NIP search and analysis in Metazoans

Jörg Lehmann

Bioinformatics Group
Department of Computer Science
University of Leipzig

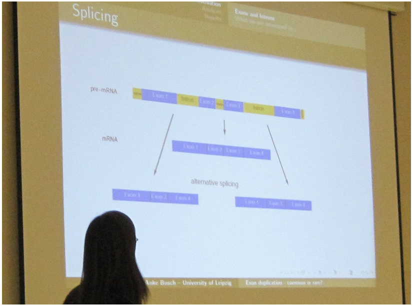
25th TBI Winterseminar in Bled, 2010

Outline

1. **Motivation**
   - The NIP Idea
   - Previous Work
2. **NIPs in Metazoa**
   - Data Compilation and NIP Extraction
   - NIP Distribution Results

Introns? Evolution?
A quick reminder.

- spliceosomal intron positions often conserved across species
- intron positions as genome-level character



(source: Langenberger 2010)

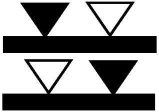
Outline

1. **Motivation**
   - The NIP Idea
   - Previous Work
2. **NIPs in Metazoa**
   - Data Compilation and NIP Extraction
   - NIP Distribution Results

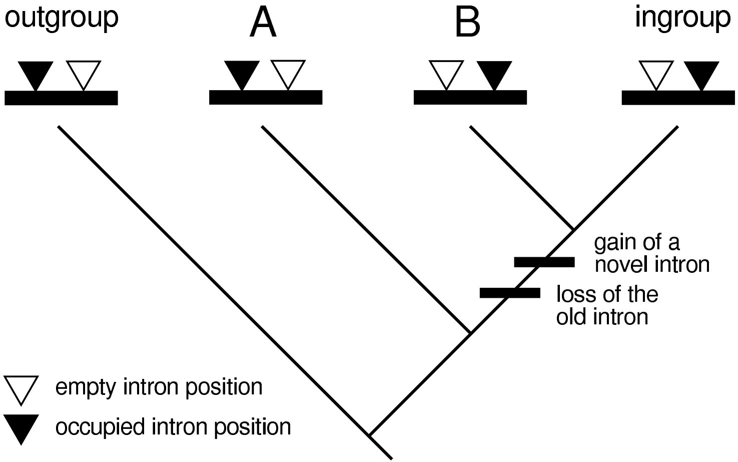
Near Intron Positions (NIPs)

```
                                          IntronPos1                            IntronPos2
ref protein                               ________119-0___________               ________124-0___________
       Dsec GCTGATTCCCCTGCCATCAAGGAGAATCGCgtaagttg//atttattcttgtagGCCTTTGGAGTGCAG                        ACCATTTCCGGCACGGGTGCCCTGCGCATT
       Dsim GCTGATTCCCCTGCCATCAAGGAGAATCGCgtaagttg//atttattcttgtagGCCTTTGGAGTGCAG                        ACCATTTCGGGCACGGGTGCCCTGCGCATT
       Dmel GCCGATTCCCCAGCCATCAAGGAGAATCGCgtaagttg//caattttcttgtagGCCTTTGGAGTGCAG                        ACCATTTCGGGCACGGGTGCCCTGCGCGTT
       Dere GCTGATTCCCCAGCCATCAAGGAGAATCGCgtaagcta//aatttttcttgtagGCTTTTGGAGTGCAG                        ACCATTTCGGGCACAGGTGCCCTGCGTATT
       Dyak GCTGATTCCCCGGCCATCAAGGAGAATCGGgtaagttg//cgattttcttgtagGCCTTTGGAGTGCAG                        ACCATTTCGGGCACGGGTGCCCTGCGTGTT
       Dana AGTGATTCTCCTGCTATTAAGGCCAATCGCgtaagtag//tttttgtctggtagGCTTTTGGAGTTCAG                        ACCATTTCGGGCACTGGTGCTCTGCGCATT
       Dper GCTGACTCCATAGCCATCAAGGAGAATCGCgtaattta//cttcttgttcgcagGCTTTTGGGGTGCAG                        ACTATATCCGGAACAGGTGCTCTGCGCGTG
       Dpse GCTGACTCCATAGCCATCAAGGAGAATCGCgtaattta//cttcttgttcgcagGCTTTTGGGGTGCAG                        ACTATATCCGGAACAGGTGCTCTGCGCGTG
       Dwil GCCGATTCCATAGCTCTTAAGGAGAAACGT                        GCCTTTGGTGTCCAGgtgagttg//gatgcatttttcagACTATATCTGGAACTGGTGCCTTGCGTGTG
       Dgri GCTGATTCCAATGCAATCAAAGAGAAACGAgttagttt//tctgtcttctgcagGCTTTTGGTGTACAA                        ACGATATCTGGAACAGGTGCAATTCGTGTC
       Dvir TCGGAATCAAAAGCATTAAAAGAGAAGCGTgtaagttt//ttattatttgcagGCTTTTGGTGTGCAA                         ACTATATCCGGCACAGGTGCGCTGCGTGTA
       Dmoj GCGGACTCAATAGCATTAAAAGAGAAACGTgtaagttt//tttttatattgcagGCCTTTGGTGTGCAA                        ACGATTTCTGGAACAGGAGCTCTGCGCGTA
```

