frontiers

Supplementary Material

Genomic Hotspots: Localized chromosome gene expansions identify lineage-specific innovations as targets for functional biodiversity and predictions of stress resilience.

Eric Edsinger[1*] and Leonid L. Moroz[1, 2]

*** Correspondence:** Eric Edsinger: 000generic@gmail.com

1 Supplementary Data

1.1 High Resolution Color Annotated Trees

Data Sheet 1: High resolution Species 16 x TRP tree - annotated genes x species color pdf.

Data Sheet 2: High resolution Species 16 x TRPM tree - annotated genes x species color pdf.

Data Sheet 3: High resolution Species 16 x TRP tree - annotated orthogroups hotspots color pdf.

Data Sheet 4: High resolution Species 16 x TRPM tree - annotated orthogroups hotspots color pdf.

Data Sheet 5: High resolution *Mytilus trossulus* x TRP tree - annotated hotspots color pdf.

Data Sheet 6: High resolution *Mytilus trossulus* x TRPM tree - annotated hotspots color pdf.

Data Sheet 7: High resolution *Mytilus trossulus* x TRP tree - annotated orthogroups hotspots pink pdf.

Data Sheet 8: High resolution *Mytilus trossulus* x TRPM tree - annotated orthogroups hotspots pink pdf.

2 Supplementary Figures and Tables

Supplementary Table 1: Species 16 genome sources. TSV files are text only while Excel files include active URLs. Filenames: Supplementary-Table-01-Genome-sources.tsv and Supplementary-Table-01-Genome-sources.xlsx.

Supplementary Table 2: T1 Proteomes BUSCO numbers and percentages. TSV files are numbers only while Excel files include calculation formulas. Filenames: Supplementary-Table-02-BUSCO-statistics.tsv and Supplementary-Table-02-BUSCO-statistics.xlsx

Supplementary Table 3: Genomic dark matter numbers and percentages. TSV files are numbers only while Excel files include calculation formulas. Filenames: Supplementary-Table-03-Genomic-Dark-Matter.tsv and Supplementary-Table-03-Genomic-Dark-Matter.xlsx

Supplementary Table 4: Hotspot numbers and percentages. TSV files are numbers only while Excel files include calculation formulas. Filenames:Supplementary-Table-04-Hotspots.tsv and Supplementary-Table-04-Hotspots.xlsx

3 3 Supplemental Text

3.1 Methods

Chromosome-scale or better assemblies, gene model proteomes, and GTF/GFF structural annotations were downloaded from NCBI and Ensembl Metazoa public genome repositories (April 2024; Supplementary Table 1) for sixteen species (Species16), including: functional annotation reference species Chordata *Homo sapiens* (human) (100,101), Arthropoda *Drosophila melanogaster* (fly) (102), and Nematoda *Caenorhabditis elegans* (worm) (103,104), and molluscan target species Cephalopoda Octopodidae *Octopus bimaculoides* (105–107), Gastropoda Patellidae *Patella caerulea* (108), Gastropoda Patellidae *Patella pellucida* (27), Gastropoda Patellidae *Patella vulgata* (26), Gastropoda Peltospiridae *Chrysomallon squamiferum* (77,78), Gastropoda Peltospiridae *Gigantopelta aegis* (77,79), Gastropoda Aplysiidae *Aplysia californica*, Bivalvia Myidae *Mya arenaria* (109,110), Bivalvia Pectinidae *Pecten maximus* (111,112), Bivalvia Mytilidae *Mytilus trossulus*, Bivalvia Ostreidae *Ostrea edulis (113)*, Bivalvia Ostreidae *Crassostrea gigas* (114), Bivalvia Ostreidae *Crassostrea virginica* (Figure 1B). Gene and coding sequence coordinates were extracted from associated GTF/GFF files for protein coding genes. Longest protein per gene was determined per proteome (T1 proteomes; sequences provided: Supplementary File 1). T1 proteomes, in addition to genome assemblies in some species (see below), were evaluated for completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO) BUSCO 5.7.1 and its Metazoa database (odb_10 n=954) (115,116). T1 proteomes were annotated by Interproscan 5.59-91.0 (InterPro 99.0, PFAM 36.0, GO 2024/03/17, Panther 18.0) (117–122). The percentage of target species proteins that were annotated vs. unannotated was determined per species. T1 proteomes were also clustered using OrthoFinder 2.5.5-2 (123) under Blast and the percentage of genes clustering with other species and with genetic models human, fly, and worm determined. T1 proteomes were additionally blasted against human, fly, and worm and percentage of genes with no hits determined per species. Sequences unannotated by all three approaches were categorized as genomic dark matter, as they were similarly unrecognized in the three standard homology-based functional annotation methods. Human, fly, and worm TRP superfamily protein sequences were collected from Uniprot and used as a reference gene set in gene family phylogenetic analyses (70). Homologs were identified in molluscan proteomes by reciprocal best hit back to any TRP reference gene sequence. Sequences were aligned using Mafft 7.525 (124), alignments trimmed using Clipkit 2.2.5 (125), and trees built from trimmed alignments using FastTree 2.1.11-3 (126) and IQTree 2.32 (127,128). Trees were assessed in iTOL 6 (129) and branches with less than 80% bootstrap (FastTree) or 95% ultrafast bootstrap (IQTree) support were collapsed. Genome proteomes were blasted against themselves and all hits to a given gene having structural positions within a window of 20 genes (10 genes to either side) and having an e-value of 1e-60 or less were collected as paralog gene sets that formed genomic hotspots in the genome. Initial hotspots were expanded in membership and physical size based on overlap until no new members were identified, at which point the hotspot was locked and assigned an identifier. Odds of a given hotspot initiating randomly is challenging to realistically calculate due to variables of gene and genome size, gene-specific numbers of hits at a given e-value threshold,

