Towards Application of Machine Learning to the Automated Assembly of Barcode Sequences

Eddie YT Ma, Sujeevan Ratnasignham, Stefan C Kremer

Agenda

- Problem Overview
  - From Samples to Contigs
  - The Need for Automation
- A Sampling of the Data
- Conclusion
- Next Steps

Problem overview

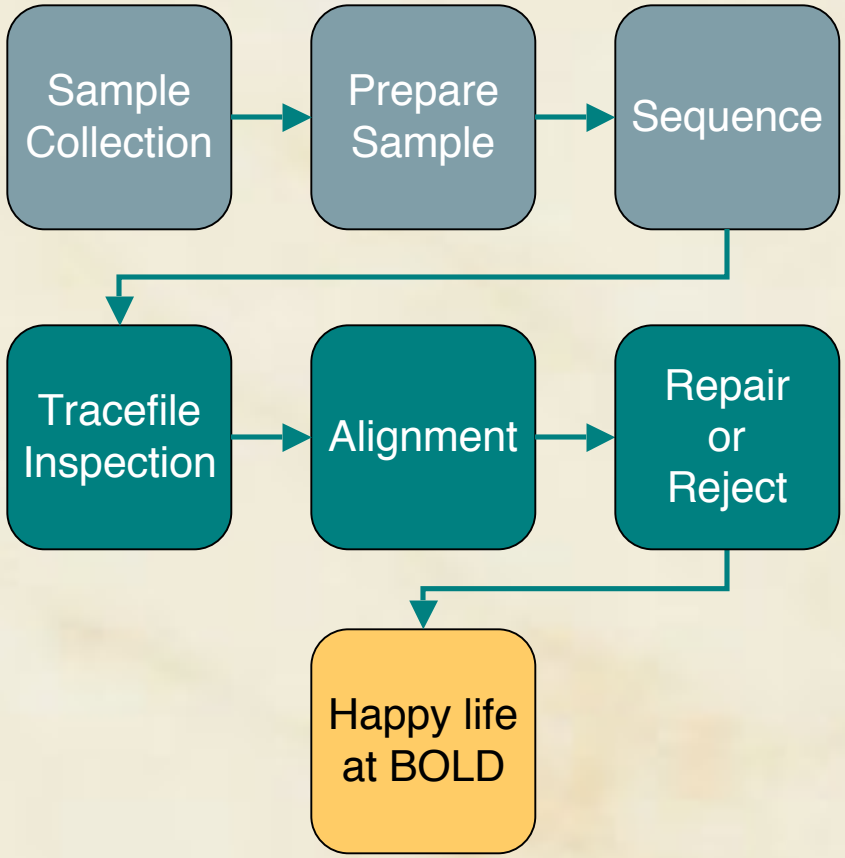
Data is growing fast ...

- 1.48M total barcode records, BOLD
- 1.8M named species described
- ~8.7M species on the planet
- iBOL to barcode 5M species by 2016

How can we keep up with the growth?

Problem overview

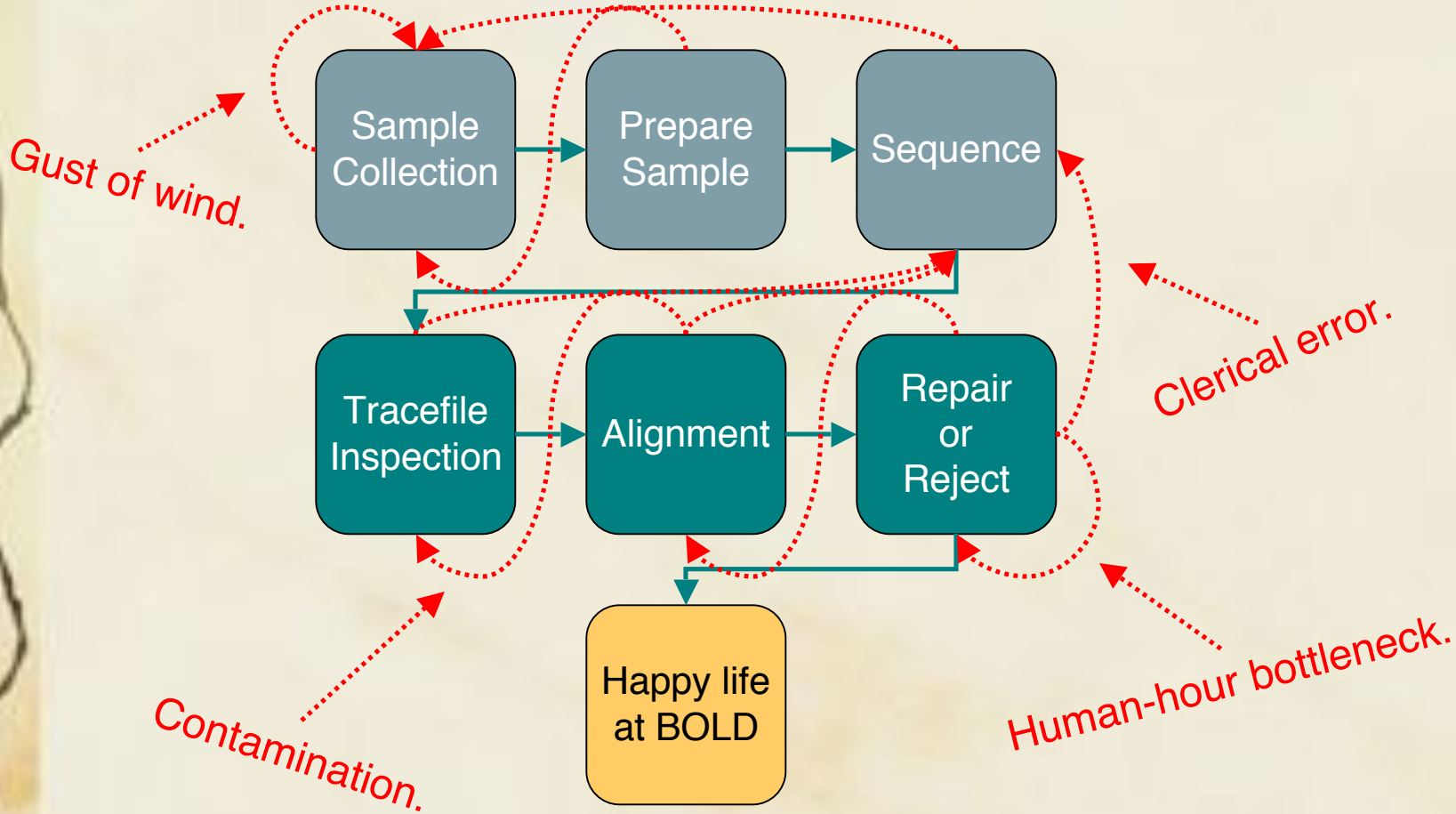
From samples to contigs



A familiar workflow …

Problem overview

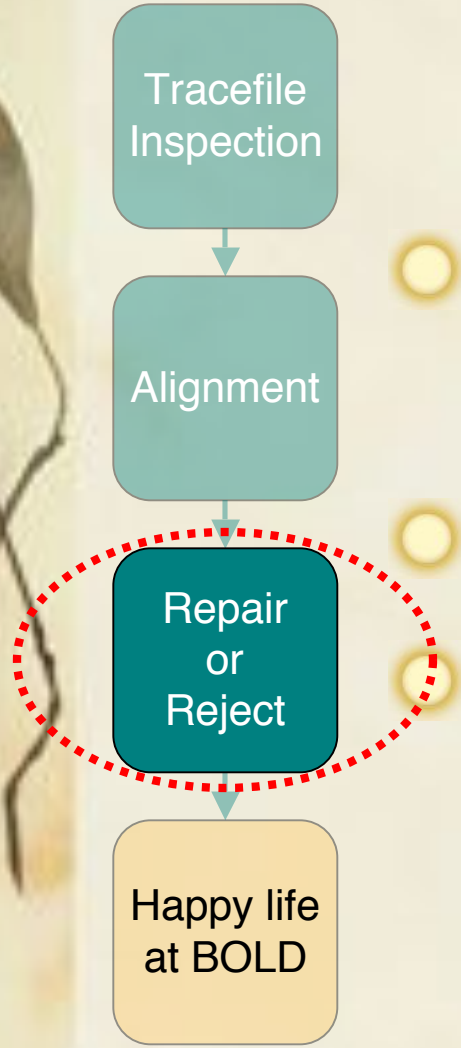
From samples to contigs



… with familiar quirks and bottlenecks

Problem overview

The need for automation



- Software can help a lot
  - Can automatically reject traces
- Still requires human intervention
- Fully automate repairs?

