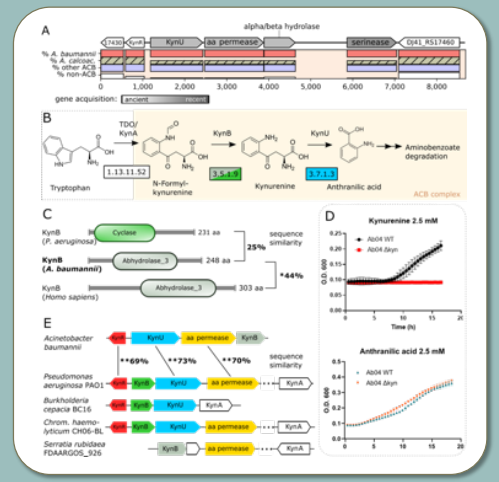
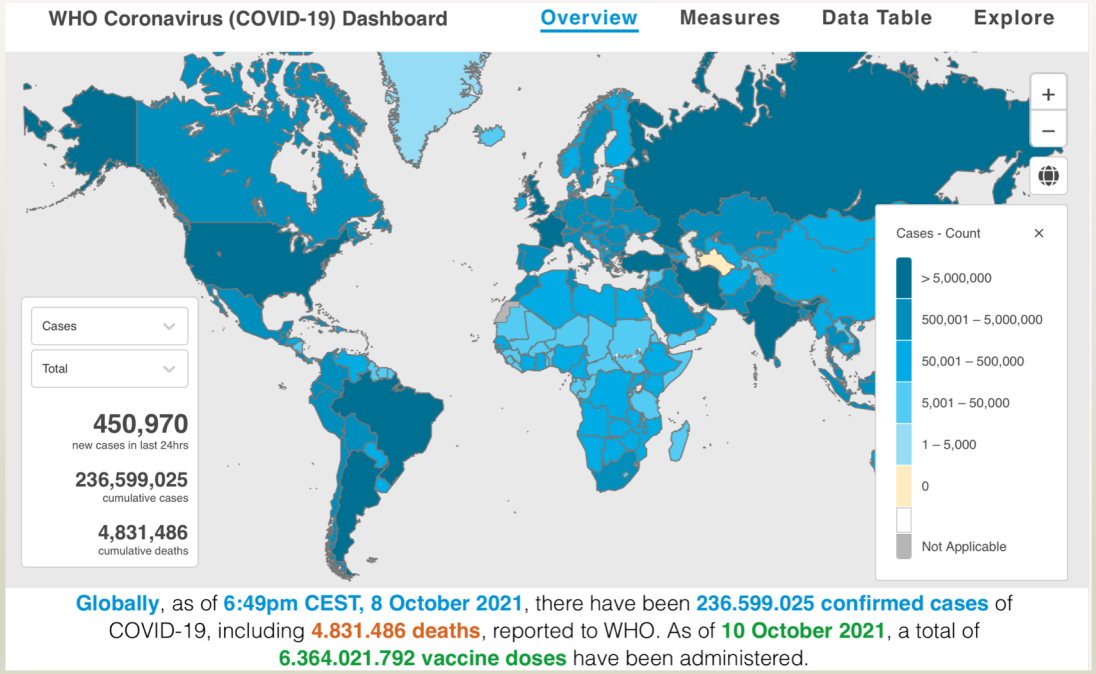
APPLIED BIOINFORMATICS FRANKFURT



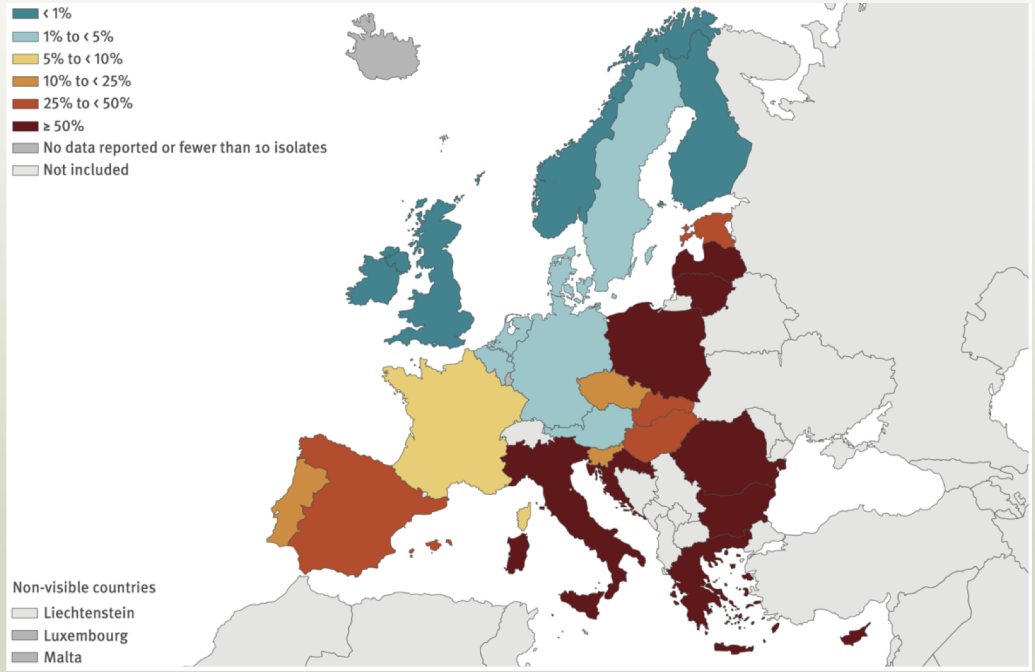
FUNCTION AND EVOLUTION OF METABOLIC PATHWAYS

A quick introduction

HUMAN PATHOGENES – A GLOBAL CHALLENGE

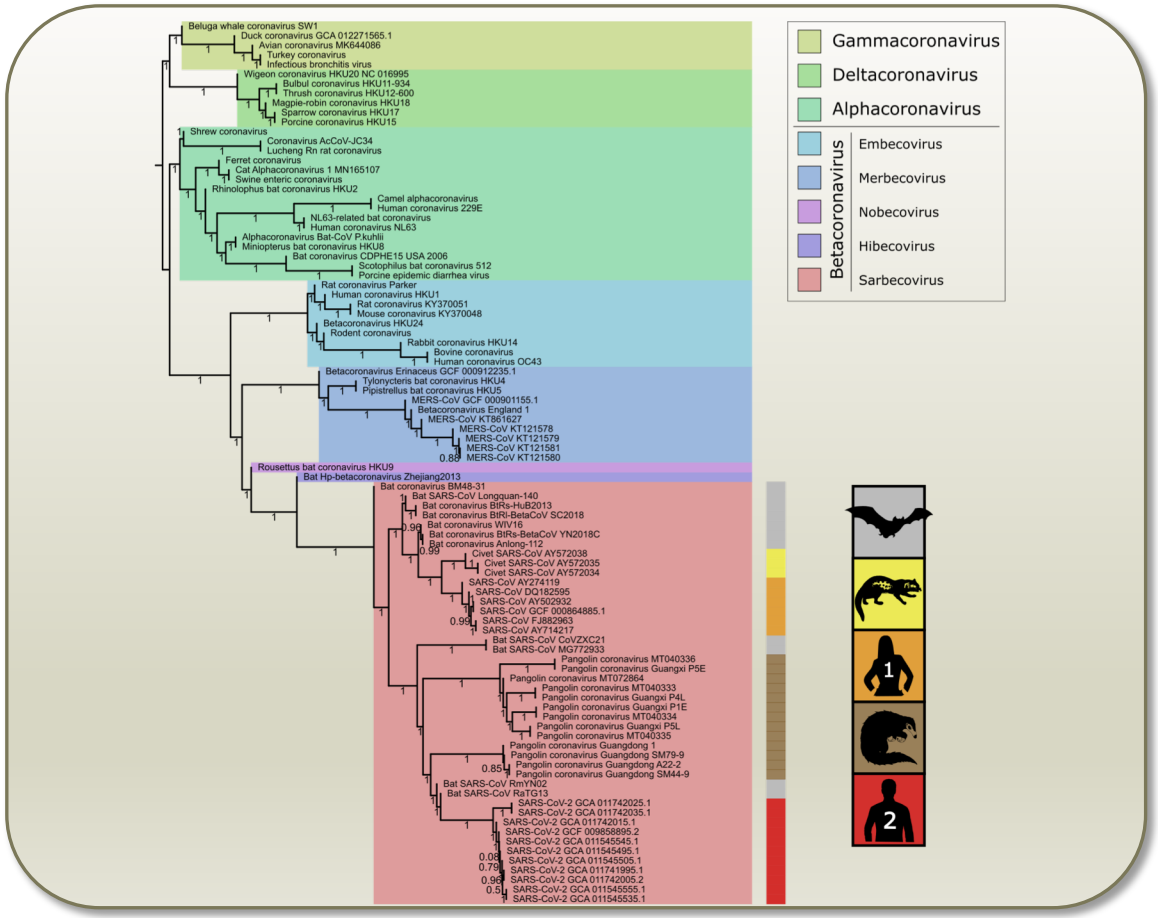


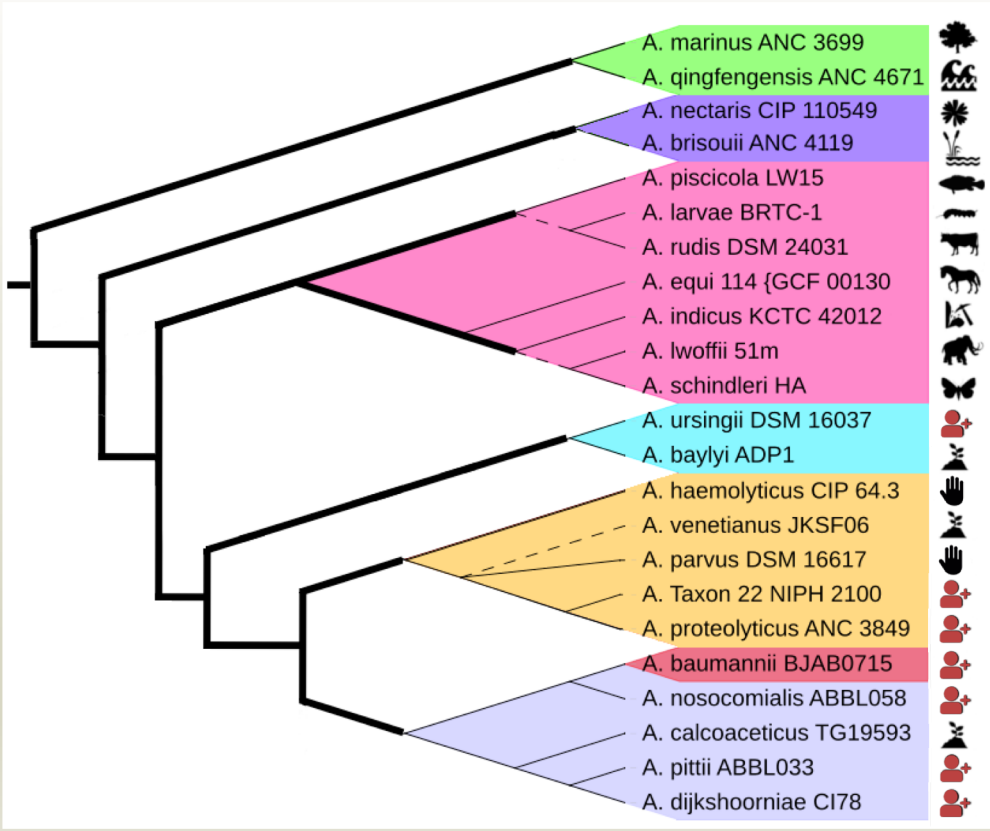
Verbreitung multi-resistenter *Acinetobacter baumannii*



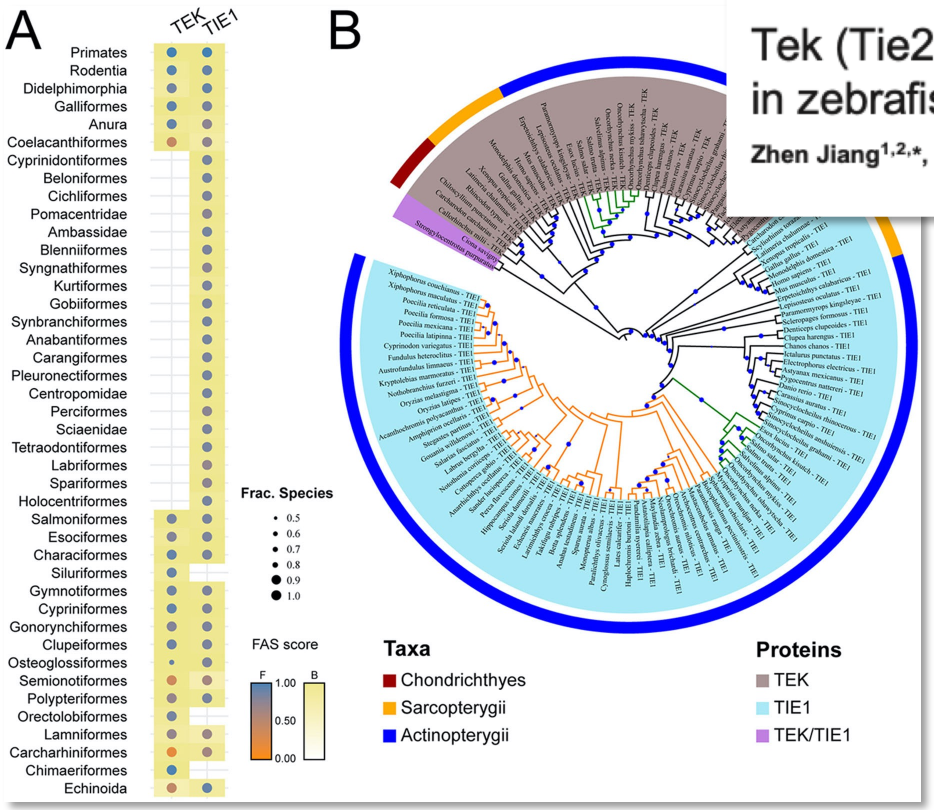
EARS Net annual report 2018 (https://www.ecdc.europa.eu)

HUMAN PATHOGENS – WHAT MAKES THEM DIFFERENT?





EVOLUTIONARY CHANGES HELP DISENTANGLING FUNCTIONAL RELATIONSHIPS IN GENERAL



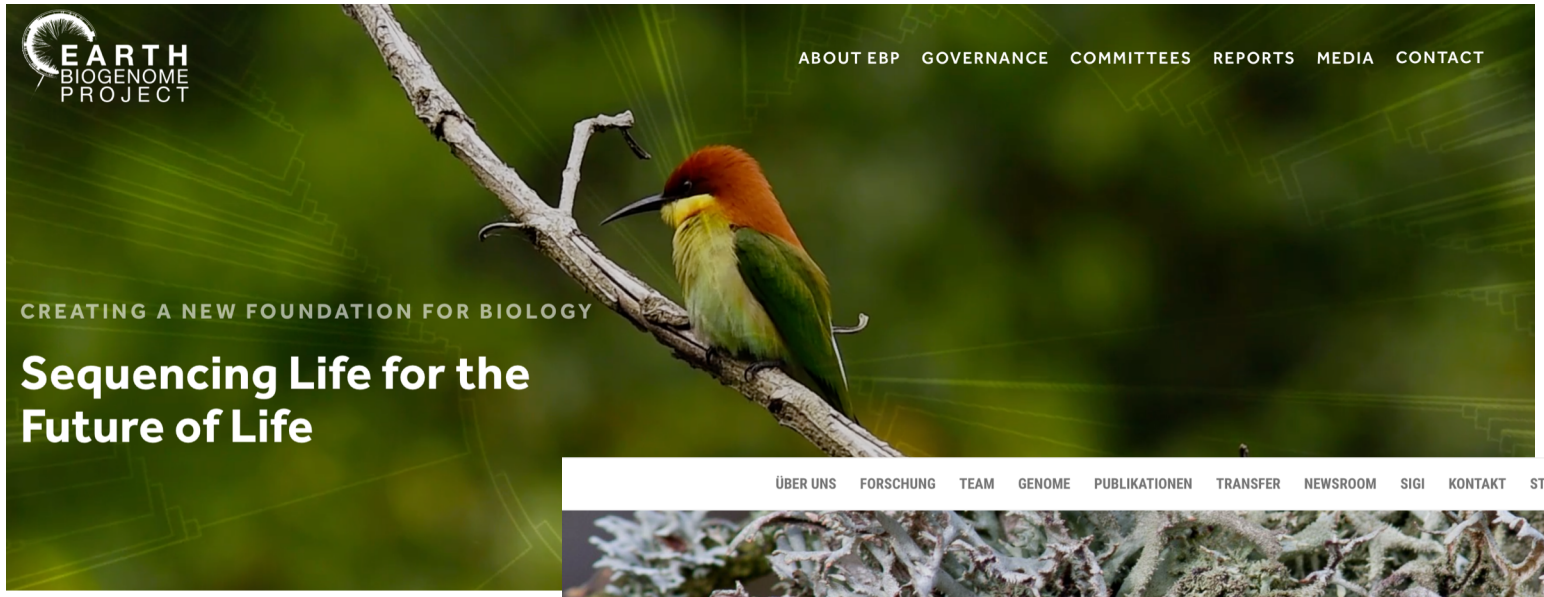
© 2020. Published by The Company of Biologists Ltd | Development (2020) 147, dev193029. doi:10.1242/dev.193029

