



Sequence-Based Patent Landscaping: The Rice Genome Landscape

Kerry Fluhr, Ph.D.
April 7, 2008

CAMBIA

What is a patent landscape?

- Patent landscapes provide information about the extent to which a technology area is encumbered by patents and patent applications.
- Patent landscapes can be broad for “scoping” purposes, narrow and detailed, or anything in between.

Outcomes of Patent Landscaping

- Provide clarity in decision making.
- Counter “perception” with facts.
- Increase the efficiency of innovation.
- Transform information into knowledge.

Landscapes Often Provide Good News

- Landscape may illuminate areas that are free of patenting activity.
- Patents are time-limited.
- Patents may be limited in scope.
- Many patents have lapsed or are no longer in force.
- Patents may not be in force in key jurisdictions.

Drawbacks to Current Landscapes

- They are complex and cumbersome to produce.
- They are out of date as soon as they are published.
- They often do not integrate informatics tools.
- They do not always tap into the collective knowledge of experts.

CAMBIA is working to improve patent landscaping methodologies.

- Development of a patent search tool: the Patent Lens (www.patentlens.net).
- Development of patent sequence analysis tools:
 - The capacity to BLAST sequences in claims of granted patents and patent applications.
 - A sequence server that allows extraction of sequences from granted patents and patent applications in a useful, digital format.
 - Incorporation of patent sequences into genome browser.
- Development of a patent family visualization tool.
- Addition of new jurisdictions.
- Development of templates and workspaces.

CAMBIA's New Patent Sequence Server: Search for Patents by Sequence

<http://www.patentlens.net/sequence/blast/blast.html>

DNA, RNA and Protein Sequences extracted from Patent Documents – serving 51,861,977 sequences

Sequence Search Facility

BLAST	By Patent	» Patent Search
Important Information		
This utility is provided by CAMBIA's Patent Lens Sequence Project . It uses NCBI's BLAST software to search sequences that are specifically listed in U.S. patents and published patent applications. Sequence data was last updated on 19/Feb/08.		
Program, Database & Sequence		
Program	<input type="text" value="blastn"/> <input type="checkbox"/> Use MegaBlast	
Database	<input type="text" value="US Grants nt in claims"/>	
<input checked="" type="radio"/> Enter sequence below in FASTA format		
<input type="text"/>		
<input type="radio"/> Or load it from disk <input type="text"/> <input <="" td="" type="button" value="Browse..."/>		
Set subsequence: From <input type="text"/> To <input type="text"/>		
<input type="button" value="Search"/> <input type="button" value="+ More Options"/> <input type="button" value="Reset"/>		

Please report any search issues to webmaster@cambia.org | [Disclaimer](#) | Patent Lens – a service of [CAMBIA](#).

CAMBIA's New Patent Sequence Server: Extract Sequences from Patent Documents

<http://www.patentlens.net/sequence/blast/seqserver.html>

DNA, RNA and Protein Sequences extracted from Patent Documents – serving 51,861,977 sequences

Sequence Search Facility

BLAST **By Patent** » Patent Search

New Search by Patent New Search by Genbank Id

Search by Patent

Enter a patent or application number to view any associated DNA/RNA and protein sequences.

Patent * :
 e.g. US5432081 or US20050289672 or EP1298222.

SEQ ID NOS:

Only sequences in the claims :

* required field

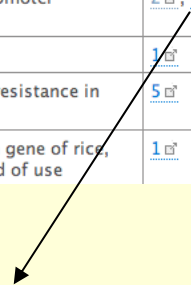
Found in US6720409	Download Sequences ¹
5 sequences	
3 DNA/RNA sequences	download
2 protein sequences	download

¹ Download sequences in [FASTA format](#) to a file on your computer. All of the sequence data can be downloaded from the [CAMBIA Sequence Project](#) page.

Please report any search issues to webmaster@cambia.org | Disclaimer | Patent Lens – a service of [CAMBIA](#).

Landscape Sequences are Linked to the Sequence Server

Patent No.	Assignee	Title	SEQ ID with homology to rice	Subject of Patent
6376750	Academia Sinica	Plant seedling and embryo promoter	2 , 5	plant seedling and embryo promoter
7291768	Academia Sinica	Plant MYB proteins	1	MYB proteins
6995253	Advanced Research & Technology Institute (Bloomington, IN, US)	Genes for regulating disease resistance in plants	5	Genes for regulating disease resistance in plants
7132587	Agency of Industrial Science and Technology	Non-autonomous transposon gene of rice, transformed plant and method of use	1	Method to detect transposon



[New Search by Patent](#)
[New Search by Genbank Id](#)
[Back to Search](#)
[Back to Sequence List](#)

