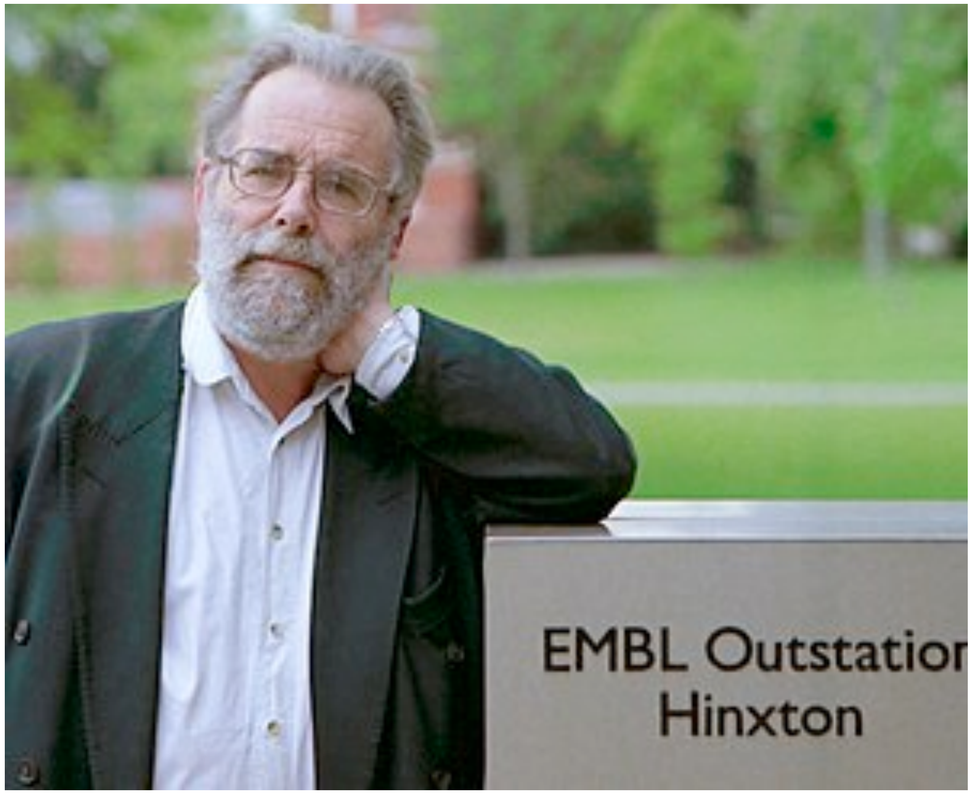
Today’s Stuff

- Go Annotation
- InterProScan
- Domain Based Functional Annotation

Nomenclature



“a scientist would rather use someone else’s toothbrush than another scientist's nomenclature”

Taxonomy and stability of names

- Hey Taxonomists have rules, why not geneticists.
- But the rules of taxonomy were established and followed almost since Linneaus' time.
- And those taxonomists rules have been in effect during the proliferation of species names
- And one of the major tenets of taxonomy is stability; the rules are quite harsh and inquisitionlike.

But geneticists went on

ranking proteins with weird names

1. Mothers against decapentaplegic: It's bizarre, it's unique, it's social satire. A name this special deserves first place.
2. Sonic Hedgehog (SHH): This one is just classic. We've *all* heard about it and we *all* love it.
3. Robotnikinin: some guys somewhere discovered an inhibitor for SHH. They called it robotnikinin. Genius. (Dr. Ivo "Eggman" Robotnik)
4. Cyclins: You might be wondering why I included this. Surely the cyclins were called that because their concentration cycles over time? Nope. The guy who named them just liked cycling a lot.
5. MAP Kinase Kinase Kinase: yeah it's a bit unwieldy (saying MAP3K is cheating) but it describes its function precisely, which is much better than most entries on this list
6. Wee1: it's called that because it's small and it's discoverer was in Scotland at the time. Eh
7. Pikachurin: it's trying too hard. Feels like the discoverers heard about SHH and thought "hey maybe if I do something similar kids will think I'm cool". For shame.

But geneticists went on

Way back in the sixties someone discovered that fruit flies with a mutation in a certain potassium channel became sensitive to ether. It made them shake their legs. That reminded the researchers of a dance that was popular at the Whiskey A Go-Go nightclub in Hollywood. So they termed the mutated gene *Ether-a-go-go*, because why not. This was California in the sixties, so assume that drugs were a factor.

It turns out that in humans there's a homologue of this channel that's like really important for cardiac function. It's called the hERG channel. Which stands for "human ether-a-go-go related gene”.

Stephen Crews, professor of biochemistry at the University of North Carolina at Chapel Hill: "If genes are named in a clever manner, then that probably helps you remember."

NOT! This name tells you nothing. You don't know what kind of protein it is, what it does and to which ion, nothing

- One name many proteins (P53)

- Many names one protein (Aceteylcholine esterase - ACE)
  - ACE1
  - ACE_HUMAN
  - angiotensin converting enzyme, somatic isoform
  - angiotensin I converting enzyme (peptidyl-dipeptidase A) 1
  - angiotensin I converting enzyme peptidyl-dipeptidase A 1 transcript
  - angiotensin-converting enzyme
  - CD143
  - CD143 antigen
  - DCP
  - DCP1
  - dipeptidyl carboxypeptidase 1
  - dipeptidyl carboxypeptidase I
  - EC 3.4.15.1
  - ICH
  - kininase II
  - MVCD

Okay as long as genetics remained a “mom and pop shop” endeavor

But it didnt did it?

Solution?

- Gene Ontology Consortium

The GOC was established in 1998 when researchers studying the genome of three model organisms — *Drosophila melanogaster* (fruit fly), *Mus musculus* (mouse), and *Saccharomyces cerevisiae* (brewer's or baker's yeast) — began to work collaboratively on a common classification scheme for gene function to compare the newly sequenced genomes of these organisms. GO grew into a large data framework adapted for all living organisms, from bacteria to human.

GO was the first of the hundreds of biomedical ontologies that currently exist, which together, aim to represent the vast amount of biomedical knowledge in a computable form. GO is a major hub within these ontologies, being linked to many other biomedical ontologies. It is widely used as a tool in scientific research, and has been cited in tens of thousands of publications.

GO

A core goal of biomedical research is to uncover how individual genes contribute to the biology of an organism, and their roles in health and disease.

The mission of the Gene Ontology Consortium (GOC) is to provide a comprehensive and up-to-date computational model of the current scientific understanding of the functions of gene products, e.g. proteins, non-coding RNAs, macromolecular complexes, or *genes* for simplicity.

GO encompasses all levels of biological systems, from molecular activities to complex cellular and organismal-level networks.

GO provides uniform descriptors applicable to gene products across the entire tree of life. Today, GO is used to represent gene function in all sequenced organisms.

- GO consists of two major things
- 1. The **GO ontology**
- 2. The corpus of **GO annotations**

- The **GO ontology**: the logical structure describing the full complexity of the biology, comprising the ‘classes’ (often referred to as ‘terms’) describing the many different types of molecular functions (Molecular Function), the pathways carrying out different biological programs (Biological Process), and the cellular locations where these occur (Cellular Component). The GO is structured by relating each class to other classes using specific relations.

GO

- The corpus of **GO annotations**: the traceable (i. e., associated with scientific articles), evidence-based statements relating a specific gene product to a specific ontology term. The set of all GO annotations associated with a gene provides a description of its biological role. As of October 2024, the GO includes experimental findings from over 180,000 published papers, representing over 1,000,000 experimentally-supported annotations.

The Ontology

Molecular Function (MF)

Cellular Component (CC)

Biological Process (BP)

Molecular Function (MF)

MF represent molecular-level activities performed by gene products, such as “catalysis” or “transcription regulator activity”.

MFs correspond to activities that can be performed by individual gene products (*i.e.* a protein or RNA), but some activities are performed by molecular complexes composed of multiple gene products, when the activity cannot be ascribed to a single gene product of the complex.

Examples of broad functional terms are *catalytic activity* and *transporter activity*; examples of narrower functional terms are *adenylate cyclase activity* or *insulin receptor activity*.

GO MF terms represent ***activities*** and not the ***entities*** that perform the actions. To avoid confusion between gene product names and their molecular activities, GO MFs are appended with the word “activity” (a *protein kinase* would have the GO MF *protein kinase activity*). Finally, MFs do not specify where, when, or in what context the action takes place.

Cellular Component (CC)

CC serves to capture the cellular location where a molecular function takes place. It includes:

- *cellular anatomical structures*, emcompassing cellular entities such as the *plasma membrane* and the *cytoskeleton*, as well as membrane-enclosed cellular compartments suc
  the *mitochondrion*.

- stable *protein-containing complexes* of which they are parts.

- *virion components*, classified separately because viruses are not cellular organisms. Examples include *viral capsid* and *viral envelope*.

Biological Process (BP)

BPs are the larger processes or ‘biological programs’ accomplished by the concerted action of multiple molecular activities.

Examples of broad BP terms are *DNA repair* or *signal transduction*.

Examples of more specific terms are *cytosine biosynthetic process* or *D-glucose transmembrane transport*.

Structure of Ontology

Each of the three GO aspects is represented by a separate root ontology term.

Moreover, the three GO aspects are *disjoint*, meaning that no relation exists between terms from the different ontology aspects.

However, other relationships such as *part of* and *occurs in* can operate between terms from different GO aspects.

For example, the MF term *cyclin-dependent protein kinase activity* is *part of* the BP *regulation of cell cycle*.

GO

https://amigo.geneontology.org/amigo/search/bioentity?q=Uroplakin%201A

Enrichment analysis tool

Users can perform enrichment analyses directly from the home page of the GOC website. This service connects to the analysis tool from the PANTHER Classification System, which is maintained up to date with GO annotations. The PANTHER classification system is explained in great detail in Mi H et al, PMID: 23868073. The list of supported gene IDs is available from the PANTHER website.

Using the GO enrichment analysis tools

1. Paste or type the names of the genes to be analyzed, one per row or separated by a comma. The tool can handle both MOD specific gene names and UniProt IDs (e.g. Rad54 or P38086).

2. Select the GO aspect (molecular function, biological process, cellular component) for your analysis (biological process is default).

3. Select the species your genes come from (Homo sapiens is default).

4. Press the submit button. Note that you will be able to upload a REFERENCE (aka "background") LIST at a later step.

5. You will be redirected to the results on the PANTHER website. These results are based on enrichment relative the set of all protein-coding genes in the genome you selected in step 3.

6. (optional but HIGHLY RECOMMENDED) Add a custom REFERENCE LIST and re-run the analysis. Press the **"change"** button on the "Reference list" line of the PANTHER analysis summary at the top of the results page, upload the reference list file, and press the **"Launch analysis"** button to re-run the analysis. The reference list should be the list of all the genes from which your smaller analysis list was selected. For example, in a list of differentially expressed genes, the reference list should only contain genes that were detected at all in the experiment, and thus potentially could have been on a list of genes derived from the experiment.

- Say you do a transcriptome on an important tissue in your favorite animal (and can quantitate the level)

- The genes to the right are the major transcripts

- They can also be listed as gene name.

