UniL SIB 3-5 September 2018

Arthropod comparative genomics with OrthoDB & BUSCO

✉ robert.waterhouse@gmail.com @rmwaterhouse www.rmwaterhouse.org

Arthropod Comparative Genomics

Questions and Approaches through Examples

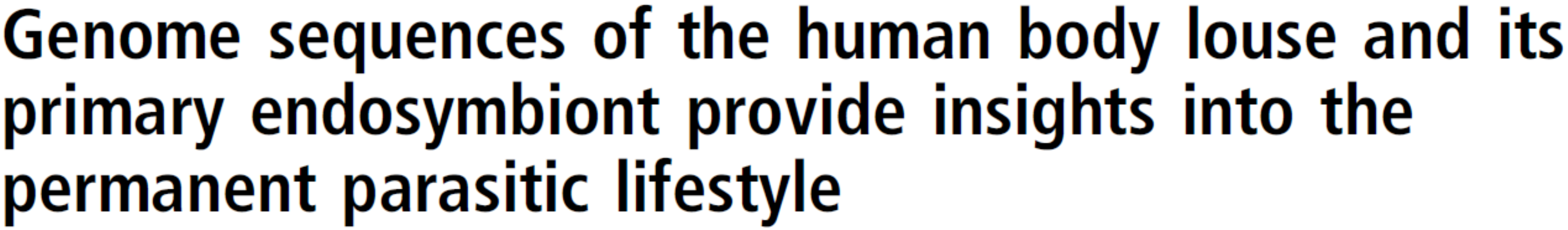
- ❖ I have predicted a small gene set - why?
- ❖ I have predicted a large gene set - why?
- ❖ Did my gene annotation upgrade work?
- ❖ Phylogenomics without genomes - how?
- ❖ *What are my species/lineage-specific genes doing?*

Pediculus humanus - body louse

Q: As an obligate parasite with a small genome, is there any evidence for the loss of genes driven by genome reduction?

Approach: orthology delineation with representatives from different insect orders and an outgroup.

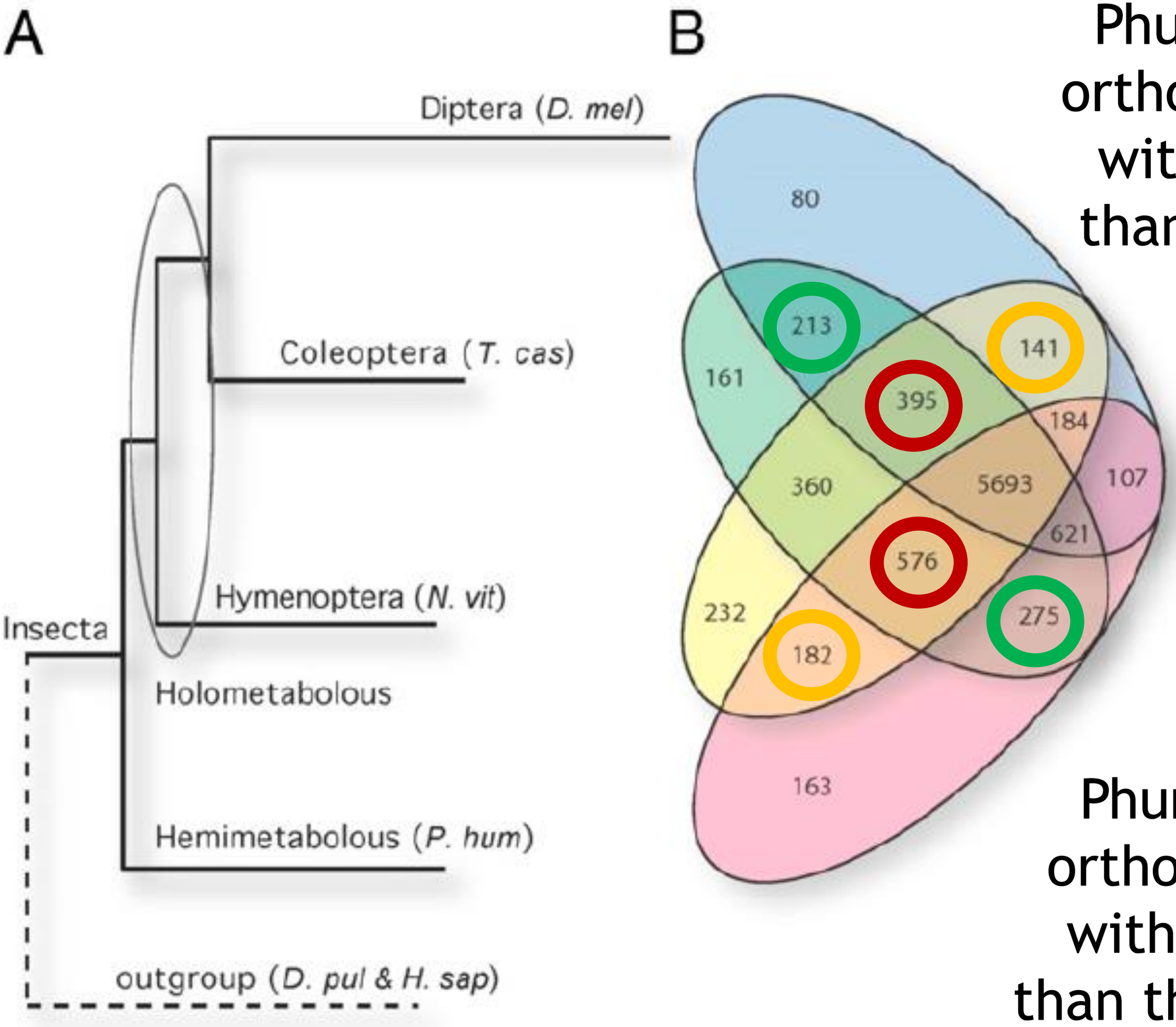
Examine the numbers of orthologs shared amongst different pairs and sets of species.



Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle

Pediculus humanus - body louse

© R.M.Waterhouse



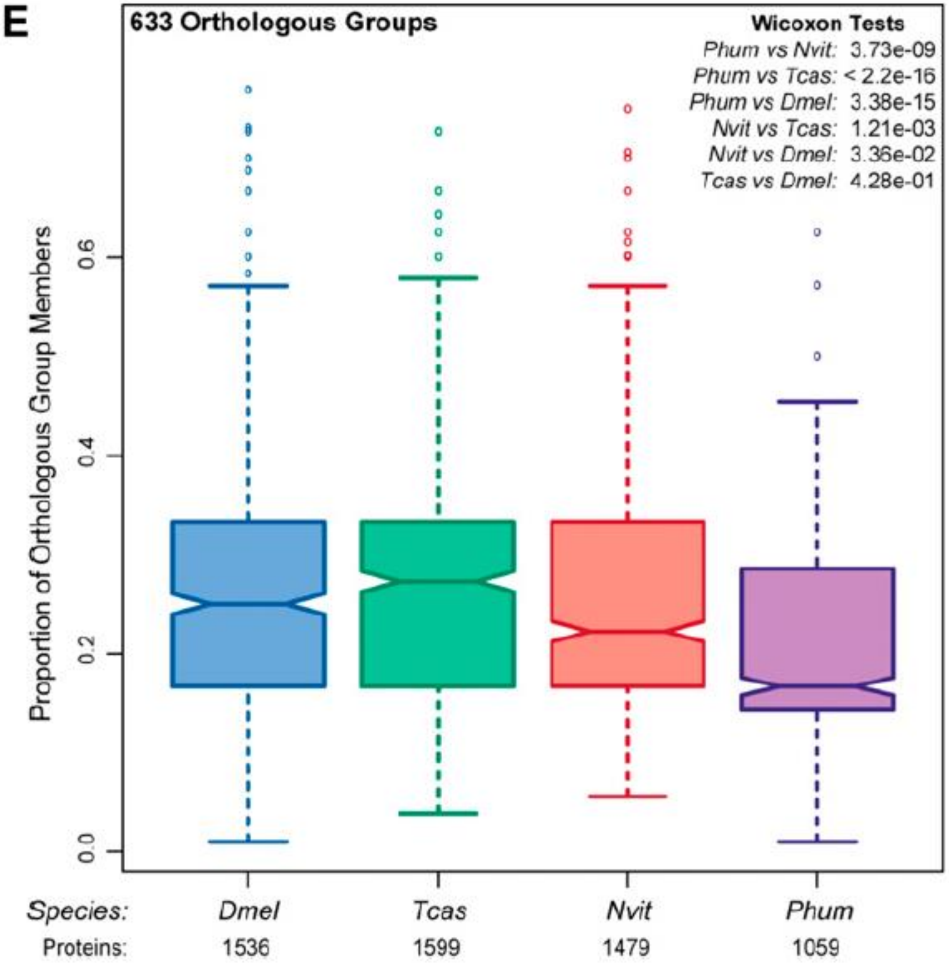
Phum shares more orthologs exclusively with Nvit OR Tcas than either do with Dmel

Phum shares more orthologs exclusively with **Nvit AND Tcas** than they do with Dmel

Pediculus humanus - body louse

© R.M.Waterhouse

This suggests that the body louse genomes has not undergone general large-scale gene loss, so perhaps the small gene set is more due to a lack of expansions ...

