

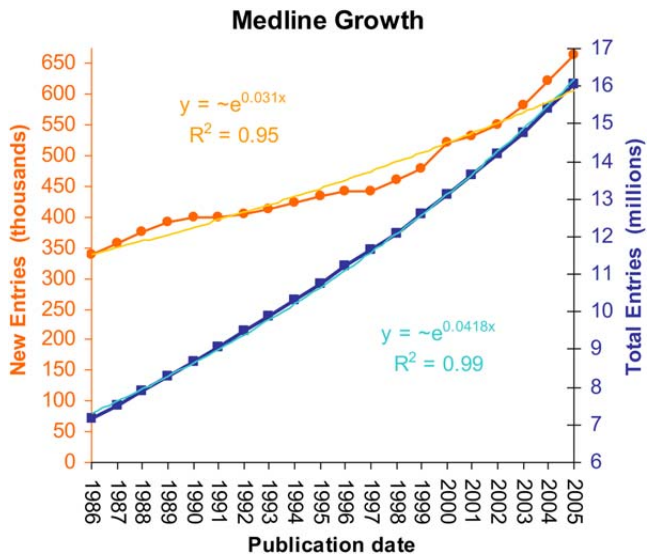
# Making the World's Scientific Information (More) Organized, Accessible, and Usable

Ted Briscoe


Natural Language and Information Processing Group  
Computer Laboratory  
University of Cambridge

Berkeley Version

# Exponential Growth of Papers



# Google Scholar




[Advanced Scholar Search](#)  
[Scholar Preferences](#)

---

**Scholar**



Results 1 - 10 of about 23,800. (0.17 sec)

## [wingless expression mediates determination of peripheral nervous system ...](#)

RG Phillips, JR Whittle - *Development*, 1993 - [dev.biologists.org](#)

The appearance of spatial patterns of cell differentiation in the epithelia of imaginal discs in *Drosophila* depends upon signalling mechanisms between adjacent cells (Haynie and Bryant, 1976; Wilcox and Smith, 1977; Mohler, 1988; Phillips et al., 1990). In particular, the specification and ...

[Cited by 152](#) - [Related articles](#) - [BL Direct](#) - [All 3 versions](#)

## [The consequences of ubiquitous expression of the wingless gene in the ...](#)

J Noordermeer, P Johnston, F Rijsewijk, R Nusse, ... - ..., 1992 - [dev.biologists.org](#)

Five hours after fertilization, the *Drosophila* embryo exhibits the first morphological signs of repeated units that will form the segments of the larva and adult fruit fly (reviewed by Lawrence, 1992).

A gene essential for correct formation of this pattern is **wingless**; in its absence, the pre- ...

[Cited by 119](#) - [Related articles](#) - [BL Direct](#) - [All 7 versions](#)

## [... wingless expression and is not required for reception of the paracrine wingless ...](#)

EJ Rulifson, SS Blair - *Development*, 1995 - [dev.biologists.org](#)

In the imaginal wing disc of *Drosophila*, sensory mother cells (SMCs), the precursors of the sensory organs, differentiate in a highly stereotyped pattern (Ghvsen and O'Kane, 1989; Huang et al., ...

## FlyBase Proforma / Information Extraction

The screenshot shows the FlyBase Proforma software interface. The window title is "File Edit Perspective". The menu bar includes "quit", "new proforma", "update proforma label", "search / replace", "tree perspective", and "plain perspective".

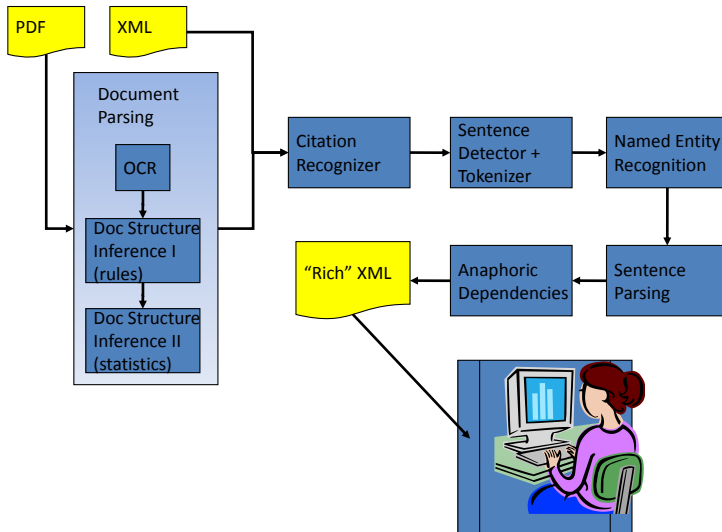
The left pane shows a tree view of the document structure:

- #document
  - P publication:
    - ::Version 17: 18 May 2004:
      - G gene:
        - ::Version 37: 5 Aug 200
          - A allele:
            - ::Version 32: 5 Aug 2
              - G gene:
                - ::Version 37: 5 Aug 200

The right pane displays a list of gene symbols and their associated information:

