Supplementary Material

The interplay of mutations and electronic properties in disease-related genes

Chi-Tin Shih[1,*], Stephen A. Wells[2], Ching-Ling Hsu[3], Yun-Yin Cheng[1], & Rudolf A. Römer[2,*]

[1]*Department of Physics, Tunghai University, 40704 Taichung, Taiwan and The National Center for Theoretical Sciences, 30013 Hsinchu, Taiwan*

[2]*Department of Physics and Centre for Scientific Computing, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK*

[3]*Department of Physics, Chung-Yuan Christian University, 32023 Chung-Li, Taiwan*

[*]*Correspondence and requests for materials should be addressed to CTS (email: ctshih@thu.edu.tw) or RAR (email: r.roemer@warwick.ac.uk).*

Comparing the Averaged Electronic Properties for the Pathogenic and Non-pathogenic Mutations for Each Gene

We denote the genomic sequence of a gene with length $\mathcal{N}$ base pairs (bps) as $(s_1, s_2, \cdots, s_{\mathcal{N}})$. Each point mutation of a given gene is characterized by the set (k, s), where k and s are the position of the point mutation in the genomic sequence and the mutant nucleotide which replaces the nucleotide s_k of normal DNA, respectively. There are totally $3\mathcal{N}$ possible point mutations of a gene with $\mathcal{N}$ bps. The sets of these $3\mathcal{N}$ mutations and the pathogenic mutations for the gene are denoted as $M_{\rm all}$ and $M_{\rm pa}$, respectively. $M_{\rm pa}$ is a subset of $M_{\rm all}$. For every possible point mutation, we compute the *mean* quantum mechanical transmission coefficient $T_L^{(k)}$ of a subsequence with length L of the *wild-type* gene. Here the mean is determined by averaging over all individual transmission coefficients $T_{j,L}$ with $j = k - L + 1, k - L + 2, \ldots, k$. In this way, the influence of the full neighborhood of hotspot k is taken into account and not just the mutation itself. The results of $T_L^{(k)}$ for $k \in M_{\text{pa}}$ already show some signatures of atypical CT reponse for the 1D model.[S41] However, the signal is much less pronounced in the 2-leg model. Hence we study the *difference* in CT between a healthy DNA base and the 3 possible mutations. For example the hotspot 14585 of $p53$ contains the correct C/G base pair in the wild but of the three possible mutations $C/G \to G/C$, $C/G \to A/T$ and $C/G \to T/A$ only the last one is know to lead to cancer.[5] Averaging again over all incident energies and subsequences of length L containing the hotspot (k, s), we can characterize the *average change* in CT as

$$\Gamma_{L,q}^{(k,s)} = \frac{1}{L} \sum_{j=k-L+1}^{k} \int_{E_0}^{E_1} \frac{|T_{j,L}(E) - T_{j,L}^{(k,s)}(E)|^q}{E_1 - E_0} dE \quad . \tag{3}$$

with $q = 1$ or 2. We find that results for $q = 1$ and 2 are similar. Hence in the manuscript we restrict our discussion to $q = 2$. We calculate such Γ estimates for all possible $3\mathcal{N}$ mutations of each gene and compare the probability distribution of CT change $\Gamma_{L,q}^{(k,s)}$ for $(k, s) \in M_{\text{all}}$ and $(k, s) \in M_{\text{pa}}$ for each gene. The result for the $p16$ gene was shown in Fig. (a) as an example. As a control group, we also shuffled the $p16$ sequence randomly under the conditions that (1) the contents of the 4 bases are not changed, and (2) the positions of the mutations can be moved but the numbers of the 12 types of mutations are not changed. The distributions of the averaged Γ for 1D and 2-leg models with $L = 40$ of the 20 shuffled sequences are shown in Fig. S1. It is clear that the distributions of Γ for the M_{all} and M_{pa} are almost identical.

CT Change for the 12 Type of Mutations

The comparison of Γ between the pathogenic and all possible mutations for the 12 types of point mutations is shown in Fig. S2. It is clear for the 1D model (a–l) Γ tends to be smaller for the pathogenic mutations. However, the difference is not visible for the 2-leg model (m–x).

Local ranking of point mutations at hotspot sites

In order to study the local effects of pathogenic mutations on CT, we compare $\Gamma_{L,2}^{(k,s)}$ of each pathogenic mutation (k, s) with the other two non-pathogenic ones at the same position k and determine the *local ranking* (LR) of CT change for (k, s). There are three possibilities of LR, namely *low*, *medium* and *high*. Note that those hotspots k with more than one pathogenic mutations are excluded in the LR analysis. As an example, percentages of the three LR for the pathogenic mutations of $p16$ are shown in the top panels of Fig. S3. The rankings of pathogenic mutations with low CT change are evidently larger than the medium and high ones for most L. A similar tendency is observed for CYP21A2. Let us again ask how significant this tendency is across all 162 genes. Figure S4 shows similar ranking analysis results as in Fig. S3 but now for *all* M_{pa}. We see that the tendency towards low CT change in the pathogenic mutations is quite strong overall. In Fig. we have sorted the LR ranking for each gene according to prevalence. We find that for $L = 20$, 40 and 60 the low CT change corresponds to 155 (95%), 148 (91%) and 140 (86%) of all 162 genes with pathogenic mutations. Note that similarly consistent is the result for large CT with only about 30 of all genes having high CT change.

