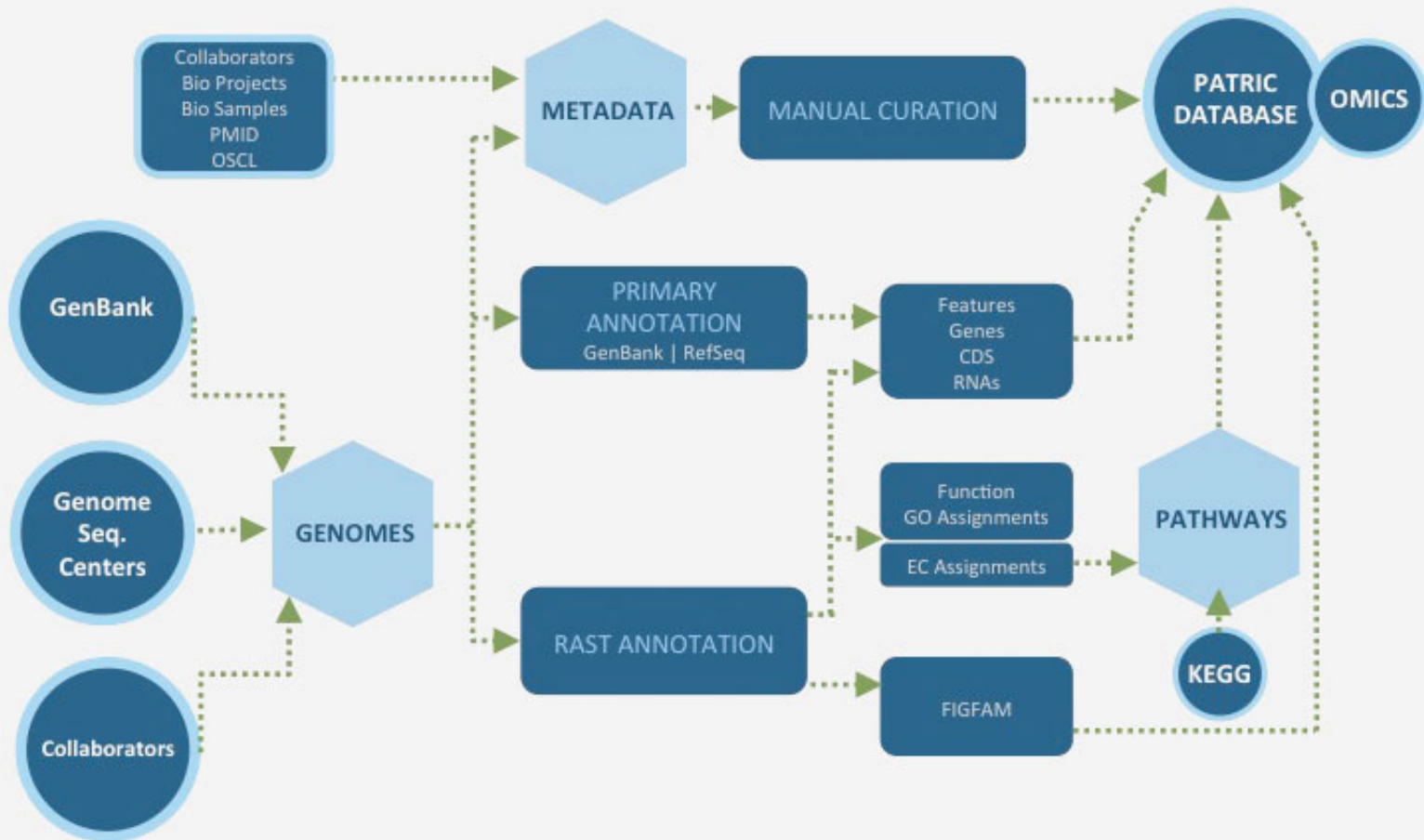


Tara Oceans Update 2/9/15

PATRIC Genomes and Annotations



Building the reference database

	start	HQ genomes	genome clustering	nr genomes	Representative genes
Genomes (1e3):	21.14	18.82	18.82	13.87	13.87
Genes (1e6):	80.03	69.52	69.52	51.26	17.08

- HQ genomes: contigs < 300; n50 > 10,000; phyeco bacterial markers > 100
- Genome clustering: average linkage clustering across 20 phyeco bacterial markers
- NR Genomes: at most 500 genomes per genome cluster
- Representative genes: used UCLUST to cluster genes within each genome cluster at 90% identity

DIAMOND vs RAPsearch2

Expected performance

Dataset	Reads	Length	Program	Speedup	% Queries (10^{-3})	% Matches (10^{-3})	% Queries (10^{-5})	% Matches (10^{-5})
Illumina (HMP)	5M	101	D-fast	19880	97.69%	79.52%	99.22%	90.25%
			D-sens	2516	99.00%	95.38%	99.60%	97.21%
			R-fast	603	96.86%	68.59%	98.43%	81.35%
			R-sens	43	99.36%	92.74%	99.88%	96.74%
Illumina (Permafrost)	6M	114	D-fast	22505	91.04%	68.65%	94.72%	80.16%
			D-sens	2120	98.54%	92.11%	99.34%	94.16%
			R-fast	582	82.59%	55.96%	87.74%	67.85%
			R-sens	45	95.85%	86.96%	98.71%	92.35%

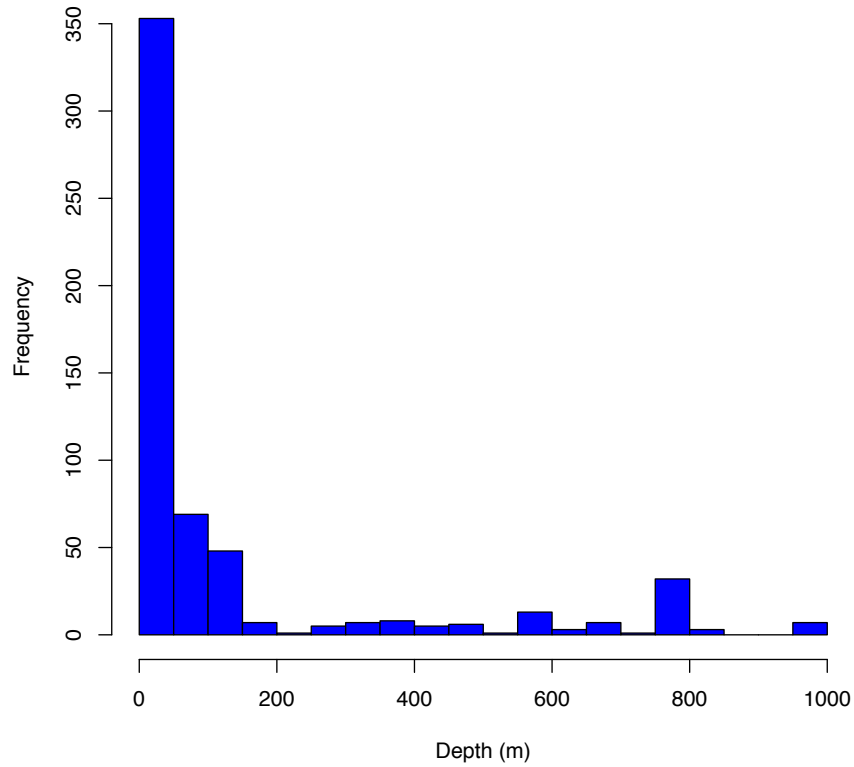
Observed performance

Reads	RAPsearch2	DIAMOND
	Time (minutes)	
10	5	271
100	6	260
1000	8	263
10000	23	266
100000	150	291