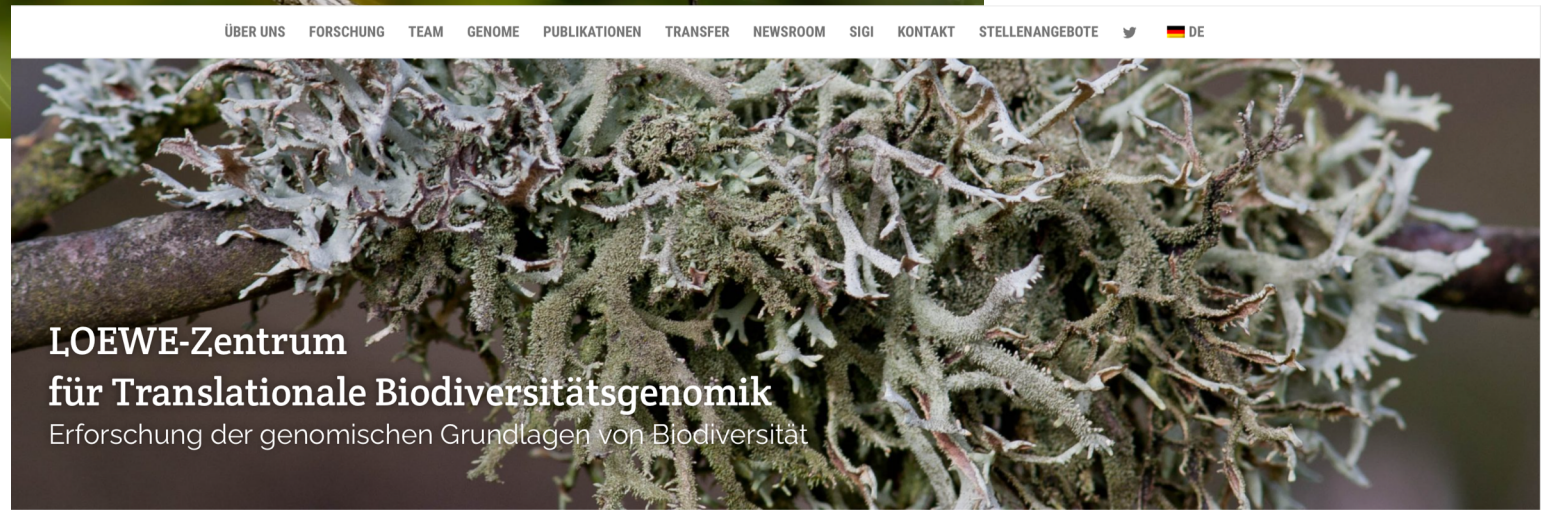
RESEARCH ARTICLE

Tek (Tie2) is not required for cardiovascular development in zebrafish

Zhen Jiang[1,2,*], Claudia Carlantoni[1,2], Srinivas Allanki[1,2], Ingo Ebersberger[3,4,5,*] and Didier Y. R. Stainier[1,2,*]

BIOLOGICAL SEQUENCES – THE KEY TO UNDERSTANDING ORGANISMIC FUNCTION AND EVOLUTION

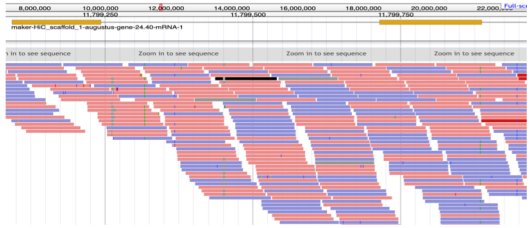




OUR DATA IS HETEROGENEOUS BOTH WITH RESPECT TO DATA TYPE AND DATA ORIGIN

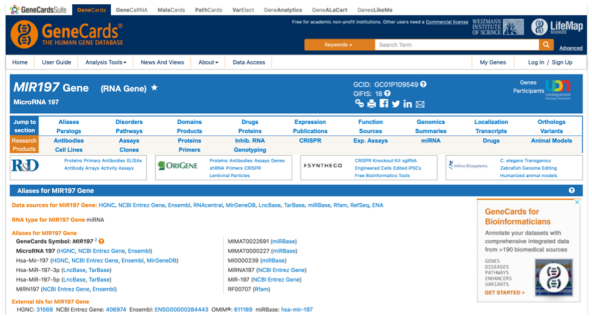
Sequence data

- Sequence reads (text)
- Sequence reads (counts)
- Genome assembly (text)
- Gene prediction (text)
- Coding sequence (text)
- translated amino acid sequence (text)



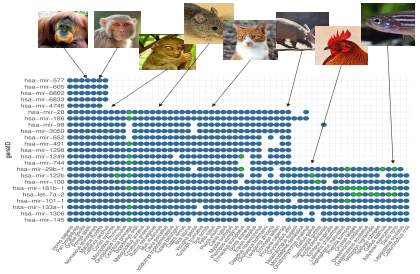
Metadata (public)

- Functional annotations (Ontologies)
- Taxon assignment (text)
- Interaction partners (text)
- …



Metadata (private)

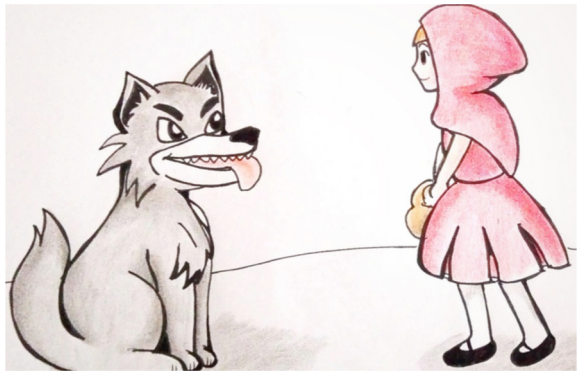
- Orthology assignments (text)
- Phylogenetic profiles (Int-n)
- Domain architectures (graph, text)
- Phylogenies (graph, text)
- Gene expression data



THE CONCEPTUAL IDEA – WORKING WITH GENOMES IS LIKE WORKING WITH TEXT

Volt egyszer egy kedves, aranyos kislány; aki csak ismerte, mindenki kedvelte, de legjobban mégis a nagymamája szerette: a világ minden kincsét neki adta volna. Egyszer vett neki egy piros bársonysapkát. A kislánynak annyira tetszett a sapka, hogy mindig csak ezt hordta; el is nevezték róla Piroskának. Piroskáék bent laktak a faluban, nagymama pedig kint az erdőben, egy takaros kis házban. Egy szép napon azt mondja Piroskának az édesanyja: – Gyere csak, kislányom! Itt van egy kalács meg egy üveg bor, vidd el a nagymamának. Beteg is, gyönge is szegényke, jól fog esni neki. Indulj szaporán, mielőtt beáll a hőség. Aztán szépen, rendesen menj, ne szaladgálj le az útról, mert elesel, és összetörik az üveg, kifolyik a bor, és akkor mit iszik a nagymama! Ha pedig odaérsz, ne bámészkodj összevissza a szobában; az legyen az első dolgod, hogy illedelmesen jó reggelt kívánj. – Bízzad csak rám, édesanyám, minden úgy lesz, ahogy mondod – felelte Piroska az intelemre, azzal karjára vette a kosárkát, és útnak indult. Átvágott a mezőn, beért az erdőbe; hát ki jön szembe vele? Nem más, mint a farkas. – Jó napot, Piroska! – köszönt rá a kislányra. Az meg mosolyogva, jó szívvel felelte: – Neked is, kedves farkas! – Nem tudta még, milyen alattomos, gonosz állattal van dolga. – Hová ilyen korán, lelkecském? – szívélyeskedett tovább a farkas. – Nagymamához. – Aztán mit viszel a kosaradban? – Bort meg kalácsot. Tegnap sütöttük; szegény jó nagymama gyönge is, beteg is, jót fog tenni neki, legalább egy kicsit erőre kap tőle. – És hol lakik a nagymama, Piroska? – Itt az erdőben, a három tölgyfa alatt. Biztosan ismered a házát, mogyorósövény van körülötte.

There was once a lovely, cute little girl; whoever knew her was loved by all, but still loved by her grandmother best: she would have given her all the treasures of the world. He once bought him a red velvet cap. The little girl liked the hat so much that she always wore it only; it was also named Piroska. The Redheads lived inside the village, and Grandma was out in the woods in a neat little house. One fine day Piroska's mother says, "Come on, baby girl!" Here's a cake and a bottle of wine, take it to your grandmother. He is both sick and weak, he will do well. Start fast before the heat sets in. Then go nicely, properly, don't run off the road because you will fall and break the bottle, the wine will flow out, and then what will your grandmother drink! And when you get there, don't stare at the room; your first thing to do is to have a decently good morning. "Trust me, my mother, everything will be as you say," Piroska replied to the rebuke, took the basket in her arms and set off. He cut through the field, entered the woods; so who will face him? He is nothing but a wolf. - Good afternoon, Little Red Riding Hood! ...



WORKING WITH GENOMES IS LIKE WORKING WITH TEXT

```
>HiC_scaffold_1
TTTAGCTCACTTGCTTATTTTGTTACATTCAGGATATACATATGGGTAAA
ACAACACTATGCGCTCTAATAACATCTCAAGTATTTTCACATAGTAAAT
TCAGTTATTTCCCGGTGCTAGACCATGAAACAATAAAACAATAAGGGTA
AATGCTAACGCTAGCTCTGGAAATACACTTGACACCCAAACAACATTC
AAACTGAGTCTTGGAACACTCTTCTTGTCTACAAAGCTAACCGAAAGG
TTTATCTGACCTTTAGGTACTATGTCAACCCATTCTTAGATGTTTTATGC
CACATTTAACGTACCTGTGGATCAGTAAGGAACGGACACAGGATTAGT
GTTATCAACTCAGGCGCTGTATTTAGTCAATACTGCCTTGAGTTGGGAG
TTTTCCTACTAGCCCACAGCGACACATGCTGTCGCACGAAGGTAACTA
CATTCAACCCAAAGGTTTGAAAAAAAGAAAACTTATTTATAAAGTAGCT
GACGATTTAGATGTTGTACAAAAAAAATAAAAATAAGGGGTTGTCAGTT
TTTATATTAAACTTCAAATAGAGGCTATTTTCCCATGCTCTTTATTAAAA
AAATAAAAAAATAAGGGGGTCACGTGATGACACCGAGTAAGATGACTG
CTTAAACCTGCGGCTGCGGCACCACTAGCTCACATTTGTGTGAACATAT
AGGCATTTGCAACTTATTACACGTTTTTTTTCCCCCACAAGTTGCCTAT
TCTAAGCGTCTTCCACTACCAATATGCCGAATCCCCGGAAAAACGAAA
ATGCGTCTCAACAACAGCAAAGCGCCAGGGCTAGCGCTAGCTTAAAAG
CCCCCCACGCCTCGCCTGATGACGCAGATTCCATCTCCACTGCTAAAAT
ACTCGAGCTTCTTCAGAAACTATCAGACGATCAAGCCAAACACTTTGC
AGACATCCGCAAGGACGTCTCAGACATAAGAAGAGCTATCGAGGCGAT
TGAGGGTAAGCTAGCCGACATGCTGACAAGGATGACTAGCGCAGAGGC
TCGCCTTGACATGCTGGAGGAGGCGGAGGGACGGCGCCGCAACGCGC
CGCCGGCCTCAGCCTCTGAAGTTGAAAAACTGAACGCCAAATTAACTG
AGGTCGAGGACCGGGAAAGGCGAGTGAATCTACGCATTTATGGTTTCC
```

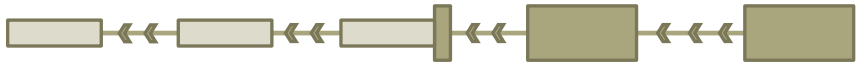
Genome Annotation

```
HiC_scaffold_1  maker  mRNA     1677  5926  .  -  .  ID=augustus-gene-0.1
HiC_scaffold_1  maker  exon     1677  1769  .  -  .  ID=augustus-gene-0.1
HiC_scaffold_1  maker  exon     1960  2079  .  -  .  ID=augustus-gene-0.1
HiC_scaffold_1  maker  exon     2168  2232  .  -  .  ID=augustus-gene-0.1
HiC_scaffold_1  maker  exon     2436  2503  .  -  .  ID=augustus-gene-0.1
HiC_scaffold_1  maker  exon     3300  3436  .  -  .  ID=augustus-gene-0.1
HiC_scaffold_1  maker  CDS      5909  5926  .  -  0  ID=augustus-gene-0.1
HiC_scaffold_1  maker  CDS      3300  3436  .  -  0  ID=augustus-gene-0.1
HiC_scaffold_1  maker  CDS      2436  2503  .  -  1  ID=augustus-gene-0.1
HiC_scaffold_1  maker  CDS      2228  2232  .  -  2  ID=augustus-gene-0.1
HiC_scaffold_1  maker  3'_UTR   2168  2227  .  -  .  ID=augustus-gene-0.1
```

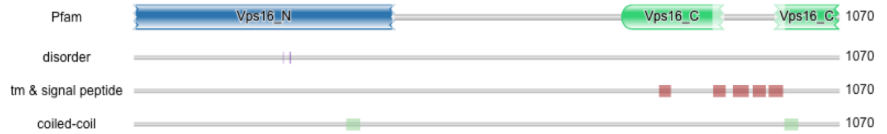
Protein Sequence

```
>maker-HiC_sc_1_gene-0.1
MCWNLKDGLRDSLVSAAPYGGPIA
LLREPHRRSPSSRPQLEIYSASGV
GIASFPWKSGPVVHLGWTVSDELL
CIQEDGSVLIYDLFGSFKRHFSMG
QEVVQSQVLEAKVFHSPYGTGVAI
VTGSSRFTLATNIDDLKLRRLPEV
PGLQGKPSCWVVLTQDRQTKVLLS
NGSELFILDSTSCTAVCPPGLGPQ
AGSVVHMSVSFSYKYLAL.....
```

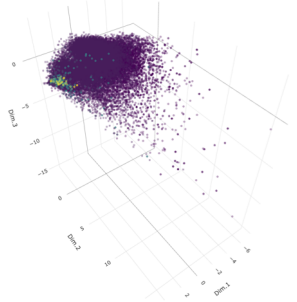
Gene Structure



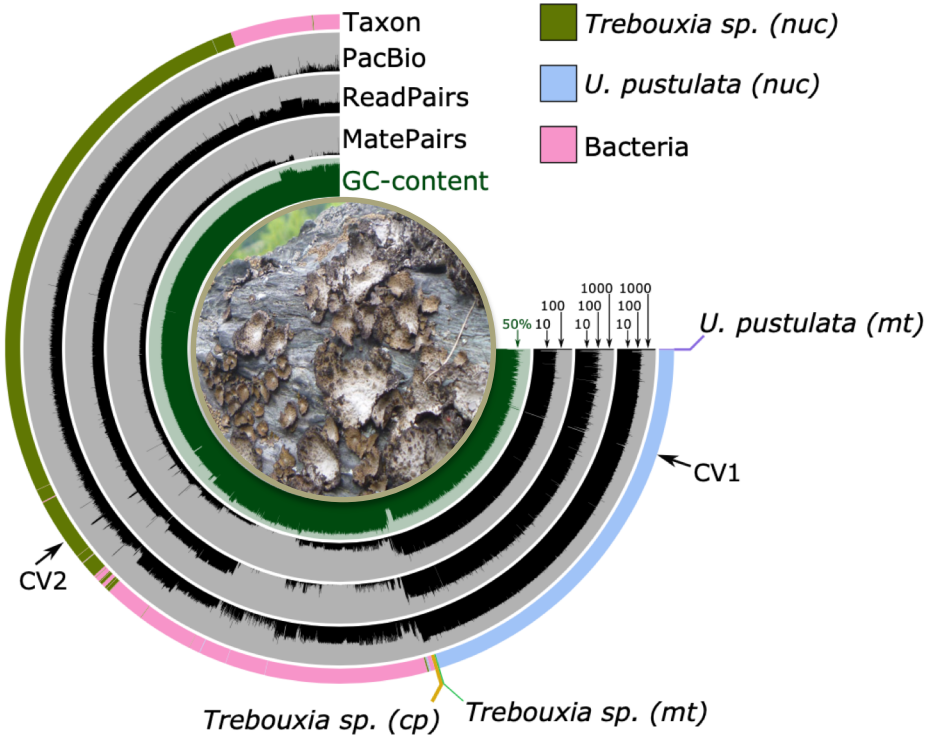
Domain Architecture



Taxonomic assignment

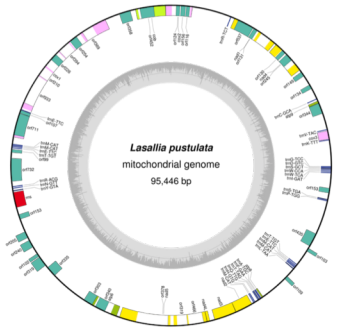


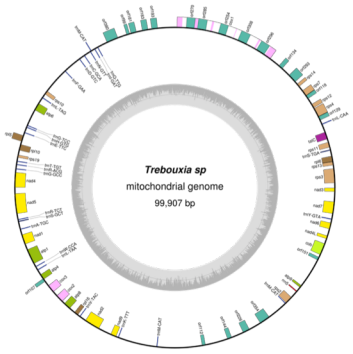
GETTING THE TEXT – RECONSTRUCTING GENOMES FROM SEQUENCE READS

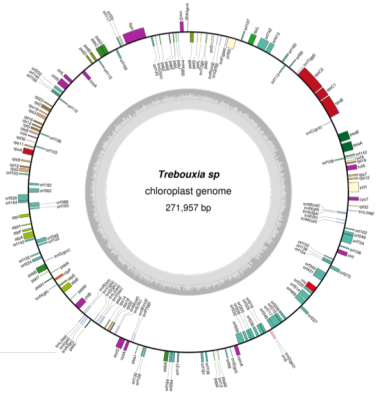


Taxonomic Assignment using *MEGAN*.
Error Correction with *PILON*
Genome Annotation with *MAKER / PROKKA*.

