The extracellular matrix phenome across species

Cyril Statzer and Collin Y. Ewald

Eidgenössische Technische Hochschule Zürich, *Department of Health Sciences and Technology, Institute of Translational Medicine, Schwerzenbach, Zürich CH-8603, Switzerland*

Correspondence to Collin Y. Ewald: *collin-ewald@ethz.ch.*
https://doi.org/10.1016/j.mbplus.2020.100039

Abstract

Extracellular matrices are essential for cellular and organismal function. Recent genome-wide and phenome-wide association studies started to reveal a broad spectrum of phenotypes associated with genetic variants. However, the phenome or spectrum of all phenotypes associated with genetic variants in extracellular matrix genes is unknown. Here, we analyzed over two million recorded genotype-to-phenotype relationships across multiple species to define their extracellular matrix phenomes. By using the previously defined matrisomes of humans, mice, zebrafish, *Drosophila,* and *C. elegans*, we found that the extracellular matrix phenome comprises of 3–10% of the entire phenome. Collagens (*COL1A1*, *COL2A1*) and fibrillin (*FBN1*) are each associated with >150 distinct phenotypes in humans, whereas collagen *COL4A1*, Wnt- and sonic hedgehog (shh) signaling are predominantly associated in other species. We determined the phenotypic fingerprints of matrisome genes and found that *MSTN, CTSD, LAMB2, HSPG2*, and *COL11A2* and their corresponding orthologues have the most phenotypes across species. Out of the 42,551 unique matrisome genotype-to-phenotype relationships across the five species with defined matrisomes, we have constructed interaction networks to identify the underlying molecular components connecting with orthologues phenotypes and with novel phenotypes. Thus, our networks provide a framework to predict unassessed phenotypes and their potential underlying molecular interactions. These frameworks inform on matrisome genotype-to-phenotype relationships and potentially provide a sophisticated choice of biological model system to study human phenotypes and diseases.

Introduction

A principle goal in biology is to link molecular mechanisms and biological pathways to phenotypes. A phenotype is the observable physical properties of an organism determined by the interplay of its genes or genotype with environmental influences. Recent systems-level approaches of genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS) have been harnessed to generate and map genotype-to-phenotype relationships. These efforts are invaluable for clinical research spearheaded by linking electronic health records to medically relevant PheWAS, thereby associating genetic variants to human phenotypes, pathologies, and diseases [1–4]. Assuming that conserved genes and pathways result in comparable phenotypes across species, knowing the genotype-to-phe1notype relationship in one species could inform on the existence of a similar genotype-to-phenotype relationship in another species. The phenome is the entire set of phenotypes resulting from all possible genetic variations [5,6]. In order to make trans-species genome-phenome inferences, the phenomes of several species need to be known. The human phenome is far from reaching saturation [6]. We estimate that the current human phenome consists of 7,037 unique phenotypes (Source: Database Monarch Initiative; timestamp 23.08.2019). From the entire human phenome, several sub-phenomes have been characterized and defined,

including the behavioral phenome [7], the aging phenome [8,9], and the disease phenome [10–12]. These genotype-to-phenotype relationships can be used to build networks and pathways.

In particular, the human disease phenome consists of over 80,000 genetic variants associated with human diseases, of which cardiovascular diseases and skin and connective diseases share the most pathways [12]. These diseases are enriched in variants closely located or found in collagens or extracellular matrix genes. Cells secrete proteins, such as collagens, glycoproteins, and proteoglycans that integrate and form matrices in the extracellular space. The extracellular matrix (ECM) provides structural support, is important for cell-cell communication, and cellular homeostasis [13,14]. The ECM has emerged as a key feature for overall health or disease, with several clinical trials targeting to modify the ECM [15–18]. Furthermore, recent approaches that combine PheWAS from mice to human or zebrafish to human revealed clinically relevant variants in collagen *Col6a5* or in *Ric1* important for collagen-secretion [19,20]. This indicates that a phenome-based approach across species can be of translational value. However, a defined phenome of collagens and other extracellular matrix genes is currently missing for any species.

Here, we mine publicly available databases to establish the phenome of extracellular matrices across species. We take advantage of the large collection of over two million genotype-to-phenotype associations that integrates data across hundreds of species from >30 different curated databases provided by the Monarch Initiative [21]. For the ECM phenome, we include all genes that assemble the matrisome. The matrisome is the compendium of all possible gene products that either form, remodel, or associate with the extracellular matrix [22]. The matrisome of humans consists of 1027 proteins [23], for mice 1110 proteins [23], for zebrafish 1002 proteins [24], for *Drosophila* 641 proteins [25], and for *C. elegans* 719 proteins [26]. This corresponds to roughly 4% of their genomes are dedicated to ECM genes. We find a similar contribution of 3–10% (average 6.5%) of the phenome consisting of 42,551 ECM-specific gene-to-phenotype associations. Across these five species, 639,272 gene-phenotype associations were recorded consisting of 34,781 distinct phenotypes. We identify the phenotypic landscape of the matrisome. Our genotype-to-phenotype interaction maps for matrisome genes reveal unstudied interaction when compared across species.

Methods and materials

Matrisome phenotype-to-genotype relationships

The phenotype-gene association table was obtained from the Monarch Initiative (Source: Database Monarch Initiative; timestamp 23.08.2019) [21]. The phenome was then filtered for associations that were documented with either a primary source or were curated when traceable support existed. Additional requirements were that the interaction was specified as phenotype and that the species in which the observation was made is provided as well. Remaining duplicates owing to gene identifier redundancies were removed (<0.002% in total). The above mentioned steps ensure unique species-gene-phenotype relationships and strike a balance between well-established and novel associations (Supplementary Table 1). The genotype-to-phenotype information was then combined with the defined matrisomes of five species: the matrisome of humans [23], mice [23], zebrafish [24], *Drosophila* [25], and for *C. elegans* [26]). To enable cross-species comparisons among the species with defined and undefined matrisomes, we mapped all genes documented in the Monarch Initiative database to their corresponding orthologues.

In a first step, the curated human orthology information of the published matrisomes was extracted and used to link model organism genes to their human orthologues yielding the highest confidence subset. For *R. norvegicus* and *S. cerevisiae* the DIOPT resource, which aggregates the orthology information of multiple prediction algorithms, was employed to identify human orthologues featuring an annotation rank of moderate to high [27]. The human orthologues of the remaining species for which no published matrisome is available and are not covered by the DIOPT database were mapped using the homologene database (version as of 22. March 2020) by the National Center for Biotechnology Information (NCBI) implemented in the homologene R package (Ogan Mancarci (2019). homologene: Quick Access to Homologene and Gene Annotation Updates. R package version 1.4.68.19.3.27. https://CRAN.R-project.org/package=homologene). The resulting human orthology information was aggregated across the different species and employed prediction methods. All data analysis was performed using the purrr and dplyr R packages (Lionel Henry and Hadley Wickham (2019). purrr: Functional Programming Tools. R package version 0.3.3. https://CRAN.R-project.org/package=purrr; Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr).

Grouping of orthologous phenotypes

To facilitate the comparison between species, the phenotypes were simplified and grouped to form larger phenotype collections (Supplementary Table 2). These phenotype collections enable both cross-species comparisons by avoiding species-specific terminology, as well as, generating a broader overview of the processes involved. Substring matching

was used to assimilate similar phenotypes to phenotype groups, while assuring that each original phenotype only belonged to a single phenotype group using the stringr R package (Hadley Wickham (2019). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr). Capitalization and terminal white spaces were ignored. To assess whether the generated phenotype collections are functionally interconnected, we employed the STRING database and analyzed all orthologues mapped to the human matrisome (version 10) to quantify the predicted degree of interaction on the protein level [28]. We utilized the human interaction information, which required a minimal interaction score of 400 to limit false positive interactions (default), and subjected all phenotypes containing at least ten genes to the interaction network analysis.

Data analysis and visualization

The circular dendrograms were generated with a single hub gene as central node connected to the species in which it was observed and the phenotypes as terminal nodes. The interaction network for each collection of gene-phenotype interactions was computed by extracting the genes, which are associated with the largest number of phenotypes followed by selecting the phenotypes, which comprise the largest number of genes. The edges of the generated network were then facetted by species and their weight set according to the number of species in which a phenotype was observed for the orthologues of a particular gene. Network analysis and visualization was performed using the igraph and ggraph R packages (Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org; Thomas Lin Pedersen (2018). ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R package version 1.0.2. https://CRAN.R-project.org/package=ggraph). Set analysis was conducted using the alluvial (Bojanowski M, Edwards R (2016). _alluvial: R Package for Creating Alluvial Diagrams_. R package version: 0.1–2, <URL: https://github.com/mbojan/alluvial>.), eulerr (Larsson J (2020). _eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses_. R package version 6.1.0, <URL: https://cran.r-project.org/package=eulerr>.) and UpSetR (Nils Gehlenborg (2019). UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets. R package version 1.4.0. https://CRAN.R-project.org/package=UpSetR) R packages. In addition, the R packages ggplot2, ggforce and ggpubr were used (H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016; Thomas Lin Pedersen (2019). ggforce: Accelerating 'ggplot2'. R package version 0.3.1. https://CRAN.R-project.org/package=ggforce; Alboukadel Kassambara (2019). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.2.2. https://CRAN.R-project.org/package=ggpubr). Illustrations were generated using Inkscape version 0.92.

