Conserved Introns Reveal Novel Transcripts in *Drosophila melanogaster*

Dominic Rose
Bioinformatics Group, University of Leipzig

Bled, Feb 2009

Outline

Outline

- Genome-wide comparative genomics approach

Outline

- Genome-wide comparative genomics approach
- Search for short conserved introns in insect genomes

Outline

- Genome-wide comparative genomics approach
- Search for short conserved introns in insect genomes
- Capable to identify conserved transcripts

Outline

- Genome-wide comparative genomics approach
- Search for short conserved introns in insect genomes
- Capable to identify conserved transcripts
- Novel conserved introns → novel cons. transcripts

Outline

- Genome-wide comparative genomics approach
- Search for short conserved introns in insect genomes
- Capable to identify conserved transcripts
- Novel conserved introns → novel cons. transcripts
- Intron detection allows to
    - ...extend annotation of existing coding or UTRs

Outline

- Genome-wide comparative genomics approach
- Search for short conserved introns in insect genomes
- Capable to identify conserved transcripts
- Novel conserved introns → novel cons. transcripts
- Intron detection allows to
  - ...extend annotation of existing coding or UTRs
  - ...identify novel protein coding genes

Outline

- Genome-wide comparative genomics approach
- Search for short conserved introns in insect genomes
- Capable to identify conserved transcripts
- Novel conserved introns → novel cons. transcripts
- Intron detection allows to
  - ...extend annotation of existing coding or UTRs
  - ...identify novel protein coding genes
  - ...identify novel mRNA-like ncRNAs

mRNA-like noncoding RNAs (mlncRNAs)

mRNA-like noncoding RNAs (mlncRNAs)

- Central topic of current RNA research

mRNA-like noncoding RNAs (mlncRNAs)

- Central topic of current RNA research
- ENCODE: Large portion of the transcriptional output of eukaryotic genomes consists of mRNA-like noncoding RNAs.

mRNA-like noncoding RNAs (mlncRNAs)

- Central topic of current RNA research
- ENCODE: Large portion of the transcriptional output of eukaryotic genomes consists of mRNA-like noncoding RNAs.
- Capped, polyadenylated, often (alternatively) spliced (just like protein-coding genes), but lack discernible open reading frames

mRNA-like noncoding RNAs (mlncRNAs)

mRNA-like noncoding RNAs (mlncRNAs)

- Gene regulators: Evf-2, Xist, roX1, Tsix, XistAS, roX2, H19, mei, LPW, KvDMR1, DGCR5, CMPD

  (e.g. Evf-2 acts as transciptional enhancer for distal-less homeobox genes)

mRNA-like noncoding RNAs (mlncRNAs)

- Gene regulators: Evf-2, Xist, roX1, Tsix, XistAS, roX2, H19, mei, LPW, KvDMR1, DGCR5, CMPD

  (e.g. Evf-2 acts as transciptional enhancer for distal-less homeobox genes)
- Some serve as precursor for miRNAs and snoRNAs

mRNA-like noncoding RNAs (mlncRNAs)

- Gene regulators: Evf-2, Xist, roX1, Tsix, XistAS, roX2, H19, mei, LPW, KvDMR1, DGCR5, CMPD

  (e.g. Evf-2 acts as transciptional enhancer for distal-less homeobox genes)
- Some serve as precursor for miRNAs and snoRNAs
- Abiotic stress signals: gadd7/adapt15, adapt33, hsrω, OxyR, DsrA, lbi, G90

  (e.g. expression caused by UV radiation)

mRNA-like noncoding RNAs (mlncRNAs)

- Gene regulators: Evf-2, Xist, roX1, Tsix, XistAS, roX2, H19, mei, LPW, KvDMR1, DGCR5, CMPD

  (e.g. Evf-2 acts as transciptional enhancer for distal-less homeobox genes)

- Some serve as precursor for miRNAs and snoRNAs
- Abiotic stress signals: gadd7/adapt15, adapt33, hsrω, OxyR, DsrA, lbi, G90

  (e.g. expression caused by UV radiation)

- Biotic stress signals: His-1, ENOD40, CR20, GUT15

  (e.g. expression correlated with viral insertion or carcinogenesis)

mRNA-like noncoding RNAs (mlncRNAs)

- Gene regulators: Evf-2, Xist, roX1, Tsix, XistAS, roX2, H19, mei, LPW, KvDMR1, DGCR5, CMPD

  (e.g. Evf-2 acts as transciptional enhancer for distal-less homeobox genes)
- Some serve as precursor for miRNAs and snoRNAs
- Abiotic stress signals: gadd7/adapt15, adapt33, hsrω, OxyR, DsrA, Ibi, G90

