Supplementary Data

Supplementary Table Titles

Table S1. List of human disease genes and the complex and/or Mendelian diseases they have been found associated with.

Table S2. Genomic location, protein network parameters and summary statistics of neutrality for all human genes.

Table S3. Odd-ratios and age of onset for all human disease genes associated to complex diseases.

Table S4. Correspondence between the complex and Mendelian diseases used in our study and the traits used in Blair et al. (1).

Table S5. List of CM genes and the complex and Mendelian traits found to co-occur in Blair et al. (1).

Supplementary Figures

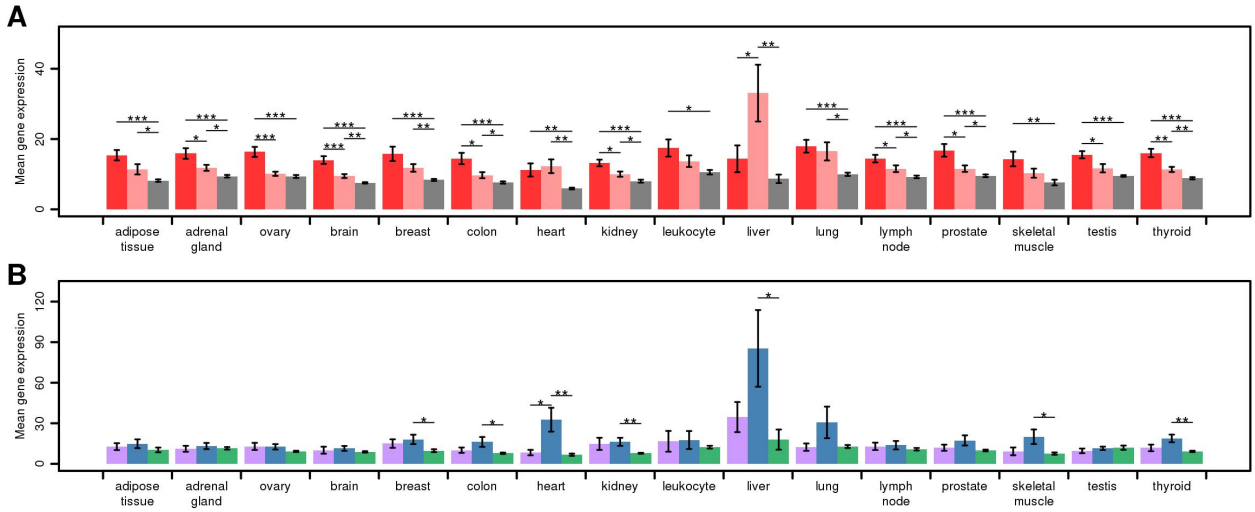

Figure S1. Mean gene expression levels in 16 human tissues. Mean and standard error expression levels at each single tissue available in the Expression Atlas database. (A) Mean expression levels in the three human gene groups considered. END genes are shown in red, HD genes in light red and NDNE genes in gray. (B) Mean expression levels in the three HD genes considered. CM genes are shown in violet, MNC genes in blue and CNM genes in green. *, ** and *** represent that for a given comparison significance is reached at 0.05, 0.005 and 0.0005 levels, respectively, for a two-sided T-test.

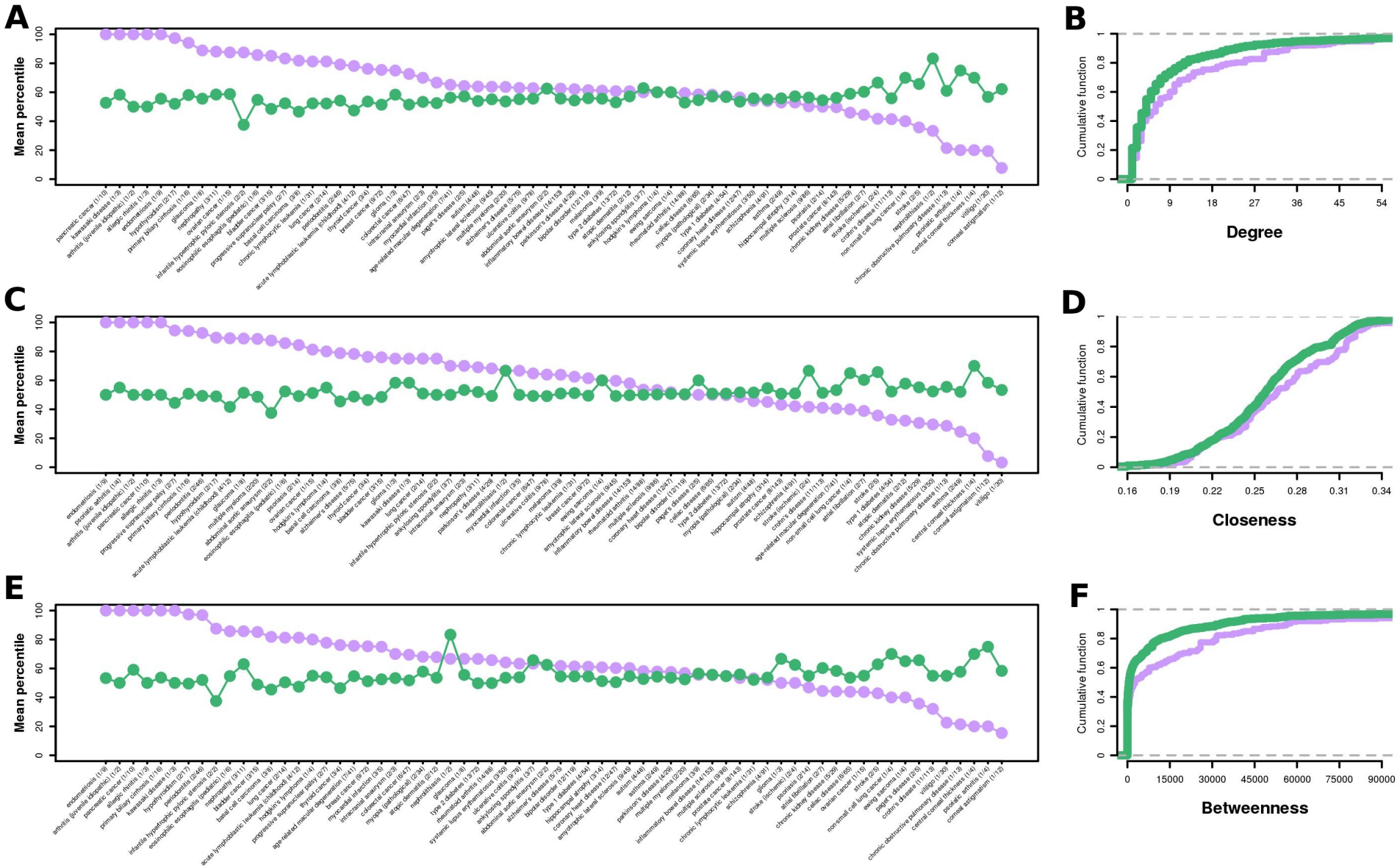
A
Mean percentile
0
20
40
60
80
100
pancreatic cancer (1/10)
kawasaki disease (1/3)
arthritis (juvenile idiopathic) (1/2)
allergic rhinitis (1/3)
endometriosis (1/9)
hypothyroidism (2/17)
primary biliary cirrhosis (1/16)
glaucoma (1/8)
nephropathy (3/11)
ovarian cancer (1/15)
infantile hypertrophic pyloric stenosis (2/2)
eosinophilic esophagitis (pediatric) (1/6)
bladder cancer (3/15)
progressive supranuclear palsy (2/7)
basal cell carcinoma (3/8)
chronic lymphocytic leukemia (1/31)
lung cancer (2/14)
acute lymphoblastic leukemia (childhood) (4/12)
periodontitis (2/46)
thyroid cancer (3/4)
breast cancer (9/72)
glioma (1/3)
colorectal cancer (6/47)
intracranial aneurysm (2/3)
myocardial infarction (3/5)
age-related macular degeneration (7/41)
psoriasis (2/25)
autism (4/48)
multiple myeloma (2/20)
alzheimer's disease (5/75)
ulcerative colitis (6/78)
abdominal aortic aneurysm (2/2)
inflammatory bowel disease (14/153)
parkinson's disease (4/29)
bipolar disorder (12/119)
melanoma (1/99)
type 2 diabetes (13/72)
atopic dermatitis (2/12)
ankylosing spondylitis (3/7)
hodgkin's lymphoma (1/4)
ewing sarcoma (1/4)
rheumatoid arthritis (14/88)
celiac disease (6/65)
myopia (pathological) (2/34)
type 1 diabetes (4/54)
coronary heart disease (12/47)
systemic lupus erythematosus (3/50)
schizophrenia (4/91)
hippocampal atrophy (3/14)
asthma (2/48)
multiple sclerosis (6/66)
prostate cancer (6/143)
psoriasis (2/14)
chronic kidney disease (6/29)
atrial fibrillation (2/7)
stroke (ischemic) (2/4)
crohn's disease (11/113)
non-small cell lung cancer (1/4)
stroke (2/5)
nephritislasis (1/2)
prostatic arthritis (1/4)
central corneal thickness (1/4)
chronic obstructive pulmonary disease (1/13)
vitiligo (1/30)
corneal astigmatism (1/12)
B
Cumulative function
0
0.2
0.4
0.6
0.8
1
0
9
18
27
36
45
54
Degree
C
Mean percentile
0
20
40
60
80
100
endometriosis (1/9)
psoriatic arthritis (1/4)
arthritis (juvenile idiopathic) (1/2)
pancreatic cancer (1/10)
allergic rhinitis (1/3)
progressive supranuclear palsy (2/7)
primary biliary cirrhosis (1/16)
periodontitis (2/46)
hypothyroidism (2/17)
acute lymphoblastic leukemia (childhood) (4/12)
multiple myeloma (2/20)
glaucoma (1/8)
abdominal aortic aneurysm (2/2)
eosinophilic esophagitis (pediatric) (1/6)
psoriasis (2/14)
ovarian cancer (1/15)
hodgkin's lymphoma (1/4)
basal cell carcinoma (3/8)
alzheimer's disease (5/75)
thyroid cancer (3/4)
bladder cancer (3/15)
glioma (1/3)
lung cancer (2/14)
kawasaki disease (1/3)
lung cancer (2/14)
infantile hypertrophic pyloric stenosis (2/2)
ankylosing spondylitis (3/7)
intracranial aneurysm (2/3)
nephropathy (3/11)
parkinson's disease (4/29)
nephritislasis (1/2)
myocardial infarction (3/5)
colorectal cancer (6/47)
ulcerative colitis (6/78)
chronic lymphocytic leukemia (1/31)
melanoma (3/99)
breast cancer (9/72)
ewing sarcoma (1/4)
amyotrophic lateral sclerosis (6/45)
inflammatory bowel disease (14/153)
rheumatoid arthritis (14/88)
multiple sclerosis (6/66)
coronary heart disease (12/47)
bipolar disorder (12/119)
psoriasis (2/25)
celiac disease (6/65)
type 2 diabetes (13/72)
myopia (pathological) (2/34)
autism (4/48)
hippocampal atrophy (3/14)
prostate cancer (6/143)
schizophrenia (4/91)
stroke (ischemic) (2/4)
crohn's disease (11/113)
age-related macular degeneration (7/41)
non-small cell lung cancer (1/4)
atrial fibrillation (2/7)
type 1 diabetes (4/54)
stroke (2/5)
atopic dermatitis (2/12)
chronic kidney disease (5/29)
systemic lupus erythematosus (3/50)
central corneal thickness (1/4)
chronic obstructive pulmonary disease (1/13)
asthma (2/48)
corneal astigmatism (1/12)
vitiligo (1/30)
D
Cumulative function
0
0.2
0.4
0.6
0.8
1
0.16
0.19
0.22
0.25
0.28
0.31
0.34
Closeness
E
Mean percentile
0
20
40
60
80
100
endometriosis (1/9)
arthritis (juvenile idiopathic) (1/2)
pancreatic cancer (1/10)
primary biliary cirrhosis (1/16)
allergic rhinitis (1/3)
kawasaki disease (1/3)
hypothyroidism (2/17)
periodontitis (2/46)
infantile hypertrophic pyloric stenosis (2/2)
eosinophilic esophagitis (pediatric) (1/6)
nephropathy (3/11)
bladder cancer (3/15)
basal cell carcinoma (3/8)
lung cancer (2/14)
acute lymphoblastic leukemia (childhood) (4/12)
hodgkin's lymphoma (1/4)
progressive supranuclear palsy (2/7)
age-related macular degeneration (7/41)
thyroid cancer (3/4)
breast cancer (9/72)
myocardial infarction (3/5)
intracranial aneurysm (2/3)
colorectal cancer (6/47)
myopia (pathological) (2/34)
atopic dermatitis (2/12)
nephritislasis (1/2)
glaucoma (1/8)
type 2 diabetes (13/72)
rheumatoid arthritis (14/88)
systemic lupus erythematosus (3/50)
ulcerative colitis (6/78)
ankylosing spondylitis (3/7)
abdominal aortic aneurysm (2/2)
alzheimer's disease (5/75)
bipolar disorder (12/119)
type 1 diabetes (4/54)
hippocampal atrophy (3/14)
coronary heart disease (12/47)
amyotrophic lateral sclerosis (6/45)
autism (4/48)
asthma (2/48)
parkinson's disease (4/29)
multiple myeloma (2/20)
melanoma (3/99)
inflammatory bowel disease (14/153)
multiple sclerosis (6/66)
prostate cancer (6/143)
chronic lymphocytic leukemia (1/31)
schizophrenia (4/91)
glioma (1/3)
stroke (ischemic) (2/4)
psoriasis (2/14)
atrial fibrillation (2/7)
chronic kidney disease (5/29)
celiac disease (6/65)
ovarian cancer (1/15)
non-small cell lung cancer (1/4)
stroke (2/5)
ewing sarcoma (1/4)
papa's disease (2/6)
crohn's disease (11/113)
chronic obstructive pulmonary disease (1/13)
central corneal thickness (1/4)
prostatic arthritis (1/4)
corneal astigmatism (1/12)
vitiligo (1/30)
F
Cumulative function
0
0.2
0.4
0.6
0.8
1
0
15000
30000
45000
60000
75000
90000
Betweenness

Figure S2. Protein network parameters of genes associated to human complex diseases. (A) Comparison of the mean degree percentiles for all the CM and CNM genes associated to a given trait. (B) Cumulative distributions of degree for the CM and CNM subgroups of HD genes. (C) Comparison of the mean closeness percentiles for all the CM and CNM genes associated to a given trait. (D) Cumulative distributions of closeness for the CM and CNM subgroups of HD genes. (E) Comparison of the mean betweenness percentiles for all the CM and CNM genes associated to a given trait. (F) Cumulative distributions of betweenness for the CM and CNM subgroups of HD genes. CM genes are shown in violet, while CNM genes in green. The number of CM and CNM genes associated to each trait is reported in brackets.