Results

The matrisome-phenome across species

To define the extracellular matrix phenome, we sought to identify the overall contribution of matrisome genes to the phenome for each species. In particular, to assess the suitability for this approach, we first plotted the phenotypic landscape in relation to genotypes across species (Fig. 1A). We found 7,044 unique phenotypes comprising the human phenome, from which 3'813 unique genes are linked to these phenotypes (Supplementary Fig. 1). Fortunately, the species with a good phenotype-to-genotype ratio are the ones that have a fully defined and characterized matrisome [23–26]. We used these defined matrisome gene lists to determine their contribution to the phenomes of humans, mice, zebrafish, *Drosophila*, and *C. elegans* (Fig. 1B-1F). The contribution of the matrisome to the human phenome is 7.6%, with a comparable 2.8–9.5% across species (Fig. 1B-F). Thus, the matrisome is involved in a number of phenotypes, but only corresponds to a maximum of 10% of the overall recorded phenome across species.

The phenotypic signatures of matrisome genes across species

Next, we determined which phenotypes are associated with the greatest number of matrisome genes. For humans the top three phenotypes are "Scoliosis", "Short stature", and "Micrognathia" (Supplementary Fig. 2). By contrast, the top three phenotypes for mice are "Decreased body weight", "Immune system", "Premature death", for zebra fish ("Decreased eye size", "Decreased length whole organism", "Lethal"), for *Drosophila* ("Viable", "Lethal", "Fertile"), and for *C. elegans* ("Locomotion variant", "Embryonic lethal", "Dumpy"; Supplementary Fig. 2). Given that certain phenotypes are species specific, for instance "dumpy" stands for short and fat *C. elegans*, we realized that we cannot compare the phenotypic landscape across species using this conventional phenotypic nomenclature. To overcome this obstacle, we grouped phenotypes into novel phenotype collections. For instance, the phenotypic category "altered body size" includes phenotypes with the key words: body weight, body size, body length, growth phenotype, dwarf, small, dumpy, stunted, tall and others (Materials and Methods, Supplementary Tables 1 and 2). With this approach, we were able to group 86.9% of the total

716,463 species-gene-phenotype associations (Supplementary Table 2). We identified sterility, development, muscle tissue, and altered body size as the most dominant phenotypic categories across species, and for instance bone, connective tissue as well as skin phenotypes for vertebrate and poor viability for invertebrates (Fig. 2). Similarly, besides skin and connective tissue, many morphological phenotypes affecting bone, eye, brain, muscle, and the cardiovascular system are represented in these top categories (Fig. 2). Unexpectedly, aging-related phenotypes, stress resilience, and immune systems phenotypes were ranked across species among these top ECM phenotypic categories (Fig. 2). Since we base our observations on five species, we extended our analysis to incorporate all known genotype-phenotype associations across 43 species. Thereby, we were able to extend the matrisome phenome by homology mapping to three additional mammalian species (*Bos taurus* (cattle), *Rattus norvegicus* (rat), *Canis lupus familiaris* (dog) (Supplementary Table 3). Then, we used these matrisome input gene lists to define their matrisome-phenome. We found similar top phenotypes either for the ungrouped or with our comparable phenotypic categories (Supplementary Figs. 3 and 4). Thus, our analysis revealed known ECM phenotypes but also unexpected phenotypes related to the ECM with high penetrance across species.

Top ranked-matrisome gene signature of the ECM-phenomes

To understand which genes are the major drivers of multiple distinct phenotypes, we ranked matrisome genes based on their number of associated phenotypes (Fig. 3, Supplementary Fig. 5). For humans, collagens (*COL2A1, COL1A1, COL5A1*), fibrillin (*FBN1*), and growth differentiation factor (*GDF5*) are each associated with >150 distinct phenotypes (Fig. 3A). By contrast, for mice, zebrafish, *Drosophila*, and *C. elegans*, Wnt- and sonic hedgehog (shh) signaling and other matrisome-secreted factors are predominantly associated with the greatest number of distinct phenotypes (Fig. 3B-E, Supplementary Fig. 5). Except for zebrafish, the highly conserved collagen type IV (*COL4A1/emb-9*) is associated with range of 20–100 phenotypes across species (Fig. 3). Overall, members of each matrisome category are represented and comparable conserved gene families are linked with numerous phenotypes.

Phenotypic finger prints associated with matrisome genes

Using our phenotype grouping approach, we observed that certain genes were associated with comparable phenotypes across species. For this comparison, we used our phenotypic categories and plotted all matrisome genes as dendrograms connecting the gene with its phenotypes through the species in which they were observed (Fig. 4, Supplementary Fig. 6). The most phenotypically conserved gene is the secreted muscle growth controlling myostatin (MSTN) displaying muscle tissue phenotypes in eight species followed by adipose and cardiovascular phenotypes both observed in three taxa each (Fig. 4A, B). This is not surprising, since either loss or pharmacological inhibition of myostatin leads to almost a doubling of muscle mass, a trait applied in livestock and a potential target for treating muscle wasting diseases [29]. Next is the cathepsin D (CTSD) gene, which is a lysosomal aspartyl protease that can be secreted into the extracellular space to remodel ECM and is involved in many physiological functions, including skin and neuronal development, lipofuscin (age-pigment) removal, and apoptosis [30]. Our analysis of CTSD showed the full phenotypic spectrum across six species with many conserved phenotypes but also species-specific phenotypes (Fig. 4C, Supplementary Fig. 6). Similarly, laminin (LAMB2), heparan sulfate proteoglycan (HSPG2), collagen (COL11A2) showed comparable orthologues phenotypes across multiple species ranging from altered body size to aging-related phenotypes (Fig. 4D-F). The phenotypic finger prints of all matrisome genes is shown in Supplementary Figs. 6 and 7. Thus, we have defined the phenotypic landscape of matrisome genes across species.

Network analysis to identify novel phenotypes and underlying molecular mechanisms

We next asked whether we could use this phenotypic landscape of matrisome genes to inform on potential phenotypes not assessed in humans. Furthermore, could we use gene-phenotype pairs in one species to infer novel interactions in another species? In order to achieve this, we built matrisome-genes-to-phenotype networks across species for all matrisome categories and divisions. These networks are shown in Supplementary Figs. 8 and 9. We chose the three core-matrisome categories, which include ECM glycoproteins, collagens, and proteoglycans, to build a network of the five most abundant interspecies gene-to-phenotype associations for *H. sapiens*, *M. musculus* and *D. rerio* (Fig. 5). We positioned these top five conserved genes and phenotypes for each species at the same location in our graphical network (Fig. 5). We identified the conserved gene-to-phenotypes interactions (bold arrows), but also interactions that only occur in a single species (light arrows; Fig. 5). For instance, genetic variants in the neuromuscular junction protein agrin are associated in humans solely with a “sterility” phenotype, whereas in mouse in addition to sterility, agrin is associated with cardiovascular impairments, altered body size and

