Supplementary data

Similarly strong purifying selection acts on human disease genes of all evolutionary ages

James J. Cai[1], Elhanan Borenstein[1,3], Rong Chen[2], and Dmitri A. Petrov[1,*]

[1]Department of Biology, Stanford University, 371 Serra Mall, Stanford, CA 94305 USA;
[2]Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University School of Medicine, 251 Campus Drive, Stanford, CA 94305 USA;
[3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501 USA.

*Corresponding author: Petrov DA (dpetrov@stanford.edu)

Figure S1. Phylogeny of 39 eukaryotic species included in Ensembl build 50. Square brackets indicate the collapses of species into the corresponding lineages in Figure 1.
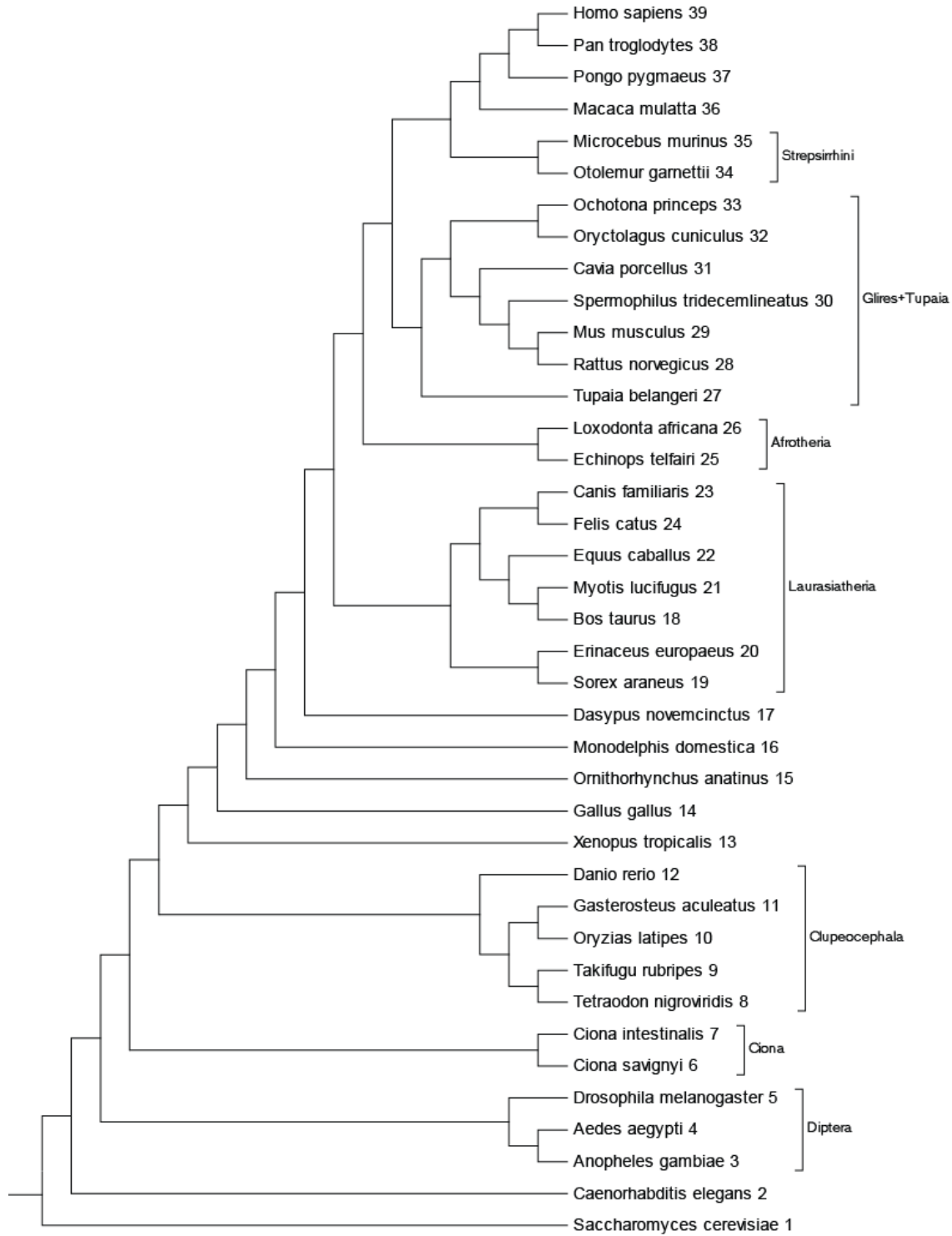
Figure S2. Significance of observed correlation coefficients between the age of Mendelian-disease genes and **(A)** Ka, **(B)** Ks, and **(C)** Ka/Ks. Observed correlation coefficients (red line) were computed for Mendelian-disease genes. Correlation coefficients for constructing the distribution were computed from 10,000 replicates of sub-sampled non-disease gene sets (see Main text for detail on the sampling). Correlation coefficients were obtained using Spearman's correlation test. Here genes were binned into equally populated bins. Using equally spaced bins produced qualitatively similar result.
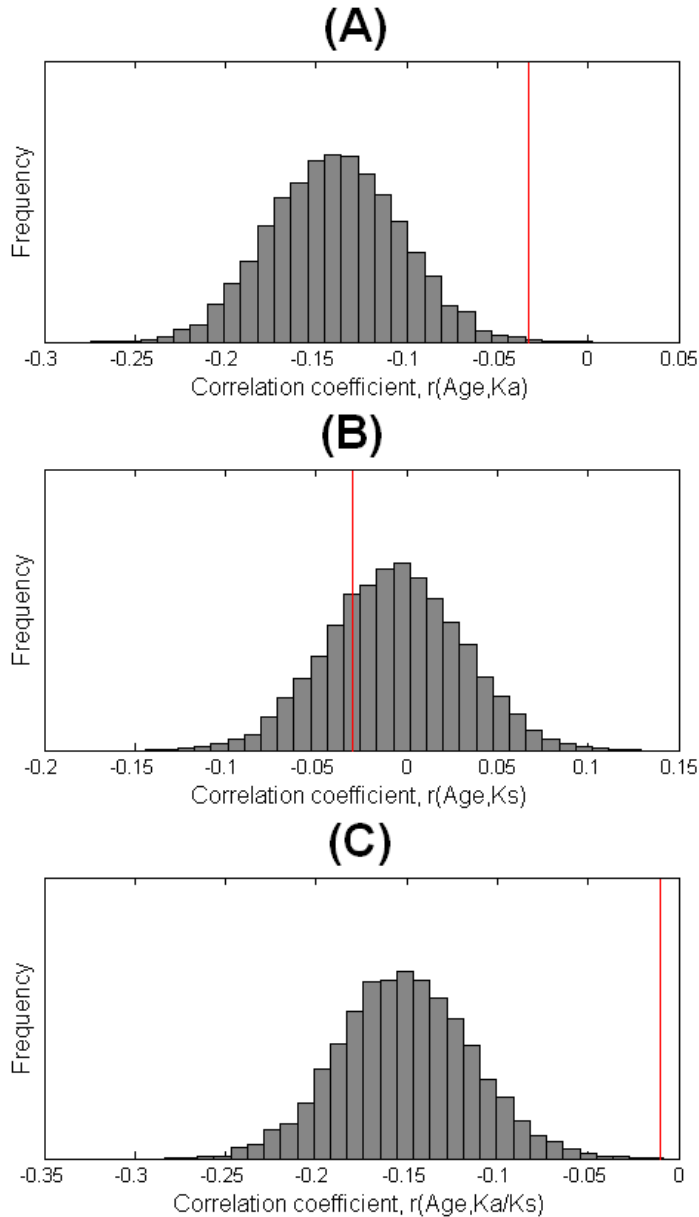
Figure S3. Significance of observed correlation coefficients between the age of complex-disease genes and **(A)** Ka, **(B)** Ks, and **(C)** Ka/Ks. Observed correlation coefficients (red line) were computed for complex-disease genes. Correlation coefficients for constructing the distribution were computed from 10,000 replicates of sub-sampled non-disease gene sets (see Main text for detail on the sampling). Correlation coefficients were obtained using Spearman's correlation test. Here genes were binned into equally populated bins. Using equally spaced bins produced qualitatively similar result.