US6376750 SEQ ID NO. 5 - DNA/RNA Sequence

SEQ ID NO	Length	In Claims	Genbank Id	BLAST	Organism
5	1730	yes	34603261	patents NCBI	N/A

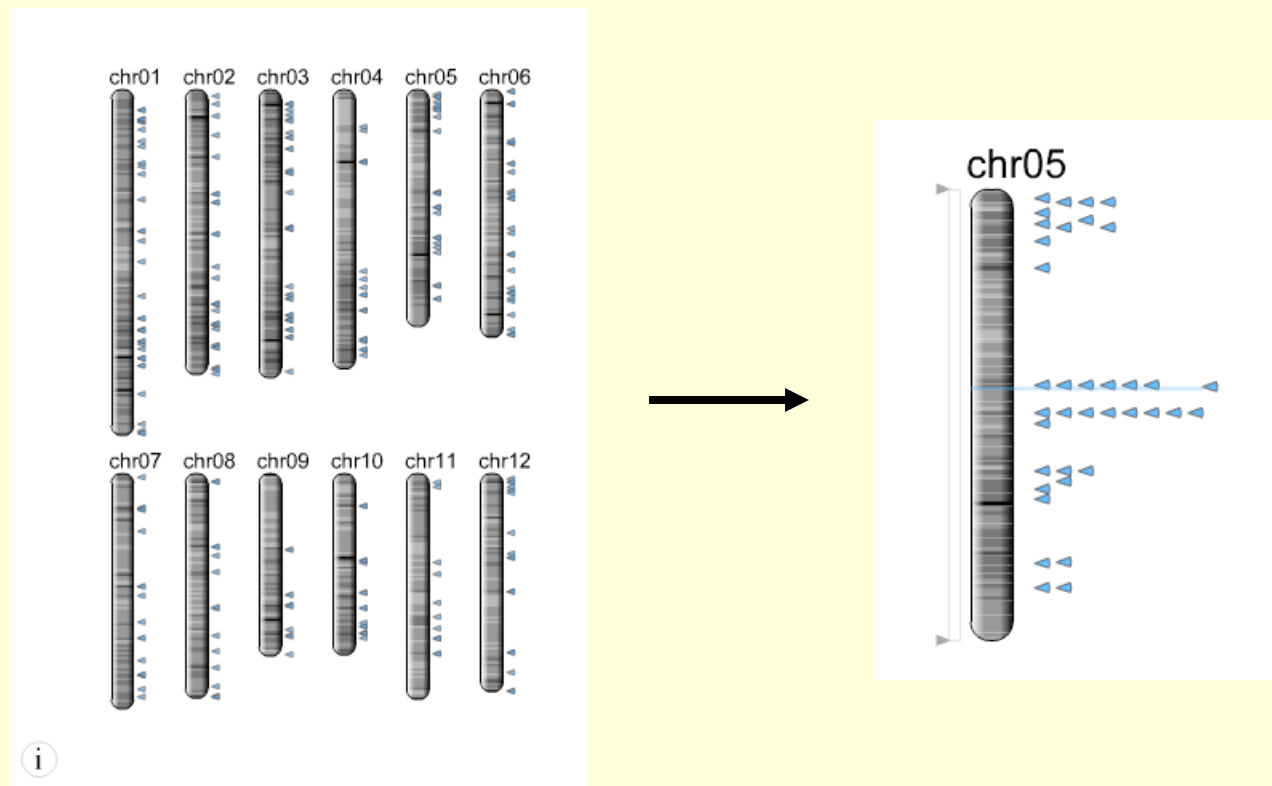
```

>US6376750_5 Sequence 5 from Patent US 6376750 inClaims gi: 34603261
TAGCTTCTAAATGTTAGTAGTATCAATAGATTGTTAATTTAACTGGCCATGGAAGAATGGTATGGCATCAATG
GCATGACCGTTTCTATAAAACCCCTCTTATTGATCAATGCATGATATCTTAAATTAATCCCTTTCCCTTTCTCTTC
TAAGGTGATGTTTGGAAACAGATACTTAACTTTAGTCTATATATTTAGACACTAATTTAGAGTATTAATATAGACTACT
TACAAAACATAATACATAAATGAAAGCTAATTTGCGAGATAAATTTTAAAGCCTAATTAATCTATAATTAGAGAATTTT
TACTGTACCATCATATAGGCATATCATGGATTAATAGGCTCAATAGATTTGCTCGCGAATAGTCCGAGATATGGAT
GAGTTTATGATAGTCTACGTTAATATTTATAATTAGTGTCCAAACATCCCATGTAATAGGACTTAAAGTTTGTAGT
CCCATCTAAACAGGGCTAAGTCTCTTAAATCTGTTACTCATATAACTGTCTAAGTAAAGTTAAGGTTGTAT
ATCATATCATCGTACGTTATATATATGATCCCTGCACCTCTCTTTTATAGAATGGACGAGACTCTTTTCTGTATAT
GTAGCGGCTTGTACTCTTGTAGTACCAATTTTGGCTCCCATTTTGACGAGACGACTGGCGTGCCATTTTGGCTCCTGGT
TCATTACAGTCTAATTTGGTGACAAACAAAGGAACAATAGGTCCTAGCGGTAGGACATGGACTCTGA
ATCCAGTAACCCGAGTTCAAATCTCGGTGGGACCTTAAATTTCTCGGTTTTATTTCTGCGCTGAGCTTATTGCTCCTC
CTGATTTTTTGTGTTGCTAATTTTCTCGCCGAAAAATGATCAAACCTCGTCAATCTACTCCTTTGAGAGCTTACTG
TGAATTTGCTCTCTCGAAGTTTCTAATTTTACTCTCTGTTATGAAAAATTTTCAATGCTAGAAATGATTACATTTG
GAAATGGAGAGAACTCGTTTGTGTTATTTAATCTTCCCTGATTTTTTCCACACCAAAACATATATTGTGATAAT
GAGTATGCTACCGGCTGACGTAACAGGTTTACTCCCTCCGCCCCAAAAAGACAAACCTGAGTTTTCATGTCCAA
TGTTTGTATATATTTTGAATAAATGAAAAATTTAAAGCCAGTACGTATAAAGTATTAATCATATTTTATCAT
ATAACAACAATGAAAAATCAATTTATAAAAAATTTTTCATATAAGACGGACAGTTAAACGTTGGACACGAAAAATCTAGGAT
TTATTTTTTTTATAGAGGGAGTACGAGGTAATAAATCGTCTCAGCGCTTCAGAAAAAAAAGGACAAAAATCCTCAGC
GCCAACCGACTCCGCTCCACAGACCACAGCCGCCAAAGTGTGCGAGGACAACGGCGGGCGGGCGGGCTAGGTTTTTG
CTGCACCCGACGCCACCGCCACCCAGCGAGTGTGGTGGGGCGGGCGCCCATAAAAAATATCTAGCGGGCCCATGTA
CTGCCAGAAAAATATCTTCTCCCGCTCGGGATCTTATCTCCGCTCGCGGGGTGCCGATCAGATCAGGAC
GGCCCGTGGGGCTATAAAGGAGGGGGGTAGGGCAAGCATGCTCTCT
    
```



