

1 SUPPLEMENTARY MATERIAL

Suppl. Table 1: Correction Coefficients.

<http://www.ncbi.nlm.nih.gov/CBBresearch/Fellows/Taher/GOCC/GOCC.txt>

<http://www.ncbi.nlm.nih.gov/CBBresearch/Fellows/Taher/GOCC/readme.txt>

Suppl. Table 2. Genes annotated to the “hemophilic cell adhesion” GO category.

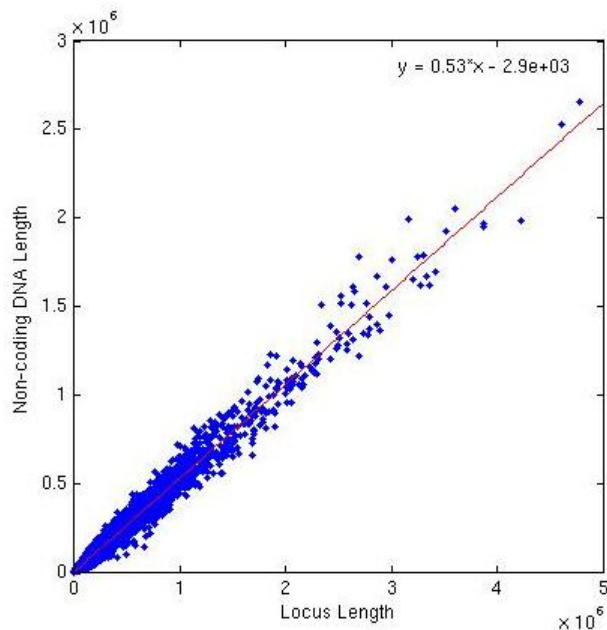
Noncoding non-repetitive (ncnr) locus length is given in the last column.

ID	Location	Gene	Locus length	Ncnr DNA length
0	chr1:9711667-9807137	CLSTN1	95471	56026
1	chr1:109594164-109619901	CELSR2	25738	25890
2	chr1:109850969-109853859	AMIGO1	2891	26302
3	chr1:157408044-157438425	IGSF4B	30382	75743
4	chr1:158782419-158815887	CD84	33469	30118
5	chr1:159541151-159546377	MPZ	5227	13536
6	chr10:33229327-33287204	ITGB1	57878	165637
7	chr10:42892523-42945803	RET	53281	165161
8	chr10:55250866-56231057	PCDH15	980192	1823029
9	chr10:72826697-73076543	CDH23	249847	243018
10	chr11:606577-615007	MUCDHL	8431	5598
11	chr11:6599134-6633650	DCHS1	34517	26400
12	chr11:91724910-92269283	FAT3	544374	1412575
13	chr11:114549555-114880451	CADM1	330897	1259828
14	chr11:117629345-117640219	EVA1	10875	18823
15	chr11:119014018-119104645	PVRL1	90628	402459
16	chr11:124128229-124137433	ESAM	9205	11153
17	chr12:7174234-7202795	CLSTN3	28562	32323
18	chr12:45755757-45759915	AMIGO2	4159	174663
19	chr12:54364623-54387894	ITGA7	23272	18465
20	chr13:38159173-38359267	FREM2	200095	321073
21	chr13:52316112-52320775	PCDH8	4664	170660
22	chr13:57103790-57201066	PCDH17	97277	1334655
23	chr13:60881822-60887282	PCDH20	5461	2797661
24	chr13:65774968-66702464	PCDH9	927497	4344407
25	chr14:22586710-22594875	CDH24	8166	10066
26	chr15:71639411-71712806	NPTN	73396	70774
27	chr16:2078712-2125900	PKD1	47189	29260
28	chr16:60244736-60627537	CDH8	382802	3212471
29	chr16:63538186-63713420	CDH11	175235	1702425
30	chr16:64958064-64996189	CDH5	38126	522781
31	chr16:65499527-65510267	CDH16	10741	20632
32	chr16:67236277-67290442	CDH3	54166	59098
33	chr16:67328696-67426945	CDH1	98250	59000

