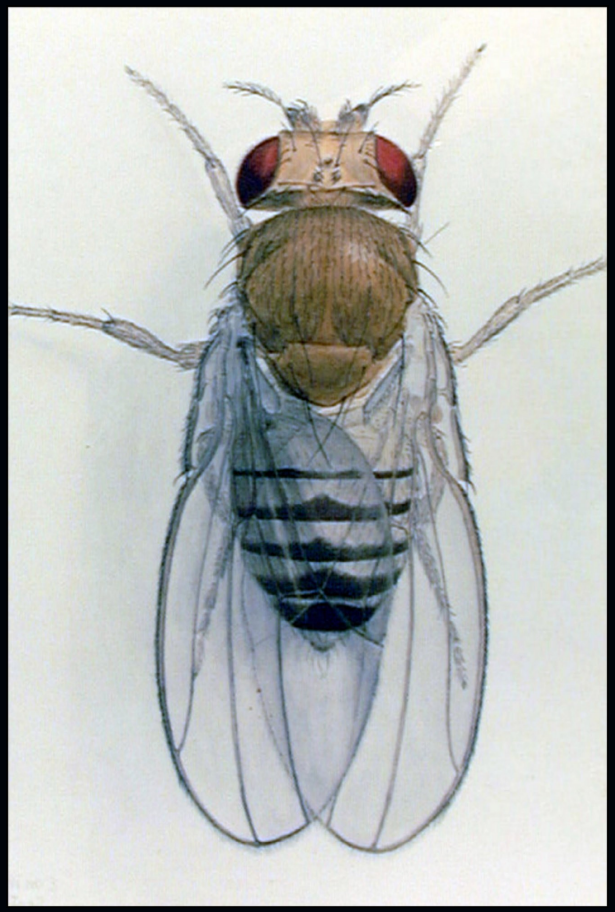
# Comprehensive Characterization of the *Drosophila* Transcriptome



49th Annual Drosophila Research Conference: modENCODE workshop

Susan Celniker
Lawrence Berkeley National Laboratory

# ModENCODE Project Goals

# How many genes in Drosophila? Current estimates

| Description | Release 5.6 |
|---|---|
| Protein-coding genes | 14,140 |
| tRNA genes | 314 |
| miRNA genes | 90 |
| snRNA genes | 47 |
| snoRNA genes | 249 |
| Pseudogenes | 88 |
| Misc. non-coding RNA | 88 |
| Transposons (and TE fragments) | 478 (5,552) |

FlyBase

# Aim 1:Expression

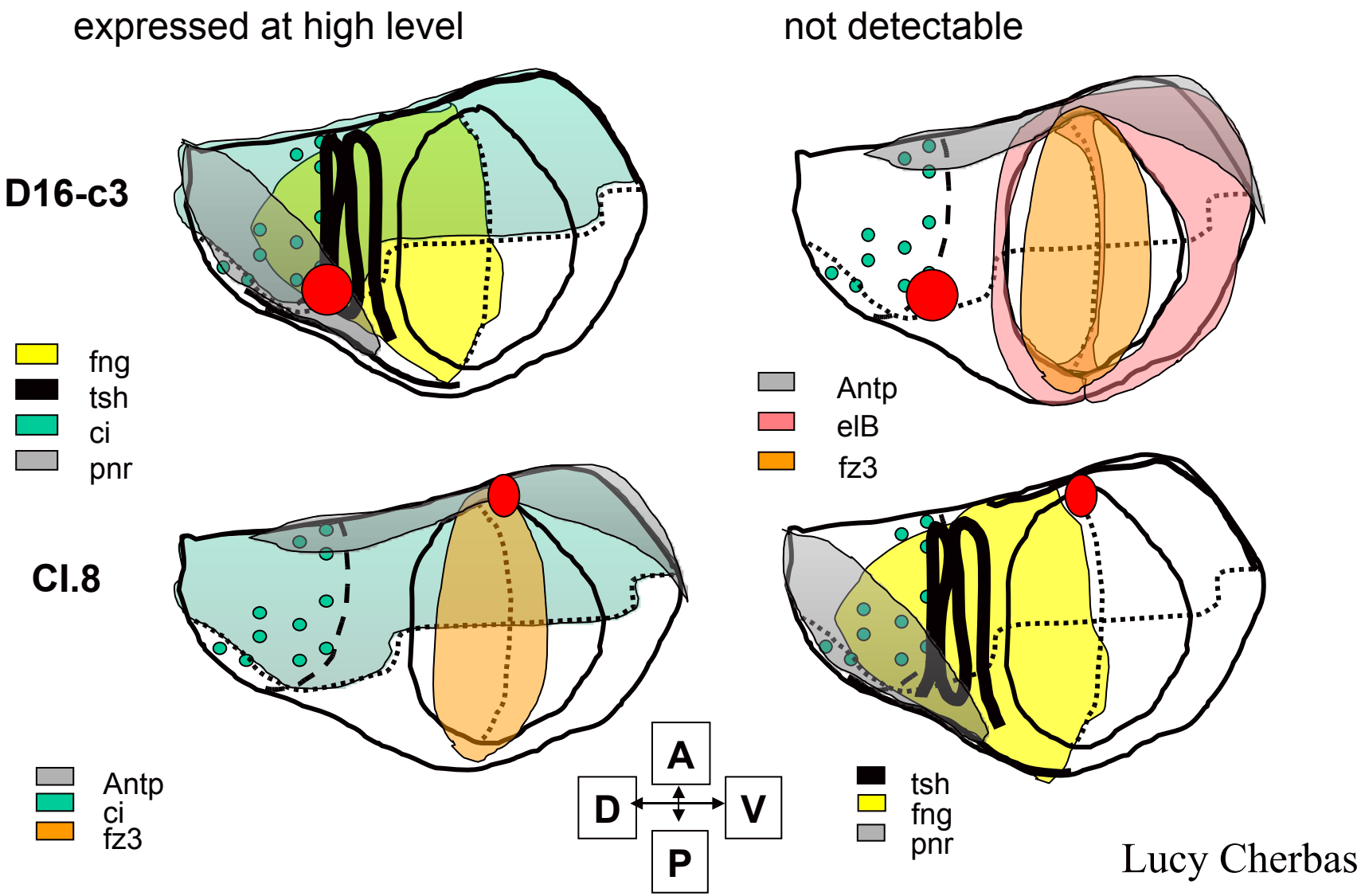
- ~300 RNA samples in biological triplicate
- ~300 samples on 38-bp genome tiling arrays
- 24 samples on 7-bp genome tiling array sets
- 160 RACE-fragment pools (16,000 prod's)

Comprehensive identification of transcribed sequences by microarray hybridization and next generation sequencing.