ENSG00000189420
ENSG00000255994
ENSG00000254726
ENSG00000205143
ENSG00000104825
ENSG00000185236
ENSG00000135094
ENSG00000162836
ENSG00000163141
ENSG00000174444
ENSG00000182776
ENSG00000196345
ENSG00000094631
ENSG00000078142
ENSG00000132466
ENSG00000176222
ENSG00000118971
ENSG00000110042
ENSG00000136450
ENSG00000054282
ENSG00000134602
ENSG00000164933
ENSG00000151876
ENSG00000186474
ENSG00000131165
ENSG00000177191
ENSG00000003056
ENSG00000119227
ENSG00000198900
ENSG00000064313

ENSEMBL Gene IDs

Keep original IDs in output?

Include descriptions in output? (slower)

Paste list of ENSEMBL Gene IDs

```
ENSG00000162836
ENSG00000163141
ENSG00000174444
ENSG00000182776
ENSG00000196345
ENSG00000094631
ENSG00000078142
ENSG00000132466
ENSG00000176222
ENSG00000118971
ENSG00000110042
ENSG00000136450
ENSG00000054282
ENSG00000134602
ENSG00000164933
ENSG00000151876
ENSG00000186474
ENSG00000131165
ENSG00000177191
ENSG00000003056
ENSG00000119227
ENSG00000198900
ENSG00000064313
```

Convert IDs

Converted Data

ENSG00000189420	ZFP92
ENSG00000255994	
ENSG00000254726	MEX3A
ENSG00000205143	ARID3C
ENSG00000104825	NFKBIB
ENSG00000185236	RAB11B
ENSG00000135094	SDS
ENSG00000162836	ACP6
ENSG00000163141	BNIPL
ENSG00000174444	RPL4
ENSG00000182776	AC239585.1
ENSG00000196345	ZKSCAN7
ENSG00000094631	HDAC6
ENSG00000078142	PIK3C3
ENSG00000132466	ANKRD17
ENSG00000176222	ZNF404
ENSG00000118971	CCND2
ENSG00000110042	DTX4
ENSG00000136450	SRSF1
ENSG00000054282	SDCCAG8
ENSG00000134602	STK26
ENSG00000164933	SLC25A32
ENSG00000151876	FBXO4
ENSG00000186474	KLK12
ENSG00000131165	CHMP1A
ENSG00000177191	B3GNT8
ENSG00000003056	M6PR
ENSG00000119227	PIGZ
ENSG00000198900	TOP1
ENSG00000064313	TAF2

- Say you do a transcriptome on an important tissue in your favorite animal (and can quantitate the level)

- The genes to the right are the major transcripts

- They can also be listed as gene name.
- https://www.biotools.fr/mouse/ensembl_symbol_converter

ZFP92

MEX3A
ARID3C
NFKBIB
RAB11B
SDS
ACP6
BNIPL
RPL4
AC239585
ZKSCAN7
HDAC6
PIK3C3
ANKRD17
ZNF404
CCND2
DTX4
SRSF1
SDCCAG8
STK26
SLC25A32
FBXO4
KLK12
CHMP1A
B3GNT8
M6PR
PIGZ
TOP1
TAF2

The GO Term Mapper tool maps the granular GO annotations for genes in a list to a set of broader, high-level parent GO slim terms, allowing you to bin your genes into broad categories. This is possible with GO because there are parent:child relationships recorded between granular terms and more general parent GO slim terms. For a better view of parent:children relationships of GO terms, please visit the AmiGO web site.

This GO Term Mapper tool serves a different function than the GO Term Finder tool. The GO Term Mapper tool simply bins the submitted gene list to a static set of ancestor GO terms. In contrast, the GO Term Finder tool finds the GO terms significantly enriched in a submitted list of genes.

Enter List of Genes

Either type the names of the genes, one per line, in the input box or upload a file that contains the gene names.
The Gene Association File Table below lists the types of identifiers in the gene association files that the GO Term Mapper program accepts as input for the list of gene names. It also provides links for tools that may help you to convert from an unsupported identifier system to one that is supported.

For example, if you have a file of gene identifiers that are a different type than those listed in the table, you can use a tool provided by UniProt that converts several types of identifiers into UniProt_IDs/Accession. You can then use the GOA gene association files that contain UniProt_IDs/Accession gene names.

Choose 1 of the 3 Ontology Aspects

Select one of the three (*biological process*, *molecular function*, or *cellular component*) ontologies by checking the appropriate radio button. The GO Term Mapper program searches ONLY 1 of the 3 ontologies at a given time.

Choose Slim Gene Association File (GO Slim)

Select a slim gene association file and associated slim ontology from the pop-up menu. If you are providing your own gene association file in the **Advanced Options**, the ontology chosen here will be used. For example, if you provide your own gene association file for S. cerevisiae, and select "SGD (Yeast - S. cerevisiae GO slim)", the yeast ontology (goslim_yeast.obo) will be used. If you select "SGD (Yeast - Generic GO slim)", the generic ontology (goslim_generic.obo) will be used. If you are providing your own ontology as well, then your selection here does not matter.
Please see Gene Association File Table for the full description of the gene association files.

The generic slim ontology contains terms that are considered generally interesting across all annotations, whereas the alternative slim ontologies are provided by individual organizations and may contain (or exclude) certain terms that are considered by them to be of particular interest (or not particularly of interest) for their specific annotation.

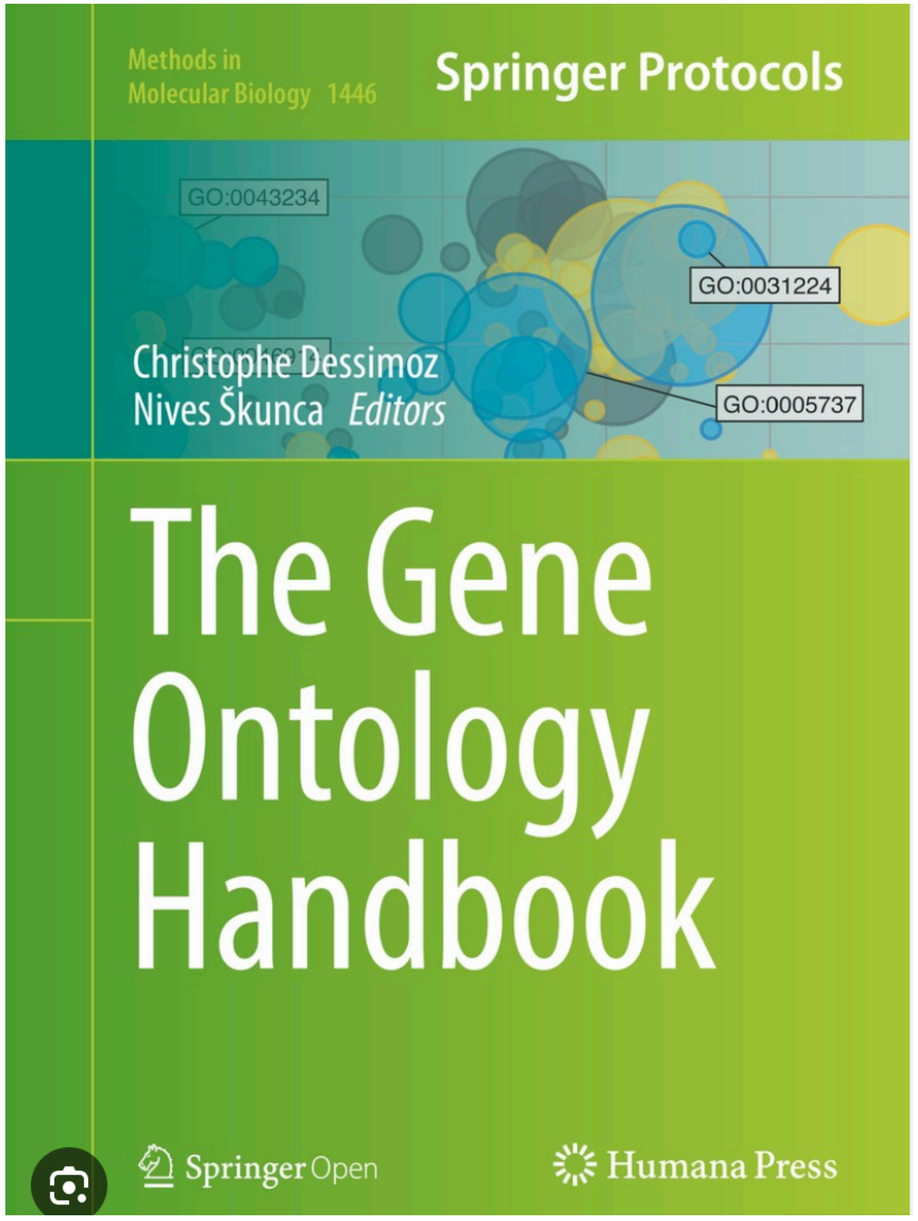
The different slim ontologies that are used include:

- Generic GO slim
- Yeast - S. cerevisiae GO slim
- S. pombe GO slim (for Biological Process only)
- GOA GO slim

Slim gene association files are created by using map2slim.pl script (part of GO::Perl). Each slim gene association file is generated using the full ontology, the slim ontology, and the annotation file.

Here is the list of go terms obtained for your gene
Transcript list.

GO:0005488
GO:0005515
GO:0003824
GO:0003676
GO:0016740
GO:0016787
GO:0016301
GO:0016853
GO:0015075
GO:0030234
GO:0016829
GO:0005198
GO:0005215
GO:0008565
GO:0003774
GO:0030528
GO:0004386
GO:0016874
GO:0016209
GO:0045182
GO:0004871
GO:0015267
GO:0016491
GO:0009055
GO:0004872
GO:0008907


Methods in
Molecular Biology 1446
Springer Protocols
GO:0043234
GO:0031224
GO:0005737
Christophe Dessimoz
Nives Škunca Editors
The Gene
Ontology
Handbook
Springer Open
Humana Press

The mission of the PANTHER knowledgebase is to support biomedical and other research by providing **comprehensive information about the evolution of protein-coding gene families**, particularly protein phylogeny, function and genetic variation impacting that function. Learn more

https://pantherdb.org/

The PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System was designed to classify proteins (and their genes) in order to facilitate high-throughput analysis. The core of PANTHER is a comprehensive, annotated "library" of gene family phylogenetic trees. All nodes in the tree have persistent identifiers that are maintained between versions of PANTHER, providing a stable substrate for annotations of protein properties like subfamily and function. Each phylogenetic tree is used to annotate each protein member of the family by its:

1. Family and Protein Class (supergrouping of protein families)
1. Subfamily (subgroup within the family phylogenetic tree)
1. Orthologs (genes in other organisms that derive from the same gene in the MRCA)
1. Paralogs (genes in the same organism that are related by gene duplication)
1. Function (using GO terms annotated on the trees by the GO Phylogenetic Annotation Project)
1. Pathways (curated by PANTHER and by Reactome)
2.

The mission of the PANTHER knowledgebase is to support biomedical and other research by providing **comprehensive information about the evolution of protein-coding gene families**, particularly protein phylogeny, function and genetic variation impacting that function. Learn more

https://pantherdb.org/

The protein-coding gene classification information can be searched on the PANTHER website, or downloaded for genes from the 144 genomes included in the PANTHER trees. For classifications of genes not in the trees, we recommend downloading the TreeGrafter tool, and running it on your own computer.

