What and where are the DNA segments that are both low in GC content and have changed little in their GC content between human and mouse?
June 09, 2002

Identification of the lowGC_noDeltaGC segments

A distinctive component of the human and mouse genomes is identified by a low GC content coupled with little change in GC content between the two species. Most segments of chromosomal DNA show a strong correlation between the GC content in human and the change in GC content between human and mouse. A graph of these two variables shows a linear, positive relationship (Fig. 1), meaning that the GC-rich regions tend to change such that they increase GC content in human, and the GC-poor regions tend to change such that they decrease GC content in human. This fits with the broader distribution of GC content in human and in mouse.

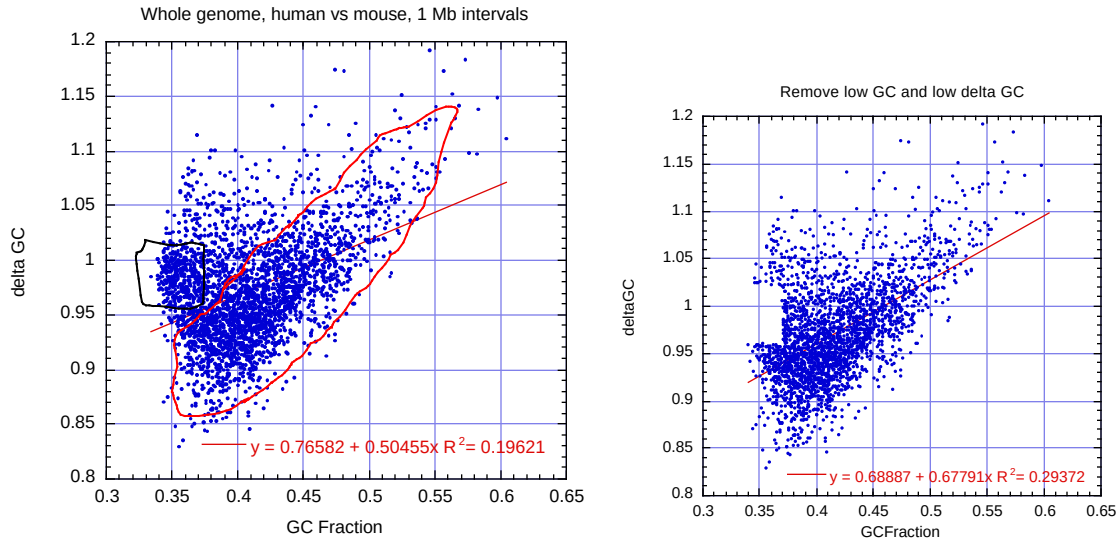

Fig. 1. Change in GC correlates with GC content for most of the genome. The left graph is for all the 1 Mb genomic DNA segments. The data cloud for the lowGC_noDeltaGC segments is outlined in black, and that for the linGC_deltaGC segments is outlined in red. The graph on the right shows the effect of removing the lowGC_noDeltaGC segments, leaving the linGC_deltaGC segments.

However, a subset of the segments segments (outlined in black in Fig. 1) change very little, despite their low GC content. To see if this portion of the genome has other properties that distinguish it from the rest, the 1 Mb segments that are low in GC content and change little in GC (lowGC_noDeltaGC segments) were extracted from the dataset. The remaining segments are referred as “linear” for the deltaGC vs GC correlation (linGC_deltaGC segments). This extraction was done by sorting the genomic DNA segments by GC content and taking all the ones between 0.33 and 0.37 for GCFraction, and then this subset was sorted by delta GC, and the sub-subset between 0.96 and 1.02 deltaGC was collected as the 306 segments in the lowGC_noDeltaGC set. All the rest of the low GC segments were rejoined with the other genomic DNA segments. The graph of

deltaGC vs GC fraction after removing the lowGC_noDeltaGC segments is shown on the right in Fig. 1. (This extraction operation could be done as a formal clustering analysis on the data. I did this crude extraction to see if there was anything to this subset right away.)

