bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Research Article

Strong spurious transcription likely a cause of DNA insert bias in typical metagenomic clone libraries

Kathy N. Lam[1] and Trevor C. Charles[1]

[1]Department of Biology, University of Waterloo, Waterloo, Ontario, Canada

Correspondence to: Trevor C. Charles, tcharles@uwaterloo.ca

Author emails:

Kathy N. Lam, k3lam@uwaterloo.ca

Trevor C. Charles, tcharles@uwaterloo.ca

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 **ABSTRACT**

2 **Background:** Clone libraries provide researchers with a powerful resource with which to study
3 nucleic acid from diverse sources. Metagenomic clone libraries in particular have aided in studies of
4 microbial biodiversity and function, as well as allowed the mining of novel enzymes for specific
5 functions of interest. These libraries are often constructed by cloning large-inserts (~30 kb) into a
6 cosmid or fosmid vector. Recently, there have been reports of GC bias in fosmid metagenomic
7 clone libraries, and it was speculated that the bias may be a result of fragmentation and loss of AT-
8 rich sequences during the cloning process. However, evidence in the literature suggests that
9 transcriptional activity or gene product toxicity may play a role in library bias.

10 **Results:** To explore the possible mechanisms responsible for sequence bias in clone libraries, and in
11 particular whether fragmentation is involved, we constructed a cosmid clone library from a human
12 microbiome sample, and sequenced DNA from three different steps of the library construction
13 process: crude extract DNA, size-selected DNA, and cosmid library DNA. We confirmed a GC bias
14 in the final constructed cosmid library, and we provide strong evidence that the sequence bias is not
15 due to fragmentation and loss of AT-rich sequences but is likely occurring after the DNA is
16 introduced into *E. coli*. To investigate the influence of strong constitutive transcription, we searched
17 the sequence data for consensus promoters and found that *rpoD*/σ^{70} promoter sequences were
18 underrepresented in the cosmid library. Furthermore, when we examined the reference genomes of
19 taxa that were differentially abundant in the cosmid library relative to the original sample, we found
20 that the bias appears to be more closely correlated with the number of *rpoD*/σ^{70} consensus sequences
21 in the genome than with simple GC content.

22 **Conclusions:** The GC bias of metagenomic clone libraries does not appear to be due to DNA
23 fragmentation. Rather, analysis of promoter consensus sequences provides support for the
24 hypothesis that strong constitutive transcription from sequences recognized as *rpoD*/σ^{70} consensus-

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

25 like in *E. coli* may lead to plasmid instability or loss of insert DNA. Our results suggest that despite
26 widespread use of *E. coli* to propagate foreign DNA, the effects of *in vivo* transcriptional activity
27 may be under-appreciated. Further work is required to tease apart the effects of transcription from
28 those of gene product toxicity.

29 **KEYWORDS**

30 metagenomics; cosmid; fosmid; cloning bias; GC bias; metagenomic libraries; *E. coli* host; spurious
31 transcription; sigma 70

32 BACKGROUND

33 Clone libraries can be generated using a range of source material, from the DNA of a single
34 organism to DNA from environmental sources representing often complex microbial communities.
35 Libraries generated from microbial communities are called metagenomic libraries and they have
36 been central to a powerful methodology contributing to understanding the diversity of microbial
37 communities, expanding the knowledge of gene function, and mining for novel sequences encoding
38 functions of interest. These activities all fall under the umbrella of functional metagenomics, and
39 require cloning the DNA, typically using low-copy vectors such as cosmids or fosmids. Cloned
40 DNA is typically propagated in *E. coli* and if the vector host range allows, the DNA can
41 subsequently be transferred to other surrogate hosts that may be more suitable for heterologous
42 expression.

