Gene patent practice across plant and human genomes

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

The uses of genetic sequences to inform, enable or create products or services for human biomedicine are substantially different from their uses in crop-based agriculture.

In human medicine, a successful new product or process can create a strong economic incentive to pay whatever it takes to be healthy, with the potential profit in conventional business models correspondingly very high. This huge value-capture opportunity is used to justify high-capital and high-risk innovation and investment, with corresponding patenting strategies. For instance, inventions that enable the development of small-molecule pharmaceuticals that associate with protein or nucleotide targets (such as receptors) or of direct biological interventions such as vaccines, RNA or protein-based therapeutics are attractive and potentially lucrative commercial pursuits. Similarly, diagnostics that detect allelic variation in human genes or proteins or detect and discriminate between genetic variants of human pathogens or beneficials have high potential value.

In plant-based agriculture, however, the unit value of a single plant or even a cultivar is generally very low, and profit margins for most commodity crops are modest. With little new acreage to cultivate, and with so much of broad-acre crops already biotech enhanced, many markets are nearly saturated, and farmers simply cannot pay much more for next-generation technologies. With current business models, therefore, and such low-margin targets, the scope of patent claiming of new inventions may need to cover an entire variety, species, very broad-use cases or new functionalities that enable potential new crop uses or novel crop-management tools such as herbicides or insecticides.

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

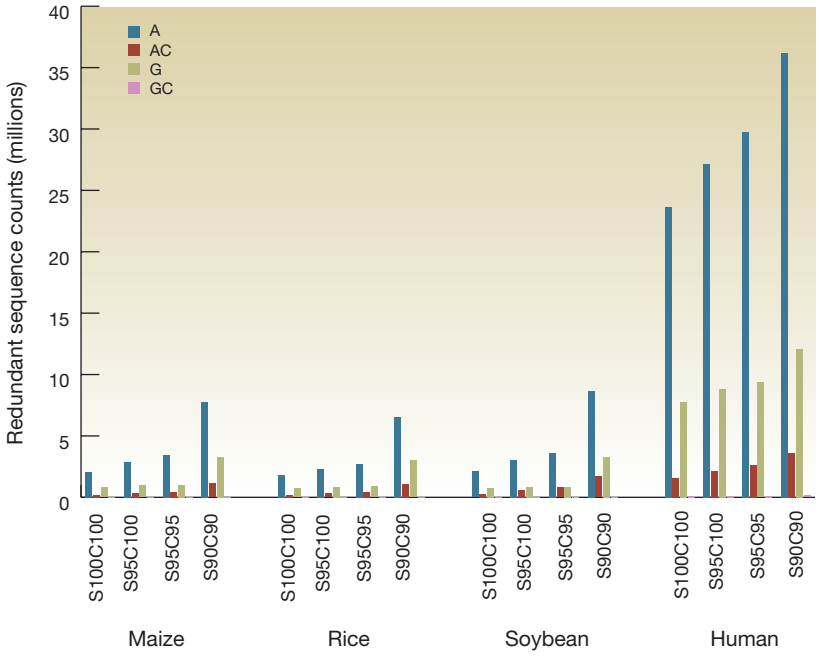

Figure 1 Mapped patent sequences on maize, rice and soybean genomes were compared to those mapped on the human genome on the basis of various similarity (S) and query length coverage (C) rates. Sequences are displayed according to their disclosure in patent document type. A, sequences disclosed in patent applications; AC, sequences referenced in the claims of patent applications; G, sequences disclosed in granted patents; GC, sequences referenced in the claims of granted patents.

For each of these uses, patenting of both nucleotide and amino acid sequences may be important but will be done with different strategies and economies in mind. We have previously described[1] the scope and type of patenting that disclosed and/or claimed genetic sequences on the human genome. Here, we explore what similarities and differences may emerge in patent use and strategies, and map patent-disclosed sequences onto three important plant genomes: maize (corn), rice and soybean. We focus on those referenced in the granted patent claims to compare their uses to the approach used in human gene patenting.