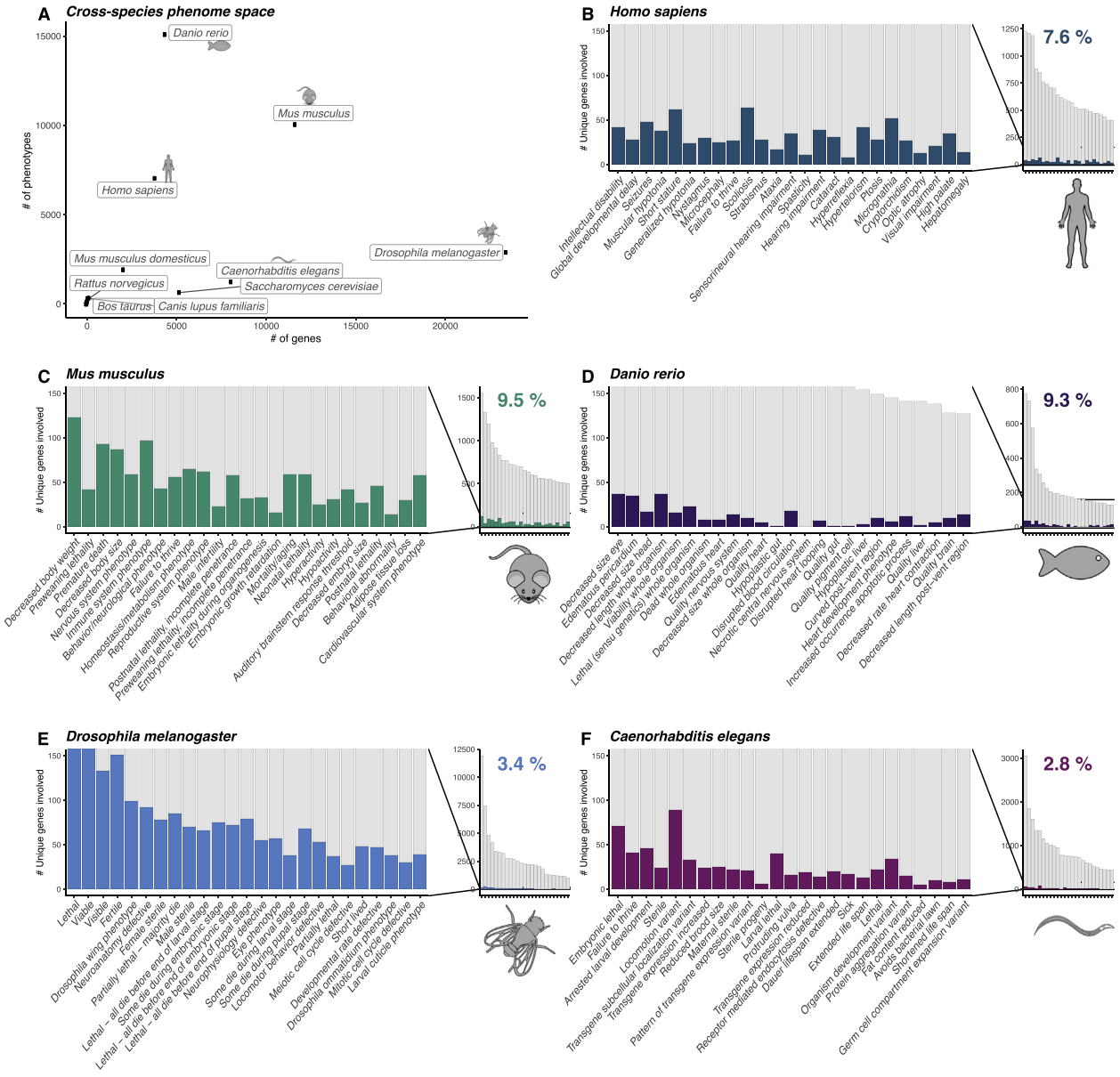

Fig. 1. The contribution of the matrisome to the phenome across species The relationship between the number of genes associated with at least one phenotype and the total number of characterized phenotypes (phenome) linked to at least one gene across species are shown (A). The species with defined matrisomes are depicted in separate panels with the number of genes associated with each phenotype shown on the y-axis and the 25 phenotypes with the highest number of total genes involved are shown on the x-axis for humans (B), for mouse (C), for zebrafish (D), for *Drosophila* (E), and for *C. elegans* (F). The fraction of genes, which are members of each species' matrisome are highlighted in color. The right panel provides an overview of the gene-to-phenotype distribution, while the left panel highlights the contribution of the matrisome by cropping the y-axis. The fraction of matrisome genes associated with the phenotypes compared to all genes is shown as a percentage.

animal morphology, the latter two can also be identified in zebrafish (Fig. 5). To gain more insights, we determined the enrichment of all matrisome category member proteins associated with each phenotype. In humans, we found an enrichment of collagens in muscle and connective tissue phenotypes as well as cardiovascular impairments (Fig. 6, Supplementary Table 6). By orthology, we observed a similar enrichment of collagens for muscle and connective tissue in mouse as well as a generally elevated enrichment of secreted matrisome factors (Supplementary Fig. 10, Supplementary Table 6). Enrichment results for the contribution of each matrisome category to each observed phenotype

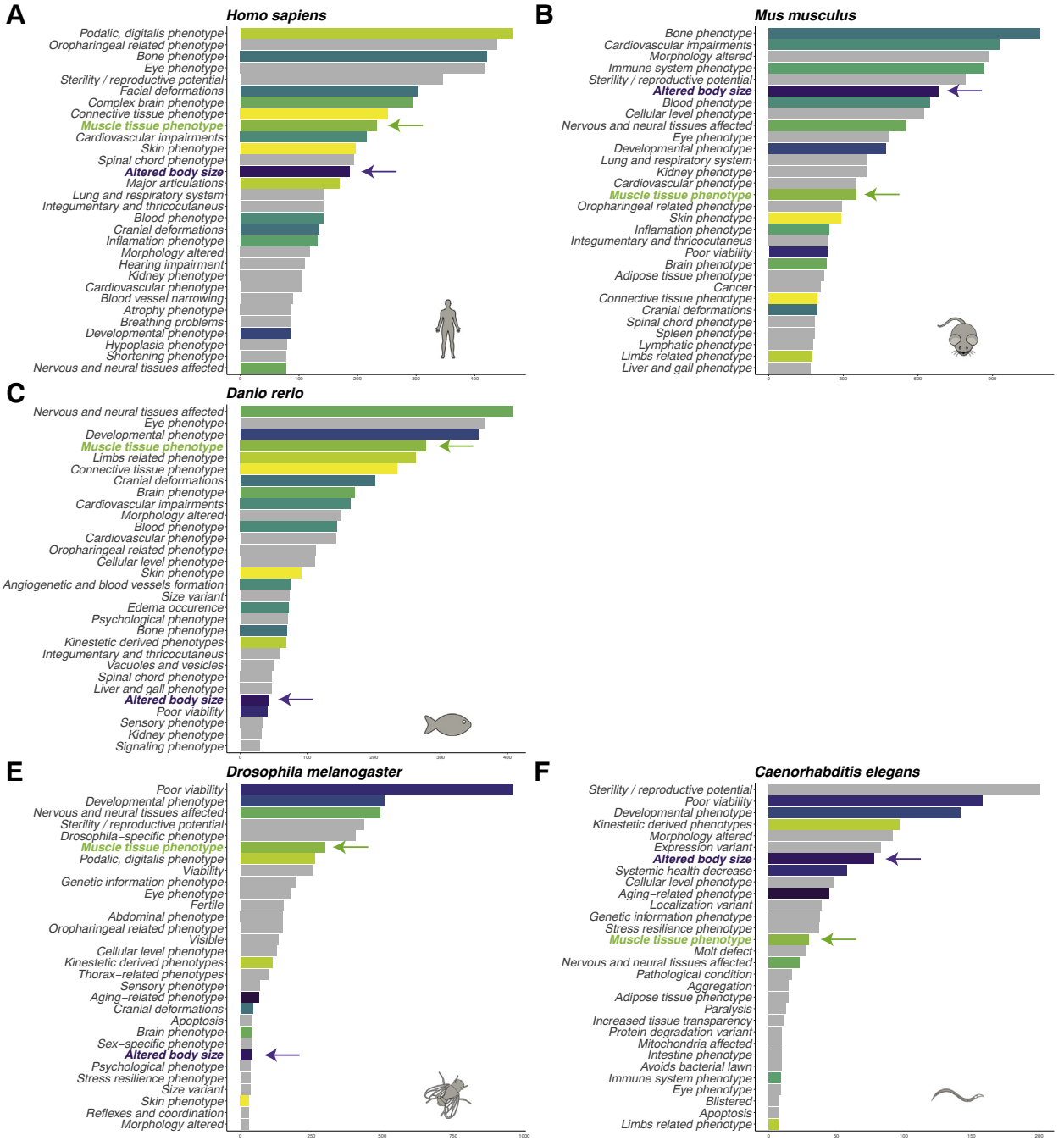

Fig. 2. The matrisome-phenome across species Using the matrisome composition for each species, we show the phenome space associated with it for humans (A), mice (B), zebra fish (C), *Drosophila* (D), and *C. elegans* (E). Each species is depicted in a separate panel with the 30 largest phenotype groups on the y-axis and the number of unique gene/phenotype associations contained in each group on the x-axis. Subsets of phenotypes and analog phenotypes across species are color-coded to ease the comparison (i.e., “altered body size” in purple). For illustration, the muscle tissue phenotype (green arrows) and altered body size (purple arrows) were highlighted across taxa. Some of the phenotypes meaning have slightly different meaning across species due to different morphologies. For instance, the *C. elegans* “eye phenotype”, which refers to anything related to phototaxis (e.g., UV-light sensing), while the “limbs-related phenotype” captures phenotypes related to the animal's tail.