  (e.g. expression caused by UV radiation)
- Biotic stress signals: His-1, ENOD40, CR20, GUT15

  (e.g. expression correlated with viral insertion or carcinogenesis)
- Others: UHG, NTT, Bsr, BC1, BC200, SRA

mRNA-like noncoding RNAs (mlncRNAs)

- Gene regulators: Evf-2, Xist, roX1, Tsix, XistAS, roX2, H19, mei, LPW, KvDMR1, DGCR5, CMPD

  (e.g. Evf-2 acts as transciptional enhancer for distal-less homeobox genes)

- Some serve as precursor for miRNAs and snoRNAs
- Abiotic stress signals: gadd7/adapt15, adapt33, hsrω, OxyR, DsrA, lbi, G90

  (e.g. expression caused by UV radiation)

- Biotic stress signals: His-1, ENOD40, CR20, GUT15

  (e.g. expression correlated with viral insertion or carcinogenesis)

- Others: UHG, NTT, Bsr, BC1, BC200, SRA

$\longrightarrow$ functionally important ncRNA class

The idea

Functional pair of donor (5’) and acceptor (3’) splice sites will be retained over long evolutionary time scales only if

The idea

Functional pair of donor (5’) and acceptor (3’) splice sites will be retained over long evolutionary time scales only if

1. The locus is transcribed into a functional transcript

The idea

Functional pair of donor (5’) and acceptor (3’) splice sites will be retained over long evolutionary time scales only if

1. The locus is transcribed into a functional transcript
2. Accurate intron removal is necessary to produce a functional transcript

The idea

Functional pair of donor (5’) and acceptor (3’) splice sites will be retained over long evolutionary time scales only if

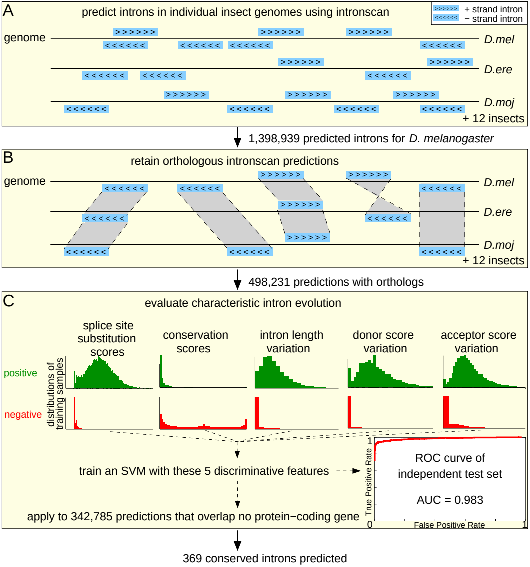
1. The locus is transcribed into a functional transcript
2. Accurate intron removal is necessary to produce a functional transcript

→ Find the intron → it guides you to your novel transcript

The data

12 drosophila genomes (fly)
+ *Anopheles gambiae* (mosquito)
+ *Tribolium castaneum* (beetle)
+ *Apis melifera* (honeybee)

The method



A predict introns in individual insect genomes using intronscan

1,398,939 predicted introns for *D. melanogaster*

B retain orthologous intronscan predictions

498,231 predictions with orthologs

C evaluate characteristic intron evolution

train an SVM with these 5 discriminative features

apply to 342,785 predictions that overlap no protein−coding gene

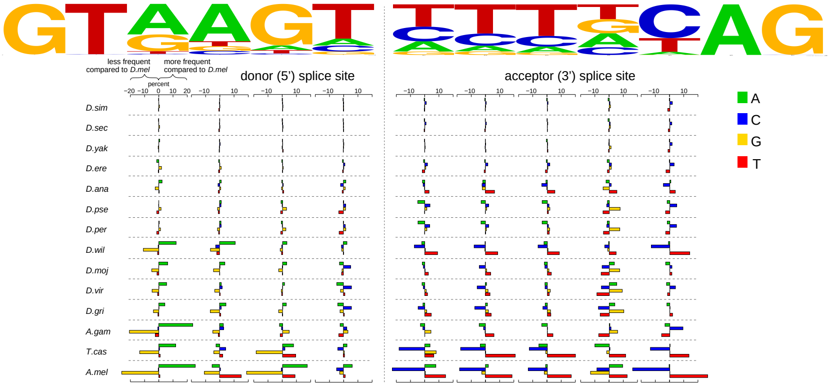
369 conserved introns predicted

intronscan

alignments

SVM

Nucleotide frequencies in SS positions differ



less frequent ← | → more frequent
(compared to Dmel)

e.g. Apis prefers A over G (donor +3) and T over C (accepor -3)

