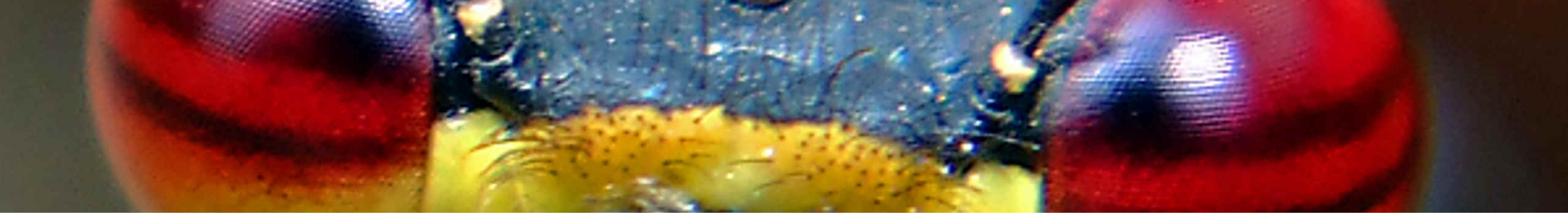


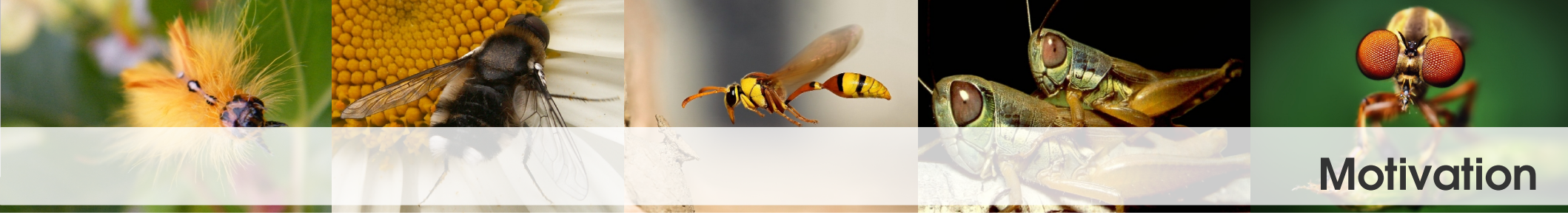
1,000 Insect Transcriptomes

Introducing the 1KITE project, taking next steps in phylogenomics and what else can be done with transcriptomic data

Karen Meusemann
on behalf of the 1KITE Consortium

CSIRO Ecosystem Sciences, Australian National Insect Collection, Canberra, Australia

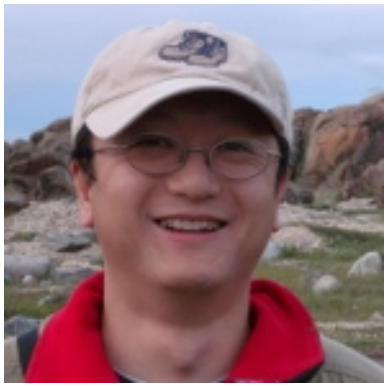
Phylomania, Hobart, Tasmania, November 08, 2013



Motivation

"We want to understand the evolution of insects in all of its aspects, using the results of the most advanced approaches in genomic and morphological research."

