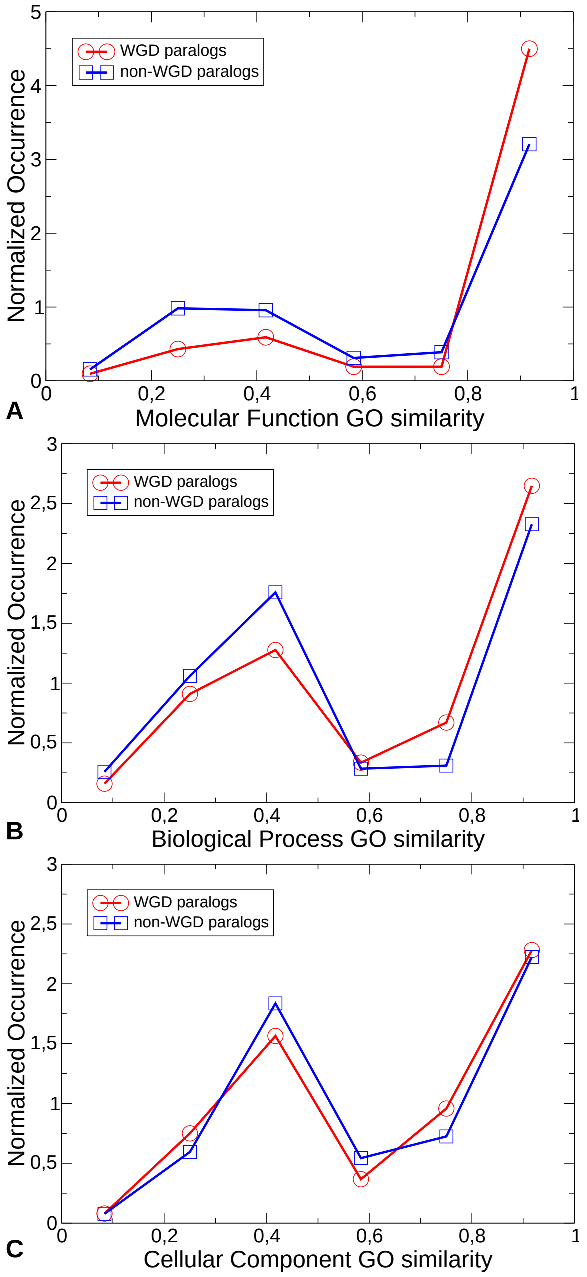

Figure S1: **Functional Similarity of WGD paralogs and non-WGD paralogs.** Normalized histograms of the Gene Ontology similarity between WGD and non-WGD duplicate pairs for the GO branches molecular function (A), biological process (B), cellular component (C). For all the three branches, WGD paralogs tend to have higher GO similarity scores than non-WGD paralogs.

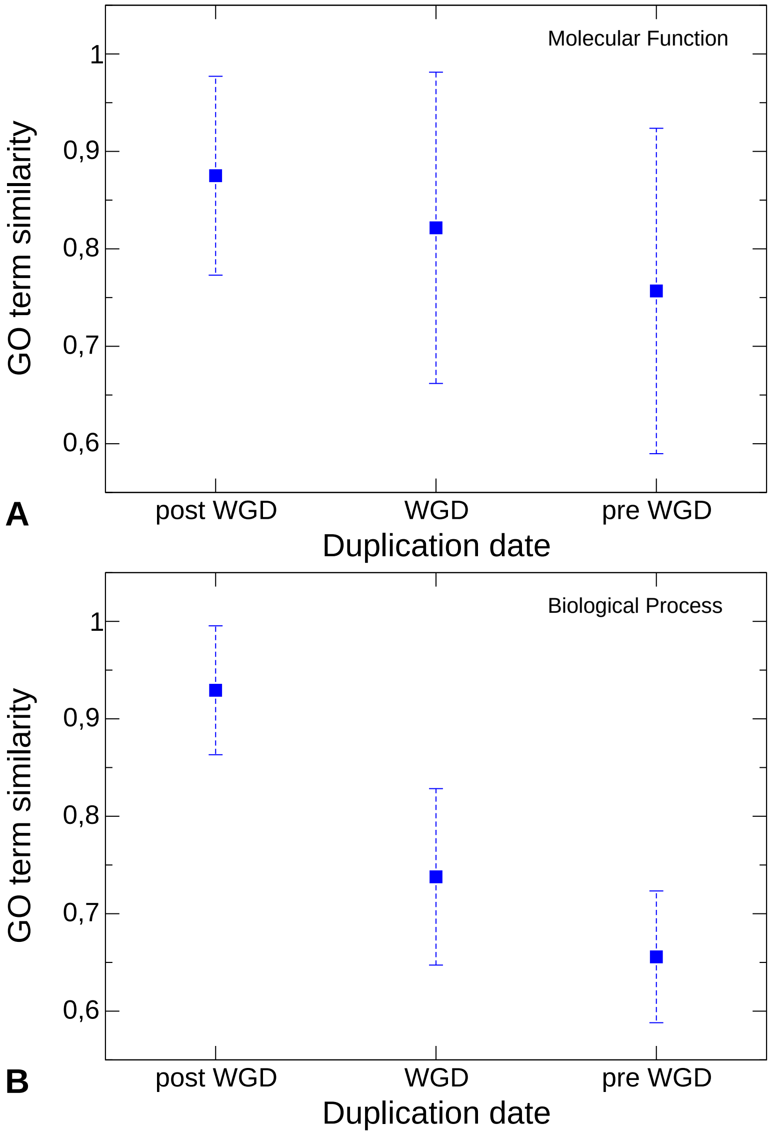

Figure S2: **Functional similarity of duplicates versus duplication age for manually curated GO annotations.** The plots report the mean (squares) and the standard deviation (error bars) of the GOsim similarity score between duplicates of the same age groups. The analysis was restricted only to the genes with experimental manually curated GO terms, grouping pre- and post-WGD duplication to gather sufficient statistics. This comparison is made for the GO branches: Biological Process (A), Molecular Function (B).

