Transparency tools in gene patenting for informing policy and practice

Osmat A Jefferson, Deniz Köllhofer, Thomas H Ehrich & Richard A Jefferson

The Supreme Court's decision in *Myriad* highlights the need for tools enabling nuanced and precise analysis of gene patents at the global level.

In the recent decision *Association for Molecular Pathology v. Myriad Genetics*[1], the US Supreme Court held that naturally occurring sequences from human genomic DNA are not patentable subject matter. Only certain complementary DNAs (cDNA), modified sequences and methods to use sequences are potentially patentable. It is likely that this distinction will hold for all DNA sequences, whether animal, plant or microbial[2]. However, it is not clear whether this means that other naturally occurring informational molecules, such as polypeptides (proteins) or polysaccharides, will also be excluded from patents.

The decision underscores a pressing need for precise analysis of patents that disclose and reference genetic sequences, especially in the claims. Similarly, data sets, standards compliance and analytical tools must be improved—in particular, data sets and analytical tools must be made openly accessible—in order to provide a basis for effective decision making and policy setting to support biological innovation. Here, we present a web-based platform that allows such data aggregation, analysis and visualization in an open, shareable facility. To demonstrate the potential for the extension of this platform to global patent jurisdictions, we discuss the results of a global survey of patent offices that shows that much progress is still needed in making these data freely available for aggregation in the first place.

Osmat A. Jefferson, Deniz Köllhofer, Thomas H. Ehrich and Richard A. Jefferson are at Cambia, Canberra, Australia, and Queensland University of Technology, Brisbane, Australia.
e-mail: osmat@cambia.org and raj@cambia.org

Mapping the patent landscape

There have been numerous studies and publications about the scope, social or economic impact, and policy and practice implications of patenting of biological sequences—commonly known as 'gene patenting'[3–6]. Many of these studies contain incomplete data sets, use analytical tools that cannot distinguish the nature of the sequence similarities or fail to parse and analyze patent claims. Few of these studies make the primary data available in a form allowing review by others. In an effort to move beyond opinion pieces and to provide a facility that can be used to ask and answer specific questions in an open, verifiable manner, we have created a biological facility (http://www.lens.org/lens/biological_search) as a public resource within 'The Lens,' an open, global cyber infrastructure dedicated to increasing the efficiency and fairness of the innovation system by making access to patent documents more transparent and inclusive. We have used this facility to create tools that allow for dynamic mapping and shared analysis of the scope of patenting over several genomes, beginning with the human genome.

The single most important consideration in gene patenting is the critical difference between disclosure of sequences and claiming of sequences. Before the recent decision in *Myriad*, the literature on gene patenting led to more confusion than insight. Although some claim that the concerns about gene patents were exaggerated and based on outliers and wrong perceptions[5,7,8], others maintain that patent protection for genetic sequences was excessive and led to obvious inventions, questionable patents and opaque innovation systems that may have harmed the integrity of the market and constrained scientific progress[9–13].

Within 'The Biological Lens' facility, the sequence database currently holds 147,565,858 million nucleotide and amino-acid sequences disclosed in 323,721 global patent documents comprising both applications and grants. Of these sequences, 67% are repeated at least once in the corpus. Some level of redundancy is to be expected, as the same sequence may be either referenced in a single patent document for different purposes or mentioned in many related or unrelated patent documents. Although a majority of patent documents list only one or a few sequences, a substantial number list thousands or even millions of sequences. For example, US Pat. No. 7,777,022 discloses 4.2 million sequences. As millions more sequences become available, patent offices face a difficult challenge to render that information accessible to and useable by the public.

Major patent offices claim to have sophisticated search tools and databases that likely comprise a very substantial set of sequences; however, information about the effectiveness of these algorithms and the scope of these sequence databases are not generally available to the public, and they may even be off limits to the dozens of patent offices in jurisdictions with emerging intellectual property (IP) protection or with limited budgets. Some commercial vendors claim to offer comprehensive data and sophisticated analysis, but this is an expensive means of accessing what is fundamentally public information, and provides one of many entry barriers that disadvantage small-to-medium enterprises (SMEs) and innovation-focused and impact-driven public sector and philanthropy. In addition, these commercial databases are incomplete[14]. For example, the millions of sequences published in the US Patent and Trademark Office's (USPTO; Washington, DC) Patent Applications since 2001 are not incorporated within GenBank

