

Multivariate analysis of functional metagenomes

*Elizabeth A. Dinsdale¹, Robert A. Edwards^{1,2,3}, Barbara Bailey⁴, Imre Tuba⁵, Sajia Akhter⁶, Katelyn McNair⁶, Robert Schmieder⁶, Naneh Apkarian⁷, Michelle Creek⁸, Eric Guan⁹, Mayra Hernandez⁴, Catherine Isaacs¹⁰, Chris Peterson⁷, Todd Regh¹¹, Vadim Ponomarenko⁴

Supplemental Online Material

Detailed Methods

Metagenomes

Publicly available metagenomes were selected from the Edwards Lab metagenome database (<http://edwards.sdsu.edu/mymgdb/>). All samples were annotated using the real-time K-mer based annotation system using a 10-amino acid word size and a requirement for at least two words per protein (<http://edwards.sdsu.edu/rtmg>). This approach, described elsewhere, (Edwards, pers. comm.) uses signature K-mers to identify the functions encoded in the metagenome sample. The K-mer based approach allows all of the samples to be annotated against the same core database, and for the annotations to be updated whenever required. The K-mer based annotation provides the number of sequences for each function, subsystem, and two level hierarchies in the subsystems ontology (Henry et al 2011). Counts were normalized by the total number of hits to account for the different sample sizes of each metagenomes and to yield percent composition by function. The functional hierarchy's *clustering-based subsystems* and *experimental subsystems* were removed from the data, leaving 27 first level functional hierarchies or functional families. The metagenomes were classified as belonging to ten different environments: hypersaline (from Solar Salterns); mat community; hydrothermal springs; human associated; other terrestrial animal associated; freshwater; and marine. Because of the abundance of marine samples (for example, because of the Global Ocean Survey data), these samples were further sub-divided into four groups: open ocean, coastal water, deep water, and coral-reef water associated samples.

References

Henry CS, Overbeek R, Xia F, Best AA, Glass E, Gilbert J, Larsen P, Edwards R, Disz T, Meyer F, Vonstein V, Dejongh M, Bartels D, Desai N, D'Souza M, Devoid S, Keegan KP, Olson R, Wilke A, Wilkening J, Stevens RL. 2011 Connecting genotype to phenotype in the era of high-throughput sequencing. *Biochim Biophys Acta* 1810(10):967-77

Supplemental Online Figures and Tables

Supplemental Table 1. Metagenomes used in the analysis.

Environment	Genome ID	Genome Name	Project	Number of sequences	Total length (bp)
Coastal water	4441143	GS009 - Coastal Block Island, NY - USA	Global Ocean Sampling	79,303	84,327,514
Coastal water	4441144	GS010 - Coastal Cape May, NJ - USA	Global Ocean Sampling	78,304	82,424,426
Coastal water	4441148	GS117b - Coastal Indian Ocean - St. Anne Island, Seychelles	Global Ocean Sampling	50,609	54,752,102
Coastal water	4441152	GS004 - Coastal Outside Halifax, Nova Scotia - Canada	Global Ocean Sampling	52,959	56,922,096
Coastal water	4441153	GS007 - Coastal Northern Gulf of Maine - Canada	Global Ocean Sampling	50,980	55,430,960
Coastal water	4441579	GS002 - Coastal Gulf of Maine - Canada	Global Ocean Sampling	121,590	128,761,768
Coastal water	4441580	GS003 - Coastal Browns Bank, Gulf of Maine - Canada	Global Ocean Sampling	61,605	66,907,344
Coastal water	4441581	GS005 - Embayment Bedford Basin, Nova Scotia Canada	Global Ocean Sampling	61,131	65,983,125
Coastal water	4441582	GS006 - Estuary -- Bay of Fundy, Nova Scotia - Canada	Global Ocean Sampling	59,679	64,615,563
Coastal water	4441583	GS008 - Coastal Newport Harbor, RI - USA	Global Ocean Sampling	129,655	137,725,898
Coastal water	4441584	GS012 - Estuary Chesapeake Bay, MD - USA	Global Ocean Sampling	126,162	136,081,077