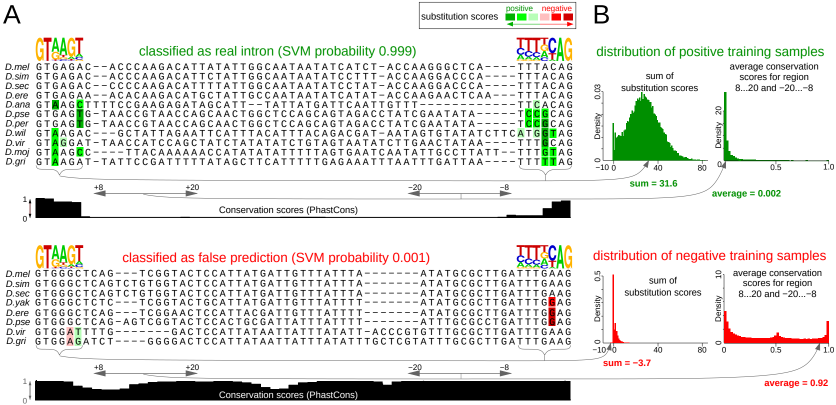
Learn log-odd substitution scores

$$log_2\left(\frac{freq_{\mathrm{pos}}(x,y)}{freq_{\mathrm{neg}}(x,y)}\right) \rightarrow \text{substitution matrix}$$

$$\forall x, y \in \{A, T, C, G\}$$

$$x \neq y$$

Evaluating intron evolution - an example



classified as real intron (SVM probability 0.999)

```
D.mel  GTGAGAC--ACCCAAGACATTATATTGGCAATAATATCATCT-ACCAAGGGCTCA----TTTACAG
D.sim  GTGAGAC--ACCCAAGACATTCTATTGGCAATAATATCCTTT-ACCAAGGACCCA----TTTACAG
D.sec  GTGAGAC--CCCCAAGACATTTTATTGGCAATAATATCCTAT-ACCAAGGACCCA----TTTACAG
D.ere  GTGAGAA--ACACAAGACATGCTATTGCCAATAATATCATAT-ACCAAGAACTCAA---TTTACAG
D.ana  GTAAGCTTTTCCGAAGAGATAGCATT--TATTATGATTCAATTGTTT------------TTCACAG
D.pse  GTGAGTG-TAACCGTAACCAGCAACTGGCTCCAGCAGTAGACCTATCGAATATA-----TCCGCAG
D.per  GTGAGTG-TAACCGTAACCAGCAACTGGCTCCAGCAGTAGACCTATCGAATATA-----TCCGCAG
D.wil  GTAAGAC---GCTATTAGAATTCATTTACATTTACAGACGAT-AATAGTGTATATCTTCATGGTAG
D.vir  GTAGGAT-TAACCATCCAGCTATCTATATATCTGTAGTAATATCTTGAACTATAA----TTTGCAG
D.moj  GTAAGCC----TTACAAAAAACCATATATATTTTTAGTGAATCAATATTGCCTTATT--TTTGTAG
D.gri  GTAAGAT-TATTCCGATTTTTATAGCTTCATTTTTGAGAAATTTAATTTGATTAA----TTTTTAG
```

Conservation scores (PhastCons)

sum = 31.6

average = 0.002

classified as false prediction (SVM probability 0.001)

```
D.mel  GTGGGCTCAG---TCGGTACTCCATTATGATTGTTTATTTA-------ATATGCGCTTGATTTGAAG
D.sim  GTGGGCTCAGTCTGTGGTACTCCATTATGATTGTTTATTTA-------ATATGCGCTTGATTTGAAG
D.sec  GTGGGCTCAGTCTGTGGTACTCCATTATGATTGTTTATTTA-------ATATGCGCTTGATTTGAAG
D.yak  GTGGGCTCTC---TCGGTACTGCATTATGATTGTTTATTTT-------ATATGCGCTTGATTTGGAG
D.ere  GTGGGCTCAG---TCGGAACTCCATTACGATTGTTTATTTT-------ATATGCGCTTGATTTGGAG
D.pse  GTGGGCTCAG-AGTCGGTACTCCACTGCGATTATTTATTTT-------ATTTGCGCCTGATTTGGAG
D.vir  GTGGATTTTG-------GACTCCATTATAATTATTTATATT-ACCCGTGTTTGCGCTTGATTTGAAG
D.gri  GTGGAGATCT----GGGGACTCCATTATAATTATTTATATTTGCTCGTATTTGCGCTTGATTTGAAG
```

