Supporting Information

Oncogenic potential is directly related to activating effect of cancer single and double somatic mutations in receptor tyrosine kinases

Kosuke Hashimoto, Igor B. Rogozin, and Anna R. Panchenko

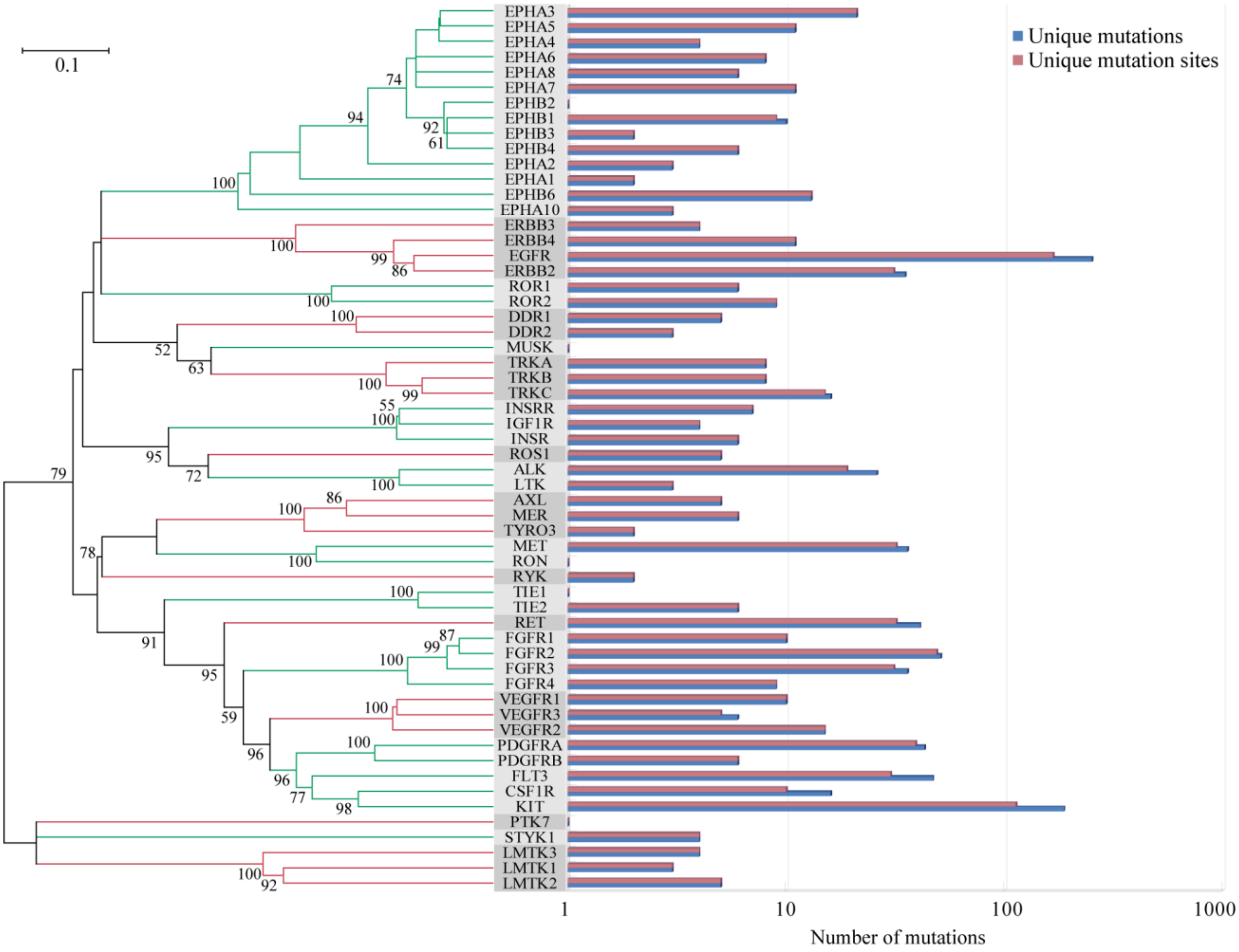

Figure S1

Supp. Figure S1. The distribution of mutations in 58 RTKs. Left panel shows a phylogenetic tree of RTKs, based on the multiple sequence alignment of the kinase domain. Bootstrap values above 50% are displayed on corresponding branches. Protein names and branches are depicted in alternate colors for each family. The right panel shows the number of unique mutations in blue and mutation sites in red on a log scale.