- :Version 37: 5 Aug 2005::
- : G1a. Gene symbol to use in database \*a ::
- : G1b. Gene symbol used in paper (if different) \*l ::
- : G1c. Database gene symbol to replace \*i ::
- : G1d. Gene category (if gene is new to FlyBase) [CV] \*t ::
- : G2a. Gene name to use in database \*e ::
- : G2b. Gene name used in paper (if different) \*v ::
- : G2c. Database gene name to replace \*v ::
- : G4a. Is the gene in FlyBase already (with any name)? NSC ::
- y
- : G4b. Other synonym(s) for gene symbol \*l ::

# The Paper Annotation Pipeline



# Evaluation Measures

Precision: 
$$\frac{TruePositives}{TruePositives + FalsePositives}$$

Recall: 
$$\frac{TruePositives}{TruePositives + FalseNegatives}$$

F-measure: 
$$\frac{Precision \times Recall \times 2}{Precision + Recall}$$

Mean Av. Prec.: 
$$\frac{\sum_{r=1}^N (Prec(r) \times TP?(r))}{TruePositives + FalseNegatives}$$

N = no. of TPs and FPs, r = rank

## PDF to (Sci)XML



Available online at www.sciencedirect.com



Developmental Biology 267 (2004) 355–368

DEVELOPMENTAL  
BIOLOGY

www.elsevier.com/locate/ydbio

## *Drosophila* Tbx6-related gene, *Dorsocross*, mediates high levels of Dpp and Scw signal required for the development of amnioserosa and wing disc primordium

Takashi Hamaguchi,<sup>a</sup> Shigeharu Yabe,<sup>b</sup> Hideho Uchiyama,<sup>b</sup> and Ryutarō Murokami<sup>a,\*</sup><sup>a</sup>Department of Physics, Biology, and Informatics, Kanagawa University, Atsugi-shi, 753-0322, Japan<sup>b</sup>Graduate School of Biological Science, Kanagawa University, 1030 Shimo-Ogino, Atsugi-shi, Kanagawa 243-0292, Japan

Received for publication 19 June 2003; revised 29 September 2003; accepted 2 October 2003

### Abstract

Regional differentiation along the dorsoventral (DV) axis of the *Drosophila* embryo primarily depends on a graded BMP signaling activity generated by Drosopodmorphogen (Dpp) and Scw (Dscw). We have identified regulated Dpp and Scw target genes (*Dorsocross*, *D* and *Dscw*, *D*), that have a conserved Thbx6 domain related to the vertebrate Thbx6 and act independently to induce dorsal structures. *D* and *Dscw* genes are expressed in the dorsal region in the early blastoderm. After gastrulation, newly expressed *D* appears in a segmental pattern in the ectoderm. This expression correlates spatially with the second phase of Dpp expression in the ectoderm. *D* expression in the early blastoderm is abolished in other dpp or scw mutant embryos, whereas the abdominal segmented expression is only on Dpp. Inactivation of *D* genes with RNAi dramatically affected the development of amnioserosa and wing disc primordia, especially focused on high levels of BMP signaling, although by *D* gene overexpression, which depends on low levels of BMP, neuronal arrest. *D* is also expressed in *Resonance* analysis induced central neurons, suppressed activin-induced events and induced *Flax* genes, which are analyzed by the effects of active Thbx6 and its upstream regulator, BMP-4. These results suggest that the Thbx6 subfamily act in the BMP signaling pathway required for embryonic patterning in both animals.

© 2004 Elsevier Inc. All rights reserved.

**Keywords:** Thbx6; *Dorsocross*; BMP; Dpp; Scw; pMad; Wing; the primordium; Amnioserosa; *Drosophila*; *Resonance*

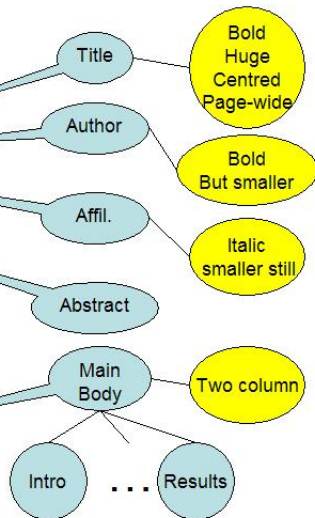
### Introduction

Thbx6 genes, which encode transcription factors characterized with a DNA-binding motif, are highly conserved across the two major animal groups, protozoans and deuterostomes, and were first identified in *Hydra* (Gallup, 1994; Hermans et al., 1990; Kasper et al., 1994; Pfleger et al., 1992b; Tachino and Biale, 1999; Yano and Satoh, 1999). Tbx6 genes have been further classified into several subfamilies, as outlined below (Papanicolaou and Wilson, 1998; Smith, 1999; Wuttler et al., 1998). The *Drosophila* gene, a member of the T subfamily that has been identified in mouse and other deuterostomes, is a founding member of this gene family and plays an essential role in the development of the axial mesoderm (Hermans et al., 1990). An ortholog of the *Drosophila* gene, *Brachymerion* or *hox* (also

known as *Pig* and *aprotaxin*), was identified in *Drosophila* and found to play an essential role in specifying the ectodermal hindgut (Kasper et al., 1994; Murokami et al., 1995; Singer et al., 1996). Three other subfamilies, Thx1, Thx2 and Thx3, are also conserved in both vertebrates and *Drosophila* (Belling et al., 1994; Brock and Cohen, 1996; Griffin et al., 2000; Pfleger et al., 1992b; Pusch et al., 1995). While the members of the Thbx6 subfamily play important roles in specifying various mesodermal and endodermal tissues in vertebrates (Cheng and Papanicolaou, 1998; Smith and Papanicolaou, 1997; Kurokawa and Griffin, 1994; Liang et al., 1996; Mitter et al., 1999; Norwood et al., 1996; Uchiyama et al., 2001; Zhang and King, 1995), until now, no corresponding genes had been identified in *Drosophila*. We have identified and analyzed *Dorsocross* (*D*), a newly named *D* subfamily gene No. AB0335412), a *Drosophila* Thbx6 gene that is related to the vertebrate Thbx6 subfamily *Dscw*1 and its homologues, *Dscw2* and *Dscw3*, have also been described in

