Supplementary document for Applied and Environmental manuscript

Metagenomics: Read length matters

K. Eric Wommack[1], Jaysheel Bhavsar[1] and Jacques Ravel[2*]

[1] Delaware Biotechnology Institute, University of Delaware, 15 Innovation Way, Newark, DE 19711, USA.

[2] Institute for Genome Sciences, Department of Microbiology & Immunology, University of Maryland School of Medicine, 20 Penn Street, Baltimore, MD 21201, USA.

[*] To whom correspondence should be addressed. E-mail: jravel@som.umaryland.edu

Cluster of orthologus groups (COG) analysis of short read data simulation experiments (Figs. S1-S3).

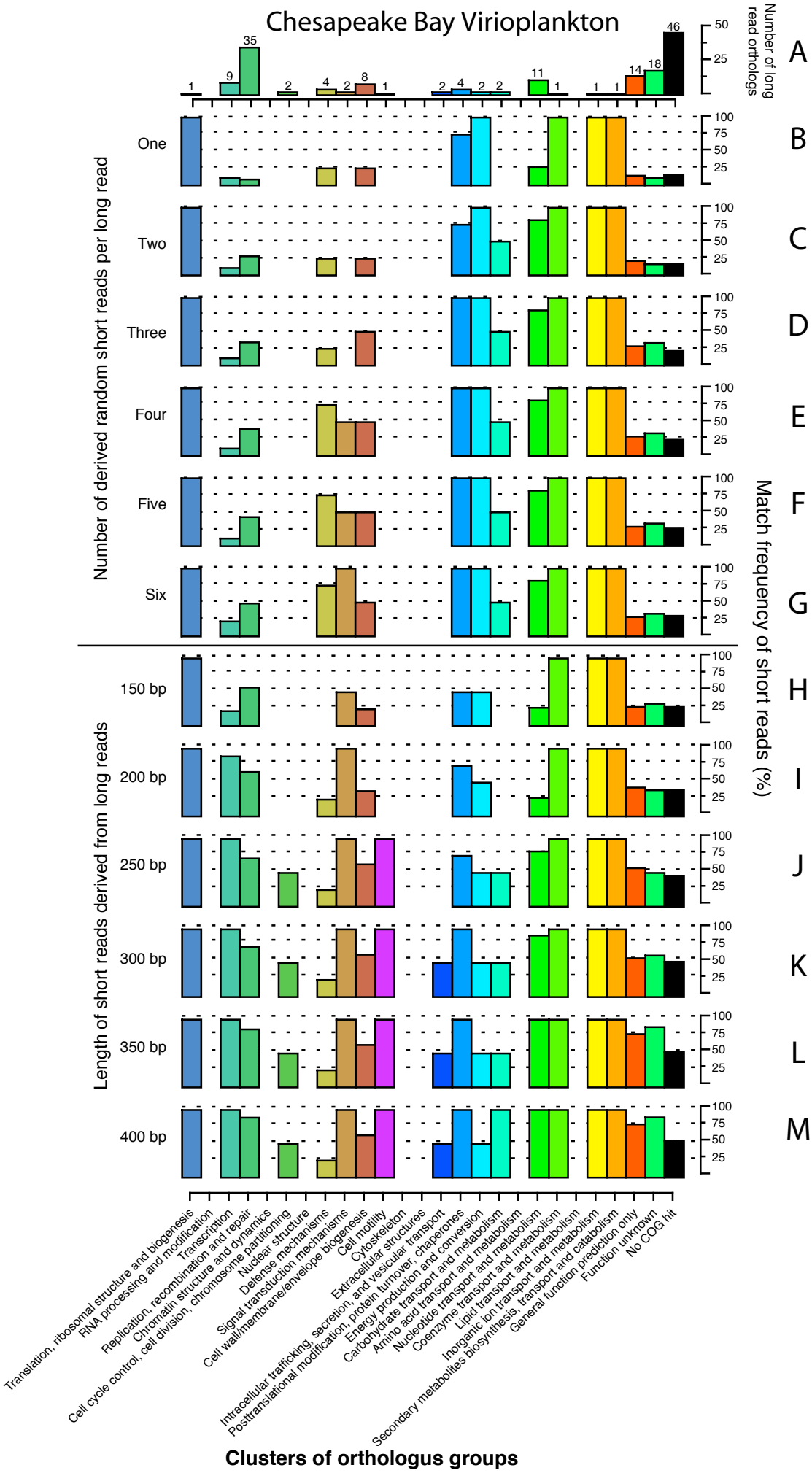
Chesapeake Bay Virioplankton
Number of long read orthologs
Number of derived random short reads per long read
One
Two
Three
Four
Five
Six
Length of short reads derived from long reads
150 bp
200 bp
250 bp
300 bp
350 bp
400 bp
Match frequency of short reads (%)
A
B
C
D
E
F
G
H
I
J
K
L
M
Translation, ribosomal structure and biogenesis
RNA processing and modification
Transcription
Replication, recombination and repair
Chromatin structure and dynamics
Cell cycle control, cell division, chromosome partitioning
Nuclear structure
Defense mechanisms
Signal transduction mechanisms
Cell wall/membrane/envelope biogenesis
Cell motility
Cytoskeleton
Extracellular structures
Intracellular trafficking, secretion, and vesicular transport
Posttranslational modification, protein turnover, chaperones
Energy production and conversion
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Nucleotide transport and metabolism
Coenzyme transport and metabolism
Lipid transport and metabolism
Inorganic ion transport and metabolism
Secondary metabolites biosynthesis, transport and catabolism
General function prediction only
Function unknown
No COG hit
Clusters of orthologus groups

