Using the Gene Ontology to Obtain a Comprehensive List of Genes and Gene Products Associated with the Vascular Extracellular Matrix of Homo *sapiens*

Brian Lee

Department of Engineering, Florida Atlantic University, Boca Raton, FL 33432 USA

(Dated: February 2, 2019)

Abstract

The extracellular matrix (ECM) is the major structural component that protects and controls the shape of vascular tissue. The concentration of components in this matrix can vary greatly from individual to individual due to age and environmental factors, and these concentrations can also vary within an individual in different regions of the body. Issues with these structural components can cause major complications such atherosclerosis. Due to these large variations in the matrix, it provides an interesting area of research in studying differential protein expression levels among individuals. There are many tools and resources available that can allow one to efficiently find meaningful targets for further research. Perhaps the two most useful are the Gene Ontology (GO) and UniProt Knowledge Base. These two tools can help provide an extensive list of potential targets and a great interface to effectively narrow the focus to the most relevant and important targets for further investigation.

I. BACKGROUND

Endothelial cells, vascular smooth muscle cells, and the extracellular matrix (ECM) are the three major components of blood vessels and their supporting structures in the vascular system. The ECM is mainly comprised of fibrous proteins and glycoproteins which are embedded in a ground substance composed of water, glycosaminoglycans, and proteoglycans. The ECM is the key structural component of the vascular wall, providing the vessel with elasticity with elastic lamellae and tensile strength through collagen fibers. It also influences cellular signaling and behavior. Under homeostatic conditions, the proteins of the ECM are quite stable: the half-life of elastic fibers in human arteries is believed to be 50–70 years [1]. The ECM and the cells that it is associated with organize vascular walls into 3 layers: the intima, media, and adventitia.

Most blood vessels have a similar basic structure. The intima is composed mostly of ground substance separating ECs from the internal elastic lamina (IEL) which begins at the media. After the IEL, the media is organized into concentric lamellar units composed of elastic fibers and smooth-muscle cells. These units are separated by an interlamellar matrix containing collagens and microfibrils as well as proteoglycans and glycoproteins which make up the ground substance. Finally, beyond the external elastic lamina of the media, is the adventitia which is composed of type I and III collagen, chondroitin sulfate and dermatan sulfate proteoglycans, fibronectin, and other proteins [2].

By examining the different components that make up the ECM, comparisons can be made between the vascular ECM that occurs in different areas of the body such as the differences between the ECM of veins and arteries. Due to the vascular ECM's role in diseases such as atherosclerosis and other complications that come from vascular aging, comparisons can be made between healthy and unhealthy tissues to help uncover the underlying mechanisms that cause these diseases. By using tools made available to us by the scientific community, we can analyze and compare these tissues using methods such as Gene Enrichment Analysis. The first step to this process, however, is finding a list of components that can later be filtered as potential candidates for further examination.

II. METHODS

By utilizing the Gene Ontology (GO), a list of vascular ECM components can be relatively easy to obtain. Amigo2 was chosen as an interface to browse through the Gene Ontology. Once understanding the logic behind GO, it becomes intuitive to navigate the ontology in order to obtain the needed information. Navigating the ontology starts at the parent level offering 3 choices of biological_process, cellular_component, molecular_function. The option cellular_component is chosen because it organizes data by cellular anatomy which is the extracellular matrix would be a part of. Once opening the list of child components, the navigation goes as follows: extracellular region, extracellular region part, extracellular matrix. From the syntax of GO, this would read as extracellular matrix is an extracellular matrix part which is part of the extracellular region which is a cellular_component.

Moving to the annotations view, Amigo2 presents a list of genes and gene products of the extracellular matrix. The filtering function of Amigo2 is then utilized to further narrow down the list to the desired results. First, the list is filtered by organism: Homo sapiens. Then, list is filtered by annotation extension: vascular system. This generates a list of 222 genes/gene products as of the time of this writing. The list contains many duplicates because the same gene may have been annotated by multiple different sources.

The data was then downloaded to a text file, using only the "Gene/product (bioentity)" as a selected field. The generated file contains UniProt Knowledge Base entries. This is desirable, because UniProt offers multitudes of information about all of these genes.