Supplementary Table 1 Comparison of four different programs and settings, D-fast (DIAMOND, fast setting), D-sens (DIAMOND, sensitive setting), R-fast (RAPSearch2, fast setting) and R-sens (RAPSearch2, sensitive setting), against BLASTX. For six different datasets, each with the specified number of reads ("# Reads") and average read length ("Length"), we report the speedup of each program over BLASTX ("Speedup"), the percentage of queries for which the program reports at least one alignment ("% Queries") and the percentage of matches recovered by the program ("% Matches"). The two latter quantities are reported both for an e-value threshold of 10^{-3} and 10^{-5} . Programs were set to report alignments for up to 250 target sequences per read. Times are wall-clock times on a server using 48 cores and exclude one-time program startup overhead, which was less than a minute for BLASTX and five minutes for DIAMOND-fast.

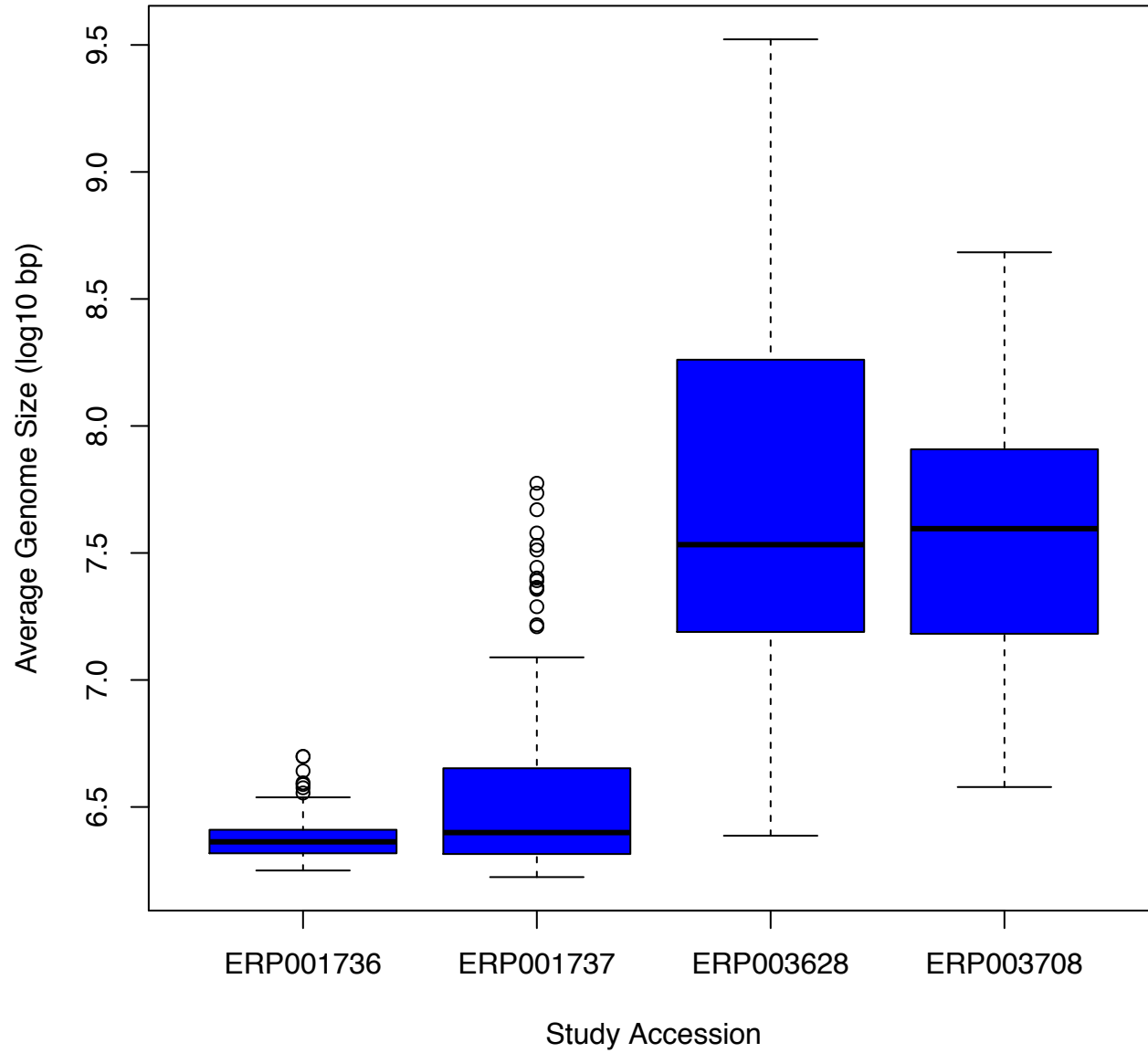
Tara Oceans Update 1/26/15

Depth

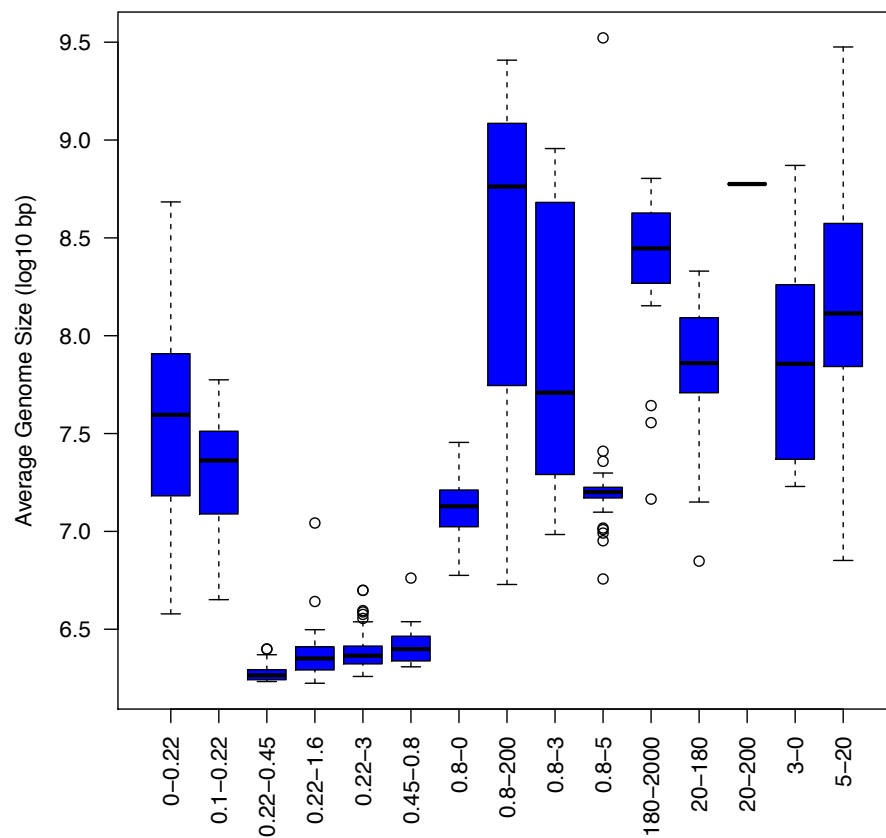
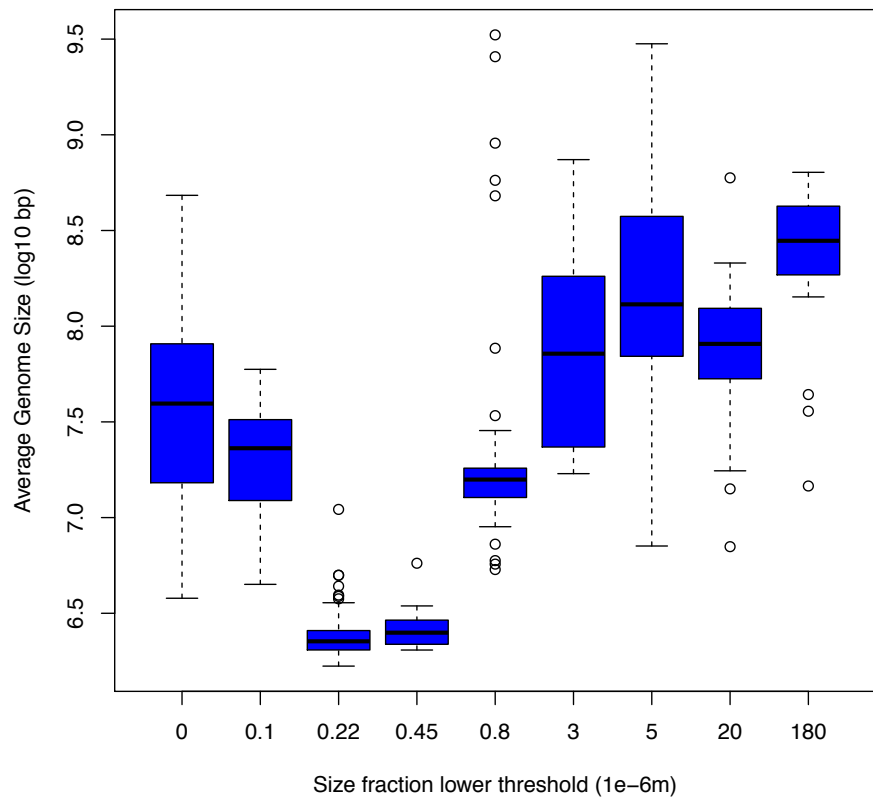


Size fraction (um)

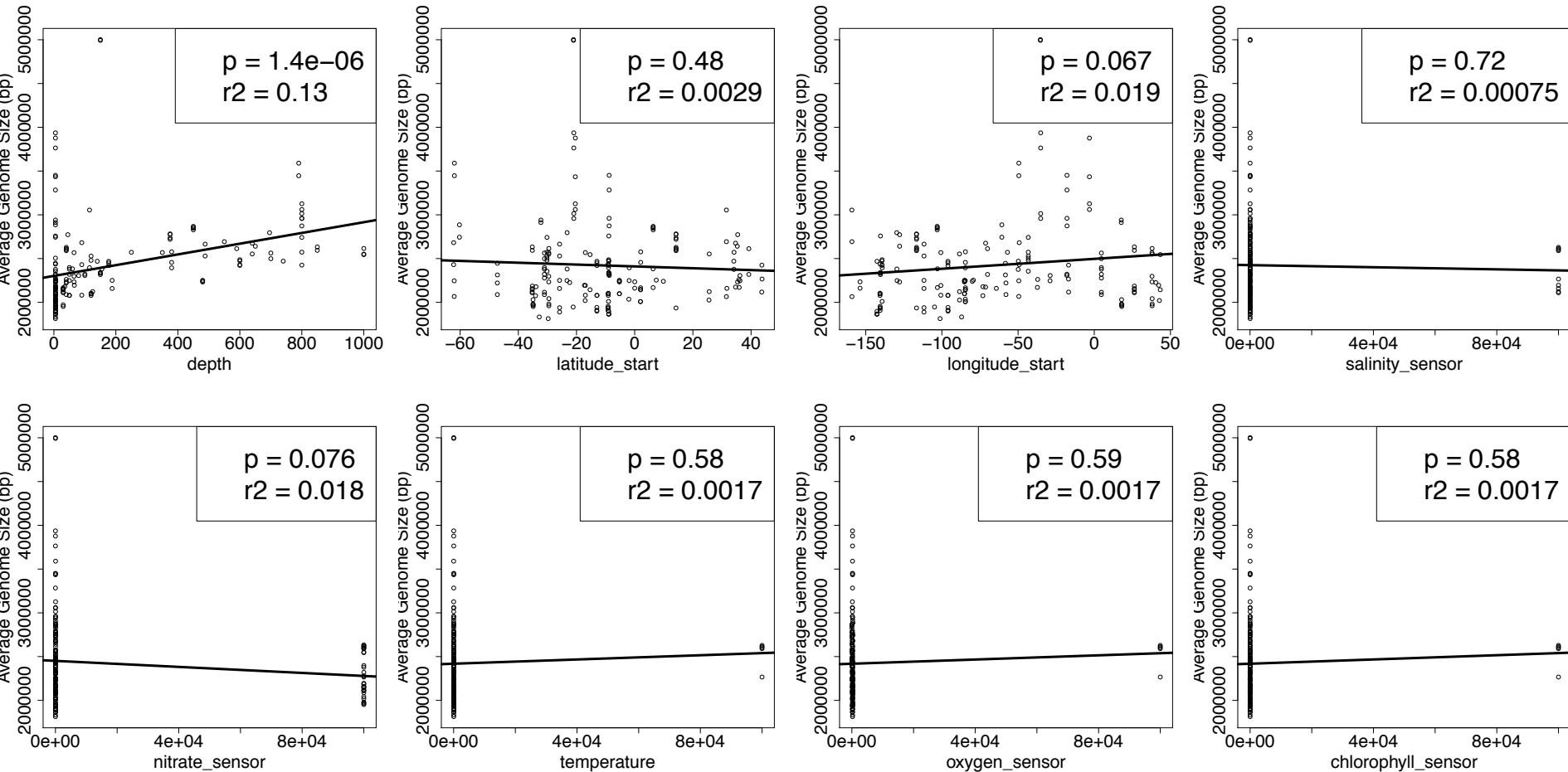
lower	upper	count
0	0.22	113
0.1	0.22	21
0.22	0.45	19
0.22	1.6	93
0.22	3	177
0.45	0.8	21
0.8	0	15
0.8	3	6
0.8	5	40
0.8	200	3
3	0	6
5	20	22
20	180	18
20	200	1
180	2000	21