Coastal water	4441585	GS013 - Coastal - Off Nags Head, NC - USA	Global Ocean Sampling	138,033	149,007,574
Coastal water	4441586	GS015 - Coastal - Caribbean Sea - Off Key West, FL - USA	Global Ocean Sampling	127,362	138,034,062
Coastal water	4441589	GS019 - Coastal - Northeast of Colon - Panama	Global Ocean Sampling	135,325	146,413,090
Coastal water	4441591	GS021 - Coastal - Gulf of Panama - Panama	Global Ocean Sampling	131,798	143,454,700
Coastal water	4441595	GS027 - Coastal - Devil's Crown, Floreana Island -Ecuador	Global Ocean Sampling	222,080	237,326,008
Coastal water	4441596	GS029 - Coastal - North James Bay, Santiago Island - Ecuador	Global Ocean Sampling	131,529	143,822,814
Coastal water	4441596	GS028 - - Coastal Floreana Ecuador	Global Ocean Sampling	189,052	205,008,796
Coastal water	4441597	GS030 - Warm Seep Upwelling, Fernandina Island	Global Ocean Sampling	436,401	461,671,889
Coastal water	4441598	GS032 - Mangrove - Mangrove on Isabella Island - Ecuador	Global Ocean Sampling	148,018	153,341,974
Coastal water	4441600	GS034 - Coastal - North Seamore Island - Ecuador	Global Ocean Sampling	134,347	142,199,308
Coastal water	4441601	GS035 - Coastal - Wolf Island - Ecuador	Global Ocean Sampling	140,814	151,840,270
Coastal water	4441602	GS036 - Coastal - Cabo Marshall, Isabella Island - Ecuador	Global Ocean Sampling	77,538	85,757,477
Coastal water	4441605	GS049 - Coastal - Moorea, Outside Cooks Bay - Fr. Polynesia	Global Ocean Sampling	92,501	94,424,378
Coastal water	4441613	GS117a - Coastal St. Anne Island, Seychelles	Global Ocean Sampling	346,952	339,868,195

Coastal water	4441618	GS149 - Harbor - West coast Zanzibar Tanzania	Global Ocean Sampling	110,984	111,178,553
Coastal water	4441658	GS011 - Estuary Delaware Bay, NJ - USA	Global Ocean Sampling	124,435	133,251,132
Coastal water	4441659	GS014 - Coastal South of Charleston, SC - USA	Global Ocean Sampling	128,885	139,914,998
Coastal water	4441660	GS016 - Coastal Sea Gulf of Mexico - USA	Global Ocean Sampling	127,122	137,479,949
Coastal water	4441662	GS030 - Warm Seep - Roca Redonda - Ecuador	Global Ocean Sampling	359,152	391,694,924
Coastal water	4440358	DMSP21SeawaterMic200511	Marine manipulated	41,461	3,882,661
Coastal water	4440359	VAN11SeawaterMic200511	Marine manipulated	29,104	2,710,130
Coastal water	4440360	DMSP2SeawaterMic200511	Marine manipulated	50,313	4,813,851
Coastal water	4440361	VAN21SeawaterMic200511	Marine manipulated	40,480	3,867,992
Coastal water	4440362	DMSP11SeawaterMic200511	Marine manipulated	44,246	4,202,321
Coastal water	4440363	VAN2SeawaterMic200511	Marine manipulated	33,773	3,269,294
Coastal water	4440364	DMSP1SeawaterMic200511	Marine manipulated	54,848	5,279,589
Coastal water	4440365	VAN1SeawaterMic200511	Marine manipulated	12,446	1,190,841
Coastal water	4443688	BBAY01	Botany Bay Metagenomic	71,068	75,802,328
Coastal water	4443689	BBAY02	Botany Bay Metagenomic	13,512	13,814,160
Coastal water	4443691	BBAY04	Botany Bay Metagenomic	14,708	15,408,753
Coastal water	4443693	BBAY15	Botany Bay Metagenomic	182,393	177,136,646
Coastal water	4443702	SRS000294	Coastal Waters Plymouth	204,693	46,327,791
Coastal water	4443703	SRS000295	Coastal Waters Plymouth	130,806	30,141,333

Coastal water	4443704	SRS000296	Coastal Waters Plymouth	326,310	56,526,614
Coastal water	4443706	SRS000299	Coastal Waters Plymouth	154,069	35,762,224
Coastal water	4443707	SRS000298	Coastal Waters Plymouth	126,086	29,082,158
Coastal water	4443708	SRS000300	Coastal Waters Plymouth	35,712	7,909,745
Coastal water	4443709	SRS000301	Coastal Waters Plymouth	99,488	22,554,197
Coastal water	4443711	SRS000536_2	Marine Synechococcus experiment	333,462	34,334,174
Coastal water	4443712	mb2000jd298_2	Monterey Bay Microbial Study	194,144	46,983,239
Coastal water	4443713	mb2000jd298_1	Monterey Bay Microbial Study	217,549	51,966,974
Coastal water	4443714	mb2001jd115_1	Monterey Bay Microbial Study	186,172	44,189,510
Coastal water	4443715	mb2001jd115_2	Monterey Bay Microbial Study	173,161	40,680,713
Coastal water	4443718	SRS000238	Sapelo Island Metagenome	49,524	4,719,520
Coastal water	4443719	SRS000239	Sapelo Island Metagenome	46,421	4,361,030
Coastal water	4443720	SRS000240	Sapelo Island Metagenome	44,317	4,209,153
Coastal water	4443721	SRS000242	Sapelo Island Metagenome	9,967	933,470
Coastal water	4443722	SRS000241	Sapelo Island Metagenome	41,537	3,890,082
Coastal water	4443724	SRS000243	Sapelo Island Metagenome	30,673	2,940,585
Deep water	4441025	Mediterranean Bathypelagic Habitat	Mediterranean Bathypelagic Habitat	9,047	7,202,361