The stream editor sed was then used to remove the "UniProtKB:" prefix from each entry. The resulting file can then be directly uploaded or copy-pasted to UniProt's "Retrieve/ID mapping" tool. This generates a personalized database which can be used to explore all types of information about these genes of interest. UniProt's interface also allows EXCEL or tab-separated information to be downloaded with customized data for each column. The results obtained provide each gene name along with the name of the protein it produces as well as the base pair length of each gene sequence.

III. RESULTS

TABLE I: UniProt Output

Entry	Protein names	Gene names	Length
P04004	Vitronectin (VN) (S-protein) (Serum-spreading factor) (V75) [Cleaved into: Vitronectin V65 subunit; Vitronectin V10 subunit; Somatomedin-B]	VTN	478
P23142	Fibulin-1 (FIBL-1)	FBLN1 PP213	703
Q9Y6C2	EMILIN-1 (Elastin microfibril interface-located protein 1) (Elastin microfibril interfacer 1)	EMILIN1 EMI	1016
P10909	Clusterin (Aging-associated gene 4 protein) (Apolipoprotein J) (Apo-J) (Complement cytolysis inhibitor) (CLI) (Complement-associated protein SP-40,40) (Ku70-binding protein 1) (NA1/NA2) (Testosterone-repressed prostate message 2) (TRPM-2) [Cleaved into: Clusterin beta chain (ApoJalpha) (Complement cytolysis inhibitor a chain); Clusterin alpha chain (ApoJbeta) (Complement cytolysis inhibitor b chain)]	CLU APOJ CLI KUB1 AAG4	449
P49747	Cartilage oligomeric matrix protein (COMP) (Thrombospondin-5) (TSP5)	COMP	757
Q6PCB0	von Willebrand factor A domain-containing protein 1	VWA1	445
O43294	Transforming growth factor beta-1-induced transcript 1 protein (Androgen receptor coactivator 55 kDa protein) (Androgen receptor-associated protein of 55 kDa) (Hydrogen peroxide-inducible clone 5 protein) (Hic-5)	TGFB1I1 ARA55	461
Q8NDA2	Hemicentin-2	HMCN2	5059
Q9UBX5	Fibulin-5 (FIBL-5) (Developmental arteries and neural crest EGF-like protein) (Dance) (Urine p50 protein) (UP50)	FBLN5 DANCE UNQ184/PRO210	448

TABLE I: UniProt Output

Entry	Protein names	Gene names	Length
P35556	Fibrillin-2 [Cleaved into: Fibrillin-2 C-terminal peptide]	FBN2	2912
P35555	Fibrillin-1 [Cleaved into: Asprosin]	FBN1 FBN	2871
Q9BQB4	Sclerostin	SOST UNQ2976/ PRO7455/PRO7476	213
Q12805	EGF-containing fibulin-like extracellular matrix protein 1 (Extracellular protein S1-5) (Fibrillin-like protein) (Fibulin-3) (FIBL-3)	EFEMP1 FBLN3 FBNL	493
Q96P44	Collagen alpha-1(XXI) chain	COL21A1 COL1AL FP633	957
O00339	Matrilin-2	MATN2 UNQ193/PRO219	956
P55268	Laminin subunit beta-2 (Laminin B1s chain) (Laminin-11 subunit beta) (Laminin-14 subunit beta) (Laminin-15 subunit beta) (Laminin-3 subunit beta) (Laminin-4 subunit beta) (Laminin-7 subunit beta) (Laminin-9 subunit beta) (S-laminin subunit beta) (S-LAM beta)	LAMB2 LAMS	1798
Q7Z5L7	Podocan	PODN SLRR5A UNQ293/PRO332	613
P03973	Antileukoproteinase (ALP) (BLPI) (HUSI-1) (Mucus proteinase inhibitor) (MPI) (Protease inhibitor WAP4) (Secretory leukocyte protease inhibitor) (Seminal proteinase inhibitor) (WAP four-disulfide core domain protein 4)	SLPI WAP4 WFDC4	132