Mapping biological sequences disclosed in patents using a 95% homology threshold shows 2.8 million patent sequences each for the maize and soybean genomes and 2.5 million on the rice genome, versus 31 million patent sequences mapped on the human genome as of 13 November 2014. We chose the 95% homology threshold to maximize the likelihood that allelic differences between a patent-disclosed sequence and a related canonical reference genome would

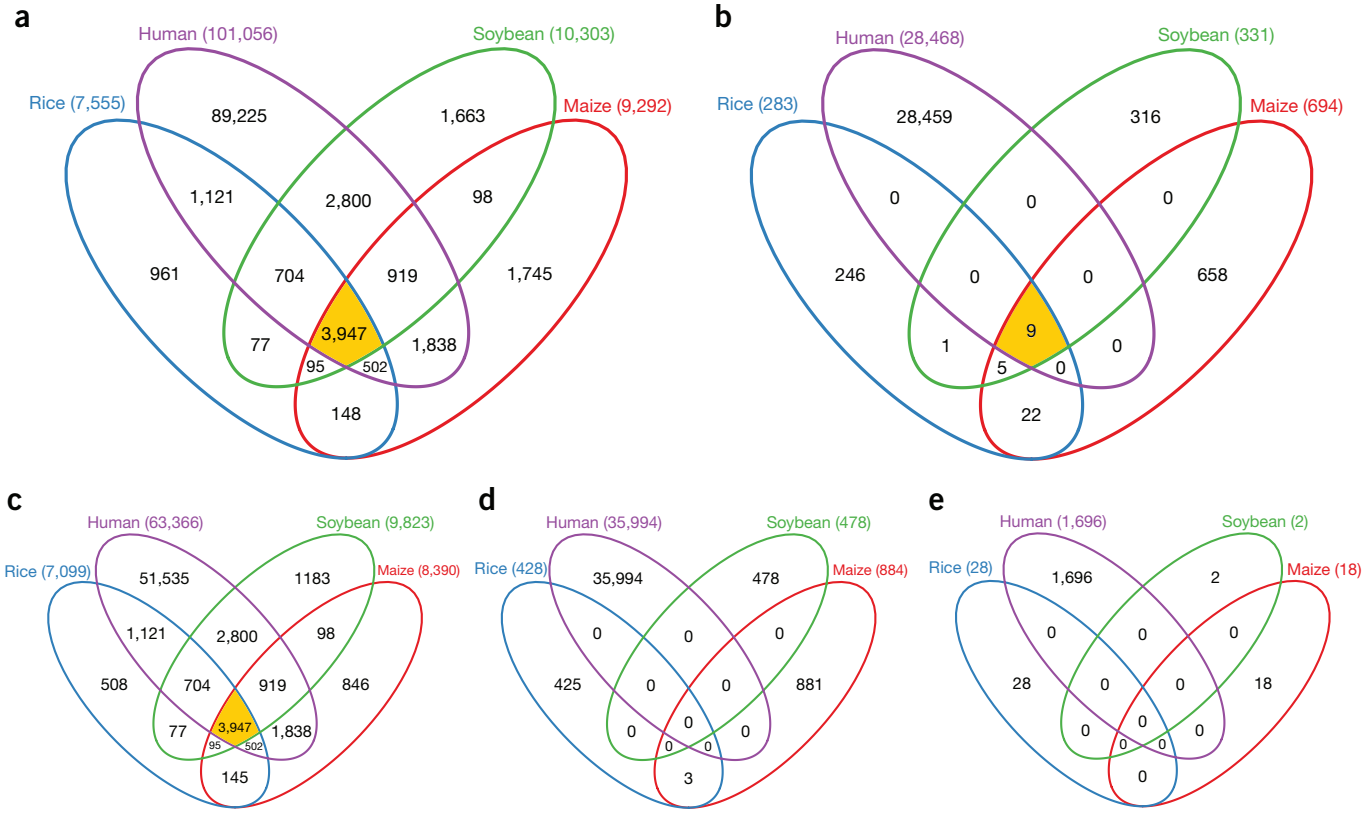

Figure 2 Unique and overlapped patent sequences that are referenced in the claims of granted patents, depicted on the basis of sequence type and length. The holdings of mapped sequence entries with 95% homology threshold against each of the four genomes are contained in color-coded ellipses, and areas of overlap are shaded. (**a**) Mapped patent nucleotide sequence entries. (**b**) Peptide patent sequence entries. (**c**–**e**) Mapped patent nucleotide sequence entries of 1–50 bp (**c**), 50–5,000 bp (**d**) and >5,000 bp (**e**).

not exclude the sequence as a likely homolog of that reference sequence.