Global CT rankings at hotspot sites

Another way to compare the CT change is a *global* ranking (GR). We have sorted the CT change $\Gamma_{L,2}^{(k,s)}$ for *all* possible $3\mathcal{N}$ mutations of a gene with $\mathcal{N}$ bps in order to get a ranking of *every* pathogenic mutation (k, s). By dividing each ranking by $3\mathcal{N}$ we compute the normalised GR $\gamma_{L,2}^{(k,s)}$ of the mutation with values between 0 and 1. As before for $\Gamma_{L,q}^{(k,s)}$, smaller values of $\gamma_{L,q}^{(k,s)}$ mean smaller CT change. To characterise the CT change in a quantitative way, we divide the $\gamma_{L,2}^{(k,s)}$ of the pathogenic mutations into again three groups as before, i.e. low ($\gamma < 33.3\%$), medium ($33.3\% \leq \gamma < 66.7\%$), and high ($\gamma \geq 66.7\%$) CT change. The distributions of the GR for the complete set of pathogenic mutations of $p16$ and CYP21A2 is shown in Fig. S3 as an example. As for the LR results, the pathogenic genes lead to many $\gamma_{L,2}^{(k,s)}$ values with low CT change. This is most pronounced in the 1D model as shown in Fig. S3(c). The results of the GR for the 162 genes are shown in the bottom row (c) and (d) of Figs. S4 and . We see that the GR results are fully consistent with the LR rankings.

Consistency of CT rankings for all DNA sequences

The prevalence ordering as shown in Fig. does not imply that the order of the genes themselves is the same in all parts (a), (b), (c) and (d) of the figure. Therefore we have calculated the correlations in the ordering and found that in both models and across models and for all $L = 20$, 40 and 60, we find positive correlation coefficients. Hence genes which have a low change in CT for, e.g., the local ranking at $L = 20$, also retain this low rank for the other L values as well as the global ranking. Similarly, this positive correlations implies that in those few case where the mutations in a gene lead to high CT change, they do so across all local as well as global rankings. This confirms that our results are internally consistent.

We graphically summarise the results for all 162 disease-related genes in Fig. S5. For each gene, we have shown a positive deviation from the 0.33 line by orange —supporting the scenario of small CT change for pathogenic mutations — and by blue when the results seem to show no or negative indication with CT change. The criteria corresponds to local and global ranking results for $L = 20$, 40 and 60 for the 1D and the 2-leg models. Similarly, in Fig. , we average of all 12 criteria and show the resulting, overall agreement with the CT hypothesis: 161 of 162 genes are above the 33% line and hence show that for both 1D and 2-leg model and averaged over lengths 20, 40 and 60, a small CT change correlates with the existence and position of pathogenic mutations. Only for STK11 do we see that there is no overall agreement.

Difference and similarities in the two models

The 2-leg model[16] allows inter-strand coupling between the purine bases in successive base pairs, in accordance with electronic structure calculations,[39] and should therefore be a better model for bulk charge transport along the DNA double helix; the 1D model, by contrast, makes use of the site energies of only the bases on the coding strand,[15] and so is most representative of the electronic environment along that strand. We also find that the 2-leg model recovers some of the coding strand dependence of the 1D model upon decreasing the diagonal hoppings. For 28 genes, we find that reducing only the diagonal hopping elements by $1/2$ leads to a much greater agreement with the 1D results similar to Fig. (c).

S41. Shih, C. T. Characteristic length scale of electric transport properties of genomes. *Phys. Rev. E* **74**, 010903(R) (2006).

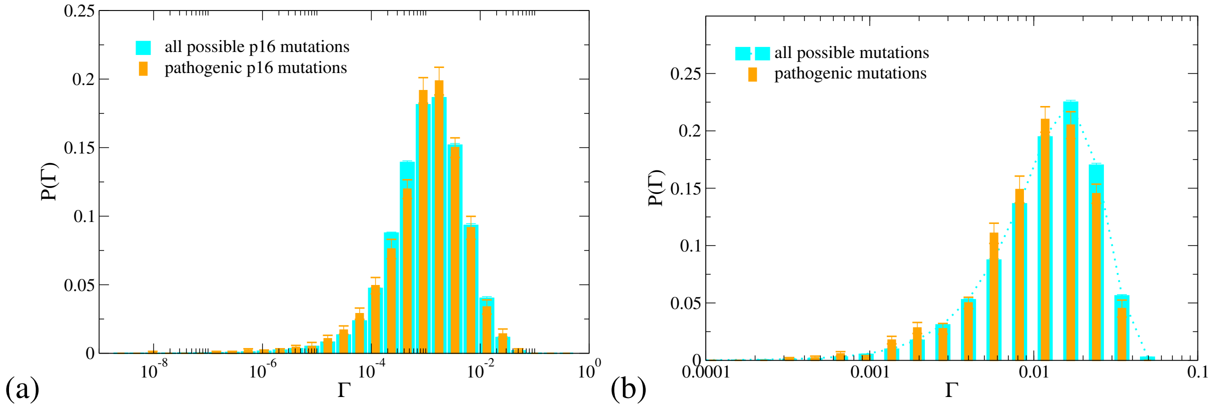

Figure S1: (Supplementary) Distribution of the change in charge transport in (a) 1D and (b) 2-leg models Γ for pathogenic (orange bars) and all possible (cyan bars) mutations averaged for the 20 shuffled $p16$ (CDKN2A) DNA strands with 26740 base pairs. All results shown are for $L = 40$, data for $L = 20$ and 60 are similar.