Problem overview

Let's decrease human effort

- Machine learning goal:
  - Imitate human intelligence to create contigs from traces
  - Human expert only needed to sample solutions

What about existing techniques?

Problem overview

Genomics vs Barcoding

- Genomics
  - Emphasize repeated coverage
  - A few large sequences
  - Need to deal with copy number variation
- Barcoding
  - At most 2x coverage
  - Thousands of sequences
  - Copy numbers within barcode consistent

Start by looking at the data …

Problem overview

Preliminary sampling of COI ...

Category	Records	Tracefiles
orders of insect		
Coleoptera	12449	26064
Diptera	2874	7670
Hymenoptera	5630	12836
Lepidoptera	37308	79944
vertebrates		
bird	3062	6301
mammal	20878	44502
fish	4169	8856

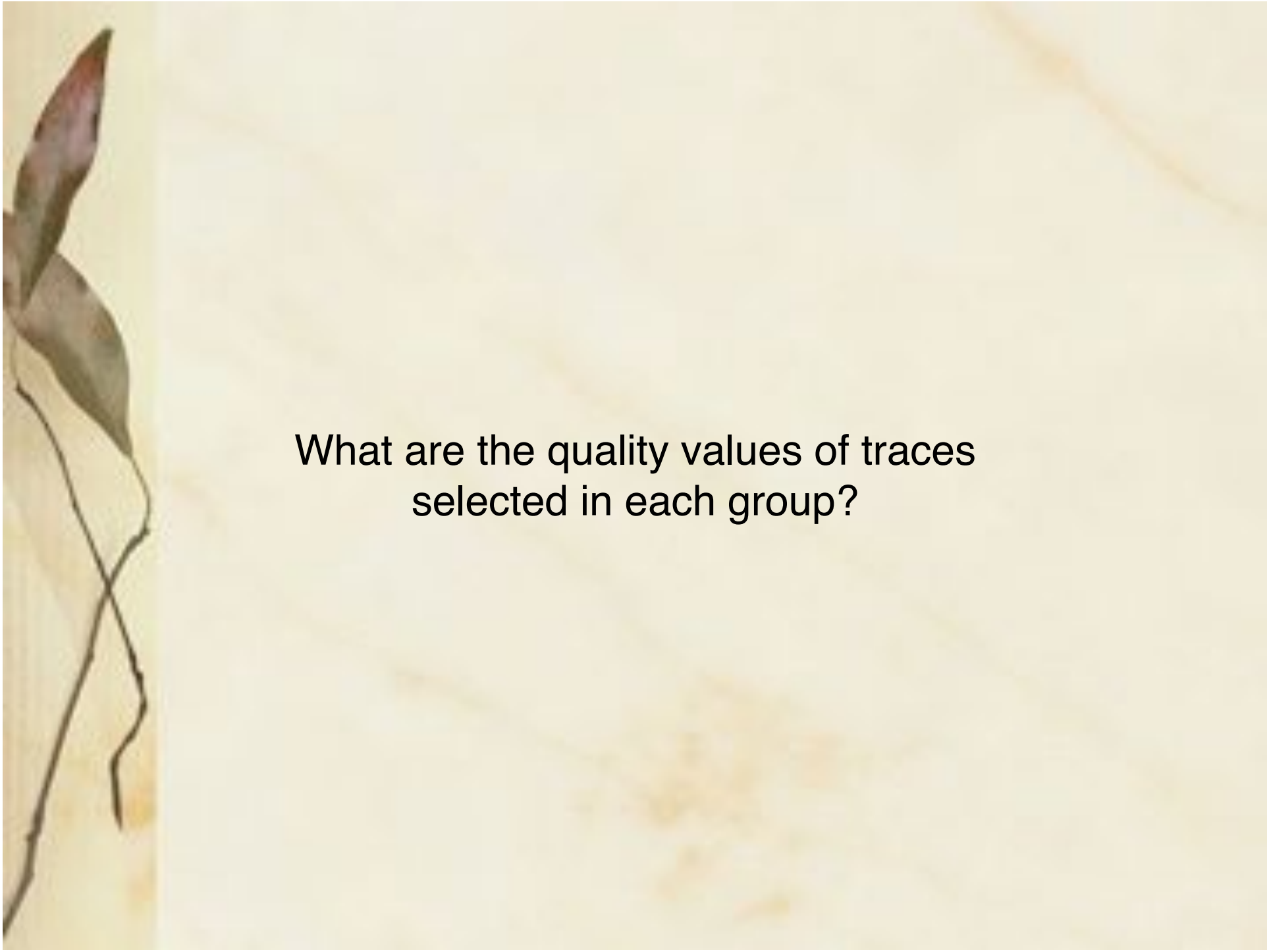
Our data is both the tracefile and the human verified contig.

Problem overview

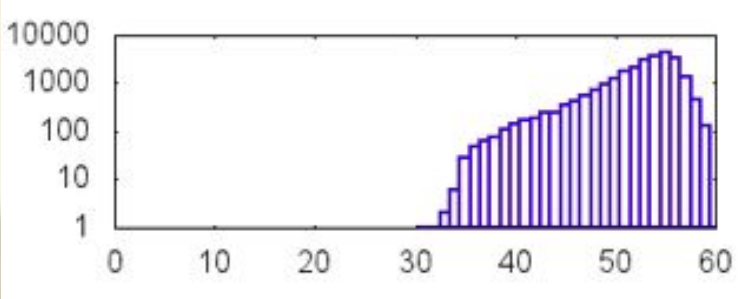
Preliminary sampling of COI ...

- Present goal:
  - Now, describe trends in the data
  - Allows us to categorize known traces
- Future goal:
  - Understand editing for known traces
  - Apply pattern of editing to new traces

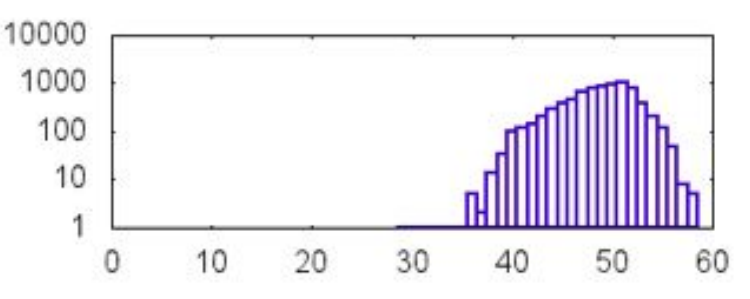
(take *editing* to mean *sequence repairing*)


What are the quality values of traces
selected in each group?

