The *Bacillus* Pangenome And the Answers Hidden Within

Ryan Yang[1], Noha Youssef[1]*, Wouter Hoff[1]**

Abstract. *Objectives*: We've been taught since we're young that bacteria are everywhere but are they really everywhere? To address this question, we created *Bacillus* pangenomes. Analysis of the pangenomes allowed us to answer questions such as whether biogeography affected the pangenome and its structure. *Material & Methods:* In this study, we relied heavily on high performance computing to generate the necessary data. Genomes were retrieved from NCIB and pangenomes were created with the micropan package for R, a software for statistical computing on Oklahoma State University's "Pete" compute cluster. Micropan and FigTree were used to create the blast distance and 16s rRNA phylogenetic trees, respectively. The calculated genomic differenced allowed us to compare how the 16s rRNA tree differed from the full genome tree. Principal Component Analysis (PCA) plots were also constructed to show the relationship between species in different environments and regions. *Results:* Our data indicated the pangenome size to differ based on environment and region. Heaps analysis showed the pangenomes to be open with an alpha value much lower than one independent from the number of genomes included in the pangenome. *Conclusion:* There is still much work that needed to be done but our preliminary results suggest that species within a genus tend to cluster together regardless of external factors and that the *Bacillus* has an open pangenome.

Introduction

The genus *Bacillus* is capable of producing spores that can be picked up by wind and can be found everywhere on Earth. Our goal here is to determine whether *Bacillus* genomes from different geographical locations and different environments differed as an adaptation or whether they remained relatively indifferent. As the cost of sequencing have dramatically decreased over the years, more and more genomes have been uploaded, enabling us to retrieve a large amount of *Bacillus* genomes from NCBI to carry out the pangenome analysis, yielding insights into evolutionary history of *Bacillus*. Pangenomic analysis would yield core and accessory genomes, or in other word would inform us about the set of genes that are common to all genomes (core genes) and these that are present in one or two genomes but not the rest (accessory genes). Pangenome is the sum of the core and accessory genes. As the size of the core genome increases, the size of the accessory genome should decrease leading to a smaller (closed) pangenome. On the other hand, open pangenomes have a small core size and a very large accessory genome size. We asked the question whether the *Bacillus* pangenome is open or closed, and whether *Bacillus* genomes coming from the same geographic location or from the same environment would have a more similar genome than these from a different location or environment. Or is the pangenome dependent on phylogeny, or in other words would *Bacillus* genomes from the same species be more similar regardless of their geographic location or environment.

[1]Oklahoma State University, Department of Microbiology & Molecular Genetics
*Thesis director, advisor
**Second reader, advisor

Methods & Materials

We began by retrieving as many *Bacillus* genomes as possible from NCBI and recorded all relevant information into a spreadsheet. These genomes originated from *Bacillus* around the globe and from a variety of different environments so genomes without location and environmental data were omitted and the remaining were assigned a genome ID (GID). Based on the country of origin, the genomes were grouped into the respective regions (see Appendix A). Regions below the threshold of 15 genomes were omitted from the study. The same genomes were also grouped together based on which environments they originated from, resulting in a total of 11 subgroups (Table 1).

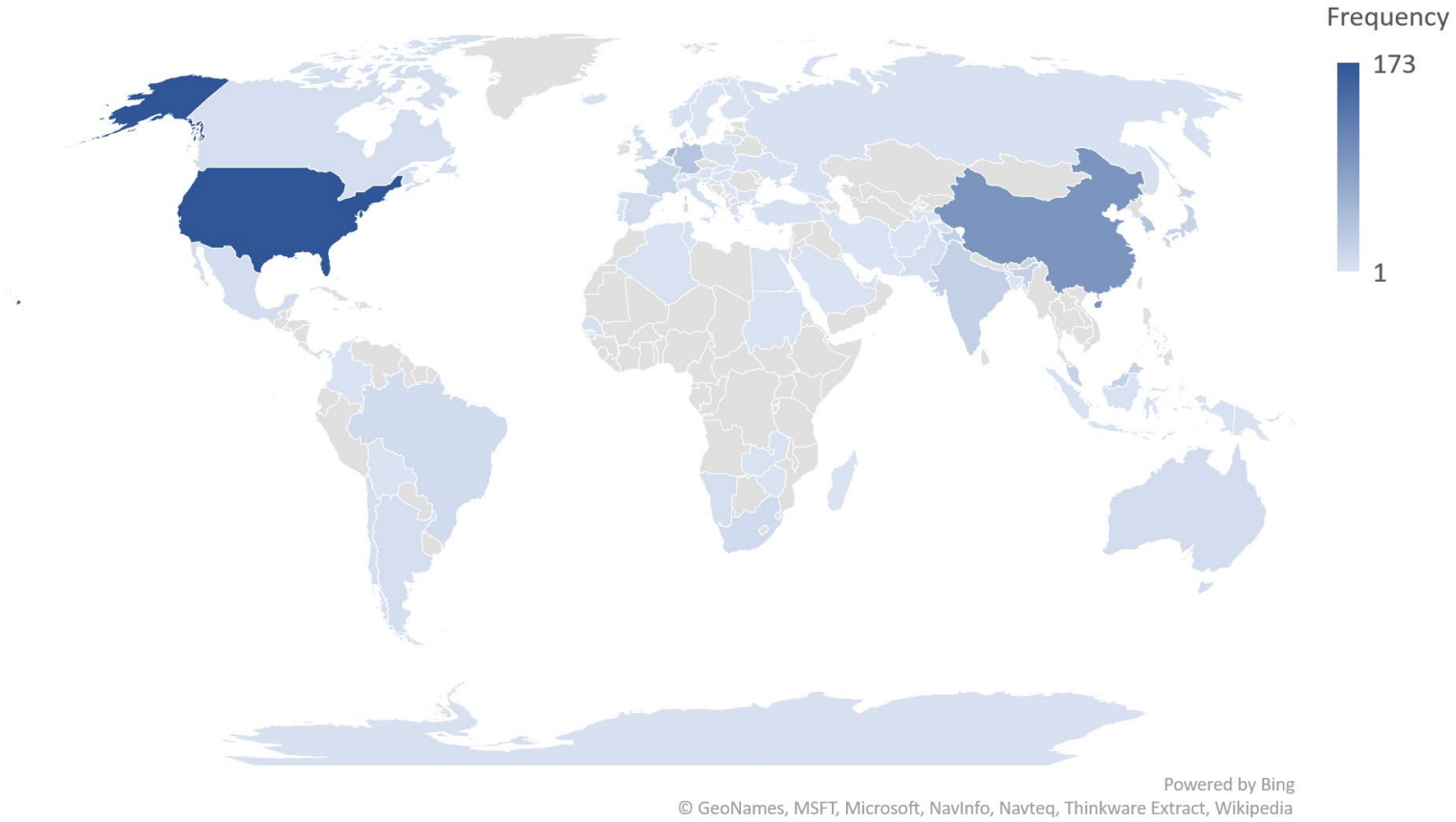


Figure 1: A world view of where the Bacillus genomes originated.

Region		Environment	
Asia	181	Engineered	32
Africa	24	Food	46
Latin America & Caribbean	26	Freshwater	32
North America	89	Host-Associated	131
Western Europe	109	Marine	18
		Terrestrial	166

TABLE 1: Regional and environmental subgroups along with how many genomes were included in each. Genomes are not mutually exclusive**.**

Pangenome creation was done via the micropan package for R following the authors' recommended pipeline. Genes were predicted with Prodigal and protein files were prepared. Next, the protein files were compared against each other. With 632 genomes, this generated 399424 (632*632) blast comparisons. Blast distances were then computed in order identify clusters and produce a pan-matrix. Subsequently, a phylogenetic tree based on blast distances was constructed. Clustering was done using the *bclust* function within the micropan package. A threshold value of 0.75 was used. Due to the size of our pangenomes and the exponential RAM requirements, the single linkage parameter was used. Core genome size and total pangenome size estimations were carried out with the binomixEstimate function, fitting binomial mixture models to the computed pan-matrix data. Principal component analysis (PCA) was performing using R. The pan-matrix data was loaded and read as a table and subsequently plotted. Unique numbers were assigned to different members of the same species.

To investigate whether or not difference in pangenome sizes were caused by having a different number of genomes making each subgroup, random trials were performed. Subsampling was done by randomly selecting 18 genomes from each subgroup (18 was chosen as it represented the size of the smallest subgroup). The entire analysis procedure was performed. This was repeated five times.