\* Corresponding author. Fax: +81-43-033-5086.

E-mail address: ymuro@sci.kanagawa-u.ac.jp (R. Murokami).



# Citation Recognition

For each paper:

- 1 Find candidate names in references section: Ashburner
- 2 Find citation dates: 19|20xx(a|b)
- 3 Mark-up occurrences of name candidates leftwards from dates: Ashburner *et al.* (1985), (see Ashburner, 1983)

97% F-meas.?



# Sentence Detection / Tokenization

- 1 Resolve abbreviatory / sentential **periods**:  
... et al. Adh vs. ... Adh. However
- 2 Separate **punctuation** / remove some hyphenation:  
Adh ., insulin-like, phosphoryl-ation
- 3 Normalize Greek **super/sub-scripts**, footnote indices, etc:  
Adh<sup>α</sup>, Adh.<sup>†</sup>

95% F-meas.

# Named Entity Recognition

- FlyBase: 18k Genes, 75k Gene Names
  - Overlap with general English: But, Can, Mad, spliced
  - Spelling variation: Fas-III, fas III
- 1 Annotate gene names in abstracts automatically using FlyBase
  - 2 Train a **Conditional Random Field** sequential classifier
  - 3 Label tokens as (part of) gene names

85% F-meas. (abstracts) 83% F-meas. (full papers)

# Sentence Parsing

- 1 Assign **Part-of-Speech** (PoS) Labels to tokens using Hidden Markov Model: `we name/VV0 ...`
- 2 Build graph of **Grammatical Relations** (GRs) between words using probabilistic LR model: `subject(name, we)`
- 3 Models trained on general English – 20% unseen words
- 4 Correct PoS labels for gene names to proper noun

75% F-meas. overall, 80% Recall for top 10 analyses

# Anaphora Resolution

- 1 Assign more **semantic classes** to biological entities:  
DNA, promoter, ... using the **Sequence Ontology**
- 2 Link **coreferential** definite descriptions / pronouns to antecedents: IL-2 promoter... This protein / It...
- 3 Link **associative** definite descriptions to antecedents:  
IL-2 is overexpressed... The promoter...
- 4 Weighted Rule-based classifier using GR-context and semantic classes

58% F-meas. (69% with correct GRs)

## PaperBrowser – Gene Mentions

JREx - FlyBase Viewer -@palermo

file:///tmp/15728670.Btml19779html#id2749817

## Materials and methods

### Fly strains

*h<sup>wd</sup>* is a weak hypomorphic allele and *h<sup>wd</sup>* is a null allele ( Wharton et al. , 1993 ) *dpp<sup>h<sup>wd</sup></sup>* was balanced over *SM6 eve-lacZ*, while *dpp<sup>H46</sup>* was balanced over *CyO23, P[dpp<sup>+</sup>]*, a chromosome that contains two copies of *dpp* ( Wharton et al. , 1993 ). **4X *dpp*** embryos are of the genotype : *CyO23, P[dpp<sup>+</sup>]/CyO23, P[dpp<sup>+</sup>]* and were derived from the *dpp<sup>H46</sup>/CyO23, P[dpp<sup>+</sup>]* stock. *so<sup>g</sup><sup>ST06</sup>* is a strong hypomorphic *so<sup>g</sup>* allele ( Ferguson and Anderson, 1992b ) balanced over *FM7c, ftz-lacZ*. *zen<sup>w36</sup>* is a null allele ( Rushlow et al. , 1987a ) balanced over *TM3, ftz-lacZ* or *TM3, hb-lacZ*. **4X *dpp*; zen<sup>-</sup>** embryos were derived from the stock *CyO23, P[dpp<sup>+</sup>]/+; zen<sup>w36</sup>/TM3, hb-lacZ* ( 1/16 of the embryos ). The double heterozygous embryos, *dpp<sup>h<sup>wd</sup></sup>/+; zen<sup>w36</sup>/+*, were identified by

Interaction -@palermo

File Menu Edit Menu

Show Images Mark as Gene Mark as NOT Gene

PaperView EntitiesView Tokens to verify Help on verify task

- paper outline
- Abstract
- Introduction
- Materials and methods
- Fly strains
  - dpp
  - dpp
  - dpp
  - dpp
  - 4X
  - dpp
  - CyO23

3) via  
el were  
anti-Zen  
tain  
on  
ground  
PCR  
al

Done.