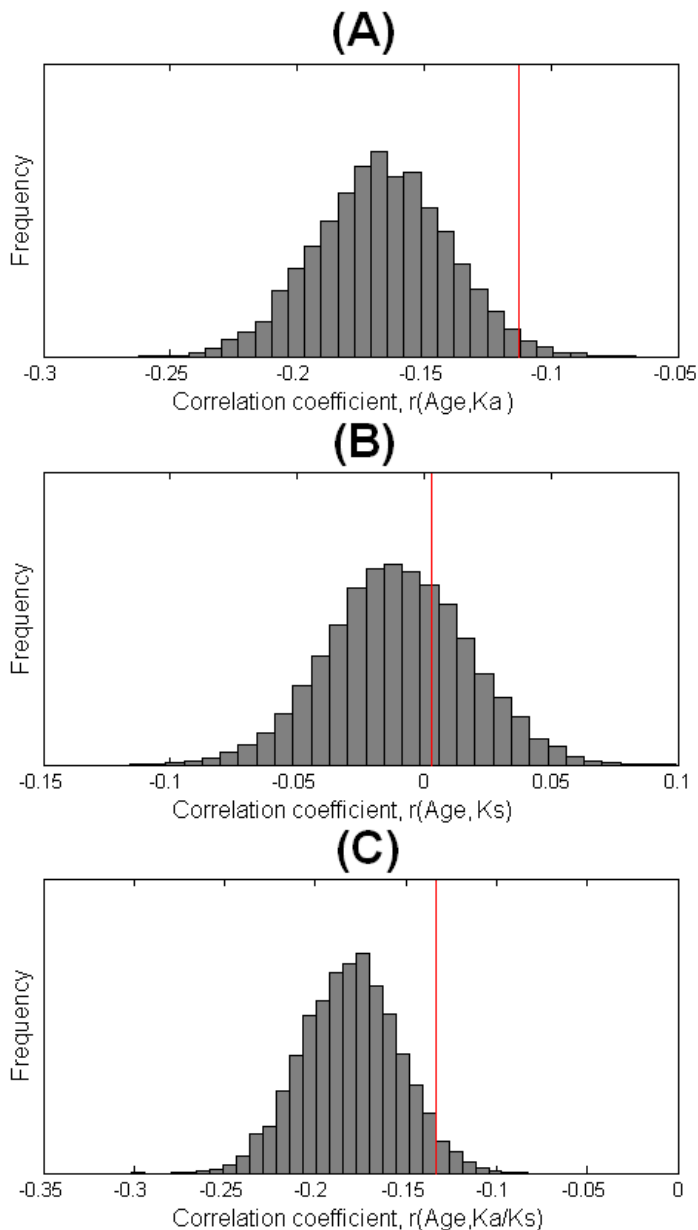
Figure S4. Ka/Ks as a function of the gene age for dominate and recessive genes.
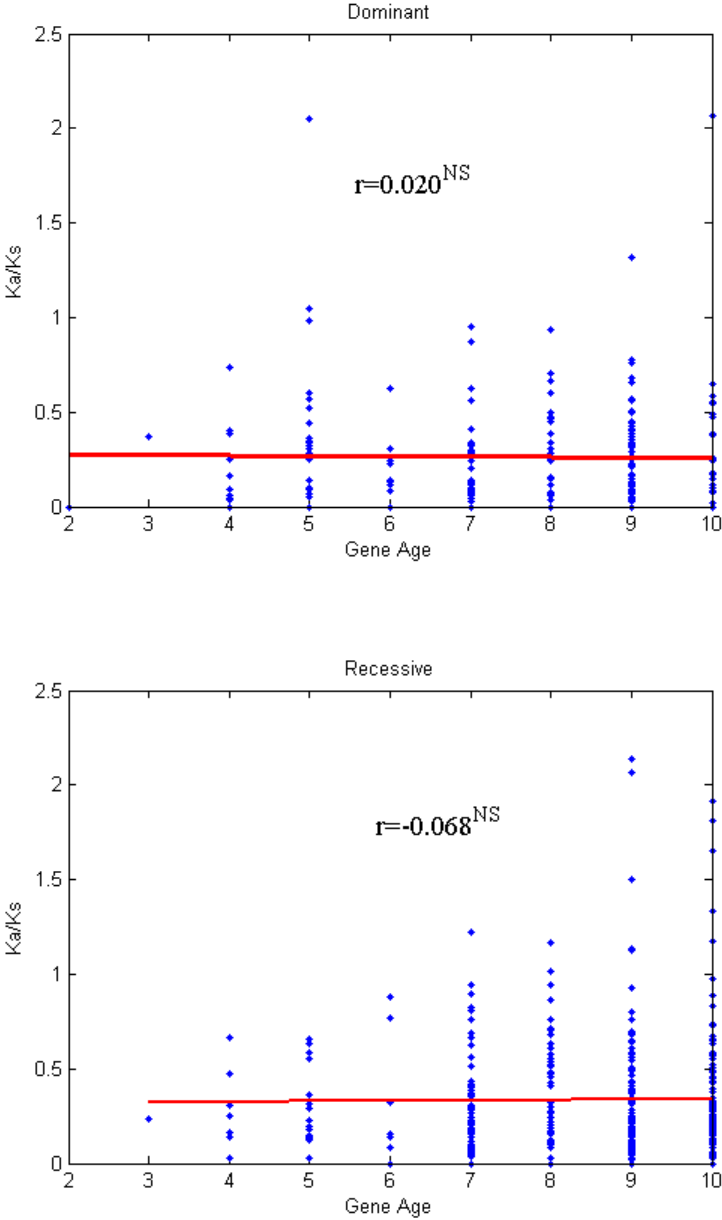
Figure S5. Curves of cumulative distribution functions of gene age. The two curves represent autosomal recessive (AR) and autosomal dominant (AD) Menelian-disease genes, respectively.