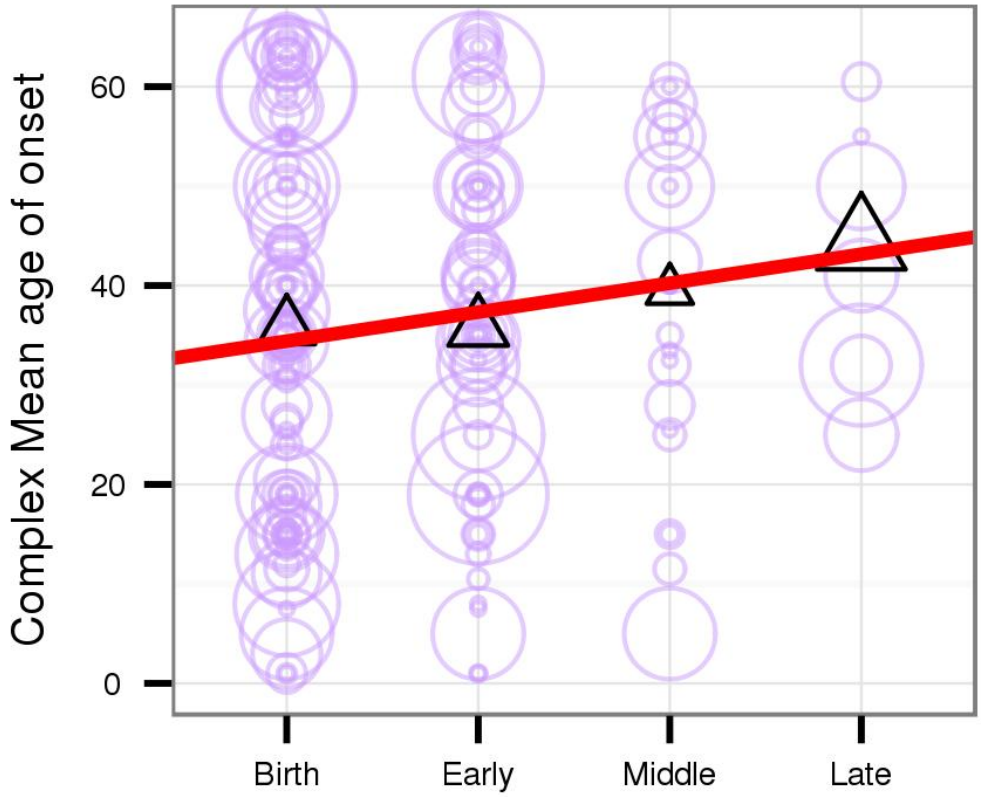

Figure S3. Evolutionary properties and age of onset of CM genes. Correlation between the age of onset for the Mendelian (x-axis) and complex diseases (y-axis) CM genes have been found to be associated with. Circle sizes are proportional to the dN/dS values observed for a given CM gene. Black triangles represent the mean dN/dS calculated over the full set of points of a given bin in the x-axis. The triangle position in the y-axis represent the mean age of onset for all complex traits the CM genes of the bin have been found associated with.

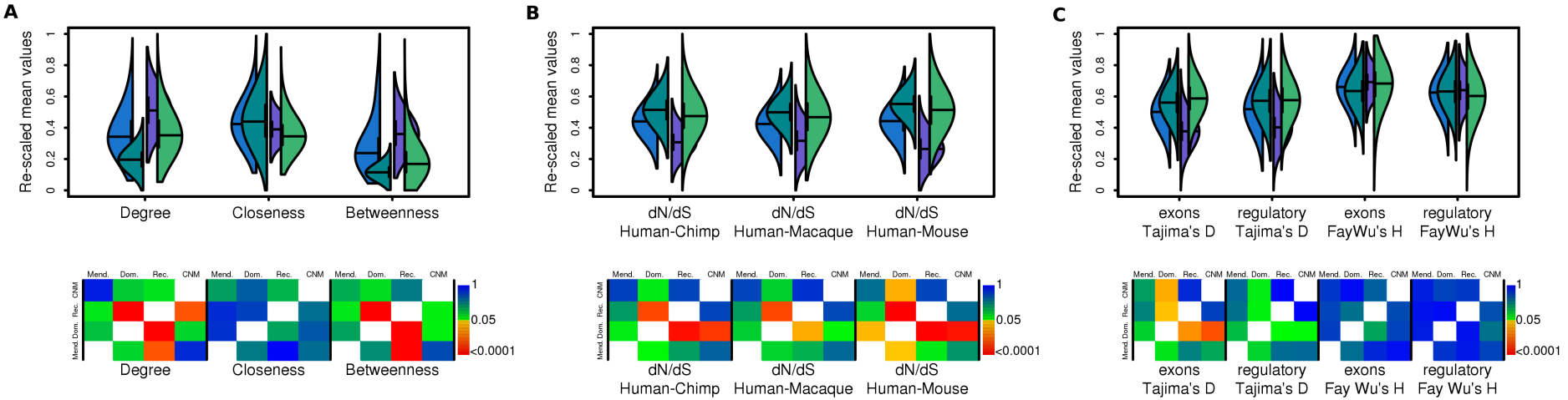
A
Re-scaled mean values
0
0.2
0.4
0.6
0.8
1
Degree
Closeness
Betweenness
Mend.
Dom.
Rec.
CNM
1
0.05
<0.0001
B
dN/dS
Human-Chimp
dN/dS
Human-Macaque
dN/dS
Human-Mouse
C
exons
Tajima's D
regulatory
Tajima's D
exons
FayWu's H
regulatory
FayWu's H
exons
Fay Wu's H
regulatory
Fay Wu's H

Figure S4. Scaled resampled mean values and resampling p-values for three different protein network parameters (A), dN/dS (B), and Tajima's D and Fay and Wu's H (C). Mendelian genes (CM+MNC) are shown in blue, dominant genes in dark violet, recessive genes in dark green, CNM genes in light green. In the panels below each figure, the corresponding p-values for the different pairwise comparisons are represented. Below the diagonal, p-values obtained when resampling the group in the column and comparing it to the mean of whole set in the row. P-values above the diagonal are calculated resampling the group in the row and comparing it to the mean of whole set in the column.

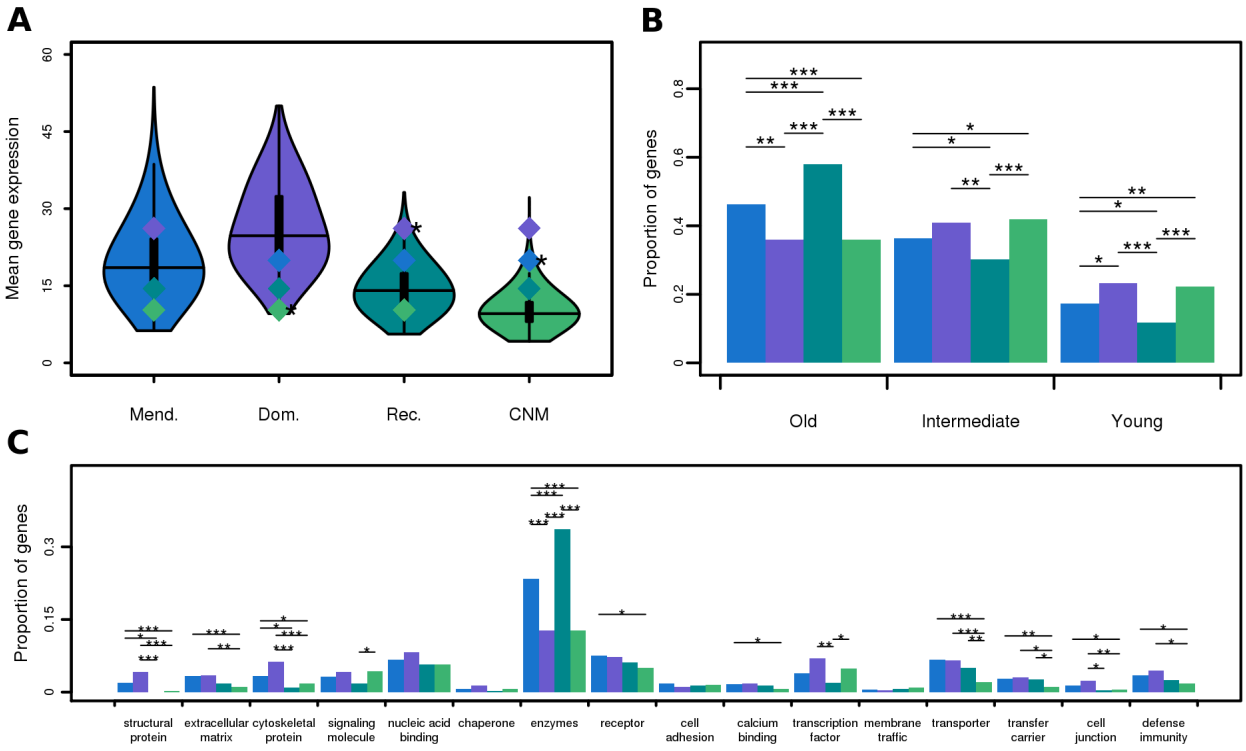

Figure S5. Biological features of different subgroups of human disease genes. (A) Resampling expression levels over 16 different human tissues reported in the Expression Atlas. Mendelian genes (CM+MNC) are shown in blue, dominant genes in dark violet, recessive genes in dark green, CNM genes in light green. At each of the 10,000 resampling, 100 genes were selected and their mean expression levels over 16 tissues was calculated. The mean expression over the 100 genes was thus calculated and the 10,000 mean values are represented in the distributions. Diamonds represent the mean expression values calculated over the whole set of genes of a group. Asterisks indicate that the mean expression is found either on the top or bottom 2.5% of the resampling distribution. (B) Proportions of genes in three age bins. Gene ages were retrieved from PhyloPat database. (C) Proportions of genes in different protein functions considered in PANTHER database. Mendelian genes (CM+MNC) are shown in blue, dominant genes are shown in dark violet, recessive genes in dark green, CNM genes in light green. *, ** and *** represent that for a given comparison significance is reached at 0.05, 0.005 and 0.0005 levels respectively for a chi-square test (B and C).

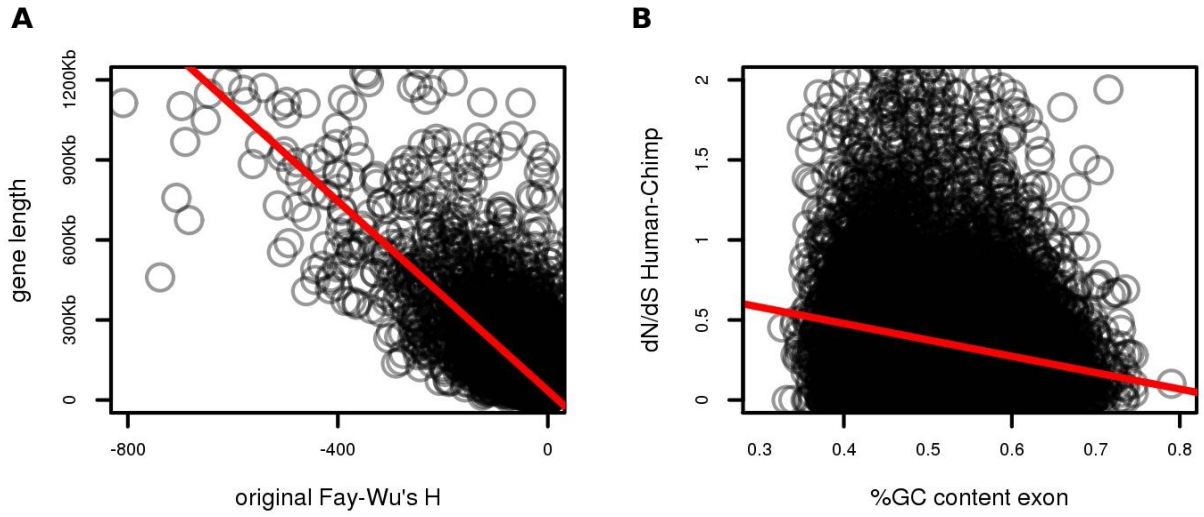

Figure S6. Parameter correlations and corrections. (A) Correlation between gene length and computed values of Fay and Wu's H (Spearman's rho=-0.43; p-value<$2.2x10^{-16}$). (B) Correlation between GC content and dN/dS values observed for all the human genes (Spearman's rho=-0.10; p-value<$2.2x10^{-16}$).

Supplementary Notes.

Supplementary Note 1. Analysis of statistical significance for comparisons of protein network parameters, summary statistics of neutrality and expression levels between gene groups and subgroups.

Statistical significance for comparisons of protein network parameters, summary statistics of neutrality and expression profiles between gene groups and subgroups was evaluated through a resampling test (as already presented in the main text) as well as using two-sided T-tests and Mann-Whitney U-tests (presented only here as Supplementary Material).

For each comparison, we ran a resampling test, on which for each pair of groups we resampled 10,000 times a fixed number of genes in one of the groups. At each resampling, the mean for a given statistic was calculated and the distribution of the 10,000 resampled means of a given group was compared to the observed mean value obtained in the non-resampled gene group of the pair under comparison. Thus, in the panels below Figures 1 and 3 in the main text, two different p-values are reported for each pair of gene groups or sub-groups comparison. Below the diagonal are indicated the p-values obtained when resampling the group in the column and comparing it to the mean of whole set in the row, while p-values above the diagonal are calculated resampling the group in the row and comparing it to the mean of whole set in the column. For comparisons involving END, HD and NDNE, 1,000 genes were sampled from each group at each resampling (Figs. 1 and 2A in the main text and Figure SN1.1); for comparisons among CNM, MNC and CM, only 100 genes were sampled from each subgroup at each resampling (Figs. 3 and 4A in the main text and Fig. SN1.2).

Additionally, comparisons of protein network parameters, summary statistics of neutrality and expression profiles were also evaluated using two-sided T-tests and Mann-Whitney U-tests.

Statistical significance for comparisons among the gene groups defined in the main text is presented in the “whole set” section of the bottom panels in Figs. SN1.1 and SN1.2. In each panel, T-test p-values are shown below the diagonal, while U-test p-values are shown above the diagonal. In the “resampling” section of the bottom panels in Figs. SN1.1 and SN1.2 are reported exactly the same p-values showed in Figs. 1 and 3 in the main text.