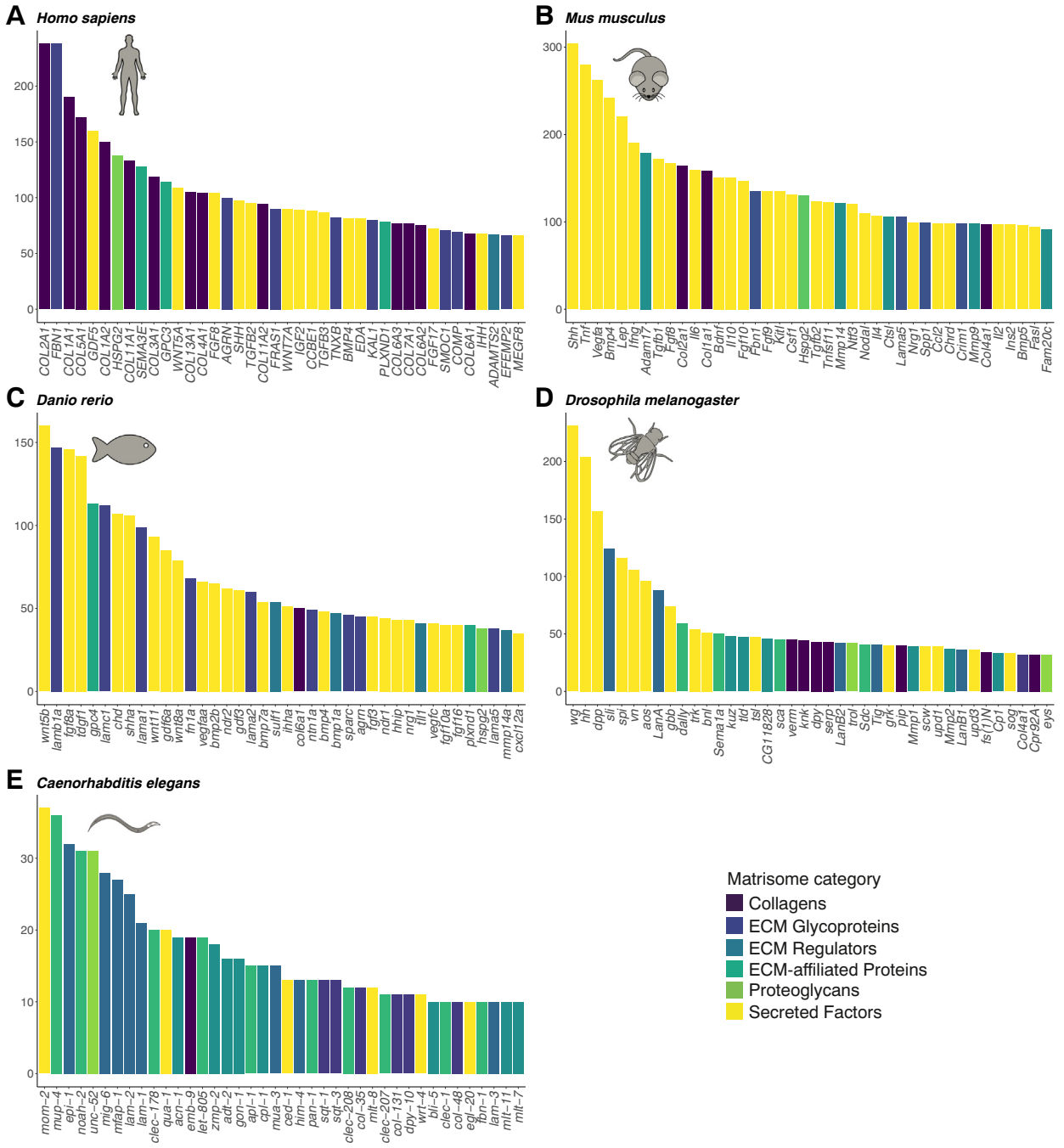

Fig. 3. The degree of phenotypic impact of matrisome genes. The number of unique phenotypes linked to individual matrisome gene is shown in decreasing order for the top 40 genes. The gene name is displayed on the x-axis and the number of unique phenotypes on the y-axis. The color of the bars refers to the matrisome category for each gene.

are provided for all species with defined matrisomes (Supplementary Table 5) as well as for the human matrisome orthologues identified from each species' phenome (Supplementary Table 6). To identify the most studied matrisome phenome independently of the individual matrisome categories, we selected the most phenotypically associated matrisome genes and mapped the gene-to-phenotype relationships across species (Fig. 7). Myostatin (*MSTN*) and collagen type VII (*COL7A1*) are the most conserved drivers for muscle or skin phenotypes, respectively (Fig. 7). Laminin (*LAMA2*), heparan sulfate proteoglycan (*HSPG2*), and sonic hedgehog (*SHH*) gene build a strong interaction network among brain, muscle, and

A

Gene	#Species	Conserved functional fingerprint	Division	Category
MSTN	8	muscle, fat & cardiovasculature	associated	secreted factor
CTSD	6	neurological & motor functions	associated	ECM regulator
COL11A2	6	body size & bone phenotypes	core	collagen
HSPG2	5	muscle, body size & aging	core	proteoglycan
LAMB2	5	muscle & development	core	ECM glycoprotein

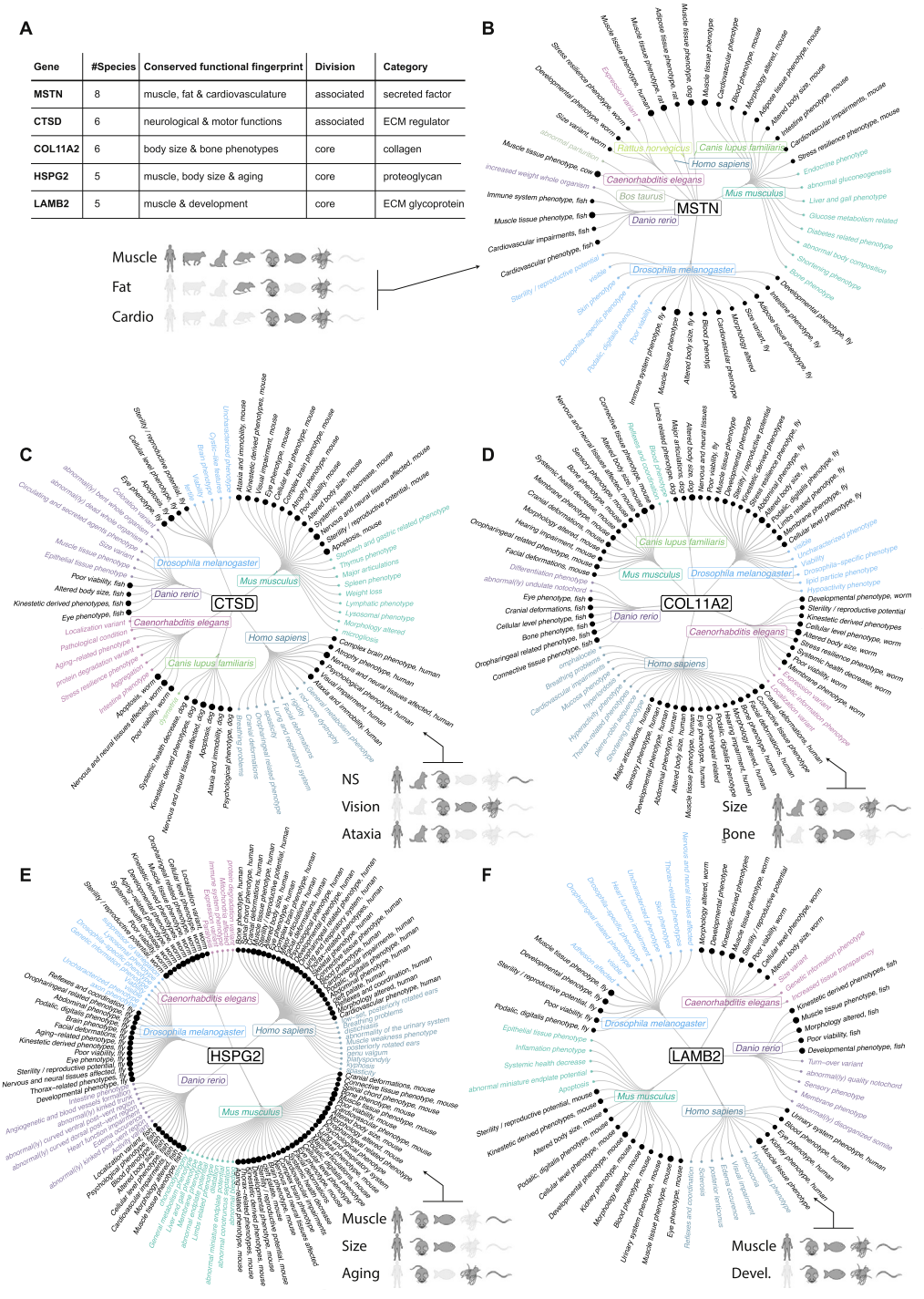

Fig. 4. Phenotypic fingerprint of the most conserved matrisome members The most conserved genes displaying phenotypes across the most species are shown in decreasing order (A). The table displays the human gene name, the number of animal species presenting a phenotype associated with the matrisome gene, the overall phenotypic signature observed for this gene, as well as the matrisome category and division are shown. Circular dendrograms depict five of the most conserved matrisome genes in more detail (B – F). The human orthologue is shown in the middle connected to the species, in which the gene was associated with a phenotype, and then each connected to terminal nodes depicting the phenotype label. Phenotypes that were observed in more than one species are shown in black and the size of the leaf node reflects the number of species this phenotype was observed in. If available, up to ten randomly sampled phenotypes are displayed for each species in addition to the cross-species phenotypes. At the bottom right of each circular dendrogram panel a graphical summary is provided highlighting the species in which by homology a phenotype group was observed (dark grey) or absent (light grey).