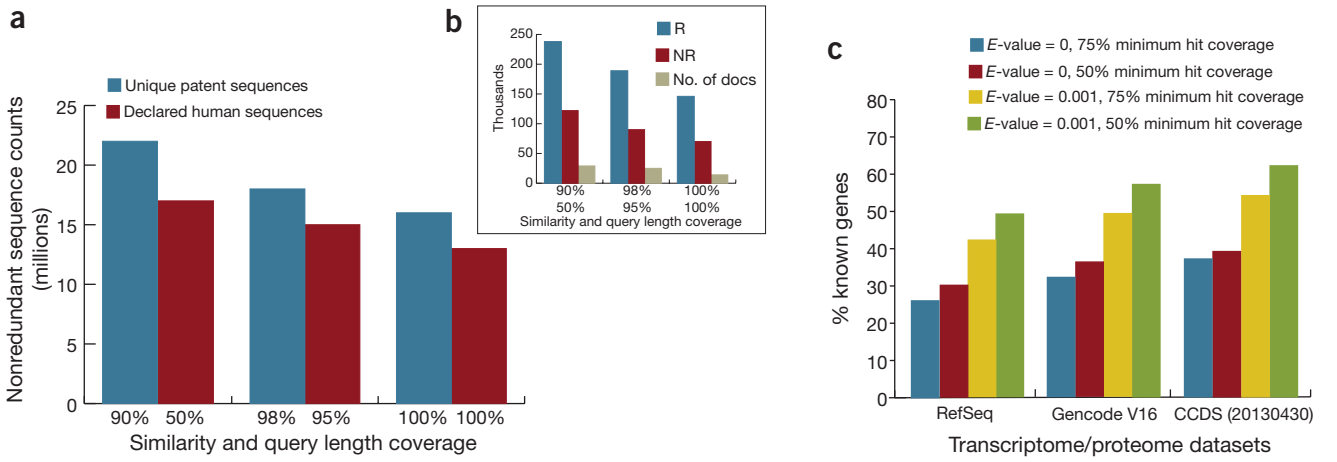

Figure 1 Patent sequences mapped on the human genome (GRCh37 at http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/index.shtml). (**a**) Mapping was based on various similarity and query length coverage rates (90% 50%, 98% 95%, and 100% 100%). Unique patent sequences refer to sequences with only unique mapping locus. Although the majority of nonredundant sequences were declared human in the patent documents, around 20% were unspecified or nonhuman. (**b**) The internal chart shows only mapped sequences that are referenced in the granted claims (1% of the data); redundant sequence counts (R), nonredundant sequence counts (NR) and their corresponding patent grants counts. (**c**) Homology-based human transcriptome and proteome analysis based on two filters of *E*-value and percentage of minimum hit coverage.

(http://www.ncbi.nlm.nih.gov/genbank/), or in any other global public facility, yet as published data they must clearly be considered as potential prior art. This lack of access is also problematic for patent applicants, who may not know whether the sequence for which they seek protection has been previously claimed or not.

To provide a basis for better understanding the complex landscape of gene patenting, we have mapped patent-disclosed sequences onto the human genome and developed a patent-sequence (PatSeq) toolkit to find, align, browse and explore these sequences. To illuminate the scope of patenting of known genes on the human genome, we selected those mapped sequences referenced in granted claims (GC) of the USPTO and performed homology-based analysis with three publicly available transcriptome or proteome data sets: RefSeq[15], GENCODE[16] and Ensembl's Consensus CDS[17]. We found that the percentage of known genes referenced—not necessarily claimed—ranges from 26% to 62%, depending on the reference data set and the homology threshold chosen.

Mapping of patent-disclosed sequences onto the human genome

Because many patent claims provide rights over sequences with as little as 70% identity to a disclosed sequence (for example, US 7,229,976, Claim 1 or US 7,919,474, Claim 2), we selected a range of homology thresholds to determine alignment and location of candidate sequences on the human genome. Homology thresholds were specified by two metrics: patent sequence similarity, and coverage in proportion to the sequence length. The similarity rate reflects the number of matching nucleotides between the patent sequence, and the reference genome and the sequence coverage reflects the proportion of the patent sequence that was included in the alignment. Because of the high repeat rate in the sequence listing corpus, a non-redundant data set of patent sequences was used for the mapping against the reference human genome (assembly GRCh37) of the GRC (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/index.shtml). For mapping highly homologous genomic sequences, we used the Burrows-Wheeler Aligner suite[18]. Potential mRNA and protein sequences were mapped to the reference genome using BLAT[19].

Under stringent conditions (100% similarity and 100% coverage rate), 15.6 million sequences were matched to the human genome at one locus or more (**Fig. 1a**). These correspond to 31.4 million sequence listing entries after reintroduction of the redundancy in the corpus. Although the majority of these sequences were declared within the patent as of human origin (**Fig. 1a**), ~20% were unspecified or annotated as unknown, artificial or sequences derived from other organisms. In the granted patents, 131,339 nucleotide and 15,054 amino acid sequence listing entries were referenced in the claims of 13,985 US patent documents by August 2012 (**Fig. 1b**).

Both sequence listing entries were then compared against three public transcriptome and proteome data sets: RefSeq (24,592 genes), GENCODE v16 (20,564 genes) and the April 30, 2013 release of CCDS (18,552 genes). The analyses revealed that 26% to 62% of known human genes are referenced in the granted claims. We obtained this wide range mainly because of the chosen Expect value (*E* value) and minimum hit coverage percentage and the data set used for comparison (**Fig. 1c**). For example, under our most stringent condition (75% minimum hit coverage and *E* value of 0), the percentage of known genes was calculated as 26% based on RefSeq, 32% based on GENENCODE v16, or 37% based on CCDS data set, whereas with 50% minimum hit coverage and 0.001 *E* value, the percentage of known genes reached 49%, 57% or 62%, respectively.

A 2005 paper by Jensen and Murray[20] forms the basis for the widespread assertion that 20% of the human genome had been patented before *Myriad*[20,21]. But as Holman observed[5], Jensen and Murray's 20% coverage conflated genetic sequences that were merely referenced in patent claims with genetic sequences that were explicitly claimed. In 2008, Cook-Deegan provided a more conservative estimate that ~3,000 to 5,000 human genes had been patented in the United States[6]. How can we differentiate disclosed from claimed sequences and provide public tools to clear perceptions and enable navigation of complex gene patents?

Claimed versus disclosed sequences

After optimizing and extending the algorithms, which select patent documents that reference a sequence in the claims[22], typically though not exclusively as 'SEQ ID NO', we analyzed initially the claims associated with the fully aligned 131,339 nucleotide sequence entries referenced in 2,716 patents. We found that 76,910 sequences mapped uniquely with 100% homology to the human genome and corresponded to 2,685 patents, whereas

the remaining 54,829 sequences were simply repeated in duplicated versions of patent documents. Claims referencing the unique sequences were then individually and manually analyzed. Analysis of grants that reference the amino-acid sequence entries in the claims and that may potentially encompass additional nucleotide sequences were not included in this analysis, but will be shared online in an upcoming patent landscape on our website.

Under *Myriad*, these fully aligned sequences would all be considered natural nucleotide sequences from the human genome, and thus potentially nonpatentable as they do not include those with excised introns (cDNAs) in the patent sequence.

