



Development of Methods for Integrating Diverse Sources of Genome-Scale Data

Daniel Dvorkin and Katerina J. Kechris
University of Colorado Denver

7th Annual Rocky Mountain Bioinformatics Conference, 2009

Introduction

- Problem: noisy genome-scale data.
- Solution: combine multiple signals to filter out noise.
- Goal: locate hot spots for more in-depth coverage.
- Application: wing shape in *D. melanogaster*.
 - Intensively studied model organism.
 - Large public databases (FlyBase, etc.)
 - Easily identified phenotype, mutant vs. wild-type.

Materials

- Conservation data (UCSC Genome Browser)
 - 12 fly species plus 2 mosquito.
 - Base pair by base pair coverage.
- Transcription factor (TF) binding data for cubitus interruptus (Ci)
 - Developmental processes including wing formation.
 - Probes at 300 bp intervals.
- Position weight matrix (PWM) scores for Ci binding
 - 10 bp binding site motif.
 - Calculated for every position.
- Custom 2-color microarray expression data for ~ 13000 genes
 - 5 gene products of interest
 - mutant vs. wild-type

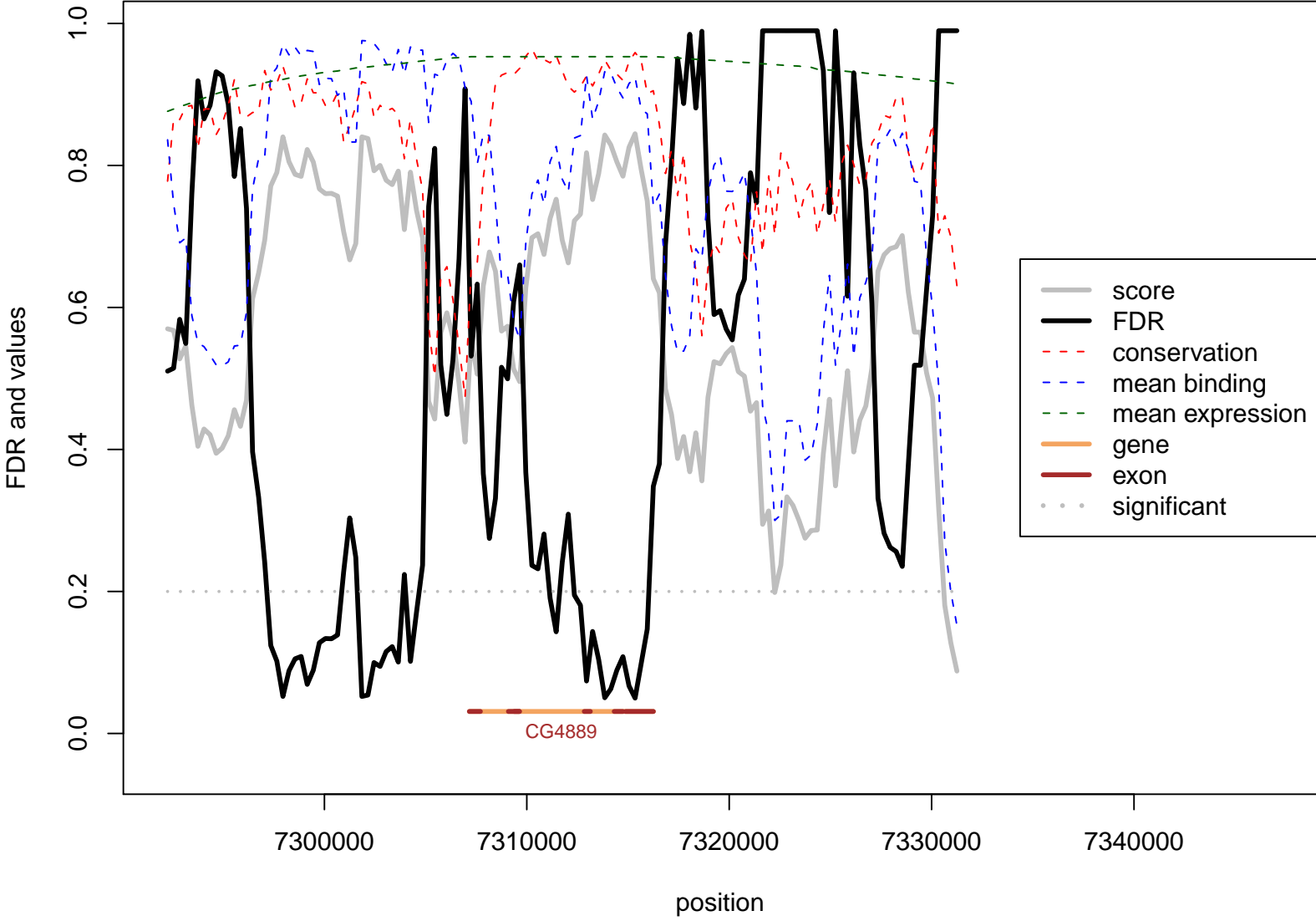
Methods

- Preprocessing: Take summary statistics within 300 bp blocks for each data type, analyze block values; rank and normalize to (0,1).
- *Union* of similar data types, *intersection* of different types.
- Three data types
 - Conservation (c).
 - Binding (b) and PWM scores (p).
 - Expressions (e_1, \dots, e_5).
- General score: $s = c \cap \cup(b, p) \cap \cup(e_1, \dots, e_5)$.
- Use product for union, mean for intersection:

$$s = c \times \frac{b + p}{2} \times \frac{e_1 + \dots + e_5}{5}.$$

- Calculate p -values by parametric bootstrap, adjust for FDR.

wingless (CG4889/wg) chr 2L



Discussion

- Promising early results
 - 4 of 19 target genes had significant results in neighborhoods.
 - 200 randomly selected genes, 2 significant results; 1 involved in cell adhesion, early development.
- Immediate future development plans
 - Systematic identification of significant genes.
 - Finer-grained normalization.
- Longer-term development plans
 - Other problems in other model organisms.
 - Sequencing-based expression analysis.
- Acknowledgements: Brian Biehs and Tom Kornberg (UCSF).
Work supported by NIH/NLM training grant T15 LM009451.