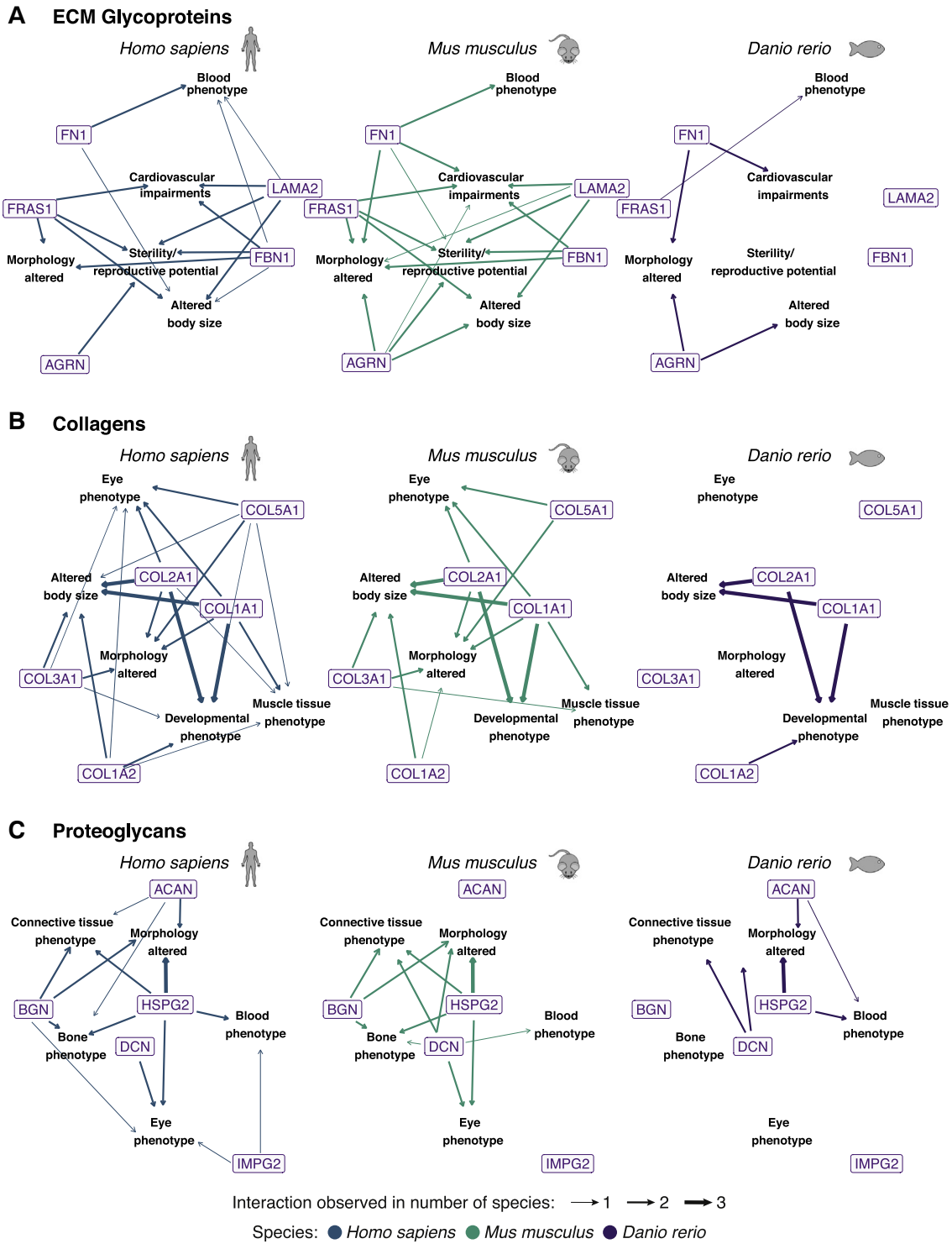

Fig. 5. Phenotypic implications of the structural matrisome categories on vertebrates The matrisome is grouped into multiple categories according to the localization and composition of the extracellular proteins. The members of the ECM glycoprotein category, collagens, and proteoglycans constitute the foundation of the matrisome responsible for the structural integrity of the extracellular matrix. Here the most abundant inter-species gene-to-phenotype associations are highlighted for *H. sapiens*, *M. musculus* and *D. rerio*. Genes are shown as human orthologues in colored labels while phenotype groups are given as plain text. The arrows connect genes to phenotype groups if the gene has been associated with at least one phenotype belonging to the phenotype group. The width of the connection reflects the degree of conservation of the gene-to-phenotype association across species.

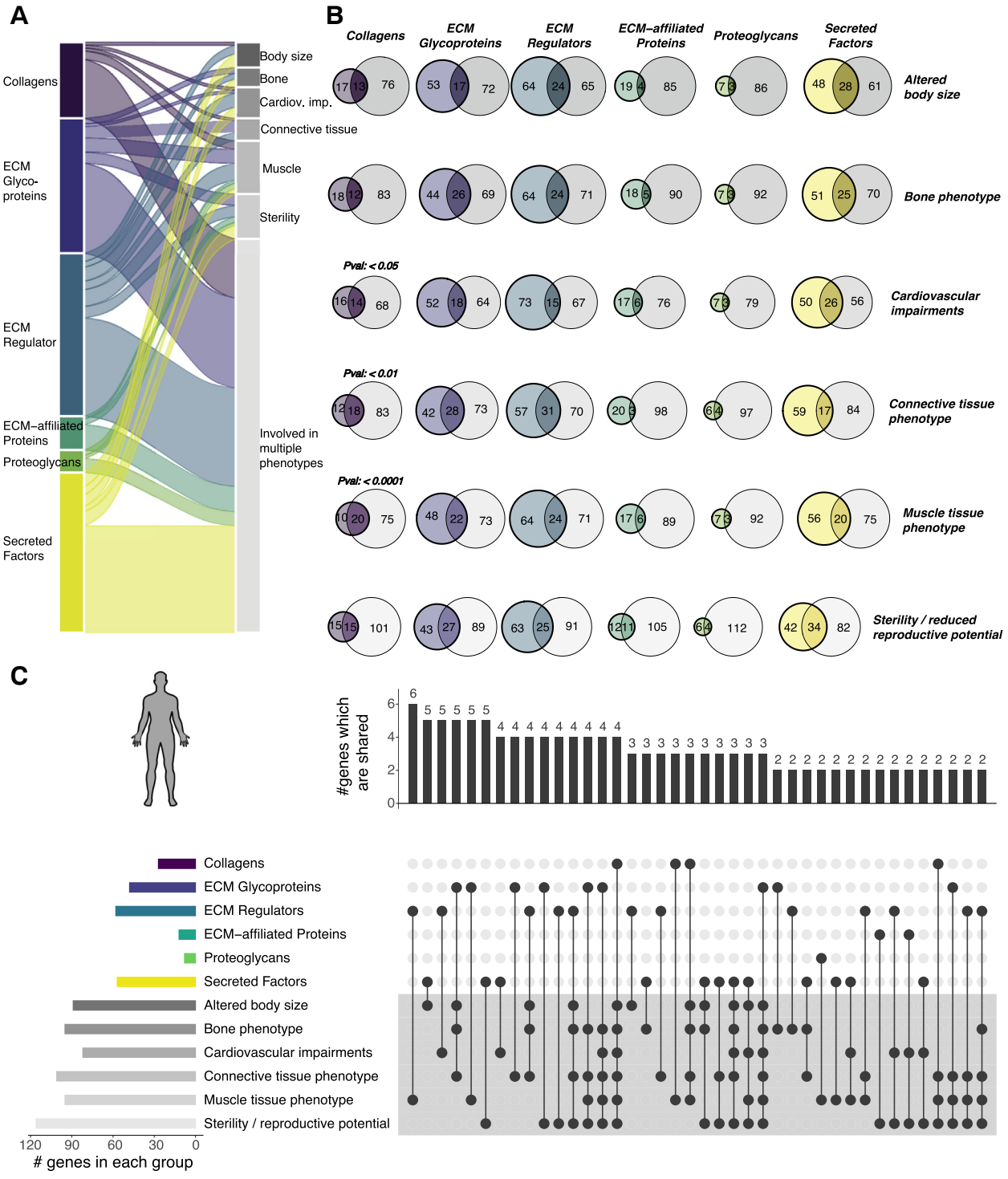

Fig. 6. Interconnection of matrisome categories across phenotypes for humans A bipartite network connecting matrisome categories with most enriched human phenotypes (A). For each phenotype the overlap of a given matrisome category with genes associated with its phenotype displayed as Venn diagrams and *P*-value was determined with one-sided Fisher's exact test (B). The overlap between phenotypes which are associated with members of multiple matrisome categories are visualized using an UpSet plot (C). Each intersecting set is displayed as vertically connected circles indicating the overlapping phenotypes and categories. The number of genes in each intersection is shown as a vertical bar and the overall sizes of the parental groups are displayed as horizontal bars.