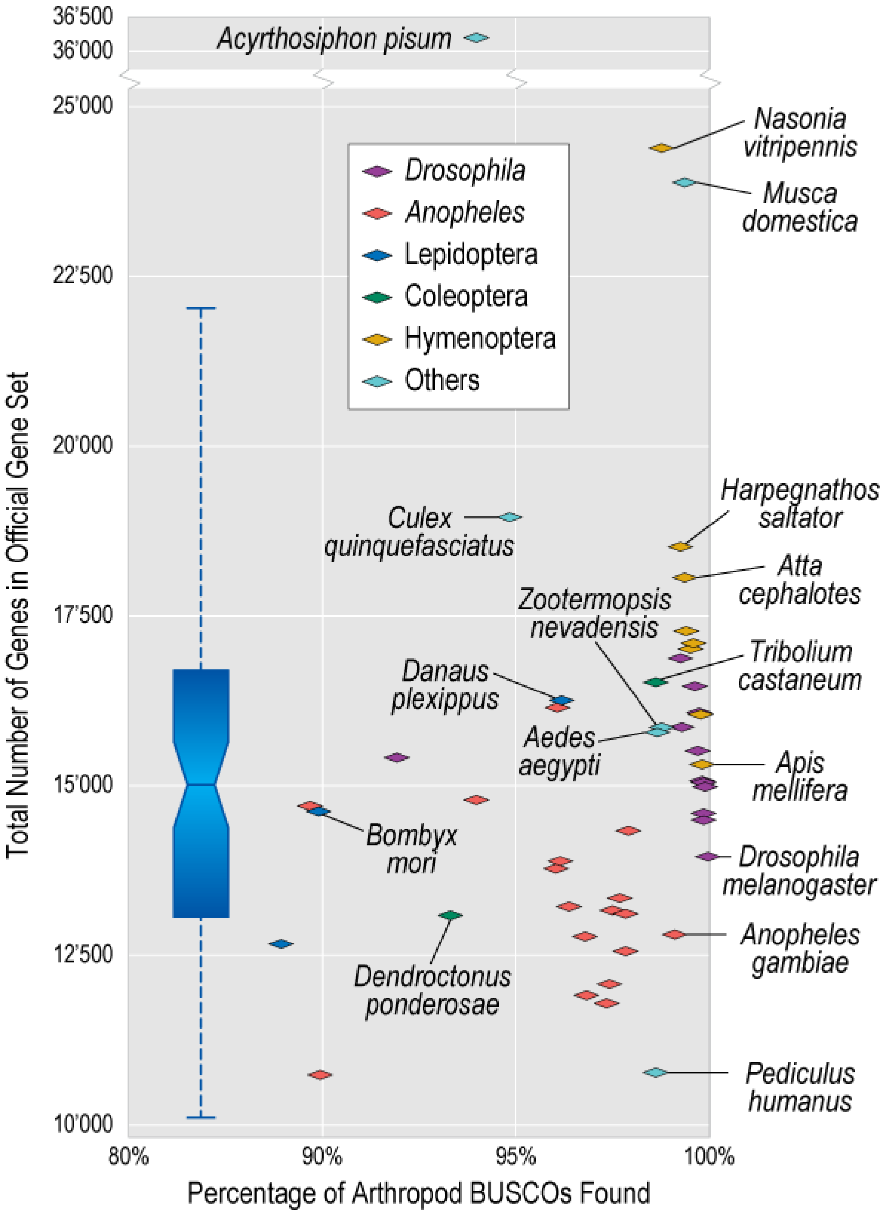


So, examine all orthologous groups with at least one ortholog in each of the 4 species, but with a total of at least 6 genes

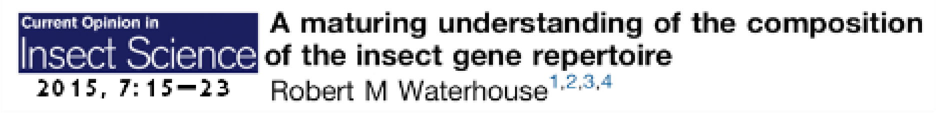
47% OGs have >1 Phum gene Nvit=59%, Tcas=70%, Dmel=64%

Phum also has lower mean & median proportions

Pediculus humanus - body louse



Large gene sets are not necessarily complete



Current Opinion in Insect Science 2015, 7:15–23

A maturing understanding of the composition of the insect gene repertoire

Robert M Waterhouse[1,2,3,4]

Small gene sets are not necessarily incomplete

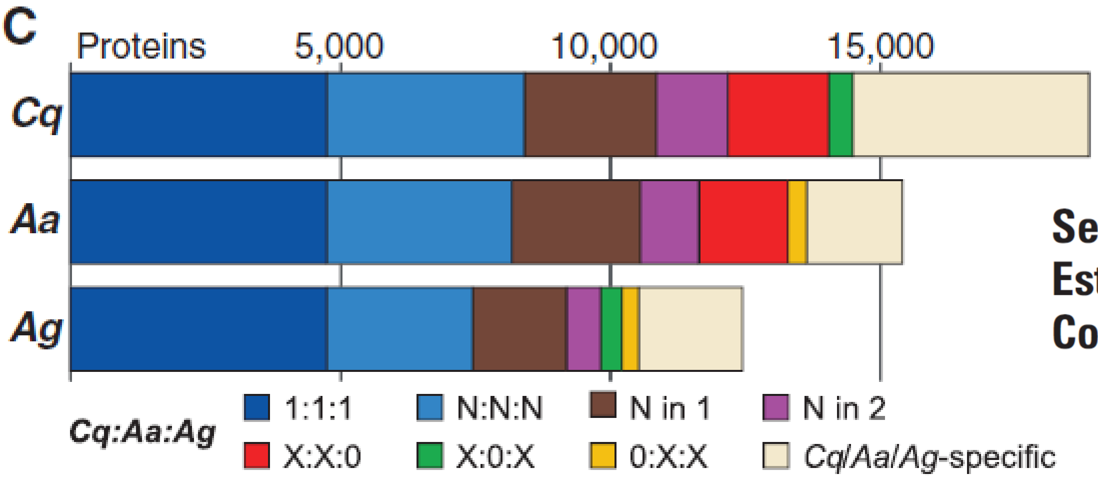
Arthropod Comparative Genomics

Questions and Approaches through Examples

- ❖ I have predicted a small gene set - why?
- > ❖ I have predicted a large gene set - why?
- ❖ Did my gene annotation upgrade work?
- ❖ Phylogenomics without genomes - how?
- ❖ *What are my species/lineage-specific genes doing?*

Culex quinquefasciatus - WNV mosquito

Q: Is the rather large predicted gene set perhaps simply due to the inclusion of many haplotype regions?



Sequencing of *Culex quinquefasciatus* Establishes a Platform for Mosquito Comparative Genomics

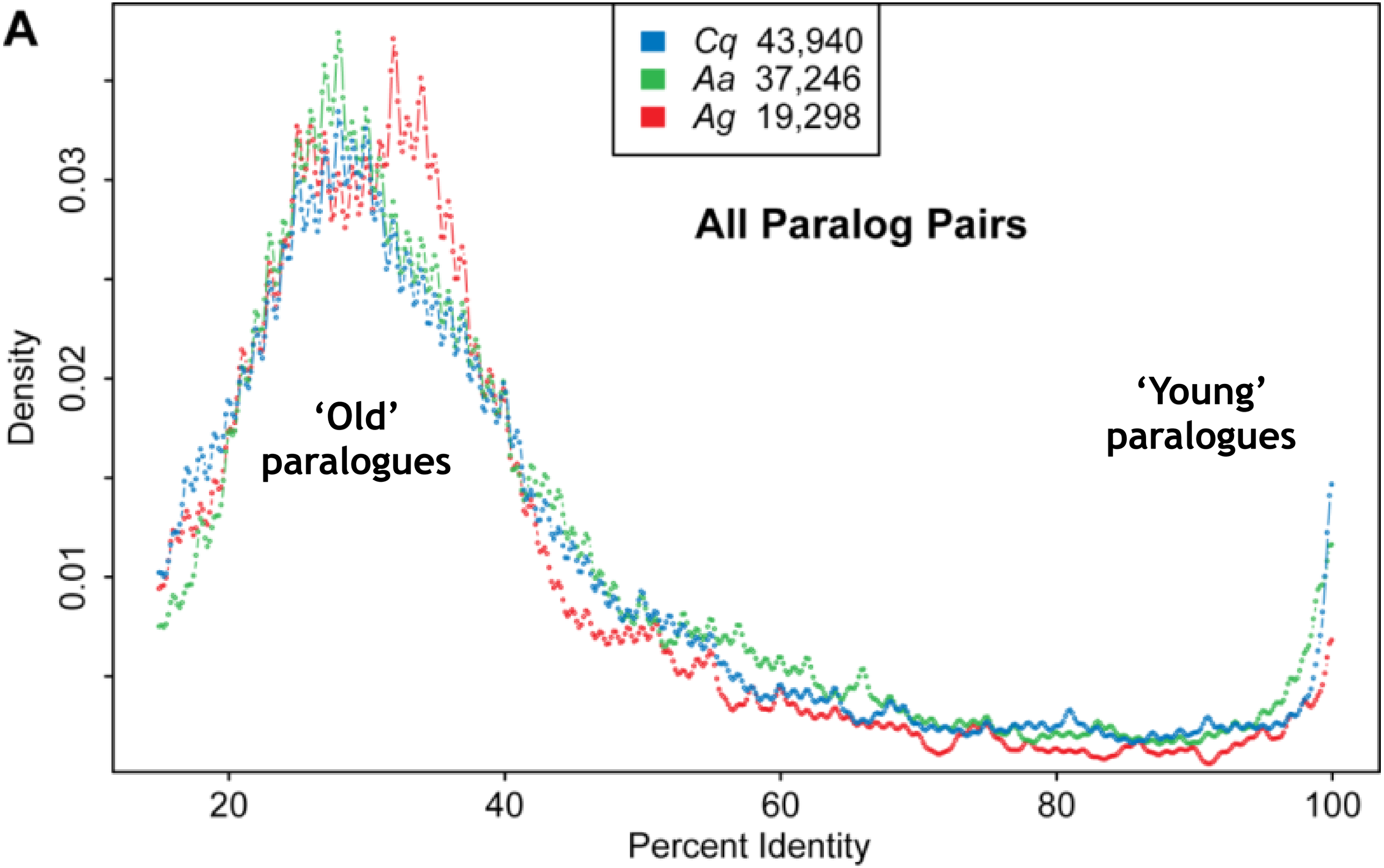
Approach: orthology delineation to identify pairs of paralogs within each of three mosquito species: *Aedes aegypti*, *Anopheles gambiae* and *Culex quinquefasciatus*.