<50nt



Phylogenetic Inference

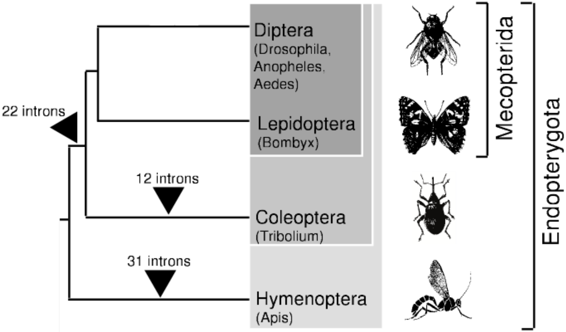


Outline

1. **Motivation**
   - The NIP Idea
   - Previous Work
2. **NIPs in Metazoa**
   - Data Compilation and NIP Extraction
   - NIP Distribution Results

NIPs in Holometabolous Insects.

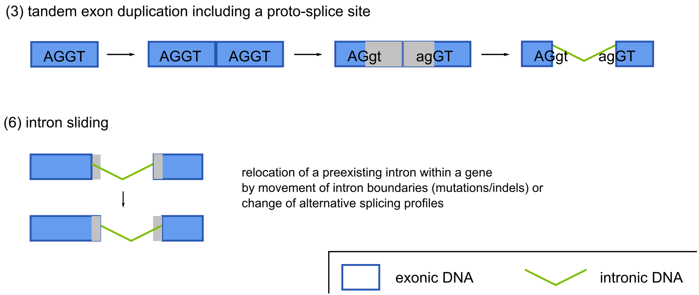
- NIP as phylogenetic marker (hypothesis testing)
- beetles more closely related to (butter-)flies than bees/wasps.



Krauss et al.: Near intron positions are reliable phylogenetic markers: an application to holometabolous insects. *Mol Biol Evol*, 25:821-830, 2008.

Novel Introns in Drosophila.

- NIPs to identify recent intron gain and its mechanisms
- intron sliding and tandem duplication are causes for novel introns in *Drosophila*