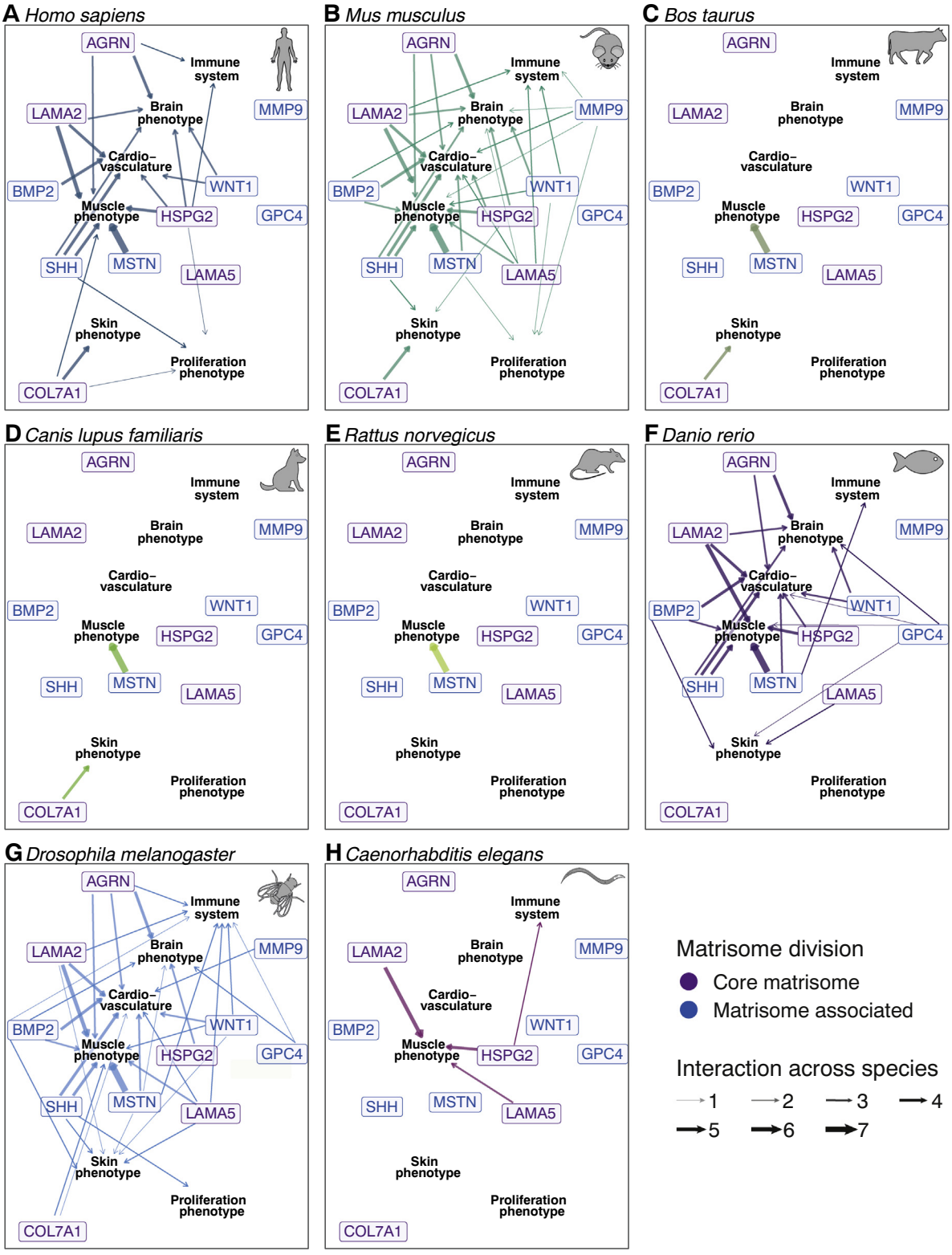

Fig. 7. Phenotypic implications of the entire matrisome A selected subset of the most connected genes of the entire matrisome and phenotypes across species are displayed. Genes are shown as human orthologues color-coded by matrisome division, while phenotype groups are shown as plain text. The arrows connect genes to phenotype groups when the gene is associated with at least one phenotype of the phenotype group. The width of the connection reflects the degree of conservation of the gene-to-phenotype association across species.

cardiovascular phenotypes (Fig. 7). Curiously, only in mice is the matrix metallopeptidase 9 (*MMP9*) implicated with these three phenotypes (brain, muscle, and cardiovascular), but also with immune system and proliferation phenotypes, while the protein is only associated with another phenotype in *D. melanogaster* (cardio vasculature) (Fig. 7). Taken together, our graphical networks build a framework to predict unassessed phenotypes and their potential underlying molecular interactions.

Shared phenotypic roles across matrisome categories

The matrisome is composed of six distinct groups: ECM glycoproteins, collagens, proteoglycans, ECM-affiliated proteins, ECM regulators, and secreted factors [23]. In our network analysis, we highlighted the phenome involvement of the most studied members within each category. To study the entire preferential phenotype association of these categories, we also analyzed their overall enrichment (one-sided Fisher's exact test) for the five species with defined matrisomes (Supplementary Table 5) as well as for all human matrisome orthologues implicated in phenotypes across species (Supplementary Table 6). Specific involvement of a single matrisome category can be observed for example in the case of collagens and skin remodeling. Pathologies like atrophic (P-value $<1.1\ 10^{-8}$) and cigarette paper scars (P-value $<6.3\ 10^{-10}$), in which the tissue is not able to regenerate functionally, are highly enriched for collagens above all other categories (Supplementary Table 5). The human matrisome phenome can be augmented for example by inspecting murine phenotypes, which highlight the specific role of secreted factors and ECM-affiliated proteins in immunological phenotypes (Supplementary Table 5). However, while these cases exist, we observed for the majority of phenotypes a shared involvement of matrisome categories. This is not unexpectantly since processes involving the extracellular matrix likely require concerted actions. Even if individual members of matrisome categories are selectively involved in a subset of phenotypes (Fig. 5), we found an extensive overlap between matrisome categories, when we included all their members. Instead of a one-phenotype to one-category relationship most categories contribute to diverse phenotypes (Supplementary Fig. 10A) and most phenotypes are influenced by multiple matrisome categories (Supplementary Fig. 10C) across species.

Case study of the secreted matrisome factor myostatin (MSTN)

Myostatin (MSTN) displays the highest gene-to-phenotype conservation of any of the tested matrisome genes. In six vertebrates and two invertebrates, the *MSTN* gene is associated with diverse muscle tissue phenotypes (Fig. 4). Using myostatin for this case study, we provided here a brief example of how additional information can be extracted for any gene and phenotype of interest. *MSTN* is a member of the TGFβ family and acts as a phylogenetically conserved negative regulator of muscle mass [31]. The MSTN protein is synthesized mostly in muscle, adipose, and cardiac tissues, where it secreted and present as an extracellular matrix-resident [32]. The clinical relevance of myostatin is predominantly in treating muscle wasting diseases, obesity and type 2 diabetes [33,34].

To be able to study the role of myostatin in muscle tissues across species, we made use of our "muscle tissue phenotype" signature, which we have built from 1808 distinct muscle-related phenotypes (Supplementary Table 2). Using this meta-phenotype, we found that all five matrisome species displayed muscle tissue traits (Fig. 2) as well as twelve additional species without defined matrisomes (8 livestock, 3 pets, 1 rodent; Supplementary Fig. 3). Myostatin and its TGFβ family related proteins are not associated with a wide range of phenotypes and cannot be observed among the genes with the broadest phenotypic impact with the exception of *gbb, dpp* and *scw* in *D. melanogaster* (Fig. 3, Supplementary Fig. 5). To be able to directly compare phenotypes across species, all genes were mapped to their human orthologues implicating 418 human proteins in muscle tissue phenotypes (23% in human, 77% by orthology; Supplementary Table 3). The resulting phenotype group unifies the information gained from human, model and non-model organisms. Its physiological relevance is substantiated by the high degree of functional interaction of its member proteins (Supplementary Table 4). Based on the orthology information, we found that *MSTN* acts as muscle tissue regulator across eight species (Fig. 4 A, B). The phenotypic fingerprint of myostatin further includes adipose tissue, cardiovascular as well as body size regulation across multiple species. From a matrisome perspective, collagens and secreted factors are the most enriched categories in muscle tissue homeostasis for *H. sapiens* and *M. musculus*. Collagens are the most involved structural category regarding the muscle tissue (Fig. 5, Supplementary Fig. 8). Apart from collagens, the ECM Glycoproteins AGRN, LAMA2 and LAMA5 as well as the proteoglycan HSPG2 are among the most involved proteins in muscle tissue regulation (Fig. 6).