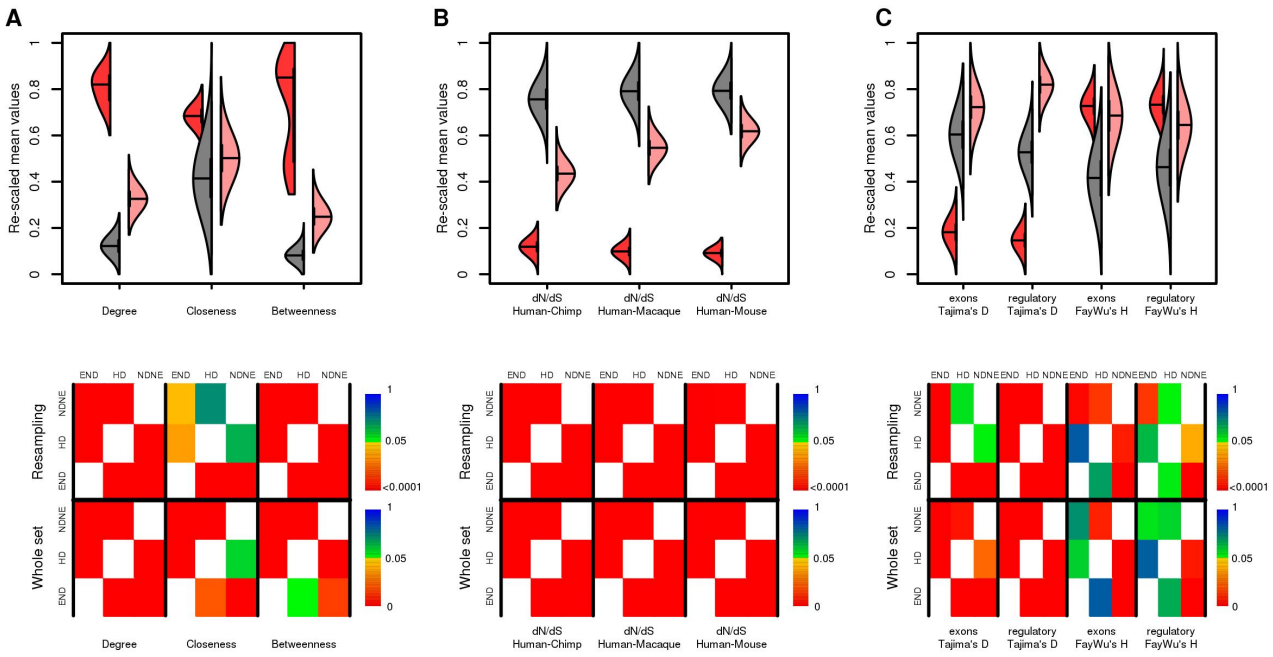

Figure SN1.1. Protein network parameters and summary statistics of neutrality for the three groups of human genes considered. Scaled resampled mean values for protein network parameters (A), dN/dS (B), Tajima's D and Fay and Wu's H (C). END genes are shown in dark red, HD genes in light red, NDNE genes in gray. P-values for protein network parameters, dN/dS values, and Tajima's D and Fay and Wu's H comparisons are shown below the corresponding parameter distribution figure.

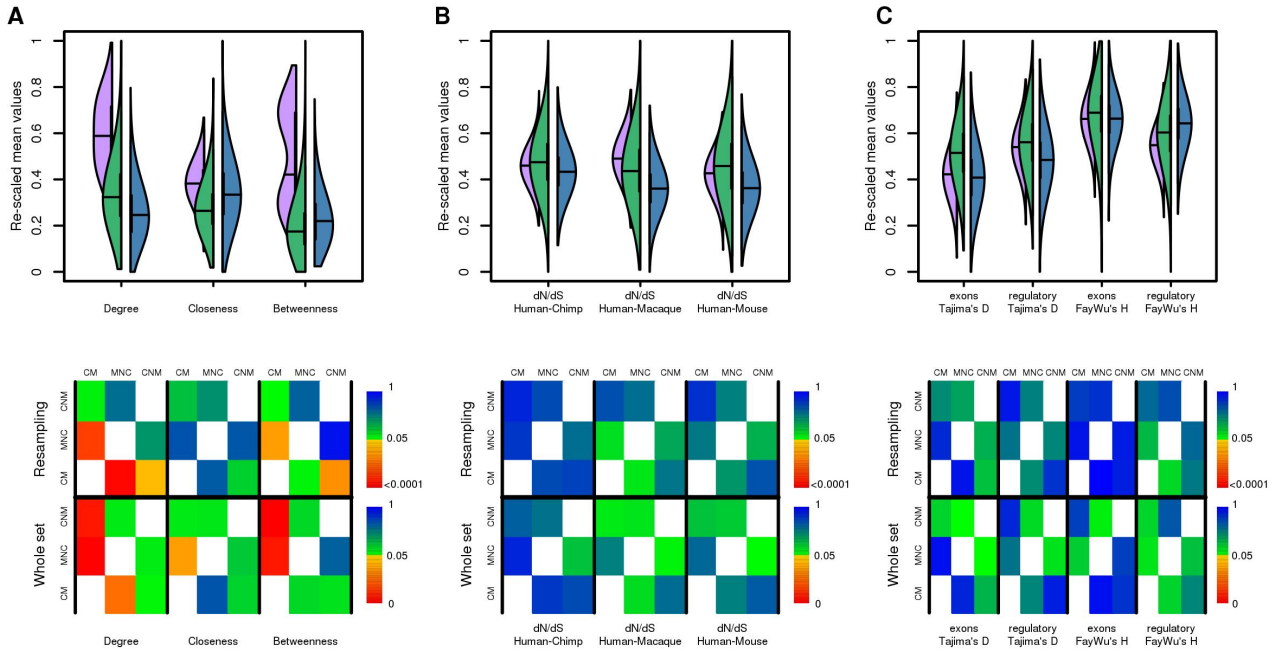

Figure SN1.2. Protein network parameters and summary statistics of neutrality for the three groups of HD genes considered. Scaled resampled mean values for protein network parameters (A), dN/dS (B), Tajima's D and Fay and Wu's H (C). CM genes are shown in violet, MNC genes in blue, CNM genes in green. P-values for protein network parameters, dN/dS values, and Tajima's D and Fay and Wu's H comparisons are shown below the corresponding parameter distribution figure.

Supplementary Note 2. Analysis of possible discovery bias affecting GWAS genes.

The power of an association study strongly depends on the frequency of both causal variants and genetic markers and on the extent of linkage disequilibrium (LD) between them. The higher the MAF and the longer the LD block length, the higher the probability to discover a significant GWAS hit (2). Thus, it may be the case that genes discovered by GWAS have site frequency spectra shifted towards higher allele frequencies and are located in genomic regions with high LD (3). In addition, since very large genes are tagged by more variants, it is more likely to detect significant signals in or around large genes. Indeed, genes reported in the GWAS catalog are longer than the rest of HD and than the rest of human genes (data not shown). Another potential source of bias is the fact that genes reported in GWAS may tend to be well-known genes, precisely because such candidates have been selected by experts according to their biological function in relation to the disease. Effectively, since the associated variants found by GWAS fall near a number of genes, picking just one of these genes as the candidate to harbor true susceptibility variants will result inevitably in the presence of false positives and false negatives in the GWAS catalog. In the following sections, we will explore these potential discovery biases and we will try to estimate how and to which extent they influence the perceived properties of HD genes.

2.1. Gene length bias

In order to evaluate whether gene length could somehow have an effect on the properties of the site frequency spectrum, we first calculated the correlation between gene length and Tajima's D values computed separately on exonic and regulatory sequences. As seen in Figure SN2.1.1, only a very slight correlation exists between gene length and Tajima's D, indicating that the site frequency spectrum of HD genes should not be substantially biased by the enrichment of long genes in the GWAS catalog. Indeed, only 0.1% of the variance in Tajima’s D can be explained by gene length.

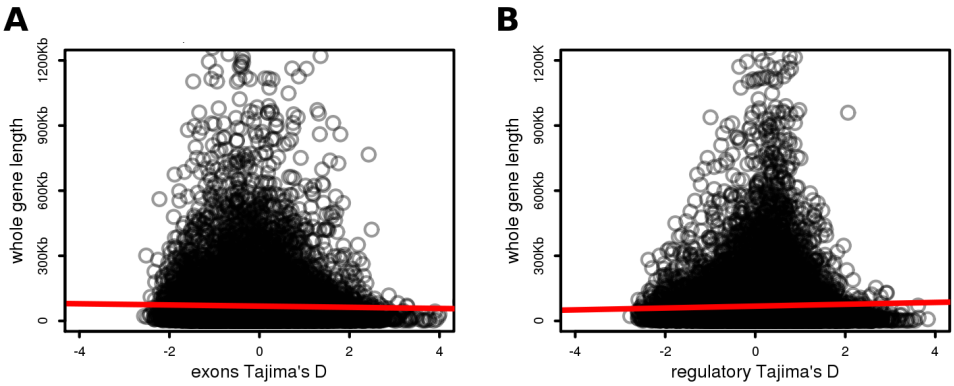

Figure SN2.1.1 Correlation between gene length and Tajima's D. (A) Correlation between the whole gene length and Tajima's D calculated on exonic sequences (Pearson's R=-0.02; p-value=0.006). (B) Correlation between the whole gene length and Tajima's D calculated on regulatory sequences (Pearson's R=0.03; p-value=2.70×10^{-5}).

Next, we repeated all the analyses concerning protein network properties, dN/dS values, neutrality summary statistics and expression levels considering only sets of HD and NDNE genes matched for gene length. We generated 1,000 resampling sets of HD and NDNE genes matching in size. At each resampling, 1,000 genes from the HD set were taken randomly and subsequently 1,000 genes matching in size were obtained from the NDNE set. To be considered a match, NDNE genes should have a length difference <10 Kb with the corresponding HD gene. Since the average and the standard deviation of the length of human genes are 68 Kb and 130 Kb, respectively, we believe that 10 Kb difference represents a good approximation to define two genes similar in length. As shown in Figure SN2.1.2, length matched HD genes continue to exhibit intermediate relevance in the protein network, intermediate rates of protein coding evolution, intermediate expression levels and a site frequency spectrum shifted toward higher amount of intermediate frequency variants. Overall, these results perfectly agree with what can be observed with the whole dataset and confirm that the length of HD genes does not affect directly their evolutionary and biological properties.

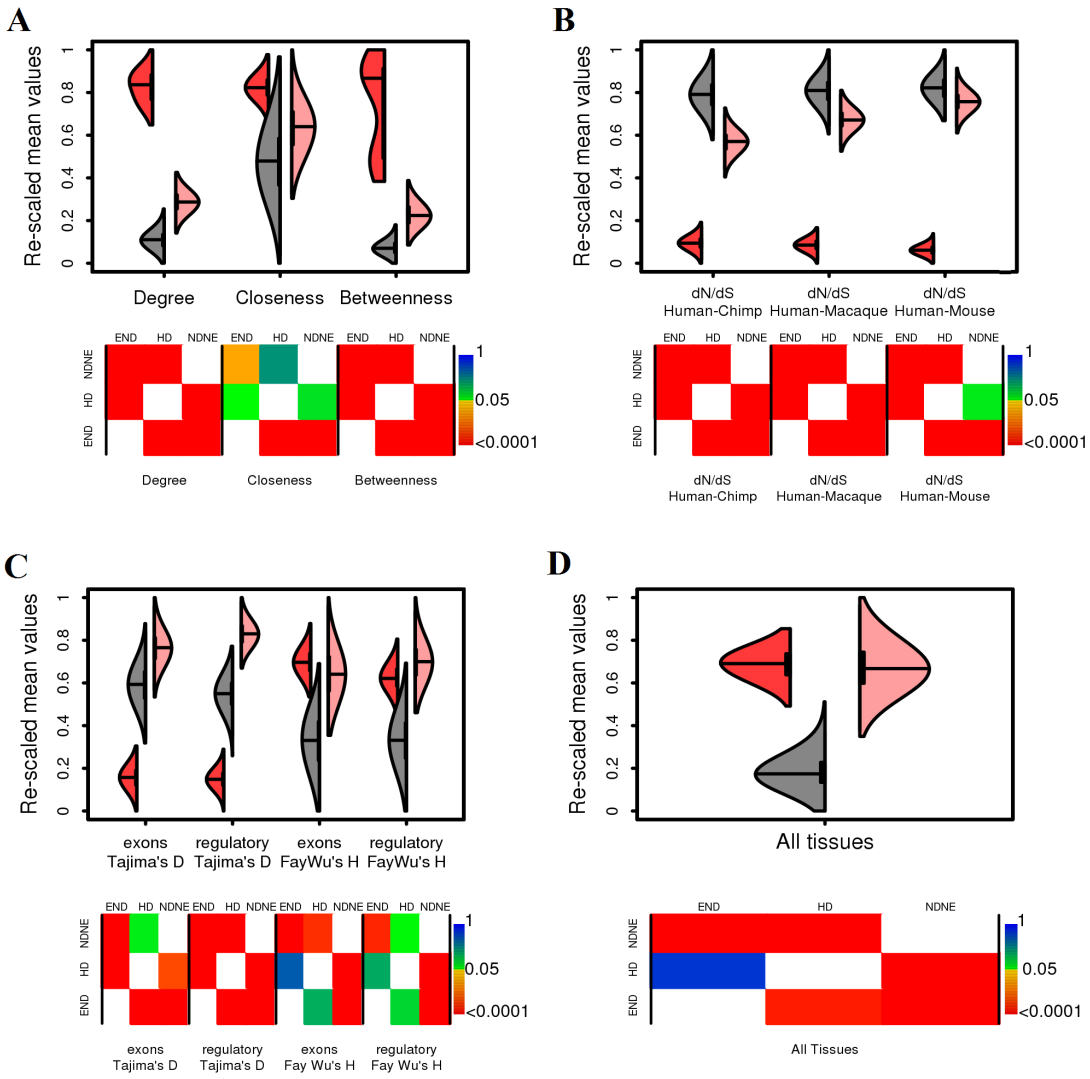

Figure SN2.1.2. Protein network parameters, evolutionary properties and expression levels of human genes matching in gene length. Scaled resampled mean values for protein network parameters (A), dN/dS values (B), Tajima's D and Fay and Wu's H (C) and expression levels (D). END genes are shown in dark red, HD genes in pink, NDNE genes in gray. P-values for protein network parameters, dN/dS values, Tajima's D and Fay and Wu's H and expression levels comparisons are shown below the corresponding parameter distribution figure. Details about p-values calculation and interpretation can be found in the Materials and Methods section of the main text and in Supplementary Note 1.