# RNA Samples

| | total RNA | A+ RNA | Nuclear RNA | Polysomal RNA | total bio samples | RNA preps (3x) |
|---|---|---|---|---|---|---|
| Embryo. time crse (12 times) | 12 | 12 | 12 | 12 | 48 | 144 |
| Larvae and puparia (10 times) | 10 | 10 | -- | -- | 20 | 60 |
| Pupae (2 samples per day) | 2 | 2 | -- | -- | 4 | 12 |
| Adults (male) (8 time pts over 16 days) | 3 | 3 | -- | -- | 6 | 18 |
| Adults (mated female) ( 8 pts over 16 d) | 3 | 3 | -- | -- | 6 | 18 |
| Cell Line Survey (70 lines) | 70 | -- | -- | -- | 70 | 210 |
| Cell Lines ⍰ select | -- | 30 | 30 | 30 | 90 | 270 |
| Dissected Tissues | 100 | 100 | -- | -- | 200 | 200 |
| avg: 10 tissues x 4 times x 3 reps | | | | | | |
| | | | | | 444 | **932** |

# Cell Lines represent specific lineages



# Expression Analysis of 15 *Drosophila* cell lines

**Base-pair coverage for individual cell lines**

| | | |
|---|---|---|
| 1182-4H | 23,932,660 | 21.5% |
| S1 | 23,057,046 | 20.7% |
| ML-DmD16c3 | 20,301,646 | 18.3% |
| ML-DmD32 | 20,157,642 | 18.1% |
| ML-DmD17c3 | 19,995,867 | 18.0% |
| ML-DmD20c5 | 19,966,250 | 18.0% |
| GM2 | 18,230,905 | 16.4% |
| Kc167 | 17,989,463 | 16.2% |
| ML-DmD20c2 | 17,920,747 | 16.1% |
| Sg4 | 16,997,360 | 15.3% |
| S2-DRSC | 16,828,540 | 15.1% |
| ML-DmD11 | 16,493,592 | 14.8% |
| S2R+ | 15,593,170 | 14.0% |
| ML-DmD8 | 15,455,167 | 13.9% |
| CME-L1 | 15,280,574 | 13.8% |

**Base-pair coverage for the union of expression**

| | | |
|---|---|---|
| 1182-4H | 23,932,660 | 21.5% |
| (+) S1 | 28,579,164 | 25.7% |
| (+) D16c3 | 32,336,076 | 29.1% |
| (+) D32 | 34,243,143 | 30.8% |
| (+) D17c3 | 37,017,601 | 33.3% |
| (+) D20c5 | 38,067,183 | 34.3% |
| (+) GM2 | 38,890,679 | 35.0% |
| (+) Kc167 | 40,919,342 | 36.8% |
| (+) D20c2 | 41,532,068 | 37.4% |
| (+) Sg4 | 41,968,002 | 37.8% |
| (+) S2-DRSC | 42,271,547 | 38.0% |
| (+) D11 | 42,613,889 | 38.3% |
| (+) S2R+ | 43,983,586 | 39.6% |
| (+) D8 | 44,258,602 | 39.8% |
| (+) CME-L1 | 44,335,437 | 39.9% |