TABLE I: UniProt Output

Entry	Protein names	Gene names	Length
O75339	Cartilage intermediate layer protein 1 (CILP-1) (Cartilage intermediate-layer protein) [Cleaved into: Cartilage intermediate layer protein 1 C1; Cartilage intermediate layer protein 1 C2]	CILP UNQ602/PRO1188	1184
Q8N2S1	Latent-transforming growth factor beta-binding protein 4 (LTBP-4)	LTBP4	1624
P08123	Collagen alpha-2(I) chain (Alpha-2 type I collagen)	COL1A2	1366
Q07092	Collagen alpha-1(XVI) chain	COL16A1 FP1572	1604
P21810	Biglycan (Bone/cartilage proteoglycan I) (PG-S1)	BGN SLRR1A	368
Q15113	Procollagen C-endopeptidase enhancer 1 (Procollagen COOH-terminal proteinase enhancer 1) (PCPE-1) (Procollagen C-proteinase enhancer 1) (Type 1 procollagen C-proteinase enhancer protein) (Type I procollagen COOH-terminal proteinase enhancer)	PCOLCE PCPE1	449
Q13361	Microfibrillar-associated protein 5 (MFAP-5) (MP25) (Microfibril-associated glycoprotein 2) (MAGP-2)	MFAP5 MAGP2	173
Q9BXN1	Asporin (Periodontal ligament-associated protein 1) (PLAP-1)	ASPN PLAP1 SLRR1C UNQ215/PRO241	380
P35625	Metalloproteinase inhibitor 3 (Protein MIG-5) (Tissue inhibitor of metalloproteinases 3) (TIMP-3)	TIMP3	211
Q05707	Collagen alpha-1(XIV) chain (Undulin)	COL14A1 UND	1796
P08311	Cathepsin G (CG) (EC 3.4.21.20)	CTSG	255
P16112	Aggrecan core protein (Cartilage-specific proteoglycan core protein) (CSPCP) (Chondroitin sulfate proteoglycan core protein 1) (Chondroitin sulfate proteoglycan 1) [Cleaved into: Aggrecan core protein 2]	ACAN AGC1 CSPG1 MSK16	2530

TABLE I: UniProt Output

Entry	Protein names	Gene names	Length
P55107	Growth/differentiation factor 10 (GDF-10) (Bone morphogenetic protein 3B) (BMP-3B) (Bone-inducing protein) (BIP)	GDF10 BMP3B	478
Q8N474	Secreted frizzled-related protein 1 (FRP-1) (sFRP-1) (Secreted apoptosis-related protein 2) (SARP-2)	SFRP1 FRP FRP1 SARP2	314
P34741	Syndecan-2 (SYND2) (Fibroglycan) (Heparan sulfate proteoglycan core protein) (HSPG) (CD antigen CD362)	SDC2 HSPG1	201
P13611	Versican core protein (Chondroitin sulfate proteoglycan core protein 2) (Chondroitin sulfate proteoglycan 2) (Glial hyaluronate-binding protein) (GHAP) (Large fibroblast proteoglycan) (PG-M)	VCAN CSPG2	3396
P27658	Collagen alpha-1(VIII) chain (Endothelial collagen) [Cleaved into: Vastatin]	COL8A1 C3orf7	744
P11047	Laminin subunit gamma-1 (Laminin B2 chain) (Laminin-1 subunit gamma) (Laminin-10 subunit gamma) (Laminin-11 subunit gamma) (Laminin-2 subunit gamma) (Laminin-3 subunit gamma) (Laminin-4 subunit gamma) (Laminin-6 subunit gamma) (Laminin-7 subunit gamma) (Laminin-8 subunit gamma) (Laminin-9 subunit gamma) (S-laminin subunit gamma) (S-LAM gamma)	LAMC1 LAMB2	1609
Q15582	Transforming growth factor-beta-induced protein ig-h3 (Beta ig-h3) (Kerato-epithelin) (RGD-containing collagen-associated protein) (RGD-CAP)	TGFBI BIGH3	683
P22105	Tenascin-X (TN-X) (Hexabrachion-like protein)	TNXB HXBL TNX TNXB1 TNXB2 XB	4244