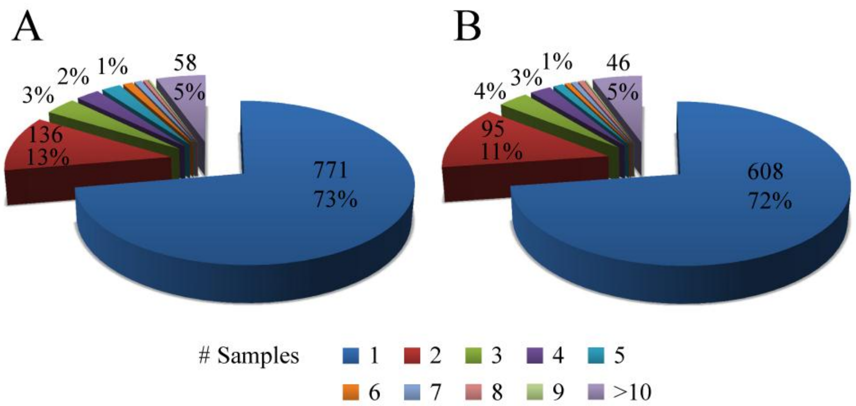

Figure S2

Supp. Figure S2. Percentage of mutations (A) and mutation sites (B) for different sample frequencies (oncogenic potential). Mutations define a particular amino acid substitution at a given protein site, there can be several mutations at a given site.

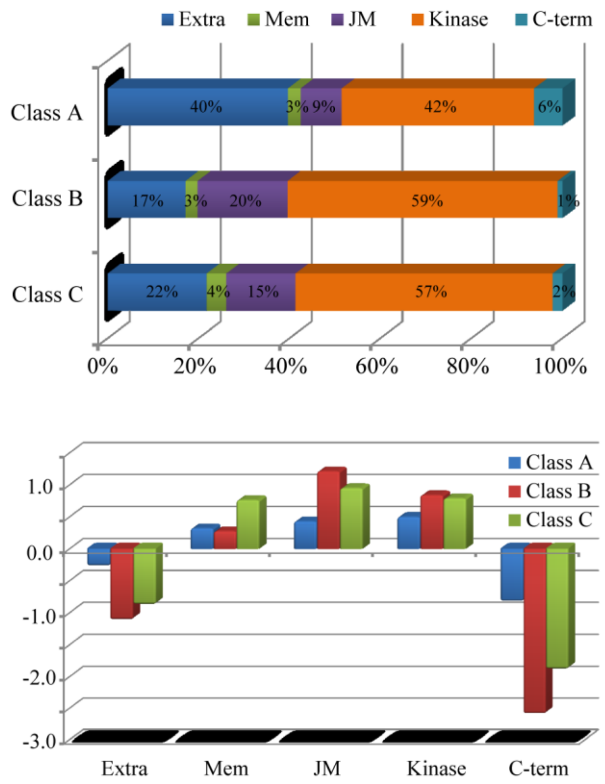

Figure S3

Supp. Figure S3. Localization of mutation sites in different regions of RTKs. (A) Percentages of mutation sites per region for different classes (A, B and C). (B) The log-ratio between mutation frequency in a given region and mutation frequency in all RTK regions. Mutation frequency is calculated as the number of mutation sites in a region divided by the length of the region.