Whereas 130,000 of the mapped human sequences were referenced in patent claims, fewer than 11,000 from each of the plant genomes were referenced. Of these, we determined that more than 80% overlapped with (i.e., were either sourced from or equivalent to homologs in) other genomes, including the human genome, and often consisted of short nucleotide sequences. In comparison, only 12% of the human nucleotide sequences overlapped with those of plant genomes. We identified 3,956 sequences overlapping across all organisms examined (maize, rice, soybean and human) and subjected their corresponding 763 patents to further characterization, including claim analysis.

In addition to a high level of sequence redundancy in the overlapping data set, we highlight three different claim types wherein the use of a short sequence could be problematic and could, hypothetically, raise infringement concerns. We review each type and provide some examples. We conclude by discussing the introduced changes to gene patent practice after the recent US Supreme Court decisions in *Association for Molecular Pathology v. Myriad Genetics, Inc.*[2] and *Mayo Collaborative Services v. Prometheus Labs, Inc.*[3] and their potential impact on the various industries and the public interest.

Mapping of plant genomes

To compare gene patent practice and its extent across plant and human genomes, we enriched our PatSeq toolkit with plant genome maps for maize, rice and soybean (http://www.lens.org/lens/bio/patseqexplorer - pse/zea_mays/latest). The selected field crops are economically important, grown widely and have diverse uses and distinct (and different) business models associated with their commercial use. Moreover, we updated the previously published human genome map[1] and created PatSeq Analyzer, a stand-alone tool wherein users can now search by sequence identifier number within each genome and download the sequences freely.

We mapped patent-disclosed sequences onto the following reference genomes: *Zea mays* (maize) assembly AGPv3.21 (http://plants.ensembl.org/Zea_mays/Info/Index), *Oryza sativa* Japonica (rice) RGSP-1.0.21 (http://plants.ensembl.org/Oryza_sativa/Info/Annotation) and *Glycine max* (soybean) assembly glyma1 V1.0.21 (http://plants.ensembl.org/Glycine_max/Info/Annotation/) using the Burrows–Wheeler Aligner suite[4]. As discussed previously, many patent claims provide rights over sequences with as little as 70% identity to a disclosed sequence; we therefore selected a range of homology thresholds to determine alignment and location of candidate sequences on the crop genomes. Homology thresholds were specified by two metrics: patent sequence similarity and coverage in proportion to the sequence length. The similarity rate (S) reflects the number of matching nucleotides between the patent sequence and the reference genome, and the sequence coverage (C) 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 nonredundant data set of patent sequences was initially used for the mapping and redundancy was later reincorporated in the tool.

Under the most stringent conditions, S100C100 (100% similarity and 100% coverage rate)—wherein the mapped sequence is

identical to that of the reference genome—we matched 2.8 million sequences on maize and soybean genomes and 2.5 million sequences on rice genome, in contrast to 31 million sequences mapped against the human genome, with the percentage of patent sequences referenced in the claims ranging 0.2–0.3% among the various genomes (**Fig. 1**).

These correspond to sequence listing entries after the reintroduction of the redundancy in the corpus. During this exercise, we observed that less than 30% of the mapped sequences against the crop genomes were declared as the corresponding plant species, whereas 71% of mapped sequences against the human genome were declared in the original patent documents as human. However, when we relaxed the homology thresholds from 100% to 95%, we saw a boost in the matching rate of the declared plant species to more than 60% in the mapped data set, indicating that a homology threshold of S95C95 would be more appropriate for further analyses.

By 13 November 2014, we had 9,986; 7,838 and 10,634 sequence listing entries mapped onto the maize, rice and soybean genomes, respectively, under the 95% homology threshold. These sequences were referenced in the claims of 2,241; 1,758 and 1,889 patents granted, most of which were US patents, and about 30% of which were lapsed or expired.