Size fraction (um)



AGS Covariates



- Conditioned on 0.22 – 3.0 μm size fraction

AGS Covariates

Covariate	Estimate	Std. Error	t value	p value	significance
nitrate_sensor	-5.0E+00	1.3E+00	-4.0E+00	9.5E-05	***
longitude_start	2.7E+03	7.7E+02	3.5E+00	5.1E-04	***
depth	4.2E+02	1.8E+02	2.3E+00	2.3E-02	*
chlorophyll_sensor	6.1E+03	6.2E+03	9.9E-01	3.3E-01	
oxygen_sensor	-5.2E+02	5.5E+02	-9.6E-01	3.4E-01	
temperature	-5.6E+03	6.1E+03	-9.1E-01	3.6E-01	
latitude_start	-1.3E+03	1.6E+03	-8.1E-01	4.2E-01	
salinity_sensor	-1.5E-01	2.4E+00	-6.4E-02	9.5E-01	

- Conditioned on 0.22 – 3.0 um size fraction
- R-squared = 0.24
- $\text{lm}(\text{formula} = \text{average_genome_size} \sim \text{depth} + \text{latitude_start} + \text{longitude_start} + \text{salinity_sensor} + \text{nitrate_sensor} + \text{temperature} + \text{oxygen_sensor} + \text{chlorophyll_sensor})$

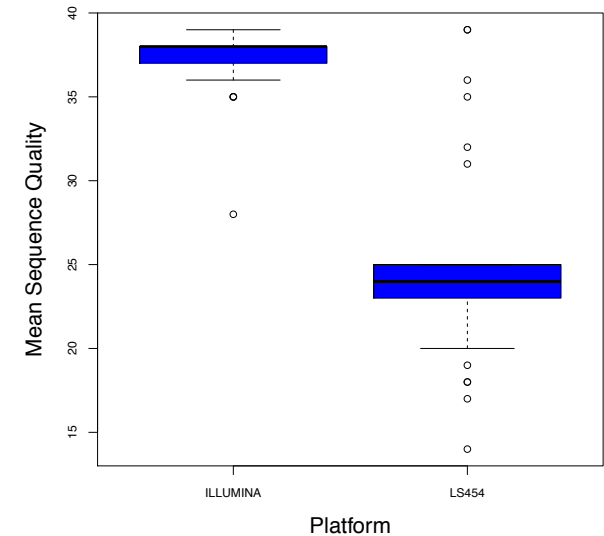
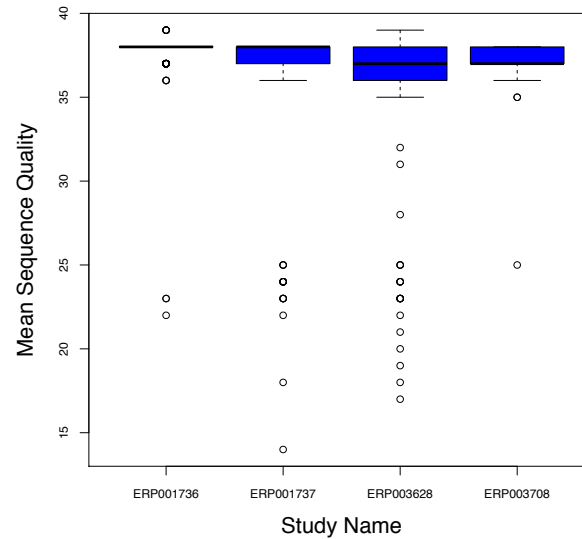
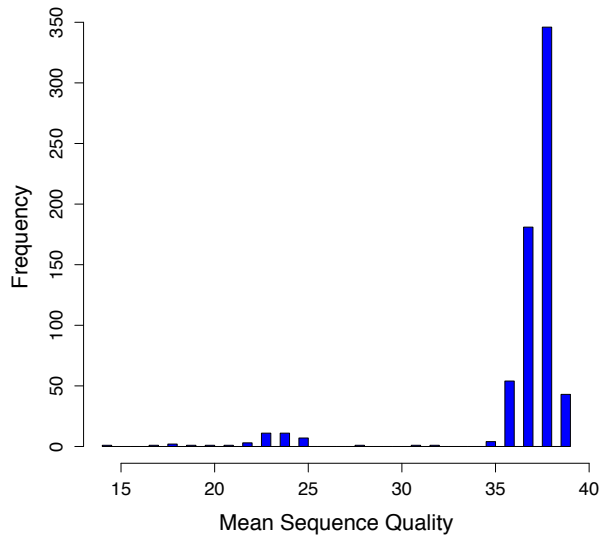
Tara Oceans Update 1/21/15

So, when was the data actually submitted...

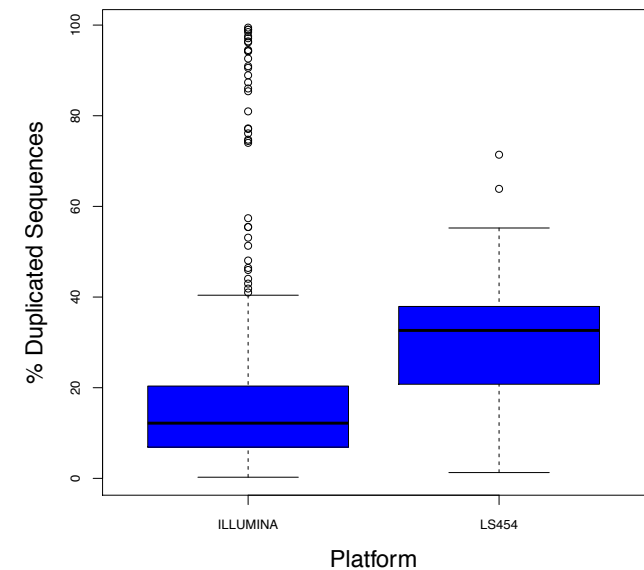
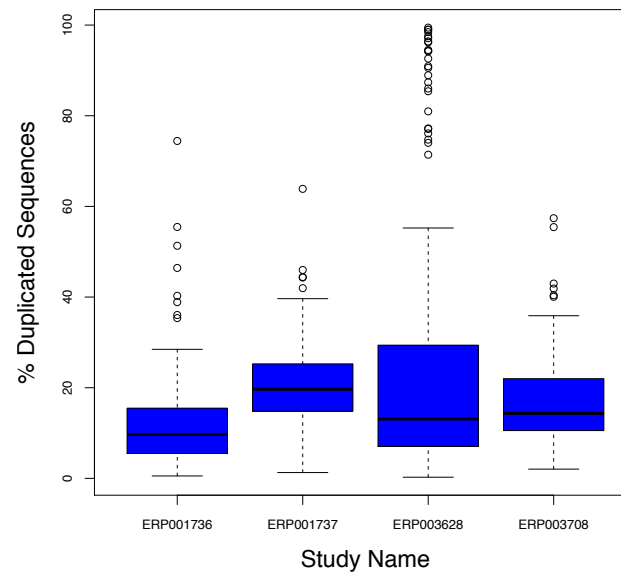
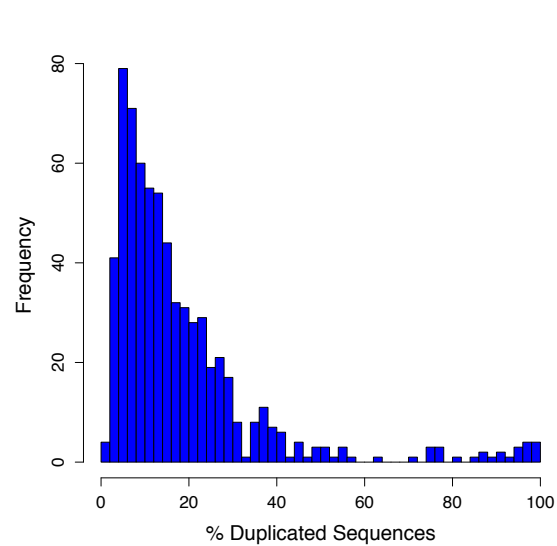
submission accession	count runs	submission date
ERA357871	235	1-Sep-14
ERA357903	104	10-Sep-14
ERA356217	66	1-Sep-14
ERA356219	64	1-Sep-14
ERA357908	48	
ERA242243	36	
ERA242028	30	
ERA357902	29	
ERA241291	23	
ERA320049	19	
ERA155563	17	
ERA242034	6	
ERA242244	4	
ERA155562	3	
ERA242033	2	
ERA245027	2	
ERA320050	1	
ERA352812	1	
ERA361805	1	

