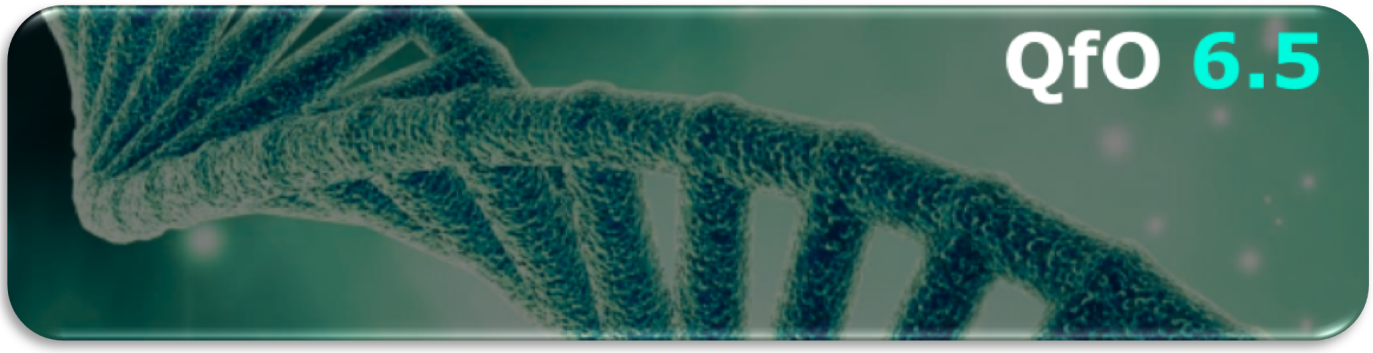


Quest for Orthologs 2021

Conference (Zoom) – https://tinyurl.com/qfo-frankfurt-2021 (Login: 12345)
Poster Session (Gather.Town) – https://gather.town/app/wqJZJU94uweeCrUt/QfO6-5

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

SENCKENBERG world of biodiversity

tbg translational biodiversity genomics

Program

Monday, August 2th 2021

Time on US West Coast	*Central European Summer time (Frankfurt)*	*Time in Japan*		
6:00	15:00	22:00	Welcome	Ingo Ebersberger
6:10	15:10	22:10	QfO past, present and future	Christophe Dessimoz
6:25	15:25	22:25	Protein length distribution is remarkably consistent across life	Yannis Nevers
6:40	15:40	22:40	Phylogenomic reconciliation links genome evolution to ecological innovation	Yuting Xiao
6:55	15:55	22:55	VOGDB - Virus Orthologous Groups database	Lovro Trgovec-Greif
7:10	*16:10*	*23:10*	*Break*	
7:25	16:25	23:25	Fast sequence and structure searching, clustering, and taxonomic analysis for the era of metagenomics	Johannes Söding
7:55	16:55	23:55	**Group discussion 1** Orthology of viruses	Benjamin Linard, Thomas Rattei
8:25	17:25	0:25	**Group discussion 2** Impact of AlphaFold2 on orthology inference	Wataru Iwasaki
8:55	*17:55*	*0:55*	*Break*	
9:10	18:10	1:10	Bacterial signaling proteins pose a unique challenge for ortholog prediction	Philip Davidson
9:25	18:25	1:25	abSENSE: a method to distinguish missing homologs from failures of homology search	Cara Weisman
9:40	18:40	1:40	Poster Session	Gather.Town

Tuesday, August 3th 2021

Time on US West Coast	*Central European Summer time (Frankfurt)*	*Time in Japan*		
August 2	**August 3**	**August 3**		
22:00	7:00	14:00	Applied Orthology: Extracting Value for Functional Inference	David S. Roos & Mark Hickman
22:15	7:15	14:15	fCAT - Assessing gene set completeness using domain-architecture aware targeted ortholog searches	Vinh Tran
22:30	7:30	14:30	Updates to HCOP: The HGNC comparison of orthology prediction tools	Tamsin Jones

22:45	7:45	14:45	Generating Reference Proteomes for QFO	Dushyanth Jyothi
23:00	*8:00*	*15:00*	*Break*	
23:15	8:15	15:15	**Group discussion 3** Ref-Proteomes; Benchmark	Adrian Altenhoff, Toni Gabaldón
23:45	8:45	15:45	**Group discussion 4** Completeness of complete genomes; Orthology value-added for functional genomics researchers	Michael Galperin David S. Roos
0:15	*9:15*	*16:15*	*Break*	
0:30	9:30	16:30	SHOOT: A online tool for gene sequence search and placement in a database of gene trees	David Emms
0:45	9:45	16:45	TOGA: a novel machine-learning approach to infer orthologs and integrate gene annotation with orthology inference at scale	Michael Hiller
1:00	10:00	17:00	Orthology ontology Applications for Life Science Database integration	Hirokazu Chiba
1:15	10:15	17:15	SonicParanoid: Machine Learning-Driven Integration of BBH and Protein Domain Analysis for Faster and More Accurate Orthology Inference	Salvatore Cosentino
1:30	10:30	17:30	An interactive visualisation tool for large gene families	Victor Rossier
1:45	10:45	17:45	Concluding Remarks	

Abstracts – Talks

Talk No.	Name	Institute	Titel
01	Yannis Nevers	Université de Lausanne	Protein length distribution is remarkably consistent across Life
02	Yuting Xiao	Carnegie Mellon University	Phylogenomic reconciliation links genome evolution to ecological innovation
03	Lovro Trgovec-Greif	University of Vienna	VOGDB - Virus Orthologous Groups database
04	Johannes Söding	Max-Planck-Institut	Fast sequence and structure searching, clustering, and taxonomic analysis for the era of metagenomics
05	Philip Davidson	Carnegie Mellon University	Bacterial signaling proteins pose a unique challenge for ortholog prediction
06	Cara Weisman	Harvard University	abSENSE: a method to distinguish missing homologs from failures of homology search
07	David S Roos & Mark Hickman	Univ Pennsylvania	Applied Orthology: Extracting Value for Functional Inference
08	Vinh Tran	Goethe University Frankfurt	fCAT - Assessing gene set completeness using domain-architecture aware targeted ortholog searches
09	Tamsin Jones	EMBL-EBI	Updates to HCOP: the HGNC comparison of orthology predictions tool
10	Dushyanth Jyothi	EMBL-EBI	Generating Reference proteomes for QFO
11	David Emms	University of Oxford	SHOOT: A online tool for gene sequence search and placement in a database of gene trees
12	Michael Hiller	Centre for Translational Biodiversity Genomics & Senckenberg Research Institute	TOGA: a novel machine-learning approach to infer orthologs and integrate gene annotation with orthology inference at scale
13	Hirokazu Chiba	Database Center for Life Science	Orthology Ontology Applications for Life Science Database Integration
14	Salvatore Cosentino	The University of Tokyo	SonicParanoid: Machine Learning-Driven Integration of BBH and Protein Domain Analysis for Faster and More Accurate Orthology Inference
15	Victor Rossier	University of Lausanne	An interactive visualisation tool for large gene families

