New genome informatics for insect genomicists

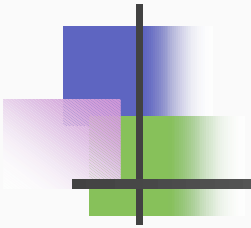
Don Gilbert

December 2009

Biology Dept., Indiana University

gilbertd@indiana.edu

wfleabase.org/docs/esangi0912.pdf



Overview

1. Current Genome Annotation recipe
2. ESTs give essential genome gene-set

   but not all genes found by ESTs
3. Next-gen base-level expression measures gene expression better

   Tiling and RNA-seq find many new, weakly expressed genes
4. RNA-Seq finds gene structures well

   the new EST? the new gene finder? maybe

Recipe for Genome Annotation

Source data:

Genome assembly, EST sequences (200+K), proteins of related species

Current Genome software (as used on Aphid & others):

PASA for EST assembly, cDNA-genes, and gene validation

NCBI BLAST to locate related proteins (tblastn), and annotating predicted genes (blastp).

Exonerate to refine gene mappings of proteins.

Augustus to predict genes, using ESTs and mapped proteins.

Other predictors (fgenesh, GeneID, SNAP, Gnomon, ...)

EvidenceModeler to combine gene models, merge EST, protein evidence for a "final" gene set.

...

Recipe for Genome Annotation

Next Generation Software (in progress):

Tophat & **Bowtie** maps RNA-Seq to genomes

Augustus predicts genes from NGS data

.. and others

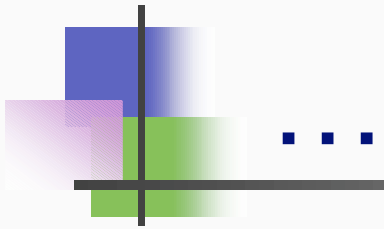
Search/Report/View tools (for web browsing results):

NCBI **BLAST** sequence search

GMOD **GBrowse** genome maps

GMOD **LuceGene** search and reports for gene annotations

...



ESTs Define Genes

EST Assembly to Genes

mRNA long reads (600 - 1000 bp) are good gene evidence

- subset of gene set (200,000 EST ~ 30%-50% of genes)

PASA, Program to Assemble Spliced Alignments
http://pasa.sourceforge.net/

- spliced alignments of expressed transcripts to model gene structures
- update gene structure annotation with new ESTs.
- classifies alternate splice variations

EST assembly to genes without genome (de novo)

RNA-Seq shorter reads as alternative ?

…

Arthropod EST Assemblies

Statistics & data of Arthropod EST assemblies including gene set validations

Acyrthosiphon pea aphid
Bombyx silkmoth
Drosophila melanogaster fruitfly
Nasonia jewel wasp
Daphnia pulex waterflea
Ixodes tick

http://insects.eugenes.org/arthropods/summaries/PASA-EST-assemblies.html

...

Arthropod EST genes

	aphid	bombyx	daphnia	drosmel	ixodes	nasonia
Total EST	168,000	246,000	166,000	568,000	194,000	176,000
Assemblies	24,700	33,200	18,200	42,600	16,200	21,900
Genes w/ EST	9,600	8,500	10,600	9,200	7,800	8,700
Gene updates	**aphid**	**bombyx**	**daphnia**	**drosmel**	**ixodes**	**nasonia**
% Incorporated	23	4	35	18	16	6
% UTR addition	10	14	13	13	8	26
% Gene extension	4	3	4	2	4	2
% Gene Merging	1	2	5	3	4	3
% Alternate splice	5	4	6	8	5	9
% New Gene	23	56	17	14	37	27
Assembly Errors	**aphid**	**bombyx**	**daphnia**	**drosmel**	**ixodes**	**nasonia**
% Low/0 identity	14	13	22	5	30	12
% Duplicates	7	8	5	2	5	2
% Split across scafs	1.0	0.7	0.3	0.2	2.9	1.0

…

wfleabase.org/docs/esangi0912.pdf

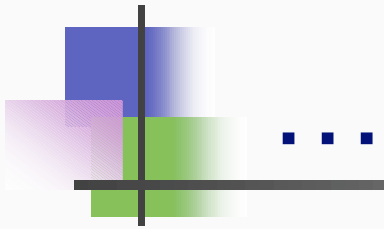
Arthropod Gene Structure

	Daphnia[1]	Aphid[1]	Bee	Wasp[1]	Moth[1]	Tick[1]	Mouse	Worm
Genome size	200	460	220	290	480	1,760	3,450	100
No. of genes	32,000	32,800	17,000	27,300	16,300	20,500	27,600	20,100
CDS size	1,360	1,340	1,690	1,620	1,460	1,070	2,140	1,300
Exons/gene	6.6	6.7	7.1	6.3	6.4	5.7	8.0	6.0
Exon size [2]	210	200	240	260	230	190	280	200
Intron size [3]	72	75/900	88/600	81/530	810/97	1730/90	1600/90	51/500
Intr > Exon	10%	41%	36%	24%	86%	87%	85%	33%
UTR size [4]	370	500	340	680	440	540	--	260
Alternate Tr.	10%	24%	--	23%	18%	15%	65%[5]	20%

	Beetle	Mosquito[1]	Fruitfly[1]
Genome size	180	580	180
No. of genes	16,400	18,900	13,700
CDS size	1,420	1,400	1,650
Exons/gene	4.5	3.5	4.0
Exon size [2]	310	420	410
Intron size [3]	51/1700	64/1900	63/750
Intr > Exon	34%	36%	27%
UTR size [4]	--	240	800
Alternate Tr.	--	--	36%