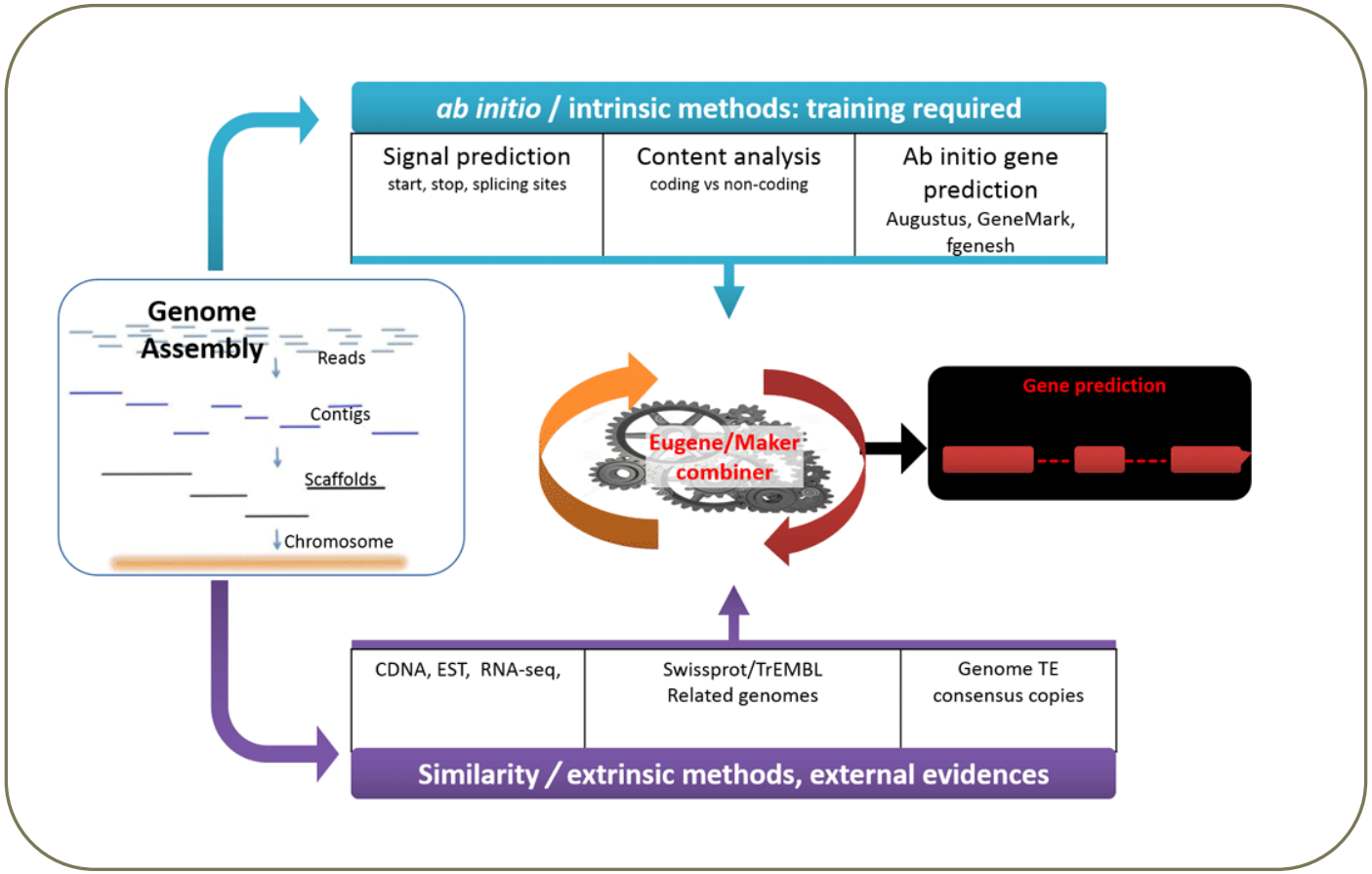
	Lasallia pustulata	***Trebouxia sp.***	**Bacteria**
# scaffolds	43	225	499
total length	33 Mbp	53 Mbp	35 Mbp
N50	1.8 Mbp	0.8 Mbp	234,856 bp
coverage	~100x	~10x	–
# of genes	9,825	13,919	~ 36,000

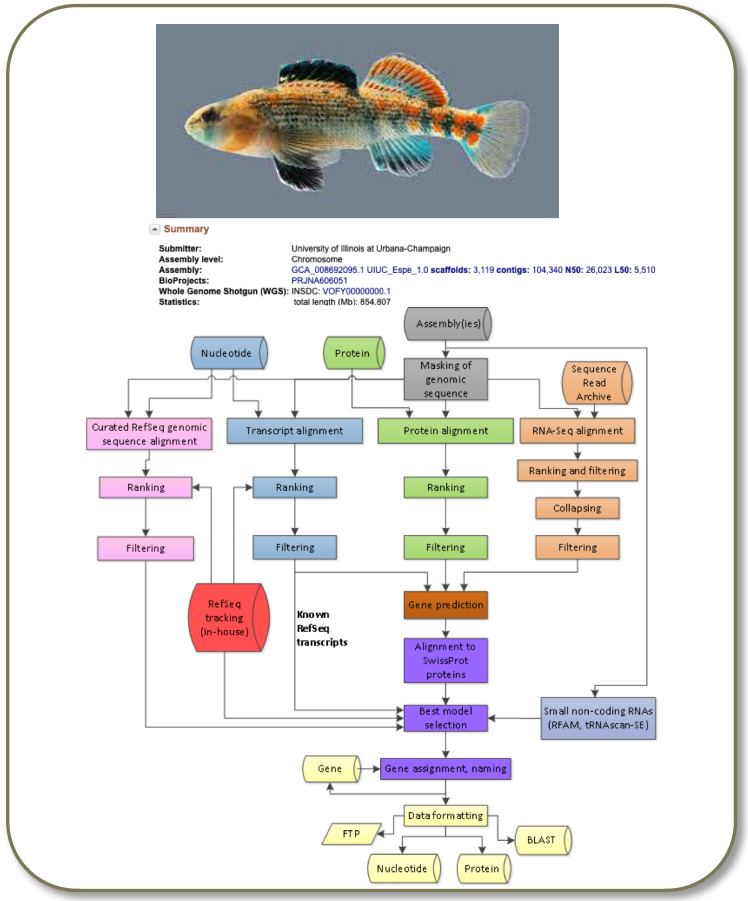




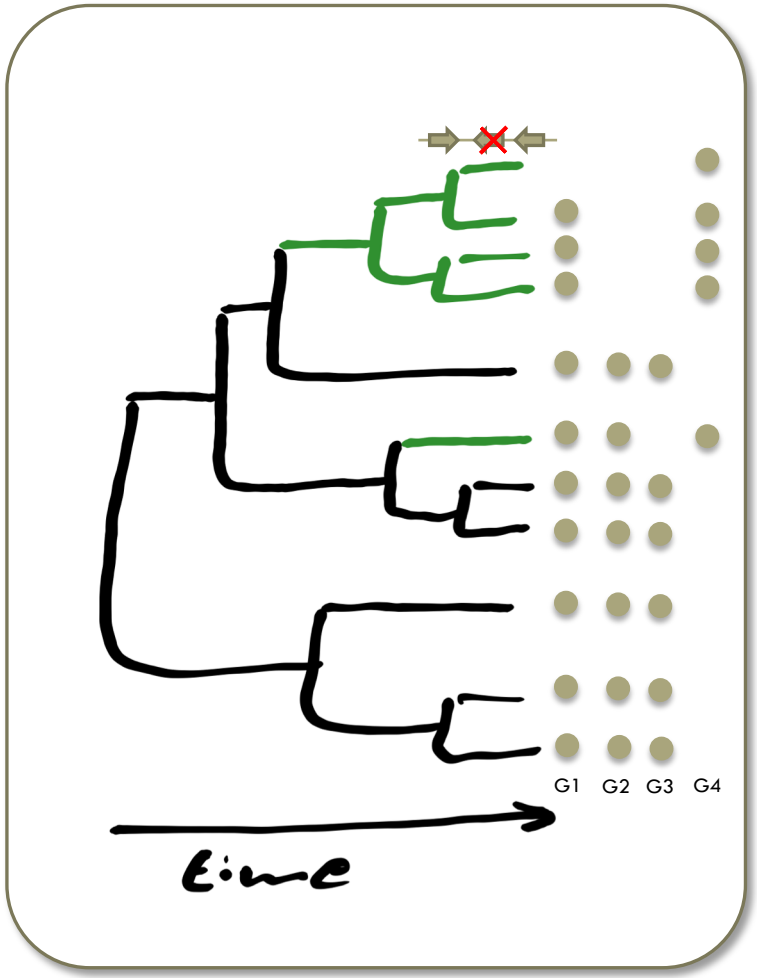


ANNOTATING GENOMES – HOW COMPREHENSIVE CAN YOU BE?

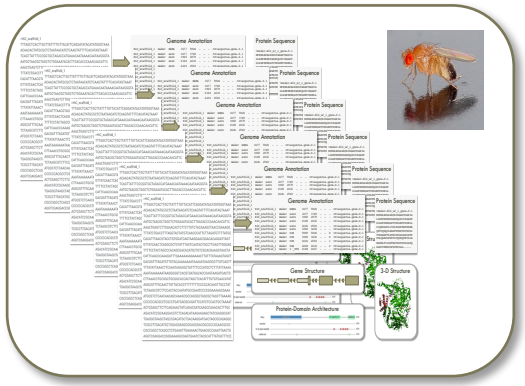


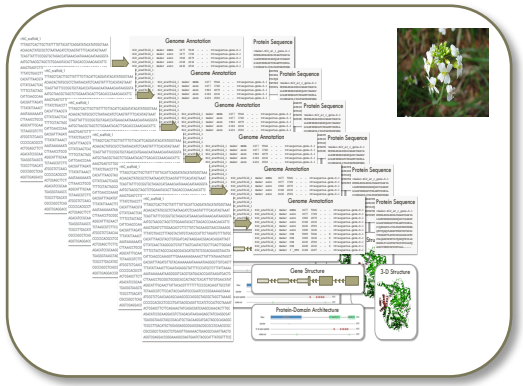


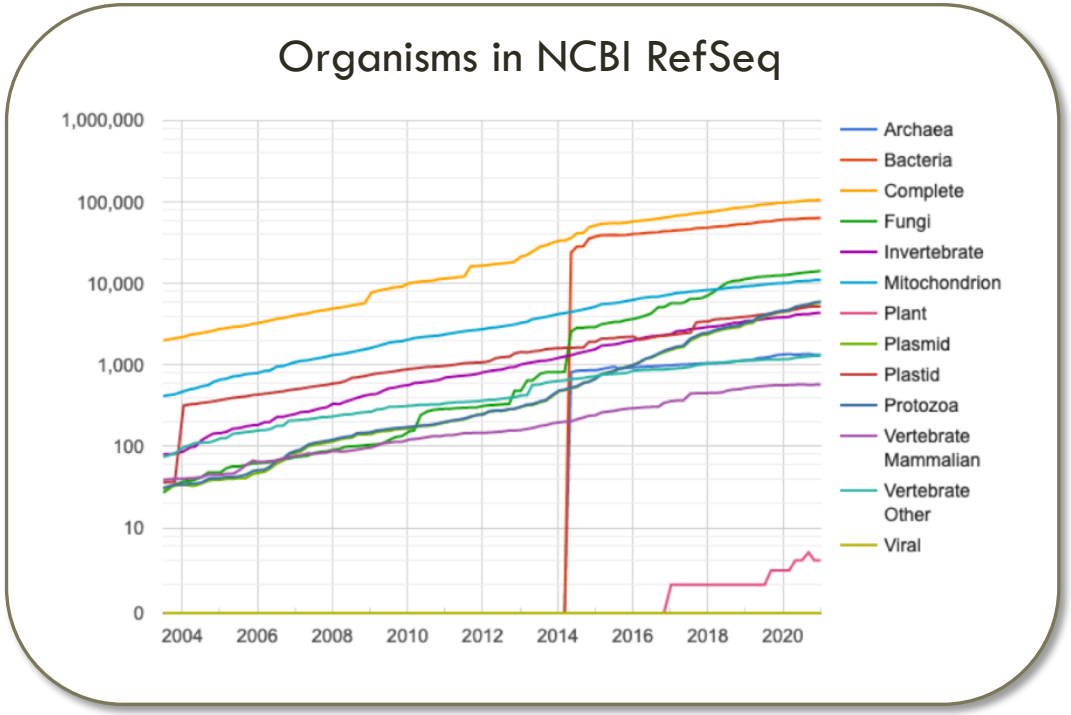
WHAT IS SPECIAL/DIFFERENT IN *OUR* RESEARCH OBJECT?

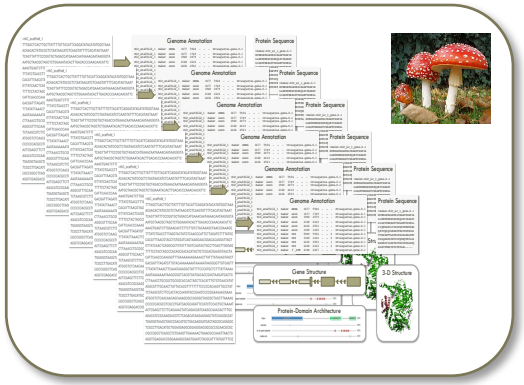


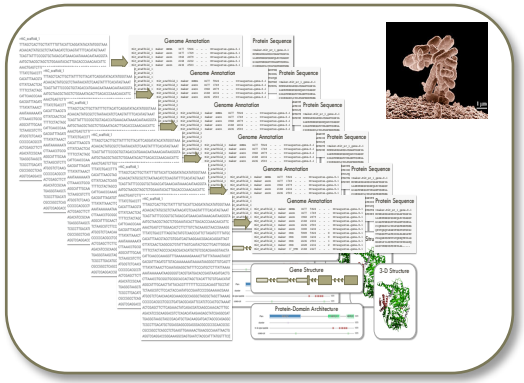
INTEGRATE OUR RESEARCH OBJECT INTO THE PUBLICLY AVAILABLE DATA



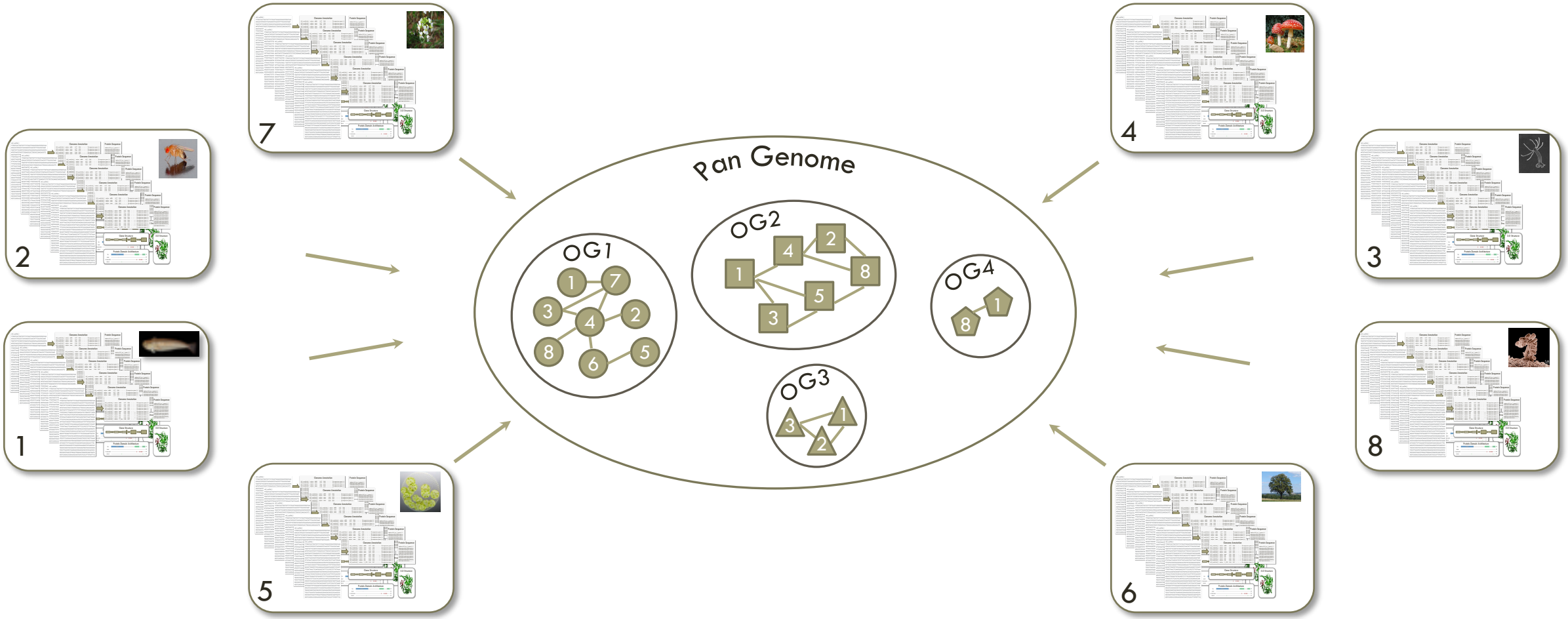






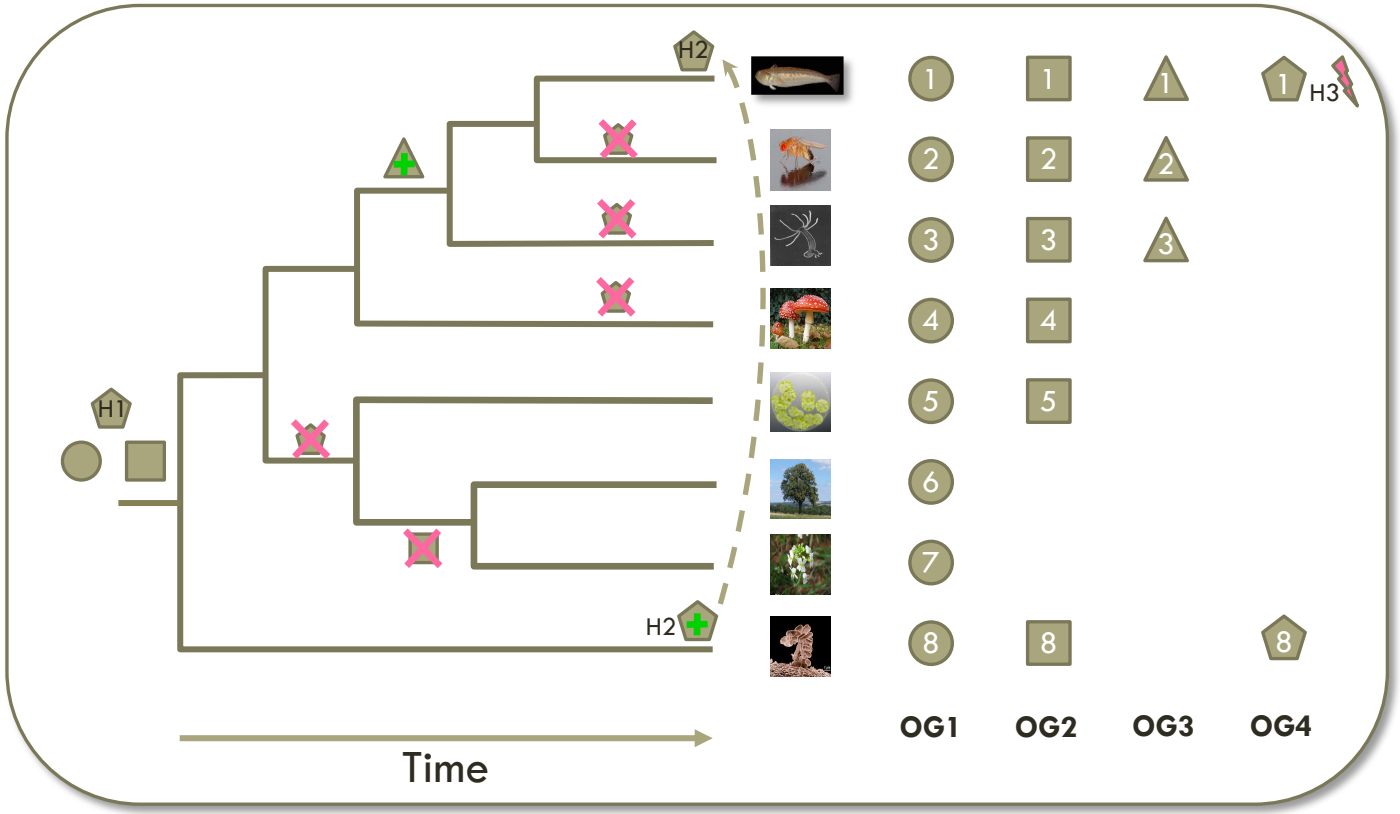


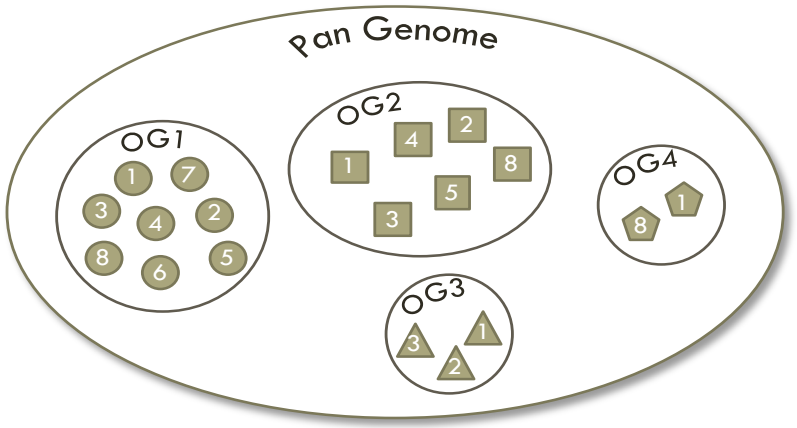
...VIA THE IDENTIFICATION OF ORTHOLOGOUS GROUPS



* in fact, we should replace ‚genome' with ‚gene set'

