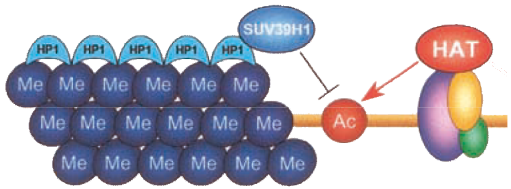
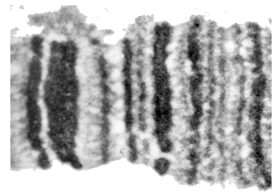
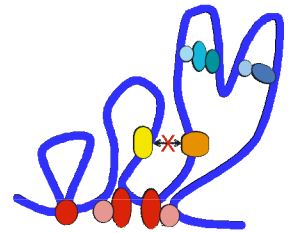
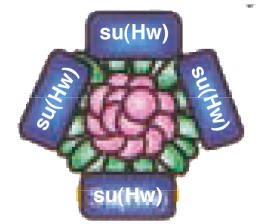
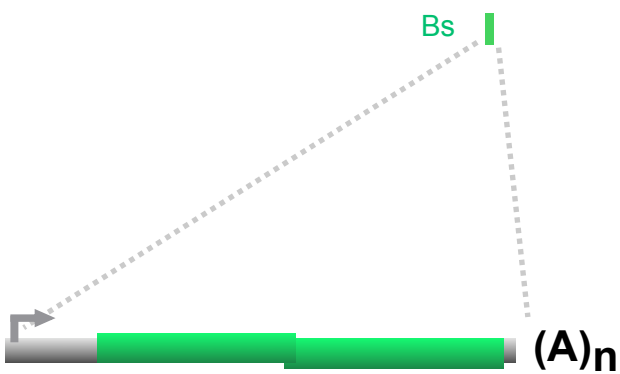
Comprehensive Functional Annotation of Metazoan Genomes

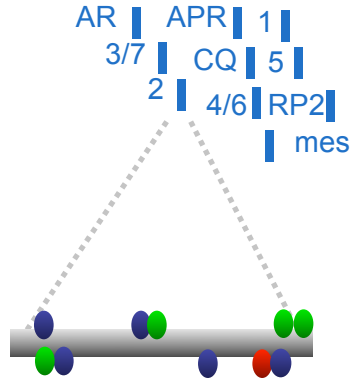


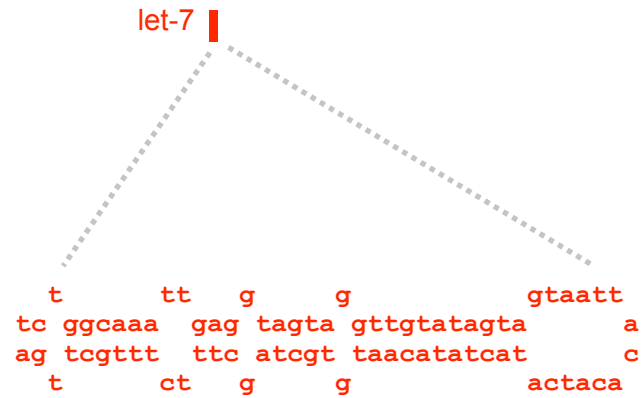












Transposable elements **Enhancers** **miRNAs, etc.**

ENCODE

- An initial list of functional elements to identify.
- An initial list of technologies/approaches that will be utilized.
- Criteria for recruiting additional participants.
- Selection of genomic targets.
- Establishment of a database and browser.
- Management of the research network.

ENCODE Participation Criteria

- Each participant will inform the NHGRI and the ENCODE Consortium members about their experimental plans.
- Each participant agrees to analyze the entire set of selected regions and not just a subset of the target sequences.
- Each participant is expected to contribute significantly to the Project, bringing his/her particular expertise to accomplish the goals of the Consortium in a timely manner.
- Participation in the Consortium should consist of more than deposition of data to the shared database, and should include substantial intellectual contributions to the Project.
- All members of the Consortium will share results according to the data sharing policy developed by the NHGRI.
- All members of the Consortium agree to participate in group activities, including attending periodic workshops to discuss the Project's progress and coordinating the publication of research results.
- The participants will fully disclose all publicly funded algorithms, software source code, and experimental methods.
- The participants agree that they will not disclose confidential information obtained from other members of the Consortium.
- The NHGRI, with the advice of the project's Scientific Advisory Panel will approve of participants for involvement in the Consortium.
- Additional criteria may be added upon recommendations of the ENCODE Scientific Advisory Panel.