…

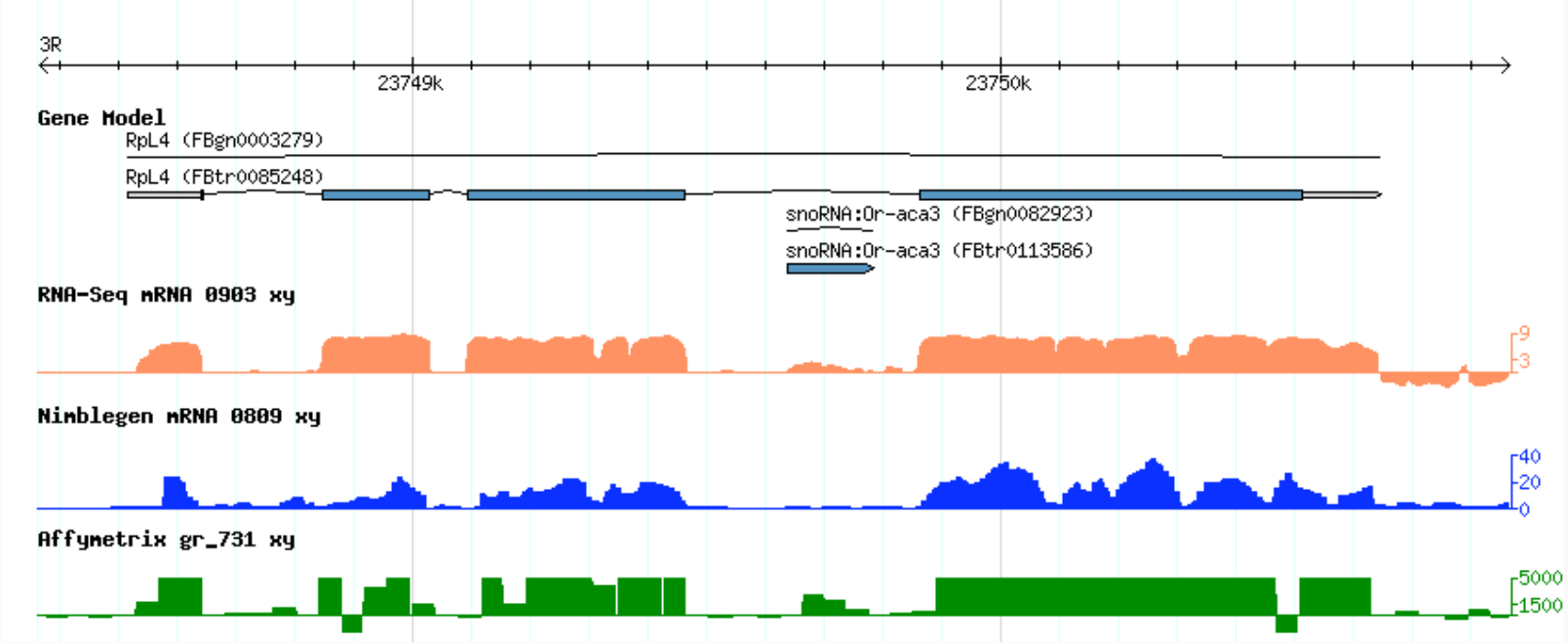
wfleabase.org/docs/esangi0912.pdf



Gene Expression from Next Gen. data

Precision improves ’06-’09

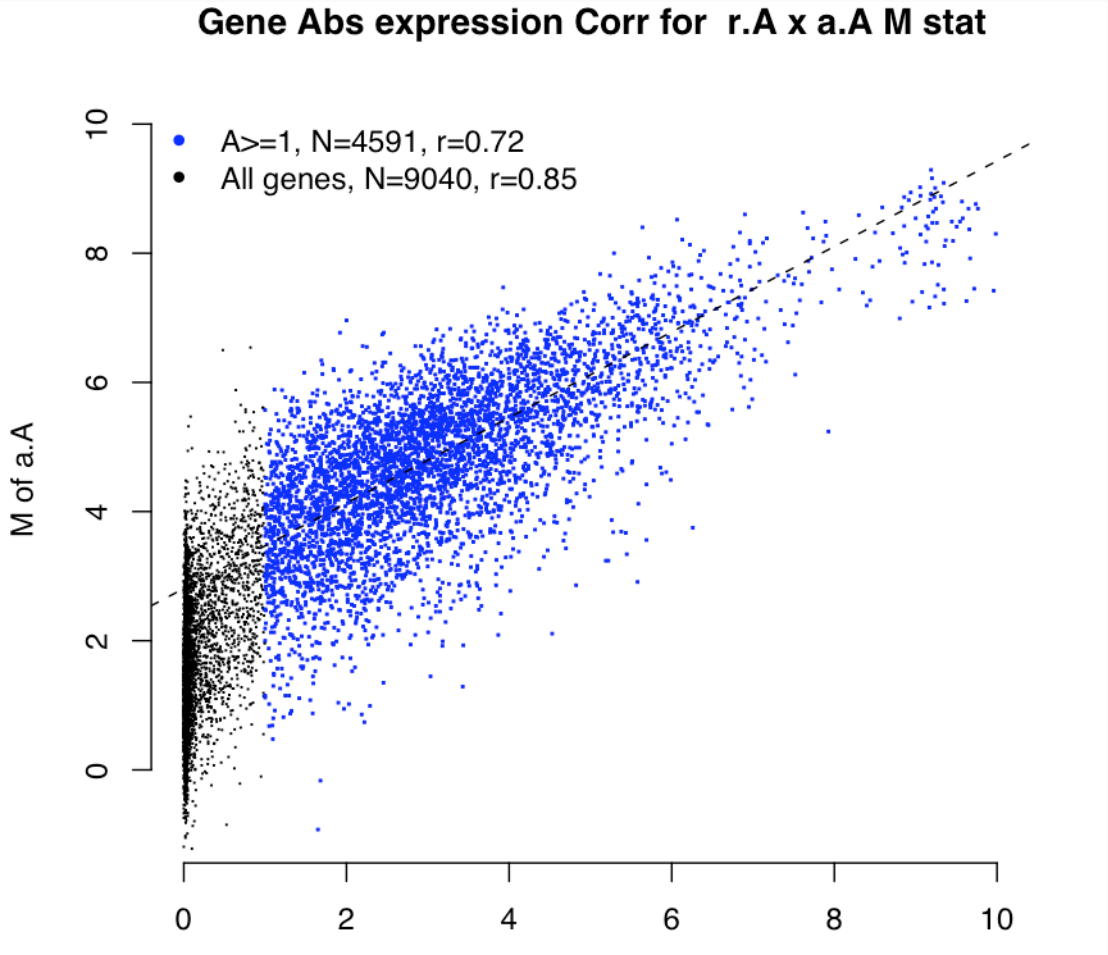
Measuring expression over gene structures,
Nimblegen (08) has higher precision than Affy (06/07)
RNA-Seq (09) has higher precision than Nimblegen



