ANNEX

Biogeography and ecological distribution of environmental microbial protein families at a global scale

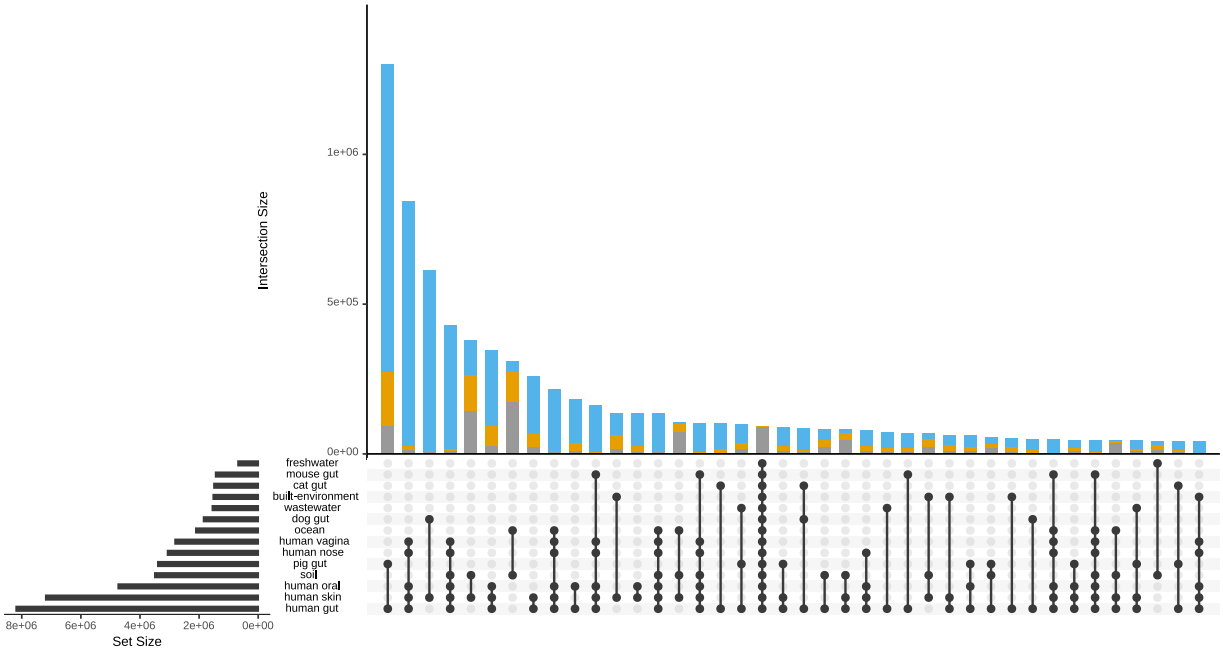

Figure 8.1: *Distribution of known function (gray), unknown function (yellow), and novel families exclusive to uncultivated taxa (blue) across the GMGC, Barcharts indicate the number of families shared by the habitats indicated below. Horizontal bars represent the number of gene families detected in each of the 14 GMGC habitats. Only families detected on more than one habitat are shown*