Speakers



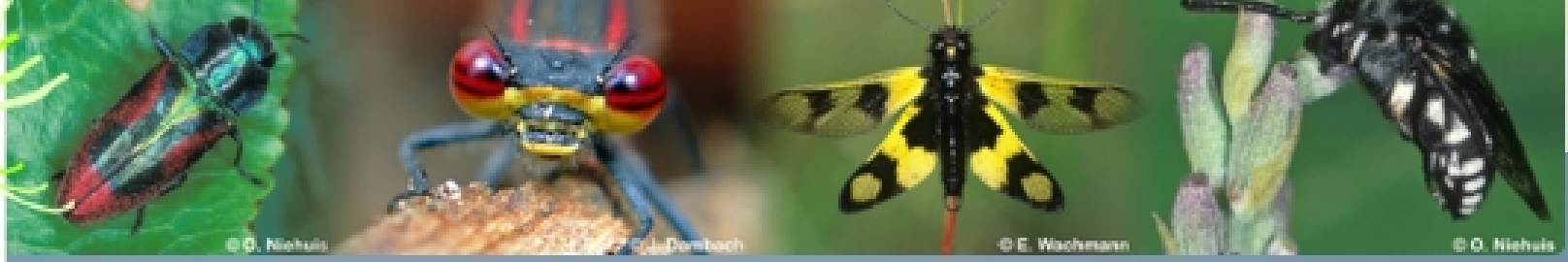
Xin Zhou,
BGI Shenzhen

Karl Kjer,
Ruttgers, NJ

Bernhard Misof,
ZMB, ZFMK Bonn

1KITE
1000 INSECT TRANSCRIPTOME EVOLUTION

& 1KITE team

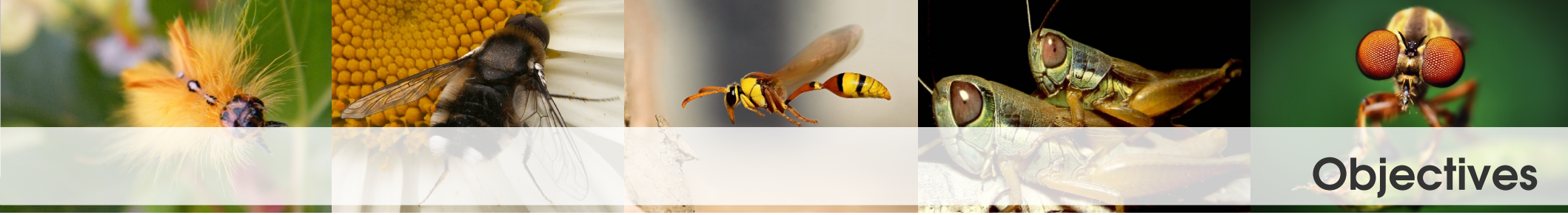


1KITE
1000 INSECT TRANSCRIPTOME EVOLUTION

Collaborators

Imprint & Copyright

Last modified 2012-01-23.

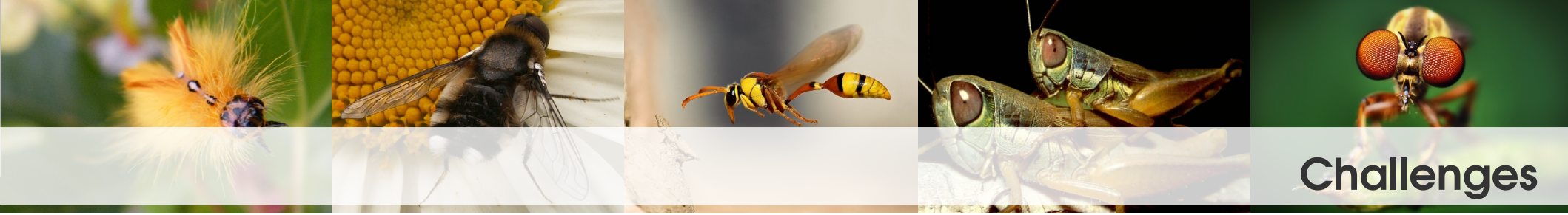


Objectives

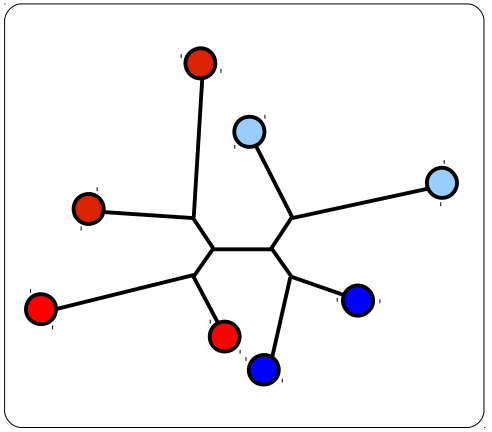
1 - Phylogenetic relationships

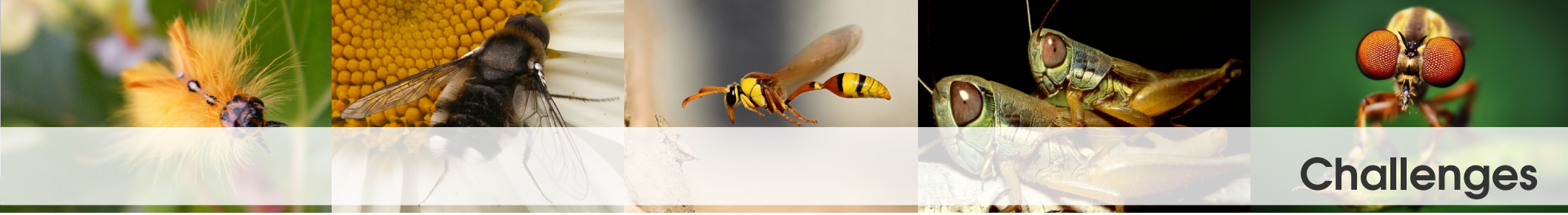
2 - Understanding phenotypic evolution

3 - Can we push molecular analyses to their limits and what do we gain?

