Table 1. Microbes, plasmids and phages with extreme values of length

The five longest and shortest values are shown in each case. G+C values have been rounded to one decimal place.

Genome		Length (bp)	G+C (mol%)
Bacterial genomes			
Longest bacterial genomes			
1	*Minicystis rosea* strain DSM 24000 (CP016211.1)	16 040 666	69.1
2	*Sorangium cellulosum* So0157-2 (CP003969.1)	14 782 125	72.1
3	*Nonomuraea* sp. ATCC 55076 (CP017717.1)	13 047 416	71.8
4	*Sorangium cellulosum* 'So ce 56' (AM746676.1)	13 033 779	71.4
5	*Archangium gephyra* strain DSM 2261 (CP011509.1)	12 489 432	69.4
Shortest bacterial genomes			
1	*Candidatus Nasuia deltocephalinicola* strain PUNC (CP013211.1)	112 031	16.6
2	*Candidatus Nasuia deltocephalinicola* str. NAS-ALF (CP006059.1)	112 091	17.1
3	*Candidatus Hodgkinia cicadicola* isolate TETUND1 (CP007232.1)	133 698	46.8
4	*Candidatus Tremblaya princeps* PCIT (CP002244.1)	138 927	58.8
5	*Candidatus Tremblaya princeps* PCVAL (CP002918.1)	138 931	58.8
Plasmid genomes			
Longest plasmids			
1	*Cupriavidus metallidurans* CH34 megaplasmid (CP000353.2)	2 580 084	63.6
2	*Burkholderia caribensis* MBA4 plasmid (CP012748.1)	2 555 069	62.4
3	*Rhizobium gallicum bv. gallicum* R602 plasmid pRgalR602c (CP006880.1)	2 466 951	59.4
4	*Sinorhizobium fredii* NGR234 plasmid pNGR234b (CP000874.1)	2 430 033	62.3
5	*Rhizobium gallicum* strain IE4872 plasmid pRgalIE4872d (CP017105.1)	2 388 366	59.2
Shortest plasmids			
1	*Candidatus Tremblaya phenacola* PAVE plasmid (CP003983.1)	744	42.2
2	*Lactococcus lactis subsp. lactis* KLDS 4.0325 plasmid 2 (CP007042.1)	870	32.6
3	*Enterococcus faecium strain* ISMMS_VRE_1 plasmid ISMMS_VRE_p5 (CP012433.1)	886	31.3
4	*Borreliella garinii* strain CIP 103362 plasmid cp32 (CP018755.1)	1 085	30.4
5	*Acinetobacter baumannii* strain JBA13 plasmid pJBA13_2 (CP020583.1)	1 109	59.1
Phage genomes			
Longest phages			
1	*Agrobacterium* phage Atu_ph07 (MF403008.1)	490 380	37.1
2	*Salicola* phage SCTP-2 (MF360958.1)	440 001	30.0
3	*Pectobacterium* phage CBB (KU574722.1)	378 379	35.9
4	*Aureococcus anophagefferens* phage *BtV-01* (NC_024697.1)	370 920	28.7
5	*Cronobacter* phage vB_CsaM_GAP32 (JN882285.1)	358 663	35.6
Shortest phages			
1	*Leuconostoc* phage L5 (L06183.1)	2 435	33.3
2	*Enterobacteria* phage M (JX625144.1)	3 405	48.0
3	*Enterobacterio* phage KU1 (AF227250.1)	3 486	46.5
4	*Enterobacteria* phage C-1 INW-2012 (JX045649.1)	3 523	48.4
5	*Enterobacterio* phage MS2 isolate DL52 (JQ966307.1)	3 525	51.0

variation in G+C started high in short genomes and decreased as genomes became longer. In keeping with previous research [13, 14], this creates a data plot of a roughly triangular shape (Fig. 1). There is a positive correlation between genomic G+C content and bacterial genome length, though this is not a simple one: length is associated more with the range of G+C content, rather with its absolute value. As noted above, small sequences accommodate the whole range of G+C content, while as length increases, G+C values tend to occupy the upper part of the range. This is in keeping with the data in Table 1, where the five longest genome sequences all have G+C values of 69 mol% or more, whilst the shortest five examples range from 16.6 to 58.8 mol%.

Therefore, trying to fit a linear regression model onto this dataset was potentially problematic. Using heteroscedasticity-robust regression, the linear model explained only a small proportion of the variation (Pearson R=0.58, P<0.001). This is equivalent to an r^2 of 0.34 and, thus, around 66 mol% of the variation in G+C content cannot be accounted by this model. The heteroskedastic pattern could