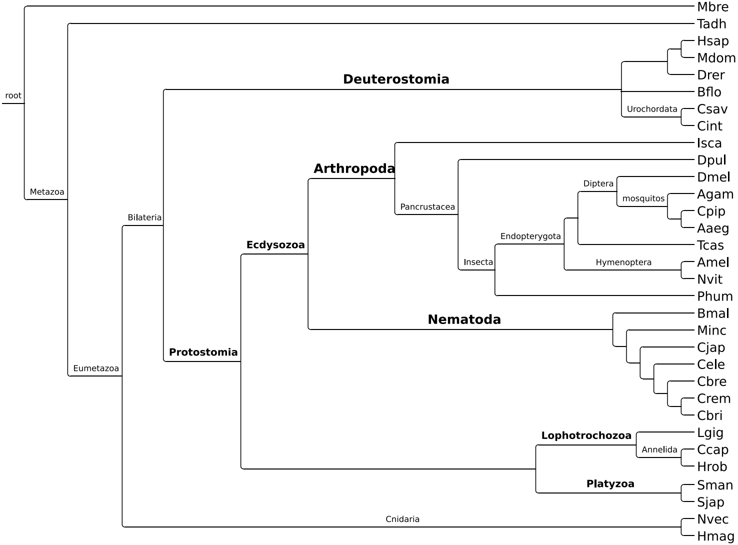


(3) tandem exon duplication including a proto-splice site

(6) intron sliding

relocation of a preexisting intron within a gene
by movement of intron boundaries (mutations/indels) or
change of alternative splicing profiles

Metazoan Species Overview.
Supposed Cladogram.



Outline

1. **Motivation**
   - The NIP Idea
   - Previous Work

2. **NIPs in Metazoa**
   - Data Compilation and NIP Extraction
   - NIP Distribution Results

Compilation of Orthologous Genes.

1. start: all *C. elegans* protein-coding genes with *D. melanogaster* 1:1 ortholog (`Ensembl/BioMart`) [**3331** genes]
2. retrieve `EnsemblCompara` orthologs [13 species]
3. `BLAST`-based ortholog annotation and CDS extraction [$>$ 28 genomes]
4. exclusion of some species (too few/no protein data available)

NIP Pipeline

1. Multiple codon alignment (`Muscle`, `transAlign.pl`, `...`)
2. Extraction of alignment regions containing NIPs of distance $<$ 32 nt (out: 45890 NIP regions from 3276 genes)
3. Filtering steps. Require:
   - amino-acid conservation score $\geq$ 0.75
   - $\geq$ 6 non-gap characters around introns
   - strict splice site consensus (`g,ct,ag:act,a,g`)

Final Dataset

- **5616 NIPs** (from 2589 aln regions and 1220 genes)
- within **32 species**
- (1349 parsimony-informative NIPs)

