

USING SOYBEAN GENE EXPRESS FOR GENE EXPRESSION PROFILING IN SOYBEAN

**BINNECK E¹, NASCIMENTO L S², ABDELNOOR R V¹, ALMEIDA A M R¹,
NEPOMUCENO A L¹, MARCELINO-GUIMARAES F C¹ AND HUNGRIAM¹**

¹*Embrapa Soybean, Londrina, PR, 86001-970, Brazil*

²*State University of Londrina (UEL), Londrina, PR, 86051-980, Brazil*

E-mail: binneck@cnpso.embrapa.br

The latest advances in understanding the composition and organization of the soybean genome have culminated in its publication. This, in conjunction with the development of next-generation sequencing or high-throughput sequencing technologies, is opening new doors into the field of molecular biology studies in soybean. Among the major applications of next generation sequencing is RNA-Seq applied to gene expression profiling with several advantages over hybridization-based methods like microarrays. The background error rate is extremely low; the capacity for data collection is greater; the results are digital and use a discrete distribution; the dynamic range of expression levels is greater; and transcripts can be detected without the prerequisite of pre-existing probes.

Here we present a case study on using Soybean Gene Express, a database system for managing, analyzing, visualizing, and extracting useful information from RNA-Seq data. The system supports Brazilian Soybean Genome Consortium (GENOSOJA, <http://bioinfo.cnpso.embrapa.br/genosoja/>) researchers, going from raw data to visual reports comparing the levels of gene expression between pairs of samples or multiple samples following different biological samples/conditions. Also, it allows exploring the power of link-based integration with other sources of biological meaning information and databases such as ontologies, KEGG pathways, Pfam, Smart, Panther and others, to uncover the biological meaning of these data. Several criteria are available in the system to explore what the data are telling us about the biological conditions being investigated. For example, using filtering thresholds we can identify groups of genes that are differentially expressed among several conditions. The ultimate goal is to facilitate the discovery of new biological insights into fundamental molecular processes as responses to biotic and abiotic stresses in soybean.

Using Soybean Gene Express for gene expression profiling in soybean



Embrapa

Soja

Soybean: The chief oil crop in the world



Genome sequence of the palaeopolyploid soybean

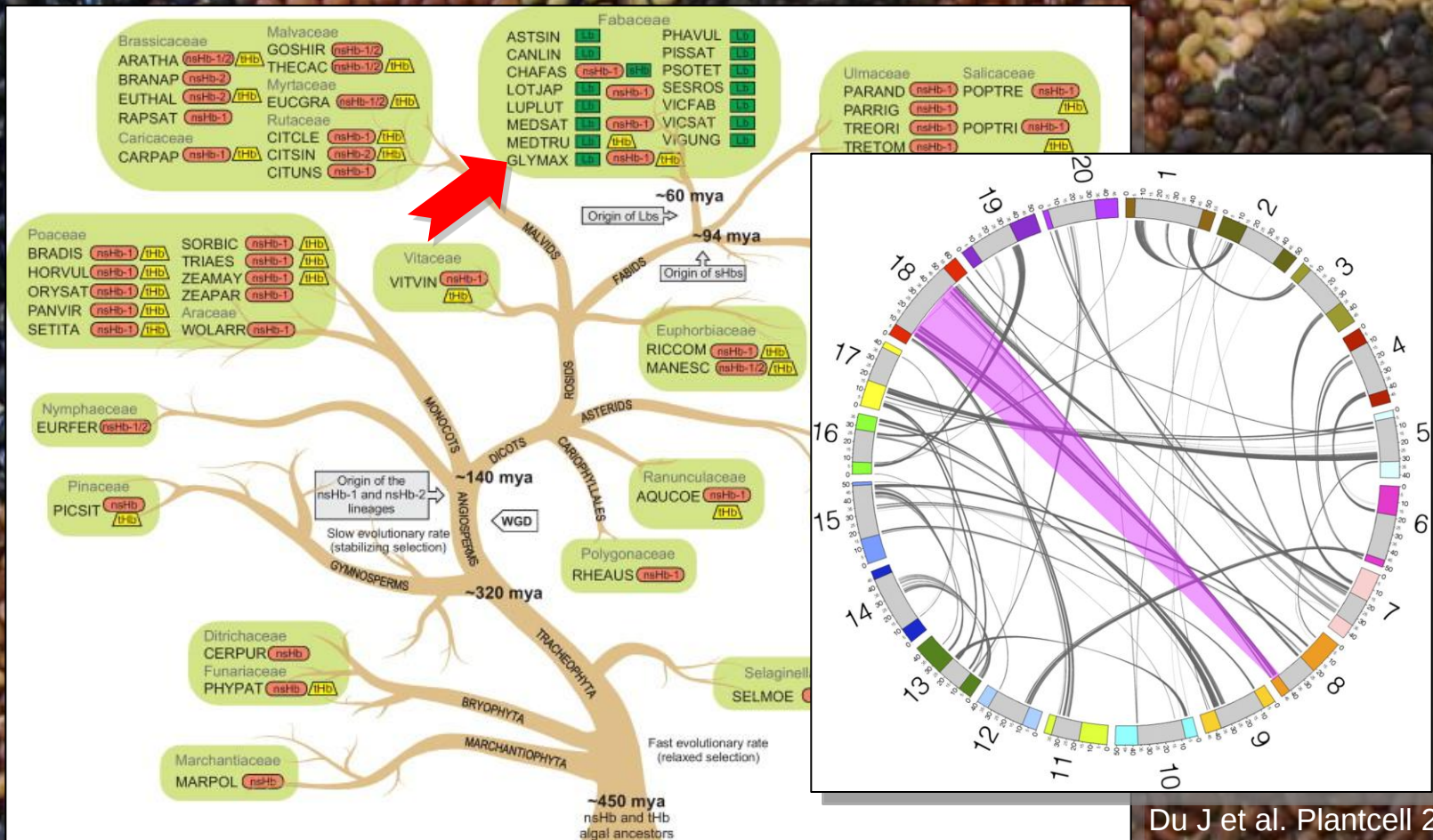
Soybean (*Glycine max*) is one of the most important crop plants for seed protein and oil content, and for its capacity to fix atmospheric nitrogen through symbioses with soil-borne microorganisms. We sequenced the 1.7-gigabase genome by a whole-genome shotgun approach and integrated it with physical and high-density genetic maps to create a chromosome-scale draft sequence assembly. We predict 46,430 protein-coding genes, 70% more than *Arabidopsis* and similar to the poplar genome which, like soybean, is an ancient polyploid (palaeopolyploid). About 78% of the predicted genes occur in chromosome ends, which comprise less than one-half of the genome but account for nearly all of the genetic recombination. Genome duplications occurred at approximately 59 and 13 million years ago, resulting in a highly duplicated genome with nearly 75% of the genes present in multiple copies. The two duplication events were followed by gene diversification and loss, and numerous chromosome rearrangements. An accurate soybean genome sequence will facilitate the identification of the genetic basis of many soybean traits, and accelerate the creation of improved soybean varieties.

