Annotation standards in ORegAnno (Draft)

Obi Griffith
The RegCreative Jamboree
Nov 29, 2006
Ghent, Belgium

Goal or purpose of annotation standards

1. Positive and negative control datasets
   - Develop motif detection algorithms
2. Training datasets for text-mining tools
   - Automate annotation
3. Resource of known regulatory sequences

Minimal information for an ORegAnno record

- Publication identifier
- Regulatory sequence type
- Species
- Target Gene
- Transcription factor
- Sequence and flanking sequence
- Experimental evidence
- Outcome
- User information

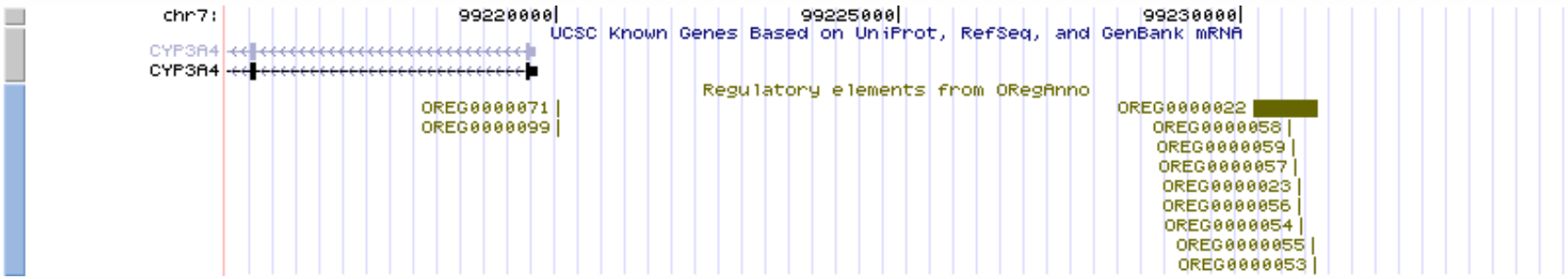
Publication identifier and Species

- Publication identifier
  - PMID
  - Must be entered into queue and checked out prior to annotation and closed after annotation
  - Ensures traceability, prevents redundancy.

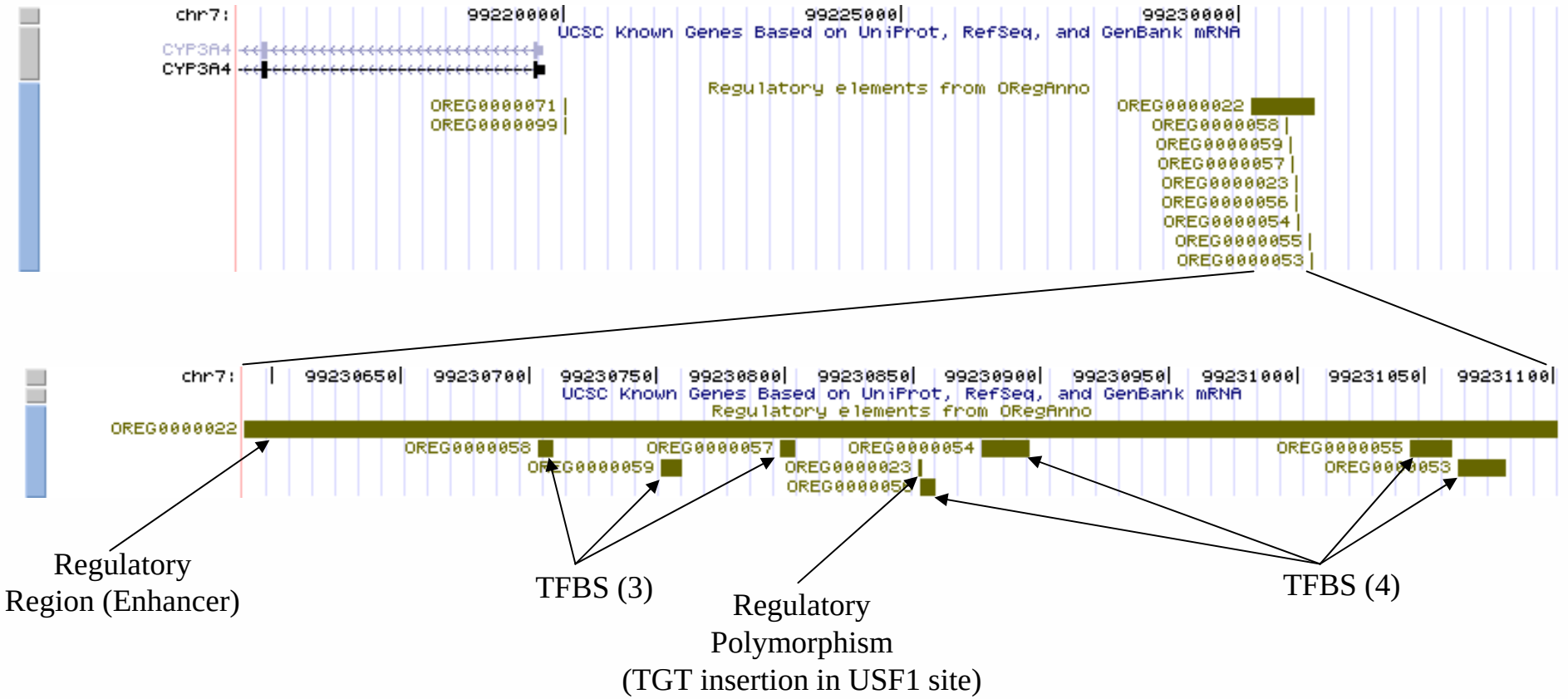
- Species
  - NCBI Taxonomy id

Regulatory sequence type

- Transcription factor binding site (TFBS)
- Regulatory region
- Regulatory polymorphism
- Regulatory haplotype