Analysis of the overlapped sequences among plant and human genomes

A substantive overlap was observed between the plant genome–mapped sequences and those mapped against the human genome. We performed an overlap analysis on all mapped sequence entries across the four genomes, focusing only on those referenced in the claims of granted patents (GC category). The analysis confirmed the presence of patent sequence overlap across all four genomes (**Fig. 2**). We selected the 3,956 sequence entries that overlapped across all four organisms for further characterization. Most (3,947) are nucleotides (DNA, RNA and cDNA) (**Fig. 2a**); nine are peptide sequences (**Fig. 2b**).

Upon splitting the nucleotide database on sequence length into three categories (1–50 bp, 50–5,000 bp and >5,000 bp), we found that almost all overlapped sequence entries were 1–50 bp in length (**Fig. 2c**), except for three overlapped sequences of the second category (**Fig. 2d**). No overlapped sequences were observed in the third category (**Fig. 2e**). To further characterize the short sequences, mainly the 3,947 nucleotide and the nine peptide sequences, we extracted and examined their associated metadata.

Sorting first by the unique fingerprint available for each sequence entry, we found that 39% of the sequences were redundant. Redundancy ranged from a simple duplication of the sequence entry within a document or between two related documents to referencing the same sequence up to 26 times across three documents (US patent 7,709,196, sequence identifiers (SEQ IDs) 33; 37; 39; 45; 105; 145; 149; 151; 157, US patent 8,293,517, SEQ IDs 18; 22; 28; 94; 134; 138; 140; 146 and US patent 8,367,319, SEQ IDs 33; 37; 39; 45; 105; 145; 149; 151; 157. These results suggest that if the sequence entry is relevant to the invention, applicants will use it in an original patent claim and probably in any of the improved or follow-on inventions, and, in a few cases, different applicants may do so as well.

Second, by aligning the overlapped data with that of human genome data analyzed previously[1], we confirmed that single sequence entries were sometimes referenced in the claims of multiple patents. For example, SEQ ID 13 of US patent 7,781,182 is a 76-amino acid (76-aa) sequence that is referenced in the claims of that patent as part of an assay for ubiquitin ligase activity. The same sequence is referenced as SEQ ID 1 in the claims of US patent 8,518,660 as part of a di-ubiquitin used in an assay for measuring the isopeptidase activity of a deubiquitinase. Finally, it is referenced as SEQ ID 15 in the claims of US patents 8,592,179; 8,790,895 as being used in methods for producing modified ubiquitin proteins and preparing a modified ubiquitin polypeptide, respectively, and of US patent 8,791,238 as the reference sequence for a modified ubiquitin protein.

According to US Patent and Trademark Office (USPTO) records, the first two patents mentioned above are assigned to different companies, but the final three are all assigned to the same company (**Supplementary Table 1** provides further detail).

Similarly, SEQ ID 5 of US patent 5,489,508 is an 18 bp–nucleic acid sequence that is referenced as a substrate for telomerase primer extension used in a method for detecting cancer in humans. It is also referenced in related US patent 5,645,986 as a telomerase substrate used as part of a method for screening for telomerase inhibitors. According to USPTO records, the first patent is assigned to the Regents of the University of Texas, and the second is assigned to the Regents of the University of California.

The same sequence, referenced as SEQ ID 6, also appears in the claims of three patents that USPTO records show as being assigned to Geron Corporation: US patents 6,545,133; 6,787,133 and 7,067,283. The sequence is used in a method for obtaining mammalian telomerase protein, in a method for identifying regulators of telomerase activity and in producing a compound that regulates telomerase activity, respectively. Finally, the same sequence also appears as SEQ ID 3 in US patent 7,781,163 as part of a compound forming a G-quadruplex structure and as SEQ ID 2 in US patent 8,053,422 in a method for treating, preventing or delaying development of papilloma (**Supplementary Table 2** provides further detail).

In a single instance, we found a sequence entry, SEQ ID 1, referenced in the claims of a plant (maize)- and *Bacteroides forsythusa* (a periodontal pathogen)-related patent (US patents 5,710,367 and 5,789,174, respectively), and in both patents the sequence was used as a primer in method claim.

Third, we checked for other potential infringement issues based on the use of sequences in claims related to both humans and plants. Relying on US classification codes, we screened the corresponding patent documents and selected relevant claims for manual analysis. We found at least three patterns of claims that may be problematic when a short sequence is referenced in the claims:

1. Wherein the applicant claims exclusively the use of the sequence as isolated polynucleotide, oligonucleotide or nucleic acid, etc., without specifying the host, modified structure or function of the sequence. In principle, the patent holder can potentially use such claim to exclude any other use of the sequence raising infringement concerns (**Table 1** shows some examples).

