Comparative Features of Multicellular Eukaryotic Genomes (2017)

(First three statistics from www.ensembl.org; other from original papers)

# of transcripts is a **Function of resources

Alternate splicing

Differences

	C. elegans	*A. thaliana*	*D. melanogaster*	*M. musculus*	*H. sapiens*
Species name	Nematode	Thale Cress	Fruit Fly	Mouse	Human
Size (Mb)	103	136	143	3,482	3,555
# Protein-coding genes	20,362	27,655 (25,498 original est.)	13,918 (13,601 original est.)	22,598 (30,000 original est.)	20,338 (30,000 original est.)
Transcripts	58,941	55,157	34,749	**131,195**	**200,310**
Gene density (#/kb)	1/5	1/4.5	1/8.8	1/83	1/97
LINE/SINE (%)	0.4	0.5	0.7	**27.4**	**33.6**
LTR (%)	0.0	4.8	1.5	9.9	8.6
DNA Elements	**5.3**	**5.1**	0.7	0.9	3.1
Total repeats	6.5	10.5	3.1	**38.6**	**46.4**
Exons					
% genome size	27	28.8	24.0		
per gene	4.0	5.4	4.1	8.4	8.7
average size (bp)		250	506		
Introns					
% genome size		15.6			
average size (bp)		168			

Arabidopsis Chromosome Structures
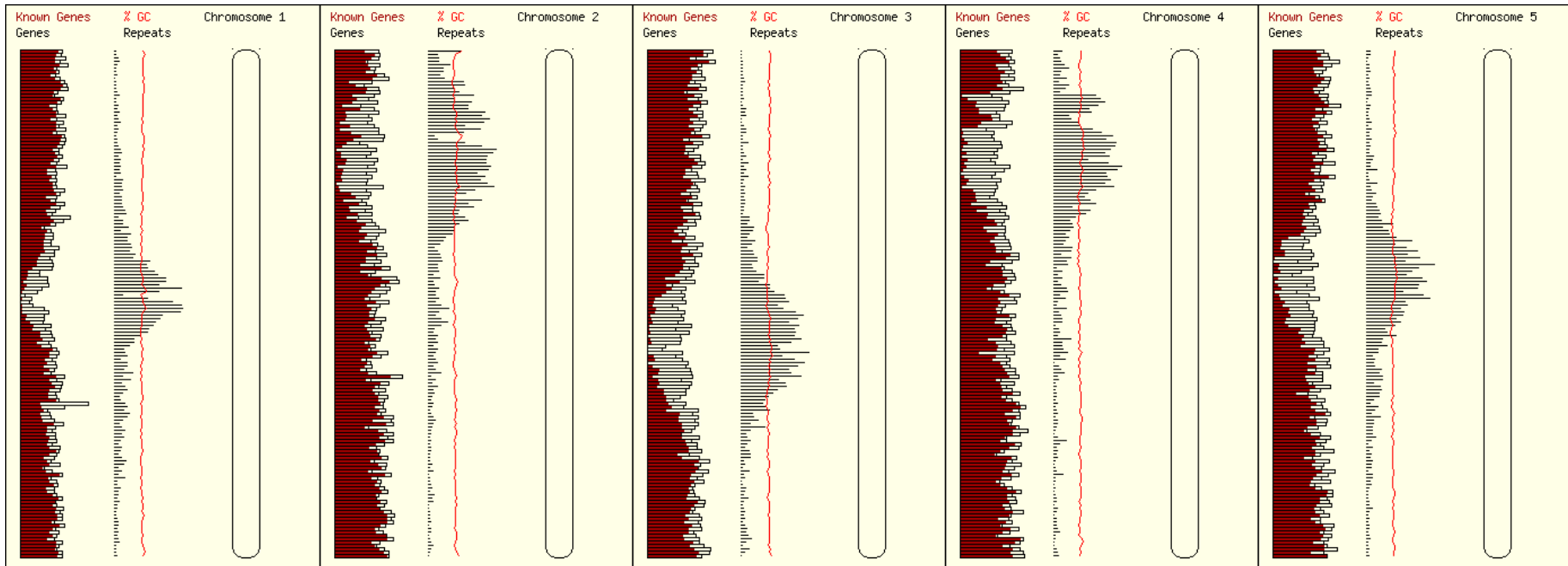
Sorghum Whole Genome Details
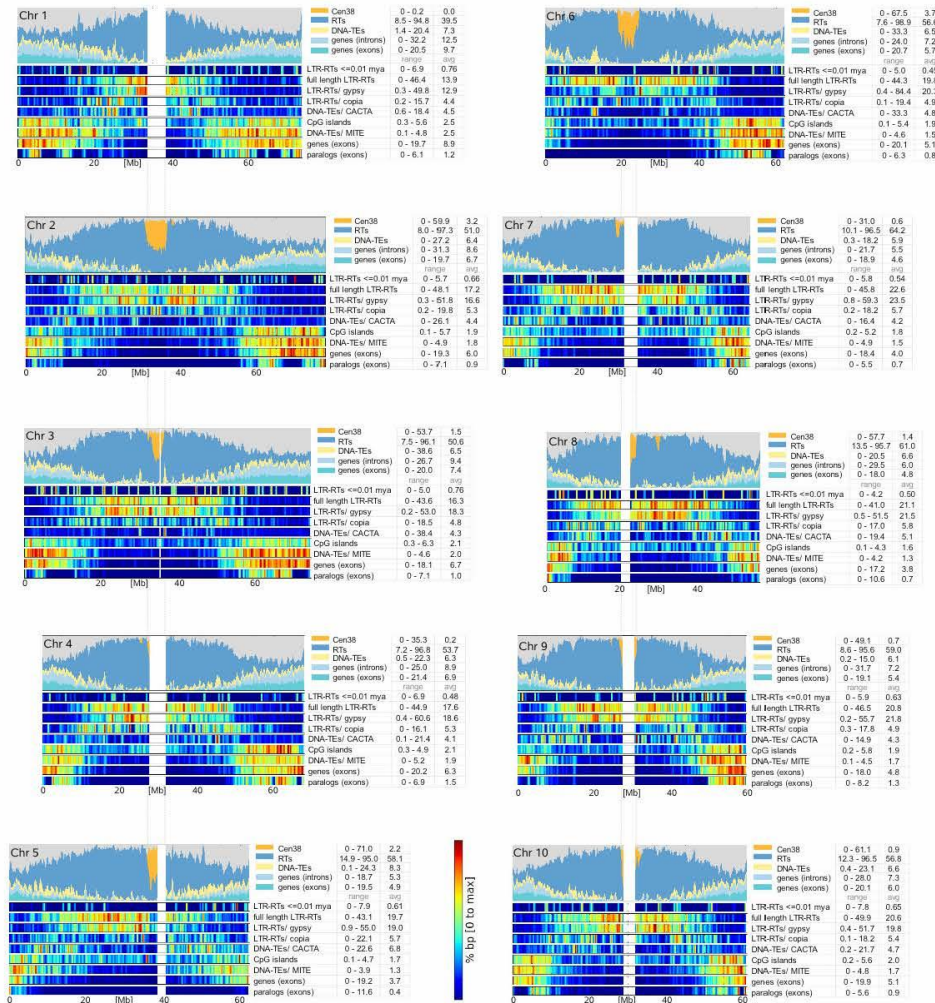
Characterizing the Proteome