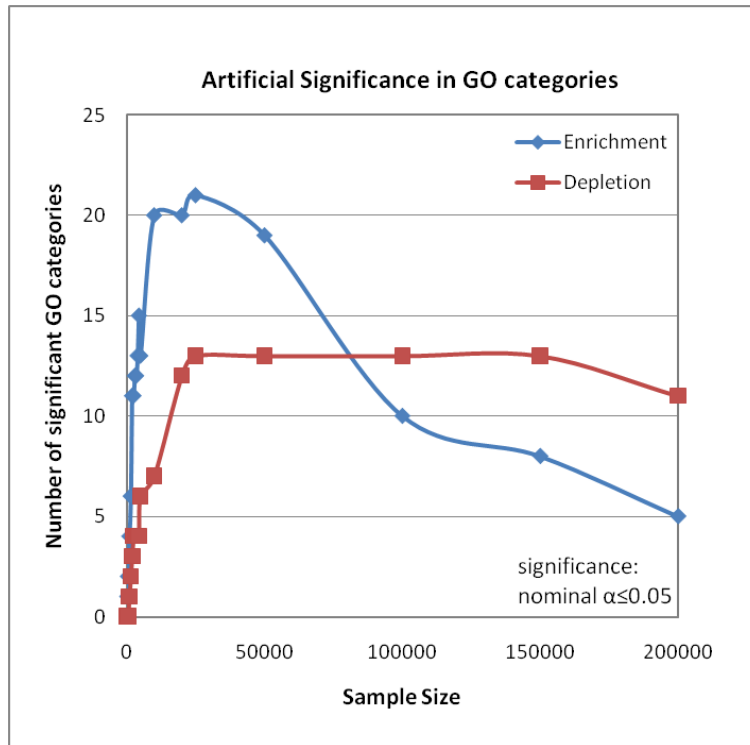
34	chr16:81218079-82387698	CDH13	1169620	940250
35	chr16:87765664-87789400	CDH15	23737	38574
36	chr18:23784934-24011189	CDH2	226256	1957946
37	chr18:26824050-26876779	DSC3	52730	1428069
38	chr18:26900005-26936375	DSC2	36371	53350
39	chr18:26963214-26996817	DSC1	33604	121959
40	chr18:27152050-27191391	DSG1	39342	118979
41	chr18:27210738-27247877	DSG4	37140	58377
42	chr18:27281730-27312663	DSG3	30934	54302
43	chr18:27332170-27380754	DSG2	48585	63429
44	chr18:57308755-57373345	CDH20	64591	790671
45	chr18:61568468-61699154	CDH7	130687	1381750
46	chr18:62322301-62422196	CDH19	99896	852894
47	chr19:47703299-47724479	CEACAM1	21181	52603
48	chr19:50041233-50084325	PVRL2	43093	25407
49	chr20:44235783-44313741	CDH22	77959	118820
50	chr20:57966877-58021562	CDH26	54686	55061
51	chr20:59260954-59945694	CDH4	684741	1185443
52	chr22:45135395-45311731	CELSR1	176337	161794
53	chr3:48638160-48647930	SLC26A6	9771	9617
54	chr3:48648900-48675352	CELSR3	26453	20500
55	chr3:77171984-77779351	ROBO2	607368	2301405
56	chr3:78729080-79721751	ROBO1	992672	2156073
57	chr3:112273555-112335752	PVRL3	62198	967856
58	chr3:139856923-139960875	PIK3CB	103953	119053
59	chr3:141136897-141769320	CLSTN2	632424	569893
60	chr4:30331135-30757519	PCDH7	426385	4812393
61	chr4:126457017-126632372	FAT4	175356	1511696
62	chr4:134289920-134332181	PCDH10	42262	4091750
63	chr4:138659522-138673079	PCDH18	13558	2493606
64	chr4:155375139-155531899	DCHS2	156761	398724
65	chr4:187745932-187881981	FAT	136050	800924
66	chr5:19508914-20017046	CDH18	508133	2257754
67	chr5:21786911-22889488	CDH12	1102578	1846474
68	chr5:24522967-24680668	CDH10	157702	1543437
69	chr5:26916466-27074446	CDH9	157981	3140504
70	chr5:31229553-31360994	CDH6	131442	2157439
71	chr5:140146060-140372113	PCDHA1,PCDHA10,PCDHA11,PCDH A12,PCDHA13,PCDHA2,PCDHA3,PC DHA4,PCDHA5,PCDHA6,PCDHA7,P CDHA8,PCDHA9,PCDHAC1,PCDHA C2	226054	131487
72	chr5:140411163-140413696	PCDHB1	2534	40917
73	chr5:140454421-140457146	PCDHB2	2726	25066
74	chr5:140460418-140463589	PCDHB3	3172	11990
75	chr5:140481765-140485385	PCDHB4	3621	11457
76	chr5:140494984-140497887	PCDHB5	2904	6632
77	chr5:140510023-140513050	PCDHB6	3028	13580
78	chr5:140532427-140536140	PCDHB7	3714	10798
79	chr5:140537614-140540203	PCDHB8	2590	1574
80	chr5:140541164-140545979	PCDHB16	4816	3665

