	1000 bp RepeatMasked Conserved Sequence			2000 bp RepeatMasked Conserved Sequence			5000 bp RepeatMasked Conserved Sequence		
Word	Up	Intronic	Down	Up	Intronic	Down	Up	Intronic	Down
CGCGCG	20.9	10.6	2.4	16.1	8.0	2.7	9.8	5.7	3.3
GCGCGC	18.7	9.9	2.5	14.4	7.5	3.0	9.0	5.5	3.5
CGCCGC	17.2	8.9	2.6	13.3	6.7	3.4	8.5	4.8	3.8
TCGCGA	17.0	6.8	1.4	13.1	5.7	2.1	8.3	4.2	2.6
CCGCGC	16.8	10.6	2.3	13.1	7.8	3.0	8.3	5.7	3.5
CGCGCC	16.8	10.9	2.4	13.0	8.0	3.0	8.4	5.8	3.6
GCGCGA	16.7	9.2	2.4	12.9	6.8	3.0	8.2	5.1	3.7
CGCGAG	16.4	8.9	2.0	12.6	6.7	2.7	8.0	4.9	3.3
CGGCGC	16.4	9.1	2.5	12.7	6.6	3.2	8.3	4.7	3.8
CGCGGC	15.9	9.9	2.4	12.5	7.4	3.1	8.0	5.3	3.7
CCGCCG	15.5	8.5	2.7	12.1	6.4	3.4	8.0	4.6	3.8
CGCGCA	15.2	8.6	2.7	11.8	6.6	3.2	7.6	5.1	3.7
CGCGAC	15.1	9.3	2.4	11.9	6.8	3.3	7.7	4.9	3.8
CGCCCG	14.9	10.2	2.4	11.9	7.6	3.1	7.8	5.4	3.6
CCGGCG	14.8	9.1	2.4	11.5	6.6	3.1	7.7	4.8	3.6
CCGCGG	14.7	10.8	2.5	11.4	8.1	3.0	7.5	5.8	3.4
CCCGCG	14.5	11.3	2.4	11.4	8.3	2.8	7.5	5.9	3.3
CGCGGA	14.4	9.6	2.3	11.2	7.1	3.0	7.3	5.2	3.4
ACGCGC	14.1	8.8	2.7	11.2	6.6	3.0	7.2	5.0	3.8
CGGCCG	13.4	9.3	2.3	10.5	6.9	3.2	6.9	4.8	3.7
CGCCCC	13.3	10.0	2.3	10.6	7.5	2.7	7.1	5.4	3.1
CGTCGC	12.4	6.4	2.6	9.7	4.9	3.3	6.8	3.6	3.9
CGACGC	12.2	6.4	2.5	9.5	4.9	3.4	6.7	3.7	4.1
GCGCCC	11.4	9.4	2.0	9.4	7.1	2.7	6.4	5.1	3.2
ACGGCG	10.4	6.1	2.6	8.2	4.6	3.4	5.8	3.4	3.9
CCGACG	9.6	5.8	2.4	7.7	4.6	3.3	5.6	3.3	3.9
CGGCCC	9.1	7.5	2.6	7.4	5.6	3.1	5.5	4.1	3.6
CGGGCC	8.9	7.3	2.6	7.4	5.5	3.2	5.4	4.0	3.8
CGTCGA	6.1	3.5	2.6	5.1	2.6	3.3	4.0	2.0	3.5
CGACGA	6.1	3.2	2.2	5.4	2.6	3.5	4.3	1.8	3.9
CCCCCC	5.4	5.9	2.6	4.8	4.7	2.7	4.0	3.6	2.8

Additional Table 1a. Word frequencies of the 10 most frequently occurring words from each of nine different classes of genomic sequence windows from the human genome. Three overlapping sequence windows of 1000, 2000 and 5000 bp from upstream, intronic and downstream regions were examined. Only RepeatMasked human sequence that was conserved in the mouse genome [35,36,76] (see **Methods**) was considered. For each word, its over-representation ratio in the given sequence window versus in all noncoding, RepeatMasked, conserved sequence in the entire human genome is given. Words are listed in descending order according to their over-representation scores in the 0-1000 bp upstream windows.