Deep water	4441041	HOT/ALOHA - Below Base of Euphotic Zone 200m	HOT/ALOHA	8,276	7,829,627
Deep water	4441056	HOT/ALOHA - Deep Abyss 4000m	HOT/ALOHA	11,223	11,028,802
Deep water	4441057	HOT/ALOHA - Well Below Upper Mesopelagic 500m	HOT/ALOHA	9,017	8,764,614
Deep water	4441062	HOT/ALOHA - Core of Dissolved Oxygen Minimum Layer 770m	HOT/ALOHA	11,478	11,811,596
Deep water	4441590	GS020 - Fresh Water - Panama Canal - Lake Gatun - Panama	Global Ocean Sampling	296,355	315,151,139
Freshwater	4443679	AntarcticaAquatic_3	Antarctica Aquatic Microbial	10,042	9,755,315
Freshwater	4443680	AntarcticaAquatic_2	Antarctica Aquatic Microbial	9,672	9,622,231
Freshwater	4443681	AntarcticaAquatic_4	Antarctica Aquatic Microbial	54,446	54,929,769
Freshwater	4443683	AntarcticaAquatic_1	Antarctica Aquatic Microbial	100,085	101,310,476
Freshwater	4443684	AntarcticaAquatic_6	Antarctica Aquatic Microbial	281,490	281,056,691
Freshwater	4443685	AntarcticaAquatic_7	Antarctica Aquatic Microbial	28,481	28,413,296
Freshwater	4443687	AntarcticaAquatic_9	Antarctica Aquatic Microbial	95,521	95,664,001
Freshwater	4440411	PrePondKentSTMic20060504	Freshwater from Aquaculture facility	44,094	4,428,989

Freshwater	4440413	TilPondKentSTMic20060504	Freshwater from Aquaculture facility	63,978	6,484,135
Freshwater	4440422	TilPondKentSTMic200608	Freshwater from Aquaculture facility	67,612	6,932,903
Freshwater	4440440	TilPondKentSTMic200511	Freshwater from Aquaculture facility	381,076	38,804,235
Human associated	4441092	Australian Phosphorus Removing (EBPR) Sludge	Phosphorus Removing (EBPR) Sludge	96,563	100,273,005
Human associated	4441093	US Phosphorus Removing (EBPR) Sludge	Phosphorus Removing (EBPR) Sludge	127,953	120,938,054
Human associated	4440453	TS1	Gut microbiome	217,386	51,708,794
Human associated	4440454	TS2	Twin Study	443,640	78,853,892
Human associated	4440461	TS4	Twin Study	414,754	95,003,113
Human associated	4440462	TS5	Twin Study	490,776	100,599,979
Human associated	4440463	TS6	Twin study	535,763	118,207,161
Human associated	4440595	TS3	Twin study	510,972	102,717,417
Human associated	4440610	TS19	Twin Study	498,880	82,117,565
Human associated	4440611	TS20	Twin Study	495,040	98,053,098
Human associated	4440613	TS28	Twin Study	302,780	101,434,082
Human associated	4440614	TS49	Twin Study	519,072	91,987,878
Human associated	4440615	TS50	Twin Study	549,700	111,999,603
Human associated	4440616	TS29	Twin Study	502,399	173,386,030
Human associated	4440639	TS21	Twin study	413,772	88,786,017