81	chr5:140547077-140551295	PCDHB9	4219	3149
82	chr5:140552136-140555397	PCDHB10	3262	3291
83	chr5:140559532-140562801	PCDHB11	3270	4013
84	chr5:140568475-140571880	PCDHB12	3406	4275
85	chr5:140573693-140577177	PCDHB13	3485	4956
86	chr5:140583262-140586042	PCDHB14	2781	16195
87	chr5:140605331-140607983	PCDHB15	2653	30103
88	chr5:140690436-140872730	PCDHGA1,PCDHGA10,PCDHGA11,PCDHGA12,PCDHGA2,PCDHGA3,PCDHGA4,PCDHGA5,PCDHGA6,PCDHGA7,PCDHGA8,PCDHGA9,PCDHGB1,PCDHGB2,PCDHGB3,PCDHGB4,PCDHGB5,PCDHGB6,PCDHGB7,PCDHGC3,PCDHGC4,PCDHGC5	182295	82895
89	chr5:141212867-141238128	PCDH1	25262	142071
90	chr5:141304715-141318811	PCDH12	14097	16590
91	chr5:150863847-150928698	FAT2	64852	77084
92	chr5:175908971-175955375	PCLKC	46405	26765
93	chr7:105390921-105461454	FLJ23834	70534	116659
94	chr8:95208568-95289986	CDH17	81419	154833

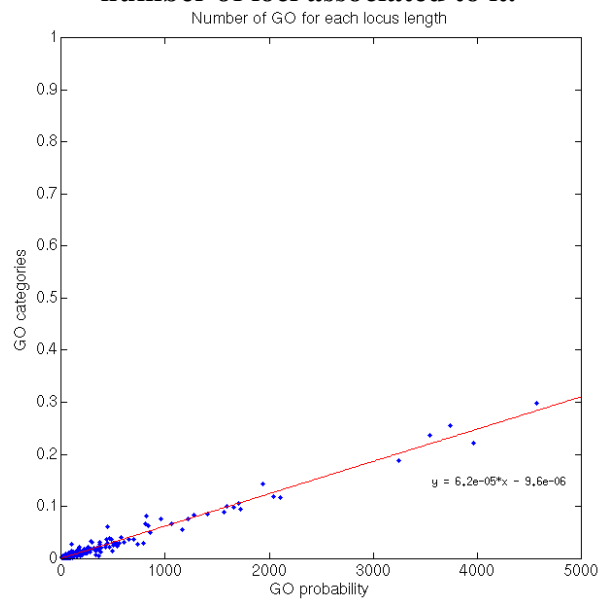
Supp. Fig. 1. Length of the locus in the human genome compared to the length of the noncoding DNA it contains.



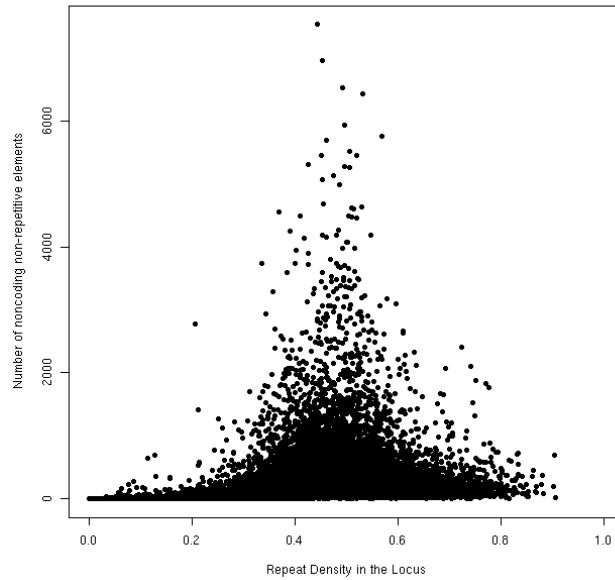
Suppl. Fig. 2. The average number of GO categories that show up as significantly over- or under-represented in experiments with random sets of noncoding elements for different sample sizes. Locus boundaries were defined using TSS locations.



Suppl. Fig. 3. Real probability \hat{p}_{GO} in the genome for each GO category, as a function of the number of loci associated to it.



Suppl. Fig. 4. Number of noncoding non-repetitive elements as a function of the repeat density of the locus.



Supp. Fig. 5. The function that samples noncoding DNA from the human genome (ncnr stands for noncoding non-repetitive elements).

Input: *ncnr_distribution* (ncnr elements probabilities, according to their length), sample size *n*

Output: sample of *n* ncnr elements

```

foreach Sample point i do
  u1 = rand();
  total = 0;
  index = 1;
  foreach ncnr element j do
    total = total + ncnr_distribution(j);
    if u1 < total then
      index = j;
      break;
    end
  end
  sample(i) = index - 1;
end

```