Similarly, we obtained 1,000 resampling sets of CNM genes matching in size MNC and CM genes. At each resampling 100 CNM genes were selected, 23 of which were matching in size CM genes and 77 were matching in size MNC genes, in order to maintain the proportion of CM and MNC genes within the whole set of Mendelian genes. As shown in Figure SN2.1.3, CM genes continue to be the most relevant in the protein-protein interaction network and exhibit intermediate expression levels, while no significant differences are reported for dN/dS, Tajima's D and Fay and Wu's H among the HD genes subgroups. Again, these results perfectly agree with the results reported without controlling for any potential gene length bias, further demonstrating that gene length does not affect the biological and evolutionary properties of the three groups of HD genes.

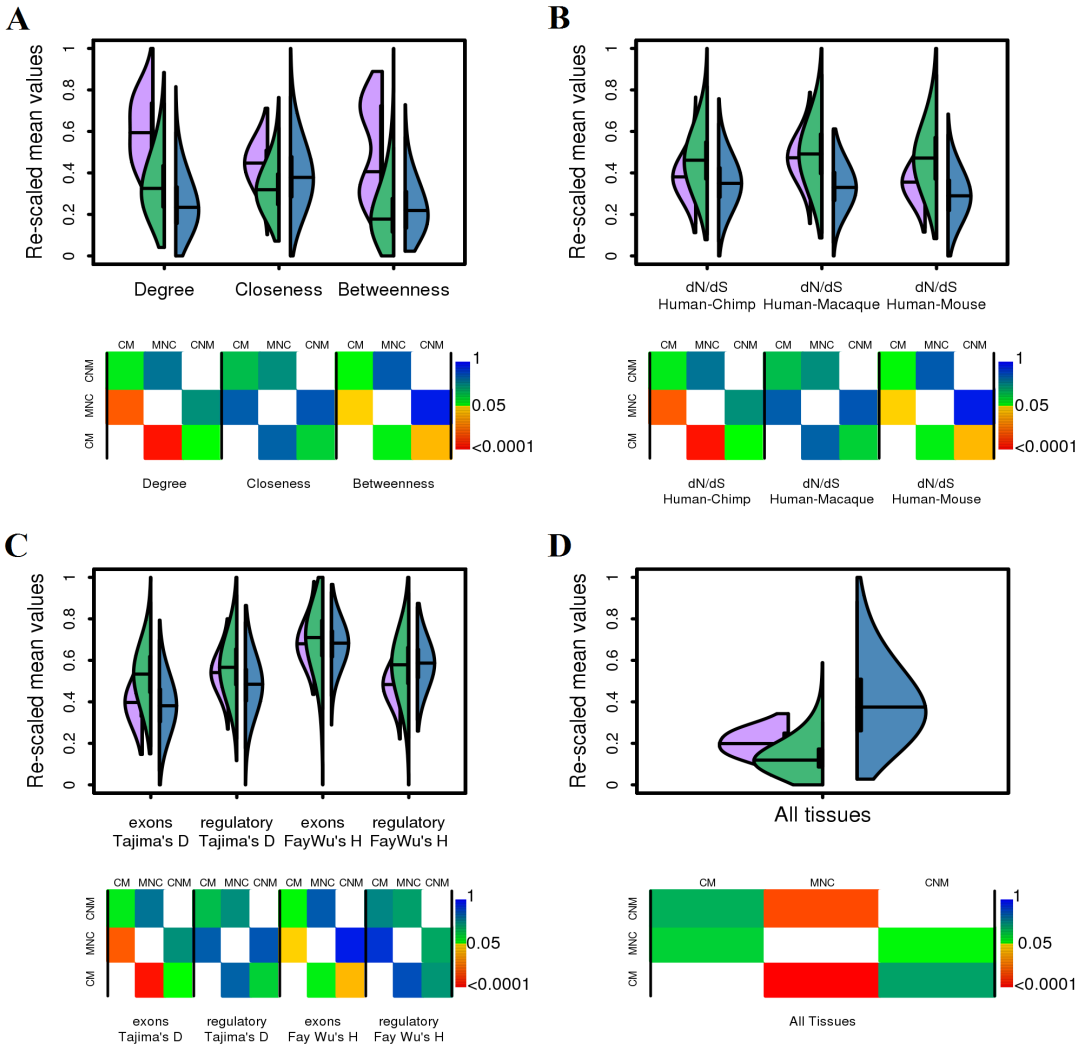

Figure SN2.1.3 Protein network parameters, evolutionary properties and expression levels for the three considered groups of HD genes matching in gene length. Scaled resampled mean values for protein network parameters (A), dN/dS values (B), Tajima's D and Fay and Wu's H (C) and expression levels (D). CM genes are shown in violet, MNC genes in blue, CNM genes in green. P-values for protein network parameters, dN/dS values, Tajima's D and Fay and Wu's H and expression levels comparisons are shown below the corresponding parameter distribution figure. Details about p-values calculation and interpretation can be found in the Materials and Methods section of the main text and in Supplementary Note 1.

2.2. Minor allele frequency and linkage disequilibrium bias

To evaluate whether the MAF and LD properties of SNPs associated to human complex diseases influence the site frequency spectrum of the nearby reported genes, we obtained a set of random SNPs matched for MAF and LD with the SNPs in the GWAS catalog associated to the considered complex diseases in populations of European ancestry. First, for the SNPs of interest stored in the GWAS catalog we calculated the MAF and the extent of the LD blocks in which they are located, using the re-sequencing data of the CEU population from the 1000 Genomes Project. The extent of the LD blocks was computed considering the leftmost and rightmost SNPs with r^2 >0.8 in a window of 500 Kb centered on the SNP of interest. Subsequently, we retrieved all the SNPs in the genome with a MAF in CEU population equal to that of the SNPs of interest in the GWAS catalog. From this subset of SNPs we then excluded all the SNPs in the genome that were found to be in direct LD with the SNPs of interest in the GWAS catalog (r^2 >0.8) or located at less than 250 Kb from any HD or END gene. From the resulting 3,709,900 SNPs, we selected randomly 100,000 SNPs maintaining the proportion of MAF values observed in the SNPs of interest in the GWAS catalog. For instance, if 27% of SNPs of interest in the GWAS catalog have a MAF of 0.15, 27% of the 100,000 random

SNPs will have also a same MAF equal to 0.15. The same procedure used to calculate the extent of the LD blocks for the SNPs of interest in the GWAS catalog was applied to compute the extent of the LD block for each of the 100,000 random SNPs.

Using the 100,000 random SNPs as input, we obtained 1,000 sets of random SNPs matching the extent of the LD block of the SNPs of interest in the GWAS catalog. Only random SNPs with a difference in the extent of the LD block <5 Kb to that of the SNPs of interest were considered when defining the 1,000 random sets of SNPs. Since the average and the standard deviation of the length of the LD blocks are 74 Kb and 85 Kb, respectively, we believe that 5 Kb difference represents a good approximation to define two LD blocks similar in length. For each selected and valid random SNP, we then obtained the NDNE genes overlapping with the specific LD block. In summary, we obtained 1,000 sets of genes located in proximity of random SNPs with similar MAF and similar extent of LD to that of the SNPs of interest in the GWAS catalog.

As suggested by the Tajima's D analysis (Fig. SN2.2.1), for both HD and NDNE genes in direct linkage with a common variant, we observed a site frequency spectrum biased toward common variants, indicating that the evolutionary properties of genes detected through GWAS could be biased by the technology used for their discovery. When considering Fay and Wu's H the results remained qualitatively similar to what was observed without matching for MAF and LD (data not shown). Overall, 1,053 of the genes reported in the GWAS catalog are in direct linkage with the associated variants, representing only 32% and 40% of HD genes and GWAS reported genes, respectively. In their regulatory sequences, the remaining 1,538 genes reported in the GWAS catalog still maintain higher Tajima's D values when compared to the NDNE genes not included among those to be in direct linkage with the 100,000 random common variants used to generate the set of NDNE genes matching in MAF and LD.

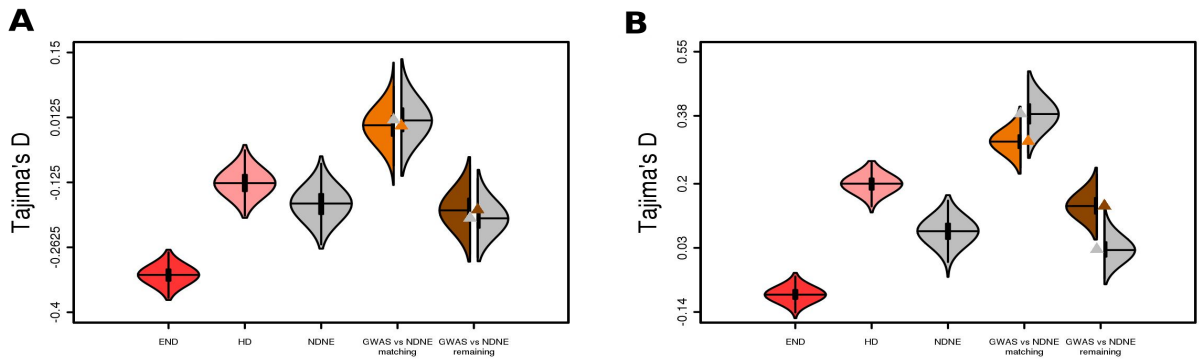

Figure SN2.2.1. Distributions of Tajima's D computed on protein coding (A) and regulatory (B) sequences for END genes (red), for HD genes (pink), for NDNE genes (gray), for GWAS reported genes and NDNE genes (orange and gray, respectively) in direct linkage with common variants matching in MAF and extent of the LD block and for the rest of GWAS reported genes and NDNE genes (dark orange and gray, respectively).

An additional alternative approach was used to test whether the site frequency spectra of the GWAS reported genes depend on the MAF and LD properties of the associated variants. Rather than calculating LD blocks, for each single associated variant we considered the whole genomic region spanning from the leftmost to the rightmost GWAS reported genes. To construct the 1,000 resampling sets of NDNE matching genes, at each resampling a set of random SNPs matching in MAF that of the SNPs of interest in the GWAS catalog was obtained. Only random SNPs that did not contain any HD and END genes in the regions equivalent in size to those of the GWAS associated variants were considered. Finally, for each valid random SNP we retrieved the list of NDNE genes located within the considered region. The slight difference in Tajima's D distribution observed when comparing the protein coding sequences of the whole sets of HD and NDNE completely disappears when focusing on the subset of NDNE and GWAS reported genes located at a similar distance from common variants with equivalent MAF. In contrast, the regulatory elements of genes reported in GWAS catalog still present a site frequency spectrum shifted toward higher Tajima's D values when compared to the matching set of NDNE genes (Fig. SN2.2.2).

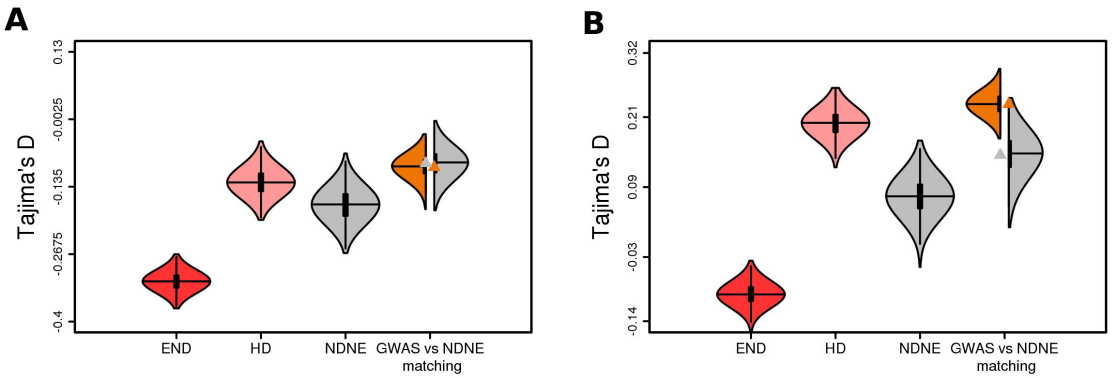

Figure SN2.2.2 Distributions of Tajima's D computed on protein coding (A) and regulatory (B) sequences for END genes (red), HD genes (pink), NDNE genes (gray), GWAS reported genes (orange) and NDNE genes located in regions of equivalent length defined by common variants with equivalent MAF (gray).

Thus, even if only the subset of HD genes in direct linkage with the associated SNPs present a biased site frequency spectrum, it seems clear that evolutionary forces other than purifying selection act on the regulatory sequences HD genes. Indeed, the distribution of Tajima's D computed on the regulatory elements of HD genes is clearly shifted toward positive values, while the distribution of Tajima's D values at protein coding sequences is visibly shifted toward negative values (Fig. SN2.2.3).

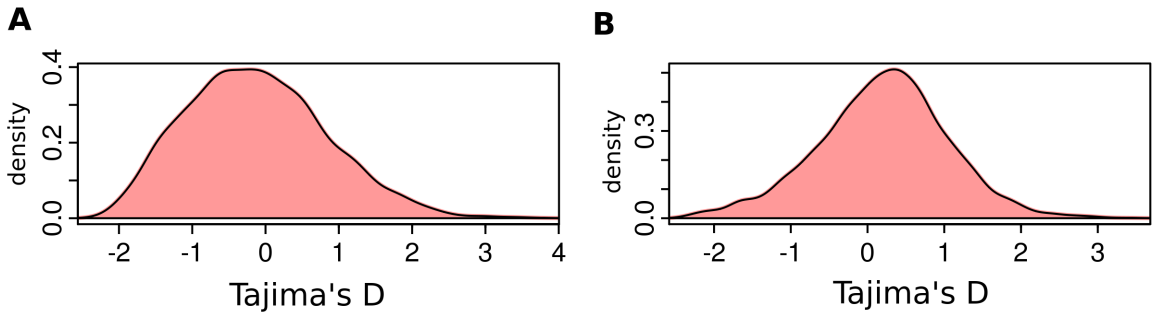

Figure SN2.2.3 Density distributions of Tajima's D values computed on protein coding (A) and regulatory (B) sequences for HD genes.

2.3. Mis-assignment bias on GWAS reported genes

As described above, the genes reported in the GWAS catalog probably suffer from both false positives and false negatives. As an approximation to assess the robustness of our results against such mis-assignment, we evaluated how our conclusions are affected by the presence of genes erroneously indicated as candidates to host the true causal variants. To do so, we devised a test in which we excluded different proportions of genes from the analysis and characterized the behavior of the remaining HD genes. To this end, we randomly removed from 10% to 90% of the genes reported in GWAS, proceeding in 10% steps. For each mis-assignment rate, we obtained 1,000 sets of GWAS-reported genes from which we removed the corresponding proportion. Subsequently, we added to these reduced sets of GWAS genes the whole MNC group, obtaining 1,000 HD sets from which a given proportion of putative false positives reported in GWAS had been removed. Of course, this procedure considers the effect of genes erroneously suggested as candidates (false positives), but cannot not take into account the effect of false negatives since these remain unknown. For the different biological and evolutionary properties under consideration, the distributions obtained for each single mis-assignment rate were then compared to a resampling distribution of HD genes. The resampling distribution of HD genes is very similar to the one extensively used in the whole study and that allowed to detect the significant differences we report in our analyses. As shown in Figure SN2.3.1, the distributions of the parameters analyzed for a mis-assignment rate of up to 40% are a subset of the corresponding HD resampling distribution.