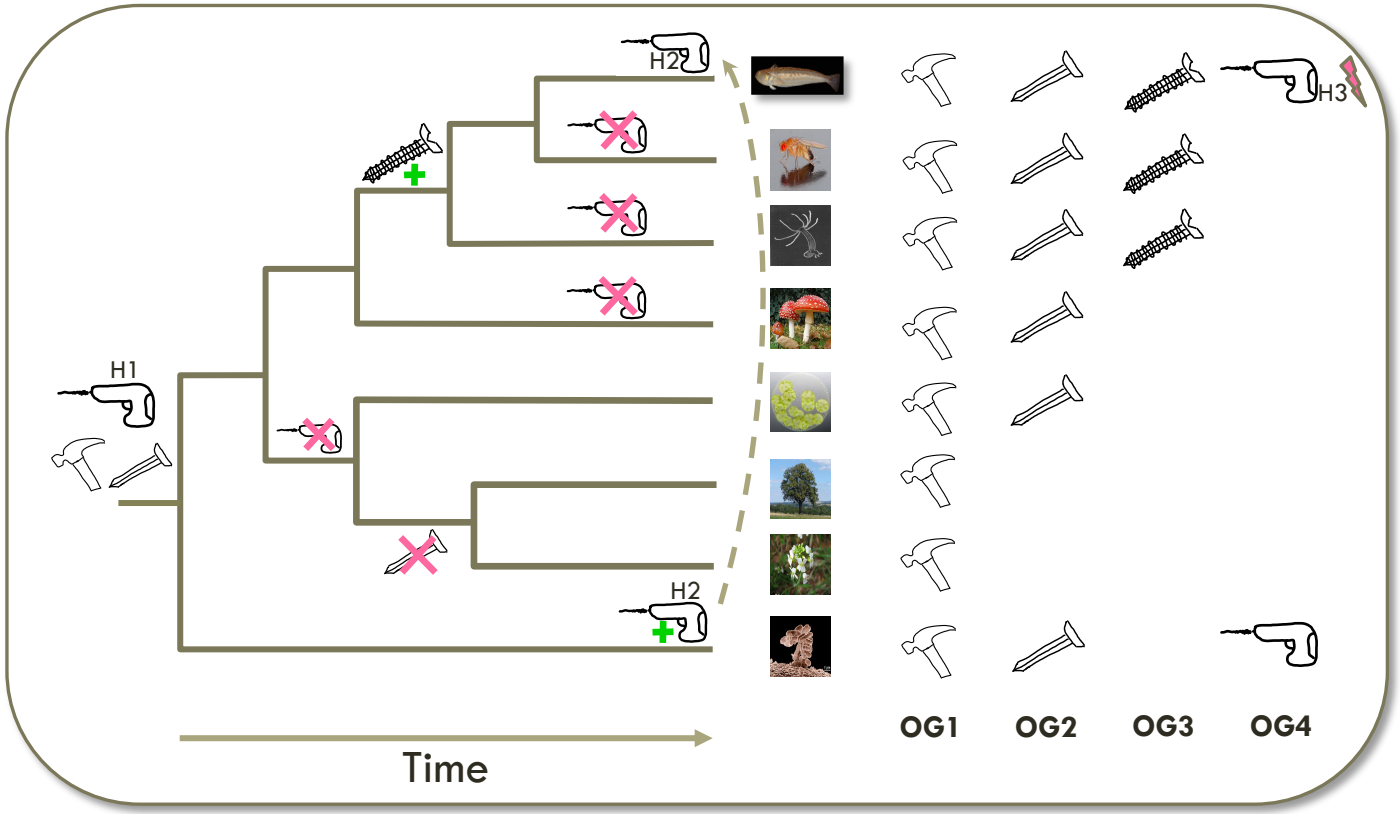
FROM ORTHOLOGOUS GROUPS TO PHYLOGENETIC PROFILES

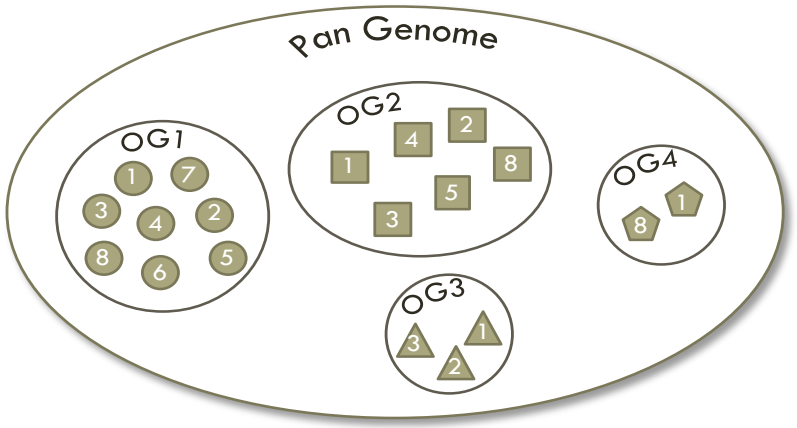




OG1 – core gene
OG2 – gene loss
OG3 – gene gain
OG4 – multiple independent gene loss (H1) / horizontal gene transfer (H2)/ contamination (H3)

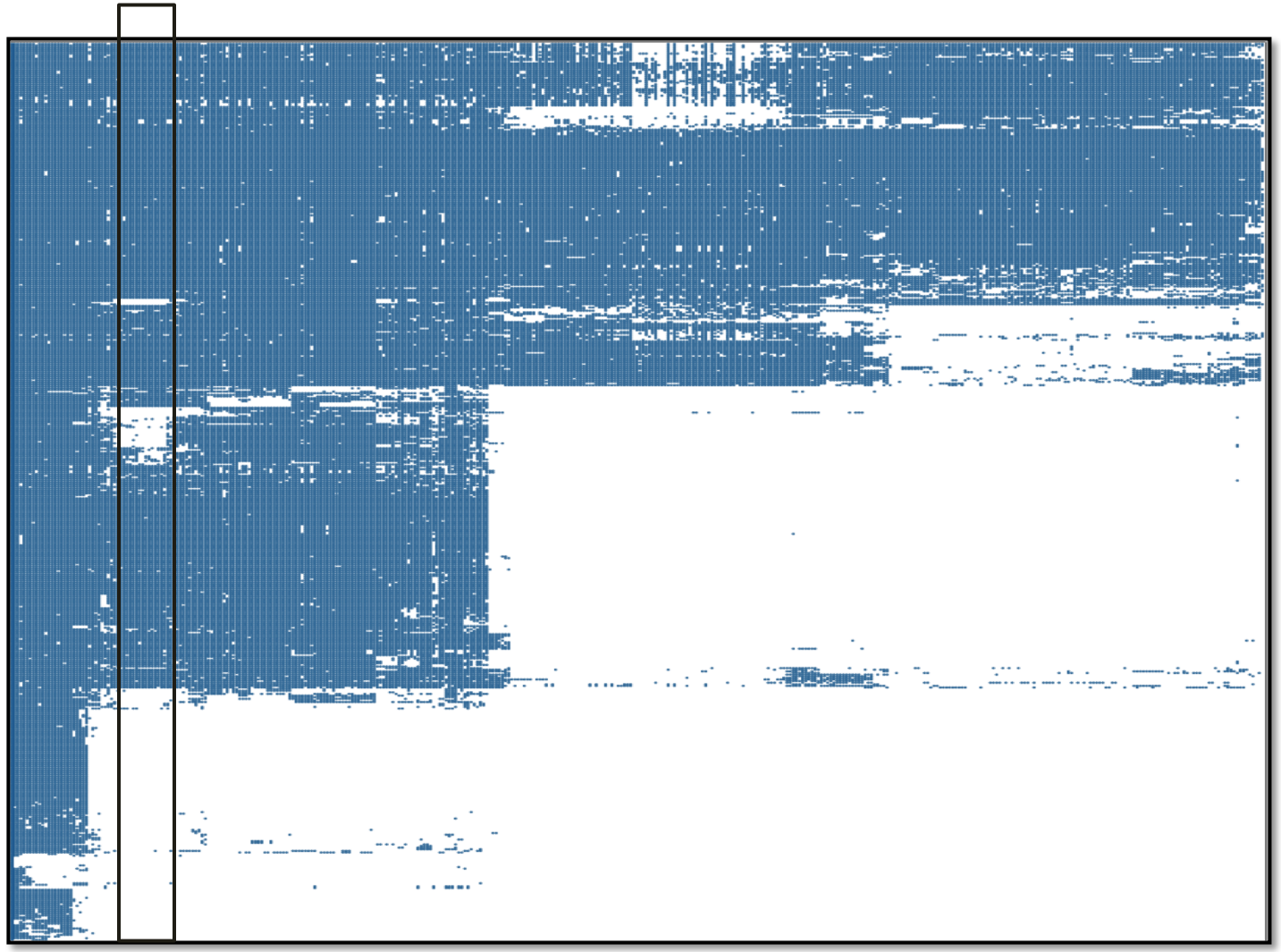
FROM PHYLOGENETIC PROFILES TO EVOLUTION OF FUNCTION



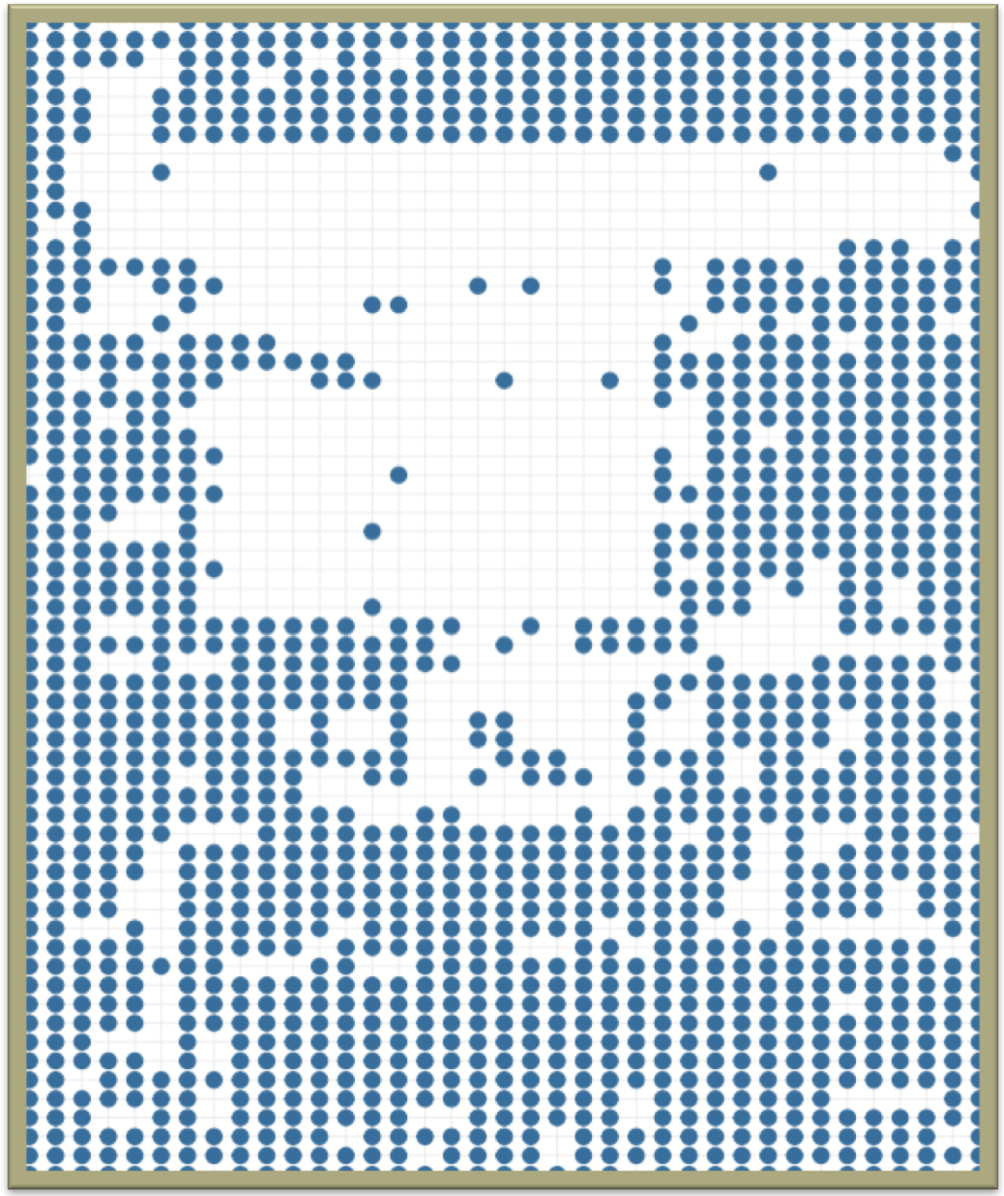


OG1 – core gene
OG2 – gene loss
OG3 – gene gain
OG4 – multiple independent gene loss (H1) / horizontal gene transfer (H2)/ contamination (H3)

WHAT IS OUR SIGNAL?

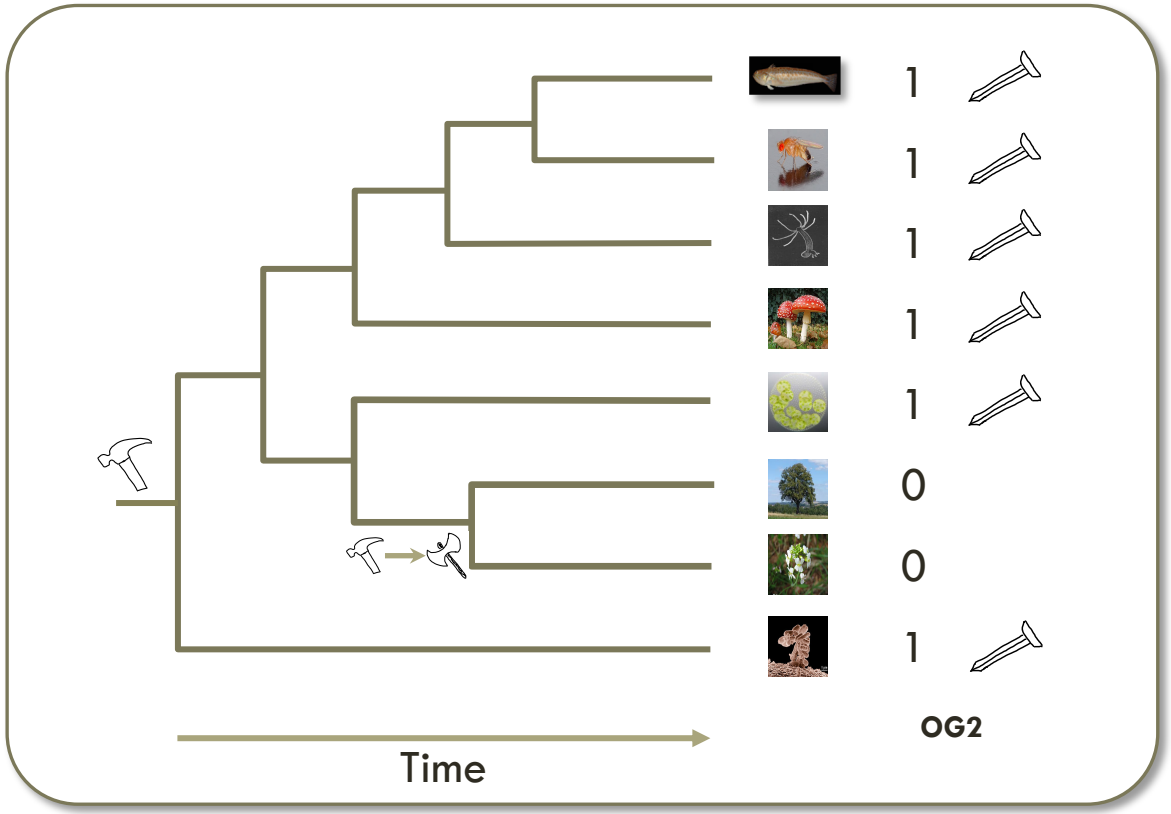


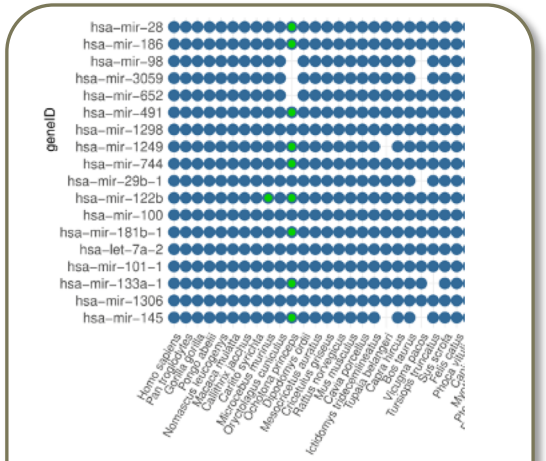
Shared presence or absence of genes by members of a systematic group



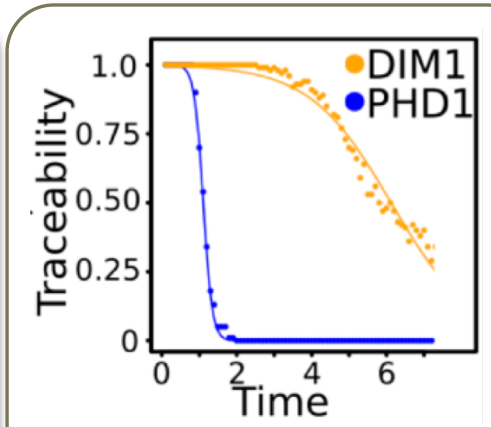
NOISE IS A PROBLEM WHEN YOUR SIGNAL IS A CHANGE!