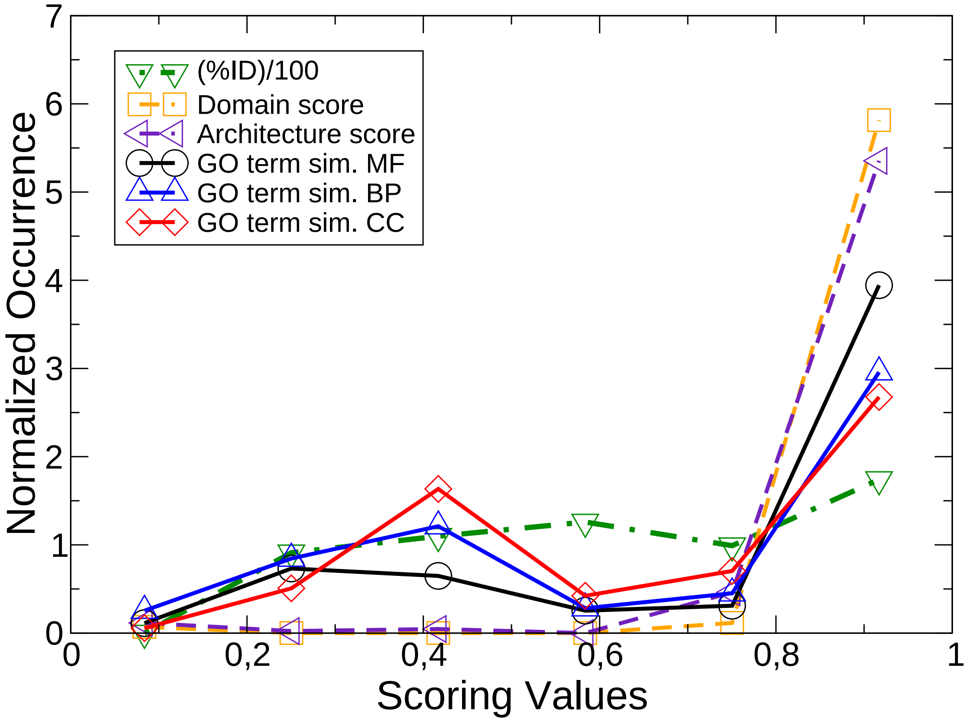

Figure S3: **Structural and functional divergence of paralogs with no gaps in the domain architecture.**. The plot reports histograms of sequence ID% retrieved from alignment, domain score, architecture score and GO term similarity (for all three branches) for all the paralog pairs with both proteins with by domain. Despite of this restriction we retrieve the same results shown in Figure 4 of the main text.

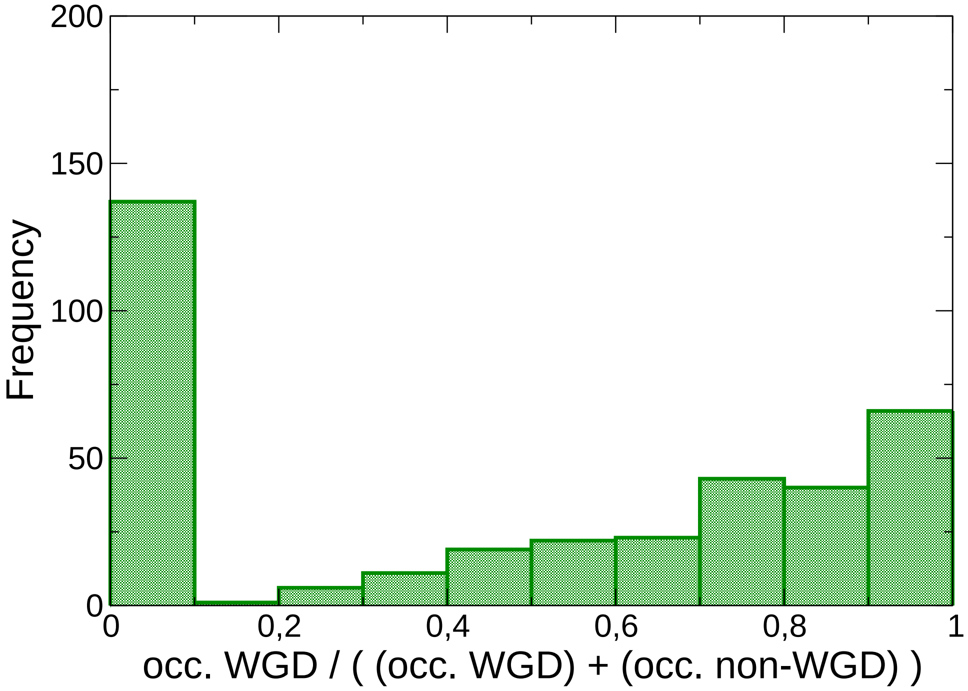

Figure S4: **Occurrence of domain topologies in WGD vs non-WGD duplicates.** For each SCOP domain, we calculated its occurrence in WGD proteins and non-WGD duplicates (normalized by the sizes of these two duplicate sets). The plot reports the histogram of the relative weight of occurrence of WGD duplicates, indicating the separation of two populations of domain topologies: domain topologies that appear in local duplications only (peak at zero), and those that appear in both the WGD and local duplications, having a preference towards the WGD (peak at one).