The Protein World

- Sequencing has defined
  - Many, many proteins
- How can we use this data to:
  - ***Define genes in new genomes***
  - ***Look for evolutionarily related genes***
  - ***Follow evolution of genes***
    - Mixing of domains to create new proteins
  - ***Uncover important subsets of genes that***
    - That deep phylogenies
      - Plants vs. animals
      - Placental vs. non-placental animals
      - Monocots vs. dicots plants
- **Common nomenclature needed**
  - Ensure consistency of interpretations

InterPro (http://www.ebi.ac.uk/interpro/)
Classification of Protein Families

- **Intergrated documentation resource for protein super families, families, domains and functional sites**
  - *Mitchell AL, Attwood TK, Babbitt PC, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research, Jan 2019, (doi: 10.1093/nar/gky1100)*
    - A database of protein families, domains, and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences
      - **A database that collapses information from 11 other databases**

- **Current Release: 7 October 2020**
  - 37,978 entries
    - Up from 33,283 in Sept 2017
  - 14,223 InterPro entries
  - Match to 33,283 GO functions
  - Based on 195,667,571 UniProtKD proteins
    - Up from 90,606,305 in 2017

Categories
**Molecular Function
**Cellular Component
**Biological Process

October 2022
****40,957 entries**

Large increase in a few years

InterPro Entries Defined as:

<div style="border: 2px solid blue;">

Homologous Superfamiliy (3,326; new entry 10/2020)

</div>

- "..signatures that match proteins sharing a common evolutionary origin, as indicated by their structural similarities."