<http://www.ebi.ac.uk/ena/data/view/ERA356219>

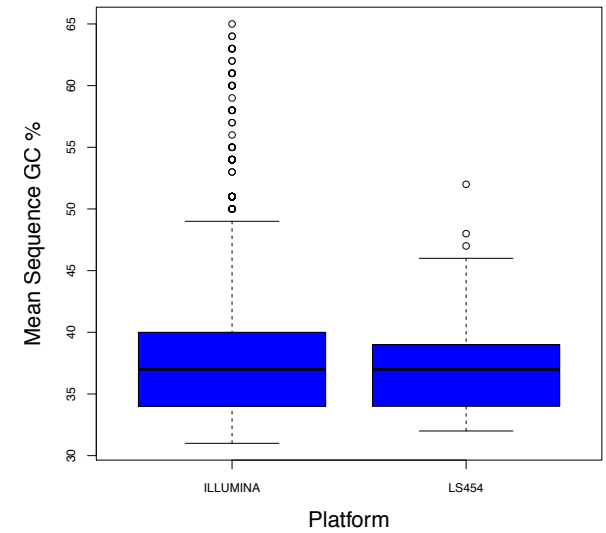
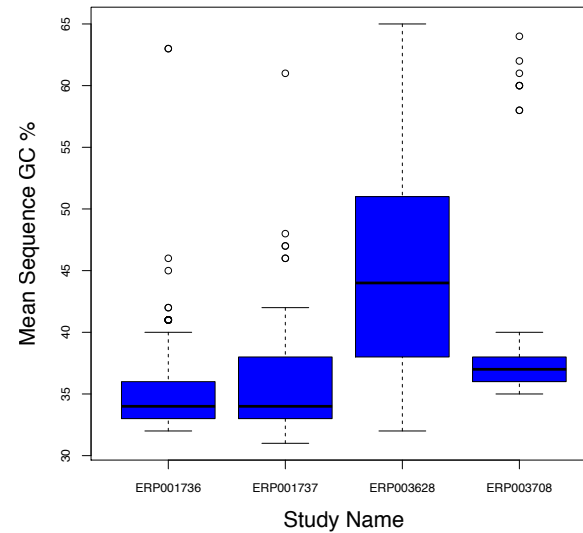
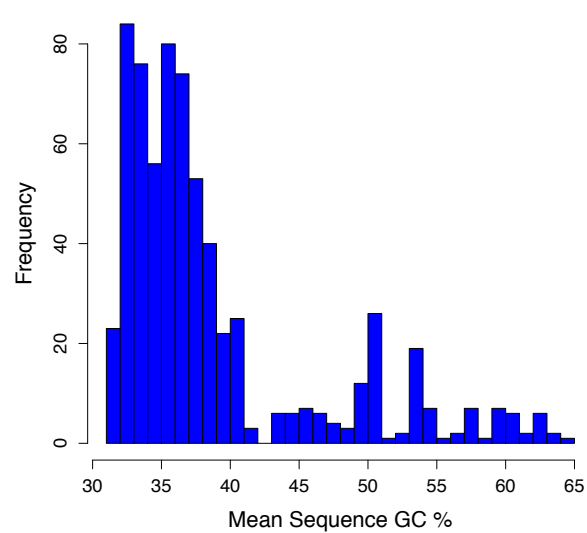
Distribution of Average Read Quality Across Runs