Gene Ontology analysis

As a control of the domain-based functional analysis of domain topologies involved in local duplications versus the WGD, we performed a more standard functional characterization based on Gene Ontology analysis on the proteins, along the lines of previous studies [23, 24, 38]. We considered the disjoint sets of WGD and non-WGD paralogs. For each set we extracted the over-represented GO terms, and we compared them looking for the terms shared between WGD and non WGD-paralogs or specifically connected to a group (over-represented in a group and not significantly present in the other). WGD and non-WGD paralogs are enriched in different GO terms. We performed the same analysis also on randomized sets. Two randomly assorted sets tend to share more over-represented GO terms than WGD paralogs and non-WGD paralogs. These results are inverted considering the terms specific for each group: differently from the random assorted groups, WGD paralogs and non-WGD paralogs have many exclusive genes (see Tables S1 and S2), indicating that WGD and non-WGD paralogs carry out different functions.

In accordance with the domain-based analysis and with the previous hierarchical analysis derived from expression profiles and functional annotations [24], we find that WGD paralogs are enriched for genes involved in ribosomes and translation, regulation of cell cycle, regulation of developmental processes, sporulation, NADP metabolic process. On the other side the non-WGD paralogs are enriched for genes involved in transport, amino acid transmembrane transport, cellular wall, vitamin metabolism.

Finally, a recent study by Guan and coworkers [19] found that WGD duplicates are more likely to share interaction partners and biological functions than non-WGD duplicates. To confirm the latter result, we analyzed the distribution of the GO similarity normalized histograms for all the pairs of the two disjoint sets. Indeed, WGD paralogs result slightly more similar than non-WGD paralogs for all the three GO branches (supplementary figure S1). On the other hand, comparing with figure 4, one notices that pre-WGD paralogs are less similar at the functional level, so that this signal might come at least in part from the functional difference of ancient non-WGD paralogs.

Gene Ontology terms exclusive of WGD-Paralogs

GO term	Number of genes	P-value	annotation
GO:0005737	571	3.62e-22	cytoplasm
GO:0009987	647	1.72e-21	cellular process
GO:0005622	675	8.80e-19	intracellular
GO:0044424	668	1.10e-17	intracellular part
GO:0005830	56	1.60e-17	cytosolic ribosome (sensu Eukaryota)
GO:0005840	97	6.97e-16	ribosome
GO:0005575	740	5.40e-15	cellular component
GO:0005829	92	5.57e-15	cytosol
GO:0044445	58	1.12e-14	cytosolic part
GO:0044464	737	2.86e-14	cell part
GO:0005623	737	3.11e-14	cell
GO:0016773	62	4.53e-14	phosphotransferase activity, alcohol group as acceptor
GO:0004674	49	5.75e-14	protein serine/threonine kinase activity
GO:0009059	138	6.01e-14	macromolecule biosynthetic process
GO:0004672	49	2.19e-13	protein kinase activity
GO:0016301	66	4.33e-13	kinase activity
GO:0003735	68	7.94e-13	structural constituent of ribosome
GO:0009058	203	1.13e-12	biosynthetic process
GO:0044262	69	1.45e-12	cellular carbohydrate metabolic process
GO:0004713	42	3.62e-12	protein-tyrosine kinase activity
GO:0065007	228	4.49e-12	biological regulation
GO:0005488	536	6.77e-12	binding
GO:0043284	31	7.92e-12	biopolymer biosynthetic process
GO:0000271	25	7.93e-12	polysaccharide biosynthetic process
GO:0006468	47	9.56e-12	protein amino acid phosphorylation
GO:0044444	383	3.28e-11	cytoplasmic part
GO:0007154	85	5.55e-11	cell communication
GO:0007165	80	9.09e-11	signal transduction
GO:0005843	26	1.60e-10	cytosolic small ribosomal subunit (sensu Eukaryota)
GO:0006412	94	3.37e-10	translation
GO:0032502	106	3.71e-10	developmental process
GO:0016051	33	5.93e-10	carbohydrate biosynthetic process
GO:0033279	56	1.01e-09	ribosomal subunit
GO:0008152	520	1.6e-09	metabolic process
GO:0050789	187	1.74e-09	regulation of biological process
GO:0046164	27	2.35e-09	alcohol catabolic process
GO:0006112	20	2.38e-09	energy reserve metabolic process
GO:0044249	152	3.72e-09	cellular biosynthetic process
GO:0044260	244	3.99e-09	cellular macromolecule metabolic process
GO:0016052	30	5.11e-09	carbohydrate catabolic process
GO:0044275	30	5.11e-09	cellular carbohydrate catabolic process
GO:0050794	181	6.51e-09	regulation of cellular process
GO:0016310	56	9.85e-09	phosphorylation
GO:0005842	27	1.09e-08	cytosolic large ribosomal subunit (sensu Eukaryota)
GO:0044237	485	1.35e-08	cellular metabolic process
GO:0006739	13	1.38e-08	NADP metabolic process
GO:0019320	24	1.41e-08	hexose catabolic process
GO:0044264	27	1.55e-08	cellular polysaccharide metabolic process
GO:0005976	27	1.55e-08	polysaccharide metabolic process
GO:0044238	478	1.57e-08	primary metabolic process
GO:0005516	11	1.86e-08	calmodulin binding
GO:0032989	62	1.91e-08	cellular structure morphogenesis
GO:0000902	62	1.91e-08	cell morphogenesis
GO:0006007	23	2.22e-08	glucose catabolic process
GO:0009250	16	2.50e-08	glucan biosynthetic process
GO:0006006	30	2.56e-08	glucose metabolic process
GO:0009653	62	2.63e-08	anatomical structure morphogenesis
GO:0005198	81	2.95e-08	structural molecule activity
GO:0005978	12	2.98e-08	glycogen biosynthetic process
GO:0006796	65	3.81e-08	phosphate metabolic process
GO:0006793	65	3.81e-08	phosphorus metabolic process
GO:0006066	52	6.20e-08	alcohol metabolic process
GO:0048856	62	7.67e-08	anatomical structure development
GO:0007242	53	7.81e-08	intracellular signaling cascade
GO:0046365	24	9.47e-08	monosaccharide catabolic process
GO:0019318	34	9.49e-08	hexose metabolic process
GO:0030529	107	1.25e-07	ribonucleoprotein complex
GO:0006073	20	1.31e-07	glucan metabolic process
GO:0007265	23	1.54e-07	Ras protein signal transduction
GO:0005977	16	1.56e-07	glycogen metabolic process
GO:0065008	74	1.59e-07	regulation of biological quality
GO:0006740	11	1.78e-07	NADPH regeneration
GO:0006897	28	2.25e-07	endocytosis
GO:0010324	30	2.48e-07	membrane invagination
GO:0019843	17	3.02e-07	rRNA binding
GO:0050793	11	4.48e-07	regulation of developmental process
GO:0016772	77	5.36e-07	transferase activity, transferring phosphorus-containing groups
GO:0005933	40	6.06e-07	cellular bud
GO:0005996	34	6.13e-07	monosaccharide metabolic process
GO:0030955	9	7.98e-07	potassium ion binding
GO:0051726	44	9.88e-07	regulation of cell cycle