43 The general assumption in cloning-based metagenomic approaches is that foreign DNA can be
44 stably maintained in *E. coli* and that the cloned DNA is a fair representation of the original sample.
45 However, it has been previously observed that fosmid libraries exhibit a GC bias [1, 2]. In general,
46 such cloning biases may affect conclusions derived from analysis of the clone libraries. The
47 observed GC bias of fosmid libraries was suggested to be due to fragmentation and subsequent loss
48 of AT-rich sequences during the cloning process, purportedly because AT-rich sequences have fewer

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

49 hydrogen bonds which makes them more vulnerable to non-perpendicular shear forces [1]. Other
50 possible reasons for the bias in libraries include transcriptional activity of the cloned DNA [3] as
51 well as toxicity from expressed genes [4, 5]. Though the exact mechanism(s) by which GC bias
52 occurs has not yet been fully elucidated, the fragmentation explanation has been echoed by others
53 [6, 7] despite being purely speculative and lacking experimental support. Indeed, from our
54 experiences constructing and screening metagenomic cosmid libraries, we believe that events
55 occurring *in vivo* may be contributing substantially to the sequence bias of libraries.

56 We investigated the nature of this GC bias, to characterize whether, and by what mechanism,
57 biases may be introduced into our own cosmid libraries. In particular, we wished to determine if
58 fragmentation was a major cause of bias, or if there is evidence that the bias was indeed occurring
59 *in vivo*. To answer this question, we constructed a cosmid library using DNA isolated from pooled
60 human fecal samples, saving a portion of the DNA from three steps of the library construction
61 process: (1) the crude extract DNA, (2) the size-selected DNA, and (3) the cloned DNA from the
62 constructed cosmid library (**Figure 1**). The DNA samples were sequenced and the resulting datasets
63 were analyzed to investigate if, where, and how any bias may have been introduced. Consistent with
64 the aforementioned studies, we observed GC bias in our constructed cosmid library. However, our
65 results indicate that fragmentation of DNA does not cause any significant bias; rather, our results
66 are consistent with the hypothesis that the bias occurs after DNA is introduced into the *E. coli* host.

67 RESULTS AND DISCUSSION

68 DNA sampling and sequencing results

69 We collected DNA at the three main steps of cosmid library construction: the crude extract DNA,
70 the size-selected DNA, and the final cosmid library DNA (**Figure 1**). Before sequencing, we first
71 checked the quality of each sample by gel electrophoresis (**Figure 2**). As expected, the crude extract
72 was the only sample that contained a heavy smear of fragmented DNA; the selection for high-

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

73 molecular-weight DNA greatly reduced fragmented DNA, as evidenced by its absence from the
74 size-selected sample. The cosmid library sample exhibited the characteristic multiple banding
75 pattern representing the various possible conformations of uncut circular DNA.
76 After confirming DNA quality, the samples were paired-end sequenced on an Illumina HiSeq
77 2000 platform, generating ~1.2 Gb of DNA sequence per sample. We expected that the cosmid
78 library would be contaminated with *E. coli* genomic DNA and cosmid vector DNA as a result of (1)
79 isolating cosmid DNA from *E. coli* cells, and (2) the fact that each and every cosmid clone
80 sequenced included its vector backbone. Thus, for fair treatment, we subtracted *E. coli* and pJC8
81 sequences from all samples (see **Methods**). For *E. coli* and pJC8, respectively: 6 701 and 164 reads
82 were removed from crude extract data (~0.05% of all reads); 9 273 and 2 410 from size-selected
83 data (~0.09%); and 851 410 and 2 130 004 from the cosmid library DNA (~23%). As expected, the
84 dataset originating from the cosmid library sample had the highest number of reads subtracted.
85 Though the crude extract and size-selected samples contained a small amount, these likely represent
86 true environmental sequences; however, their subtraction was necessary for equal treatment of all
87 samples, and the small fraction removed should not affect overall conclusions from the data.
88 After host and vector sequence subtraction, we used Nonpareil [8] to estimate the overall
89 sequencing coverage of the samples, which was ~85% for the crude extract and size-selected
90 samples and ~95% for the cosmid library sample (**Additonal file 3: Figure S3**). This relatively high
91 sequencing coverage was sufficient for our comparative sequence analyses; for all subsequent
92 results discussed in this paper, the forward and reverse sequencing reads for the three samples were
93 analyzed separately.