A striking feature of the soybean genome is that 57% of the genomic sequence occurs in repeat-rich, low-recombination heterochromatic regions surrounding the centromeres. The average ratio of genetic-physical distance is 1 cM per 197 kb in euchromatic regions, and 1 cM per 3.5 Mb in heterochromatic regions (see Supplementary Information section 1.8). For reference, these proportions are similar to those in *Sorghum*, in which 62% of the sequence is heterochromatic, and different than in rice, with 15% in heterochromatin⁸. In



1700 Km

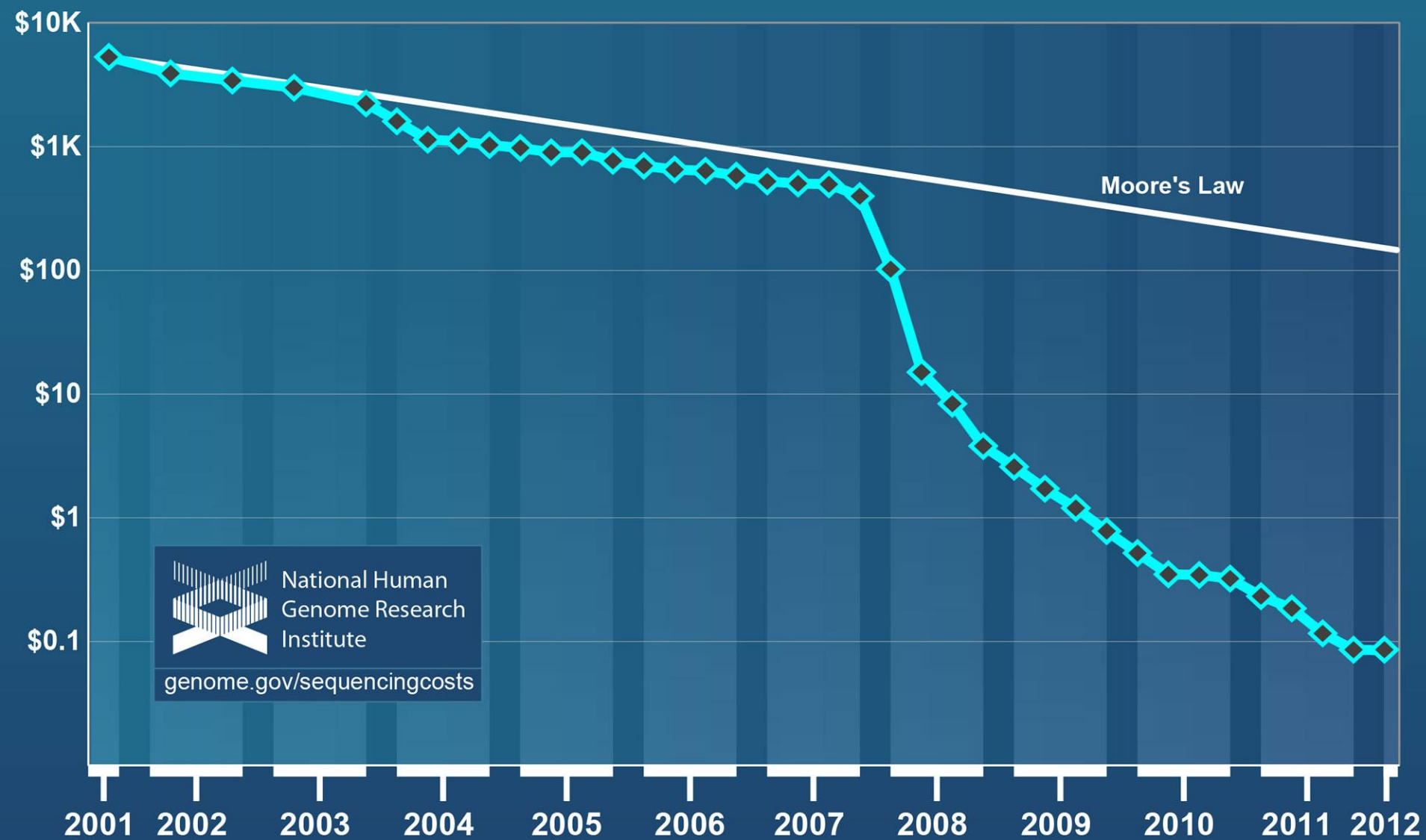




Du J et al. Plantcell 2012;24:21-32

Vázquez-Limón C, Hoogewijs D, Vinogradov SN, Arredondo-Peter R. The evolution of land plant hemoglobins. Plant Sci. 2012 Aug; 91-192:71-81.

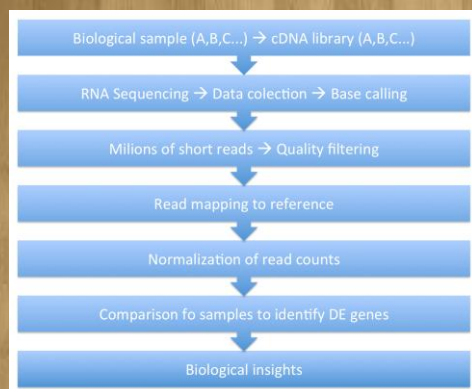
Cost per Raw Megabase of DNA Sequence



Next Generation Sequencing



**Transcriptome
profiling by
RNA-Seq**



Pipeline

Show 10 entries

Gene	Transcript	Chromosome Name	Strand	Gene start bp	Gene end bp	Pfam Description	Panther Description	GGW 21	GGW 22	GGW 23	GO:BP
☐ Glyma13g41330	Glyma13g41330.1	Gm13	-1	41655421	41657296	ZIP Zinc transporter	ZINC-IRON TRANSPORTER	187.305	127.685	257.233	GO:0006313 GO:0006829
☐ Glyma20g02770	Glyma20g02770.1	Gm20	1	2356369	2358876	ZIP Zinc transporter	ZINC-IRON TRANSPORTER	174.299	28.6965	209.145	GO:0006826 GO:0006828 GO:0006829 GO:0006879 GO:0015691