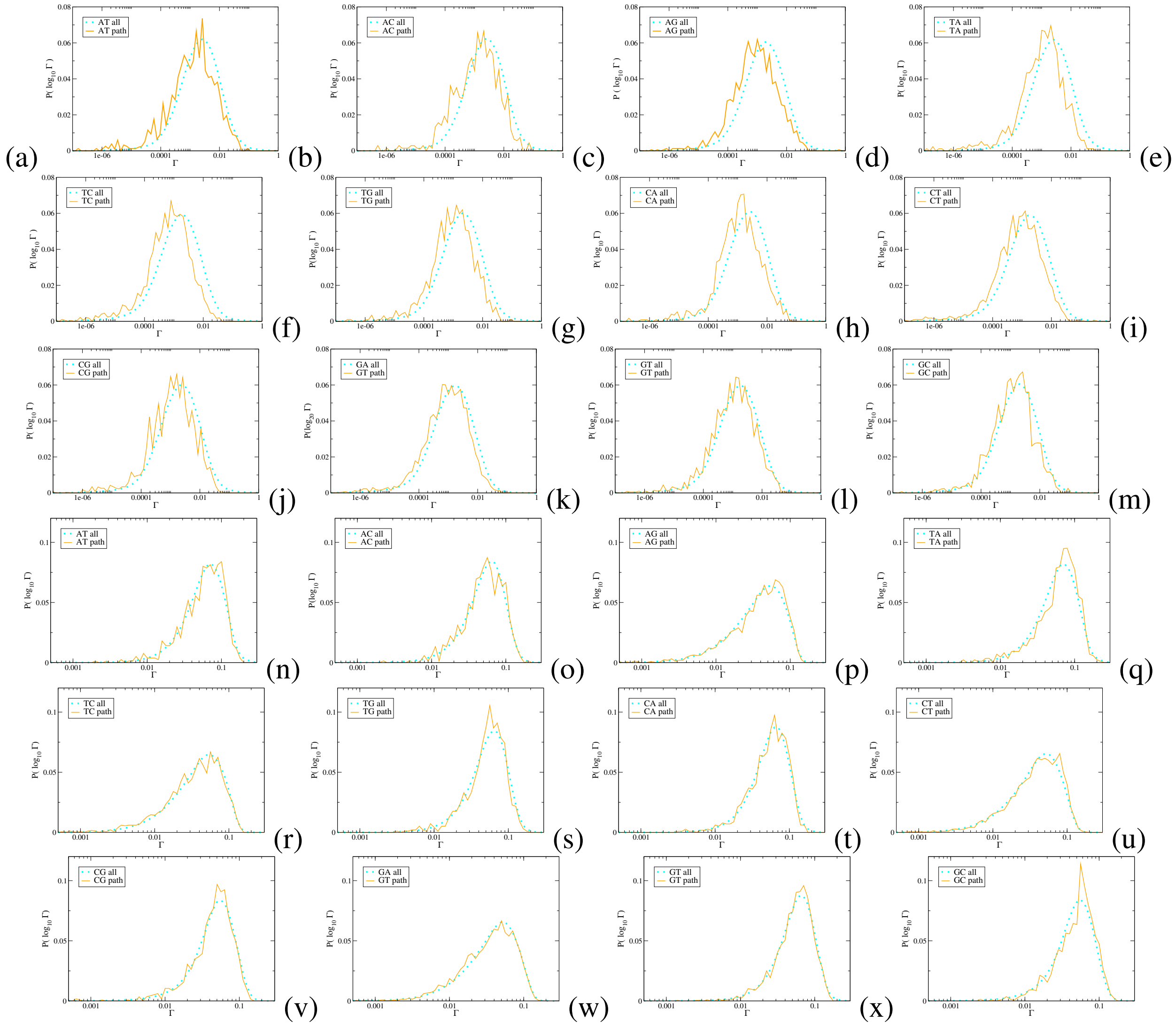

Figure S2: (Supplementary) Panels a-l: 1D model, results divided into the twelve subtypes of mutation. The shift for pathogenic mutations is clearly present in every case. Panels m-x: 2-leg model, results divided into the twelve subtypes of mutation. There is no consistent shift for pathogenic mutations.