## PaperBrowser – Associated Entities

The screenshot displays the JReX FlyBase Viewer interface. The main window shows a paper abstract titled "Ectopic expression of zen and zen-Del". The abstract text describes the cloning of the zen cDNA and the creation of zen-Del mutants, followed by transgenic fly generation and expression analysis. A secondary window titled "Interaction" is open, showing a list of entities related to the paper, including "zen", "The zen cDNA", "the UAS-zen", "a zen w36", "the zen mutant rescue", and "the zen pattern". The "zen" entity is highlighted in the list.

**Ectopic expression of zen and zen-Del**

The *zen* cDNA (+10 to +1234 from the transcription starting site) (Rushlow et al., 1987a) was cloned into pUAST (Brand and Perrimon, 1993) via *Eco* RI and *Xba*I sites on the 5' and 3' ends, respectively. The *zen-Del* cDNA was made by PCR mutagenesis (Expand High Fidelity PCR system, Roche Applied Science) using oligos spanning the deletion region (amino acids 152-198) and cloned into pUAST via the *Eco* RI and *Xba*I sites. Transgenic flies were generated by the standard transformation protocol (Spradling and Rubin, 1982). Flies carrying UAS-*zen* and UAS-*zen-Del* were crossed to stripe-2 *eve-Gal4* drivers (gift from S. Small) and the expression of ectopic *zen* or *zen-Del* proteins was confirmed by staining with anti-Zen antibodies. Guinea pig or rabbit anti-Zen antibodies were generated (Covance) as described by Rushlow et al. (Rushlow et al., 1987b). To obtain uniform early embryonic expression of the UAS-*zen* and UAS-*zen-Del* transgenes, a maternal *Gal4* driver was used in which the *GALA-VP16* fusion protein is expressed maternally under the control of the  $\alpha$ -*tubulin 67C* promoter. These were further crossed into a *zen<sup>w36</sup>/TM3, hb-lacZ* background for the *zen* mutant rescue experiments.

**In vitro mutagenesis and transgenic analysis**

The *Race* 533 bp enhancer DNA was kindly provided by P. ten Dijke (P. ten Dijke et al., 1992). An internal deletion of 66 bp was made by PCR mutagenesis. Two proximal Zen-binding sites were mutated as follows: TAGAAAATAACTGCA. Constructs were transformed into flies using a standard transformation vector that contains the *UAS* promoter (Brand and Perrimon, 1993). At least three transgenic lines were generated for each construct.

**In situ hybridization and antibody staining**

Wild-type, mutant and transgenic embryos were stained with anti-Zen antibodies (P. ten Dijke et al., 1992), dehydrated and mounted on slides. The slides were stained with anti-Zen antibodies (P. ten Dijke et al., 1992) and developed using the Vectastain ABC kit (Pierce and Warriner, 1991). The slides were stained with anti-Zen antibodies (P. ten Dijke et al., 1992) and developed using the Vectastain ABC kit (Pierce and Warriner, 1991).

**Smad/Zen-mediated activation of *Race* 1**

The *Race* 1 promoter was cloned into a pUAST vector (Brand and Perrimon, 1993) via the *Eco* RI and *Xba*I sites. Transgenic flies were generated by the standard transformation protocol (Spradling and Rubin, 1982). Flies carrying UAS-*Race* 1 were crossed to stripe-2 *eve-Gal4* drivers (gift from S. Small) and the expression of UAS-*Race* 1 was confirmed by staining with anti-Race 1 antibodies. Guinea pig or rabbit anti-Race 1 antibodies were generated (Covance) as described by P. ten Dijke et al. (P. ten Dijke et al., 1992). To obtain uniform early embryonic expression of the UAS-*Race* 1 transgene, a maternal *Gal4* driver was used in which the *GALA-VP16* fusion protein is expressed maternally under the control of the  $\alpha$ -*tubulin 67C* promoter. These were further crossed into a *zen<sup>w36</sup>/TM3, hb-lacZ* background for the *zen* mutant rescue experiments.

**Interaction**

File Menu Edit Menu

Show Images Mark as Gene Mark as NOT Gene

PaperView EntitiesView Tokens to verify Help on verify task

- zen
- C zen
- C zen
- C zen
- The zen cDNA
- C zen
- C the UAS-zen
- a zen w36
- the zen mutant rescue
- C zen
- C zen
- the zen pattern

Created by PCR mutagenesis. Two proximal Zen-binding sites were mutated as follows: TAGAAAATAACTGCA. Constructs were transformed into flies using a standard transformation vector that contains the UAS promoter (Brand and Perrimon, 1993). At least three transgenic lines were generated for each construct.

Wild-type, mutant and transgenic embryos were stained with anti-Zen antibodies (P. ten Dijke et al., 1992), dehydrated and mounted on slides. The slides were stained with anti-Zen antibodies (P. ten Dijke et al., 1992) and developed using the Vectastain ABC kit (Pierce and Warriner, 1991).

The *Race* 1 promoter was cloned into a pUAST vector (Brand and Perrimon, 1993) via the *Eco* RI and *Xba*I sites. Transgenic flies were generated by the standard transformation protocol (Spradling and Rubin, 1982). Flies carrying UAS-*Race* 1 were crossed to stripe-2 *eve-Gal4* drivers (gift from S. Small) and the expression of UAS-*Race* 1 was confirmed by staining with anti-Race 1 antibodies. Guinea pig or rabbit anti-Race 1 antibodies were generated (Covance) as described by P. ten Dijke et al. (P. ten Dijke et al., 1992). To obtain uniform early embryonic expression of the UAS-*Race* 1 transgene, a maternal *Gal4* driver was used in which the *GALA-VP16* fusion protein is expressed maternally under the control of the  $\alpha$ -*tubulin 67C* promoter. These were further crossed into a *zen<sup>w36</sup>/TM3, hb-lacZ* background for the *zen* mutant rescue experiments.