Distribution of % Duplicated Across Runs

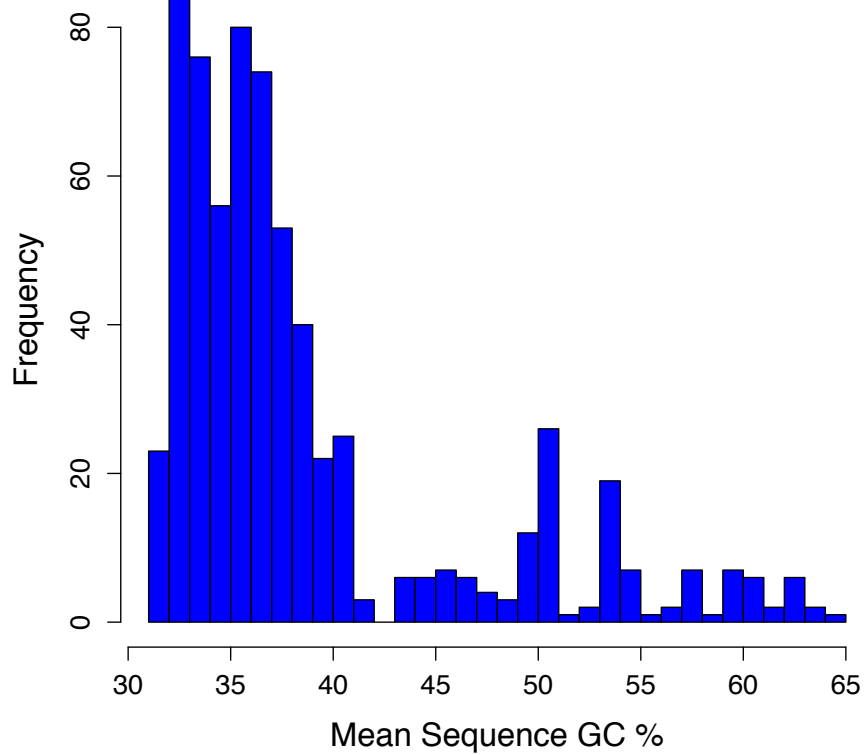


Distribution of Average % GC Across Runs

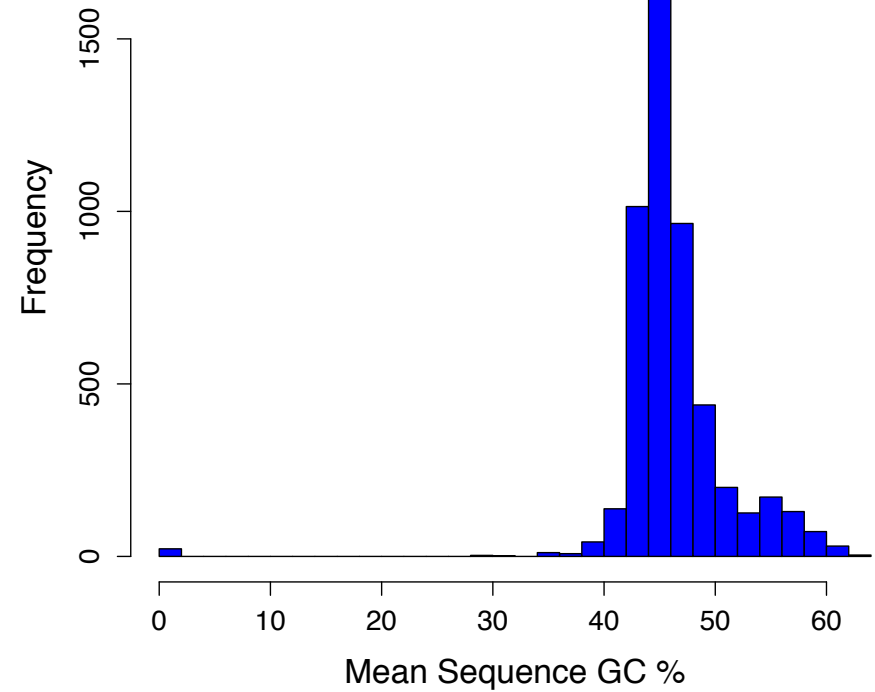


Different %GC Between Marine and Human Gut

Tara Oceans



Human Gut



Fraction of runs which failed FastQC checks

study_accession	ERP001736	ERP001737	ERP003628	ERP003708
count_runs	248	85	223	114
Basic Statistics	0.00	0.00	0.00	0.00
Per base sequence quality	0.01	0.25	0.19	0.02
Per sequence quality scores	0.00	0.02	0.02	0.00
Per base sequence content	0.25	0.41	0.10	0.04
Per base GC content	0.02	0.00	0.12	0.04
Per sequence GC content	0.64	0.41	0.26	0.50
Per base N content	0.00	0.01	0.01	0.00
Sequence Length Distribution	0.00	0.00	0.00	0.00
Sequence Duplication Levels	0.01	0.01	0.13	0.02
Overrepresented sequences	0.00	0.00	0.03	0.00
Kmer Content	0.00	0.14	0.16	0.02

Tara Oceans Update 1/12/15

Summary of Available Tara Oceans Data

study title	data type	size fraction	study accession	reads_1e6	bases_1e9	runs	experiments	samples
Shotgun Sequencing of size fractions for prokaryotes and large DNA viruses	Metagenome	prokaryotes	ERP001736	24,764	4,892	248	248	135
Shotgun Sequencing of size fractions for prokaryotes and large DNA viruses	Metagenome	prokaryotes	ERP001737	8,972	1,801	85	85	75
Shotgun Sequencing of size fractions for protist	Metagenome	protists	ERP003628	23,518	4,660	243	243	159
Shotgun Sequencing of size fractions for DNA viruses	Metagenome	DNA viruses	ERP003708	4,489	863	115	114	92
Metatranscriptome sequencing of size fractions for protists.	Metatranscriptome	protists	ERP006152	284	52	27	27	22
Metatranscriptome sequencing of size fractions for protists.	Metatranscriptome	protists	ERP006156	1	0	29	29	24
MAST4 Single Cell Whole Genome Amplification	Single-cell		ERP003711	125	12	4	4	2
16S Sequencing of size fractions for prokaryotes	16S	prokaryotes	ERP003634	0	0	6	6	6
Amplicon sequencing of size fractions for large DNA viruses.	16S	DNA viruses	ERP006153	2	1	35	35	19
Amplicon sequencing of size fractions for protists.	16S	protists	ERP006157	993	293	370	370	334
Amplicon sequencing of size fractions for protists.	16S	protists	ERP007024	24	6	20	20	14

“This project will rely on samples collected during the scientific expedition Tara-Oceans (2009-2012). By March 2012, the schooner Tara, equipped with innovative systems for sampling of 11 organismal size-ranges covering entire planktonic communities from viruses to animals, has collected standardized genetic (total DNA/RNA), morphological, and physico-chemical (contextual) samples from 153 sites across the world oceans, locations carefully selected with input from near-real-time remote sensing and in-situ hydrographic criteria. Overall, a total of ~50,000 biological samples and ~13,000 contextual measures from 3 depths will be analysed. The metagenomics component of the project consists of size fractionated plankton samples that are submitted to barcoding and shotgun sequencing, as well as isolated single-cell amplified protists and single-organisms isolated metazoans that are sequenced as reference genomes.”