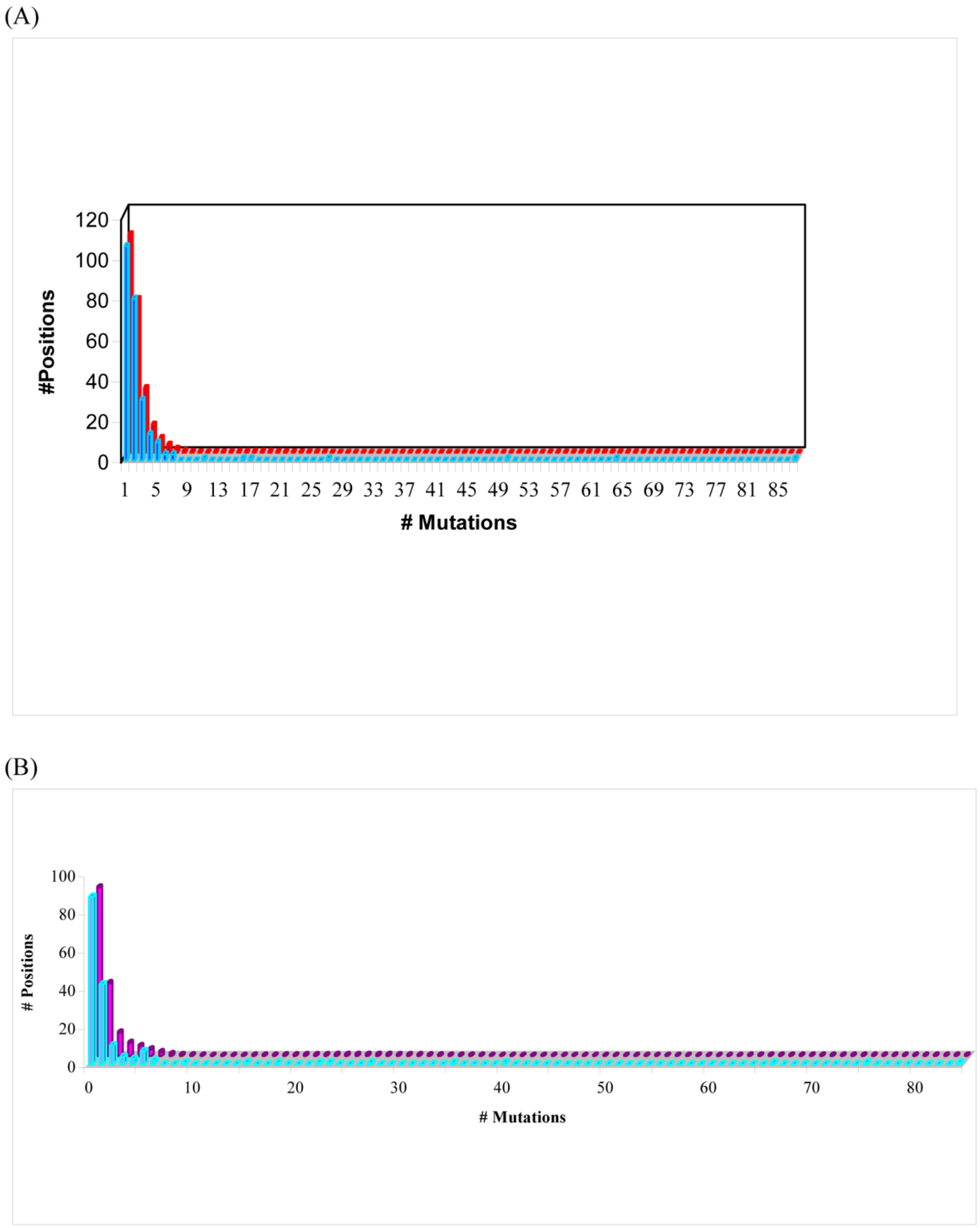

Figure S4

Supp. Figure S4. Mutation spectra for EGFR (A) and KIT (B). The observed number of mutations is shown in blue, the expected number of mutations is shown in red.