We categorized the sequences in the claims on the basis of the role of that sequence in the claim. We created ten claim categories; their detailed description and distribution by patent as well as by sequence are depicted in **Figure 2**. Among granted US patents, the distribution is fairly even (**Fig. 2a**). Patents that actually claim, as opposed to merely reference, the sequence, comprise the largest single category of patents analyzed, but this accounts for only a third of all documents. The remaining two-thirds of the patents analyzed use the referenced sequences in a claimed method, or claim them in combination with other sequences or compositions, but do not claim the sequences by themselves.

The distribution of sequences per category revealed that only 13% of the examined sequences are actually claimed as sequence (**Fig. 2b**). Some categories show a relatively higher percentage of sequences referenced because of a small number of patents that each reference (but do not directly claim) a very large number of sequences in the claims. For example, 40% of the patent sequences were categorized as 'Comparison/target,' largely because of the effect of a single patent, US 7,510,834. Claim 1 of this patent references 27,088 sequences, of which 23,820 mapped with 100% homology to the human genome and so were included in our analysis. Similarly, many of the sequences categorized as 'Subpart' and 'Alter phenotype' come from the claims of three related patents: US 6,936,467, US 7,226,785 and US 7,258,854. Each of these three patents reference the same series of 4,192 sequences in the claims, of which 4,188 mapped with 100% homology to the human genome and so were included in our analysis.

To illustrate the complex and evolving dynamic in patent sequence claiming, we plotted the major categories as they have risen and fallen over the past 20 years using either publication date (**Fig. 2c**) or filing date (**Fig. 2d**). As speculated in the literature[23–25], we found that most sequences were claimed around the time when the human genome sequencing project was being completed but just before the public release of its complete sequence in GenBank, after which the number of claims made was dramatically reduced.

We were particularly interested in the 'Sequence claimed' category because this category consists of claims that are potentially invalid in view of *Myriad*. As noted above, in this study we only looked at sequences that map completely to the genome with no gaps, meaning that none of the sequences we looked at fall within the Supreme Court's exception for cDNA that spans an intron. The examined sequences in this category were often claimed as primer sequences or probe sequences, but the putative use of a claimed sequence would not have mattered in an infringement context before *Myriad*, and now it may not matter for purposes of validity. Similarly, many of the sequences are antisense sequences that are probably ineligible for patenting as the Supreme Court made no distinction between forward-reading and antisense sequences in its analysis.

To provide some context, we plotted claimed sequence counts by applicant and publication date along with events that may have affected genetic sequence patents in the United States across the *x* axis. Moreover, we examined the legal status of these patents, determined the percentage of those that have already expired, and displayed the information as a timeline (with sequence counts indicated for each year; **Fig. 3**).

A comprehensive treatment of these events is beyond the scope of our study, but in the early 1990s, with the controversy of expressed sequence tag (EST) patenting, the number of sequences being claimed remained low until early 1996, when it started to increase at a slower rate. It picked up again in 1999 and reached a peak in 2002.

Once the Human Genome Project (HGP) was announced to be complete in 2003 (ref. 26), the number of sequence claims began to decrease. Of course, one interpretation is that the perceived high-value genes and loci had been patented before this time, as they would have been subject to intensive investment and scrutiny. An alternative explanation is that new business values became possible then, because of the sequence annotations efforts that led to increased claiming in other categories, such as altered phenotype, subpart or comparison/target (**Fig. 2c**,**d**).

Although all the 927 patents we analyzed that contain claimed sequences are probably now invalid under *Myriad*, we were curious about the perception of value that the owners of these patents previously assigned to them. We looked at failure to pay maintenance fees as a proxy for the patent owner's perception[27]. In the United States, maintenance fees are due at about 4-year intervals after the patent is granted and registered. We examined the percentage of expired US patents at 4, 8 or 12 years (whenever information was available) post granting and found that 30–33% of sequence patents were not maintained for their full potential lifetime. The percentage reported in **Figure 3** reflects expirations per year at 4, 8 or 12 years post granting. To investigate the matter further, we grouped the patents on the basis of the applicant type (corporation, government, hospital, university or individual) and inferred the value of claimed sequences from these patents to their applicants. Our findings indicate that 47% of the sequences claimed by hospitals or 43% of the sequences claimed by universities are associated with expired patents, whereas only 13% of the sequences claimed by corporations are from expired patents (**Table 1**).

These results likely reflect the differing perceptions of patent value and models for their use between public and private institutions. Although the former commonly aspire to direct revenue generation from such patents—either through licensing or spinoff—private enterprises must consider additional values, such as defensive, deterrent, cross-licensing and signaling drivers in their business positioning and partnership development, and thus are more likely to bear the costs of maintaining patents during their whole life cycle.

Isis Pharmaceuticals (Carlsbad, CA) applied for and received patent protection for far more nucleotide sequences than any other entity from 1999 until 2003, accounting for at least 61% of the total claimed sequences, with 2,285 sequences granted in 2002 alone. Unfortunately, until recently the USPTO rules did not obligate public registry of patent assignments; it is therefore very difficult to ascertain who owns what patents and whether all Isis patents, for example, are actually owned by Isis; licensing status is even more obscure. However, initiatives such as USPTO's request for comments for more complete patent assignment information[28] and the Executive Order from the US President to render patent ownership more clear[29] could substantially change this situation.