Example **P-type ATPase, A domain superfamily**

P-ATPases (also known as E1-E2 ATPases) ([intenz:3.6.3.-]) are found in bacteria and in a number of eukaryotic plasma membranes and organelles
[1]

P-ATPases function to transport a variety of different compounds, including ions and phospholipids, across a membrane using ATP hydrolysis for energy. There are many different classes of P-ATPases, which transport specific types of ion: H^+, Na^+, K^+, Mg^{2+}, Ca^{2+}, Ag^+ and Ag^{2+}, Zn^{2+}, Co^{2+}, Pb^{2+}, Ni^{2+}, Cd^{2+}, Cu^+ and Cu^{2+}. P-ATPases can be composed of one or two polypeptides, and can usually assume two main conformations called E1 and E2.

Transmembrane ATPases are membrane-bound enzyme complexes/ion transporters that use ATP hydrolysis to drive the transport of protons across a membrane. Some transmembrane ATPases also work in reverse, harnessing the energy from a proton gradient, using the flux of ions across the membrane via the ATPase proton channel to drive the synthesis of ATP.

Family (22,422 on 10/20)

- **Evolutionarily related proteins**
- **Share architecture base on:**
  - **Domain signature**
  - **Repeat signature**
- Multiple signatures can define a family
- 7/2003 release: 6.416 families
- Examples

IPR000003	Retinoid X Receptor
IPR000006	Vertebrate metallothionein
IPR000007	Tubby
IPR000009	Protein phosphate 2A regulatory subunit PR55
IPR000010	Cysteine protease inhibitor

Example Families

Cytochrome P450

The cytochrome P450 enzymes constitute a superfamily of haem-thiolate proteins. P450 enzymes usually act as terminal oxidases in multicomponent electron transfer chains, called P450-containing monooxygenase systems and are involved in metabolism of a plethora of both exogenous and endogenous compounds. P450-containing monooxygenase systems primarily fall into two major classes: bacterial/mitochondrial (type I), and microsomal (type II). All P450 enzymes can be categorised into two main groups, the so-called B- and E-classes: P450 proteins of prokaryotic 3-component systems and fungal P450nor (CYP55) belong to the B-class; all other known P450 proteins from distinct systems are of the E-class [1].

Disease resistance protein

Plants are attacked by a range of phytopathegenic organisms, including viruses, mycoplasma, bacteria, fungi, nematodes, protozoa and parasites. Resistance to a pathogen is manifested in several ways and is often correlated with a hypersensitive response (HR), localised induced cell death in the host plant at the site of infection [1]. The induction of the plant defense response that leads to HR is initiated by the plants recognition of specific signal molecules (elicitors) produced by the pathogen; R genes are thought to encode receptors for these elicitors. RPS2, N and L6 genes confer resistance to bacterial, viral and fungal pathogens.

- **Independent structural unit**
- **Can be associated with other**
  - Domains
  - Repeat
- Proteins with the domain are evolutionarily related
- 7/2003 release: 1,902 domains
- Examples

IPR000001	Kringle
IPR000002	Cdc20/Fizzy
IPR000005	Helix-turn-helix, AraCtype
IPR000008	C2 domain
IPR000014	PAS domain

Example domains

Protein kinase-like; Protein kinase

==Protein kinases (<u>IPR000719</u>) catalyze the phosphotransfer reaction fundamental to most signalling and regulatory processes in the eukaryotic cell== [1]. The catalytic subunit contains a core that is common to both ==serine/threonine and tyrosine protein kinases.== The catalytic domain contains the nucleotide-binding site and the catalytic apparatus in an inter-lobe cleft. Structurally it shares functional and structural similarities with the ATP-grasp fold, which is found in enzymes that catalyse the formation of an amide bond, and with PIPK (phosphoinositol phosphate kinase). The three-dimensional fold of the protein kinase catalytic domain is similar to domains found in several other proteins. These include the catalytic domain of actin-fragmin kinase, an atypical protein kinase that regulates the F-actin capping activity in plasmodia [2]; the catalytic domain of phosphoinositide-3-kinase (PI3K), which phosphorylates phosphoinositides and as such is involved in a number of fundamental cellular processes such as apoptosis, proliferation, motility and adhesion [3]; the catalytic domain of the MHCK/EF2 kinase, an atypical protein kinase that includes the TRP (transient channel potential) calcium-channel kinase involved in the modulation of calcium channels in eukaryotic cells in response to external signals [4]; choline kinase, which catalyses the ATP-dependent phosphorylation of choline during the biosynthesis of phosphatidylcholine [5]; and 3',5'-aminoglycoside phosphotransferase type IIIa, a bacterial enzyme that confers resistance to a range of aminoglycoside antibiotics [6].