Results

Pangenome	number of genomes	pangenome size	Estimated core size (clusters)	closed/ open	alpha	Jaccard
Africa	24	26491	621	open	0.680735	0.488927
Asia	181	71758	10	open	0.393525	0.583958
Latin America & Caribbean	26	23404	769	open	0.731071	0.437667
North America	89	41417	475	open	0.481170	0.490172
Western Europe	109	43772	457	open	0.456434	0.573861
Engineered	32	30510	851	open	0.661623	0.542938
Food	46	27722	825	open	0.641062	0.516747
Freshwater	32	42465	641	open	0.415799	0.607756
Host	131	43067	493	open	0.528638	0.506518
Marine	18	39092	272	Open	0.46046	0.550491
Terrestrial	166	60758	1	open	0.458459	0.586594

TABLE 2: SUBGROUP DATA. (A CLUSTER IS A GENE FAMILY)

Pangenome sizes across locations and environments. Our data indicated pangenome sizes differed by location and environment (Table 2 and Appendix B). By performing random

subsampling, we were able to rule out the differing number of genomes as the cause. Furthermore, subsampling was able to confirm that the pangenomes are indeed open. The calculated alpha values were far below 1 for all subsamples except for one North American subsample. In addition, when comparing the alpha and Jaccard values of subgroups (Table 2) to the respective subsample averages (Appendix B), the differences between values were small. This is also clear when the number of clusters are plotted against the number of genomes both for the subsamples (Figure 2) and the total (Figure 3), where the collector's curves are not showing a plateau. These results indicate that much more sampling (or in other words much more genomes) is required in order to see a closed pangenome.

Which factors affected genome clustering; geographic location, environment, or phylogeny? To answer this question, we used Principal Component Analyses (PCA) based on the clustering information. We plotted PCA for each subgroup including the five location and 6 environment subgroups shown in the first column of Table 2 above. We then examined each of these PCA to see whether the genomes clustered by their geography, their environment, or merely by their phylogeny. The analysis was also repeated using the clustering information from the 5 random subsamples (with 18 genomes each) for each of these subgroups. After examining these PCA plots, it was evident that clustering was mainly based on phylogeny as genomes from the same species or from supergroups always clustered closely together regardless of the location or the environment from which they were obtained (Appendix C).

Conclusion

Our analysis suggests that the *Bacillus* pangenome is an open pangenome. The alpha values were well below the threshold of 1. As we added more genomes, the slope of the clusters vs number of genomes line does decrease. It is possible that our study did not include enough genomes to tell the whole story. It was also obvious that the there was an uneven geographical and environmental distribution of *Bacillus* genomes uploaded to NCBI. Further studies with more genomes with need to be done. However, with an increasing number of genomes, exponentially more computational resources are required unless more efficient methods are discovered. Time also scales exponentially unless advances are made to enable parallelization.

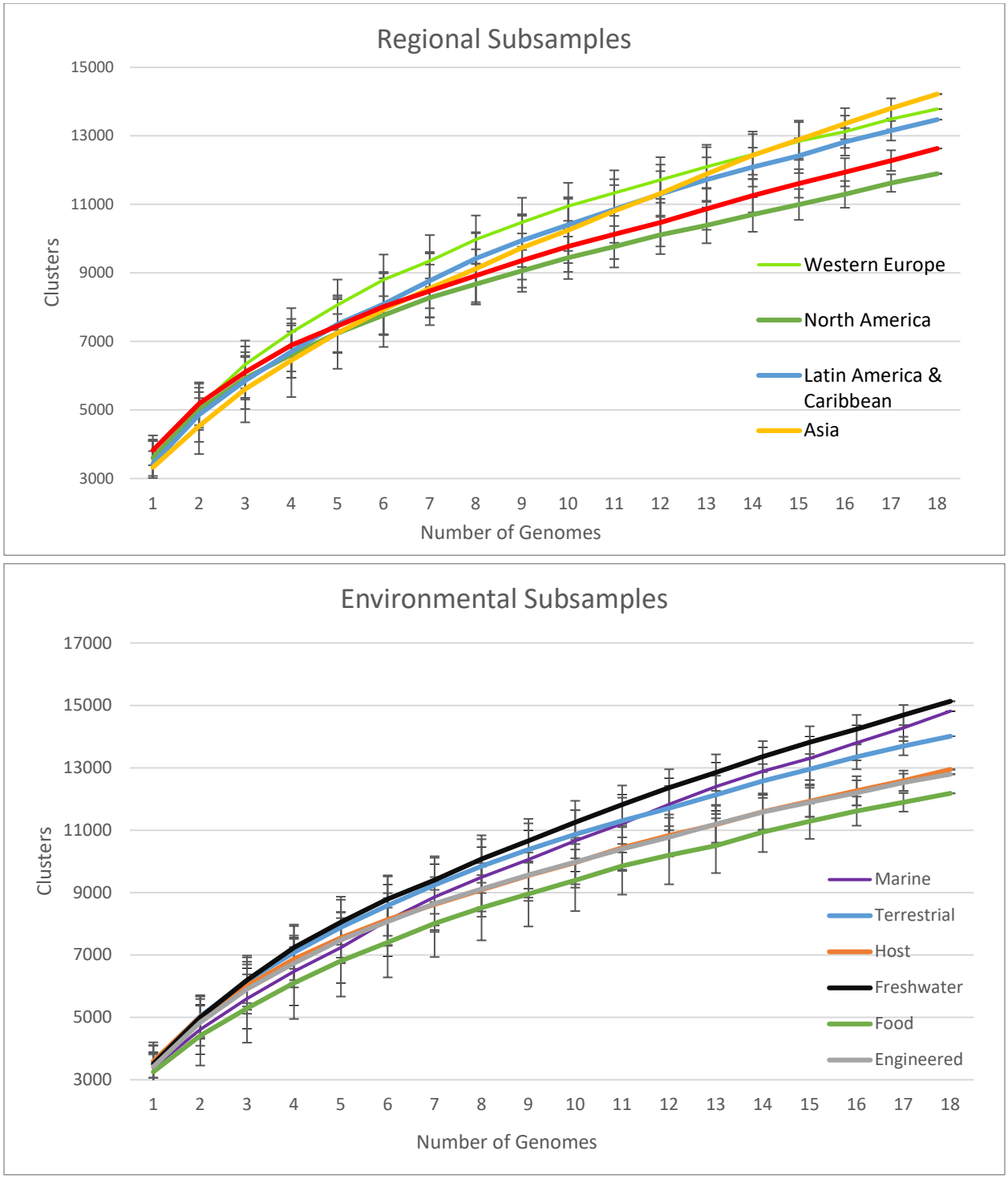


Figure 2: As the number of genomes increased, the numbers also increased. Unlike a closed pangenome, the slopes of the lines presented here does not appear to be approaching a limit. Vertical bars represent standard deviations at each x value.