Examine percent identity distributions of these pairs of paralogs differentiating between pairs on the same scaffold and pairs of different scaffolds.

Culex quinquefasciatus - WNV mosquito

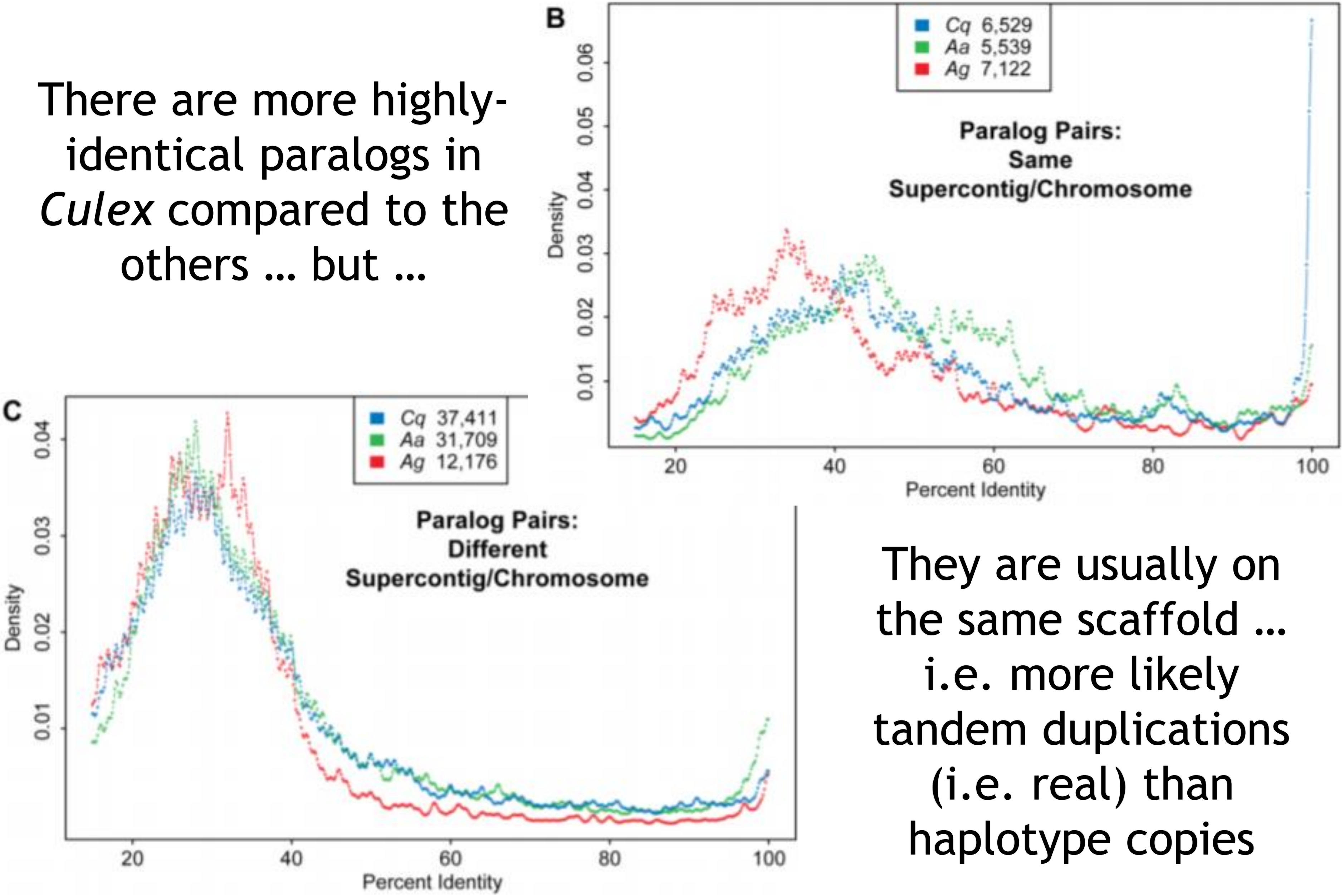
© R.M.Waterhouse

Culex does have slightly more highly-similar paralogues



Culex quinquefasciatus - WNV mosquito

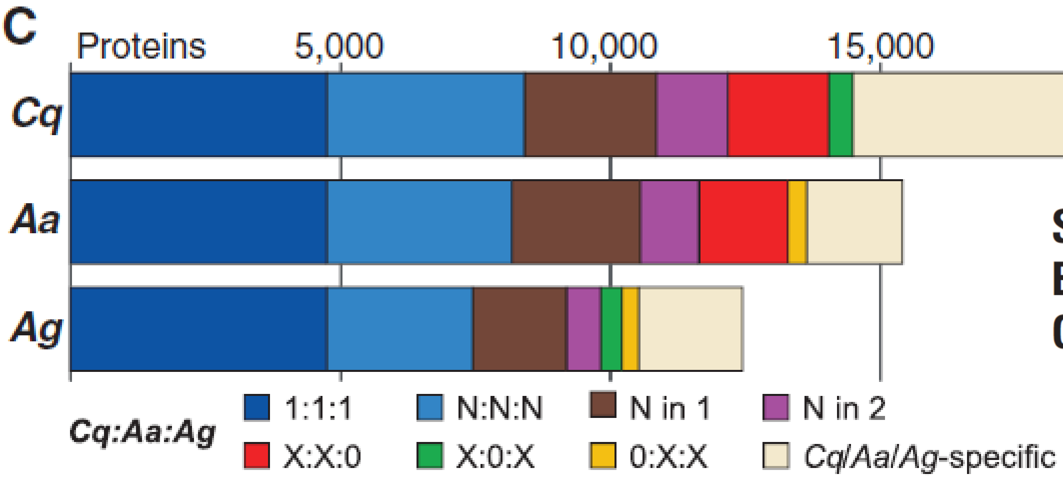
There are more highly-identical paralogs in *Culex* compared to the others … but …



They are usually on the same scaffold … i.e. more likely tandem duplications (i.e. real) than haplotype copies

Culex quinquefasciatus - WNV mosquito

Q: Is the rather large predicted gene set perhaps simply due to the inclusion of many haplotype regions?



Sequencing of *Culex quinquefasciatus* Establishes a Platform for Mosquito Comparative Genomics

N (duplicated) categories are all greater in *Culex quinquefasciatus*

Culex-specific fraction is large – real genes?

Arthropod Comparative Genomics

Questions and Approaches through Examples

❖ I have predicted a small gene set - why?

❖ I have predicted a large gene set - why?

➤ ❖ Did my gene annotation upgrade work?