Signaling pathways

- **Region not expected to form a globular structure on their own**
  - WD40
    - Six to eight copies needed for globular domain to form
- 7/2003 release: 163 repeats
- Examples

IPR000033	Low-density lipoprotein receptor, YWTD repeat
IPR000102	Neuraxin/MAP1B repeat
IPR000127	Ubiquitin-activating enzyme repeat
IPR000225	Armadillo repeat
IPR000258	Bacterial ice-nucleation proteins octamer repeat

Example Repeats

Leucine-rich repeat

Leucine-rich repeats (LRR) consist of 2-45 motifs of 20-30 amino acids in length that generally folds into an arc or horseshoe shape [1]. LRRs occur in proteins ranging from viruses to eukaryotes, and appear to provide a structural framework for the formation of protein-protein interactions [2]. Proteins containing LRRs include tyrosine kinase receptors, cell-adhesion molecules, virulence factors, and extracellular matrix-binding glycoproteins, and are involved in a variety of biological processes, including signal transduction, cell adhesion, DNA repair, recombination, transcription, RNA processing, disease resistance, apoptosis, and the immune response.

Sequence analyses of LRR proteins suggested the existence of several different subfamilies of LRRs. The significance of this classification is that repeats from different subfamilies never occur simultaneously and have most probably evolved independently. It is, however, now clear that all major classes of LRR have curved horseshoe structures with a parallel beta sheet on the concave side and mostly helical elements on the convex side. At least six families of LRR proteins, characterized by different lengths and consensus sequences of the repeats, have been identified. Eleven-residue segments of the LRRs (LxxLxLxxN/CxL), corresponding to the ß-strand and adjacent loop regions, are conserved in LRR proteins, whereas the remaining parts of the repeats (herein termed variable) may be very different. Despite the differences, each of the variable parts contains two half-turns at both ends and a "linear" segment (as the chain follows a linear path overall), usually formed by a helix, in the middle. The concave face and the adjacent loops are the most common protein interaction surfaces on LRR proteins. 3D structure of some LRR proteins-ligand complexes show that the concave surface of LRR domain is ideal for interaction with alpha-helix, thus supporting earlier conclusions that the elongated and curved LRR structure provides an outstanding framework for achieving diverse protein-protein interactions [2]. Molecular modeling suggests that the conserved pattern LxxLxL, which is shorter than the previously proposed LxxLxLxxN/CxL is sufficient to impart the characteristic horseshoe curvature to proteins with 20- to 30-residue repeats [3].

PPR repeat

Pentatricopeptide repeat proteins are characterised by the presence of a tandem array of repeats, where the number of PPR motifs controls the affinity and specificity of the PPR protein for RNA. These proteins occur predominantly in plants, where they appear to play essential roles in RNA/DNA metabolism in mitochondria and chloroplasts [1]. It has been suggested that each of the highly variable PPR proteins is a gene-specific regulator of plant organellar RNA metabolism. PPR proteins may also play a role in organelle biogenesis, probably via binding to organellar transcripts [2]. Examples of PPR repeat-containing proteins include PET309 P32522 , which may be involved in RNA stabilisation [3], and crp1, which is involved in RNA processing [4]. The repeat is associated with a predicted plant protein O49549 that has a domain organization similar to the human BRCA1 protein.

Site (812 on 9/17)

==Post-translational modification site== **(17 on 10/20; same as 9/17)**
- ○ **Modifies primary protein structure**
  - ▪ ==Glycosylation==
  - ▪ Splicing
  - ▪ ==Phosphorylation==
- 7/2003 release: 20 post-translational modifications
- Examples

IPR000042	N-glycosylation site
IPR000134	Amidation site
IPR000152	Aspartic acid and asparagin hydroxylation site
IPR000220	Tyrosine kinase phosphorylation site
IPR000338	N-myristoylation site

==Binding site== **(76 on 10/20; same as 9/17)**
- ○ **Binds chemical compounds**
  - ▪ Compounds not substrates
  - ▪ Cofactors required for protein function
- 7/2003 release: 20 binding sites
- Examples

IPR000205	NAD-binding site
IPR000214	Formamidopyrimidine-DNA glycolase, Zn-binding site
IPR000345	Cytochrome c heme-binding site
IPR000634	Serine/threonine dhydrate, pyridoxal-phosphate-binding site
IPR001216	Cysteine synthase/cystathionine beta-synthase P-phosphate binding site