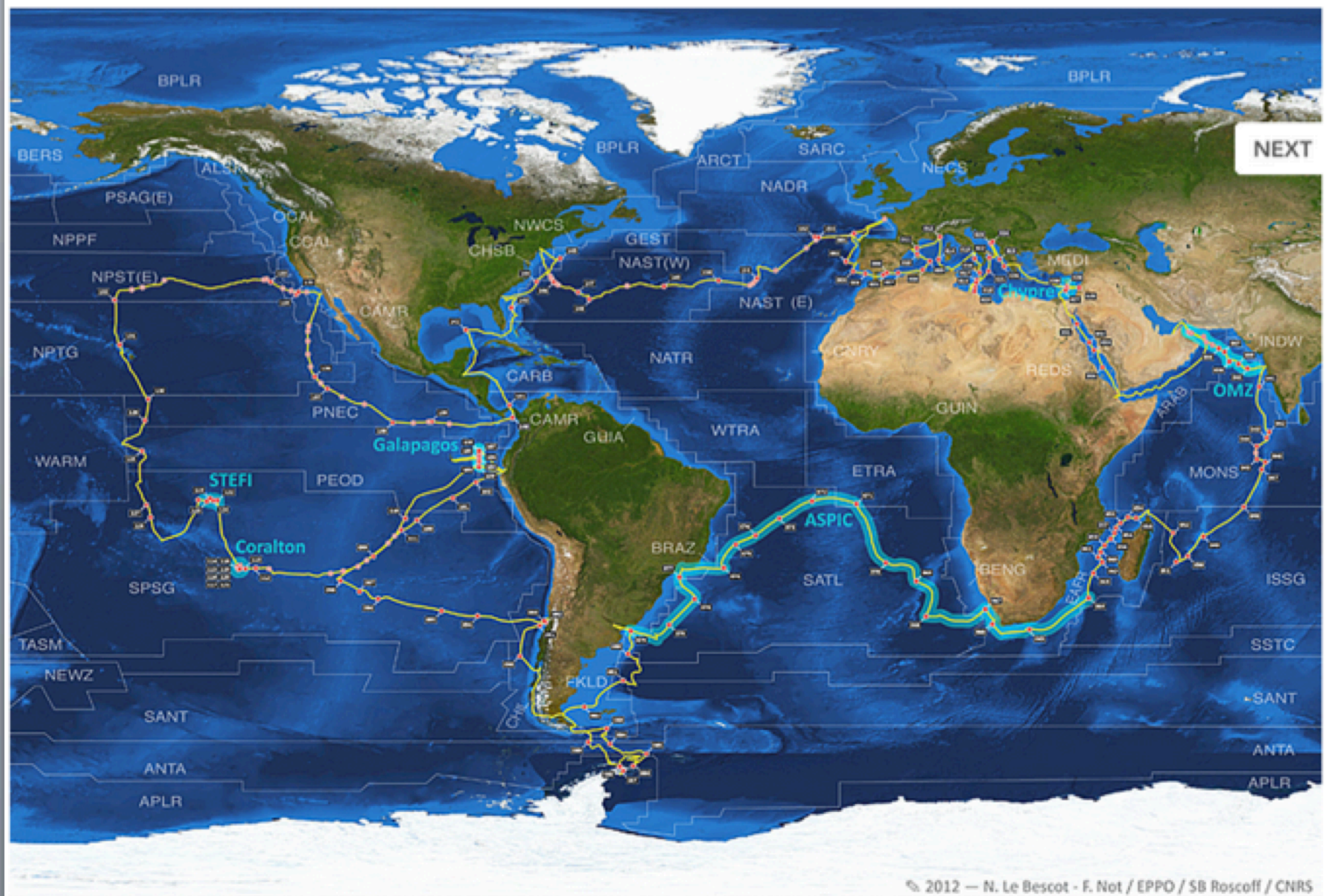
Data release policy

(http://www.embl.de/tara-oceans/start/facts_figures/Tara-Oceans-Data-Release-Policy.pdf)

The Tara Oceans Consortium has established a policy for pre-publication data usage with the aim to make quickly and broadly available the global data generated during the project for advancing science, while maintaining a priority for analyses by the laboratories that participate in the data production effort. **During one year after release in a public database, the data will be considered usable with the following restriction:** specifically, intended analyses may not compete with those reserved for the Tara Oceans Consortium, including global analyses at the genome and community level, and comparisons between samples. Therefore, **before starting any study on these reserved data sets, investigators are required to contact the Tara Oceans Consortium to obtain an agreement.**

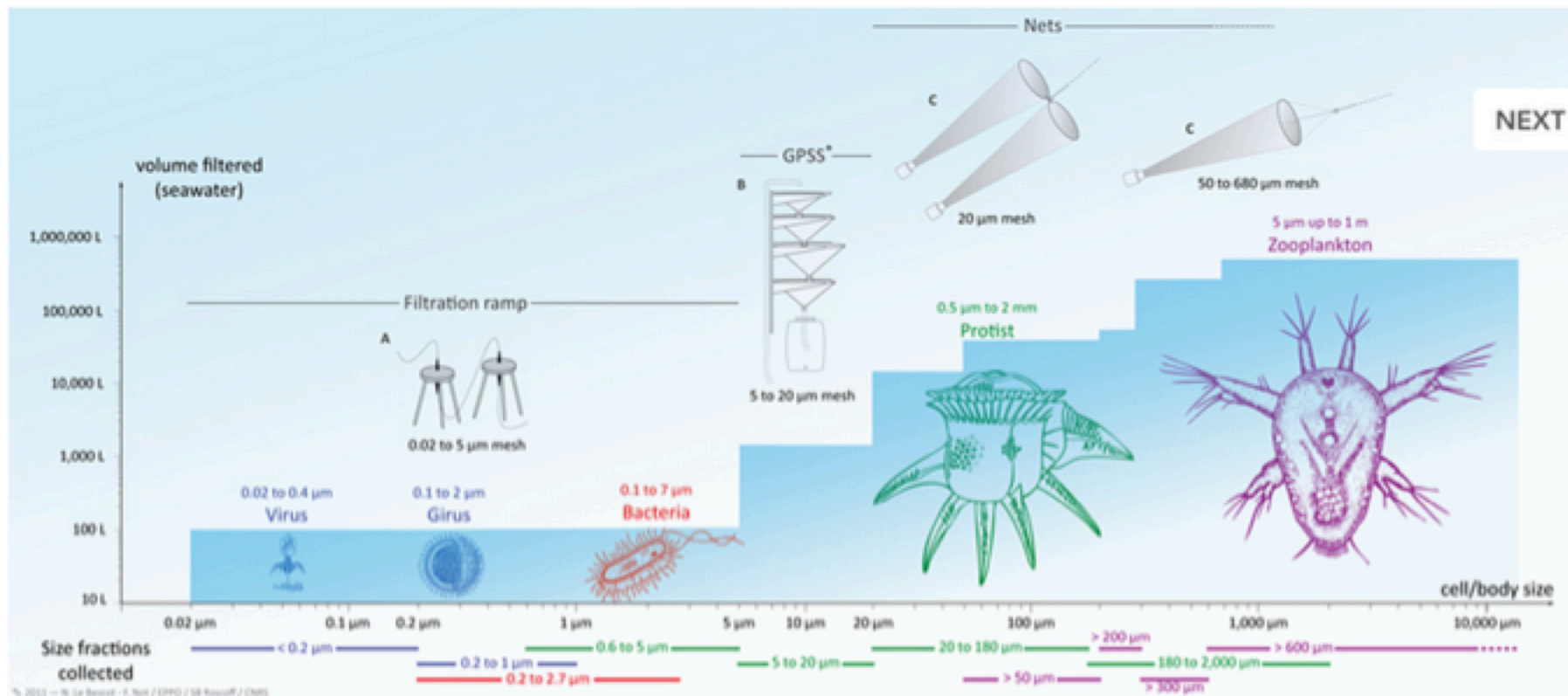
This restriction will be lifted after the publication of analyses on the considered data sets by the Tara Oceans Consortium, or automatically one year after the release of the data. If a publication is made from the released data sets, it is expected that the Tara Oceans Consortium is cited in the acknowledgment section. Raw data sets have the same limits for use than the advanced data sets derived from them. Download of data from the Tara Oceans project implies a full acceptance of the policy usage described in this paragraph.

- All runs in the SRA were updated at some point in 2014
- Run dates range from 3/19/2010 to 11/20/2013 (one in 2014)



Voyage of the schooner TARA during the TARA Oceans expedition, showing station numbers (red dots) and specific experiments (in blue and named)





TARA OCEANS organisms
Image 1 of 4



<http://www.embl.de/tara-oceans/start/publications/index.html>

Karsenti, E. et al. A Holistic Approach to Marine Eco-Systems Biology. *Plos Biol* 9, doi:ARTN e1001177 DOI 10.1371/journal.pbio.1001177 (2011).