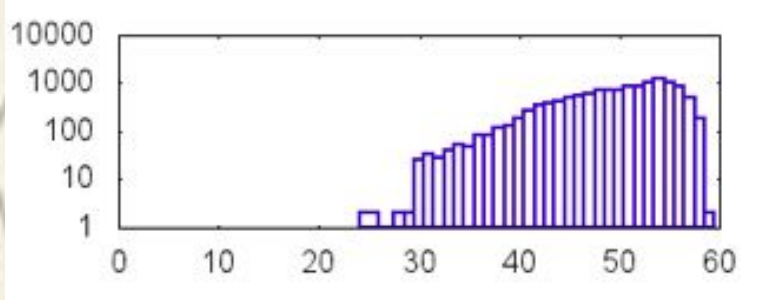
Quality values in each group



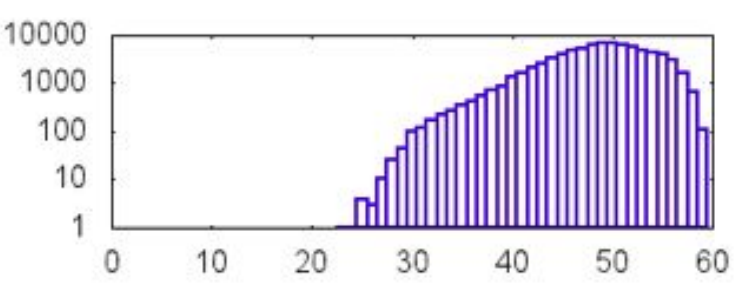
Coleoptera



Diptera



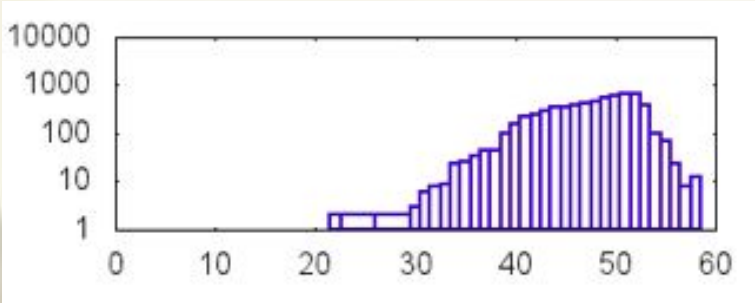
Hymenoptera



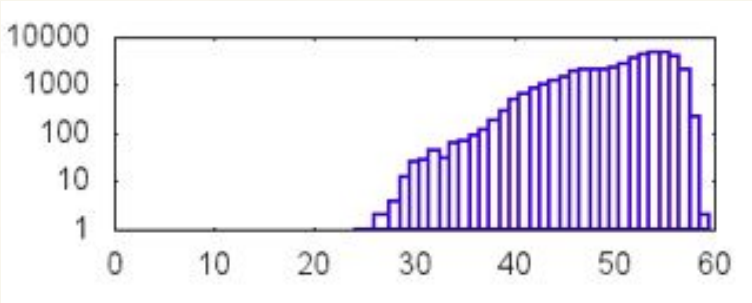
Lepidoptera

x = average quality value over all aligned positions
y = log count of occurrences

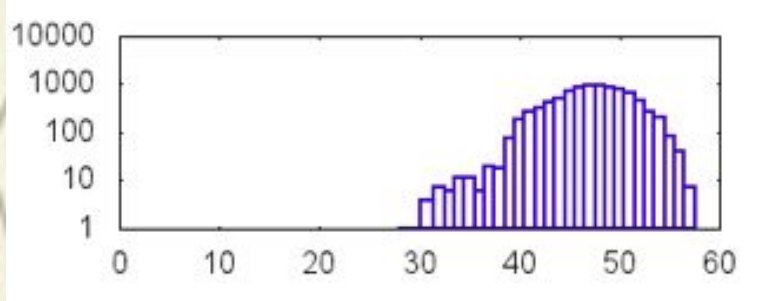
Quality values in each group



birds

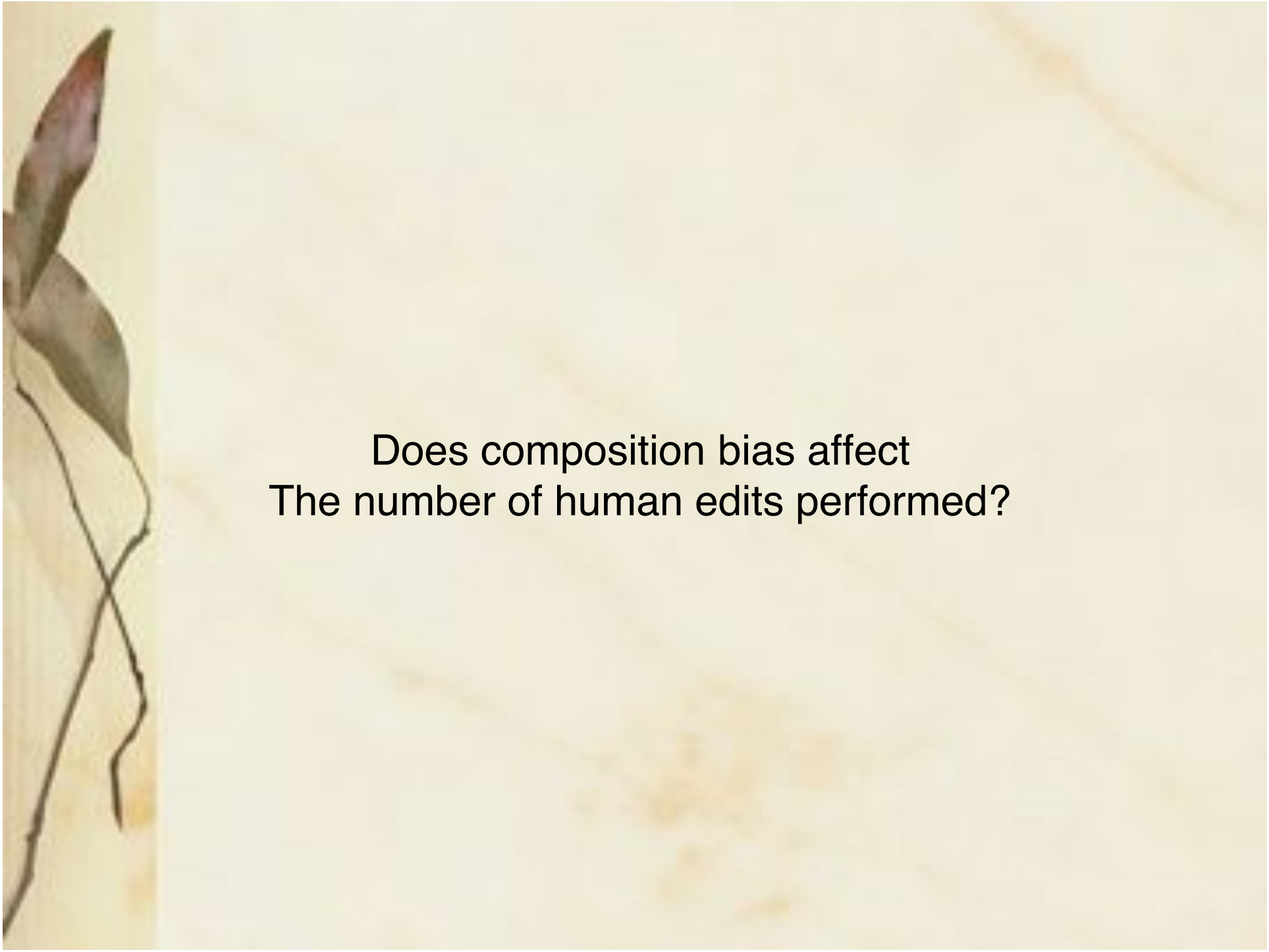


mammal

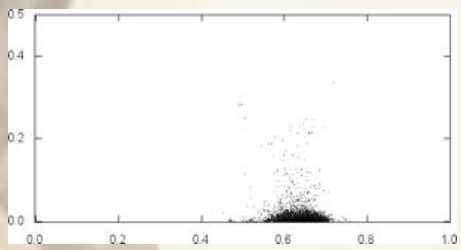


fish

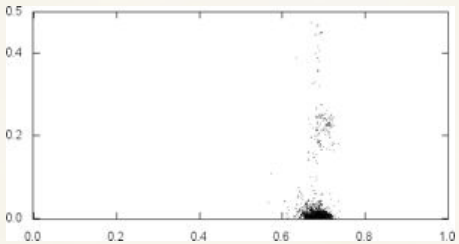
x = average quality value over all aligned positions
y = log count of occurrences


Does composition bias affect
The number of human edits performed?