ERROR SOURCE 1 – FALSE NEGATIVES

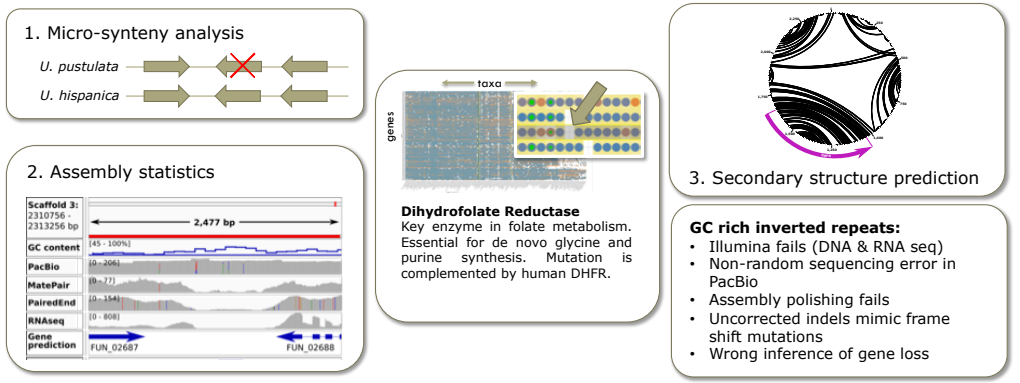




Increase Taxon sampling

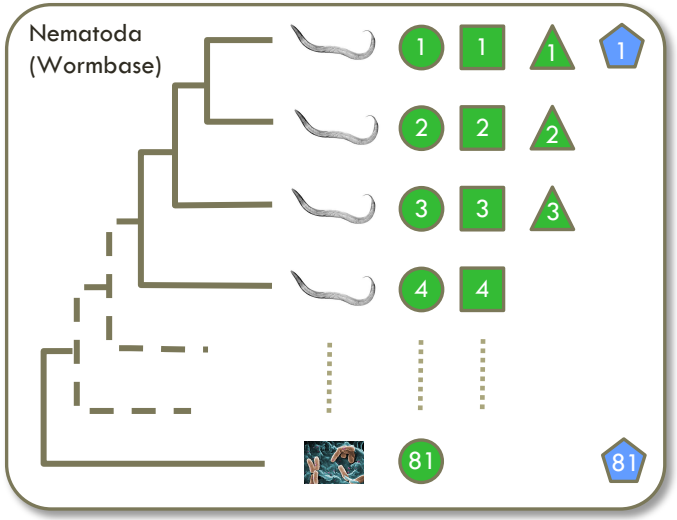


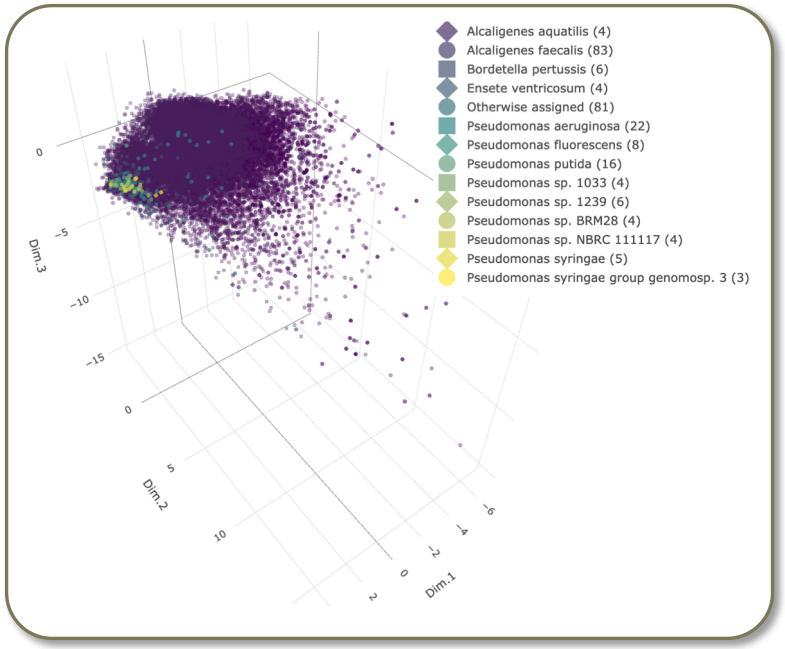
Infer limits of sensitivity



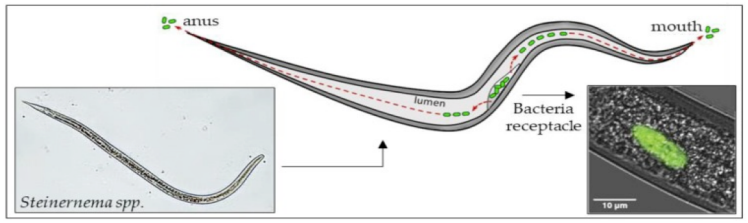
Look hard at the data

ERROR SOURCE 2 – FALSE POSITIVES





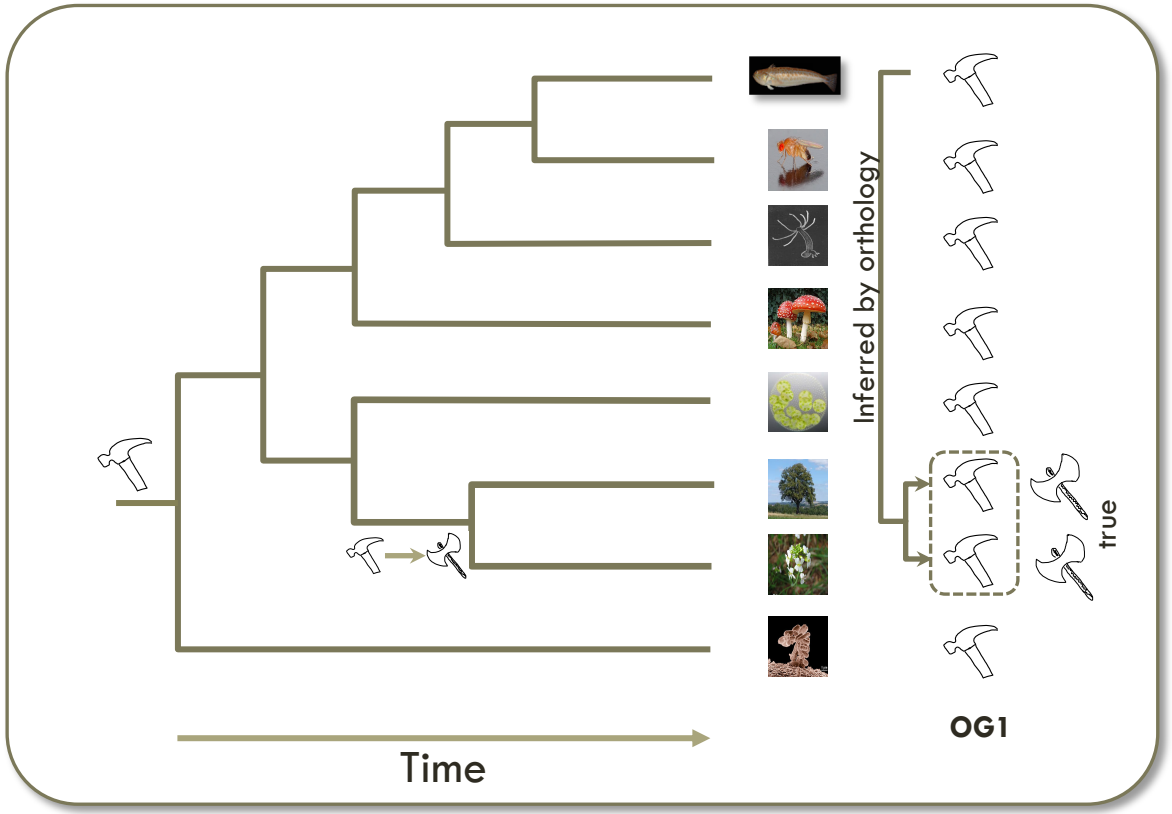
The pathobiome of *S. glaseri*

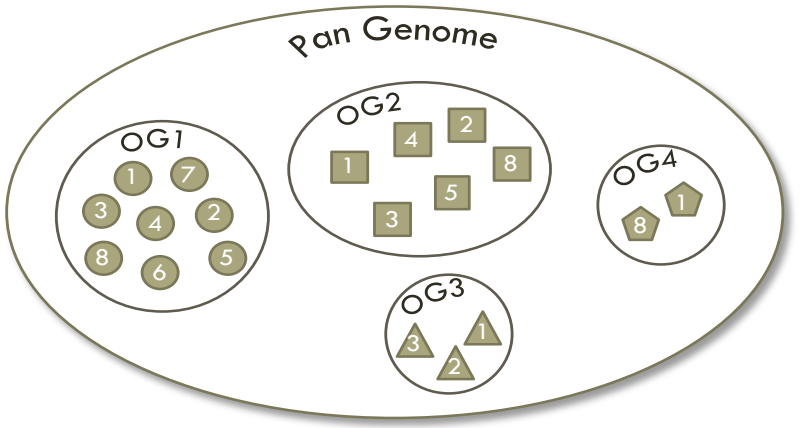


Brivio et al. (2018) Insects, 11 Sep 2018, 9(3)

Perisa et al. Contaminated gene sets inform about ecology and evolution of nematodes (in prep)

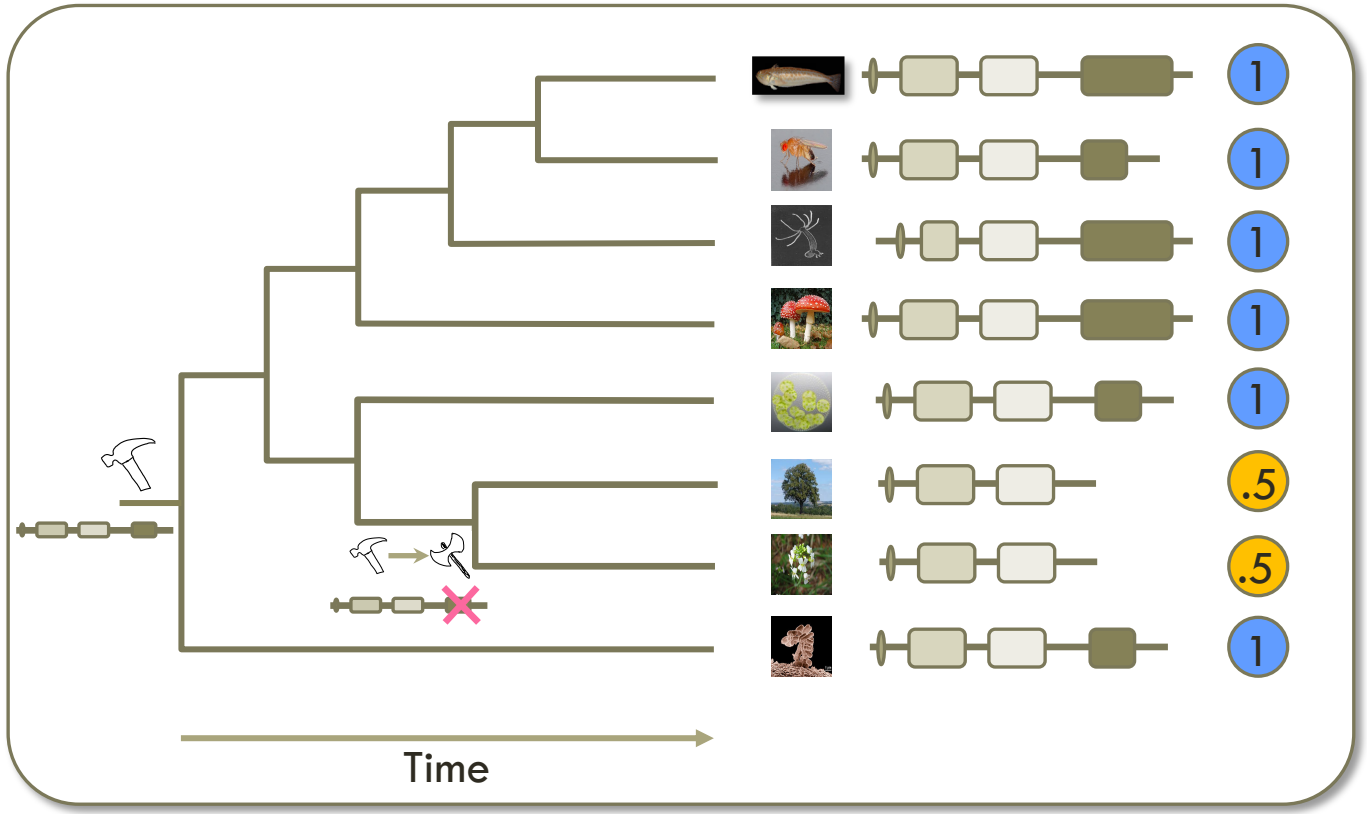
OBSERVED DIFFERENCES – EVOLUTIONARY SIGNAL OR ARTEFACT? EXAMPLE 3: WRONG CONCLUSIONS





H0 – Orthologous proteins are functionally equivalent

FEATURE ARCHITECTURES CAPTURE SIGNALS OF FUNCTIONAL DIVERGENCE



Abundance Score

$$MS(S,O) = \sum_{i=1}^{N^S} (\omega_i * \min\left(\frac{N_i^S * N_i^O}{(N_i^S)^2}, 1\right)$$

Positional Score

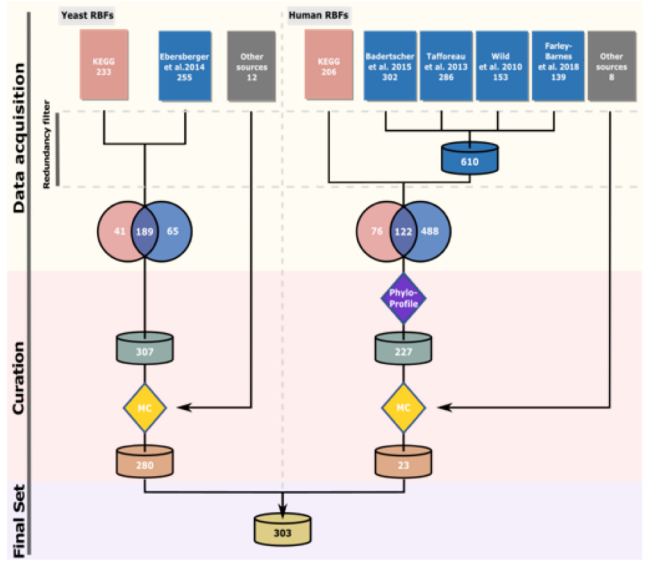
$$PS(S,O) = \sum_{i=1}^{N^S} \frac{\omega_i}{N_i^S} * \sum_{j=1}^{N_i^S} (1 - min_{1 \le l \le N_i^O} |P_i^S, j - P_i^O, l|)$$