2. Wherein the applicant claims nonexclusively the use of the overlapped sequence in a core technology applicable at the research phase. As the technology has potential to be applied to all sequence-based inventions, such as "a kit for gene expression" (US patent 6,221,600) or "a screening method for genomic polynucleotide library" (US patent 8,846,403), the patent holder can monopolize the use of such a technology at the research phase across all fields of use and potentially delay or block improvements.

3. Wherein the applicant uses overlapped and conserved short sequences to target a specific technique that alters the genomic makeup of both human and plant components (US patents 6,936,467; 7,226,785 and 7,258,854), and thus claiming broadly based on the use of that sequence across many species. A single patent rights holder can then potentially block

Table 1 Examples of potentially broad sequence claims

US patent number and SEQ ID	Selected claims
Patent 8,785,611; SEQ ID 31	Claim 1: "An isolated polynucleotide sequence comprising at least two mRNA translational enhancer elements (TEE), wherein at least one TEE consists of a full-length sequence that is selected from the group consisting of 5′-CGCGGCTGA-3′ (SEQ ID NO: 31), 5′-AGCCGCCGCA-3′ (SEQ ID NO: 34) and 5′-ACGCCGCCGA-3′ (SEQ ID 35)."
Patent 8,288,355; SEQ ID 27	Claim 1: "An isolated polynucleotide of 15 to 49 bases in length comprising at least 15 contiguous bases in the nucleotide sequence of SEQ ID NO:26 that include the nucleotide sequence of SEQ ID NO:27."
Patent 7,807,817; SEQ ID 151	Claim 2: "An isolated catalytic DNA molecule having sequence-specific endonuclease activity, wherein said molecule comprises a conserved core, wherein said conserved core comprises the sequence CCGAGCCGGACGA (SEQ ID NO:151), or wherein said conserved core comprises the sequence CCGAGCCGGACGA (SEQ ID NO:151) having one to three residues substituted by G, A, T, or C."
Patent 7,456,273; SEQ ID 40	Claim 1: "An isolated transcriptional regulatory element selected from any of SEQ ID NOS: 40." Claim 2: "A recombinant nucleic acid molecule comprising a plurality of operatively linked isolated transcriptional regulatory elements of claim 1." Claim 3: "A kit, comprising an isolated transcriptional regulatory element according to claim 1." Claim 4: "The kit of claim 3, further comprising a vector for containing the regulatory element." Claim 5: "The kit of claim 3, comprising a plurality of isolated transcriptional regulatory elements."
Patent 5,707,803; SEQ ID 1	Claim 1: "An isolated DNA molecule 13 to 200 nucleotides in length comprising at least one regulatory element that binds to an activated transcriptional regulatory protein, said regulatory element comprising a nucleotide sequence of TATTCCTGGAAGT (SEQ ID NO: 1), TATTCCGGTAAGT (SEQ ID NO: 2), TCTTCCTGTAAGT (SEQ ID NO: 3), TATTCCCGTAAGT (SEQ ID NO: 6), or TATTCCTATAAGT (SEQ ID NO: 7)."

further experimentation and follow-on innovations in a cumulative-type innovation.

Referenced sequences that are plant related

To examine mapped sequences that are referenced in plant-related inventions, we first had to hand-edit each of the three plant genome data sets—in particular granted patents that reference these mapped sequences in the claims as their 'declared species' metadata information was often lacking or inadequate—disambiguate applicant and owner names and analyze them within the context of other intellectual property (IP) rights.

State of affairs after *Myriad*

After decades of allowing patent protection for a broad range of biotechnology tools and materials—including genetically modified cells, plants and animals—the US Supreme Court recently narrowed the scope of patent protection in its *Myriad* and *Mayo* decisions by holding that naturally occurring DNA sequences are unpatentable, as are laws of nature and abstract mental processes.

