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

Griffith OL†, Montgomery SB†, 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

† These authors contributed equally to this work

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?

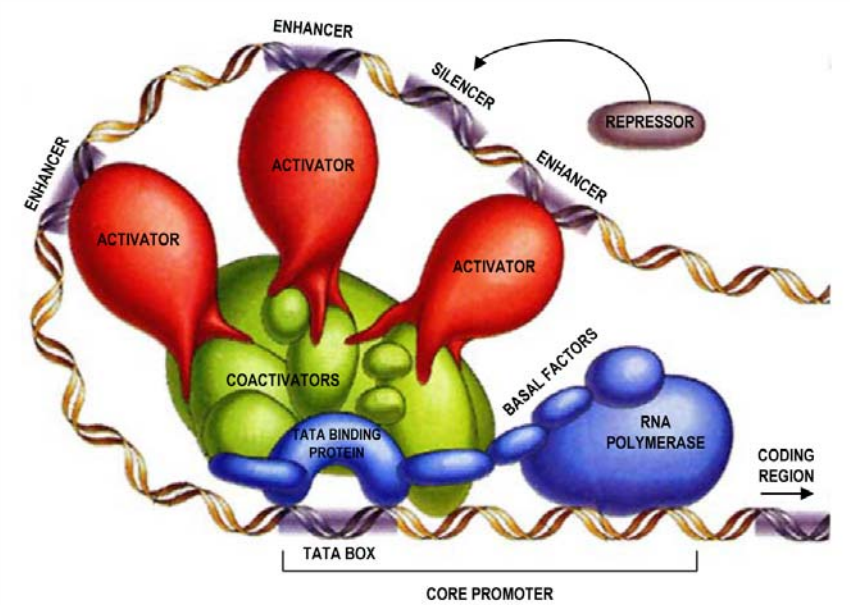
Why should we care?

- 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.

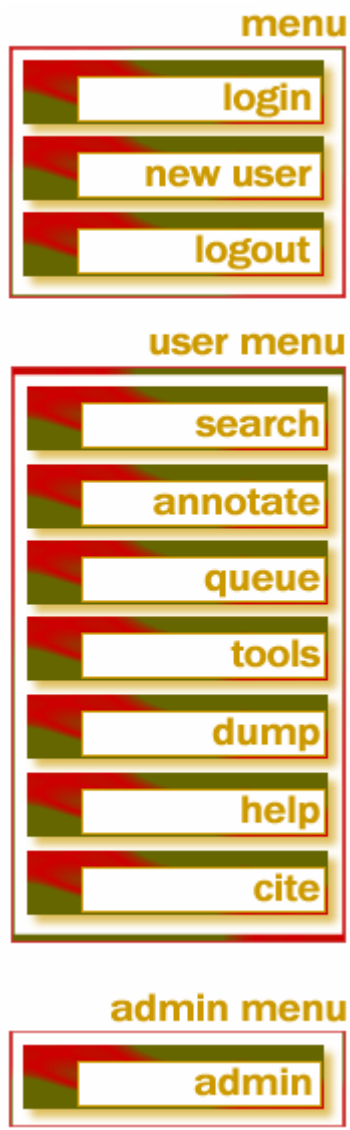
Open Regulatory Annotation Database (www.oreganno.org)

- 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 (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