Corse, E. et al. Phylogenetic Analysis of Thecosomata Blainville, 1824 (Holoplanktonic Opisthobranchia) Using Morphological and Molecular Data. *Plos One* 8, doi:ARTN e59439 DOI 10.1371/journal.pone.0059439 (2013).

Hingamp, P. et al. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *Isme J* 7, 1678-1695, doi:DOI 10.1038/ismej.2013.59 (2013).

Brum, J. R. et al. Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *Isme J* 7, 1738-1751, doi:DOI 10.1038/ismej.2013.67 (2013).

Solonenko, S. A. et al. Sequencing platform and library preparation choices impact viral metagenomes. *Bmc Genomics* 14, doi:Artn 320 Doi 10.1186/1471-2164-14-320 (2013).

Swan, B. K. et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *P Natl Acad Sci USA* 110, 11463-11468, doi:DOI 10.1073/pnas.1304246110 (2013).

Benzoni, F. *Echinophyllia tarae* sp n. (Cnidaria, Anthozoa, Scleractinia), a new reef coral species from the Gambier Islands, French Polynesia. *Zookeys*, 59-79, doi:DOI 10.3897/zookeys.318.5351 (2013).

Logares, R. et al. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*, n/a-n/a, doi:10.1111/1462-2920.12250 (2013).

Werdell, P. J. et al. Underway sampling of marine inherent optical properties on the Tara Oceans expedition as a novel resource for ocean color satellite data product validation. *Methods in Oceanography* 7, 40-51, doi:http://dx.doi.org/10.1016/j.mio.2013.09.001 (2013).

Abida, H. et al. Bioprospecting Marine Plankton. *Mar Drugs* 11, 4594-4611, doi:Doi 10.3390/Md11114594 (2013).

Boss, E. et al. The characteristics of particulate absorption, scattering and attenuation coefficients in the surface ocean; Contribution of the Tara Oceans expedition. *Methods in Oceanography* 7, 52-62, doi:http://dx.doi.org/10.1016/j.mio.2013.11.002 (2013).

Chase, A. et al. Decomposition of in situ particulate absorption spectra. *Methods in Oceanography* 7, 110-124, doi:http://dx.doi.org/10.1016/j.mio.2014.02.002 (2013).

Clerissi, C. et al. Unveiling of the Diversity of Prasinoviruses (Phycodnaviridae) in Marine Samples by Using High-Throughput Sequencing Analyses of PCR-Amplified DNA Polymerase and Major Capsid Protein Genes. *Appl Environ Microb* 80, 3150-3160, doi:Doi 10.1128/Aem.00123-14 (2014).

Benzoni, F. et al. Phylogenetic relationships and revision of the genus *Blastomussa* (Cnidaria: Anthozoa: Scleractinia) with description of a new species. *RAFFLES BULLETIN OF ZOOLOGY*, 358–378 (2014).

Cunningham, B. R. et al.

Acinas, S. G. et al. Validation of a new catalysed reporter deposition-fluorescence in situ hybridization probe for the accurate quantification of marine Bacteroidetes populations. *Environ Microbiol*, doi:10.1111/1462-2920.12517 (2014).

Arrigoni, R. et al. A phylogeny reconstruction of the Dendrophylliidae (Cnidaria, Scleractinia) based on molecular and micromorphological criteria, and its ecological implications. *Zoologica Scripta*, n/a-n/a, doi:10.1111/zsc.12072 (2014).

Arrigoni, R. et al. Phylogenetic relationships and taxonomy of the coral genera *Australomussa* and *Parascolumia* (Scleractinia, Lobophylliidae). *Contributions to Zoology* (2014).

Gasmi, S. et al. Evolutionary history of Chaetognatha inferred from molecular and morphological data: a case study for body plan simplification. *Frontiers in zoology* 11, 84, doi:10.1186/s12983-014-0084-7 (2014).

Roullier, F. et al. Particle size distribution and estimated carbon flux across the Arabian Sea oxygen minimum zone. *Biogeosciences* 11, 4541-4557, doi:DOI 10.5194/bg-11-4541-2014 (2014).

Environmental data stored at Pangaea

<http://www.pangaea.de/search?env=All&q=project%3Alabel%3ATara+Oceans>

Not logged in ([log in](#) or [sign up](#))



All Water Sediment Ice Atmosphere
project:label:Tara Oceans Search
Help Advanced Search Preferences more...

Always quote citation when using data!

More than 10000 datasets found on search for »project:label:Tara Oceans«

[Show Map](#) [Google Earth](#) [Data Warehouse](#)

Your query requires all search terms to be in the results, but produced no hits. Because of this, the behaviour was changed to show a ranked list of hits with **any of the terms**. To enforce query terms in the results, you may prefix them by "+".

<< PREV | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | NEXT >>

1. **Tara Oceans Expedition, P; Tara Oceans Consortium, C (2014):** Registry of selected samples from the Tara Oceans Expedition (2009-2013)

Related to: Brum, J; Tara Oceans Consortium, C; et al.: Global patterns and ecological drivers of ocean viral communities. *Science*
de Vargas, C; Tara Oceans Consortium, C; et al.: Sea change in eukaryotic plankton diversity. *Science*
Lima Mendez, G; Tara Oceans Consortium, C; et al.: Top-down determinants of community structure in the global plankton interactome. *Science*
(and more)
Size: 10784 data points
Dataset #840721 (DOI registration in progress) - Score: 100% - [Similar datasets](#)

2. **Morozov, NP (1968):** Lithium and rubidium concentrations in waters of seas and oceans

Supplement to: Morozov, NP (1968): Geochemistry of rare alkaline elements in the oceans and seas. *Oceanology*
Size: 4 datasets
doi:10.1594/PANGAEA.753510 - Score: 71% - [Similar datasets](#)

3. **Johnson, DA; Nigrini, CA (1985):** Neogene radiolarian datum levels in the equatorial Indian and Pacific Oceans

Supplement to: Johnson, DA; Nigrini, CA (1985): Synchronous and time-transgressive Neogene radiolarian datum levels in the Equatorial Indian and Pacific Oceans. *Marine Micropaleontology*
Size: 9 datasets
doi:10.1594/PANGAEA.690483 - Score: 58% - [Similar datasets](#)

4. **Schmiedl, G; Mackensen, A (1993):** Observational data around large miliolid foraminifera from the polar oceans

Supplement to: Schmiedl, G; Mackensen, A (1993): Cornuspiroides striolatus (BRADY) and C. rotundus nov. spec.: Large miliolid foraminifera from Arctic and Antarctic Oceans. *Journal of Foraminiferal Research*
Size: 2 datasets
doi:10.1594/PANGAEA.715971 - Score: 58% - [Similar datasets](#)

Questions

- Data embargos?
- Which studies?
- What's already been done?
- Analysis plan?
- Data QC?