Supplementary Table 1: Descriptions of the proto-oncoproteins and housekeeping proteins used in the category 1 comparison. Protein sequences were gathered from the Swiss-Prot database (www.expasy.org).

Proto-oncoproteins		**Housekeeping proteins**	
Accession	**Acronym and description**	**Accession**	**Acronym and description**
P51587	BRCA2 Breast cancer susceptibility protein	P04075	ALDOA Aldolase A
P10721	KIT Mast/stem cell growth factor receptor	P06744	GPI Glucose phosphate isomerase
P08581	MET Hepatocyte growth factor receptor	P00558	PGK1 Phosphoglycerate kinase 1
P07949	RET Tyrosine-protein kinase receptor	O14548	COX7A2L Cytochrome c oxidase
P04637	TP53 Cellular tumor antigen p53	P84098	RPL19 Ribosomal protein L19
P07332	FES Tyrosine-protein kinase	P60709	ACTB Actin, beta
P06241	FYN Tyrosine-protein kinase	P08133	ANXA6 Annexin A6
Q9UM73	ALK Tyrosine kinase receptor	O75506	HSBP1 HSF binding protein 1
P04629	NTRK1 Nerve growth factor receptor	P07199	CENPB Centromere protein B
Q05397	FAK1 Focal adhesion kinase 1	P12109	COL6A1 Collagen, type VI
P08922	ROS Tyrosine-protein kinase	P15954	COX7C Cytochrome c oxidase
P04626	ERBB2 Receptor tyrosine-protein kinase	P35222	CTNNB1 Catenin
P16591	FER Tyrosine-protein kinase	P04406	GAPDH G3P dehydrogenase
P42684	ABL2 Tyrosine-protein kinase	P35579	MYH9 Myosin, non-muscle
P09769	FGR Tyrosine-protein kinase	P62988	RPS27A Ribosomal protein S27a
P12931	SRC Tyrosine-protein kinase	Q5QNW6	HIST1H2BC Histone H2bc
P07947	YES1 Tyrosine-protein kinase	O15371	EIF3S7 Translation IF
P38398	BRCA1 Breast cancer susceptibility protein	O43707	ACTN4 Actinin, alpha 4
P11362	FGFR1 Fibroblast growth factor receptor 1	P00338	LDHA Lactate dehydrogenase A
O60674	JAK2 Tyrosine-protein kinase	P08238	HSP90AB1 Heat shock protein

Supplementary Table 2: GO terms describing the housekeeping proteins used in the category 1 comparison. The number of human proteins described by each GO term is indicated. The final column indicates whether or not a given GO term was used in one of the category 2 comparisons (those with ten or more corresponding proteins).

Accession	**Description**	**# of proteins**	
GO0000122	negative regulation of transcription from RNA polymerase II promoter	127	Yes
GO0000910	cytokinesis	19	Yes
GO0001525	angiogenesis	78	Yes
GO0001837	epithelial to mesenchymal transition	11	Yes
GO0005975	carbohydrate metabolic process	211	Yes
GO0006091	generation of precursor metabolites and energy	83	Yes
GO0006096	glycolysis	59	Yes
GO0006412	translation	277	Yes
GO0006446	regulation of translational initiation	20	Yes
GO0006509	membrane protein ectodomain proteolysis	11	Yes
GO0006928	cell motility	126	Yes
GO0006941	striated muscle contraction	32	Yes
GO0006959	humoral immune response	32	Yes
GO0006986	response to unfolded protein	49	Yes
GO0007049	cell cycle	232	Yes
GO0007155	cell adhesion	421	Yes
GO0007229	integrin-mediated signaling pathway	62	Yes
GO0007411	axon guidance	47	Yes
GO0007599	hemostasis	6	
GO0008360	regulation of cell shape	26	Yes
GO0015031	protein transport	284	Yes
GO0016310	phosphorylation	13	Yes
GO0016567	protein ubiquitination	37	Yes
GO0030048	actin filament-based movement	12	Yes
GO0030220	platelet formation	2	
GO0030224	monocyte differentiation	3	
GO0030433	ER-associated protein catabolic process	11	Yes
GO0030521	androgen receptor signaling pathway	35	Yes
GO0031532	actin cytoskeleton reorganization	8	
GO0032417	positive regulation of sodium-hydrogen antiporter activity	1	
GO0042062	long-term strengthening of neuromuscular junction	2	
GO0042981	regulation of apoptosis	84	Yes
GO0043534	blood vessel endothelial cell migration	1	
GO0045429	positive regulation of nitric oxide biosynthetic process	8	
GO0045893	positive regulation of transcription, DNA-dependent	23	Yes
GO0045941	positive regulation of transcription	32	Yes
GO0048167	regulation of synaptic plasticity	6	
GO0050900	leukocyte migration	7	
GO0051272	positive regulation of cell motility	4	
GO0060070	Wnt receptor signaling pathway through beta-catenin	5	