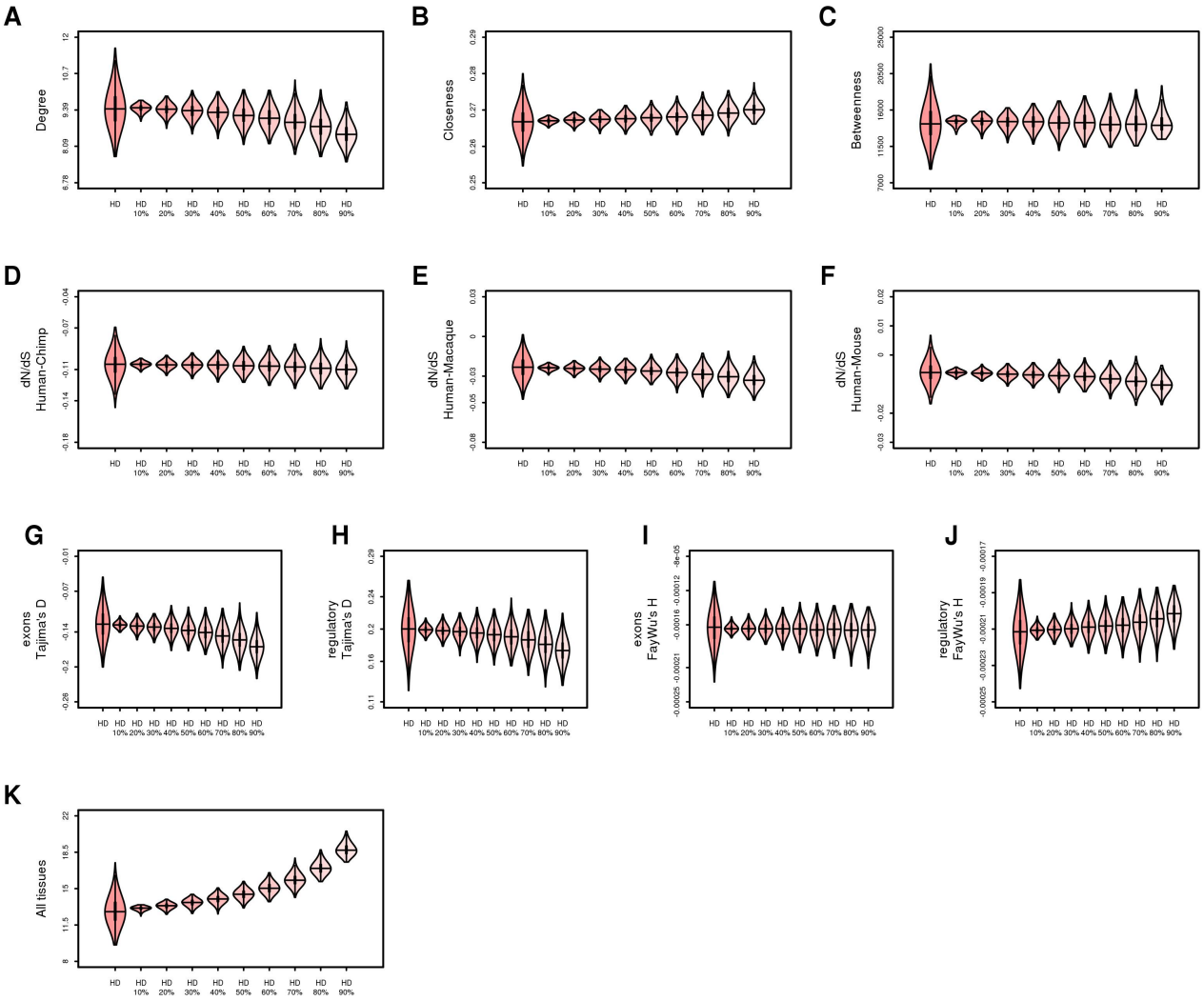

Figure SN2.3.1. Distributions for protein network parameters (from A to C), evolutionary properties (from D to J) and expression levels (K) computed on the whole set of HD genes and for the HD gene sets obtained considering the different mis-assignment rates on the GWAS reported genes.

In order to test specifically how the behavior of GWAS genes (CM+CNM) changes across the different mis-assignment rates, we removed randomly from 10% to 90% of the GWAS reported genes, proceeding in 10% steps. For each mis-assignment rate, we obtained 1,000 sets of GWAS

reported genes from which we removed the corresponding proportion, but this time we did not add MNC in the obtained resampling sets. For the different considered biological and evolutionary properties, the distributions obtained for the used single mis-assignment rates were then compared to a resampling distribution of the full set of GWAS genes. The resampling distribution of GWAS genes is obtained selecting randomly 1,000 GWAS reported genes for each of the 1,000 resampling steps. As reported above for the mis-assignment analysis concerning HD genes, for most of the biological and evolutionary parameters considered the distributions for a mis-assignment rate up to 50% are subsets of the GWAS resampling distribution (Figure SN2.3.2).

Since only one causal gene probably exists per each GWAS locus, for each association of interest in the GWAS catalog, we also selected randomly 1,000 times one single reported gene per GWAS locus. Next we computed the average values over each of these 1,000 random sets for all considered biological and evolutionary properties and compared the distribution of average values to the resampling distribution of the full set of GWAS genes. Out of the 3,544 considered GWAS entries, 2,798 had only one single reported gene, 746 entries presented multiple reported genes, and only 74 entries presented more than four different reported genes. Interestingly, the sets obtained by selecting randomly only one single gene for each GWAS entry showed significantly lower dN/dS values at the three considered time-scales, while at both exonic and regulatory sequences their Tajima's D was significantly higher than that of the resampling distribution of the full set of GWAS genes. No differences in the protein network properties nor in the gene expression levels were observed between the sets obtained selecting randomly one single gene for each GWAS entry and the resampling distribution of the full set of GWAS genes (Figure SN2.3.2). These results indicate that if we assume the existence of a single causal gene per GWAS locus, the observations concerning the comparisons of the evolutionary properties between HD and NDNE genes would be more extreme than those reported in the main text.

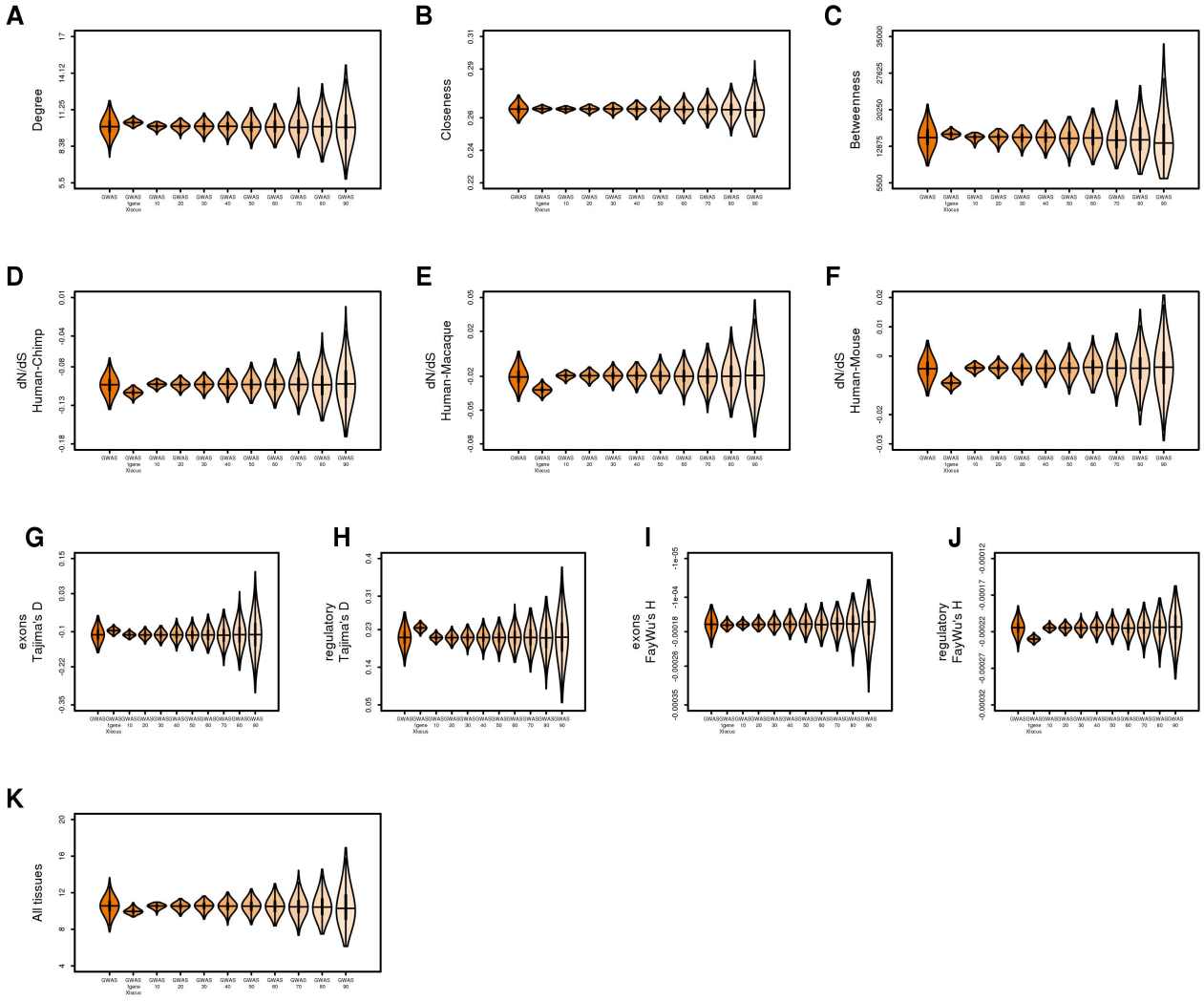

Figure SN2.3.2. Distributions for protein network parameters (from A to C), evolutionary properties (from D to J) and expression levels (K) computed on the resampling set of GWAS reported genes, the random GWAS sets obtained considering only one causal gene per GWAS locus and for the different considered mis-assignment rates on the GWAS reported genes.

Finally, we also controlled how the different mis-assignment rates could specifically affect the CNM genes, assuming that among GWAS genes those contained in CM group are correctly associated to the complex diseases. Again, for the different mis-assignment rates we obtained 1,000 resampling sets removing the corresponding proportion of genes from the CNM set. These mis-assignment distributions were then compared to a resampling CNM distribution that is very similar

to that used through the whole study. As shown in Figure SN2.3.3, all the distributions for the different mis-assignment rates are a subset of the CNM resampling distribution.

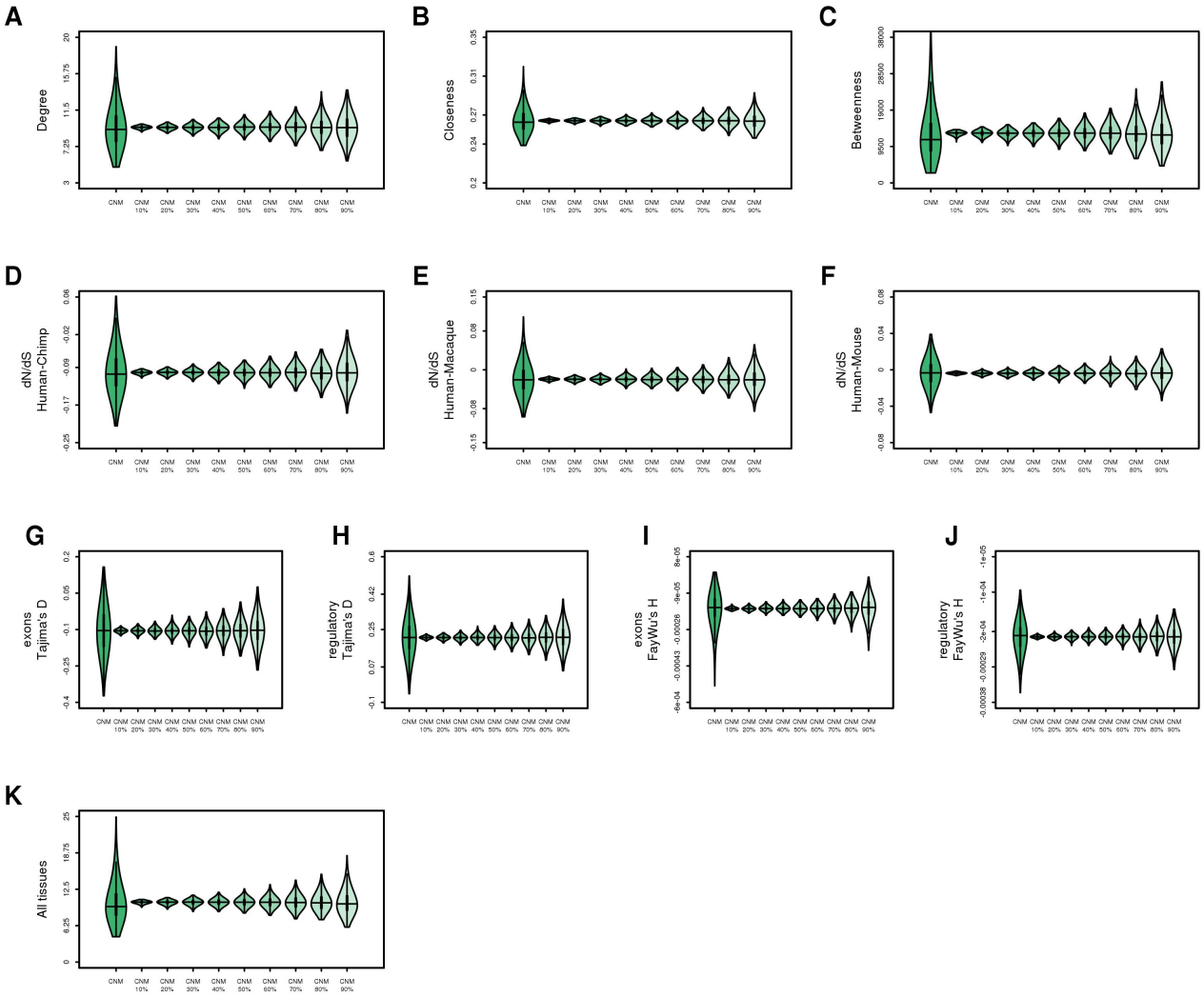

Figure SN2.3.3 Distributions for protein network parameters (from A to C), evolutionary properties (from D to J) and expression levels (K) computed on the whole set of CNM genes and for the different considered mis-assignment rates on the CNM genes.