==Active site== **(132 on 9/17; same as 9/17)**
- ○ **Catalytic sites of protein**
  - ▪ Site where substrate binds
- 7/2003 release: 26 active sites
- Examples

IPR000126	Serine protease V8, active site
IPR000138	Hydroxymethylglutaryl-coenzyme A lyase, active site
IPR000180	Peptidase M19, renal dipeptidase, active site
IPR000189	Prokaryotic transglycosylase, active site
IPR000590	Hydroxymethylglutaryl-coenzyme A synthase, active site

==Conserved Site== (687 on 9/17)

- o **A site motif specific defined by PROSITE that characterizes a protein or protein family**
    - ▪ Not an PTM, binding site, or active site
- 9/2/08 release: 166 conserved sites
- Examples

IPR000035	Alkylbase DNA glycosidase, conserved site
IPR000059	NUDIX hydrolase, conserved site
IPR000132	Nitrilase/cyanide hydratase, conserved site
IPR000283	NADH dehydrogenase 75 kDa subunit, conserved sit
IPR000291	D-alanine--D-alanine ligase/VANA/B/C, conserved site

InterPro Strategy
(https://www.ebi.ac.uk/interpro/about.html#about_08)

- Uses nomenclature developed by other databases
- Collapses those nomenclatures to define each InterPro entry

The InterPro Consortium
(https://www.ebi.ac.uk/interpro/about.html)

The following databases make up the InterPro Consortium:

GENE3D	CATH-Gene3D database describes protein families and domain architectures in complete genomes. Protein families are formed using a Markov clustering algorithm, followed by multi-linkage clustering according to sequence identity. Mapping of predicted structure and sequence domains is undertaken using hidden Markov models libraries representing CATH and Pfam domains. CATH-Gene3D is based at University College, London, UK.
Conserved Domains	CDD is a protein annotation resource that consists of a collection of annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domain models, which use 3D-structure information to explicitly define domain boundaries and provide insights into sequence/structure/function relationships, as well as domain models imported from a number of external source databases.
MobiDB	MobiDB offers a centralized resource for annotations of intrinsic protein disorder. The database features three levels of annotation: manually curated, indirect and predicted. The different sources present a clear tradeoff between quality and coverage. By combining them all into a consensus annotation, MobiDB aims at giving the best possible picture of the "disorder landscape" of a given protein of interest.
HAMAP	HAMAP stands for High-quality Automated and Manual Annotation of Proteins. HAMAP profiles are manually created by expert curators. They identify proteins that are part of well-conserved proteins families or subfamilies. HAMAP is based at the SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.
PANTHER	PANTHER is a large collection of protein families that have been subdivided into functionally related subfamilies, using human expertise. These subfamilies model the divergence of specific functions within protein families, allowing more accurate association with function, as well as inference of amino acids important for functional specificity. Hidden Markov models (HMMs) are built for each family and subfamily for classifying additional protein sequences. PANTHER is based at at University of Southern California, CA, US.
Pfam	Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains. Pfam is based at EMBL-EBI, Hinxton, UK.
PIRSF	PIRSF protein classification system is a network with multiple levels of sequence diversity from superfamilies to subfamilies that reflects the evolutionary relationship of full-length proteins and domains. PIRSF is based at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, US.

	PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family or domain. PRINTS is based at the University of Manchester, UK.
	ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches. ProDom is based at PRABI Villeurbanne, France.
	PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family a new sequence belongs. PROSITE is base at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland.
	SFLD (Structure-Function Linkage Database) is a hierarchical classification of enzymes that relates specific sequence-structure features to specific chemical capabilities.
	SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. SMART is based at at EMBL, Heidelberg, Germany.
	SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure. The library is based on the SCOP classification of proteins: each model corresponds to a SCOP domain and aims to represent the entire SCOP superfamily that the domain belongs to. SUPERFAMILY is based at the University of Bristol, UK.
	TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, hidden Markov models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology. TIGRFAMs is based at the J. Craig Venter Institute, Rockville, MD, US.

Pfam Database
http://pfam.xfam.org/

The protein world is vast
- GenBank RefSeq
- Proteins 45,166,402 proteins

Pfam – the Protein Family Database

The new standard for defining what families a protein belongs to
Helpful to determine the potential function of a protein

Pfam Families
- Entry types
  - **Family**
    - A group of proteins that are related by structure
  - **Domain**
    - A short amino acid sequence
  - **Repeat**
    - A short amino acid sequence that functions when a group of the sequences are clustered
  - **Motif**
    - A short amino acid that defines a specific feature

Two families
- Pfam-A
  - High reliable family domain structure
  - Repeated over many proteins and considered to be "real" and associated with function
  - Annotated
- Pfam-B (discontinued in 2016 release)
  - Family domains of lower quality
  - Fewer proteins contain the family domain
  - Not annotated

Common bean FLS2 ortholog Pfam structure

Sequence search results

Show the detailed description of this results page.