PatSeq toolkits

We have developed a suite of evidence-based public tools that will allow any interested party to investigate and navigate patent-disclosed sequences within the context of their metadata and patent claim rights. We have introduced into 'the Lens' several indicators to identify patent documents containing a sequence listing. After a search is carried out and if a patent document contains a sequence listing, we insert a sequence

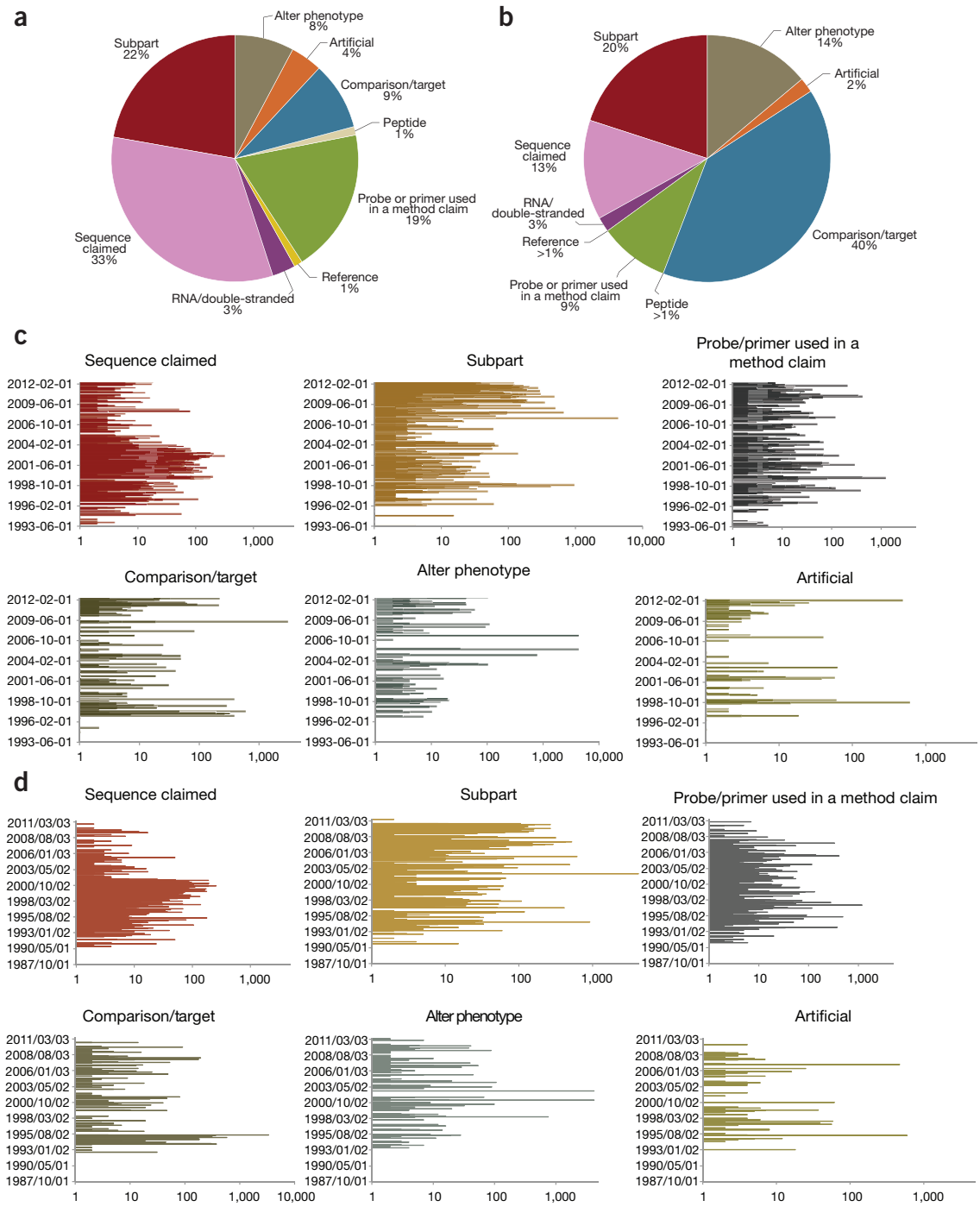

Figure 2 Categorization of patent sequences referenced in the claims. (**a**,**b**) We categorized 76,910 unique patent sequences that map with 100% homology to the human genome and are referenced in the claims of 2,685 granted US patents according to the role of the sequence in the claim. In the distribution based on patents (**a**), if a patent had different sequences in multiple categories, the patent was counted more than once. However, for the distribution based on sequences (**b**), if the sequences were referenced in two different contexts in the same set of claims, they were categorized only once, according to the broadest category. For example, a sequence referenced both as a primer for use in a method claim and also claimed as a primer sequence would be categorized as 'Sequence claimed.' We used ten categories. 'Sequence claimed' includes claims for isolated nucleotide sequences, sequences specifically claimed as primer or probe sequences and antisense sequences. 'Subpart' includes sequences that are part of a larger sequence, sequences that are one of several sequences claimed as a set, and sequences claimed alongside nonsequence substances, such as a pharmaceutical carrier. 'Alter phenotype' includes sequences in method claims in which the sequence is used to alter a cell, tissue or organism. 'Comparison/target' includes sequences in claims that employ the sequence in a comparison (e.g., of methylation or expression), in a screening assay or as a target for some claimed product or method. 'Probe or primer used in a method claim' includes sequences that are referenced as a probe or primer to be used in a claimed method. 'RNA/double-stranded' includes sequences that are specifically claimed as RNA sequences along with sequences that are claimed as double-stranded. 'Artificial' includes sequences claimed that differ from the wild-type version of the sequence (so that manipulation of the wild-type sequence would be unlikely to infringe the claim), as well as sequences that are identical to the wild-type sequence, but with chemical modifications to the backbone or sugar residues. 'Peptide' includes claims for a peptide made with reference to the coding nucleotide. 'Reference' includes sequences that are generally referenced either as a placeholder or in the negative. 'Submarine' sequences are claims wherein the sequence ID is not referenced in the claim but the patent claims a broad set of sequences (e.g., US 6258540). By way of comparison, Merz *et al*. write that "Gene patents cover three distinct types of invention: diagnostics, compositions of matter and functional uses."[10] Our categories do not correspond perfectly to Merz's, but generally Sequence claimed, Subpart, RNA, Artificial, Peptide and Submarine categories all correspond to compositions of matter. Comparison/target corresponds to diagnostics, Alter phenotype corresponds to functional uses, and Probe or primer used in a method corresponds to either diagnostics or functional, depending on the specific claim. (**c**,**d**) Profiles of the major categories are depicted based on publication date (**c**) or filing date (**d**).

tab within the document portfolio that clarifies the nature of the disclosed sequences and provides information, if available, about their metadata (nature of sequence, length, origin of organism) and potentially their redundancy level, location within the document where the sequence is referenced, and the source from which the sequence was downloaded. For further analyses, we have also created the PatSeq Finder, PatSeq Explorer and PatSeq Analyzer tools.