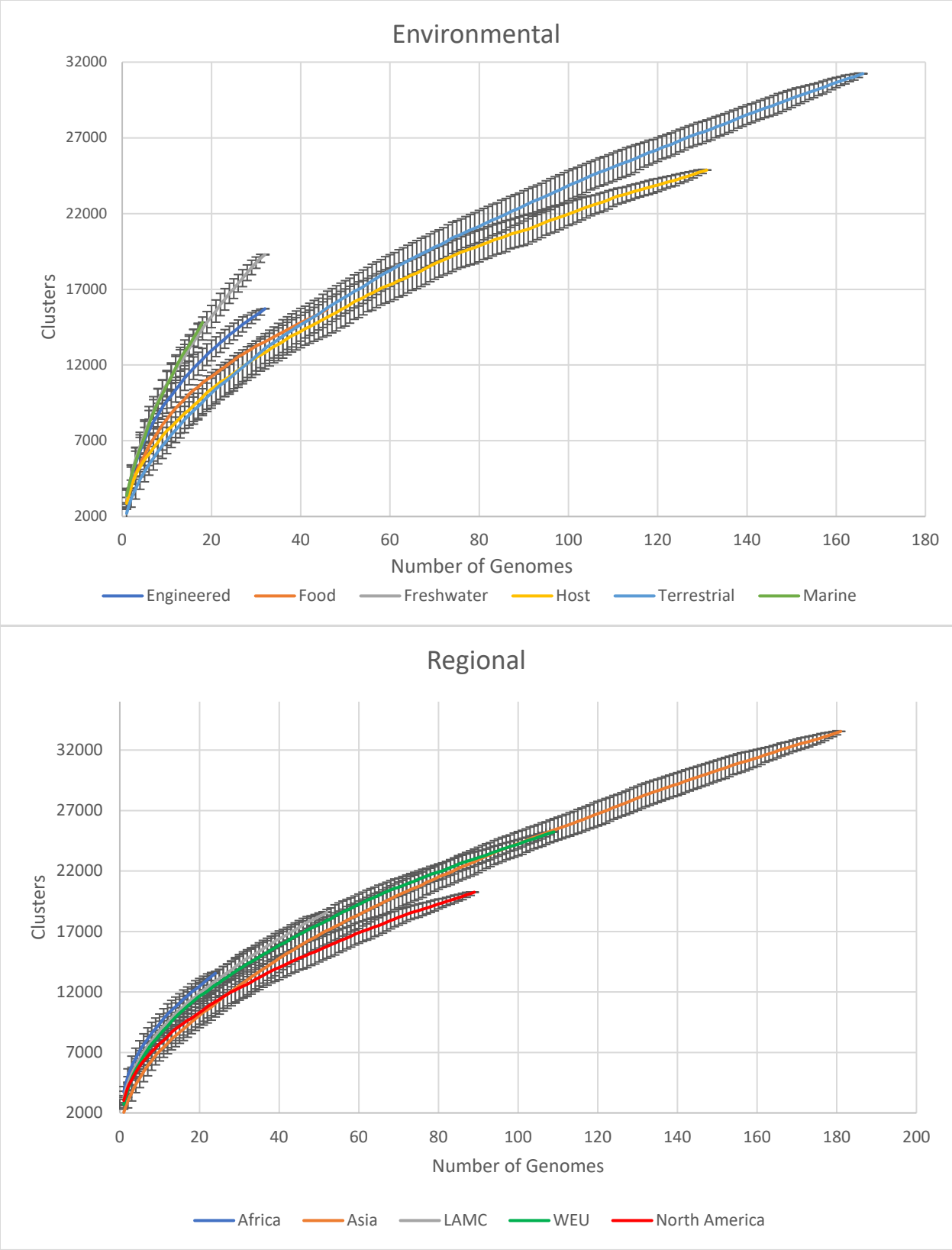


Figure 3: When looking at the clusters vs number of genome graphs for the subgroups, we see the number of clusters rapidly increase then slow down. However, the slopes still do not appear to be approaching a limit, in line with the calculated alpha values. Vertical bars represent standard deviations at each x value.

Appendix A: Region classification by Country of Origin

Asia	**Baltics**	**Latin America & Caribbean**	**Africa**	**Western Europe**
Afghanistan	Estonia	Anguilla	Algeria	Andorra
Bangladesh	Latvia	Antigua & Barbuda	Angola	Austria
Bhutan	Lithuania	Argentina	Benin	Belgium
Brunei		Aruba	Botswana	Denmark
Burma	**C.W. OF IND. STATES**	Bahamas, The	Burkina Faso	Faroe Islands
Cambodia	Armenia	Barbados	Burundi	Finland
China	Azerbaijan	Belize	Cameroon	France
East Timor	Belarus	Bolivia	Cape Verde	Germany
Hong Kong	Georgia	Brazil	Central African Rep.	Gibraltar
India	Kazakhstan	British Virgin Is.	Chad	Greece
Indonesia	Kyrgyzstan	Cayman Islands	Comoros	Guernsey
Iran	Moldova	Chile	Congo, Dem. Rep.	Iceland
Japan	Russia	Colombia	Congo, Dem. Rep.	Ireland
North Korea	Tajikistan	Costa Rica	Cote d'Ivoire	Isle of Man
South Korea	Turkmenistan	Cuba	Djibouti	Italy
Laos	Ukraine	Dominica	Egypt	Jersey
Macau	Uzbekistan	Dominican Republic	Equatorial Guinea	Liechtenstein
Malaysia	Former Soviet Union	Ecuador	Eritrea	Luxembourg
Maldives		El Salvador	Ethiopia	Malta
Mongolia	**Near East**	French Guiana	Gabon	Monaco
Nepal	Bahrain	Grenada	Gambia, The	Netherlands
Pakistan	Cyprus	Guadeloupe	Ghana	Norway
Philippines	Gaza Strip	Guatemala	Guinea	Portugal
Singapore	Iraq	Guyana	Guinea-Bissau	San Marino
Sri Lanka	Israel	Haiti	Kenya	Spain
Taiwan	Jordan	Honduras	Lesotho	Sweden
Thailand	Kuwait	Jamaica	Liberia	Switzerland
Vietnam	Lebanon	Martinique	Libya	
	Oman	Mexico	Madagascar	Scotland
	Qatar	Montserrat	Malawi	United Kingdom
	Saudi Arabia	Netherlands Antilles	Mali	
	Syria	Nicaragua	Mauritania	**Eastern Europe**
	Turkey	Panama	Mauritius	Albania
	United Arab Emirates	Paraguay	Mayotte	Bosnia & Herzegovina
	West Bank	Peru	Morocco	Bulgaria
	Yemen	Puerto Rico	Mozambique	Croatia
		Saint Kitts & Nevis	Namibia	Czechoslovakia
	North America	Saint Lucia	Niger	Czech Republic
	Bermuda	Saint Vincent and the	Nigeria	Hungary
	Canada	Grenadines	Reunion	Macedonia
	Greenland	Suriname	Rwanda	Poland
	St Pierre & Miquelon	Trinidad & Tobago	Saint Helena	Romania
	United States	Turks & Caicos Is	Sao Tome & Principe	Serbia
		Uruguay	Senegal	Slovakia
		Venezuela	Seychelles	Slovenia
		Virgin Islands	Sierra Leone	
			Somalia	
			South Africa	
			Sudan	
			Swaziland	
			Tanzania	
			Togo	
			Tunisia	
			Uganda	
			Western Sahara	
			Zambia	
			Zimbabwe	

Appendix B: Subsample Data

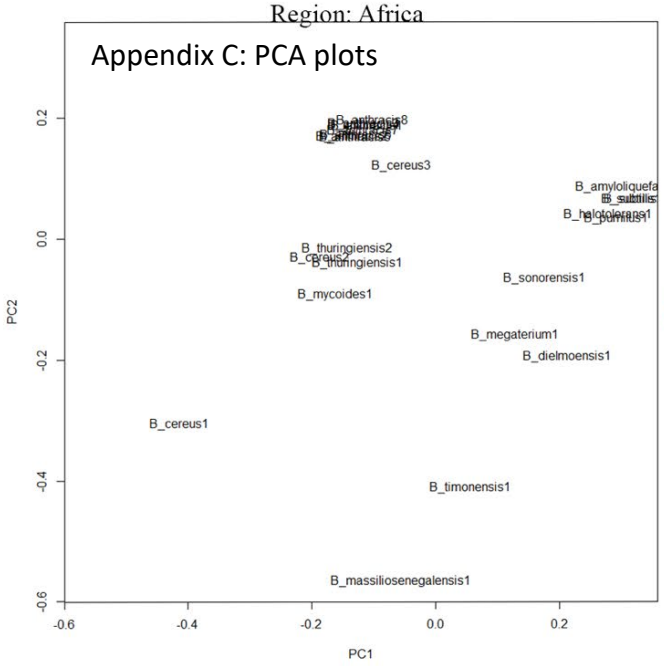
Pangenome	pangenome size	Estimated core size (clusters)	closed/ open	alpha	Jaccard	Average Pan size	Average core	Average alpha	Average Jaccard	StDev PanSize	StDev core	Stdev Alpha	StDev Jaccard
Africa 1	22005	593	open	0.4553537	0.433882								
Africa 2	27792	345	open	0.4563117	0.529493								
Africa 3	26928	1103	open	0.4521538	0.512742								
Africa 4	26443	273	open	0.7050855	0.529934								
Africa 5	19900	744	open	0.7785115	0.47349	24613.6	611.6	0.56948324	0.4959081	3457.896658	333.5697828	0.15943676	0.041589815
Asia 1	34580	695	open	0.6721211	0.567932								
Asia 2	19485	1019	open	0.6301683	0.495048								
Asia 3	27507	853	open	0.5016329	0.556166								
Asia 4	21472	898	open	0.3543851	0.487318								
Asia 5	25378	799	open	0.479846	0.554265	25684.4	852.8	0.52763068	0.5321 4588	5889.064552	119.8549123	0.12685103	0.037857103
Engineered 1	23590	877	open	0.708943	0.572166								
Engineered 2	22064	950	open	0.7259468	0.556165								
Engineered 3	15704	1104	open	0.6709187	0.530243								
Engineered 4	20686	350	open	0.6923687	0.533681								
Engineered 5	26098	929	open	0.6502747	0.526212	21628.4	842	0.68969038	0.5436933	3873.034314	287.7351212	0.02999306	0.019703817
Food 1	11252	1316	open	0.8309799	0.468074								
Food 2	24048	761	open	0.7139766	0.563266								
Food 3	15704	1104	open	0.6709187	0.530243								
Food 4	18110	1075	open	0.642987	0.553348								
Food 5	17942	1179	open	0.5452709	0.493223	17411.2	1087	0.68082662	0.5216 3068	4627.835909	204.703444	0.10434352	0.040276893
Freshwater 1	37187	722	open	0.3632943	0.618665								
Freshwater 2	28207	467	open	0.3768284	0.62452								
Freshwater 3	24184	803	open	0.6409266	0.593168								
Freshwater 4	34266	846	open	0.4247286	0.609472								
Freshwater 5	24670	829	open	0.6489047	0.582388	29702.8	733.4	0.49093652	0.6056 4266	5806.455261	156.3339375	0.14243214	0.017585572
Host 1	27560	1067	open	0.6964876	0.542498								
Host 2	32549	84	open	0.6314323	0.547085								
Host 3	23338	1100	open	0.7242915	0.544012								
Host 4	17639	895	open	0.777438	0.464703								
Host 5	30771	958	open	0.4669348	0.499903	26371.4	820.8	0.65931684	0.5196 401	6008.084495	420.0508303	0.11973343	0.036322513
Latin America & Caribbean 1	23499	1096	open	0.6838306	0.556342								
Latin America & Caribbean 2	28157	1096	open	0.7711738	0.550643								
Latin America &	19191	712	open	0.8188463	0.544046								