Done.

# Image Processing

- Low-dimensional **feature vector** to summarise content of each image
- **Colour and Intensity** global bitstring, concatenated with:
  - Wavelet decomposition for **edge information**
  - Project vectors to randomly generated **hyperplanes**
  - Use their signs as key for **locality sensitive hashing**

# Indexing for Search

- **Lucene** – open source IR library, native XML handling, scalable
- **Fields**: word stems & lemmas, GRs, and named entities
- **Ranked search** overlaid with Boolean operators that alter rank
- **Search** by word stems and named entity (classes) in search box
- **Refine** search over sentences using lemmas and GR-patterns



## Distributed Paper Recovery and Annotation

- Each paper takes av. **10mins** to run thru' pipeline
- Use (UK part of) **Grid** (for LHC data processing) 200K CPUs
- **15K FlyBase papers**, 8K hours CPU, 3 days, max 100 jobs
- **Ganga**: error handling and job resubmission
- **Distrbuted Spider**: retrieved over 350K PDFs for papers

## PaperSearch: Example Query Session

**Goal:** Find out which genes are involved in eye development and what they do.

**Query:** Find all sentences in figure captions within the document collection which contain any gene name premodifying the term *expression*, where the figure is a picture of an eye.

**Method:** Incrementally and interactively combine term search, image clustering, and pattern search over GRs to realize this query.

### Screenshots:

Highlighted search terms, Gene names, Gene products

## Step1: Captions containing eye

Flybase search

http://beta.camtology.com/sci/

Camtology  
content sensitive search

Flybase search

Welcome, Andrew | profile | log out

Science | DVDs | Wine

1 2 3 4 5 ... next

Search Browse Options

Help

Search text  Search captions

**FBRf0155717** Benassayag, Plaza, Callaerts, Clements, Romeo, Gehring and Cribbs. (2003)

Fig. 1. **ey** alleles isolated as dominant **enhancers** of **eye** loss induced by ectopic expression from the **HSPB** sy **transgene**. Dose-sensitive **eye** loss induced by **HSPB** sy.

(A-C) Heads of flies carrying one, two or four **HSPB** sy copies, respectively.

(B) Two copies; slightly reduced **eye** (arrowhead).

(C) Four copies; complete **eye** deletion (arrowhead).

The genetic screen selected for female progeny carrying two **HSPB** sy copies (B) or an interacting Enhancer locus, which yields an **eye** loss resembling four copies (C).

(D) Strength and specificity of the **pb** - **ey** genetic interaction.

Four new alleles (**ey** JD, **ey** D1Da, **ey** I1, **ey** EH), two previously isolated **ey** loss-of-function alleles [**ey** 2 and Df(4) BA], and several alleles of other genes implicated in **eye** differentiation, were tested for dose-sensitive interactions with **HSPB** sy. Males harboring one **HSPB** sy copy and heterozygous for mutant alleles of the **eye** development genes **silene oculis** (**so**), **eyes absent** (**eya**), **eye** gene (**ey**) or **ey** were crossed with homozygous **HSPB** sy females.

Maximal **eye** loss expected is 50 % in the resulting female progeny (harboring two copies of **HSPB** sy) because half should carry the **eye** gene mutation. Several allele names are shortened: **eya** c1 is *eya* c1R1 (from L. Zipursky), **eyg** M is *eyg* M3-12 (from H. Sun) and **ey** D1 is *ey* D1Da (this paper) (E,F).

All four newly isolated **ey** alleles should yield truncated forms of the **EY** protein.

(E) Representations of wild-type **EY** protein and of the proteins encoded by **ey** JD, **ey** D1Da (Callaerts et al., 2001), or predicted based on the sequences of **ey** I1 and **ey** EH (this paper).

HD, homeodomain; PD, paired domain.

(F) Sequences of wild-type **ey** and of the mutant lesions in **ey** I1 and **ey** EH.

Modifications in the mutant sequences are underlined, exon or intron boundaries are indicated by arrows, and a new stop codon by \*.

**FBRf0103272** authors unknown

Fig. 1.

The adult **ey** 2 phenotype and rescue by an **eyeless** minigene.

(A-C) Scanning electron micrographs of heads of (A) a wild-type fly and (B,C) **ey** 2 flies with moderate and strong **eye** phenotypes, respectively.

Anterior is to the left.

The fly with the strong **eyeless** phenotype has a small head and completely lacks the compound **eyes** (C).



Bristles normally surrounding the **eye** are also missing.

The ocelli on the dorsal head are not affected in **ey** 2 flies (arrowheads).

(D) The **eye** enhancer located in the first **intron** of the **ey** gene drives expression in the **eye** disc. -galactosidase activity staining of an eye-antennal disc from a third instar larva carrying an **ey** enhancer lacZ transgene. -galactosidase activity is detected in the entire **eye** disc (to the right), barely in the

**ey** (2072)

View the location of search results on the world map. View large map.