The overall distributions of several parameters are distinctive for the two sets of genomic DNA fragments (Table 1). The lower %id and aln_nr in lowGC_noDeltaGC are actually rather striking (see Fig. 2). In addition, the density of exons is lower in lowGC_noDeltaGC. Repeats show some distinctive patterns, e.g. higher L1 density and lower AluY density in the lowGC_noDeltaGC subset. The fact that other genomic parameters differ for these two sets of genomic DNA segments indicates that their evolutionary history has differed not just in the forces related to changing GC. It also raises the possibility that these sets may differ in their functional properties.

Table 1. A. LowGC_noDeltaGC segments: statistics on genomic parameters and alignments

Variable	Minimum	Maximum	Sum	Points	Mean	Median	StDev	Std Error
%id	0.6646	0.7304	210.4943	306	0.6879	0.6858	0.0113	0.0006
aln_tot	0.0288	0.5706	66.4529	306	0.2172	0.2067	0.0905	0.0052
aln_nr	0.0902	0.7063	109.0868	306	0.3565	0.3551	0.1146	0.0066
inteRep	0.2446	0.7703	142.1346	306	0.4645	0.4673	0.0675	0.0039
AluY	0.0021	0.0250	2.8301	306	0.0092	0.0088	0.0027	0.0002
AluOther	0.0135	0.0817	11.5629	306	0.0378	0.0363	0.0090	0.0005
L1	0.0701	0.5098	70.6961	306	0.2310	0.2292	0.0610	0.0035
LTR	0.0299	0.2384	34.3598	306	0.1123	0.1105	0.0305	0.0017
GCFraction	0.3345	0.3700	108.7341	306	0.3553	0.3549	0.0082	0.0005
deltaGC	0.9604	1.0179	301.8662	306	0.9865	0.9853	0.0146	0.0008
reXnRate	0.0000	4.5305	246.6363	306	0.8060	0.5296	0.8740	0.0500
exon	0.0000	0.0141	0.4455	306	0.0015	0.0000	0.0026	0.0002
snpNIH	0.0000	0.0012	0.0771	306	0.0003	0.0002	0.0002	0.0000
snpTsc	0.0001	0.0009	0.1436	306	0.0005	0.0005	0.0001	0.0000

Table 1. B. LinGC_deltaGC segments: statistics on genomic parameters and alignments

Variable	Minimum	Maximum	Sum	Points	Mean	Median	Stdev	Std Error
%id	0.6281	0.7564	1834.2897	2627	0.6982	0.6971	0.0157	0.0003
aln_tot	0.0072	0.6341	756.2407	2627	0.2879	0.2875	0.0963	0.0019
aln_nr	0.0114	0.7577	1185.7228	2627	0.4514	0.4613	0.1143	0.0022
inteRep	0.2019	0.8656	1197.9185	2627	0.4560	0.4547	0.0707	0.0014
AluY	0.0014	0.0617	36.8468	2627	0.0140	0.0115	0.0081	0.0002
AluOther	0.0156	0.3891	266.0337	2627	0.1013	0.0825	0.0647	0.0013
L1	0.0213	0.7149	437.4228	2627	0.1665	0.1582	0.0725	0.0014
LTR	0.0000	0.2584	213.4509	2627	0.0813	0.0783	0.0297	0.0006
GCFraction	0.3400	0.6046	1092.9950	2627	0.4161	0.4086	0.0415	0.0008
deltaGC	0.8290	1.1917	2550.6223	2627	0.9709	0.9638	0.0519	0.0010
reXn	0.0000	13.3126	3267.7112	2627	1.2439	0.9322	1.2668	0.0247
exon	0.0000	0.1762	31.4072	2627	0.0120	0.0086	0.0126	0.0002
snpNIH	0.0000	0.0035	0.9723	2627	0.0004	0.0003	0.0003	0.0000
snpTsc	0.0000	0.0012	1.0148	2627	0.0004	0.0004	0.0002	0.0000