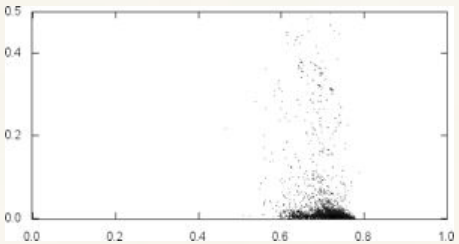
Composition bias and human edits (substitutions)



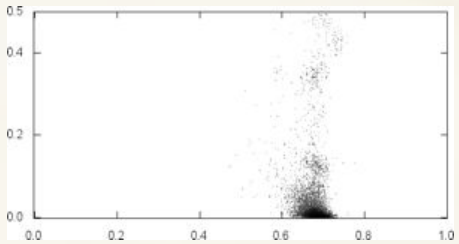
Coleoptera



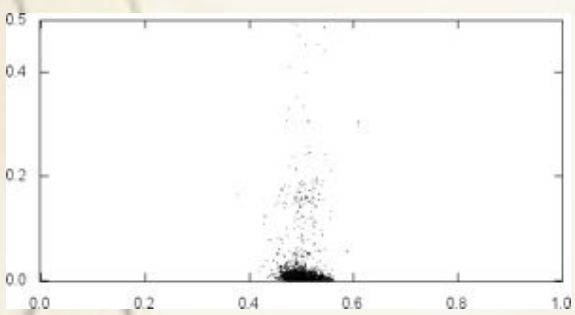
Diptera



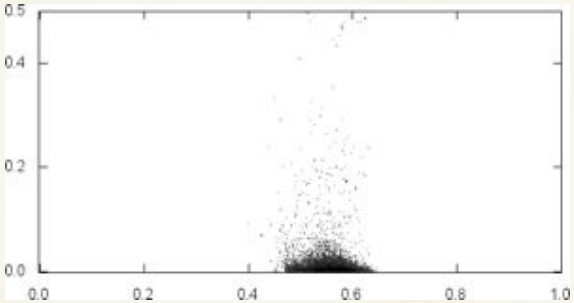
Hymenoptera



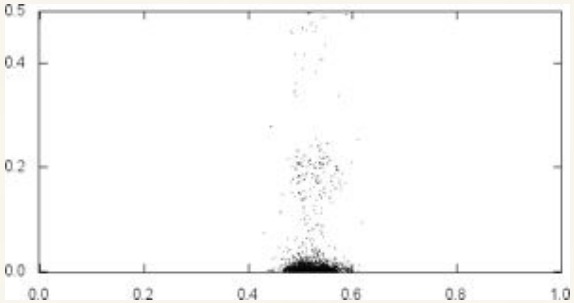
Lepidoptera



birds



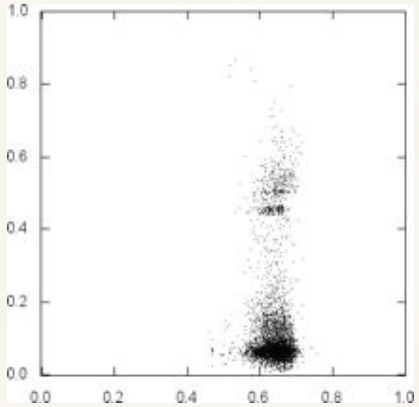
mammals



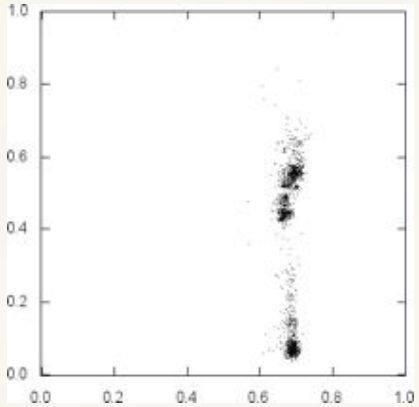
fish

x = %AT-composition over accepted length of raw trace
y = %Edits (substitutions) over length of alignment

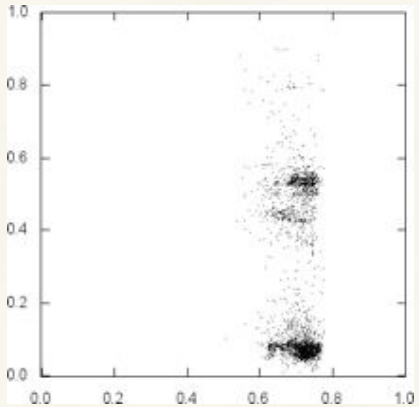
Composition bias and human edits (indels)



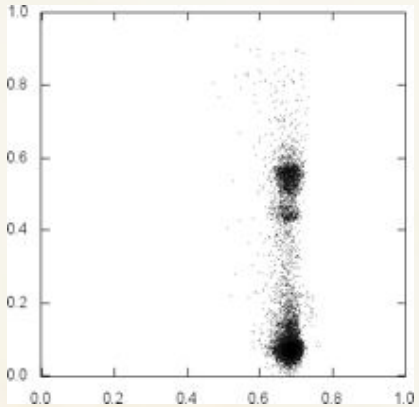
Coleoptera



Diptera



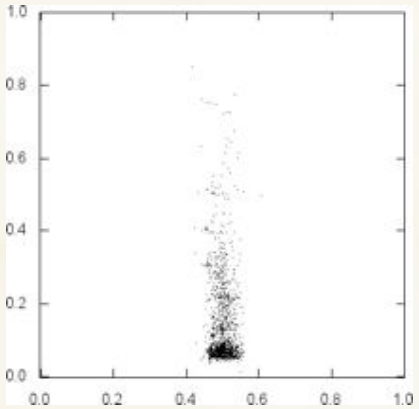
Hymenoptera



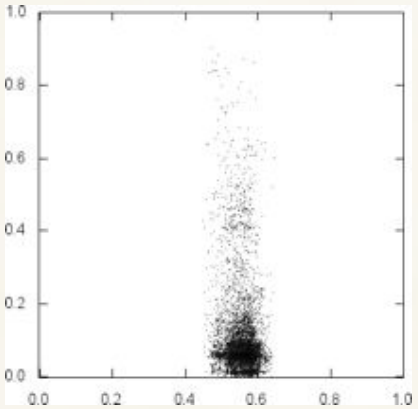
Lepidoptera

x = %AT-composition over accepted length of raw trace
y = %Edits (indels) over length of alignment