Metazoan NIPs: A Nice Example Alignment.

```
(Dmel protein)                               ________149-1__________          ________152-1__________
          Mbre ATGCCATGCTGGACGCCCCCGTATCTGGCGgttcgtag//tcgatgcgtggcagGTGTCGGAG                        GCGCTGAAGCTGGCACTCTGACCTTTATGG
          Tadh GTACTTTTATGGACGCTCCAGTTTCTGGAGgtacaact//attcccgtaaatagGAGTTGTAG                        CGGCTCGCGATGCGTTATTAACATTTATGG
          Hsap CAGTTTTCATGGATGCCCCTGTTTCTGGTGgtaagtgg//cccaaatgattcagGTGTAGGAG                        CTGCACGATCTGGGAACCTCACGTTTATGG
          Mdom CACTTTTCATGGATGCCCCTGTTTCTGGTGgtgagtaa//atctttttctccagGGGTGGGAG                        CTGCTCGAGCTGGGAACCTGACATTCATGG
         Mdom1 CACTTTTCCTGGATGTCCCTGTTTCTGGTG                        GG-------                        --------CGTGGGAACCTGGCATTCATGG
          Drer CTGTTTTCATGGATGCCCCGGTGTCTGGAGgtgagatg//ttaatcttcctcagGTGTTGGTG                        CTGCCACTTCTGGTAAGCTCACTTTTATGG
         Drer1 CAGTGTTCATGGACGCGCCGGTGTCAGGAGgtatgatg//tgtgttctgtctagGTGTTGGTG                        CAGCCAGTTTGGCTAAACTCACTTTCTTGG
          Bflo GAGTCTACATGGACGCACCTGTGTCTGGGGgtatggta//tctacacattacagGTATTAAGG                        CAGCTGCTGCAGCCACACTAACCTTCATGG
          Csav CCAAGTATCTGGATGCACCAGTTTCTGGAGgtttgcat//attgttctacacagGTGTGGGTG                        CCGCTCAAGCAGGAACTTTAACTTTCATGG
          Isca --------------------CTTTTACGGGgtatcacc//cttttatattctagGTGTGAATG                        CAGCCAAGGCTGGGACACTTACATTCATGG
          Dmel CTCGGTTTATCGATGCCCCCGTTTCCGGCG                        GAGTTCCCG                        GCGCCGAGCAGGCCACCCTCACCTTCATGG
          Agam CAACGTTCGTCGATGCGCCCGTGTCCGGCG                        GTGTGCCCG                        GTGCAAAGAACGCGACGCTCACGTTCATGG
          Aaeg CTACCTTCGTCGATGCCCCGGTTTCCGGTG                        GTGTTCCCG                        GAGCTCAGAATGCTACTCTGACCTTCATGG
          Cpip CGACCTTTGTCGATGCGCCCGTTTCCGGTG                        GAGTTCCCG                        GAGCTAAGAACGCCACGCTGACGTTCATGG
          Tcas TCAAGTTTCTCGACGCTCCGGTCTCAGGTG                        GAGTCACGG                        GGGCGGAGGCCGGGACCCTCACTTTTATGG
          Phum CTTCATTTGTTGATGCTCCTGTGTCTGGAGgtgagtta//cggtttttttaagGTGTCATTG                         GCGCAAAAGATGGTACTCTGACTTTTATGG
          Dpul TGATTTTCATTGATGCTCCAGTTTCTGGAGgtattaaa//acggttattgacagGTGTGATGG                        CTGCTAAAGCTGGAACTTTGACTTTCATGG
          Lgig CAGTTTATCTTGATGCACCAGTCTCAGGGGgtaagatc//taatttctttccagGTGTAAATG                        CAGCTAGAGATGCTTTACTCACATTCATGG
          Ccap CGGTCTTCATGGATGCCCCTGTGTCTGGAGgtaggccg//tttgcatcgtgcagGAGTGGTGG                        CCGCTAGAGACGCCCTCTTGACCTTCATGG
          Hrob CCACGTTTATTGATGCCCCTGTTTCTGGAGgtaacata//tgattgaacaaaagGTGTGACTG                        CAGCCAAGGAAGGTACTTTGACGTTTATGG
          Minc CTAGTTTTTTGGATGCCCCTGTATCTGGGG                        GAGTTTTAGgttcctta//cttttaatttgtagGTGCAGAAAAGGCTACTTTAACGTTTATGA
          Cjap CCGAGTACATTGATGCGCCGATTTCGGGAG                        GAGTTACAGgtgggaaa//ccttttattccagGCGCCCAAAACGCCACACTAACATTCATGG
          Cele CCGAGTATATCGACGCCCCAATCTCCGGTG                        GAGTCACTGgtacggga//tcataattttccagGCGCCCAGCAAGCTACACTAACATTTATGG
          Cbre CTGAATACATCGATGCTCCGATTTCCGGAG                        GGGTTACAGgtttgtta//acttaaaacttcagGAGCCCAACAAGCCACACTCACTTTTATGG
          Cbri CTGAATACATCGATTCTCCAATTTCCGGAG                        GCGTCACTG                        GAGCTCAACAAGCCACGTTGACATTTATGG
          Crem CAGAGTACATTGATGCACCGATTTCTGGTG                        GTGTCACTGgtacggag//ccaacaatttccagGCGCCCAACAAGCGACTCTCACTTTCATGG
          Nvec CAACATATTTAGATGCACCTGTTTCGGGAGgtattttt//tgttctatctttagGGATTACAG                        CAGCAAAAGCAGGTACACTGACATTTATGG
          Hmag GTTCATATGTGGATGCACCTGTTTCAGGGGgtaaattt//tatcatatgtttagGTGTTAACG                        CTGCAAAAGAAGGAACATTGACAATTATGG

from ortholog set B0250.5 (Cele)
ref protein is Dmel (FBpp0085821/FBgn0034390)
```

Metazoan NIPs: A Nice Example Alignment.