## Step2: Select an image of an eye

Camtology  
content sensitive search

Flybase search

Welcome, Andrew | profile | log out  
Science | DVDs | Wine

1 2 3 4 5 ... next

Search Browse Options

View the location of search results on the world map. [View large map.](#)

**ey**  
eye (2972)

Go to [http://beta.camtology.com/medial/flybase/0150000-0159999/FB0155717/FB0155717\\_fig\\_2.png](http://beta.camtology.com/medial/flybase/0150000-0159999/FB0155717/FB0155717_fig_2.png)

Flybase search

http://beta.camtology.com/sci/

Google

Help

eye

Search text  Search captions

**FB0155717 Benassayag, Plaza, Callaerts, Clements, Romeo, Gehring and Cribbs. (2003)**

**Title** Evidence for a direct functional antagonism of the selector genes proboscipedia and eyeless in *Drosophila* head development.

**Authors** Benassayag, Plaza, Callaerts, Clements, Romeo, Gehring and Cribbs.

**Abstract** Diversification of *Drosophila* segmental and cellular identities both require the combinatorial function of homeodomain-containing transcription factors. Ectopic expression of the mouthparts selector proboscipedia (*pb*) directs a homeotic antenna-to-maxillary palp transformation. It also induces a dosage-sensitive eye loss that we used to screen for dominant Enhancer mutations. Four such Enhancer mutations were alleles of the *eyeless* (*ey*) gene that encode truncated EY proteins. Apart from eye loss, these new *eyeless* alleles lead to defects in the adult olfactory appendages: the maxillary palps and antennae. In support of these observations, both *ey* and *pb* are expressed in cell subsets of the prepupal maxillary primordium of the antennal imaginal disc, beginning early in pupal development. Transient co-expression is detected early after this onset, but is apparently resolved to yield exclusive groups of cells expressing either *PB* or *EY* proteins. A combination of *in vivo* and *in vitro* approaches indicates that *PB* suppresses *EY* transactivation activity via protein-protein contacts of the *PB* homeodomain and *EY* Paired domain. The direct functional antagonism between *PB* and *EY* proteins suggests a novel crosstalk mechanism integrating known selector functions in *Drosophila* head morphogenesis.

**Journal** Development:130.3

**Pubmed** 12490563

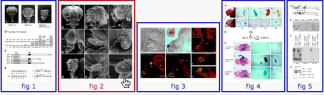

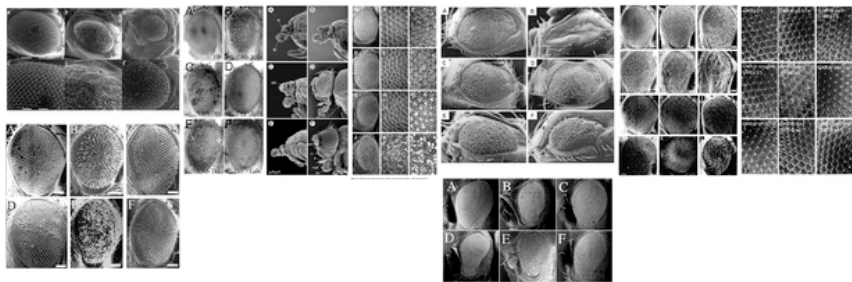


fig 1 fig 2 fig 3 fig 4 fig 5

Fig. 1. **ey alleles** isolated as dominant **Enhancers** of **eye** loss induced by ectopic expression from the **HSPB sy transgene**.  
Dose-sensitive **eye** loss induced by **HSPB sy**.  
(A-C) Heads of flies carrying one, two or four **HSPB sy** copies, respectively.  
(B) Two copies; slightly reduced **eye** (arrowhead).  
(C) Four copies; complete **eye** deletion (arrowhead).  
The genetic screen selected for female progeny carrying two **HSPB sy** copies (B) or an interacting Enhancer locus, which yields an **eye** loss resembling four copies (C).  
(D) Strength and specificity of the **pb-ey** genetic interaction.  
Four new **alleles** (**ey JD**, **ey D1Da**, **ey 11**, **ey EH**), two previously isolated **ey** loss-of-function **alleles** [**ey 2** and Df(4) BA], and several **alleles** of *eyeless* were included in this screen. **ey JD** and **ey D1Da** were the only alleles that were capable of interacting with **HSPB sy**.



# Step3: Clustered images (captions not shown)



## Step4: Refine text search within caption for one image

Safari File Edit View History Bookmarks Develop Window Help

http://beta.camtology.com/sci/ Flybase search

Welcome, **Andrew** | profile | log out

Science | DVDs | Wine

prev 1 2 3 4 5 6 ... next

Search Browse Options

Help! eye Go! Search text Search captions

**FBfr0162225 Delon, Chanut-Delalande and Payre. (2003)**

Fig. 7. **eyo/svb expression** can promote formation of cytoplasmic extensions when directed in the **eye**. Scanning electron micrographs of adult **eyes** at high magnification. (a) The wild-type **eye** is composed of a regular arrangement of ommatidia and interommatidial sensory bristles. Expression of **OvoA** using the **GMR-GAL4**-driver results in a diminution of the size of the **eye**, with ommatidial fusion and absence of bristle (b). Expression of **OvoB** disrupts the regular ommatidial structure and triggers the formation of ectopic cuticular extensions at the **eye** surface (c). Bristle formation is also affected. **Svb** overexpression leads to similar results, with smaller ectopic extensions (d). (e, f) **Eyes** were dissected at 96 % of pupal development and stained for F-actin. The regular arrangement of actin-rich rhabdomeres from the seven photoreceptors observed in a wild-type ommatidium (e) is disrupted in **eyes** expressing **OvoB** from the **GMR-Gal4** driver (f). **OvoB** expression leads to a strong increase of F-actin in rhabdomeres, which appear amalgamated in an highly abnormal structure.