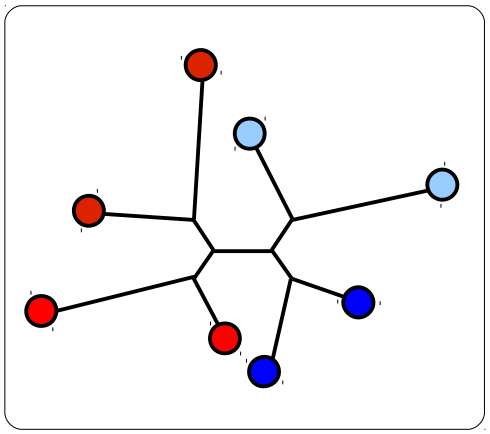


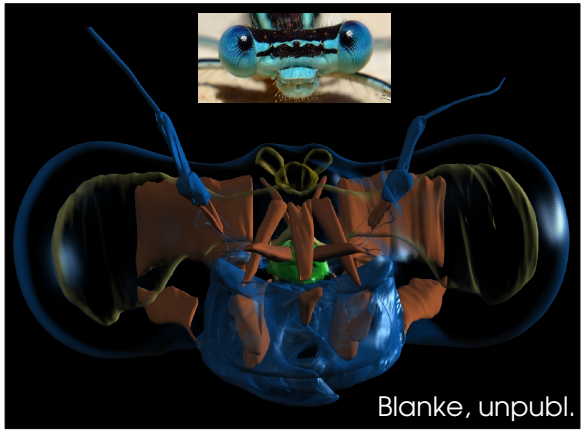
Challenges

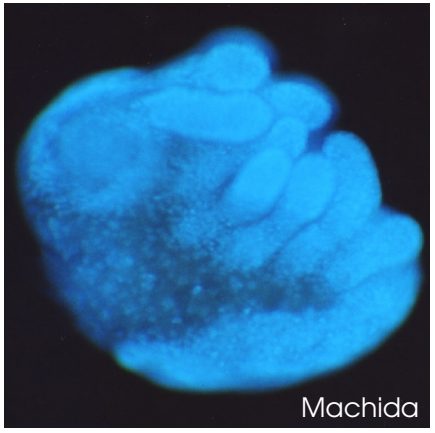




Challenges







– Infer robust phylogeny

– Understand evolution

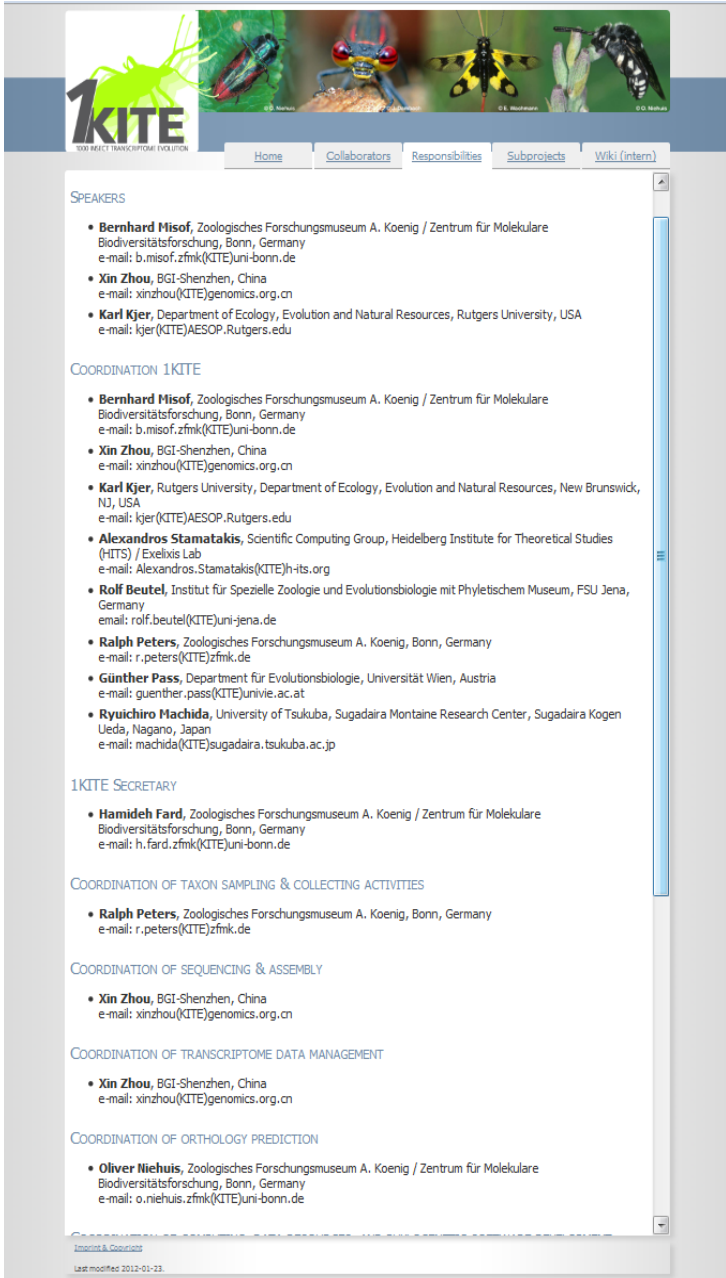
– Explain biodiversity

Responsibilities

Coordination of

- taxon sampling and collecting activities
- data management
- sequencing and assembly
- orthology prediction
- computing, data resources, and phylogenetic software development
- phylogenetic workflow
- plausibility assessment
- taxonomic subprojects

see www.1kite.org



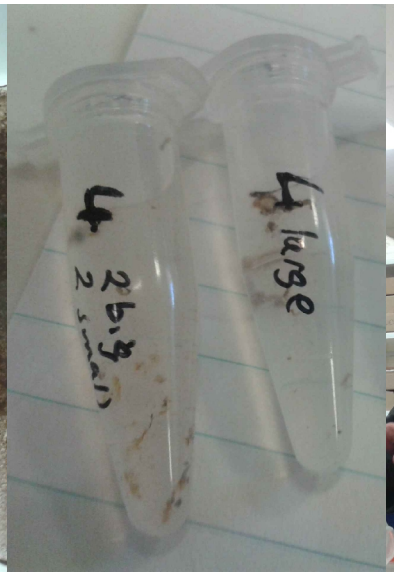
Taxonomic Subprojects

11 Subprojects

- **Basal Hexapods (109)**
- **TransOdonata (120)**
- **Polyneoptera (131)**
- **Dictyoptera (96)**
- **Paraneoptera (72)**
- **Hymenoptera (250)**
- **Neuropterida (95)**
- **Coleoptera (62)**
- **Trichoptera (60)**
- **Lepidoptera (70)**
- **Antliophora (124)**