The PANTHER Classifications are the result of human curation as well as sophisticated bioinformatics algorithms. Details of the methods can be found in (Mi et al. NAR 2013; Thomas et al., Protein Science 2022).

1.

PANTHER Classification System

The mission of the PANTHER knowledgebase is to support biomedical and other research by providing **comprehensive information about the evolution of protein-coding gene families**, particularly protein phylogeny, function and genetic variation impacting that function. Learn more

PANTHER19.0 Released. Click for more details.

search keyword | All | Go

Home | About | Data Version | Tools | API/Services | Publications | Workspace | Downloads | FAQ/Help/Tutorial | Login | Register | Contact us

Current Release: PANTHER 19.0 | 15,683 family phylogenetic trees | 144 species | News
Whole genome function views

PANTHER™ website news

June 20, 2024

PANTHER19.0 Released.

- PANTHER19.0 is generated from the 2023_03 and 2023_05 release of ReferenceProteome dataset. Here is the composition of all genomes.
  - 144 total genomes
  - 35 bacteria
  - 8 archaea
  - 15 fungus
  - 40 plants
  - 8 protista and alveolata
  - 3 amoebazoa
  - 15 invertebrate
  - 20 vertebrate
  - 2,692,827 total genes
- 2,017,190 genes in PANTHER™ families with phylogenetic trees, multiple sequence alignments and HMMs
  - 15,683 PANTHER™ families
  - 128,012 subfamilies
  - 177 pathways
  - 3092 pathway components
  - 51028 sequences associated to pathways
  - 5996 references captured for the pathways
- PANTHER19.0 is indexed by PANTHER GO slim and an updated PANTHER Protein Class. PANTHER GO slim is based on Gene Ontology phylogenetic annotations to over 8000 PANTHER™ families. The GO slim ontology contains:
  - 3348 total terms
  - 2200 biological process terms
  - 537 cellular component terms
  - 611 molecular function terms
- PANTHER™ Protein Class contains a total of 210 terms.

Last release of PANTHER 18.0 can be accessed via https://archive.pantherdb.org until September 16th, 2025

The mission of the PANTHER knowledgebase is to support biomedical and other research by providing **comprehensive information about the evolution of protein-coding gene families**, particularly protein phylogeny, function and genetic variation impacting that function. Learn more

PANTHER19.0 Released. Click for more details.

search keyword | All | Go

Home | About | Data Version | Tools | API/Services | Publications | Workspace | Downloads | FAQ/Help/Tutorial | Login | Register | Contact us

Current Release: PANTHER 19.0 | 15,683 family phylogenetic trees | 144 species | News
Whole genome function views

RESEARCH TOOLS

Score proteins against the PANTHER HMM library and download PANTHER tools and data.

Gene List Analysis
Analyze gene lists, and expression data files with PANTHER. Map lists to multiple annotation data sources from PANTHER and the Gene Ontology Consortium, as well as biological pathways. Overlay your results on pathway diagrams to visualize the relationships between genes/proteins in known pathways.

Evolutionary analysis of coding SNPs
Estimates the likelihood that a particular nonsynonymous coding SNP will cause a functional impact on the protein, as described in Tang H & Thomas PD, 2016.

PANTHER Grafting or PANTHER scoring
Graft proteins against the PANTHER library of trees or score proteins against the entire PANTHER library of over 38,000 HMMs to obtain PANTHER classifications and alignments.

PANTHER® Evolutionary Species Tree Explorer
View genomes supported by PANTHER and also generate custom list of families and taxonomies to be used with the PANTHER API

About | Release Information | Contact Us | System Requirements | Privacy Policy | Disclaimer


GLOBAL CORE BIODATA RESOURCE

https://pantherdb.org/tools/index.jsp

I am studying this protein and its functional evolution in aligators.

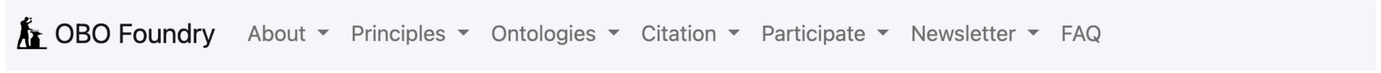
```
MKVLWAALLVTFLAGCQAKVEQAVETEPEPELRQQ
TEWQSGQRWELALGRFWDYLRWVQTLSEQVQEELL
SSQVTQELRALMDETMKELKAYKSELEEQLTPVAE
ETRARLSKELQAAQARLGADMEDVCGRLVQYRGEV
QAMLGQSTEELRVRLASHLRKLRKRLLRDADDLQK
RLAVYQAGAREGAERGLSAIRERLGPLVEQGRVRA
ATVGSLAGQPLQERAQAWGERLRARMEEMGSRTRD
RLDEVKEQVAEVRAKLEEQAQQIRLQAEAFQARLK
SWFEPLVEDMQRQWAGLVEKVQAAVGTSAAPVPSD
NH
```

I had enough funds to sequence 8 individuals genomes' and
found the following 4 SNPs in the aligator genomes I sequenced that produce a change in amino acid.

L46P
C130R
R163C
R176C

Do I have anything good going on here??
https://pantherdb.org/tools/csnpScoreForm.jsp?

OBO FOUNDRY https://obofoundry.org/



Open Biological and Biomedical Ontology Foundry

Community development of interoperable ontologies for the biological sciences

Learn about OBO best practices and community resources

- OBO Foundry principles
- OBO tutorial
- Ontology browsers, tutorials, and tools

Participate

- Code of Conduct
- Join the OBO mailing list and the OBO Community Slack workspace
- OBO Foundry Operations and Working Groups
- Submit bug reports or suggestions for improvement via GitHub
- Submit your ontology to be considered for inclusion in the OBO Foundry

OBO Library: find, use, and contribute to community ontologies

Download table as: [YAML | JSON-LD | RDF/Turtle]

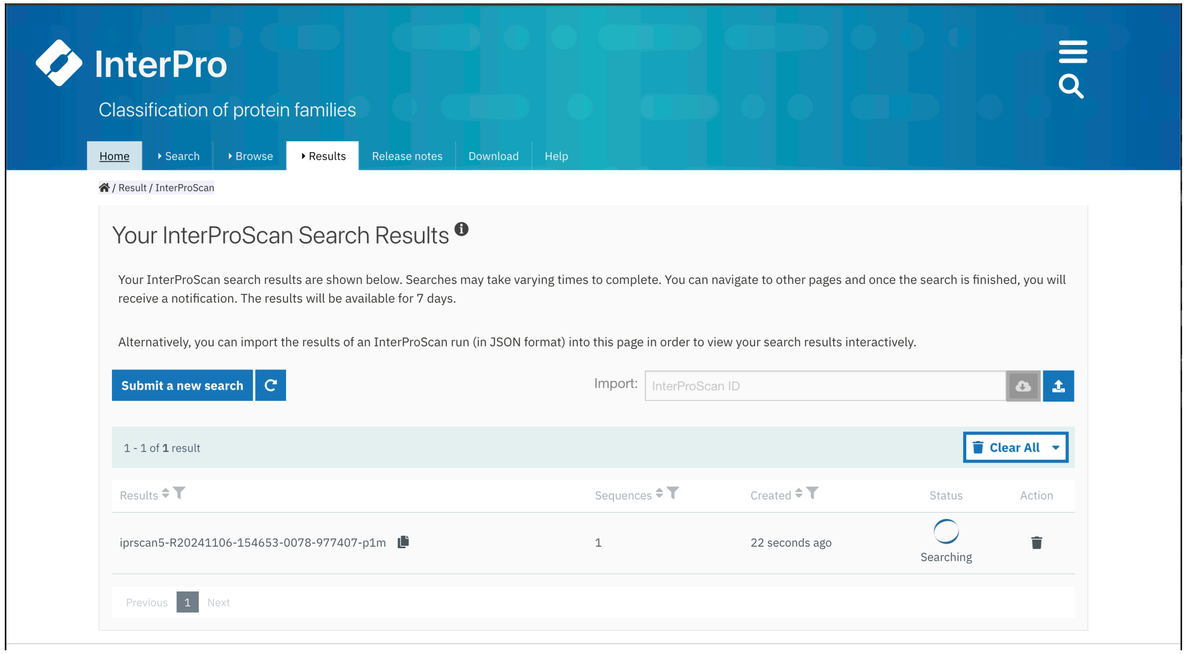
OBO FOUNDRY **https://obofoundry.org/**

Other **Ontology Browsers**

- OBO Foundry
- BioPortal: has over 700 biomedical ontologies, including the OBO Foundry ontologies. Users can search for terms or browse individual ontologies. BioPortal also provides tools to aid in increasing ontology interoperability.
- Ontology Lookup Service (OLS): lets users search for terms across all ontologies, or individual ontologies, as well as view term hierarchies. OLS also has an API so that ontologies can be browsed programmatically.
- Ontobee: Part of the Onto-Animal tool collection. Users can search all available ontologies and browse detailed descriptions of individiual terms. OntoBee is the default server for most OBO Foundry ontology IRIs.
- AberOWL: A repository of biological ontologies and access to reasoned versions of those ontologies. Ontolgies can be browsed by term labels or queried by Manchester OWL syntax statements.
- QuickGO
- AmiGO
- Linked Open Vocabularies