Talk No.: 01

Protein length distribution is remarkably consistent across Life

Yannis Nevers
Université de Lausanne, Switzerland

In every living species, the function of a protein depends on its organisation of structural domains, and the length of a protein is a direct reflection of this. Because every species evolved under different evolutionary pressures, it is expected that the protein length distribution, much like other genomic features, varies across species. Here we evaluated this diversity by comparing protein length distribution across 2326 species (1688 bacteria, 153 archaea and 485 eukaryotes). We found that proteins tend to be on average slightly longer in Eukaryotes than in the other Domains, but that the variation of length distribution across species â€" in terms of shape and median value â€" is low, especially compared to the variation of other genomic features: genome size, number of proteins, gene length, distribution of GC content of genes and distribution of isoelectric points of proteins. Strikingly, the most extreme divergence observed in terms of protein length distribution, across all Domains of life, are likely due to artifacts in sequence annotation, most often leading to an overrepresentation of small proteins. Importantly, some of these problematic proteomes are not detected as such by current methods of genome quality annotation. This last result points to an unrecognised issue in the state of quality of available gene annotation data, and the need to develop new indicators of data quality, complementary to the existing ones.

Talk No.: 02

Phylogenomic reconciliation links genome evolution to ecological innovation

Yuting Xiao
Carnegie Mellon University, United States

Reconstructing the evolution of orthologous families in the context of species evolution links new molecular functions to ecological, geological, and morphological changes. By fitting a gene family tree into its corresponding species tree, phylogenetic reconciliation infers the ancestral family sizes and the duplications, losses and horizontal transfers in the history of the family. Advances in whole-genome sequencing and computational power have enabled phylogenomic reconciliation of large collections of gene trees. Notung-3.0 facilitates phylogenomic reconciliation via support for automated analysis pipelines and summaries of inferred events and ancestral states, aggregated over all gene families in a multi-tree analysis. Our analysis of 10,000 cyanobacterial gene trees with Notung-3.0 illustrates the power of phylogenomic reconciliation. Our reconstruction of genome dynamics reveals a history of family turnover, functional specialization, and genome streamlining. An analogous reconstruction with Wagner parsimony fails to discover these trends, illustrating the inferential advantage of phylogenetic over trait-based approaches. Reconciliation is also a powerful approach to studying the molecular basis of adaptation. In our dataset, the toxic bloom former, Microcystis aeruginosa, is an example. Reconciliation identifies gains and losses in more than a third of gene families in the Microcystis lineage, potential candidates for metabolic restructuring associated with Microcystis' unique ecological strategy. These include gain of the gas vesicle operons and loss of nitrogen fixation genes, both of which are known to be associated with the functional adaptation of M. aeruginosa. Phylogenomic reconciliation with Notung-3.0 recapitulates known biology in this well-studied species, highlighting the promise of this approach for novel discoveries.

Talk No.: 03

VOGDB - Virus Orthologous Groups database

Lovro Trgovec-Greif
University of Vienna, Austria

Viruses are a very abundant group of biological entities and the real viral omnipresence is only being discovered with the help of high-throughput sequencing technologies. Besides being abundant, they are also genetically very diverse and without a single common ancestor. Current metagenomic experiments are producing lots of metagenome assembled viral genomes or viral fragments and the most of them will never be characterized experimentally. In order to work with unknown viral genomes and to predict their function, gene orthology relationships are of very high value. We developed the database VOGDB, which represents Virus Orthologous Groups. VOGDB contains groups of viral proteins with even distant evolutionary relationship. In order to achieve highest quality, viral genomes are extracted from RefSeq and are filtered by their annotation quality. Polyproteins are segregated into mature peptides, if these are not yet annotated. Clusters of orthologous groups for viral proteins are calculated separately for phage and non-phage genomes because phages have little recent evolutionary relationship with eukaryotic viruses. Profile Hidden Markov models (HMMs) are created from the clusters and HMMs are clustered in order to capture distant evolutionary relationships and create VOGs. VOGs are subsequently functionally annotated by transferring function from proteins in VOGs that are present in UniProt/Swiss-Prot database. Clusters are validated with two approaches. First, we checked for the homogeneity of functions within VOGs and second we assessed the homogeneity of structures using SCOPe superfamilies. Both show high degree of homogeneity and separation. VOGDB is updated with every release of RefSeq.

Talk No.: 04

Fast sequence and structure searching, clustering, and taxonomic analysis for the era of metagenomics

Johannes Söding
Max-Planck-Institut, Germany

To keep pace with the fast-growing amount of protein sequences extracted from genomic and metagenomic data, our group develops fast software tools and algorithms for large-scale protein sequence analysis. Our core library, on which many of our tools are based, is MMseqs2. I give an overview of our current tools for sequence searching and clustering, genomic assembly and taxonomic annotation and will report on the latest unpublished developments: fast profile-profile searches and fast structure searches.