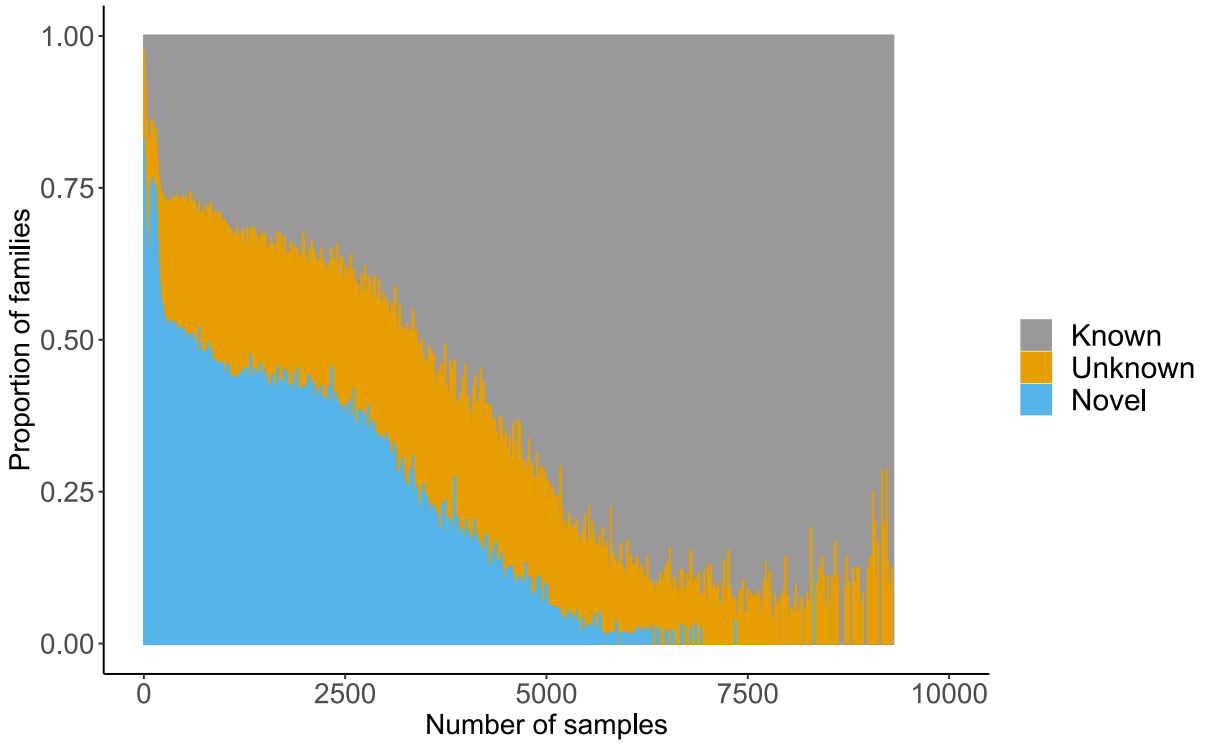

Figure 8.2: *Proportion of known function, unknown function and novel gene families detected in the number of samples indicated*

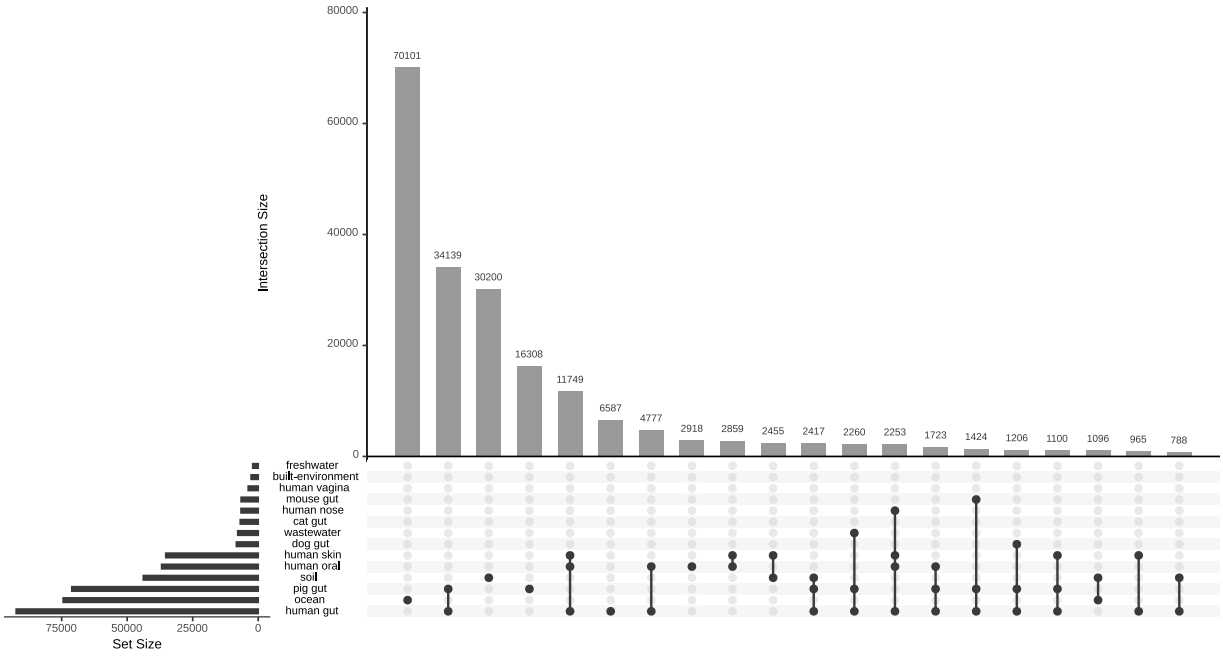

Figure 8.3: Distribution of filtered novel gene families in the GMGC habitats. Barcharts indicate the number of families shared by the habitats indicated below. Horizontal bars represent the number of curated novel gene families detected in each of the 14 GMGC habitats

Table 8.1: *Fitting of the observed gene frequency distributions and the neutral inifinite gene model at the lower-end (k = 1,...,10, the genes that we call rare)*

Habitat	**r-value**	**p-value**
all	0.999463	3.645764e-13
soil	0.999443	4.221371e-13
human gut	0.998952	5.266045e-12
cat gut	0.998720	1.171272e-11
mouse gut	0.998390	2.933164e-11
dog gut	0.997189	2.724043e-10
human oral	0.996557	6.121141e-10
built-environment	0.996032	1.079534e-09
pig gut	0.993670	6.970725e-09
human skin	0.990273	3.871034e-08
marine	0.989412	5.428386e-08
freshwater	0.967956	4.437601e-06

