QFO-7 Sitges Spain

Conference Notes

Natasha Glover - Bringing orthology to the public in the light of evolution (9:10).

Gap in communication between scientists and public

- misconceptions
- distrust
- sensationalism by journalists
- lack in skills by scientists
- stigma from scientists for writing “popular science”

Example: “we share 50% of our dna with bananas, making us half cannibals”.Came from a misquote of that 7000 proteins are 40% identical. However, only 25% of proteins have orthologs as measured by Natasha.
Funding for science communication -> Agora project funding
Outreach website : https://lightofevolution.org/ > several stories to explain research on evolution
Practical exercises to outline idea of sequence comparisons
Marbles comparative genomics kit
Meme generation webserver based on real data (from OMAdb): https://ohmygenes.org/
Comment by Paul: This all addresses people that take evolution for granted, how to approach people that do not

Yannis Nevers: Multifacet quality assessment of gene repertoire annotation with OMArk (9:35)

- coding gene sets referred to as proteomes
- OMArk - based on hierarchical orthologous groups
- Gene to HOG association via kmer-based approach
- clade assignment first -> dynamic LCA set assessment: id of missing, duplicated, single copy
- Assignment of consistent, contaminating, inconsistent and unkown genes => assignment of full set
- P. megacephalum as an example for a gene set with issues
- Simulation based benchmark:
    - Completeness is overestimated when there is high amount of duplication
- Identified contamination in many reference proteomes, 115 cases from bacteria to fungi. Some are however HGT but one has to check the assembly. Paul thinks this could be semi-automated.
- use cases> proteome selection> which one to use,
- error propagation of gene annotation in b10k bird assemblies
- how to differentiate between HGT and contamination. Issue about the use on fungi or bacteria

Diego Fuentes: Phylome DB v5: an exanding repository for genome-wide catalogues of annotated gene phylogenies (10:00)

23 M proteins (2.3 x increase)
540 public phylomes
8M gene trees
Phylomizer phylogenetic tree reconstruction pipeline
Seed based phylome reconstruction
Blast> evalue / 1e-5; 50% overlap, 150 hits
Muscle, Mafft, KALIGN - forward and reverse -> M-coffee; TrimAL
ML Tree - IQTree; 1000 rapid bootstraps
Gene tree - species tree reconciliation - how to root the gene tree?
Species overlap to infer duplication events - not the species tree reconciliation (might ameliorate gene tree reconstruction artefacts?
consistency score to improve performance on tree distance challenge in benchmark

- Orthology Consistency Score = number of trees confirming a given relationship/ the number of trees used to infer the relationship between a particular pair. Cutoff for orthology predictions >= 0.5 CS

Metaphors 2.5
MetaDatabase

Classic Challenges

- exponential data growth -> scalability issue
- Computational bottlenecks (data integrity vs performance)
- **how to deal with strains??**
    - collapse it
    - metaproteome
- crosslinking of proteomes is an issue
- **Reference proteome selection issues** (QFO ref proteomes?)

Dannie Durand: Modeling the Evolution of Multidomain Architectures (11:25)

Multidomain families are those which have at least one domain in common, but the rest can be variable. The variability can come from domain/gene duplication losses, insertions, deletions, HGT?
Event catalogue of domain architecture evolution
constraints of domain order and
Domain adjacency are constrained and not random
From species tree to the evolution of genes and of their sequences -> extend to domain gain&loss
Rule - how to maintain constraints on domain order
Simulation tool: “DomArchov” (http://www.cs.cmu.edu/~durand/DomArchov/)
Markov model of domain architecture evolution

implicit learning of rate parameters using Metropolis Hastings Algo
Probabilty of a DA, where to get this from?
Comparison of genuine vs. simulated domain architecture in terms of domain length, promiscuity metrics (unique neighbors, weighted bigram frequency), tandem repeats (mean tandem array length). Simulations also mimic genuine in that they gains and losses of domains occur at ends. In general, the simulations they did match what is observed in the real data.

Comparison with natural language models
represenation of DA in higher dimension space
architecture as point in n/dimensional space with n the numbers of domains
overlay of oberved (genuine) vs simulated architectures - nice overlay. No pronounced overlay with randomized data (what does this mean)
Natural language models to capture domain order. word2vec. Context domain directly before and after. Trained with genuine data only
Probability estimates differ with the training data. How to get aournd this

David Moi: Reconstructing protein interactions across time using phylogeny-aware graph neural networks (11:50)

HOGPROF for phylogenetic profiling
Interaction is an evolving trait - rewiring of networks over time
STRING COGs as training data -> interacting COGs can be mapped to the species they're interacting in
graph reconstructing whether two COGs are interacting or not (along a tree)
ML problem / graph neural networks (pytorch geometric)
Graph structure is necessary, feed forward strategy performs poorly (AUC 0.68)
Illustration using Actin interacting genes String vs prediction: Example that ML works in the right direction

Improvments

- NCBI tree is an issue
- string is incomplete/unreliable
- remove species tree constraint?