Genome Browser for Rice Genome Landscape

- Incorporated results into Gbrowse genome browser.
- Bands on chromosomes represent applications and blue arrows represent grants.
- Shift-click to access positional data.

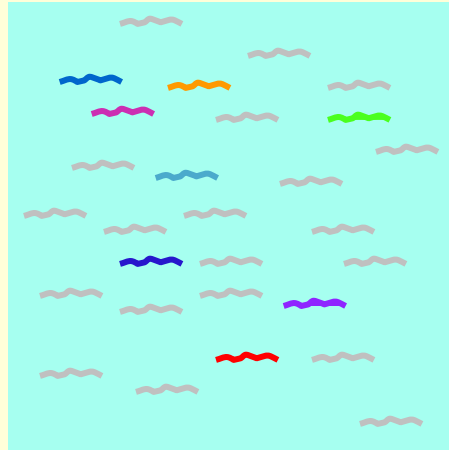


Sequence-Based Patent Landscapes

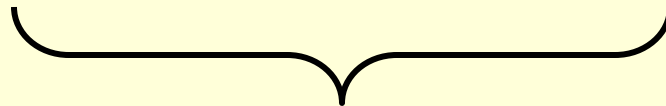
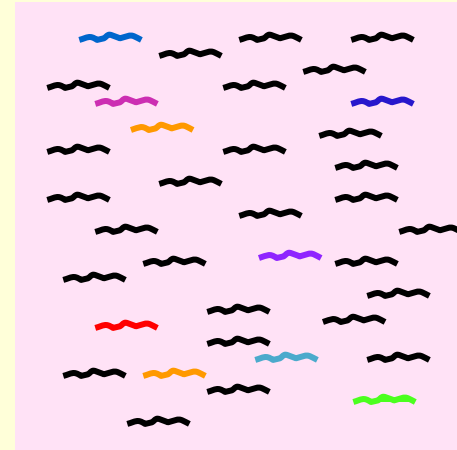
- Use a program that identifies SEQ ID numbers in the claims of patents.
- Correlate SEQ ID numbers with patent sequences to produce a database that contains *only* sequences that are recited in claims.
- Use sequence alignment tools such as BLAST to identify genome or proteome sequences that match the sequences recited in claims.
- Based on the methodology of Jensen and Murray (Science 2005;□ 310, 239-240).

