Metagenome skimming of species-rich lineages

Benjamin Linard
b.linard@nhm.ac.uk

Follow me on ResearchGate

Who is this guy ?

Thesis in bioinformatics: IGBMC Strasbourg, bioinfo lab of Julie Thompson, Olivier Poch

→ Java software development
→ Orthology/paralogy inference via algorithmics
→ Online databases development
→ Cluster and network-based visualization of protein-based evolutionary histories

→ comparative genomics of mammals proteomes

Post-doc in environmental genomics: NHM London, phylogenetics lab of Alfried Vogler

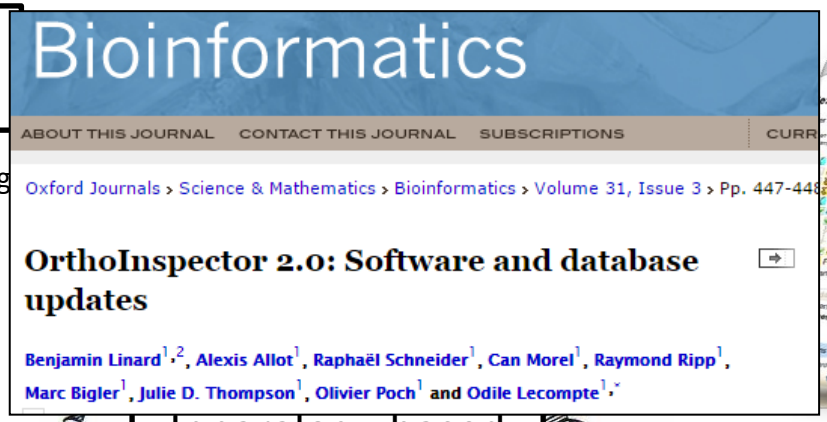
→ NGS data processing
→ metagenomics
→ large-scale taxonomic assignations
→ mitochondria analyses and annotations
→ insect genomics and comparative genomics

→ analyses of species-rich communities (insects and bacterial symbionts)

Thesis work: orthology and paralogy

Interoperability

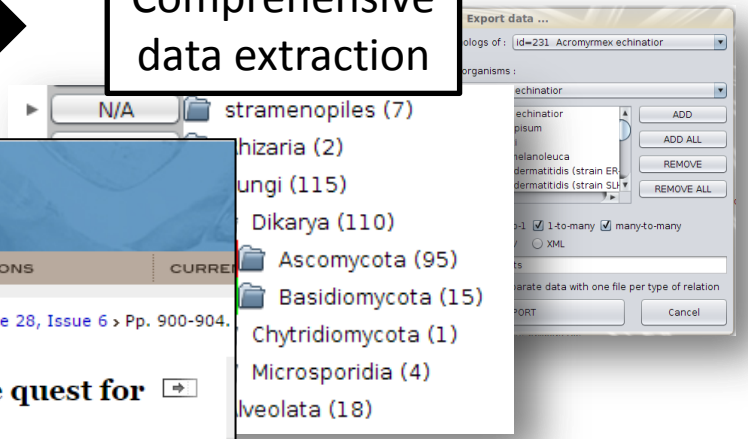
- Compatible with any SQL eng
- Plateform independant
- OrthoXML specifications
- Pipeline friendly



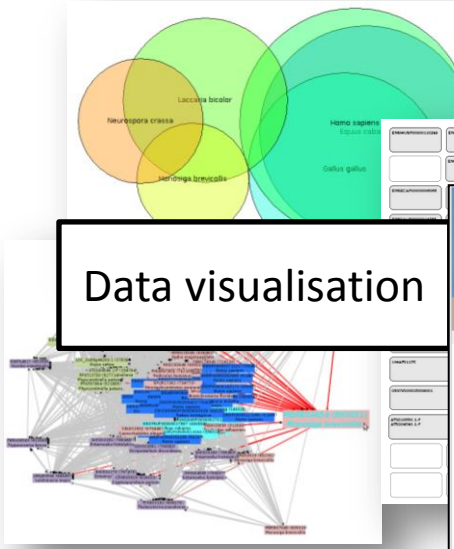
lbgi.fr/orthoinspector/

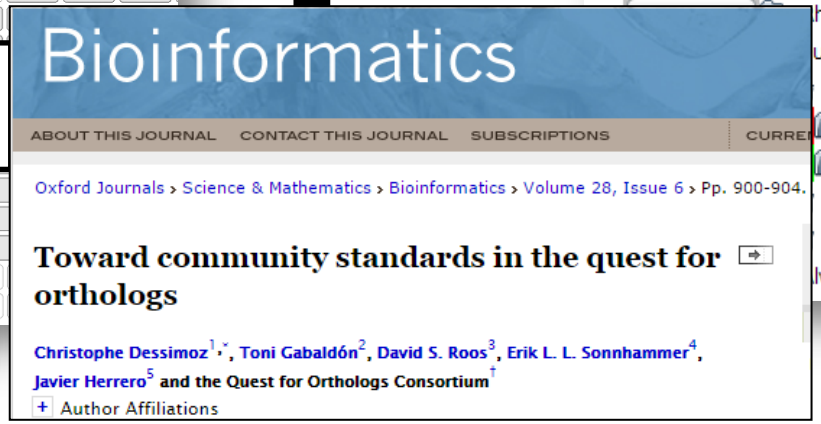
Inparalogy-based Core algorithm

Comprehensive data extraction



Data visualisation

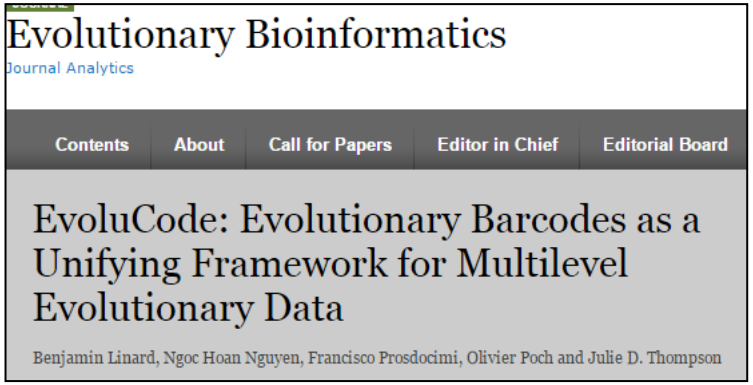


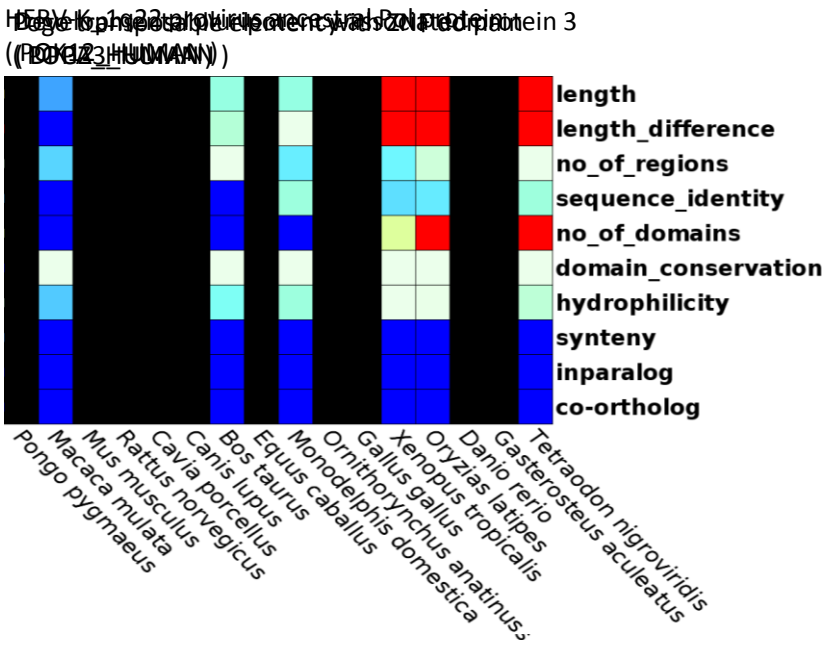


Thesis work: visualization tools

EvoluCodes : Evolutionary Barcodes

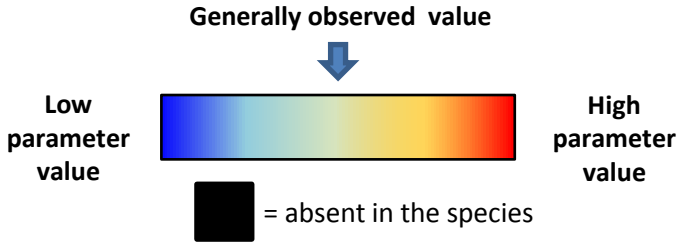
1 gene
1 evolutionary history
1 barcode



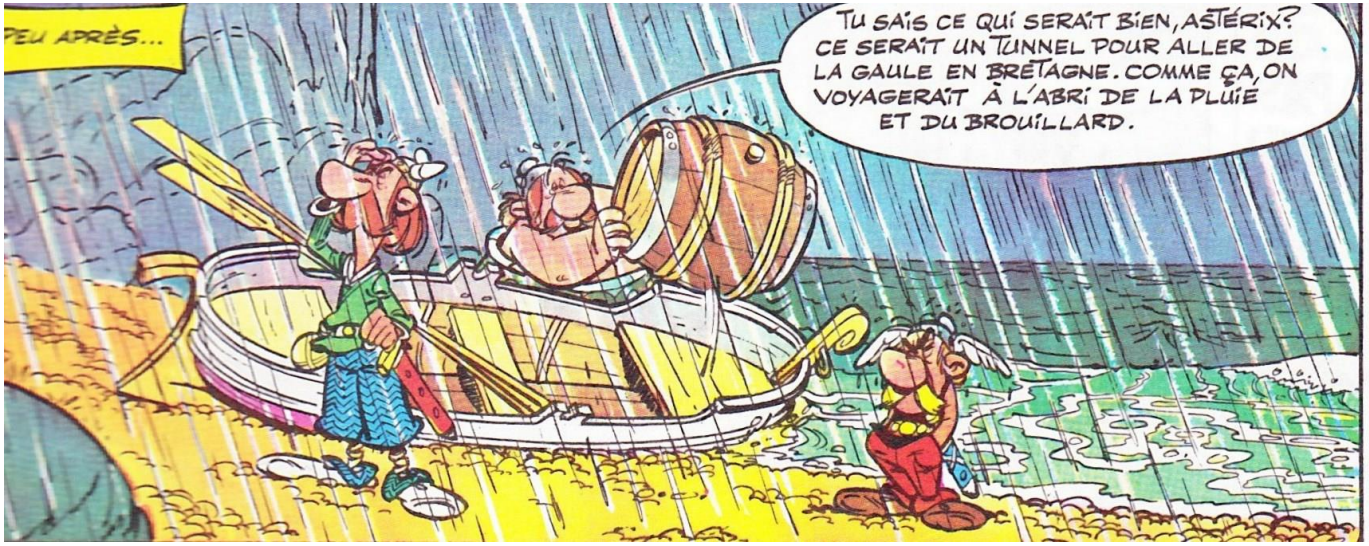


- Integrates multi-scale data
- Describes the variation of a gene
- A framework for knowledge extraction
- Facilitates visualisation of biological messages

Variable repartition in vertebrates, viral DNA integration



Moving to the Natural History Museum



(publié en 1966, l'eurotunnel n'existait pas encore !)

From algorithmics, Java software development and comparative genomics …

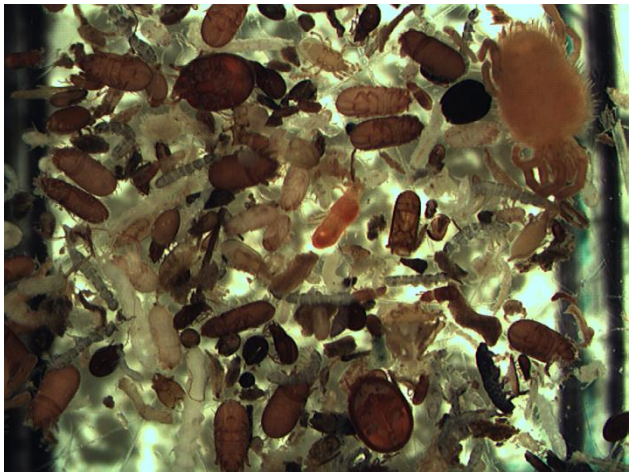
…to the development of new methodologies to generate de-novo RAW data
(and understand them via comparative genomics)

Context

Arthropods in current ecosystems

→ 1,300,000 arthropods species described (80% of all described animals)
estimates suggest 2 to 20 million species (Basset et al. 2012; Zhang 2011)

→ Virtually present in all ecosystems. A tremendous source of biodiversity and genomic diversity!