InterProScan

https://www.ebi.ac.uk/interpro/search/sequence/



InterProScan

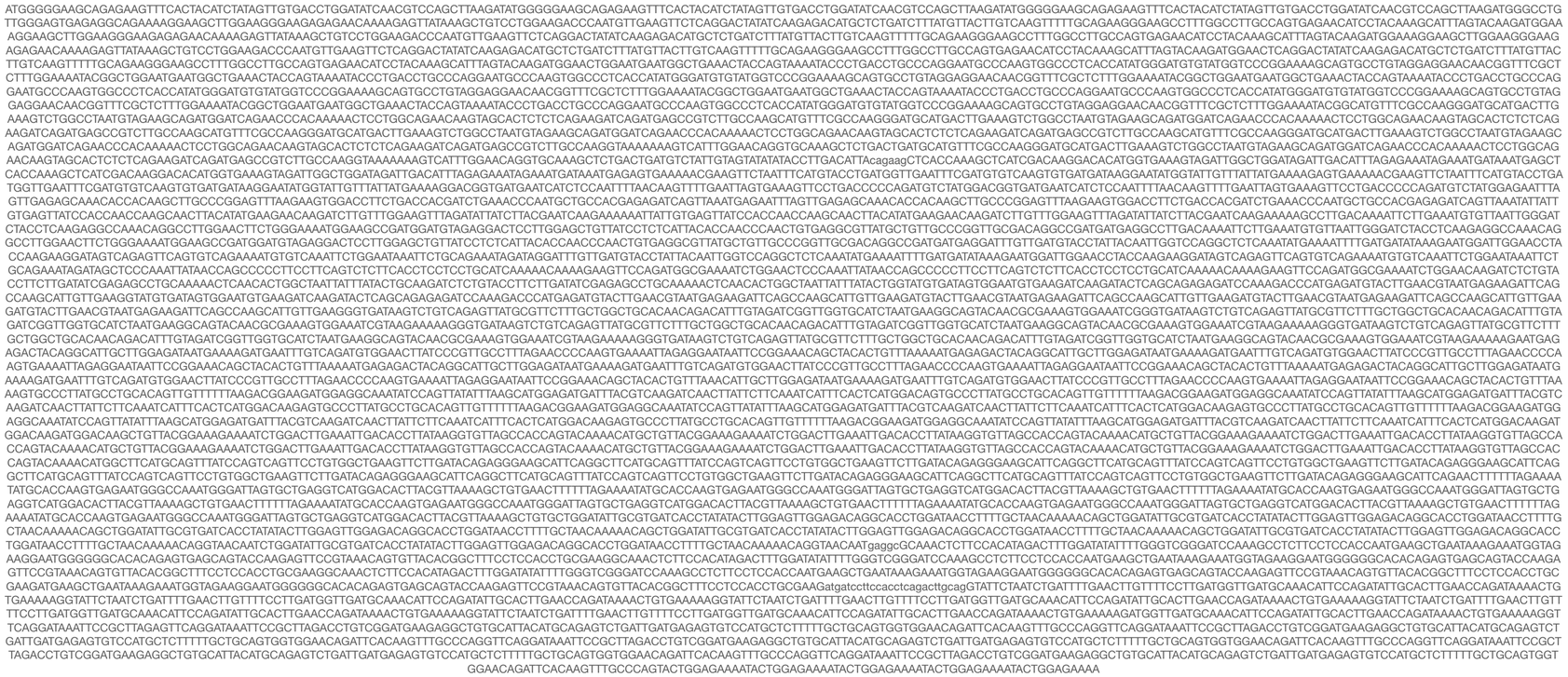
Classification of protein families

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites.

To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium.

By combining protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.

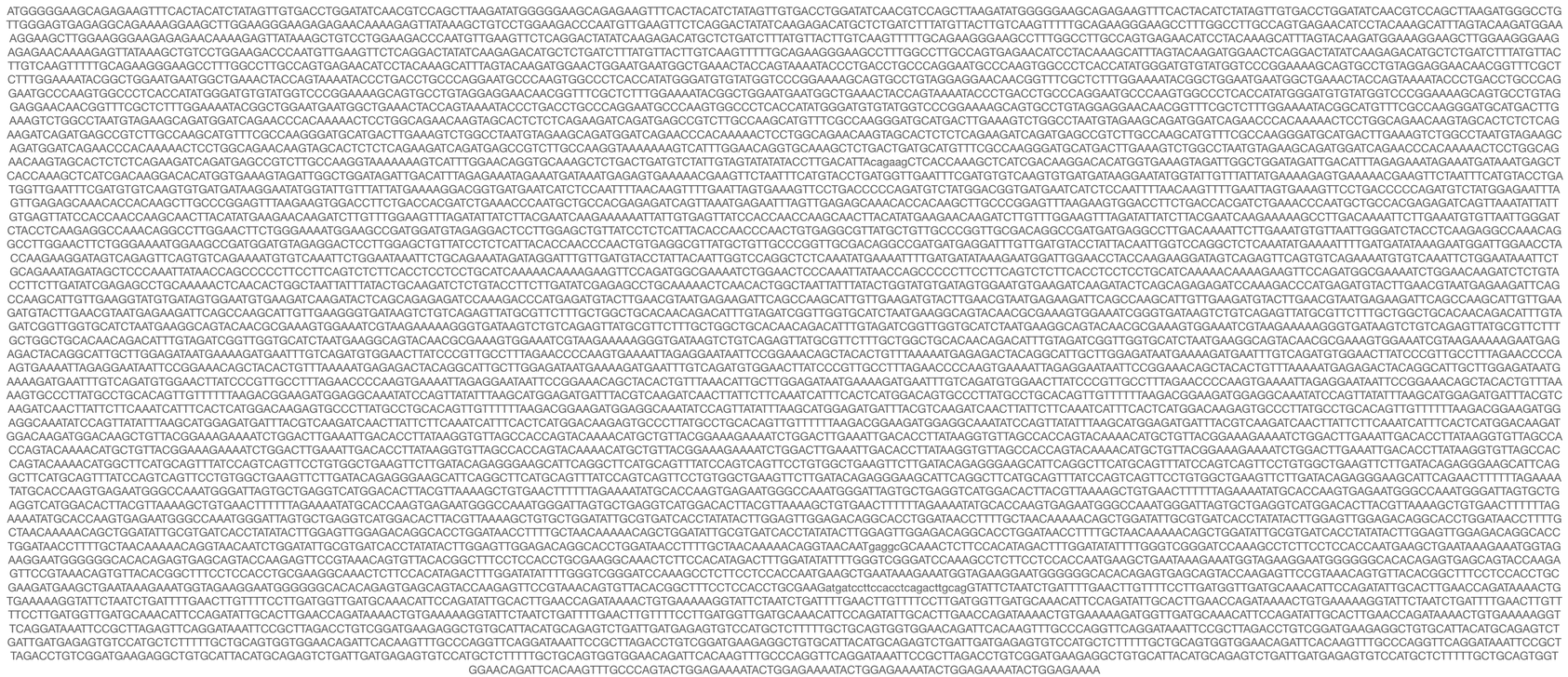
Here is a pretty big protein gene I found.



https://www.ebi.ac.uk/interpro/search/sequence/

Here is a pretty big protein gene I found.

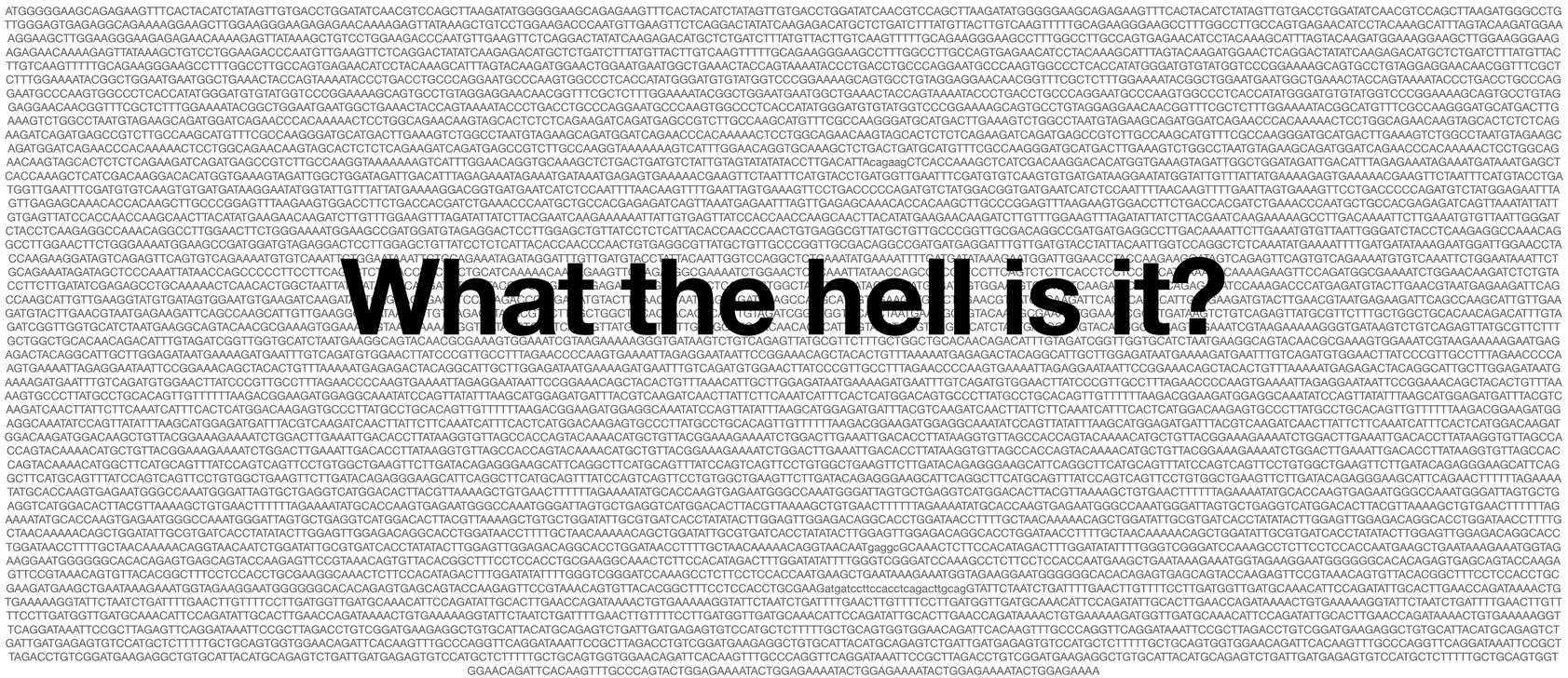
I BLASTED it and got no good hits



https://www.ebi.ac.uk/interpro/search/sequence/

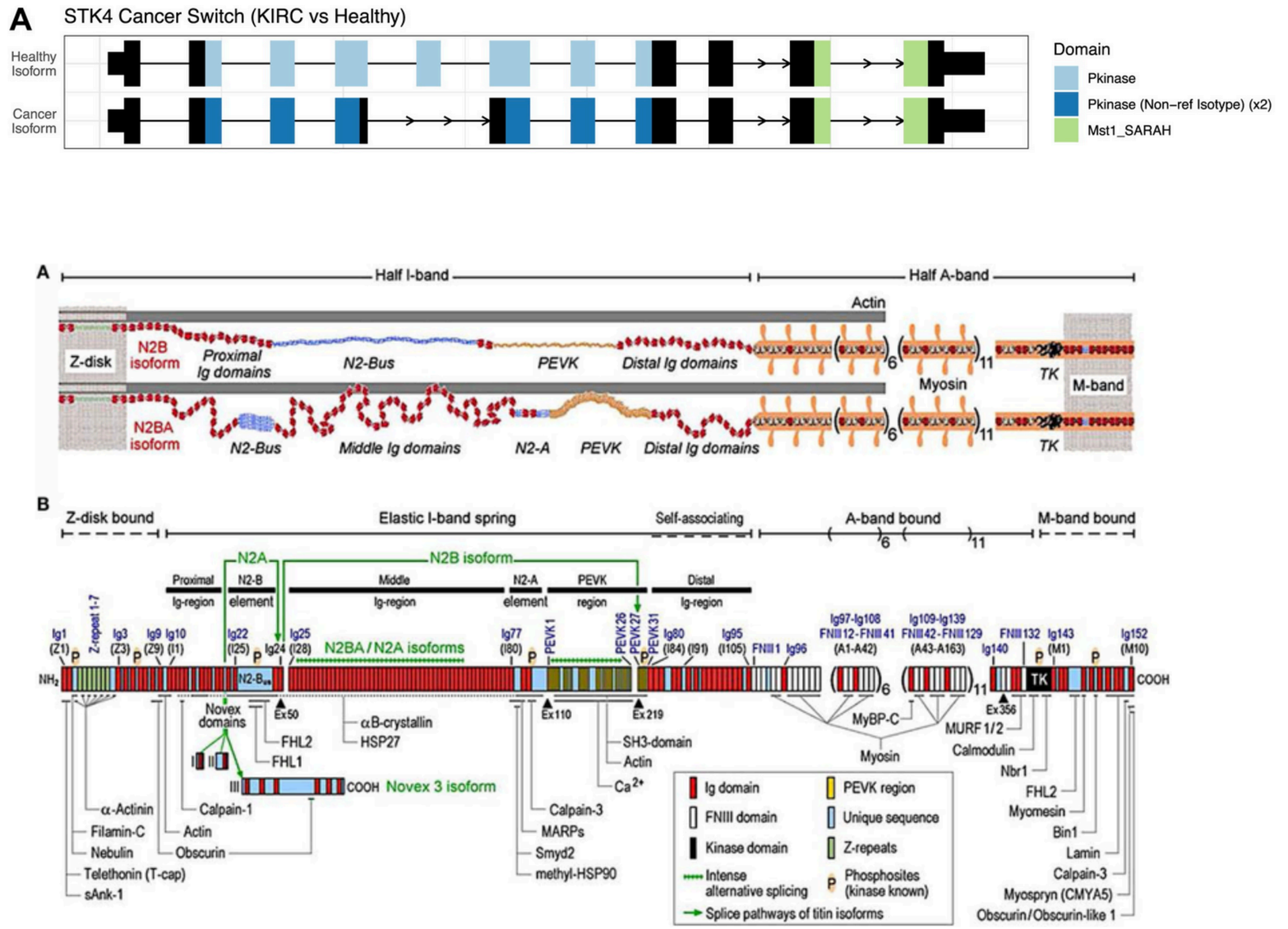
Here is a pretty big protein gene I found.