Table 8.2: *AMPs predicted on the GMGC high quality novel gene families*

Gene family code	Representative sequence	AMP type	AMP probability	Hemolytic / non-hemolytic	Hemolytic probability
005_046_665	KVMRLAKAKAGEKNYCCNTFTGGCCNTYSTTGC	CDP	0.683	NonHemo	0.416
006_348_558	GLIAVTAAFAIVGLATGIALLATVRRHHRRRPPRP	CLP	0.554	Hemo	0.663
065_281_253	KKNGDYKGYPNPKPAAKKAPTGTSAKRGGKPKQGTVKLRGY	CLP	0.653	Hemo	0.663
070_112_934	KTAKCGASNPASGSAKKKAKGYAYGGMPMNPKKKATMGGGGYAKKKK	CDP	0.644	Hemo	0.634
078_259_111	VTGYAVSMKKKDPKNDKLFMWGTAIFFLVFFAFIKVLIS	CLP	0.505	Hemo	0.802
080_886_431	RERRRECSSARMFYEGRRRERERERERGGGGGEGGGGMCACFKEEEGR	CDP	0.515	NonHemo	0.069
082_172_685	KCGGINQMGMKGMYKKKPGARLPKKKKPAKKGKKKPSYGY	CDP	0.564	Hemo	0.782
084_985_735	VFSLVGVKGGKTVGIAKGGKPSYKRKKKKKKKKKNGRI	CLP	0.545	Hemo	0.782
087_116_850	AKAKGKNKGVAVFIAIGMPESKLLKNLKRRKNGKSKKK	CLP	0.505	Hemo	0.901
090_504_354	VQQQRLVRVRVRVWVRVRVRVRVRVRVRGAAAAPSACSSP	CDP	0.545	Hemo	0.822
090_824_004	RVRVRVRVRVRVRVRQPPRLGIVLRHTLAVVVHVTKAVLRLCGR	CDP	0.584	Hemo	0.812
091_427_716	GKLIIKLGEAVLKLENFLKSKWNSFIKVLMFKEKGTK	CLP	0.733	Hemo	0.782
093_504_939	GVIMLIYGKTPNDYLKLAKTHKKKVVVGLIIVVAVLSFIF	CLP	0.545	Hemo	0.941
116_360_395	AKAKKAVKKVAKTVKKTVKKAEPVVAEVSRGVYSQRGK	CLP	0.535	Hemo	0.792
124_411_894	QTIKDFFWWYKHLTSQGHGYIACIDYALYNCRRFNRDGSYKKNTPGGS	CDP	0.564	NonHemo	0.307
125_919_848	KKFKKIMETIIIFLVGFVVGILVGRKHTGAVESVVGLAKKLIGKK	CLP	0.564	Hemo	0.861
126_329_857	ALKKAHKEMLKGAGVTFGIAVFMKQQFDQWVASASRMIGRG	CLP	0.564	Hemo	0.802
132_350_966	AKAKKATKKVVKKTVKKIKVEAPKVVRGEYSQRGK	CLP	0.525	Hemo	0.713
133_220_372	VLGLMKKRFGKYHDFIVLWTMFWTAISVGLILGYLLN	CLP	0.673	Hemo	0.861
147_112_069	IKFKKILNFERRKNLIEIKEKVLKYMKKNNFNGIILKKEIQSFG	CLP	0.545	Hemo	0.653
147_362_560	LILINEIIMNLVRAAIFILIAIAGIKTGKKFRDYKDAQKK	CLP	0.683	Hemo	0.822
149_780_564	NVFKLNALLSVIIKTKKLIKDIKKKIKIKVIRNEK	CLP	0.584	Hemo	0.931
154_942_319	DPRKLIKQLFKKKAKPTWPKPPKPPFPDTSFGETGPPTRH	CLP	0.525	NonHemo	0.465
157_012_981	AAFASSAPRLTKTLLPSGRVYRLERRSHKSLVRRLRSIWRWRQG	CLP	0.535	Hemo	0.911
157_228_026	ELLLALAVIAGLDVLALRFGRDSRHVADRERRRAGF	CLP	0.683	NonHemo	0.178
157_907_294	GIFVLSVVVMIVVVGGLAGWAFRAEKRRIRQAGNR	CLP	0.554	Hemo	0.802
164_032_778	GLEIIGGVVVGLALTAFVLMKILKRGRGRRGDRL	CLP	0.545	Hemo	0.673
184_185_534	KIICITAPKFLNGLIKKISGKGGKKTVEKKLVK	CDP	0.653	Hemo	0.901
188_535_092	IMAITAAVAGISTAVEAIAAGVTTVVTIYGIAKTGKKPKK	CLP	0.594	Hemo	0.861
188_808_372	RQKKAMKNKQQSNPVAAICAAATLVVIHLAVRHCMRKYGRK	CDP	0.505	Hemo	0.723
191_995_198	VIGGEIMERIYKPTNSGTQKVAGQPKKSSTVKVVKSTGKGK	CLP	0.537	Hemo	0.673
192_207_689	VKWKGVDKREKGGVKMVYIGMFVACIGIGICIKALLKLK	CDP	0.535	Hemo	0.782
195_840_607	VEIVAILVVEGKILKNIKKKLRKIWKLEKKLISLHYQNKRKD	CLP	0.693	Hemo	0.713
195_916_215	FLLKNTNIEGQIMFTITINGRVYRGTRKALKAFGNPRLTIFKKGVKK	CLP	0.614	Hemo	0.683
208_713_664	LGTGRKKKVLGYLDFFTIGIRLFKSKYFSFHFLLIKLNQGLKVIDYE	CLP	0.644	Hemo	0.871
209_699_760	SYIGYILFIVVLVVVGVRTNKKSRYYLSTPLPRKK	CLP	0.574	Hemo	0.802
223_554_421	VIVMIARQLLKLALVPLDAAGPSRSRLALRRLGLRLKPNHSF	CLP	0.505	Hemo	0.752
228_559_806	PAAKKAAAKPAAKKPAPAKKAPAKKAPAKKVARKAASPVKKSAVKK	CLP	0.540	Hemo	0.743
245_937_732	ALAALLGVFFIVMVAIVLFAFPPNKIVEWMSKKDK	CLP	0.515	Hemo	0.525
249_888_139	LWLAVAFVLGAASAIAFIAFGALDGVFNFLEKRR	CLP	0.515	NonHemo	0.485
265_947_488	VIVLYSCKKSPKKKHKKNKPVGEIKNLNTTTPGLFENMFKSKSN	CDP	0.535	Hemo	0.535
291_106_072	GKFFKGFLFGYLGMKGKKMATKKANKSHSKILANKPVKKNWGG	CLP	0.535	Hemo	0.950
293_561_511	KNIFKTVRKIIVVSGVAALGAAGYFVYKKFIKK	CLP	0.663	Hemo	0.871
306_213_191	NNFLLNKLLYIFVGKTVKRGQRRTVWNGIRLAKSLFYMF	CLP	0.535	Hemo	0.644
309_440_233	HLKGGKDMWREFKELPLKDKVDVIVKIIGSIIAIITTIKHL	CLP	0.535	Hemo	0.634