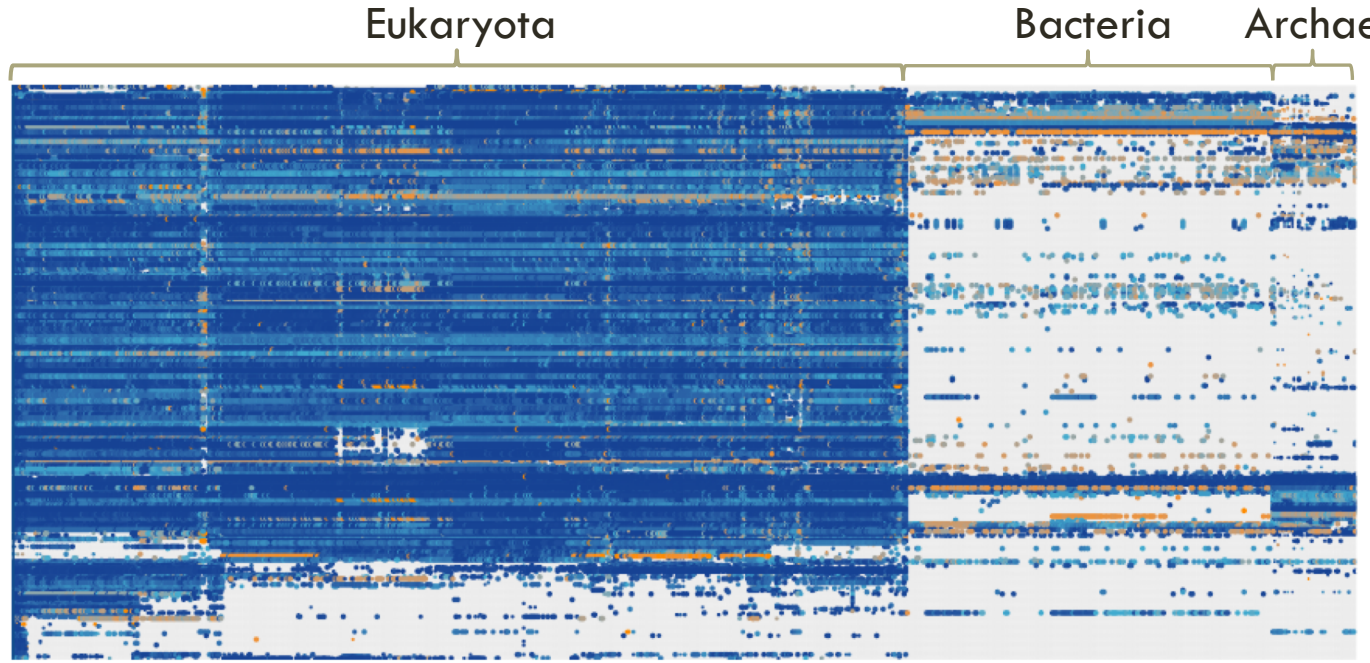
Feature Architecture Similarity Score [0,1]

$$FAS(S,O) = \alpha * MS + \beta * PS$$

Koestler et al. (2010) BMC Bioinf 11:417; https://www.github.com/BIONF/FAS

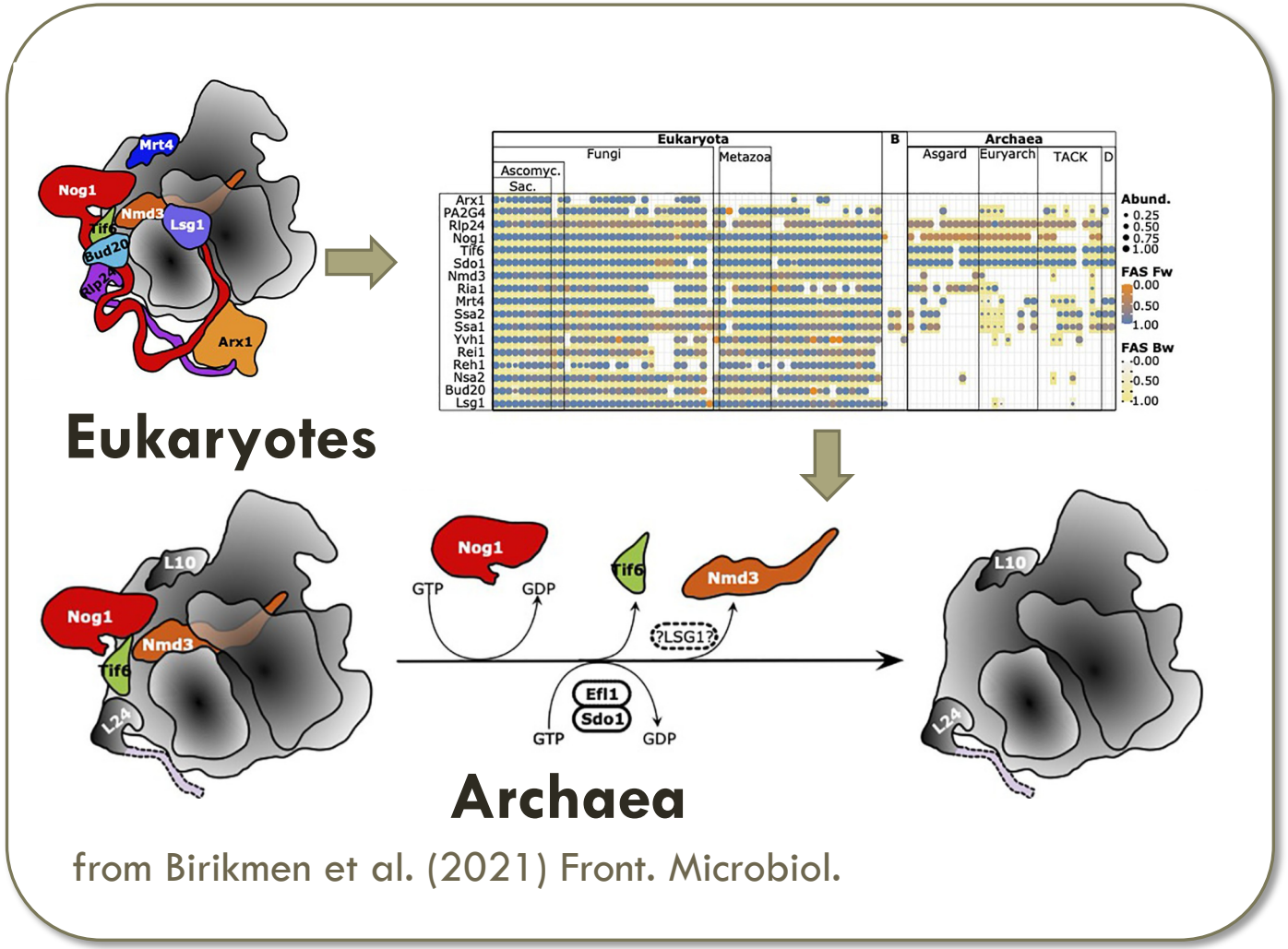
EXAMPLE APPLICATION 1 – THE EVOLUTION OF EUKARYOTIC RIBOSOME BIOGENESIS





Summarized on family level

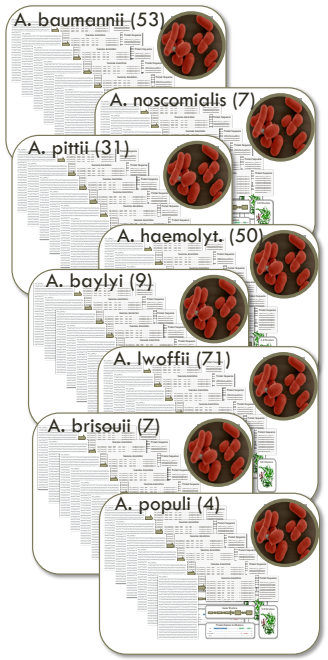
THE INTERPRETATION OF PHYLOGENETIC PROFILES – FUNCTIONAL CONSERVATION VS EVOLUTIONARY CHANGE



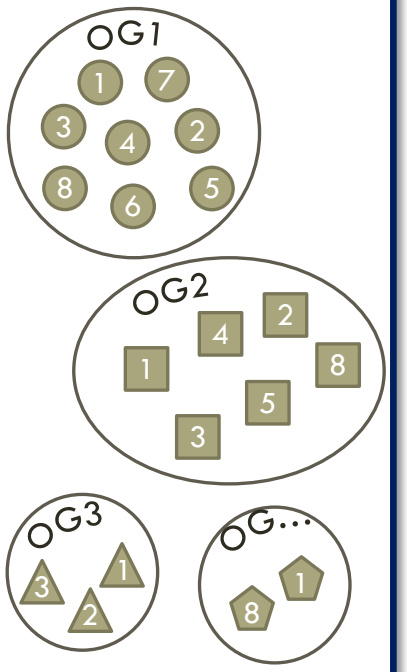
from Birikmen et al. (2021) Front. Microbiol.

EXAMPLE 2 – WHAT CHARACTERIZES PATHOGENIC ACINETOBACTER?

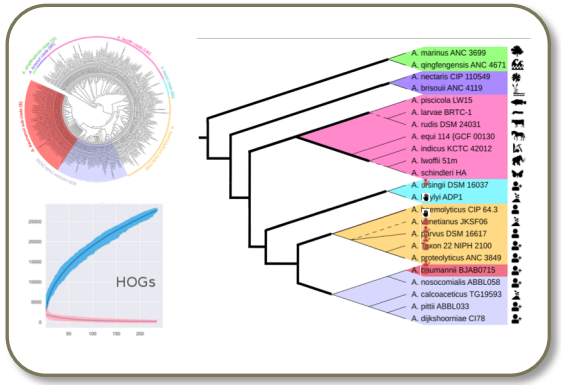
Priming

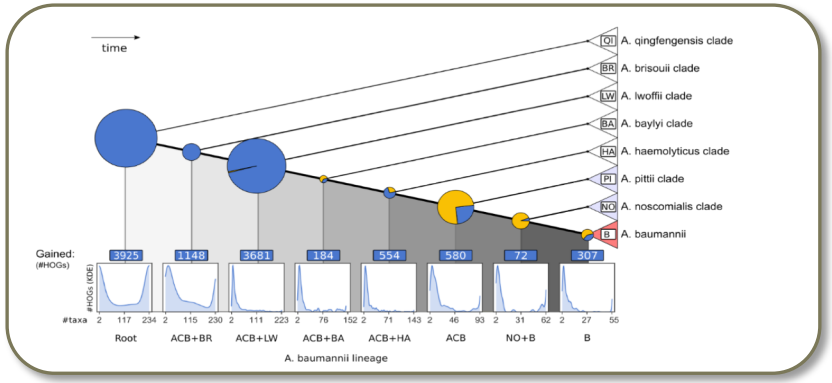


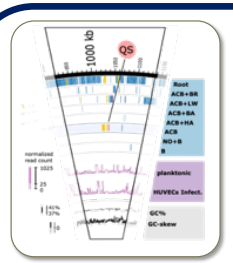
232 genomes
62+x species

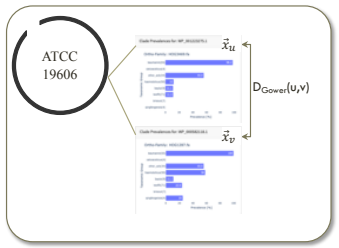


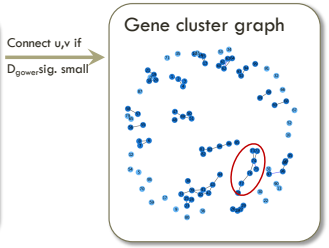
Pan genome:
14,390 HOGs
22,350 OGs

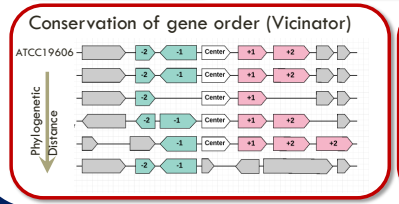


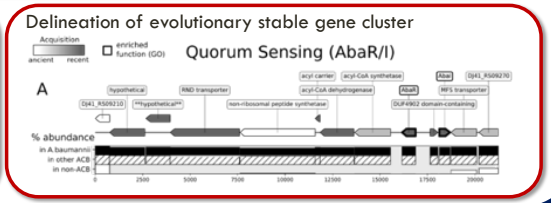










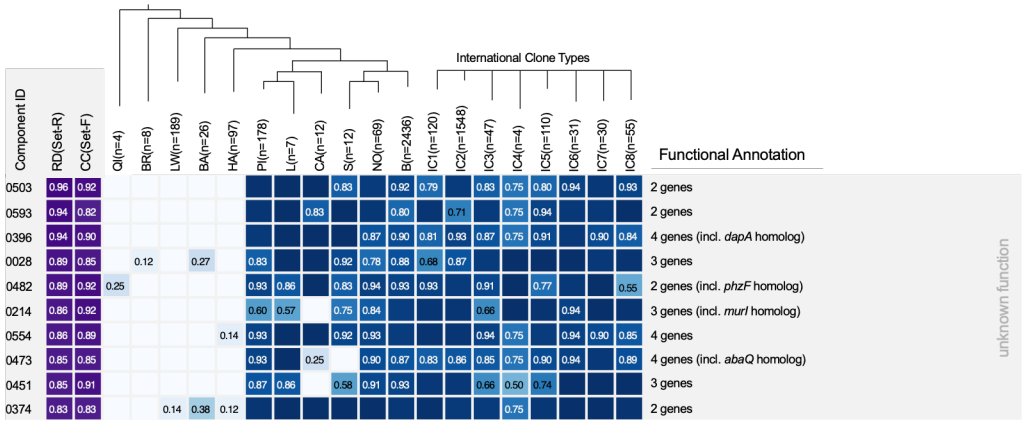


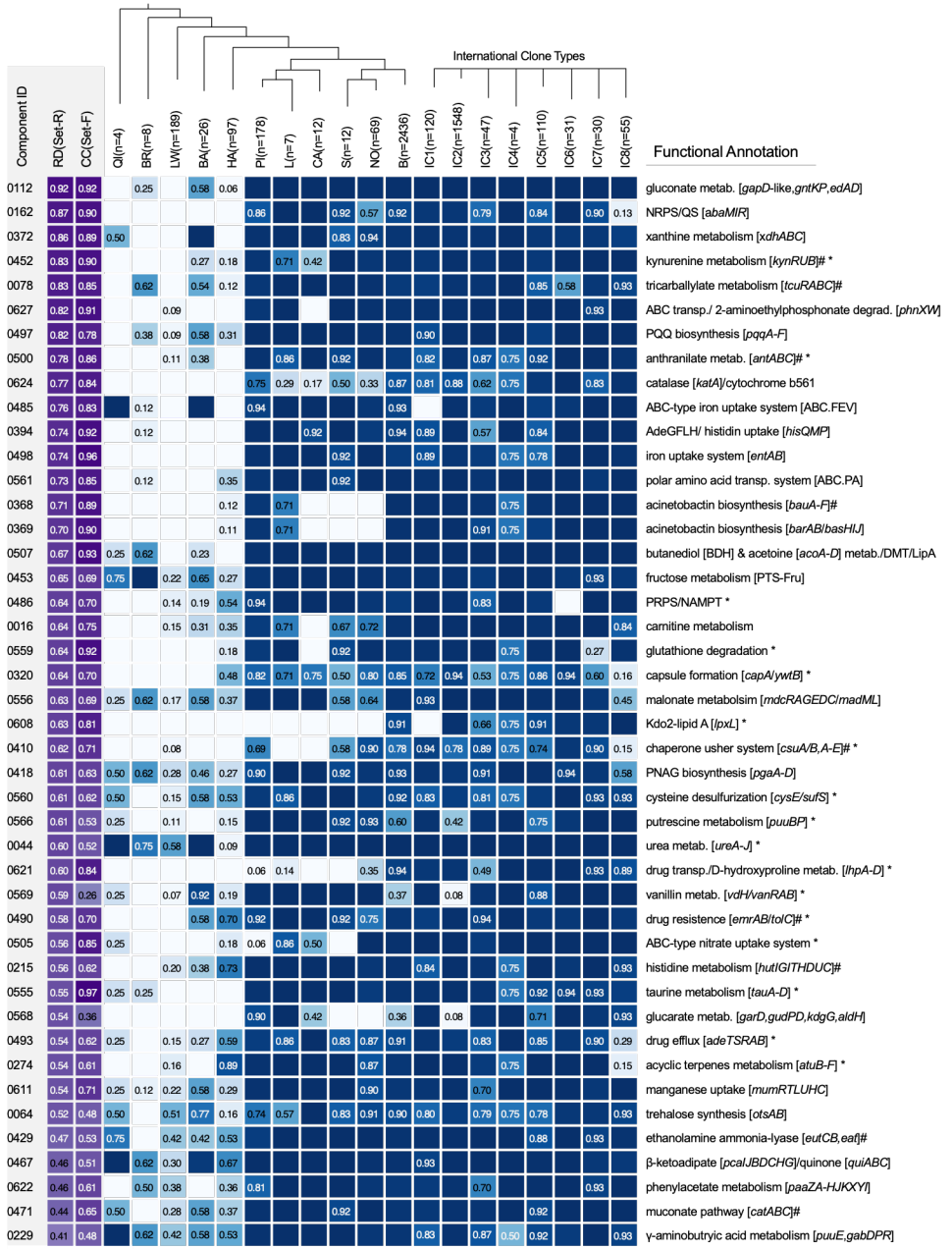
Extension

3,052 genomes

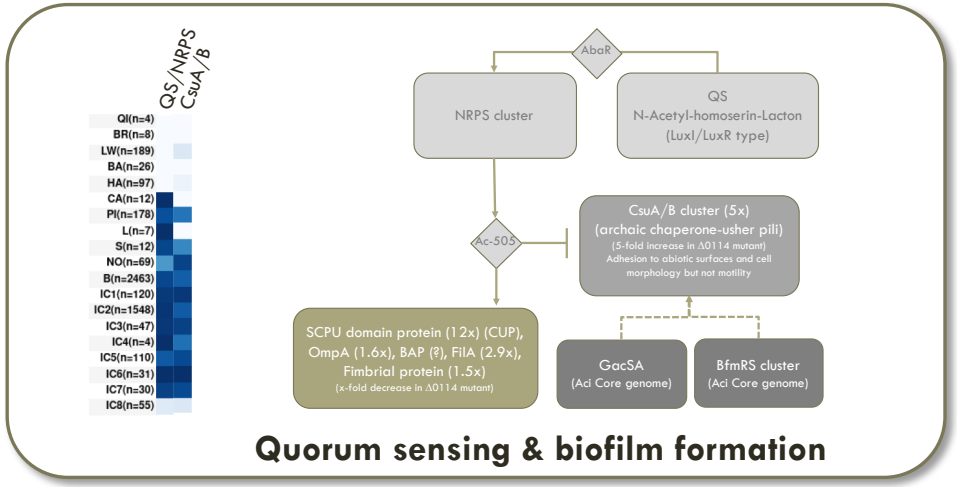
WHAT CHARACTERIZES PATHOGENIC ACINETOBACTER?