On 4 March 2014, the USPTO proposed a new procedure to apply the court decisions and invited members of the public to comment[6]. The published reactions encompassed divergent arguments for or against the guidance and revealed old, unsettled tensions and anxiety from national and international groups[7–10]. For example, the Association for Molecular Pathology commended the USPTO for implementing the court decisions; confirmed that threshold determinations are for all claims involving natural laws, principles, phenomena and products; and recommended that the guidance provide an even clearer qualitative standard of "markedly different" to maintain patent ineligibility for homologous sequences and any "claimed associations between genetic mutations and their relationships to medically relevant physical or physiologic effects."[11] But the Association of University Technology Managers, the Council on Governmental Relations, the Association of American Universities and the Association of Public and Land-grant Universities challenged the legality of expanding court decisions to all claims and expressed strong concerns about the proposed legal standard and its potential negative impact on pending or existing university patents on natural products[12]. Law associations and professionals in the industry sector raised similar concerns about the applicability of *Myriad* and *Mayo* to other claims.

As these fierce legal debates go on, patent examiners have, since March 2014 and in some cases since the *Myriad* decision in June 2013, been narrowing some pending application claims that involve nucleotide sequences during prosecution. For example, below we compare an early version of claim 1 of the application that issued as US patent 8,772,024 on 8 July 2014 with a later version.

Before March 2014, claim 1 read: "An isolated nucleic acid molecule selected from the group consisting of: a nucleotide sequence consisting of the polynucleotide sequence of SEQ ID 1 or 2; wherein the sequence initiates transcription in a plant cell."

After March 2014, the claim read: "An isolated nucleic acid molecule selected from the group consisting of: a nucleotide sequence consisting of the polynucleotide sequence of SEQ ID 1 or 2; wherein the sequence initiates transcription in a plant cell, *and wherein said nucleic acid molecule is linked to a heterologous nucleic acid molecule* [italics added]."

The italicized clause was added by the USPTO examiner specifically for the purpose of complying with *Myriad*, and the applicants consented to this amendment, but this part was omitted from the claims as published. A notice of correction needs to be filed to have this fixed. Claim 1 as amended complies with *Myriad*.

Another example is US patent 8,692,076 issued on 8 April 2014. In the application before *Myriad*, claim 1 read: "A DNA molecule comprising a sequence which is, or is complementary to, a DNA sequence selected from the groups consisting of SEQ ID NO: 1 and SEQ ID NO: 2." Claim 2 read: "An isolated DNA molecule for use as a DNA probe that is diagnostic for soybean event MON87769 DNA, comprising at least 11 contiguous nucleotides of SEQ ID NO: 1, or complement thereof."

After *Myriad*, the claims were as follows. Claim 1: "A DNA molecule comprising SEQ ID NO: 1 or SEQ ID NO: 2, *the DNA molecule further comprising a nucleic acid molecule encoding Primula juliae delta 6 desaturase, or the full complement thereof* [italics added]." Claim 2: "*The* DNA molecule *of Claim 1*, comprising SEQ ID NO: 6, or a *full* complement thereof [italics added]."

The introduced changes, italicized here, were added by the USPTO examiner specifically for the purpose of complying with *Myriad* and written description requirements. In the application, claims 1 and 2 were "broadly drawn" to cover "any DNA molecule which comprises any DNA nucleotide residue which is complementary to only a single nucleic acid residue of SEQ ID NO: 1 or SEQ ID NO: 2." The amended claims had a narrower scope. The applicants consented to these amendments.

Conclusions

As in the human genome map, the three plant genome maps reveal physical locations of patent sequences that were either sourced