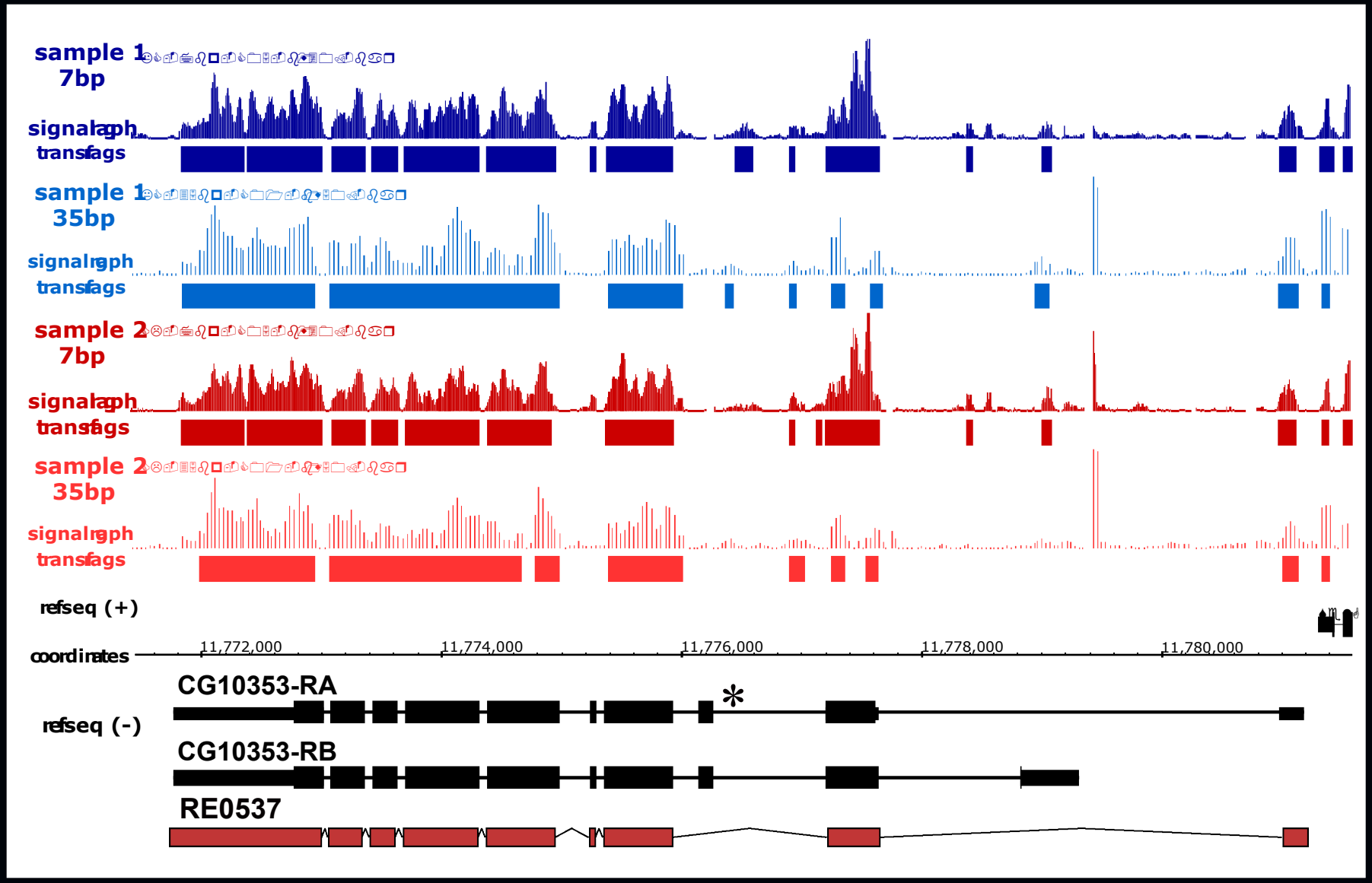
**Unique transcription per cell line**

| | | |
|---|---|---|
| ML-DmD8 | 267,558 | 0.24% |
| ML-DmD17c3 | 1,933,843 | 1.74% |
| ML-DmD32 | 832,618 | 0.75% |
| ML-DmD16c3 | 1,799,278 | 1.62% |
| 1182-4H | 1,089,389 | 0.98% |
| S1 | 1,426,649 | 1.28% |
| S2R+ | 1,353,685 | 1.22% |
| S2-DRSC | 246,568 | 0.22% |
| GM2 | 526,401 | 0.47% |
| Kc167 | 1,716,663 | 1.54% |
| ML-DmD20c5 | 512,226 | 0.46% |
| ML-DmD20c2 | 453,141 | 0.41% |
| ML-DmD11 | 306,839 | 0.28% |
| Sg4 | 278,251 | 0.25% |
| CME-L1 | 76,835 | 0.07% |

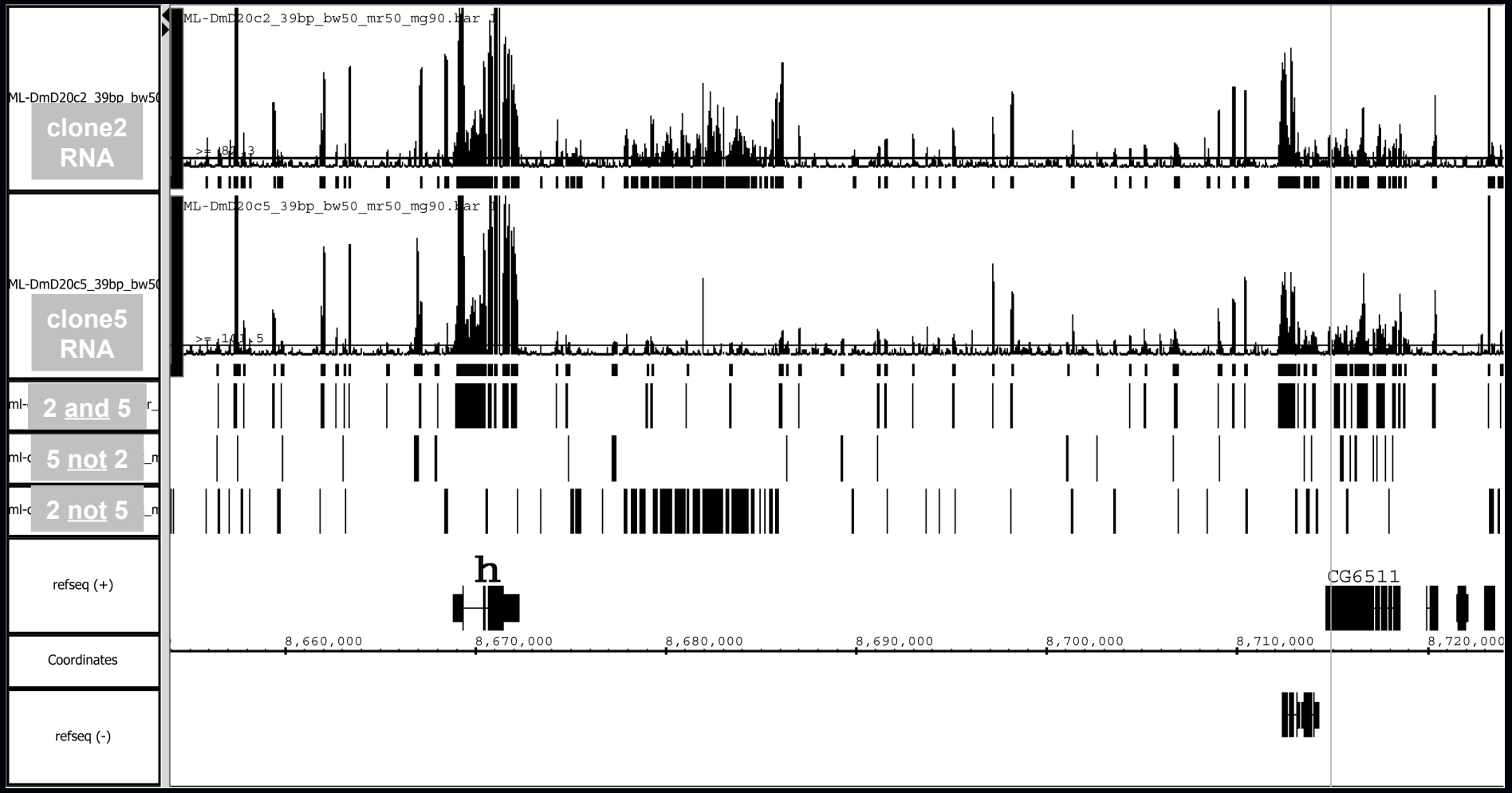
- all transfrags created with bandwidth 50, min-run 90, max-gap 90
- interrogated genomic space is calculated using “blanket transfrags” which are created assuming all probes are above threshold