For example:

patent claim sequences



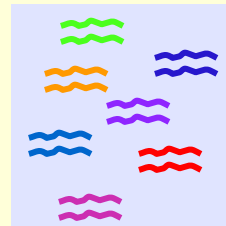
rice genome sequences



BLAST



sequence matches



CAMBIA

Advantages of Sequence-Based Landscaping Methodologies

- Allows focus on patents that specifically claim DNA or protein sequences.
- Allows identification of patents that claim specific sequences.
- Not biased by the text that is used to describe the invention.
- *Complementary* to key-word based searches.

Limitations of Sequence-Based Landscaping Methodologies

- May miss claims that only *refer* to sequences in the specification but do not have recognizable “SEQ ID NO” language.

An isolated nucleic acid molecule comprising a sequence encoding an influenza type hemagglutinin polypeptide.

- May miss claims that recite short sequences.

An isolated nucleic acid molecule comprising the sequence of SEQ ID NO:2. (where SEQ ID NO:2 is only 20 nt in length)

- May include sequences from unrelated species that should be excluded due to claim language or the specification.

An isolated nucleic acid molecule comprising the sequence of SEQ ID NO:2 wherein the the sequence encodes a maize protein.

- **Bottom line:** still need a human to analyze claims!

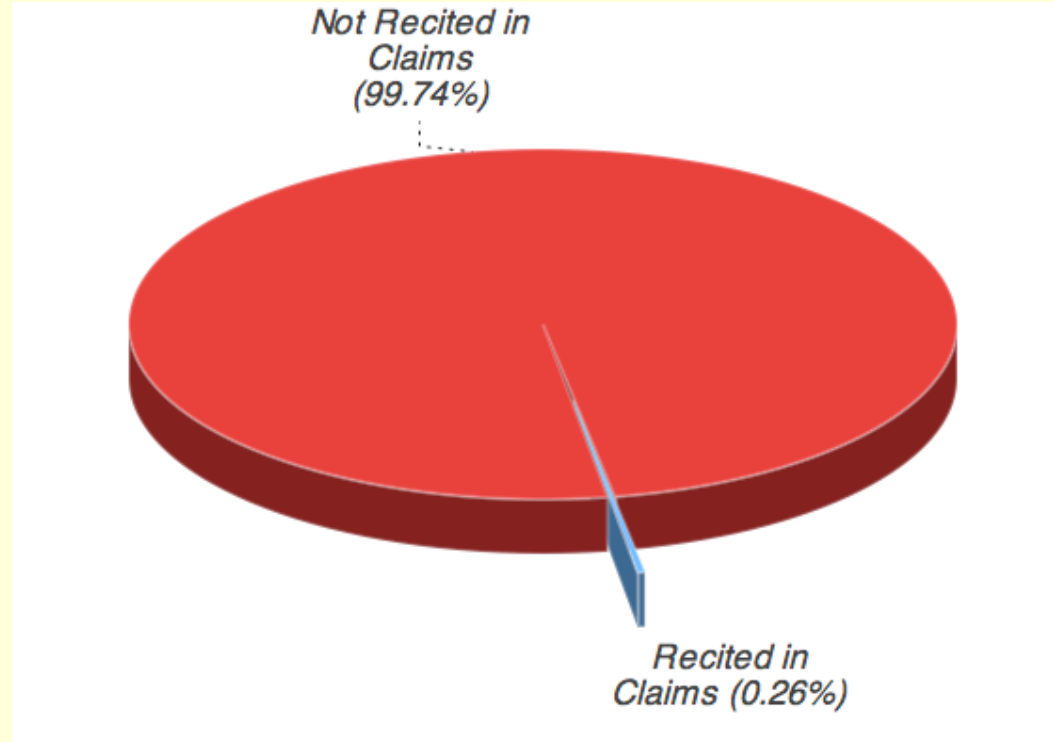
CAMBIA's Rice Genome Analysis

<http://www.patentlens.net/daisy/RiceGenome/3648.html>

- Analyzed both granted U.S. patents and U.S. patent applications
- Used the *Oryza sativa* genome database from TIGR.
- Only analyzed nucleotide sequences.
- Included the entire genome--both coding and non-coding sequences (e.g., promoter, activator sequences).
- Did *not* exclude highly similar sequences from other plant species.
- Sequence matches had an E value of $1e^{-200}$ or less, and were at least 150 nucleotides in length.
- Mapped sequence matches onto the rice genome.
- Created genome browser to display results.