Caribbean 3													
Latin America & Caribbean 4	27717	16	open	0.709396	0.568584								
Latin America & Caribbean 5	35354	0	open	0.6625861	0.56735	26783.6	584	0.7291 6656	0.5573 9298	6013.4 30618	548.71 48622	0.064 58762	0.0105 97225
Marine 1	39092	272	open	0.4854578	0.550491								
Marine 2	39092	272	open	0.4854578	0.550491								
Marine 3	39092	272	open	0.4854578	0.550491								
Marine 4	39092	272	open	0.4854578	0.550491								
Marine 5	39092	272	open	0.4854578	0.550491	39092							
North America 1	28228	0	open	0.7383063	0.513465								
North America 2	25896	1	open	0.6940126	0.548129								
North America 3	25808	578	open	0.7419565	0.537815								
North America 4	11975	1160	closed	1.124703	0.440259								
North America 5	24451	711	open	0.7981726	0.525639	23271.6	490	0.8194 302	0.5130 613	6459.3 12216	496.16 1768	0.174 61114	0.0427 23121
Terrestrial 1	33683	857	open	0.5857039	0.56566								
Terrestrial 2	31464	819	open	0.6379748	0.572156								
Terrestrial 3	25344	0	open	0.6593896	0.568993								
Terrestrial 4	30102	953	open	0.678333	0.557079								
Terrestrial 5	29035	23	open	0.4257995	0.566002	29925.6	530.4	0.5974 4016	0.5659 781	3094.3 52646	476.26 86217	0.102 00820	0.0056 2509
Western Europe 1	27628	991	open	0.5932333	0.609213								
Western Europe 2	19534	1027	open	0.7556711	0.589953								
Western Europe 3	31016	896	open	0.4100924	0.565506								
Western Europe 4	25065	993	open	0.6962406	0.572009								
Western Europe 5	27838	1072	open	0.7294422	0.570945	26216.2	995.8	0.6369 3592	0.5815 2506	4290.6 45429	64.751 06177	0.141 01454	0.0180 05398

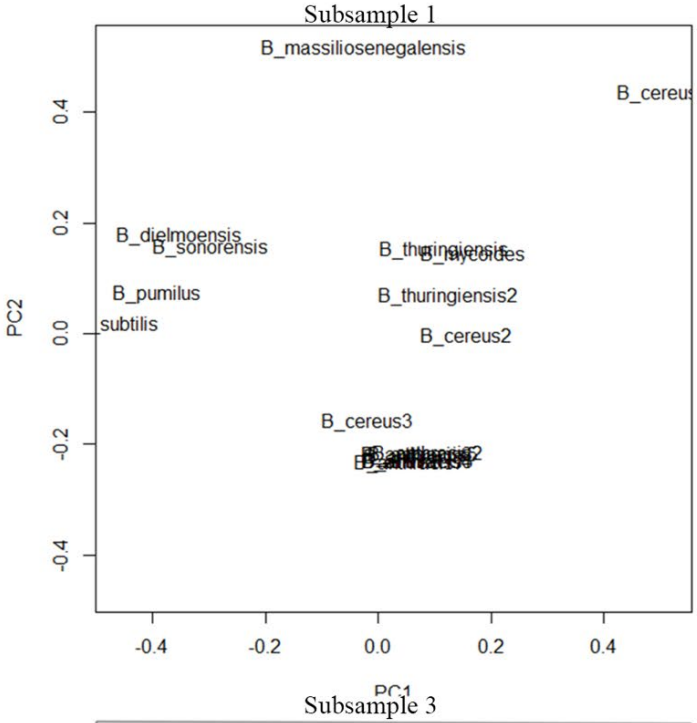
Each subset contained 18 genomes. Marine 1-5 are identical because it was the smallest pangenome, which only contained 18 genomes.