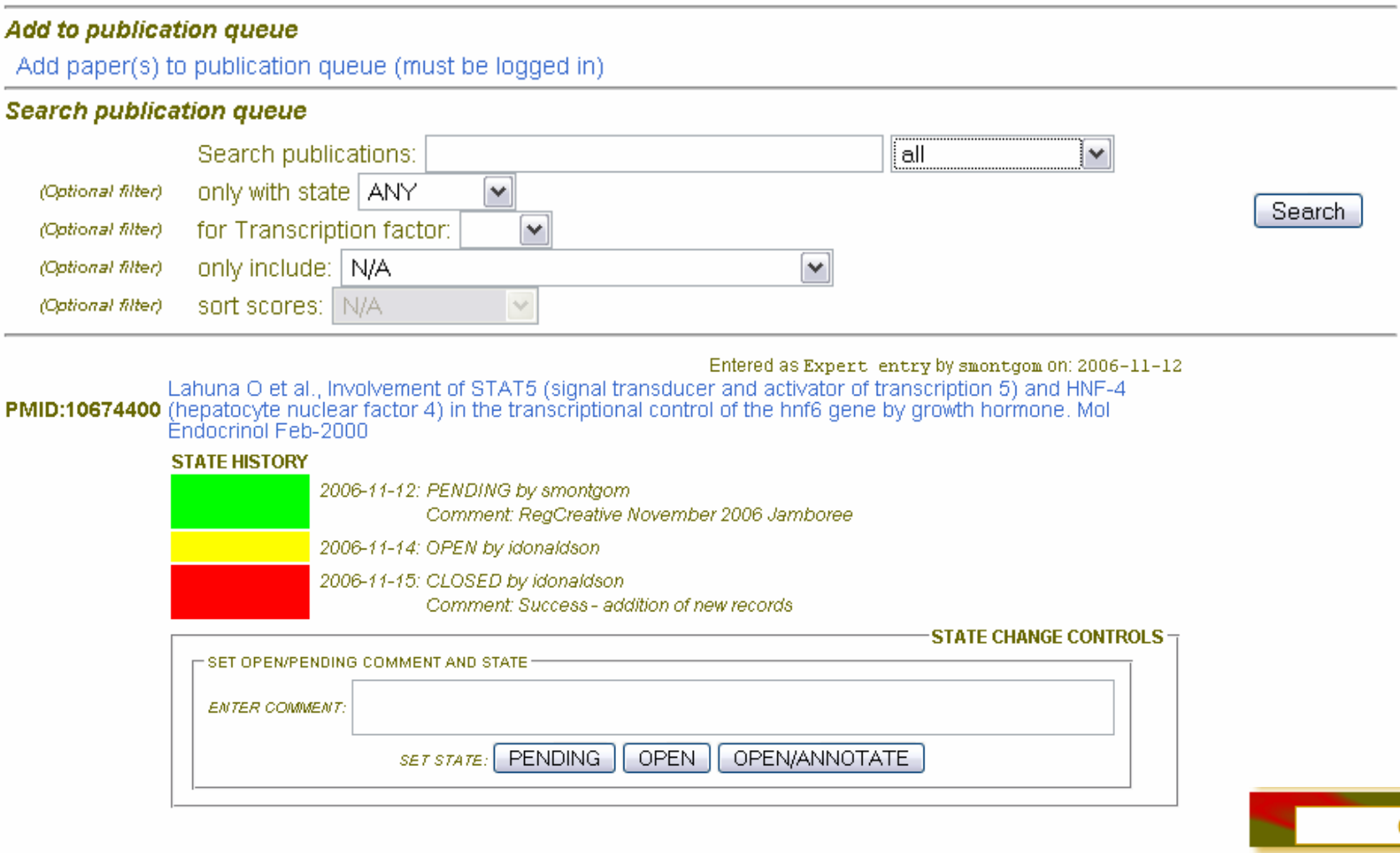


User Management and Quality Control

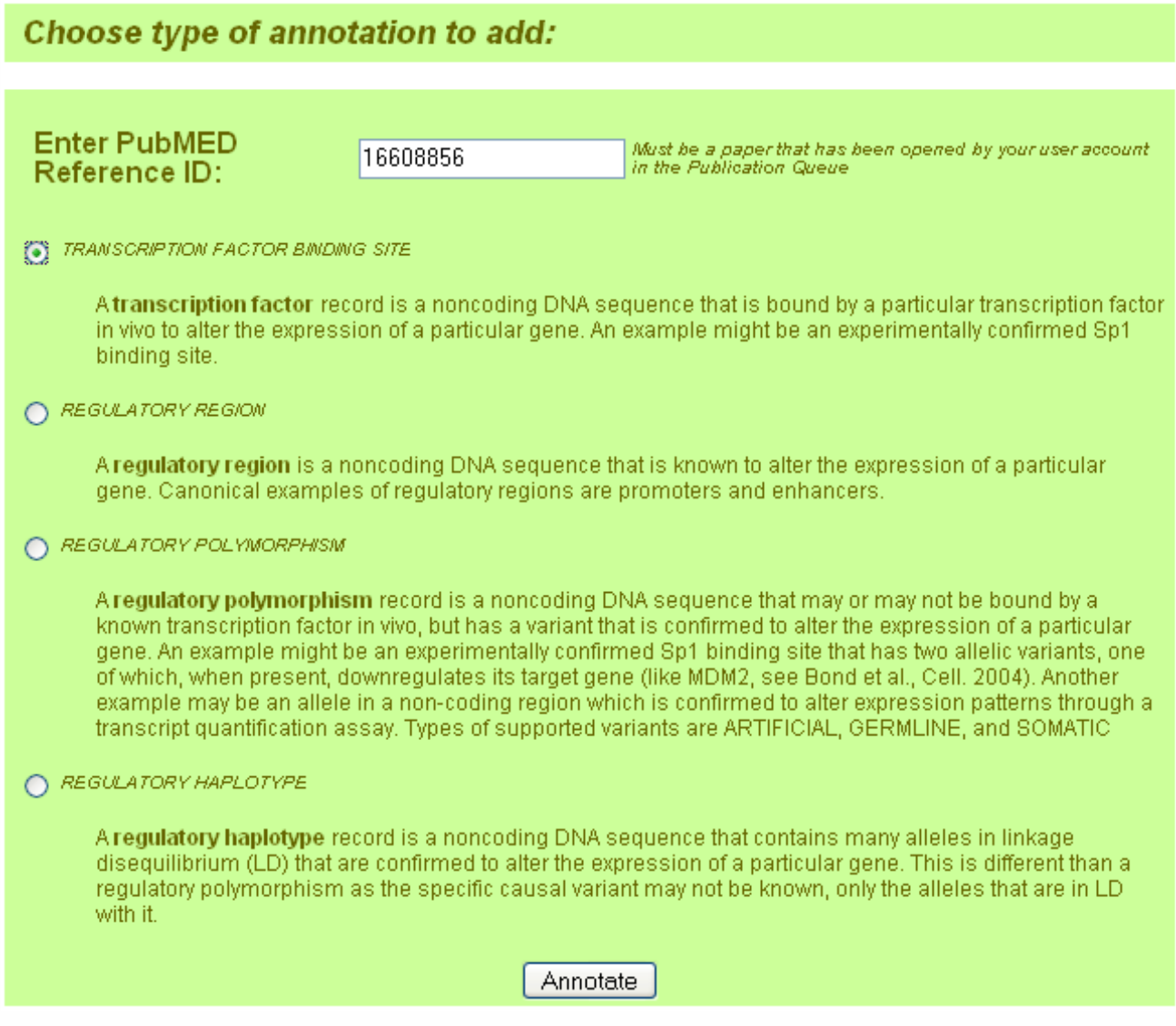
- User roles: USER, VALIDATOR, ADMIN

- 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

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



Annotation System



annotate

GENOME Sciences Centre

BC Cancer Agency
CARE & RESEARCH

Annotation System (cont'd)

Add a new record:

1. STABLE ID:

Regenerate Stable ID | OREG0002889 | *Unique record identifier EXAMPLE: OREG0000001, OREG1234567*

2. DATASET:

Enter Dataset: ORegAnno | *The dataset to associate this annotation with.*

3. GENE:

Choose Source: ○ USER DEFINED ○ ENSEMBL ◉ NCBI

User-defined gene ID: 4613 | *A user-defined gene or UNKNOWN*

3b. TRANSCRIPTION FACTOR:

Choose Source: ◉ USER DEFINED ○ ENSEMBL ○ NCBI

User-defined tf name: SP1/SP3 | *A user-defined transcription factor or UNKNOWN*

4. LOCI NAME:

Enter Regulatory Region Loci 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: 9606 | *HELP: Find NCBI Taxonomy IDs EXAMPLE: 9606 (Homo sapiens), 10090 (Mus musculus)*

Annotation System (cont'd)

6. SEQUENCE:

For help obtaining sequence or sequence with flank, try the ORegAnno fetcher/scanner tools