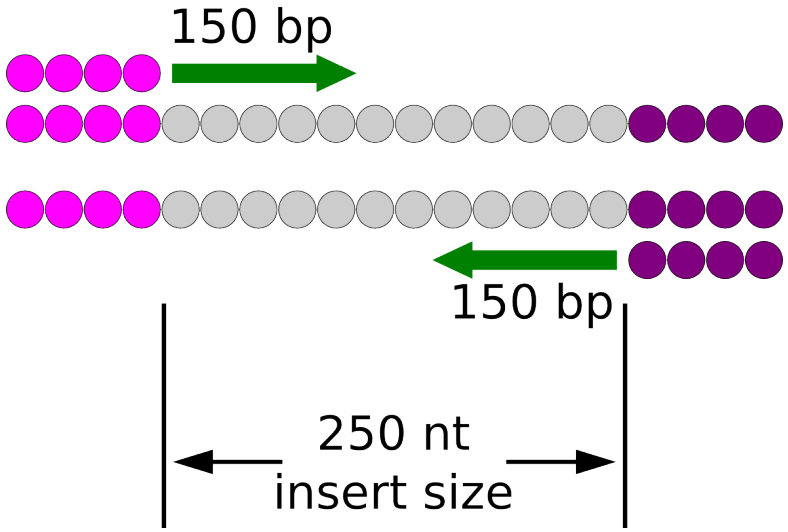
Collecting more than 1,000 Insect Species

garden springtail *Bourletiella hortensis*, 06/2011



enigmatic mecoperans: *Nannochorista*, 10/2013

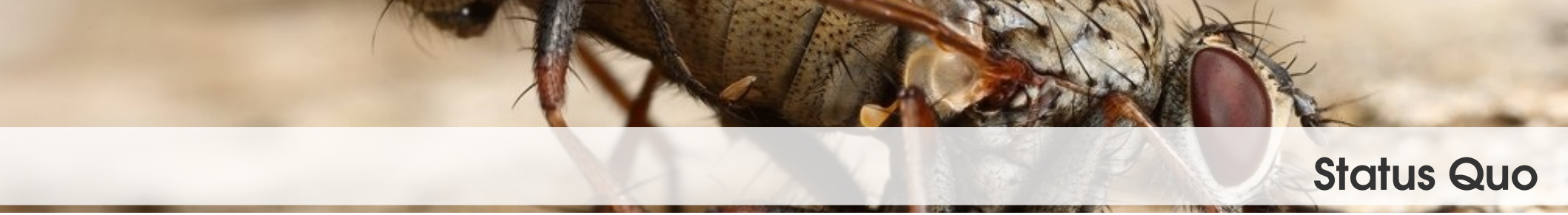
Sequencing and Assembly



- 250 bp insert-size libraries
- 150 bp paired-end sequencing
- 2.5 GB raw data per species
- ILLUMINA HiSeq 2000 sequencing
- > 5,000 full length genes

SOAP Short Oligonucleotide Analysis Package SOAPdenovo-Trans

av. contig #	= 92533
av. scaffold #	= 14569
av. N50	= 629bp

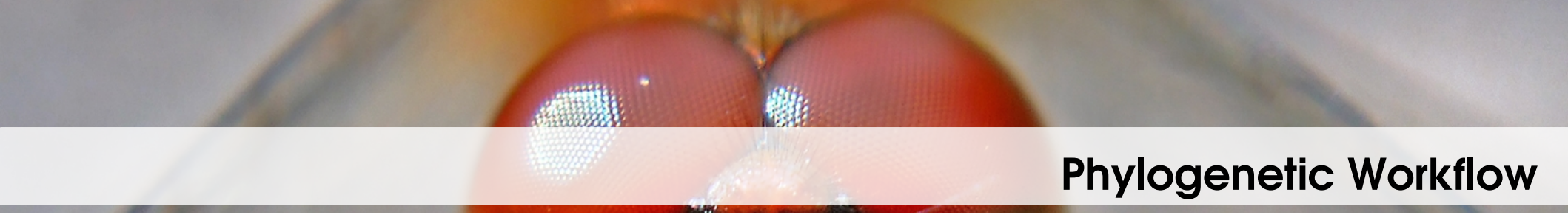


Status Quo

Struck et al. (2011)

1KITE, (in prep.)

x 10 by the end of 2013



Phylogenetic Workflow

- orthology prediction and assignement of transcripts
- annotation, alignment, refinement and alignment masking
- identification of protein domains
- selection of a „decisive" dataset (reduce missing data)
- sophisticated partitioning (coll. Rob Lanfear, Brett Calcott, ANU)
- exploration of datasets (coll. Lars Jermiin & group)
- evaluate and reduce heterogeneity (coll. Lars Jermiin & group)
- tree reconstruction, dating
- **Four-cluster Likelihood mapping** (test single nodes, detect „hidden" signal)

Specific Hypothesis Testing - FcLM

- single splits – is there alternative signal present?
- FcLM implemented in RAxML
- new mapping tool
- are the results affected by non-phylogenetic signal