We found **3** Pfam-A matches to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.

Show the search options and sequence that you submitted.

[Return](#) to the search form to look for Pfam domains on a new sequence.

Output from HMMR3

(Bold lines are those outputted from the Pfam site. Pfam just returns "high" quality hits.)

seq_id	alignment _start	alignment_ end	envelope _start	envelope _end	hmm_acc	hmm_name	type	hmm_start	hmm_end	hmm_length	bit_score	E-value
Phvul.002G196200	**31**	**68**	**28**	**68**	**PF08263**	**LRRNT_2**	**Family**	**4**	**43**	**43**	**44.8**	**7.60E-12**
Phvul.002G196200	120	156	120	164	PF12799	LRR_4	Family	1	36	44	25.6	5.90E-06
Phvul.002G196200	218	253	217	259	PF12799	LRR_4	Family	3	37	44	23.3	3.10E-05
Phvul.002G196200	264	324	264	324	PF13855	LRR_8	Repeat	1	61	61	23.9	2.30E-05
Phvul.002G196200	337	373	336	381	PF12799	LRR_4	Family	2	37	44	18.6	0.00096
Phvul.002G196200	385	421	384	427	PF12799	LRR_4	Family	2	37	44	21	0.00016
Phvul.002G196200	432	471	432	481	PF12799	LRR_4	Family	1	39	44	22.6	5.00E-05
Phvul.002G196200	481	518	480	526	PF12799	LRR_4	Family	2	38	44	21.7	0.0001
Phvul.002G196200	578	638	578	638	PF13855	LRR_8	Repeat	3	61	61	29.8	3.20E-07
Phvul.002G196200	**654**	**687**	**650**	**695**	**PF12799**	**LRR_4**	**Family**	**5**	**37**	**44**	**21.8**	**9.30E-05**
Phvul.002G196200	701	759	699	759	PF13855	LRR_8	Repeat	3	61	61	35.5	5.50E-09
Phvul.002G196200	**870**	**1153**	**867**	**1156**	**PF00069**	**Pkinase**	**Domain**	**4**	**256**	**260**	**145.2**	**1.70E-42**

[The Pfam protein families database](#): R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L.L. Sonnhammer, J. Tate, M. Punta Nucleic Acids Research (2014) Database Issue 42:D222-D230

Phytozome: Integrates Multiple Data Into Display
Phaseolus vulgaris FLS2 gene output

InterPro Data

Arabidopsis Top 15 Entries (10/2003)

Rank	No. proteins (% proteome)	InterPro Entry (type)
1	1050 (4.0)	Protein kinase (family)
2	796 (3.1)	Serine/threonine protein kinase active site (active site)
3	648 (2.5)	Cyclin-like F-box (domain)
4	518 (2.0)	Leucine-rich repeat (repeat)
5	491 (1.9)	Zn-finger, RING (domain)
6	454 (1.7)	PPR repeat (Repeat)
7	345 (1.3)	Myb DNA-binding protein (domain)
8	342 (1.3)	Tyrosine protein kinase (family)
9	332 (1.3)	Leucine-rich repeat, plant specific (repeat)
10	314 (1.2)	AAA ATPase (domain)
11	300 (1.2)	RNA-binding region RNP-1 (RNA recognition motif) (domain)
12	267 (1.0)	G-protein beta WD40 repeat (repeat)
13	252 (1.0)	Cytochromoe P450 (family)
14	233 (0.9)	Zn-finger, CCHC type (domain)
15	227 (0.9)	E-class P450, group I (family)

Arabidopsis Top 15 Families (10/2003)

Rank	No. proteins	InterPro Entry
1	1050	Protein kinase
2	252	Cytochrome P450
3	157	Disease resistance protein
4	114	2OG-Fe(II) oxygenase superfamily
5	110	Retrotransposon gag protein
6	110	Lipolytic enzyme, G-DE-S-L family
7	109	UDG-glucoronosyl/UDP-glucosyl transferase
8	108	No apical meristem (NAM) protein
9	106	Ras GTPase superfamily
10	96	Haem peroxidase
11	91	Short-chain dehyrogenase/reductase SDR
12	91	General substrate transporter
13	85	Mitochondrial substrate carrier
14	84	Major facilitator superfamily
15	81	Transcriptional factor B3