Round table discussion

First question : what is the purpose of QfO and why do we want to study orthologs? (What taxononic scale)

Ingo: Everything. Lot of différents challenges interests different member of the community

Christophe: originally, qfo was also to bring together expert so they could agree on common ground.

“Orthology” means different things to different people, depending on the use case. We often come from the evolutionary perspective, but others come from a more functional perspective.

opening up novel issues the community should think about / be concerned about

Diego suggests to focus on different strains/accessions/varieties and how they can be incorporated into orthology

- possibility of using pan-genomes
- issue of using sequencing of lab strains with often streamlined genomes. Similar to plant domestication: lose a lot of genes (in comparison with wild types). Importance e.g. climate change, drought resistance
- Ingo: pan genomes could be made by combining assemblies
- scaling issue of data integration on the strain / species diversity level
- orthology on different levels
- enduser should be able to resolve to the desired level
- Christophe advocates the use of alternative spliceforms instead of only relying on longest
- Paul: balancing orthology inference accuracy with practicality. For example using 1 isoform rather than all, orthology on the gene level rather than domain, etc. But what is the biggest problem and what should we address next?
- **What is the biggest problem we are facing** ? 1.handling isoforms (for multidomains studies; although domain orthology and isoforms orthology are not exactly the same problem). Not necessarily for orthology(?). 2. Increasing amount of novel orthologous group (hidden in pangenome, metagenomics) -> how to handle this computationally.
- Dave: already facing bottleneck with sequence comparison and we'll have to deal with structure level comparisons. He advocates using structure for homology detection and not sequence alone
- Erik: should we extend the QFO Reference Proteomes since a lot of resources now can handle much more? Should these be pangenomes?
- Nicola: 3D bioinfo network - annotation community. European network, core resources accross europe, cath, pdbe, -> how does ‘data’ talk to each other. Helping in setup of standardized formats
- Christophe/Paul: sharing of resources and results -> green bioinformatics
- Luis: issue of small proteins up to 100 aa down to oligopeptides
- Will evolution of data quality (e.g. transcript assembly vs IsoSeq full length RNA seq) solve issues automatically (stated by Thomas R.)? **What then are then the remaining challenges beyond data quality**
- Diego: we need to tackle the issues of assembly and annotation before we can make reliable orthology inferences
- Jaime: we need to add genomes from new/underrepresented phyla to improve orthology inference
- How to prioritize species / genomes to be added? In essence, what does ‘underrepresented’ actually stand for -> maybe do it on the gene level not on the species level (add diversity not redundancy). Statement comes from the microbial metagenomics community
- Jaime / structural similarity as a proxy for functional similarity but probably not as a proxy for homology -> issue of convergence

- big data vs. small data discussion. what is the best data amount / quality trade off /. Nicola: when it comes to structure, more sequence is better

String database
curated knowledge, gene neighborhood, operons, coexpression (variation autoencoders), experiments, text mining
Physical string network / functional string network
String clusters - as higher order modules
most of data in string based on orthology - use of EggNOG
integration of novel species - stepwise transfer from taxa with increasing evolutionary distance

- user species -> what about question
- proteome as input
- no longer EggNOG mapper
- fasta header parsing included
- DIAMOND as a search tool (sensitive setting) -> map proteome against 12 Million proteins in String
- projection of annotation (KEGG, Tissue, etc) from LCA of search taxon and best hit