I BLASTED it and got no good hits



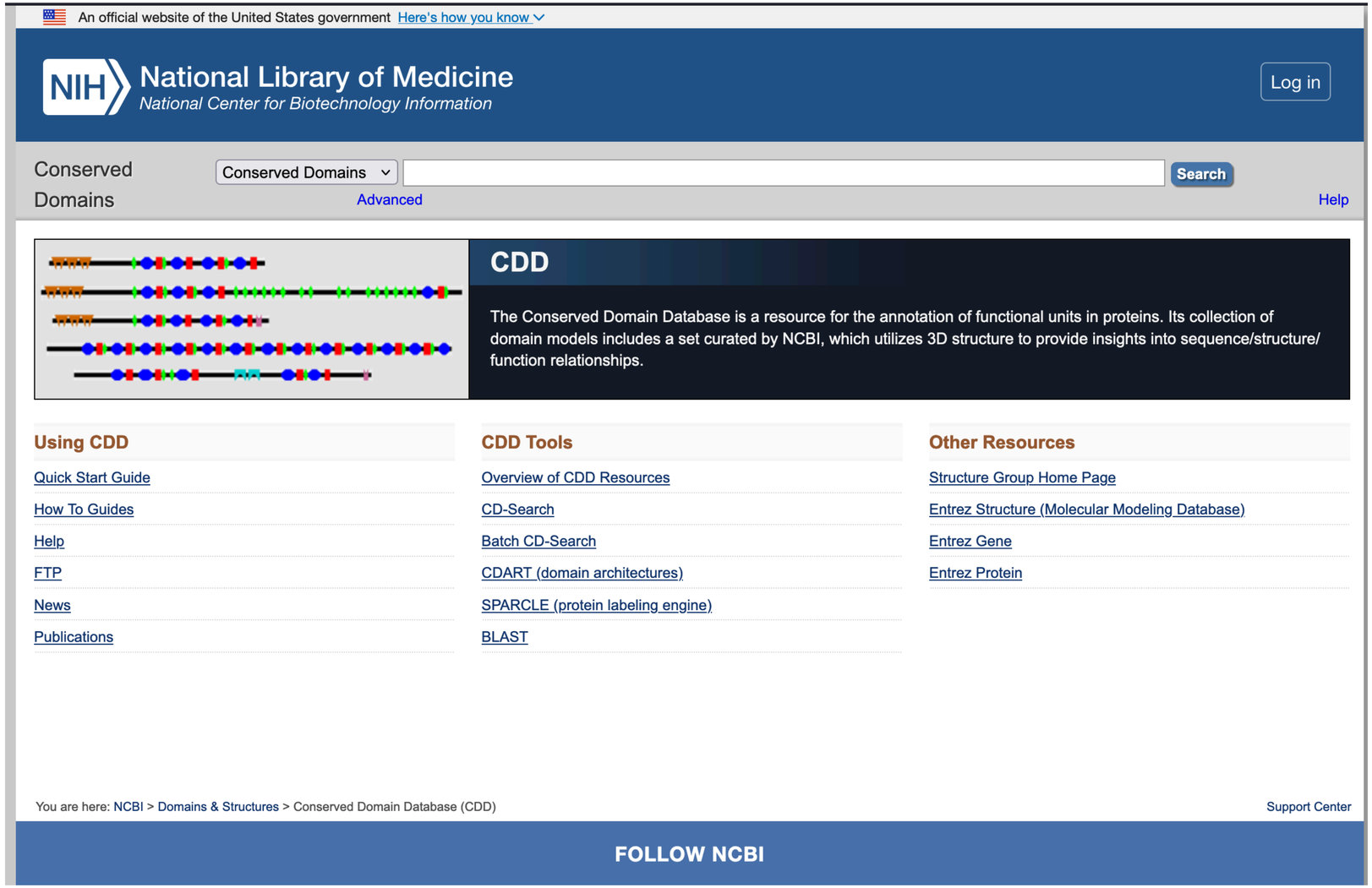
What the hell is it?

https://www.ebi.ac.uk/interpro/search/sequence/



Domain structure of titin isoforms and binding sites of titin ligands. (A) N2B and N2BA titin isoforms represented in the cardiac half-sarcomere, (B) Domain structure of titin sequence, Q8WZ42-1, with ligand binding sites represented (from Linke and Hamdani, with permission) (60).

- Domain Based Functional Annotation



MyCLADE
An online server that uses a large dataset of probabilistic models to annotate domains. It can find domains for proteins that are annotated for the first time, and can enrich known architectures with new domains.

GO FEAT
A free, online tool that can annotate and enrich genomic and transcriptomic data. It integrates with several databases, including UniProt, InterPro, KEGG, Pfam, and NCBI.

DAVID
A bioinformatics resource system that provides tools for functional annotation, functional enrichment analysis, and ID conversion of gene lists. It can identify enriched biological themes, discover enriched functional-related gene groups, and more.

eggNOG-mapper
Provides protein domain predictions, including PFAM and SMART. It can annotate functional terms per query, and can transfer PFAM domain annotations from inferred orthologs.

Domainator
A flexible and modular tool developed on Linux, and tested on Mac. It probably won't work on Windows, except through WSL.

MyCLADE: an accurate multi-source domain annotation server designed for a fast exploration of genomic and metagenomic sets of sequences

The understanding of the ever-increasing number of genomic and metagenomic sequences accumulating in our databases demands for approaches that rapidly "explore" the content of sets of (fragmented) protein sequences with respect to specific domain targets, avoiding full domain annotation and full assembly. MyCLADE performs a multi-source domain annotation strategy based on a library of probabilistic domain models associated to each domain. It works in two modes:

1. It explores large datasets of a few thousands amino-acid sequences and extracts those sequences containing few targeted domains.
2. It annotates small datasets of a few hundreds amino-acid sequences with the full set of Pfam domains.

If sequences are sufficiently long, DAMA can be used to accurately resolve protein domain architectures.

MyCLADE annotates protein sequences with a library of more than 2.5 million probabilistic models for the whole Pfam32 database (17,929 domains). For this reason, keep in mind that a domain annotation of a single protein sequence could take the same time as for 50-100 sequences.

Input

MyCLADE can be run on three different library types defined by:

- up to 10 domains chosen by the user
- all Pfam domains
- domains in a clan

The first and third library types require a list of up to 2000 sequences in FASTA format (possibly uploaded). The second library type requires a smaller dataset of up to 200 sequences.

The list of input sequences is checked for format requirements. Several **error messages** suggest the user how to correct the FASTA file given as input:

- *name* should start with a '>'
- Your sequence *name* misses its sequence
- Your sequence *name* contains characters that are not amino acids
- There are more than *max_seq* sequences in your input data
- The sequence *n°seq_nb* should have an ID starting with a ">"
- You have a trailing whitespace before your sequence *n°seq_nb*

Several **parameter values** can be chosen:

Model library: each library type is characterized by either 350 or 50 models per domain in the set. The option allows the user to decide on the number of models per domain.

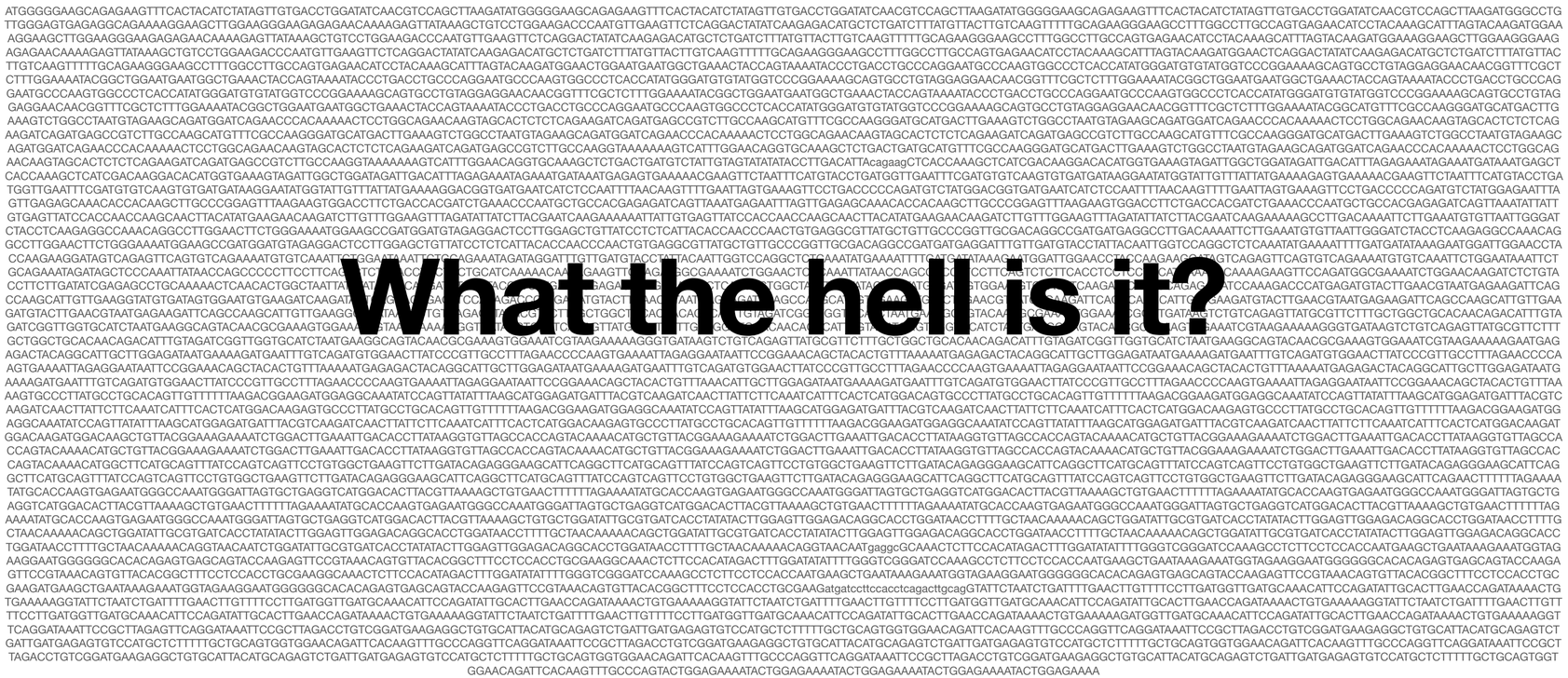
E-value threshold: 1e-3 (default value). The user can filter out all hits with an E-value which is greater than the chosen threshold

The reconstruction of **the best domain architecture** is possible by selecting DAMA together with its three parameters:

- DAMA E-value: 1e-10 (default value)
- number of amino acids allowed in domain overlapping: ≤ 30aa (default value)
- domain matches must cover at least 50% (default value) of the domain average size

Here is a pretty big protein gene I found.