GO:0000074	44	9.88e-07	regulation of progression through cell cycle
GO:0006098	10	1.03e-06	pentose-phosphate shunt
GO:0009117	41	1.15e-06	nucleotide metabolic process
GO:0007264	34	1.76e-06	small GTPase mediated signal transduction
GO:0005979	6	2.99e-06	regulation of glycogen biosynthetic process
GO:0051278	12	3.25e-06	chitin- and beta-glucan-containing cell wall polysaccharide biosynthetic process
GO:0008360	8	5.04e-06	regulation of cell shape
GO:0006038	8	5.04e-06	cell wall chitin biosynthetic process
GO:0022603	8	5.04e-06	regulation of anatomical structure morphogenesis
GO:0022604	8	5.04e-06	regulation of cell morphogenesis
GO:0006769	17	5.74e-06	nicotinamide metabolic process
GO:0044267	220	7.05e-06	cellular protein metabolic process
GO:0015935	26	7.80e-06	small ribosomal subunit
GO:0005935	31	8.82e-06	cellular bud neck
GO:0019362	17	1.16e-05	pyridine nucleotide metabolic process
GO:0006031	9	1.29e-05	chitin biosynthetic process
GO:0006037	8	1.35e-05	cell wall chitin metabolic process
GO:0000028	8	1.35e-05	ribosomal small subunit assembly and maintenance
GO:0048610	36	1.53e-05	reproductive cellular process
GO:0022413	36	1.53e-05	reproductive process in single-celled organism
GO:0030427	37	1.59e-05	site of polarized growth
GO:0016192	70	1.61e-05	vesicle-mediated transport
GO:0005934	18	1.83e-05	cellular bud tip
GO:0005498	6	1.88e-05	sterol carrier activity
GO:0005496	6	1.88e-05	steroid binding
GO:0032934	6	1.88e-05	sterol binding
GO:0006887	17	2.22e-05	exocytosis
GO:0015934	30	2.95e-05	large ribosomal subunit
GO:0008361	33	3.01e-05	regulation of cell size
GO:0015980	36	3.91e-05	energy derivation by oxidation of organic compounds
GO:0009272	13	3.91e-05	chitin- and beta-glucan-containing cell wall biogenesis
GO:0040007	34	4.31e-05	growth
GO:0065009	21	4.50e-05	regulation of a molecular function
GO:0042546	13	5.74e-05	cell wall biogenesis
GO:0006665	12	6.26e-05	sphingolipid metabolic process
GO:0010383	8	6.56e-05	cell wall polysaccharide metabolic process
GO:0030011	6	6.75e-05	maintenance of cell polarity
GO:0006869	14	7.15e-05	lipid transport
GO:0050790	20	7.36e-05	regulation of catalytic activity
GO:0031505	15	8.24e-05	chitin- and beta-glucan-containing cell wall organization and biogenesis
GO:0006042	9	8.97e-05	glucosamine biosynthetic process
GO:0006045	9	8.97e-05	N-acetylglucosamine biosynthetic process
GO:0046349	9	8.97e-05	amino sugar biosynthetic process
GO:0006893	12	9.31e-05	Golgi to plasma membrane transport

Table S1: **Gene Ontology terms exclusive of WGD paralogs.** The table reports the results of the enrichment analysis for Gene Ontology terms exclusive of non-WGD duplicates, with populations of functional categories (column two) and P-values from hypergeometric testing (column three).