.

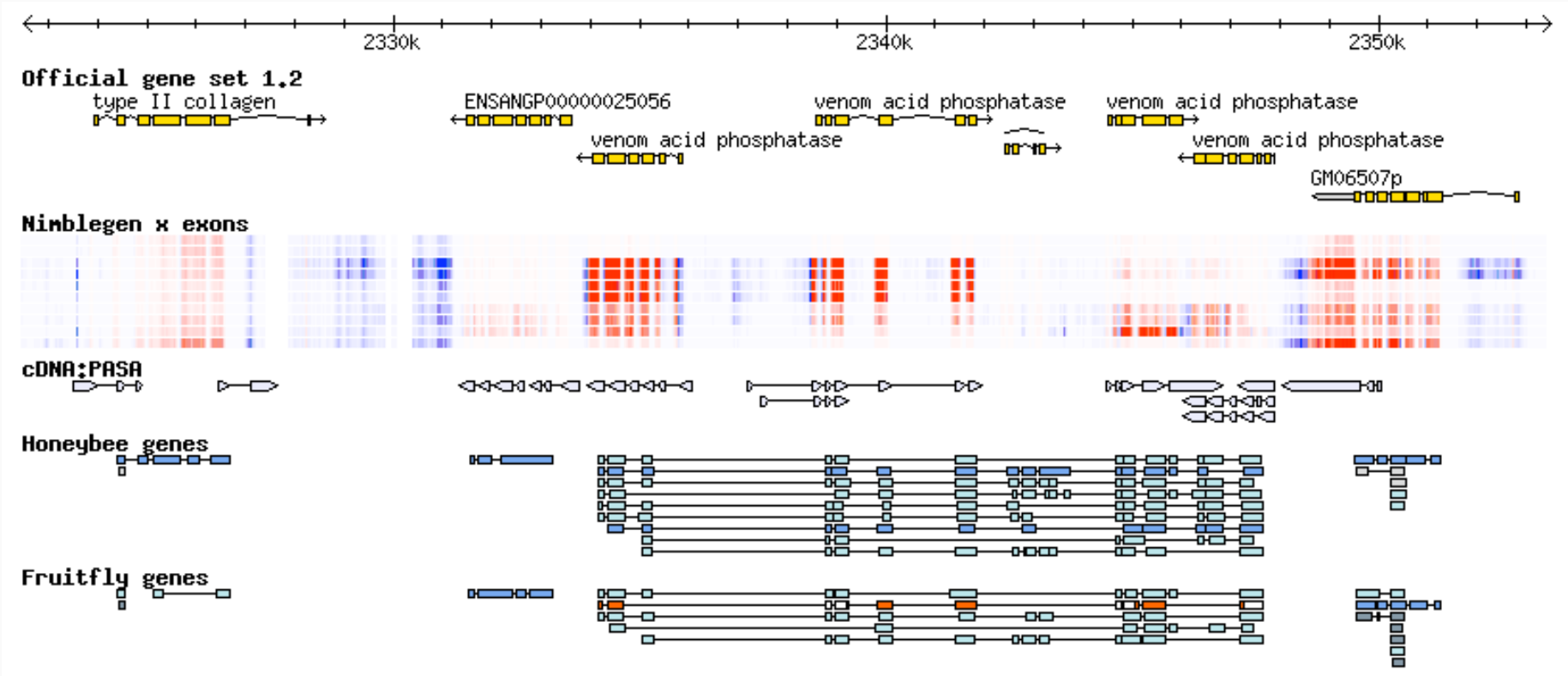
RNA-Seq & Tile arrays agree

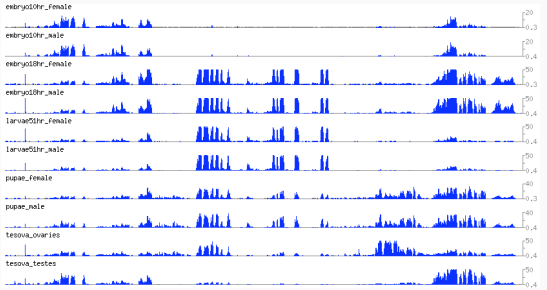
Gene expression correlation for RNA-Seq (x) by Affymetrix tiles (y) for *Drosophila* (modENCODE)



wfleabase.org/docs/esangi0912.pdf

Expression tiling in *Nasonia*





wfleabase.org/docs/esangi0912.pdf

NGS Expression Recipe

Find Base-level expression first, then gene level

1. Design treatments & replicates as for microarrays
2. Calculate differential expression per tile/read
   - R: LIMMA package (bioinf.wehi.edu.au/limma/)
3. Combine tiles/reads over gene
   - Gene to tile/read mapping table
   - Independent observations, but biologically related
   - R tilegenex software (D. Gilbert) used for daphnia, fruitfly, wasp