TABLE I: UniProt Output

Entry	Protein names	Gene names	Length
O00468	Agrin [Cleaved into: Agrin N-terminal 110 kDa subunit; Agrin C-terminal 110 kDa subunit; Agrin C-terminal 90 kDa fragment (C90); Agrin C-terminal 22 kDa fragment (C22)]	AGRN AGRIN	2068
Q9HCB6	Spondin-1 (F-spondin) (Vascular smooth muscle cell growth-promoting factor)	SPON1 KIAA0762 VSGP	807
Q15063	Periostin (PN) (Osteoblast-specific factor 2) (OSF-2)	POSTN OSF2	836
P08572	Collagen alpha-2(IV) chain [Cleaved into: Canstatin]	COL4A2	1712
P36955	Pigment epithelium-derived factor (PEDF) (Cell proliferation-inducing gene 35 protein) (EPC-1) (Serpin F1)	SERPINF1 PEDF PIG35	418
P51888	Prolargin (Proline-arginine-rich end leucine-rich repeat protein)	PRELP SLRR2A	382
P51884	Lumican (Keratan sulfate proteoglycan lumican) (KSPG lumican)	LUM LDC SLRR2D	338
Q8IVN8	Somatomedin-B and thrombospondin type-1 domain-containing protein (RPE-spondin)	SBSPON C8orf84 RPESP	264
P08493	Matrix Gla protein (MGP) (Cell growth-inhibiting gene 36 protein)	MGP MGLAP GIG36	103
Q8IUX7	Adipocyte enhancer-binding protein 1 (AE-binding protein 1) (Aortic carboxypeptidase-like protein)	AEBP1 ACLP	1158
P24821	Tenascin (TN) (Cytotactin) (GMEM) (GP 150-225) (Glioma-associated-extracellular matrix antigen) (Hexabrachion) (JI) (Myotendinous antigen) (Neuronectin) (Tenascin-C) (TN-C)	TNC HXB	2201
P14543	Nidogen-1 (NID-1) (Entactin)	NID1 NID	1247

TABLE I: UniProt Output

Entry	Protein names	Gene names	Length
P23946	Chymase (EC 3.4.21.39) (Alpha-chymase) (Mast cell protease I)	CMA1 CYH CYM	247
P24158	Myeloblastin (EC 3.4.21.76) (AGP7) (C-ANCA antigen) (Leukocyte proteinase 3) (PR-3) (PR3) (Neutrophil proteinase 4) (NP-4) (P29) (Wegener autoantigen)	PRTN3 MBN	256
P98160	Basement membrane-specific heparan sulfate proteoglycan core protein (HSPG) (Perlecan) (PLC) [Cleaved into: Endorepellin; LG3 peptide]	HSPG2	4391
P02462	Collagen alpha-1(IV) chain [Cleaved into: Arresten]	COL4A1	1669
P02452	Collagen alpha-1(I) chain (Alpha-1 type I collagen)	COL1A1	1464
O15230	Laminin subunit alpha-5 (Laminin-10 subunit alpha) (Laminin-11 subunit alpha) (Laminin-15 subunit alpha)	LAMA5 KIAA0533 KIAA1907	3695
Q9GZM7	Tubulointerstitial nephritis antigen-like (Glucocorticoid-inducible protein 5) (Oxidized LDL-responsive gene 2 protein) (OLRG-2) (Tubulointerstitial nephritis antigen-related protein) (TIN Ag-related protein) (TIN-Ag-RP)	TINAGL1 GIS5 LCN7 OLRG2 TINAGL PP6614 PSEC0088 UNQ204/PRO230	467
P02751	Fibronectin (FN) (Cold-insoluble globulin) (CIG) [Cleaved into: Anastellin; Ugl-Y1; Ugl-Y2; Ugl-Y3]	FN1 FN	2386
P02749	Beta-2-glycoprotein 1 (APC inhibitor) (Activated protein C-binding protein) (Anticardiolipin cofactor) (Apolipoprotein H) (Apo-H) (Beta-2-glycoprotein I) (B2GPI) (Beta(2)GPI)	APOH B2G1	345