different rates of movement within vs. between chromosomes but should be significantly less than 1 in 1000, which would be the odds of a second gene being within a 20 gene window of a first gene in a genome of 20,000 genes, assuming all genes are above threshold and all rate of movement within the genome are equivalent. Orthogroups, hotspots, and gene sizes were mapped onto TRP gene family trees. Trees were color annotated for different features in FigTree 1.4.4 (http://tree.bio.ed.ac.uk/software/**figtree**/) and Keynote 14.0.

3.2 **Results**

3.2.1 **Simplifying homology to be more flexible and encompassing**

We define homology as a state shared between biological features that originate from a feature in a common ancestor. This definition is largely consistent with common definitions of homology (91,92,130–133) but avoids explicit statements of similarity, as seemingly dissimilar features can be homologous. It also generalizes often stated specifics that refer to structures, physiology, and/or development, for example, and simply uses the generic term feature. We define a feature as biological and composed of components and/or processes. Different components and processes within the full set of components and processes that describe a given feature in a given species may have different evolutionary histories, meaning features can be evolutionary mosaics with complex patterns of homology. This situation is commonly referred to as the hierarchical nature of homology (90–94) but we specifically use the term mosaic, as there is no reason homology should be strictly hierarchical in general, even if it happens to be on occasion. Instead, patterns of homology within a feature can be complex and varied and mosaic seems to be a more accurate flexible term to describe this. We avoid use of mosaic in the definition itself, as a given homology of any feature is not mosaicism though the feature may be a mosaic of homologies. Finally, it does not reference individuals or higher levels of organismal comparison, as homology can exit between features within an organism (92,133). In general, the definition is simple and generalized to readily encompass biological complexity in evolution, including molecular genetic to organismal. Overall, it provides a simplified encompassing framework that may clarify and can help facilitate integration of fields, approaches, data, and experiments that leverage biodiversity and evolution.

Definitions used here for gene homologs, orthologs, paralogs, orthogroups, and gene families are standard (91,92,123,130,134–141) but are provided in Supplementary materials (Supplementary Material 1) and overviewed in Figure 1 (Figure 1A) to be explicit.

3.2.2 **Expanding genomic dark matter to include unannotated genomic features**

The term “genomic dark matter” is variously used to described regions of a genome that lack known function, including dark matter of sequencing or assembly-resistant regions, non-coding DNA, introns, and genes more generally (95–98). **Here, we expand the concept of genomic dark matter to include genomic features resistant to functional annotation based on sequence homology to characterized sequences in other species** (Figure 1A). Homology-based functional annotation transfers functions known in reference sequences to their homologs in uncharacterized species. It is a comparative approach commonly done using sequence similarity assessments and tools such as Blast or HMMer for homolog identification, in addition to new structural machine learning and artificial intelligence approaches that improve accuracy and deeper detection of remote sequences (35,39,121,142–150). It works best when reference and target species are 1:1 orthologs, as gene function is more likely to remain similar (60,151–153) but is commonly used with simple one-

direction blast hits of high evalues (1e-3 to 1e-10 is common), where sequences are likely distantly related or share only a domain or motif and function can be likely but unknowingly divergent. Functional annotations to reference species, such the genetic models human (*Homo sapiens*), fly (*Drosophila melanogaster*), and worm (*Caenorhabditis elegans*), are especially critical to genome projects of species in biodiversity, including, for example molluscs and other spiralian taxa, as often little, if any, functional work on target species or even phyla has been done (77–79,105,109,154–158).