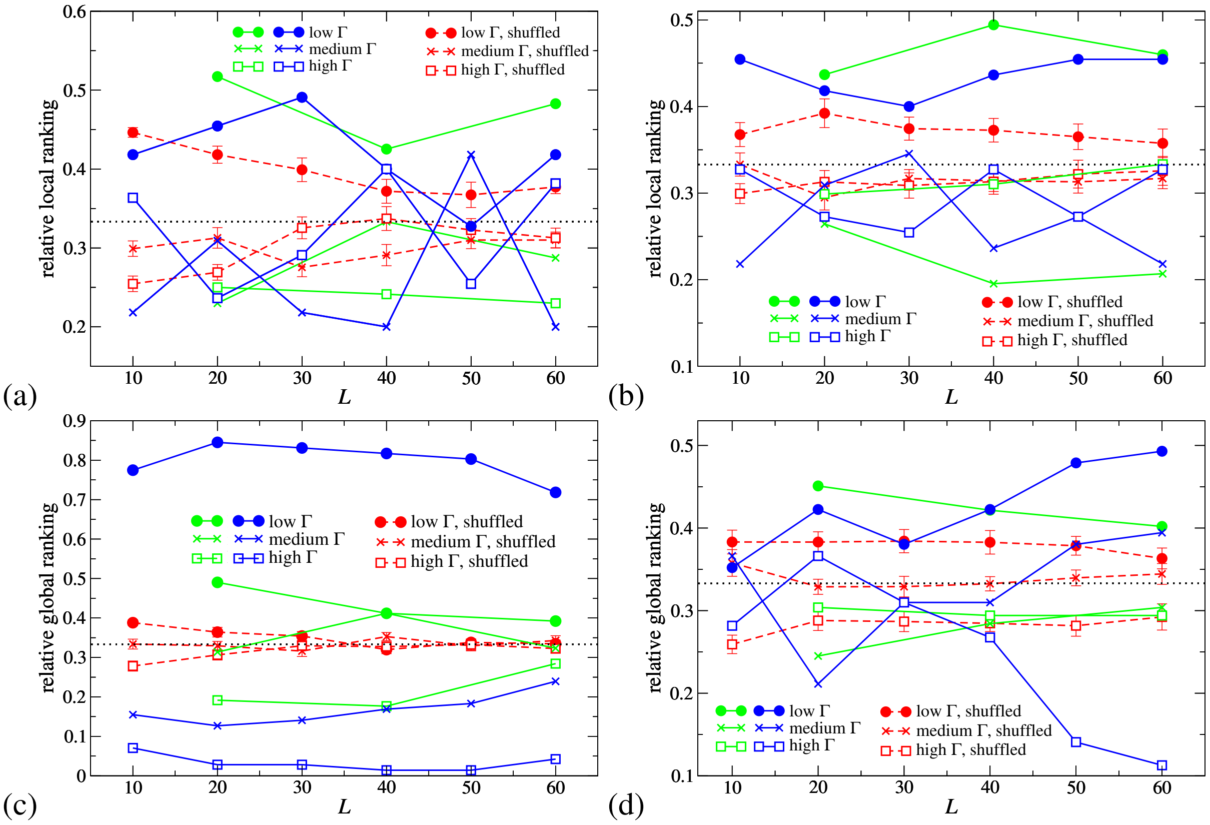

Figure S3: (Supplementary) Distribution of the *local* (a+b) and *global* (c+d) ranking results of pathogenic mutations of $p16$ (CDKN2A) (blue solid lines) and CYP21A2 (green) as a function of window lengths L. The dashed lines indicate averaged results for 20 randomly shuffled $p16$ sequences. The left/right columns distinguish results for the 1D/2-leg models. The dashed horizontal line shows the 33% mark expected for a completely random sequence. All lines are guides to the eyes only. Error bars are within symbol size.

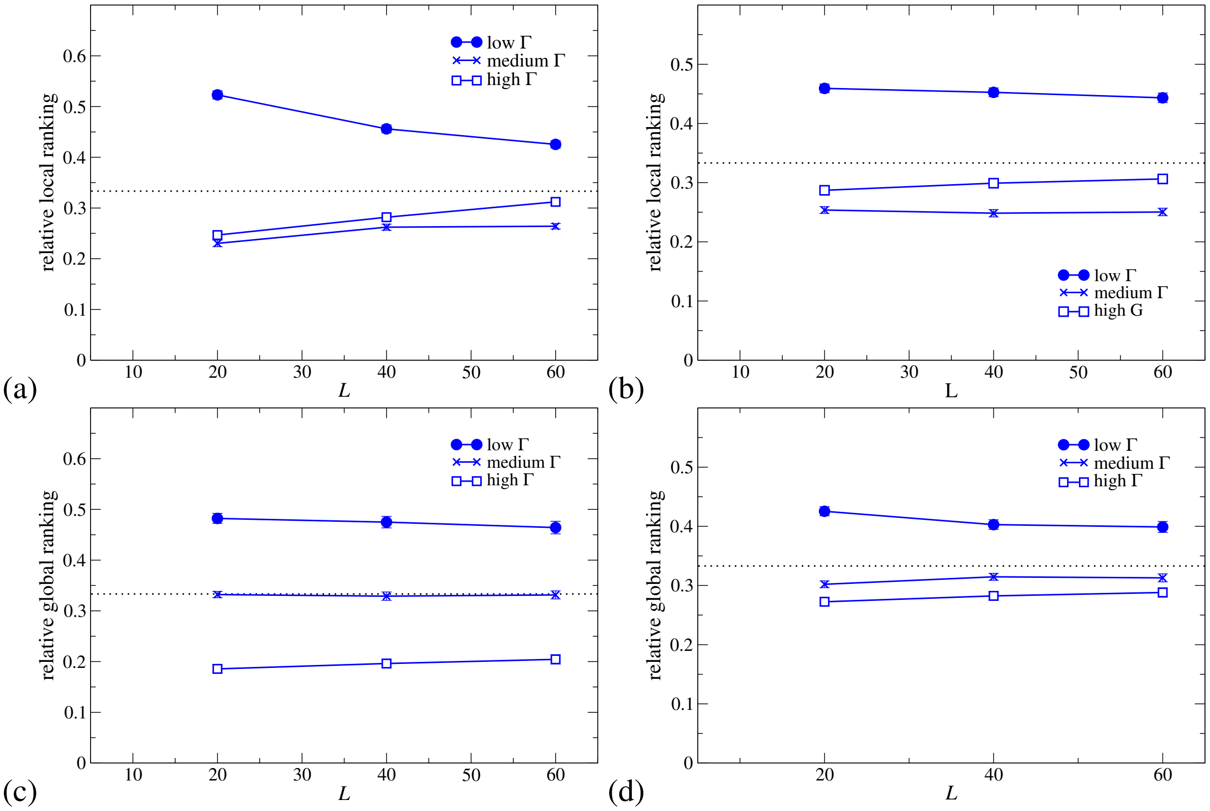

Figure S4: (Supplementary) Distribution of the *local* (a+b) and *global* (c+d) ranking results of *all* 19882 pathogenic mutations of the 162 genes as a function of window lengths L. The left/right columns distinguish results for the 1D/2-leg models. The dashed horizontal lines show the 33% mark of a completely random sequence. All lines are guides to the eyes only.