Arabidopsis Top 15 Domains (10/2003)

Rank	Proteins matched (Matches/genome)	InterPro Entry (type)
1	648 (1487)	Cyclin-like F-box
2	491 (1287)	Zn-finger, RING
3	345 (1535)	Myb DNA-binding domain
4	314 (390)	AAA ATPase
5	300 (1302)	RNA-binding region RNP-1 (RNA recognition motif
6	233 (923)	Zn-finger, CCHC type
7	224 (647)	Basic helix-loop-helix dimerization domain bHLH
8	222 (1546)	Calcium-binding EF-hand
9	213 (223)	F-box protein interaction domain
10	210 (213)	Esterase/lipase/thioesterase
11	196 (852)	Zn-finger,C2H2 type
12	164 (472)	Zn-finger-like, PHD finger
13	159 (192)	FBD
14	154 (157)	NB-ARC domain
15	153 (158)	Integrase, catalytic domain

Arabidopsis Top 15 Repeats (10/2003)

Rank	No. proteins	InterPro Entry
1	518	Leucine-rich repeat
2	454	PPR repeat
3	267	G-protein beta WD-40 repeat
4	136	TPR repeat
5	117	Ankyrin
6	115	Kelch repeat
7	70	Armadillo repeat
8	62	Parallel beta-helix repeat
9	28	Paired amphipathic helix
10	26	Bacterail transferase hexapeptide repeat
11	26	Leucine-rich repeat, cysteine-containing
12	24	Protein of unknown function DUF321
13	20	Prenyltransferae/squalene oxidase
14	18	Late embryogenesis abundant protein
15	17	HEAT repeat

Are their plant specific families?

Rank	InterPro entry	*A. thaliana* # proteins	*C. elegans* # proteins	*D. melanogaster* # proteins	*H. sapiens* # proteins	*S. cervisiae* # proteins
1	Protein kinase	1050	528	303	694	118
2	Cytochrome P450	252	83	88	90	3
3	*Disease resistance protein*	*157*	*1*	*0*	*4*	*2*
4	2OG-Fe(II) oxygenase superfamily	114	11	26	19	1
5	*Retrotransposon gag protein*	*110*	*5*	*3*	*4*	*3*
6	Lipolytic enzyme, G-DE-S-L family	110	11	2	5	1
7	UDG-glucoronosyl/UDP-glucosyl transferase	109	76	35	29	0
8	*No apical meristem (NAM) protein*	*108*	*0*	*0*	*0*	*0*
9	Ras GTPase superfamily	106	82	86	118	34
10	Haem peroxidase	96	26	16	32	4
11	Short-chain dehyrogenase/reductase SDR	91	92	60	80	13
12	General substrate transporter	91	96	118	77	54
13	Mitochondrial substrate carrier	85	65	74	93	38
14	Major facilitator superfamily	84	128	123	106	670
15	*Transcriptional factor B3*	*81*	*0*	*0*	*0*	*0*

Are their plant specific domains? [No. proteins (Matches/proteome)]

Rank	InterPro entry	*A. thaliana* # proteins	*C. elegans* # proteins	*D. melanogaster* # proteins	*H. sapiens* # proteins	*S. cervisiae* # proteins
1	Cyclin-like F-box	648 (1487)	441 (957)	37 (78)	86 (204)	14 (32)
2	Zn-finger, RING	491 (1287)	175 (504)	150 (376)	387 (1194)	40 (108)
3	Myb DNA-binding domain	345 (1535)	39 (112)	50 (103)	92 (276)	21 (91)
4	AAA ATPase	314 (390)	124 (173)	131 (193)	176 (251)	89 (125)
5	RNA-binding region RNP-1 (RNA recognition motif	300 (1302)	131 (193)	176 (251)	346 (1809)	58 (307)
6	Zn-finger, CCHC type	233 (923)	47 (210)	27 (111)	56 (308)	14 (89)
7	Basic helix-loop-helix dimerization domain bHLH	224 (647)	101 (271)	93 (307)	168 (635)	9 (36)
8	Calcium-binding EF-hand	222 (1546)	133 (813)	134 (873)	316 (1945)	17 (102)
9	*F-box protein interaction domain*	*213 (223)*	*0 (0)*	*0 (0)*	*0 (0)*	*0 (0)*
10	Esterase/lipase/thioesterase	210 (213)	134 (142)	148 (149)	121 (121)	37 (37)
11	Zn-finger,C2H2 type	196 (852)	251 (2042)	397 (7249)	962 (29,334)	54 (381)
12	Zn-finger-like, PHD finger	164 (472)	46 (146)	51 (203)	126 (529)	16 (57)
13	*FBD*	*159 (192)*	*1 (1)*	*2 (2)*	*1 (1)*	*0 (0)*
14	*NB-ARC domain*	*154 (157)*	*0 (0)*	*0 (0)*	*0 (0)*	*0 (0)*
15	Integrase, catalytic domain	153 (158)	30 (31)	6 (6)	15 (15)	40 (40)

Are their plant specific repeats?