PatSeq Finder allows users to query any sequence against the PatSeq databases and conduct sequence similarity searches based on BLAST version 2.2.28 (http://www.ncbi.nlm.nih.gov/books/NBK131777/). Search results are aligned based on a score of relatedness to the original query and sequence information is depicted with that of the corresponding patent document. Users can view patent document's attributes including patent claims, selected alignment views and sequence annotation, if available, and embed or download results in various formats.

PatSeq Explorer enables a multi-level visualization and navigation of patent-disclosed sequences that map under various homology thresholds to a reference genome. At the genome and chromosome levels, users can investigate overall patenting trends, filter, and search sequence and patent attributes, and link to various sets of patent documents in the Lens (**Fig. 4**). Mapped sequence entries are displayed based on their location in the patent document (Grants in claims, Grants, Applications in claims and Applications) and their type (nucleotide or peptide), along with a summary statistic view for the overall coverage per jurisdiction. All views are embeddable.

PatSeq Analyzer allows users to zoom in to the details of a particular sequence entry and enables comparative analysis within the context of a gene region. The tool is a modified genome viewer built and integrated into PatSeq Explorer based on the open source HTML5/SVG genome maps browser by the Computational Medicine Institute, Prince Felipe Research Centre, Valencia, Spain[30]. In addition to the dedicated patent sequence tracks, PatSeq Analyzer provides feature tracks from public genome annotation datasets (including SNPs and gene/transcripts). Conversely, all views in PatSeq Analyzer are embeddable.

Global patent offices and sequence listings

Although highly controversial, patents on isolated genomic sequences are still allowed in many countries, including within Europe, Canada, Australia and Japan[31]. In addition to providing more precise analytic tools, with the increasingly globalized markets and innovation, there is now an urgent need for shared, harmonized data to guide decision-making and to accommodate differences in patent practice and policy in diverse jurisdictions. What are the standards and practices regarding patentability of nucleotide or amino-acid sequences in the various patent offices? And how do they make sequence listings data available?

We have carried out a survey of patent offices (**Table 2** and **Supplementary Table 1**) to ascertain the standards and practices regarding the patentability of nucleotide and amino-acid sequences. Between July and October 2011, we mailed 55 patent offices around the world a series of questions modeled on those from the 2001 survey conducted by the World

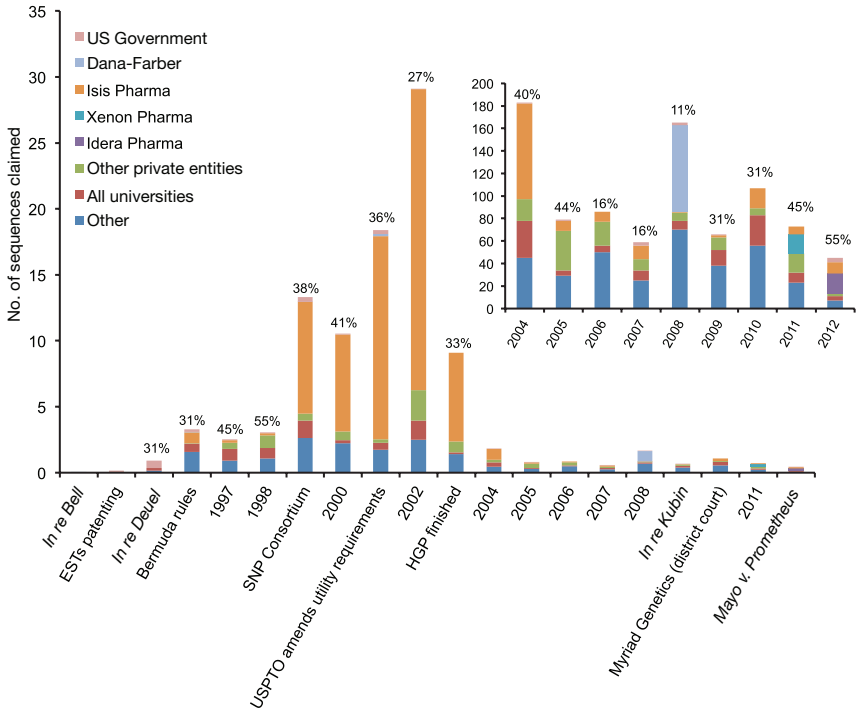

Figure 3 Sequences claimed by applicant and year. Patent sequences, which were categorized as 'Sequence claimed,' are shown by applicant and by the year in which the patent claiming the sequence was granted by the US Patent and Trademark Office. Some relevant legal and regulatory events affecting the patenting of genetic sequence in the US are also shown across the *x*-axis, and percent inactive patents per year is also depicted on top of each bar. Patent sequences claimed from 2004–2012 are also shown in the internal chart to allow for differences in scale between the number of claimed sequences before and after 2003.