To gain a more fine-grained understanding of how myostatin is involved in muscle physiology, we needed to move away from the phenotype group and investigate the individual phenotypes in the ungrouped dataset. The ungrouped dataset can be inspected visually (Supplementary Fig. 7) or in tabular format discussed here (Supplementary Table 3). For mammals, we observed a strong association with increased muscle mass including "skeletal muscle hypertrophy" (*H. sapiens*), "increase in muscle weight" (*R. norvegicus*), "increase in muscle mass" (*B. taurus*), "skeletal

muscle tissue phenotype (hypertrophy)" (*C. lupus familiaris*), and for *M. musculus* 17 ambivalent muscle-related phenotypes, including "increased skeletal muscle fiber number". In the case of *D. rerio,* the available evidence highlights the morphology of the myocardium and for *D. melanogaster* the neuromuscular junction. Hence, the most resourceful species to dissect the role of MSTN in muscle tissue is *M. musculus*. To examine the evidence underlying the claimed increase of muscle fibers, we utilized the listed source publications and/or the MGI database identifiers. Using this information, we determined the parental strain ((129 × 1/SvJ x 129S1/Sv)F1-Kitl^{+}) and the specific genetic variation (replacement of the C-terminal protein sequence). Similarly, the observed adipose tissue signature (Fig. 4 A, B) can be deconstructed into epididymal, inguinal, parametrial and retroperitoneal fat pad (*M. musculus*), embryonic fat (*D. melanogaster*) as well as abdominal fat (*R. norvegicus*) changes. Thus, our output lists or graphical networks provided here as Supplementary Tables and Figures can be used to generated novel hypothesis and/or identify species to study either a given matrisome gene or a phenotype of interest.

Discussion

The recent collective scientific efforts focusing on deep phenotyping and phenome-wide association studies revealed many novel genotype-to-phenotype relationships [2,3,35]. Extracellular matrices are important for cellular and organismal function [13,17]. However, the contribution of ECM gene variants to the phenotypic landscape is unknown.

Here, we defined the ECM phenome of humans, mice, zebrafish, *Drosophila*, and *C. elegans*, with extrapolations to three additional animal species. We found over 42,551 matrisome genotype-to-phenotype relationships across the five defined matrisome species and an additional 108 genotype-to-phenotype relationships in the three undefined-matrisome species. We identified the phenotypic landscape of the matrisome that is conserved among species and also species-specific phenomes. Our analysis of the phenotypic fingerprints of genes revealed *MSTN, CTSD, LAMB2, HSPG2*, and *COL11A2* matrisome genes bearing the most associated phenotypes. From these genotype-to-phenotype interactions, we have built networks linking analogue phenotypes that might be driven by similar underlying molecular mechanisms involving matrisome genes. We provide the ECM phenome as a platform to use information gained from model organisms to implicate novel genotype-to-phenotype relationships for humans.

One limitation of our analysis is our nearly-completed phenotype categorical collections, which include the most frequent occurring phenotypes, but ignores certain infrequently occurring phenotypes during our manual curation process due to the large number phenotypes and sometimes jargon-specific phenotype naming. We aimed to functionally group phenotypes so that these become more comparable across species. This includes also species-specific phenotypes. We manually added 157 broadly defined phenotype categories, which captured 86.9% of the original phenome. However, while the grouping achieves its primary goal to unify the most frequent phenotypes some of the less abundant phenotypes might be falsely grouped or remain ungrouped. Modifications in the grouping system achieved a better assignment of misclassified infrequent phenotypes, with little or no changes in the top phenotypic categories, which adds reliability to our findings and interpretations in this present study. To corroborate that the phenotype groups are functionally connected, we determined their degree of protein-protein interaction using the STRING database. We compared the interaction network of each phenotype group to one hundred randomly sampled interaction networks of the same size and from the same gene background to compare network density. 98.4% of all phenotypes displayed a higher connectivity between their member proteins compared to the randomly sampled group as quantified by the edge-to-node ratio (98.5% of grouped phenotypes). The phenotypes with a lower connectivity than expected by chance can mostly be found in very small communities, where in certain cases only few interactions can be identified. Furthermore, 53.3% of protein networks associated with a phenotype displayed a shorter mean distance compared to background (71.2% of grouped phenotypes; Supplementary Table 4). This confirms that the original and grouped phenotypes utilized in this study are functionally connected and comparable to the ungrouped phenotypes. All analysis including grouped or ungrouped are provided in the supplementary sections, allowing researchers to build upon.

In our phenotypic analysis, we are limited by the phenotypes that have been studied and linked to gene products to date. Using *MSTN* as an example, the gene likely owes its position as the most phenotypically conserved gene to its clearly visible phenotype, it's often beneficial traits in domesticated species leading to selective breeding, which probably increased its likelihood to be identified in genetic studies and its high conservation, resulting in successful orthologue mapping. Many other proteins could be driving phenotypes across even more species but remain to be identified. Overall, we faced a bias towards more easily measured phenotypes in the context of each species and more studied genes (e.g., driven by reverse genetics approaches) over more subtle phenotypes and lesser known genes. While phenotypes studied in humans are directly disease-relevant, the majority of associations is generated through epidemiological studies biased towards selected phenotypes and not studied in functional

experiments. The four model organisms with defined matrisomes offer to augment the human ECM-phenotype landscape through the assessment of orthologues genes. Each model organism contributes a specific part of the phenotypic landscape that would be near impossible to study in another. The contribution of our study is to reduce the species-specific obstacles using our phenotype grouping strategy and orthology mapping approach to allow cross-species inferences. Taken together, we can greatly expanded our understanding of the matrisome phenome by aligning the multiple species-specific perspectives, while keeping in mind that each species is biased towards the phenotypes that researchers are able and interested to study.

The strongest points provided in the present study are our networks of matrisome genes mapped to recorded phenotypes, which offers a conceptual framework for further investigation. This phenome-based network allows the identification of conserved underlying molecular interactions of matrisome genes across comparable phenotypes and species. In particular, it enables querying on given human phenotypes or ECM genes, which are provided with genotype-to-phenotype relationships across other species, suggesting potential novel molecular targets or read-outs for phenotypic high-throughput screens. For instance, early developmental or lethal genes that lead to spontaneous abortion in humans, can be studied with their orthologue gene counterparts in other species [10]. We provide the most conserved subset of the matrisome-interactome dataset in graphical format highlighting the genes and phenotypes with the highest degree of connectivity (Figs. 5 and 6; Supplementary Figs. 8 and 9). Such graphical interaction frameworks have been invaluable for all known human gene associated phenotypes [10].

In our phenome analysis across species, we identified novel aging-related phenotypes associated with matrisome gene variants in the top categories (Fig. 2). Decline and remodeling of the matrisome is a major driver of aging and longevity [18,36–38]. Recently, the human aging phenome has been defined revealing that phenotypes associated with collagens or the matrisome, such as facial wrinkles, kyphosis, arthritis, and osteoporosis are under the most prevalent phenotypes recorded from 77 million elderly [8]. Furthermore, connective tissue diseases, such as Marfan syndrome caused by mutation in *FBN1* and Ehlers-Danlos syndrome, were clustered with features of premature aging diseases (progeria) and aging [8]. In our analysis, we found that the *FBN1* gene has the most associated human phenotypes (Fig. 3 and Supplementary Fig. 5).

The breath of the phenotypic fingerprints of some matrisome genes hints at the conserved modularity of gene systems. Some of these orthologue phenotypes across species suggest a repurposing of the underlying molecular systems through evolution. Our network of these matrisome genes associated with these orthologues phenotypes facilitate studying human phenotypes or even diseases in nonobvious model systems. Thus, our computational analysis provides a systems-level approach to facilitate mechanistic discoveries underlying ECM phenomes across species.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mbplus.2020.100039.

Author contributions

All authors participated in analyzing and interpreting the data. CYE and CS designed the study and wrote the manuscript, CS performed the computational analysis.

Declaration of competing interest

The authors have no competing interests to declare. Correspondence should be addressed to C. Y. E.

Acknowledgement

We thank members of the Ewald lab for critical reading of the manuscript, Davide Vitiello for his help with the phenotypic grouping, and the Monarch Initiative (https://monarchinitiative.org) for providing the indispensable resources for the genotype-phenotype analysis performed in this work. Funding for this study was from the Swiss National Science Foundation grant number PP00P3_163898 to CS and CYE.

Received 6 March 2020;
Received in revised form 11 May 2020;
Accepted 26 May 2020
Available online 23 June 2020

Keywords:
Phenome;
Genotype-to-phenotype;
Matrisome;
Extracellular matrix;
Collagen;
Data mining

References

[1] J.C. Denny, L. Bastarache, D.M. Roden, Phenome-wide association studies as a tool to advance precision medicine, Annu. Rev. Genomics Hum. Genet. 17 (2016) 353–373, https://doi.org/10.1146/annurev-genom-090314-024956.