Talk No.: 05

Bacterial signaling proteins pose a unique challenge for ortholog prediction

Philip Davidson (Presenter), Dannie Durand
Carnegie Mellon University, USA

The signaling proteins that allow bacteria to sense and respond to environmental cues are a substantial component of the bacterial proteome and important targets for functional and evolutionary analyses. However, ongoing turnover in the signaling repertoire obscures orthologous relationships. Bacterial signaling proteins are easily identified by their ubiquitous interaction domains, but correct assignment to orthologous groups is a challenge for sequence-based orthology prediction. Our analysis of the Firmicute endosporulation signaling pathway illustrates the extent of this problem. Endospore formation is the characteristic survival response in Bacilli and Clostridia, the two major classes in Firmicutes. In Bacilli, the master regulator of endsporulation is activated by a three-step BBphosphorelayBB pathway. In Clostridia, sequence-based methods identified distantly related homologs, but not unambiguous orthologs of Bacillar phosphorelay proteins. The apparent lack of Clostridial orthologs led to the widely accepted view that different pathways initiate endosporulation in Bacilli and Clostridia. In contrast, by comparing domain content in genomic neighborhoods, we identified putative phosphorelay orthologs in Clostridia, a prediction that we confirmed experimentally using cross-species complementation. In the study that followed, we overturned the prevailing hypothesis concerning the evolution of endosporulation during the Great Oxidation Event. Given the importance of Clostridia to a healthy gut microbiome and to synthetic biofuel and solvent production, the impact of finding phosphorelay orthologs in numerous Clostridial species far exceeds the scope of a single study. In summary, orthology prediction methods that address the unusual characteristics of bacterial signaling proteins are urgently needed. Leveraging genomic neighborhood information offers a potential solution.

Talk No.: 06

abSENSE: a method to distinguish missing homologs from failures of homology search

Cara Weisman
Harvard University, United States

Biologists regularly search for homologs of genes in species throughout the tree of life. Sometimes, such homology searches fail to find a homolog in a given taxon. How should this be interpreted? There are at least two possibilities. First, the homolog could truly be missing, as a result of events like gene birth or gene loss, which may have biological interest. Second, the absence of the search method in question may simply have failed to detect a homolog even in the absence of evolutionarily interesting events: the homolog could be present, but merely have diverged too far from the queried sequence to be detected. We present a tool, abSENSE, to help distinguish between these two possibilities. abSENSE predicts whether, under a null evolutionary model in which homologs are present and evolving in identical fashion in all species, homologs of an input gene in a target taxon would be successfully detected by BLASTP. If abSENSE predicts that homologs would be undetected in this scenario, one need not invoke evolutionary events like gene loss or gain to explain an apparent absence. We demonstrate the utility of abSENSE by applying it to the case of lineage-specific genes, commonly interpreted as 'novel' genes because they lack homologs in all species outside of a narrow taxon, and show that a majority can be explained by homology detection failure, cautioning against the assumption of novelty.

Talk No.: 07

Applied Orthology: Extracting Value for Functional Inference

David S Roos & Mark Hickman
Univ Pennsylvania, United States

In recent years, technological advances have provided a wealth of well-annotated organismal genomes, improved methods for assessing deep phylogenetic relationships, detailed functional genomics datasets (e.g. comprehensive transcriptional & proteomic profiles), and genomic-scale phenotypic characterization (of essentiality data, subcellular localization, etc), even for non-model organisms. Genome database resources serve as a primary entry point through which a large and growing community of biomedical researchers gains access to this information … and exploit orthologous relationships for functional inference across species. For example, the Eukaryotic Pathogen & Vector Bioinformatics Resource Center (VEuPathDB.org) is used by ~40K unique users per month, from 100+ countries, with the average user returning approximately weekly … and virtually all of these users take advantage of orthology data: as anchors for genome alignments and synteny assessment; for automated annotation of EC numbers, GO terms, etc; and for transitive functional inference. A user might wish to identify genes based on expression patterns in the species of interest, subcellular location of orthologs in closely related species, ortholog essentiality in model organisms, and annotation of manually-curated orthologs across the tree of life. We will highlight changes in the BLAST-based OrthoMCL algorithm used by VEuPathDB for ortholog identification, extending scalability from hundreds to thousands of species, and improving users' ability to assess phyletic patterns of paralogous gene amplification and loss. We also hope to stimulate discussion on how the Q4O initiative can increase the value of orthology data by providing the broader research community with integrated insights based on multiple assessment methods.

Talk No.: 08

fCAT - Assessing gene set completeness using domain-architecture aware targeted ortholog searches

Vinh Tran
Goethe University Frankfurt, Germany

The assessment of gene set completeness is a routine task in genome analysis. The standard workflow starts with the identification of a set of single copy core genes for the taxonomic group the newly sequenced species, the target, is part of. The fraction of missing core genes serves then as a proxy of the target gene set completeness. Genes that are represented but differ significantly in length from the expectation provide information about the fragmentation status of the predicted gene models, and ultimately gene duplication levels can be assessed. Though well established, this approach comes along with a number of restrictions of which the focus on single copy genes, the use of an error-prone unidirectional ortholog search, and the application of a simple length cutoff criterion are the most prominent ones.
Here we present fCAT, a novel algorithm for assessing gene set completeness using a targeted and domain architecture-aware ortholog search. As a consequence, fCAT's core sets are not limited to single-copy orthologs by that providing a comprehensive overview of the core gene set. Next to the conventional length difference assessment, fCAT identifies target genes that significantly differ in their domain architectures from the core genes allowing an alternative view on the accuracy of target gene models. Phylogenetic profiles resulting from the analysis can be visualized and explored in the context of the entire orthologous groups which provides the necessary information to identify and ultimately correct erroneous gene annotations.

Talk No.: 09

Updates to HCOP: the HGNC comparison of orthology predictions tool

Tamsin Jones
EMBL-EBI, UK

Multiple resources currently exist that predict orthologous relationships between genes. These resources differ both in the methodologies used and in the species they make predictions for. The HGNC Comparison of Orthology Predictions search tool (HCOP, https://www.genenames.org/tools/hcop) was created in the early 2000s to aggregate, display and simplify searching of orthology assertions from a variety of orthology prediction tools, for a specified human gene or set of genes. Since its original inception, HCOP has undergone a series of changes as new orthology resources were released and computer technologies improved, culminating in a complete reimplementation of the HCOP pipeline in 2014 to better utilize the available data and computational methods. The changes made to HCOP include the addition of several new species, increasing from just 4 species in 2006 to a total of 20 in 2021. The number of orthology sources aggregated by HCOP has increased to 14 from an initial 6, with some of the original orthology sources being removed due to the resource no longer being maintained or updated. The decision to add a new orthology source involves several factors, including the availability of the orthology data, update frequency, species coverage and methods used to create the orthology assertions. Data from HCOP are used extensively in our work naming genes as the Vertebrate Gene Nomenclature Committee (https://vertebrate.genenames.org), and by a number of external resources such as MGI (http://www.informatics.jax.org/) and RGD (https://rgd.mcw.edu/).