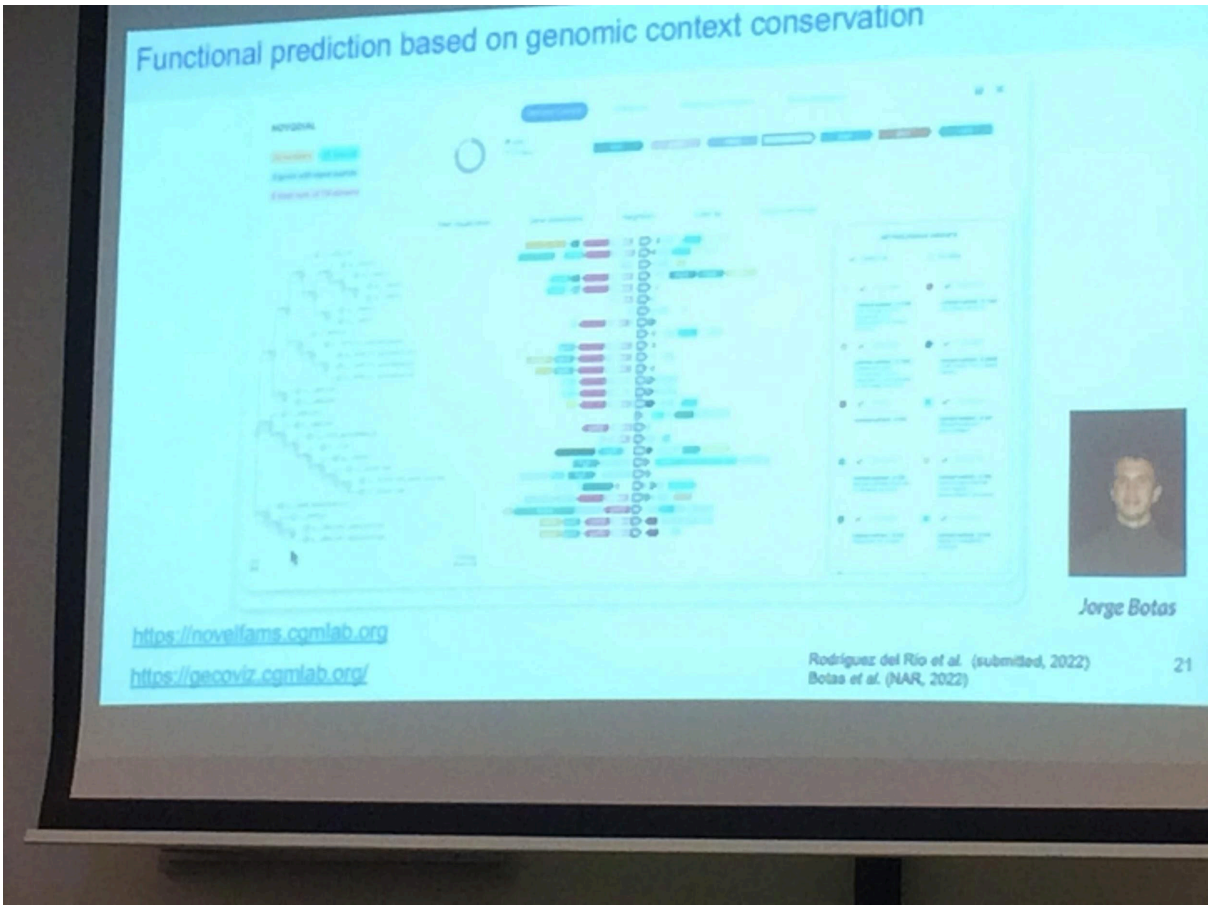
New release is pending

Paul Thomas: A complete draft human functionome as determined by the Gene Ontology Phylogenetic Annotation Project (15:45)

Human Functionome
GO - universe of possible functions for gene products (protein, ncRNA)
BP - Biological Process becomes now more a Biological Program
GO terms are descriptors of functions, each term covers part of the function of a gene.
Every GO statement as some kind of evidence (with evidence code)
GO milestones outlined from start in 1998 first release to 2022
85% of experimental annotations are from studies in 10 model organisms
Phylogeny based GO annotation is based on fully reconciled trees, each node having a persistent identifier to track version change.
manual review of all ~9000 protein families with human genes and EXP annotations
How often functional change along a speciation branch vs along a duplication branch
change of function after duplication more common than after speciation
general pattern: function conservation is more common than change (both after speciation and duplication)
Function evolution correlates with sequence subst. rate
least diverged orthologs tend to conserve the ancestral function
GO term subselection process - describe gene products with a minimum number of terms that still convey the full info
function similarity of paralogs depend on duplication age
Even though human genes have the most annotation derived from experimentation, lots of GO terms come from model species

Jaime Huerta Cepas

functional and evol significance of unknown genes from uncultivated taxa
upscale of sequence comparison methods (MMseq, Diamond, etc)
Working with metagenomics: High proportion of environmental sequences have no detectable homologs in cultivated species: these gene families get “thrown in the trash”
are novel gene families relevant? Genes represented only in few or even a single taxon
workflow for id of unknown protein families with evolutionary/functional significance > segments with 20 or more conserved amino acids in 3 or more taxa w dn/ds <0.5 as evidence for purifying selection; no sig hit to pfam refseq, eggnog
413 K putative novel gene families
doubles the number of orthologous groups currently available in the databases
980 novel OGs are synapomorphic of 16 uncultivated phyla - evidence for stronger purifying selection compared to other OGs
What do they do? Functional annotation done by: synteny (neighboring genes in same kegg pathway)
Again, a problem with big data and scalability. PhyloCloud: an online platform for making sense of phylo genomic data. Also a new version of ETE toolkit



Ikuo Uchiyama

MBGD update
DomClust - domain aware ortholog clustering
DomRefine - to remove spurious domains based on MSA
stepwise protocol for ortholog grouping with pangenomes
2022 release: 15,000 genomes 4747 species comprising bacteria (mainly), Archaea, and unicellular eukaryotes
DomRefine resulted in fewer OGs overall (removed some of the splitting of groups)
Growth of MBGD resolved on the level of #proteins on pan-genome level, OGs shared among n members in genus(?). Shows impact of rare genes
Used a synteny approach to define the core genomes
In MyMGDB (mode for analysis of user genomes): upload user genomes, profile-based orthology assignment, ortholog clustering
GenomeMaple: Annotation based on KEGG module - evaluation of module completion ratio -> compare MCR across genomes
Used Dollo parsimony compressed profiles (DPCP) to avoid taxonomic bias with phylogenetic profiles
Phylogenetic profiles can be used to infer missing genes in KEGG modules
how broad is the vocabulary used for phenotype-genotype association