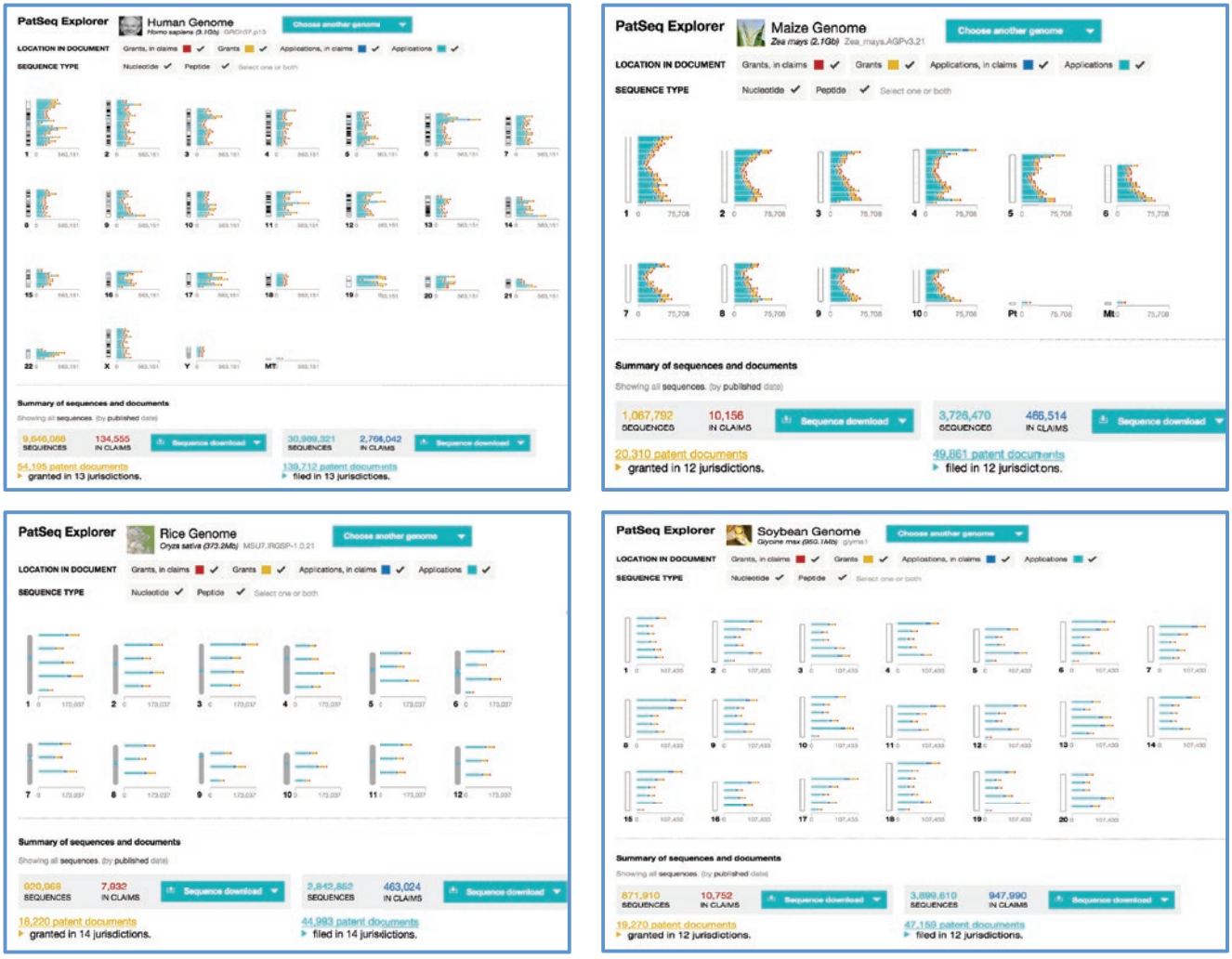

Figure 3 A screenshot of the three plant genome maps each compared to that of human in PatSeq Explorer. Users can access and navigate mapped sequences on the basis of sequence type (nucleotide or peptide) and/or location in a patent document (referenced in patent claims or not) and download relevant patent sequences freely.

from the corresponding genome or derived from other organisms but showed homology during the mapping process (**Fig. 3**). Having such maps can be extremely useful in precision plant breeding, especially as additional tracks of various markers are linked. In addition to tracking inheritance patterns of important traits, particularly quantitative trait loci, trends of gene expression or regulation in commercial traits can be now visualized across these three genome maps using the PatSeq toolkit.

Although far from exhaustive, our claim analysis of the overlapping sequences between human and plant genomes shows that such sequences are referenced relatively frequently in patent claims, often without explicit reference to a host, function or chemical modifications. Such practice could, in principle, raise infringement concerns—for example, if an agribusiness and a medical diagnostic company used the same DNA primers for polymerase chain reaction–based genetic testing.