Talk No.: 10

Generating Reference proteomes for QFO

Dushyanth Jyothi
EMBL-EBI, United Kingdom

Advancements in the sequencing technologies led to the rapid growth of sequenced genomes. This is further propelled by the massive sequencing projects currently underway such as the Darwin Tree of Life Project. Organising, identifying and annotating proteomes of important organisms is a huge challenge in the ever-growing proteome space. UniProt addresses this by defining a set of "Reference Proteomes" as "landmarks" considering various factors including well-studied model organisms, organisms of interest for biomedical and biotechnological research and organisms representing taxonomic diversity. This is achieved by both expert manual curation and advanced computational methods. "Reference Proteomes for QFO" are a subset of reference proteomes, particularly chosen by the QFO community, which are used as a common dataset for comparing orthology inference methods. These are provided annually in the established data file formats and standards. In this talk I will provide an overview of the process involved in generating reference proteomes and highlight some of the challenges associated with production.

Talk No.: 11

SHOOT: A online tool for gene sequence search and placement in a database of gene trees

David Emms
University of Oxford, United Kingdom

BLAST identifies the biological sequences with regions of local similarity to a given query sequence. The output of BLAST is a similarity-ranked list of biological sequences and other pairwise-alignment statistics. Although these similarity statistics are of great utility, a frequent use of BLAST is to identify orthologous sequences in other organisms. Such orthology relationships are poorly inferred from a ranked list of similar sequences and are more accurately determined and interpreted using gene trees.
We present SHOOT, a website which searches a novel query sequence against a database of gene trees and returns a gene tree with the given query sequence correctly grafted within it. We show that SHOOT can perform this search and placement step rapidly enough to facilitate general use. The key to achieving this is that SHOOT builds on OrthoFinder to pre-compute the relationships between the sequences in its database in advance. In summary, SHOOT is an accurate and fast tool for complete phylogenetic analysis of novel query sequences. It is available online at www.shoot.bio.

Talk No.: 12

TOGA: a novel machine-learning approach to infer orthologs and integrate gene annotation with orthology inference at scale

Michael Hiller
Centre for Translational Biodiversity Genomics & Senckenberg Research Institute, Germany

We present TOGA (Tool to infer Orthologs from Genome Alignments), the first method that integrates gene annotation and ortholog inference. TOGA implements a novel methodology to infer orthologous genes between related species that does not rely on protein or coding exon sequences. Instead TOGA utilizes information contained in whole genome alignments and uses machine learning to accurately distinguish orthologs from paralogs or processed pseudogenes based on alignments of intronic and intergenic regions. TOGA scales to many genomes, which we show by applying it to annotating genes and inferring orthologs across more than 450 mammals and 400 birds, creating the largest comparative datasets for these clades so far. We also use TOGA to show that a subset of mammalian BUSCO genes is not well conserved and highlight other highly-conserved single copy genes as replacements. Together, TOGA is a powerful and scalable method to annotate and compare genes and infer orthologs in the genomic era.

Talk No.: 13

Orthology Ontology Applications for Life Science Database Integration

Hirokazu Chiba
Database Center for Life Science, Japan

The need of a common ontology for describing orthology information in the life science research domain has led to the creation of the Orthology Ontology (ORTH). This ontology structures orthology-related resources in heterogeneous databases. Specifically, it can be used to represent different types of orthology information such as pairwise relationships, hierarchical and non-hierarchical ortholog groups. The benefit of employing ORTH is the interoperable data model that enables querying over relevant distributed databases on the web with SPARQL. A practical example of this benefit is the SwissOrthology project (see https://swissorthology.ch). So far, this ontology is employed for describing several orthology databases including OMA, MBGD and reusable resources published on the web such as HomoloGene. Recently, it has been adopted by the BioResource Metadatabase, a data repository for metadata of bioresources at RIKEN, a research institute in Japan. For example, OMA orthology data was used to relate different gene-centric RIKEN datasets. The application of other orthology data sources besides OMA is straightforward using the ORTH ontology, and this approach not only avoids ambiguities but also enables the reuse of existing queries that are written using ORTH. In future work, we will also consider adopting requested concepts such as pangenomes to extend the application. The ORTH ontology was developed with the Web Ontology Language and it is publicly available at http://purl.org/net/orth . Moreover, the ontology is published in ontology sharing services including BioPortal and the EBI Ontology Lookup Service.

Talk No.: 14

SonicParanoid: Machine Learning-Driven Integration of BBH and Protein Domain Analysis for Faster and More Accurate Orthology Inference

Salvatore Cosentino
The University of Tokyo, Japan

Accurate inference of orthologous genes constitutes a prerequisite for genomic and evolutionary studies. SonicParanoid is one of the fastest methods for orthology inference and comparably accurate to well-established methods despite being orders of magnitude faster. Nevertheless its scalability is hampered by the lengthy all-vs-all alignments, and sequence-similarity search alone is not enough to predict very distant orthologs. In this work we try to tackle these two limitations using machine learning. We substantially reduced the all-versus-all alignment execution time using an AdaBoost model which exploits the properties of the Bidirectional-Best-Hit and the factors affecting the computational time in local sequence alignment. Evaluation based on multiple datasets showed reductions in execution time up to 50% without negative effects on the accuracy of the orthology inference. To address the second limitation we trained a doc2vec model with domain-architectures extracted from the input proteins, and we used it to infer orthologs based on domain-architecture similarities, which resulted in an increase of one-third in the number of predicted orthologs. The way we reduced all-vs-all execution time could be used by other graph-based methods, while the domain-based approach could in the future, thanks to its scalability, eliminate the need for all-vs-all alignments in orthology inference

Talk No.: 15

An interactive visualisation tool for large gene families

Victor Rossier
University of Lausanne, Switzerland

Gene families are sets of genes that evolved through duplications and speciations. These evolutionary histories can be complex to interpret and thus greatly benefit from an appropriate visualization. However, existing viewers struggle to handle large gene families resulting from the growing number of genomes processed in comparative genomic pipelines. For example, the Ensembl viewer collapses all subtrees around the focal gene to reduce the gene family complexity, thus losing information associated with these subtrees (Herrero et al. 2016). Here, we introduce a gene family viewer with a unique two dimensional structure to visualize and interpret large gene families.