Salvatore Consentino

SonicParanoid - update
Domain architecture based orthology assignment
Faster graph-based orthology

- fast alignment - prediction of which direction of pairwise protein alignment works faster
- BBA running on a subset of sequences - avoiding unnecessary alignments
- comparison to orthoFInder and Broccoli in speed
- EC challenge not affected by speedup
- 2000 MAGs as test set -> Broccoli could not compute
- The closer related the genomes, the less time save

Domain-based orthology

- NLP based approach
- uses doc2vec
- cosine similarity between architecture pairs
- integration of graph- and domain architecture-based orthology assignment

Victor Rossier

OMAmer https://academic.oup.com/bioinformatics/article/37/18/2866/6206361
achieving scalability with placement approaches
linear scaling
phylogenetic placement using Max Lik - sensitive and specific but not precise
alternative - find most similar sequence

Issue - closest sequence not part of the sub-family
OMAmer - tree driven and alignment free
conserved k-mers mapped to ancestral nodes
more precise that closest sequence approach
OMAmer works better with recently diverged species in densely sampled clades

case study
gene content evolution with bird phenotype
correlation of gene family expansions and contractions with convergent phenotypes
Diving as the phenotype

- expansion of hemoglobin family correlates with diving phenotype
- loss of flight: contraction of gene families -> Bone Morphogenetic signalling

Benefits of recent and densely sampled clades

needs a sufficient representation of the taxonomic group for which novel members should be added. OMA originally had 4 birds, that was not enough. Increasing to 8 seemed to have resolved the issue

Sina Majidian

BIOQA - human-machine interaction
Natural language -> SPARQL query
pipeline: survey literature -> searching for relevant papers -> export -> score relevance -> database assignment -> question design
DB contains questions and answers

- application, evaluation, improvement of QA systems
    - in context of orthology data
    - increase ortho-db use
    - How are people using ortho dbs
- what are the questions

Challenges

- finding the relevant papers
- summarising the right questions
- finding diverse range of questions
- automatizing the process
- conversion of question to database query
- Interface of dataset Q/A

Relevant to survey the literature to create a corpus of information about how and why people use orthologs
Inferring the question from the statement in the paper

Maria Martin
annotating canonical proteins in proteomes
Intro into uniprot
External sources -> uniparc sequence archiver (unique sequences only; 100/100 rule)
-> UniProtKB (SwissProt (manually curated)/TrEMBL (automatic)) -> proteomes; uniRef

from proteome to reference proteome
e! / refseq /wormbase -> manual selection
INSDC - all proteomes
=> 451,601 ‘raw proteomes’

- manual selection
- automatic filter
    - quality criteria
    - redundant proteomes
    - RPG grouping and selection

20,800 reference proteomes as of now

Filters:
RefSeq provided information but partly self-computed (e.g. BUSCO)

- abnormal gene to sequence ratio
- genome too large or too small
- low quality sequence
- many frameshifted proteins
- chimeric
- contaminated
-
- hybrid (from NCBI)
- missassembled
- mixed culture
- sequence duplications
- unverified source organism

Redundancy removal module

- novel versions of a genome/proteome do not automatically enter the Reference Proteomes

Criteria for choosing representative within a group (genus level)
in 55% uniref clusters

- manual selection
- QfO reference selection
- number of publications
- proteome annotation score
- try to keep proteomes with SwissProt entries

Issue of BUSCO result interpretation - a relative clade-specific cutoff rather than an absolute value

Reference proteomes are gene-centric -> one protein per gene
isoforms in swiss prot with partly identical id (xxx_2; xxx_3, etc), in trembl not
trembl -> take the longest isoform
for sp -> take the canonical isoform

Stats files always an issue for the community: What has happened to my refProteom. Stats files gives basically all the information about changes. Additional statistics like BUSCO would be helpful to interpret the effect of the change

3 major sources of changes: new assembly, change of source (e.g. Ensmbl/Refseq) or reannotation

issue of tracking which genes have been modified / added / removed. We have the info in principle about the phylogenetic profile of each protein but the protein id is hard to track across different refproteome versions

Slides would be great to upload on the WIKI here:
https://questfororthologs.org/intranet/slides_qfo7

Round table discussion 2
Working groups: most active -> Reference Proteomes & Benchmark. People can and should sign up. Setting up a novel working group is possible. Just initiate it and organize

Meeting Report - QfO 6.5 and 7 together ->
Ingo, Luis (Metagenomics), Paul, Felix, Natasha, Nicola (Structure), Christian (Viruses), Sina, Yannis, Erik, Salvo -> including other papers in the field / review ok, to include things that were not discussed,