Supplementary Table 3: Tyrosine kinases used in the category 3 comparison. All protein sequences were gathered from the Swiss-Prot database (www.expasy.org). The table includes the Swiss-Prot accession number and a short description of the protein.

Proto-oncogenic tyrosine kinases	
Accession	**Acronym and description**
P00519	ABL1 Proto-oncogene
P42684	ABL2 Tyrosine-protein kinase
Q07912	ACK1 activated CDC42 kinase 1
Q9UM73	ALK tyrosine kinase receptor
P30530	AXL/UFO Tyrosine-protein kinase receptor
P51813	BMX Cytoplasmic tyrosine-protein kinase
Q13882	PTK6/BRK Tyrosine-protein kinase 6
P41240	CSK Tyrosine-protein kinase
P42679	MATK/CTK Megakaryocyte-associated tyrosine-protein kinase
Q08345	DDR1 Epithelial discoidin domain-containing receptor 1
Q16832	DDR2 Discoidin domain-containing receptor 2
P00533	EGFR Epidermal growth factor receptor
P21709	EPHA1 Ephrin type-A receptor 1
P29317	EPHA2 Ephrin type-A receptor 2
P29320	EPHA3 Ephrin type-A receptor 3
P54764	EPHA4 Ephrin type-A receptor 4
P54756	EPHA5 Ephrin type-A receptor 5
Q15375	EPHA7 Ephrin type-A receptor 7
P54760	EPHB4 Ephrin type-B receptor 4
Q05397	FAK1 Focal adhesion kinase 1
P16591	FER Proto-oncogene tyrosine-protein kinase
P07332	FES roto-oncogene tyrosine-protein kinase Fes/Fps
P11362	FGFR1 Basic fibroblast growth factor receptor 1
P21802	FGFR2 Fibroblast growth factor receptor 2
P22607	FGFR3 Fibroblast growth factor receptor 3
P22455	FGFR4 Fibroblast growth factor receptor 4
P09769	FGR Proto-oncogene tyrosine-protein kinase
P17948	VGFR1 Vascular endothelial growth factor receptor 1
P36888	FLT3 FL cytokine receptor
P35916	VGFR3 Vascular endothelial growth factor receptor 3
P07333	CSF1R Macrophage colony-stimulating factor 1 receptor
P06241	FYN Proto-oncogene tyrosine-protein kinase
P04626	ERBB2 Receptor tyrosine-protein kinase
P21860	ERBB3 Receptor tyrosine-protein kinase
Q15303	ERBB4 Receptor tyrosine-protein kinase
P08069	IGF1R Insulin-like growth factor 1 receptor
P23458	JAK1 Tyrosine-protein kinase
O60674	JAK2 Tyrosine-protein kinase
P52333	JAK3 Tyrosine-protein kinase
P35968	VGFR2 Vascular endothelial growth factor receptor 2
P10721	KIT Mast/stem cell growth factor receptor
P06239	LCK Proto-oncogene tyrosine-protein kinase
P07948	LYN Tyrosine-protein kinase

Supplementary Table 3 (continued)