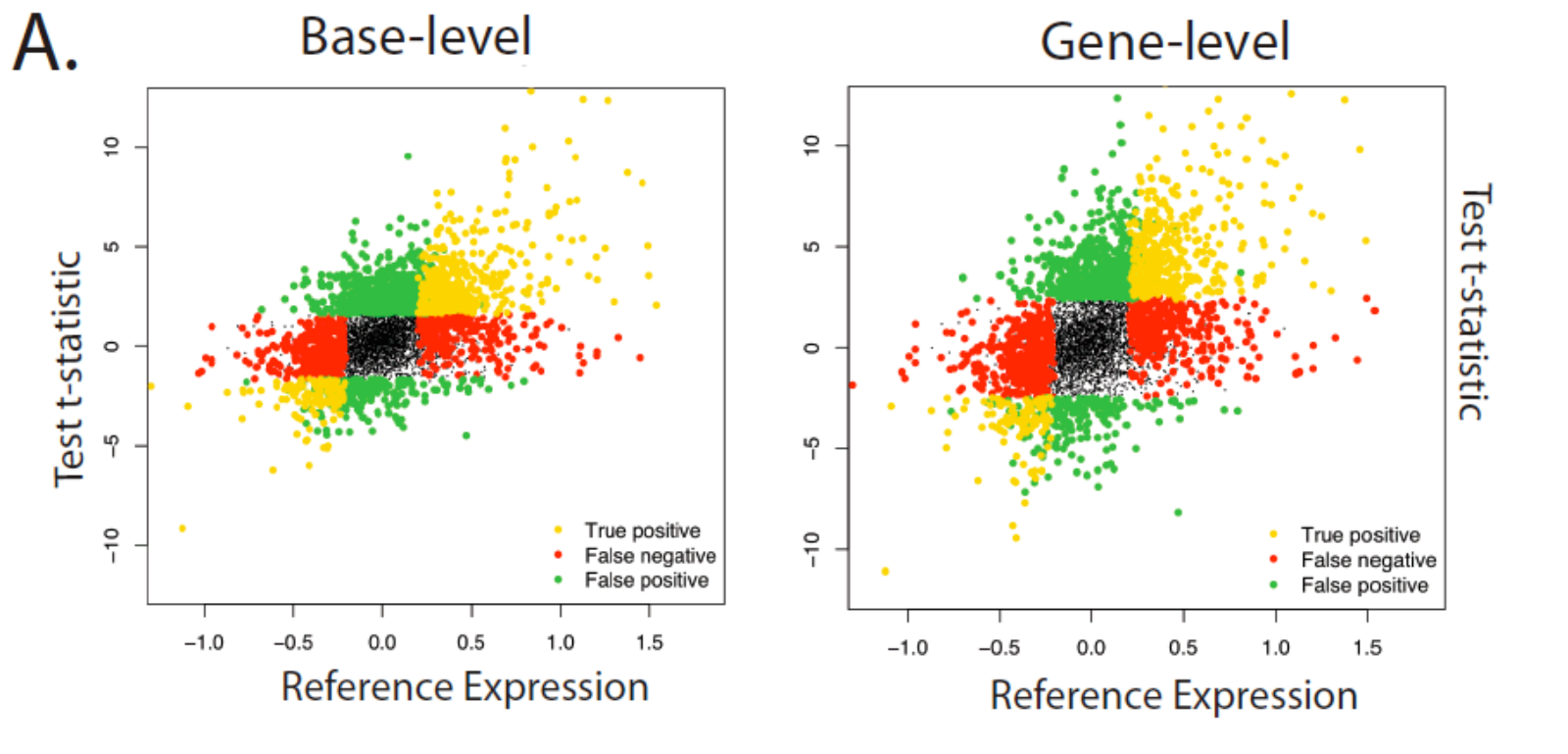
Base-level expression statistics have higher sensitivity and specificity than gene statistics.

Assess differential expression outside known genes

…

Gene or Base expression?

Base level tests find expression better than gene average

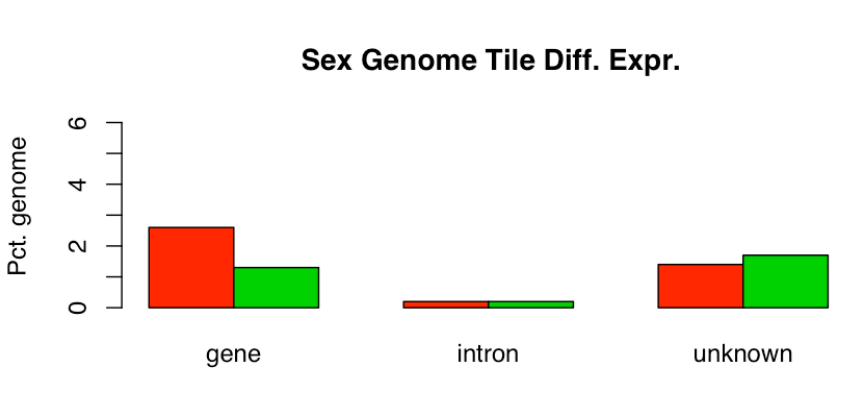


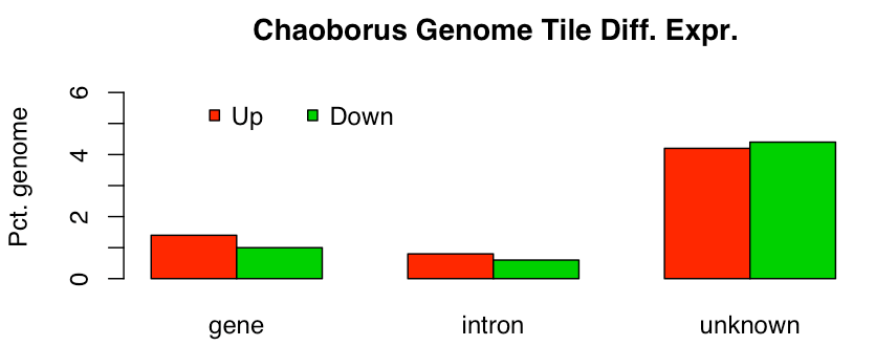
Base level sensitivity= 42%, Gene level sensitivity= 38% Both have specificity= 37%
Sensitivity = 1 - false rejection; Specificity = 1 - false discovery