Benchmark paper: 2 year cycle for NAR DB issue. Planned for 2024 -> 2022 reference proteomes -> feature architecture similarities as a novel challenge (Ingo)

- Benchmarking starting now - with the current version of 2022 (August 2022). We need to make sure that zebrafish is fixed since it was badly broken, and otherwise make a fixed release.
- what gives the paper more impact?
- inclusion of paralog assignments -> Request by Claire Hu
- suggestion of adding a bit more biology -> suggestion by Nicola Bordin
    - fly examples possibly provided by Claire
- **Erik will take the lead of this group.** Other people to include: Jaime Huerta Cepas, Odile Lecompte, Paul Thomas, Adrian Altenhoff, Yannis Nevers

QfO 8 - 2024: Dannie Durand: Either Austin-Texas; or Edinburgh -> local organizer / partner necessary. Funding: Could be SMBE Satellite Meeting; co-localization with other orthology; Maria suggested Hinxton.
Dave can provide contact to people in Montevideo/Uruguay who might be interested in hosting.

Tom Richards suggested to get a Marie Curie training grant to fund meetings.

QFO 9 - 2026 probably in Straßburg (Odile Lecompte)

Nicola Bordin

CATH workflow weekly processing of new chains
cut a PDB chain and then assign domains to superfamilies - expert curation
impact of Alphafold2 -> latest release covers all uniprot

CATH workflow not scalable to deal with AlphaFold data (even if they only take close homology, there's still 32 million sequences to deal with)
structural similarity search as an issue
quality assessment of alphafold models: half are of poor quality
only about half of domains can be identified by hmms
Can they use embeddings from Protein Language Models (pLMs) to find very remote homologs? -> function prediction & structure prediction
quality of model depends on the availability of training data (cath - > .. -> phmm ->)
PROTBert-Tucker
CATHe: were able to assign 8% more of low %id sequences in CATH superfamilies
Foldseek ultrafast protein structure search retaining precision
Benchmarking the threshold for foldseek. Empirical determination of a cutoff score beyond which a group assignment is meaningful

CATH does not distinguish between ortho- and paralogs
classifies functional & structurally similar groups

FunFams -> functional families: cluster sequences at 90% similarity, extract clusters with at least 1 GO, encode pattern in HMM…

FunFHMMer - conserved positions - important for folding, variable positions (conserved within clade) - important for function?

Inference of functionally important positions using alphafold2

8% of unassigned sequences (new structures in AlphaFold)

identification of 26 novel folds in the Alphafold2 structures

Claire Hu

DIOPT
DRSC/TRiP Functional Genomics
shift from RNAi to CRISPR
D. melanogaster, Mosquito, tick recently
75% of fly genes can be mapped to human (not necessarily orthologs), but interested in functional equivalents
Different wet-lab people have different preferences on which orthologs they want, which tools used to get them, if they want the union or the intersection, etc.

Small and dedicated tools to address individual/specific questions - no metaplatform
DIOPT integrates many different algorithms on ortholog assignments. 'likes them all'. No weighting here
Can filter DIOPT output based on score and ranking
Output: Protein alignment, domains for each pair, conservation among species (heatmap)

User can provide new ortholog pairs

functional orthologs - is that a reasonable term? Based on complementation in experimental studies (replacement of ortholog rescues the phenotype)

Timeline of tool integration - is there a tool saturation at one point?
Gene2Function: query the gene of interest, output is integrated information about the orthologs, including publications, direction evidence-based GO, interactions, phenotype, expression, 3D structure, researchers, human disease annotation…

BioLitMine: summary of PI - gene relationship -> enhance collaboration between groups

Public resources for interaction data -> combine them into MIST Molecular Interaction Search Tool
Interologs: mapped PPI interaction from one species to another

iProteinDB: integrated protein database of post-translational modifications. Aligns sequences and highlight phosphorylation sites among different species

DRscDB (DRSC scRNA-seq Database)
overlap of celltype specific markers - high overlap for certain tissues between human and drosophila

Alliance of genome resources is DIOPT in 'disguise'

- integration of tree based (treefam), graph-based (inparanoid), hybrid (panther), and manually curated (ZFIN, HGNC, SGD, ECGA) ortholog assignments

Entrez gene id as the main ID

Felix Langschied

orthology of non-coding RNAs (ncOrtho)
How to detect? Challenging because pre-miRNA are short (~60nt), can be located in repeat regions, and miRNA annotation is incomplete for non-model species
Targeted search to find miRNAs: select a reference species and train Covariance Models with high-confidence orthologs, then use the models to find orthologs. Can scale computatiojnaliy.
First shrink the search space by using microsynteny then do a sequence similarity search in only those regions. Identify orthologs using BBH. Only these high confidence orhtologs are used for training the models
After getting models, can search for orthologs in full taxon set
Benchmarked with MirGeneDB (experimentally validated and manually curated). 93% accuracy and 97% sensitivity in vertebrates
Phylogenetic profiles of hundreds of vertebrates (to see gene gains and losses)
Seed sequence in less conserved in younger miRNA genes
Can confirm WGDs by looking at the copy number of miRNA genes (co-orthologs)
For miRNA-associated regions, less SNPs than in regions associated with protein-coding genes. Because diversity is so low, does that reduce the effects of ILS? Yes, pre-miRNAs resolve classic clades affet by ILS.

