

# ORegAnno: A Community-Based Annotation System for Literature-Derived Regulatory Sequences

Griffith OL<sup>†</sup>, Montgomery SB<sup>†</sup>, Bergman CM, Bilenky M, Chu B, Pleasance ED, Prychyna Y, Sleumer MC, Zhang X and Jones SJM



Advances in Genome Biology and Technology

Feb 9, 2007, Marco Island, FL, USA

# Why should we care?

- New genomes sequenced continuously
- Genomes need to be annotated
- medterms.com defines ‘genome annotation’ as:  
“The process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do.”

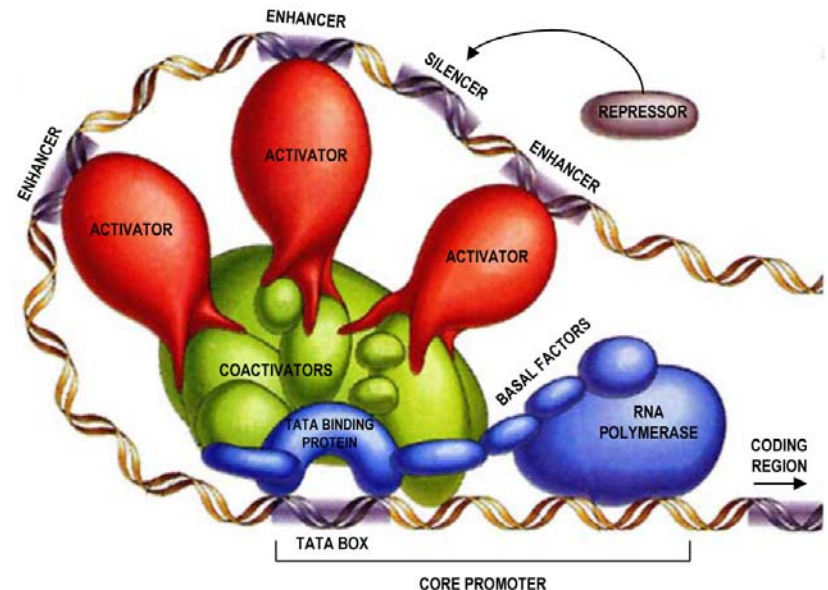
What’s missing here?

# Why should we care?

- New genomes sequenced continuously
- Genomes need to be annotated
- medterms.com defines ‘genome annotation’ as:  
“The process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do **and what sequences regulate their expression.**”

# Regulatory sequence annotation is difficult

- Gene regulation is complex
- Binding models are ill-defined
- Multiple TFs bind same sequence
- Sequences are degenerate
- Located at great distance, upstream, downstream or intronic
- Signal to noise problem
- Very few known binding sites for most known TFs
- Known sites that do exist are “hidden” in literature



Tjian, R. (1995) “Molecular Machines That Control Genes”; Scientific American, Feb 1995, p. 38.

# Open Regulatory Annotation Database ([www.oreganno.org](http://www.oreganno.org))

- Community-driven annotation of regulatory sequences reported in scientific literature.
- Open access and open source
- A large resource of experimentally proven regulatory regions, binding sites, and regulatory polymorphisms
- Positive and negative control/training datasets
  - Develop motif detection algorithms
  - Automate annotation
- [www.cisRED.org](http://www.cisRED.org) (Robertson et al. 2006)
- rSNP detection methods (Montgomery et al., under review)

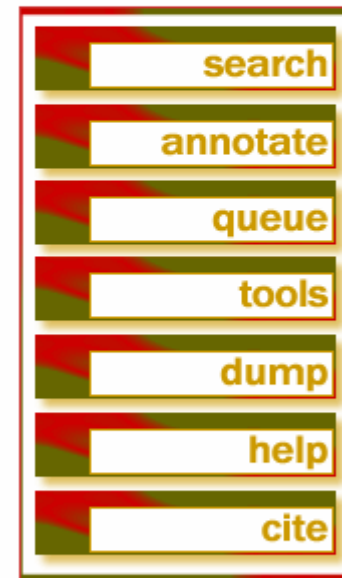
# Overview

- User Management and Quality Control
- Publication queue
- Annotation System
- Data contents
- Data Access and Visualization