Human associated	4440640	TS51	Twin study	434,187	81,330,211
Human associated	4440823	TS7	Twin study	555,853	134,889,015
Human associated	4440824	TS8	Twin study	414,497	100,520,072
Human associated	4440825	TS30	Twin study	495,865	94,405,318
Human associated	4440826	TS9	Twin study	499,499	124,768,172
Human associated	4440939	human F1-S	Human feces - Kurokawa	28,900	38,010,851
Human associated	4440940	human F1-U	Human feces - Kurokawa	16,539	24,369,492
Human associated	4440941	human F1-T	Human feces - Kurokawa	36,326	43,259,070
Human associated	4440942	human F2-V	Human feces - Kurokawa	36,455	45,906,118
Human associated	4440943	human F2-W	Human feces - Kurokawa	30,198	40,076,128
Human associated	4440944	human F2-X	Human feces - Kurokawa	31,237	39,071,077
Human associated	4440945	human In-B	Human feces - Kurokawa	9,958	14,499,070
Human associated	4440946	human In-A	Human feces - Kurokawa	20,226	29,296,224
Human associated	4440947	human F2-Y	Human feces - Kurokawa	35,177	45,480,292
Human associated	4440948	human In-D	Human feces - Kurokawa	37,296	46,397,089
Human associated	4440949	human In-M	Human feces - Kurokawa	16,164	25,941,797
Human associated	4440950	human In-E	Human feces - Kurokawa	20,532	27,208,886
Human associated	4440951	human In-R	Human feces - Kurokawa	34,797	43,473,860
Hypersaline	4441050	Marine NaCl-Saturated Brine	Marine NaCl-Saturated Brine	2,947	2,380,900
Hypersaline	4441599	GS033 - Hypersaline Floreana Island - Ecuador	Global Ocean Sampling	692,255	729,708,089

Hypersaline	4440324	LowSalternSDBayMic20051110	Solar Saltern	49,074	4,632,200
Hypersaline	4440329	SaltonSeaMic20060823	Solar Saltern	178,407	18,876,339
Hypersaline	4440416	MedSalterSDBayMic20051128	Solar Saltern	8,062	705,995
Hypersaline	4440419	HighSalternSDBayMic20051128	Solar Saltern	35,446	3,711,295
Hypersaline	4440425	MedSalternSDBayMic20051116	Solar Saltern	120,987	11,867,028
Hypersaline	4440426	LowSalternSDBayMic20051128	Solar Saltern	34,296	3,453,306
Hypersaline	4440429	HighSalternSDBayMicB200407	Solar Saltern	39,553	4,028,912
Hypersaline	4440430	HighSalternSDBayMicA200407	Solar Saltern	78,524	7,982,909
Hypersaline	4440433	HighSalternSDBayMicC200407	Solar Saltern	123,879	12,641,571
Hypersaline	4440434	MedSalternSDBayMic20051111	Solar Saltern	23,261	2,323,241
Hypersaline	4440435	MedSalternSDBayMic20051110	Solar Saltern	38,929	3,905,955
Hypersaline	4440437	LowSalternSDBayMic200407	Solar Saltern	268,206	25,280,522
Hypersaline	4440438	HighSalternSDBayMicD200407	Solar Saltern	340,725	34,806,789
Mat community	4440963	Guerrero Negro 1-2mm	Hypersaline Guerro Negro	11,562	7,469,278
Mat community	4440964	Guerrero Negro 0-1mm	Hypersaline Guerro Negro	12,213	8,596,197
Mat community	4440965	Guerrero Negro 2-3mm	Hypersaline Guerro Negro	12,407	8,286,254
Mat community	4440966	Guerrero Negro 3-4mm	Hypersaline Guerro Negro	12,821	8,214,974
Mat community	4440967	Guerrero Negro 4-5mm	Hypersaline Guerro Negro	15,652	9,803,688
Mat community	4440968	Guerrero Negro 10-22mm	Hypersaline Guerro Negro	12,686	8,016,534
Mat community	4440969	Guerrero Negro 5-6mm	Hypersaline Guerro Negro	12,525	8,376,984
Mat community	4440970	Guerrero Negro 6-10mm	Hypersaline Guerro Negro	15,048	9,863,015

Mat community	4440971	Guerrero Negro 22-34mm	Hypersaline Guerro Negro	12,522	8,382,531
Mat community	4440972	Guerrero Negro 34-49mm	Hypersaline Guerro Negro	11,627	7,240,219
Open water	4441051	HOT/ALOHA - Upper Euphotic 10m	HOT/ALOHA	7,837	7,482,115
Open water	4441055	HOT/ALOHA - Base of Chlorophyll Maximum 130m	HOT/ALOHA	6,797	6,091,740
Open water	4441057	HOT/ALOHA - Upper Euphotic 70m	HOT/ALOHA	10,992	10,828,356
Open water	4441125	GS040 - Open Ocean - Tropical South Pacific	Global Ocean Sampling	736	772,365
Open water	4441126	GS041 - Open Ocean - Tropical South Pacific	Global Ocean Sampling	678	739,958
Open water	4441127	GS042 - Open Ocean - Tropical South Pacific	Global Ocean Sampling	699	788,466
Open water	4441128	GS043 - Open Ocean - Tropical South Pacific	Global Ocean Sampling	711	789,468
Open water	4441129	GS044 - Open Ocean - Tropical South Pacific	Global Ocean Sampling	678	714,813
Open water	4441130	GS045 - Open Ocean - Tropical South Pacific	Global Ocean Sampling	730	796,793
Open water	4441131	GS046 - Open Ocean - Tropical South Pacific	Global Ocean Sampling	626	683,240
Open water	4441134	GS110b - Open Ocean - Indian Ocean -	Global Ocean Sampling	49,597	53,607,277
Open water	4441135	GS120 - Open Ocean - Indian Ocean - Madagascar	Global Ocean Sampling	46,052	45,710,196
Open water	4441136	GS039 - Open Ocean - Tropical South	Global Ocean Sampling	759	866,795