3.2.3 **Sequence Homology Terms And Definitions**

Gene homologs are genes descended from a single gene in a common ancestor (54,56,134,139,140). Based on their origins and species evolutionary relationships, gene homologs can be classified as orthologs, paralogs, and orthogroups and can be considered in contexts of gene and gene family or superfamily trees (91,92,123,130,134–141). Paralogs are gene copies of a given gene arising from gene duplication events (Gene paralogs in a single species: 1A1, 1A2, 1A3 - read as: Species 1 Gene A Paralog 1) (54,56,134,139,140). Generally speaking, all gene family members within a species are paralogs (1A1, 1A2, 1A3, 1B1, 1C1, 1C2, 1C3). When paralogs exist prior to speciation (0A1, 0A2, 0A3), paralog copies within each extant species are in-paralogs (1A1, 1A2, 1A3) and (2A1, 2A2, 2A3) while paralog copies between the species are out-paralogs (1A1, 2A3) (54,56,134,139,140). Orthologs are copies of the same gene in different species with the gene copies arising from speciation (1A1, 2A1) (54,56,134,139,140). If paralogs of the gene have arisen in one or both species post-speciation, ortholog relationships between two species can be 1 to 1, 1 to many, many to 1, and many to many (54,56,123,134,136,139,140). Orthogroups are the set of orthologs in two or more species (36,123,136,137,141,159,160). Gene families and superfamilies can include phylogenetically deeper gene homologs that all share a deep common single ancestral sequence (Figure 1A) (121,134,141,161).

3.2.4 **Genome Assembly Assessments**

The lowest BUSCO score coming from *Caenorhabditis* may be surprising (Figure 1C), as its assembly is the first and oldest available genome of any animal and it is of known high-quality (103,104). On the other hand, and more generally, the *Caenorhabditis* and *Drosophila* genomes are well-known for having markedly fewer human homologs than other similarly distantly-related species. However, given the importance of the two heavily studied genetic models, we highlight here that the two genomes likely have fewer human homologs than other species due to very different mechanisms in their genome evolution. *Drosophila* has a relatively small genome of almost 14,000 genes and a high BUSCO score of 98%, suggesting substantial gene loss but a retention of gene families and/or of core genes common to animal biological function, which would include BUSCO genes. In contrast and typical of most animals, *Caenorhabditis* has a human-sized number of genes at over 19,000 but a surprisingly low BUSCO score of 75%, despite its well assembled genome. This suggests gene and possibly gene family loss in parallel to high levels of novelty with gene and gene family innovation. Of note, *Caenorhabditis* homologs in gene family trees are often oddities on the tree, exhibiting very long-branches when most others do not, including in the TRP ion channel superfamily trees below. This suggests an additional or alternative situation of there being cryptic homologs residing in the midnight zone of sequence homology (41,162,163) and going undetected due to high rates of sequence evolution and/or other mechanisms of genetic divergence in the

Caenorhabditis lineage relative to the last common ancestor of human and other phyla. This predicts that Drosophila is likely to have fewer total and potential relative hotspots than *Caenorhabditis* and others and that *Caenorhabditis* may have higher numbers of hotspots but potentially low numbers with identified homologs in human.

Lower BUSCO scores in *Patella vulgata*, *Patella pellucida*, and *Chrysomallon* are surprising, as their genome publications indicate chromosome-scale assemblies and BUSCO Metazoa scores of 97% (26,27,78,108)). However, BUSCO was run on the assemblies themselves and not on the gene model proteomes in publication. We obtained similar BUSCO Metazoa scores (98%) for all three when run directly on the assemblies, suggesting high-quality chromosome-scale assemblies but incomplete structural annotation of genes in the assemblies. Overall, lower T1 proteome but higher genome BUSCO Metazoa scores for the three mollusc species suggests that actual levels of hotspots in their genomes will be under-reported in our pipelines due to incompleteness in gene model calls but not due to genome assembly. Given the comparative phylogenetic and biological importance of the three species and completeness of their assemblies, they were retained in this study.