Credits: C. Andujar, P. Arribas

→ Arthropod diversity is modulated by : habitat health, pollution, climate change.
→ They are an “easy to reach” bioindicator

Context

Arthropod genomics

→ Only dozens of insect complete genome and transcriptomes, most of them Diptera or relevant to human health / pest control

→ only two complete Coleopteran complete nuclear genomes in 2015 :
Tribolium castaneum (Tenebrionoidea) and *Dendroctonus ponderosae* (Curculionoidea) (Friedrich & Muqim 2003; Keeling et al. 2013).

Understanding the black box between
"molecular community" and "ecological community" (Huttenhower & Hofmann 2010)

Big Initiatives…
But slow process

Transcriptomes Initiative

Genomes Initiative

→**Arthropod biodiversity studies generally focus on specific loci, which aretargeted mostly through Metabarcoding approaches**.

Context

What about metagenomics approaches ?

Genome skimming:
Non-targeted and non-selective

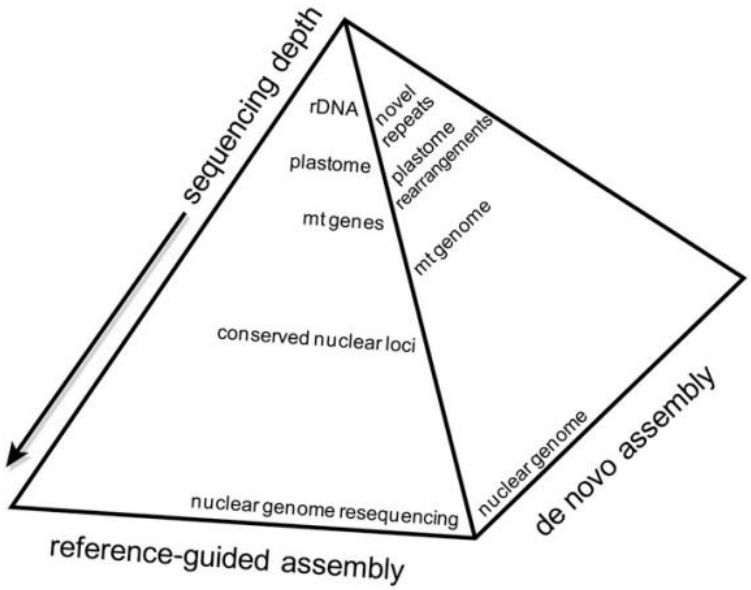
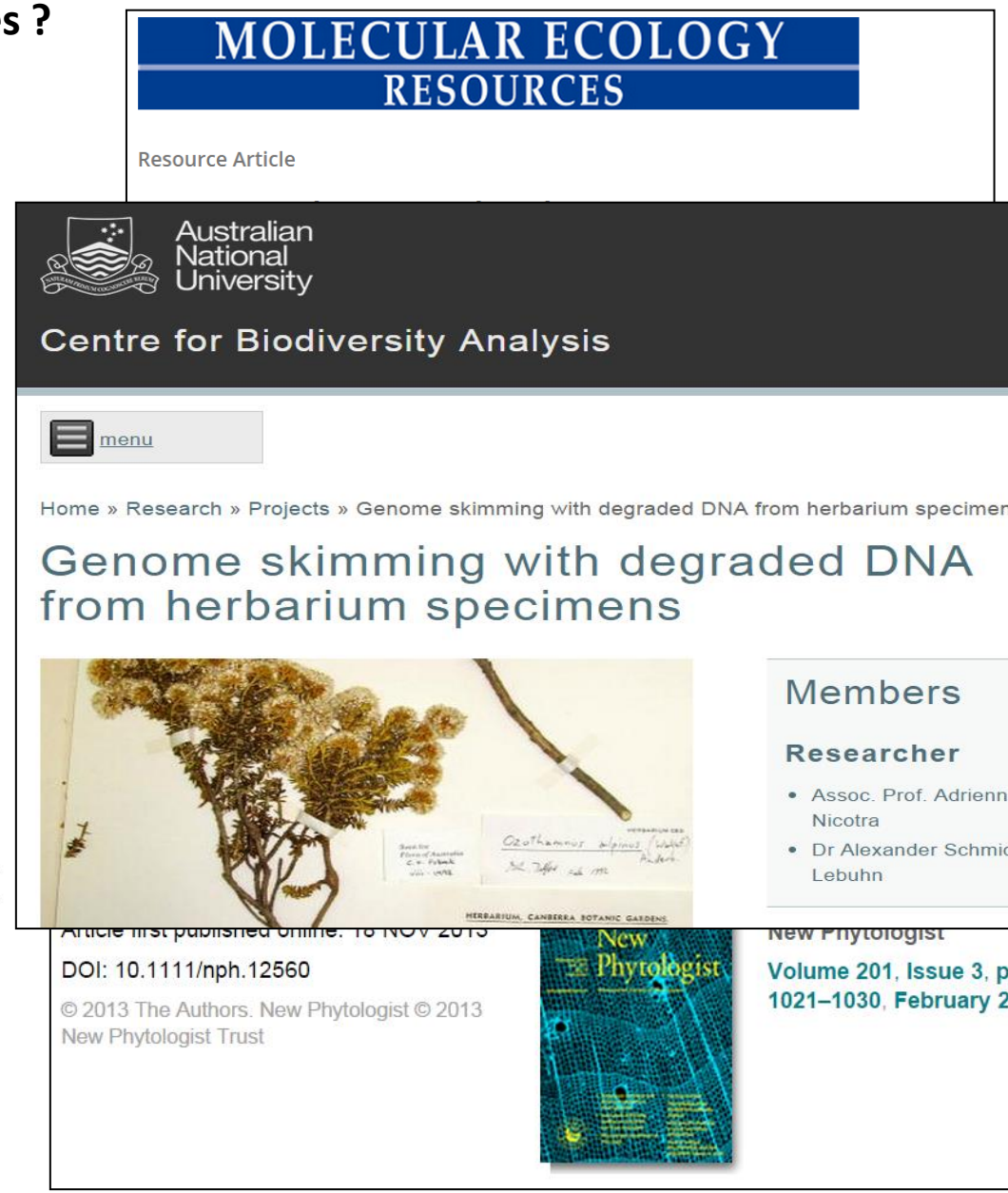


Fig. 1. The genomic iceberg: the relationship between genomic targets, the sequencing depth required to obtain them, and the most appropriate method of sequence assembly.

(Straub et al., 2012)



Context

Genome skimming:
→ shallow sequencing of direct DNA extractions

Non-targeted and non-selective

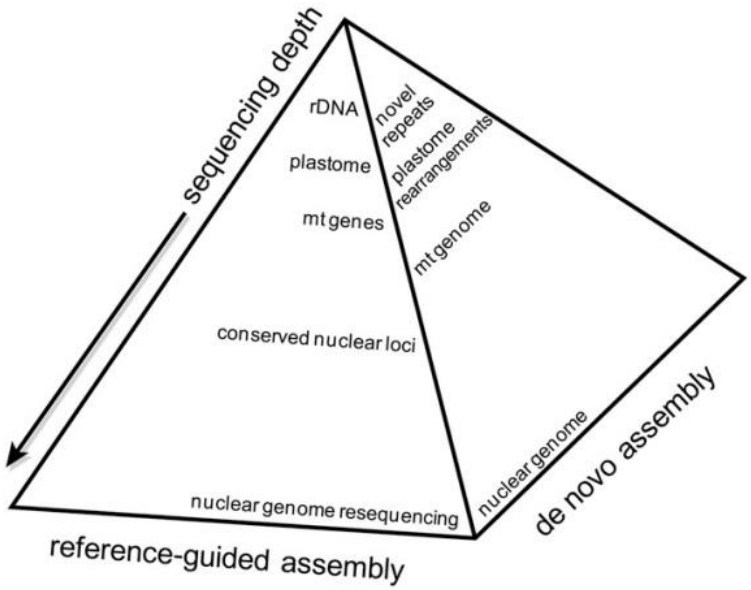
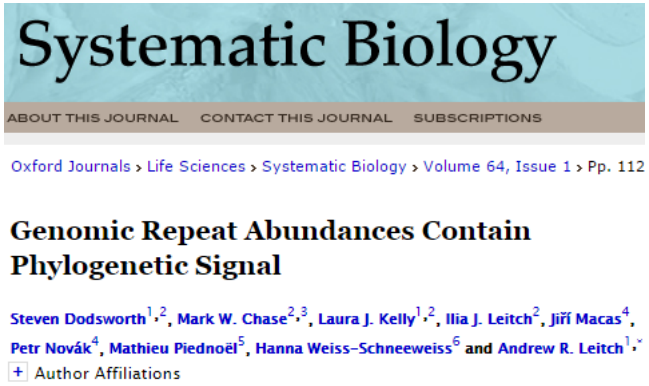


Fig. 1. The genomic iceberg: the relationship between genomic targets, the sequencing depth required to obtain them, and the most appropriate method of sequence assembly.

(Straub et al., 2012)

- **De-novo assembly of the reads**
- Only the most abundant DNA motifs are assembled.
- Organelles are the first obvious outcome (many genome copies per cell)
- Genomic repeats were recently proposed as another source of phylogenetic signal



Mitochondrial metagenomics

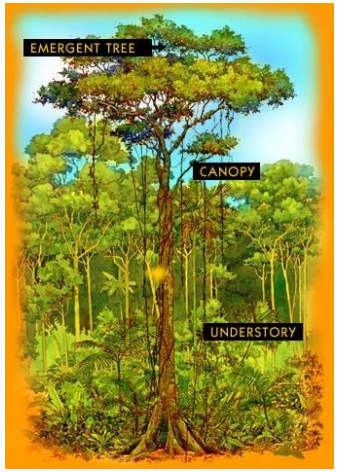
(or mito-metagenomics or mitogenomics…)

Mitochondria DNA is abundant in animal cells, untargeted NGS of arthropods samples around 1% of mitochondrial reads.

Pool of 50 to 300 species
(hundreds of specimens)

→ goal: complete mitochondrial genomes
a proxy to biodiversity and phylogenetic signal

The NHM Biodiversity Initiative:
Reaching rapidly patterns of arthropod biodiversity



n traps per site:
(soil, canopy,
Ground…)

×

N plots

=

Many soup samples...

Mitochondrial metagenomics

Mitochondria DNA is abundant in animal cells, untargeted NGS of arthropods samples around 1% of mitochondrial reads.

Mixed sample

Pool of 50 to 300 species (hundreds of specimens)

All 480 beetles

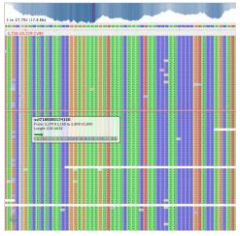
DNA extraction

PCR barcodes

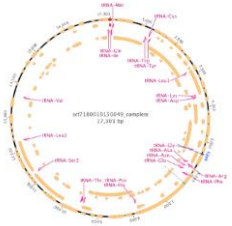
Estimation
→288 species

Pool DNA by volume

1x Illumina MiSeq run (15Gb)



De novo assembly into contigs



Mitochondrial contigs

Assemblers

Some Chimera

Annotations

Mitochondrial metagenomics

(or mito-metagenomics or mitogenomics…)

Mitochondria DNA is abundant in animal cells, untargeted NGS of arthropods samples around 1% of mitochondrial reads.