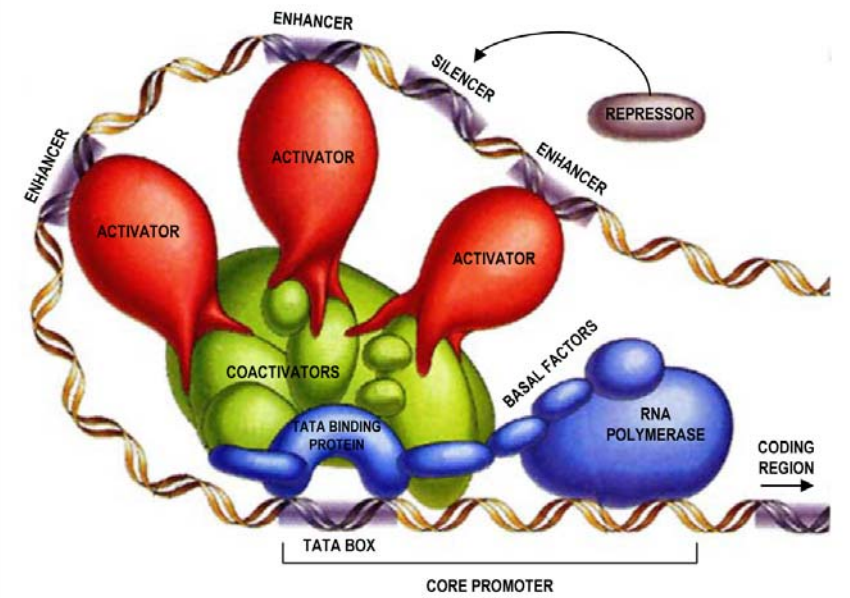
Regulatory elements for CYP3A4 in ORegAnno



Matsumura et al. 2004. Identification of a novel polymorphic enhancer of the human CYP3A4 gene. Mol Pharmacol. 65(2):326-34

Target gene and transcription factor (TF)

- Each record can be linked to one gene and one TF
- Entrez gene id, Ensembl gene id, or user-defined



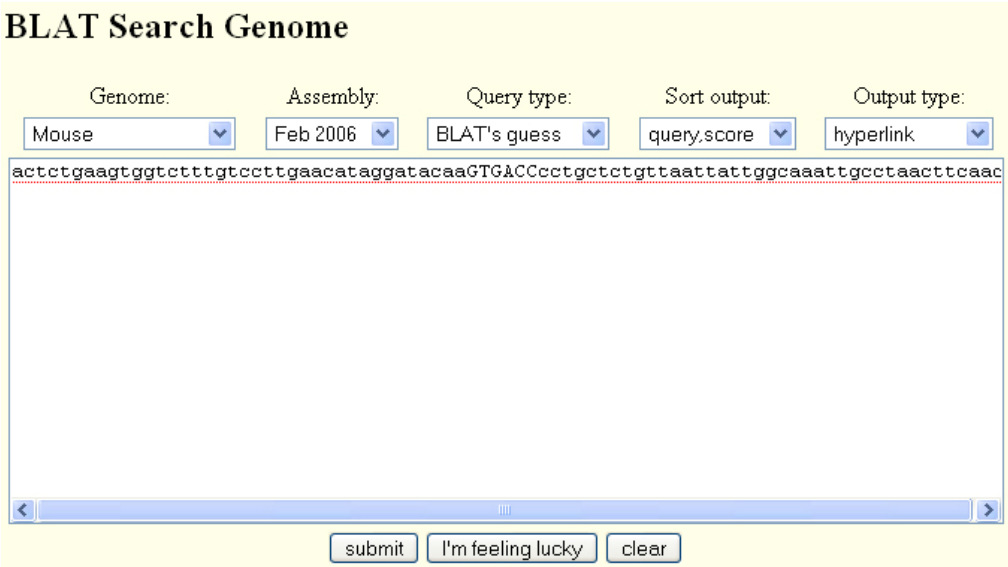
Tjian, R. (1995) “Molecular Machines That Control Genes”; Scientific American, Feb 1995, p. 38.