☐ Glyma06g05460	Glyma06g05460.1	Gm06	-1	3877551	3881484	ZIP Zinc transporter	ZINC-IRON TRANSPORTER	0.591063	0.0432569	0.0449852	GO:0006829 GO:0009624
☐ Glyma15g04100	Glyma15g04100.1	Gm15	1	2872362	2873863	ZIP Zinc transporter	ZINC-IRON TRANSPORTER	0	0	0	GO:0006829
☐ Glyma02g13950	Glyma02g13950.1	Gm02	1	12266532	12268224	ZIP Zinc transporter	ZINC-IRON TRANSPORTER	0	0	0	GO:0006826 GO:0006828 GO:0006829 GO:0006879 GO:0015691
☐ Glyma14g37560	Glyma14g37560.1	Gm14	1	12931405	12931974	ZIP Zinc transporter	ZINC-IRON TRANSPORTER	0	0	7.59743	GO:0006826 GO:0006828 GO:0006829 GO:0006879 GO:0009624 GO:0015691
☐ Glyma13g10780	Glyma13g10780.1	Gm13	1	38642348	38645535	ZIP Zinc transporter	ZINC-IRON TRANSPORTER	42.4781	93.0256	117.871	GO:0006829 GO:0009624
☐ Glyma17g34660	Glyma17g34660.1	Gm17	1	7515288	7518066	ZIP Zinc transporter	ZINC-IRON TRANSPORTER	7.99329	0	3.63353	GO:0006313 GO:0006829 GO:0006879 GO:0015674
☐ Glyma18g08760	Glyma18g08760.1	Gm18	1	13016630	13019134	ZIP Zinc transporter	ZINC-IRON TRANSPORTER	4.12923	0	10.48	GO:0006829 GO:0010043
☐ Glyma08g17530	Glyma08g17530.1	Gm08	-1								

Showing 1 to 10 of 55,787 entries

Utilize os operadores lógicos para refinar filtro

- < : Menor (Ex.: <123)
- <= : Menor igual (Ex.: <=123)
- > : Maior (Ex.: >123)
- >= : Maior igual (Ex.: >=123)
- <> : Diferente (Ex.: <>123)

Web Interface



Database

**A tool for *in silico* studies
about the data.**

Biological sample (A,B,C...) → cDNA library (A,B,C...)