The viewer structure consists of a two dimensional matrix with a gene tree on the left and a species tree on top. The matrix rows match genes and subfamilies in the gene tree, while columns match species and taxa in the species tree. Gene copy numbers are displayed per subfamily and taxon in the matrix squares. When collapsing a subfamily in the gene tree, the corresponding rows are summed to summarize the number of copies per species in that subfamily. By contrast, when collapsing a taxon in the species tree, two actions are triggered. First, all subfamilies defined for that taxon are collapsed. Secondly, the corresponding columns are merged by averaging the number of copies of that taxon in each corresponding subfamily.

This viewer allows the study of large gene families through three key aspects: (i) By treating the species tree as a separate dimension, its inherent redundancy across the gene tree is removed when collapsing speciation nodes like a factorization operation. (ii) Dynamic collapsing and unfolding of subfamilies and taxa allows users to navigate complex gene families by summarizing irrelevant subtrees and exploring relevant ones. (iii) The biological interpretation of gene families is facilitated with the design of the viewer. For instance, lineage-specific expansions are made striking with the heatmap-like matrix coloring, while gene losses are represented by grey squares.

Herrero, Javier, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J. Vilella, et al. 2016. “Ensembl Comparative Genomics Resources.” Database: The Journal of Biological Databases and Curation 2016 (February). https://doi.org/10.1093/database/bav096.

Abstracts – Posters

Poster No.	Name	Institute	Titel
01	Ivana Piližota	EMBL-EBI	Scaling Homology Inference in Ensembl
02	Erik Sonnhammer	Stockholm University	InParadiam: Faster orthology analysis with the InParanoid algorithm
03	Uciel Chorostecki	Barcelona SuperComputing Center	MetaPhOrs 2.0: phylogeny-based inference of orthology and paralogy across different species
04	Michael Y. Galperin	NCBI, NLM, National Institutes of Health	Using COGs to identify missing orthologs of widespread genes
05	Rachael Cox	University of Texas at Austin	Orthology models enable large-scale comparative proteomics
06	Diego Fuentes	IRB Barcelona	PhylomeDB v5: An updated site to browse and mine genome-wide catalogs of gene phylogenies
07	Ikuo Uchiyama	National Institute for Basic Biology	Functional inference of microbial genomes based on orthology assignment in MBGD
08	Claire McWhite	Princeton University	Application of language models of proteins to sequence analysis
09	Hannah Mülbaier	Goethe University Frankfurt	Searching for orthologs in un-annotated genome assemblies with fDOG – Assembly
10	Frank D'Agostino	Harvard University	Automating Evolutionary Distance Computation to Interpret Homology Search Error Detection
11	Arpit Jain	Goethe University Frankfurt	The evolutionary traceability of a protein

Poster No.: 01

Scaling Homology Inference in Ensembl

Ivana Piližota, Jorge Alvarez-Jarreta, Carla Cummins, Thiago Genez, Cristina Guijarro-Clarke, Arthur Gymer, Matthieu Muffato, Thomas Walsh, David Thybert, Kevin Howe

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI)

Ensembl (www.ensembl.org) provides comprehensive homology data (gene trees, orthology and paralogy) across the eukaryote tree of life. Recently, large-scale sequencing initiatives such as the Darwin Tree of Life (https://www.darwintreeoflife.org) have begun generating genome assemblies at an unprecedented pace. Ensembl will have to adapt its comparative genomics pipelines to process such high volumes of genomic data in a timely manner.
As a scaling strategy, we have developed a high throughput homology annotation pipeline comparing a query genome to a set of ~40 reference genomes using best BLAST hit and reciprocal best BLAST hit criteria. Moving from quadratic to linear scaling will enable provision of homology predictions on the Ensembl Rapid Release site soon after a new genome has been annotated. Reference genomes are selected according to community needs, genome availability and the tree of life coverage.
Current developments include the incorporation of well-established gene tree inference methods and the application of deep learning methods to improve the sensitivity and specificity of our homology predictions. Our large number of gene trees spanning the eukaryotic tree of life will also enable us to explore methods for accurately inserting new sequences onto already computed gene trees, and further increase the efficiency of the approach.

Poster No.: 02

InParadiam: Faster orthology analysis with the InParanoid algorithm

Erik Sonnhammer
Stockholm University, Sweden

InParanoid is a popular algorithm for orthology analysis, but based on BLAST it suffers from long runtimes on large datasets. Here, we present an update to the InParanoid algorithm that can use the faster tool DIAMOND instead of BLAST for the homolog search step. We show that InParadiam reduces the runtime by 94%, while still performing equally well in the Quest for Orthologs benchmark.

Poster No.: 03

MetaPhOrs 2.0: phylogeny-based inference of orthology and paralogy across different species

Uciel Chorostecki, Anna Vlasova, Diego Fuentes, Manu Molina, Toni Gabaldon
Barcelona SuperComputing Center, Spain

Defining homology relationships across genes in different species is a central task in comparative genomics. Homologous genes can be orthologs or paralogs, depending on whether they diverged from their common ancestor through speciation or duplication, respectively, and this is best inferred through phylogenetic analysis. Different databases provide access to thousands of gene phylogenies across different taxa, but they often do not provide specific orthology or paralogy information. MetaPhOrs (from Meta-predictions of Phylogeny-based Orthologs) is a free and unique web-server providing phylogeny-based orthology and paralogy predictions computed from phylogenetic data available in 13 different large-scale databases and associated with a consistency-based confidence score. MetaPhoRs was first described a decade ago, and it has been regularly updated and expanded. Here we developed an improved version of the web-server which includes major new implementations and provides orthology and paralogy relationships derived from ~11 million gene family trees -from 13 different source repositories- across ~4,000 different fully-sequenced species. MetaPhOrs has been benchmarked alongside other methods, showing the superiority of the integrative approach over the individual methods or databases. Altogether we think that MetaPhOrs constitutes a resource of broad interest and applicability. MetaPhOrs server is freely available, without registration, at http://orthology.phylomedb.org/ .