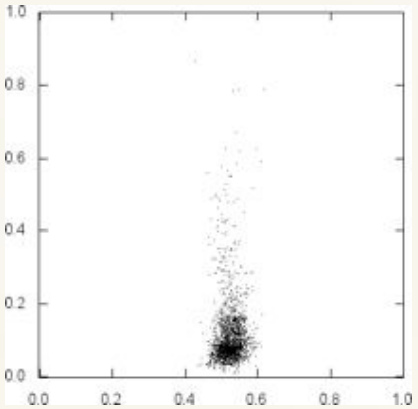
Composition bias and human edits (indels)



birds

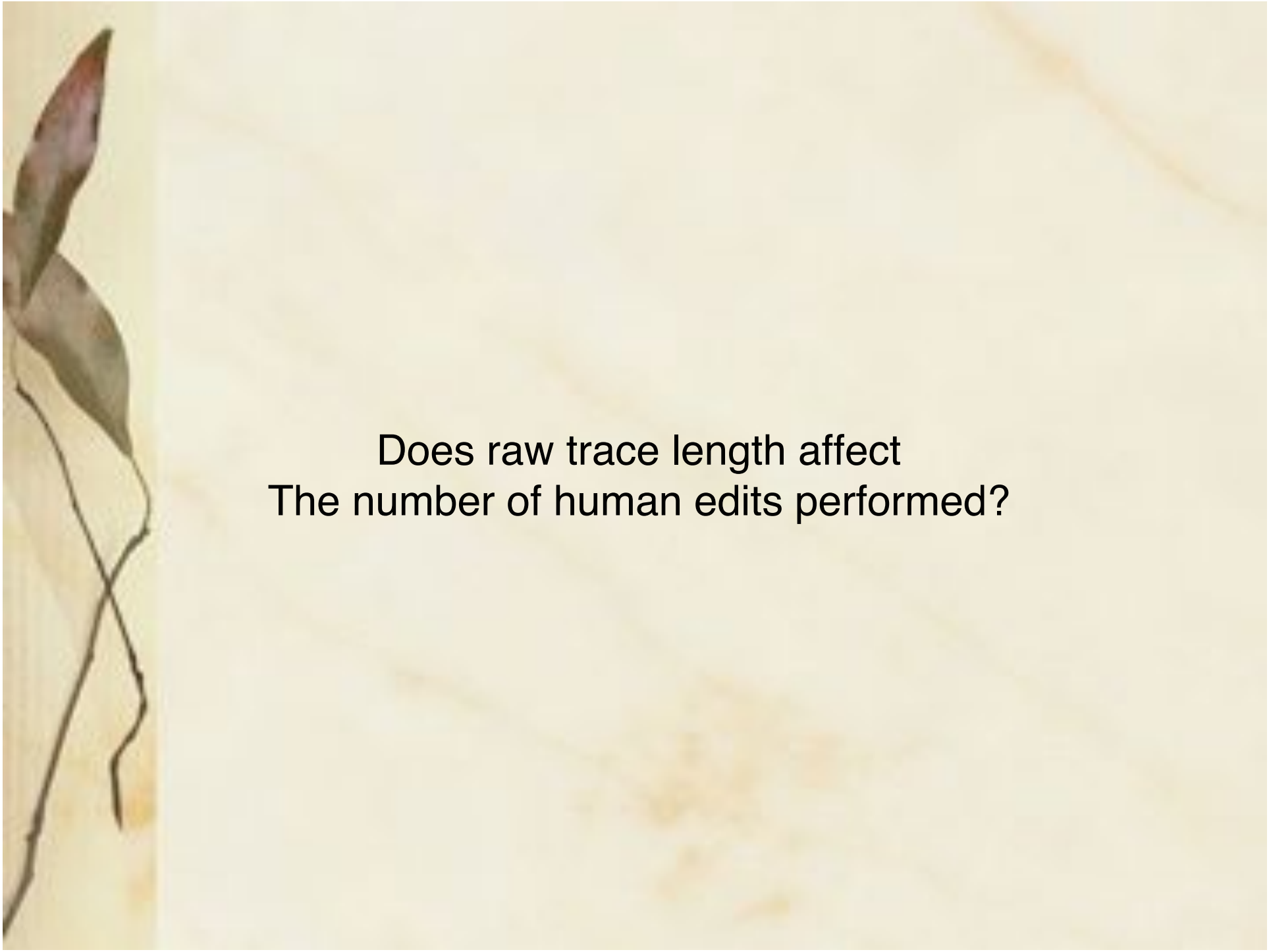


mammals

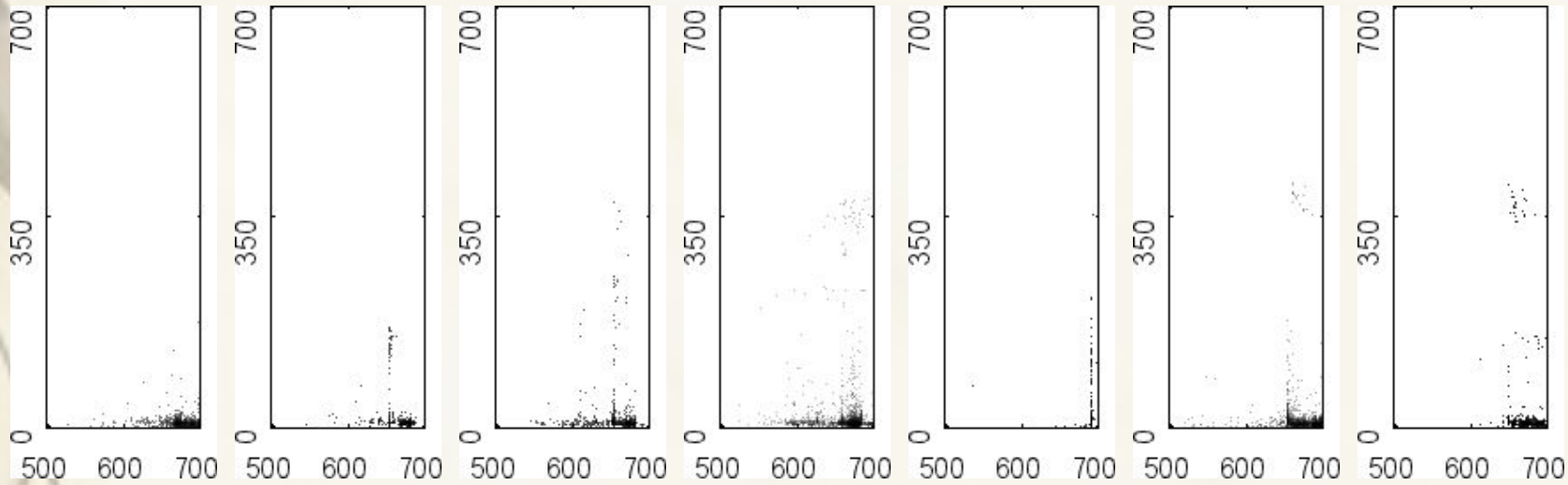


fish

x = %AT-composition over accepted length of raw trace
y = %Edits (indels) over length of alignment

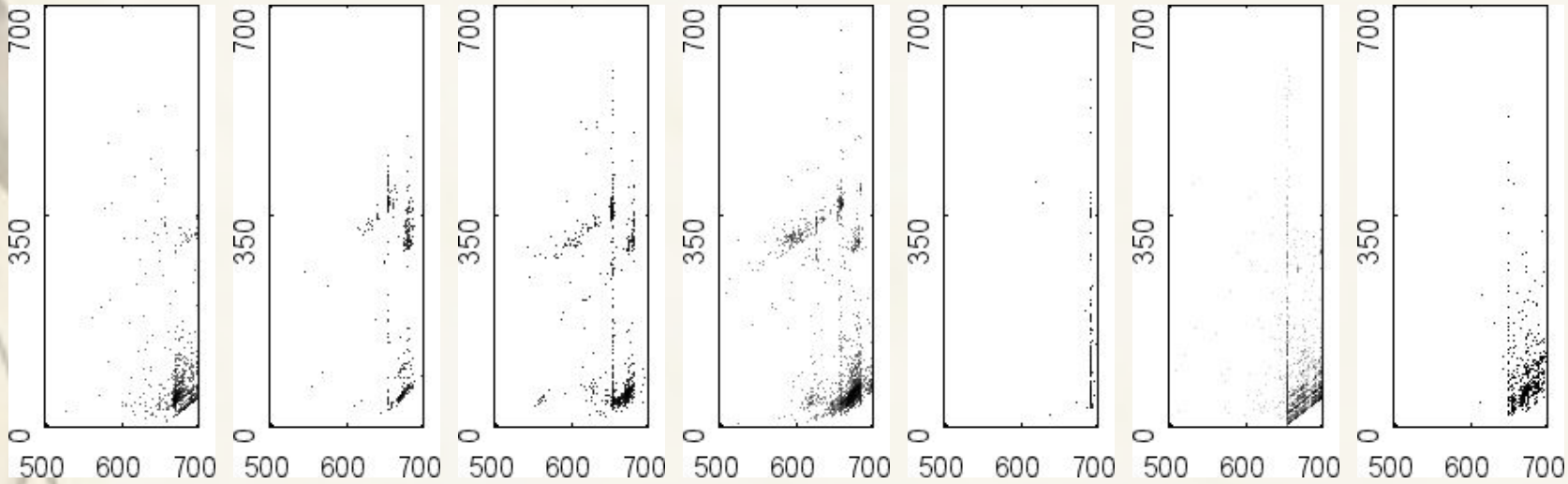

Does raw trace length affect
The number of human edits performed?