		Pacific			
Open water	4441139	GS122b - Open Ocean Madagascar and South Africa	Global Ocean Sampling	50,096	52,667,848
Open water	4441145	GS037 - Open Ocean - Eastern Tropical Pacific	Global Ocean Sampling	65,670	68,651,473
Open water	4441146	GS047 - Open Ocean - Tropical South Pacific	Global Ocean Sampling	66,023	68,340,256
Open water	4441147	GS112b - Open Ocean - Indian Ocean	Global Ocean Sampling	52,118	55,638,894
Open water	4441149	GS116 - Open Ocean - Indian Ocean	Global Ocean Sampling	60,932	64,223,447
Open water	4441150	GS115 - Open Ocean - Indian Ocean	Global Ocean Sampling	61,020	64,230,062
Open water	4441151	GS119 - Open Ocean - Indian Ocean	Global Ocean Sampling	60,987	65,055,874
Open water	4441155	GS109 - Open Ocean - Indian Ocean	Global Ocean Sampling	59,813	62,752,349
Open water	4441156	GS111 - Open Ocean - Indian Ocean	Global Ocean Sampling	59,080	62,072,289
Open water	4441570	GS000a - Open Ocean - Sargasso Sea	Global Ocean Sampling	644,551	658,755,696
Open water	4441573	GS000b - Open Ocean - Sargasso Sea	Global Ocean Sampling	317,180	321,026,307
Open water	4441574	GS000c - Open Ocean - Sargasso Sea	Global Ocean Sampling	368,835	371,688,861
Open water	4441575	GS000d - Open Ocean - Sargasso Sea	Global Ocean Sampling	332,240	335,939,509
Open water	4441576	GS001a - Open Ocean - Sargasso Sea	Global Ocean Sampling	142,352	143,316,448
Open water	4441577	GS001b - Open Ocean - Sargasso Sea	Global Ocean Sampling	90,901	90,951,299
Open water	4441578	GS001c - Open Ocean - Sargasso Sea	Global Ocean Sampling	92,351	92,688,958
Open water	4441587	GS017 - Open Ocean - Yucatan Channel - Mexico	Global Ocean Sampling	257,581	281,259,325
Open water	4441588	GS018 - Open Ocean - Rosario Bank -	Global Ocean Sampling	142,743	156,474,992

Honduras					
Open water	4441592	GS022 - Open Ocean - Eastern Tropical Pacific	Global Ocean Sampling	121,662	131,079,270
Open water	4441594	GS026 - Open Ocean - Galapagos Islands	Global Ocean Sampling	102,708	109,049,397
Open water	4441607	GS110a - Open Ocean - Indian Ocean	Global Ocean Sampling	99,288	100,097,831
Open water	4441609	GS110a - Open Ocean - Indian Ocean	Global Ocean Sampling	99,781	101,818,659
Open water	4441610	GS110a - Open Ocean - Indian Ocean	Global Ocean Sampling	109,700	118,339,154
Open water	4441611	GS110a - Open Ocean - Indian Ocean	Global Ocean Sampling	348,823	345,285,679
Open water	4441614	GS110a - Open Ocean - Indian Ocean	Global Ocean Sampling	110,720	119,426,081
Open water	4441615	GS110a - Open Ocean - Indian Ocean	Global Ocean Sampling	101,558	105,196,135
Open water	4441616	GS110a - Open Ocean - Indian Ocean	Global Ocean Sampling	107,966	115,611,614
Open water	4441661	GS023 - Open Ocean - Eastern Tropical Pacific	Global Ocean Sampling	133,051	143,626,589
Open water	4443740	TA_34838	Sargasso Sea Bacterioplankton	94,851	16,575,969
Reef water	4441121	GS050 - Coral Atoll - Tikehau Lagoon - Fr. Polynesia	Global Ocean Sampling	715	755,429
Reef water	4441133	GS108b - Lagoon Reef -Coccos Keeling, Inside Lagoon - Australia	Global Ocean Sampling	49,595	53,530,124
Reef water	4441139	GS108a - Lagoon Reef Coccos Keeling, Inside Lagoon - Australia	Global Ocean Sampling	51,788	50,890,568
Reef water	4441167	GS048b - Coral Reef Moorea, Cooks Bay - Fr. Polynesia	Global Ocean Sampling	47,692	50,969,448