Conservation scores (PhastCons)

distribution of positive training samples

distribution of negative training samples

sum = −3.7

average = 0.92

Results

Results

- intronscan: $\sim$1.4 Mio introns in Dmel

Results

- intronscan: $\sim$1.4 Mio introns in Dmel
- alignments: 498k loci

Results

- intronscan: $\sim$1.4 Mio introns in Dmel
- alignments: 498k loci
- 155.5k overlap annotated protein-coding transcripts

Results

- intronscan: $\sim$1.4 Mio introns in Dmel
- alignments: 498k loci
- 155.5k overlap annotated protein-coding transcripts
- Two SS: 23.5k (positive sample), one SS: 14.5k (ommited), no SS: 117.5k (negative samples)

Results

- intronscan: $\sim$1.4 Mio introns in Dmel
- alignments: 498k loci
- 155.5k overlap annotated protein-coding transcripts
- Two SS: 23.5k (positive sample), one SS: 14.5k (ommited), no SS: 117.5k (negative samples)
- SVM training: 95%

Results

- intronscan: $\sim$1.4 Mio introns in Dmel
- alignments: 498k loci
- 155.5k overlap annotated protein-coding transcripts
- Two SS: 23.5k (positive sample), one SS: 14.5k (ommited), no SS: 117.5k (negative samples)
- SVM training: 95%
- SVM testing: 5%

Results

- intronscan: $\sim$1.4 Mio introns in Dmel
- alignments: 498k loci
- 155.5k overlap annotated protein-coding transcripts
- Two SS: 23.5k (positive sample), one SS: 14.5k (ommited), no SS: 117.5k (negative samples)
- SVM training: 95%
- SVM testing: 5%
- area under ROC: 0.983

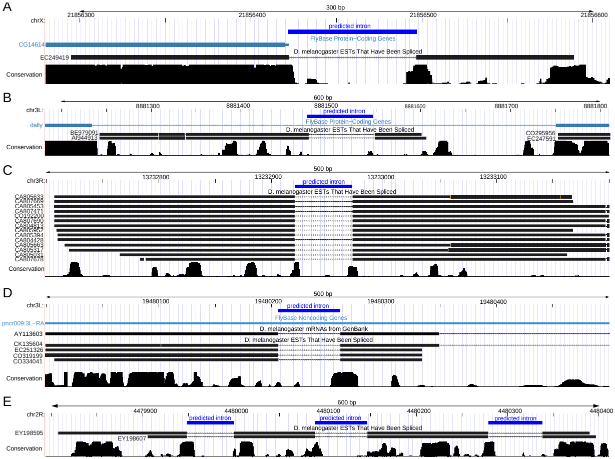
Results

- intronscan: $\sim$1.4 Mio introns in Dmel
- alignments: 498k loci
- 155.5k overlap annotated protein-coding transcripts
- Two SS: 23.5k (positive sample), one SS: 14.5k (ommited), no SS: 117.5k (negative samples)
- SVM training: 95%
- SVM testing: 5%
- area under ROC: 0.983
- p$>$0.95: 80% true positives at 0.12% false positives

Results

- intronscan: $\sim$1.4 Mio introns in Dmel
- alignments: 498k loci
- 155.5k overlap annotated protein-coding transcripts
- Two SS: 23.5k (positive sample), one SS: 14.5k (ommited), no SS: 117.5k (negative samples)
- SVM training: 95%
- SVM testing: 5%
- area under ROC: 0.983
- p$>$0.95: 80% true positives at 0.12% false positives
- p$>$0.99: 72% true positives at 0.07% false positives (4 FP, manual inspection: 3 are true introns → 1 FP)

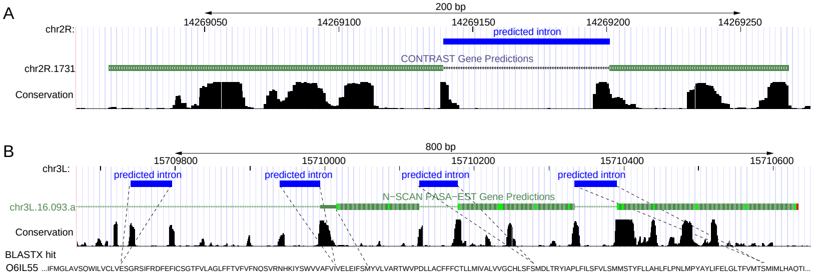
Novel spliced transcripts



369 predictions outside of protein-cod. genes ($p > 0.95$)

131 EST/FlyBase-transcript confirmed introns, 238 unconfirmed

Novel protein-coding genes



A) CONTRAST predicted coding gene, B) NSCAN coding gene

- 20/238 located within 100nt upstream of cod. genes
- 14/20 no annotated 5’UTR

  (in contrast to 77/218, Fischer’s exact test, p=0.005)
- 23 extend CDS, 30 belong to novel CDS

Novel spliced non-coding RNAs

Novel spliced non-coding RNAs

- remove everything protein-coding

Novel spliced non-coding RNAs

- remove everything protein-coding
- remove repeats

Novel spliced non-coding RNAs

- remove everything protein-coding
- remove repeats

→ Heureka! You’ve found mlncRNAs.