Function unknown

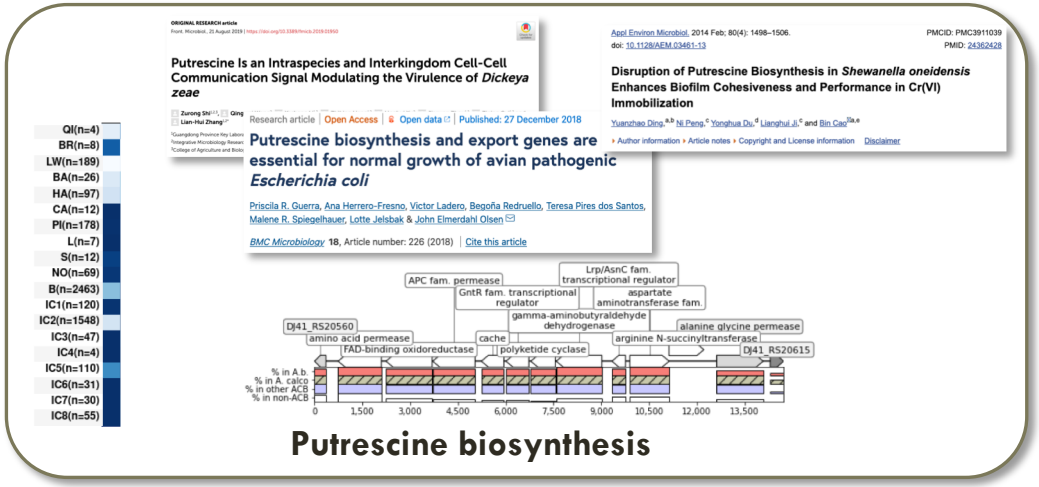




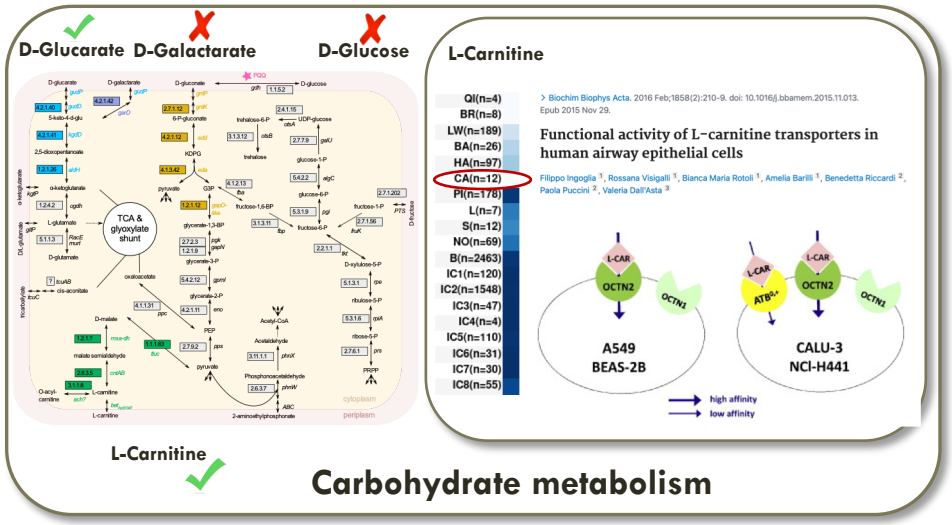
WHAT CHARACTERIZES PATHOGENIC ACINETOBACTER – FEW HIGHLIGHTS



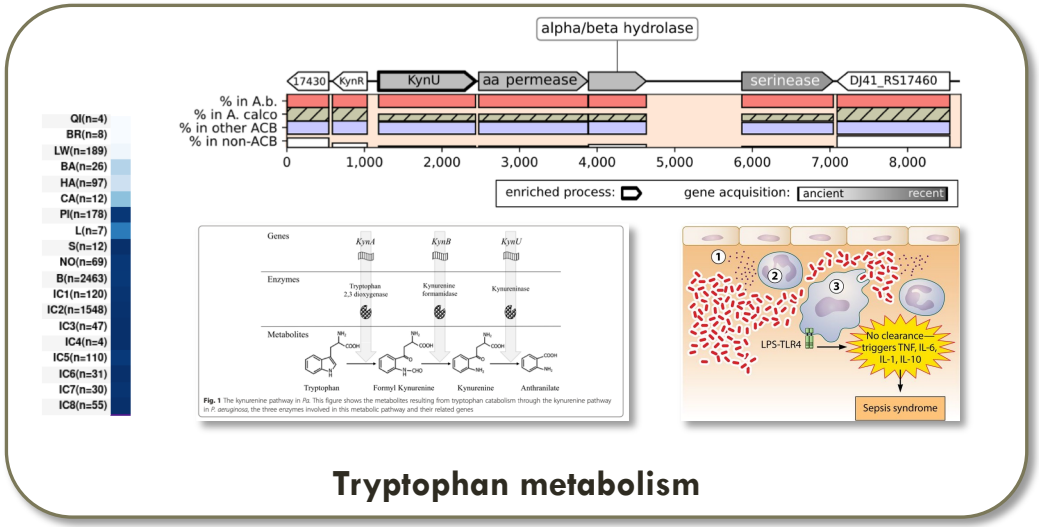
Quorum sensing & biofilm formation



Putrescine biosynthesis

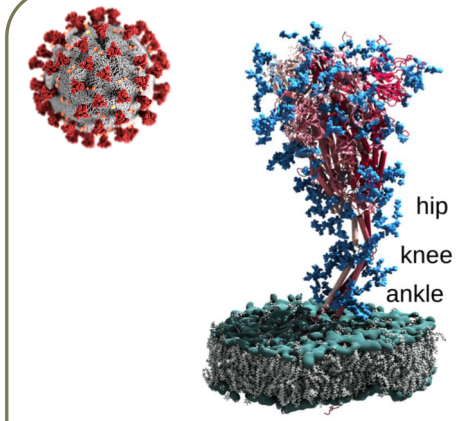


Carbohydrate metabolism

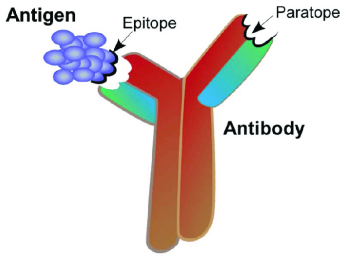


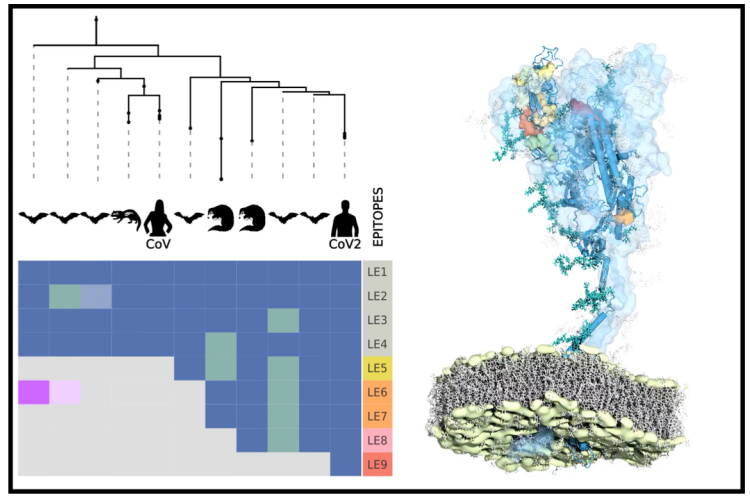
Tryptophan metabolism

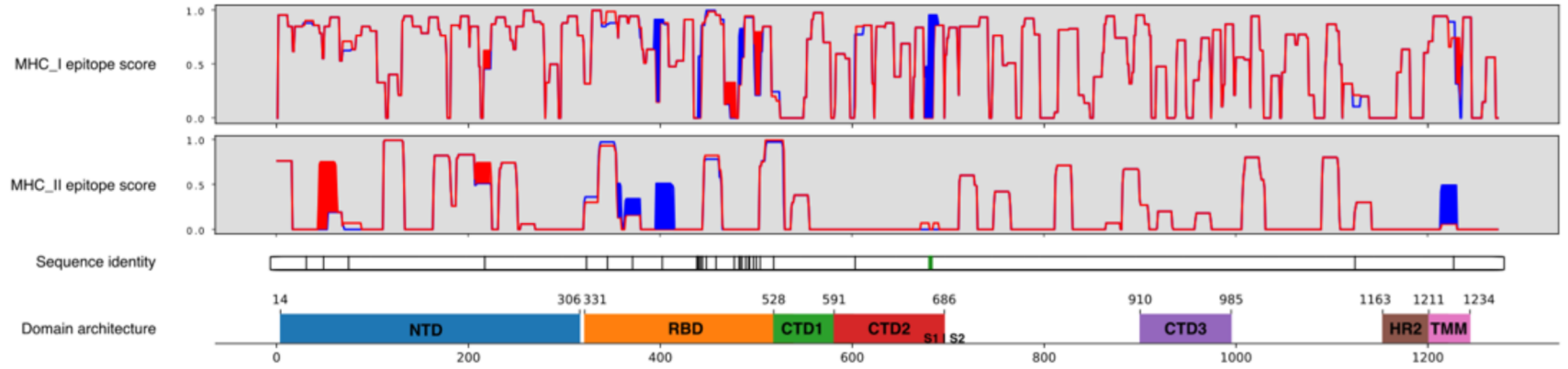
EXAMPLE 3 – WHAT IS SPECIAL ABOUT SARS-COV-2?



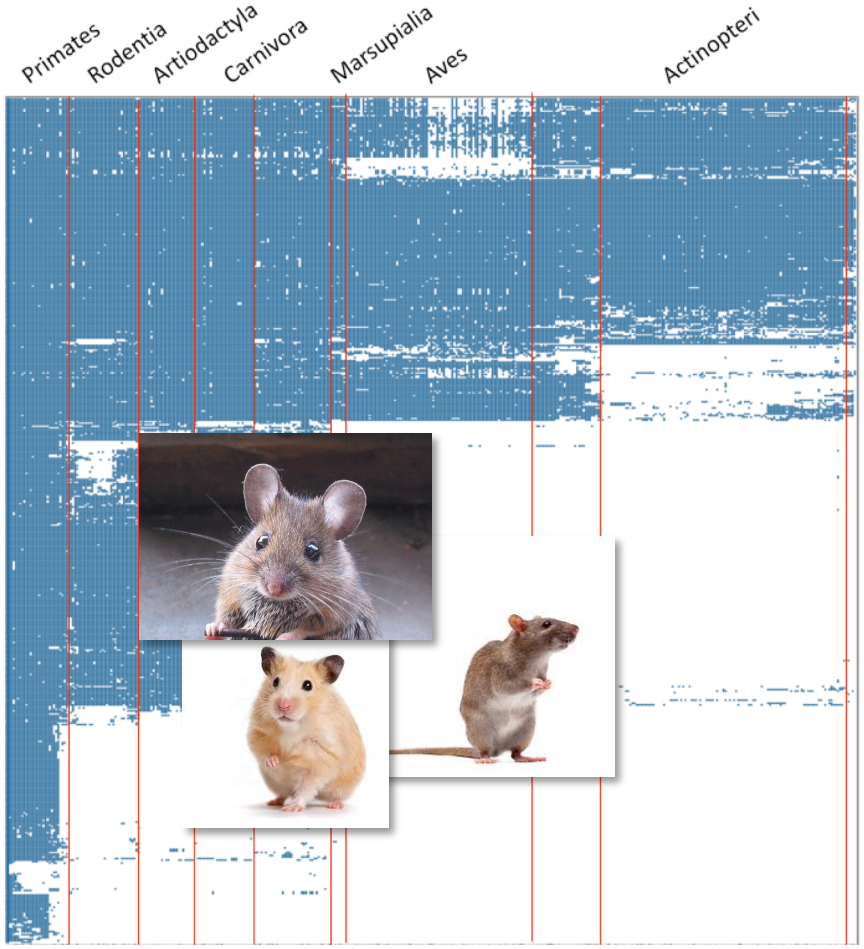
Turonova et al. Science (2020) 370 (6513): 203-208

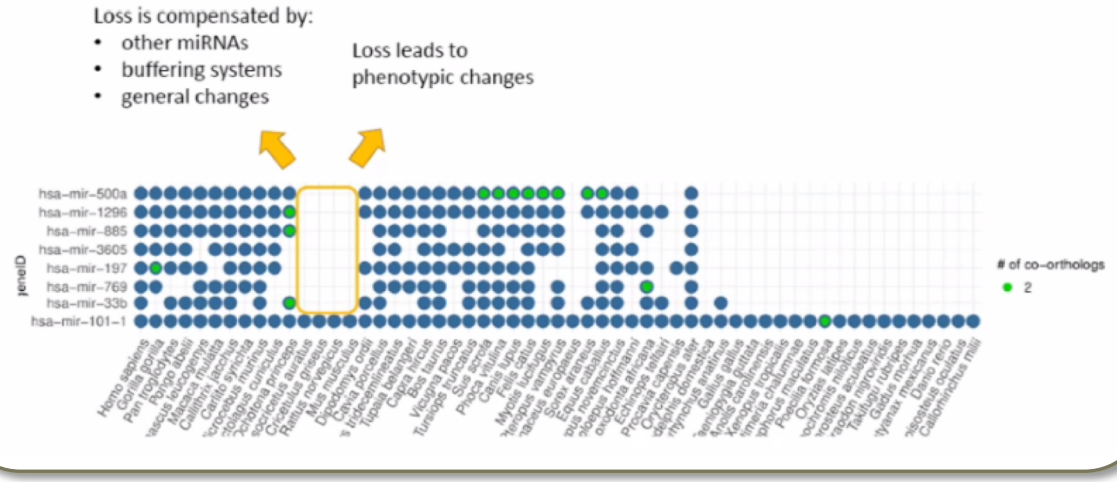


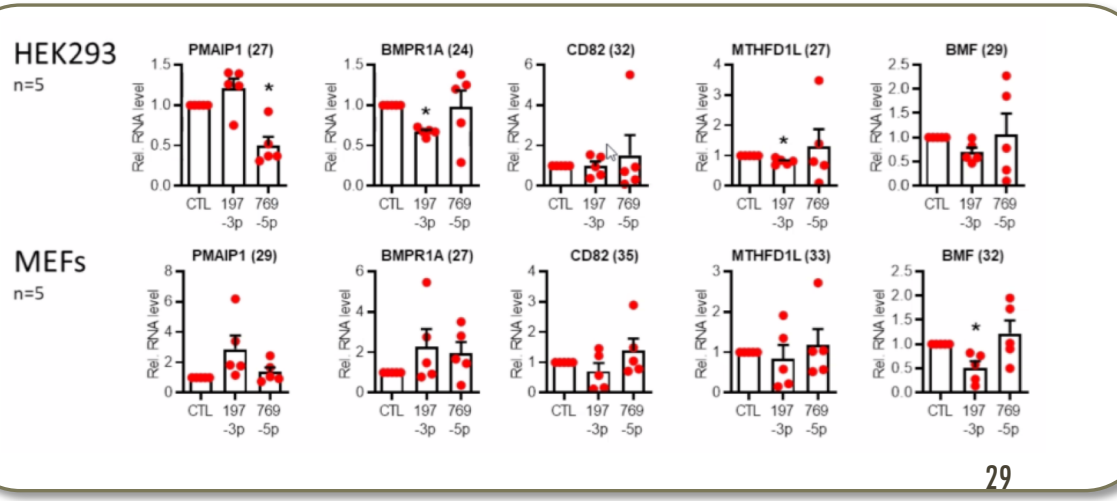




EXAMPLE 4 – EVOLUTION AND FUNCTION OF NON-CODING RNAS





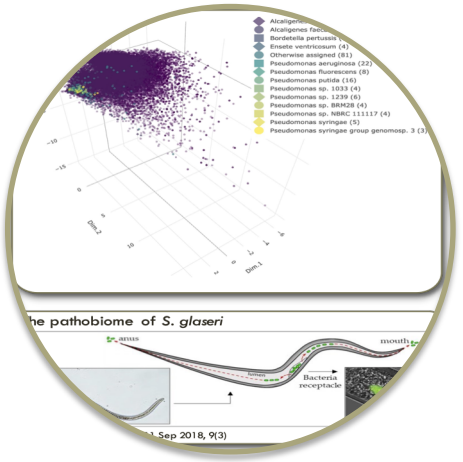


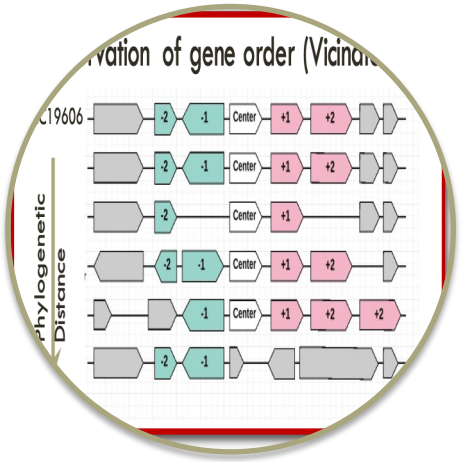
MODUL 16: FUNCTION AND EVOLUTION OF METABOLIC PATHWAYS

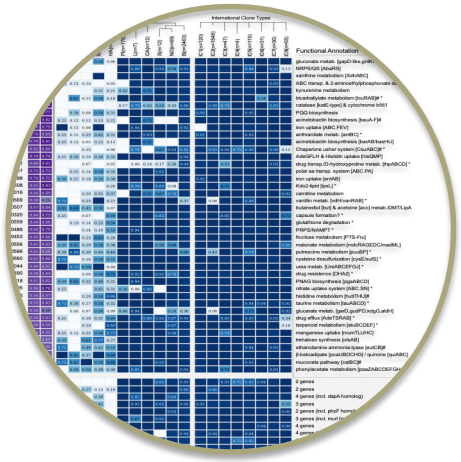
Course contents:

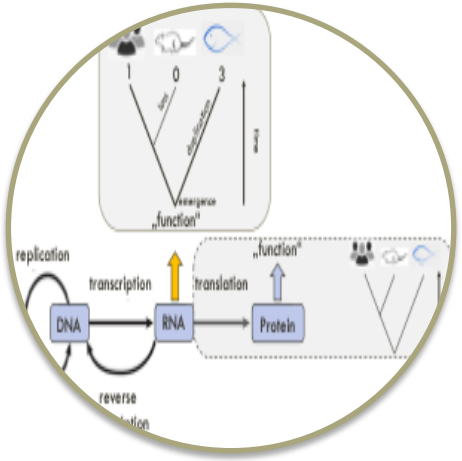
- Theory and praxis in the **sequence analysis** in a functional and evolutionary context
- Functional characterisation and reconstruction of the **evolution of a metabolic pathway** or a protein complex **-> link to other groups is appreciated**
- Functional annotation of protein sequences and **functional annotation transfer**
- Data mining and reconstruction of **phylogenetic trees**
- Presentation of a paper from the area of applied bioinformatics
- Summary of the research results at the end of the module in a protocol/poster and a presentation in the group seminar.