I BLASTED it and got no good hits



What the hell is it?

https://www.ebi.ac.uk/interpro/search/sequence/

Using Phylogenomic Patterns and Gene Ontology to Identify Proteins of Importance in Plant Evolution

Angélica Cibrián-Jaramillo*,†,[1,2], Jose E. De la Torre-Bárcena†,[3], Ernest K. Lee[1], Manpreet S. Katari[3], Damon P. Little[2], Dennis W. Stevenson[2], Rob Martienssen[4], Gloria M. Coruzzi[3], and Rob DeSalle[1]

[1]Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, New York

[2]Molecular Systematics, The New York Botanical Garden, Bronx, New York

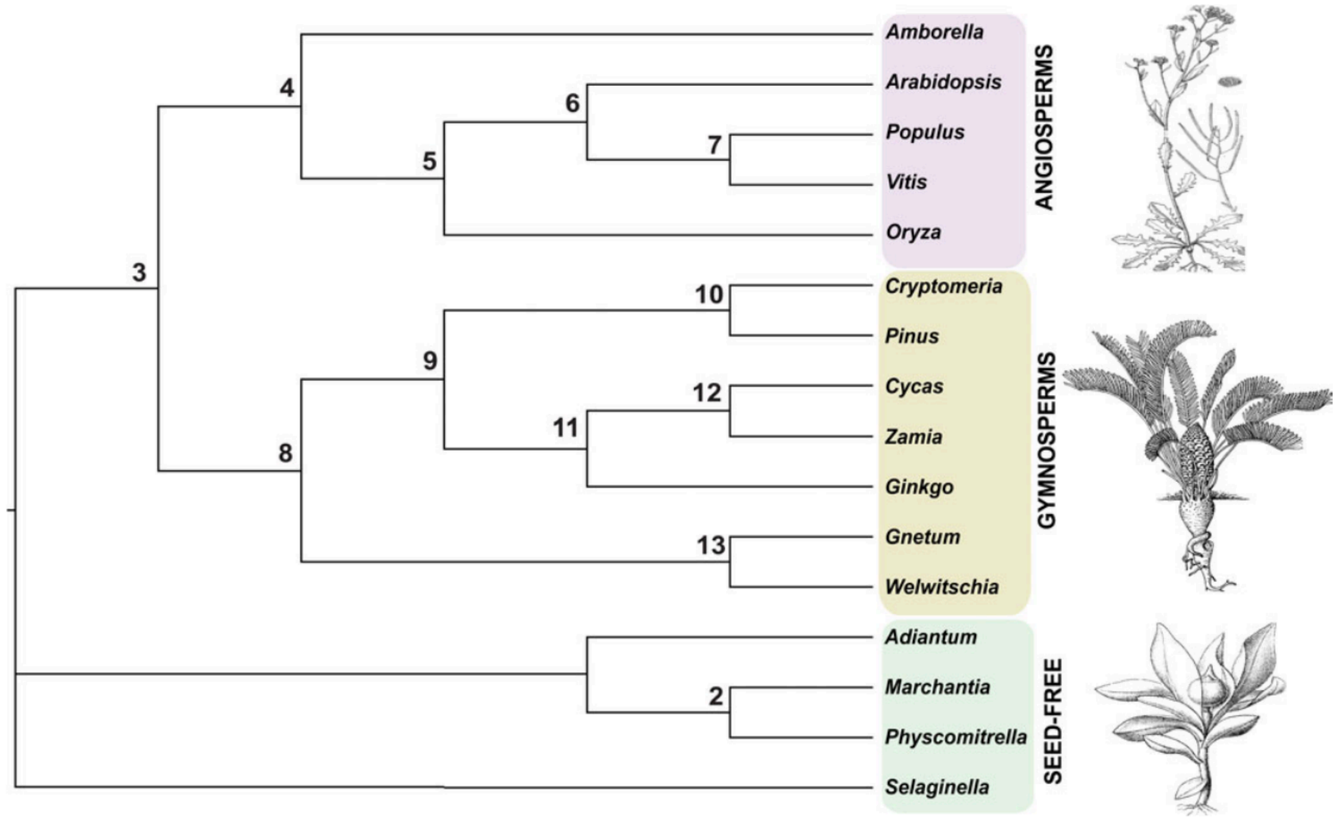
[3]Center for Genomics and Systems Biology, Department of Biology, New York University

[4]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York

*Corresponding author: E-mail: acibrian@amnh.org.

†These authors contributed equally to this work.

Accepted: 14 March 2010



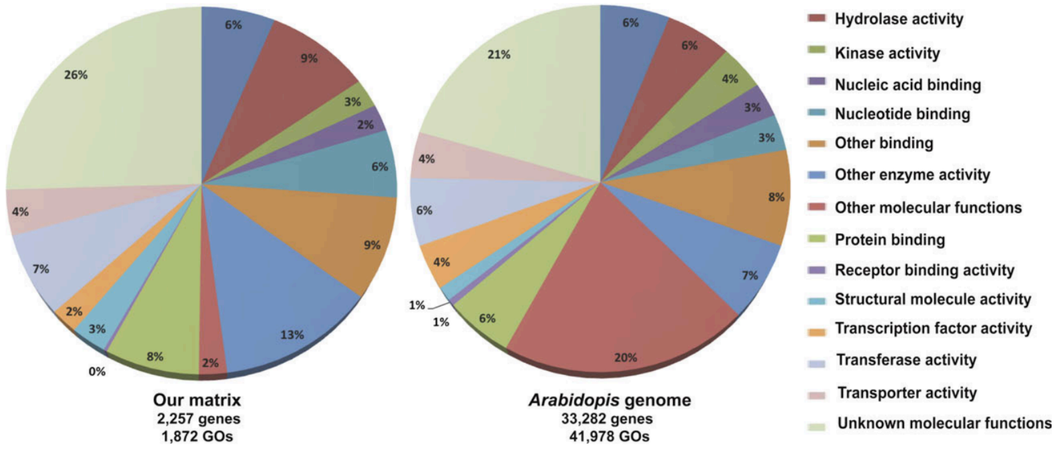
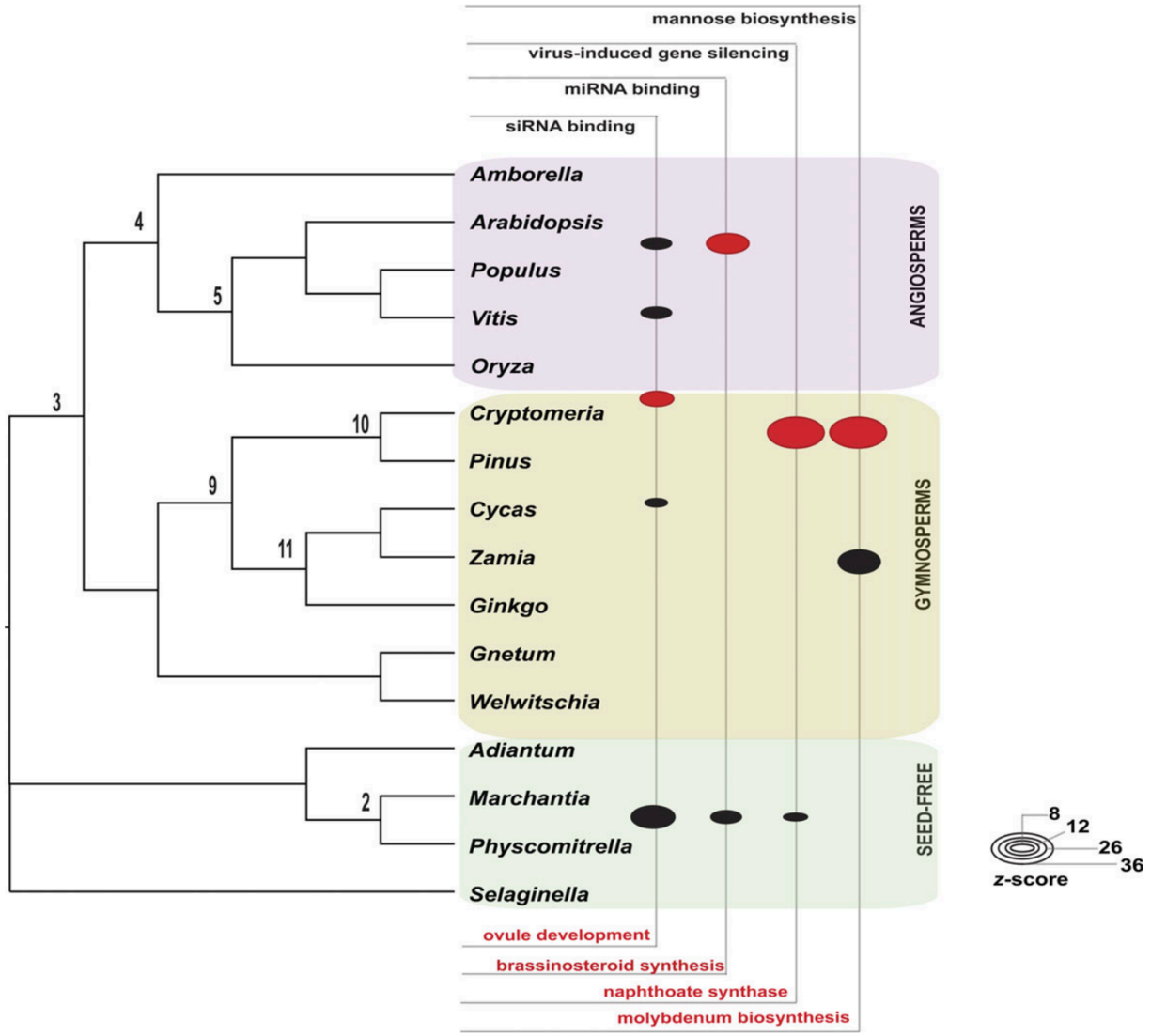


Fig. 2.—Distribution of GO molecular function categories.


mannose biosynthesis
virus-induced gene silencing
miRNA binding
siRNA binding
Amborella
Arabidopsis
Populus
Vitis
Oryza
Cryptomeria
Pinus
Cycas
Zamia
Ginkgo
Gnetum
Welwitschia
Adiantum
Marchantia
Physcomitrella
Selaginella
ANGIOSPERMS
GYMNOSPERMS
SEED-FREE
4
5
3
10
9
11
2
ovule development
brassinosteroid synthesis
naphthoate synthase
molybdenum biosynthesis
8
12
26
36
z-score

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04136-5 OPEN

Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty

Jordi Paps[1,2] & Peter W.H. Holland[2]

Understanding the emergence of the Animal Kingdom is one of the major challenges of modern evolutionary biology. Many genomic changes took place along the evolutionary lineage that gave rise to the Metazoa. Recent research has revealed the role that co-option of old genes played during this transition, but the contribution of genomic novelty has not been fully assessed. Here, using extensive genome comparisons between metazoans and multiple outgroups, we infer the minimal protein-coding genome of the first animal, in addition to other eukaryotic ancestors, and estimate the proportion of novelties in these ancient genomes. Contrary to the prevailing view, this uncovers an unprecedented increase in the extent of genomic novelty during the origin of metazoans, and identifies 25 groups of metazoan-specific genes that are essential across the Animal Kingdom. We argue that internal genomic changes were as important as external factors in the emergence of animals.