Accession	Acronym and description
Q12866	MERTK Proto-oncogene tyrosine-protein kinase
P08581	MET Hepatocyte growth factor receptor
P16234	PGFRA Alpha platelet-derived growth factor receptor
P09619	PGFRB Beta platelet-derived growth factor receptor
Q14289	FAK2 Protein tyrosine kinase 2 beta
P07949	RET Proto-oncogene tyrosine-protein kinase receptor
Q04912	RON Macrophage-stimulating protein receptor
P08922	ROS Proto-oncogene tyrosine-protein kinase
P12931	SRC Proto-oncogene tyrosine-protein kinase
P43405	KSYK Tyrosine-protein kinase
Q6J9G0	STYK1 Tyrosine protein-kinase
P04629	NTRK1 High affinity nerve growth factor receptor
Q16620	NTRK2 BDNF/NT-3 growth factors receptor
Q16288	NTRK3 NT-3 growth factor receptor
P07947	YES Proto-oncogene tyrosine-protein kinase
Non-proto-oncogenic tyrosine kinases	
P51451	BLK Tyrosine-protein kinase
Q06187	BTK Tyrosine-protein kinase
Q13308	PTK7 Tyrosine-protein kinase-like 7
Q5JZY3	EPHAA Ephrin type-A receptor 10
Q9UF33	EPHA6 Ephrin type-A receptor 6
P29322	EPHA8 Ephrin type-A receptor 8
P54762	EPHB1 Ephrin type-B receptor 1
P29323	EPHB2 Ephrin type-B receptor 2
P54753	EPHB3 Ephrin type-B receptor 3
O15197	EPHB6 Ephrin type-B receptor 6
P42685	FRK Tyrosine-protein kinase
P08631	HCK Tyrosine-protein kinase
P06213	INSR Insulin receptor
P14616	INSRR Insulin receptor-related protein
Q08881	ITK Tyrosine-protein kinase
Q6ZMQ8	LMTK1 Serine/threonine-protein kinase / Apoptosis-associated tyrosine kinase 1
Q8IWU2	LMTK2 Serine/threonine-protein kinase / Apoptosis-associated tyrosine kinase 2
Q96Q04	LMTK3 Serine/threonine-protein kinase / Apoptosis-associated tyrosine kinase 3
P29376	LTK Leukocyte tyrosine kinase receptor
O15146	MUSK Muscle, skeletal receptor tyrosine protein kinase
Q01973	ROR1 Tyrosine-protein kinase transmembrane receptor
Q01974	ROR2 Tyrosine-protein kinase transmembrane receptor
P34925	RYK Tyrosine-protein kinase
Q9H3Y6	SRMS Tyrosine-protein kinase
P42680	TEC Tyrosine-protein kinase
P35590	TIE1 Tyrosine-protein kinase receptor
Q02763	TIE2 Angiopoietin-1 receptor
Q13470	TNK1 Non-receptor tyrosine-protein kinase
P42681	TXK Tyrosine-protein kinase
P29597	TYK2 Non-receptor tyrosine-protein kinase
Q06418	TYRO3 Tyrosine-protein kinase receptor
P43403	ZAP70 Tyrosine-protein kinase

Supplementary Table 4: Complete data for the category 1 comparison. For each definition of a rare k-mer, and for each value of k (5, 6, and 7), the P-value is given, which shows whether there was a statistically significant difference in rare k-mer frequency between the control set and the proto-oncoprotein set. The P/C columns indicate which set contained a higher frequency of rare k-mers—the proto-oncoprotein set (P) or the control set (C).

	0 times		**≤ 2 times**		**≤ 5 times**	
	P-value	**P/C**	**P-value**	**P/C**	**P-value**	**P/C**
5-mers	0.0165	P	5.78e-07	P	3.85e-06	P
6-mers	0.00473	P	6.97e-05	P	3.68e-06	P
7-mers	8.19e-08	P	6.64e-07	P	4.24e-06	P

Supplementary Table 5: Individual results for each category 2 comparison for 5-mers. For each definition of a rare 5-mer, the P-value is given, which shows whether there was a statistically significant difference in rare 5-mer frequency between that GO protein set and the proto-oncoprotein set. The P/C columns indicate which set contained a higher frequency of rare 5-mers—the proto-oncoprotein set (P) or the control set (C). If this difference was not statistically significant, (n.s.) is appended.