What We Found

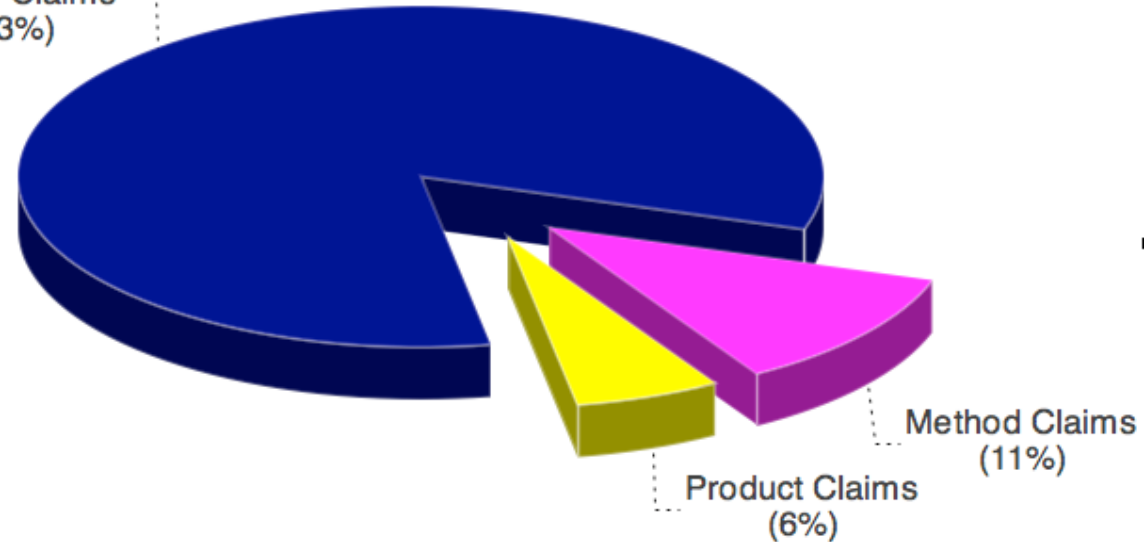
Only 0.26% of the rice genome and less than 1.0% of coding sequences are claimed in U.S. patents.