menu



user menu



admin menu



# User Management and Quality Control

- User roles: USER, VALIDATOR, ADMIN

- USER

- Add new records
  - Comment on all records

- VALIDATOR

- Score records
  - Deprecate records

- ADMIN

- Add new evidence and meta terms
  - Add new datasets
  - Batch upload data



The image shows a user interface for a validator. It consists of two main sections. The top section is titled 'VALIDATOR MENU' and contains three buttons: '[RECORD VALID (+1)]' in green, '[RECORD INDETERMINATE (0)]' in yellow, and '[RECORD INVALID (-1)]' in red. The bottom section is titled 'SCORE' and contains two text boxes: 'SCORE: 0' and 'ACTIVITY: 0'.

[new user](#)

# A Publication queue allows paper to be added to the system for annotation

## Add to publication queue

Add paper(s) to publication queue (must be logged in)

## Search publication queue

Search publications:  all ▼

(Optional filter) only with state ANY ▼

(Optional filter) for Transcription factor:  ▼

(Optional filter) only include: N/A ▼

(Optional filter) sort scores: N/A ▼

Entered as Expert entry by smontgom on: 2006-11-12

**PMID:10674400** Lahuna O et al., Involvement of STAT5 (signal transducer and activator of transcription 5) and HNF-4 (hepatocyte nuclear factor 4) in the transcriptional control of the hnf6 gene by growth hormone. Mol Endocrinol Feb-2000

### STATE HISTORY

	2006-11-12: PENDING by smontgom Comment: RegCreative November 2006 Jamboree
	2006-11-14: OPEN by idonaldson
	2006-11-15: CLOSED by idonaldson Comment: Success - addition of new records

### STATE CHANGE CONTROLS

SET OPEN/PENDING COMMENT AND STATE

ENTER COMMENT:

SET STATE:

queue



# Annotation System

Choose type of annotation to add:

Enter PubMed  
Reference ID:

16608856

*Must be a paper that has been opened by your user account  
in the Publication Queue*

☒ TRANSCRIPTION FACTOR BINDING SITE

A **transcription factor** record is a noncoding DNA sequence that is bound by a particular transcription factor in vivo to alter the expression of a particular gene. An example might be an experimentally confirmed Sp1 binding site.

☐ REGULATORY REGION

A **regulatory region** is a noncoding DNA sequence that is known to alter the expression of a particular gene. Canonical examples of regulatory regions are promoters and enhancers.

☐ REGULATORY POLYMORPHISM

A **regulatory polymorphism** record is a noncoding DNA sequence that may or may not be bound by a known transcription factor in vivo, but has a variant that is confirmed to alter the expression of a particular gene. An example might be an experimentally confirmed Sp1 binding site that has two allelic variants, one of which, when present, downregulates its target gene (like MDM2, see Bond et al., Cell. 2004). Another example may be an allele in a non-coding region which is confirmed to alter expression patterns through a transcript quantification assay. Types of supported variants are ARTIFICIAL, GERMLINE, and SOMATIC

☐ REGULATORY HAPLOTYPE

A **regulatory haplotype** record is a noncoding DNA sequence that contains many alleles in linkage disequilibrium (LD) that are confirmed to alter the expression of a particular gene. This is different than a regulatory polymorphism as the specific causal variant may not be known, only the alleles that are in LD with it.

Annotate

annotate

# Annotation System (cont'd)

**Add a new record:**

**1. STABLE ID:**

Regenerate Stable ID:  Unique record identifier. EXAMPLE: OREG0000001, OREG1234567

**2. DATASET:**

Enter Dataset:  The dataset to associate this annotation with.

**3. GENE:**

Choose Source: ☐ USER\_DEFINED ☐ ENSEMBL ☒ NCBI

User-defined gene ID:  A user-defined gene or UNKNOWN

**3b. TRANSCRIPTION FACTOR:**

Choose Source: ☒ USER\_DEFINED ☐ ENSEMBL ☐ NCBI

User-defined tf name:  A user-defined transcription factor or UNKNOWN

**4. LOCI NAME:**

Enter Regulatory Region Loc Name (if known):

Some regulatory regions have loci names associated to them (particularly in fruitfly) or other ORegAnno ID's, enter them here.