	0 times		**$\leq$ 2 times**		**$\leq$ 5 times**	
Accession	**P-value**	**P/C**	**P-value**	**P/C**	**P-value**	**P/C**
GO0000122	2.57e-05	P	1.23e-11	P	8.98e-24	P
GO0000910	7.04e-08	P	1.37e-21	P	9.57e-43	P
GO0001525	0.85	C (n.s.)	0.592	P (n.s.)	0.0694	P (n.s.)
GO0001837	0.948	C (n.s.)	0.167	C (n.s.)	0.114	C (n.s.)
GO0005975	3.26e-06	C	2.26e-12	C	1.65e-15	C
GO0006091	0.948	C (n.s.)	0.303	C (n.s.)	0.844	P (n.s.)
GO0006096	0.217	P (n.s.)	0.136	P (n.s.)	0.802	C (n.s.)
GO0006412	0.0478	P	0.0424	P	0.00959	P
GO0006446	2.98e-10	P	5.4e-20	P	1.36e-27	P
GO0006509	1.49e-05	P	2.8e-05	P	2.88e-07	P
GO0006928	0.00211	P	3.24e-05	P	8.09e-07	P
GO0006941	1.42e-16	P	1.59e-27	P	1.57e-30	P
GO0006959	1.29e-07	P	3.91e-14	P	1.51e-16	P
GO0006986	1.39e-19	P	1.19e-26	P	6.32e-26	P
GO0007049	4.52e-20	P	6.52e-29	P	3.13e-36	P
GO0007155	0.000458	P	7.74e-09	P	1.54e-11	P
GO0007229	0.00108	P	0.0244	P	0.364	P (n.s.)
GO0007411	0.666	C (n.s.)	0.71	P (n.s.)	0.927	C (n.s.)
GO0008360	3.68e-16	P	1.13e-34	P	1.09e-45	P
GO0015031	6.86e-10	P	8.21e-17	P	2.33e-20	P
GO0016310	0.212	P (n.s.)	0.00772	P	0.0347	P
GO0016567	0.401	P (n.s.)	0.511	P (n.s.)	0.435	P (n.s.)
GO0030048	2.96e-12	P	8.3e-18	P	3.92e-27	P
GO0030433	0.0154	P	0.000284	P	1.27e-05	P
GO0030521	2.22e-07	P	5.89e-10	P	1.16e-15	P
GO0042981	5.01e-11	P	3.29e-19	P	2.76e-29	P
GO0045893	0.0226	P	0.000335	P	2.33e-09	P
GO0045941	1.34e-08	P	1.32e-17	P	1.13e-34	P

Supplementary Table 6: Individual results for each category 2 comparison for 6-mers. For each definition of a rare 6-mer, the P-value is given, which shows whether there was a statistically significant difference in rare 6-mer frequency between that GO protein set and the proto-oncoprotein set. The P/C columns indicate which set contained a higher frequency of rare 6-mers—the proto-oncoprotein set (P) or the control set (C). If this difference was not statistically significant, (n.s.) is appended.

	0 times		**≤ 2 times**		**≤ 5 times**	
Accession	**P-value**	**P/C**	**P-value**	**P/C**	**P-value**	**P/C**
GO0000122	1.71e-31	P	3.5e-49	P	4.59e-52	P
GO0000910	1.87e-33	P	4.54e-35	P	1.21e-19	P
GO0001525	2.6e-08	P	2.06e-21	P	1.05e-28	P
GO0001837	0.963	P (n.s.)	0.000831	P	2.38e-10	P
GO0005975	9.82e-10	C	3.76e-05	C	0.184	C (n.s.)
GO0006091	0.25	P (n.s.)	0.00558	P	2.5e-06	P
GO0006096	0.000226	C	0.000867	C	0.00209	C
GO0006412	0.614	P (n.s.)	0.000394	P	2.69e-06	P
GO0006446	7.78e-17	P	4.52e-30	P	1.11e-34	P
GO0006509	6.71e-07	P	1.52e-18	P	3.39e-14	P
GO0006928	4.85e-08	P	5.3e-12	P	8.98e-14	P
GO0006941	1.04e-09	P	7.9e-06	P	0.00618	P
GO0006959	4.91e-22	P	5.29e-23	P	2.06e-31	P
GO0006986	6.34e-12	P	9.93e-15	P	3.55e-12	P
GO0007049	2.51e-25	P	4.9e-18	P	2.4e-11	P
GO0007155	3.85e-09	P	1.14e-13	P	3.6e-12	P
GO0007229	0.0136	P	2.93e-06	P	3.22e-14	P
GO0007411	0.166	P (n.s.)	1.09e-05	P	6.31e-10	P
GO0008360	3.63e-48	P	1.19e-47	P	5.68e-42	P
GO0015031	2.02e-13	P	1.42e-12	P	3.29e-12	P
GO0016310	0.133	P (n.s.)	0.000593	P	0.000414	P
GO0016567	0.153	P (n.s.)	0.0039	P	3.43e-06	P
GO0030048	3.97e-26	P	2.18e-29	P	1.68e-29	P
GO0030433	3.02e-05	P	3.77e-10	P	3.98e-10	P
GO0030521	2.7e-14	P	4.31e-25	P	5.53e-32	P
GO0042981	8.1e-19	P	4.21e-19	P	1.39e-10	P
GO0045893	1.61e-13	P	2.7e-18	P	1.72e-22	P
GO0045941	2.17e-46	P	1.21e-69	P	5.82e-71	P