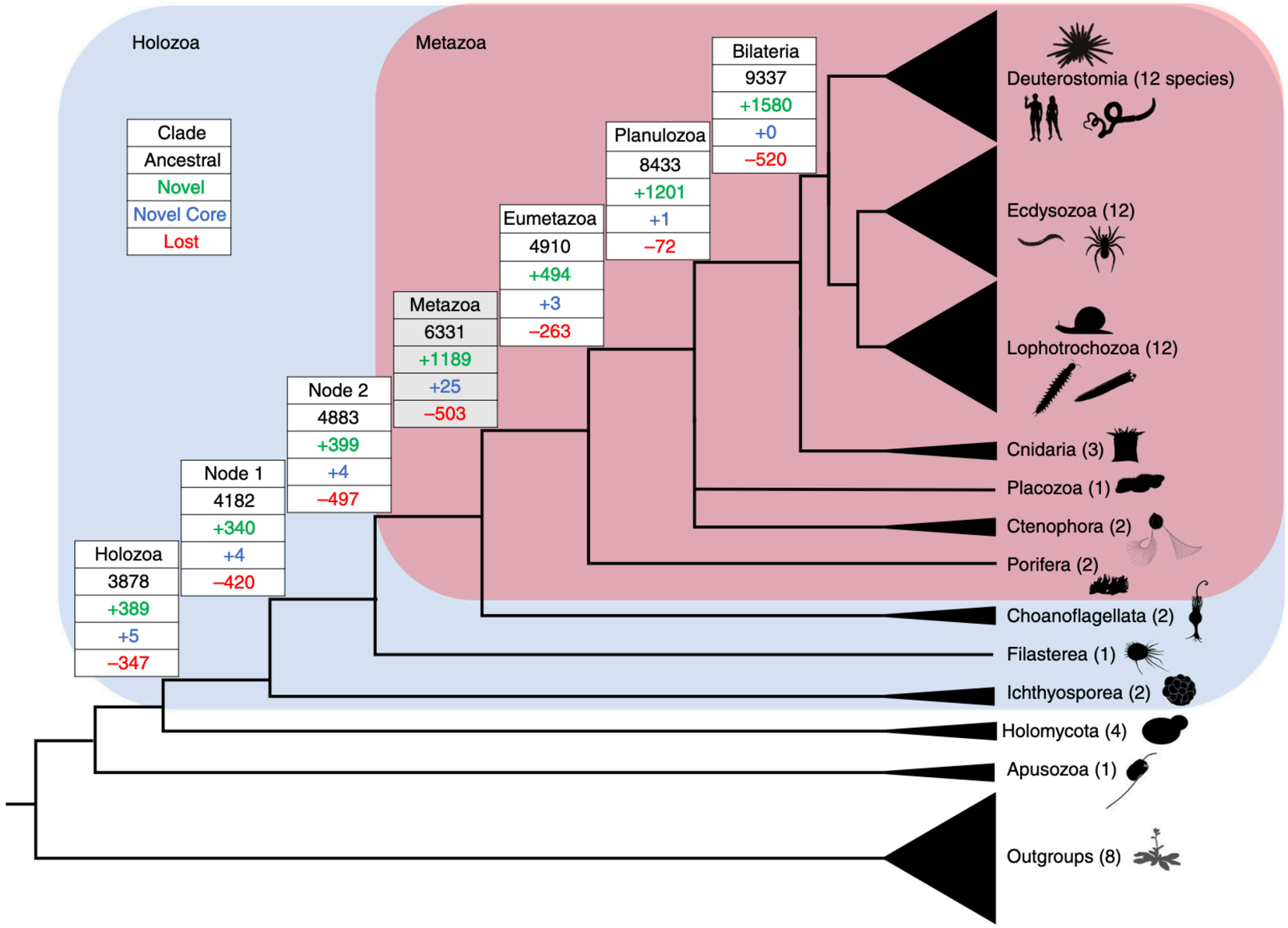
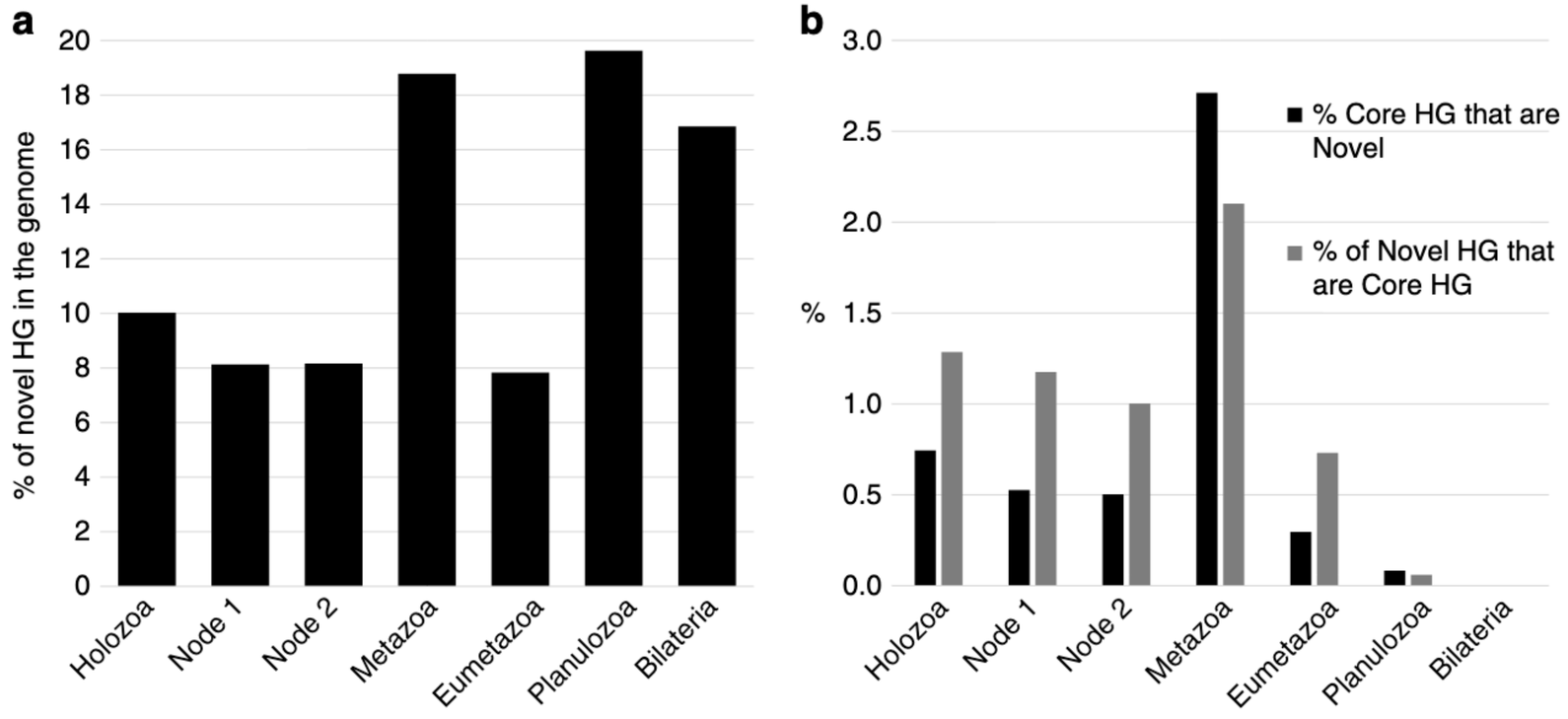
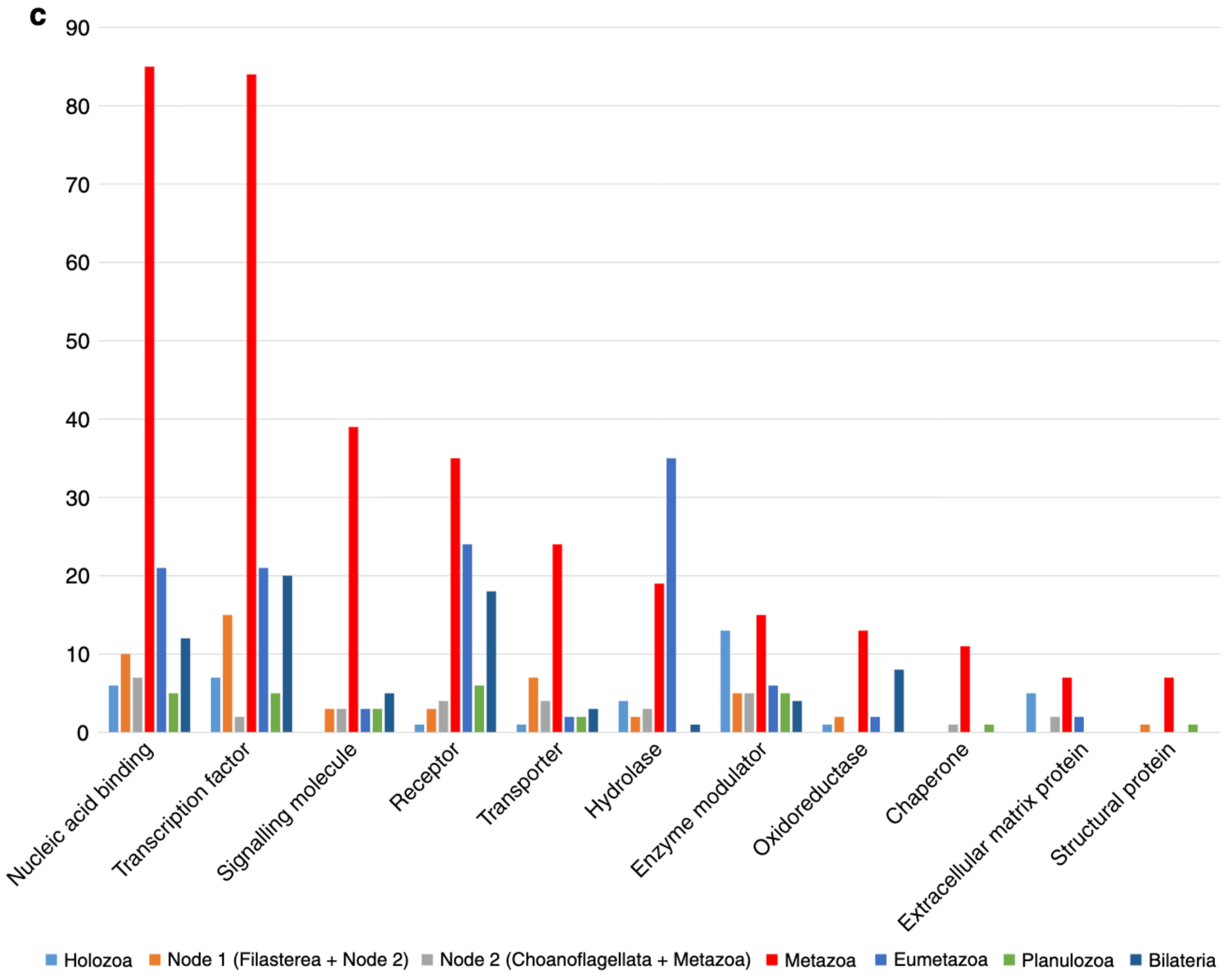
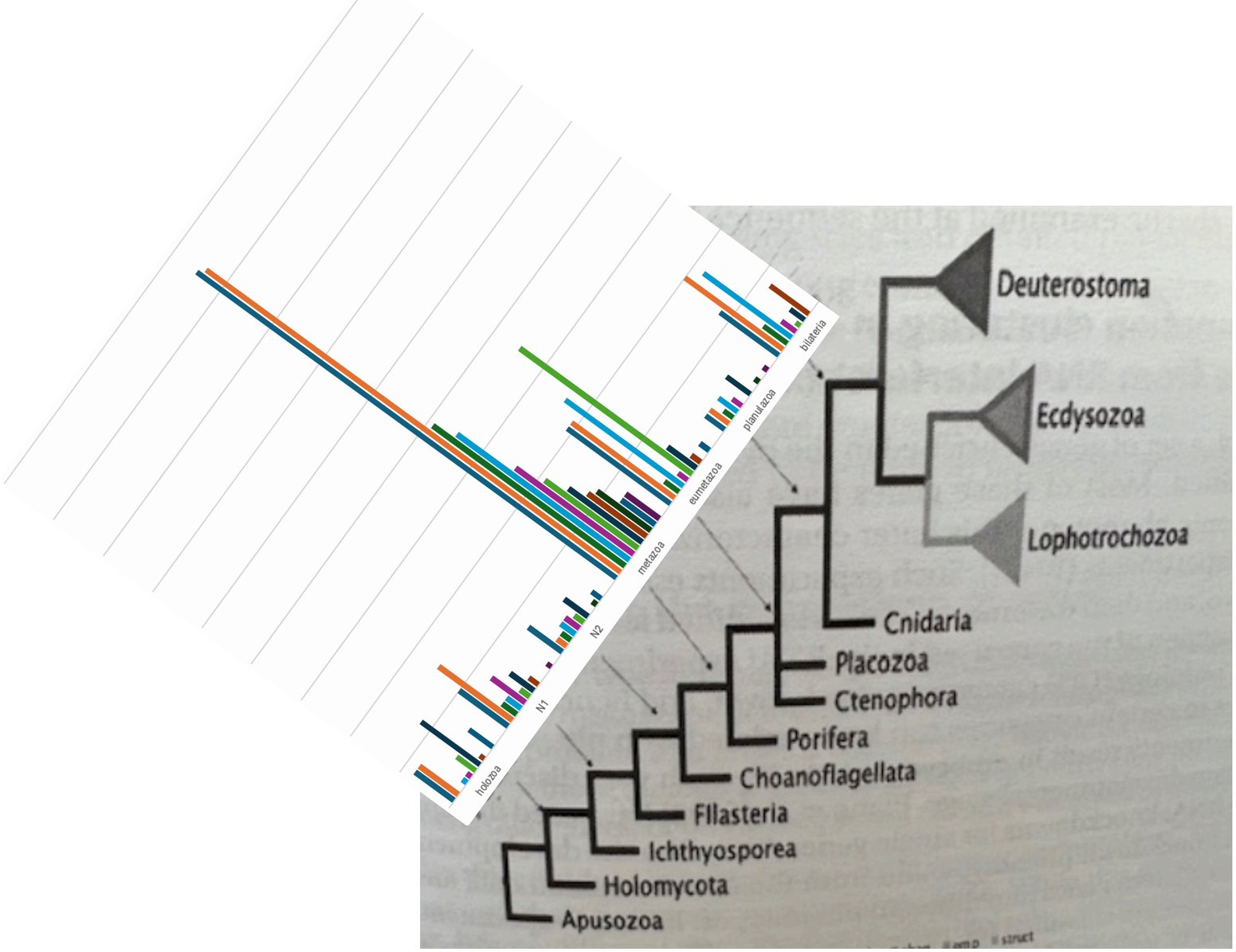
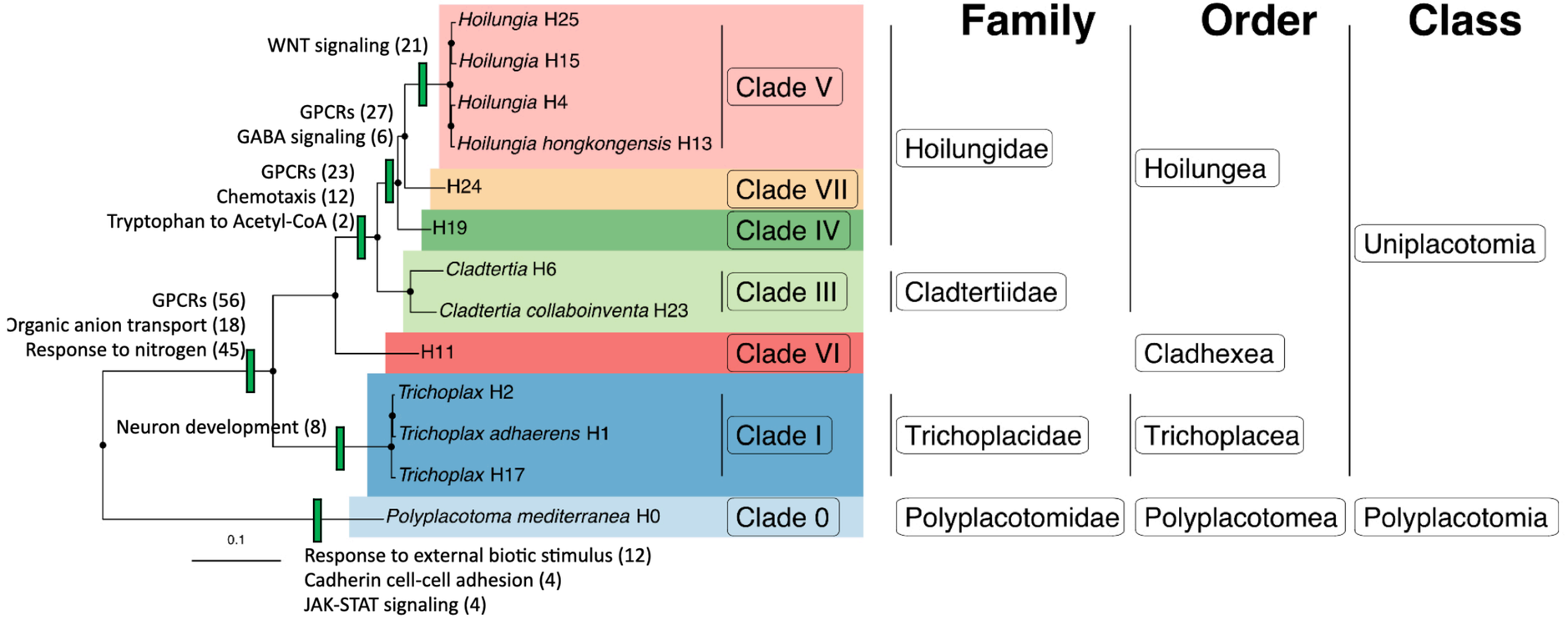