Biodiversity recovery

Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification

Xin Zhou[123*], Yiyuan Li[123], Shanlin Liu[123], Qing Yang[1], Xu Su[123], Lili Zhou[12], Min Tang[12], Ribei Fu[1],

Oxford Journals › Science & Mathematics › Nucleic Acids Research › Volume 42, Issue

Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics

Min Tang[1], Meihua Tan[1,2], Guanliang Meng[1,3], Shenzhou Yang[1], Xu Su[1], Shanlin Liu[1], Wenhui Song[1], Yiyuan Li[1], Qiong Wu[1], Aibing Zhang[4] and Xin Zhou[1,*]

Community ecology

Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics

Carmelo Andújar[1,2,*], Paula Arribas[1,2], Filip Ruzicka[1,3], Alex Crampton-Platt[1,3], Martijn J.T.N. Timmermans[1,2,†] and Alfried P. Vogler[1,2]

Issue

Molecular Ecology

Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages

Carola Gómez-Rodríguez[1,2,*], Alex Crampton-Platt[1,3], Martijn J. T. N. Timmermans[1,4,5], Andrés Baselga[2] and Alfried P. Vogler[1,4]

Issue

Methods in Ecology and Evolution

Volume 6, Issue 8, pages 883–894, August 2015

Large-scale phylogenetics

Bulk De Novo Mitogenome Assembly from Pooled Total DNA Elucidates the Phylogeny of Weevils (Coleoptera: Curculionoidea)

Conrad P.D.T. Gillett[*,1,2], Alex Crampton-Platt[1,3], Martijn J.T.N. Timmermans[1,4], Bjarte

Soup to Tree: The Phylogeny of Beetles Inferred by Mitochondrial Metagenomics of a Bornean Rainforest Sample

Alex Crampton-Platt[*,1,2], Martijn J.T.N. Timmermans[†,1,3], Matthew L. Gimmel[4], Sujatha Narayanan Kutty[‡,1], Timothy D. Cockerill[1,5], Chey Vun Khen[6] and Alfried P. Vogler[1,3,*]

Pollinators Monitoring

High-throughput monitoring of wild bee diversity and abundance via mitogenomics

Min Tang[1,†], Chloe J. Hardman[2,†], Yinqiu Ji[3,†], Guanliang Meng[1], Shanlin Liu[1], Meihua Tan[1,4], Shenzhou Yang[1], Ellen D. Moss[5], Jiaxin Wang[3], Chenxue Yang[3], Catharine Bruce[6], Tim Nevard[7,8], Simon G. Potts[2], Xin Zhou[1,*] and Douglas W. Yu[3,6,*]

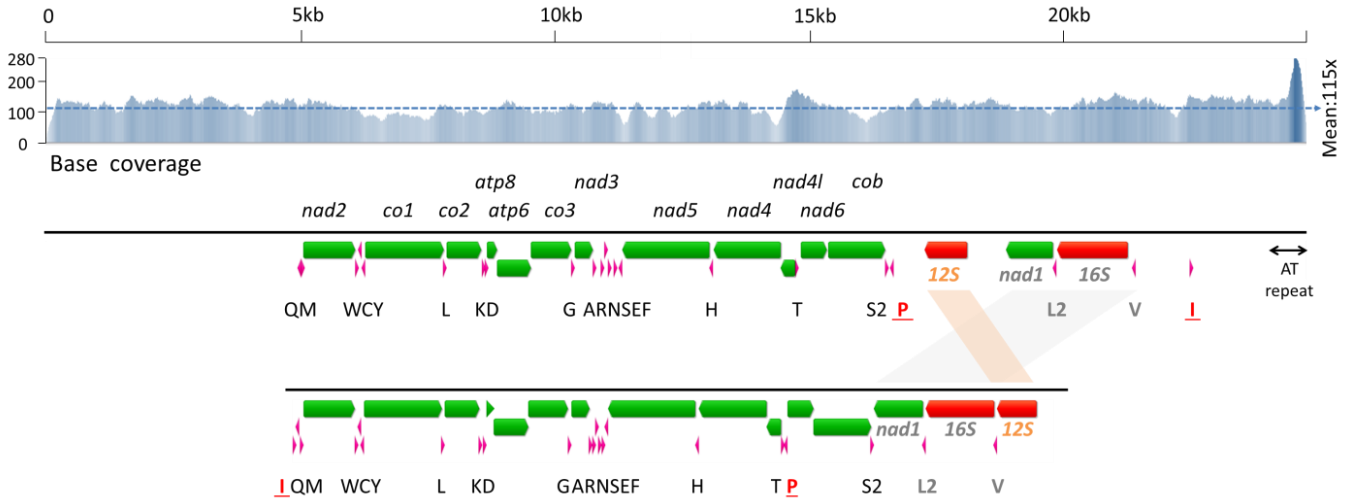
Issue

Methods in Ecolo Evolution

Volume 6, Issue 9 1034–1043, Septe

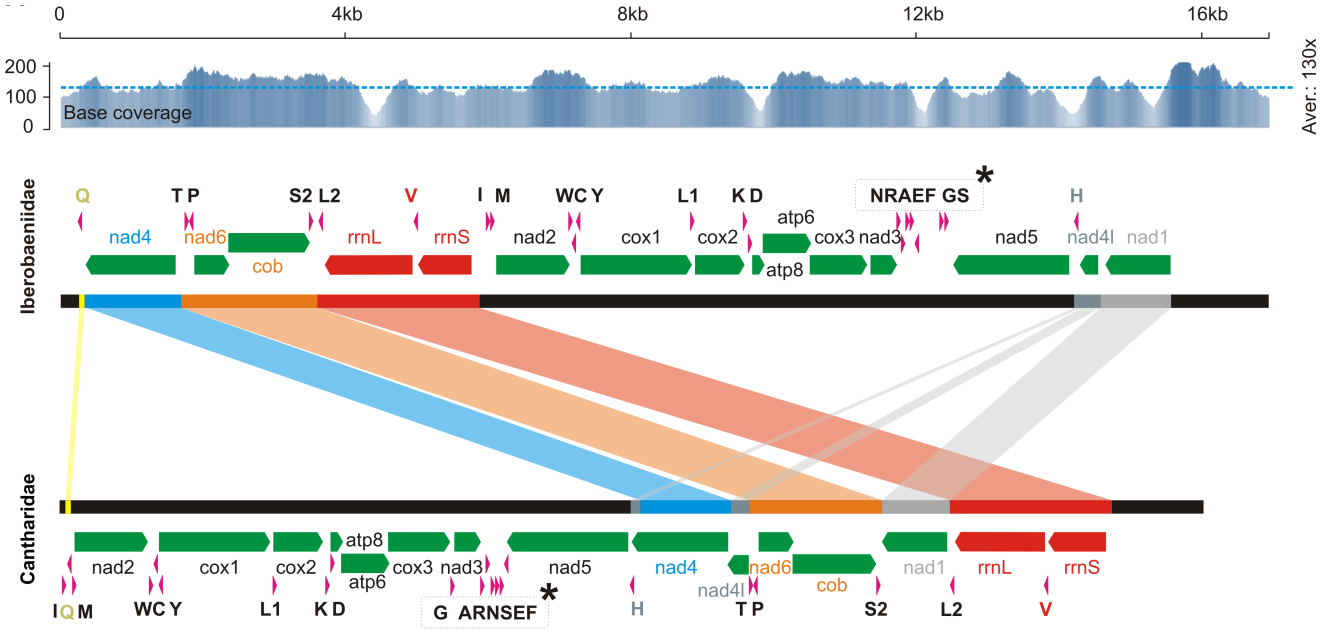
(parenthesis) Mitochondrial genome evolution

Detection of complex mitochondrial rearrangements that explain some Barcoding failures



Hydropsyche pellucidula
(Hydropsychidae)
first gene rearrangement in the insect order Trichoptera

Linard B, et al.
Mitochondrial DNA, 2015



Iberobaenia
(Coleoptera: iberobaeniidae)
First rearrangement of protein coding genes in the beetles

Andújar C et al.
(in review)

Skimming arthropod communities

Skimming an animal community will use the same sampling principles (high-copy motifs first) but has also specific characteristics … **it's a META Genomic Skimming (MGS)**

	Genome skimming on plant pools	Meta-mitogenomics on insect pools
PCR-free sequencing (organelles & genome)	○	○
Shallow sequencing	○	○
Strategy	**(Generally) Multiplexed sequencing**	**1 extraction for the whole pool, Anonymous reads**
# of morphospecies	1	Many
Phyletic diversity	Low	High
Genome complexity (per specimen)	High	Low
Chloroplast	○	✗
Gut content	✗	○

Problematic

Our approach is similar to previous Genome Skimming works,
but with specific characteristics: "Metagenome skimming" (MGS) …

Which genomic elements are extensively sampled from an arthropod metagenome ?

Can we recover information from gut contents ?

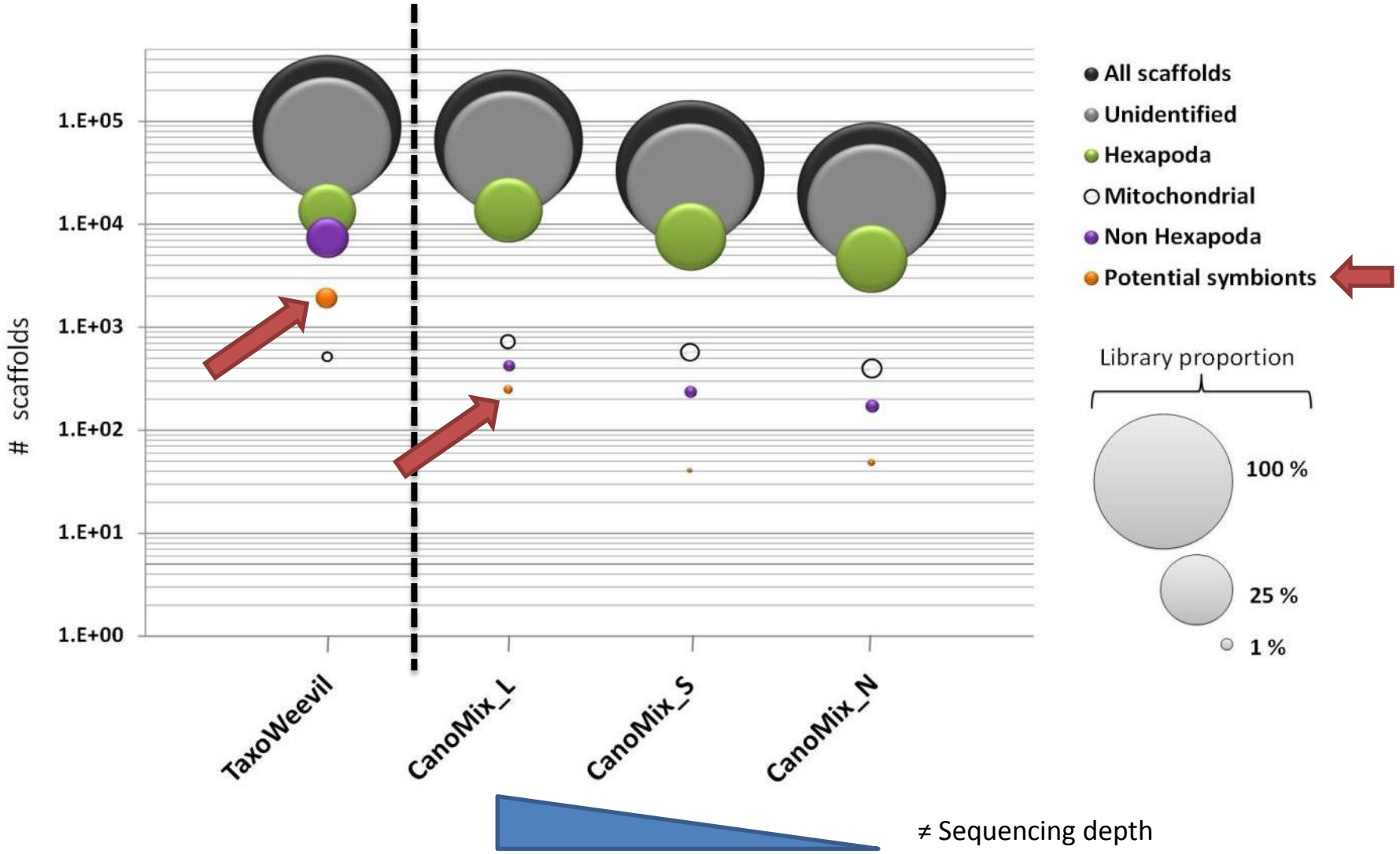
Can MGS teach us something about arthropod genomes and evolution ?