Region: Africa

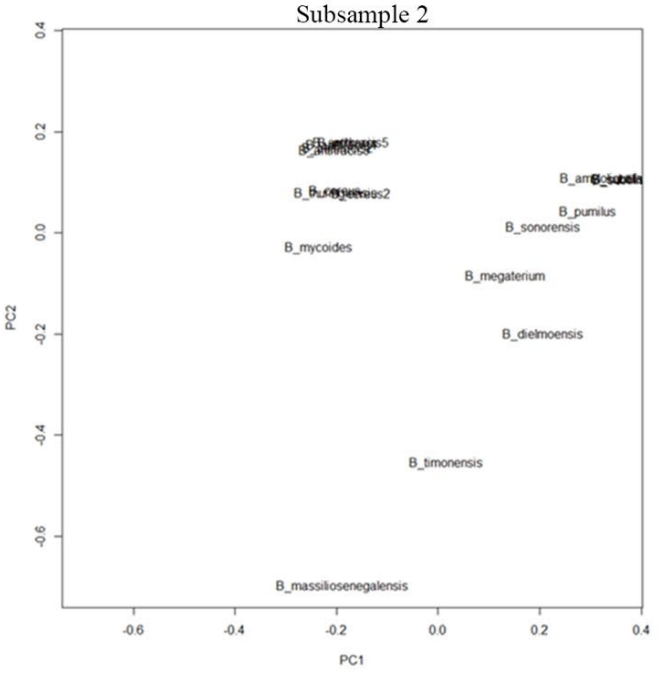
Appendix C: PCA plots



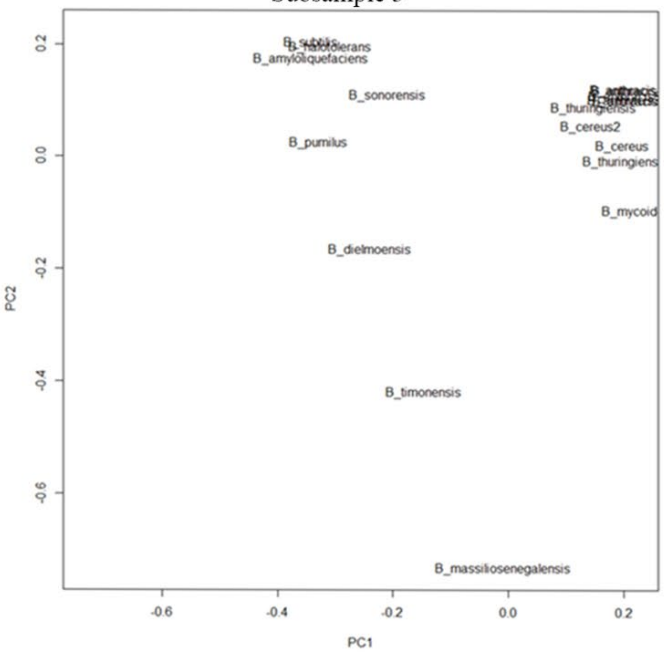
Subsample 1



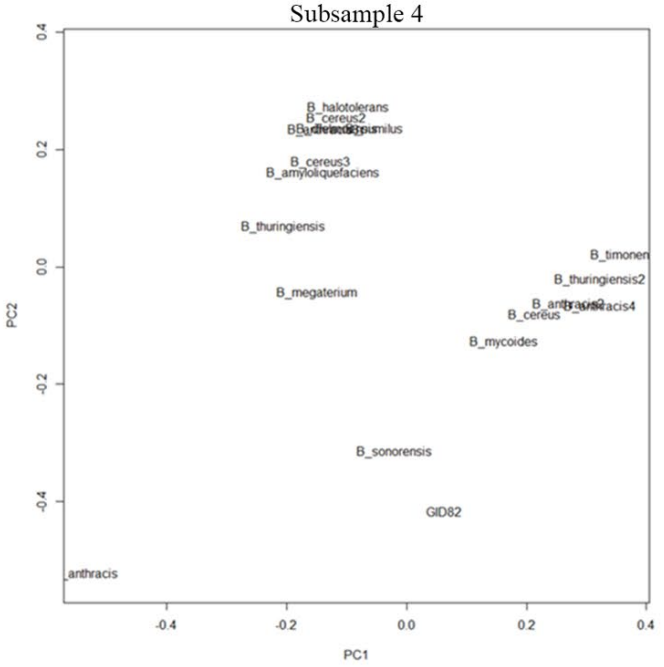
Subsample 2



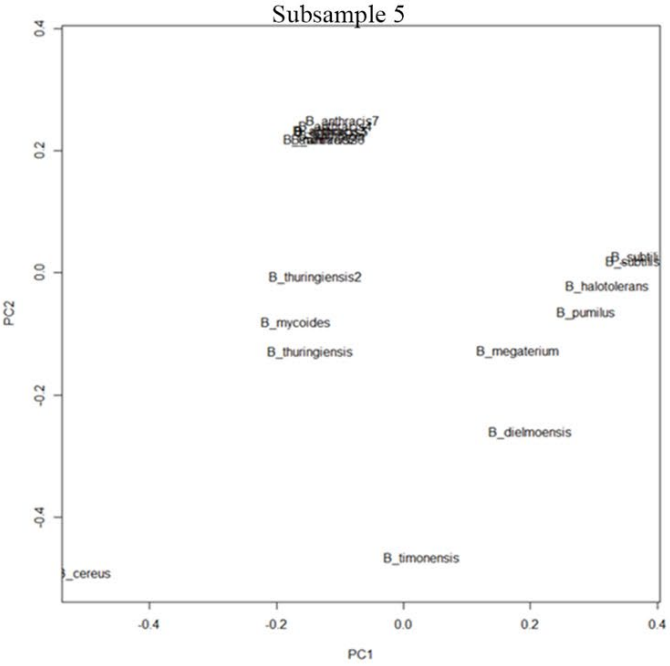
Subsample 3

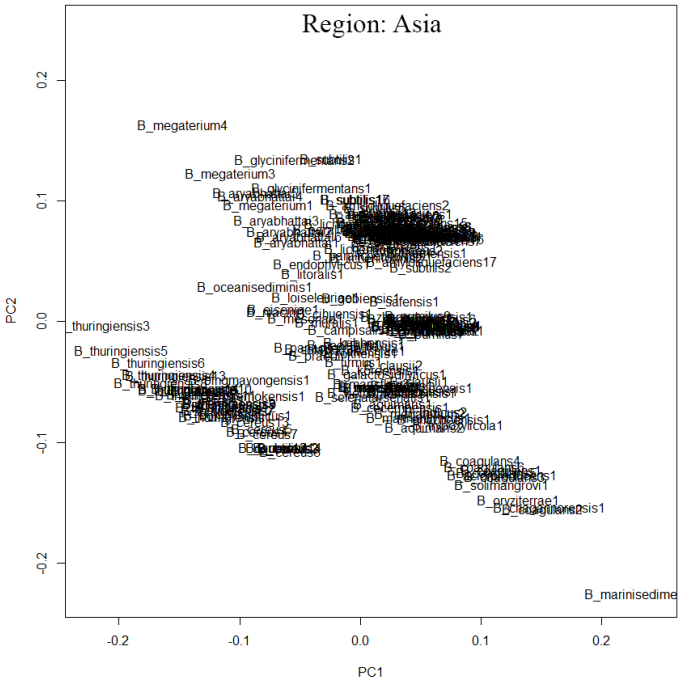


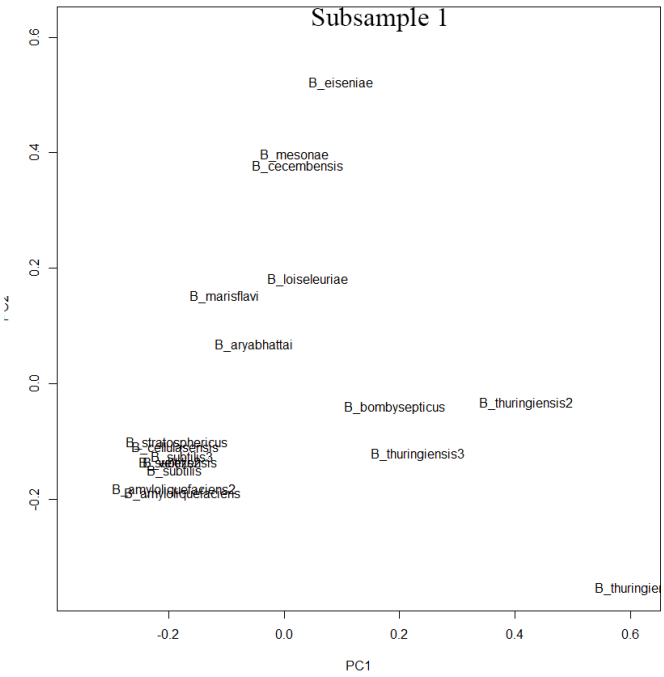
Subsample 4

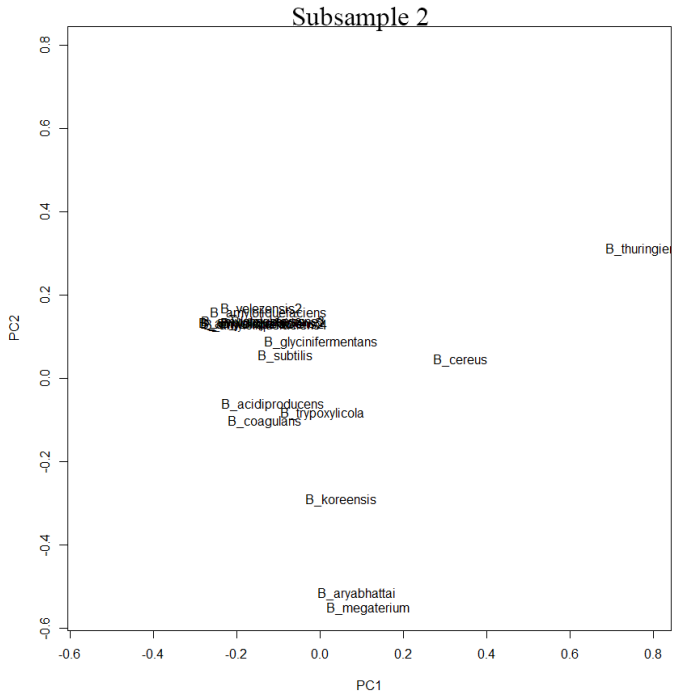


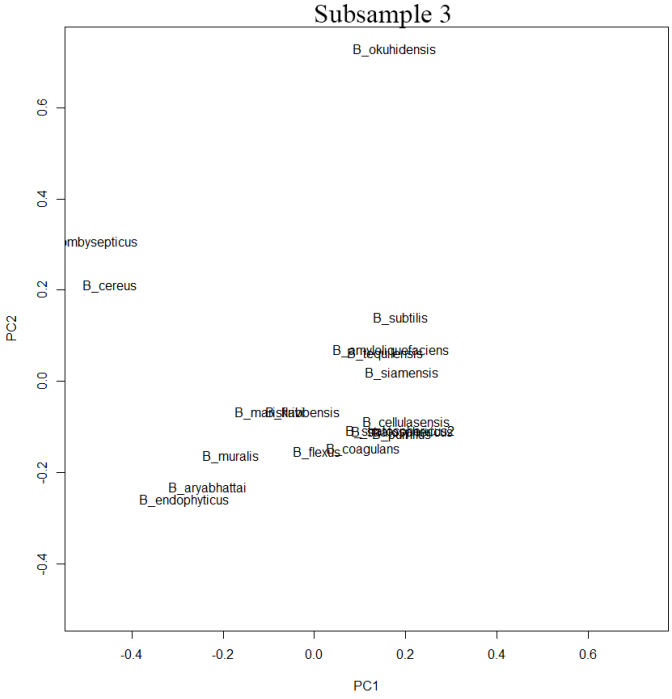
Subsample 5

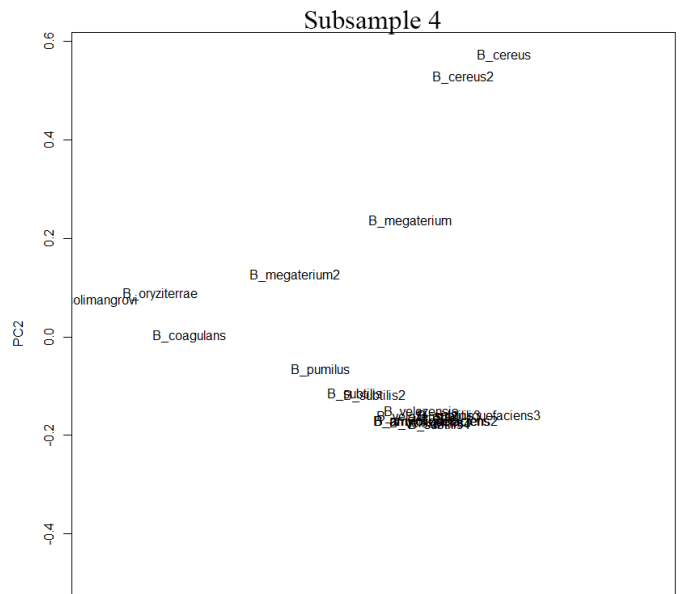


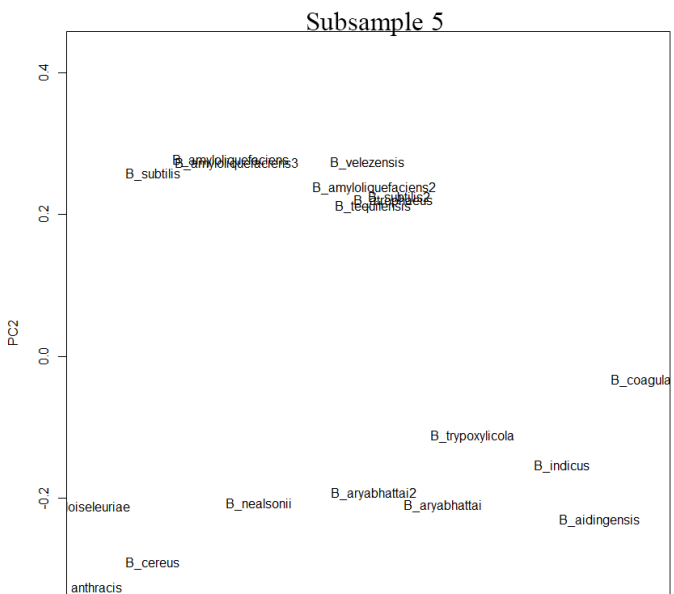

Region: Asia
B_megaterium4
B_megaterium3
B_megaterium1
B_endophyticus
B_litoralis1
B_oceanisediminis1
B_thuringiensis3
B_thuringiensis5
B_thuringiensis6
B_coagulans4
B_solimangrovi1
B_oryziterrae1
B_marinisedime
PC1
PC2


Subsample 1
B_eiseniae
B_mesonae
B_cecembensis
B_loiseleuriae
B_marisflavi