94 **GC bias is not caused by fragmentation of AT-rich DNA during cloning**

95 In our experience, extracting high-molecular-weight genomic DNA from low-GC organisms is no
96 more difficult than from *E. coli*. We have previously constructed genomic libraries in cosmid
97 vectors using DNA from *Bacteroides thetaiotaomicron* and *Bacteroides fragilis* (both ~43% GC)

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

98 with no difficulties obtaining high-quality DNA [9]. Therefore, it seemed to us that the suggestion
99 by Temperton *et al.* [1] that the GC bias in cosmid/fosmid libraries might be due to fragmentation of
100 AT-rich sequences was unlikely to be true. Our experimental design was such that we could address
101 this specific point: because we sequenced both crude extract and size-selected samples, we could
102 determine whether the removed fragmented DNA from the crude extract (visible in **Figure 2**) led to
103 a bias in the size-selected DNA sample. We examined percent GC in each of the three datasets and
104 found that the GC bias was only present in the final cosmid library and not the size-selected sample
105 (**Table 1**), effectively ruling out fragmentation as the mechanism for cosmid library bias.
106 After confirming that the bias occurs post-size selection, we next asked if certain taxa were
107 differentially represented across the samples to see if this would point to a possible reason for
108 library sequence bias. We used Taxy [10] as well as Taxy-Pro [11] as part of the CoMet web server
109 [12] to do a fast preliminary comparison of taxa abundance across the three different samples. Taxy
110 calculates *k*-mer frequencies for the dataset and then uses mixture modelling of *k*-mer frequencies
111 of sequenced genomes to obtain a profile similar to that of the sample, whereas Taxy-Pro has a
112 similar modelling approach but uses protein domains rather than *k*-mer frequencies. Both tools
113 generated very similar profiles for the crude extract and the size-selected DNA, but a very different
114 profile for the cosmid library DNA (data not shown), supporting the percent GC results. With
115 positive results from this preliminary work, we then performed more thorough taxonomic analyses
116 using two different approaches; in the first, all sequencing reads were used, and in the second, only
117 the 16S rRNA gene-containing reads were used (see **Methods**).
118 In the first approach, we used MetaPhlAn, a profiling tool that maps reads against clade-
119 specific marker sequences [13], to estimate sample composition down to the species level
120 (**Additonal file 4: Table S4**). We examined the abundance of the top four most common phyla in
121 human gut metagenomes to see whether there were large overall changes in taxa abundance across
122 the samples (**Figure 3**). The crude extract and size-selected samples showed high Firmicutes and

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

123 Bacteroidetes content with lower levels of Actinobacteria and Proteobacteria, compositions that are
124 typical of gut-derived samples [14–16]. Notably, our results indicated that the cosmid library
125 sample underwent a substantial decrease in the Firmicutes, accompanied by a comparably
126 substantial increase in the Actinobacteria. These results were consistent with the percent GC
127 analysis, as members of the Firmicutes phylum are generally known to be low-GC, and those of the
128 Actinobacteria, high-GC. We also examined the MetaPhlAn results at the species level to see which
129 genomes may be under- or over-represented in the cosmid library, choosing to examine the top 50
130 most differentially abundant species (**Figure 4**). Several members of the *Bifidobacterium* genus
131 were substantially over-represented in the cosmid library while many members of the Firmicutes
132 were completely or very nearly lost; for example, *Eubacterium rectale*, *Ruminococcus bromii*, and
133 *Faecalibacterium prausnitzii* were all highly abundant in the original sample.