- https://dgrc.cgb.indiana.edu/cells/store/catalog.html

# Improved exon discrimination using 7bp arrays



# Comparison of ML-DmD20 cell lines clone2 versus clone5

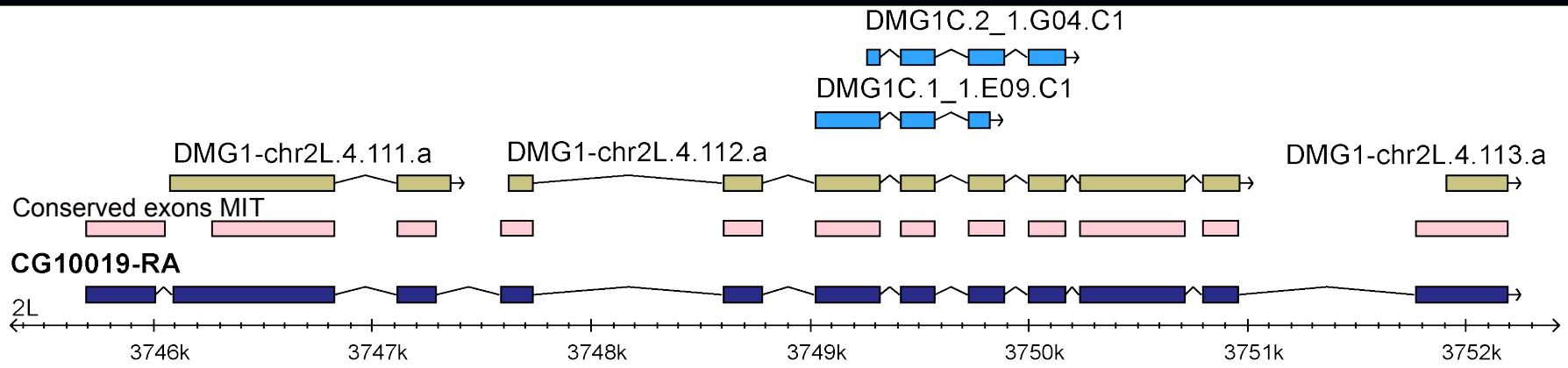



- Intergenic transcription downstream of *hairy (h)*

# Specific Aim 2: Synthesis and Validation

- Synthesis of RNA expression data, comparative data and gene predictions.
- 20,000 short RT-PCRs
- 20,000 RACE experiments
- Small RNA sequencing on 454: 16 runs
- 6,000 cDNA screens & 3,000 long RT-PCRs
- RNAi of 120 RNA binding proteins on arrays
- Identify *cis*-reg. elements in control of splicing

# Modeling and Validation




| Unique Splice Junctions in DMG2 | |
|---|---|
| | |
| Verified Splice Sites | 42,138 |
| Verified by cDNA and EST | 41,079 |
| Verified by RT - PCR | 1,059 |
| | |
| Unverified Splice Sites | 8,006 |
| No overlapping cDNAs | 5,717 |
| Overlapping cDNAs differ from Predicted | 2,289 |
| | |
| Total | 50,144 |

GenBank Accession #s 49077286 - 49077870

Michael Brent, Charlie Comstock, Laura Langton and Jeltje van Baren

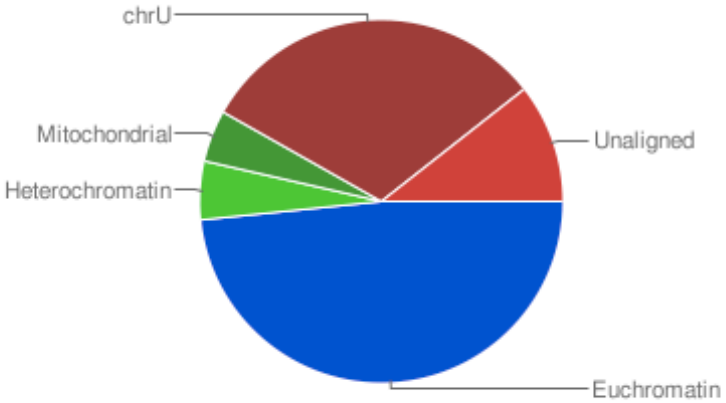
# Solexa sequencing to verify splice sites

## Run 6-6

Whole Fly, Single Stranded cDNA, Oligo-DT Primed, run using a 2 p-mol concentration

| | | Reads |
|---|---|---|
| **Euchromatin** | 48.37% | 536124 |
| **Heterochromatin** | 5.20% | 57626 |
| **Mitochondrial** | 4.61% | 51116 |
| **chrU** | 31.22% | 346076 |
| **Unaligned** | 10.60% | 117462 |
| **Total** | | **1108404** |