B_aryabhattai
B_bombysepticus
B_thuringiensis2
B_thuringiensis3
B_amyloliquefaciens2
B_thuringie
PC1

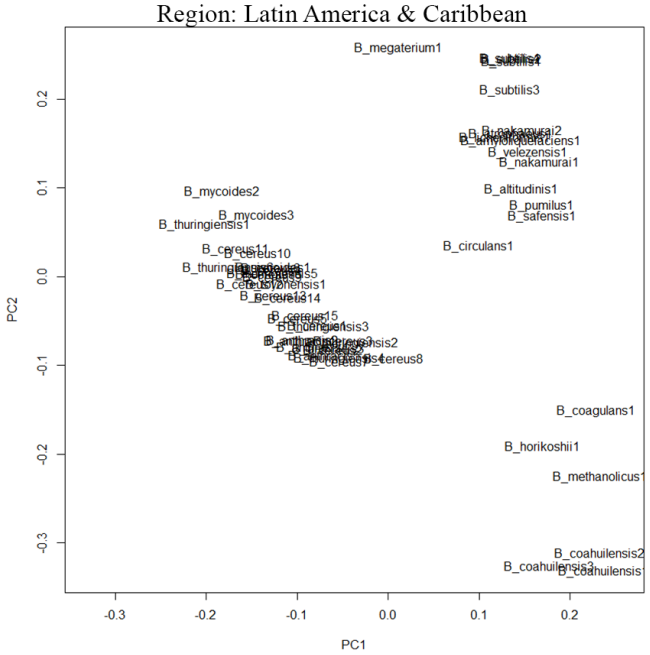

Subsample 2
B_thuringie
B_velezensis2
B_glycinifermentans
B_subtilis
B_cereus
B_acidiproducens
B_trypoxylicola
B_coagulans
B_koreensis
B_aryabhattai
B_megaterium
PC1
PC2


Subsample 3
B_okuhidensis
mbysepticus
B_cereus
B_subtilis
B_siamensis
B_cellulasensis
B_coagulans
B_muralis
B_aryabhattai
B_endophyticus
PC1
PC2

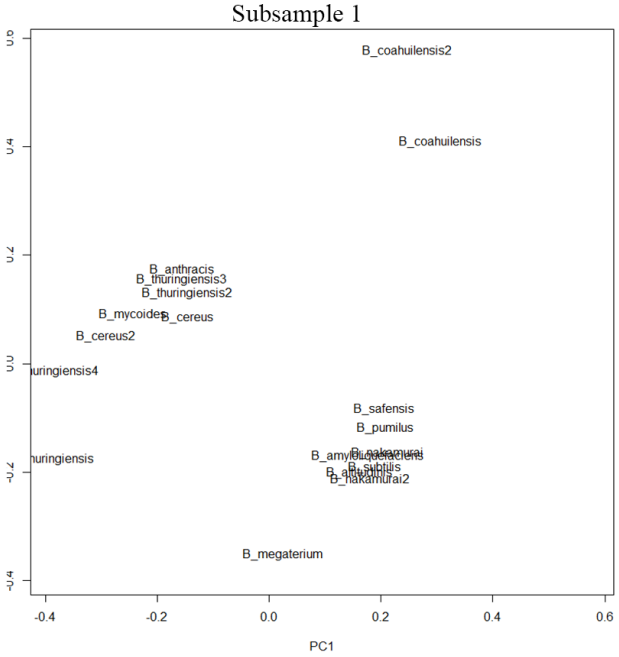

Subsample 4
B_cereus
B_cereus2
B_megaterium
B_megaterium2
B_oryziterrae
olimangrovi
B_coagulans
B_pumilus
B_subtilis2
PC1
PC2


Subsample 5
B_velezensis
B_subtilis
B_amyloliquefaciens2
B_coagula
B_trypoxylicola
B_indicus
B_aryabhattai2
B_nealsonii
B_aryabhattai
B_aidingensis
loiseleuriae
B_cereus
anthracis
PC1
PC2