Reef water	4441593	GS025 - Fringing Reef - Dirty Rock, Cocos Island – Costa Rica	Global Ocean Sampling	120,671	129,781,299
Reef water	4441603	GS048a - Coral Reef -Moorea, Cooks Bay - Fr. Polynesia	Global Ocean Sampling	90,515	92,813,604
Reef water	4441604	GS051 - Coral Reef Atoll - Rangirora Atoll - Fr. Polynesia	Global Ocean Sampling	128,982	140,497,312
Reef water	4441617	GS148 - Fringing Reef East coast Zanzibar Tanzania	Global Ocean Sampling	107,741	107,616,215
Reef water	4442642	King14LIMic20070829	Northern Line Islands	108029	31667620
Reef water	4442643	King2LIMic20070817	Northern Line Islands	97767	37285824
Reef water	4442647	Xmas16LIMic20070729	Northern Line Islands	53169	19900801
Reef water	4442648	Xmas29LIMic20070805	Northern Line Islands	111061	38238805
Reef water	4442649	Xmas35LIMic20070808	Northern Line Islands	44544	15484390
Reef water	4442650	Xmas6LIMic20070721	Northern Line Islands	118943	39280406
Reef water	4442651	XmasLag1LIMic20070720	Northern Line Islands	60531	21801386
Reef water	4442652	King7LIMic20070821	Northern Line Islands	181525	42145245
Reef water	4442653	King8LIMic20070823	Northern Line Islands	119830	37606997
Reef water	4440037	KingLIMic20050821	Northern Line Islands	188,445	19,753,735
Reef water	4440039	PalmLIMic20050818	Northern Line Islands	289,723	30,795,962
Reef water	4440041	XmasLIMic20050805	Northern Line Islands	227,542	23,693,344
Spring	4442583	OCTOPUS	Yellowstone National Park	22,272	22,557,192
Spring	4443746	MushroomSpringsMatCoreB	Yellowstone National Park	2,708	2,713,791
Spring	4443747	MushroomSpringsMatCoreD	Yellowstone National Park	320	325,932

Spring	4443749	OctopusSpringsMatCoreF	Yellowstone National Park	19,124	18,644,488
Spring	4443750	OctopusSpringsMatCoreR	Yellowstone National Park	1,266	1,328,730
Spring	4443762	MushroomSpringsMatCoreF	Yellowstone National Park	6,521	6,493,181
Animal associated	4441679	Cow rumen -- 640F6	Cow rumen	264,849	26,644,817
Animal associated	4441680	Cow rumen -- 80F6	Cow rumen	178,713	18,153,371
Animal associated	4441681	Cow rumen -- 710F6	Cow rumen	345,317	35,115,534
Animal associated	4441682	Cow rumen -- Pooled Planktonic	Cow rumen	236,830	24,016,021
Animal associated	4440283	Chicken Cecum A	FS-CAP	294,682	30,657,259
Animal associated	4440284	Chicken Cecum B	FS-CAP	237,940	24,707,007
Animal associated	4440463	LeanMouseCecumMic2005	Human feces - Turnbaugh	10,845	8,478,662
Animal associated	4440464	ObeseMouseCecumMic2005	Human feces - Turnbaugh	11,857	9,067,143
Animal associated	4440056	FishMorGutKentSTMIC20060504	Fish stomach	60,311	5,956,666

Supplemental Table 2. The samples present in each of the clustered identified by the *K*-means analysis with *K* of six. This was chosen because the silhouette analysis suggested that six clusters were the most appropriate (**Supplemental Fig. 2**). There were 33 human, 9 terrestrial animal, 10 mat community, 42 open water, 20 reef water, 60 coastal water, 5 deep water, 7 fresh water, 15 hypersaline, and 6 hot spring samples in total.

Cluster	Number of metagenomes	Original metagenome classification
1	52	31 human 5 terrestrial animals 6 mat community Water samples: <ul style="list-style-type: none"> • 4 open • 3 reef • 2 coastal • 1 fresh
2	1	1 reef water sample
3	1	1 reef water sample
4	3	1 human 1 fresh water 1 reef water
5	149	4 mat 4 terrestrial animals 1 human Water samples: <ul style="list-style-type: none"> • 56 coastal • 5 deep • 15 hypersaline • 6 spring • 38 open • 13 reef • 7 fresh
6	6	Water samples: <ul style="list-style-type: none"> • 2 coastal • 3 fresh • 1 reef

Supplementary Table 3: Tree size and average deviance from a series of tree cross-validation experiments.