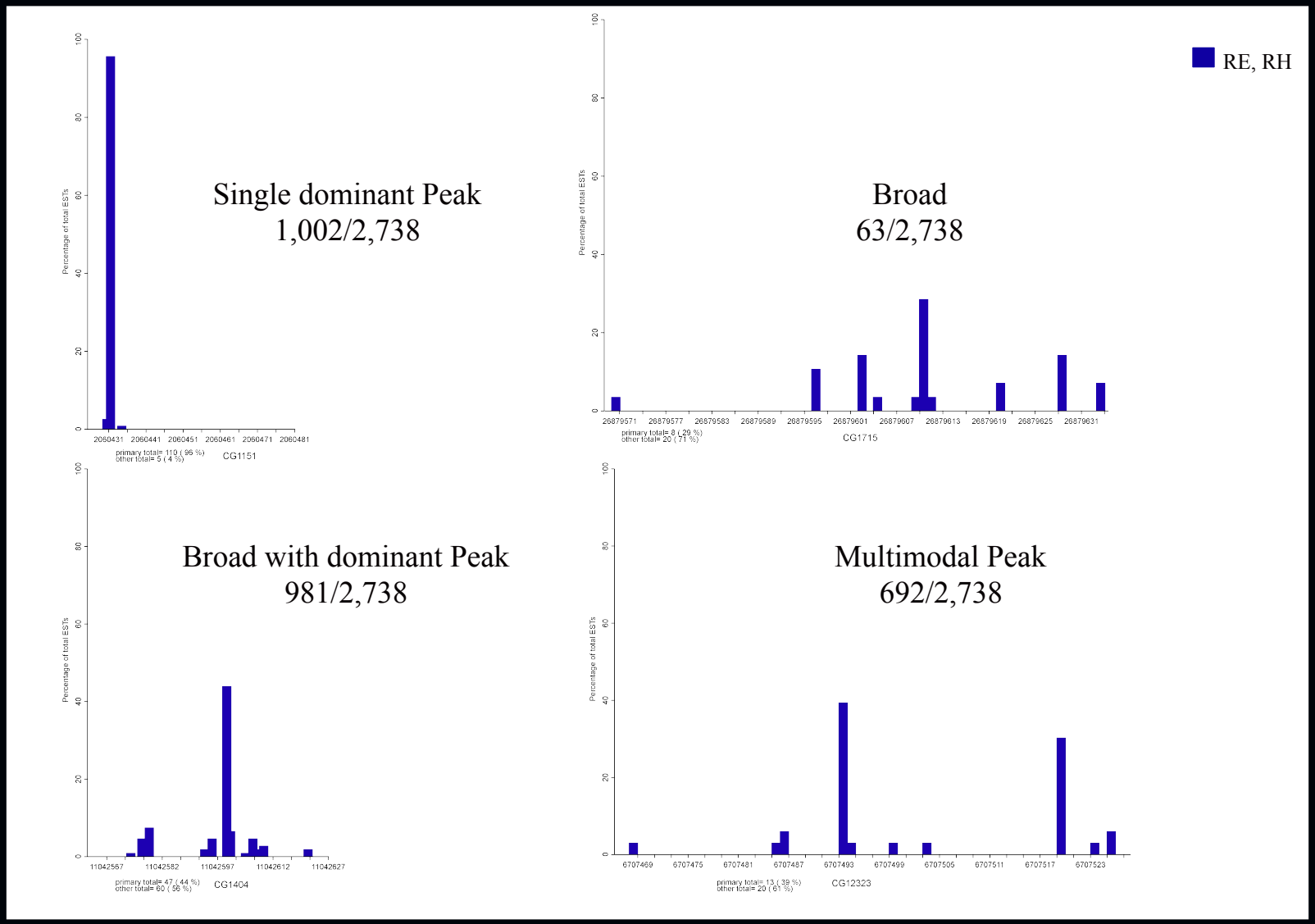
scarf | alignment | unaligned




## Intron Statistics

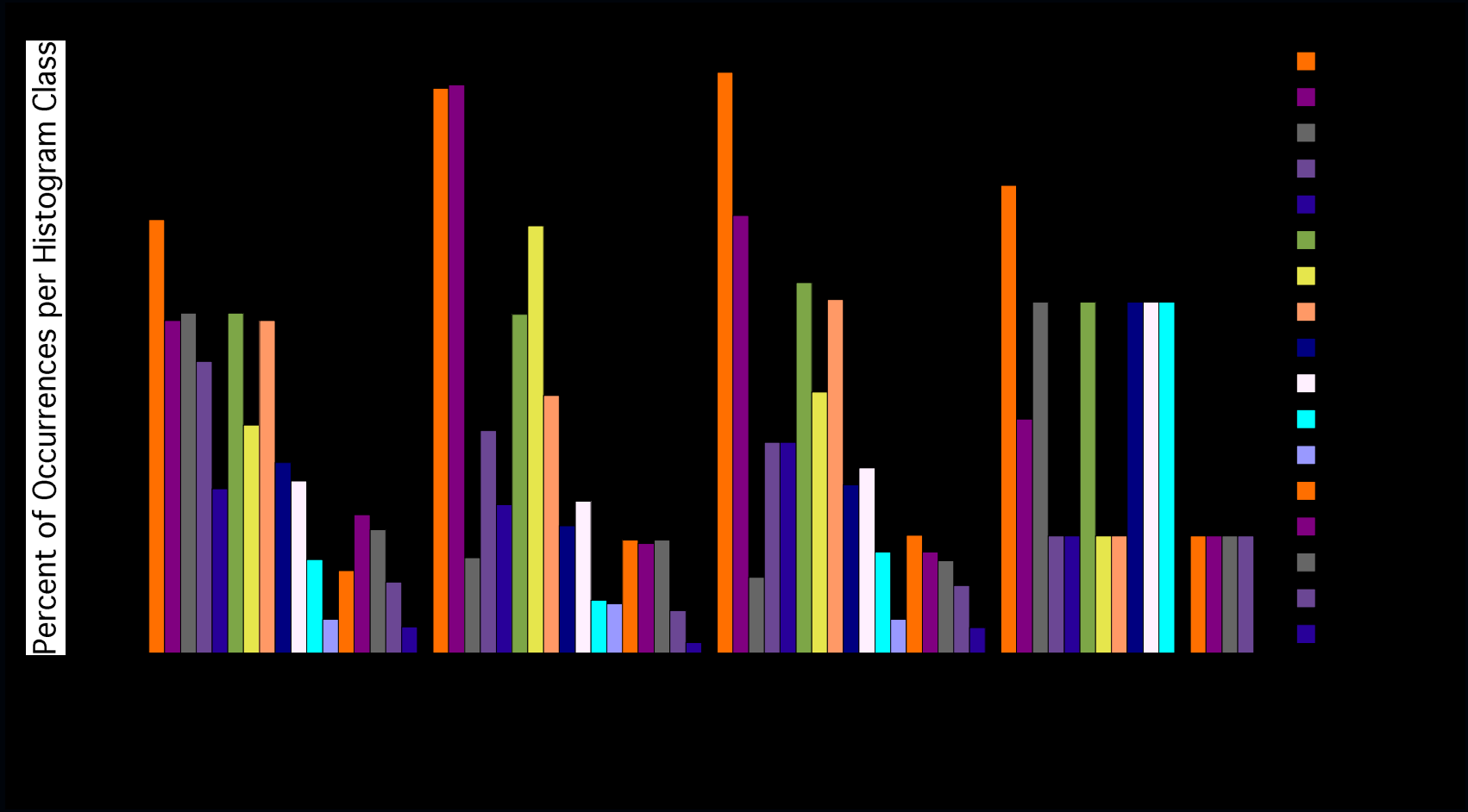
| | DMG2 | DMG3 |
|---|---|---|
| **6-6** | novel: 145/5125<br>alternative: 146/2416<br>known: 7742/42603<br>total: 8033/50144 | novel: 169/3329<br>alternative: 78/1225<br>known: 7653/41299<br>total: 7900/45853 |