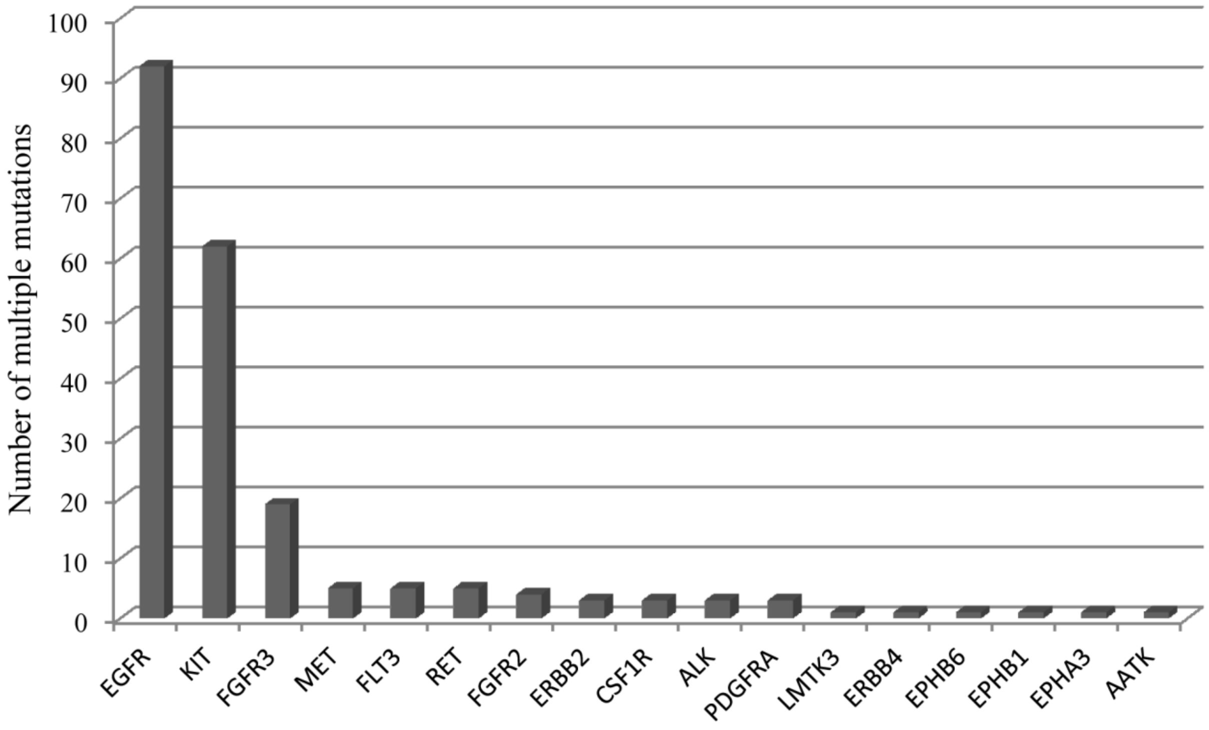

Figure S5

Supp. Figure S5. Number of multiple mutations in different RTK families.