RNA Sequencing → Data collection → Base calling

Milions of short reads → Quality filtering

Read mapping to reference

Normalization of read counts

Comparison fo samples to identify DE genes

Biological insights

Step 1 - Select your options for **Experiment/Library**:
 Step 2 - Set your parameters for **GO query** (optional):
 Step 3 - Select your options for **Annotation** to be shown in the table of results (optional):
 Step 4 - Submit your job:

- ☒ GGW-21
- ☒ GGW-22
- ☒ GGW-23
- ☐ GGW-24
- ☐ GGW-25
- ☐ GGW-26
- ☐ GGW-27
- ☐ GGW-28
- ☐ GGW-29
- ☐ GGW-30
- ☐ GGW-31
- ☐ GGW-32

Search all columns:

Show entries

☐ Gene	Transcript	Chromosome Name	Strand	Gene start bp	Gene end bp	GGW 21	GGW 22	GGW 23
☐ Glyma10g39450	Glyma10g39450.1	Gm10	1	47119499	47121519	3494.88	1500.47	625.199
☐ Glyma17g03350	Glyma17g03350.1	Gm17	-1	2222082	2223155	3017.49	1391.28	400.9
☐ Glyma17g15690	Glyma17g15690.1	Gm17	-1	12412680	12415378	2511.98	2490.6	1514.85
☐ Glyma07g16420	Glyma07g16420.1	Gm07	1	16137826	16138557	2495.83	2928.36	1767.01
☐ Glyma16g06520	Glyma16g06520.1	Gm16	-1	5857348	5858417	2405.36	1685.21	395.939
☐ Glyma15g08300	Glyma15g08300.2	Gm15	-1	5871373	5873068	2274.5	833.729	264.008
☐ Glyma02g42860	Glyma02g42860.1	Gm02	-1	47801570	47802378	2268.43	4142.62	1258.4
☐ Glyma08g24760	Glyma08g24760.1	Gm08	-1	18917235	18918826	2149.63	1742.79	1534.94
☐ Glyma09g09430	Glyma09g09430.1	Gm09	-1	9260555	9264639	2149.38	539.062	34.506
☐ Glyma15g20970	Glyma15g20970.2	Gm15	1	19077254	19079891	2070.81	533.031	24.8701

Showing 1 to 10 of 55,787 entries

Fluoding.csv

Search in Sheet

Home Layout Tables Charts SmartArt Formulas Data Review

File Home Insert References Send To View

Font Color Background Color Borders Styles

Number Alignment Conditional Formatting

General Layout Styles Insert Delete Format Themes

Gene	Transcript	Chromosome	Strand	Gene start bp	Gene end bp	Plasmid	Plasmid Descr	GGW 21	GGW 22	GGW 23	GGW 24	GGW 25	GGW 26	GGW 27	GGW 28	GGW 29	GGW 30	GGW 31	GGW 32	GO-BP	GO-MF	GO-CC	Blat
1	Glyma08g13	Glyma08g13	Gm08	1	43094849	43094849	WIKY DNA	25.1282	24.0668	0	8.91806	0	0	0	0	0	0	0	0	0	0	0	0
2	Glyma16g10	Glyma16g10	Gm16	-1	13444507	13444507	GATA zinc fr	15.8898	0	10.1508	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Glyma16g10	Glyma16g10	Gm16	-1	13444507	13444507	GATA zinc fr	15.8898	0	10.1508	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Glyma16g10	Glyma16g10	Gm16	-1	13444507	13444507	GATA zinc fr	15.8898	0	10.1508	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Glyma16g10	Glyma16g10	Gm16	-1	13444507	13444507	GATA zinc fr	15.8898	0	10.1508	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Glyma16g10	Glyma16g10	Gm16	-1	13444507	13444507	GATA zinc fr	15.8898	0	10.1508	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Glyma16g10	Glyma16g10	Gm16	-1	13444507	13444507	GATA zinc fr	15.8898	0	10.1508	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Glyma16g10	Glyma16g10	Gm16	-1	13444507	13444507	GATA zinc fr	15.8898	0	10.1508	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Glyma16g10	Glyma16g10	Gm16	-1	13444507	13444507	GATA zinc fr	15.8898	0	10.1508	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Glyma16g10	Glyma16g10	Gm16	-1	13444507	13444507	GATA zinc fr	15.8898	0	10.1508	0	0	0	0	0	0	0	0	0	0	0	0	0

Normal View Ready

Step 1 - Select your options for **Experiment/Library**:
 Step 2 - Set your parameters for **GO query** (optional):
 Step 3 - Select your options for **Annotation** to be shown in the table of results (optional):
 Step 4 - Submit your job:
 Submit

PFAM
☒ PFAM id
☒ PFAM description
SMART
☐ Smart id
☐ Smart description
PANTHER
☐ Panther id
☐ Panther descriptionn
KOG
☐ KOG id
☐ KOG descriptionn
EC
☐ EC id
☐ EC descriptionn
KO
☐ KO id
☐ KO descriptionn

