

Additional files

Figure S1. Comparison of amino acid frequencies of indel sequences in “all indels”(All), “naturally occurring indels”(Natural) and reference (Background) datasets.

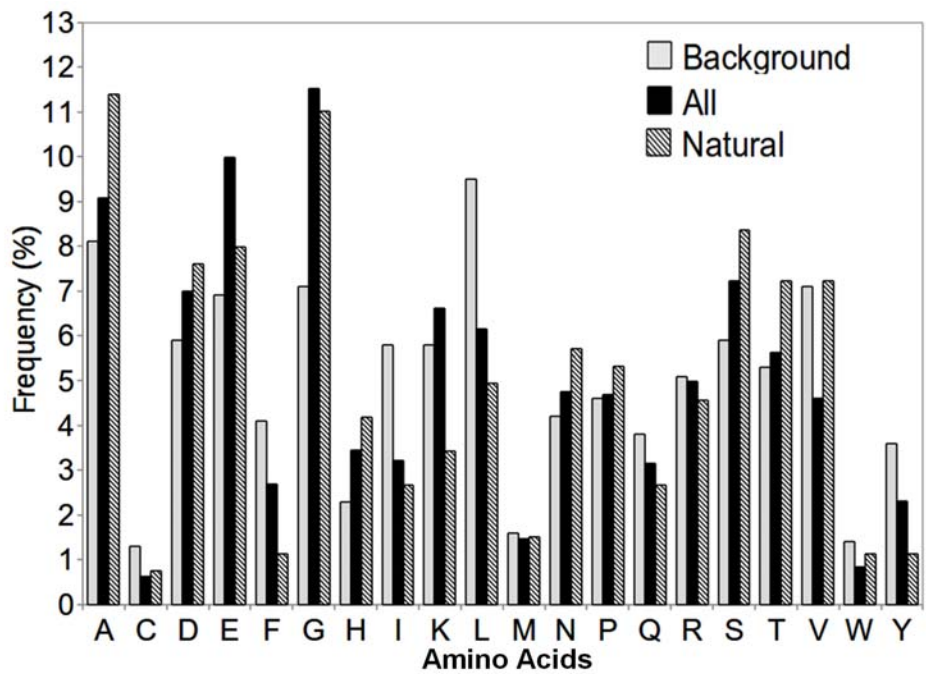


Figure S2. Frequencies of secondary structure types of residues flanking indel sequences.

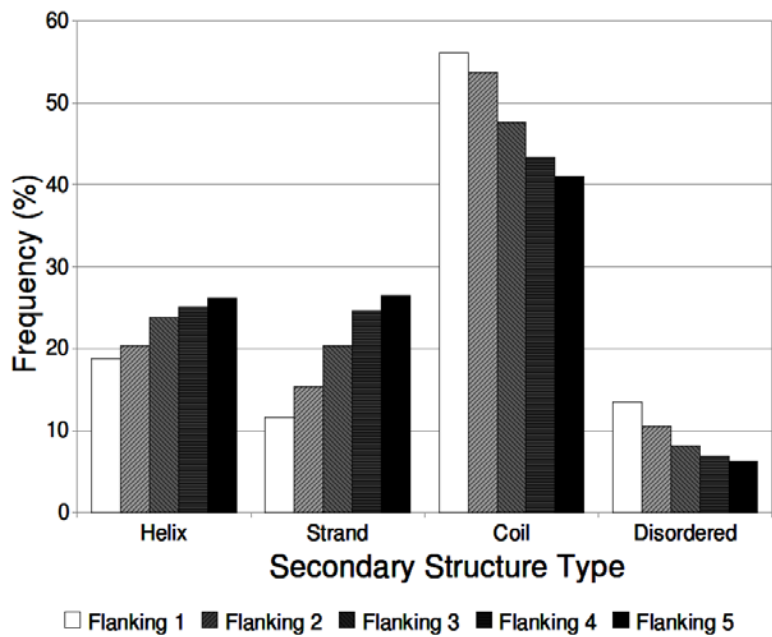


Table S1. χ^2 analysis of amino acids frequencies, secondary structure types, and relative solvent accessibility of “All” indels

Amino acids	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	<i>p-value</i>
observed	118	8	91	130	35	150	45	42	86	80	19	62	61	41	55	94	73	60	11	30	1.36e ⁻¹⁹
expected¹	105	17	77	90	53	92	30	75	75	124	21	55	60	49	66	77	69	92	18	47	

Secondary structure types	helix	strand	coil	disordered	<i>p-value</i>
observed	276	94	548	383	0.0
expected¹	468	270	490	73	

Relative solvent accessibility	buried	intermediate	exposed	<i>p-value</i>
observed	115	292	894	1.28e ⁻¹³⁰
expected¹	408	411	482	

¹Expected numbers are calculated using the frequencies from the background dataset as described in Methods

Table S2. χ^2 analysis of amino acids frequencies, secondary structure types, and relative solvent accessibility of “Natural” indels

Amino acids	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	<i>p-value</i>
observed	30	2	20	21	3	29	11	7	9	13	4	15	14	7	12	22	19	19	3	3	5.1e ⁻⁴
expected¹	21	3	16	18	11	19	6	15	15	25	4	11	12	10	13	16	14	19	4	9	
expected²	24	2	18	26	7	30	9	8	17	16	4	13	12	8	13	19	15	12	2	6	0.46

Secondary structure types	helix	strand	coil	disordered	<i>p-value</i>
observed	51	20	114	78	2.6e ⁻⁶⁸
expected¹	95	55	99	15	
expected²	56	19	111	77	

Relative solvent accessibility	buried	intermediate	exposed	<i>p-value</i>
observed	20	40	203	9.3e ⁻⁴¹
expected¹	83	83	97	
expected²	23	59	181	

¹Expected numbers are calculated using the frequencies in the “Background” dataset as described in Methods

²Expected numbers are calculated using the frequencies of “All” indels.