Gene Ontology terms exclusive of non-WGD paralogs			
GO term	Number of genes	P-value	annotation
GO:0022891	60	4.99e-16	substrate-specific transmembrane transporter activity
GO:0022857	64	6.24e-16	transmembrane transporter activity
GO:0022892	65	1.36e-13	substrate-specific transporter activity
GO:0005215	71	2.38e-13	transporter activity
GO:0005353	11	4.78e-11	fructose transmembrane transporter activity
GO:0015578	11	4.78e-11	mannose transmembrane transporter activity
GO:0005355	11	1.44e-10	glucose transmembrane transporter activity
GO:0015149	11	3.86e-10	hexose transmembrane transporter activity
GO:0015145	11	3.86e-10	monosaccharide transmembrane transporter activity
GO:0015291	25	1.17e-09	secondary active transmembrane transporter activity
GO:0015293	19	1.36e-09	symporter activity
GO:0022804	35	3.71e-09	active transmembrane transporter activity
GO:0015171	14	1.02e-08	amino acid transmembrane transporter activity
GO:0015837	17	1.13e-08	amine transport
GO:0051119	14	1.55e-08	sugar transmembrane transporter activity
GO:0005351	14	1.55e-08	sugar:hydrogen ion symporter activity
GO:0005342	19	1.83e-08	organic acid transmembrane transporter activity
GO:0046943	18	3.04e-08	carboxylic acid transmembrane transporter activity
GO:0015144	14	3.42e-08	carbohydrate transmembrane transporter activity
GO:0006865	15	4.90e-08	amino acid transport
GO:0046942	19	5e-08	carboxylic acid transport
GO:0015849	19	6.35e-08	organic acid transport
GO:0000023	8	7.87e-08	maltose metabolic process
GO:0008615	8	7.87e-08	pyridoxine biosynthetic process

GO:0042819	8	7.87e-08	vitamin B6 biosynthetic process
GO:0008614	8	1.93e-07	pyridoxine metabolic process
GO:0042816	8	1.94e-07	vitamin B6 metabolic process
GO:0009277	19	1.42e-06	chitin- and beta-glucan-containing cell wall
GO:0048503	13	3.21e-06	GPI anchor binding
GO:0015205	6	9.08e-06	nucleobase transmembrane transporter activity
GO:0015174	6	9.08e-06	basic amino acid transmembrane transporter activity
GO:0042402	6	9.084e-06	biogenic amine catabolic process
GO:0016020	168	1.22e-05	membrane
GO:0005984	8	1.29e-05	disaccharide metabolic process
GO:0015075	29	1.82e-05	ion transmembrane transporter activity
GO:0042219	6	3.59e-05	amino acid derivative catabolic process
GO:0015175	5	4.20e-05	neutral amino acid transmembrane transporter activity
GO:0030976	5	4.20e-05	thiamin pyrophosphate binding
GO:0019660	5	4.20e-05	glycolytic fermentation
GO:0006559	4	6.82e-05	L-phenylalanine catabolic process
GO:0031224	124	7.03e-05	intrinsic to membrane
GO:0030287	5	8.98e-05	cell wall-bounded periplasmic space
GO:0009083	5	8.98e-05	branched chain family amino acid catabolic process
GO:0044270	9	9.37e-05	nitrogen compound catabolic process
GO:0009310	9	9.37e-05	amine catabolic process
GO:0016021	123	9.81e-05	integral to membrane

Table S2: **Gene Ontology terms exclusively found in non-WGD paralogs.** The table reports the results of the enrichment analysis for Gene Ontology terms exclusive of non-WGD duplicates, with populations of functional categories (column two) and P-values from hypergeometric testing (column three).

SCOP superfamily domain occurrence		
Domain	Occurrence in WGD proteins	Occurrence in non-WGD proteins
46561	2	0
46565	0	16
46579	0	7
46589	0	2
46626	2	0
46689	8	14
46774	0	2
46785	8	13
46906	2	0
46934	2	3
46938	2	2
46946	2	1
46955	0	2
46977	2	0
47060	0	2
47072	0	2
47095	4	3
47113	2	22
47212	2	0
47240	2	1
47323	2	2
47370	4	2
47459	0	8
47473	2	10
47576	0	2
47592	4	0
47616	0	5
47661	2	3
47672	1	0
47694	0	2
47769	2	2
47807	2	1
47819	0	2
47923	4	5
47954	10	10
47973	0	2
48019	0	4
48065	2	2
48097	0	2
48140	2	0
48150	2	1
48168	2	0
48179	2	5
48208	2	6
48225	0	2
48239	0	4
48256	2	1
48264	0	3
48317	2	4
48334	0	2
48350	6	4
48366	2	1
48371	8	57
48403	6	6
48425	2	2
48431	1	0
48439	0	6
48445	2	0
48452	6	24
48464	6	6
48557	0	3
48576	0	3
48592	0	6
48613	0	5
48695	2	0
49348	2	0
49354	2	0
49447	0	2
49493	0	2
49562	2	2
49764	0	3
49777	0	3
49785	0	3
49863	1	0
49879	6	4
49899	4	4
50044	9	11
50104	6	1
50129	0	4
50182	0	16
50193	2	1