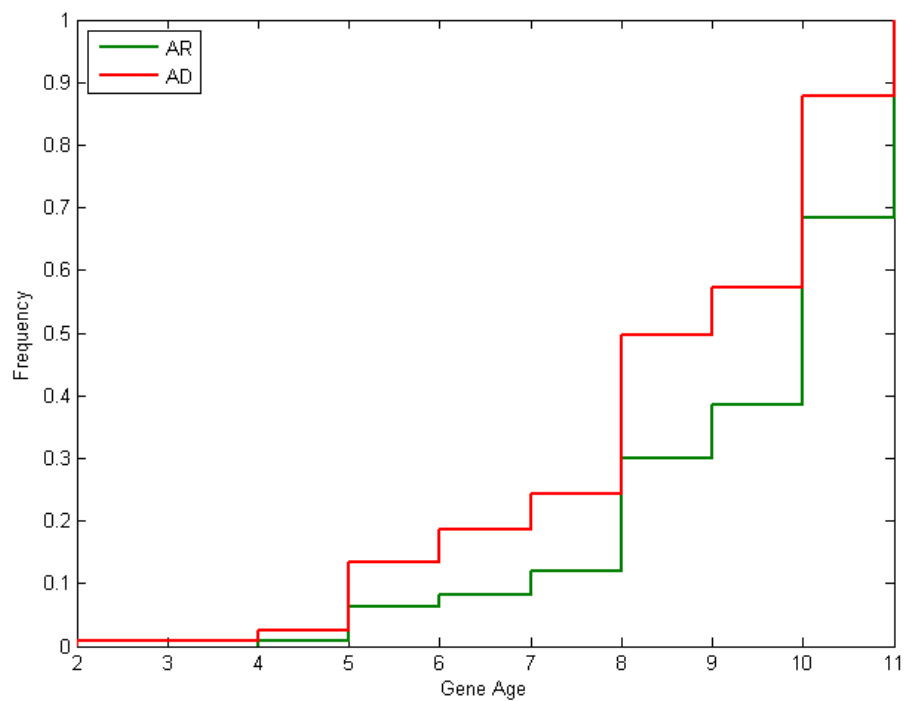
Figure S6. Distribution and portion of genes showing peak expression in tissues. **(A)** Numbers of non-disease, Mendelian-disease, and complex-disease genes that show peak expression in 54 tissues; **(B)** Portions of Mendelian-disease genes in all genes showing peak expression in the corresponding tissues; **(C)** Portions of complex-disease genes in all genes showing peak expression in the corresponding tissues.
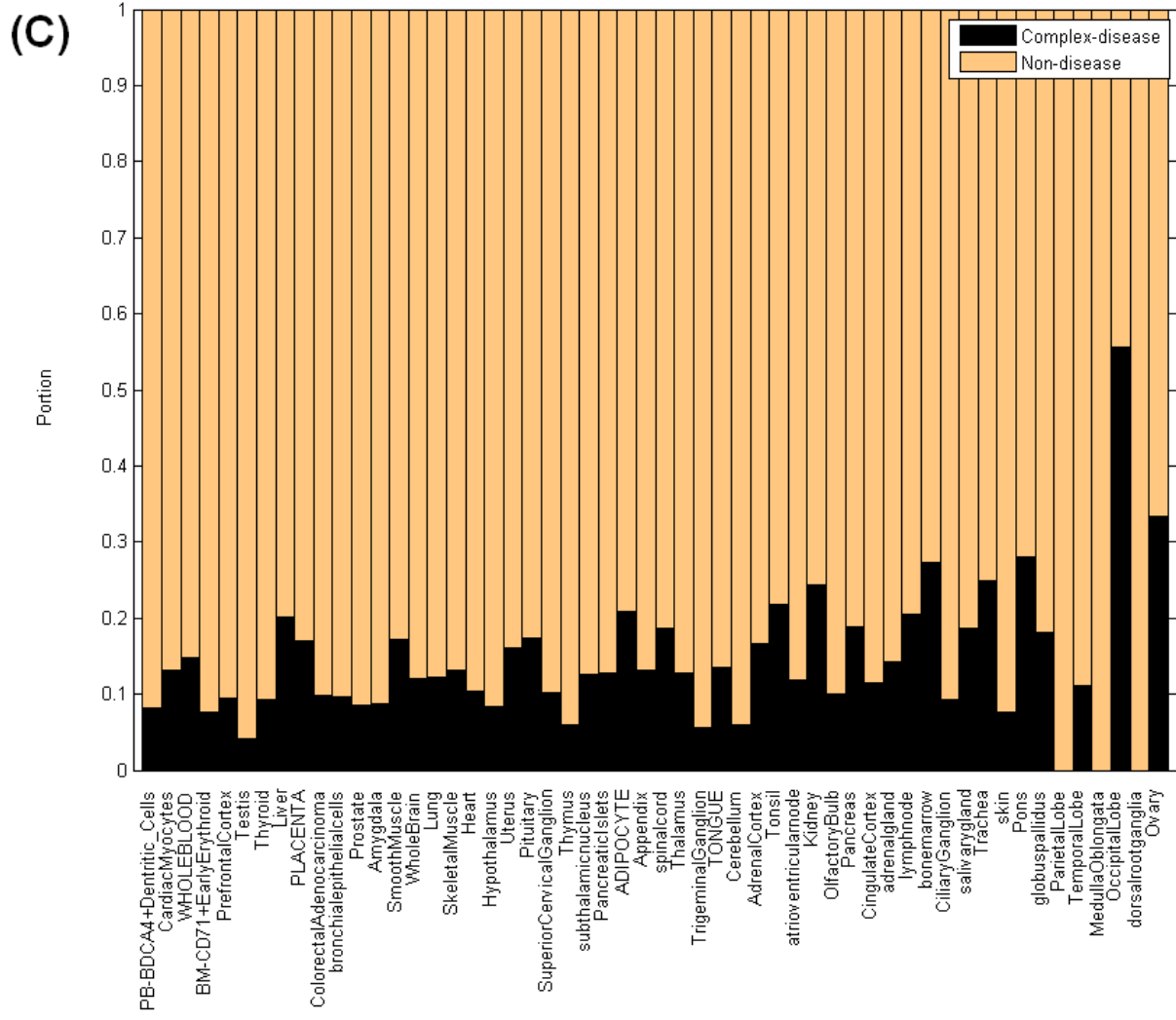
(B)

(C)
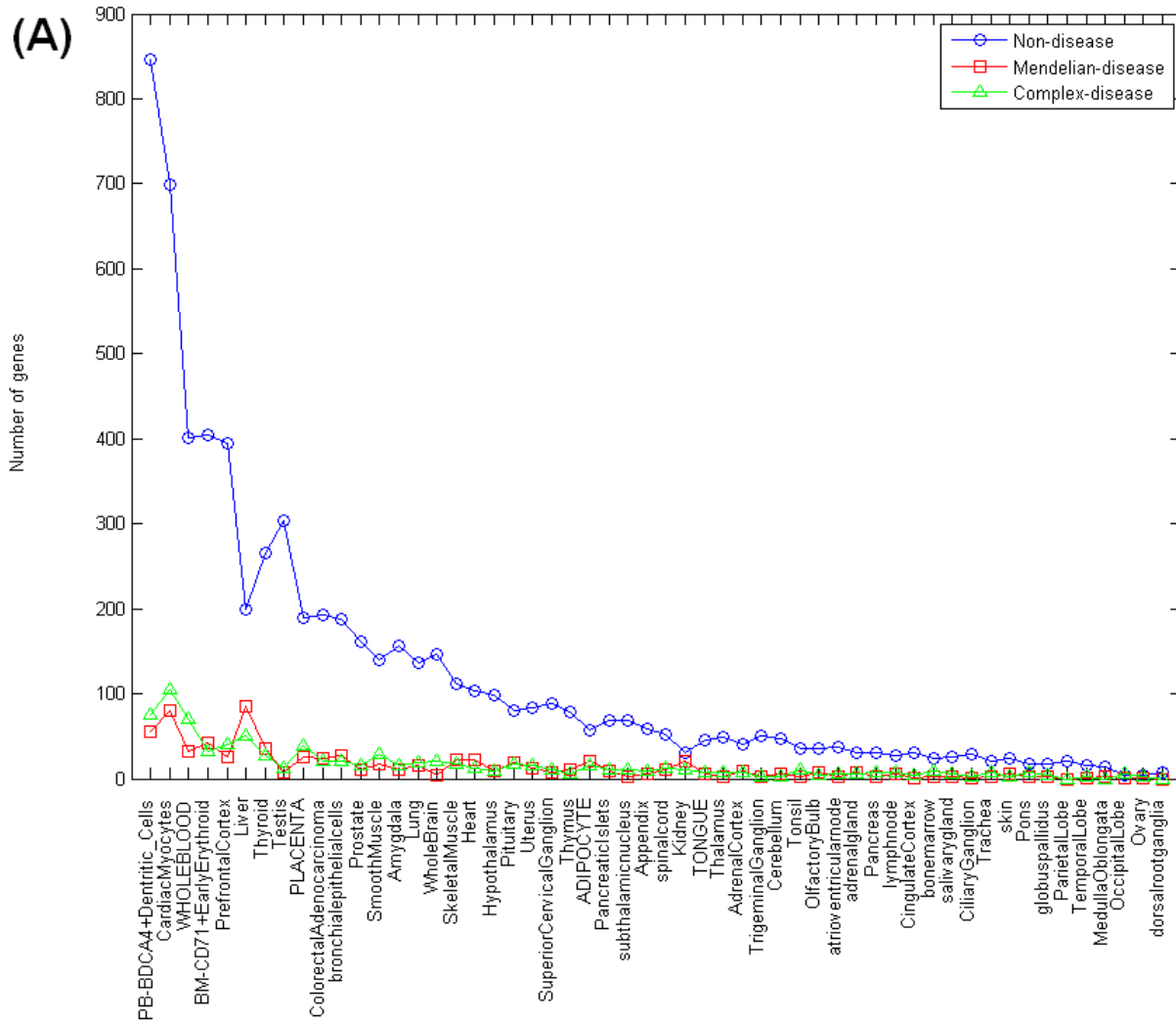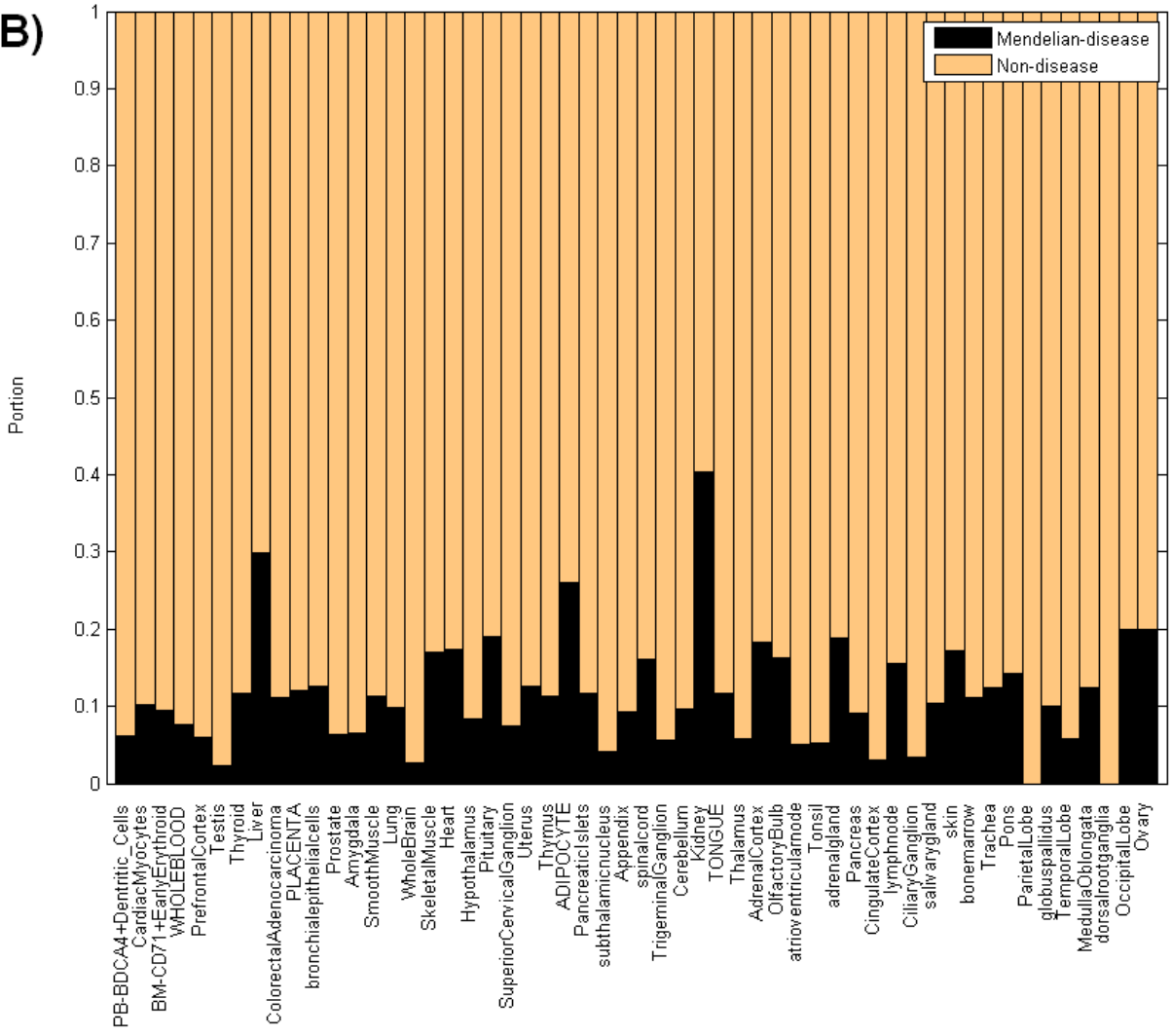
Figure S7. Correlations between Ka/Ks and gene expression for non-disease genes. Non-diseases genes are grouped into Young-, Middle-, and Old-aged groups. For each group, scatter plots of Ka/Ks vs. each of three gene-expression parameters, aveExp, hetExp, and maxExp, are shown. Spearman's correlation coefficients between Ka/Ks and gene expression are given. The symbol ** indicates P<0.001, * indicates P<0.01, and NS (not significant) indicates P≥0.01. The three age groups were constructed using **(A)** Equally populated bins **(B)** Equally spaced bins.