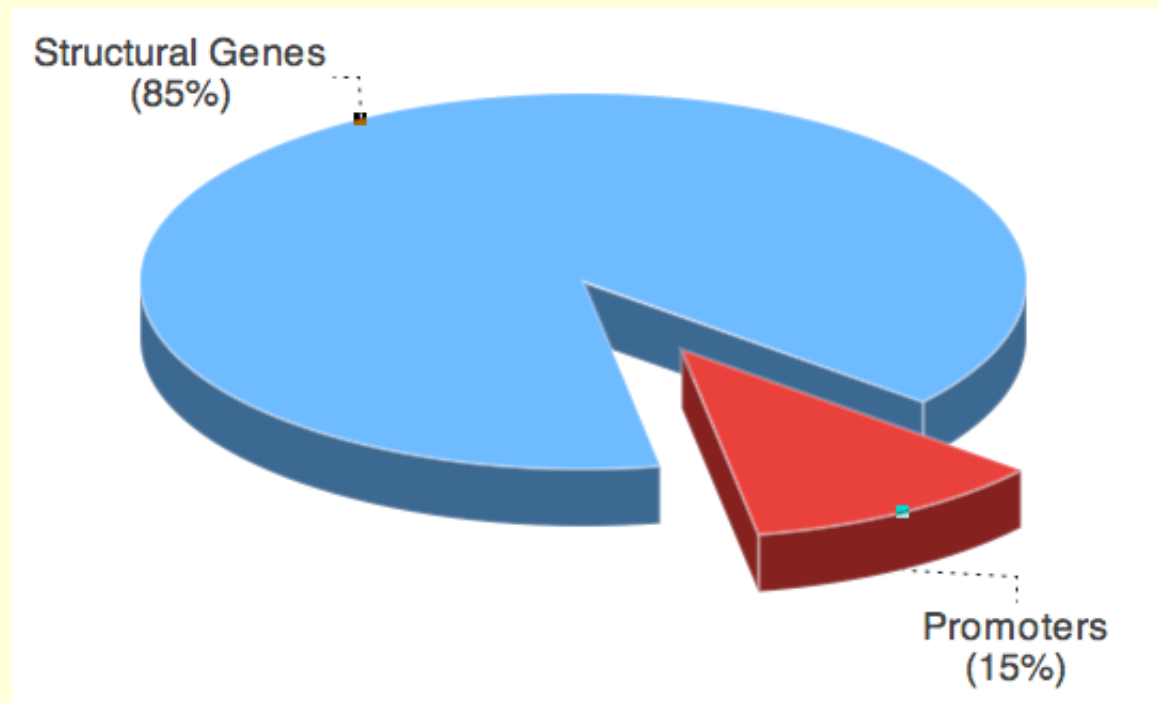
Types of Claims

- 182 granted U.S. patents recite rice sequences; of these, 151 (83%) have claims that explicitly claim rice sequences or sequences highly similar to rice.

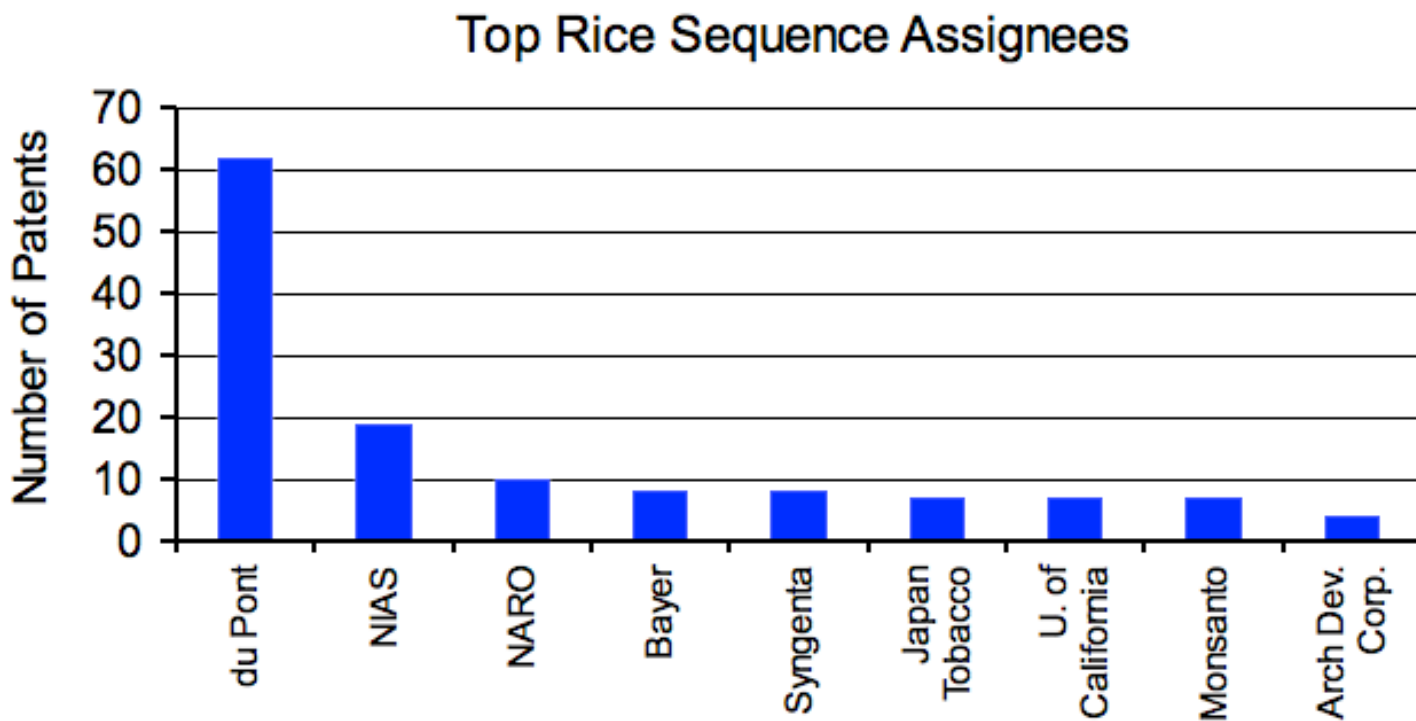
Composition of
Matter Claims
(83%)



85% of the composition of matter claims were to structural genes.