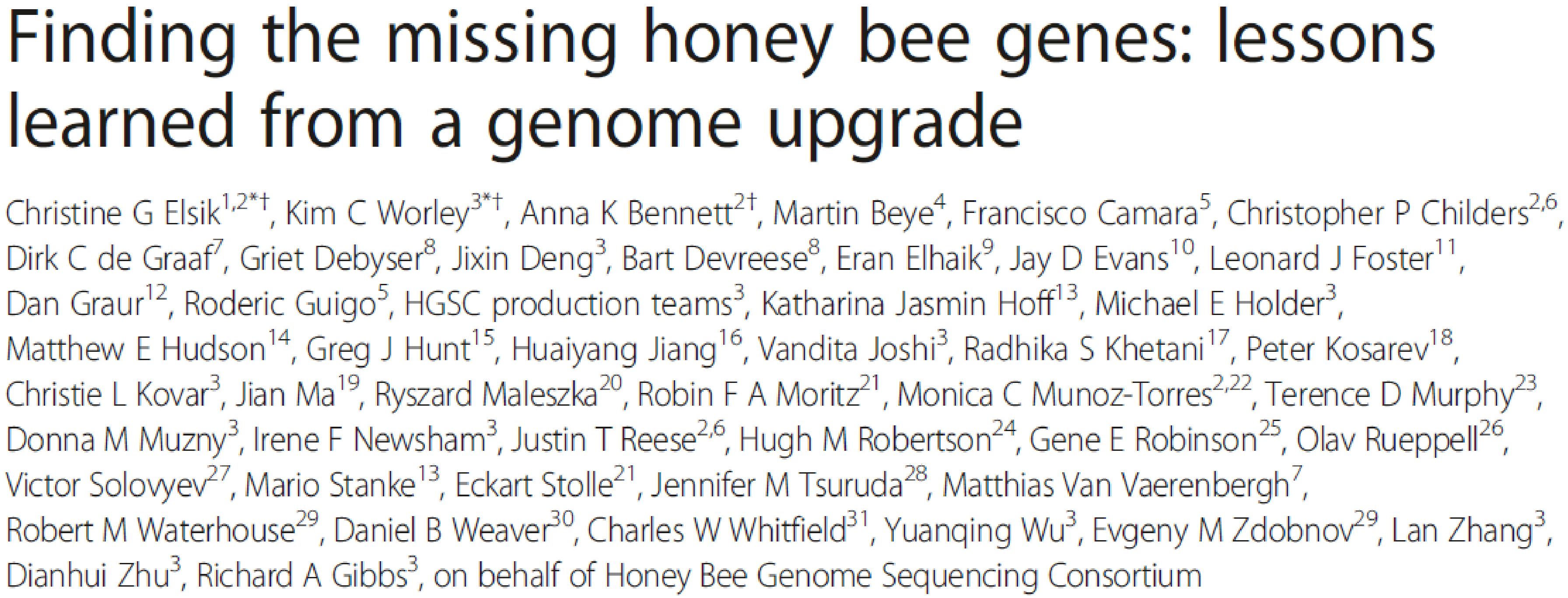
❖ Phylogenomics without genomes - how?

❖ *What are my species/lineage-specific genes doing?*

Improving honeybee gene annotations

Q: How to improve a genome annotation and has all the effort paid off?

Approach: everything you can think of*!*

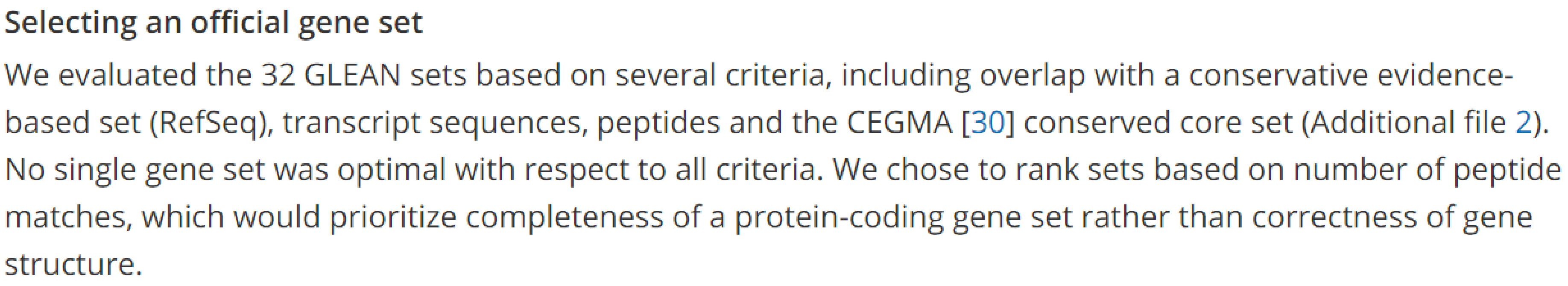


Finding the missing honey bee genes: lessons learned from a genome upgrade

Christine G Elsik[1,2*†], Kim C Worley[3*†], Anna K Bennett[2†], Martin Beye[4], Francisco Camara[5], Christopher P Childers[2,6], Dirk C de Graaf[7], Griet Debyser[8], Jixin Deng[3], Bart Devreese[8], Eran Elhaik[9], Jay D Evans[10], Leonard J Foster[11], Dan Graur[12], Roderic Guigo[5], HGSC production teams[3], Katharina Jasmin Hoff[13], Michael E Holder[3], Matthew E Hudson[14], Greg J Hunt[15], Huaiyang Jiang[16], Vandita Joshi[3], Radhika S Khetani[17], Peter Kosarev[18], Christie L Kovar[3], Jian Ma[19], Ryszard Maleszka[20], Robin F A Moritz[21], Monica C Munoz-Torres[2,22], Terence D Murphy[23], Donna M Muzny[3], Irene F Newsham[3], Justin T Reese[2,6], Hugh M Robertson[24], Gene E Robinson[25], Olav Rueppell[26], Victor Solovyev[27], Mario Stanke[13], Eckart Stolle[21], Jennifer M Tsuruda[28], Matthias Van Vaerenbergh[7], Robert M Waterhouse[29], Daniel B Weaver[30], Charles W Whitfield[31], Yuanqing Wu[3], Evgeny M Zdobnov[29], Lan Zhang[3], Dianhui Zhu[3], Richard A Gibbs[3], on behalf of Honey Bee Genome Sequencing Consortium

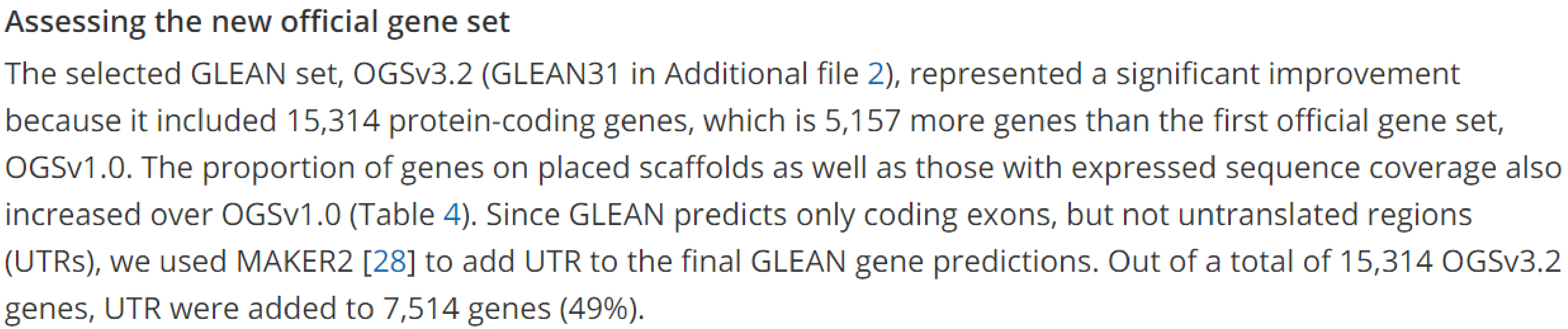
Improving honeybee gene annotations

Everything but the kitchen sink …



Selecting an official gene set

We evaluated the 32 GLEAN sets based on several criteria, including overlap with a conservative evidence-based set (RefSeq), transcript sequences, peptides and the CEGMA [30] conserved core set (Additional file 2). No single gene set was optimal with respect to all criteria. We chose to rank sets based on number of peptide matches, which would prioritize completeness of a protein-coding gene set rather than correctness of gene structure.



Assessing the new official gene set

The selected GLEAN set, OGSv3.2 (GLEAN31 in Additional file 2), represented a significant improvement because it included 15,314 protein-coding genes, which is 5,157 more genes than the first official gene set, OGSv1.0. The proportion of genes on placed scaffolds as well as those with expressed sequence coverage also increased over OGSv1.0 (Table 4). Since GLEAN predicts only coding exons, but not untranslated regions (UTRs), we used MAKER2 [28] to add UTR to the final GLEAN gene predictions. Out of a total of 15,314 OGSv3.2 genes, UTR were added to 7,514 genes (49%).