Pos0 <32nt Pos1

```
(Dmel protein)                               ________149-1___________         ________152-1___________
          Mbre ATGCCATGCTGGACGCCCCCGTATCTGGCGgttcgtag//tcgatgcgtggcagGTGTCGGAG                        GCGCTGAAGCTGGCACTCTGACCTTTATGG
          Tadh GTACTTTTATGGACGCTCCAGTTTCTGGAGgtacaact//attcccgtaaatagGAGTTGTAG                        CGGCTCGCGATGCGTTATTAACATTTATGG
          Hsap CAGTTTTCATGGATGCCCCTGTTTCTGGTGgtaagtgg//cccaaatgattcagGTGTAGGAG                        CTGCACGATCTGGGAACCTCACGTTTATGG
          Mdom CACTTTTCATGGATGCCCCTGTTTCTGGTGgtgagtaa//atctttttctccagGGGTGGGAG                        CTGCTCGAGCTGGGAACCTGACATTCATGG
         Mdom1 CACTTTTCCTGGATGTCCCTGTTTCTGGTG                        GG-------                        --------CGTGGGAACCTGGCATTCATGG
          Drer CTGTTTTCATGGATGCCCCGGTGTCTGGAGgtgagatg//ttaatcttcctcagGTGTTGGTG                        CTGCCACTTCTGGTAAGCTCACTTTTATGG
         Drer1 CAGTGTTCATGGACGCGCCGGTGTCAGGAGgtatgatg//tgtgttctgtctagGTGTTGGTG                        CAGCCAGTTTGGCTAAACTCACTTTCTTGG
          Bflo GAGTCTACATGGACGCACCTGTGTCTGGGGgtatggta//tctacacattacagGTATTAAGG                        CAGCTGCTGCAGCCACACTAACCTTCATGG
          Csav CCAAGTATCTGGATGCACCAGTTTCTGGAGgtttgcat//attgttctacacagGTGTGGGTG                        CCGCTCAAGCAGGAACTTTAACTTTCATGG
          Isca --------------------CTTTTACGGGgtatcacc//cttttatattctagGTGTGAATG                        CAGCCAAGGCTGGGACACTTACATTCATGG
          Dmel CTCGGTTTATCGATGCCCCCGTTTCCGGCG                        GAGTTCCCG                        GCGCCGAGCAGGCCACCCTCACCTTCATGG
          Agam CAACGTTCGTCGATGCGCCCGTGTCCGGCG                        GTGTGCCCG                        GTGCAAAGAACGCGACGCTCACGTTCATGG
          Aaeg CTACCTTCGTCGATGCCCCGGTTTCCGGTG                        GTGTTCCCG                        GAGCTCAGAATGCTACTCTGACCTTCATGG
          Cpip CGACCTTTGTCGATGCGCCCGTTTCCGGTG                        GAGTTCCCG                        GAGCTAAGAACGCCACGCTGACGTTCATGG
          Tcas TCAAGTTTCTCGACGCTCCGGTCTCAGGTG                        GAGTCACGG                        GGGCGGAGGCCGGGACCCTCACTTTTATGG
          Phum CTTCATTTGTTGATGCTCCTGTGTCTGGAGgtgagtta//cggttttttttaagGTGTCATTG                        GCGCAAAAGATGGTACTCTGACTTTTATGG
          Dpul TGATTTTCATTGATGCTCCAGTTTCTGGAGgtattaaa//acggttattgacagGTGTGATGG                        CTGCTAAAGCTGGAACTTTGACTTTCATGG
          Lgig CAGTTTATCTTGATGCACCAGTCTCAGGGGgtaagatc//taatttctttccagGTGTAAATG                        CAGCTAGAGATGCTTTACTCACATTCATGG
          Ccap CGGTCTTCATGGATGCCCCTGTGTCTGGAGgtaggccg//tttgcatcgtgcagGAGTGGTGG                        CCGCTAGAGACGCCCTCTTGACCTTCATGG
          Hrob CCACGTTTATTGATGCCCCTGTTTCTGGAGgtaacata//tgattgaacaaaagGTGTGACTG                        CAGCCAAGGAAGGTACTTTGACGTTTATGG
          Minc CTAGTTTTTTGGATGCCCCTGTATCTGGGG                        GAGTTTTAGgttcctta//cttttaatttgtagGTGCAGAAAAGGCTACTTTAACGTTTATGA
          Cjap CCGAGTACATTGATGCGCCGATTTCGGGAG                        GAGTTACAGgtgggaaa//cctttttattccagGCGCCCAAAACGCCACACTAACATTCATGG
          Cele CCGAGTATATCGACGCCCCAATCTCCGGTG                        GAGTCACTGgtacggga//tcataattttccagGCGCCCAGCAAGCTACACTAACATTTATGG
          Cbre CTGAATACATCGATGCTCCGATTTCCGGAG                        GGGTTACAGgtttgtta//acttaaaacttcagGAGCCCAACAAGCCACACTCACTTTTATGG
          Cbri CTGAATACATCGATTCTCCAATTTCCGGAG                        GCGTCACTG                        GAGCTCAACAAGCCACGTTGACATTTATGG
          Crem CAGAGTACATTGATGCACCGATTTCTGGTG                        GTGTCACTGgtacggag//ccaacaatttccagGCGCCCAACAAGCGACTCTCACTTTCATGG
          Nvec CAACATATTTAGATGCACCTGTTTCGGGAGgtattttt//tgttctatctttagGGATTACAG                        CAGCAAAAGCAGGTACACTGACATTTATGG
          Hmag GTTCATATGTGGATGCACCTGTTTCAGGGGgtaaattt//tatcatatgtttagGTGTTAACG                        CTGCAAAAGAAGGAACATTGACAATTATGG

from ortholog set B0250.5 (Cele)
ref protein is Dmel (FBpp0085821/FBgn0034390)
```