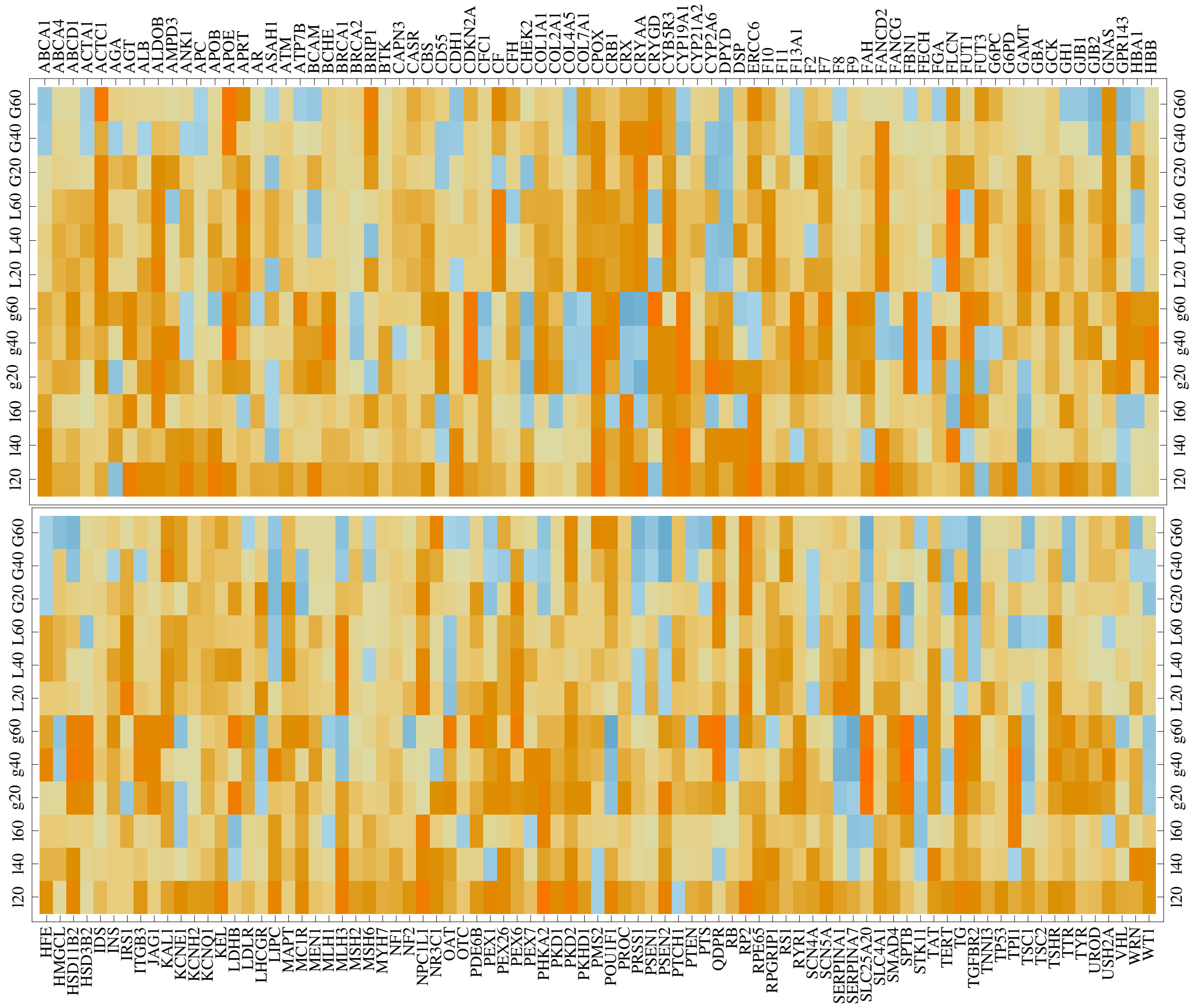

Figure S5: (Supplementary) Numerical representation of the 12 criteria for all 162 genes, i.e. deviation from the 0.33 line for the *local* rankings (li, Li) and the *global* rankings (gi, Gi) corresponding to the sorted prevalence for $L = 20, 40$ and 60, respectively. The lower case (l,g) indicates results for the 1D model, uppercase (L,G) refers to the 2-leg model. The genes are named according to the usage in the DNA databases.[3–6] The orange shading corresponds to an agreement with the CT hypothesis while the blue shading denotes disagreement. The first (last) column in the top (bottom) row gives the scale from 0 to 1 with 0.33 corresponding to the white square.