50249	10	16
50324	0	2
50447	5	4
50465	3	2
50475	2	2
50630	2	10
50677	0	2
50729	12	8
50800	2	0
50891	4	3
50965	4	1
50978	9	83
50985	0	3
51011	0	7
51161	0	2
51182	0	3
51206	0	2
51230	4	2
51246	4	0
51306	0	3
51316	0	3
51366	2	5
51395	0	7
51412	2	4
51419	0	2
51430	2	14
51445	4	18
51556	1	5
51569	6	4
51604	2	3
51621	2	3
51645	0	2
51726	0	2
51730	0	3
51735	12	61
51905	10	10
51998	2	0
52016	0	4
52025	2	1
52047	2	6
52058	4	3
52080	2	2
52087	2	2
52096	4	0
52113	0	3
52151	4	3
52161	2	1
52166	2	1
52172	1	0
52218	2	2
52283	2	3
52313	2	1
52317	4	8
52335	2	0
52343	4	3
52374	4	11
52402	1	6
52440	4	0
52467	2	7
52490	2	2
52507	2	2
52518	2	6
52540	32	121
52743	2	0
52768	0	6
52777	0	4
52799	2	10
52821	2	6
52833	16	24
52922	2	0
52935	2	0
52949	0	2
52954	2	0
52972	0	2
53032	0	2
53067	12	23
53092	0	2
53098	2	54
53137	2	2
53167	0	2
53187	2	7
53223	0	4
53244	2	1
53254	4	13

53271	6	6
53328	0	2
53335	0	45
53383	4	30
53448	12	12
53474	9	31
53613	0	9
53623	2	1
53633	2	0
53649	2	4
53659	2	5
53686	0	6
53697	1	2
53720	0	11
53732	0	4
53738	2	1
53756	4	7
53774	2	5
53850	0	4
53901	0	4
53927	1	5
54001	6	17
54189	2	2
54197	3	4
54211	6	15
54236	4	12
54427	0	3
54495	2	13
54534	2	2
54570	0	2
54575	2	0
54616	2	0
54626	0	2
54631	2	2
54637	0	5
54686	0	2
54695	2	2
54747	2	0
54768	0	5
54791	0	3
54826	2	3
54843	2	1
54849	0	6
54897	2	2
54928	10	40
54980	2	0
54999	0	2
55021	2	2
55035	2	0
55060	2	3
55103	0	2
55120	4	4
55129	2	2
55154	2	0
55174	2	3
55190	2	0
55205	0	2
55257	0	4
55277	2	0
55282	2	1
55298	2	0
55307	2	2
55315	4	4
55424	2	1
55455	2	2
55469	0	2
55486	2	4
55608	0	6
55666	0	2
55681	2	5
55729	0	18
55753	0	3
55797	2	1
55811	0	7
55821	0	2
55856	0	5
55874	2	0
55920	0	6
55957	2	1
55973	2	0
55979	0	2
56019	0	3
56047	0	3
56053	0	3

56059	4	0
56104	0	4
56112	55	2
56204	0	2
56219	4	5
56235	4	15
56281	3	5
56300	6	14
56317	0	5
56425	4	0
56542	2	0
56634	0	2
56655	0	4
56672	0	8
56752	2	1
56784	10	18
56801	2	6
56808	2	3
56815	0	4
56988	0	6
57196	1	0
57667	19	15
57701	12	41
57716	4	10
57756	2	2
57783	0	5
57829	8	0
57850	4	25
57863	2	4
57868	0	2
57879	2	1
57903	2	11
63380	2	4
63393	0	2
63411	0	7
63737	2	1
63748	0	3
64005	0	3
64153	0	2
64197	2	0
64268	1	9
64356	0	12
64484	0	6
68906	2	2
69000	0	2
69322	1	0
69572	2	7
69593	2	3
69645	0	2
74650	0	3
74924	0	3
75217	2	1
75304	0	2
75553	0	4
75620	0	2
75632	1	0
81271	0	2
81296	6	2
81321	2	1
81333	0	4
81338	0	5
81342	0	2
81343	2	1
81383	0	4
81406	2	1
81442	0	4
81606	2	5
81631	2	0
81653	2	2
81660	2	2
81665	0	2
81811	2	0
81901	2	3
81995	2	0
82061	1	0
82109	2	5
82199	2	9
82215	2	1
82282	2	1
82549	0	2
82649	2	0
82657	0	3
82754	2	0
82919	2	0

88697	1	2
88713	0	2
88723	0	6
88798	0	2
89000	4	0
89009	4	1
89124	0	3
89360	0	2
89942	0	2
90096	0	2
90123	2	0
90229	2	0
100920	6	1
100934	2	3
100950	4	6
101152	0	2
101447	0	3
101473	0	2
101489	2	0
101576	2	1
102114	0	2
102712	0	2
102860	2	0
103111	0	2
103243	2	0
103473	22	68
103481	3	3
103506	10	24
109993	0	2
110296	0	6
110921	2	0
110942	2	0
111331	2	1
111352	2	1
111430	2	1

Table S3: **List of the SCOP superfamily domains appearing in duplications and their relative population in the WGD and non-WGD sets of duplicates.**

	Domain score	Arch. score	GO sim MF	GO sim BP	GO sim CC
Domain score		0.97	0.16	0.07	0.1
Arch. score			0.16	0.09	0.11
GO sim MF				0.44	0.24
GO sim BP					0.34

Table S4: **Spearman's rank correlation coefficient of the different scores used to compare paralogs -** Domain score and Architecture score have a strong positive correlation while only weak positive correlation is found between other scores.