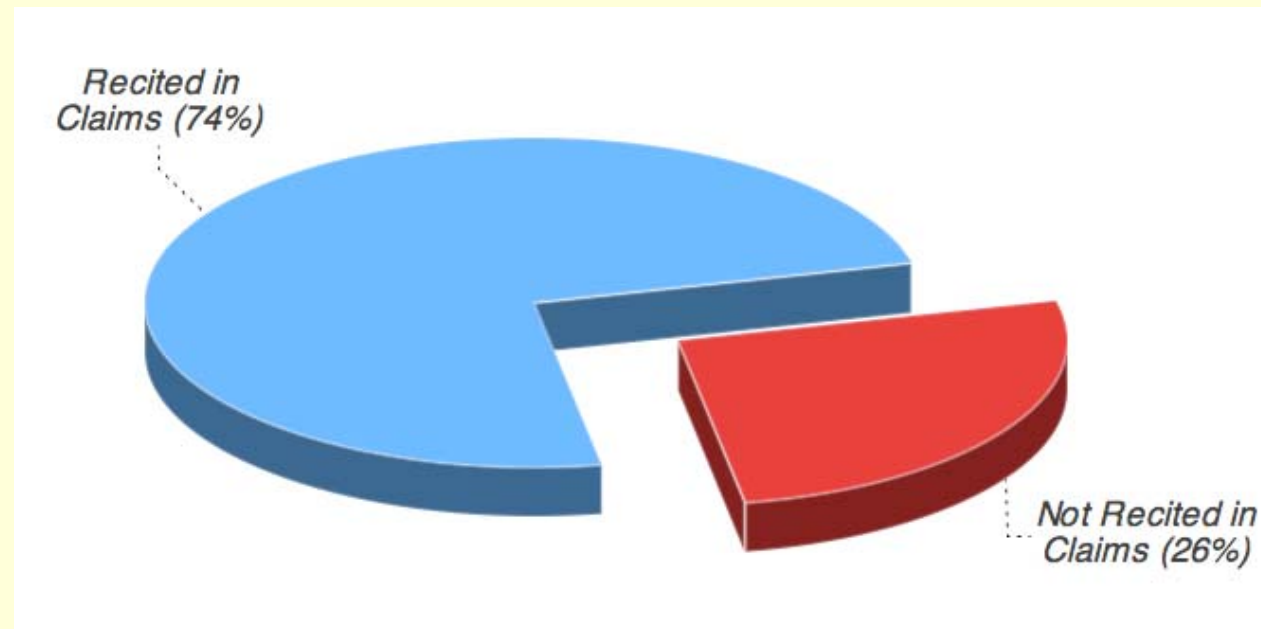
DuPont is the top assignee for granted U.S. patents.



* DuPont includes Pioneer Hi-Bred.

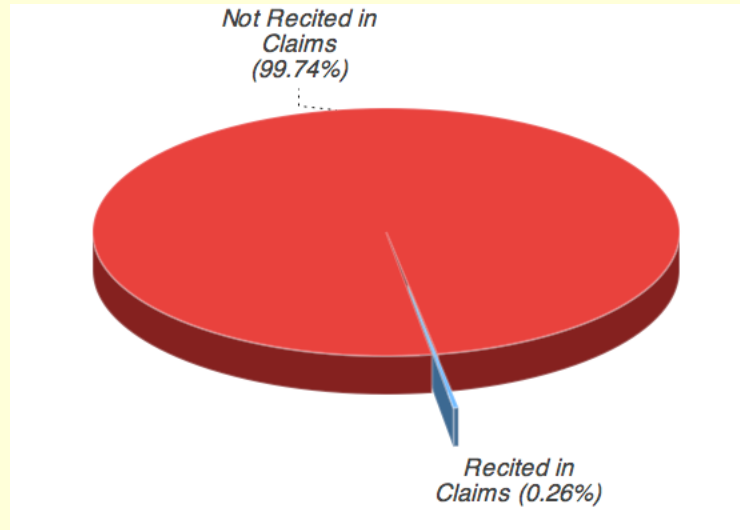
The applications tell a different story.

- While there are more patent applications - 313 U.S. applications - that recited rice sequences in claims, the sequences encompass about 74% of the rice genome.

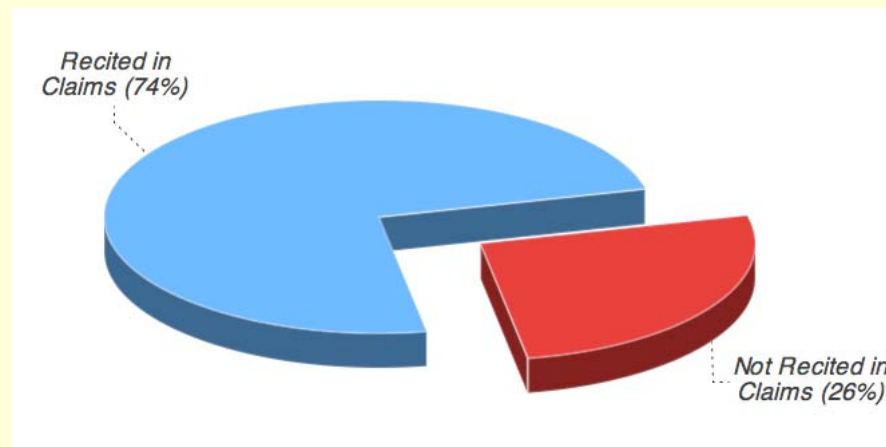


Rice Grants vs. Applications

Grants:



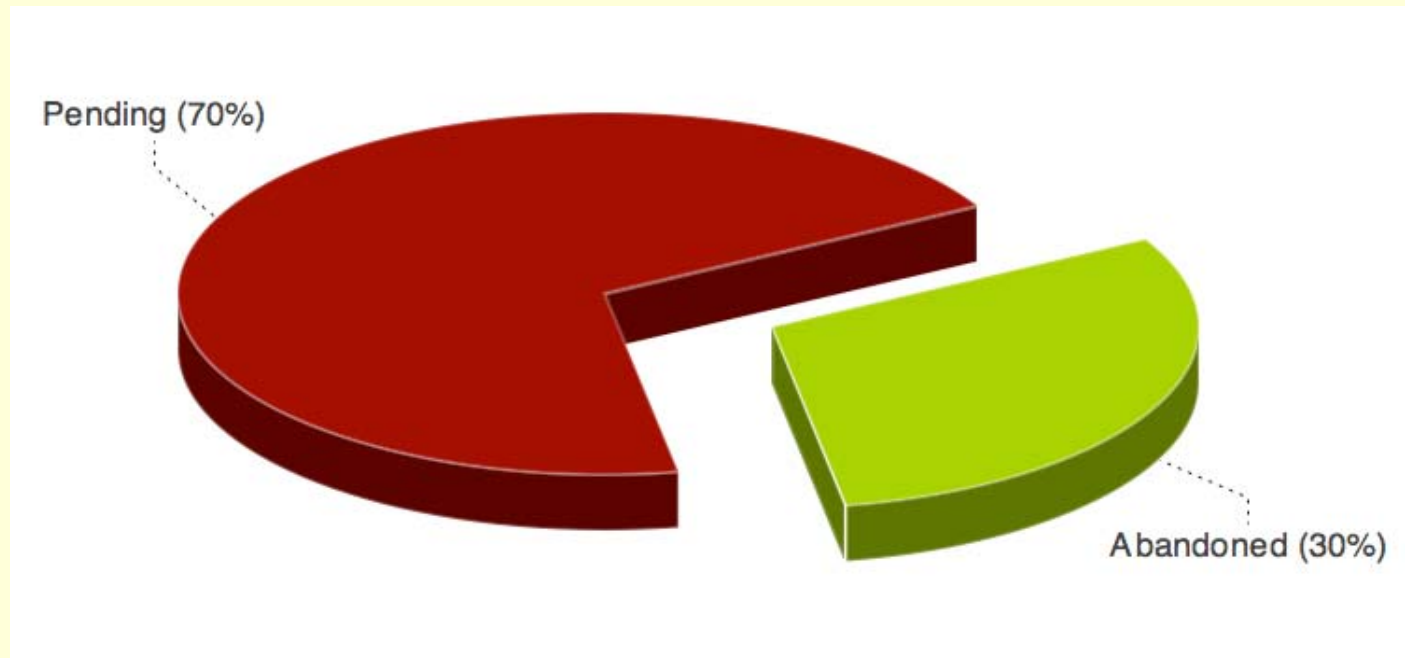
Applications:



Degree of Patent Application Coverage Largely Due to “Bulk Sequence Applications”

- Bulk sequence applications are applications that initially claim a large number of nucleotide and/or amino acid sequences, e.g., more than 1,000.
- Bulk sequence applications do not result in bulk sequence *patents*.

30% of the Rice Genome Applications are Abandoned



Our results are both an overcount and an undercount.

- *Overcount* because they include patents that recite sequences in claims but do not explicitly claim the sequences.
- *Undercount* because many nucleotide sequences are claimed based on the amino acid sequences that they encode, for example, a claim that reads:

Any nucleotide sequence that encodes the amino acid sequence of SEQ ID NO:2.

- This type of claim would not be included in the analysis because it recites an *amino acid* sequence.

Rice Genome Landscape Conclusions

- Only 0.26% of the rice genome is claimed in granted U.S. patents.
- In contrast, 74% of the genome is claimed in U.S. patent applications.
- The percent genome coverage is unlikely to increase dramatically.

Next step:

- Analysis in key rice-growing jurisdictions.



Live Demonstration of the Rice Genome Landscape