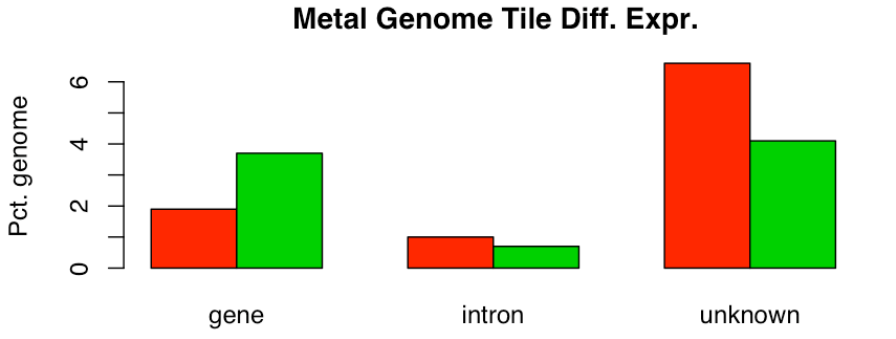
wfleabase.org/docs/esangi0912.pdf

Expression beyond known genes

Gene models miss much expression





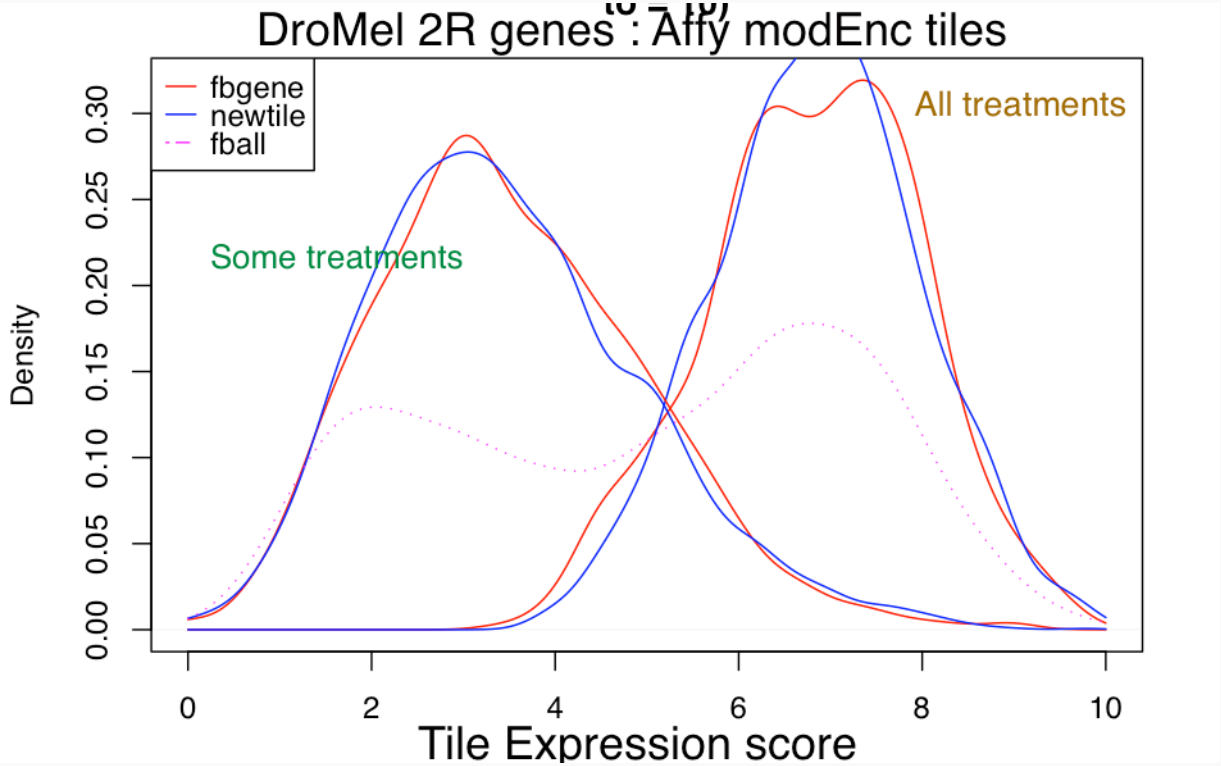


Known sex genes capture DE, but unknown regions capture environmental stress expression, in *Daphnia*.