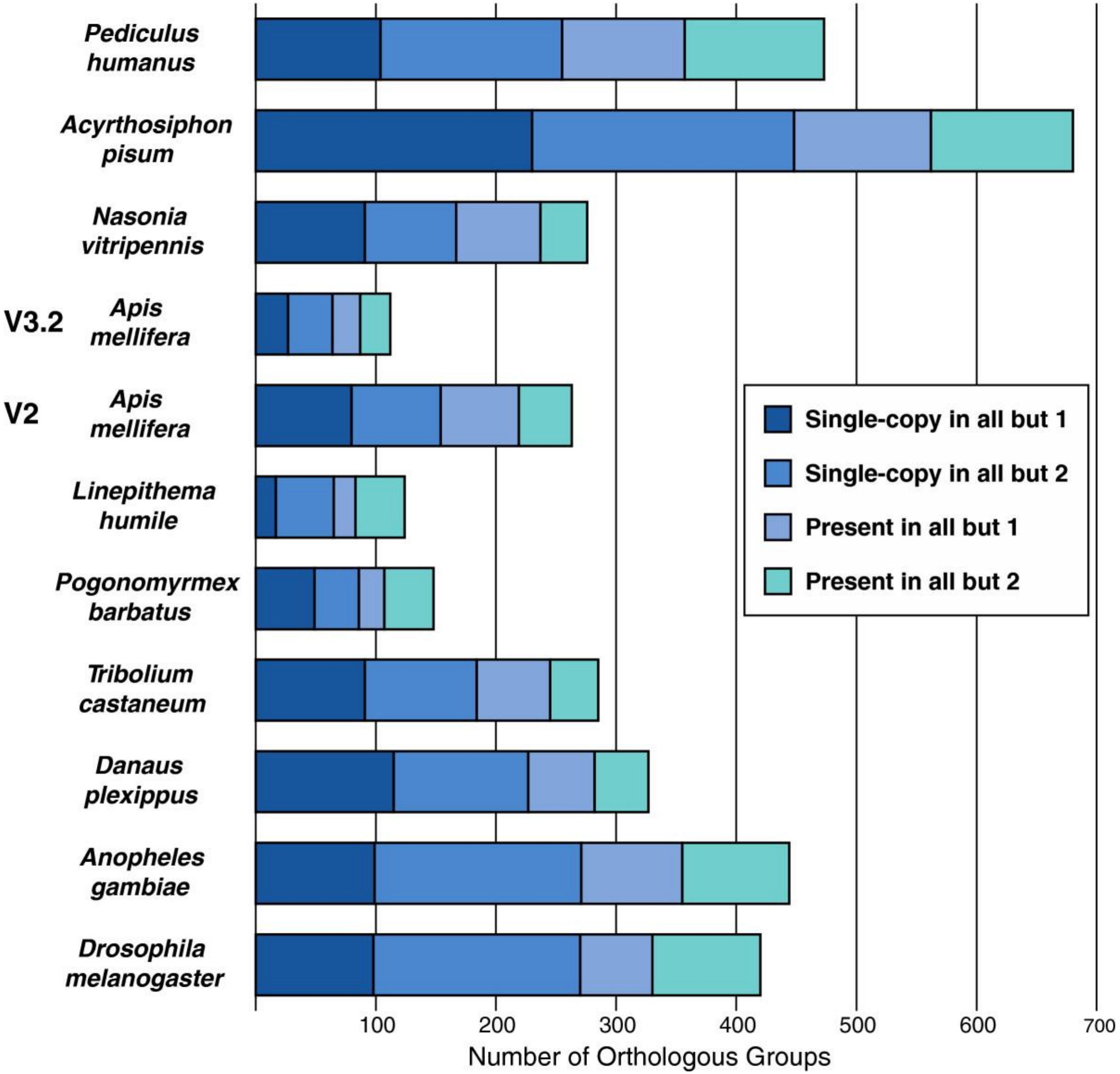
Improving honeybee gene annotations

Using orthology to assess the quality a new annotation set

I.e. you could use BUSCO on the different versions, or in this case you can try to count the numbers of ‘**rare gene losses**’ across a number of species and your versions

‘**rare gene losses**’ can and do happen, but they can also be a proxy for estimating numbers of genes missing from your annotation set as they have orthologs in almost all other species and therefore it is likely that the annotation pipeline missed the gene model rather than the gene being lost

Improving honeybee gene annotations



Using orthology to assess the quality of the new Honeybee gene set annotation:

Fewer 'missing' in all but 1 or 2 species orthologues in new gene set

Arthropod Comparative Genomics

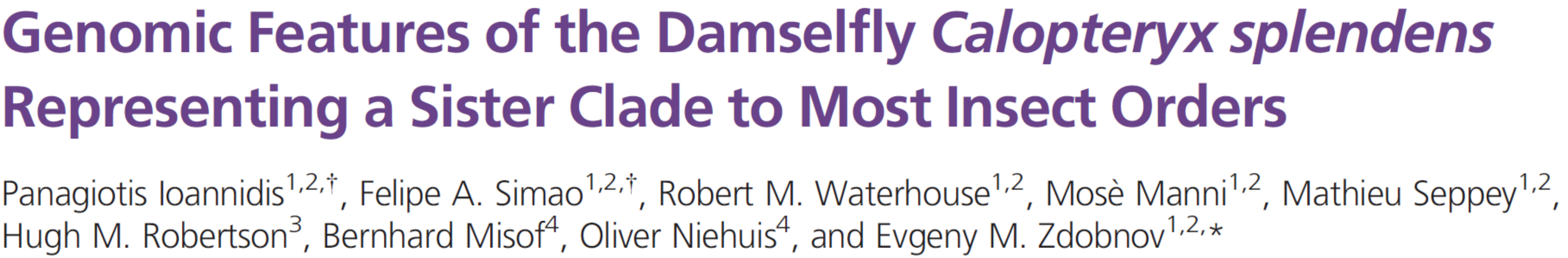
Questions and Approaches through Examples

- ❖ I have predicted a small gene set - why?
- ❖ I have predicted a large gene set - why?
- ❖ Did my gene annotation upgrade work?
- ❖ Phylogenomics without genomes - how?
- ❖ *What are my species/lineage-specific genes doing?*

Odonata phylogeny

Q: What to do when I need a species phylogeny but there are only transcriptomes from other species from my genus/lineage of interest?

Approach: BUSCO genome mode + BUSCO transcriptome mode



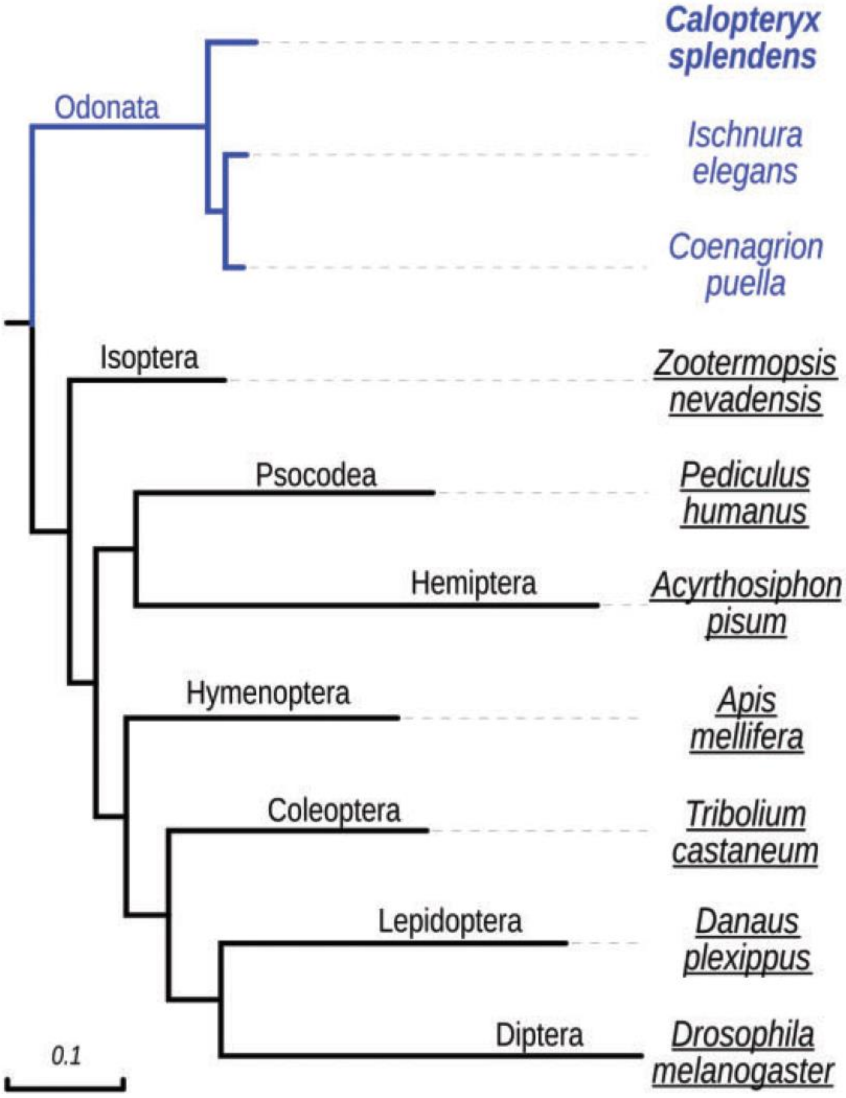
Genomic Features of the Damselfly *Calopteryx splendens* Representing a Sister Clade to Most Insect Orders

Panagiotis Ioannidis[1,2,†], Felipe A. Simao[1,2,†], Robert M. Waterhouse[1,2], Mosè Manni[1,2], Mathieu Seppey[1,2], Hugh M. Robertson[3], Bernhard Misof[4], Oliver Niehuis[4], and Evgeny M. Zdobnov[1,2,*]

Odonata phylogeny

© R.M.Waterhouse

Calopteryx splendens, the banded demoiselle



Single-copy orthologs present in *C. splendens* and each of 8 other arthropods identified from OrthoDB.