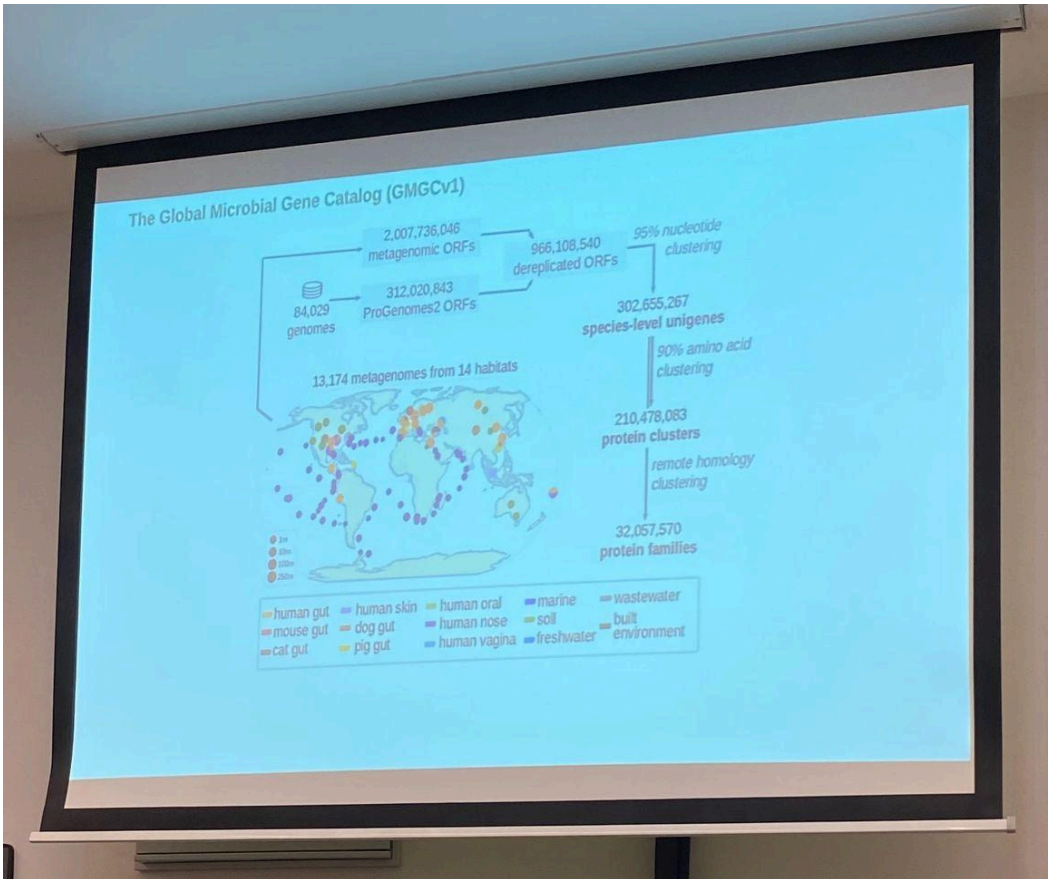
Large taxon sets differentiate between signal and noise. Can identify lineage-specific loss of miRNAs in certain clades. Concerted loss implies functional integration (i.e. phylogenetic profiling)
Did experimental miRNA knockouts to look at downregulation of associated targets. The miRNA target sites can be conserved, even though miRNAs are lost.

Luis Pedro Coelho

large catalogues of small genes
global microbiome 13,174 metagenomes from 14 habitats

metagenomes -> assembly -> contigs
mgRAST -> meta-GeneMarks
Collapsed ORFs into 303 million species-level unigenes
annotation with eggNOG mapper

2 Billion ORFs -> 80,000 genomes ProGenomes2 (312,000,000 ORFs)
966 e6 dereplicated ORFs, 95% nuc clustering -> 302 e6 species-level unigenes -> 90% aa clustering ß: 200 e6 protein clusters -> 32 e6 protein families



GMGCv1 as a resource

id of multi-habitat genes -> few unigenes shared between different environments

unigenes often shared between similar environments, e.g. found in human and dog gut

most genes are rare -> issue of an open global pan-genome. This is what is predicted by (nearly) neutral models of evolution
most genes and esp. rare ones are not under strong pos. selection
Most genes group into a (relatively) small number of families. 0.6% of families contain 50% of unigenes

habitat signature in the ratio of orthologs / unigenes (i.e. number of local pangenome genes). Environmental samples have different distributions than gut samples. Made up of a mixture of subpatterns (bimodal or with multiple distinct peaks). However, the different distributions correspond to different sub-habitats.