Fig. 1 Reconstruction of ancestral genomes. Evolutionary relationships of the major groups included in his study[2]. Different categories of HG are indicated in each node, from top to bottom, Ancestral HG, Novel HG, Novel Core HG, and Lost HG. Values assume sponges as the sister group to other animals, and placozoans as sister group to Planulozoa (=Cnidaria + Bilateria); alternative phylogenetic hypotheses are explored in Supplementary Data 3-8. Organism outlines from phylopic.org and the authors


a
% of novel HG in the genome
20
18
16
14
12
10
8
6
4
2
0
Holozoa
Node 1
Node 2
Metazoa
Eumetazoa
Planulozoa
Bilateria
b
%
3.0
2.5
2.0
1.5
1.0
0.5
0.0
% Core HG that are Novel
% of Novel HG that are Core HG
Holozoa
Node 1
Node 2
Metazoa
Eumetazoa
Planulozoa
Bilateria


c
90
80
70
60
50
40
30
20
10
0
Nucleic acid binding
Transcription factor
Signalling molecule
Receptor
Transporter
Hydrolase
Enzyme modulator
Oxidoreductase
Chaperone
Extracellular matrix protein
Structural protein
Holozoa
Node 1 (Filasterea + Node 2)
Node 2 (Choanoflagellata + Metazoa)
Metazoa
Eumetazoa
Planulozoa
Bilateria


holozoa
N1
N2
metazoa
eumetazoa
planulazoa
bilateria
Deuterostoma
Ecdysozoa
Lophotrochozoa
Cnidaria
Placozoa
Ctenophora
Porifera
Choanoflagellata
Fllasteria
Ichthyosporea
Holomycota
Apusozoa



Family	Order	Class
Hoilungidae	Hoilungea	Uniplacotomia
Cladtertiidae		
	Cladhexea	
Trichoplacidae	Trichoplacea	
Polyplacotomidae	Polyplacotomea	Polyplacotomia

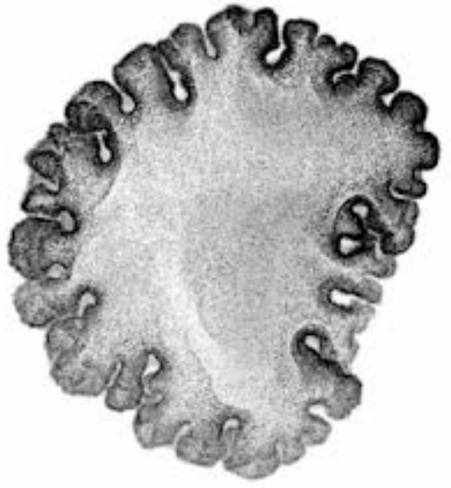


Figure 9: Transcription factor losses, gains and duplications mapped on the placozoan phylogenetic tree