TABLE I: UniProt Output

Entry	Protein names	Gene names	Length
P02743	Serum amyloid P-component (SAP) (9.5S alpha-1-glycoprotein) [Cleaved into: Serum amyloid P-component(1-203)]	APCS PTX2	223
Q08397	Lysyl oxidase homolog 1 (EC 1.4.3.-) (Lysyl oxidase-like protein 1) (LOL)	LOXL1 LOXL	574
P17931	Galectin-3 (Gal-3) (35 kDa lectin) (Carbohydrate-binding protein 35) (CBP 35) (Galactose-specific lectin 3) (Galactoside-binding protein) (GALBP) (IgE-binding protein) (L-31) (Laminin-binding protein) (Lectin L-29) (Mac-2 antigen)	LGALS3 MAC2	250
P02656	Apolipoprotein C-III (Apo-CIII) (ApoC-III) (Apolipoprotein C3)	APOC3	99
P09382	Galectin-1 (Gal-1) (14 kDa laminin-binding protein) (HLBP14) (14 kDa lectin) (Beta-galactoside-binding lectin L-14-I) (Galaptin) (HBL) (HPL) (Lactose-binding lectin 1) (Lectin galactoside-binding soluble 1) (Putative MAPK-activating protein PM12) (S-Lac lectin 1)	LGALS1	135
P24043	Laminin subunit alpha-2 (Laminin M chain) (Laminin-12 subunit alpha) (Laminin-2 subunit alpha) (Laminin-4 subunit alpha) (Merosin heavy chain)	LAMA2 LAMM	3122
P07996	Thrombospondin-1 (Glycoprotein G)	THBS1 TSP TSP1	1170
P07942	Laminin subunit beta-1 (Laminin B1 chain) (Laminin-1 subunit beta) (Laminin-10 subunit beta) (Laminin-12 subunit beta) (Laminin-2 subunit beta) (Laminin-6 subunit beta) (Laminin-8 subunit beta)	LAMB1	1786

Entry	Protein names	Gene names	Length
Q06828	Fibromodulin (FM) (Collagen-binding 59 kDa protein) (Keratan sulfate proteoglycan fibromodulin) (KSPG fibromodulin)	FMOD FM SLRR2E	376
Q14112	Nidogen-2 (NID-2) (Osteonidogen)	NID2	1375
Q6UVK1	Chondroitin sulfate proteoglycan 4 (Chondroitin sulfate proteoglycan NG2) (Melanoma chondroitin sulfate proteoglycan) (Melanoma-associated chondroitin sulfate proteoglycan)	CSPG4 MCSP	2322
Q92743	Serine protease HTRA1 (EC 3.4.21.-) (High-temperature requirement A serine peptidase 1) (L56) (Serine protease 11)	HTRA1 HTRA PRSS11	480
P31151	Protein S100-A7 (Psoriasin) (S100 calcium-binding protein A7)	S100A7 PSOR1 S100A7C	101
P08294	Extracellular superoxide dismutase [Cu-Zn] (EC-SOD) (EC 1.15.1.1)	SOD3	240
P10915	Hyaluronan and proteoglycan link protein 1 (Cartilage-linking protein 1) (Cartilage-link protein) (Proteoglycan link protein)	HAPLN1 CRTL1	354
P55083	Microfibril-associated glycoprotein 4	MFAP4	255
Q99715	Collagen alpha-1(XII) chain	COL12A1 COL12A1L	3063
Q15661	Tryptase alpha/beta-1 (Tryptase-1) (EC 3.4.21.59) (Tryptase I) (Tryptase alpha-1)	TPSAB1 TPS1 TPS2 TPSB1	275
Q7Z7G0	Target of Nesh-SH3 (Tarsh) (ABI gene family member 3-binding protein) (Nesh-binding protein) (NeshBP)	ABI3BP NESHBP TARSH	1075
Q07507	Dermatopontin (Tyrosine-rich acidic matrix protein) (TRAMP)	DPT	201