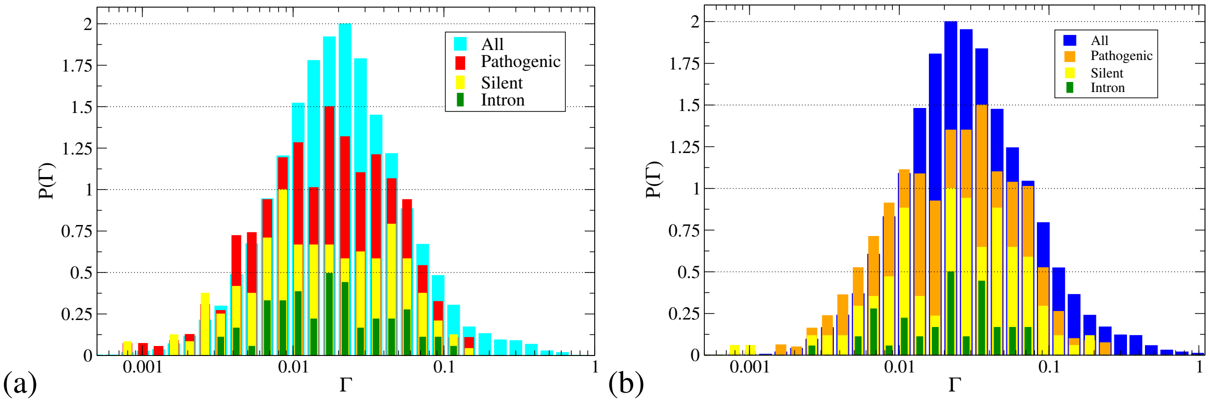

Figure S6: Histograms of Γ distributions for (a) transitions and (b) transversions in TP53, simulated using the 1D model and $L = 20$. Histograms are shown for all possible mutations and for pathogenic, silent and intronic subsets. The maximum heights of the populations are scaled to be 2, 1.5, 1 and 0.5 to ease comparison. The scales factors are indicated by the dotted horizontal lines.

Table S1: (Supplementary) List of the 162 genes with their lengths (bps), number of all point mutations (N_{pa}), and their numbers of the 12 types of point mutations. For example, N_{At} means the number of $A \rightarrow T$ substitution.

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
ABCA1	147154	87	0	4	9	2	7	2	4	24	3	18	7	7
ABCA4	128313	382	11	9	21	13	51	21	27	73	19	99	23	15
ABCD1	19894	223	8	7	14	6	31	3	15	46	17	47	13	16
ACTA1	2852	164	10	7	22	5	13	6	13	12	11	29	17	19
ACTC1	7631	14	0	1	3	0	0	0	0	4	1	2	1	2
AGA	11668	19	0	0	0	1	3	1	0	2	0	8	2	2
AGT	11673	10	0	0	1	0	1	1	0	5	0	1	0	1
ALB	17127	63	3	2	13	2	1	0	1	6	1	24	4	6
ALDOB	14448	28	0	0	0	1	9	1	3	5	3	3	1	2
AMPD3	56903	11	0	1	0	1	1	0	0	6	1	0	0	1
ANK1	144397	18	0	0	1	0	2	0	1	7	1	4	2	0
APC	108353	222	10	0	4	18	1	8	21	83	28	18	28	3
APOB	42645	51	0	0	2	4	1	1	3	26	2	8	3	1
APOE	3612	33	0	1	1	0	2	0	2	9	2	9	2	5
APRT	2466	13	2	0	1	0	3	0	0	1	0	4	1	1
AR	180246	299	11	6	24	11	31	12	22	53	25	56	31	17

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
ASAH1	28574	12	1	0	3	1	0	0	1	0	3	1	0	2
ATM	146268	169	8	3	20	9	11	15	5	55	10	19	8	6
ATP7B	78826	315	10	14	25	14	27	10	17	62	16	68	30	22
BCAM	12341	14	1	0	1	1	0	0	1	4	1	5	0	0
BCHE	64562	58	6	2	6	3	6	3	2	12	0	8	5	5
BRCA1	81155	301	12	6	30	14	29	23	12	63	15	38	50	9
BRCA2	84193	162	12	9	20	8	11	8	12	33	13	15	19	2
BRIP1	180771	13	1	0	0	0	1	1	0	3	2	2	1	2
BTK	36741	329	15	14	29	19	47	23	26	44	14	48	32	18
CAPN3	64215	213	2	9	18	5	23	6	10	45	19	48	14	14
CASR	102813	144	2	5	12	4	21	7	8	20	10	38	12	5
CBS	23121	107	2	1	6	4	10	0	4	24	7	39	2	8
CD55	38983	14	0	0	1	2	0	1	0	4	0	3	1	2
CDH1	98250	30	0	1	2	0	2	2	0	9	1	8	4	1
CDKN2A	26740	71	1	3	4	2	6	6	5	12	3	11	8	10
CFC1	6748	10	0	0	0	0	1	0	0	4	1	4	0	0
CF	188699	828	35	31	103	50	85	54	47	117	41	136	84	45
CFH	95494	83	3	3	8	6	10	5	2	10	6	14	13	3
CHEK2	54092	20	1	1	2	0	1	0	2	4	0	7	1	1
COL1A1	17544	292	0	2	2	0	1	2	1	21	4	134	79	46
COL2A1	31538	124	0	1	2	1	1	1	5	26	0	53	19	15