Table 1 Percentage of claimed sequences in expired patents in the collection of 927 patents that contain claimed sequences and which map with 100% homology onto human genome

Applicant category	Total number of sequences	Sequences in expired patents	Percentage of sequences in expired patents
Corporation	7,584 (554 patents)	982 (154 patents)	13%
Corporation & research institute	70 (9)	14 (3)	20%
Corporation & university	265 (16)	17 (4)	6%
Government	210 (38)	80 (14)	38%
Hospital	38 (14)	18 (7)	47%
Hospital & research institute	289(18)	54 (6)	19%
Individual	228 (27)	10 (6)	4%
Research institute	359 (100)	86 (30)	24%
University	846 (149)	367 (55)	43%

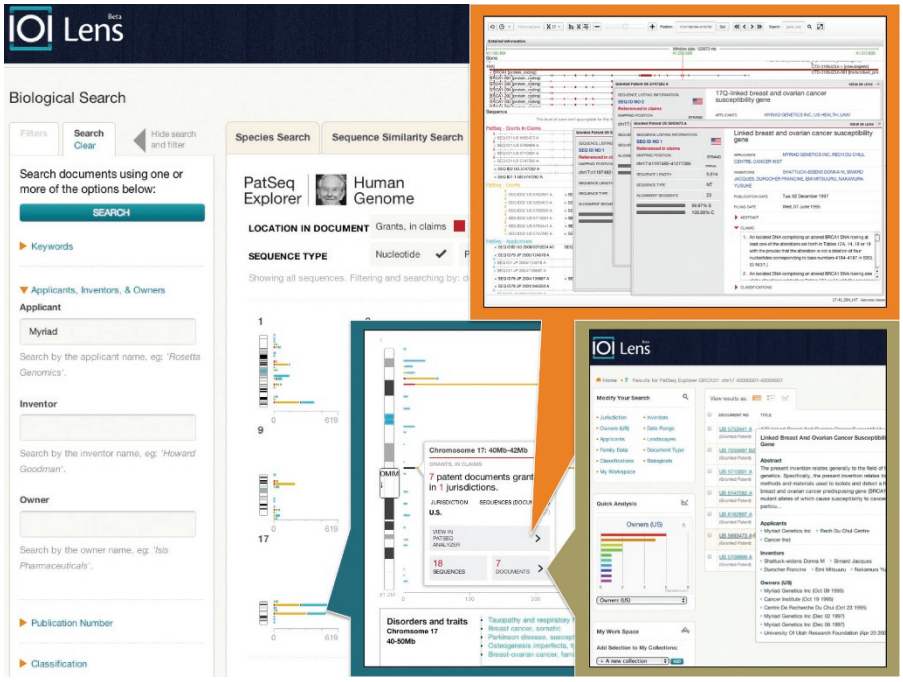

Figure 4 Patent sequences, which map to the human genome with various homology thresholds, can now be investigated using PatSeq Explorer-Human genome (http://www.lens.org/lens/bio/patseqexplorer). In this example, disclosed sequences in patents with applicant "Myriad" and which map to the human genome with 90–100% similarity and 100% coverage are displayed in PatSeq Explorer. Under 'Filters' option, users can view patenting trends based on either publication or filing dates, or filter based on jurisdiction, sequence length, species or document type. Under 'Search' option, users can interrogate the data based on patent attributes such as claims, applicant, owner, inventor and classification. In the chromosome view, added features include linking to the OMIM database (turquoise panel) for associated disorders and traits on that particular position, viewing the document collection in the Lens at http://www.lens.org (brown panel), and analyzing the data at the loci/gene/sequence regions using PatSeq Analyzer.(orange panel).

Intellectual Property Organization (WIPO) Intergovernmental Committee on Intellectual Property and Genetic Resources, Traditional Knowledge and Folklore[32], together with some additional queries on the public availability of sequence listings data. **Supplementary Table 1** displays the information received from various patent offices based on 2011 Cambia's surveys and compared with WIPO 2001 survey results. **Table 2** displays the 33 responses received on the public availability of sequence listings data.

We did not get a direct response from patent offices in the following countries: Argentina, Brazil, China, Colombia, Costa Rica, Cyprus, Dominican Republic, Guatemala, Honduras, India, Italy, Luxembourg, Mexico, Monaco, Korea, Hong Kong, Philippines, Singapore, Sweden or the USA. However, we were able to gather some information on the practices of some of these offices by alternative correspondence and investigations of their public websites.

Almost all patent offices—with the exception of Israel's—indicated that they comply with the Standard ST.25, which is the agreed standard for disclosure of sequences associated with patent filings. Unfortunately, this standard does not stipulate machine readability. So even though most offices make sequence listings available as part of the published patent document, these listings are mostly pdfs or images and hence not in a machine-readable form. Only patent offices in the USA, Canada, Germany, Hungary, Japan and to a certain extent Korea provide machine-searchable sequence listings through third-party providers or electronic downloadable files through their own websites, often for a fee.

Although we were able to obtain accurate counts on the total collection of sequence listings from some jurisdictions, it was difficult to get information from many others because they do not keep records on submitted sequence listings in their jurisdiction or they rely on the regional patent offices such as the European Patent Office (EPO; Munich, Germany) and WIPO for that information. Even in the United States, where compliance with sequence rules is more rigorously observed, we found several thousand sequence listings cited in patents published since 1990 that were not included in the GB-PAT database (**Table 3**), perhaps because of a lack of machine-readable forms, improper standards compliance or errors in sequence processing. Although some commercial entities are presumably able to negotiate access to patent sequence information from some jurisdictions, render it machine searchable and provide it on a fee-based service to the public

Table 2 Survey results on public availability of biological patent sequences from patent offices in 2011

Jurisdiction	Public availability of sequence listings	Format provided	Coverage
Australia	Sequence listings can be accessed through AusPat (http://www.ipaustralia.gov.au/auspat/), or from WIPO through PatentScope (http://www.wipo.int/patentscope/en/) if through the PCT route.	Mostly in pdf format	Since 1978
Belgium	Access is through a paid service of 0.36 euro per (scanned) page.	Scanned images	Since 2005 and only from granted patents
Bulgaria	Sequence listings are provided only as a part of a patent or application. It is a paid service.	Scanned pdfs	Since September 1, 2007
Canada	Electronic files and scanned images of sequence listings are available on the internet from the Canadian Patents Database (CPD). There is no fee applicable for the retrieval of files.	Multiple formats	Since October 1, 1996 and maintains full records
Chile	No available sequence listings yet.	Not applicable	Not applicable
China	Some sequence listings are available as part of patent documents.	Mostly in pdf and scanned images	No response