(A)
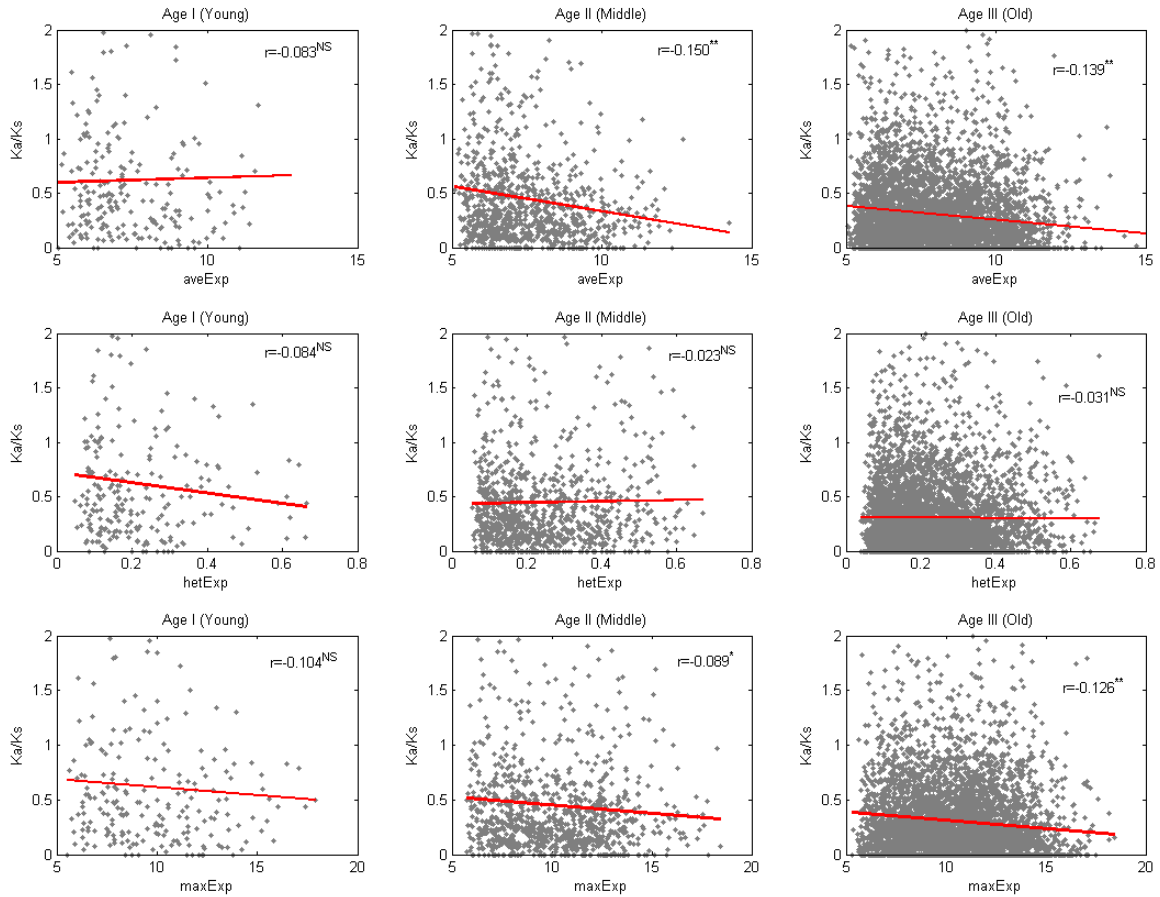
(B)

Figure S8. Correlation between Ka/Ks and gene expression for Mendelian-disease genes. Legends and symbols are identical to Figure S6. The three age groups were constructed using **(A)** Equally populated bins **(B)** Equally spaced bins.