Enter Bound Sequence (min 5 nt):	CCTCCC	Description: Enter the sequence bound by the transcription factor if one is bound
Enter Sequence with Flank (min 40 nt):	cttctctctgcaaagaaaagcaagtggcttttggcgcgaaagc	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: 16608856 — 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:	hepatocyte (EV:0200061) — 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 Evidence Description | Remove Evidence Description

Annotation System (cont'd)

9. EXPERIMENTAL OUTCOME:

Enter Reported Outcome: POSITIVE 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:

```
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
```

11. META DATA (OPTIONAL):

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

Add Meta Data | Remove Meta Data

12. ANNOTATED BY:

User: obig

Entry Date: 2-Feb-2007

13. REVIEW RECORD (before commit):

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:

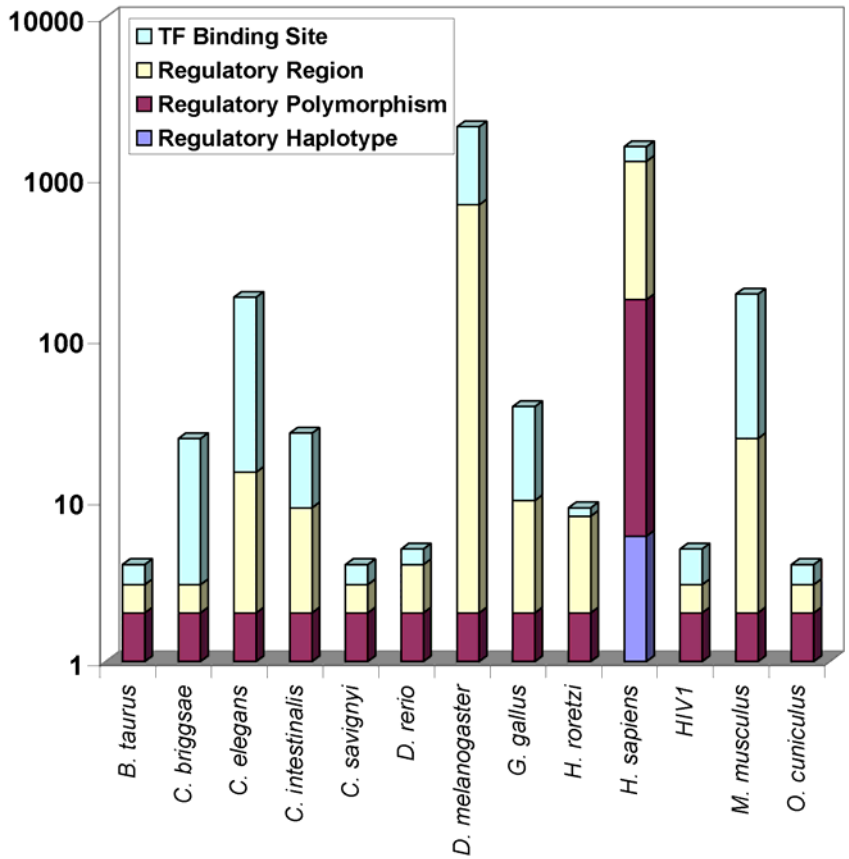
SET STATE: PENDING | OPEN | OPEN/ANNOTATE

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!