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
COL4A5	257622	244	2	0	4	2	2	5	4	20	1	117	55	32
COL7A1	31088	265	0	3	6	2	1	0	1	56	7	122	34	33
CPOX	14152	36	0	2	1	0	3	1	0	14	2	9	3	1
CRB1	210178	91	3	1	2	8	16	7	3	11	2	22	11	5
CRX	21483	18	0	1	1	0	0	1	2	4	0	8	0	1
CRYAA	3773	10	0	0	0	0	0	0	1	5	0	3	1	0
CRYGD	2882	12	0	1	0	0	0	0	4	3	1	2	1	0
CYB5R3	30587	35	0	0	3	0	6	2	2	12	0	10	0	0
CYP19A1	129126	13	0	0	0	0	2	1	0	5	0	5	0	0
CYP21A2	3338	102	4	4	5	7	8	4	6	23	2	25	4	10
CYP2A6	6897	12	1	0	1	1	2	0	0	2	0	2	2	1
DPYD	843317	34	2	3	7	2	0	1	2	7	0	5	4	1
DSP	45077	20	0	0	2	1	1	1	0	6	1	6	1	1
ERCC6	80364	18	1	0	2	1	1	1	0	10	1	1	0	0
F10	26731	81	1	4	5	1	6	2	4	11	3	33	5	6
F11	23718	131	2	5	6	3	17	3	9	28	2	29	13	14
F13A1	176614	55	1	0	2	0	6	4	4	12	3	14	8	1
F2	20301	42	0	3	3	0	1	1	0	11	1	17	3	2
F7	14891	164	4	1	13	1	17	4	9	30	6	55	13	11
F8	186936	1168	79	47	124	56	117	78	55	153	72	198	112	77
F9	32723	707	31	26	55	58	69	52	42	54	28	135	95	62

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
FAH	33342	26	2	1	2	0	1	3	2	6	0	5	4	0
FANCD2	75502	14	0	0	0	0	3	3	0	4	0	4	0	0
FANCG	6179	16	0	0	0	0	2	1	0	7	0	2	2	2
FBN1	237414	640	18	12	52	32	88	37	21	63	32	173	68	44
FECH	38454	49	2	1	2	3	7	3	1	11	1	11	4	3
FGA	7618	45	3	1	3	3	1	2	3	12	2	7	7	1
FLCN	24971	11	0	0	1	0	0	0	0	4	2	3	1	0
FUT1	7380	22	0	0	1	2	2	1	2	5	1	4	1	3
FUT3	8587	11	0	0	0	1	0	2	2	0	0	5	0	1
G6PC	12572	66	2	2	3	2	8	3	3	13	2	15	5	8
G6PD	16182	163	3	3	21	4	15	4	8	27	15	39	13	11
GAMT	4465	11	0	2	0	0	1	0	0	1	1	3	1	2
GBA	10246	259	8	11	25	8	32	19	14	42	10	53	19	18
GCK	45153	255	5	13	15	7	32	8	19	40	11	64	23	18
GH1	1636	35	2	2	7	0	3	1	1	5	2	7	3	2
GJB1	10004	240	4	5	25	18	31	12	10	39	24	39	17	16
GJB2	5513	208	8	9	19	5	28	8	12	23	15	49	19	13
GNAS	71456	51	2	2	2	1	6	2	1	17	4	9	3	2
GPR143	40464	43	2	0	3	2	4	3	4	6	1	10	4	4
HBA1	842	73	2	5	9	2	5	2	7	6	9	8	7	11
HBB	1606	263	15	20	20	21	23	16	22	26	18	38	20	24

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
HFE	9612	27	1	2	0	0	4	1	0	3	2	7	3	4
HMGCL	23583	27	2	0	2	0	3	1	1	4	1	8	3	2
HSD11B2	6421	24	1	0	1	0	3	2	1	12	1	3	0	0
HSD3B2	7879	32	0	1	1	1	2	3	3	8	3	6	2	2
IDS	26493	203	15	8	15	2	16	13	17	31	19	32	20	15
INS	1431	30	0	0	2	0	3	2	1	3	6	6	4	3
IRS1	64538	14	0	1	3	0	1	0	1	2	1	3	0	2
ITGB3	58870	53	2	2	3	1	10	4	1	12	1	11	5	1
JAG1	36257	131	2	0	3	6	11	6	11	30	12	28	16	6
KAL1	203313	25	0	0	1	2	1	1	1	9	2	6	1	1
KCNE1	65586	17	0	0	1	0	2	0	1	5	0	6	1	1
KCNH2	32966	266	8	11	27	5	19	12	15	61	9	43	35	21
KCNQ1	404120	226	3	2	19	8	24	5	12	44	13	61	11	24
KEL	21303	33	2	0	3	1	3	0	0	9	1	13	0	1
LDHB	22501	11	1	1	1	0	1	2	1	1	0	2	0	1
LDLR	44450	741	23	31	48	31	84	35	51	88	48	168	92	42
LHCGR	68951	37	2	3	3	3	7	3	2	7	1	3	2	1
LIPC	136898	11	0	1	2	0	0	1	0	2	0	4	0	1
MAPT	133924	36	3	2	2	0	3	2	2	6	1	9	5	1
MC1R	2360	24	0	1	1	0	4	0	3	8	0	5	1	1
MEN1	7779	239	10	7	8	9	26	11	19	44	14	38	33	20