wfleabase.org/docs/esangi0912.pdf

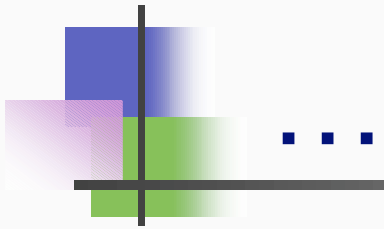
NGS find rarely used genes

Average expression is high for genes found in all treatments



Accurate gene detection at low level expression remains a challenge

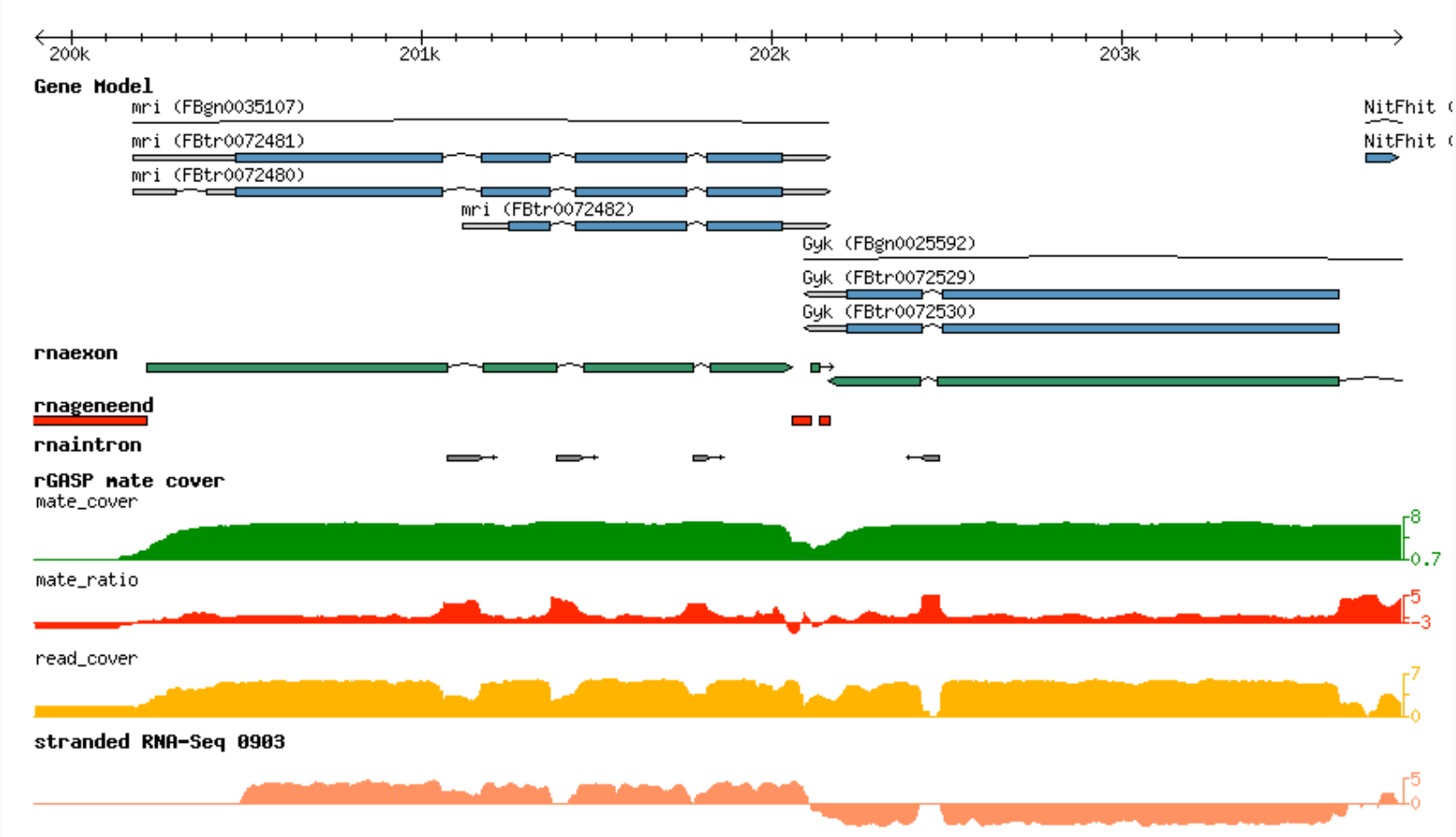
wfleabase.org/docs/esangi0912.pdf



Next Gen. Expression Finds Genes

RNA-Seq Mates find Genes

Gene ends are at Mate/Read drop in DrosMel



wfleabase.org/docs/esangi0912.pdf

RNA-Seq Genes Recipe

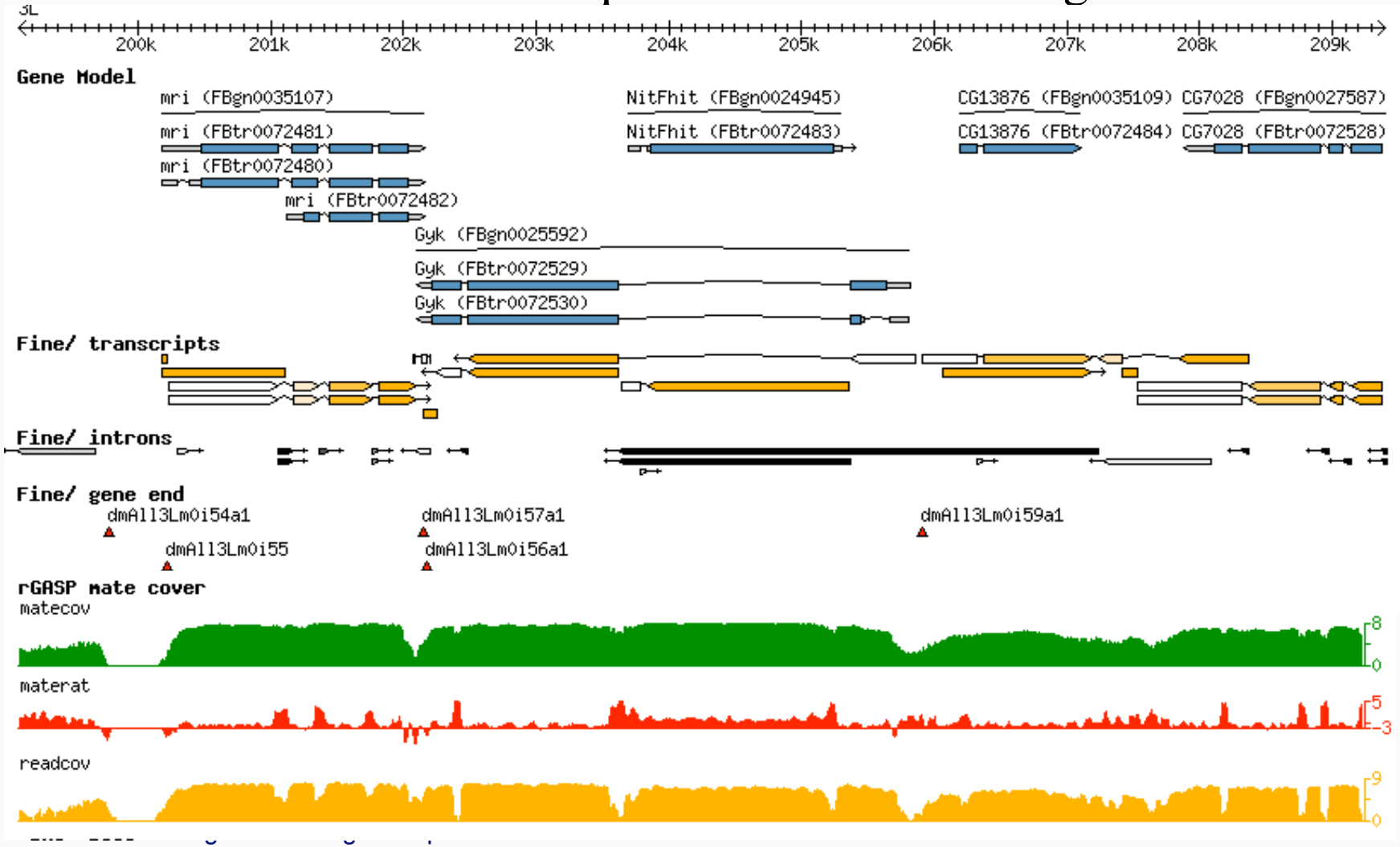
RNA-Seq can be assembled to gene models, as for ESTs, rather than rely on assumptions of gene prediction