Functional and evolutionary significance of unknown genes from uncultivated taxa

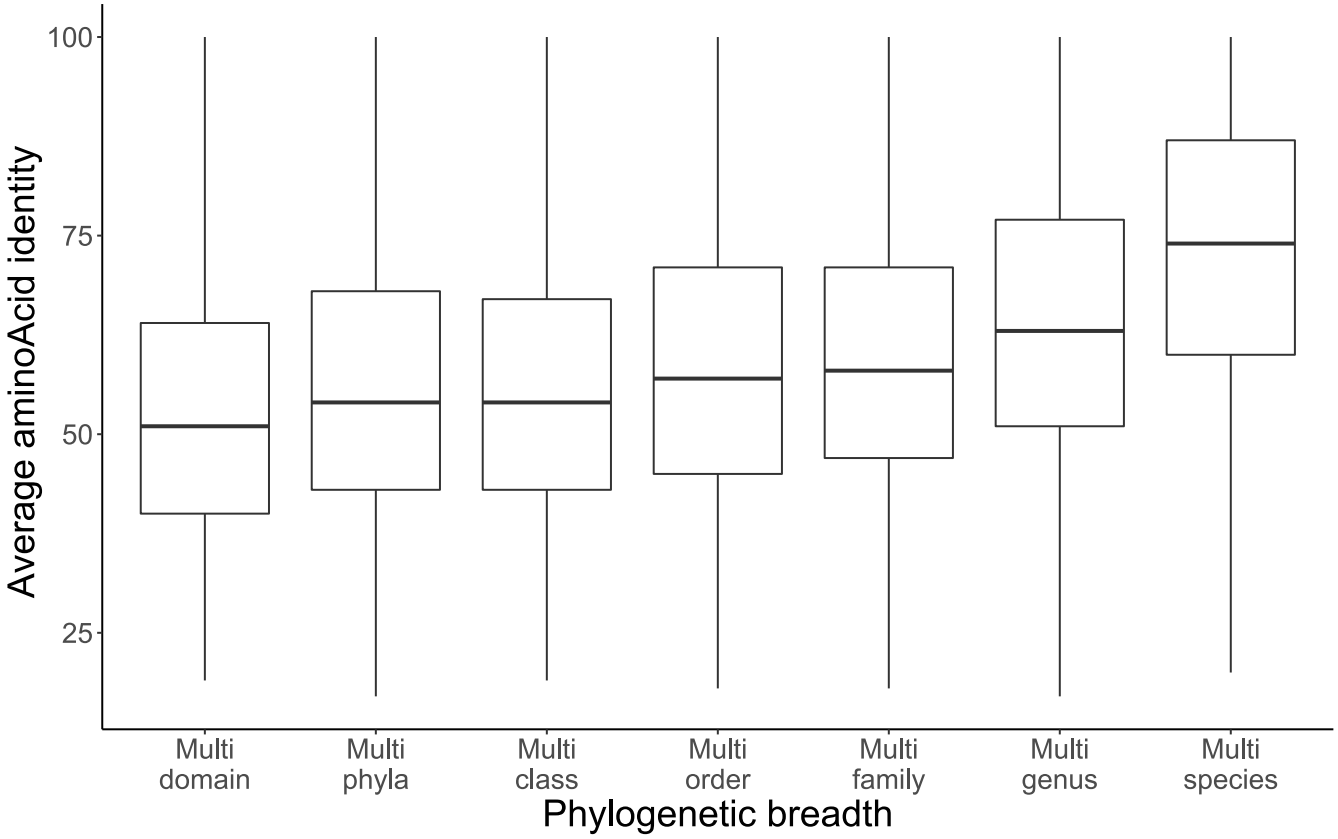

Figure 8.4: *Average amino acid identity of the families, stratified by taxonomic breadth*

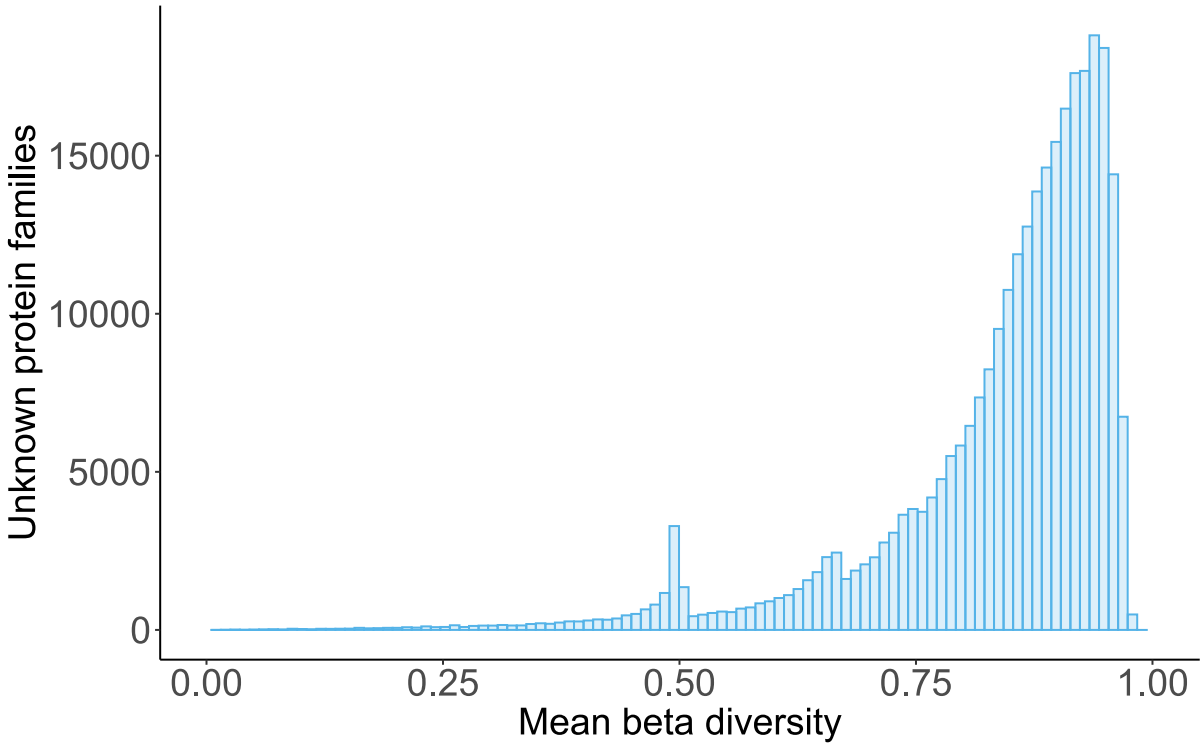

Figure 8.5: *Distribution of beta diversity values calculated for the novel gene families*

