Table 1. Range of function prediction protocols in a sampling of metagenomics publications to date

							Gene calling				Functional annotation				
Pub. Yr.	Environment (Location)	No. of ORFs	(Mbp)	No. of novel ORFs	(%)	No. of COGs*	Procedure	Sequence comparison algorithm	Parameters and cutoffs	Sequence DB searched	Procedure	Parameters	Functional DB searched	Org.	Ref.
2004	Acid Mine (California)	46,862	(76)	34,301	73.20	1,824	FGENESB pipeline (Softberry Inc)	DBScan	1E-10, >100 bp	nr	blastp against COG	manual refinement using ARTEMIS tool	COG, nr	UCB	Tyson et al. 2004. Nature 428:37–43.
2004	Surface Sea Water (Sargasso Sea, samples 1-4)	1,001,987	(779)	649,608	64.83	3,714	Evidence-based, using translation start & stop sites	tblastn	1E-03	Genes: Bacterial portion of nraa	BLAST against TIGR	>40bp, 3 or more hits to a role category	TIGR Role Category	Venter Inst.	Venter et al. 2004. Science 304:66–74.
								tblastx	1E-04	rRNA: nraa	blastn	1E-40	Sargasso (self-blast)		
2005	Deep-Sea Whalefall (Pacific, Antarctic)	122,147	(75)	63,021	51.59	3,332	FGENESB pipeline (Softberry Inc)	DBScan	Default parameters of software	nraa	blastp against COG, KEGG	low-complexity filtering disabled, 60 bits cutoff equiv. to 1E-8	extCOG** v 6, KEGG	JGI	Tringe et al. 2005. Science 308:554–557.
2005	Farm Soil (Minnesota)	183,536	(100)	114,301	62.28	3,394	FGENESB pipeline (Softberry Inc)	DBScan	Default parameters of software	nraa	blastp against COG, KEGG	low-complexity filtering disabled, 60 bits cutoff equiv. to 1E-8	extCOG** v 6, KEGG	JGI	Tringe et al. 2005. Science 308:554–557.
Total		1,354,532	1,030	861,231	63.58										

* ACE/Chao1 estimates of community richness.

** extCOG, extended COG database in STRING.