Two other types of short sequence–based claims may be problematic: those extending to core research technologies and those targeting wide fields of use (plant, fungi, bacteria and animal). Although not involving specific sequences, the best example of the former are US patents 4,816,567; 6,331,415; and 7,923,221, known collectively as the Cabilly patents after one of the inventors[13]. The Cabilly patents are directed to key steps in the manufacture of therapeutic antibodies and as such have been immensely valuable to their owner, Genentech, as well as the subject of numerous lawsuits between Genentech and various other biotechnology companies. An example of the latter is US patent 8,273,954 directed to *Agrobacterium* transformation of dicotyledonous plant cells. The patent was issued in 2012 after almost 30 years (priority filing 1983), and granted Monsanto, its owner, exclusive rights on this technology until about year 2029 (https://www.lens.org/lens/patent/US_8273954_B1). This is one of the broad patents on "*Agrobacterium*-mediated transformation methods" that—after a series of lawsuits, oppositions, mergers and acquisitions—asserted Monsanto's dominance in the genetically modified crop business[14].

Although narrowing the scope of gene patentability is commendable and may address some of these issues, more can be done to stimulate biological innovation. Further work is under way to uncover plant-related inventions based on these plant genome maps, analyze broadly the patent claims and explore their ownership and its impact in the context of other plant intellectual properties.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nbt.3364).

ACKNOWLEDGMENTS

We thank I. Medina and his team at the University of Cambridge for ongoing support of Genome Maps and making the crop genomes available in Cellbase; the Lens team at Cambia for continually improving the Lens features; and Small Multiples (Sydney) for help in improving the stand-alone PatSeq toolkit. We are particularly grateful for review by S. Hughes at the early stages of this project. This work was supported by the Bill and Melinda Gates Foundation 2013 Global Health Grant–Cambia Lens Accelerated Grant OPP1104285 (R.A.J.); Gordon and Betty Moore Foundation Grant GBMF 3465 (R.A.J.); and Queensland University Technology Grant 321121-0023/08 (O.A.J.).

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Jefferson, O.A., Köllhofer, D., Ehrich, T.H. & Jefferson, R.A. *Nat. Biotechnol.* **31**, 1086–1093 (2013).
2. *Association for Molecular Pathology* et al. *v. Myriad Genetics, Inc.* et al. 569 US 12–398 (2013).
3. *Mayo Collaborative Services v. Prometheus Labs, Inc.* 566 US ___, 132 S. Ct. 1289 (2012).
4. Li, H. & Durbin, R. *Bioinformatics* **26**, 589–595 (2010).
5. Redundant and overlapping patent sequences can be viewed in PatSeq Analyzer (http://www.lens.org/lens/bio/patseqanalyzer#psa/glycine_max/latest/chromosome/15/12825152–12827178?grant=1&similarity=90&coverage=50).
6. US Patent and Trademark Office. Guidance for determining subject matter eligibility of claims reciting or involving laws of nature, natural phenomena, and/or natural products (2014). http://www.uspto.gov/patents/law/exam/myriad-mayo_guidance.pdf
7. Harrison, C. *Nat. Biotechnol.* **32**, 403–404 (2014).
8. Check-Hayden, E.C. *Nature* **511**, 138 (2014).
9. Holman, C. *Biotechnol. Law Rep.* **32**, 289–293 (2013).
10. Public comments on guidance for determining subject matter eligibility of claims reciting or involving laws of nature, natural phenomena, and natural products (2014). http://www.uspto.gov/patents/law/comments/myriad-mayo_guidance_comments.jsp
11. The Association for Molecular Pathology. Re: guidance for determining subject matter eligibility of claims reciting or involving laws of nature, natural phenomena, and natural products (2014). http://www.uspto.gov/sites/default/files/patents/law/comments/mm-a-amp20140730.pdf
12. Association of University Technology Managers, the Council on Governmental Relations, the Association of American Universities and the Association of Public and Land-grant Universities. Comments on the USPTO's guidance for determining subject matter eligibility of claims reciting or involving laws of nature, natural phenomena, and natural products (2014). http://www.uspto.gov/sites/default/files/patents/law/comments/mm-f-autm-cogr-aau-aplu20140729.pdf
13. Storz, U. *MAbs* **4**, 274–280 (2012).
14. Nottenburg, C. & Rodriguez, C.R. in *Agrobacterium: From Biology to Biotechnology* (eds. Tzfira, T. & Citovsky, V.) 699–735 (Springer, 2008).