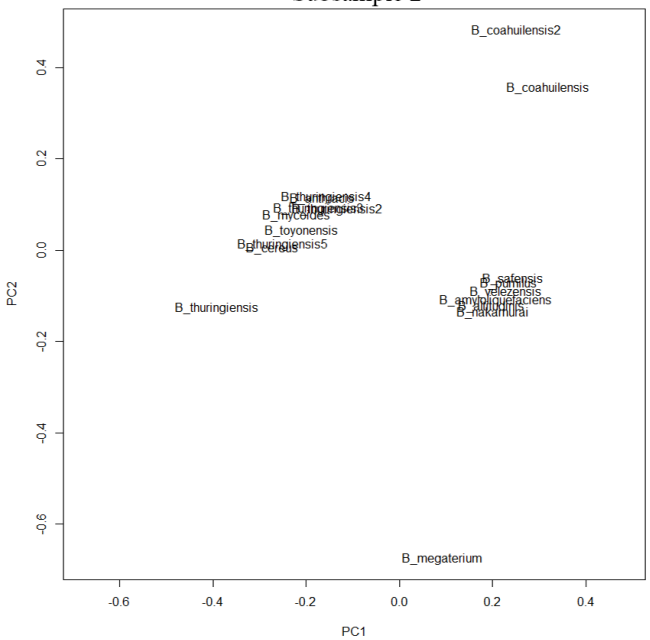
Region: Latin America & Caribbean



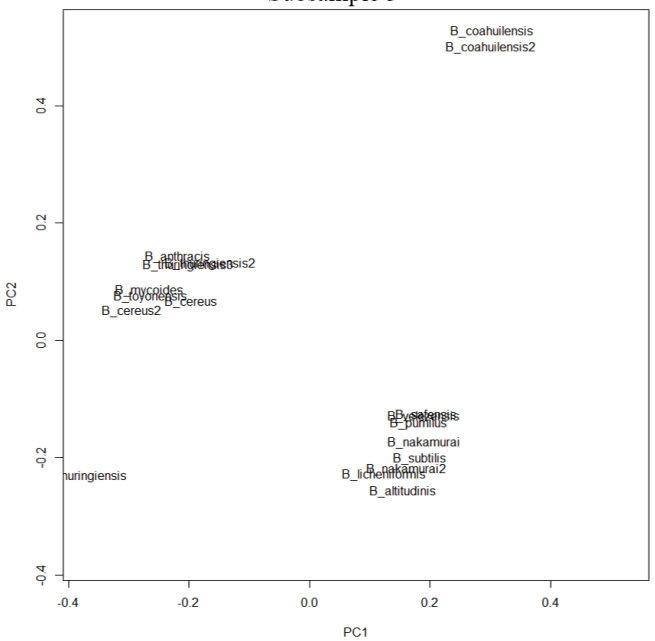
Subsample 1



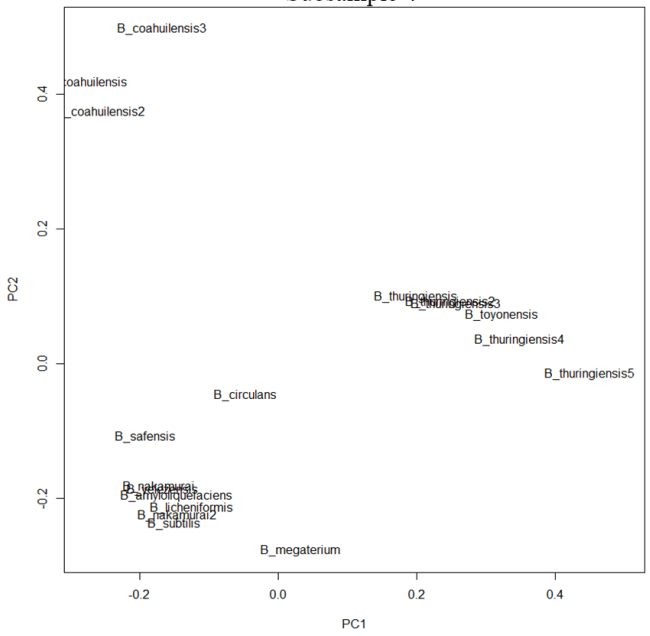
Subsample 2



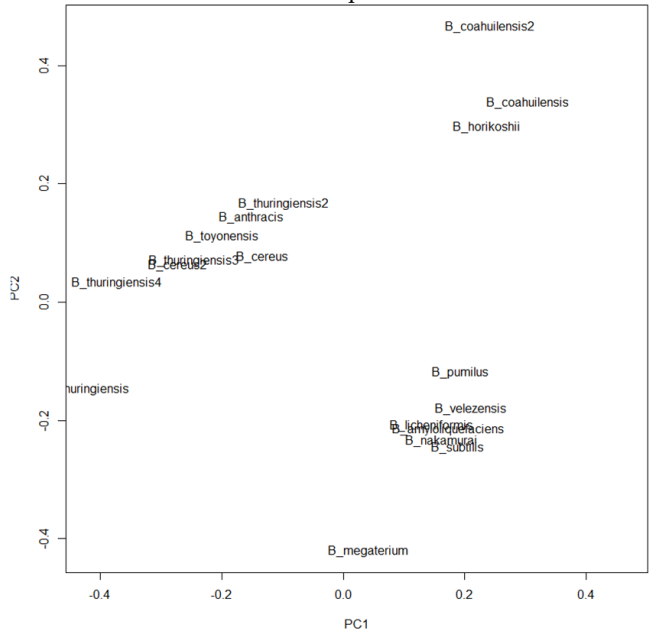
Subsample 3



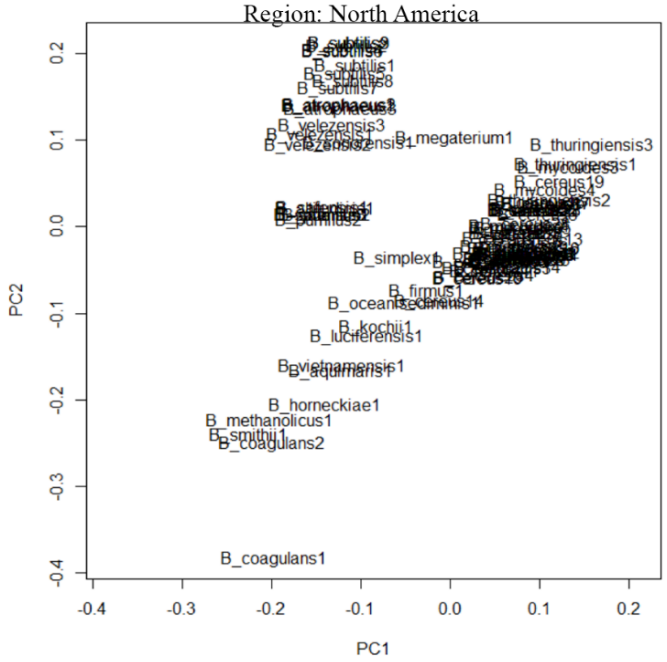
Subsample 4



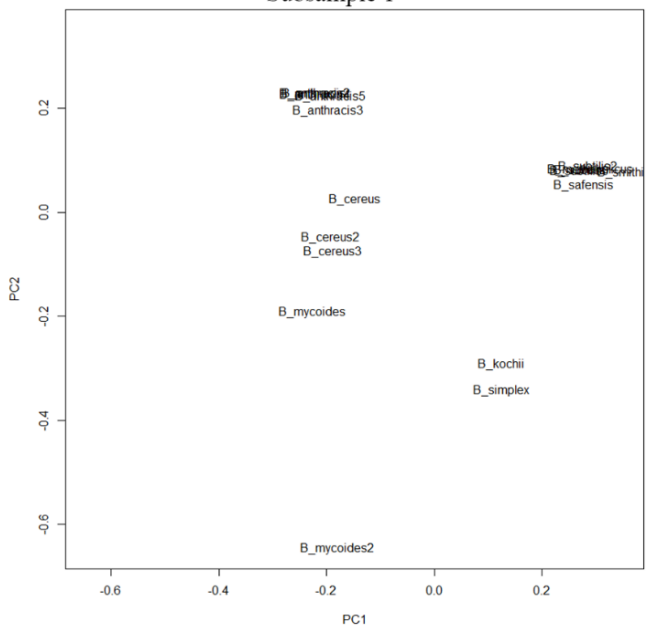
Subsample 5



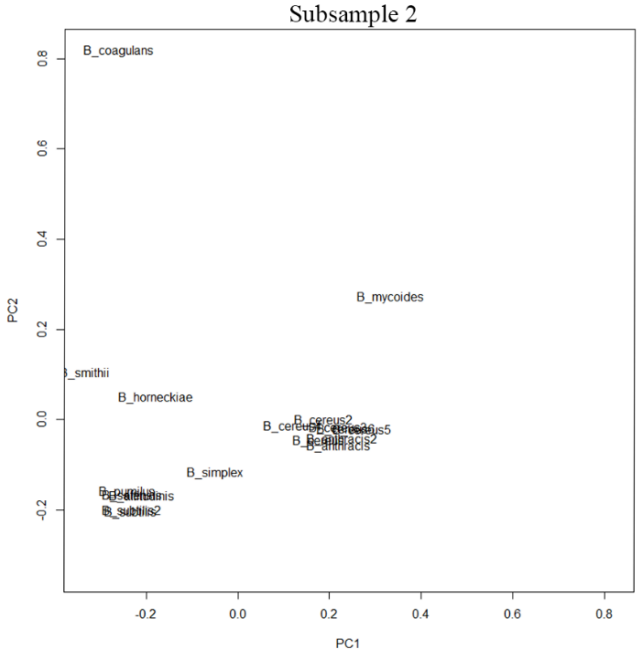
Region: North America



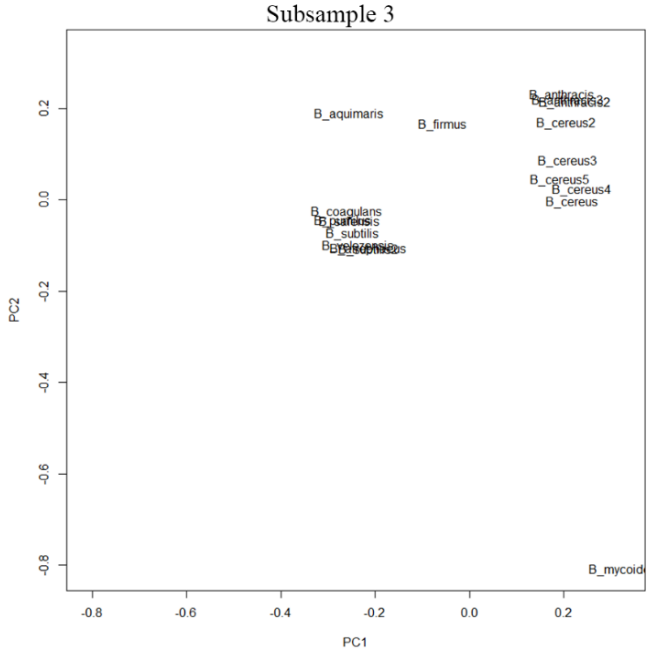