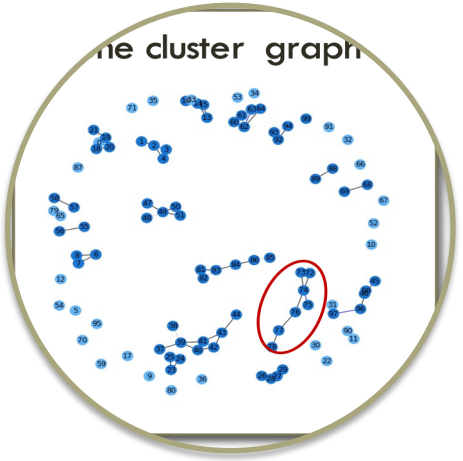
WHERE YOU CAN CONTRIBUTE

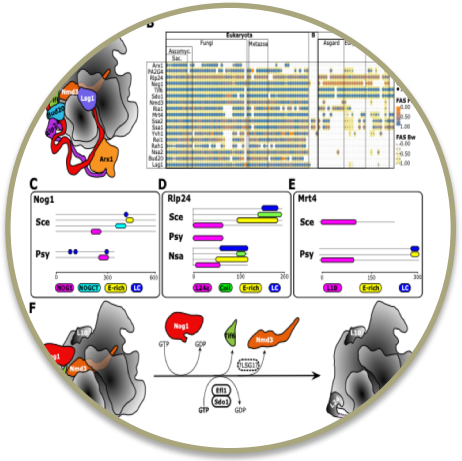


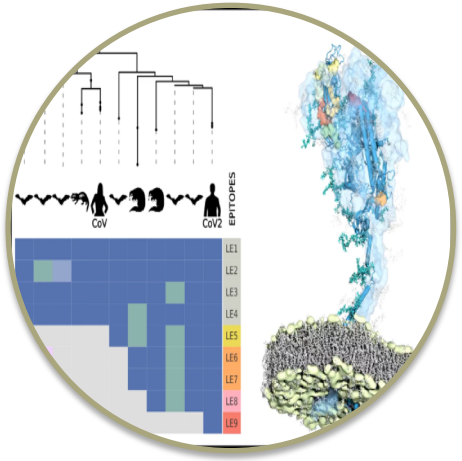




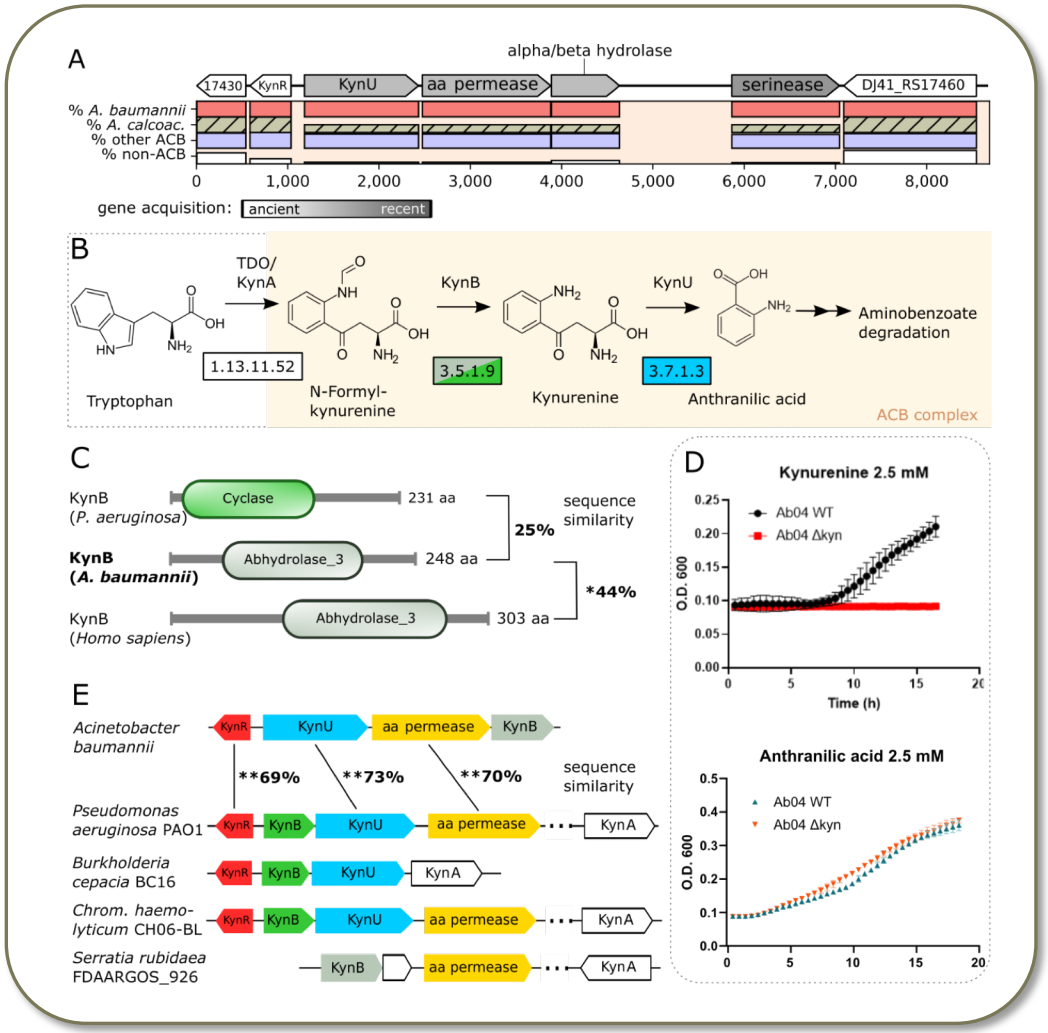




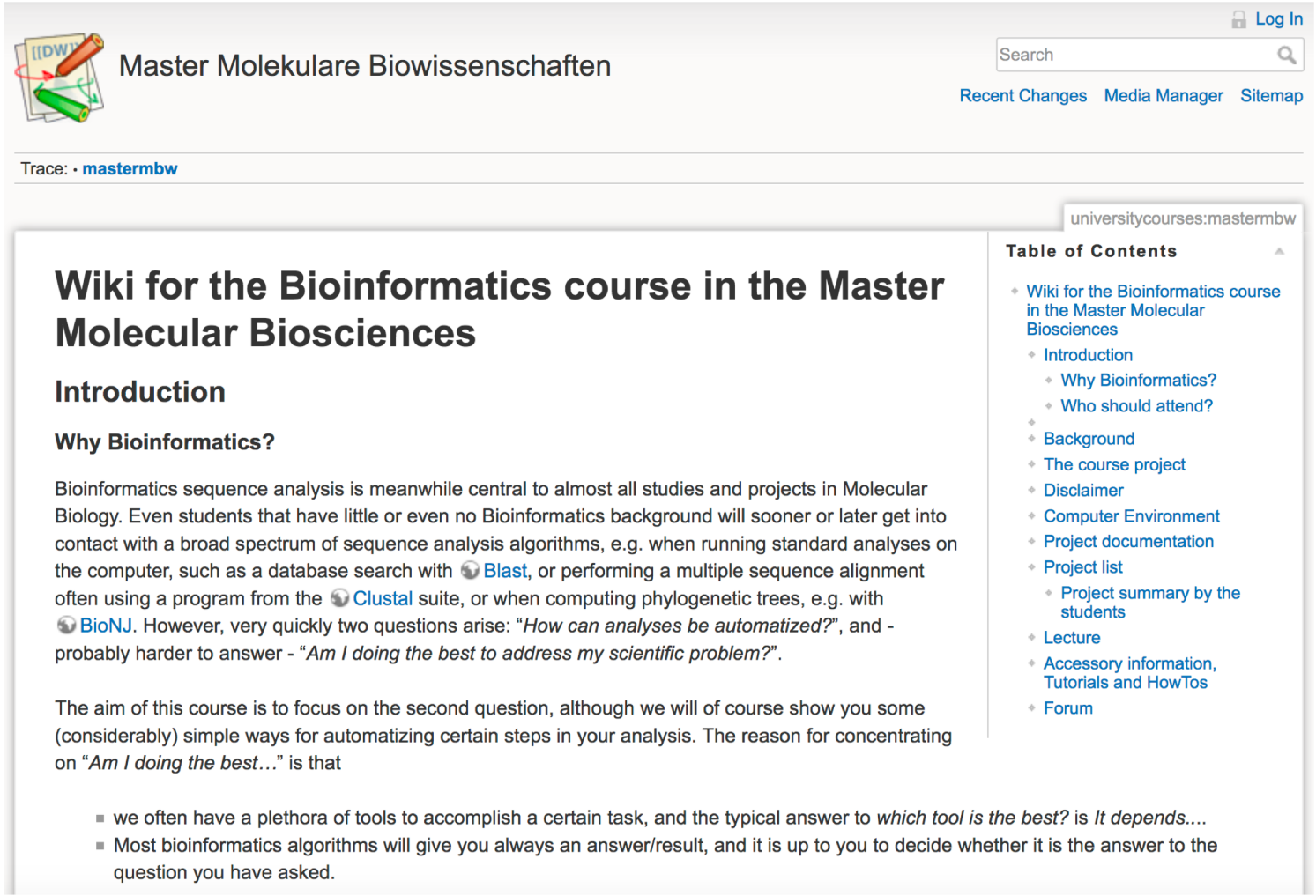




THE KYNURENINE PATHWAY – A WAY TO MODULATE HOST IMMUNE RESPONSE?



ELECTRONIC DOCUMENTATION USING A DOKUWIKI



Log In

Master Molekulare Biowissenschaften

Search

Recent Changes Media Manager Sitemap

Trace: • mastermbw

universitycourses:mastermbw

Wiki for the Bioinformatics course in the Master Molecular Biosciences

Introduction

Why Bioinformatics?

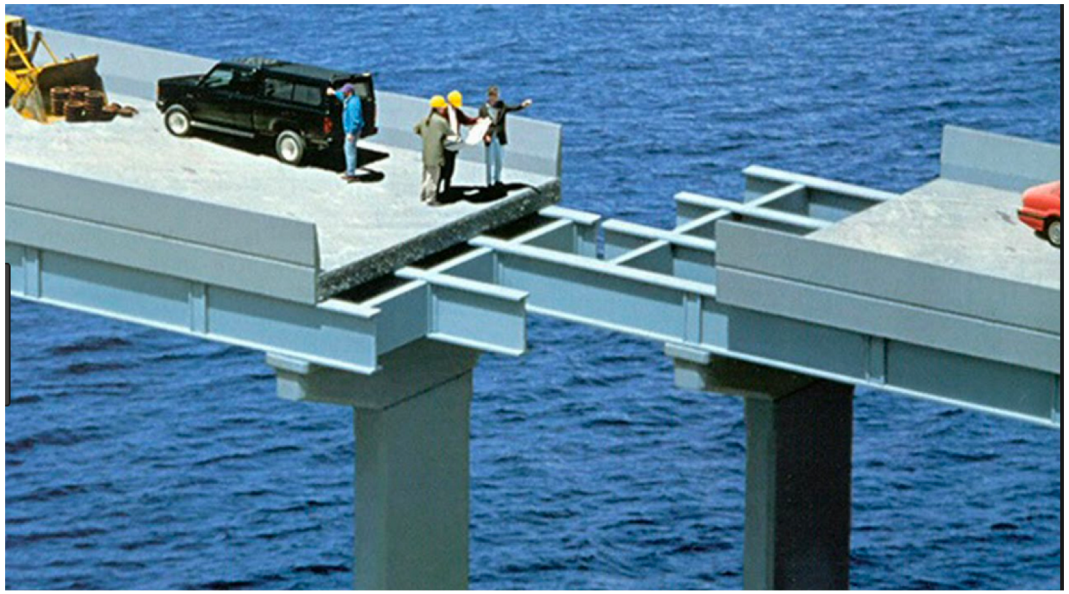
Bioinformatics sequence analysis is meanwhile central to almost all studies and projects in Molecular Biology. Even students that have little or even no Bioinformatics background will sooner or later get into contact with a broad spectrum of sequence analysis algorithms, e.g. when running standard analyses on the computer, such as a database search with Blast, or performing a multiple sequence alignment often using a program from the Clustal suite, or when computing phylogenetic trees, e.g. with BioNJ. However, very quickly two questions arise: "*How can analyses be automatized?*", and - probably harder to answer - "*Am I doing the best to address my scientific problem?*".

The aim of this course is to focus on the second question, although we will of course show you some (considerably) simple ways for automatizing certain steps in your analysis. The reason for concentrating on "*Am I doing the best…*" is that

- we often have a plethora of tools to accomplish a certain task, and the typical answer to *which tool is the best?* is *It depends....*
- Most bioinformatics algorithms will give you always an answer/result, and it is up to you to decide whether it is the answer to the question you have asked.

Table of Contents

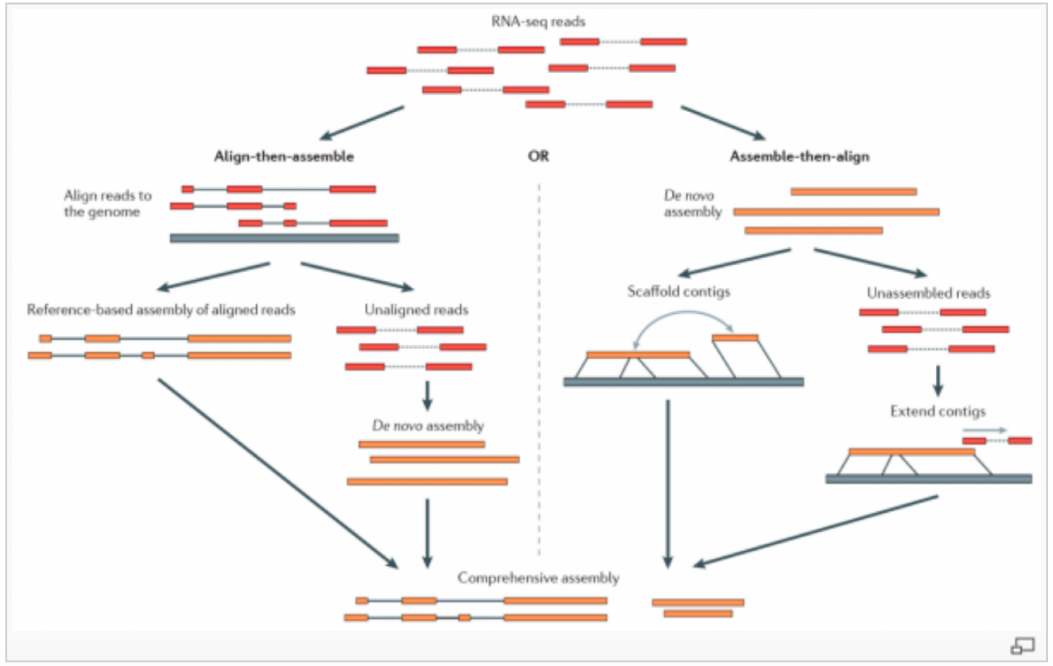
THANK YOU...



WORK PACKAGE – AN EXAMPLE

De novo sequence assembly

If no reference genome is available, sequence similarity between reads is the only information that helps in identifying reads stemming from the same transcript (Fig. 2). The rationale is simple, reads that cover overlapping regions of the same transcript must have a local sequence similarity that is higher than it is expected by chance. In a perfect world, reads covering the same region should be identical. However, errors during base calling and amplification artifacts during template amplification (jointly referred to as *sequencing errors*), together with genetic diversity between alleles of the same gene - of course this requires at least diploid organisms - can result in differences between the reads, despite that they cover overlapping regions of the same gene or transcript. At the same time, reads representing different, but closely related genes (paralogs), can display a high local sequence similarity despite stemming from different transcripts. Obviously, a trade off is required that aims at maximizing length and extent of local sequence similarity to consider two sequence reads overlapping, while still maintaining a sufficiently high sensitivity given sequencing error and genetic diversity.



WORK PACKAGE – AN EXAMPLE

Project outline

In this part of our course project you will use your pre-processed algal RNA seq reads and perform a Trinity assembly (Haas et al. 2013). Once this is completed, the interesting part of the analysis starts. We will then have to reconcile the results from the assembly with our prior expectation. In essence, we will have to answer the question

Do our results somehow make sense, and what are the additional assumptions we have to make?

This all sounds rather cryptic at the moment, but hopefully it will be come clearer as the analysis proceeds. At the end of this project, we will have our assembled transcriptome, we can distinguish between different splice variants for a gene, and we have predicted ORFs in the individual transcripts. If all runs well, we will end up with individual files that can be uploaded into a project database that will help us in managing data and associated information.

Before you start with this project, make sure that you have the following information and data at hand. If something is missing, make sure to discuss this with the other course members and/or with your tutors.

1. You are familiar with the concepts and methodology of RNA sequencing, and sequence read pre-processing
2. You understand nature and relevance of paired end reads during sequence assembly
3. You can explain the biological concepts and the functional relevance of alternative splicing
4. You are familiar with the standard summary statistics used in the context of sequence assembly, such as N50, read coverage, insert size distribution, and the like
5. You have access to your pre-processed sequence reads from the previous project
6. You have access to Trinity and you approximately comprehend the parameter that you have to hand over Trinity, together with your sequence reads
7. You have access to the Transdecoder software that we will use for ORF prediction.

WORK PACKAGE – AN EXAMPLE

Tasks and Questions

1. Perform the Trinity assembly. Depending on the number of sequence reads, this task can take up to six hours. **HINT**: If it is the first time you run this software, and you basically have no idea what you can expect, and if the program runs smoothly at all with your data, it might be a good idea to try an example run with just a subset of your sequence reads. A good start would be something like 1 Million read pairs. If this test run is successful, you can start the full assembly. While this is running, you can make yourself familiar with the Trinity output. Otherwise, you can start with the troubleshooting a bit earlier. Once you've made sure that the test run works, you can submit the command to the cluster using a syntax similar to the following file. Make sure to set the parameters correctly and to change the file extension to .sh.
   a. What kind of output files are generated, and what information do they store? (test assembly is sufficient). The following link to an explanation of the Trinity output might help.
   b. What information does the contig header give you? (test assembly is sufficient)
   c. How many contigs do you get? (full assembly only)
   d. How many genes are represented, and how many splice variants[1)] ? (full assembly only). Concerning the number of genes, was this what you expected? Explain!
2. Run a rnaQUAST analysis on your assembly and document the results. If you are encountering error messages everywhere or you don't know where to start, feel free to check the step by step guide.
   a. In the first round, just take a look at the summary statistics. Document and discuss.
   b. In the second round, add the Busco analysis, using the PlantSet.
3. Run Transdecoder on the Trinity assembly to predict the ORFs in your transcripts.
   a. How many ORFs do you obtain? Discuss the findings. Is this what you expected?
4. Prepare the data files for the upload into the database (Project 2)

Once you have successfully completed the tasks above, you are ready for the next step in the analysis workflow: To find an efficient way for storing and managing your data, such that all information stays consistent.