[2] D.M. Roden, Phenome-wide association studies: a new method for functional genomics in humans, J. Physiol. Lond. 595 (2017) 4109–4115, https://doi.org/10.1113/JP273122.
[3] W.S. Bush, M.T. Oetjens, D.C. Crawford, Unravelling the human genome-phenome relationship using phenome-wide association studies, Nat Rev Genet. 17 (2016) 129–145, https://doi.org/10.1038/nrg.2015.36.
[4] I. Lappalainen, J. Almeida-King, V. Kumanduri, A. Senf, J.D. Spalding, S. Ur-Rehman, et al., The European genome-phenome archive of human data consented for biomedical research, Nat. Genet. 47 (2015) 692–695, https://doi.org/10.1038/ng.3312.
[5] N. Freimer, C. Sabatti, The human phenome project, Nat. Genet. 34 (2003) 15–21, https://doi.org/10.1038/ng0503-15.
[6] M.E. Samuels, Saturation of the human phenome, Curr. Genomics. 11 (2010) 482–499, https://doi.org/10.2174/138920210793175886.
[7] E. Krapohl, J. Euesden, D. Zabaneh, J.-B. Pingault, K. Rimfeld, S. von Stumm, et al., Phenome-wide analysis of genome-wide polygenic scores, Mol. Psychiatry 21 (2016) 1188–1193, https://doi.org/10.1038/mp.2015.126.
[8] M. Ben Ezra, A defined human aging phenome, Aging (Albany NY). 11 (2019) 5786–5806. doi:10.18632/aging.102166.
[9] P. Eline Slagboom, N. van den Berg, J. Deelen, Phenome and genome based studies into human ageing and longevity: an overview, Biochim. Biophys. Acta Mol. Basis Dis. 1864 (2018) 2742–2751, https://doi.org/10.1016/j.bbadis.2017.09.017.
[10] K.-I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, A.-L. Barabási, The human disease network, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 8685–8690, https://doi.org/10.1073/pnas.0701361104.
[11] T.A. Peterson, Towards precision medicine: advances in computational approaches for the analysis of human variants, J. Mol. Biol. 425 (2013) 4047–4063, https://doi.org/10.1016/j.jmb.2013.08.008.
[12] A.G. Cirincione, K.L. Clark, Pathway networks generated from human disease phenome, BMC Med. Genet. 11 (2018) 75, https://doi.org/10.1186/s12920-018-0386-2.
[13] R.O. Hynes, The extracellular matrix: not just pretty fibrils, Science. 326 (2009) 1216–1219, https://doi.org/10.1126/science.1176009.
[14] C. Frantz, K.M. Stewart, V.M. Weaver, The extracellular matrix at a glance, J. Cell Sci. 123 (2010) 4195–4200, https://doi.org/10.1242/jcs.023820.
[15] M.C. Lampi, C.A. Reinhart-King, Targeting extracellular matrix stiffness to attenuate disease: From molecular mechanisms to clinical trials, Sci Transl Med. 10 (2018) eaao0475. https://doi.org/10.1126/scitranslmed.aao0475.
[16] I.N. Taha, A. Naba, Exploring the extracellular matrix in health and disease using proteomics, Essays Biochem. 63 (2019) 417–432, https://doi.org/10.1042/EBC20190001.
[17] C. Bonnans, J. Chou, Z. Werb, Remodelling the extracellular matrix in development and disease, Nat Rev Mol Cell Biol. 15 (2014) 786–801, https://doi.org/10.1038/nrm3904.
[18] C.Y. Ewald, The Matrisome during aging and longevity: a systems-level approach toward defining Matreotypes promoting healthy aging, Gerontology (2019) 1–9, https://doi.org/10.1159/000504295.
[19] X. Wang, A.K. Pandey, M.K. Mulligan, E.G. Williams, K. Mozhui, Z. Li, et al., Joint mouse-human phenome-wide association to test gene function and disease risk, Nat. Commun. 7 (2016) 10464, https://doi.org/10.1038/ncomms10464.
[20] G. Unlu, X. Qi, E.R. Gamazon, D.B. Melville, N. Patel, A.R. Rushing, et al., Phenome-based approach identifies RIC1-linked Mendelian syndrome through zebrafish models, biobank associations and clinical studies, Nat. Med. 26 (2020) 98–109, https://doi.org/10.1038/s41591-019-0705-y.
[21] K.A. Shefchek, N.L. Harris, M. Gargano, N. Matentzoglu, D. Unni, M. Brush, et al., The monarch initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species, Nucleic Acids Res. 48 (2020) D704–D715, https://doi.org/10.1093/nar/gkz997.
[22] R.O. Hynes, A. Naba, Overview of the matrisome—an inventory of extracellular matrix constituents and functions, Cold Spring Harb. Perspect. Biol. 4 (2012) a004903, https://doi.org/10.1101/cshperspect.a004903.
[23] A. Naba, K.R. Clauser, H. Ding, C.A. Whittaker, S.A. Carr, R. O. Hynes, The extracellular matrix: tools and insights for the "omics" era, Matrix Biol. 49 (2016) 10–24, https://doi.org/10.1016/j.matbio.2015.06.003.
[24] P. Nauroy, S. Hughes, A. Naba, F. Ruggiero, The *in-silico* zebrafish matrisome: a new tool to study extracellular matrix gene and protein functions, Matrix Biol. 65 (2018) 5–13, https://doi.org/10.1016/j.matbio.2017.07.001.
[25] M.N. Davis, S. Horne-Badovinac, A. Naba, *In-silico* definition of the *Drosophila melanogaster* matrisome, Matrix Biology Plus. 4 (2019) 100015, https://doi.org/10.1016/j.mbplus.2019.100015.
[26] A.C. Teuscher, E. Jongsma, M.N. Davis, C. Statzer, J.M. Gebauer, A. Naba, et al., The *in-silico* characterization of the *Caenorhabditis elegans* matrisome and proposal of a novel collagen classification, Matrix Biology Plus. (2019) 1–13, https://doi.org/10.1016/j.mbplus.2018.11.001.
[27] Y. Hu, I. Flockhart, A. Vinayagam, C. Bergwitz, B. Berger, N. Perrimon, et al., An integrative approach to ortholog prediction for disease-focused and other functional studies, BMC Bioinformatics. 12 (2011) 357, https://doi.org/10.1186/1471-2105-12-357.
[28] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, et al., STRING v9.1: protein-protein interaction networks, with increased coverage and integration, Nucleic Acids Res. 41 (2013) D808–15. https://doi.org/10.1093/nar/gks1094.
[29] B.D. Rodgers, D.K. Garikipati, Clinical, agricultural, and evolutionary biology of myostatin: a comparative review, Endocr. Rev. 29 (2008) 513–534, https://doi.org/10.1210/er.2008-0003.
[30] P. Benes, V. Vetvicka, M. Fusek, Cathepsin D—many functions of one aspartic protease, Crit. Rev. Oncol. Hematol. 68 (2008) 12–28, https://doi.org/10.1016/j.critrevonc.2008.02.008.
[31] A.C. McPherron, A.M. Lawler, S.J. Lee, Regulation of skeletal muscle mass in mice by a new TGF-beta superfamily member, Nature. 387 (1997) 83–90, https://doi.org/10.1038/387083a0.
[32] C.V.C. Grade, C.S. Mantovani, L.E. Alvares, Myostatin gene promoter: structure, conservation and importance as a target for muscle modulation, J Anim Sci Biotechnol. 10 (2019) 32, https://doi.org/10.1186/s40104-019-0338-5.
[33] J. Jespersen, M. Kjaer, P. Schjerling, The possible role of myostatin in skeletal muscle atrophy and cachexia, Scand. J. Med. Sci. Sports 16 (2006) 74–82, https://doi.org/10.1111/j.1600-0838.2005.00498.x.
[34] S.-J. Lee, Extracellular regulation of Myostatin: a molecular rheostat for muscle mass, Immunol Endocr Metab Agents Med Chem. 10 (2010) 183–194, https://doi.org/10.2174/187152210793663748.

[35] H. Li, J. Auwerx, Mouse systems genetics as a prelude to precision medicine, Trends Genet. 36 (2020) 259–272, https://doi.org/10.1016/j.tig.2020.01.004.

[36] C.Y. Ewald, J.N. Landis, J. Porter Abate, C.T. Murphy, T.K. Blackwell, Dauer-independent insulin/IGF-1-signalling implicates collagen remodelling in longevity, Nature. 519 (2015) 97–101, https://doi.org/10.1038/nature14021.

[37] D. Bakula, A.M. Aliper, P. Mamoshina, M.A. Petr, A. Teklu, J. A. Baur, et al., Aging and drug discovery, Aging (Albany NY). 10 (2018) 3079–3088. doi:10.18632/aging.101646.

[38] A.C. Teuscher, C. Statzer, S. Pantasis, M.R. Bordoli, C.Y. Ewald, Assessing collagen deposition during aging in mammalian tissue and in *Caenorhabditis elegans*, Methods Mol. Biol. 1944 (2019) 169–188, https://doi.org/10.1007/978-1-4939-9095-5_13.