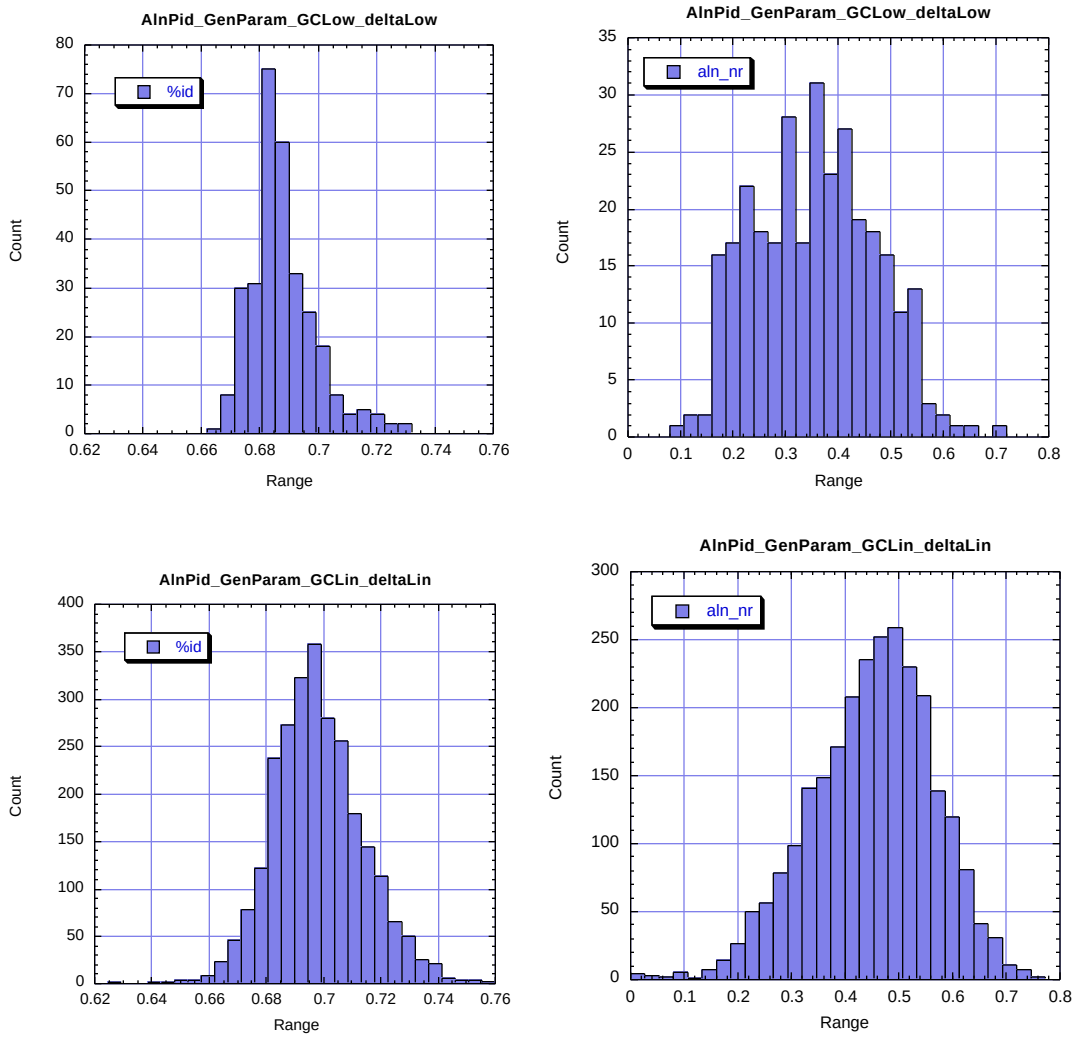

Fig. 2. Comparison of distributions of scores for percent identity in alignments (left 2 graphs) and coverage of nonrepetitive human sequence by alignments (aln_NR, right 2 graphs) for lowGC_noDeltaGC segments (top 2 graphs) and linGC_deltaGC segments (bottom 2 graphs).

It is important to realize that it is the lack of change in GC that distinguishes lowGC_noDeltaGC from the linGC_deltaGC set. The latter set has many low GC segments but these have changed by lowering their GC content between mouse and human. Thus some features such as the enrichment of the lowGC_noDeltaGC set in L1 repeats need more investigation. Is this necessarily a trivial expectation given the enrichment of L1's in DNA with low GC content? It would be interesting to look at the low GC subset of the linGC_deltaGC set to see if it differs from the lowGC_noDeltaGC set for L1 content.