Generation of Domain Architectures

In this section, we describe in more detail the algorithms used for the construction of homology classes. To give a clear and complete description we will employ pseudocode. A brief summary of standard conventions is given here for reference.

- *Hash Tables.* A hash table, or a hash map, is a data structure that associates keys with values. The primary operation it must supports efficiently is a lookup: given a key (a given gene, for example), find the corresponding value or values (in this example, its architecture). Hash tables are written with capital boldface letters: for example, **DAG** refers to a hash table called DAG (in the following, the one storing the Domain Architectures for the Genes).
- Variables are not declared. Variables, which may or may not be keys of a hash table, are usually indicated with lowercase boldface letters, as in **g** or **d**.
- The “pertaining to set” ($\in$) symbol has the conventional set-theoretical meaning. The value or values for a given key are always an homogenous set of some kind: these might be numbers, names or, more in general, strings. The pseudocode $\mathbf{g} \in \mathbf{DAG}$ indicates that the specific gene **g** is a key of the hash table **DAG**.
- If the hash table **H** contains more than a value for a given key (say **k**), then $\mathbf{H[k]}$ is defined as the set of all the values for the given key **k**. For example, let **DA** be the hash table consisting of a given number of distinct Domain Architecture as keys, whose values are (for each given architecture) the genes with that distinct architecture; let **g** be a gene, and **d** an architecture. Then $\mathbf{g} \in \mathbf{DA[d]}$ means that the gene **g** is a value for the key **d** in the hash table **DA**.
- The “absolute value” symbol (|) expresses the value of the current key or variable, or the dimension (number of keys) of the hash table. For example, $|\mathbf{DA}| = N$ means that the hash table **DA** is composed of N keys; otherwise $|\mathbf{g}| = \mathrm{d}$ means that the gene **g** has architecture d. In this case d is considered a value.
- The “equality” symbol (=) has two distinct meanings. It might refer to the mathematical equality: $|\mathrm{H}| = N$ means that N and the number of keys for hash table **H** are the same number. For non-numerical arguments, = symbol can imply a broader meaning of similitude; this is specially true when equating names or strings. In this latter case, higher forms of equality may be represented with other symbols, such as “$\equiv$”, “$\simeq$” or other easily recognizable symbols which must be completely defined.
- In the specific case of our work, each architecture is an ordered string of domain assignment codes. Were needed, one may indicate with Length[**d**] or L[**d**] the number of domains the architecture is composed of. The symbol **d**[i] may then be used to indicate the i^{th} domain of architecture **d**.
- Attribution of numerical values is usually represented with “$\leftarrow$” symbol (but never with the = symbol). The code $\mathbf{X[i]} \leftarrow 1$ assigns the value 1 to the i^{th} component of object **X**.

```
for each $\mathbf{g} \in \mathbf{DAG}$                    # considering each gene $\mathbf{g}$
  for each $\mathbf{d} \in \mathbf{DA}$                   # considering all the domain architectures $\mathbf{d}$
    HOMOLOGY := TRUE                        # will $\mathbf{g}$ have architecture $\mathbf{d}$ ?
    if Length[$\mathbf{d}$] = Length[$|\mathbf{g}|$]               # they have the same length ?
    then                                    #
      for i = 0 to Length[$\mathbf{d}$]                 # considering all the domains
        if $|\mathbf{g}[i]| \neq \mathbf{d}[i]$                   # are they all equal, in sequence ?
          HOMOLOGY := FALSE                 #
      end for                               #
    else HOMOLOGY := FALSE                  # they do not have the same length
    if HOMOLOGY = TRUE                      #
    then $\mathbf{g} \in \mathbf{DA}[\mathbf{d}]$                       # $\mathbf{g}$ has domain architecture $\mathbf{d}$
  end for
end for
```

Table S5: Pseudocode describing the algorithm for Homology Criterion **A**. The domain architecture of each gene is compared to all the different domain architectures. When the algorithm is complete, the genes results aggregated in sets, depending on their architecture. In other words, each set of keys for an element (domain architecture) of the hash table **DA**, represent an equivalence class of genes.

- Attribution of non-numerical values is represented with "$\cup$" symbol, as in $|\mathbf{g}| = |\mathbf{g}| \cup \mathbf{d}$: add the value $\mathbf{d}$ to the (set of) values of key $\mathbf{g}$. Usually the set-theoretical properties of $\cup$ are implied. This means that this will not be used on objects where order matters (strings).
- Pseudocode was kept as simple as possible, but sometimes the employ of flow control is necessary. "For" cycles will begin with a lowercase bold **for** keyword followed by the control statement. In the most complex cases, another keyword **endfor** will be provided for clarity. The same applies to "while" cycles and "if", "case" or "switch" statements.

S1 Homology criteria

Criterion A.

This criterion implements the simple requirement that two protein architectures must be exactly matching in order to be considered as being coded by paralogs. This so that it generates equivalence classes: each protein appears in only one class, together with all the other homologous proteins. Therefore, the classes form a partition of the set of all proteins.

Criterion B.

This criterion relaxes the previous one, considering two proteins as homologous if their architectures are equal, or if one can be seen as a multiple repetition of the other, ignoring possible gap mismatches. This criterion is also implemented so to generate equivalence classes.