To conclude, the analysis of the mis-assignment bias on GWAS reported genes provides an approximate measure to the robustness of our results. Based on this analysis, we can say that our

resampling procedures guarantee the validity of our observations even for unrealistically high rates of genes erroneously reported as biological candidates in the GWAS catalog and that our results are robust to the unavoidable noise produced by the inclusion of all the reported genes in the GWAS catalog.

2.4. Effect of the significance threshold on GWAS reported genes

As described in Materials and Methods in the main text, the GWAS catalog compiles all variants reported in GWAS studies with a significance level $\leq 9x10^{-6}$. We next investigated whether considering only SNPs reaching the genome wide significance level (p-value$<5x10^{-8}$) could somehow affect the observed differences among the groups of genes considered. To this end, for all complex diseases of interest we obtained the list of associated SNPs, the p-value of their association and the list of reported genes for each association hit. From the original 3,544 association entries for populations of European ancestry, 1,884 were associated at genome wide significance level. From these genome-wide significance entries we next retrieved a list of 1,488 unique genes , while the original set of complex disease genes (CM+CNM) used in our study comprised up to 2,591 unique genes. Similarly, when considering only genes reported to be associated to SNPs reaching genome wide significance, the number of genes in the remaining datasets changed from 3,275 to 2,287 for HD, from 1,572 to 1,739 for END and from 13,135 to 13,956 for the NDNE group.

For the three new groups of genes obtained we repeated again all the analyses performed regarding the different biological and evolutionary properties considered, using the same statistical strategy described in Materials and Methods and Supplementary Note 1 to test for significance.

As shown in Figure SN2.4.1, all differences previously observed among END, HD and NDNE genes remained significant when considering only those genes related to SNPs associated at genome wide significant level in the GWAS catalog.

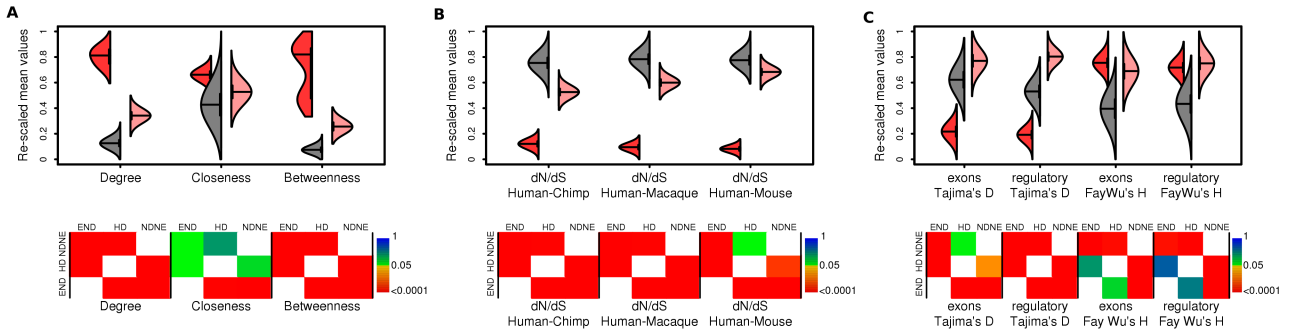

Figure SN2.4.1. Protein network parameters and summary statistics of neutrality for the three groups of human genes considered, obtained when considering only SNPs reaching genome wide significance in the GWAS catalog. Scaled resampled mean values for protein network parameters (A), dN/dS (B), Tajima's D and Fay and Wu's H (C). END genes are shown in dark red, HD genes in light red, NDNE genes in gray. In the panels below each figure, the corresponding p-values for the different pairwise comparisons are represented. Below the diagonal, p-values obtained when resampling the group in the column and comparing it to the mean of whole set in the row. P-values above the diagonal are calculated resampling the group in the row and comparing it to the mean of whole set in the column.

Similarly, the new subsets of END and HD genes still show higher expression levels when compared to NDNE. In contrast, the differences previously described between END and HD disappear, suggesting that genes reported to be associated to SNPs reaching genome wide significance have higher expression levels compared to the remaining genes reported in the GWAS catalog (Figure SN2.4.2A). When exploring the corresponding phylogenetic-based age for the three groups, results remained identical to what observed when considering all the reported genes in the GWAS catalog (Figure SN2.4.2B). When considering only SNPs reaching genome wide significance in the GWAS catalog, the analysis of protein function enrichment reveals that: i) the previous enrichment of END genes compared to HD genes in calcium binding protein category

disappear, ii) HD genes protein products are enriched among structural proteins, receptors and transporters compared to the proteins encoded by END genes, while no differences were reported between the original END and HD groups for these functional categories, and iii) HD genes code for higher amount of structural proteins, cytoskeletal proteins and nucleic acids binding proteins compared to those coded by NDNE, while these differences were not observed in the original groups. Apart from these exceptions, for the protein function enrichment analysis all the previously reported differences among the three groups were also observed for the groups obtained considering only SNPs reaching genome wide significance in the GWAS catalog (Figure SN2.4.2C).

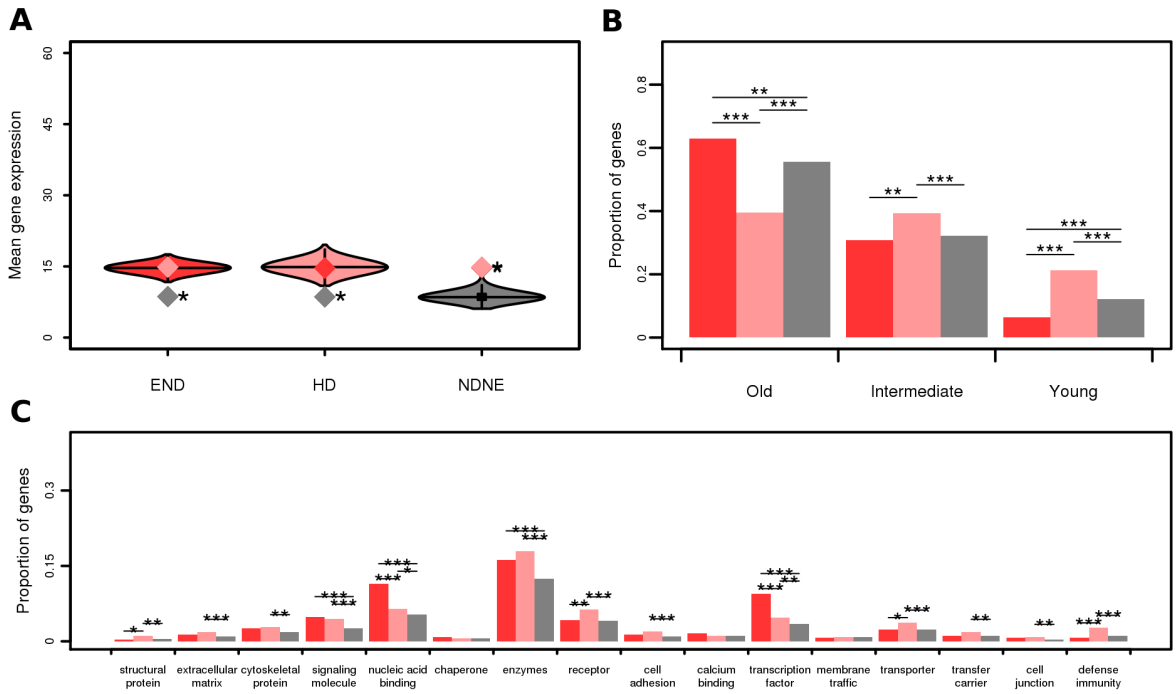

Figure SN2.4.2. Expression levels, gene age and protein functions categories enrichment for the three groups of human genes obtained when considering only SNPs reaching genome wide significance in the GWAS catalog. (A) Resampling expression levels over 16 different human tissues reported in the Expression Atlas. END genes are shown in dark red, HD genes in light red, NDNE genes in gray. At each of the 10,000 resamplings, 1,000 genes were selected and their mean expression levels over 16 tissues was calculated. The mean expression over the 1,000 genes was thus calculated and the 10,000 mean values are represented in the distributions. Diamonds represent

the mean expression values calculated over the whole set of genes of a group. Asterisks indicate that the mean expression is found either on the top or bottom 2.5% of the resampling distribution. (B) Proportions of genes in three age bins. Gene ages were retrieved from PhyloPat database. (C) Proportions of genes in different protein functions considered in PANTHER database. END genes are shown in dark red, HD genes in light red, NDNE genes in gray. *, ** and *** represent that for a given comparison significance is reached at 0.05, 0.005 and 0.0005 levels respectively for a chi-square test (B and C).

The number of genes in the different HD subgroups considered changed from 203 to 126 for CM, from 684 to 799 for MNC and from 2,388 to 1,362 for CNM when considering only genes related to genome wide significantly associated SNPs. As previously reported, CM genes continue to be the most relevant disease genes in the protein-protein interaction network. Remarkably, the statistical support for the CM gene group is stronger when considering only SNPs that reached genome wide significance in the GWAS catalog (Figure SN2.4.3A). No significant differences in the rate of protein evolution were observed for the three subsets of human disease genes considered throughout the study, even if MNC genes tended to show lower dN/dS values compared to CNM genes. Similarly, the MNC group continues to show a similar tendency and only when considering the Human-Mouse dN/dS values MNC genes have significantly lower dN/dS values compared to CNM genes (Figure SN2.4.3B). At intra-species level, none of the comparisons reached significance in the three groups of disease genes considered throughout the study. Similarly no differences were found for the three subsets of disease genes obtained considering only SNPs that reached genome wide significance in the GWAS catalog, even if MNC genes seem to be enriched of rare variants in their regulatory elements compared to CM genes (Figure SN2.4.3C).

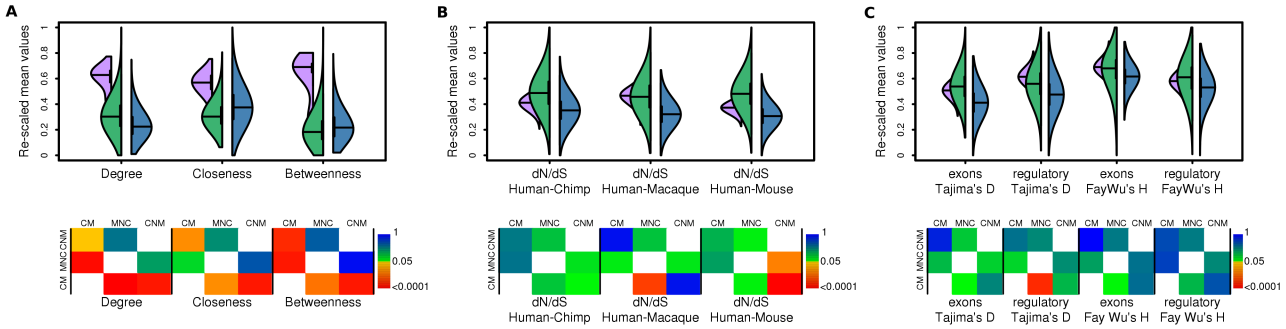

Figure SN2.4.3. Protein network parameters and summary neutrality statistics for the three groups of human disease genes considered, obtained when considering only SNPs reaching genome wide significance in the GWAS catalog. Scaled resampled mean values for protein network parameters (A), dN/dS (B), Tajima's D and Fay and Wu's H (C). CM genes are shown in violet, MNC genes in blue, CNM genes in green. In the panels below each figure, the corresponding p-values for the different pairwise comparisons are represented. Below the diagonal, p-values obtained when resampling the group in the column and comparing it to the mean of whole set in the row. P-values above the diagonal are calculated resampling the group in the row and comparing it to the mean of whole set in the column.

As previously reported, MNC genes are significantly more expressed than CNM genes, and CM genes show intermediate expression levels compared to both MNC and CNM, even if no significance is reached with resampling tests. The only difference with the original analysis is that when comparing the average expression for CNM genes to the resampling distribution of MNC genes significance is not reached (Figure SN2.4.4A); nevertheless significance was maintained when performing a T-test to assess differences in gene expression among the whole MNC and CNM groups (data not shown). When exploring the corresponding phylogenetic-based age for the three groups of human disease genes, most of the results remained identical to what observed when considering all the reported genes in the GWAS catalog. The only exception is found in the MNC

genes that previously were depleted in the “Intermediate” age class compared to CM genes, while for the groups obtained considering only SNPs that reached genome significance in the GWAS catalog this difference disappears (Figure SN2.4.4B). For protein function enrichment analysis, all the previous comparisons remained valid, with the only exception that MNC genes are enriched for cell junction proteins and defence immunity proteins when compared to CNM genes (Figure SN2.4.4C)