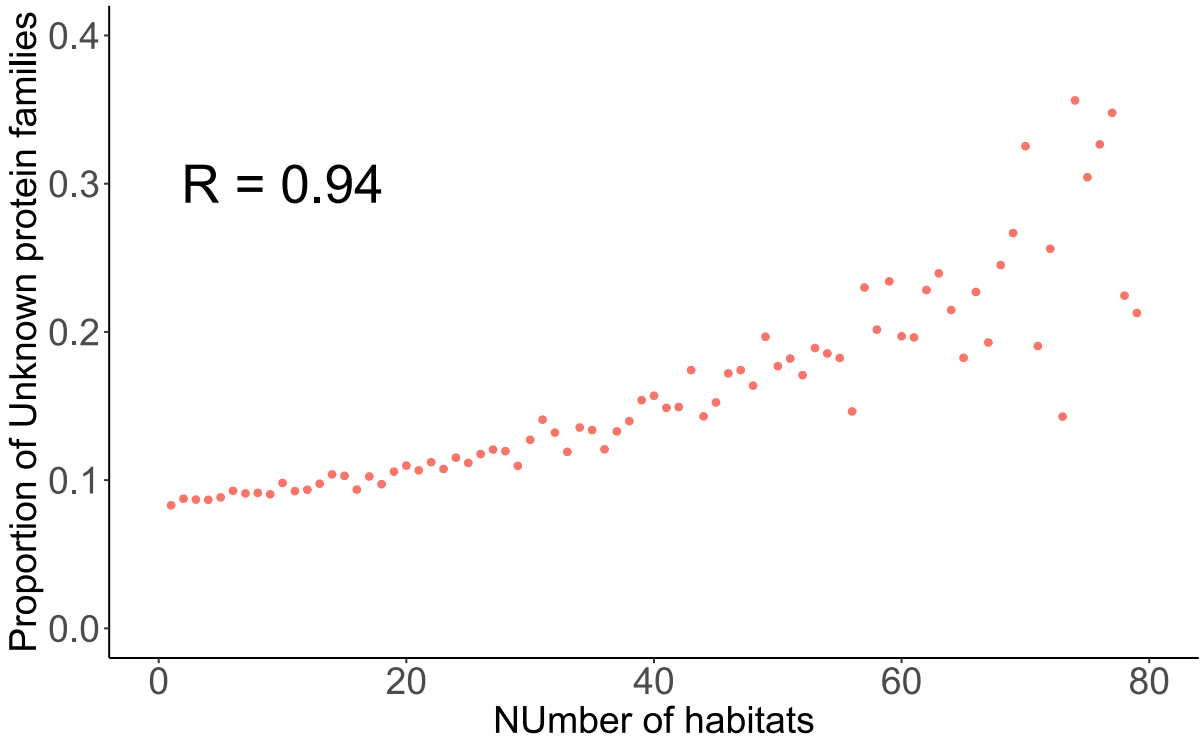

Figure 8.6: *Proportion of protein families linked to plasmids or viral contigs with relation to the number of habitats they were detected in. R was calculated as the Spearman correlation coefficient*