Samples: Field capture or taxon assemblage

	Content			
Sample	Specim-ens	Morpho-species	Super-families	Read pair
TaxoWeevils	173	173	1	17389929
CanoMix	480	212	14	23922520

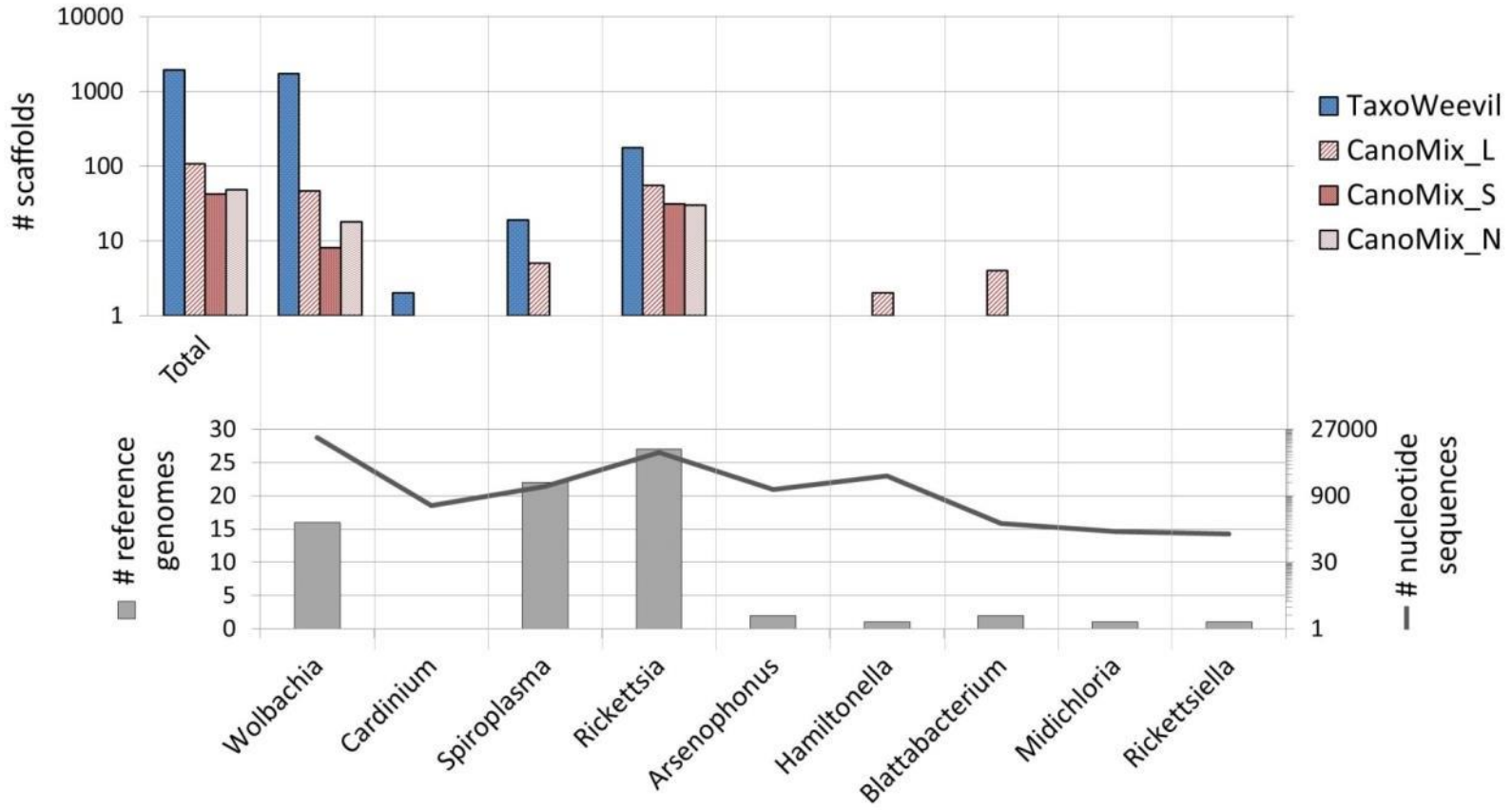
DNA Scaffolds identification

- Annotation by homology to 3 complete NCBI databases (nt, est, genomes)
- Categorized by their best blast hits



Metagenome skimming and associated fauna

- Large sampling of associated bacterial symbionts :



- (Canopy sample only) cherry on the cake … Plastids and rRNAs scaffolds

Diet remnants ?

* Dozens of Cocoa family chloroplasts in Canopy sample

Parasites ?

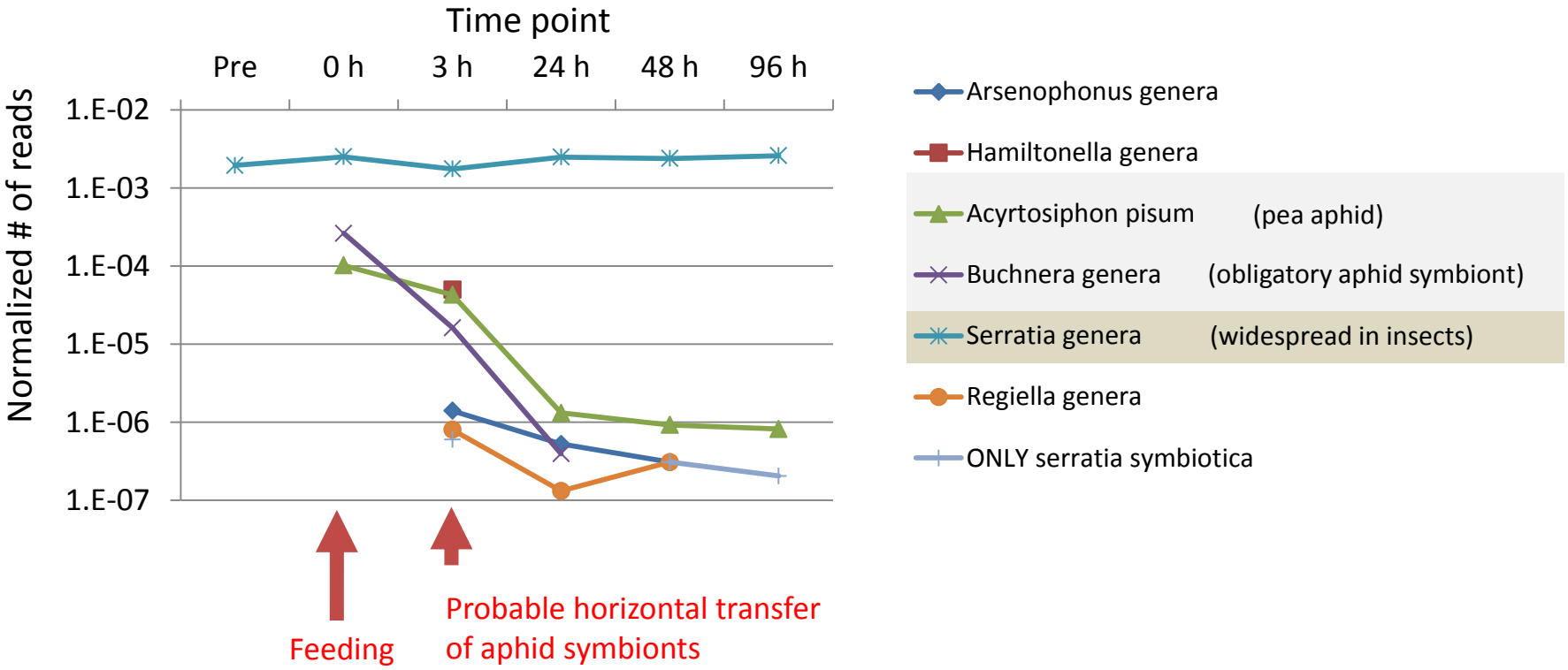
* 7kb nematode scaffold
* rRNA 99.5% similar to *Glarea lozoyensis* (fungi, insect pathogen)

Detecting bacterial symbionts via total DNA extraction

→ **Harmonia axyridis (coleoptera:coccinellidae)**

- Lady bird fed with 1 aphid at 0 h
- DNA decay monitored in the gut
  by shallow sequencing at different time points

Collaboration with
Debora Pires-Paula,
Embrapa Genetic Resources and Biotechnology, Brazilia



Bayesian models of food decay

	Elapsed time after feeding					
	Pre	0 h	3 h	24 h	48 h	96 h
A. pisum mtDNA	0	27	15	0	0	0
A. pisum nuclear DNA	0	624	**214**	**10**	**6**	**4**
B. aphidicola	0	**1,651**	171	2	0	0

Detection and decay rates of prey and prey symbionts in the gut of a predator through metagenomics

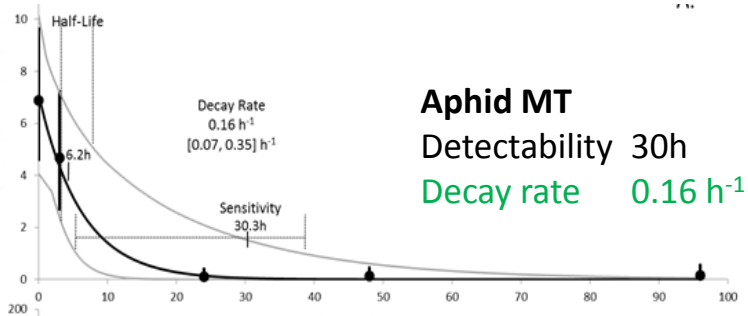
Débora P. Paula[1,2,*], Benjamin Linard[2], David A. Andow[3], Edison R. Sujii[1], Carmen S. S. Pires[1] and Alfried P. Vogler[2,4]

Issue

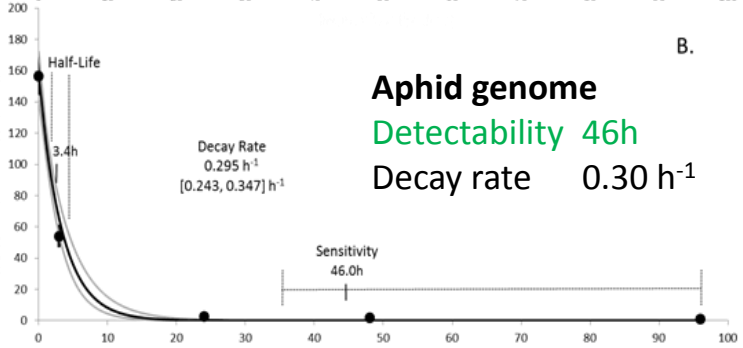
MOLECULAR ECOLOGY RESOURCES

Molecular Ecolo Resources

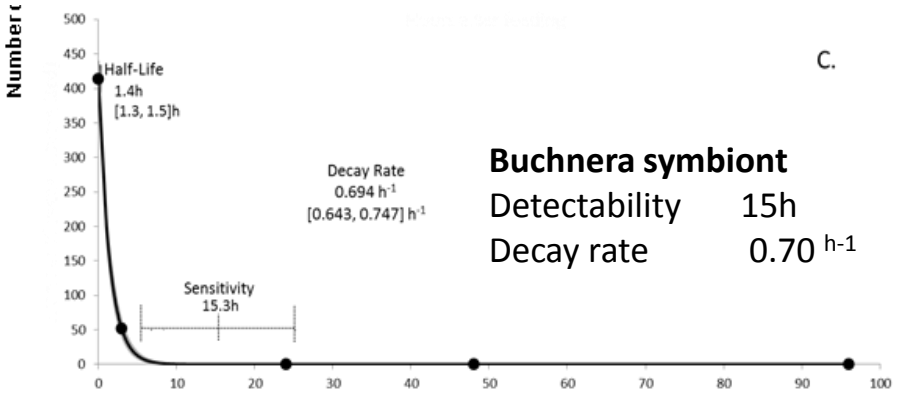
Volume 15, Issue



Aphid MT
Detectability 30h
Decay rate 0.16 h^{-1}



Aphid genome
Detectability 46h
Decay rate 0.30 h^{-1}



Buchnera symbiont
Detectability 15h
Decay rate 0.70 h^{-1}

Gut content used for intraguild predation analysis

Debora's second project ***Trophic interaction between carnivorous Ladybirds From Brazilian agrocultures***