Word	1000 bp RepeatMasked Conserved Sequence			1000-2000 bp RepeatMasked Conserved Sequence			2000-5000 bp RepeatMasked Conserved Sequence		
	Up	Intronic	Down	Up	Intronic	Down	Up	Intronic	Down
CGCGCG	20.9	10.6	2.4	6.1	2.8	3.1	2.5	1.3	3.8
GCGCGC	18.7	9.9	2.5	5.7	2.9	3.5	2.7	1.7	3.9
CGCCGC	17.2	8.9	2.6	5.2	2.4	4.2	3.1	1.1	4.2
TCGCGA	17.0	6.8	1.4	5.2	3.5	2.9	2.8	1.3	2.9
CCGCGC	16.8	10.6	2.3	5.5	2.4	3.7	2.9	1.6	3.8
CGCGCC	16.8	10.9	2.4	5.3	2.4	3.8	3.0	1.5	4.0
GCGCGA	16.7	9.2	2.4	5.0	2.3	3.6	2.8	1.8	4.3
CGCGAG	16.4	8.9	2.0	4.9	2.4	3.5	2.7	1.4	3.7
CGGCGC	16.4	9.1	2.5	5.1	1.8	3.9	3.2	1.2	4.3
CGCGGC	15.9	9.9	2.4	5.4	2.6	3.8	3.0	1.3	4.1
CCGCCG	15.5	8.5	2.7	5.0	2.4	4.0	3.3	1.0	4.1
CGCGCA	15.2	8.6	2.7	4.9	2.6	3.7	2.7	2.2	4.0
CGCGAC	15.1	9.3	2.4	5.5	1.8	4.1	2.9	1.3	4.1
CGCCCG	14.9	10.2	2.4	5.7	2.6	3.8	3.1	1.2	3.9
AGCGCG	14.8	8.4	2.3	5.3	2.6	3.8	2.7	2.2	4.2
CCGGCG	14.8	9.1	2.4	4.8	1.9	3.8	3.5	1.2	4.1
CCGCGG	14.7	10.8	2.5	4.8	3.0	3.5	2.9	1.4	3.8
CCCGCG	14.5	11.3	2.4	5.1	2.6	3.3	2.9	1.3	3.7
CGAGCG	14.5	7.4	2.1	5.0	2.6	3.4	3.0	1.0	3.9
CGCGGA	14.4	9.6	2.3	4.8	2.3	3.7	2.8	1.6	3.6
ACGCGC	14.1	8.8	2.7	5.2	2.4	3.3	2.6	2.0	4.4
CCGCCC	13.6	8.9	2.5	4.9	2.6	3.3	3.0	1.4	3.5
CGGCCG	13.4	9.3	2.3	4.6	2.1	4.2	2.9	0.9	4.0
CGCCCC	13.3	10.0	2.3	5.1	2.5	3.1	3.0	1.4	3.3
CCGCGA	13.1	9.0	2.4	5.4	2.1	3.7	2.9	1.8	4.0
CCCGCC	12.7	8.4	2.4	4.7	2.7	3.4	3.2	1.4	3.4
ACCGCG	12.4	8.9	2.5	4.5	2.1	3.8	3.0	2.2	4.2
CGTCGC	12.4	6.4	2.6	4.3	2.0	4.0	3.4	1.2	4.4
CGACGC	12.2	6.4	2.5	3.9	2.1	4.4	3.4	1.4	4.6
CGGACG	12.1	7.3	2.4	4.5	1.8	3.8	2.7	1.1	4.3
CCCCGC	11.7	8.9	2.3	4.6	2.6	3.1	3.0	1.4	3.3
CGACCG	11.5	7.5	2.4	4.5	2.7	3.0	2.9	2.1	3.9
GCGCCC	11.4	9.4	2.0	5.2	2.5	3.4	3.0	1.3	3.7
CGCGAA	11.4	7.0	1.9	5.8	2.0	2.9	2.7	2.2	3.1
ACGCGG	11.1	8.7	2.4	4.2	1.7	3.6	3.1	1.1	4.4
ACGGCG	10.4	6.1	2.6	3.7	1.8	4.1	3.0	1.2	4.4
CTCCGC	10.3	7.1	2.0	4.0	2.3	2.7	2.7	1.7	3.0
GCGGCC	10.3	7.4	2.4	4.2	2.2	4.0	3.1	1.2	4.1
ATCGCG	9.8	6.5	2.0	4.1	3.6	2.5	2.4	1.8	3.8
CCGACG	9.6	5.8	2.4	3.8	2.2	4.2	3.1	1.0	4.4
CGGCCC	9.1	7.5	2.6	3.9	2.1	3.7	3.3	1.1	4.0
CGGGCC	8.9	7.3	2.6	4.1	1.9	3.9	3.2	1.1	4.3
CCCCCG	8.8	7.7	2.4	4.3	2.6	3.1	3.0	1.2	3.6
CGGACC	7.6	6.5	2.5	3.6	1.6	3.6	3.2	1.3	3.8
CGTCGA	6.1	3.5	2.6	3.2	0.7	4.1	2.7	0.9	3.7
CGACGA	6.1	3.2	2.2	3.9	1.4	4.8	3.2	0.4	4.2
CCCCCC	5.4	5.9	2.6	3.8	2.4	2.9	2.9	1.5	2.9

Additional Table 1b. Word over-representation ratios for non-overlapping human genomic sequence windows. Non-overlapping sequence windows of 0-1000 bp, 1000-2000 bp and 2000-5000 bp were considered for upstream, intronic and downstream regions.