1. Best data are paired-end mate reads, 75+ bp long
   Shorter and single reads are useful, but less information
2. Map reads to genome, with bowtie and tophat software
   bowtie-bio.sourceforge.net, tophat.cbcb.umd.edu
3. Process accepted_hits.sam to signal location table
   With count of Reads, Mates, Introns
4. Produce rough gene models from signal table using run-length encoding of signals
5. Refine rough models, produce alternate transcripts, and resolve problems

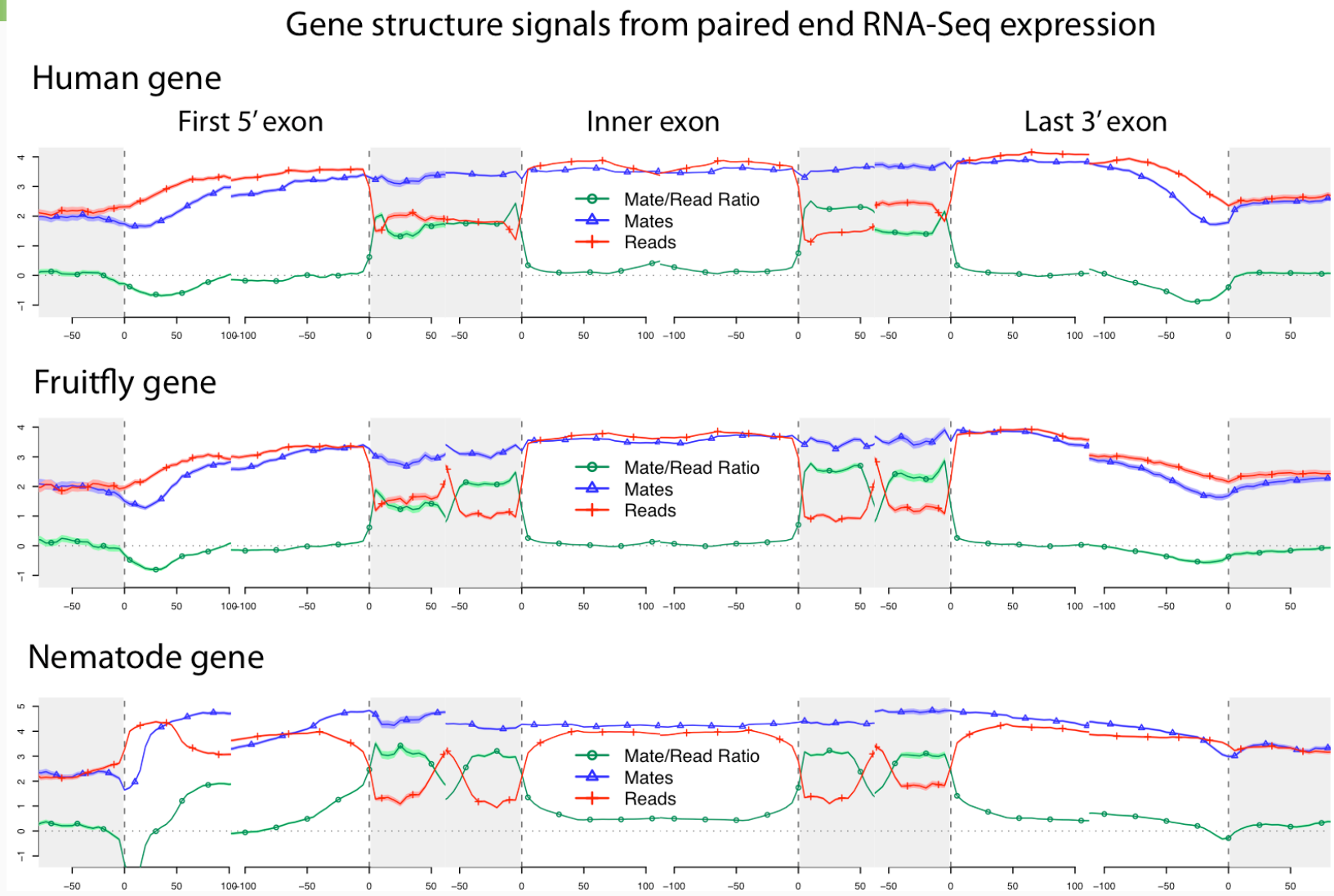
http://insects.eugenes.org/species/data/dmel5/modencode/rgaspdg/