Fig. S1. Distribution of Chesapeake Bay virioplankton BLAST-positive sequences according to clusters of orthologus groups of proteins (COGs). A) Distribution of long read sequences within each COG. Numbers above each bar are the number of long read BLAST-positive sequences within the COG. B-G) Increasing number of 100 bp random derived short read sequences per long read. H-M) Increasing length of single random derived short reads. BLAST-positive short derived sequences expressed as a percent of the long read sequences within each COG. The frequency of long read BLAST-positive sequences that did not belong to a COG was 56% (209 sequences).

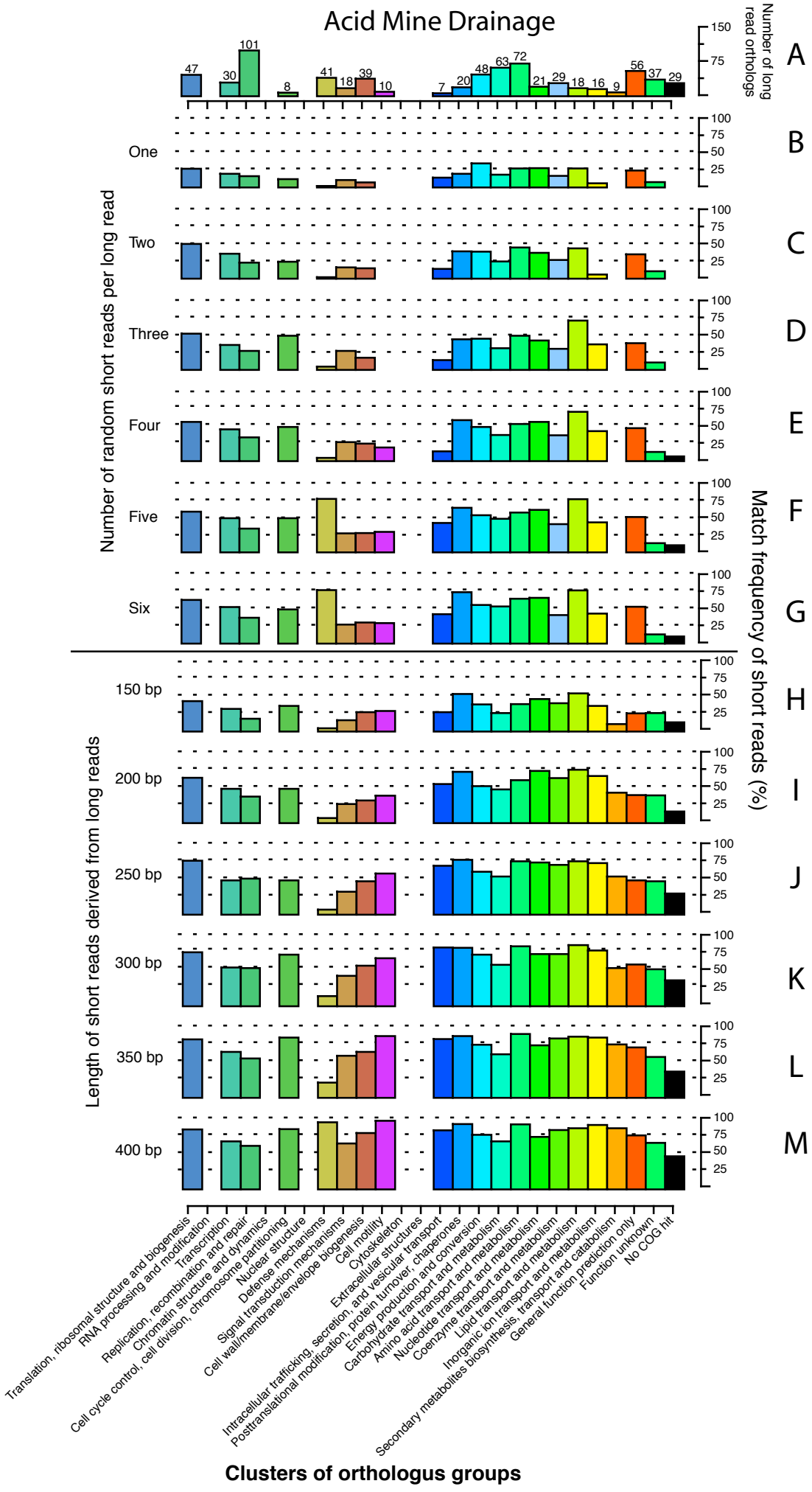
Acid Mine Drainage
Number of long read orthologs
A
B
C
D
E
F
G
H
I
J
K
L
M
One
Two
Three
Four
Five
Six
150 bp
200 bp
250 bp
300 bp
350 bp
400 bp
Number of random short reads per long read
Length of short reads derived from long reads
Match frequency of short reads (%)
Translation, ribosomal structure and biogenesis
RNA processing and modification
Transcription
Replication, recombination and repair
Chromatin structure and dynamics
Cell cycle control, cell division, chromosome partitioning
Nuclear structure
Defense mechanisms
Signal transduction mechanisms
Cell wall/membrane/envelope biogenesis
Cell motility
Cytoskeleton
Extracellular structures
Intracellular trafficking, secretion, and vesicular transport
Posttranslational modification, protein turnover, chaperones
Energy production and conversion
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Nucleotide transport and metabolism
Coenzyme transport and metabolism
Lipid transport and metabolism
Inorganic ion transport and metabolism
Secondary metabolites biosynthesis, transport and catabolism
General function prediction only
Function unknown
No COG hit
Clusters of orthologus groups

Fig. S2. Distribution of Acid mine drainage microbial metagenome BLAST-positive sequences according to clusters of orthologus proteins (COGs). A) Distribution of long read sequences within each COG. Numbers above each bar are the number of long read BLAST-positive sequences within the COG. B-G) Increasing number of 100 bp random derived short read sequences per long read. H-M) Increasing length of single random derived short reads. BLAST-positive short derived sequences expressed as a percent of the long read sequences within each COG. The frequency of long read BLAST-positive sequences that did not belong to a COG was 13% (107 sequences).