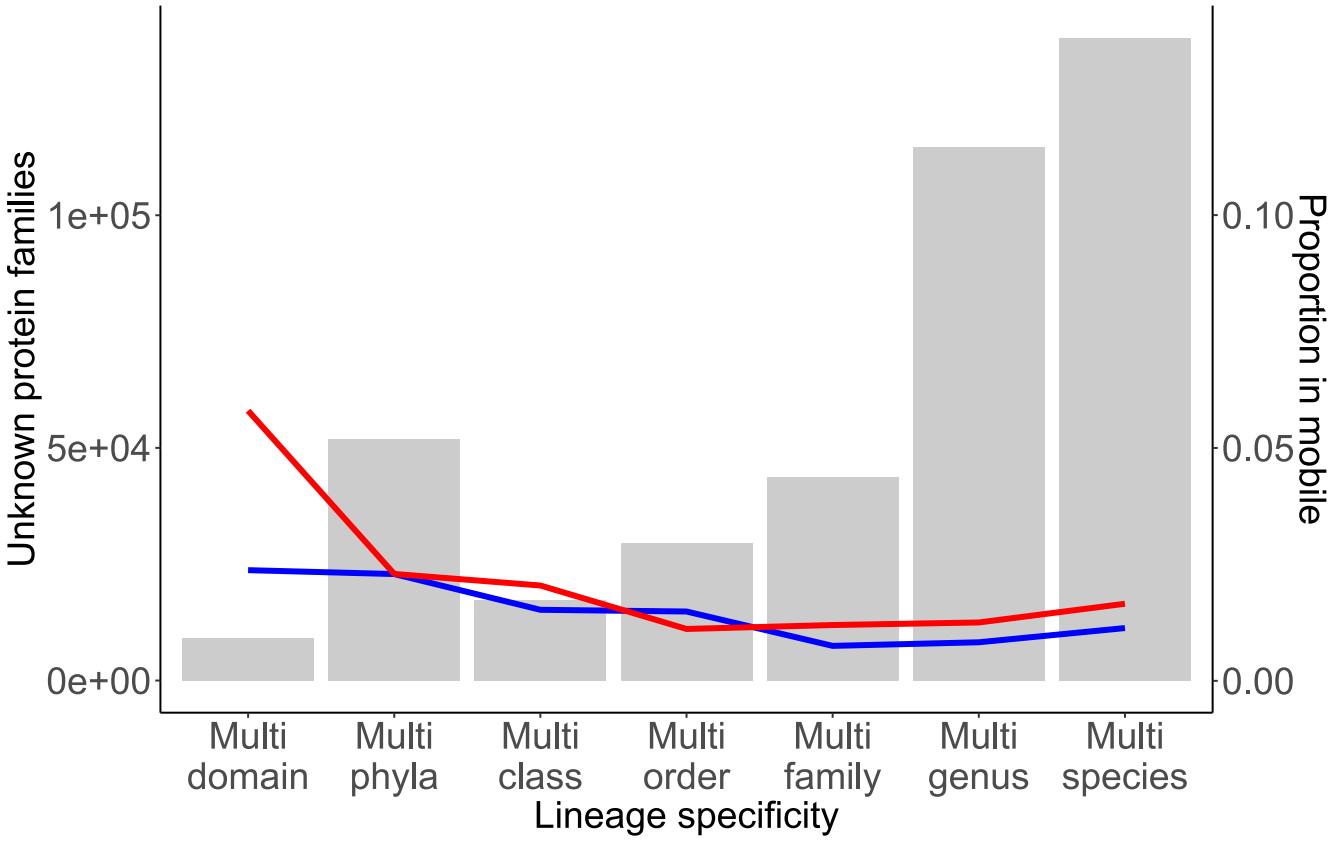

Figure 8.7: *Number of unknown protein families confined to each taxonomic rank. The term genus in the x-axis indicates the number of protein families detected in multiple species from the same genus, while the domain bar indicates families spanning more than one phylum from the same domain. The blue and red lines indicate the proportion of protein families predicted as mobile in plasmids and viral contigs respectively. Equivalent to Figure 4.15C but requiring 30% of the members to be present in mobile elements)*