Gene	Transcript	Chromosome Name	Strand	Gene start bp	Gene end bp	Pfam Id	Pfam Description	GGW 24	GGW 25	GGW 36	GGW 30	GGW 31	GGW 32	GO:BP
Glyma0448s00210	Glyma0448s00210.1	scaffold_448	-1	4843	10153	PF01416	tRNA pseudouridine synthase	15.2445	2.26022	0	19.1212	3.0627	4.08209	GO:0001522 GO:0008033
Glyma12g28830	Glyma12g28830.1	Gm12	-1	32195886	32201611	PF01416	tRNA pseudouridine synthase	0.0110072	0.0451851	0.101697	0.40917	0.507509	0.0430514	GO:0001522 GO:0008033
Glyma16g00480	Glyma16g00480.1	Gm16	-1	169847	172509	PF01416	tRNA pseudouridine synthase	0	4.41412	0	5.41661	4.56746	7.36685	GO:0001522 GO:0008033
Glyma08g03090	Glyma08g03090.1	Gm08	1	2181894	2186503	PF01416	tRNA pseudouridine synthase	0	3.08255	0	0	4.00064	4.11257	GO:0001522 GO:0008033
Glyma02g39980	Glyma02g39980.1	Gm02	-1	45185801	45189290	PF01416	tRNA pseudouridine synthase	0	0	0	0	0	0	GO:0001522 GO:0008033
Glyma17g04570	Glyma17g04570.1	Gm17	1	3033354	3035721	PF01416	tRNA pseudouridine synthase	0	0	0	0	0	0	GO:0001522 GO:0008033
Glyma05g36480	Glyma05g36480.1	Gm05												
Glyma02g39970	Glyma02g39970.1	Gm02												
Glyma13g17960	Glyma13g17960.1	Gm13												

Showing 1 to 9 of 9 entries (filtered from 55,787 total entries)

PF01416 trna

Summary: tRNA pseudouridine synthase

PFAM includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

No Wikipedia article | [PFAM](#) | [InterPro](#)

This tab holds the annotation information that is stored in the Pfam database. As we move to using Wikipedia as our main source of annotation, the contents of this tab will be gradually replaced by the Wikipedia tab.

tRNA pseudouridine synthase [Get protein](#)

Involved in the formation of pseudouridine at the anticodon stem and loop of transfer RNA. Pseudouridine is an isomer of uridine (5 (beta-D-ribofuranosyl) uracil), and is the most abundant modified nucleoside found in all cellular RNAs. The TruA-like proteins also exhibit a conserved sequence with a strictly conserved aspartic acid, likely involved in catalysis.

Literature references

1. Anisimov V, Houtenberg C, Robert B, Grosjean H. Biochemistry 1998;37:7268-7276. Transfer RNA-pseudouridine synthase Pua1 of Saccharomyces cerevisiae contains one atom of zinc essential for its native conformation and RNA recognition. [PubMed:9383509](#)

Internal database links

SCOP: [DUF2344](#)

External database links

Pfam: [PF01416](#)
 PseudoFam: [PF01416](#)
 SCOP: [1JBB](#)
 Rfam: [RF00118](#)

Extraneous structure

[PDB entry 2Z8D](#): Crystal structure of tRNA pseudouridine synthase TruA in complex with a different substrate. [EMSD: 1](#)

9 entries (filtered from 55,787 total entries)

>> Step 1 - Select your options for **Experiment/Library**:
 >> Step 2 - Set your parameters for **GO query** (optional):
 >> Step 3 - Select your options for **Annotation** to be shown in the table of results (optional):
 >> Step 4 - Submit your job:

☒ GO ☒ AND ☐ OR (Select the logic)
☒ GO number ☐ GO description
 Insert your query term or a comma separated list of terms (Ex.: 0003700,0043565 for GO number or

Search all columns:

strand	Gene start bp	Gene end bp	Pfam Id	Pfam Description	GGW 24	GGW 25	GGW 36	GGW 30	GGW 31	GGW 32	GO:BP	GO:MF	GO:CC	Blast
1	43094949	43098475	PF03106	WRKY DNA-binding domain	25.1282	24.0668	0	8.91806	0	0	GO:0006355	GO:0003700 GO:0043565 GO:0046872	GO:0005634 GO:0016020	Blast
-1	7334635	7335187	PF00808	Histone-like transcription factor (CBF/NF-Y) and archaeal histone	15.9856	0	0	39.479	14.1226	0	GO:0006355 GO:0009414	GO:0003700 GO:0043565	GO:0016602	Blast
-1	13444507	13445181	PF00320	GATA zinc finger	15.8898	0	0	10.1508	0	0	GO:0		GO:0005730	Blast
1	5308573	5311368	PF03106	WRKY DNA-binding domain	11.4433	4.1579	0	11.2949	0	0	GO:0		GO:0005634	Blast
-1	21617824	21623330	PF00664!PF00005	ABC transporter transmembrane region!ABC transporter	10.4354	2.0788	0	9.45493	2.79465	0	GO:0		GO:0005634 GO:0005886 GO:0016021	Blast
-1	11960949	11964356	PF03106	WRKY DNA-binding domain	9.19135	0	0	13.2573	0	0	GO:0006355	GO:0003700 GO:0043565	GO:0005634	Blast
-1	48986718	48990305			8.62246	0	0	9.11343	0	0	GO:0006355 GO:0009409 GO:0009862 GO:0042742	GO:0003700 GO:0005516 GO:0016563 GO:0016564 GO:0043565 GO:0046983	GO:0005634 GO:0005737	Blast