no match  
match lemma only  
match anything  
match lemma and biotype

**FBfr0155717** **Jan, Zou, Frei and Noll. (2001)**

Headless phenotypes of **ey** alleles with **ey** functions, which depends on DNA-binding activities different from that of **Ey**. Left **eyes** of flies are shown in scanning electron micrographs. (A) UAS-Ey rescues the headless phenotype in **ey-Gal4/UAS-Gab-7**; **UAS-Ey** / + flies almost completely to a small-eye phenotype. (B) A different small-eye phenotype is produced in **ey-Gal4** / +; **UAS-Ey** / + flies. (C) **ey-Gal4** / +; **UAS-GE-8** / + flies, which carry mutations in amino acids 42 (Q mutated to I), 44 (R to Q) and 47 (H to N) in the paired domain of **UAS-Gab** changing its DNA-binding specificity to that of the **Ey** paired domain, exhibit little or no interference with **ey** functions and display, in four out of six lines, a phenotype similar to wild type (D) or, in two lines, a weak phenotype similar to **ey-Gal4** / +; **UAS-Ey** / + flies (B).

**FBfr0144814** **Kronhamn, Frei, Daube, Zhao, Shi, Noll and Rasmuson-Lestander. (2002)**

Fig. 7. Headless phenotype of **ey** D pharates and their partial rescue by inhibition of apoptosis. Scanning electron micrographs of the anterior portion (A-D,I,J) or left **eyes** (E-H) of pharate (B-D) or viable (A,E-J) adults of the genotype indicated are compared. Note that, in contrast to the headless phenotype of **ey** hdl flies, the penetrance and expressivity of the headless phenotype of **ey** D pharates is the same at 18°C and 25°C with about 50 % of the pharates exhibiting no (B) or only few (C) structures derived from the eye-antennal discs, while the phenotype of most pharates is much stronger than that shown in D. The variability of heterozygous **ey** D phenotypes (E-I) presumably reflects a strong influence of the genetic background as illustrated by the **eyeless** phenotype obtained after several generations of selections for

View the location of search results on the world map. View large map.

**ey** FBfr0155717\_fig\_2.png  
q:ey  
l:0.1 img:200.0 inv:true (443)

**ey** eye (2972)

## Results of refined text search

Camtology  
21st century search

Flybase search

Welcome, Andrew | profile | log out  
Science | DVDs | Wine

1 2 3 4 5 next


Search Browse Options

View the location of search results on the world map. [View large map.](#)

Help    Search text  Search captions


**FBrf0192127**

The embryonic expression patterns of SIP1 were determined using the C-terminal regions of **SIP1** ( probe F ). a Embryonic day ( E ) 9.5 transverse sections showing low-level SIP1 in the neural tube ( n ) and notochord ( thin arrow ). b E10.5 day sagittal section showing strong expression of **SIP1** in neural tissues, dorsal root ganglia ( small arrows ), trigeminal ganglia ( arrowheads ), and myotomes ( open arrow ). c E11.5 day oblique sections showing **SIP1** expression in other non-neural tissues and dorsal root ganglia ( small arrow ). d Sagittal section of E12.5 day embryo showing expression of SIP1 in the ependymal layer of the neural tube and the ventricular zones of the brain, both regions of dividing neurons, and dorsal root ganglia ( small arrow ). e Transverse section of E12.5 days embryo showing **SIP1** expression in umbilical cord ( U ) and blood vessels and weak positive expression of **SIP1** in developing endocardium ( open arrow ) in the heart. f Frontal section of E14.5 day embryo showing through the brain, eyes, and nasal sinuses showing **SIP1** expression in the cortex ( C ), the mesenchyme around the whisker follicles ( w )




**FBrf0152356** (2002)

Lenses appear fused and bristles are misplaced and reduced in numbers. c.f : **eyeless-GAL4** driving UAS- **disco C** expression anterior to the morphogenetic furrow of the eye imaginal disc causes a severe reduction of compound eyes.




**FBrf0108521** (1999)

While **ELAV** expression in R2/R5 and R8 was unaffected in **sevE-GAL4/UAS-spry** eye discs, we observe loss of **ELAV** expression from one or two cells of the R3/R4/R1/R6 cell group in a majority of clusters ( indicated by **white** arrowheads ).



**FBrf0183217**

The effect of dominant-negative **Myosin VI** expression .