Coccinellini

Cycloneda sanguinea

Coccinellini

Harmonia axyridis

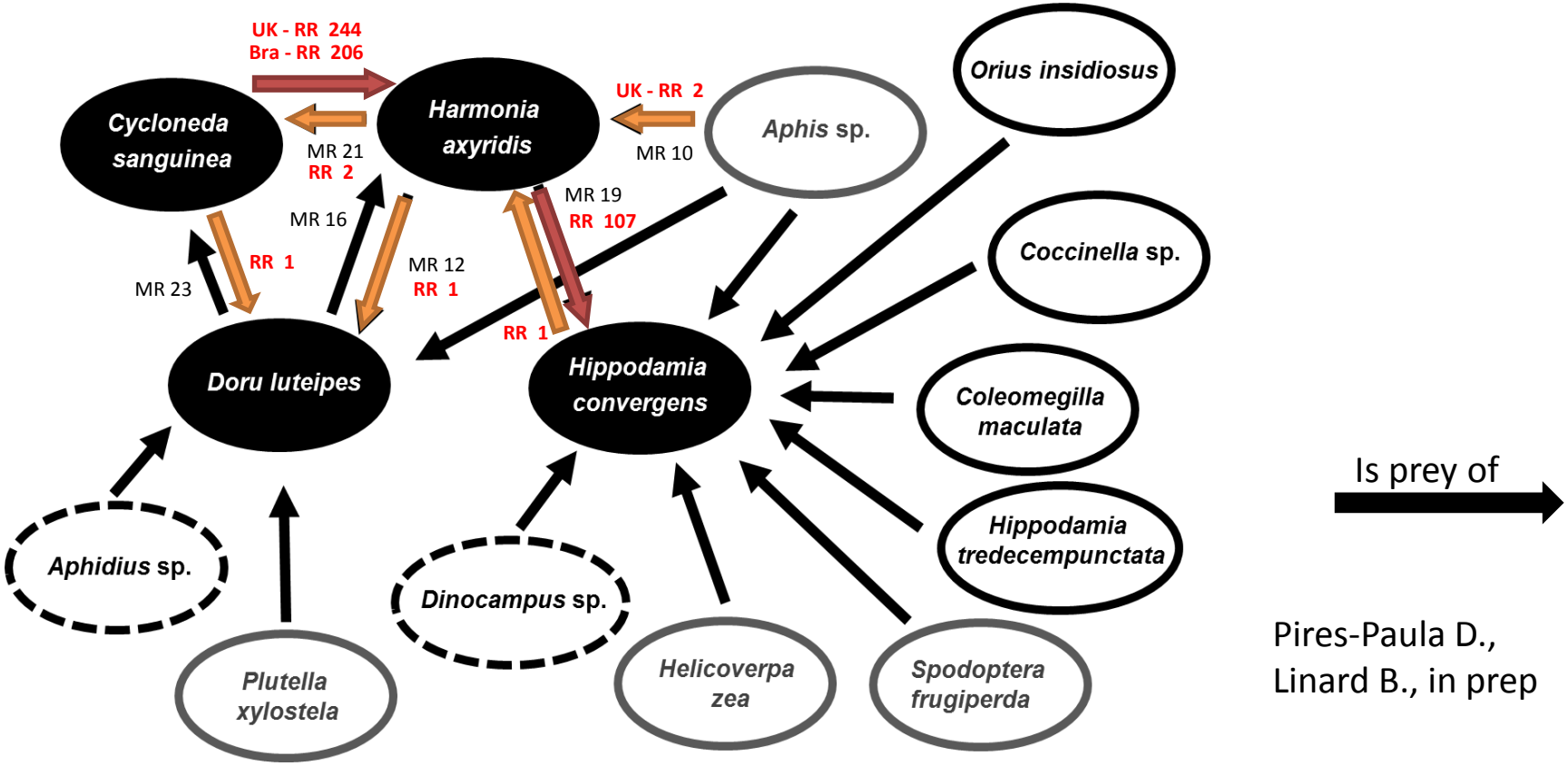
Doru luteipes

Earwig
(Dermaptera)

Hippodamia convergens

Coccinellini

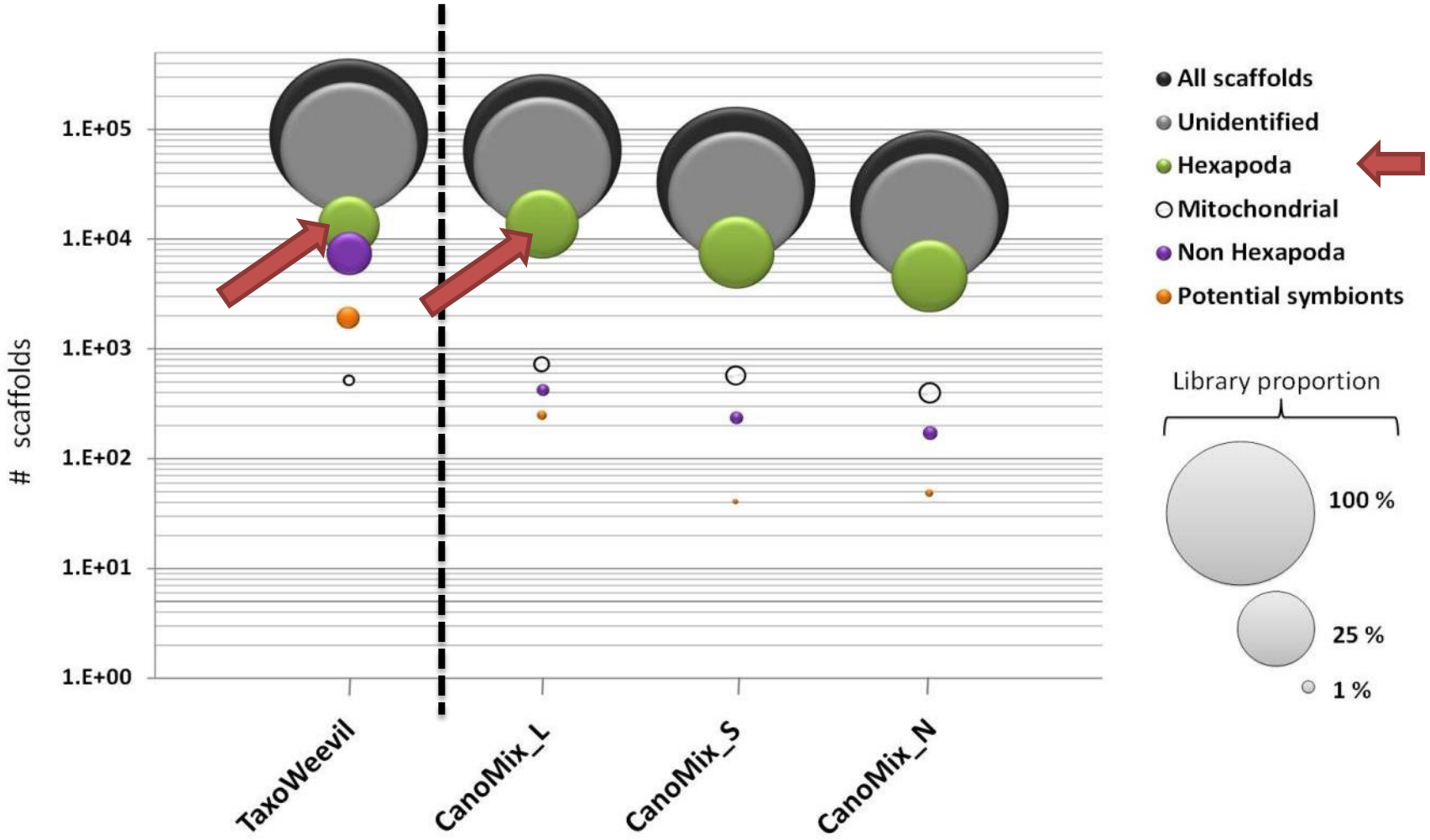
Building an exploratory food web



Pires-Paula D.,
Linard B., in prep

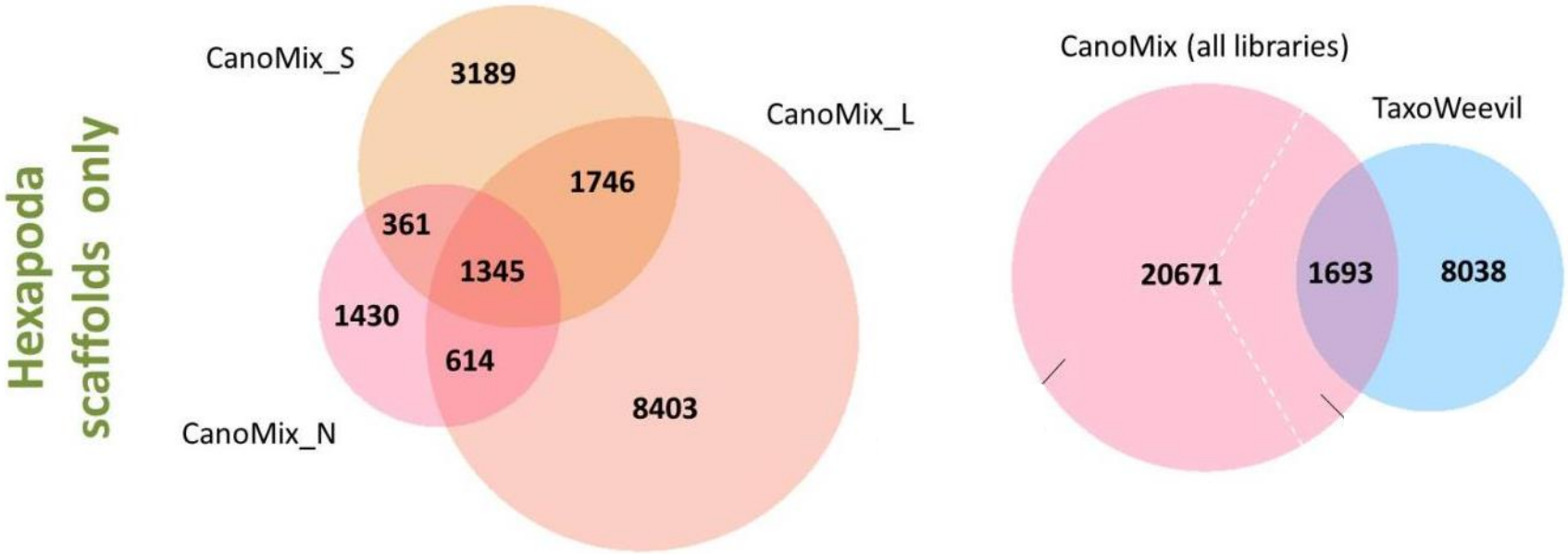
DNA Scaffolds identification

- Annotation by homology to 3 complete NCBI databases (nt, est, genomes)
- Categorized by their best blast hits



Nature of Hexapoda scaffolds

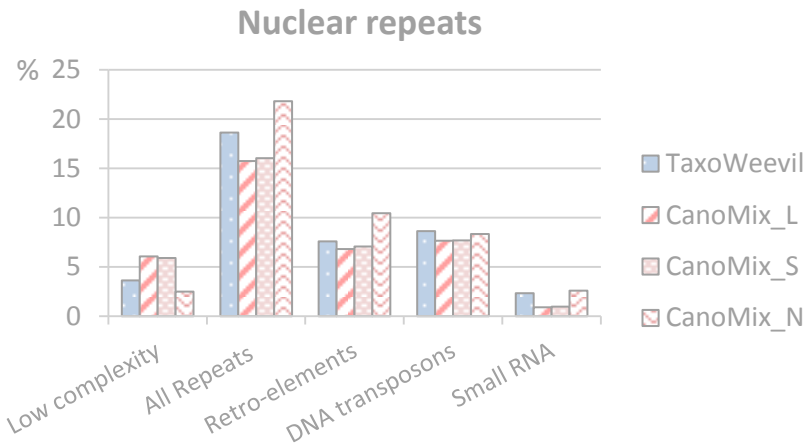
- Library intersections:

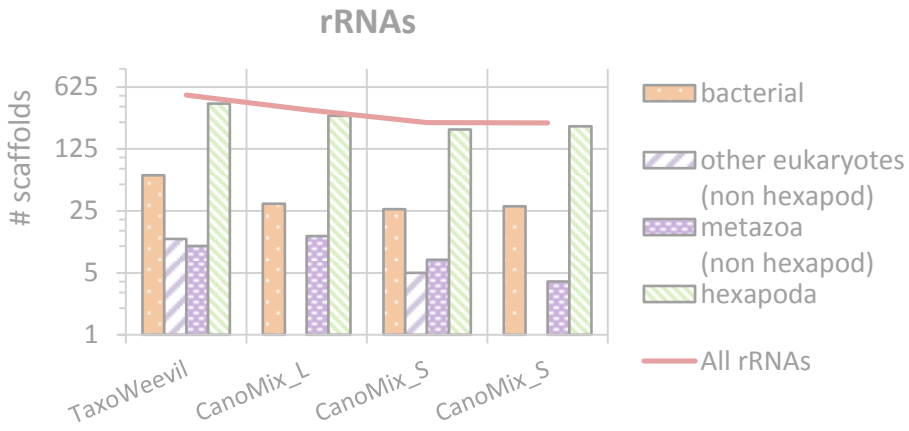


→ **A core set of sequences is systematically recovered**

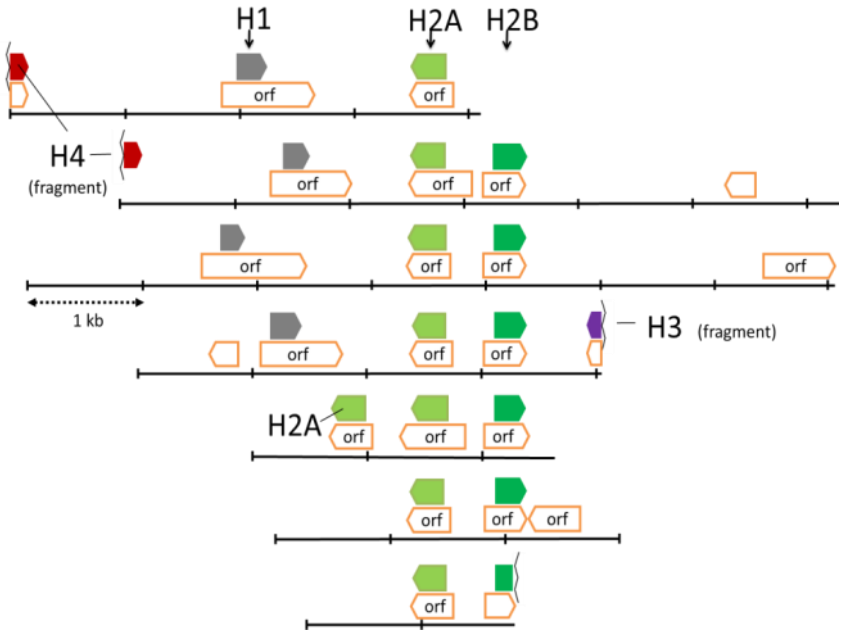
Nature of Hexapoda scaffolds

- Library intersections: → **A core set of sequences is systematically recovered**
- Nature of the sequences: during MGS, we expect to sample multicopy elements.





Most abundant coding genes:
Histone gene clusters



→Potential phylogenetic marker (Talbert et al. 2012)

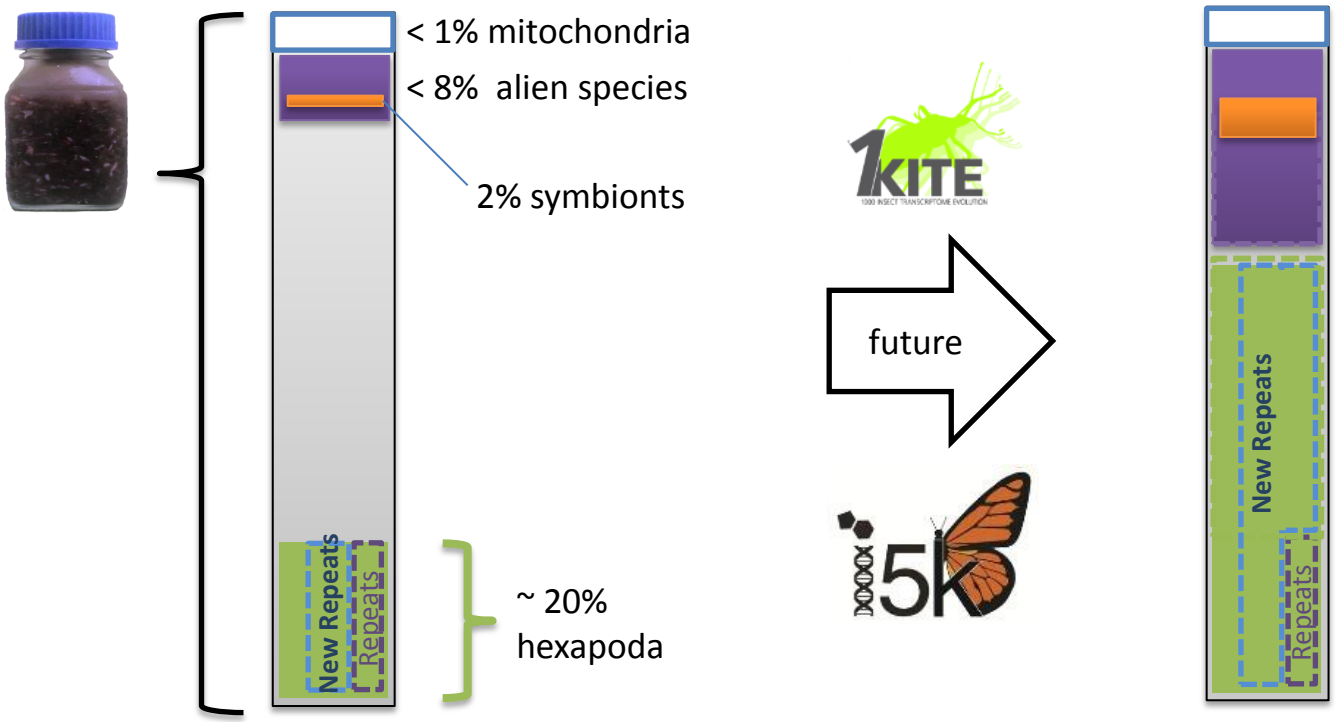
→Gene sequence + cluster rearrangements

Metagenome skimming of arthropod specimen pools

What we identify in one pool:

Metagenome Skimming of Insect Specimen Pools: Potential for Comparative Genomics

Benjamin Linard[1], Alex Crampton-Platt[1,2], Conrad P.D.T. Gillett[1], Martijn J.T.N. Timmermans[1,3] and Alfried P. Vogler[1,3,*]



- Up to 42% of insect genomes are repetitive (Wang et al. 2008)
- Less than 0.15% of repeats found to be homologous between the two coleopteran genomes sequenced to date (Keeling et al. 2013)

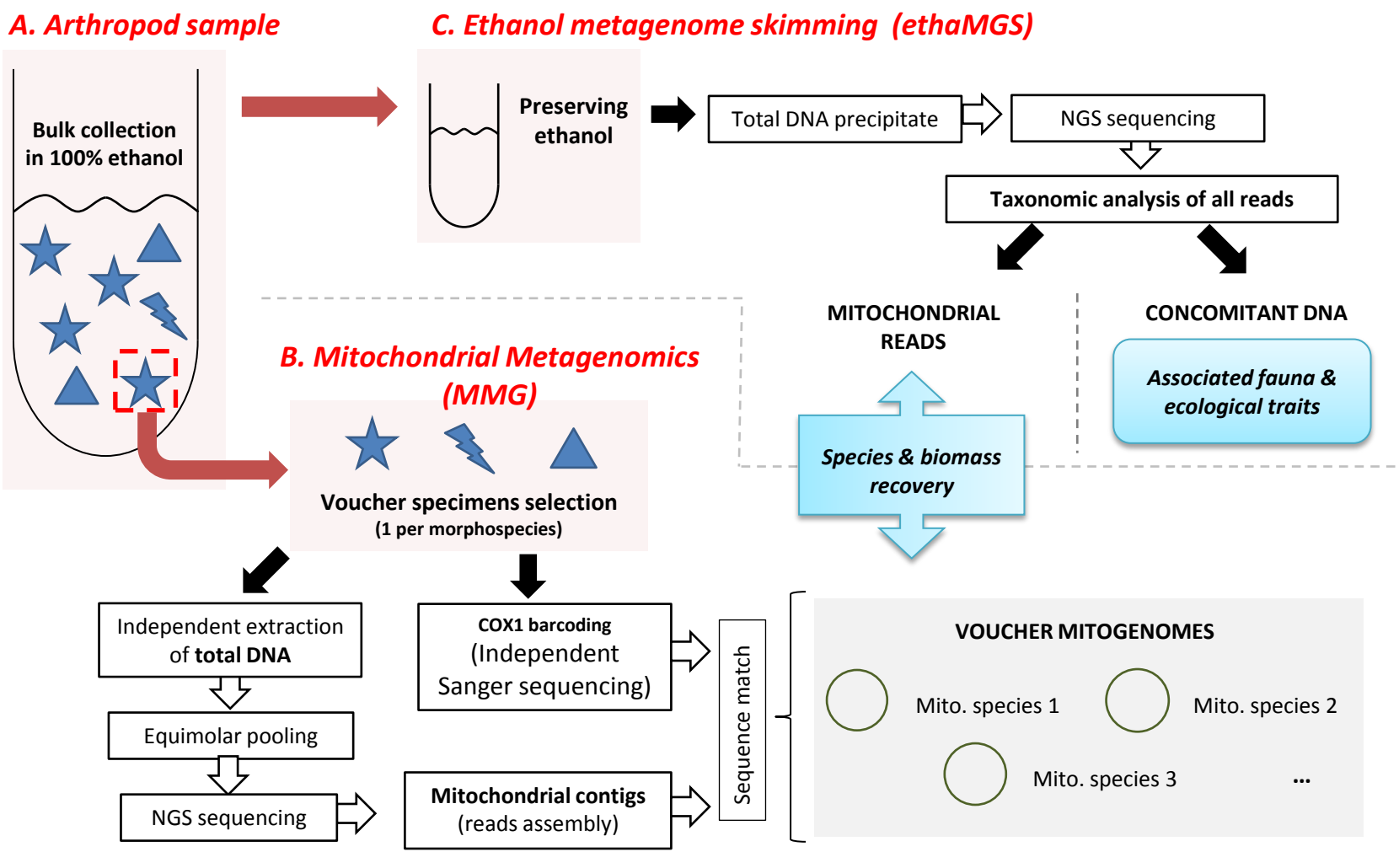
→ clade-specific nuclear repeats are the most sampled sequences

Metagenome skimming of preserving ethanol

MGS is semi-destructive, what if new undescribed species were collected ?