**5. TARGET SPECIES:**

Enter Taxon ID:  HELP: Find NCBI Taxonomy IDs. EXAMPLE: 9606 (Homo sapiens), 10090 (Mus musculus)

**6. SEQUENCE:**

For help obtaining sequence or sequence with flanks, try the [ORegAnno Fetcher](#) [Sequence Tools](#)

Enter Bound Sequence (gen 5' to 3')

Enter Sequence with Flank (gen 5' to 3')

Enter Sequence (gen 5' to 3')

**7. REFERENCE:**

Enter PubMed ID:  HELP: Search PubMed. EXAMPLE: 11511759 (Mullins et al. 2001)

**8. EVIDENCE:**

For help, click [here](#) for an explanation of available evidence types

Enter Record Evidence Type:

Enter Record Evidence Settings:

Enter Record Evidence Class:

Enter cell type:  ▼ VDC Cell type ontology

Evidence Comment:

**9. EXPERIMENTAL OUTCOME:**

Enter Reported Outcome:  EXAMPLE: Results which confirm regulatory properties of sequence are positive. This field should be translated directly from the reference.

**10. COMMENT (OPTIONAL):**

Enter a comment:

**11. META DATA (OPTIONAL):**

For help, click [here](#) for an explanation of available meta data types

**12. ANNOTATED BY:**

User:

Entry Date:

## Add a new record:

### 1. STABLE ID:

Regenerate Stable ID

Unique record identifier  
EXAMPLE:  
OREG0000001,  
OREG1234567

### 2. DATASET:

Enter Dataset:

The dataset to associate this annotation with.

### 3. GENE:

Choose Source: ☐ USER\_DEFINED ☐ ENSEMBL ☒ NCBI

User-defined gene ID:

A user-defined gene or UNKNOWN

### 3b. TRANSCRIPTION FACTOR:

Choose Source: ☒ USER\_DEFINED ☐ ENSEMBL ☐ NCBI

User-defined tf name:

A user-defined transcription factor or UNKNOWN

### 4. LOCI NAME:

Enter Regulatory Region Loc Name (if known):

Some regulatory regions have loci names associated to them (particularly in fruitfly) or other ORegAnno ID's, enter them here.

### 5. TARGET SPECIES:

Enter Taxon ID:

HELP: Find NCBI Taxonomy IDs  
EXAMPLE:  
9606 (Homo sapiens),  
10090 (Mus musculus)

# Annotation System (cont'd)

**Add a new record:**

1. STABLE ID:  
 Regenerate Stable ID:  (Click here to regenerate a new stable ID)

2. DATASET:  
 Enter Dataset:  The dataset to associate this annotation with

3. GENE:  
 Choose Source: ☒ USER DEFINED ☐ ENSEMBL ☐ NCBI  
 Enter defined gene ID:  A user-defined gene or UNIGENCODE

3b. TRANSCRIPTION FACTOR:  
 Choose Source: ☒ USER DEFINED ☐ ENSEMBL ☐ NCBI  
 Enter defined TF name:  A transcription factor or TRANSFAC

4. LOC NAME:  
 Enter Regulatory Region Loc:  Gene regulatory region from 5' to 3' (transcription start site) or 3' to 5' (polyadenylation site)

5. TARGET SPECIES:  
 Enter Tissue ID:  HELP: Find NCBI Tissue IDs  
 EXAMPLE: 15672001588 (Montgomery et al.)  
 12581799 (Griffith et al.)

6. SEQUENCE:  
 For help obtaining sequence or sequence with flank, try the [ORegAnno fetcher/scanner tools](#)

Enter Bound Sequence (min 5 nt):

Enter Sequence with Flank (min 40 nt):

Enter Sequence Search Space:

7. REFERENCE:  
 Enter PubMed ID:  HELP: Search PubMed  
 EXAMPLE: 15123592 (Montgomery et al.),  
 12581799 (Griffith et al.)