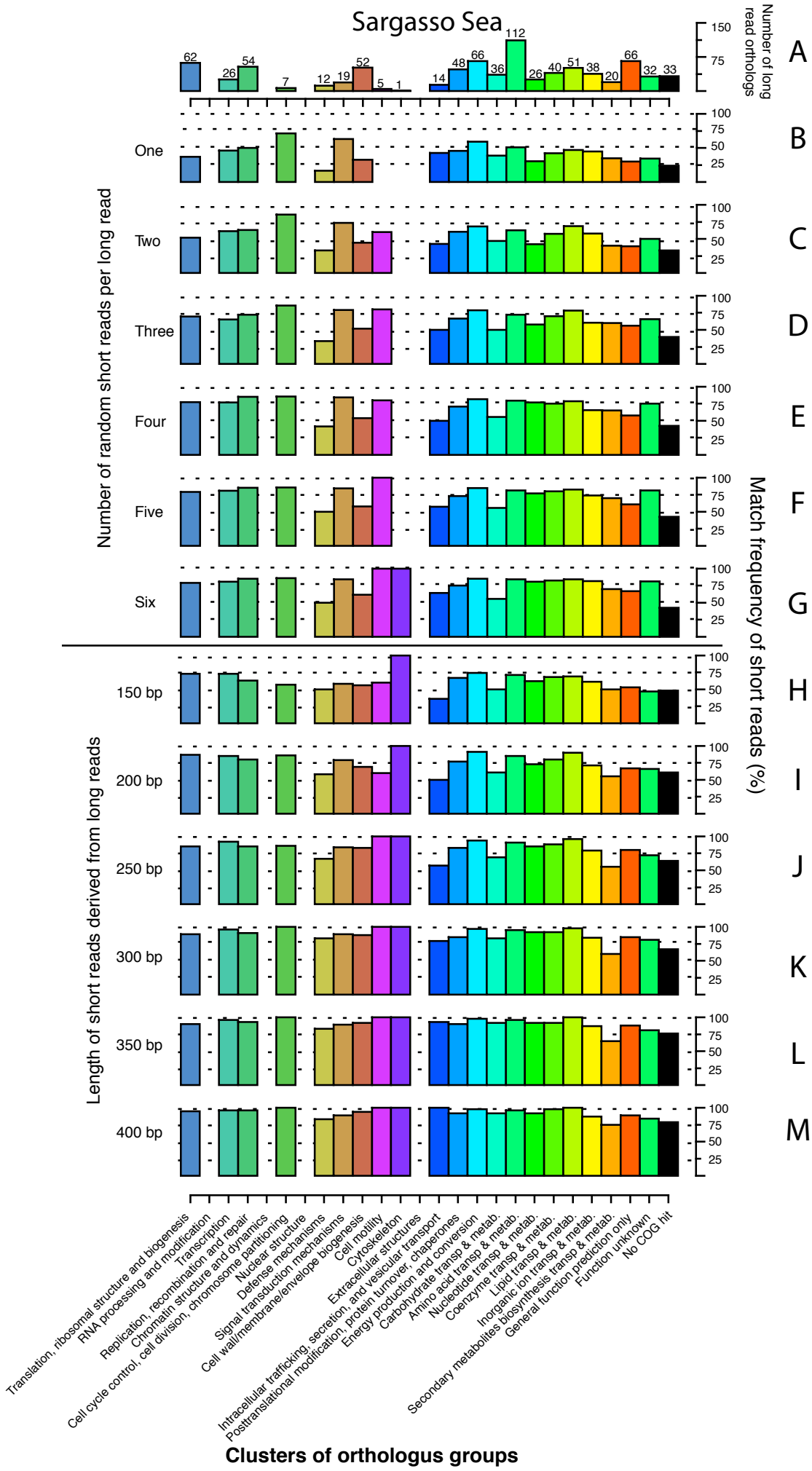
Sargasso Sea
Number of long read orthologs
A
B
C
D
E
F
G
H
I
J
K
L
M
One
Two
Three
Four
Five
Six
150 bp
200 bp
250 bp
300 bp
350 bp
400 bp
Number of random short reads per long read
Length of short reads derived from long reads
Match frequency of short reads (%)
Translation, ribosomal structure and biogenesis
RNA processing and modification
Transcription
Replication, recombination and repair
Chromatin structure and dynamics
Cell cycle control, cell division, chromosome partitioning
Nuclear structure
Defense mechanisms
Signal transduction mechanisms
Cell wall/membrane/envelope biogenesis
Cell motility
Cytoskeleton
Extracellular structures
Intracellular trafficking, secretion, and vesicular transport
Posttranslational modification, protein turnover, chaperones
Energy production and conversion
Carbohydrate transp & metab.
Amino acid transp & metab.
Nucleotide transp & metab.
Coenzyme transp & metab.
Lipid transp & metab.
Inorganic ion transp & metab.
Secondary metabolites biosynthesis transp & metab.
General function prediction only
Function unknown
No COG hit
Clusters of orthologus groups

Fig. S3. Distribution of Sargasso Sea microbial metagenome BLAST-positive sequences according to clusters of orthologus proteins (COGs). A) Distribution of long read sequences within each COG. Numbers above each bar are the number of long read BLAST-positive sequences within the COG. B-G) Increasing number of 100 bp random derived short read sequences per long read. H-M) Increasing length of single random derived short reads. BLAST-positive short derived sequences expressed as a percent of the long read sequences within each COG. The frequency of long read BLAST-positive sequences that did not belong to a COG was 5.4% (47 sequences).