Novel spliced non-coding RNAs

- remove everything protein-coding
- remove repeats

→ Heureka! You’ve found mlncRNAs.

- 129 *bona fide* mlncRNAs

Novel mRNA-like noncoding RNAs

Novel mRNA-like noncoding RNAs

- 29/129 have predicted orthologous introns outside Sophophora subgenus (*D. virilis*, *D. mojavensis*, *D. grimshawi*)
  → conserved exon-intron structure over 63 My years

Novel mRNA-like noncoding RNAs

- 29/129 have predicted orthologous introns outside Sophophora subgenus (*D. virilis*, *D. mojavensis*, *D. grimshawi*)
  → conserved exon-intron structure over 63 My years
- Mostly unstructured (just 2 transcripts have RNAz hit)

Experimental verification

Experimental verification

- RT-PCR, 5 different develomental stages of Dmel: embryo, larva, pupa, male, female

Experimental verification

- RT-PCR, 5 different develomental stages of Dmel: embryo, larva, pupa, male, female
- 18/29 (62%) experimentally validated:

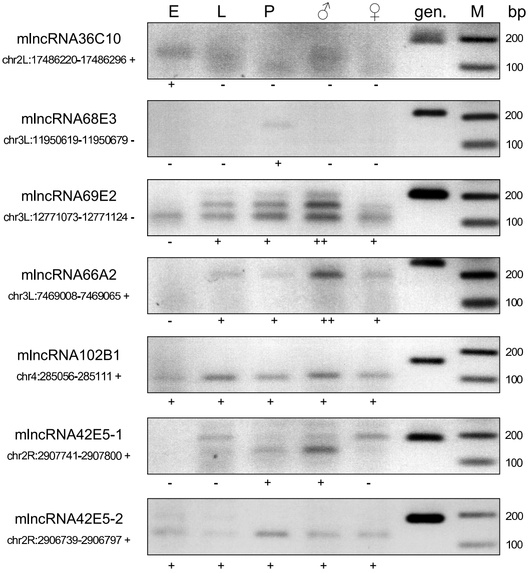
Experimental verification

- RT-PCR, 5 different develomental stages of Dmel: embryo, larva, pupa, male, female
- 18/29 (62%) experimentally validated: mlncRNAs: 7/12

Experimental verification

- RT-PCR, 5 different develomental stages of Dmel: embryo, larva, pupa, male, female
- 18/29 (62%) experimentally validated:
  mlncRNAs: 7/12
  introns in putative coding transcripts: 11/17

Experimental verification of mlncRNAs



Summary

Summary

- Novel method that predicts intron-containing transcripts

Summary

- Novel method that predicts intron-containing transcripts
- We solely use intron information for prediction

Summary

- Novel method that predicts intron-containing transcripts
- We solely use intron information for prediction

We identify novel...

Summary

- Novel method that predicts intron-containing transcripts
- We solely use intron information for prediction

We identify novel...

- transcripts coding for proteins or mlncRNAs

Summary

- Novel method that predicts intron-containing transcripts
- We solely use intron information for prediction

We identify novel...

- transcripts coding for proteins or mlncRNAs
- transcripts without conserved secondary structures

Summary

- Novel method that predicts intron-containing transcripts
- We solely use intron information for prediction

We identify novel...

- transcripts coding for proteins or mlncRNAs
- transcripts without conserved secondary structures
- transcripts with low sequence conservation

Summary

- Novel method that predicts intron-containing transcripts
- We solely use intron information for prediction

We identify novel...

- transcripts coding for proteins or mlncRNAs
- transcripts without conserved secondary structures
- transcripts with low sequence conservation

Limitations: Transcript start, transcript end?

Thank you

Michael Hiller (Stanford)

Leipzig:
Sven Findeiß, Manja Marz, Christine Schulz, Sonja J. Prohaska

Halle:
Sandro Lein, Claudia Nickel, Gunter Reuter

Freiburg:
Rolf Backofen

Various cities, countries, and continents:
Peter F. Stadler