TABLE I: UniProt Output

Entry	Protein names	Gene names	Length
O43854	EGF-like repeat and discoidin I-like domain-containing protein 3 (Developmentally-regulated endothelial cell locus 1 protein) (Integrin-binding protein DEL1)	EDIL3 DEL1	480
Q9H239	Matrix metalloproteinase-28 (MMP-28) (EC 3.4.24.-) (Epilysin)	MMP28 MMP25 UNQ1893/PRO4339	520
P07339	Cathepsin D (EC 3.4.23.5) [Cleaved into: Cathepsin D light chain; Cathepsin D heavy chain]	CTSD CPSD	412
P39060	Collagen alpha-1(XVIII) chain [Cleaved into: Endostatin; Non-collagenous domain 1 (NC1)]	COL18A1	1754
P39059	Collagen alpha-1(XV) chain [Cleaved into: Restin (Endostatin-XV) (Related to endostatin) (Restin-I); Restin-2 (Restin-II); Restin-3 (Restin-III); Restin-4 (Restin-IV)]	COL15A1	1388
Q16363	Laminin subunit alpha-4 (Laminin-14 subunit alpha) (Laminin-8 subunit alpha) (Laminin-9 subunit alpha)	LAMA4	1823
P05997	Collagen alpha-2(V) chain	COL5A2	1499
P07585	Decorin (Bone proteoglycan II) (PG-S2) (PG40)	DCN SLRR1B	359
Q16787	Laminin subunit alpha-3 (Epiligrin 170 kDa subunit) (E170) (Epiligrin subunit alpha) (Kalinin subunit alpha) (Laminin-5 subunit alpha) (Laminin-6 subunit alpha) (Laminin-7 subunit alpha) (Nicein subunit alpha)	LAMA3 LAMNA	3333
P12111	Collagen alpha-3(VI) chain	COL6A3	3177
P12110	Collagen alpha-2(VI) chain	COL6A2	1019
P27918	Properdin (Complement factor P)	CFP PFC	469
P12109	Collagen alpha-1(VI) chain	COL6A1	1028

TABLE I: UniProt Output

Entry	Protein names	Gene names	Length
Q16610	Extracellular matrix protein 1 (Secretory component p85)	ECM1	540
P59665	Neutrophil defensin 1 (Defensin, alpha 1) (HNP-1) (HP-1) (HP1) [Cleaved into: HP 1-56; Neutrophil defensin 2 (HNP-2) (HP-2) (HP2)]	DEFA1 DEF1 DEFA2 MRS; DEFA1B	94
P20774	Mimecan (Osteoglycin) (Osteoinductive factor) (OIF)	OGN OIF SLRR3A	298
P20908	Collagen alpha-1(V) chain	COL5A1	1838
Q16270	Insulin-like growth factor-binding protein 7 (IBP-7) (IGF-binding protein 7) (IGFBP-7) (IGFBP-rP1) (MAC25 protein) (PGI2-stimulating factor) (Prostacyclin-stimulating factor) (Tumor-derived adhesion factor) (TAF)	IGFBP7 MAC25 PSF	282
Q08431	Lactadherin (Breast epithelial antigen BA46) (HMFG) (MFGM) (Milk fat globule-EGF factor 8) (MFG-E8) (SED1) [Cleaved into: Lactadherin short form; Medin]	MFGE8	387
P25067	Collagen alpha-2(VIII) chain (Endothelial collagen)	COL8A2	703
O75056	Syndecan-3 (SYND3)	SDC3 KIAA0468	442
Q14767	Latent-transforming growth factor beta-binding protein 2 (LTBP-2)	LTBP2 C14orf141 LTBP3	1821
Q14766	Latent-transforming growth factor beta-binding protein 1 (LTBP-1) (Transforming growth factor beta-1-binding protein 1) (TGF-beta1-BP-1)	LTBP1	1721

IV. DISCUSSION

With this centralized personal knowledge base, information is readily available to find gene and gene product interactions in the ECM and surrounding cells that can be analyzed as targets for further research. Information about their respective roles in disease such as

13

atherosclerosis can be efficiently and effectively gleaned because UniProt links dozens of databases and tools together in one interface. After clinical or wet-lab research is completed, the data obtained from these experiments can then be used to further expand the knowledge base. Tools such as GO Enrichment Analysis provide easy ways to draw even more conclusions from clinical research as well.

V. CONCLUSION

Once a workflow is created, relevant data can be quickly obtained for any research area of interest by properly utilizing tools that have been made available by the scientific community. This personalized database can be used as a central knowledge base to effectively explore potential routes for research. This information is indispensable for either finding limitations in current knowledge and performing experiments that fills in these knowledge gaps, or using this centralized knowledge to draw conclusions that can lead to further progressive research.

---

[1] R Mazurek, JM Dave, RR Chandran, A Misra, AQ Sheikh, and DM Greif. Vascular cells in blood vessel wall development and disease. *Advances in pharmacology (San Diego, Calif.)*, 78:323—350, 2017.

[2] Junyan Xu and Guo-Ping Shi. Vascular wall extracellular matrix proteins and vascular diseases. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(11):2106–2119, nov 2014.