Future - more data and microproteins. Currently ignore proteins under 32 aa.
72 habitats, 60,000 samples

small genes are not just smaller

- hard to predict
- hard to assign function by homology
- modules, families harder to predict
- weak evolutionary signal

Tool to predict antimicrobial peptides (MACREL)
looking for significant Sequence similarity (BLAST) often does not work for small proteins.
Example on antimicrobial peptides -> AMPSphere 1.0
length distribution: mean 40 bp
Bacteria in host associated habitats produce more AMPs

How can we organize these sequences? Sequence identity based clustering does not work as well. Cannot cluster 47% of families (singletons)

-> are the gene intact -> have the relevant regulatory elements for transcription / translation
AMP genes as eroded/truncated versions of longer genes

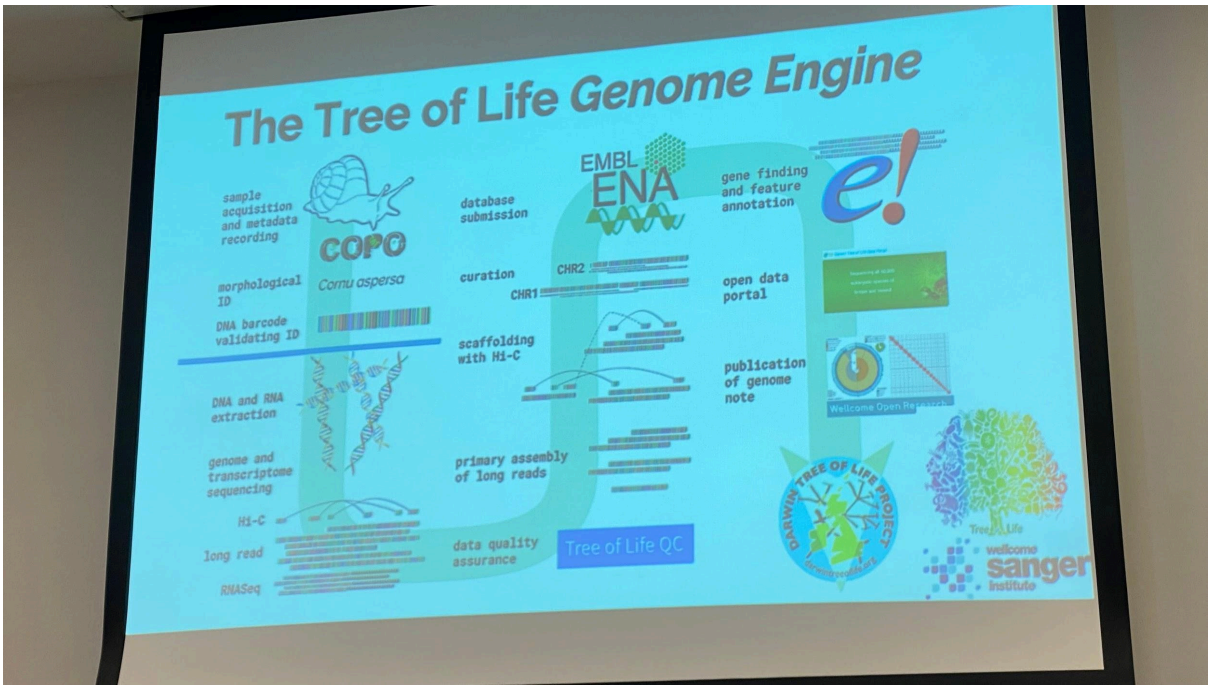
Christian Zmasek

Herpesviridae and coronaviridae as examples

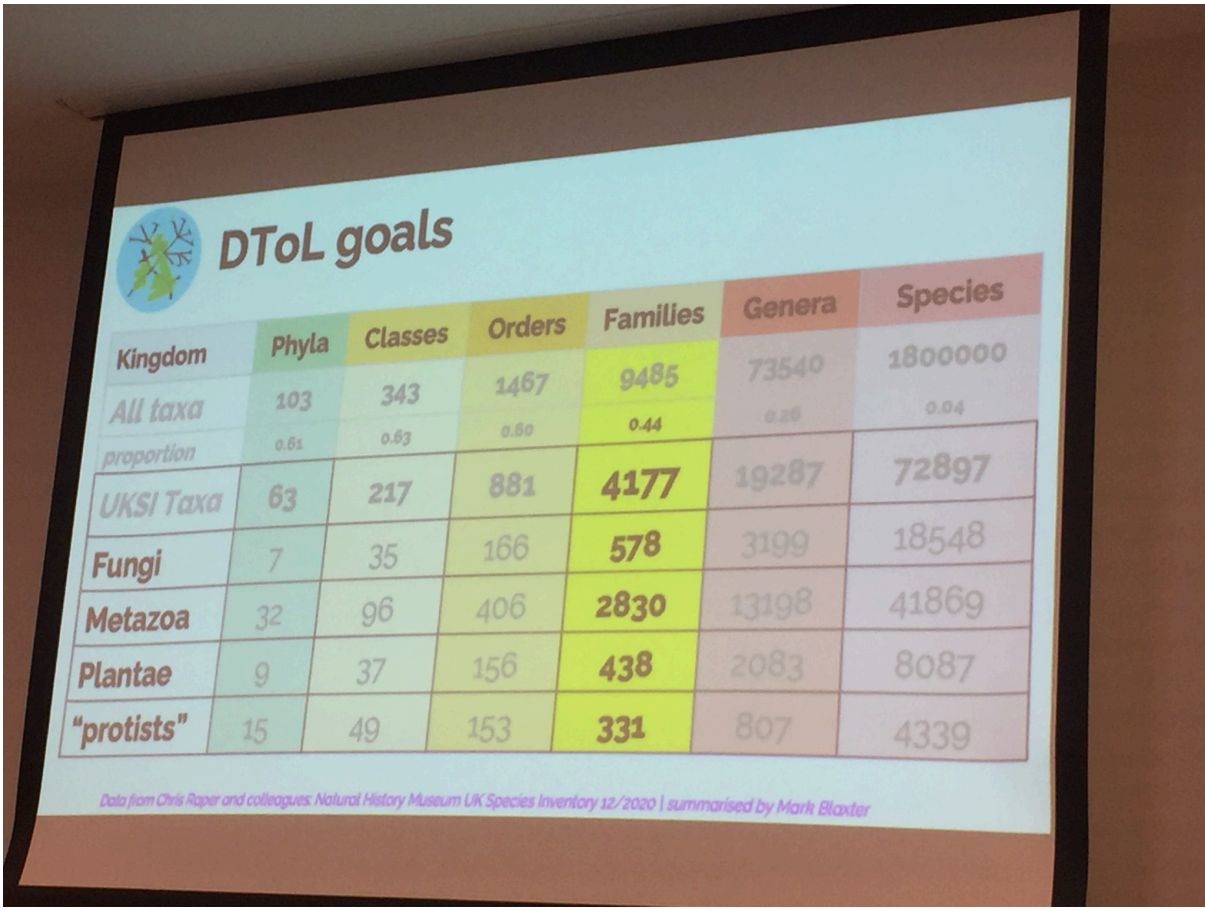
- easily compare viral proteins across different genera
- Genes with the same function do not have the same names in different viruses (e.g. ORF7)
- concept of orthology in multi-domain proteins is an issue. In a strict sense, orthology can only be assigned if all domains are orthologous
- Goal: classify viral proteins into groups with exact same domain architecture *and* they are orthologous (SOGs = Strict Ortholog Groups). New naming system where suffixes indicated taxonomic distribution
- works on sequence of domains / PFAM
- change of domain architecture along a tree

- duplication of viral genes a rare event? Anecdotal statement
- SOGs - cliques of orthologous groups and identical domain architecture
- computational approach -> what is the concept of a viral species?
- issue of e-value cutoffs, ad/hoc decision on evalue cutoff> e10^-6 and e10^-9; domain length cutoff