As before, let L[**g**] be defined as the total number of domains present in gene architecture, and let us suppose to consider a pair of gene architectures, **a** and **b**. We have three cases:

1. L[**a**] = L[**b**]
2. L[**a**] > L[**b**] but L[**a**] < 2 × L[**b**]
3. L[**a**] > 2 × L[**b**]

In the first case the algorithm follows exactly criterion A: if each pair of corresponding domains is equal between the genes A and B, the two genes show homology.

In the second case, the two genes are considered equivalent only if the whole string of the shorter can be found in the longer, and the excess domains in the longer are gaps.

Lastly, in case number three, the algorithm performs an integer division and computes how many times the shorter architecture may fit in the longer one (quotient). The remainder of this division, if nonzero, is used to offset the beginning domains of the longer string. For each of the possible values of the offsetting value, $0 \leq \text{offset} \leq \text{remainder}$, the short architecture is repeated *remainder* times in the longer, starting from the offsetted domain **g**[i]. If match is found AND the offset domains are gaps, the two genes are considered matching. In case still no match is found, the algorithm repeats the procedure, assuming again that the shorter string is repeated in the longer, but also assuming that the repetitions of the shorter string are intervalled by gap domains. This is done considering the shorter domain architecture as it was one (_gap_) domain longer. Again, if match is found AND the offsetted domains are gaps, the architectures are equivalent.

```
for each $\mathbf{g} \in \mathbf{DAG}$
  for each $\mathbf{d} \in \mathbf{DA}$
    if $L[|\mathbf{g}|] = L[\mathbf{d}]$
      hard criterion on $|\mathbf{g}|$ and $\mathbf{d}$
      if match then HOMOLOGY := TRUE; break
    if $L[|\mathbf{g}|] > L[\mathbf{d}]$ and if $L[|\mathbf{g}|] < 2 \times L[|\mathbf{g}|]$
      for i=0 to $(L[|\mathbf{g}|]) - L[\mathbf{d}]$
        hard criterion on $|\mathbf{g}[j+i]|$ and $\mathbf{d}[j]$
        if (match) and (discarded = _gap_)
          then HOMOLOGY := TRUE; break
      end for
    if $L[|\mathbf{g}|] > 2 \times L[\mathbf{d}]$
      $L[|\mathbf{g}|] = \mathbf{QUOT} \times L[\mathbf{d}] + \mathbf{REM}$
      for each i=0 to REM
        for each j=0 to QUOT
          hard criterion on $|\mathbf{g}|[j \times QUOT+i]$ and $\mathbf{d}[j]$
        end for
        if (match) and (discarded = _gap_)
        then HOMOLOGY := TRUE; break
      end for
      $L[|\mathbf{g}|] = \mathbf{QUOT} \times L[\mathbf{d} \cup 0] + \mathbf{REM}$
      for each i=0 to REM
        for each j=0 to QUOT
          hard criterion on $|\mathbf{g}|[j \times QUOT+i]$ and $\mathbf{d}[j]$
        end for
        if (match) and (discarded = _gap_)
        then HOMOLOGY := TRUE
      end for
  end for
end for
```

Table S6: Algorithm for Criterion **B**. In this case a multiple repetition of an architecture is allowed. However the duplication must be retrieved completely, without exceptions. It can be noted that the presence of the gaps is allowed in principle, but it happens that gap domains *inside* the sequences are almost absent in most datasets.

```
for each $\mathbf{g} \in \mathbf{DAG}$
  for each $\mathbf{d} \in \mathbf{DA}$
    if $L[|\mathbf{g}|] = L[\mathbf{d}]$
      hard criterion on $|\mathbf{g}|$ and $\mathbf{d}$
      if match then HOMOLOGY := TRUE; break
    if $L[|\mathbf{g}|] > L[\mathbf{d}]$ and if $L[|\mathbf{g}|] < 2 \times L[|\mathbf{g}|]$
      for i=0 to $(L[|\mathbf{g}|]) - L[\mathbf{d}]$
        hard criterion on $|\mathbf{g}[j+i]|$ and $\mathbf{d}[j]$
        if (match)
          then HOMOLOGY := TRUE; break
      end for
    if $L[|\mathbf{g}|] > 2 \times L[\mathbf{d}]$
      $L[|\mathbf{g}|] = \mathbf{QUOT} \times L[\mathbf{d}] + \mathbf{REM}$
      for each i=0 to REM
        for each j=0 to QUOT
          hard criterion on $|\mathbf{g}|[j \times QUOT+i]$ and $\mathbf{d}[j]$
        end for
        if (match) and (discarded = _gap_)
        then HOMOLOGY := TRUE; break
      end for
      $L[|\mathbf{g}|] = \mathbf{QUOT} \times L[\mathbf{d} \cup 0] + \mathbf{REM}$
      for each i=0 to REM
        for each j=0 to QUOT
          hard criterion on $|\mathbf{g}|[j \times QUOT+i]$ and $\mathbf{d}[j]$
        end for
        if (match)
        then HOMOLOGY := TRUE
      end for
  end for
end for
```

Table S7: Algorithm for Criterion **C**.

Criterion C.

This last criterion is obtained through further relaxing of the conditions considered in criterion **B**. Two protein architectures are considered as homologous if they are equal, or if one of them can be seen as an *approximate* repetition of the other. With approximate we mean that the repeated architecture domain sequences can be interspaced by gaps *or* other domains.