical, biotech, and agricultural companies in the world, by specialized law firms, and even by patent offices themselves. We would love to help you protect your IP as well.

References:
1. Genbank patent https://www.ncbi.nlm.nih.gov/education/patent_and_ip_faqs/
2. Jefferson, Osmat et al, Nature Biotechnology 31, 1086–1093 (2013) http://www.nature.com/nbt/journal/v31/n12/full/nbt.2755.html
3. Patent Lens https://www.lens.org/lens/bio/patseqdata#table/
4. Sherin, Ellen; Bayada, Denis. Erroneous SEQ ID Numbers in Major Public Sequence Databases, PIUG Biotechnology Conference 2014.) https://www.piug.org/biotech14program#Sherin

CONTACT

For more information on how GQ Life Sciences can help you with IP Searches, contact us at: www.gqlifesciences.com.

GQ Life Sciences
4325 Alexander Drive Suite 100
Alpharetta, GA 30022