8. EVIDENCE:  
 For help, click here for an explanation of available evidence types

Enter Record Evidence Type:  UNKNOWN

Enter Record Evidence Subtype:  UNKNOWN

Enter Record Evidence Class:  UNKNOWN

Enter cell type:  eVOC: Cell type ontology

Evidence Comment:

9. EXPERIMENTAL OUTCOME:  
 Enter Reported Outcome:  EXAMPLE: Results which confirm regulatory properties of sequence are positive. This field should be translated directly from the reference.

10. COMMENT (OPTIONAL):  
 Enter a comment:

11. META DATA (OPTIONAL):  
 For help, click here for an explanation of available meta data types

12. ANNOTATED BY:  
 User:   
 Entry Date:

## 6. SEQUENCE:

For help obtaining sequence or sequence with flank, try the [ORegAnno fetcher/scanner tools](#)

Enter Bound Sequence (min 5 nt):  Description: Enter the sequence bound by the transcription factor if one is bound

Enter Sequence with Flank (min 40 nt):  Description: Enter the sequence above with flank for mapping purposes.

Enter Sequence Search Space:  Description: Enter the sequence that was assayed. Not necessarily functional.

## 7. REFERENCE:

Enter PubMed ID:  HELP: Search PubMed  
 EXAMPLE: 15123592 (Montgomery et al.),  
 12581799 (Griffith et al.)

## 8. EVIDENCE:

For help, click here for an explanation of available evidence types

Enter Record Evidence Type:  Electrophoretic mobility shift assay (EMSA)

Enter Record Evidence Subtype:  Direct gel shift

Enter Record Evidence Class:  Transcription regulator site

Enter cell type:  eVOC: Cell type ontology

Evidence Comment:  Gel shift analysis performed using nuclear extracts and RARR oligonucleotide probes containing the same mutations within or outside of core

**Add a new record:**

---

**1. STABLE ID:**

Fragmentable Stable ID:  Uniquely identify  
Example:  
CFMGK00000000  
CFMG-K1234567

---

**2. DATASET:**

Enter Dataset:  The dataset to associate this annotation with.

---

**3. GENE:**

Choose Source: ☒ USER DEFINED ☐ EXTERNAL ☐ NCBI  
 Unidentified gene ID:  # associated gene or UNPROCESSED

---

**2b. TRANSFORMATION FACTOR:**

Choose Source: ☒ HELP DEFINED ☐ EXTERNAL ☐ NCBI  
 Unidentified Factor:  # associated transformation factor or UNPROCESSED

---

**4. LOC NAME:**

Enter Regulatory Region Last Name (if known):   
Name required if you have found related sequences.  
If they are particularly relevant, please include them in your other data.

---

**5. TARGET SPECIES:**

Enter Target ID:  HELP: List NCBI Taxonomy IDs.  
Example:  
Homo sapiens (taxid: 9606)  
Yeast (taxid: 4915)

---

**6. SEQUENCE:**

**For help obtaining sequence or sequence with flanks, try the [CFmgkms fetcher script](#) tools**

Enter Bound Sequence from EMBL: <input type="text"/>  Enter Sequence with flanks (EMBL 42-42): <input type="text"/>  Enter Sequence (SwissProt) (name): <input type="text"/>	<b>Description:</b> Enter the sequence bound at the reference position in Bound.  <b>Description:</b> Enter the sequence aligned with flanks for regulatory purposes.  <b>Description:</b> Enter the sequence that was assigned (not necessarily functional).
--	--

---

**7. REFERENCE:**

Enter PubMed ID:  HELP: Search PubMed  
Example:  
18121652 (Regulatory et al., 2006)  
12561759 (Smith et al.)

---

**8. EVIDENCE:**

**For help, click here for an explanation of available evidence types**

Enter Record Evidence Type:  UNKNOWN  
 Enter Record Evidence Tagline:  UNKNOWN  
 Enter Record Evidence Class:  UNKNOWN +IOC: Out type ontology  
 Enter ref type:  UNKNOWN  
 Evidence Comment:

---

**9. EXPERIMENTAL OUTCOME:**

Enter Reported Outcome:  EXAMPLE:  
Results which confirm regulatory properties of sequence can be used in this field should be provided directly from the reference.