REDfly Regulatory Element Database for Drosophila

PENNSTATE HBB CRMs

VISTA Enhancer Browser
whole genome enhancer browser

Drosophila DNase I Footprint Database

Stanford ENCODE Promoters

GENOME Sciences Centre

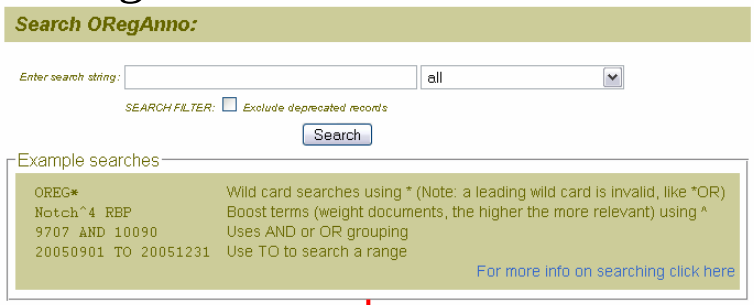
BC Cancer Agency
CARE & RESEARCH

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

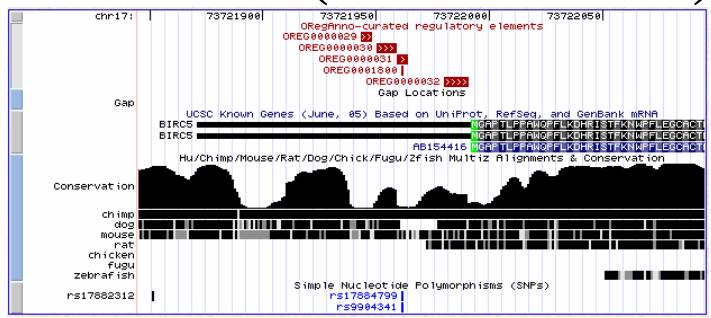
Data Access and Visualization

ORegAnno search

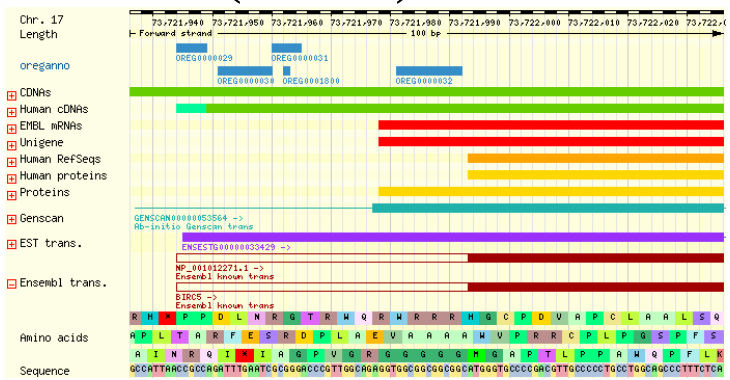


Sample record

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

Coop students:
- Yuliya Prychyna
- Maggie Zhang
- Bryan Chu
- Bridget Bernier

Oreganno curators
Regcreative participants

www.oreganno.org

Montgomery SB*, Griffith OL*, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJM. 2006. **Bioinformatics. 22(5):637-40.**