Trace length and human edits (substitutions)



x = Length of raw trace
y = Edits (substitutions) over length of alignment

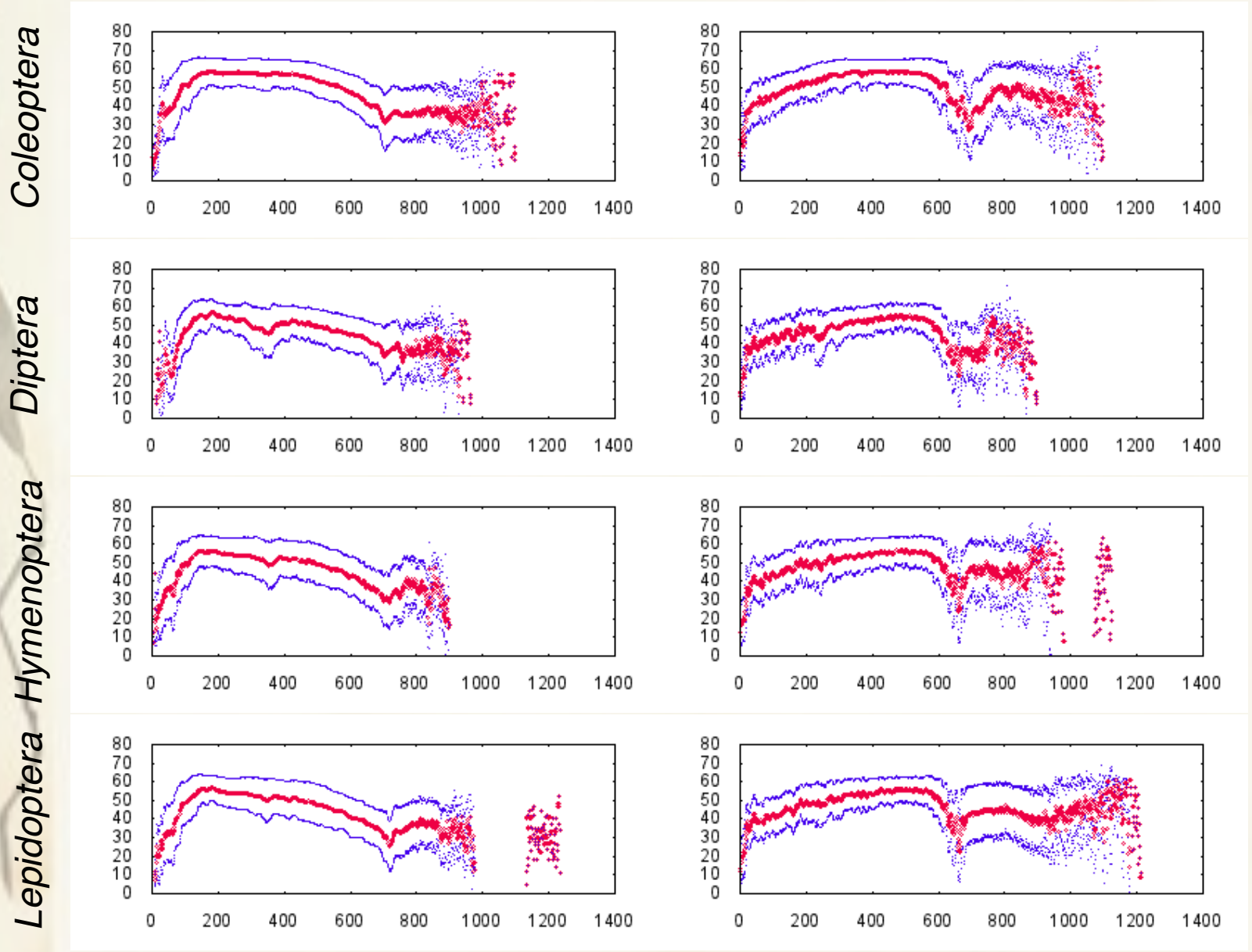
Trace length and human edits (indels)



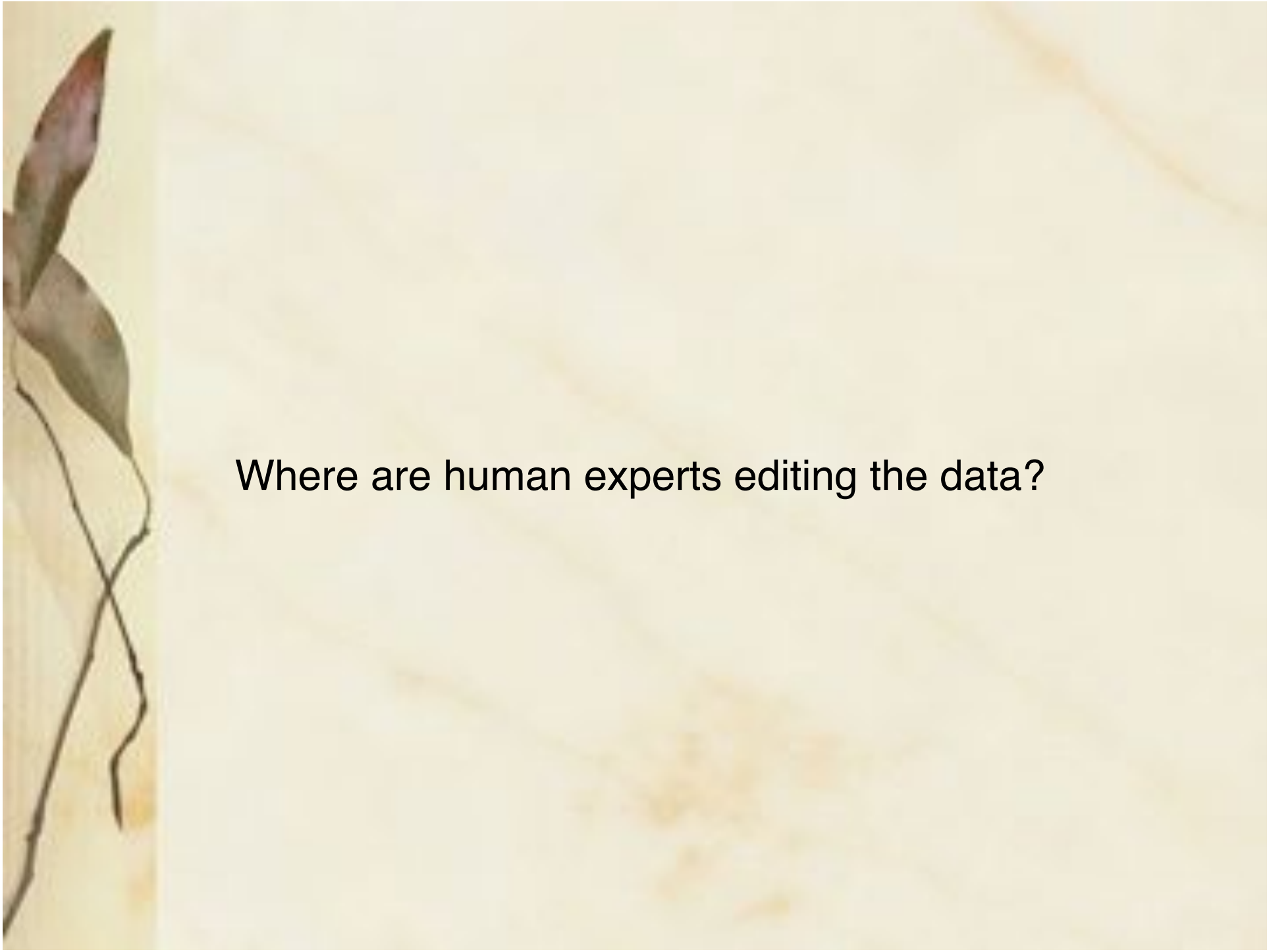
x = Length of raw trace
y = Edits (indels) over length of alignment


Is there a trend in quality values
over aligned positions?