Continued

Table 2 Survey results on public availability of biological patent sequences from patent offices in 2011 (continued)

Jurisdiction	Public availability of sequence listings	Format provided	Coverage
Czech Republic	A sequence listing is made available as part of the patent document (grant or published application), which can be accessed online free of charge (http://isdv.upv.cz/portal/pls/portal/portlets.pta.formular?lok=en).	Unsearchable pdf files	Since 1991
Denmark	As of July 13, 2011, Danish patent office had 5 granted patents with sequence listings and 25 published applications with sequence listings.	Unsearchable scanned copies	Since 1987 and maintains full records
Estonia	Available as part of the patent publication.	Multiple formats	No response yet.
Eurasian Patent Organization	Published Eurasian applications and patents may be viewed in EAPO Gazette and on EAPO website (http://www.eapo.org), or via the Eurasian information system EAPATIS. Access to EAPO Gazette is free; access to EAPATIS is free only in guest mode.	Scanned copies	Since 1996 for patent applications and 1997 for grants
Finland	Accessible online (http://patent.prh.fi/patinfo/default2.asp?Etsi1=Etsi / Search). Only searchable by application number.	pdf format	Since 2013 for patent applications
France	Sequence listings are part of the description in French patent documents. They are accessible at http://fr.espacenet.com.	pdf format	Since 1978 for patent applications and 1989 for granted patents
Germany	Granted patents and published patent applications as of 1991 may be researched through the office's website (http://depatisnet.dpma.de/DepatisNet/depatisnet?action=einsteiger).	Computer searchable	Since 1991 with full records kept
Great Britain	Sequence listings are published with a patent application as it becomes open to the public inspection on the publication date.	Multiple formats	Since 2006, but sequence listings were accepted well before this date
Greece	Sequences are provided as part of patent documents and by a formal request. This is a paid service in a case-by-case decision of the Administrative Council of the Industrial Property Organization.	Multiple formats	Since 1988 for patent applications
Hungary	Sequence listings of all published patent applications are available online (http://www.sztnh.gov.hu/English/ IP-SEARCH – patents).	Searchable pdf format	Since January 1, 2003 with records kept
Iceland	If a sequence listing forms part of the issued patent or the application, a paper copy of the list is joining the document and available upon request.	pdf format	Since 1984 with full records kept
Ireland	The office only provides the specifications disclosed. The sequences therein are not independently obtainable.	Scanned images	Since 1970 for granted patents and since 1991 for applications
Israel	Sequence listings are not published along with the patent.	Not applicable	Not available to the public
Japan	JPO provides gazettes for granted patents and published patent applications via the Industrial Property Digital Library (IPDL; http://www.inpit.go.jp/english/distri/ipdl/). Also, JPO deposits sequence listings in DNA Data Bank of Japan (DDBJ).	Multiple formats	
Liechtenstein	See responses from Switzerland.		
Republic of Lithuania	Sequence listings present in patent specification are available on internet together with the whole specification (espacenet and http://www.vpb.lt/index.php?l=en&n=333).	Scanned images but no computer readable files	Since 1994 for full-text grants and since January 1, 2010 for applications
Korea	Sequence listings are provided as part of patent documents as bulk download.	Multiple formats	
Netherlands	Sequence listings form part of the patent dossier and are public from the day the patent is signed in and granted (http://register.octrooicentrum.nl/register/zoekformulier).	pdf or tif image	
Peru	Available on internet for the years 2000–2006 at Indecopi (http://www.indecopi.gob.pe/0/home.aspx?PFL=0&ARE=0).	pdf format	There is no registry of the patent applications and/or patents that have sequence listings, but there is a registry for biotech patents
Poland	A sequence listing is made available as part of the granted patent document only and can be accessed either by official request or online (http://pubserv.uprp.pl/PublicationServer/index.php?strona=index).	pdf format	Since 1990
Portugal	A copy of the granted patent (or the published patent) that contains the sequence listings can be provided upon request.	pdf format	No response yet
Romania	Patents containing sequence listings are published within 4 months after the grant (http://bd.osim.ro/cgi-bin/invsearch8).	Multiple formats	No response yet
Russia	Information regarding nucleic and amino acid sequences is available only within the description of issued patents.	Scanned copies	No information
Slovak Republic	Sequence listings available online (http://www.indprop.gov.sk/?introduction).	pdf format	No response yet
Slovenia	Available online (http://www3.uil-sipo.si/PublicationServer/).	pdf format	Since 1992
Spain	Sequence listings can be obtained from published patent documents.	Scanned image	No response yet
Switzerland	Available via espacenet or searchable online (http://www.swissreg.ch).	pdf and scanned images	Since 1995
Taiwan	Available as part of patent specifications or published application specifications from the patent search database.	Scanned image	From May 1, 2003
Turkey	Available as part of patent documents.	Scanned image	Since 1995 for applications and issued patents
USA	Available online as part of patent documents, as downloadable bulk, and deposits in GenBank.	Multiple formats	Since 1982 for grants and 2001 for applications

Table 3 Number of patent documents, which contain formatted sequence listings, as extracted from United States Patent and Trademark Office (USPTO) full text and bulk listings data sets and compared to those available at GenBank-Patent division (GB-PAT) from 1982–2013

Publication date	USPTO documents	Gb-PAT documents	Corresponding sequences missing from GB-PAT
Mar 1982–Jun 1990	Selected grants were sent to NCBI	1,005 (23,850 sequences)	Not known
Jun 1990–Nov 1992	2	0	4
Nov 1992–Dec 2001	18,391 grants	16,220 grants	127,588
Jan 2001–April 2013	97,567 applications	0	78,556,258

The last column shows the number of patent sequences that were missing in GB-PAT division and thus not available in machine-searchable form to the public. Individually, an electronic copy of the sequence listing can be requested from USPTO based on a fee set under 37 CFR 1.19(b)(3) (http://www.uspto.gov/web/offices/pac/mpep/s2435.html).