Michael Smith Foundation for Health Research

CIHR IRSC

NSERC CRSNG

GenomeBritishColumbia

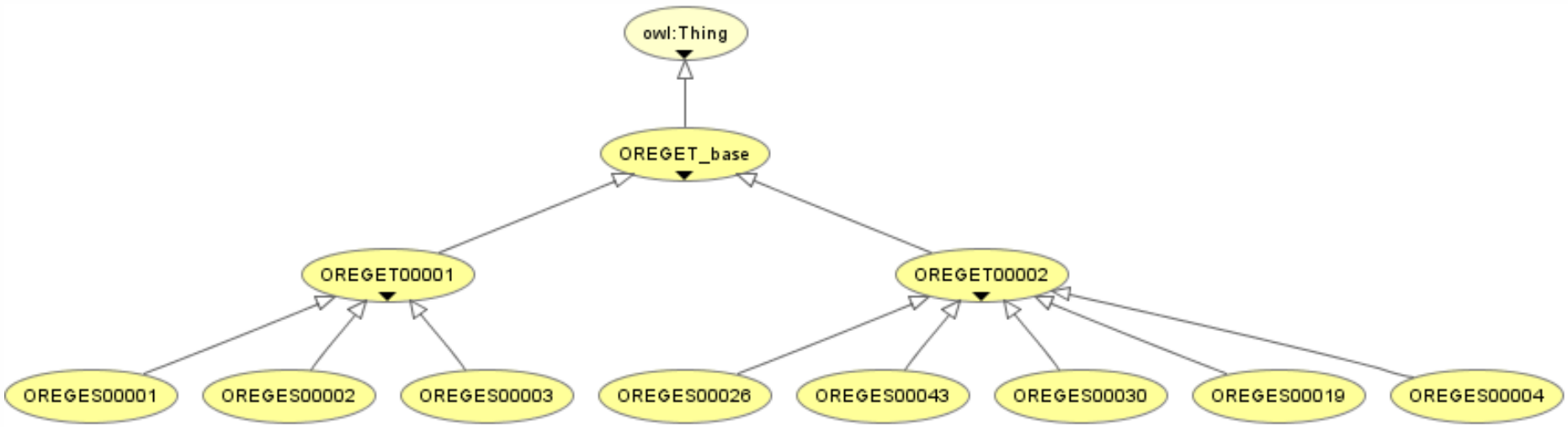
GENOME Sciences Centre

BC Cancer Agency CARE & RESEARCH

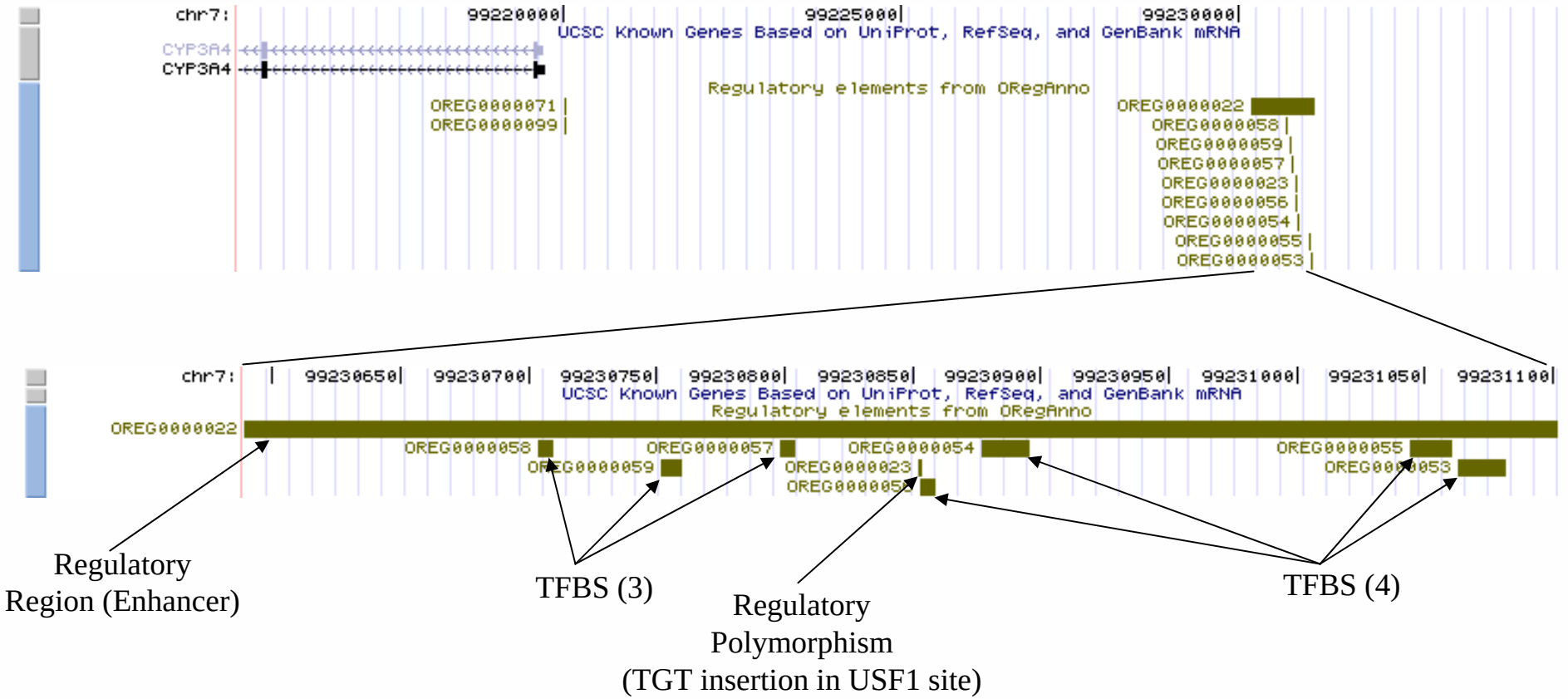
Experimental evidence

- 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.