Metazoan NIPs: A Nice Example Alignment.

Pos0 **Pos1**

```
(Dmel protein)                               ________149-1___________         ________152-1___________
          Mbre H__A__M__L__D__A__P__V__S__G__gttcgtag//tcgatgcgtggcagG__V__G__                        G__A__E__A__G__T__L__T__F__M__
          Tadh S__T__F__M__D__A__P__V__S__G__gtacaact//attcccgtaaatagG__V__V__                        A__A__R__D__A__L__L__T__F__M__
          Hsap A__V__F__M__D__A__P__V__S__G__gtaagtgg//cccaaatgattcagG__V__G__                        A__A__R__S__G__N__L__T__F__M__
          Mdom A__L__F__M__D__A__P__V__S__G__gtgagtaa//atctttttctccagG__V__G__                        A__A__R__A__G__N__L__T__F__M__
         Mdom1 P__L__F__L__D__V__P__V__S__G__                        G__-__-__                        -__-__-__R__G__N__L__A__F__M__
          Drer A__V__F__M__D__A__P__V__S__G__gtgagatg//ttaatcttcctcagG__V__G__                        A__A__T__S__G__K__L__T__F__M__
         Drer1 A__V__F__M__D__A__P__V__S__G__gtatgatg//tgtgttctgtctagG__V__G__                        A__A__S__L__A__K__L__T__F__L__
          Bflo G__V__Y__M__D__A__P__V__S__G__gtatggta//tctacacattacagG__I__K__                        A__A__A__A__A__T__L__T__F__M__
          Csav A__K__Y__L__D__A__P__V__S__G__gtttgcat//attgttctacacagG__V__G__                        A__A__Q__A__G__T__L__T__F__M__
          Isca -__-__-__-__-__-__-__L__L__R__gtatcacc//cttttatattctagG__V__N__                        A__A__K__A__G__T__L__T__F__M__
          Dmel A__R__F__I__D__A__P__V__S__G__                        G__V__P__                        G__A__E__Q__A__T__L__T__F__M__
          Agam A__T__F__V__D__A__P__V__S__G__                        G__V__P__                        G__A__K__N__A__T__L__T__F__M__
          Aaeg A__T__F__V__D__A__P__V__S__G__                        G__V__P__                        G__A__Q__N__A__T__L__T__F__M__
          Cpip A__T__F__V__D__A__P__V__S__G__                        G__V__P__                        G__A__K__N__A__T__L__T__F__M__
          Tcas F__K__F__L__D__A__P__V__S__G__                        G__V__T__                        G__A__E__A__G__T__L__T__F__M__