Trace quality values at each aligned position

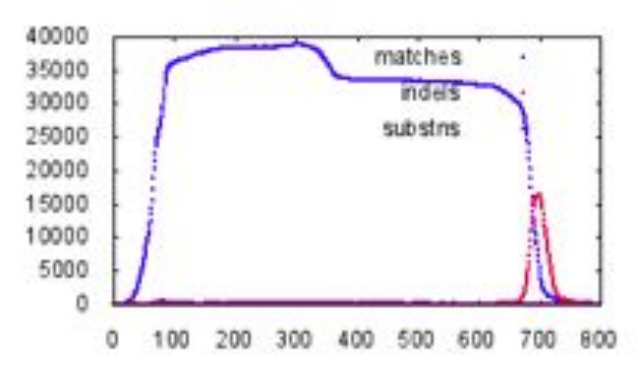


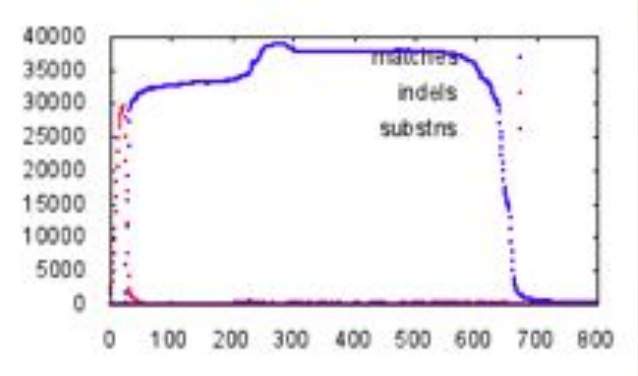
x = aligned position, y = mean trace quality (blue = s.d.)


Where are human experts editing the data?

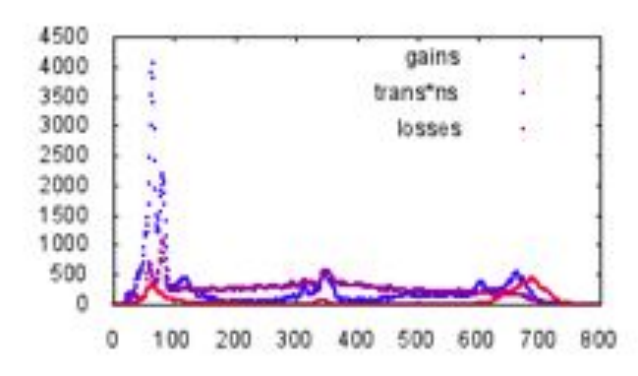
Where are edits occurring? (*Lepidoptera*)

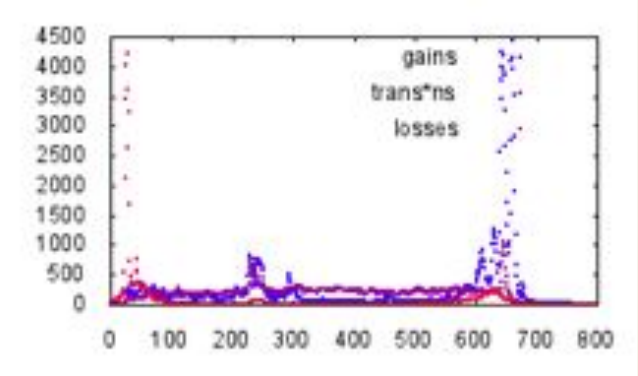
All edits



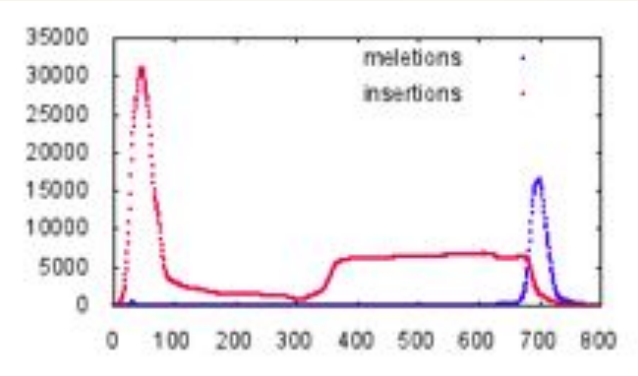


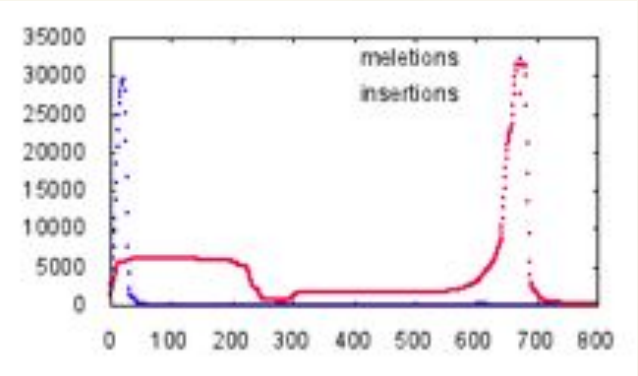
Substitutions





Indels



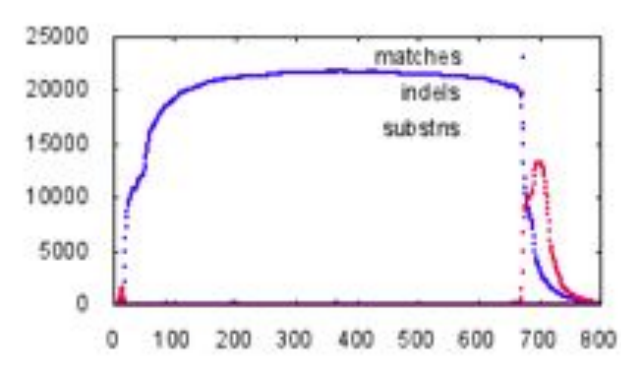


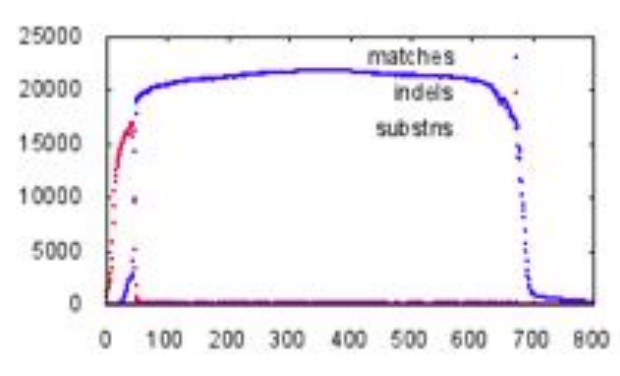
Forward traces

Reverse traces

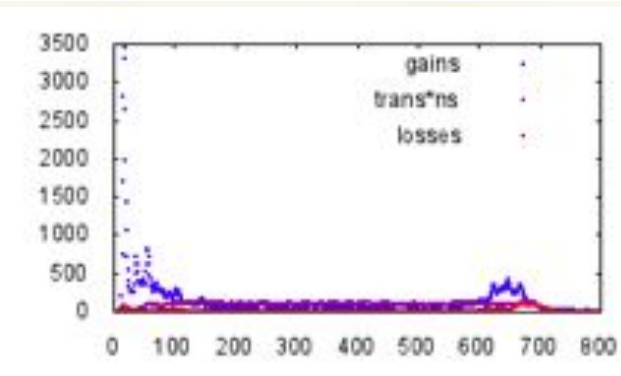
Where are edits occurring? (mammals)