- phylogenetic analysis- semi automatised

Tom Richards

- “Go see the ortholog people and scare the living hell out of them”— Mark Blaxter
- sequence everything to full completion (chromosome level assembly) freely available
- 77,000 genomes
- 

- aquatic symbiosis genomics - 500 symbiotic systems; 1000 genomes
    - 171 species acquired 86 genomes in progress
- Darwin tree of life
    - Goal: 9485 families in the UK /British isles (mostly metazoa)



For each novel genome assembly a paper will be automatically written with minimal human interaction and then ‘published’

- Seuencing protists - 331 out of cultures; 10,000 single cell genomes picked from the environment. Latter good for getting organellar genomes specifically
- UK provenance of cultured samples has to be proven, family representative & priority species
- parallel genome and transcriptome sequencing from the same cell.
    - connect metagenomics and single cell sequencing from the same environmental example
    - single cell sorting using FACS melody -> heterotrophs are hard to find in this approach
    - cultivate the cell to get micro-culture
    - single cells give too little data quality - combine up to ten clonal cells will increase BUSCO score substantially
- many do not use standard genetic code - stop is not a stop
- Only 0.5% of know ciliates have been sequenced
- LECA gene set reconstruction - 64 proteomes -> orthoMCL -> TreeBuilding -> manual refinement -> group splitting -> Gain/loss parsimony / ML -> community annotation -> dataset assessment
- Inferring LECA gene content very sensitive to poor assemblies and annotations
- eukaryotes are chimera ->
- model organisms Paramecium bursaria -> RNAi, etc. funded for 10,000 KOs
- feedback loops have been shaped by HGT

- Claim that HGT is occuring in Eukaryotes -> paper HGT eukaryote <- virus Nature microbiol Irwin et al.
-