Four-cluster Likelihood Mapping - *Specific* Hypothesis Testing

Meusemann *et al.* 2010

v. Reumont *et al.* 2012

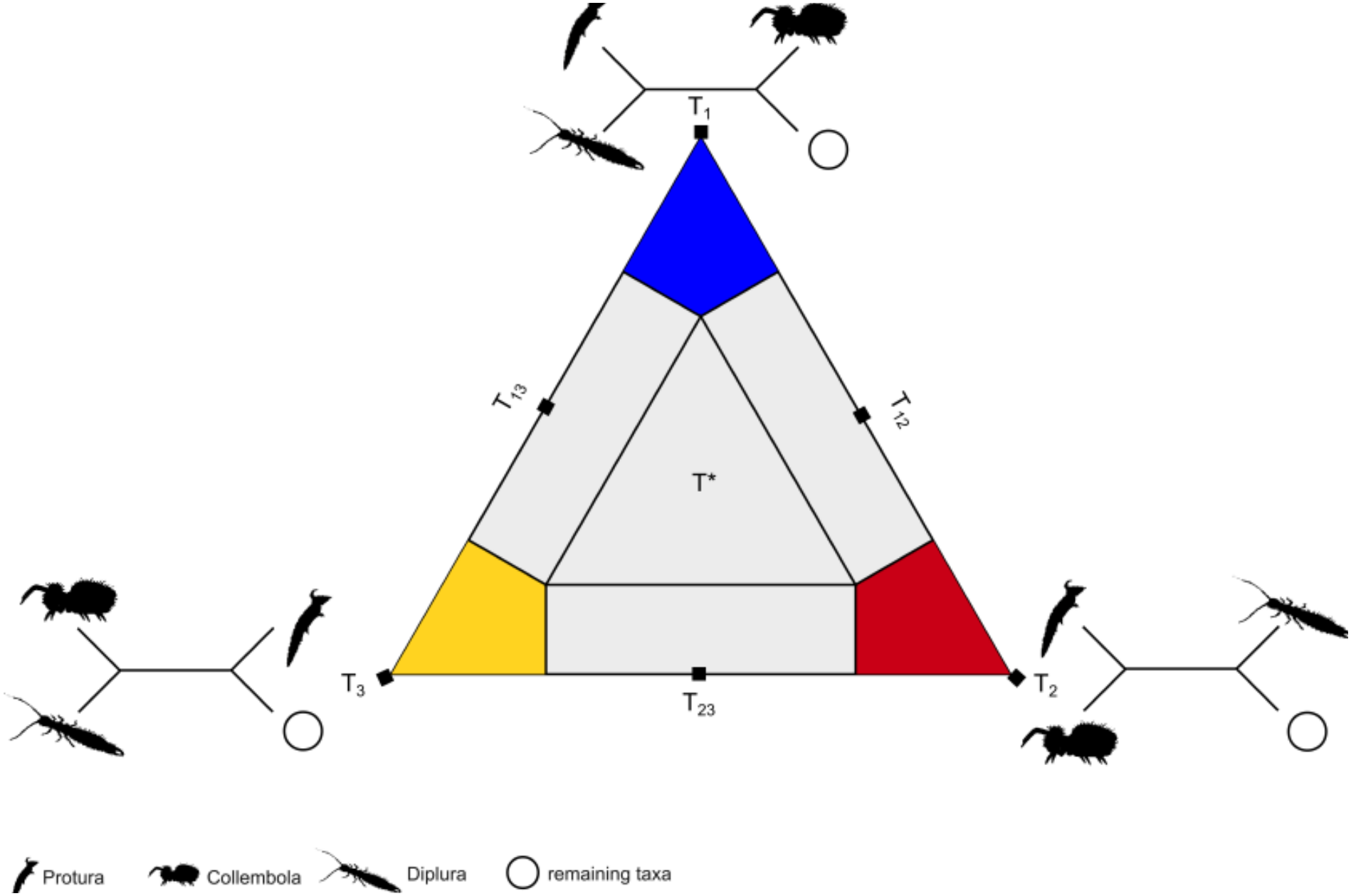
Example: Basal Hexapods:

Protura + Diplura (Nonoculata)

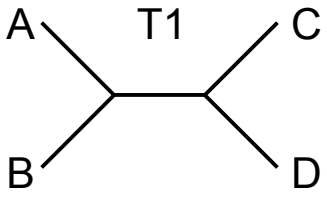
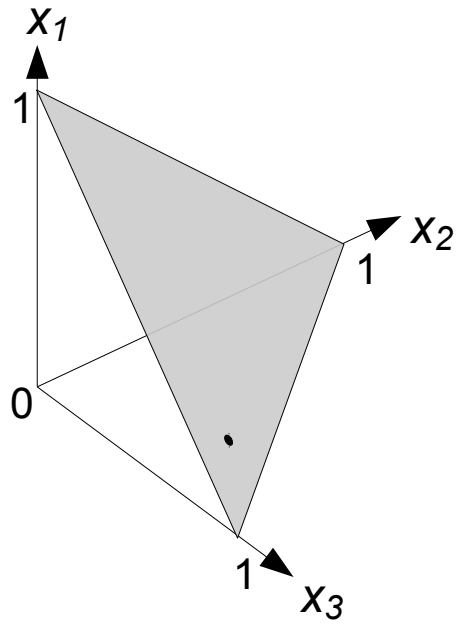
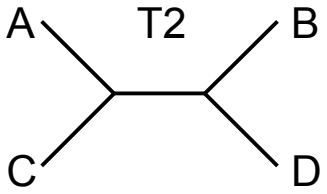
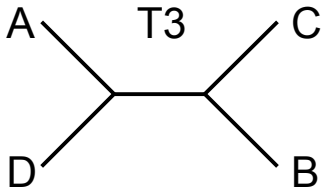
versus