# Analysis of transcription start sites

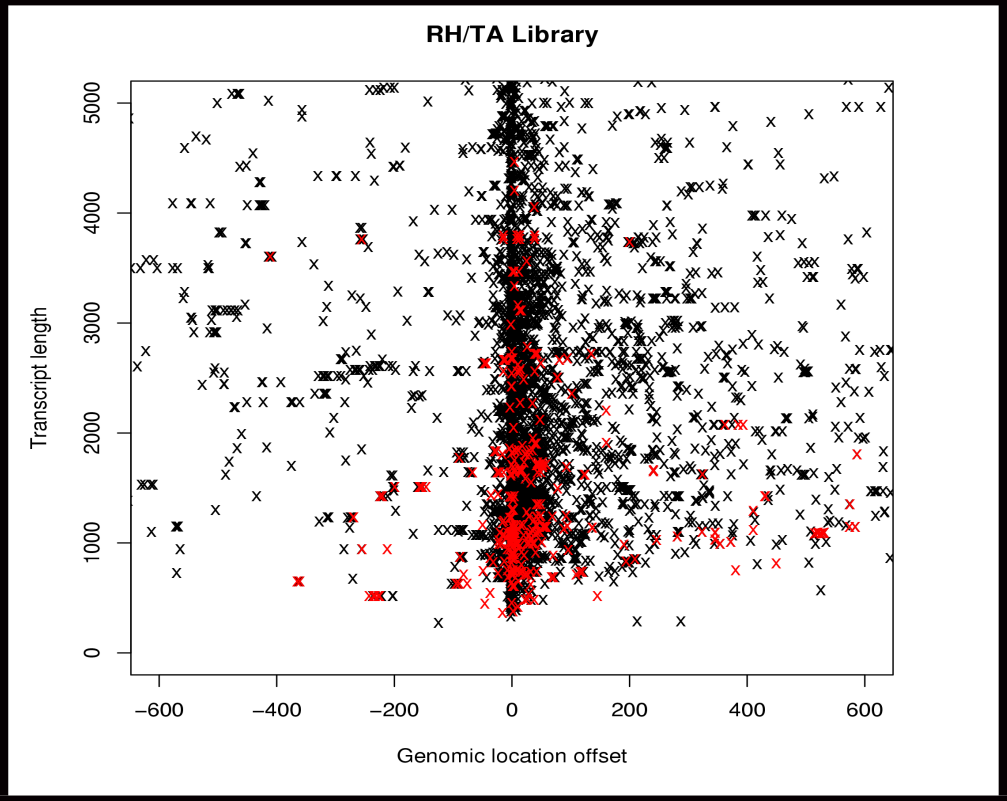


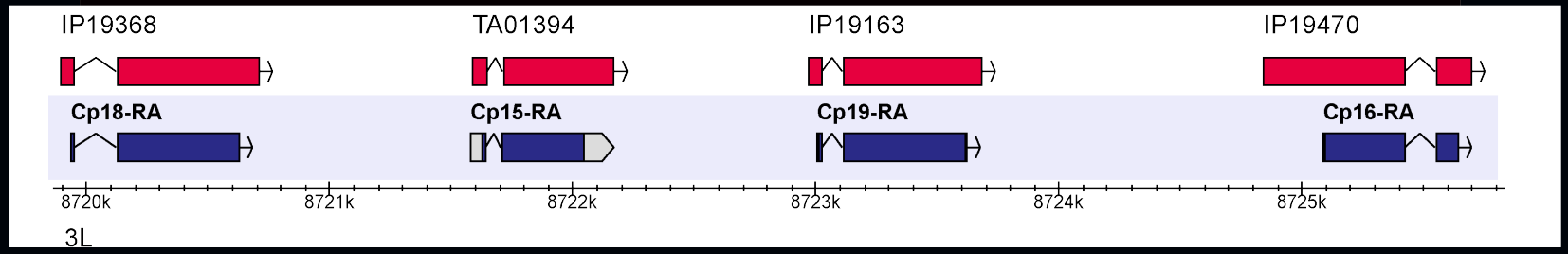

**Ben Booth and Joseph Carlson, LBNL; Classification Carninci et al., '06**

# Motif Abundance



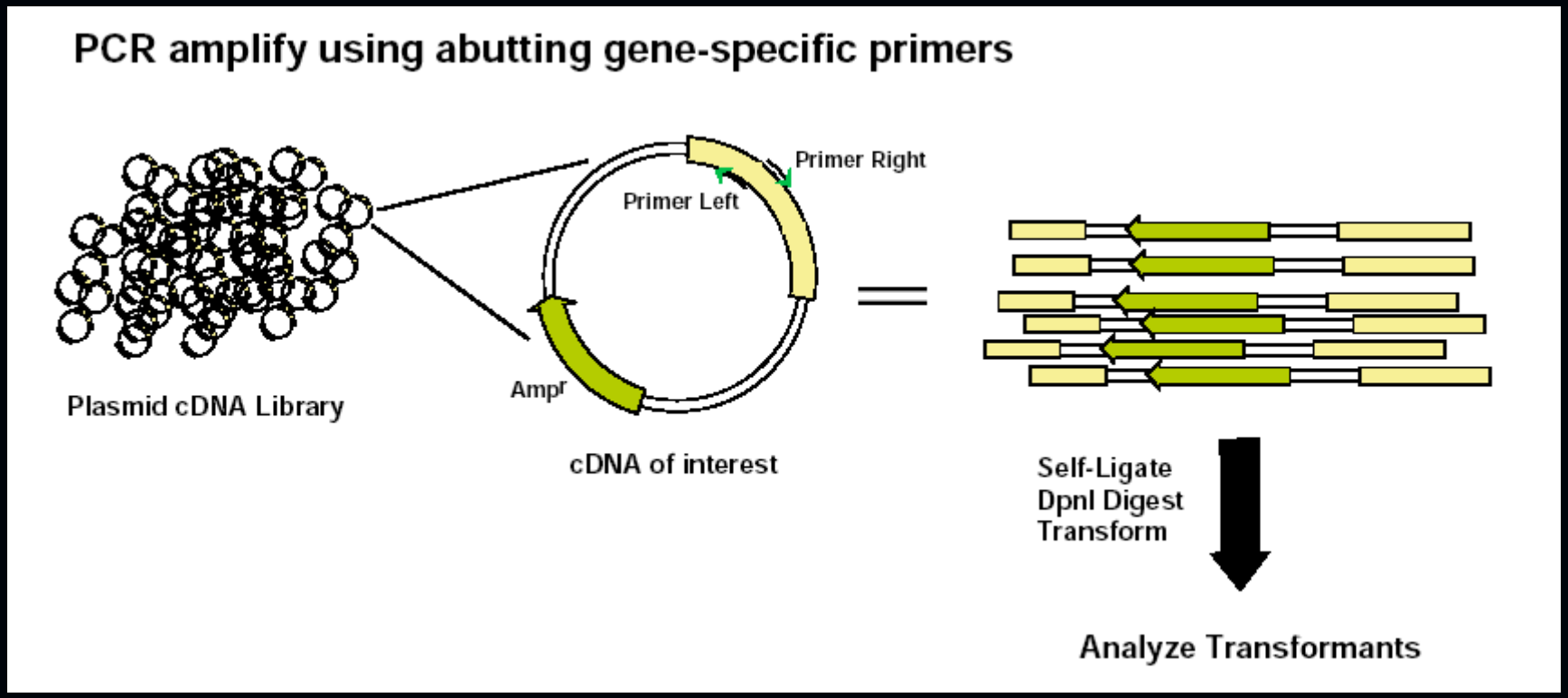
# RLM cDNA Library




Charles Yu, Roger Hoskins and Joseph Carlson, LBNL

# cDNA Library Screening Using iPCR




**Advantages over RT PCR**

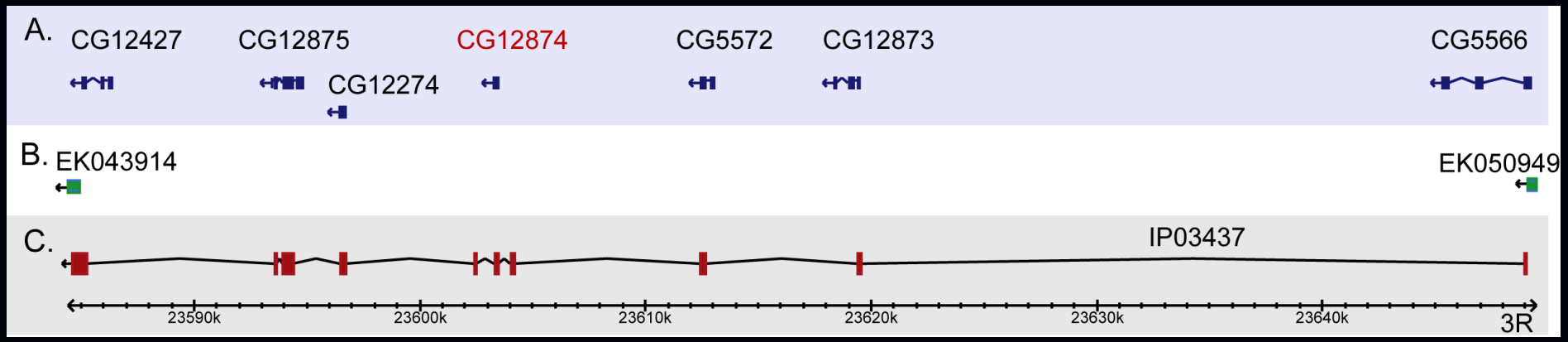
- Captures 5' and 3' UTRs
- Captures splice variants
- Extends predictions