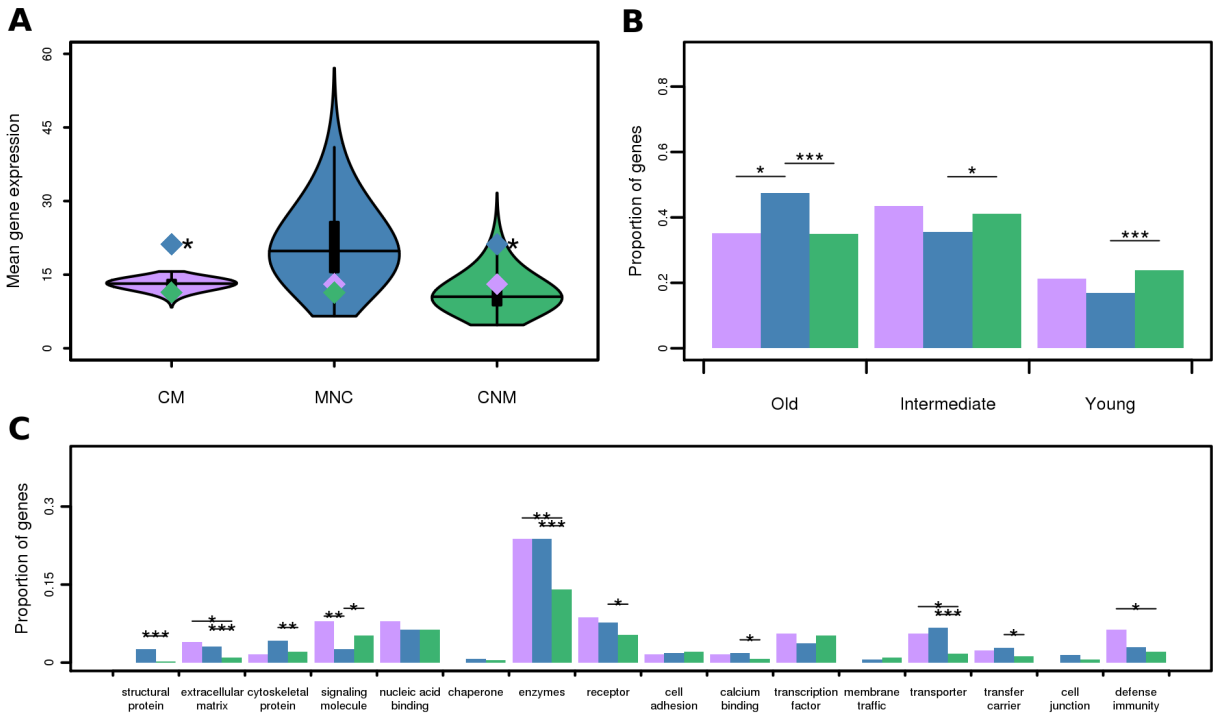

Figure SN2.4.4. Expression levels, gene age and protein functions categories enrichment for the three groups of human genes obtained when considering only SNPs reaching genome wide significance in the GWAS catalog. (A) Resampling expression levels over 16 different human tissues reported in the Expression Atlas. CM genes are shown in violet, MNC genes in blue, CNM genes in green. At each of the 10,000 resamplings, 100 genes were selected and their mean expression levels over 16 tissues was calculated. The mean expression over the 100 genes was thus calculated and the 10,000 mean values are represented in the distributions. Diamonds represent the mean expression values calculated over the whole set of genes of a group. Asterisks indicate that the

mean expression is found either on the top or bottom 2.5% of the resampling distribution. (B) Proportions of genes in three age bins. Gene ages were retrieved from PhyloPat database. (C) Proportions of genes in different protein functions considered in PANTHER database. CM genes are shown in violet, MNC genes in blue, CNM genes in green. *, ** and *** represent that for a given comparison significance is reached at 0.05, 0.005 and 0.0005 levels respectively for a chi-square test (B and C).

Overall, these results indicate that the inclusion of genes related to SNPs that do not reach genome wide significance does not account for the observed differences among the different groups of genes and subsets of human disease genes used throughout the study. Thus, we conclude that our resampling procedure guarantees the validity of our observations and is robust to the possible bias produced by the consideration of genes located nearby loci with borderline associations to the complex diseases of interest.

Supplementary Note 3. Odd-Ratios analysis and comparisons among genes associated to complex diseases.

To assess potential differences between the two subgroups of genes associated to complex diseases (CM and CNM), we compared the allelic Odds-Ratios (ORs) of all associations so far reported in these genes. For each single SNP reported in the GWAS catalog to be significantly associated to the considered human diseases, we retrieved the list of reported genes and the OR of the given SNP. For each single gene was thus possible to obtain a list of OR values, representing the different traits the gene has been found to be associated and the corresponding effect size. In order to compare ORs for CM and CNM genes, we calculated for each single gene the mean and maximum OR (Table S3 in Supplementary Material). Interestingly, we found that CM genes tend to have higher ORs than those genes associated only to complex disorders (Fig. 5A). In particular, when the mean OR for each single gene was considered, the median ORs for each subgroup was of 1.22 and 1.18 for CM and CNM genes, respectively (p-value=3.0×10^{-4} for Mann-Whitney two-sided test); whereas when considering the maximum OR for each single gene, the median ORs within each subgroup were 1.26 and 1.21 for CM and CNM genes, respectively (p-value=1.53×10^{-4} for Mann-Whitney two-sided test).

Note that 99.5% of maximum OR values were lower than 7.95 and 8.36 for CM and CNM genes, respectively; while the maximum OR values were 9.62 and 66.8 for CM and CNM, respectively. These observations suggest that only a few extreme values on CNM genes are probably shifting the whole distribution towards higher values and could explain the small difference in the median OR observed between the two sub-groups of disease genes. Effectively, five different genes (*TCF19*, *POU5F1*, *CCHCR1*, *PSORS1*, *PSORS1C1*) showed the most extreme OR value observed for CNM genes but actually reflected a single association, that of SNP rs2734583 (which is in the vicinity of all five genes) with Stevens-Johnson syndrome and toxic epidermal necrolysis (4). To avoid the

effect of such extreme outliers, we repeated the analysis resampling 10,000 times a subset of 100 CM and CNM genes and reporting for each resampled gene the mean and the maximum OR values. Subsequently, for each resampled set of 100 genes we calculated the mean OR and the 10,000 mean resampling values for CM and CNM were compared (Fig. SN3.1).

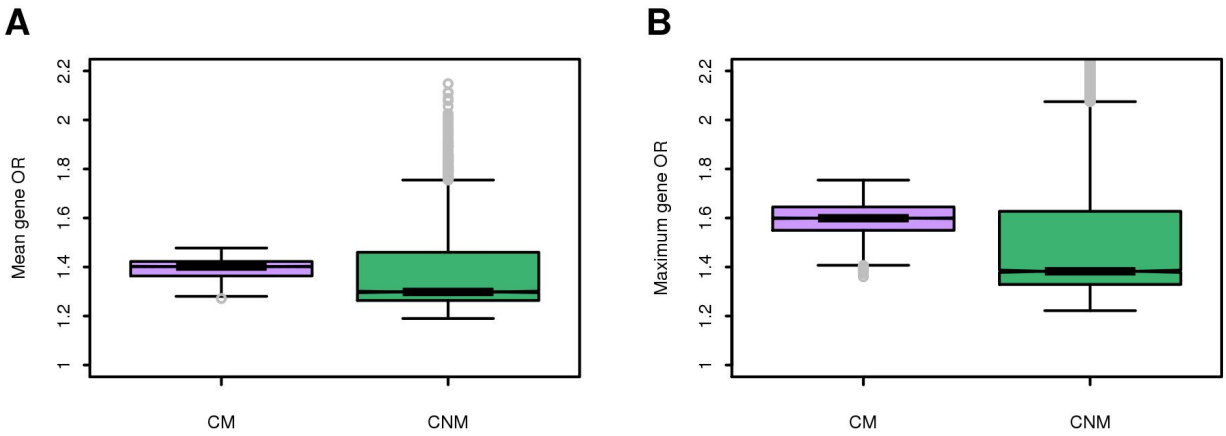

Figure SN3.1. Odd-ratios of resampled genes associated to complex diseases. Distributions of resampling means ORs when considering the mean OR (A) or the maximum OR (B) for each single gene. CM and CNM genes are shown in violet and green, respectively.

When considering the mean OR value for each single gene (Fig. SN3.1A), we observed a median of resampling means of 1.40 and 1.29 for CM and CNM genes, respectively (p-value<$2.2x10^{-16}$ for Mann-Whitney two-sided test). Conversely, when considering the maximum OR value for each single gene (Fig. SN3.1B), we reported a median of resampling means of 1.59 and 1.38 for CM and CNM genes, respectively (p-value<$2.2x10^{-16}$ for Mann-Whitney two-sided test). Thus, genetic variants involved in susceptibility for complex diseases shows a general trend to present higher effect sizes when found on CM genes than if found in genes associated only to complex diseases.

Supplementary Note 4. Human disease genes and essential genes.

In our study, we have shown that human disease (HD) genes present several intermediate biological and evolutionary features comprised between the extremes represented by Essential Non-Disease (END) genes and Non-Disease Non-Essential (NDNE) genes. Furthermore, we suggest a model on which HD genes need to be functionally relevant to be associated to a disease phenotype, but not as much as the END genes. The set of HD genes included all the genes reported to be associated to any human disease in the GWAS catalog (5, 6) or in the hOMIM dataset (7). However, since some HD genes were reported as essential in mice by Georgi et al. (8), we set out to investigate whether the overlap of essential and disease genes (*i.e.* "ED", 822 genes) has specific features compared to the rest of HD genes (*i.e.* "DNE", 2453 genes). We thus repeated the analyses performed in the main text considering END, NDNE and the two subsets of HD genes.

Results regarding protein network parameters, dN/dS values and summary statistics of neutrality are reported in Fig. SN4.1. We used the same rationale and statistical methodology previously described in the main text and in Supplementary Note 1 with the exception that we used a fixed number of 500 genes in all the resampling procedures. Comparisons of gene expression, gene age and protein function categories among the four groups of genes are shown in Fig. SN4.2. As expected, END genes were found to be the most relevant genes in the genome, showing the highest protein network relevance, the lowest rate of protein evolution and the lowest Tajima's D when considering intra-species variation data (Fig. SN4.1). Additionally, END genes are the most expressed (Fig. SN4.2A) and the oldest genes in the human genome (Fig. SN4.2B). As described in the main text, the whole set of HD genes have intermediate functional relevance in the protein network and dN/dS values comprised between the extremes of END and NDNE genes (Fig. SN4.1 A and B). Among HD genes, ED genes have higher functional importance in the protein network and lower rates of protein coding evolution than DNE (Fig. SN4.1 A and B). Additionally, ED genes

also display higher gene expression and tend to be slightly older than DNE genes (Fig. SN4.2 A and B). Similarly to the rest of essential genes (*i.e.* the END gene group), ED genes encode for protein products directly interacting with the genetic material (Fig. SN4.2 C). By contrast, DNE genes have very similar protein network features and rates of protein evolution to NDNE genes (Fig. SN4.1 A and B). Overall these results, suggest that there is a continuous gradient of evolutionary constraint and biological relevance with the two subgroups of HD genes comprised between the extremes defined by the END and NDNE subsets.

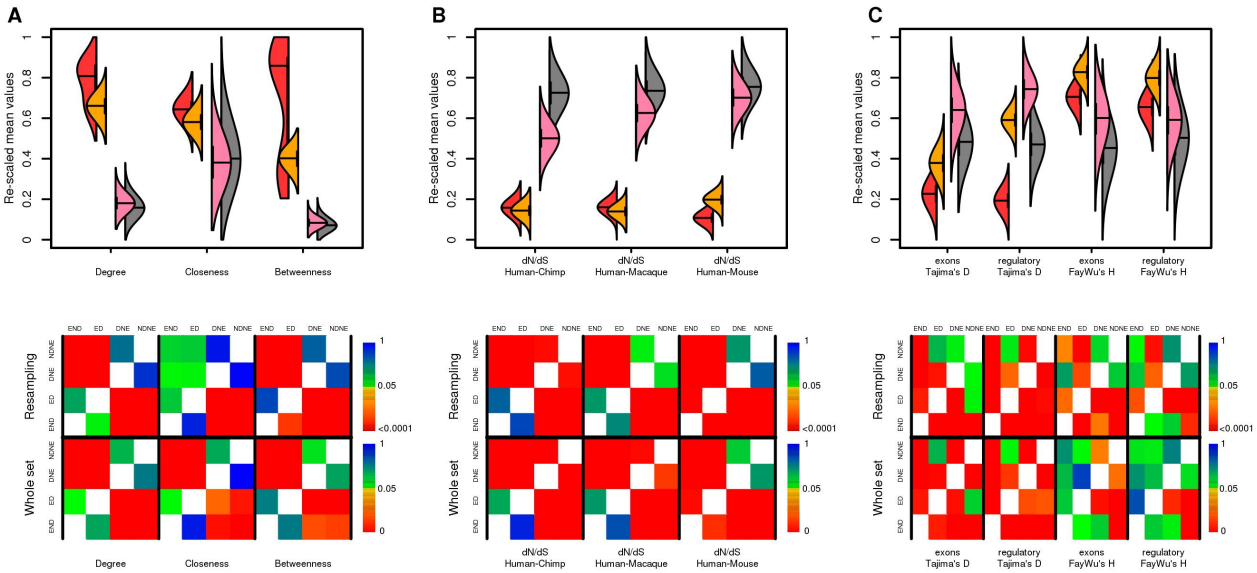

Figure SN4.1. Protein network parameters and evolutionary properties of human genes. Scaled resampled mean values for protein network parameters (A), dN/dS values (B), and Tajima's D and Fay and Wu's H (C). END genes are shown in dark red, ED genes in orange, DNE genes in pink, NDNE genes in gray. P-values for protein network parameters, dN/dS values, and Tajima's D and Fay and Wu's H comparisons are shown below the corresponding parameter distribution figure. Details about p-values calculation and interpretation can be found in the Materials and Methods section of the main text and in Supplementary Note 1.

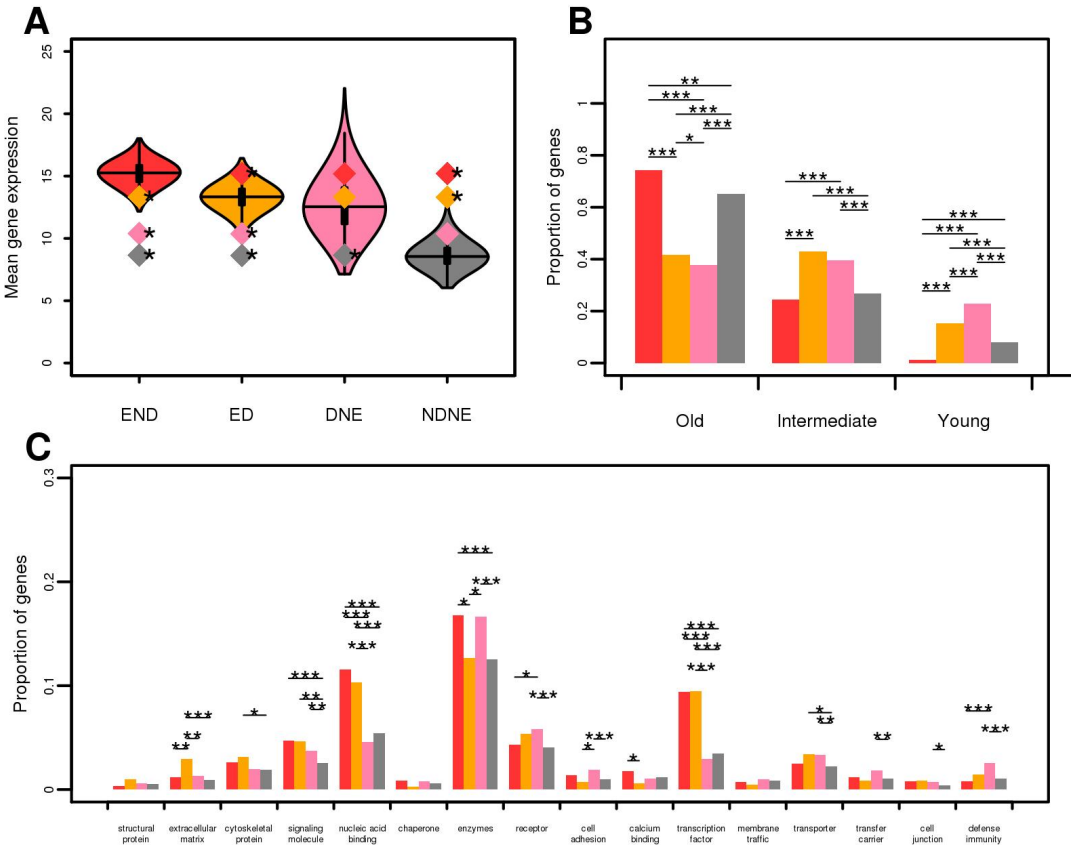

Figure SN4.2. Biological features of different groups of human genes. (A) Resampling expression levels over 16 different human tissues reported in the Expression Atlas (9,10). END genes are shown in red, ED genes in orange, DNE genes in pink, NDNE genes in gray. At each of the 10,000 resamplings, 500 genes were selected and their mean expression levels over 16 tissues was calculated. The mean expression over the 500 genes was thus calculated and the 10,000 means values are represented in the distributions. Diamonds represent the mean expression values calculated over the whole set of genes of a group. Asterisks indicate that the mean expression is found either on the top or bottom 2.5% of the resampling distribution. (B) Proportions of genes in three age bins. Gene ages were retrieved from PhyloPat database (11). (C) Proportions of genes in different protein functions considered in PANTHER database (12). *, **, *** represent that for a given comparison significance is reached at 0.05, 0.005 and 0.0005 levels, respectively, for a chi-square test (B and C).