Collembola + Protura (Ellipura)

Four-cluster Likelihood Mapping - *S*pecific Hypothesis Testing



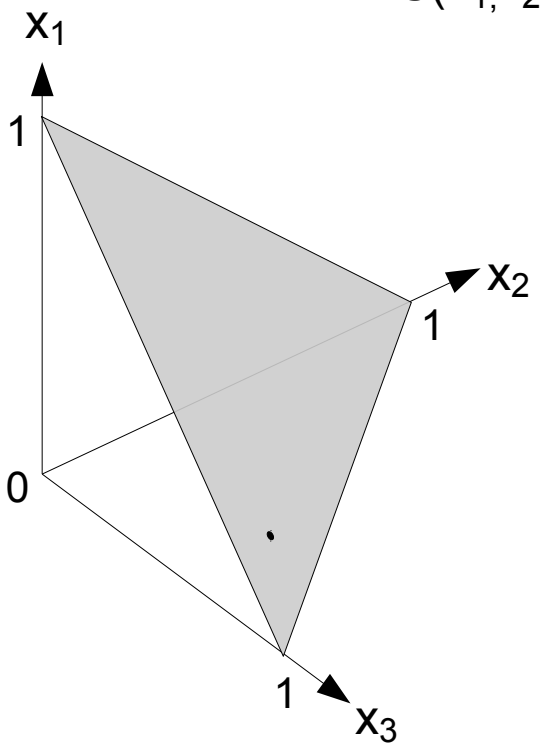
Four-cluster Likelihood Mapping - *Specific* Hypothesis Testing

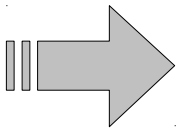
Topology	Support		Relative Support	2D-Simplex
	$s_1 = \sum_N (AB \mid CD)$	⇒	$x_1 = \frac{s_1}{s_1 + s_2 + s_3} = 0.1$	
	$s_2 = \sum_N (AC \mid BD)$	⇒	$x_2 = 0.1$	⇒
	$s_3 = \sum_N (AD \mid CB)$	⇒	$x_3 = 0.8$	$S(x_1,x_2,x_3) = (\ 0.1,\ 0.1,\ 0.8\ )$

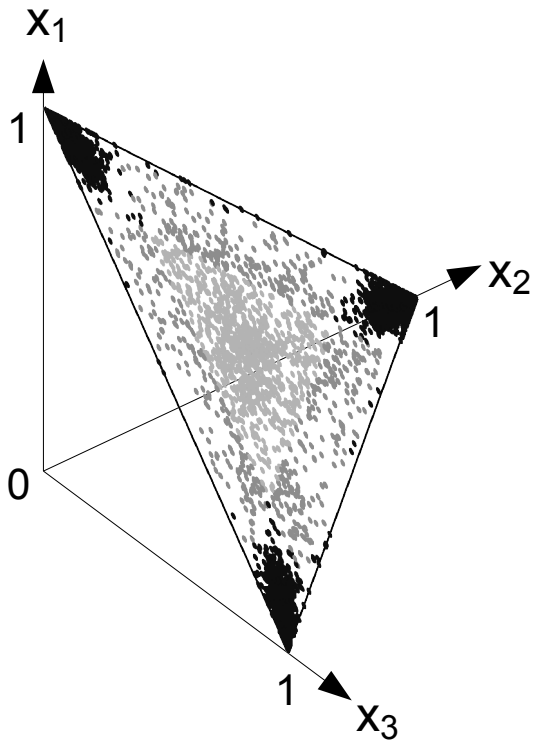
Four-cluster Likelihood Mapping - *S*pecific Hypothesis Testing

$S(x_1,x_2,x_3) = (0.1, 0.1, 0.8)$
$S(x_1,x_2,x_3) = (0.3, 0.3, 0.3)$
$S(x_1,x_2,x_3) = (0.7, 0.1, 0.2)$

.
.
.