Tree Size	Average CV Deviance
1	152.014
2	122.432
3	102.636
4	99.642
6	92.762
8	92.970
9	92.812
14	95.848
16	98.342
17	98.622

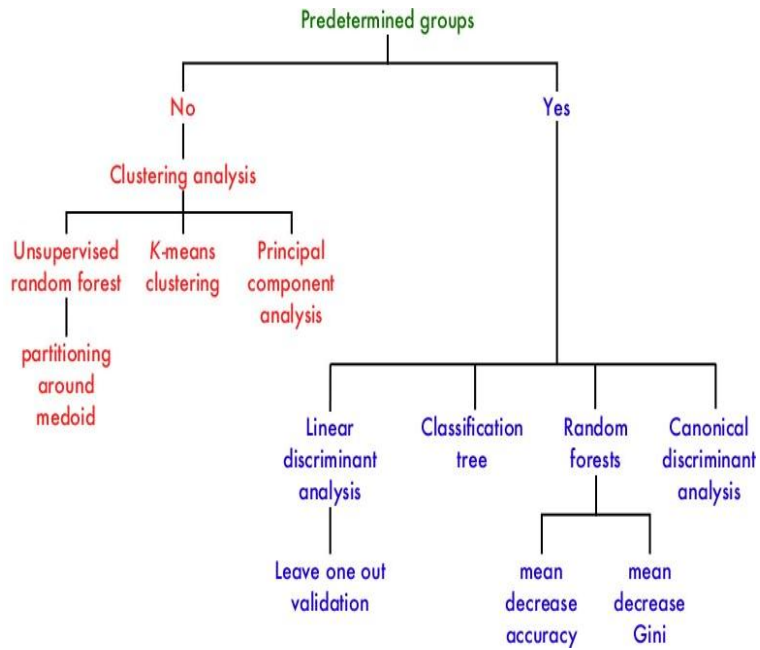
Supplementary Table 4. The group that each metagenome was assigned to by the random forest analysis.

Initial classification	Classification from the random forest								
	Mixed marine	Deep water	Coastal marine	Open marine	Spring water	Terrestrial animals	Human associated	Fresh water	Hyper-saline
Freshwater	3				1	1			
Open marine	6	1	1	31					2
Spring water	1				5				
Coastal marine	6	1	43	8	2				
Terrestrial animal						5 cow 2 mice	3 mice 1 fish		
Human associated	1		1				32		
Mat community	4	1						4	
Deep marine		4	1						
Reef water	4	1		15					
hypersaline	4				1				9
Total	29	8	47	44	8	8	36	10	11

Supplementary Table 5. The misclassification table generated by the canonical discriminant analysis.

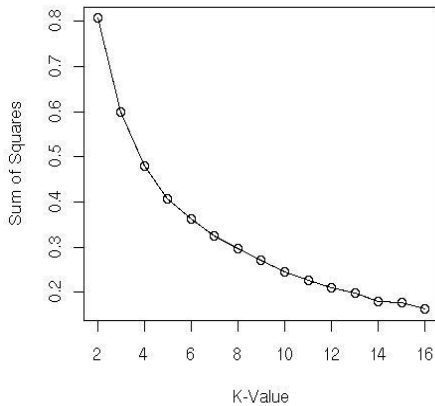
	coastal	deep	fresh	human	hypersaline	mat	open	reef	spring	Terrestrial animal	Class error
coastal	9.820	0.000	0.301	0.391	0.000	0.226	0.962	0.009	0.127	0.160	0.181
deep	0.990	0.004	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.995
fresh	0.816	0.000	0.433	0.231	0.000	0.235	0.081	0.028	0.160	0.075	0.783
human	0.000	0.000	0.207	6.268	0.000	0.457	0.014	0.051	0.000	0.000	0.104
hypersaline	1.231	0.000	0.000	0.000	1.485	0.000	0.283	0.000	0.000	0.000	0.504
mat community	0.382	0.000	0.000	0.004	0.000	1.613	0.000	0.000	0.000	0.000	0.193
open	4.377	0.009	0.033	0.448	0.169	0.349	2.410	0.169	0.014	0.018	0.698
reef	1.509	0.009	0.283	0.429	0.000	0.226	1.117	0.235	0.023	0.377	0.994
spring	0.047	0.000	0.000	0.000	0.000	0.000	0.113	0.004	0.834	0.000	0.165
terrestrial	0.287	0.000	0.108	1.193	0.000	0.216	0.000	0.000	0.000	0.193	0.903

Supplemental Figure 1. A diagram of the relationship between the seven statistical methods evaluated.