---

**10. COMMENT (OPTIONAL):**

Enter a comment:

---

**11. META DATA (OPTIONAL):**

**For help, click here for an explanation of available meta data types**

---

**12. ANNOTATED BY:**

User:   
 Entry Date:

Enter Reported Outcome:

Transient transfections of S2 cells showed that both individually and together, Sp1 and Sp3 were able to trans-activate a wild type CT-box-driven luciferase reporter construct in a dose-dependent manner. Transfection of

For help, [click here for an explanation of available meta data types](#)

User: obig  
Entry Date: 2-Feb-2007

[Review record](#)

# Once annotated the publication is closed in the queue

Entered as Expert entry by smontgom on: 2006-11-12

PMID:9171244

Chen CY et al., Competition between negative acting YY1 versus positive acting serum response factor and tinman homologue Nkx-2.5 regulates cardiac alpha-actin promoter activity. Mol Endocrinol Jun-1997

## STATE HISTORY



2006-11-12: PENDING by smontgom

Comment: RegCreative November 2006 Jamboree



2006-11-21: OPEN by eblanco

## STATE CHANGE CONTROLS

### SET OPEN/PENDING COMMENT AND STATE

ENTER COMMENT:

SET STATE:

PENDING

OPEN

OPEN/ANNOTATE

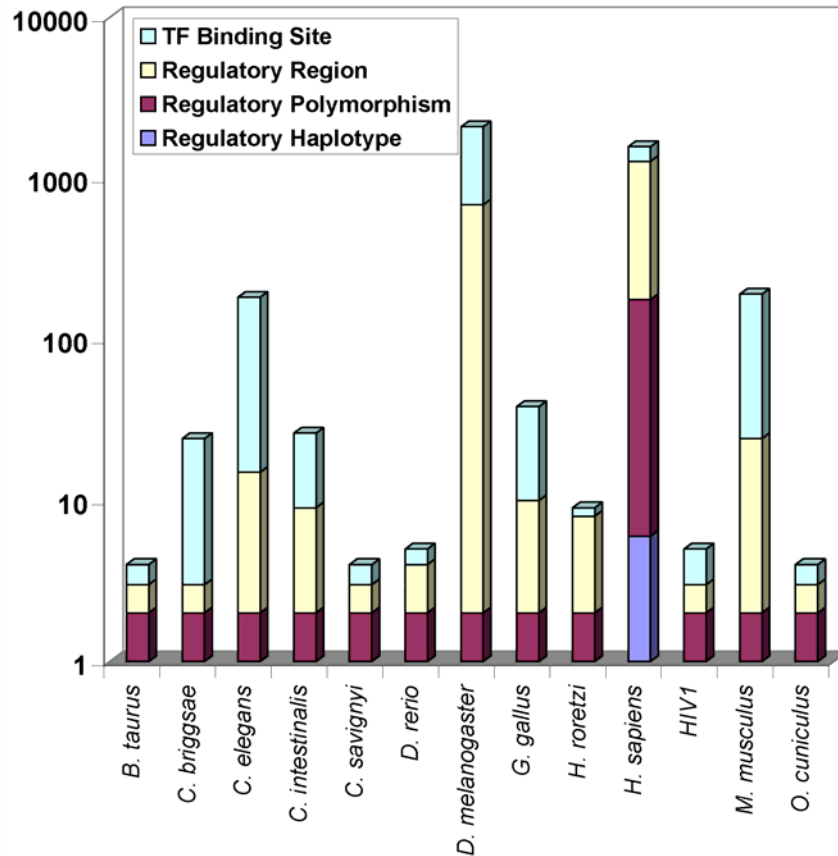
### SET CLOSED COMMENT AND STATE

- ☐ Success - addition of new records
- ☐ Failure - did not describe regulatory element
- ☐ Failure - publication describes regulatory element but there is insufficient information to annotate it
- ☐ Failure - paper describes regulatory element but has been closed without annotation

SET STATE:

CLOSED

# ORegAnno Database Contents



- 4229 publications in queue
- 832 publications curated
- 4234 regulatory sequences
  - 1853 regulatory regions
  - 2204 TFBSs
  - 177 regulatory mutations
- 17 species
- 217 registered users
- 5 datasets and lots of manual curation!