Four-cluster Likelihood Mapping - *S*pecific Hypothesis Testing

Example: Basal Hexapods

G1: Protura, G2: Diplura, G3: Collembola, G4: Ectognatha *partim*

Possible unambiguous topologies:

T1: G1G2|G3G4 = Protura+Diplura
T2: G1G3|G2G4 = Protura+Collembola
T3: G1G4|G2G3 = Diplura+Collembola

Possible ambiguous topologies:

T1T2: G1G2|G3G4 or G1G3|G2G4
T1T3: G1G2|G3G4 or G1G4|G2G3
T2T3: G1G3|G2G4 or G1G4|G2G3
T*: G1G2|G3G4 or G1G3|G2G4 or G1G4|G2G3

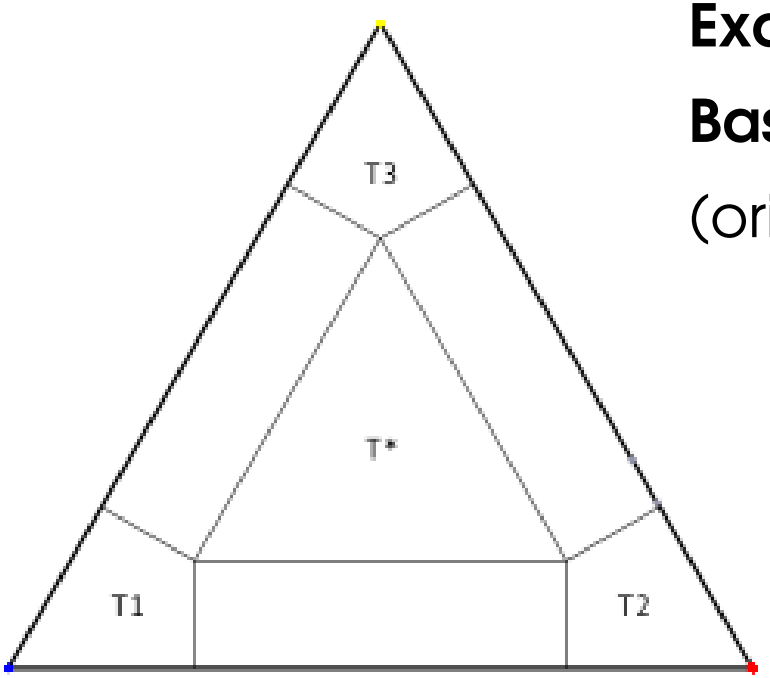
Number of all possible quartets: 120

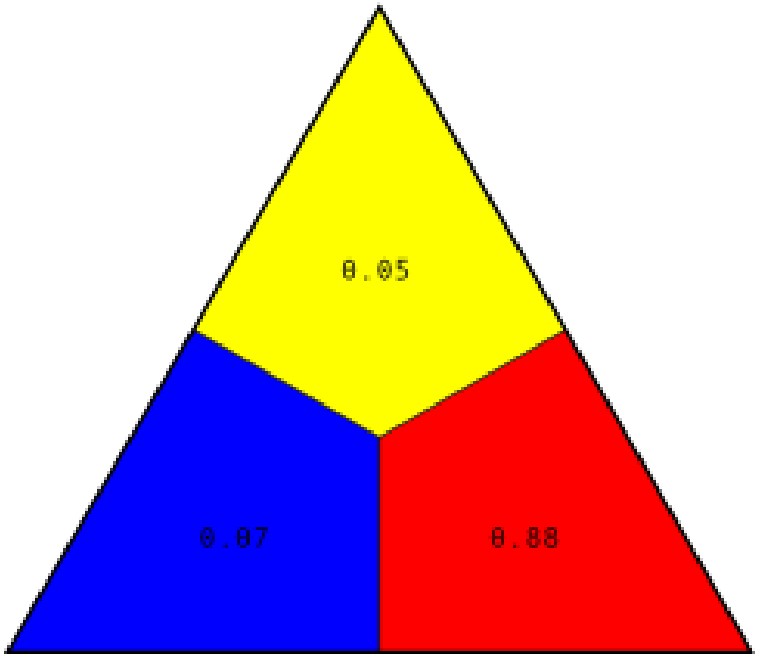
```
LM_simplex
number of quartets: 120
```

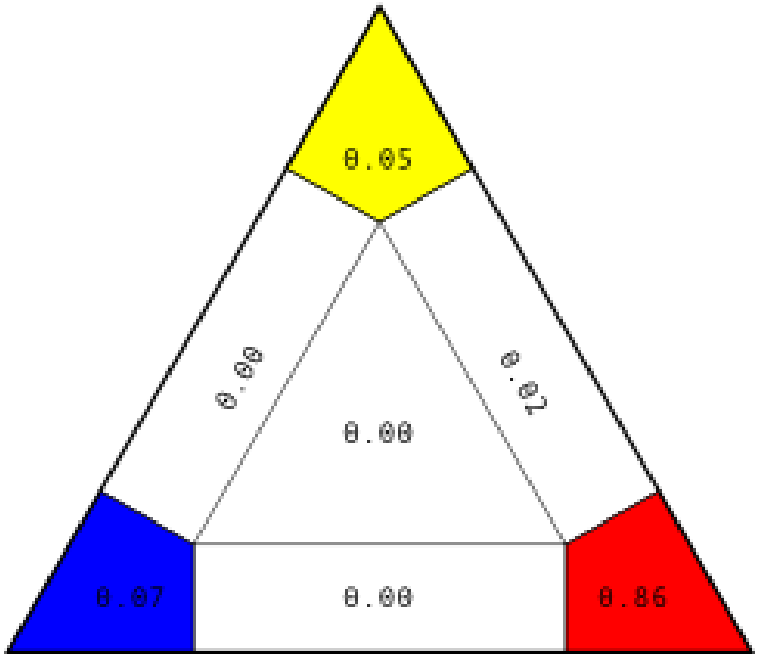
Example:

Basal Hexapods

(original)