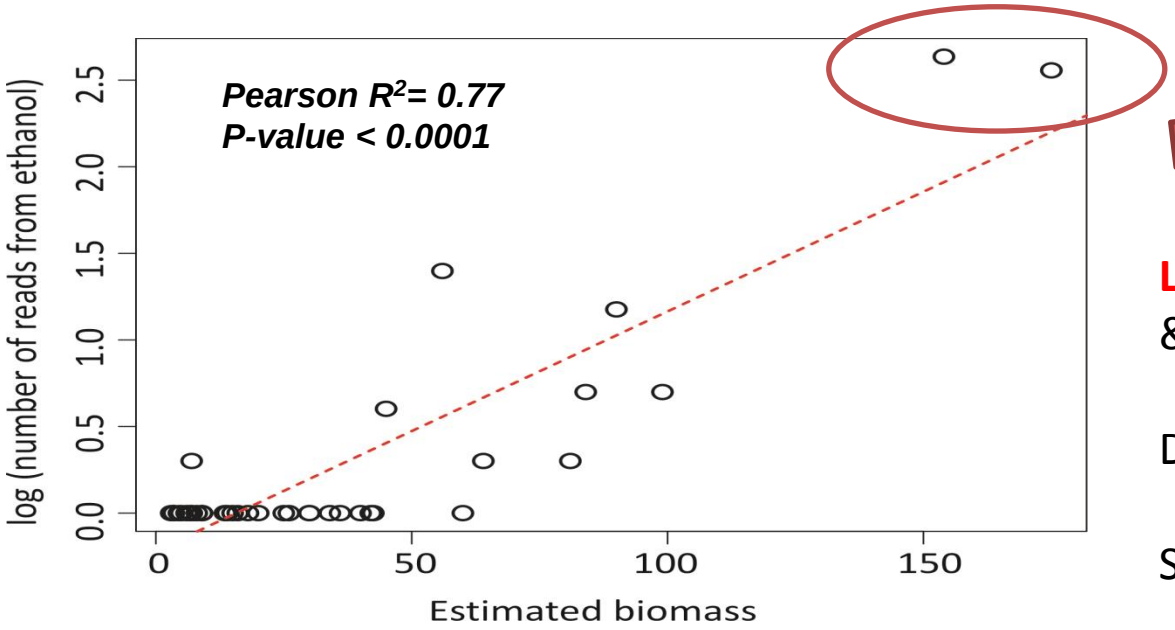
Wait… all our communities are collected in 100% ethanol !

Hajibabei et al., 2012 -> high species recovery via metabarcoding of preserving ethanol



Metagenome skimming of preserving ethanol

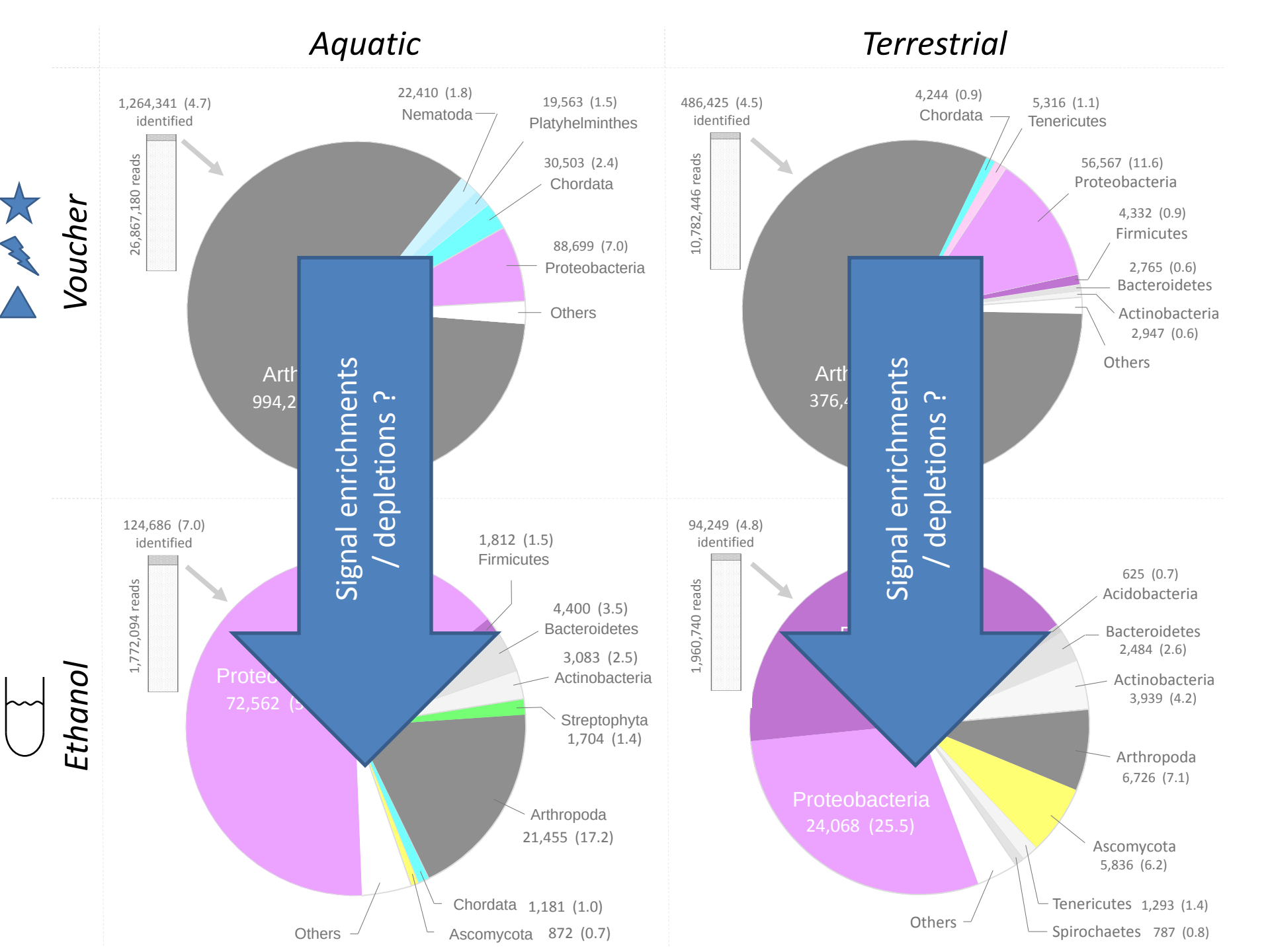
		Vouchers (MMG)		Ethanol reads (ethaMGS)		
	Recovery	**cox1 Sanger**	**Mitogenome**	**matching cox1**	**matching complete mitogenome**	**matching prot-coding regions**
Aquatic (21 species)	Species	20	21	2	9	7
	Proportion	95%	100%	9.5%	43%	33%
Ground (19 species)	Species	17	18	2	6	6
	Proportion	89%	95%	11%	32%	32%



Larva ! , high biomass
& lots of transfer to the media

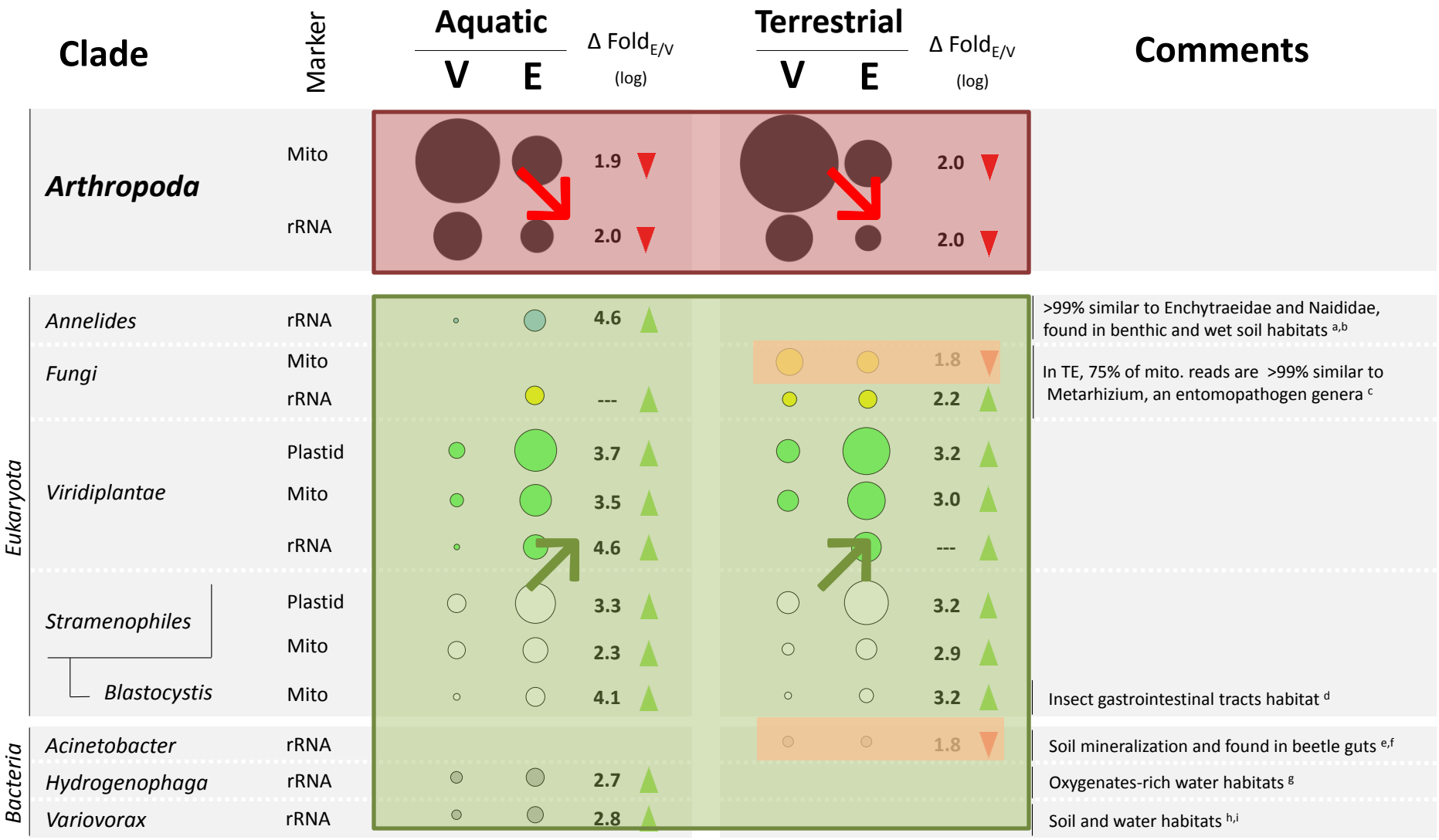
DNA Reads abundances

Sampling bias to specific species


Aquatic
Terrestrial
Voucher
Ethanol
1,264,341 (4.7) identified
26,867,180 reads
22,410 (1.8) Nematoda
19,563 (1.5) Platyhelminthes
30,503 (2.4) Chordata
88,699 (7.0) Proteobacteria
Others
Arth
994,2
Signal enrichments / depletions ?
486,425 (4.5) identified
10,782,446 reads
4,244 (0.9) Chordata
5,316 (1.1) Tenericutes
56,567 (11.6) Proteobacteria
4,332 (0.9) Firmicutes
2,765 (0.6) Bacteroidetes
Actinobacteria 2,947 (0.6)
Others
Arth
376,
Signal enrichments / depletions ?
124,686 (7.0) identified
1,772,094 reads
1,812 (1.5) Firmicutes
4,400 (3.5) Bacteroidetes
3,083 (2.5) Actinobacteria
Streptophyta 1,704 (1.4)
Arthropoda 21,455 (17.2)
Chordata 1,181 (1.0)
Ascomycota 872 (0.7)
Others
Proteo
72,562 (
94,249 (4.8) identified
1,960,740 reads
625 (0.7) Acidobacteria
Bacteroidetes 2,484 (2.6)
Actinobacteria 3,939 (4.2)
Arthropoda 6,726 (7.1)
Ascomycota 5,836 (6.2)
Tenericutes 1,293 (1.4)
Spirochaetes 787 (0.8)
Others
Proteobacteria 24,068 (25.5)

General patterns of concomitant DNAs

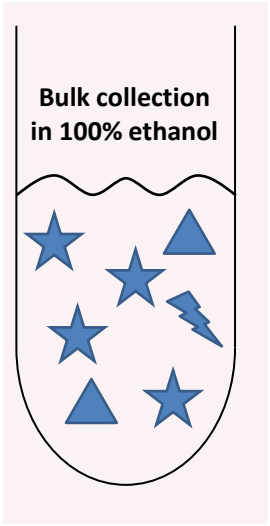
→ deep taxonomic analysis of mitochondrial, rRNA, chloroplastic and symbiontic markers.
→ Understanding the potential of the ethanol before going back to targeted approaches



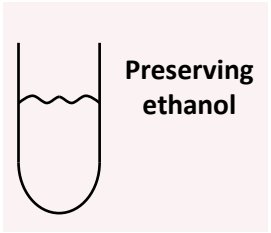
Ethanol MGS : lessons learned

(manuscript in preparation)

Arthropod sample



Ethanol metagenome skimming (ethaMGS)



Think twice before throwing the ethanol !
It contains ecological traits.