          Phum T__S__F__V__D__A__P__V__S__G__gtgagtta//cggtttttttaagG__V__I__                        G__A__K__D__G__T__L__T__F__M__
          Dpul M__I__F__I__D__A__P__V__S__G__gtattaaa//acggttattgacagG__V__M__                        A__A__K__A__G__T__L__T__F__M__
          Lgig S__V__Y__L__D__A__P__V__S__G__gtaagatc//taatttctttccagG__V__N__                        A__A__R__D__A__L__L__T__F__M__
          Ccap S__V__F__M__D__A__P__V__S__G__gtaggccg//tttgcatcgtgcagG__V__V__                        A__A__R__D__A__L__L__T__F__M__
          Hrob S__T__F__I__D__A__P__V__S__G__gtaacata//tgattgaacaaaagG__V__T__                        A__A__K__E__G__T__L__T__F__M__
          Minc A__S__F__L__D__A__P__V__S__G__                        G__V__L__gttcctta//cttttaatttgtagG__A__E__K__A__T__L__T__F__M__
          Cjap A__E__Y__I__D__A__P__I__S__G__                        G__V__T__gtgggaaa//cctttttattccagG__A__Q__N__A__T__L__T__F__M__
          Cele A__E__Y__I__D__A__P__I__S__G__                        G__V__T__gtacggga//tcataattttccagG__A__Q__Q__A__T__L__T__F__M__
          Cbre A__E__Y__I__D__A__P__I__S__G__                        G__V__T__gtttgtta//acttaaaacttcagG__A__Q__Q__A__T__L__T__F__M__
          Cbri A__E__Y__I__D__S__P__I__S__G__                        G__V__T__                        G__A__Q__Q__A__T__L__T__F__M__
          Crem A__E__Y__I__D__A__P__I__S__G__                        G__V__T__gtacggag//ccaacaatttccagG__A__Q__Q__A__T__L__T__F__M__
          Nvec A__T__Y__L__D__A__P__V__S__G__gtattttt//tgttctatctttagG__I__T__                        A__A__K__A__G__T__L__T__F__M__
          Hmag C__S__Y__V__D__A__P__V__S__G__gtaaattt//tatcatatgtttagG__V__N__                        A__A__K__E__G__T__L__T__I__M__

from ortholog set B0250.5 (Cele)
ref protein is Dmel (FBpp0085821/FBgn0034390)
```

Metazoan NIPs: Example Character Matrix.

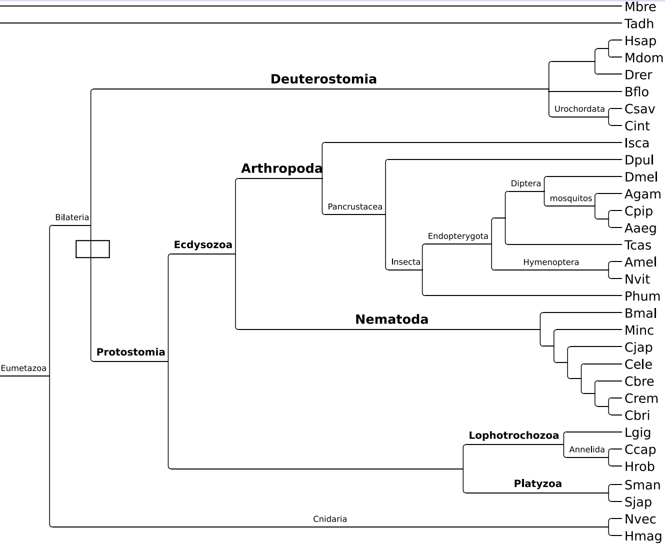
```
 'B0250.1.47-0_48-1' 'B0
 'B0250.5.149-1_152-1' 'B
MATRIX
Mbre    1???????
Tadh    ?00?0???

  ...

Hrob    ????0???
Bmal    ?00?????
Cjap    ?00?1???
Cele    ????100?
Cbre    ?00??00?
Crem    ?00?100?
Cbri    ?00??00?
Sman    ?1?0????
Nvec    ?00?0?11
Hmag    ?00?0???
Mdom    000?0?11
Drer    000?0???
:
```