dENCODE

- **Provide relevant input to ENCODE: methods and data.**
- **An interactive consortium of research groups.**
  - **Small core of groups (Genome Centre Model)?**
  - **Larger collection of groups?**
- **Focused research themes to catalogue genomic sequence elements.**
- **Rigorous and comprehensive analysis:**
  - **Entire *melanogaster* genome.**
  - **Selected regions of other *Drosophila* species.**
- **Assessment of HTP experimental methods.**
- **Development of data validation criteria.**
- **Development of data curation standards.**
- **Development of data integration and visualization tools.**
- **Development and production of common resources.**

White Paper: Fly Contribution to ENCODE

- Inventory of elements
  - What elements are we looking for?
- Inventory of approaches
  - Experimental
    - Array Based
    - Transgenic
    - Other
  - Computational
    - Prediction
    - Algorithm development
    - Analysis
- Curation
  - Warehousing data
  - Annotation and meta-data standards
  - Data presentation
- Governance
  - Project management
  - Consortium rules
  - Dissemination

dENCODE ?

Parts List

- Protein coding genes.
  - Transcripts, splice forms
- Non-protein coding genes
  - tRNA, rRNA, snRNA, siRNA, miRNA
- Transcriptional regulatory regions
  - Promoters, enhancers
- Processing regulators
  - Splice site, splice enhancers
- Chromatin regulatory structures
  - insulators, silencers, boundaries, scaffold
- Transposable/mobile elements
- Unknown sequence features

Tools

- 12 Drosophila genomes
- 50 melanogaster genomes
- 367570 ESTs, 11,203 cDNAs
- cDNA arrays
- Amplicon arrays
- Long oligo arrays
- Short oligo arrays
- Custom arrays
- Genome Tile arrays
- Other

Drosophila DNase I Footprint Database

Drosophila DNase I Footprint Database

http://www.flyreg.org/

Regular Expr...ions in Perl | Google | Entrez-PubMed | NCBI BLAST | Apple | Amazon | News ▾ | FlyBase @ fb...n.cam.ac.uk

Drosophila DNase I Footprint Database

Home | Browse by Target | Browse by Factor

Background

This page provides access to results of the systematic curation and genome annotation of 1,367 DNase I footprints for the fruitfly *D. melanogaster* reported in Bergman, Carlson and Celniker (200X) *Bioinformatics*. These data have been extracted from 201 primary references and provide a non-redundant set of high quality binding site information for 87 transcription factors and 101 target genes. Unlike previous work, this dataset has been generated from a single experimental data type, represents all available developmental stages (including anterior-posterior, mesoderm and imaginal disk patterning), and is linked explicitly to finished genome sequence coordinates. This dataset should provide a useful resource for computational analyses of transcription factor binding site biology in the genus *Drosophila*.

Browse the *Drosophila* DNase I Footprint Database (v1.0)

Browse by Target: A hypertext document sorted by target gene with links to Flybase, PubMed and the UCSC genome browser showing alignments of footprints with *D. yakuba*, *D. pseudoobscura*, and *Anopheles gambiae*.

Browse by Factor: A hypertext document sorted by binding factor with links to Flybase, PubMed and the UCSC genome browser alignments of footprints with *D. yakuba*, *D. pseudoobscura*, and *Anopheles gambiae*.

Download the *Drosophila* DNase I Footprint Database (v1.0)

Download GFF: A text file in Sanger Institute's GFF v2 format with additional target, factor, PubMed ID fields. This file is identical to Supplemental File 3 in Bergman, Carlson and Celniker (200X) *Bioinformatics* <link>. The coordinate system is half-open base {e.g. a (start, end] pair of (100,200] represents bases 101-200 on the genome sequence}. The strand for genome sequence coordinates is purposefully not reported since a binding site is a double-stranded feature.

Download Sequences: A text file of sequences in FASTA format corresponding to footprinted regions on the plus strand of the *D. melanogaster* Release 3 genome sequence. A text file of file of sequences in FASTA format corresponding to footprinted regions extended by ±30 bp on the plus strand of the *D. melanogaster* Release 3 genome sequence can be found here.

Download SQL: A dump of the mySQL database (and documentation) used for data entry and storage derived from the GadFly schema.