**Summary**

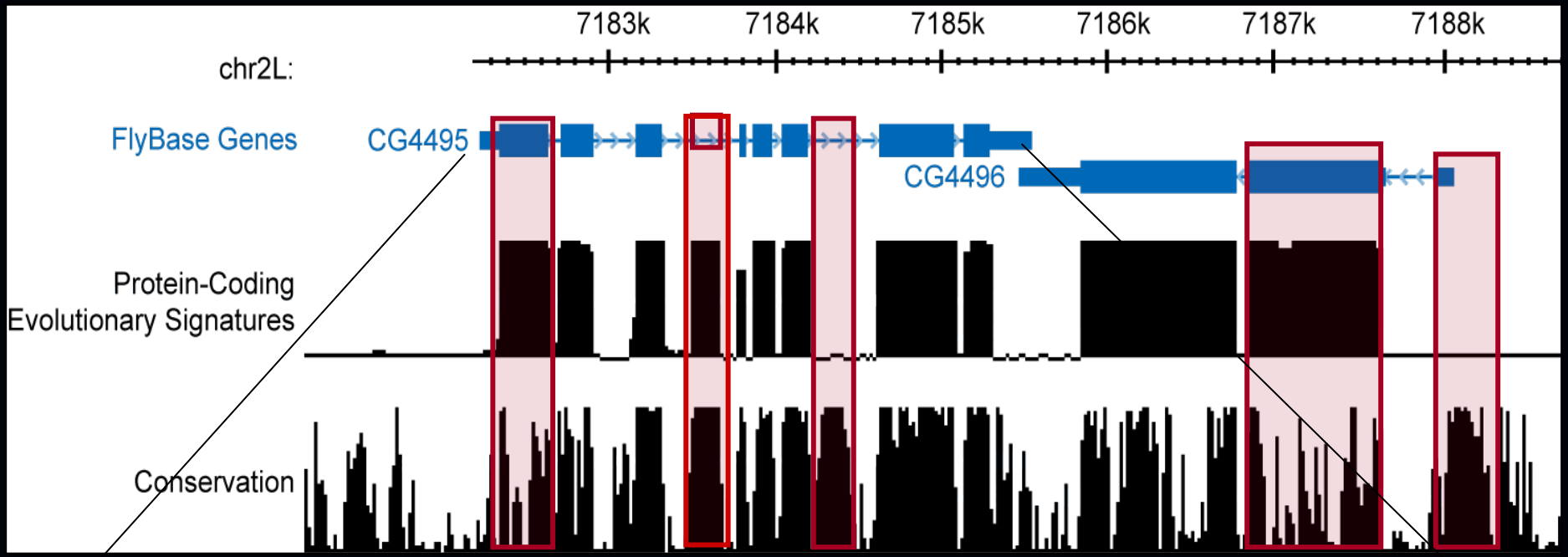
| | |
|---|---|
| • Attempts | 3,829 |
| • Recovered | 2,047 |
| • Success rate | 53% |

Hoskins *et al.*, (2005) *NAR* 33(21):e185
Wan *et al.*, (2006) *Nat Proto* 1:624

# cDNA Sequencing Corrects Gene Models

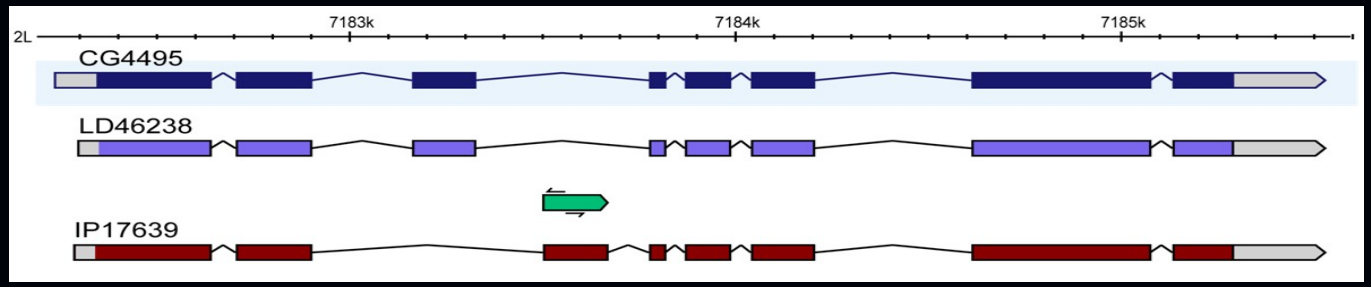


# Power of Evolutionary Signatures for Exon Identification




High protein-coding signal, low conservation
Ability to recognize fast-evolving exons

High conservation, but not protein-coding
Evolutionary signatures specific to function




Collaboration with Manolis Kellis

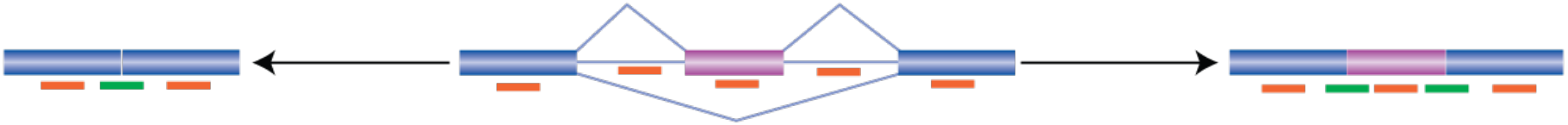
Stark et al, Nature 2007 450:219: Lin et al., Genome Res. 2007