1. Understanding which signal could hold your community of interest → metagenomics, deep taxonomic analysis

   Warning: Larva vs adults, biomass influence, pooling design

2. Eventually, use the ethanol to add value to your study by using targeted approaches (ethanol metabarcoding)

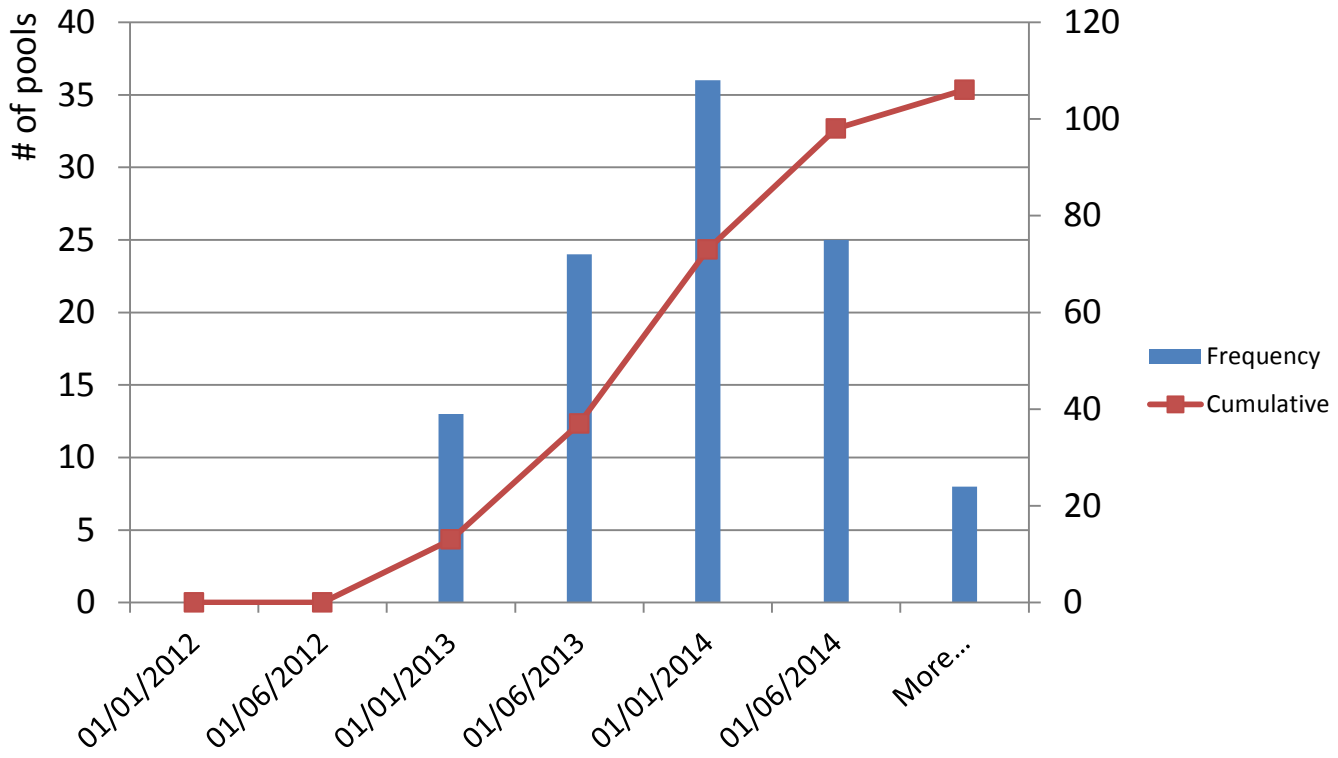
3. If you are interested in symbionts, only open associations are likely to be transfered to the media (vomit effect)

Hey ! You trashed our lunch and symbionts !
Stupid postdoc !

Perspectives: integrative analysis of insect pools

1. A collection of pools ? Some challenges…

→ Pools from different projects and build for different purposes



MODELING : predicting sequencing outcomes

My question: I do a low coverage metagenome of a specimen pool.
(Mito-metagenomics, metagenome skimming…)

What read outcome should I expect

1. for different targets (rRNA, mito, symbionts, histones, repeats…)
2. using a given sequencing depth (# bp sequenced)
3. on a pool of given taxonomic diversity (# specimens/diversity)

Inspired from the future Directions described in the review

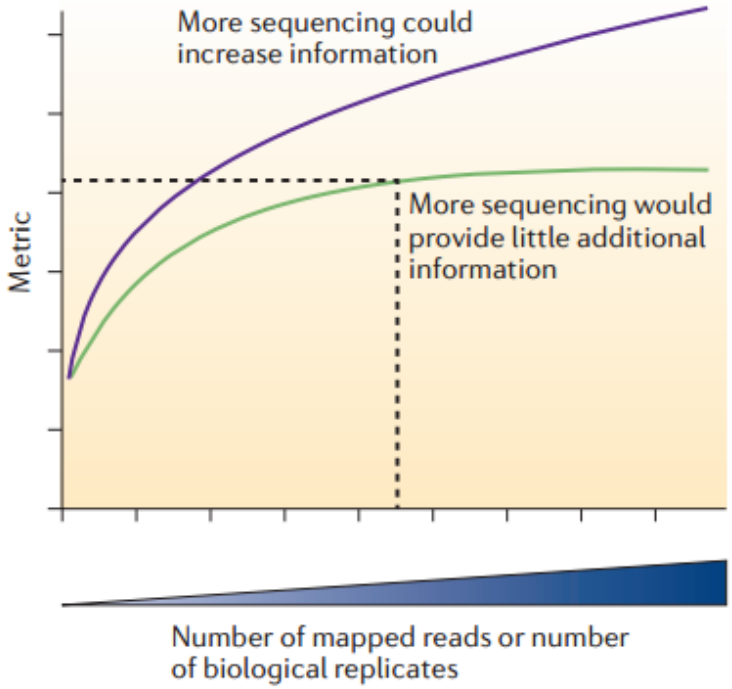
Sequencing depth and coverage: key considerations in genomic analyses

David Sims, 2014, Nature reviews

Box 3 | **Staged sequencing for predicting sequencing requirements**

Possible metrics:

- General transcriptome coverage: percentage of genes covered over 90% at a given expression level
- Differential expression: number of differentially expressed genes
- Alternative isoform detection: percentage of split reads (that is, junction that spans reads)
- ChIP–seq peak detection: number of enriched loci



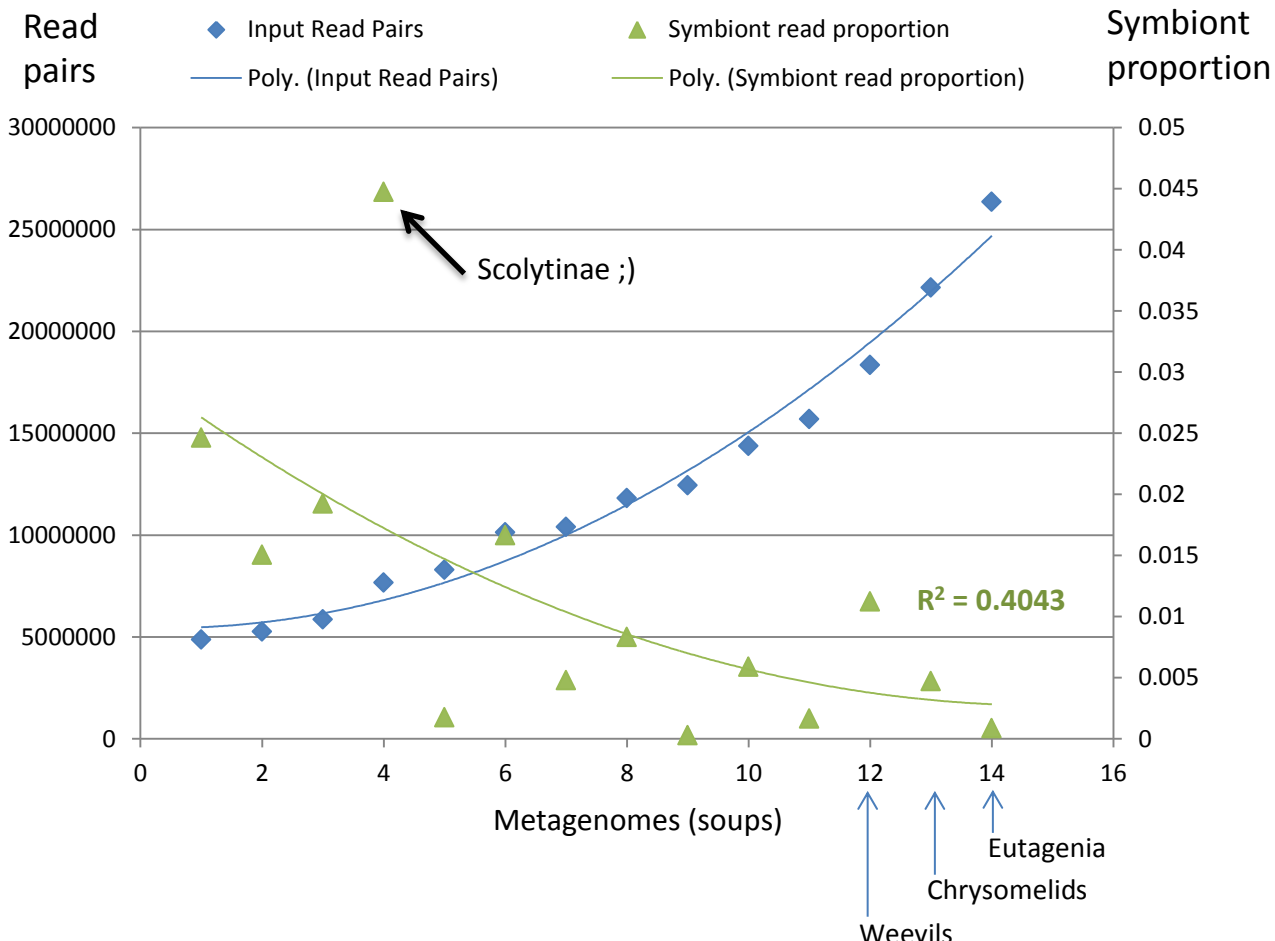
Read level MODELING : some symbiont results

My data : Reads of all libraries mentionned previously.
High thresholds for IDENTIFICATION (>99% identity on 90% of the read)

Very basic example:
(not well normalized)

Will be completed
with the taxo diversity
of the samples…

Lower sequencing depth
= more symbionts
sampled ???



Conclusions

For questions related to species-rich and relatively unknown clades, metagenomic approaches based on organelles show lots of potential.

Mitochondrial mitogenomics can be seen as a « superbarcoding » needing very few wet lab work to build a large genomic reference database.

When enrichment technics will be optimized, thousands of mitochondria per run

Genomics and « large-scale » comparative genomics will be the new core of environmental studies, but new challenges and bioinfo developments are now needed, even for a « well-known » marker like the mitochondria.

Hey ! You trashed our lunch and symbionts !
Stupid postdoc !

All NHM Biodiversity Initiative members

&
Kevin Hoptkins for NGS advices and support
Peter Foster for bioinfo support
Paul Ward for IT infrastructure

&
NHM/Imperial molecular labs
NHM/Oxford sequencing platforms

Alfried Vogler
Alex Crampton-Platt
Martijn Timmermans
Conrad Gillett
Carmelo Andujar
Chris Barton
Belen Arias
Hannah Norman
Chris Phipps

Paula Arribas Blázquez
Kirsten Miller
Debora Pires Paula
Rosli Kasah
Emeline Favreau
Carola Gomez
Andres Baselga
CT Tang

Thank you for your attention.