(A)

(B)
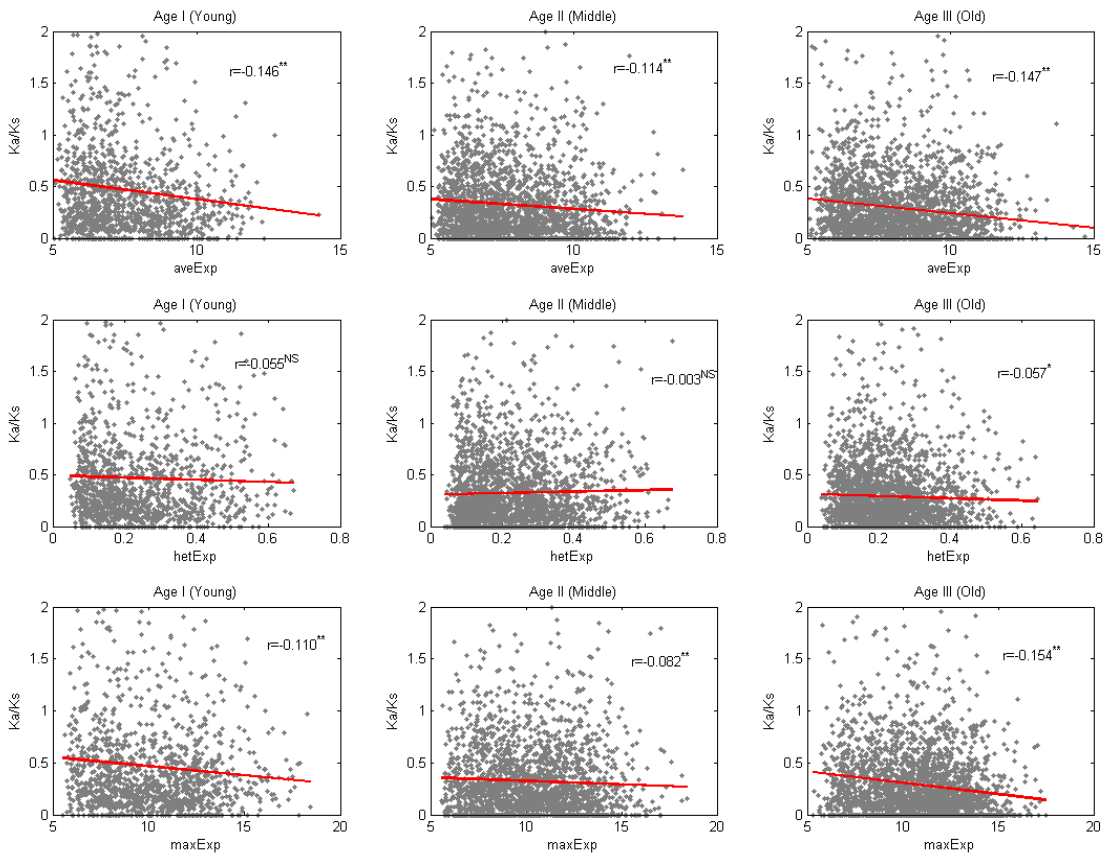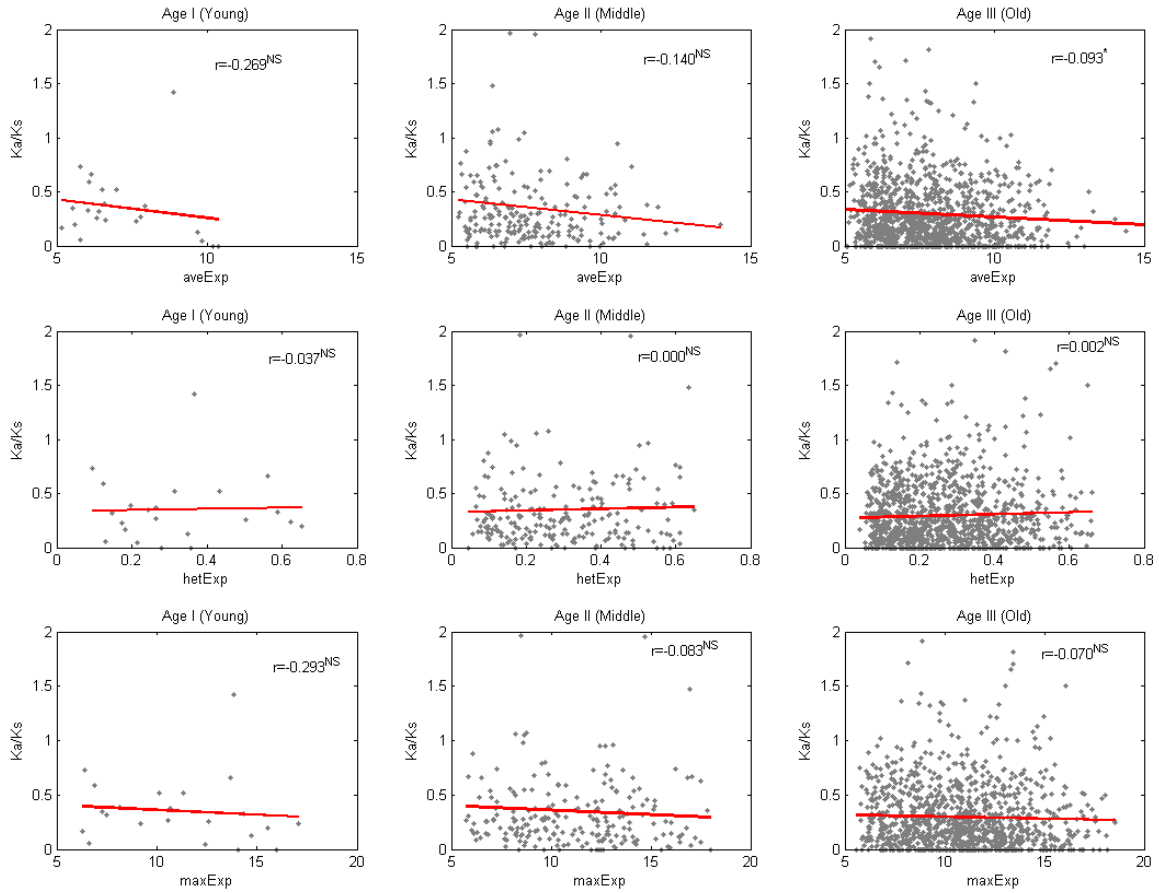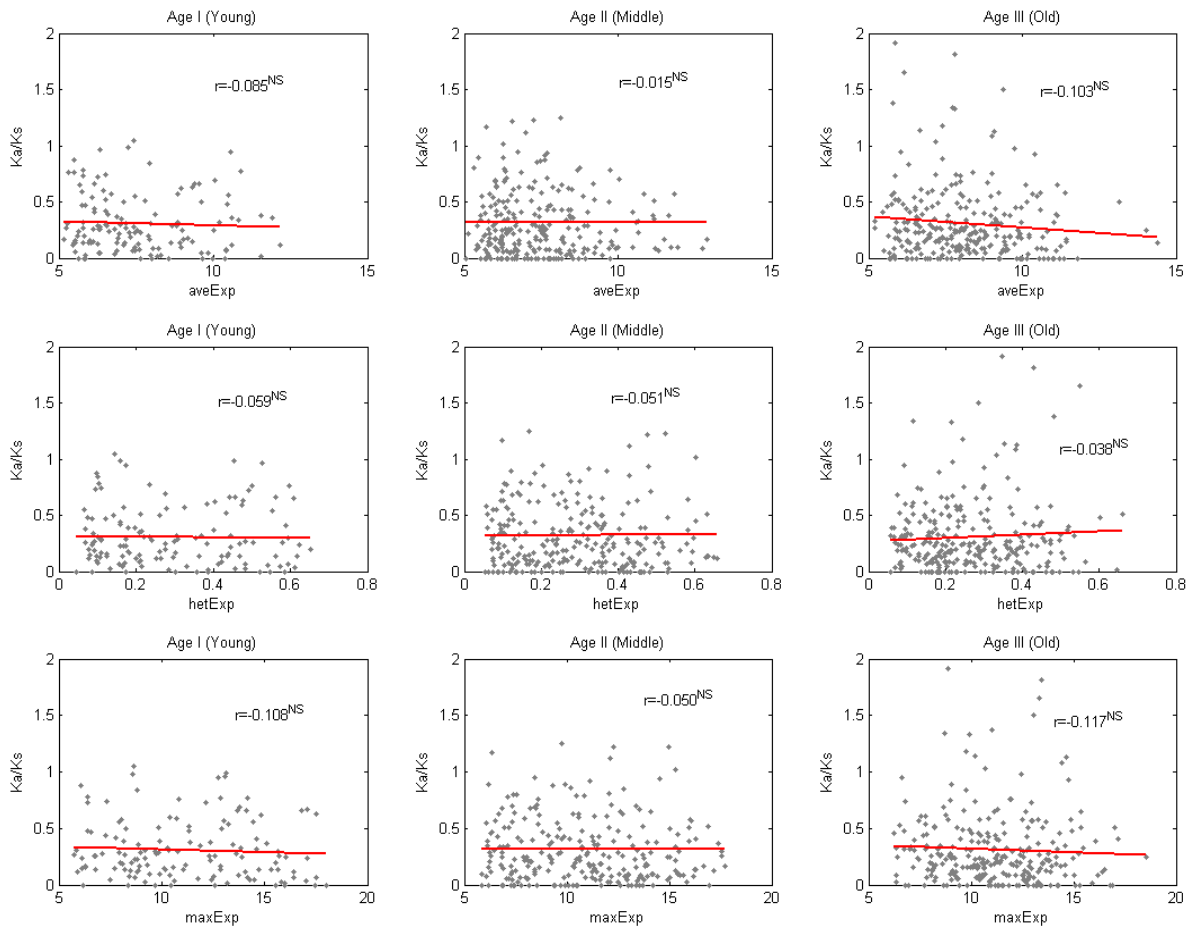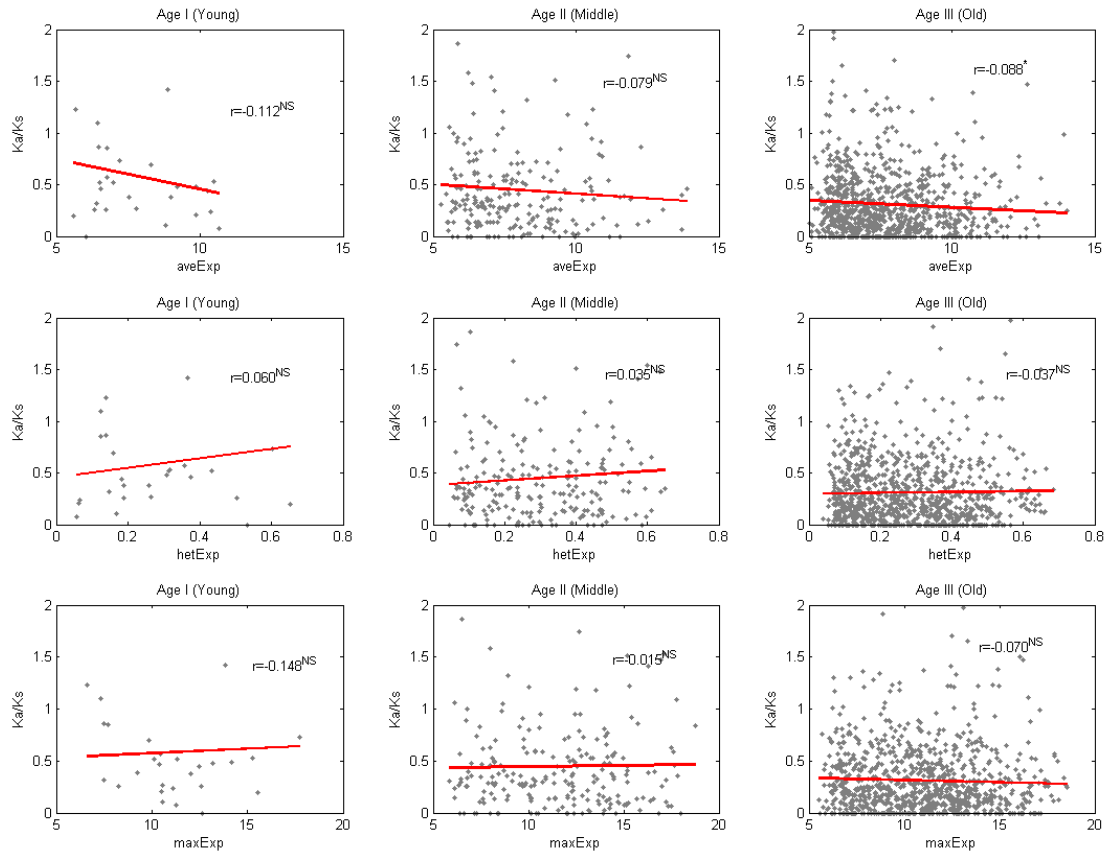
Figure S9. Correlation between Ka/Ks and gene expression for complex-disease genes. Legends and symbols are identical