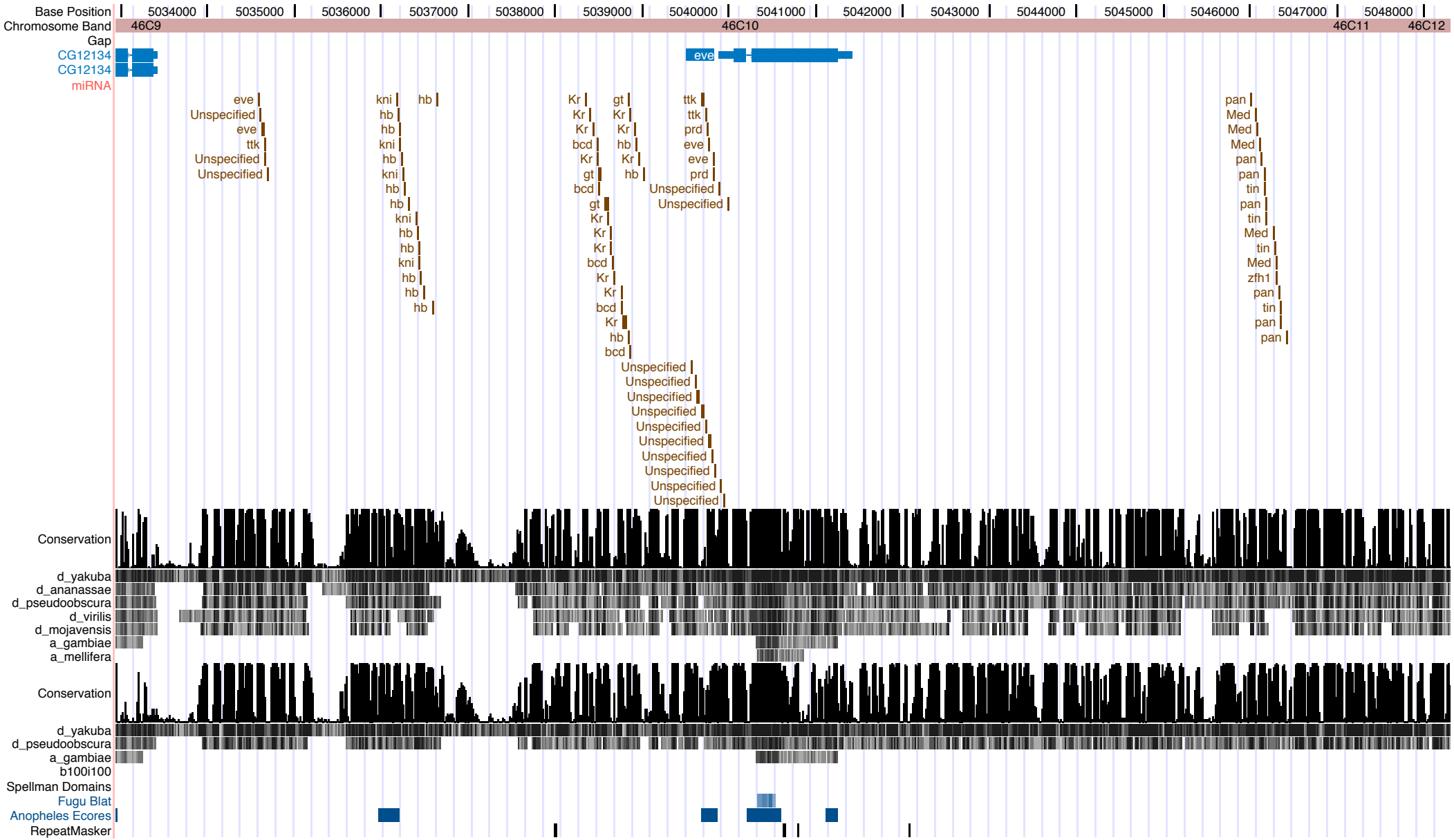
NB: The GFF file (not the Sequences or SQL database dump) defines the information contained in v1.0.