Poster No.: 04

Using COGs to identify missing orthologs of widespread genes

Michael Y. Galperin
NCBI, NLM, National Institutes of Health, United States

The recent release of the Clusters of Orthologous Genes database, https://www.ncbi.nlm.nih.gov/research/COG , covers complete genomes from 1,187 bacteria and 122 archaea from1,234 genera. While COGs could be used for a variety of comparative genome analyses, one of its unique traits is the availability of COG-specific patterns of the presence and absence of the evolutionarily conserved genes in the respective organisms. This allows one to identify protein families (COGs) that are not encoded in the given genome and, conversely, the genomes that do not encode the given gene. This feature has been previously used to analyze metabolic pathways and predict candidate ORFs for the missing enzymatic function. This tool, however, has a more pragmatic application: checking whether the given genome sequence carries the full set of the conserved genes whose protein products are fully expected to be encoded by this genome. Our recent analysis of the distribution of the ribosomal genes identified some ribosomal genes that were genuinely lost in the course of evolution but also a number of missing genes whose absence appeared extremely unlikely. One of the reasons for that was the presence of frameshifts and nonsense mutations, some of which reflected likely sequencing errors. Others resulted from annotation errors: some short genes were overlooked and some longer one were listed as pseudogenes owing to unrealistic sequence models. A recent analysis of conserved sporulation genes revealed a similar picture. These observations show that certain genes may be missed in the GenBank entries even for the relatively small prokaryotic genomes.

Poster No.: 05

Orthology models enable large-scale comparative proteomics

Rachael Cox
University of Texas at Austin, United States

Molecular characterization of the last eukaryotic common ancestor (LECA) would give a unique view of the composition and cellular organization of a key progenitor that gave rise to all extant eukaryotes. Most of what we know thus far about LECA's cellular organization stems from studies of gene content and phylogeny across current-day species. LECA likely existed ~2 billion years ago and possessed sophisticated cellular machinery (e.g., at least one nucleus, endoplasmic reticulum, Golgi apparatus, endosomes, mitochondria, and at least one cilium). However, a more detailed look at LECA's makeup remains a significant challenge. To address this challenge, we combine phylostratigraphy on Quest for Ortholog-affiliated orthogroups and mass spectrometry proteomics across 33 eukaryotes (with 2 bacterial and 2 archaeal species acting as prokaryotic outgroups) in order to reconstruct LECA's likely protein complement and those proteins likely organization into multiprotein assemblies. The resulting view of LECA helps to annotate ancient protein complexes and illuminate the molecular organization and evolution of the fundamental biochemical machinery shared broadly across eukaryotic clades.

Poster No.: 06

PhylomeDB v5: An updated site to browse and mine genome-wide catalogs of gene phylogenies

Diego Fuentes
IRB Barcelona, Spain

Gene phylogenies represent the evolutionary relationships across genes in different species. These phylogenetic trees are commonly used to aid in the inference of homology relationships (i.e. orthology and paralogy) as well as of evolutionary relevant events such as family expansions, recombination and horizontal gene transfer. The plurality of evolutionary histories of genes encoded by an organism's genome is best represented by a genome-wide collection of phylogenetic trees (i.e.: a phylome).

Poster No.: 07

Functional inference of microbial genomes based on orthology assignment in MBGD

Ikuo Uchiyama
National Institute for Basic Biology, Japan

Microbial genome sequencing has recently been expanded to uncultured microbes by sequencing DNAs extracted from the environment and determining genomes of dominant species as metagenome assembled genomes, revealing the great diversity of the microbial world. Predicting metabolic potential of microbes from genomic information is thus important subject to utilize this information. We are developing a tool to predict functional potential of novel microbial genomes through orthology assignment based on the MBGD ortholog table. For this purpose, we have expanded the MyMBGD functionality to accept user genomes and assign an MBGD ortholog group to each gene using the profile search functionality of MMseqs. To evaluate the metabolic potential of the query genome from this assignment, we refer to the cross-reference to the KEGG Orthology database and utilize the Genomaple software (formerly MAPLE Takami et al. 2016) to calculate the module completion ratio (MCR) for each KEGG Module entry. The result is displayed on the entire list of KEGG Modules with MCRs in the query genome. Although, ideally, the MCRs of the KEGG modules that present in the query genome should be 100%, there are often the cases where MCR is high enough but less than 100%. To identify candidate genes to fulfill the missing genes in such a case, we utilized several MBGD search functions including those for homologous orthologs, neighborhood orthologs, and orthologs with similar phylogenetic profiles.

Poster No.: 08

Application of language models of proteins to sequence analysis

Claire McWhite
Princeton University, United States

Sequence similarity measures and multiple sequence alignments underlie ortholog inference. Enhancements in these areas could directly improve ortholog detection, along with many other standard processes in computational biology, including motif finding, amino acid conservation analysis, structure prediction, and molecular phylogenetics. In the past several years, natural language Transformer models have been found to be highly successful at capturing subtleties of human language. Interestingly, models of protein language can be also constructed, where a protein sequence is considered as a sentence, and each amino acid is a word. As in human language, where a word's context matters strongly to its meaning, protein language models capture not only an amino acid position’s variability, but also its contextual relationship with other amino acids in the sequence. Protein language models have been previously shown to remarkably capture features of protein structure and function. Here, we demonstrate two applications of these language models of proteins. First, fast all-by-all sequence similarity search. Second, a novel clustering-based multiple sequence alignment algorithm based on matching amino acid positions with similar contexts.