to Figure S6. The three age groups were constructed using **(A)** Equally populated bins **(B)** Equally spaced bins.
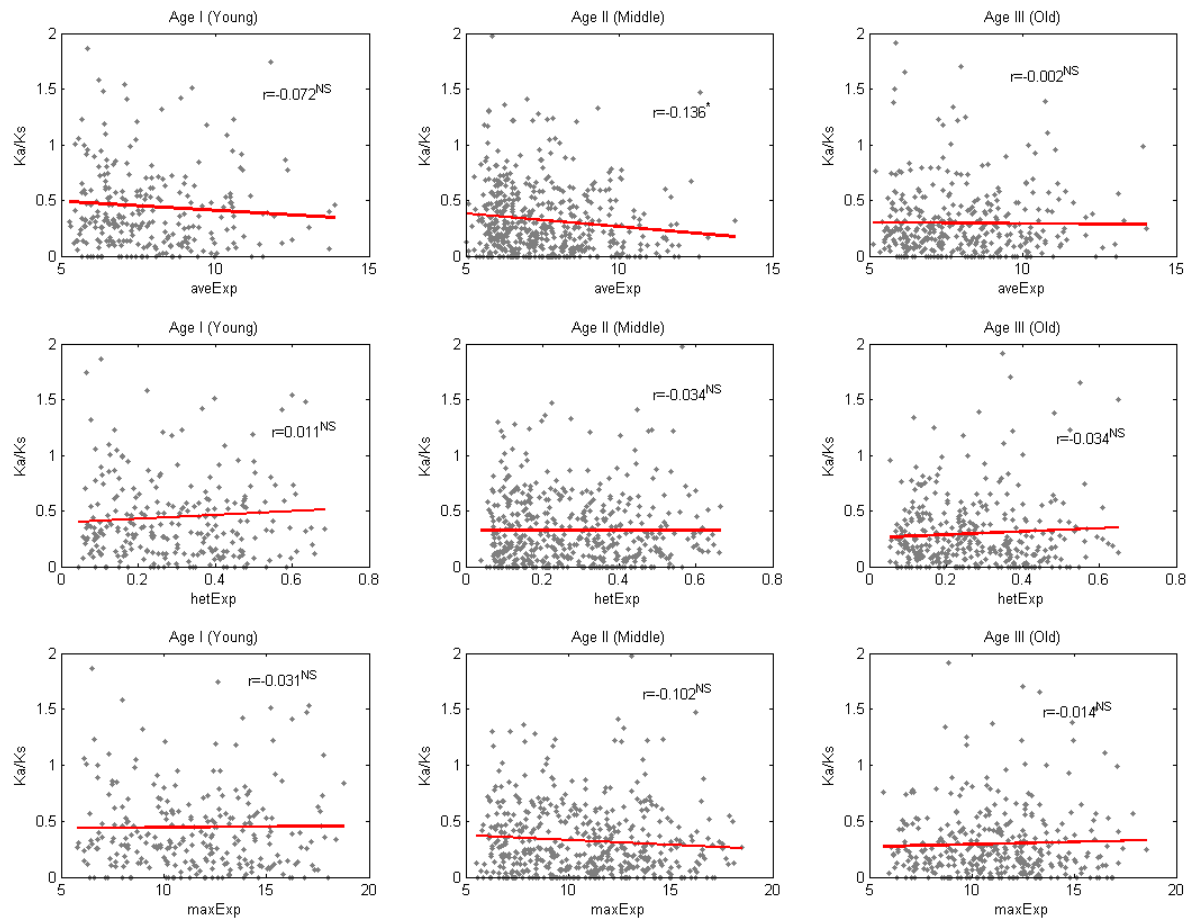
(A)

(B)

Table S1. GO categories that are overrepresented in "Young Mendelian-disease genes" versus "Young non-disease genes" sets.