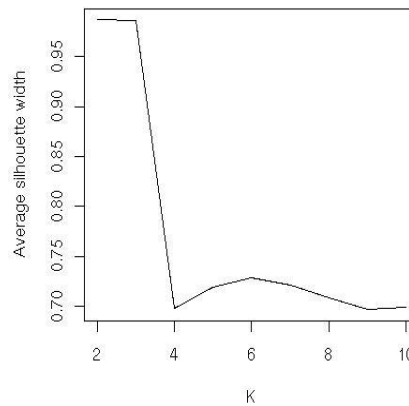


Supplemental Fig. 2. (a) The sums of squares and K -value used to identify the number of groups that the samples should be split into. No clear elbow was evident; therefore silhouette plots were used to examine the data. **(b)** A silhouette plot showing how it creates metagenomic groups in the data. The most favorable grouping number is where the average silhouette width is nearest to one. **(c)** The variation of average silhouette width and K . There is a peak at $K=6$ and an uptick at $K=10$.

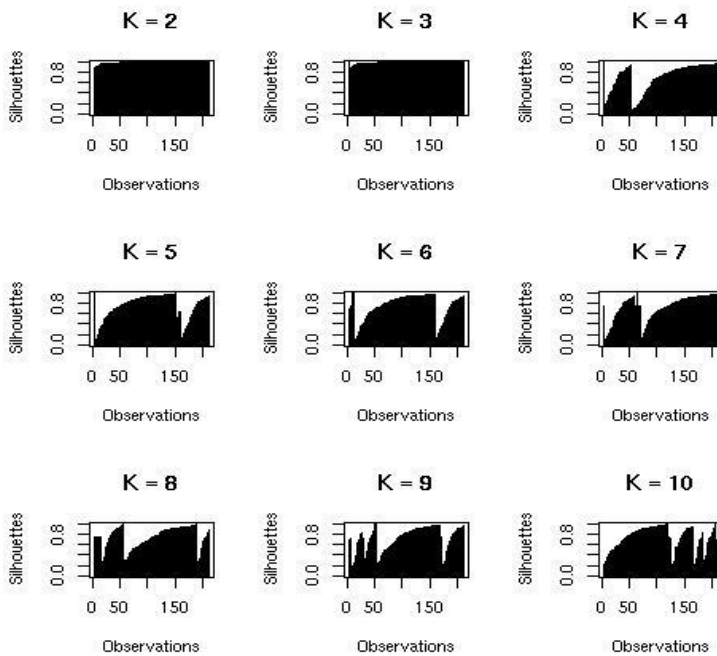
A)



C)

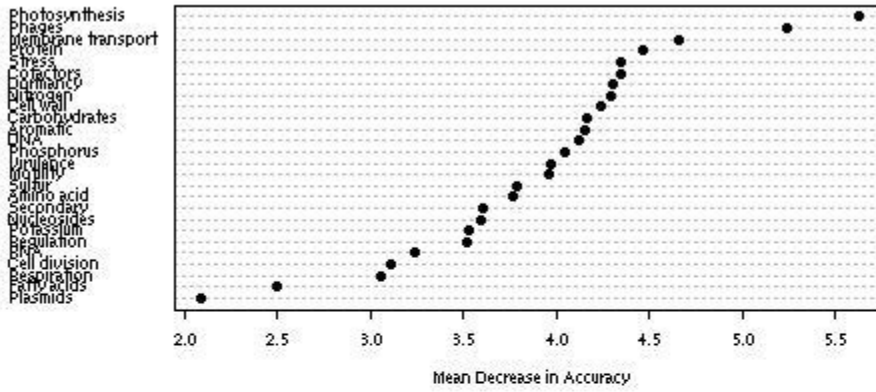


B)

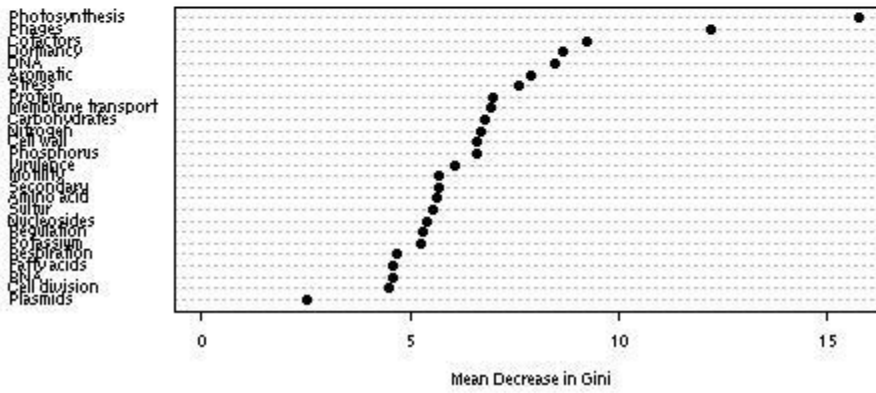


Supplemental Fig. 3 Mean decrease in (a) accuracy and (b) Gini determined by the random forest analysis for the variables.

A



B



Supplemental Fig. 4. Linear discriminant analysis of the environmental samples.