Location of the lowGC_noDeltaGC genomic DNA segments

To ascertain whether the lowGC_noDeltaGC segments clustered in distinctive regions, the data were sorted by chromosome. The result indicates that the 1 Mb lowGC_noDeltaGC are not randomly distributed, since none are found on chromosomes 19, 20 or 22, but some are on chr 21. Only one 1 Mb lowGC_noDeltaGC segment is on chromosomes 15 and 17, two are on chr 12, and only three are on chromosomes 10 and 16.

The segments were then sorted by location on each chromosome, and this showed that they are often contiguous for intervals as large as 5 or even 10 Mb. Examination of these intervals in the Genome Browser showed several examples where they comprise a Giemsa dark cytogenetic band. The fact that they are in Giemsa dark bands is expected from the low GC content, but it is striking that they often comprise the entire band. Adjacent linGC_deltaGC segments are in different bands. Thus two independent methods (cytogenetic staining and comparative genomic DNA analysis) has come up with similar segmentation patterns.

The lowGC_noDeltaGC intervals invariably have no or very few RefSeq genes, and they tend to have very sparse gene predictions, e.g. by Acembly and Twinscan. This fits with the lower density of exons seen in this fraction overall.

Some of the intervals are adjacent to centromeres.

Table 2. Correspondence of lowGC_noDeltaGC intervals with Giemsa dark bands with few genes.

chrom	chromStart	chromEnd	band	bandStart-bandEnd	genes	interval
chr1	107000000	113000000	1p21.1	107.8-113.4Mb	3	6Mb
chr1	189000000	193000000	1q13.1	187.6-192.9Mb	0	4Mb
chr1	196000000	201000000	1q13.3	196.3-202.6Mb	4	5Mb
chr2	82000000	85000000	2p12	76.6-84.4Mb	0	3Mb
chr2	137000000	140000000	2q22.1&2q 22.2*		1	3Mb
chr2	142000000	144000000	2q22.3	140.2-143.1Mb	0	2Mb
chr2	152000000	155000000	2q24.1	151.6-157.3Mb	2	2Mb
chr2	157000000	158000000	2q24.1	151.6-157.3Mb	2	2Mb
chr3	80000000	87000000	3p12.1&3p 11.2	80.3-84.5&84.5-87.5Mb	1	8Mb
chr3	98000000	100000000	3q11.2	93.1-100.7Mb	0	2Mb
chr3	170000000	175000000	3q26.1	167.6-175.4Mb	2	5 Mb
chr4	30000000	34000000	4p15.1	30.3-37.3Mb	2	8Mb
chr4	35000000	38000000				
chr4	45000000	48000000	4p13	43.3-47.6Mb	1	3Mb
chr4	59000000	63000000	4q13.1	60.5-67.2Mb	1	10Mb
chr4	64000000	69000000				
chr4	131000000	133000000	4q28.2&4q 28.3		0	2Mb
chr4	164000000	166000000	4q33	163.6-166Mb	0	2Mb
chr4	166000000	169000000	4q34.1	166-175.8Mb	0	3Mb

chr4	170000000	171000000	4q34.1	166-175.8Mb	3	1Mb
chr4	172000000	174000000	4q34.1	166-175.8Mb	genes	2Mb
chr4	175000000	176000000	4q34.1	166-175.8Mb	genes	1Mb
chr4	176000000	178000000	4q34.2	175.8-178.4Mb	few	2Mb
chr4	180000000	182000000	4q34.4	178.4-183.5Mb	few	2Mb
chr5	18000000	22000000	5p14.3	17.6-21.8Mb	3	4Mb
chr5	22000000	23000000	5p14.2	21.8-22.2Mb	0	1Mb
chr5	23000000	25000000	5p14.1	22.2-26.3Mb	0	2Mb
chr5	84000000	89000000	5q14.3	84.5-91.0Mb	9	7Mb
chr5	90000000	91000000	5q14.3			
chr5	97000000	98000000	5q15	91.0-98.6Mb	1	1Mb
chr5	98000000	102000000	5q21.1	98.6-104.8Mb	1	4Mb

* start at breakpoint in conserved synteny

Significance

I think the significance of this finding is that we can identify a distinctive component of the genome that is evolving differently from the rest. It has not diverged to the point that it no longer aligns, and indeed it is distinctive in keeping it's GC content low in both human and mouse. It has few genes. These are properties one may expect for genomic DNA with a role in chromosome mechanics or interphase nuclear structure.