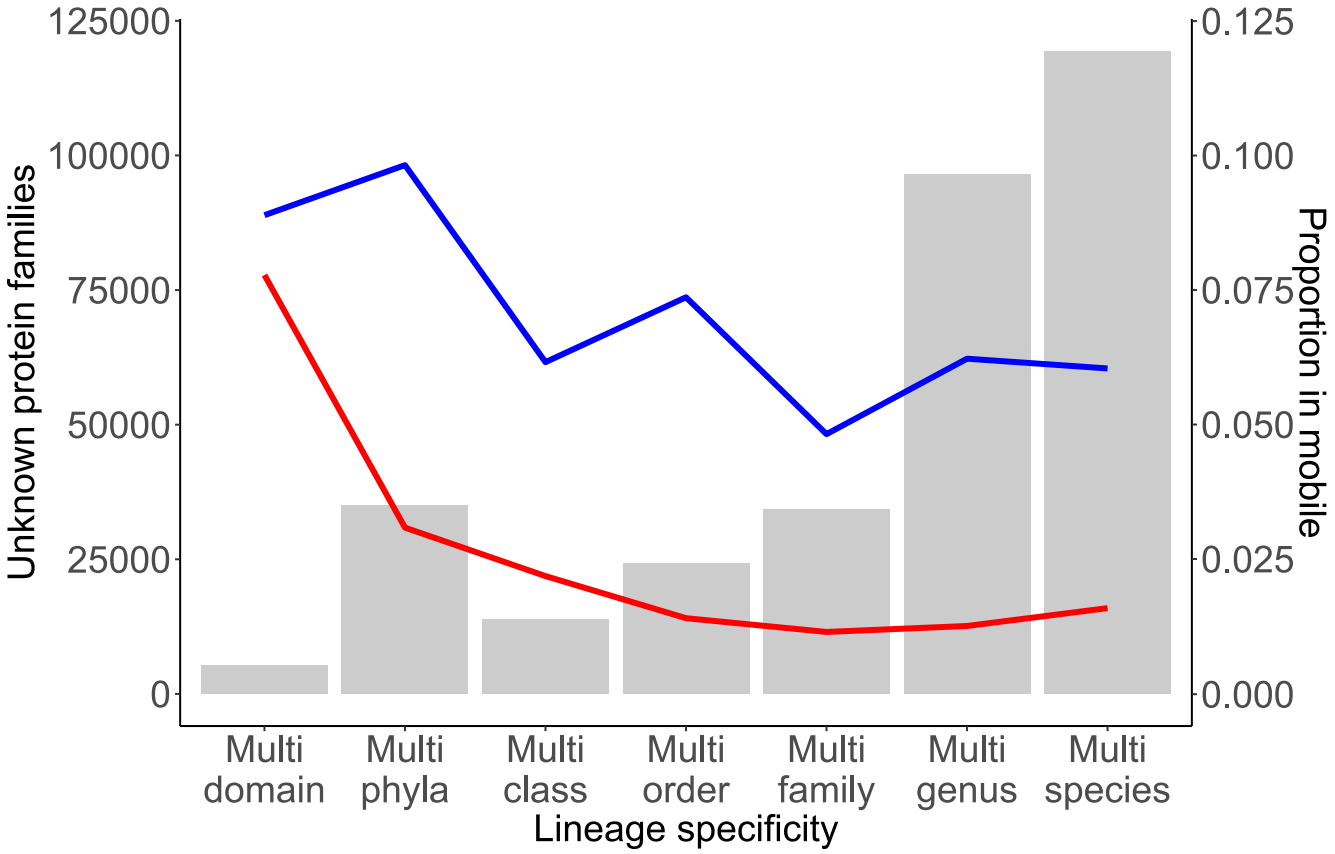

Figure 8.8: Equivalent to Fig. 3 4.15C but calculating LCAs as the most basal taxonomic group gathering 50% of the members of the family. Red: proportion of families in viral contigs. Blue: proportion of families in plasmids.

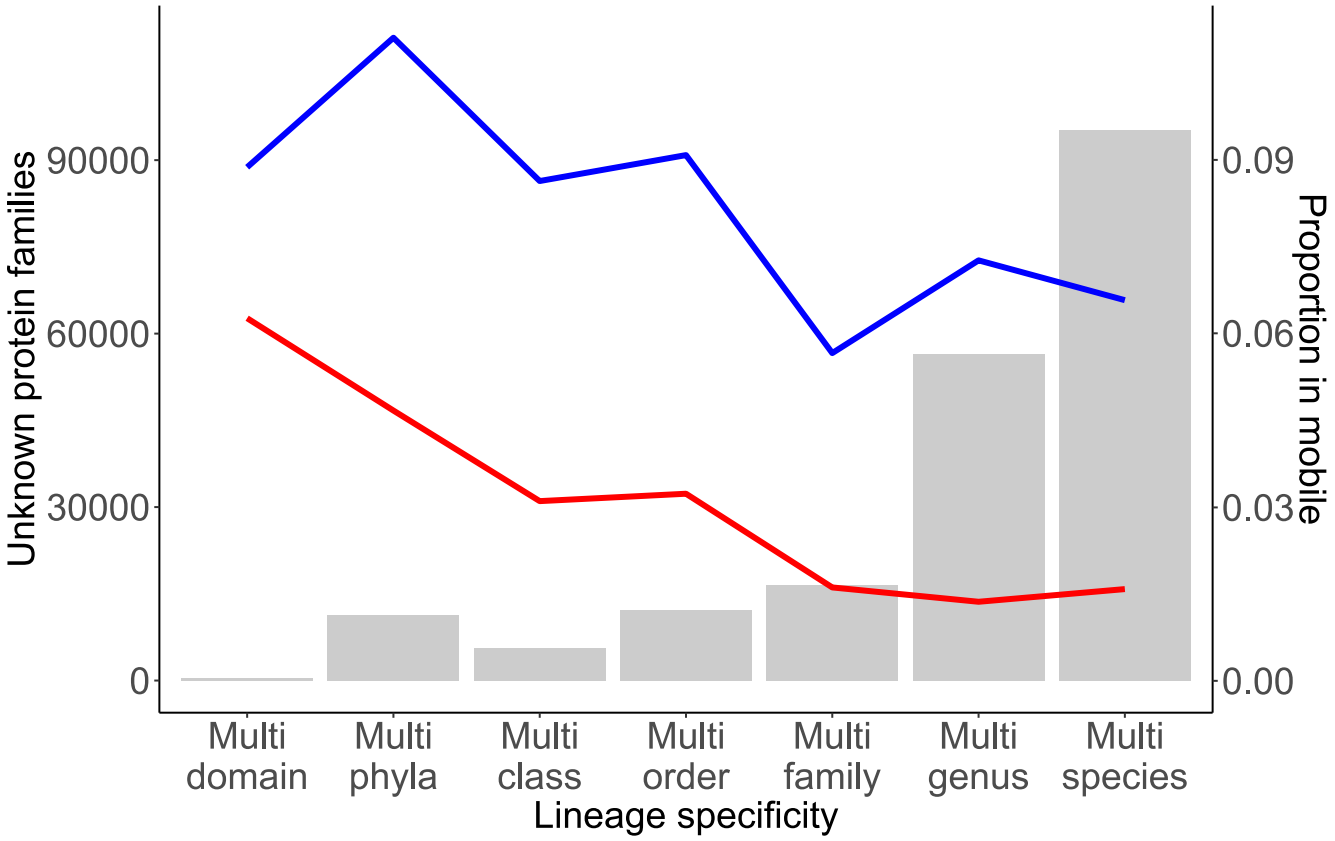

Figure 8.9: Equivalent to Fig. 3 4.15C but calculating LCAs as the most basal taxonomic group gathering 80% of the members of the family. Red: proportion of families in viral contigs. Blue: proportion of families in plasmids.

The code used for generating these results and the supplementary tables were uploaded to https://github.com/AlvaroRodriguezDelRio/NovFamilies.