Rank	InterPro entry	*A. thaliana* # proteins	*C. elegans* # proteins	*D. melanogaster* # proteins	*H. sapiens* # proteins	*S. cervisiae* # proteins
1	Leucine-rich repeat	518	85	127	269	11
2	*PPR repeat*	*454*	*1*	*4*	*5*	*2*
3	G-protein beta WD-40 repeat	267	170	195	405	100
4	TPR repeat	136	64	92	189	33
5	Ankyrin	117	116	104	306	21
6	Kelch repeat	115	17	22	97	6
7	Armadillo repeat	70	7	11	47	2
8	*Parallel beta-helix repeat*	*62*	*3*	*2*	*9*	*1*
9	Paired amphipathic helix	28	1	3	3	1
10	Bacterail transferase hexapeptide repeat	26	7	5	8	4
11	Leucine-rich repeat, cysteine-containing	26	7	10	21	3
12	*Protein of unknown function DUF321*	*24*	*0*	*0*	*0*	*0*
13	Prenyltransferae/squalene oxidase	20	3	3	4	4
14	*Late embryogenesis abundant protein*	*18*	*4*	*1*	*2*	*0*
15	HEAT repeat	17	14	6	30	10

Examples of Arabidopsis Genes With Known Phenotypic Functions (Same structure; many functions)

Family	Gene	Phenotype
Protein kinase	*YDA* (At1g63700)	Defective embryo and seedling
Cytochrome P450	*REF8* (At2g40890)	Reduced phenylpropanoid metabolites and lignin
Disease resistance (LRR)	*RPM1* (At3g07040)	Resistance to bacterial pathogens
Glycosyl Transferase	*QUA1* (At3g25410)	Dwarf plant reduced pectin

Domain		
Myb-transcription factor	*AS1* (At2g37630)	Altered leaf morphology
	CPC (At2g46410)	Reduced root hairs
	LAF1 (At4g25560)	Elongated hypocotyls in far-red light
Basic helix-loop-helix	*HFR1* (At1g02340)	Elongated hypocotyls in far-red light
	SPT (At4g36930)	Altered carpel development

Repeats		
Leucine-rice repeat	*BRL1* (At1g55610)	Altered vascular cell differentiation
	RPP1 (At3g44480)	Fungal disease resistance
Ankyrin	*EMB506* (At5g40160)	Embryo defective
	CAO (At2g47450)	Pale, chlorotic plants
TPR	*HBT* (At2g20000)	Altered root meristem development
	SPY (At3g11540)	Spindly plants
WD40	*TTG1* (At5g24520)	Yellow seed coat

from: Meinke, D.W., L.K. Meinke, T.C. Showalter, A.M. Schissel, L.A. Mueller, and I. Tzafrir. 2003. A sequence-based map of Arabidopsis genes with mutant phenotypes. Pl. Physiol. 131:409-418.

Gene Ontology (GO) (http://www.geneontology.org/)

"The goal of the Gene Ontology Consortium is to produce a controlled vocabulary that be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing."

Major Categories

Molecular Function Ontology
- "The action characteristic of a gene product."

Biological Process Ontology
- A phenomenon marked by changes that lead to a particular result, mediated by one or more gene products.

Cellular Component Ontology
- "The part of a cell of which a gene product is a component; for purpose of GO includes the extracellular environment of cells; a gene product may be a component of one or more parts of a cell; this term includes gene products that are parts of macromolecular complexes, by the definition that all members of a complex normally copurify under all except extreme conditions."

GO Annotations for Eukaryotic Species

Species	Biological process	Molecular function	Cellular component	Total genes associated
A. thaliana	6,712	7,967	13,532	18,572
C. elegans	5,115	5,755	3,057	6,925
D. melanogaster	4,439	6,795	3,942	7,938
H. sapiens	16,643	18.772	13,880	20,687
S. cerevasiae	6,646	6,434	6,435	6,448

Note: Each of these annotation was performed by a different group.