`<key> 0:[*any*+bio(gene)]  
1:[expression] 2:(ncmod  
1:[expression]  
0:[*any*+bio(gene)])  
2:(ncmod 1:[expression]  
0:[*any*+bio(gene)]) (41)`

ey  
FBrf0155717\_fg\_2.png  
q:ey  
L:0.1 Imp:200.0 inv:true (443)

ey  
eye (2972)

# Gene Expression

Query 1: `express AND Adh`

Query 2: `express →+ Adh`

Query 3: Query 1 + OR `overexpress...` CG32954...

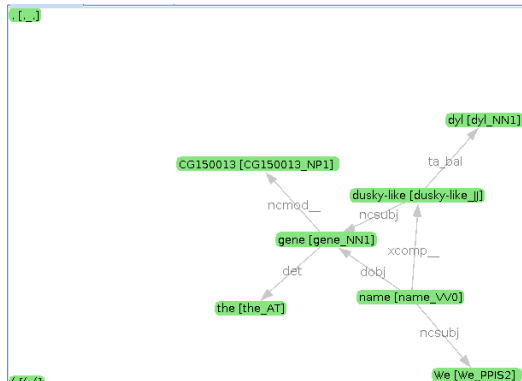
Query 4: Query 2 + OR `overexpress...` CG32954...

- `express Adh`
- `expression of Adh`
- **Adh** is one of the most highly **expressed** genes

Query	1	2	3	4
MAP	0.735	0.758	0.855	0.933



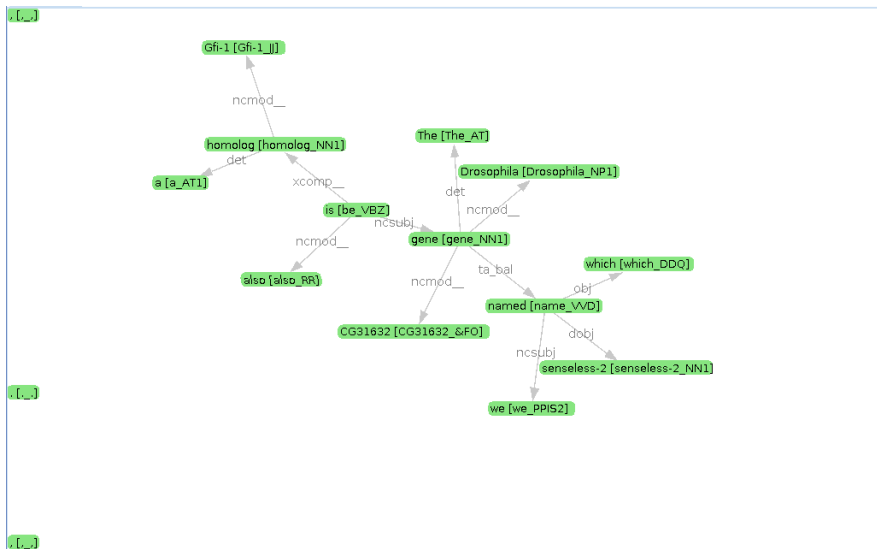
# GRs and Gene Naming



( [ ( ] )

S2P0: We name the gene CG150013 dusky-like ( dyl ) .

## GRs, anaphora and naming



S3P0: The Drosophila gene CG31632, which we named senseless-2, is also a Gfi-1 homolog.



# Gene Naming Queries

Query 1: bioG:CG\* AND name

Query 2: bioG:CG\* AND (name OR call OR refer OR ...)

Query 3: Query 2 + (CGid 'refer to as' GENE) OR ('name' CGid GENE) OR (CGid '(' GENE ')') ...)

Query 4: Query 2 + CGid  $\rightarrow^+$  GENE

Query 5: Queries 2, 3 + 4

Query	1	2	3	4	5
MAP	0.116	0.461	0.552	0.512	0.562

## User Interface and Usability

- **Term/class queries** over sentences useful
- **Image** handling useful, clustering unintuitive
- Intuitive construction of **GR-patterns**
- **But** complex patterns cannot be easily constructed
- **Ranking** of complex (refined) query results often unintuitive
- **3/3 Curators** are enthusiastic, but often frustrated...

## Conclusions and Further Work

- 1 From PDF to SciXML using NLP
- 2 Integration of image and text search
- 3 Generic: domain-independent or weakly-supervised
  - Make it all work better!
  - IR to IE: Saving searches and search results
  - Inference: e.g. transitivity (genes  $\rightarrow$  proteins  $\rightarrow$  diseases)

# Acknowledgements

Contributors	Affiliation	Funding
Rachel Drysdale	Cambridge Univ	BBSRC
Caroline Gasperin	Cambridge Univ	CAPES
Karl Harrison	Cambridge Univ	STFC
Nikiforos Karamanis	Cambridge Univ	BBSRC
Ian Lewin	Cambridge Univ	BBSRC
Andrew Naish	Camtology Ltd	Camtology
Andrew Parker	Cambridge Univ	STFC
Marek Rei	Cambridge Univ	EPSRC
Advait Siddharthan	Cambridge Univ	STFC
David Sinclair	Imense Ltd	Imense
Simone Teufel	Cambridge Univ	BBSRC
Andreas Vlachos	Cambridge Univ	BBSRC
Rebecca Watson	iLexIR Ltd	iLexIR

Papers: [‘FlySlip Project’](#) / [‘Ted Briscoe’](#) / Questions: [Ask](#), [Email...](#)