Poster No.: 09

Searching for orthologs in un-annotated genome assemblies with fDOG – Assembly

Hannah Mülbaier
Goethe University Frankfurt, Germany

The identification of orthologs in the genomes of newly sequenced species is a relevant step for their integration into a broad range of evolutionary and functional studies. Numerous approaches varying in computational complexity, sensitivity and specificity have been developed for this purpose. However, one dependency is common to all tools: they require comprehensively annotated gene sets as input where any overlooked gene will result in a missed ortholog. Here, we present fDOG – Assembly, a targeted profile-based ortholog search tool that can identify orthologs in unannotated genome assemblies. Using an aligned set of pre-computed core orthologs as a start, the algorithm generates a consensus sequence that serves as query for a tblastn search. Hit regions are scanned for the presence of a gene using Augustus guided by a block profile, which was computed from the core ortholog alignment. In case a gene is annotated, the encoded protein is tested for orthology and, upon success, will be added to the core orthologous group. An assessment of domain architecture similarity to a reference protein in the core set is optional. We initially benchmarked fDOG – Assembly, which revealed a performance that is comparable to the ortholog search in fully annotated gene sets. We envision that fDOG – Assembly will be helpful for closing gaps in phylogenetic profiles due to annotation artefacts, but even more for studies requiring the identification of candidate genes in genomes irrespective of their annotation status.

Poster No.: 10

Automating Evolutionary Distance Computation to Interpret Homology Search Error Detection

Frank A. D'Agostino[1], Caroline M. Weisman[1], Sean R. Eddy[1,2,3]
Harvard University, United States

1 Department of Molecular & Cellular Biology,
2 Howard Hughes Medical Institute,
3 John A. Paulson School of Engineering and Applied Sciences,
Harvard University, Cambridge MA, USA

Sequence homology searches with computational tools like BLAST and HMMER are fundamental, but there is a limit to the power of any homology search tool to successfully find homologs. When homologs are not detected, there remains a question of how to interpret this absence; for example, a lack of detectable homologs in species outside some clade is often interpreted as support for de novo origination of a gene. A new computational method from our lab, abSENSE, determines the probability that a homology search program is expected to fail to find an ortholog for a given gene simply from expected lack of search sensitivity (e.g. for rapidly evolving or short genes). Currently, abSENSE requires user input of precalculated evolutionary distances between the source and target species, and the original paper only precalculated distances for two well-studied clades (yeasts and *Drosophila*). Here, we developed a fully automated method to precompute the necessary evolutionary distances among any set of input species. This method identifies available protein annotations from reference databases, locally downloads the necessary files, identifies well-conserved genes, and uses them to calculate the desired evolutionary distances. This addition makes it easier to apply abSENSE analyses to different target species genomes.

Poster No.: 11

The evolutionary traceability of a protein

Arpit Jain, Dominik Perisa, Arndt von Haeseler, Ingo Ebersberger
Goethe University, Frankfurt, Germany; LOEWE Centre for Translational Biodiversity Genomics, Frankfurt, Germany

Orthologs document the evolution of genes and metabolic capacities encoded in extant and ancient genomes. However, the similarity between orthologs decays with time, and ultimately it becomes insufficient to infer common ancestry. This leaves ancient gene set reconstructions incomplete and distorted to an unknown extent. Here we introduce the “evolutionary traceability” as a measure that quantifies, for each protein, the evolutionary distance beyond which the sensitivity of the ortholog search becomes limiting. Using yeast, we show that genes that were thought to date back to the last universal common ancestor are of high traceability. Their functions mostly involve catalysis, ion transport, and ribonucleoprotein complex assembly. In turn, the fraction of yeast genes whose traceability is not sufficient to infer their presence in last universal common ancestor is enriched for regulatory functions. Computing the traceabilities of genes that have been experimentally characterized as being essential for a self-replicating cell reveals that many of the genes that lack orthologs outside bacteria have low traceability. This leaves open whether their orthologs in the eukaryotic and archaeal domains have been overlooked. Looking at the example of REC8, a protein essential for chromosome cohesion, we demonstrate how a traceability-informed adjustment of the search sensitivity identifies hitherto missed orthologs in the fast-evolving microsporidia. Taken together, the evolutionary traceability helps to differentiate between true absence and non- detection of orthologs, and thus improves our understanding about the evolutionary conservation of functional protein networks. “protTrace,” a software tool for computing evolutionary traceability, is available at https://github.com/BIONF/protTrace.git

Poster Session via Gather.Town

Gather.Town is a “video chat platform designed to make virtual interactions more human.” We have customized Gather.Town for the QfO6.5 meeting with various spaces to better spend time with our scientific community. It is probably a good idea to **close ZOOM** before using Gather.Town.

Video tutorial

There is brief video tutorial on how to use Gather.Town (sorry for being not a professional Vlogger...). Click here to see it.

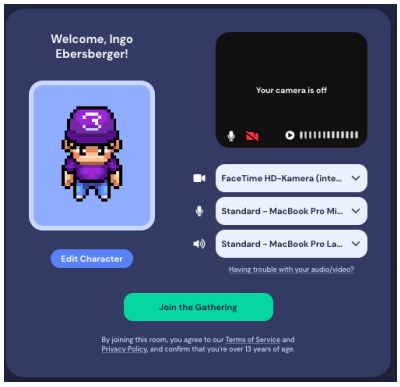
Quick start

1. Please use **Google Chrome** (best) or Firefox (we had difficulties) when using Gather. Town. Mac users should avoid using Safari.

2. Link to the QfO6.5 Gather.Town: https://gather.town/app/wqJZJU94uweeCrUt/QfO6-5

3. Sign-In: All registered participants are on the guest list of the QfO6.5 Gather.Town. Since we have to pay per user, we have to restrict access to the spaces, unfortunately. Upon clicking the access link you will be asked to sign in with your email address. Please use the one that you provided with the meeting registration

4. You will be asked to create an avatar. Your name should already be set. If not, enter your full name (first and last names). Make sure to enter your real name so people can find you in the space.



5. You will also be asked to activate your video and audio inputs, similar to any zoom call.

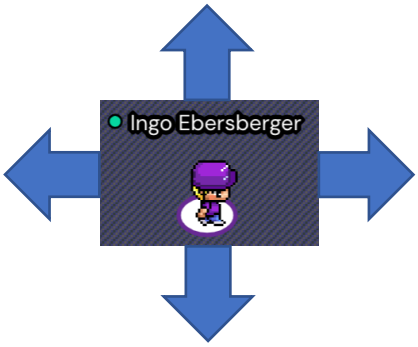
6. After you join, you will arrive in the QfO6.5 conference area. We have four main rooms,

Main hall **Poster hall**

Meeting room 1 **Meeting room 2**

How do I navigate the site?

Simply use the arrow keys on your keyboard to move your Avatar.



How do I speak to someone?