Supplementary Table 7: Individual results for each category 2 comparison for 7-mers. For each definition of a rare 7-mer, the P-value is given, which shows whether there was a statistically significant difference in rare 7-mer frequency between that GO protein set and the proto-oncoprotein set. The P/C columns indicate which set contained a higher frequency of rare 7-mers—the proto-oncoprotein set (P) or the control set (C). If this difference was not statistically significant, (n.s.) is appended.

	0 times		**≤ 2 times**		**≤ 5 times**	
Accession	**P-value**	**P/C**	**P-value**	**P/C**	**P-value**	**P/C**
GO0000122	4.61e-28	P	2.97e-31	P	7.99e-25	P
GO0000910	8.39e-11	P	0.00129	P	0.00883	P
GO0001525	8.29e-14	P	9.82e-19	P	6.85e-16	P
GO0001837	0.000189	P	8.59e-13	P	1.11e-10	P
GO0005975	0.000462	C	0.278	C (n.s.)	0.266	C (n.s.)
GO0006091	0.138	P (n.s.)	0.0867	P (n.s.)	0.721	P (n.s.)
GO0006096	6.2e-05	C	0.0245	C	0.0224	C
GO0006412	0.107	P (n.s.)	0.00359	P	2e-05	P
GO0006446	1.54e-15	P	3.35e-14	P	4.26e-07	P
GO0006509	0.000294	P	0.0133	P	0.746	C (n.s.)
GO0006928	7.52e-08	P	4.97e-05	P	0.00022	P
GO0006941	0.038	P	0.305	P (n.s.)	0.351	C (n.s.)
GO0006959	1.47e-11	P	1.42e-12	P	1.87e-12	P
GO0006986	0.000271	P	0.000175	P	0.0178	P
GO0007049	1.9e-08	P	0.00905	P	0.0241	P
GO0007155	6.3e-07	P	0.000738	P	0.00928	P
GO0007229	0.000107	P	6.97e-10	P	7.54e-07	P
GO0007411	0.000432	P	7.01e-07	P	6.13e-05	P
GO0008360	4.45e-21	P	4.28e-13	P	3.25e-06	P
GO0015031	5.11e-05	P	0.00176	P	0.0888	P (n.s.)
GO0016310	0.432	P (n.s.)	0.0274	P	0.854	C (n.s.)
GO0016567	0.00287	P	4.19e-09	P	1.37e-05	P
GO0030048	1.85e-18	P	1.04e-10	P	6.97e-08	P
GO0030433	0.00102	P	0.000451	P	0.705	P (n.s.)
GO0030521	2.83e-16	P	1.25e-23	P	1.67e-22	P
GO0042981	9.24e-07	P	0.00567	P	0.0283	P
GO0045893	1.24e-12	P	1.92e-15	P	1.62e-08	P
GO0045941	1.45e-44	P	7.72e-41	P	7.07e-33	P

Supplementary Table 8: Complete data for the category 3 comparison. For each definition of a rare k-mer, and for each value of k (5, 6, and 7), the P-value is given, which shows whether there was a statistically significant difference in rare k-mer frequency between the control set and the proto-oncoprotein set. The P/C columns indicate which set contained a higher frequency of rare k-mers—the proto-oncoprotein set (P) or the control set (C).

	0 times		≤ 2 **times**		≤ 5 **times**	
	P-value	**P/C**	**P-value**	**P/C**	**P-value**	**P/C**
5-mers	0.00144	P	5.16e-09	P	8.89e-11	P
6-mers	8.84e-10	P	1.46e-13	P	1.33e-15	P
7-mers	2.75e-09	P	3.22e-12	P	2.79e-08	P