1 Data

1.1 Multi-Species Genome for Pre-Training

Table 1 lists the 135 species in 7 categories for genome foundation model pre-training and presents the number of nucleotides we achieved from each species.

Category	Species	Num. of Nucleotides (M)
Fungi	Ceratobasidium	655.37
	Claviceps Maximensis	329.79
	Fusarium Annulatum	449.98
	Melampsora	699.52
	Metschnikowia	109.36
	Mucor Saturninus	391.17
	Penicillium Chermesinum	275.81
	Saccharomyces Cerevisiae	121.54
	Sporopachydermia Quercuum	155.71
	Tranzscheliella Williamsii	184.77
	Xylariales	399.96
Protozoa	Phytophthora Sojae	792.65
	Pythium Apiculatum	450.99
Mammalian	Bubalus Bubalis	28768.00
	Camelus Dromedarius	19757.02
	Human	31372.10
	Macaca Assamensis	27593.76
	Macaca Nigra	28217.13
	Mus Musculus	26545.98
	Peromyscus Californicus	24677.56
Invertebrate	Brachionus Rubens	1327.37
	Ceroptres Masudai	12.95
	Cotesia Typhae	1866.62
	Croniades Pieria	3889.85
	Drosophila Athabasca	1221.16
	Emesis Russula	4848.08
	Hydra Oligactis	12597.75
	Meganola Albula	3604.25
	Oscheius	383.21
	Rutpela Maculata	20213.33
Other Vertebrate	Anas Zonorhyncha	11697.08
	Coregonus Clupeaformis	26824.02
	Gnathonemus Longibarbis	7314.74
	Myxocyprinus Asiaticus	23407.19
	Rhipidura Dahli	10112.96
	Aeromonas	47.33
	Agrobacterium	97.22
	Alcaligenaceae Bacterium	20.88
	Aliivibrio	46.48
	Alphaproteobacteria Bacterium	14.22
	Amycolatopsis Antarctica	63.43
	Anaerostipes Faecis	32.00
	Arthrobacter	36.27
	Atopobium	28.63
	Bacillus Bc15	57.34
	Bacillus Bs3 2021	43.51
	Bacterium	7.54

(Continued on next page)

(Continued from previous page)

Category	Species	Nucleotides (M)
Bacteria	Bacteroidetes Bacterium Qs	8.99
	Breoghania Corrubedonensis	53.32
	Caldicoprobacter Oshimai	27.25
	Candidatus Cryptobacteroides Excrementipullorum	27.63
	Candidatus Dadabacteria Bacterium Rbg Combo	11.49
	Candidatus Dwaynia Gallinarum	16.82
	Candidatus Falkowbacteria Bacterium	13.88
	Candidatus Geothermincola Secundus	24.76
	Candidatus Gottesmanbacteria Bacterium	11.08
	Candidatus Nomurabacteria Bacterium Full	6.29
	Candidatus Portnoybacteria Bacterium Big Fil Rev	8.17
	Candidatus Regiella Insecticola	20.62
	Candidatus Roizmanbacteria Bacterium Combo All	11.13
	Candidatus Rokubacteria Bacterium	22.06
	Candidatus Saccharibacteria Bacterium	6.55
	Candidatus Staskawiczbacteria Bacterium Full	6.79
	Christensenella	18.75
	Clostridiaceae Bacterium	29.62
	Clostridiales Bacterium	16.59
	Clostridium Cag 505	21.26
	Clostridium Mcc328	36.43
	Clostridium Nexile	38.43
	Clostridium Uba3521	25.99
	Collinsella Urealyticum	19.45
	Coprobacillus Cateniformis	38.38
	Cyanobium	40.33
	Dehalococcoidia Bacterium	17.59
	Enterobacteriaceae Bacterium	41.46
	Evtepia Gabavorous	24.94
	Firmicutes Bacterium	36.66
	Fulvivirga	65.24
	Jeongeupia Chitinilytica	39.11
	Legionella Endosymbiont Of Polyplax Serrata	5.30
	Listeria Ilorinensis	30.31
	Maribacter Cobaltidurans	46.40
	Marinomonas	37.73
	Mesorhizobium	65.15
	Methyloceanibacter Caenitepidi	34.25
	Microvirga	68.63
	Mycolicibacter Engbaekii	45.21
	Novosphingobium	46.18
	Omnitrophica Wor Bacterium Rbg	12.52
	Pantoea	43.14
	Paraburkholderia Edwinii	82.99
	Parerythrobacter Lutipelagi	30.98
	Paulownia Witches Phytoplasma	8.92
	Polaromonas Eurypsychrophila	41.61
	Prevotella Ag 487 50 53	29.63
	Prevotella Uba3619	31.72
	Prevotella Uba634	18.51
	Prochlorococcus Ag-321-I09	3.29
	Prochlorococcus Ag-363-B18	15.54
	Prochlorococcus Ag-402-L19	11.17
	Prochlorococcus Scb243 498N4	14.12
	Providencia	41.89

(Continued on next page)

(Continued from previous page)

Category	**Species**	**Nucleotides (M)**
Bacteria	Pseudomonas 35 E 8	63.56
	Pseudomonas Bigb0408	59.52
	Pseudomonas P867	62.01
	Pseudomonas Promysalinigenes	50.47
	Roseobacter	44.14
	Salinicola Peritrichatus	46.19
	Salmonella S096 02912	48.09
	Salmonella Zj-F75	47.87
	Sinorhizobium	65.53
	Sodalis Ligni	63.85
	Sphaerochaeta	28.61
	Sphingobacterium	36.55
	Sphingomonas Carotinifaciens	37.53
	Sphingomonas Mesophila	22.91
	Sporosarcina Jiandibaonis	36.30
	Sporosarcina Ureilytica	34.37
	Staphylococcus Gdq20D1P	28.50
	Staphylococcus M0911	24.38
	Streptococcus	22.18
	Streptomyces 8401	88.39
	Streptomyces Di166	88.71
	Streptomyces Durbertensis	59.24
	Streptomyces Neau-Yj-81	118.84
	Streptomyces Rk74B	87.36
	Thermopetrobacter	26.06
	Uncultured Kushneria	35.31
	Uncultured Phascolarctobacterium	17.95
	Uncultured Proteus	35.66
	Verrucomicrobiales Bacterium	3.15
	Vibrio	41.47
	Victivallis Lenta	55.45
	Virgibacillus Salexigens	44.18
	Xanthomonadales Bacterium	37.47

Table 1: Details statistics of the multi-species genome dataset for pre-training.