Browse & Download the Literature

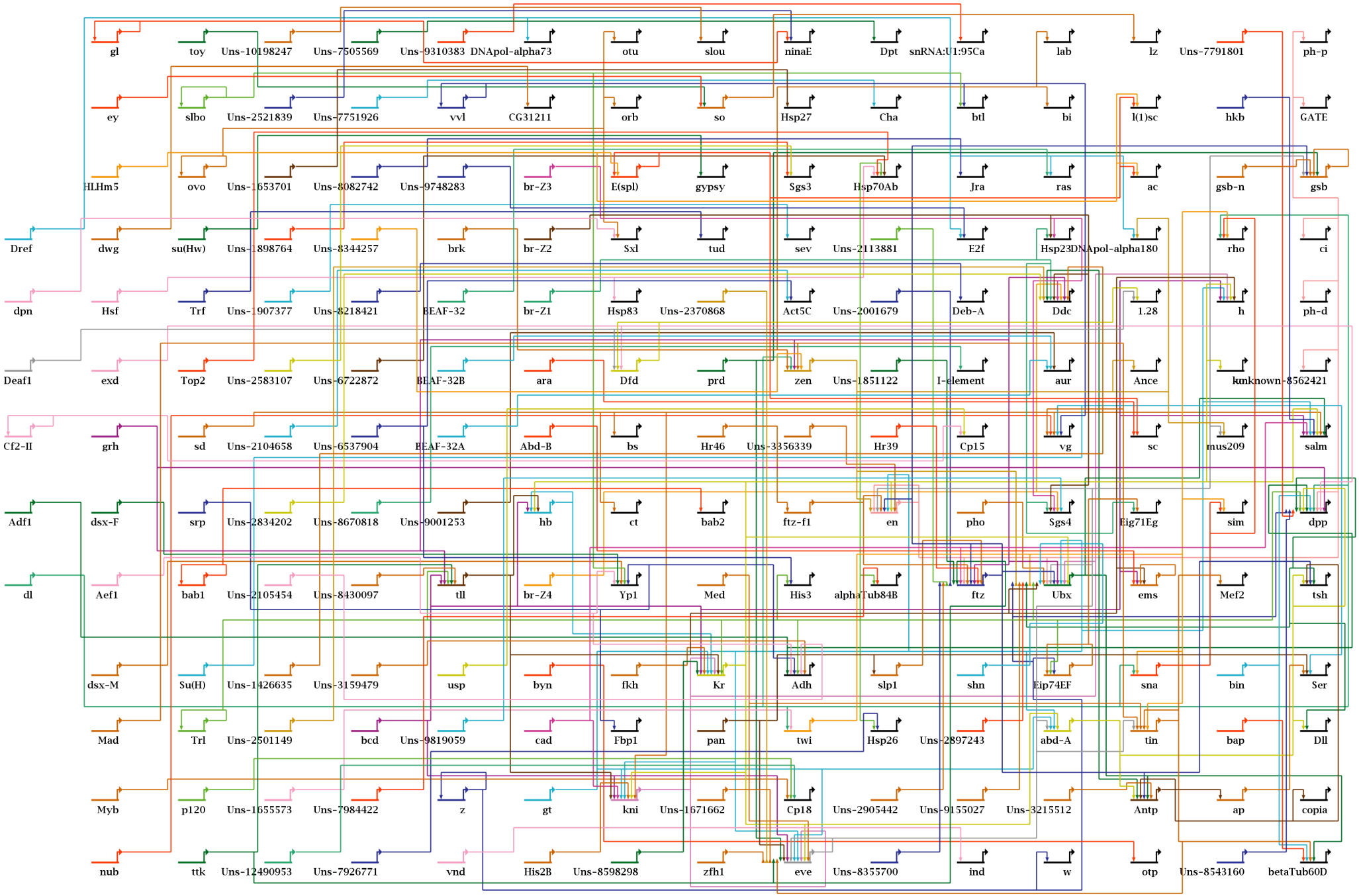
Browse PubMed (1984-1994) (1995-2004): Hypertext summaries of the 201 primary references that form the experimental foundation for this database.

Download XML: A tar.gz text file of 201 PubMed abstracts for literature mining.