VISTA Enhancer Browser  
whole genome enhancer browser



Drosophila  
DNase I Footprint Database

PENNSTATE  
HBB CRM

Stanford  
ENCODE Promoters

# The RegCreative Jamboree

- Belgium, Nov-Dec, 2006.
- ~50 attendees
- 150+ papers processed
- One dataset (Vista Enhancers)
- 600 sequences
- System improvements
- Annotation standards
- Ontologies
- Text-mining





# Data Access and Visualization

## ORegAnno search

### Search ORegAnno:

Enter search string:  all

SEARCH FILTER: ☐ Exclude deprecated records

#### Example searches

OReg\* Wild card searches using \* (Note: a leading wild card is invalid, like \*OR)  
Notch~4 RBP Boost terms (weight documents, the higher the more relevant) using ^  
9707 AND 10090 Uses AND or OR grouping  
20050901 TO 20051231 Use TO to search a range

[For more info on searching click here](#)

## Sample record

ID: **OReg0000033** [\[VIEW COMMENTS\]](#) [\[VIEW SCORES\]](#) [\[VIEW EVIDENCE\]](#)

Record type: TRANSCRIPTION FACTOR BINDING SITE  
Outcome: POSITIVE OUTCOME  
Gene Source: ENSEMBL  
Gene ID: ENSG00000082196  
Gene name: AMACR  
Gene version: homo\_sapiens\_core\_29\_35b  
TFBS Source: ENSEMBL  
TFBS Gene ID: ENSG00000172216  
TFBS Name: CCAAT enhancer binding protein B  
TFBS Version: homo\_sapiens\_core\_29\_35b  
Species: Homo sapiens  
Species Taxon ID: 9606  
PubMed Reference ID: 15755877  
Entry Date: 2005-08-16  
Sequence: GTGCGCAGAA  
Sequence with Flank: ogggggtgggggaagococaaGTGCGCAGAAactocgggggtggcgaagc

#### User Information

Added by: obig ([Click for user information](#))  
User email: obig@bcgsa.ca

#### Record Details: Record Evidence

Evidence class: Transcription regulator site (OREGEC00001)  
Evidence type: Reporter gene assay (OREGET00002)  
Evidence subtype: Transient transfection luciferase assay (OREGES00004)  
Evidence comment: Transfection and Luciferase Assays with 2 single bp mutations in element and ectopic expression of CEBPB

SCORE

SCORE: 0

ACTIVITY: 0

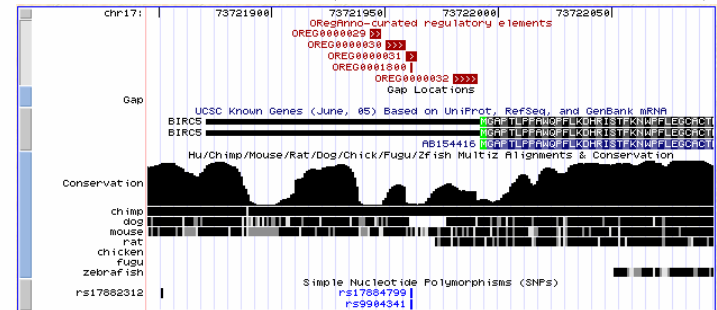
[How are scores defined](#)

GENOME MAPPINGS

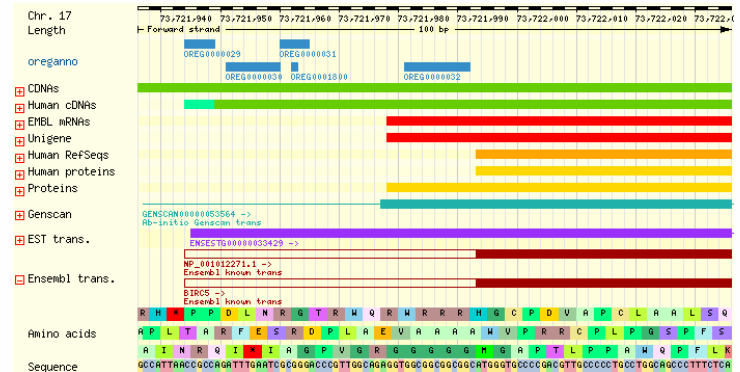
UCSC genome browser BUILD: H9.17

e! BUILD: H9.17

## UCSC tracks (custom + official)



## Ensembl (custom)



- Also via: MySQL, XML file dumps, Web Services (SOAP)