Term Information Close

name: sequence-specific DNA binding transcription factor activity

namespace: molecular_function

def: Interacting selectively and non-covalently with a specific DNA sequence in order to modulate transcription. The transcription factor may or may not also interact selectively with a protein or macromolecular complex. [GOC:curators, GOC:txnOH]

xref: Reactome:4325 sequence-specific DNA binding transcription factor activity

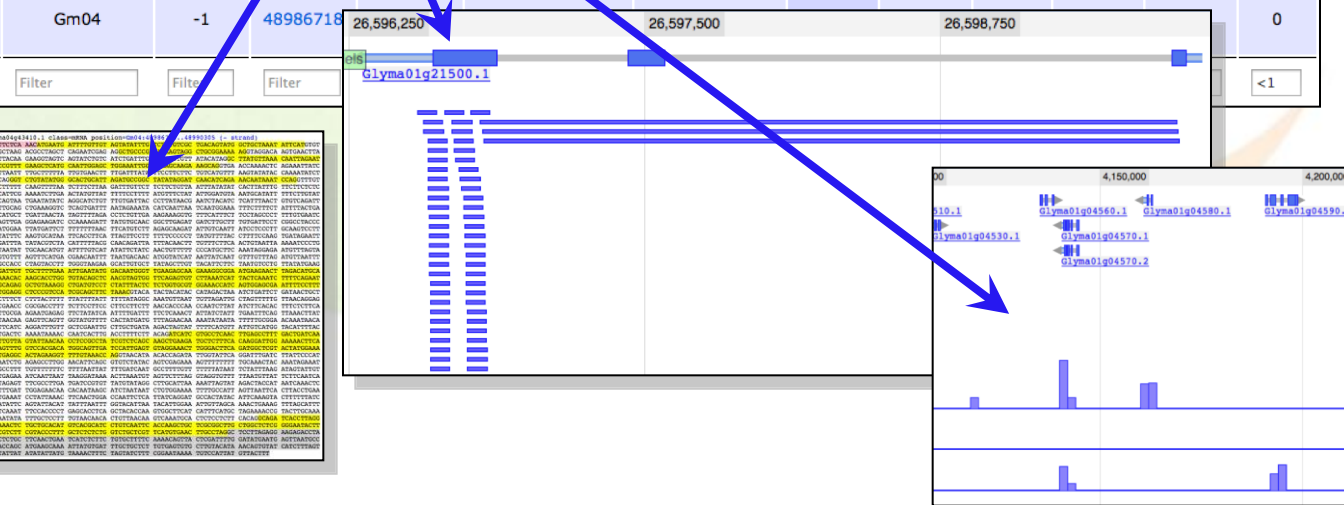
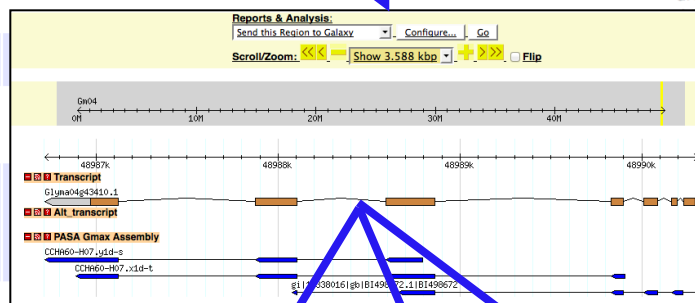
Show cDNA sequence data of the selected items