Four-cluster Likelihood Mapping - *S*pecific Hypothesis Testing

Evaluate result with the help of permutations:

compare original data with

- permutation I

→ permute non-ambiguous amino acid characters, leaving the distribution of missing characters untouched

- permutation II

→ see permutation I but, use the WAG matrix to generate partitions that match stationary, homogeneous and reversible (SHR) conditions

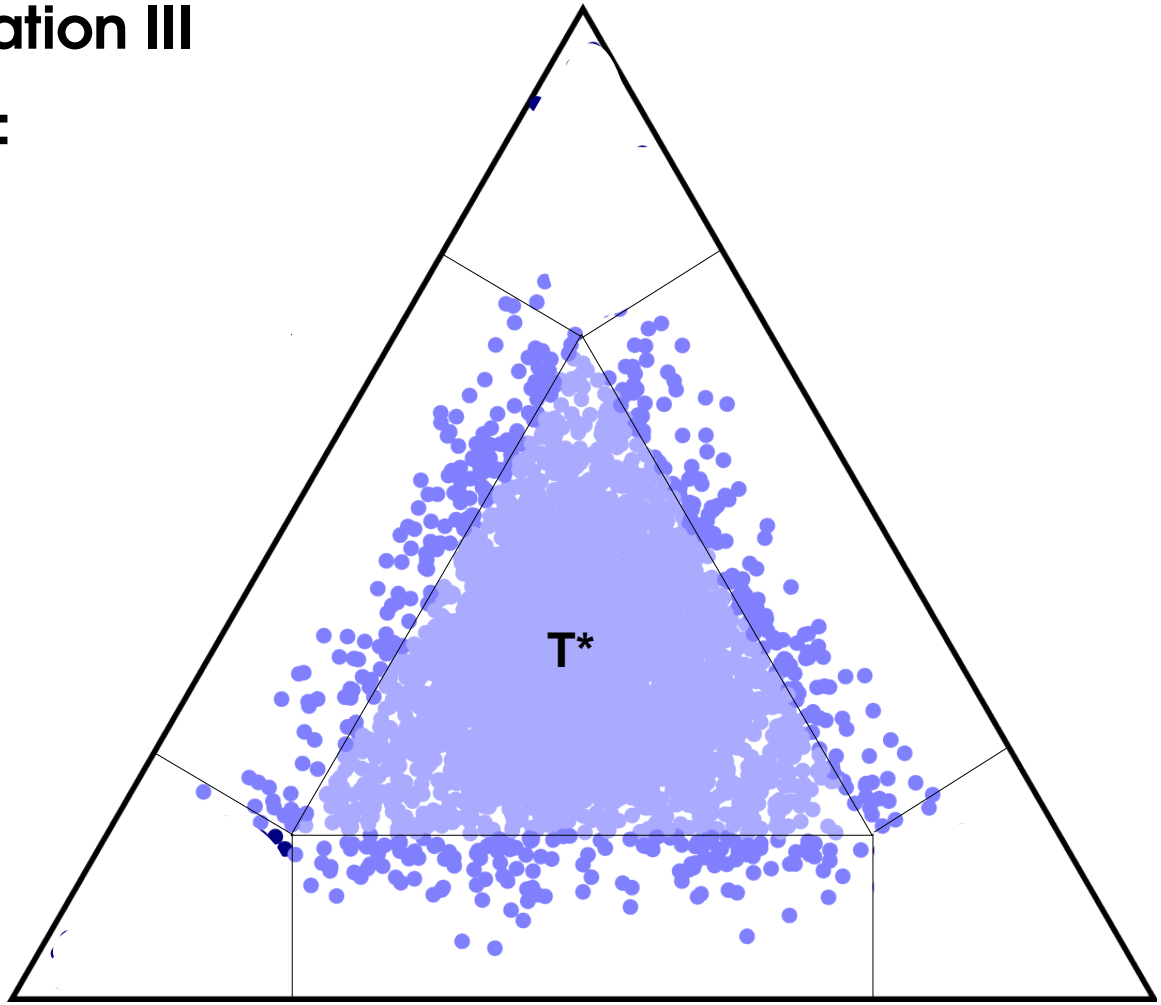
- permutation III

→ see permutation II but, additionally with randomised missing data to obtain a matrix that matches SHH conditionsand with randomly distributed missing data.

Four-cluster Likelihood Mapping - *S*pecific Hypothesis Testing

for permutation III

we expect:



Four-cluster Likelihood Mapping - *Specific* Hypothesis Testing

Supermatrix C: Topology	original (support in %):	permutation I (support in %):	permutation II (support in %):	permutation III (support in %):
T1	7	9	47	12
T2	**86**	23	7	27
T3	5	54	4	5
T1T2	0	4	12	12
T1T3	0	3	15	3
T2T3	2	5	4	23
T*	0	1	12	21

T2: G1G3|G2G4 = Ellipura

Acknowledgements

- David Yeates, ANIC, CES
- the 1KITE consortium, especially Bernhard Misof (zmb)
- the Beijing Genomics Institute (BGI)
- Lars Jermiin & group, Bioinformatics & Phylogenomics, CES
- Rob Lanfear, Brett Calcott, ANU
- Barbara Holland, Jeremy Sumner
- *all further collaborators, collectors, identifiers, etc that are continuously in this projects...*

Thank you!

www.1kite.org