Term	#1 vs #2	p value	Adjusted p value
#(Terms of DB: GO biological process at level 3)			
anatomical structure development (GO:0048856)	78.08% 21.92%	1.26E-27	8.31E-26
multicellular organismal development (GO:0007275)	74.4% 25.6%	4.77E-22	1.57E-20
circulation (GO:0008015)	95.69% 4.31%	4.73E-14	1.04E-12
response to stress (GO:0006950)	77.89% 22.11%	5.08E-12	8.38E-11
cellular component organization and biogenesis (GO:0016043)	71.72% 28.28%	1.24E-11	1.64E-10
response to external stimulus (GO:0009605)	78.48% 21.52%	1.45E-10	1.60E-09
coagulation (GO:0050817)	94.07% 5.93%	1.87E-09	1.76E-08
cellular developmental process (GO:0048869)	68.62% 31.38%	1.31E-08	1.08E-07
regulation of body fluids (GO:0050878)	91.35% 8.65%	3.05E-08	2.01E-07
establishment of localization (GO:0051234)	70.17% 29.83%	2.86E-08	2.01E-07
muscle contraction (GO:0006936)	90.49% 9.51%	6.70E-08	4.02E-07
regulation of biological quality (GO:0065008)	81.55% 18.45%	2.46E-07	1.35E-06
cell adhesion (GO:0007155)	69.57% 30.43%	3.58E-07	1.82E-06
cell proliferation (GO:0008283)	73.99% 26.01%	7.42E-07	3.50E-06
cell activation (GO:0001775)	80.62% 19.38%	1.72E-06	7.56E-06
regulation of a molecular function (GO:0065009)	77.89% 22.11%	1.96E-05	8.10E-05
death (GO:0016265)	70.92% 29.08%	6.73E-05	0.000261
#(Terms of DB: GO biological process at level 4)			
system development (GO:0048731)	78.54% 21.46%	4.89E-25	7.96E-23
protein metabolic process (GO:0019538)	68.56% 31.44%	2.20E-12	1.19E-10
regulation of multicellular organismal process (GO:0051239)	86.76% 13.24%	1.66E-12	1.19E-10
positive regulation of biological process (GO:0048518)	75% 25%	1.22E-10	4.96E-09
response to wounding (GO:0009611)	80.17% 19.83%	2.28E-10	7.43E-09
negative regulation of biological process (GO:0048519)	75.65% 24.35%	3.88E-10	1.05E-08
anatomical structure morphogenesis (GO:0009653)	75.63% 24.37%	9.92E-10	2.31E-08
hemostasis (GO:0007599)	94.06% 5.94%	1.85E-09	3.78E-08
cellular macromolecule metabolic process (GO:0044260)	66.65% 33.35%	3.40E-09	6.16E-08
cell differentiation (GO:0030154)	68.61% 31.39%	1.19E-08	1.94E-07

	transport (GO:0006810)	70.8% 29.2%	2.08E-08	3.09E-07
	homeostatic process (GO:0042592)	85.15% 14.85%	1.78E-07	2.42E-06
	regulation of coagulation (GO:0050818)	100% 0%	8.17E-07	1.02E-05
	immune effector process (GO:0002252)	87.22% 12.78%	2.13E-06	2.48E-05
	adaptive immune response (GO:0002250)	89.42% 10.58%	3.07E-06	3.33E-05
	blood pressure regulation (GO:0008217)	100% 0%	6.11E-06	6.22E-05
	lipid metabolic process (GO:0006629)	84.08% 15.92%	7.48E-06	6.77E-05
	embryonic development (GO:0009790)	91.35% 8.65%	7.13E-06	6.77E-05
	alcohol metabolic process (GO:0006066)	94.41% 5.59%	1.23E-05	0.0001
	regulation of developmental process (GO:0050793)	83.35% 16.65%	1.18E-05	0.0001
	regulation of catalytic activity (GO:0050790)	77.99% 22.01%	2.98E-05	0.000231
	humoral immune response (GO:0006959)	79.38% 20.62%	3.63E-05	0.000269
	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO:0006139)	38.93% 61.07%	7.90E-05	0.00056
	organic acid metabolic process (GO:0006082)	87.7% 12.3%	0.000128	0.000868
	heart contraction (GO:0060047)	95% 5%	0.000134	0.000874
	organelle organization and biogenesis (GO:0006996)	68.77% 31.23%	0.000148	0.000927
	phosphorus metabolic process (GO:0006793)	67.14% 32.86%	0.000158	0.000957
#(Terms of DB: GO molecular function at level 3)				
	protein binding (GO:0005515)	64.86% 35.14%	2.02E-21	1.46E-19
	extracellular matrix structural constituent (GO:0005201)	91.69% 8.31%	1.48E-10	5.34E-09
	lipid transporter activity (GO:0005319)	97.21% 2.79%	4.40E-08	1.06E-06
	structural constituent of cytoskeleton (GO:0005200)	88.19% 11.81%	3.20E-07	5.75E-06
	nucleic acid binding (GO:0003676)	35.17% 64.83%	7.38E-07	1.06E-05
	pattern binding (GO:0001871)	83.94% 16.06%	5.74E-05	0.000689
#(Terms of DB: GO molecular function at level 4)				
	cytoskeletal protein binding (GO:0008092)	81.42% 18.58%	1.85E-07	2.57E-05
	enzyme binding (GO:0019899)	82.99% 17.01%	1.41E-06	9.78E-05
	purine nucleotide binding (GO:0017076)	67.19% 32.81%	5.94E-06	0.000275