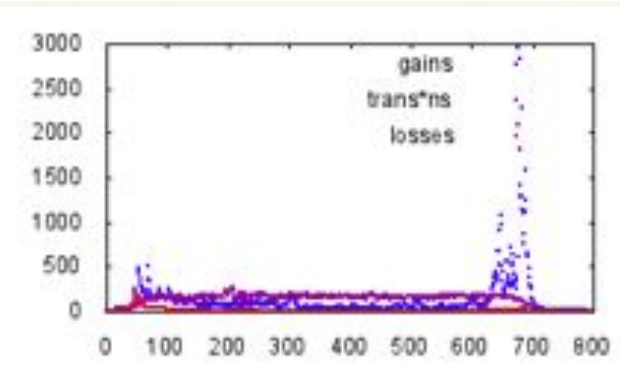
All edits



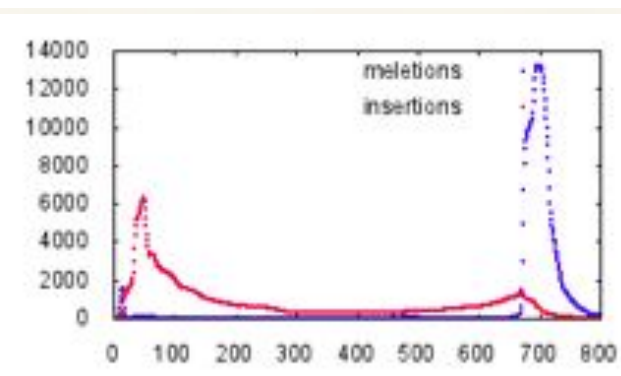


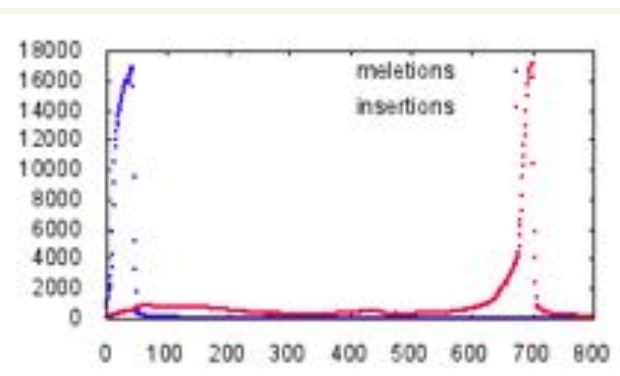
Substitutions





Indels





Forward traces

Reverse traces

Conclusion

- Many bottlenecks during barcoding
- Automation would help build contigs
- Taxonomy of edits as varied as the barcodes they describe
- Identified some trends in the data, visible from mass sampling

Next Steps

(1) Continuing work with the data

- Create a suite of characteristics for categories {phylum, class, order} depending on identity of COI
- Use this in a vector categorizing strategy to match new trace files

Vector classifying: neural networks, Bayesian classifier, etc.

Next Steps

(2) Tracefile to contig automation

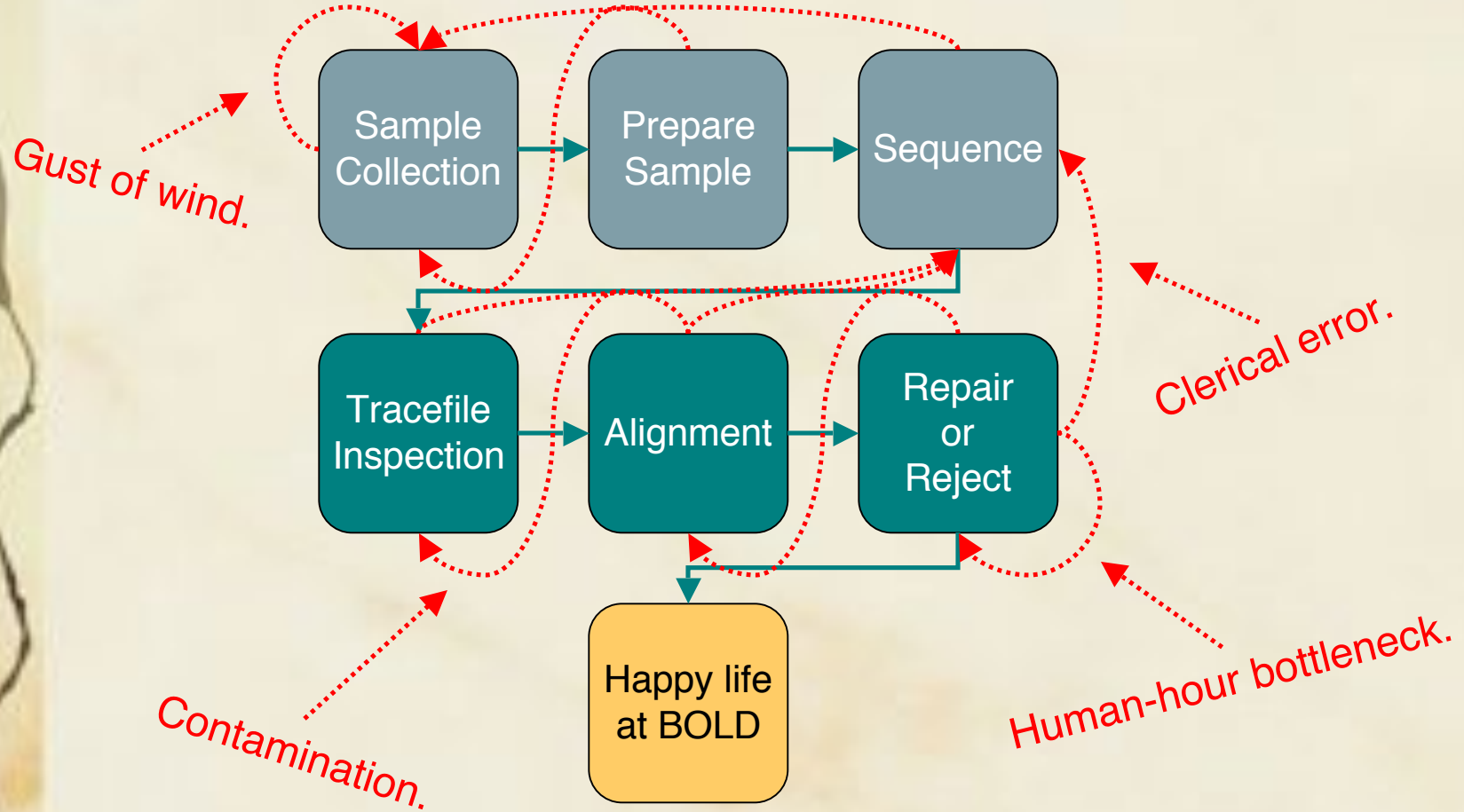
- Develop an editing profile for each of the categories based on {wave, call, quality value} characters
- Use a sequence to sequence mapping strategy to emit repaired calls based on above characters

Sequence mapping: neural networks, hidden Markov models, etc.


Until then …

Thanks for attending.

Happy barcoding!



\EOF