In spite of the different network properties and protein evolution rates observed between ED and DNE, the two subsets have been considered together in the HD gene group in all the analyses in the main text, because both contain genes linked and/or associated to different human diseases. Indeed, 364 over 822 ED genes were detected by linkage studies to harbor genetic variants causing Mendelian disorders, representing ~41% of the total Mendelian genes considered in our study; while the remaining 458 ED genes only represent 17.7% of the total genes associated to complex disease. In particular, the two HD groups that include genes linked to Mendelian disorders (CM and MNC) present higher proportions of putative essential genes than the CNM group (Table SN4.1). Overall, these observations indicate that genes linked to Mendelian diseases are enriched in ED genes. In turn, the higher proportion of ED genes observed in the CM group when compared to MNC agrees with the distinctive protein network features of CM genes.

	MNC (267/684) - 39.04%	**CNM** (458/2388) - 19.18%
CM (97/203) - 47.78%	1.11×10^{-3}	3.47×10^{-21}
MNC (267/684) - 39.04%	*	1.12×10^{-100}

Table SN4.1. Enrichment of essential genes among different groups of HD genes. In brackets are indicated the number of genes overlapping with the original set of essential genes proposed by Georgi et al. (8) and the total number of genes of each subgroup of HD genes. P-values for chi-square test are indicated for each pair.

Interestingly, both subgroups of HD genes (*i.e.* DNE and ED) show higher amounts of intermediate frequency variants at their regulatory elements, when compared to NDNE genes (Fig. SN4.1C). This observation indicates that the whole set of HD genes is enriched of intermediate frequency variants and that intra-species adaptive events could also target genes that were reported as essential in mice and that show high functional relevance in humans.

Supplementary Note 5. Human disease and Mendelian genes versus other gene groups.

In our study, we have shown that HD genes present several intermediate biological and evolutionary features comprised between the extremes represented by END genes and NDNE genes. The set of HD genes included all the genes reported to be associated to any human disease in the GWAS catalog (5, 6) or in the hOMIM dataset (7). The genes included in the hOMIM dataset represent a subset of disease genes that have been detected through linkage studies to be related to simple monogenic disorders. Their involvement in Mendelian diseases has been functionally proved and they represent the most reliable subset of HD genes. By contrast, genes reported in the GWAS catalog could suffer from several biases. Even though we show that these confounding factors do not seem to account for the differences observed between HD genes and the rest of the genome (see Supplementary Note 2) we also tested whether our observations regarding the full set of HD genes could be biased by the inclusion of genes reported in the GWAS catalog. To this end, we repeated all the analyses performed in the main text considering the full set of Mendelian genes (CM+MNC) together with the previously analyzed groups (END, HD, and NDNE).

For most protein network properties, dN/dS values and neutrality summary statistics, most of the differences found when comparing HD genes with END and NDNE genes are also observed for the full set of Mendelian genes. Only for closeness, Mendelian genes behave more similarly to END genes, while HD genes show a distribution more similar to that of NDNE genes. In addition, Mendelian genes show a slight enrichment of rare variants at their exonic sequences compared to HD genes, as suggested by Tajima's D. Nevertheless, both HD and Mendelian genes have: i) similar Tajima's D distributions to that of NDNE at their exonic sequences and ii) higher Tajima's D in their regulatory elements compared to NDNE genes (Figure SN5.1). These results highlight that the observations concerning the full set of HD genes do not seem to be biased by the presence of GWAS genes and that the enrichment of intermediate frequency variants at their regulatory sequences is a pattern shared among different subsets of human disease genes.

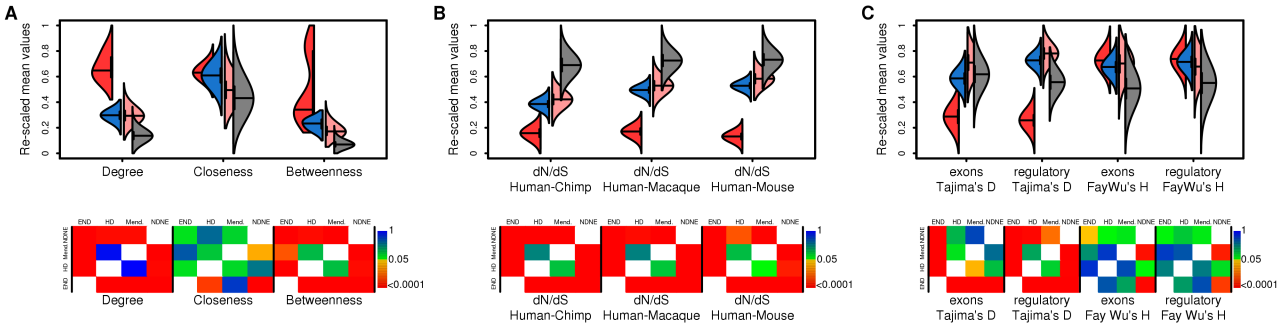

Figure SN5.1. Scaled resampled mean values for protein network parameters (A), dN/dS (B), Tajima's D and Fay and Wu's H (C). END genes are shown in dark red, HD genes in light red, Mendelian genes in blue, NDNE genes in gray. At each of the 10,000 resamplings, 500 genes were selected for each group and the average values were calculated for the different biological and evolutionary properties in order to obtain the resampling distributions of 10,000 average values. In the panels below each figure, the corresponding p-values for the different pairwise comparisons are represented. Below the diagonal, p-values obtained when resampling the group in the column and comparing it to the mean of whole set in the row. P-values above the diagonal are calculated resampling the group in the row and comparing it to the mean of whole set in the column.

Interestingly, Mendelian genes are the most expressed genes in the genome, while HD genes have expression levels that are intermediate between those of END and NDNE genes (Figure SN5.2A). By contrast, the gene age of Mendelian genes is in perfect agreement with the full set of HD genes; indeed both Mendelian and HD genes are depleted in the “Old” age category and are enriched in the “Intermediate” and “Young” categories compared to both END and NDNE (Figure SN5.2B). The analysis of the protein functional categories reveals that the proteins encoded by both HD and Mendelian genes are enriched in eight different functional categories compared to NDNE (extracellular matrix components, enzymes, receptors, cell adhesion proteins, transporters, transfer carriers, cell junction proteins and defense immunity related proteins). Compared to NDNE, protein

products encoded by Mendelian genes are also enriched among structural and cytoskeletal proteins encoded proteins, while no differences are reported for these categories when comparing the protein products of HD and NDNE genes. Similarly, compared to NDNE genes the protein products encoded by HD genes are enriched among signaling molecules and transcription factors, while no differences are reported in these categories for the proteins encoded by Mendelian and NDNE genes. Finally, the proteins encoded by Mendelian genes are enriched in five different functional categories when compared to the full set of HD genes (structural proteins, extracellular matrix components, enzymes, transporters and transfer carriers) (Figure SN5.2C).

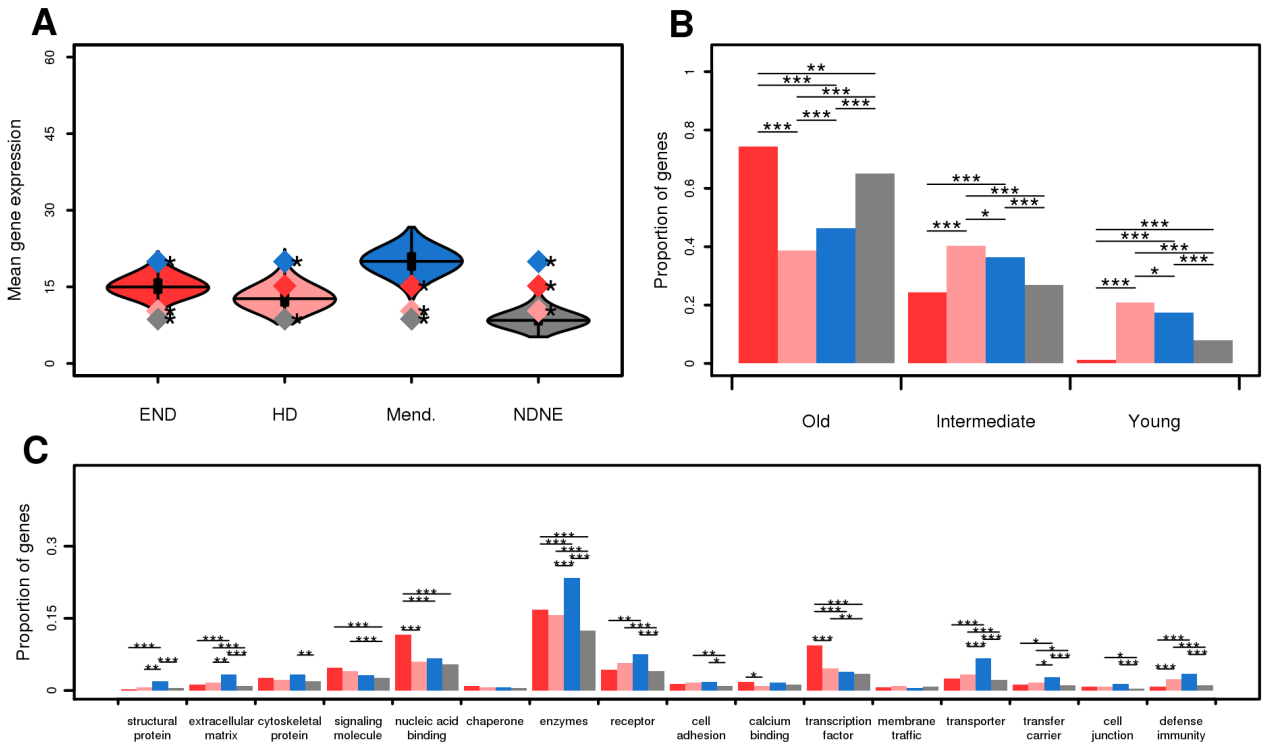

Figure SN5.2. Biological features of different groups of human genes. (A) Resampling expression levels over 16 different human tissues reported in the Expression Atlas (9,10). END genes are shown in dark red, HD genes in light red, Mendelian genes in blue, NDNE genes in gray. At each of the 10,000 resamplings, 500 genes were selected and their mean expression levels over 16 tissues was calculated. The mean expression over the 500 genes was thus calculated and the 10,000 means values are represented in the distributions. Diamonds represent the mean expression values calculated over the whole set of genes of a group. Asterisks indicate that the mean expression is

found either on the top or bottom 2.5% of the resampling distribution. (B) Proportions of genes in three age bins. Gene ages were retrieved from the PhyloPat database (11). (C) Proportions of genes in different protein functions considered in PANTHER database (12). *, **, *** represent that for a given comparison significance is reached at 0.05, 0.005 and 0.0005 levels, respectively, for a chi-square test (B and C).

Overall these results suggest that Mendelian genes behave similarly to the full set of HD genes and that most of the evolutionary and biological properties reported in our study for human disease genes should not be biased by the inclusion of GWAS genes. Moreover, as already expected given the patterns observed when comparing the three HD gene subgroups (CM, MNC and CNM), the only differential behaviours between Mendelian and HD genes are found at the level of their expression patterns and for the protein functions they encoded.

REFERENCES

1. Blair, D.R., Lyttle, C.S., Mortensen, J.M., Bearden, C.F., Jensen, A.B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam, E. V, Brunak, S., et al. (2013) A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. *Cell*, **155,** 70–80.
2. Hong, E.P. and Park, J.W. (2012) Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genomics Inf.*, **10**, 117–122.
3. Lachance, J. and Tishkoff, S.A. (2013) SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays*, **35**, 780–786.
4. Tohkin, M., Kaniwa, N., Saito, Y., Sugiyama, E., Kurose, K., Nishikawa, J., Hasegawa, R., Aihara, M., Matsunaga, K., Abe, M., *et al.* (2011) A whole-genome association study of major determinants for allopurinol-related Stevens-Johnson syndrome and toxic epidermal necrolysis in Japanese patients. *Pharmacogenomics J.*, **13**, 60–69.
5. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33,** 514–517.
6. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.*, **33,** 9362-9367.
7. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42,** 1001–1006.
8. Georgi, B., Voight, B.F. and Buc, M. (2013) From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. *PLoS Genet.*, **9,** e1003484.
9. Petryszak, R., Burdett, T., Fiorelli, B., Fonseca, N. A., Gonzalez-Porta, M., Hastings, E., Huber,

W., Jupp, S., Keays, M., Kryvych, N., *et al.* (2014) Expression Atlas update - A database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **42**, 926–932.

10. Fonseca, N. A., Marioni, J. and Brazma, A. (2014) RNA-Seq Gene Profiling - A Systematic Empirical Comparison. *PLoS One*, **9**, e107026.
11. Hulsen, T., Groenen, P.M.A., de Vlieg, J. and Alkema, W. (2009) PhyloPat: An updated version of the phylogenetic pattern database contains gene neighborhood. *Nucleic Acids Res.*, **37,** 731–737.
12. Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S. and Thomas, P.D. (2009) PANTHER version 7: Improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.*, **38,** 204–210.