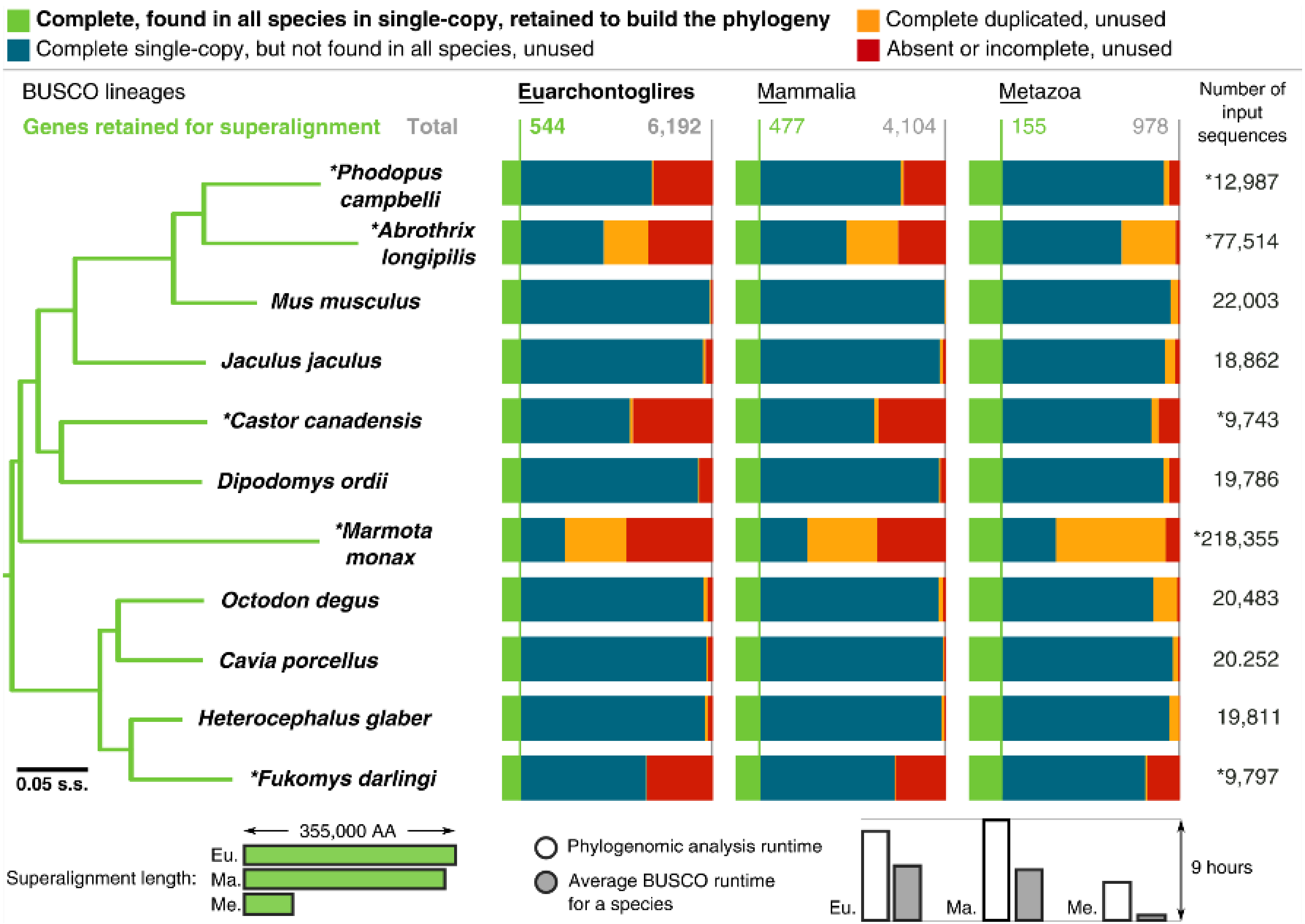
+

BUSCOs from published transcriptomes of azure damselfly, *Coenagrion puella* and the blue-tailed damselfly, *Ischnura elegans*.

154’159 aa alignment for RAxML phylogeny

Phylogenomic Analyses

© R.M.Waterhouse



Mathieu Seppey

Arthropod Comparative Genomics

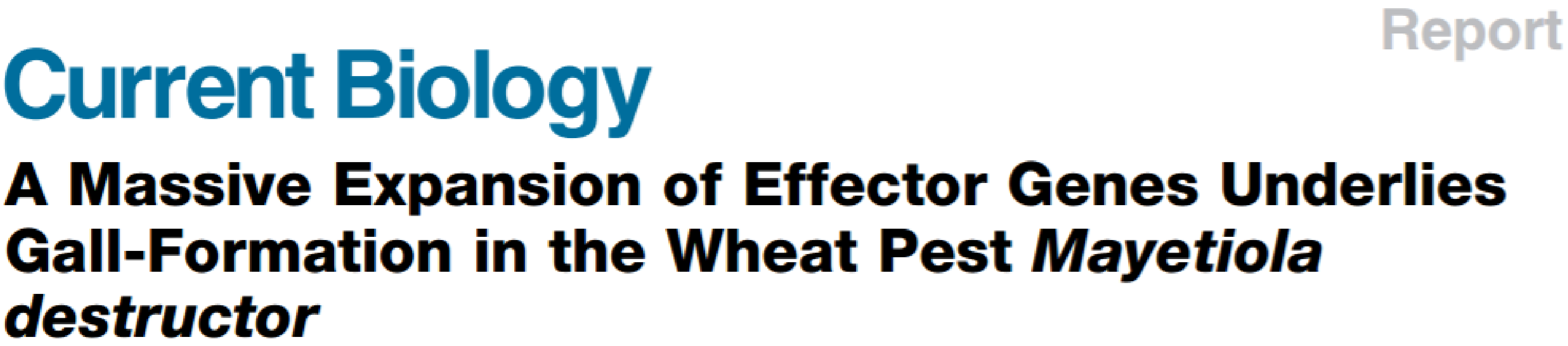
Questions and Approaches through Examples

- ❖ I have predicted a small gene set - why?
- ❖ I have predicted a large gene set - why?
- ❖ Did my gene annotation upgrade work?
- ❖ Phylogenomics without genomes - how?
- ❖ *What are my species/lineage-specific genes doing?*

Hessian fly saliva

Q: What does the hessian fly secrete into its saliva to manipulate wheat?