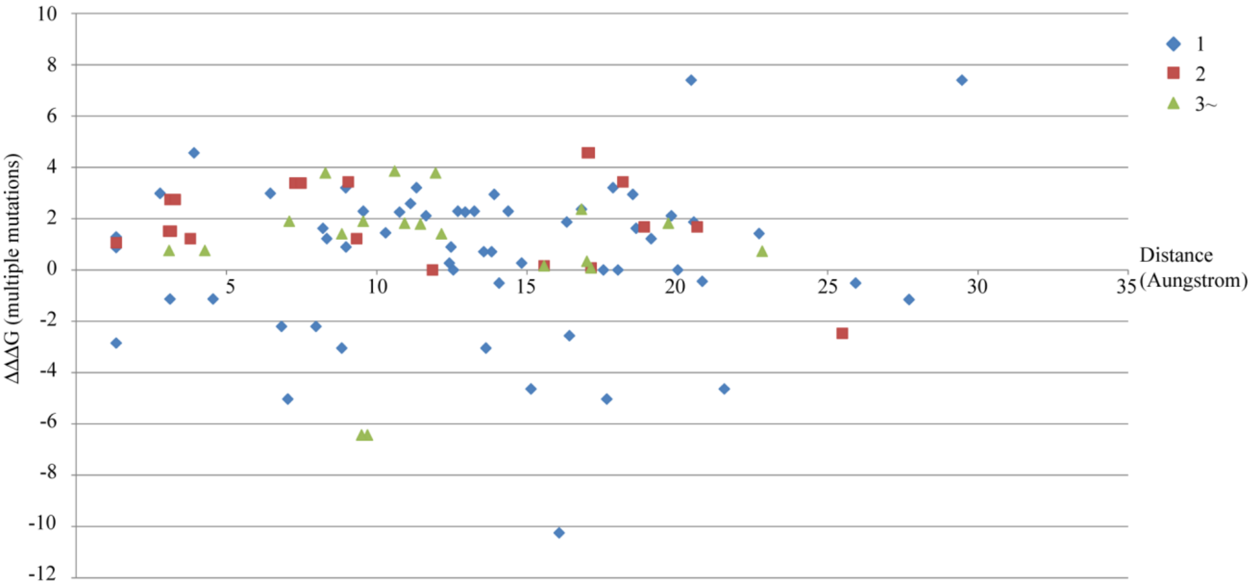

Figure S6

Supp. Figure S6. Correlation between activation effect of doublets (ΔΔΔG values) and spatial distances between two mutation sites in a protein molecule. Double mutations derived from one sample are shown as blue diamonds, from two samples as red squares and from more than three samples as green triangles.

Supp. Table S1. RTK mutations mapped on crystal structures in active and inactive states

Name	Protein accession	Active state	#mutations in active	Inactive state	#mutations in inactive	# mutations in both
EGFR	NM_005228	2GS6	207	2GS7	201	199
KIT	NM_000222	1PKG	94	1T45	150	94
FGFR2	NM_000141.2	2PVF	18	2PSQ	18	18
FGFR1	NM_000604	3GQI	4	1FGK	4	4
ERBB4	NM_005235	3BCE	2	3BBW	2	2
IGF1R	NM_000875	1K3A	1	1M7N	1	1
RET	NM_020975	2IVT	18			
MET	NM_000245			2G15	21	
ALK	NM_004304			3LCS	16	
MER	NM_006343			2P0C	3	
TIE2	NM_000459			1FVR	2	
EPHA2	NM_004431			1MQB	1	

Supp. Table S2. Comparison of non-synonymous and synonymous substitutions for double mutations in EGFR, KIT and TP53 gene

	Two non-synonymous substitutions	One non-synonymous substitution + one synonymous substitution	Two synonymous substitutions
EGFR and KIT	566	51	3
TP53*	9	12	7
	$P < 10^{-10}$		
Hardy-Weinberg model# for EGFR and KIT	p^2	2pq	q^2
	p = 23.8	2pq = 82	q = 1.73
	P = 0.0045 according to the binomial test		

# - in the Hardy-Weinberg model, the mathematical relation between the allele frequencies and the genotype frequencies is given by AA : p2; Aa : 2pq; aa : q2; in which p2; 2pq; q2 are the frequencies of the genotypes AA; Aa; aa in zygotes and p; q are the allele frequencies of A and a in gametes in the previous generation and p + q = 1.

* - frequencies of substitutions of P53 gene are taken from Meng et al, “Multiple mutations of the p53 gene in human mammary carcinoma”. Mutat Res 435(3):263-9 (1999).