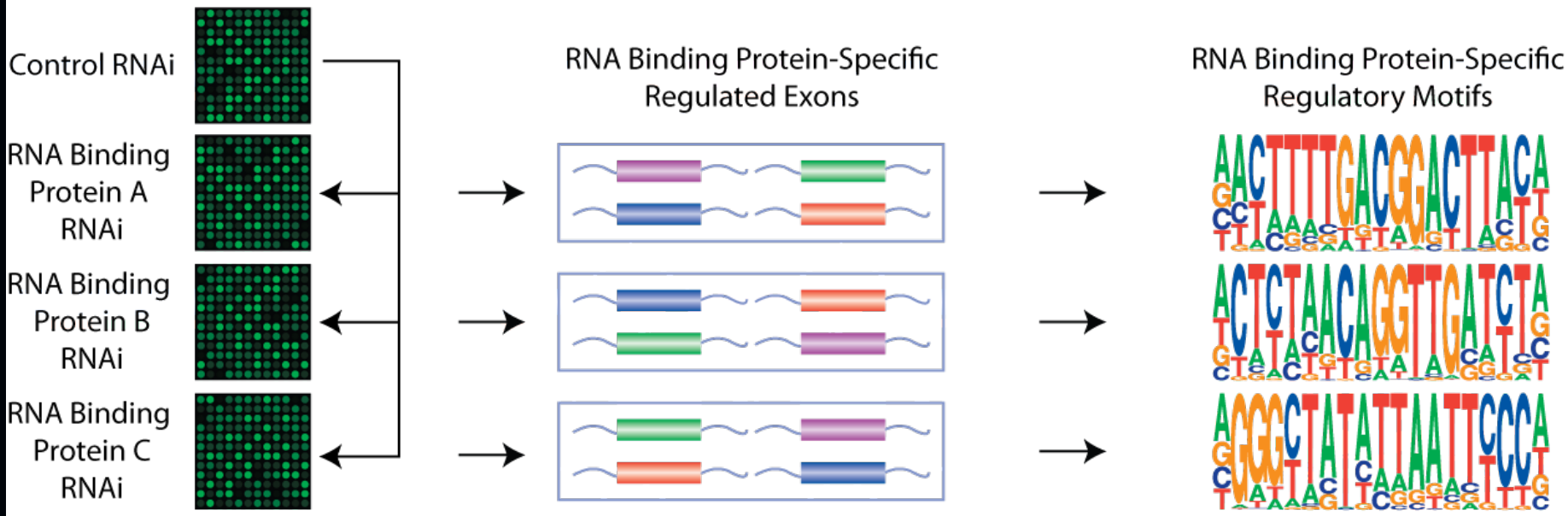
# Validation of the Transcriptome

| | Transcript Statistics | | | | | Status of Experimental Validation* | | |
|---|---|---|---|---|---|---|---|---|
| | Transcripts | Single Exon Transcripts | Multiple Exon Transcripts | Exons of multiple exon transcripts | Introns | Introns Spanned by cDNAs | Introns Spanned by ESTs | Unconfirmed Introns |
| Release 5.2 | 21,660 | 3,553 | 18,107 | 64,169 | 50,594 | 34,353 | 6,325 | 9,916 |
| | | | | | | | | |
| Release 5.5 | 21,853 | 3,536 | 18,313 | 64,445 | 50,766 | 38,322 | 5,754 | 6,690 |

* Comparison of FlyBase Release 5.2, 5.5 Annotations and BDGP and Exelixis ESTs, BDGP cDNA and modENCODE RT-PCR data

# Custom Microarray with Exon, Intron, and Splice-Junction Probes





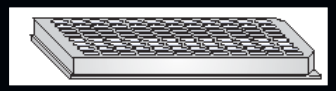

Brenton Graveley (UConn Health Center), Steven Brenner (UC Berkeley), Sandrine Dudoit (UC Berkeley)

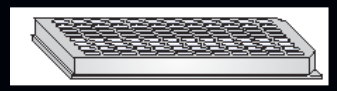
# Plans for demonstrating biological relevance of ncRNAs RNAi screens

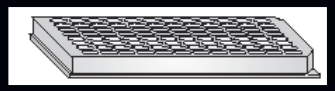
DRSC dsRNAs arrayed in 384-well plates

Plate reader-based assays

Microscopy-base assays



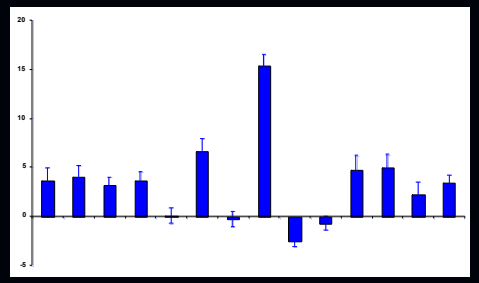


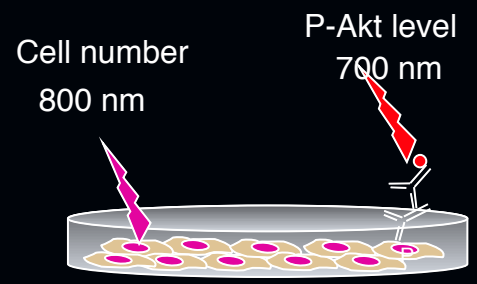


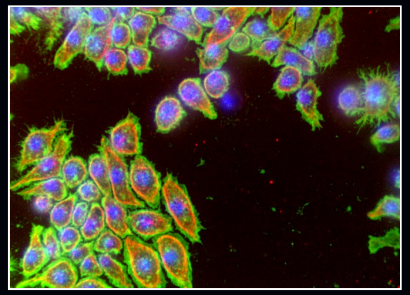
Transcriptional-Luciferase Reporter Assays

Protein modification (phospho-specific antibodies) (Aerius)

GFP or antibodies





**DRSC: Drosophila RNAi Screening Center, Harvard Medical School**
**http://flyrnai.org/ - Mathey-Prevot and Perrimon**

# Acknowledgements


## modENCODE Drosophila Transcriptome Project

- **UCB:** Angela N. Brooks, Kasper D. Hansen, Sandrine Dudoit and Steven E. Brenner
- **LBNL:** Roger Hoskins, Ann S. Hammonds, Joseph W. Carlson, Kenneth H. Wan, Charles Yu and Benjamin Booth
- **IU:** Peter Cherbas, Justen Andrews, Lucy Cherbas, Dayu Zhang, David Miller, Andreas Rechsteiner, Thomas C. Kaufman and Justin P. Kumar
- **WashU:** Laura Langton, Marijke J. van Baren, Aaron E. Tenney, Charles L. G. Comstock and Michael Brent
- **Affymetrix and CSH:** Aarron T. Willingham, Philipp Kapranov, Srinka Ghosh and Thomas R.Gingeras
- **UCHC:** Michael O. Duff, Li Yang, and Brenton R. Graveley
- **Harvard:** Norbert Perrimon, Stephanie Mohr and Bernard Mathey-Prevot
- **Funding: modENCODE NHGRI, expression NHGMS**