Approach: salivary-gland transcriptome and comparative genomics



Report

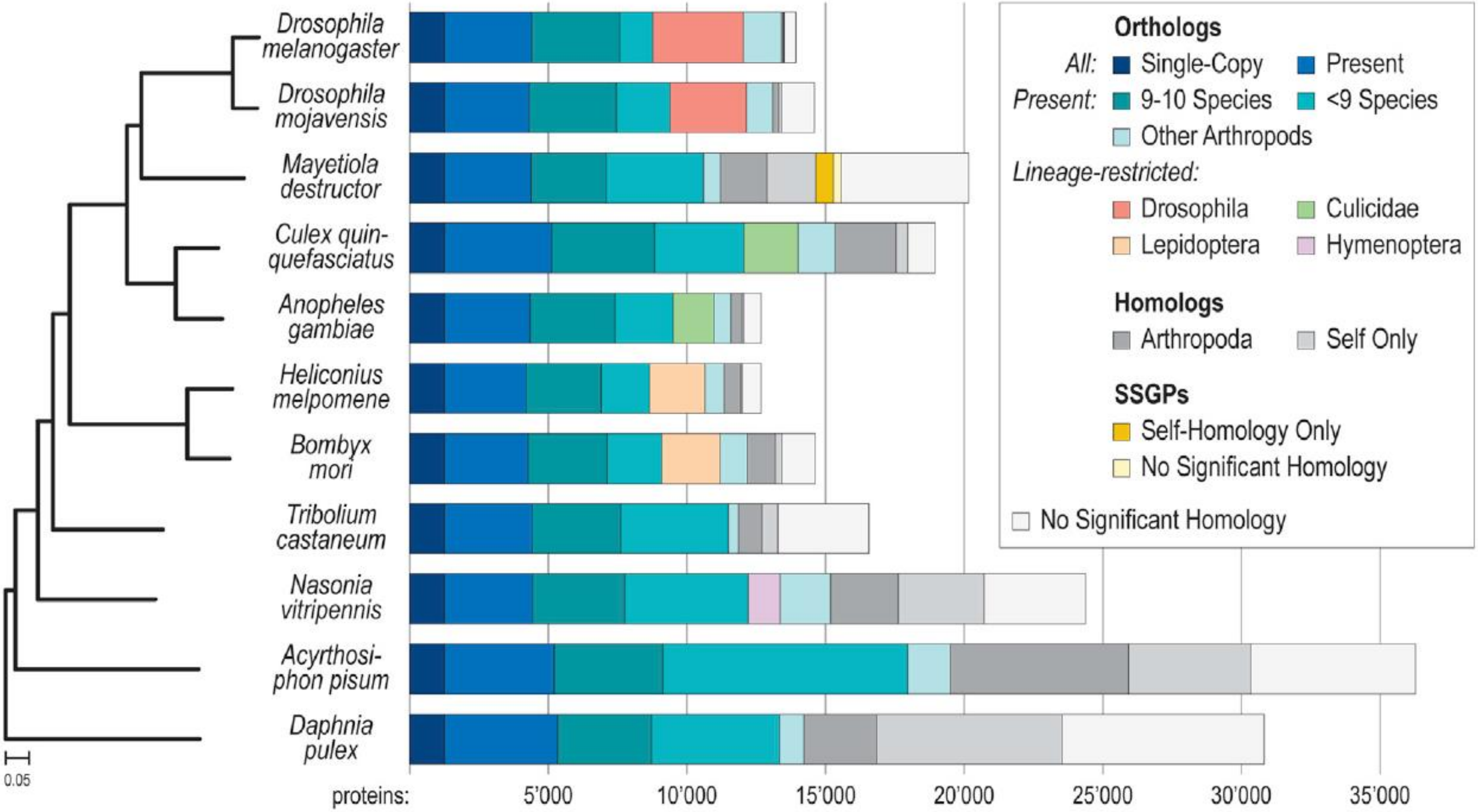
Current Biology

A Massive Expansion of Effector Genes Underlies Gall-Formation in the Wheat Pest *Mayetiola destructor*

Hessian fly saliva

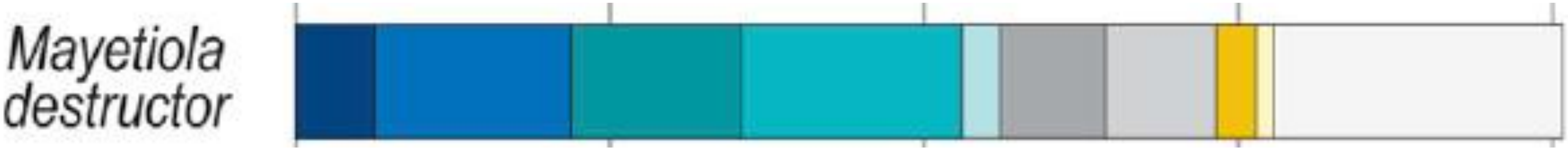
© R.M.Waterhouse

15% of no-orthology genes, i.e. species or lineage –specific genes are homologous to the major salivary gland product



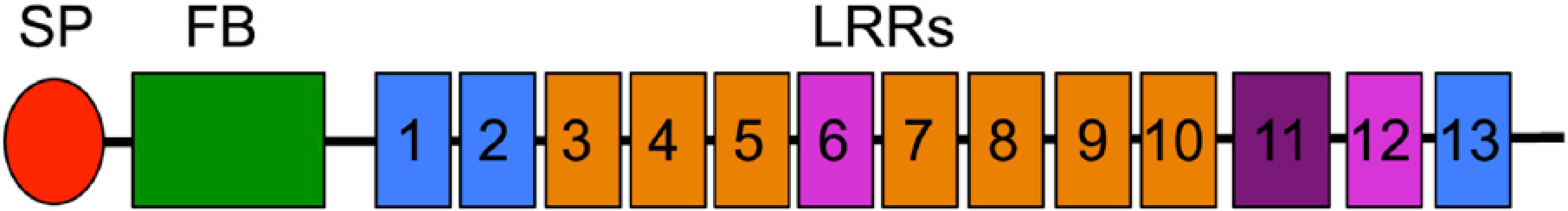
Hessian fly saliva

© R.M.Waterhouse



Compared to most sequenced insect genomes *M. destructor* has a large fraction of genes (34%) lacking homologs in other organisms. Within this fraction, 919 SSGPs had a perfect match with a MAKER2 gene model; 284 were in the single-copy "no-homology" fraction, and 635 were in the multi-copy "self-homology-only" fraction of *M. destructor* genes.

A large reservoir of effector genes to manipulate the host plant.



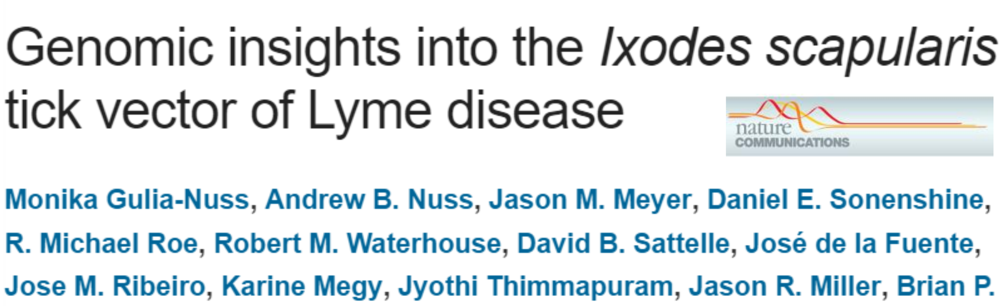
Arthropod Comparative Genomics

Questions and Approaches through Examples

- ❖ I have predicted a small gene set - why?
- ❖ I have predicted a large gene set - why?
- ❖ Did my gene annotation upgrade work?
- ❖ Phylogenomics without genomes - how?
- ❖ *What are my species/lineage-specific genes doing?*

➤ Tick and Mite intron evolution

Tick and Mite intron evolution

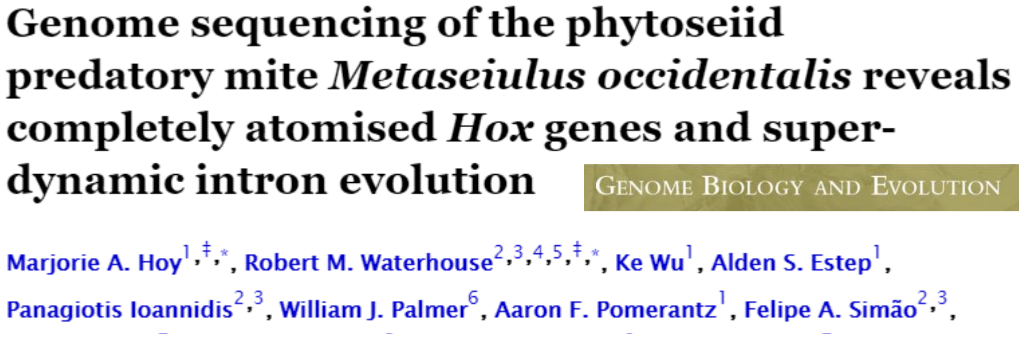


Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease

nature COMMUNICATIONS

Monika Gulia-Nuss, Andrew B. Nuss, Jason M. Meyer, Daniel E. Sonenshine, R. Michael Roe, Robert M. Waterhouse, David B. Sattelle, José de la Fuente, Jose M. Ribeiro, Karine Megy, Jyothi Thimmapuram, Jason R. Miller, Brian P.

The tick genome, therefore, supports an **intron-rich gene architecture** at the base of the arthropod radiation and more similar to that of **ancestral metazoans** than extant pancrustaceans.



Genome sequencing of the phytoseiid predatory mite *Metaseiulus occidentalis* reveals completely atomised *Hox* genes and super-dynamic intron evolution

GENOME BIOLOGY AND EVOLUTION

Marjorie A. Hoy[1,‡,*], Robert M. Waterhouse[2,3,4,5,‡,*], Ke Wu[1], Alden S. Estep[1], Panagiotis Ioannidis[2,3], William J. Palmer[6], Aaron F. Pomerantz[1], Felipe A. Simão[2,3],

Examining gene architectures of ancient universal orthologues to identify shared and unique intron positions revealed **dramatic intron losses** from *M. occidentalis* genes accompanied by **striking numbers of intron gains**.

Tick and Mite intron evolution

Approach:

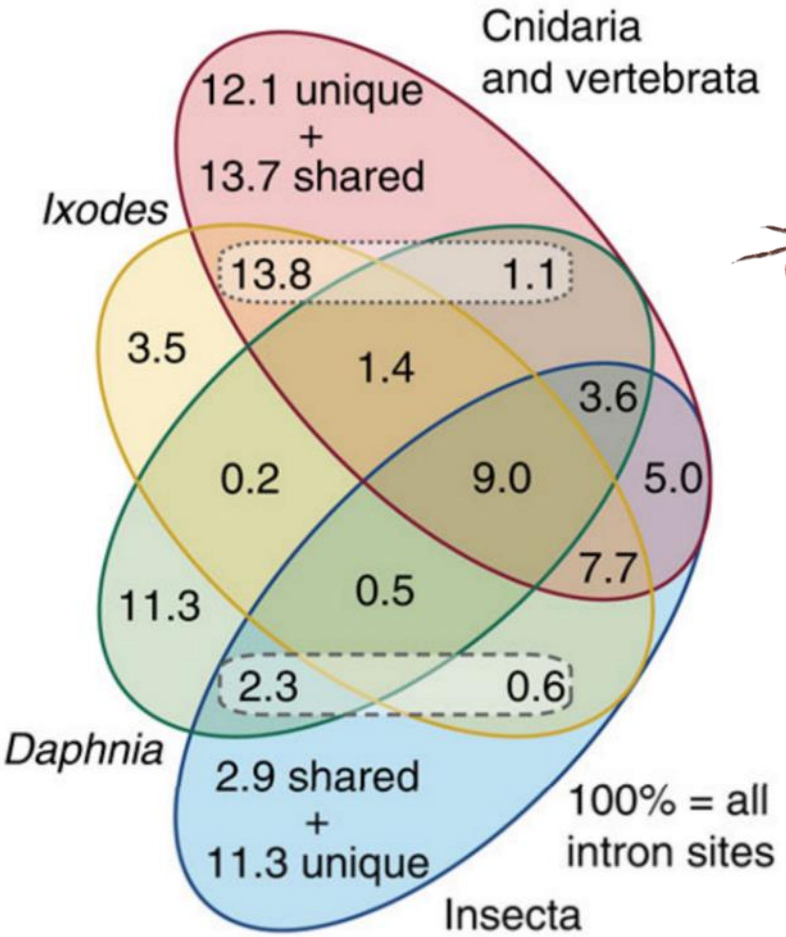
Identify all near-universal single-copy orthologs across a set of species with representatives from different clades