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
MLH1	57359	275	16	15	26	18	19	17	18	42	20	36	28	20
MLH3	37769	17	0	1	5	0	1	0	0	2	1	4	2	1
MSH2	80098	238	16	11	25	8	9	14	11	62	14	30	25	13
MSH6	23872	54	3	1	5	2	3	0	3	17	6	7	4	3
MYH7	22924	268	8	10	20	4	19	8	16	47	17	80	16	23
NF1	282701	338	22	4	24	20	35	26	14	82	24	44	29	14
NF2	95023	72	5	2	5	2	6	1	2	25	4	7	11	2
NPC1L1	28781	26	0	0	3	2	0	0	0	11	1	8	1	0
NR3C1	157582	14	1	0	1	1	4	1	0	1	0	4	0	1
OAT	21580	42	0	0	2	2	4	0	3	9	2	11	5	4
OTC	68968	276	16	11	28	9	31	18	17	36	15	44	27	24
PDE6B	45199	20	1	0	0	3	3	1	2	5	1	4	0	0
PEX1	41509	24	0	0	0	0	4	1	2	7	3	6	0	1
PEX26	11503	10	0	0	0	0	3	0	0	3	2	2	0	0
PEX6	15143	18	0	1	0	0	3	1	1	7	0	5	0	0
PEX7	91337	24	1	2	2	1	1	3	2	6	1	4	1	0
PHKA2	91305	23	0	1	2	0	0	1	1	11	0	5	2	0
PKD1	47189	149	2	3	6	5	12	4	8	59	10	27	8	5
PKD2	70110	35	1	0	1	1	1	1	2	17	0	7	3	1
PKHD1	472279	213	8	10	22	7	29	9	7	50	7	38	17	9
PMS2	35868	21	3	1	1	1	0	0	0	6	0	5	4	0

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
POU1F1	16954	22	1	0	2	1	3	1	1	6	0	4	2	1
PROC	10802	203	6	6	10	3	21	6	15	40	8	55	13	20
PRSS1	3592	26	1	2	2	2	2	0	2	5	1	5	2	2
PSEN1	83931	154	6	8	13	8	22	11	7	21	12	19	16	11
PSEN2	25532	18	2	2	3	0	0	0	0	5	1	5	0	0
PTCH1	73984	59	3	2	1	2	4	2	7	15	2	11	8	2
PTEN	105338	98	2	2	10	9	13	11	5	15	8	15	6	2
PTS	7595	27	1	0	8	1	0	2	0	6	2	4	2	1
QDPR	57702	20	0	1	2	0	3	3	0	3	0	6	1	1
RB	180388	226	9	8	18	12	16	11	10	38	8	51	28	17
RP2	45418	17	0	0	1	0	1	2	0	5	2	4	2	0
RPE65	21136	42	1	1	2	1	5	3	3	9	1	7	7	2
RPGRIP1	63325	24	0	2	5	1	0	0	0	7	0	5	3	1
RS1	32422	93	3	0	7	5	11	4	5	15	5	19	7	12
RYR1	153865	244	5	4	21	9	20	6	10	56	14	63	17	19
SCN4A	34365	43	1	0	5	2	3	1	4	7	3	12	2	3
SCN5A	101611	226	0	2	18	9	16	6	13	49	10	77	15	11
SERPINA1	12332	29	4	1	0	1	2	1	2	8	1	9	0	0
SERPINA7	3870	16	1	0	0	2	1	0	1	4	0	5	1	1
SLC25A20	41966	11	0	0	1	0	0	0	0	4	1	3	1	1
SLC4A1	18428	65	1	1	3	2	5	0	6	20	4	20	1	2

Name	Length	N_{pa}	N_{At}	N_{Ac}	N_{Ag}	N_{Ta}	N_{Tc}	N_{Tg}	N_{Ca}	N_{Ct}	N_{Cg}	N_{Ga}	N_{Gt}	N_{Gc}
SMAD4	49535	20	0	1	1	0	0	2	1	6	3	4	1	1
SPTB	76865	18	0	0	2	2	2	2	0	6	1	0	1	2
STK11	22637	62	4	4	2	1	4	5	7	12	5	8	8	2
TAT	10242	11	0	0	0	0	1	1	0	5	1	2	1	0
TERT	41881	30	0	1	3	1	2	0	0	10	3	8	0	2
TG	267939	33	0	1	2	1	2	1	1	7	0	14	4	0
TGFBR2	87641	14	0	0	1	1	1	0	0	5	0	3	1	2
TNNI3	5966	30	0	1	5	1	1	0	0	8	2	10	0	2
TP53	20303	2003	137	113	158	121	142	109	165	284	156	252	202	164
TPI1	3287	11	0	0	1	1	1	0	0	1	0	4	1	2
TSC1	53285	44	2	0	1	1	1	2	5	19	5	5	3	0
TSC2	40724	165	7	4	6	5	13	5	22	48	18	22	10	5
TSHR	190778	45	1	0	3	1	9	2	3	8	1	12	2	3
TTR	6944	98	4	5	10	6	15	9	6	5	1	19	11	7
TYR	117888	205	10	10	22	6	16	6	16	27	13	42	26	11
UROD	3512	45	0	1	2	5	6	2	3	9	2	11	2	2
USH2A	800503	66	0	3	1	1	2	3	6	24	3	8	10	5
VHL	10444	172	5	7	12	13	22	15	7	22	21	17	18	13
WRN	140499	22	3	1	1	1	1	0	0	11	2	1	1	0
WT1	47763	56	1	2	5	1	6	3	3	13	4	11	5	2