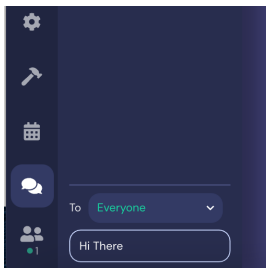
When your avatar is near another avatar an automatic video live chat will begin, much like a Zoom call. To stop the chat, simply walk away.

How do you find someone you in Gather Town?

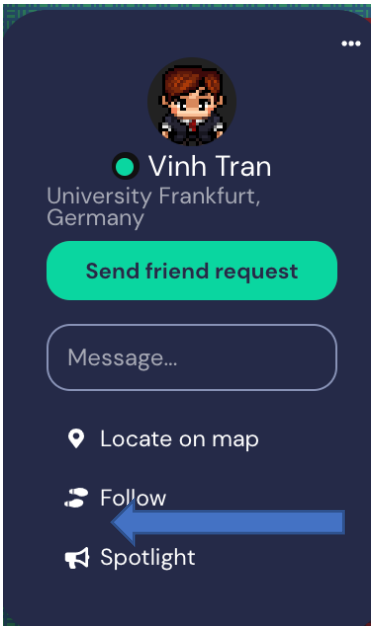
Perhaps you want to discuss their presentation, ask a question about their poster, or schedule a chat about a job opportunity.

Option 0: You are lucky and you 'see' the person close by. Just walk over…

Option 1: Start a private chat, using the chat feature at the bottom left of your screen. There are multiple options for chats (everyone, people nearby, or you can select an individual).



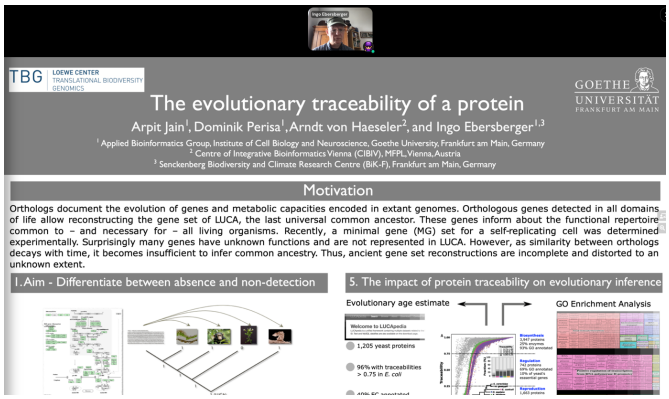
Option 2: Use the follow feature. To activate “follow” mode, click on the person’s name in the participants list (bottom left of screen) and click “follow”. Your avatar will now walk directly to the person you are looking for even if the person is moving. Once you arrive, you will continue to follow the participant without using the arrow keys. To stop following, press any arrow key to move away from the participant.



Interactive objects

Posters

Posters are stored as PNGs on ‘poster walls’. If you approach a poster, a preview is displayed at the bottom of your screen. Once you hit the ‘x’ key on your keyboard, you will see the poster in full detail. At the same time, you will connect to other people looking at the same poster in a conference call.

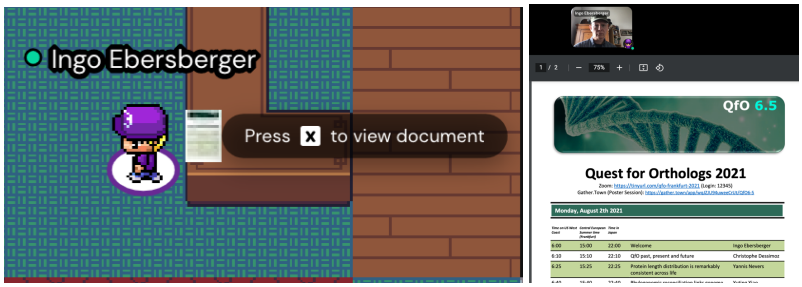


You can use your mouse to move around at the poster. To the right, there is also an icon to activate a ‘laser pointer’ and a zoom button. To leave the poster, just click on the close button in the top right corner.

Documents

You will find the Abstract books distributed across the tables. Approach them and hit the ‘x’ key to look at them.

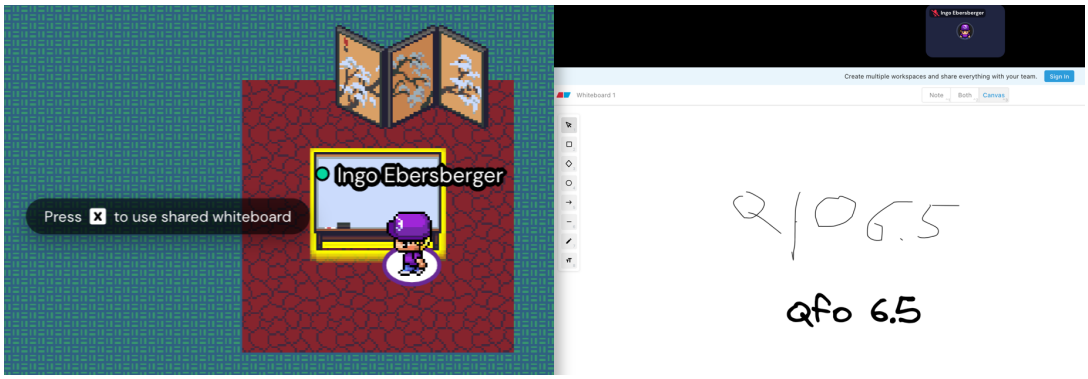
Preview **Full view**



To close the view, click on the ‘x’ in the top right corner

Whiteboard

Sometimes, it is helpful for a discussion to sketch something on a whiteboard. You will find interactive whiteboards distributed in the poster hall and in the two meeting rooms. To activate the whiteboard, hit the ‘x’ key on your keyboard.



The whiteboard will pop up, and all people participating in the discussion can look at it, and also draw & write on it. **Note, like in real life, if you don’t clean the board after using it, your drawings will remain.** To clean: use the arrow to activate objects and then hit the delete button on your keyboard.

Click the ‘x’ in the top right corner to close the whiteboard.

What did I miss…

Probably many things, so feel free to look at the various online resources that explain the use of Gather.Town, and in particular the help pages by the developers.