Subsample 1



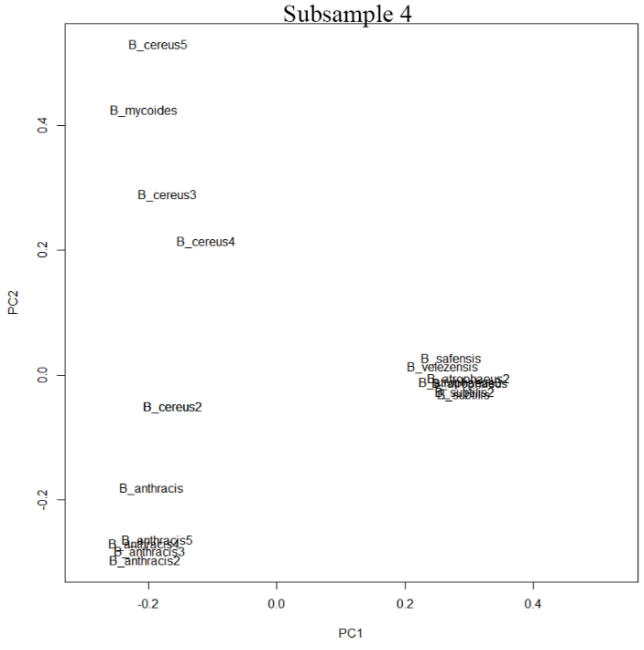
Subsample 2



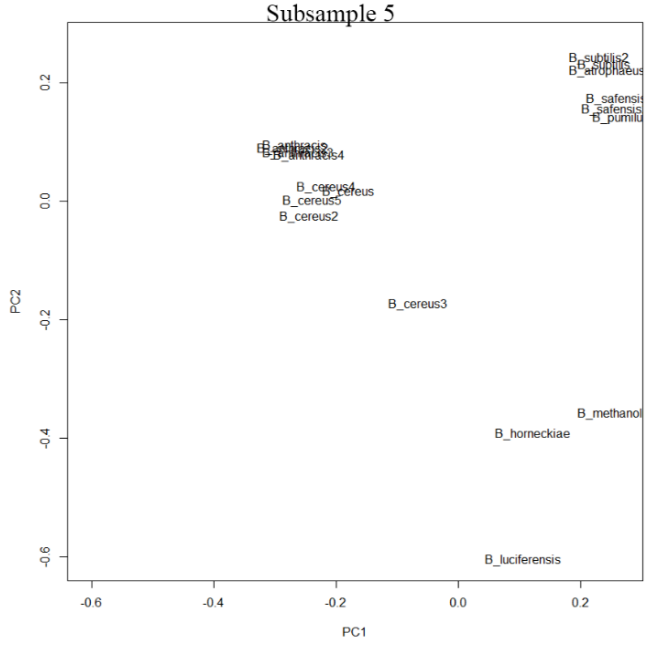
Subsample 3



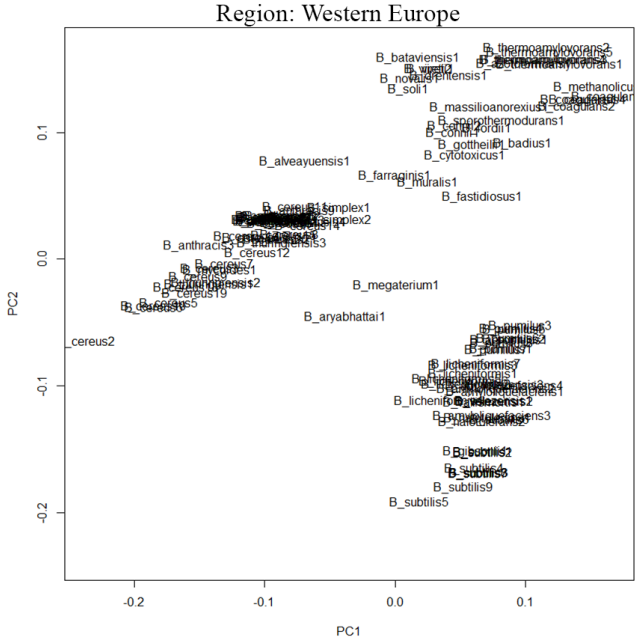
Subsample 4



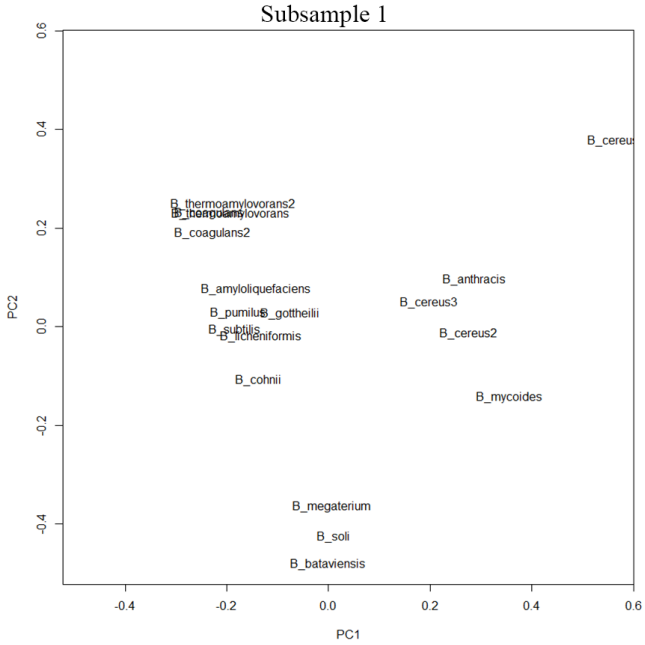
Subsample 5



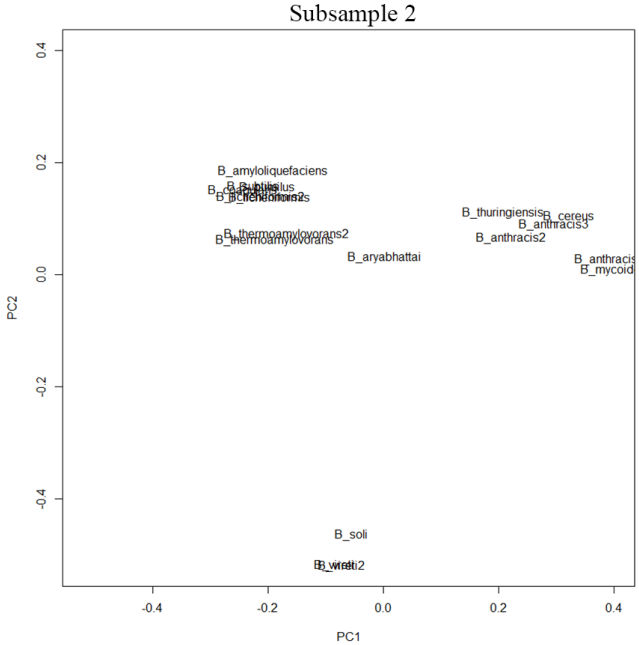
Region: Western Europe



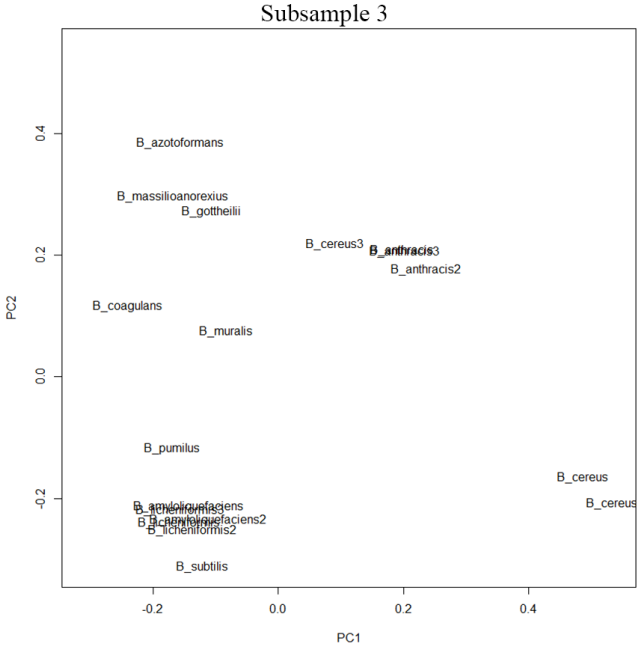
Subsample 1



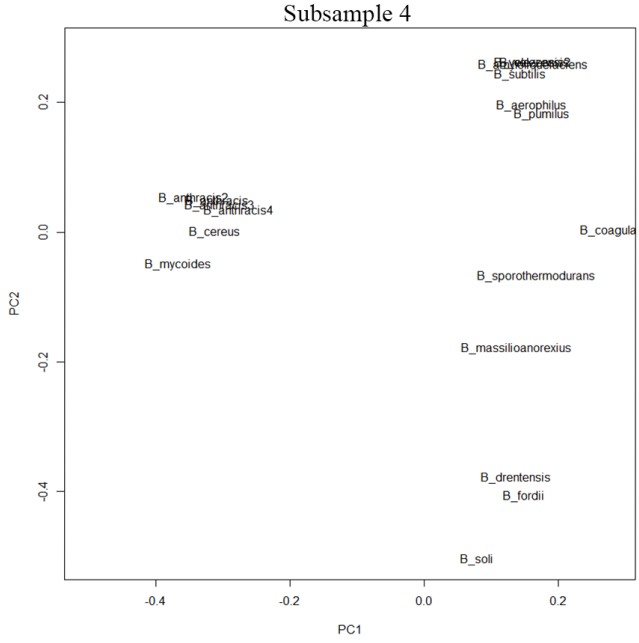
Subsample 2



Subsample 3



Subsample 4



Subsample 5