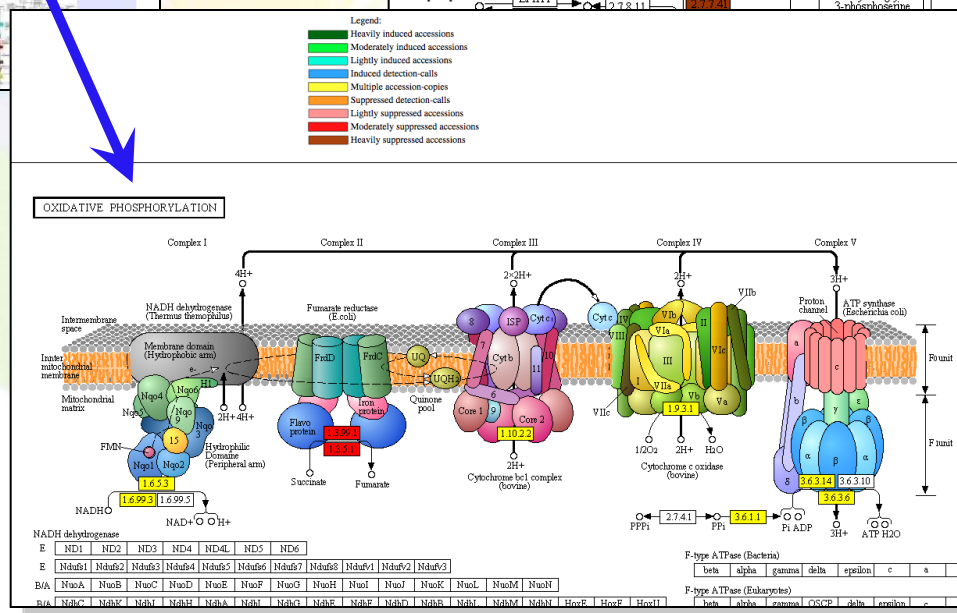
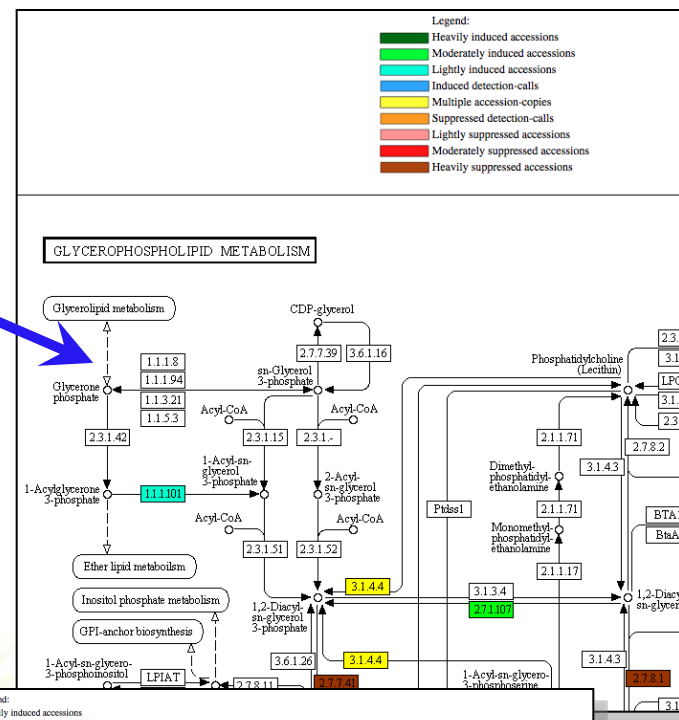
Copy CSV Excel PDF Print

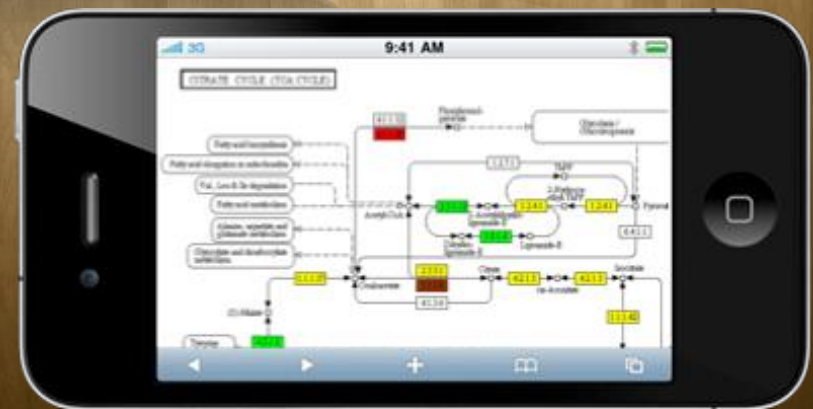
Search all columns:

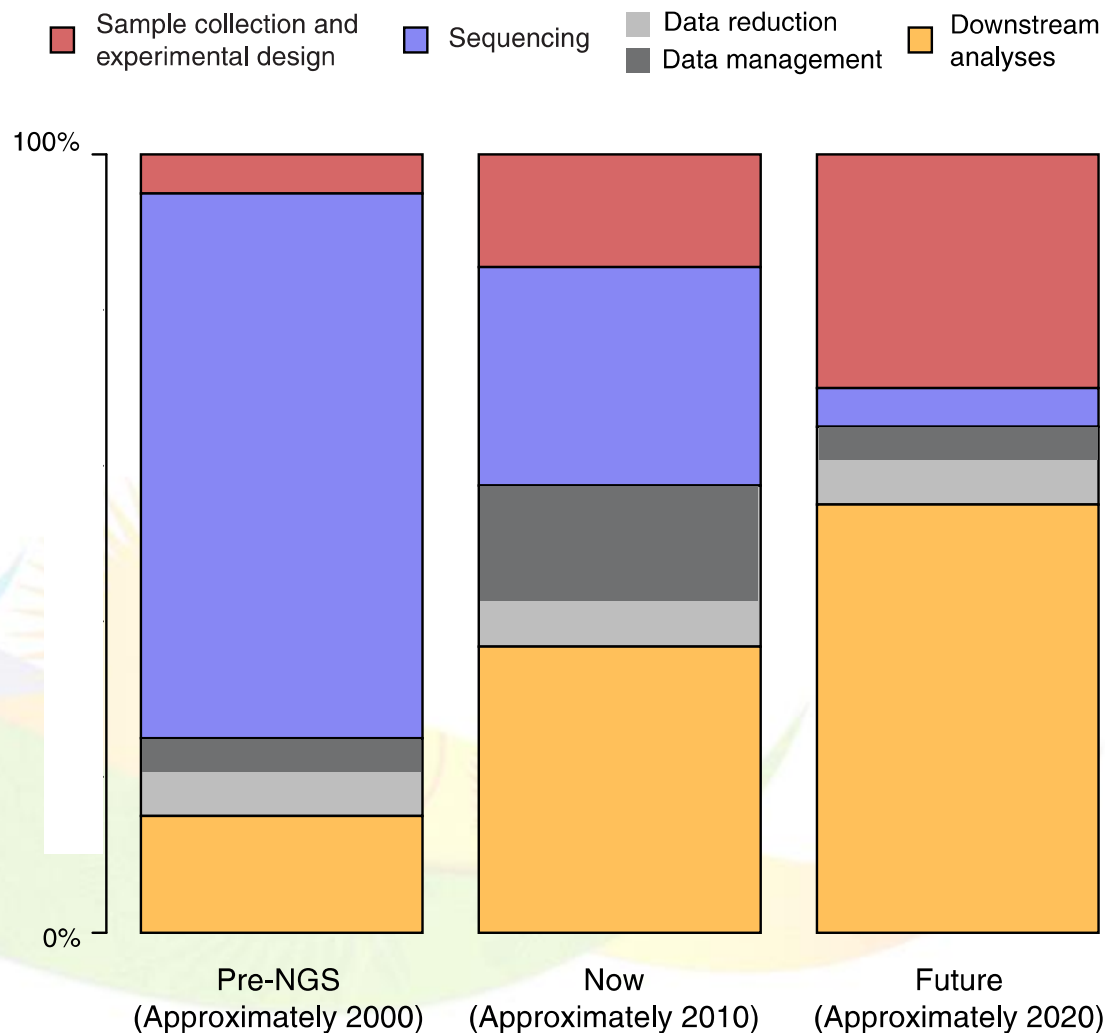
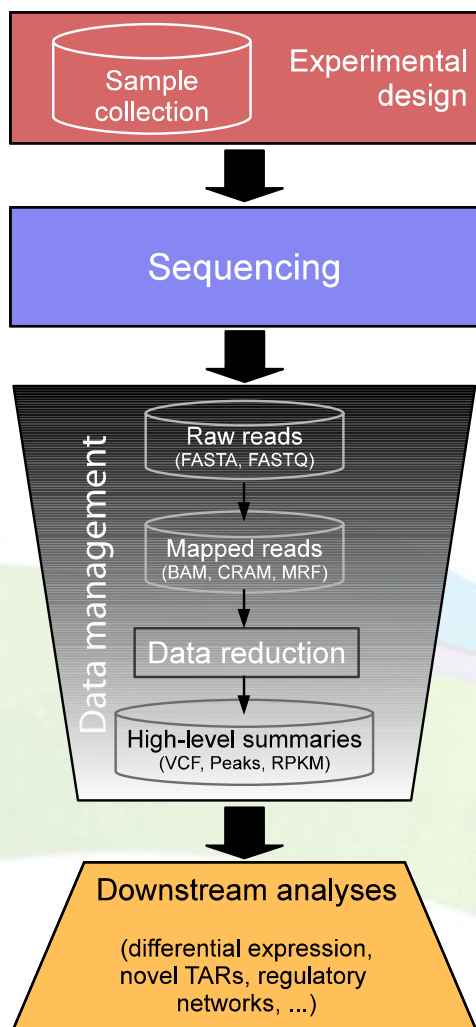
Show 10 entries

☐ Gene	Transcript	Chromosome Name	Strand	Gene start bp	Gene end bp	Pfam Id	Pfam Description	GGW 24	GGW 25	GGW 36	GGW 30	GGW 31	GGW 32
☐ Glyma08g43260	Glyma08g43260.1	Gm08	1	43094949	43098475	PF03106	WRKY DNA-binding domain	25.1282	24.0668	0	8.91806	0	0
☒ Glyma18g08620	Glyma18g08620.1	Gm18	-1	7334635	7335187	PF00808	Histone-like transcription factor (CBF/NF-Y) and archaeal	15.9856	0	0	39.479	14.1226	0
☒ Glyma01g10390	Glyma01g10390.1							15.8898	0	0	10.1508	0	0
☐ Glyma16g05880	Glyma16g05880.1							11.4433	4.1579	0	11.2949	0	0
☐ Glyma13g17930	Glyma13g17930.1							10.4354	2.0788	0	9.45493	2.79465	0
☒ Glyma06g15200	Glyma06g15200.1	Gm06	-1	11960949	11964356	PF03106	-binding domain	9.19135	0	0	13.2573	0	0
☒ Glyma04g43410	Glyma04g43410.1	Gm04	-1	48986718	26,596,250								0









2000

2010

2020



**This work was supported by National Council
for Scientific and Technological Development
(CNPq grant 552735) and Brazilian
Agricultural Research Corporation
(Embrapa).**

**We are thankful to all GenoSoja team
members and numerous collaborators and
data contributors.**

Contact

Eliseu Binneck

Embrapa Soybean

Londrina, PR, Brazil

eliseu.binneck@embrapa.br

+55 (43) 3371-6223