Outline

1. Motivation
   - The NIP Idea
   - Previous Work
2. NIPs in Metazoa
   - Data Compilation and NIP Extraction
   - NIP Distribution Results

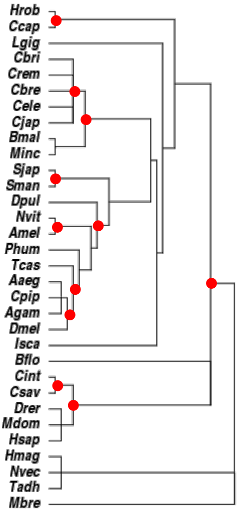
Parsimony Tree Search with PAUP*.

```
hsearch addseq=random swap=tbr nreps=1000
```



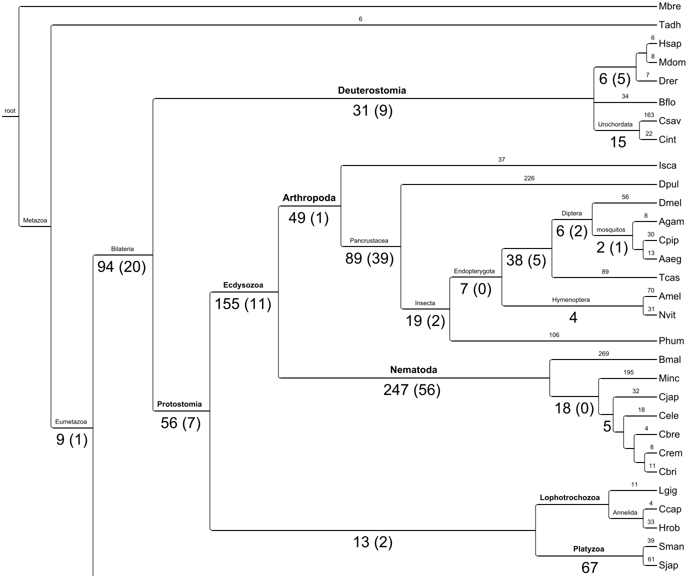
data: 1349 informative character

supposed tree: 1538 steps



our MajRule consensus tree
(17010 trees with 1510 steps)

Synapomorphic and Autapomorphic NIPs for Supposed Tree



Open Questions.

- position of Amphioxus (*Bflo*)
- position of the flatworms (Platyhelminthes, *Sjap*/*Sman*/*Smed*)

Limitations with MP and NIP character

- **Long Branch Attraction** - solution: dense taxon sampling
- limited amount of informative character changes in some branches vs
- conflicting (but frequent) character changes leading to unresolved clades

Where To Go Next?

- improve/extend ortholog dataset
- evaluate specific parts of the Metazoan tree (testing phylogenetic hypotheses), e.g. within Caernorhabditis

Summary

- NIP concept and application as phylogenetic marker and novel intron search
- NIP data for Metazoan seems promising

- To improve
  - more reliable and complete ortholog assignments (increase data and taxa)
  - extend automated NIP validation

Thank You!

Many thanks to:

- Carina Eisenhardt (Genetics Group Uni Leipzig)
- Veiko Krauss
- Peter F. Stadler

Thank you for your attention!

Any questions?

For Further Reading I

Krauss V, Thümmler C, Georgi F, Lehmann J, Stadler PF, Eisenhardt C.
Near intron positions are reliable phylogenetic markers: an application to holometabolous insects.
Mol Biol Evol, 25:821-830, 2008.

Lehmann J, Eisenhardt C, Stadler PF, Krauss V.
Some novel intron positions in conserved Drosophila genes are caused by intron sliding or tandem duplication.
BMC Evol Biol, in revision.