# Conclusions

- A large resource of experimentally proven regulatory sequences
- Open access and open source
- Sign up for an account today
  - Add papers to queue
  - Annotate a paper
- Grab the data!

# Acknowledgements

**Supervisor: Steven Jones**

**Oreganno developers**

- **Stephen Montgomery**
- Casey Bergman
- Monica Sleumer
- Misha Bilenky
- Erin Pleasance

[www.oreganno.org](http://www.oreganno.org)

**Coop students:**

- Yuliya Prychyna
- Maggie Zhang
- Bryan Chu
- Bridget Bernier

**Oreganno curators**

**Recreative participants**

Montgomery SB\*, Griffith OL\*, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJM. 2006.

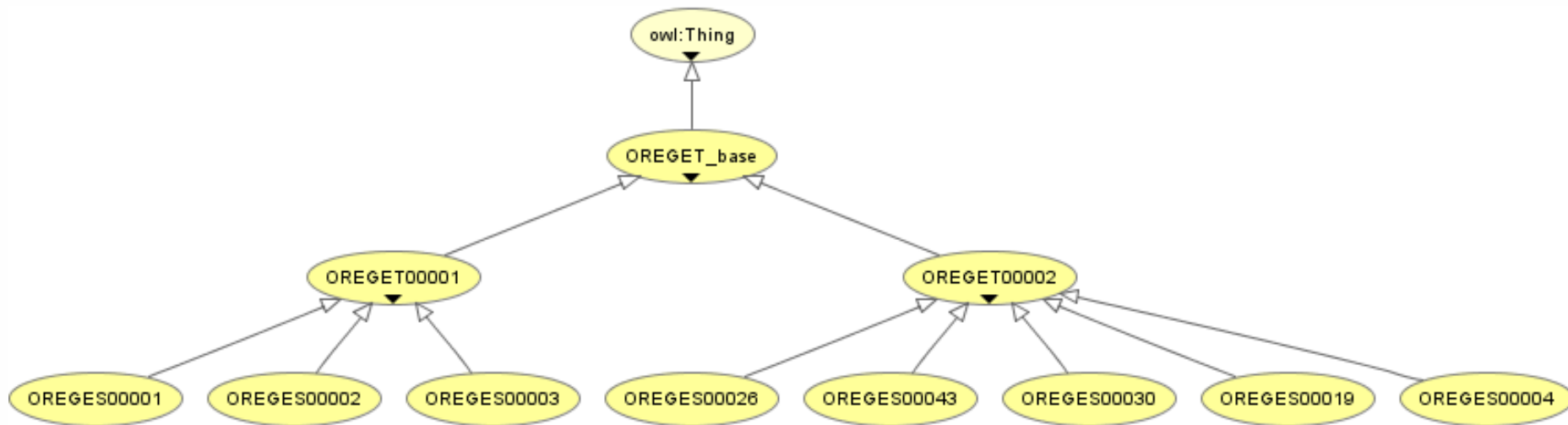
**Bioinformatics. 22(5):637-40.**



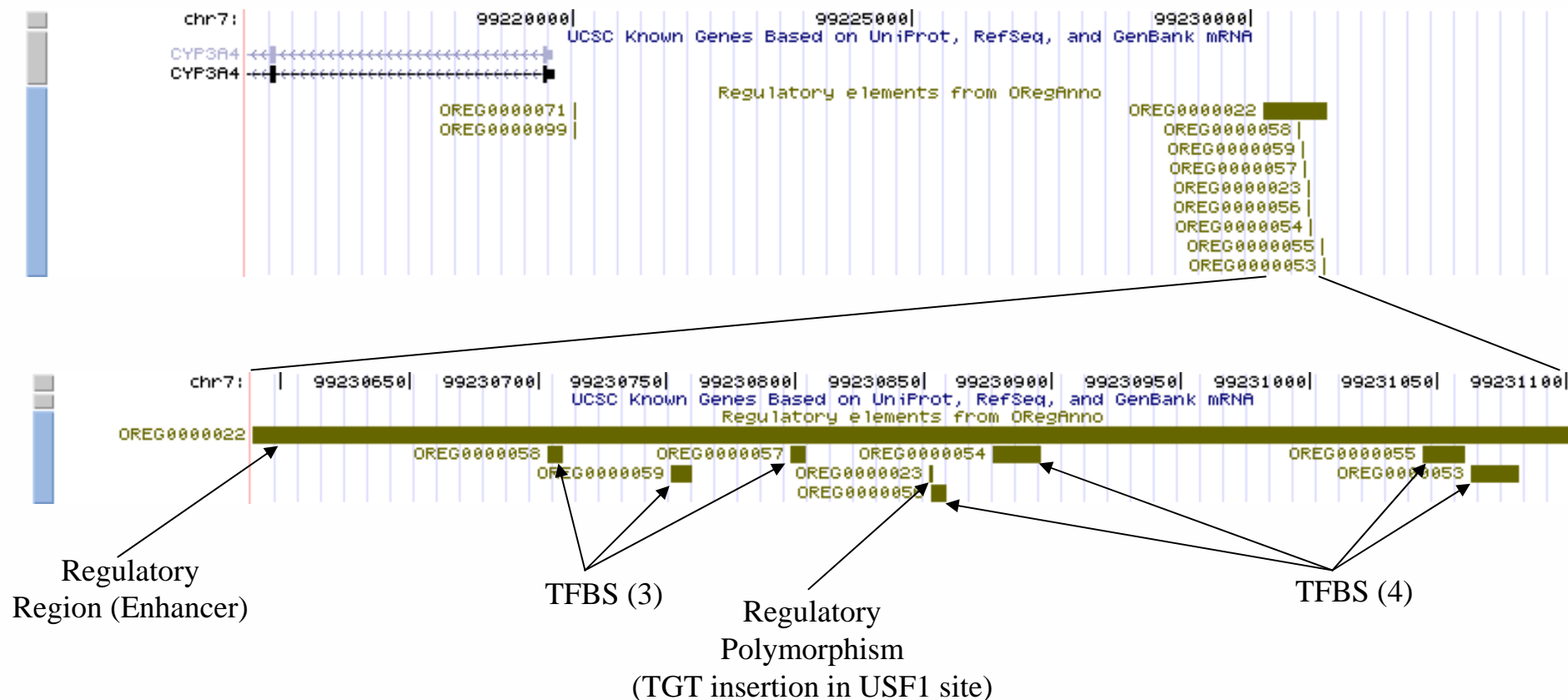
# Experimental evidence

- Evidence Class
  - Regulator (protein) or Regulator Site (sequence)
  - Transcription, Transcript stability, Translation
- Evidence type
  - Evidence subtype
- Evidence comment
- Cell type (eVOC ontology)
- BRENDA for tissues and cell lines soon

Evidence type	Evidence subtype
EMSA (OREGET00001)	Direct gel shift (OREGES00001)
	Supershift (OREGES00002)
Reporter Gene Assay (OREGET00002)	Transient transfection luciferase assay (OREGES00004)
	Chloramphenicol acetyltransferase (CAT) Assay (OREGES00019)



# Example: Regulatory elements for CYP3A4 in ORegAnno



- Matsumura et al. (2004) identified a novel polymorphic enhancer.
- Rodriguez-Antona et al. (2005) and Amirimani et al. (2003) both identified a functional rSNP closer to TSS

# Extra assistance

- Tools
  - Basic tools for fetching sequence data from NCBI and Ensembl.
  - Example: TFBS exists at -543 to -538, sequence CCGCCC, use NCBIFETCH or ENSFETCH
- Help
  - Contains walkthroughs, case studies, and descriptions of various components of ORegAnno.