Sequence and flanking sequence

- Bound sequence in upper case, flanking sequence in lower case
- Minimum 40bp total flanking sequence (recommended: ~100bp)
- Use flanking sequence from current genome, not paper
- Verify final sequence for unambiguous mapping

GTGACC

actctgaagtggtctttgtccttgaacataggatacaaGTGACCcctgctctgttaattattggcaaattgcctaacttcaac



Mouse BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	YourSeq	83	1	83	83	100.0%	5	+	91565757	91565839	83
browser details	YourSeq	20	29	48	83	100.0%	9	+	60468928	60468947	20

Side by Side Alignment

```
00000001 actctgaagtggtctttgtccttgaacataggatacaagtgacccctgct 00000050
>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>
91565757 actctgaagtggtctttgtccttgaacataggatacaagtgacccctgct 91565806

00000051 ctgttaattattggcaaattgcctaacttcaac 00000083
>>>>>>>> |||||||||||||||||||||||||||||||| >>>>>>>>
91565807 ctgttaattattggcaaattgcctaacttcaac 91565839
```

- Unambiguous hit target genome
- Perfect alignment
- Expected position relative to gene

Experimental evidence

- Evidence Class
  - Regulator (protein) or Regulator Site (sequence)
  - Transcription, Transcript stability, Translation
- Evidence type and subtype
  - Type: Reporter Gene Assay
    - Subtype: Transient transfection luciferase assay
- Cell type
  - eVOC cell type ontology
- Evidence comment

Experimental evidence (cont'd)

- Multiple evidence lines
- Minimum: one line of evidence
- *In silico* alone not sufficient
- For regulatory polymorphisms: Association study alone not sufficient
- Avoid use of "literature-derived" evidence type

Record Details: Record Evidence

Evidence class:	**Transcription regulator site** (OREGEC00001)	
Evidence type:	**Protein Binding Assay** (OREGET00003)	
Evidence subtype:	**DNase Footprinting Assay** (OREGES00015)	
Cell type:	**hepatocyte (EV:0200061)**	*eVOC: Cell type ontology*
Evidence comment:	From Results: "DNAse I protection assays were used to identify protected DNA sequences and the proteins that bind to them (Fig 1). The sequences of the eight different protected regions and the locations of the mutations are shown in Fig. 1. The mutated sequences were carefully chosen (a) to change only the sequence similar to the consensus sequence of that site and (b) not to interfere with the overlapping binding of other factors. The effect of these mutations on the interactions between protein-binding factors and the AFP promoter was checked by DNase I protection analysis using the mutated AFP promoter as a probe. The footprinting patterns indicated that DNase I protection was abolished by the site-specific mutations in all regions except II and II' (Fig. 5). In this case, the mutation in region II was made in one of the two sequences of the NF-1 dyad so that partial binding can apparently still occur in that region, i.e. from -128 to -113 bp; mutation of region II' resulted in an altered protection that extended from nucleotides -115 to -98 bp." The sequence in this ORegAnno record is for the site referred to as III.	

Evidence class:	**Transcription regulator site** (OREGEC00001)	
Evidence type:	**Reporter gene assay** (OREGET00002)	
Evidence subtype:	**Chloramphenicol acetyltransferase (CAT) Assay** (OREGES00019)	
Cell type:	**hepatocyte (EV:0200061)**	*eVOC: Cell type ontology*
Evidence comment:	From Results: "As shown in Fig. 6, the mutations in regions Ia, Ib, II, II', and IV decrease the promoter activity significantly. The mutation in region III increases the AFP promoter activity and the mutations in regions Va and Vb had a slight inhibitory effect on the AFP promoter activity." The sequence in this ORegAnno record is for the site referred to as III.	

Outcome and User info

- Outcome
  - Positive, neutral or negative
  - Was the sequence proven functional? Yes, no, or uncertain
  - Sequences can only be considered negative or positive for the conditions under which they were tested
- User information
  - Encourages ownership and accountability
  - Associated with every record, comment, and validation a user creates
  - Three roles: ‘User’, ‘Validator’, ‘Administrator’

Optional information for an ORegAnno record

- Dataset id
- Locus name
- Sequence search space
- For regulatory polymorphism records:
  - Variant sequence and identifier
  - Type of variant
- Meta-data
- Comment

Discussion items

- *Discussion item: Should a record reference only one publication? What should be done in cases where several papers describe experimental validation of the same regulatory sequence?*
- *Discussion item: Should further sub-categorization of regulatory regions be allowed (e.g. Silencer, enhancer, locus-control region, etc)*
- *Discussion item: In a case where sequence conservation is perfect between the species of interest and model organism (e.g. both mouse and human have identical regulatory sequence upstream of an orthologous gene) could an assay in one be considered evidence for function of the sequence in both species?*
- *Discussion item: Should multiple TFs be allowed for a single record? Or should this form a second record?*
- *Discussion item: Should TF complexes be allowed?*
- *Discussion item: What is the minimal evidence we should allow?*
- *Discussion item: Are there other evidence classes that should be included?*
- *Discussion item: Should ORegAnno migrate to a more complex or formal ontology system for evidence.*

Acknowledgements

Supervisor: Steven Jones

Oreganno developers and co-authors:
- Stephen Montgomery
- Monica Sleumer
- Casey Bergman
- Misha Bilenky
- Erin Pleasance

Regcreative organizers and participants

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