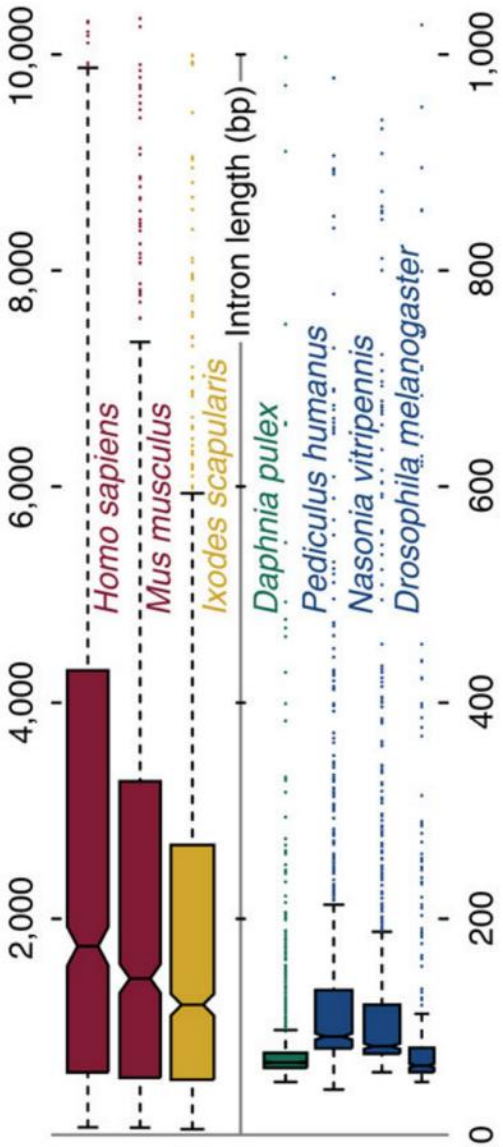
Align the protein sequences and annotate the locations of all underlying intron sites in each protein

Use MALIN: www.iro.umontreal.ca/~csuros/introns/malin

Identification of homologous splice sites in annotated protein sequence alignments.
Computation of primary statistics about introns in homologous sites (shared introns).
Estimation of ancestral intron content, intron losses and gains by Dollo parsimony.
Estimation of intron loss and gain rates in a probabilistic model.
Estimation of ancestral intron content, intron losses and gains in a probabilistic model.
Inference of evolutionary histories at individual sites.
Error estimation for rates and histories by bootstrap.

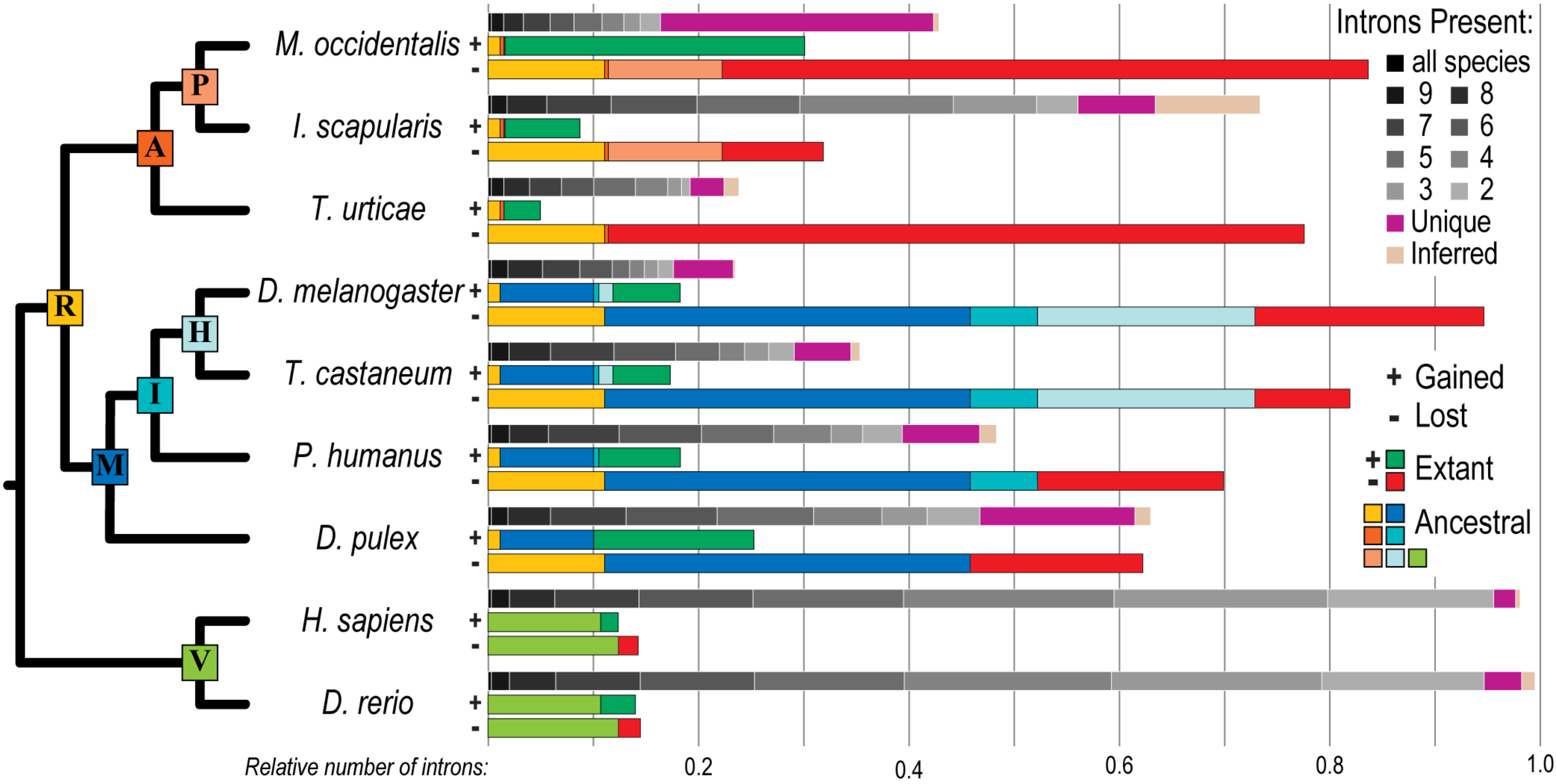
Tick and Mite intron evolution

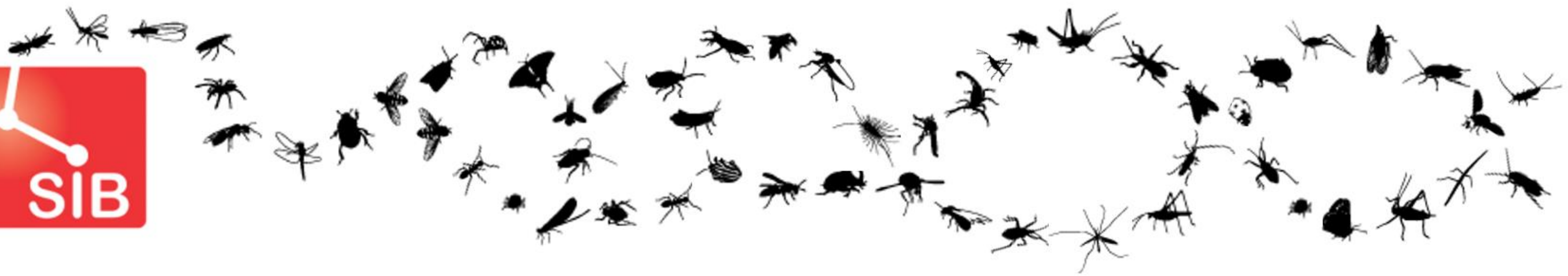




Tick and Mite intron evolution

© R.M.Waterhouse





UniL SIB 3-5 September 2018

Arthropod comparative genomics with OrthoDB & BUSCO

✉ robert.waterhouse@gmail.com @rmwaterhouse www.rmwaterhouse.org