RNA-Seq Mates find Genes

New software for Rna-seq matches and extends gene models



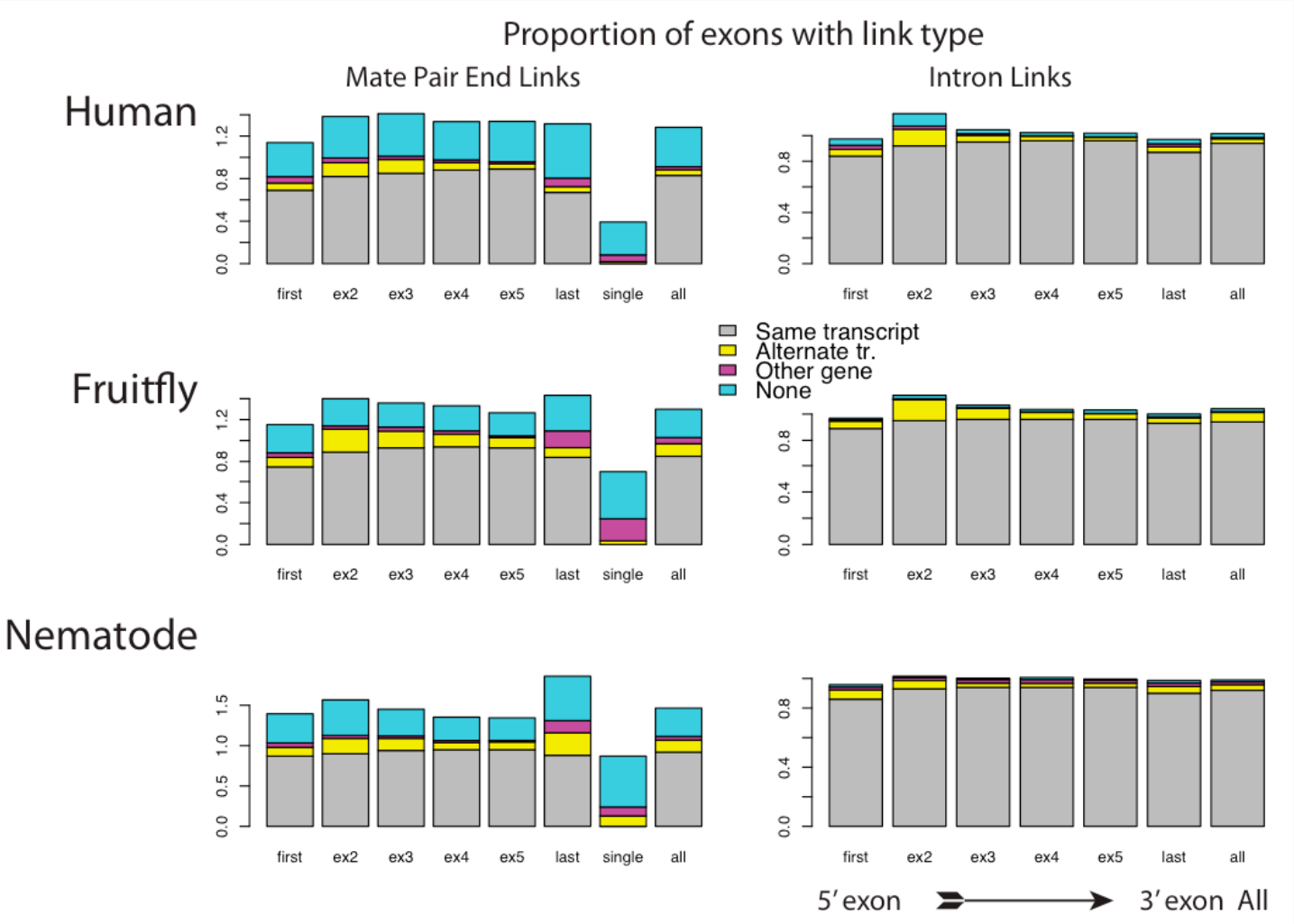
RNA-Seq & Gene Structure



wfleabase.org/docs/esangi0912.pdf

RNA-Seq Mates link Genes

RNA-Seq mates and introns show linkage of exons; also to other genes & alternate transcripts.



wfleabase.org/docs/esangi0912.pdf

End note

gilbertd@indiana.edu

Genome collaborators and data providers

12 Drosophila Genomes collaboration	rana.lbl.gov/drosophila/
Daphnia Genome Consortium	daphnia.cgb.indiana.edu
Generic Model Organism Database	GMOD.org
International Aphid Genomics Consortium	www.as.miami.edu/iagc/
modENCODE project	www.modencode.org
Nasonia Genome project	www.rochester.edu/College/BIO/labs/WerrenLab/
VectorBase project	vectorbase.org

... plus several others

Links to this work

insects.eugenes.org/DroSpeGe/	12 Drosophila
insects.eugenes.org/arthropods/	14+ bug genomes
insects.eugenes.org/species/data/dmel5/modencode/	RNA-Seq software
www.bio.net	Arthropod news/discussion list