(http://www.prweb.com/releases/2010/12/prweb4865134.htm; and WIPO with STN online patent information database), such initiatives require subscription fees.

In conclusion, although *Myriad* clarified the position of the United States on gene sequence claiming, the court decision also highlighted the pressing need for nuanced and precise analysis of gene patents at the global level. Our survey results confirm that public tools are not yet available in many of the emerging patent offices; thus, biological innovations that rely on genetic sequences can be severely affected when reaching global markets. In this article, we present a carefully designed public platform that can be a valuable alternative to the commercial services that serve only a few elite innovators in biological sciences.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nbt.2755).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

ACKNOWLEDGMENTS

This work was supported, in part, by the Bill & Melinda Gates Foundation, Global Health Grant ID 52239; Gordon and Betty Moore Foundation "Grant GBMF3465"; Lemelson Foundation "Grant 2008-731-1036", and Queensland University Technology "Grant 321121-0023/08". We thank I.M. Castello, head of the computational biology unit at the Computational Medicine Institute of Prince Felipe Research Centre, Spain, for enabling us to access their service "CELLBASE" and guiding us through the integration of genome maps tools into our facility; Small Multiples, a private visualization company in Sydney, Australia for implementing PatSeq Explorer platform design, L. de Vine for helping with the extraction of US-patent sequences between 1992 and 2001, B. Warren for improving the parser to locate sequences within the patent document, and N. Prasolova for establishing endnote libraries for the project. The assistance of S. Rajamanikam, I. Epichev and M. O'Neill in the survey and lapsed patent data is highly appreciated and we are particularly grateful for the constructive reviews by M. Rabson, senior partner at Wilson Sonsin Goodrich and Rosati and C. Nottenburg, owner of Cougar Patent Law at the early stages of this project.

1. *Association for Molecular Pathology v. Myriad Genetics, Inc. et al.* 569 US 12–398 (2013).
2. Than, K. *National Geographic Daily News* http://news.nationalgeographic.com/news/2013/06/130614-supreme-court-gene-patent-ruling-human-genome-science/ (June 14, 2013).
3. Hemphill, T.A. *Sci. Public Policy* **39**, 815–826 (2012).
4. Dobson, A.W. & Evans, J.P. *Genome Biol.* **13**, 161–167 (2012).
5. Holman, C. *UMKC Law Rev.* **80**, 563–605 (2012).
6. Cook-Deegan, R. in *From Birth to Death and Bench to Clinic: The Hastings Center Bioethics Briefing Book for Journalists, Policymakers, and Campaigns* (ed. Crowley, M.) 69–72 (The Hastings Center, Garrison, NY; 2008).
7. Caulfield, T. *et al. Nat. Biotechnol.* **24**, 1091–1094 (2006).
8. Holman, C. *UMKC Law Rev.* **76**, 295–361 (2007).
9. Amani, B. & Coombe, R.J. *Law Policy* **27**, 152–188 (2005).
10. Merz, J.F. & Cho, M.K. *Public Health Genomics* **8**, 203–208 (2005).
11. Lauren, C. *et al. Sci. Transl. Med.* **4**, 1–3 (2012).
12. Merges, R.P. *Berkeley Technol. Law J.* **24**, 1583–1614 (2009).
13. Williams, H.L. *J. Polit. Econ.* **121**, 1–27 (2013).
14. Andree, P.J. *et al. World Pat. Inf.* **30**, 300–308 (2008).
15. Pruitt K.D. *et al. Nucleic Acids Res.* **40** (database issue), D130–D135 (2012).
16. Harrow, J. *et al. Genome Res.* **22**, 1760–1774 (2012).
17. Pruitt, K.D. *et al. Genome Res.* **19**, 1316–1323 (2009).
18. Li, H.D. *Bioinformatics* **26**, 589–595 (2010).
19. Kent, W.J. *Genome Res.* **12**, 656–664 (2002).
20. Jensen, K. & Murray, F. *Science* **310**, 239–240 (2005).
21. Cook-Deegan, R. & Heaney, C. *Annu. Rev. Genomics Hum. Genet.* **11**, 383–425 (2010).
22. Bacon, N., Ashton, D., Jefferson, R.A. & Connett, M.B. CAMBIA Patent Lens OS4 Initiative, http://www.patentlens.net/daisy/patentlens/2205.html (2006).
23. Rimmer, M. in *Intellectual Property and Biotechnology: Biological Inventions* (ed. Rimmer, M.) 138–163 (Edward Elgar Publishing, Cheltenham; 2008).
24. Aggarwal, S. *et al. Nat. Biotechnol.* **24**, 643–653 (2006).
25. Adelman, D.E. & DeAngelis, K.L. *Tex. Law Rev.* **85**, 1677–1744 (2006).
26. Leite, M. *Genet. Mol. Res.* **3**, 575–581 (2004).
27. Moore, K.A. *Berkeley Technol. Law J.* **20**, 1521–1549 (2005).
28. http://www.uspto.gov/patents/law/comments/patent_assignment_information.jsp
29. http://www.whitehouse.gov/the-press-office/2013/06/04/fact-sheet-white-house-task-force-high-tech-patent-issues
30. Medina I. *et al. Nucleic Acids Res.* **41** (Web server issue), W41–46 (2013).
31. Gold, E.R. *et al. Genet. Med.* **12**, S39–S70 (2010).
32. Secretariat, W.I.P.O. 48 (WIPO, Geneva; 2001).