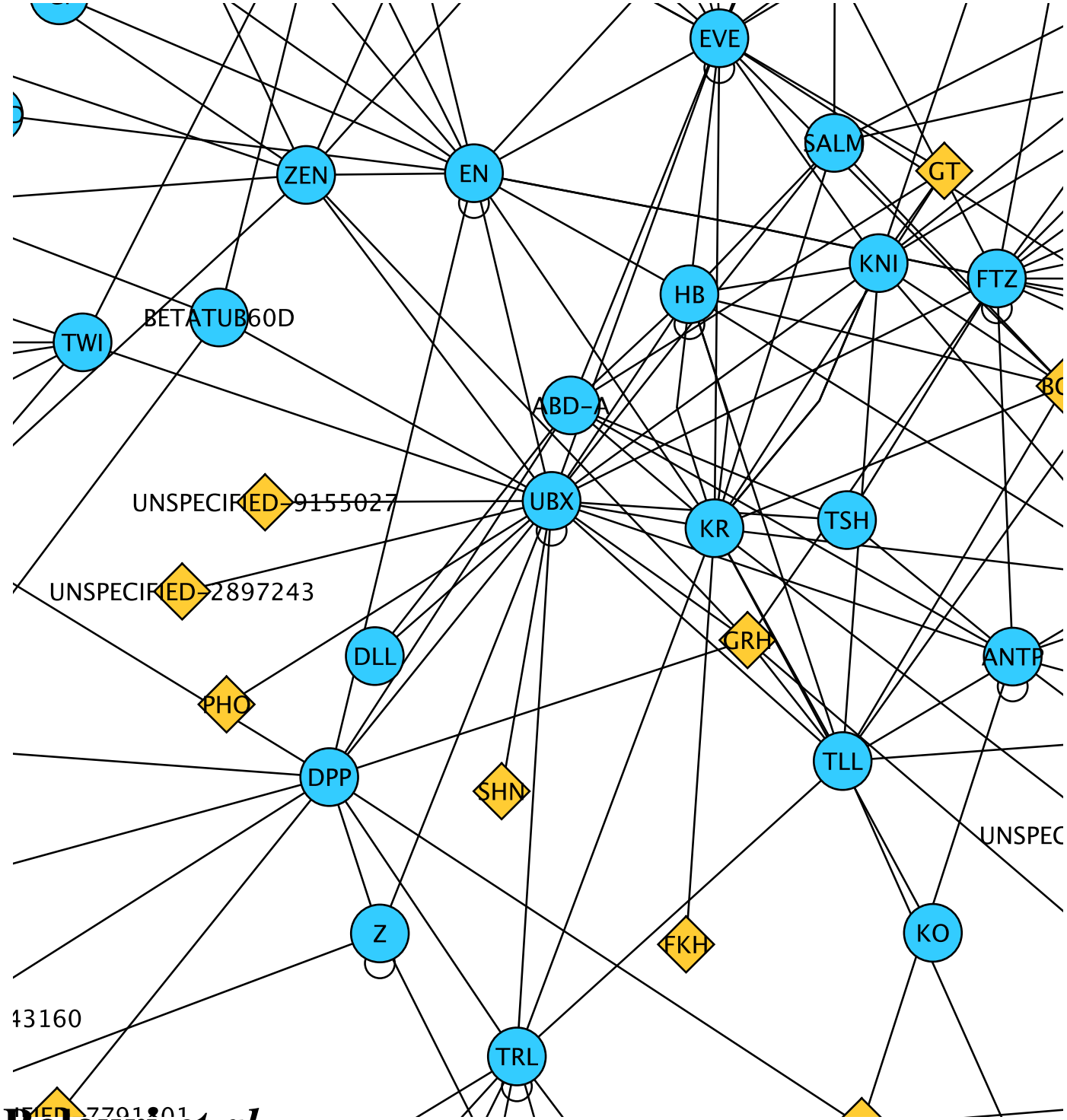
http://www.flyreg.org


Base Position
5034000
5035000
5036000
5037000
5038000
5039000
5040000
5041000
5042000
5043000
5044000
5045000
5046000
5047000
5048000
Chromosome Band
46C9
46C10
46C11
46C12
Gap
CG12134
CG12134
miRNA
eve
Conservation
d_yakuba
d_ananassae
d_pseudoobscura
d_virilis
d_mojavensis
a_gambiae
a_mellifera
Conservation
d_yakuba
d_pseudoobscura
a_gambiae
b100i100
Spellman Domains
Fugu Blat
Anopheles Ecores
RepeatMasker

From Footprints to Networks



BioTapestry: Bolouri *et al*



CytoScape: Bolouri *et al*

dENCODE
Meeting Plan

- **Functional & Comparative Genomics Tools (Greg Gibson)**
- **Computational DNA Element Discovery & Refinement I (Gene Myers)**
- **Experimental DNA Element Discovery & Refinement II (Sue Celniker)**
- **DNA Element Function (Hugo Bellen)**
- **dENCODE Database Structure & Data Serving (Bill Gelbart)**
- **Data Release, Recognition, & Publication (Mike Eisen)**
- **Panel discussions and conclusions with each group (Kevin White).**
- **White Paper draft**