134 In our second approach, we identified reads in the datasets that were from the 16S rRNA gene,
135 and used the RDP classifier to classify these to the genus level (**Additonal file 5: Figure S5**). We
136 observed that analyses using only 16S rRNA gene-containing reads showed high agreement with
137 analyses carried out using all reads (i.e., **Figure 4**), indicating that 16S rRNA gene content tracks
138 well with genomic content in large-insert cosmid libraries. Both of our approaches provided similar
139 results, and both were in agreement with percent GC, Taxy, and Taxy-Pro results, all of which
140 provide compelling evidence that cosmid library biases are not due to fragmentation of AT-rich
141 sequences during the cloning process.

142 GC content may be merely a proxy for *E. coli* constitutive promoter content

143 From these results, our own previous experiences, and what was previously known in the literature,
144 we had reason to suspect that the cause of the bias occurred *in vivo*. We are not the first to suggest
145 that sequences from AT-rich genomes may resemble the constitutive *E. coli* promoter [17, 18],
146 particularly the −10 Pribnow box. To investigate whether transcription of the insert may be having a
147 negative effect on its maintenance by the host cell, we analyzed the sequence data from the three

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

148 samples for *E. coli* consensus promoter sequences; in particular, we were interested in examining
149 the data for differential abundance of the *rpoD*/σ^{70} consensus sequence, as σ^{70} is the "house-
150 keeping" sigma factor whose promoters are constitutive.

151 We used the known promoter consensus sequence for *rpoD*/σ^{70} [19], and, as negative controls,
152 we used the consensus sequence for: *rpoE*/σ^{24} [20]; *rpoH*/σ^{32} [21]; *rpoN*/σ^{54} , which has a GC-rich
153 consensus [22]; as well as the primary sigma factor of *Bacteroides* species, σ^{ABfr} [23] because the
154 *Bacteroides* genus had comparable abundance across the three samples (**Additonal file 5: Figure**
155 **S5**) and because *Bacteroides* constitutive promoters are not recognized by *E. coli* [24]. We
156 examined each of the three samples for relative abundance of these five consensus sequences (see
157 **Methods** for details). Our results showed that while the crude extract and size-selected samples had
158 similar promoter content profiles, the cosmid library exhibited a deviation (**Figure 5**). Supporting
159 our hypothesis, only the *rpoD* consensus content was considerably different in abundance, by about
160 an order of magnitude when compared to either the crude extract or size-selected sample. The loss
161 of these specific sequences from the cosmid library suggests that the widely used cloning host *E.*
162 *coli* may be problematic for cosmid-cloned foreign fragments of DNA that are incidentally
163 transcriptionally active in a constitutive manner, and indeed, these findings are supported by
164 previous reports in the literature, which we discuss in more detail in the following section.

165 Given that *rpoD* promoter sequences were under-represented in the cosmid library and that
166 certain species appear to be over- or under-represented, we next asked whether a species' abundance
167 in the cosmid library be predicted from the *rpoD* consensus content of its genome? And in
168 particular, is *rpoD* consensus content more predictive of library abundance than GC content? To
169 answer our questions, we turned to the results of our MetaPhlAn analysis, which gave us a list of
170 the top 50 most differentially abundant species (**Figure 4**). To analyze the genomes of the species
171 for possible sequence determinants of library abundance, we used the NCBI Genome database to
172 find sequenced representatives of each species where possible, and retrieved 46 genomes (complete,

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

173 draft, or whole genome shotgun sequences; see **Methods** for details); for each genome, we
174 calculated the percent GC as well as the number of *rpoD* consensus promoter sequences present
175 (**Additonal file 6: Table S6**). Next, to quantify bias in the cosmid library relative to the original
176 sample (the crude extract), we calculated the fold change in abundance of the 46 species (using the
177 average abundance of the forward and reverse datasets). We then plotted the fold change in
178 abundance first against genome percent GC (**Figure 6A**) and second, against *rpoD* consensus
179 content, normalizing to genome size (**Figure 6B**). Our results show that while library bias only
180 generally correlates with GC content, library bias correlates surprisingly well with the *rpoD*
181 consensus content of the genome.

182 These results suggest that GC content may be only a rough proxy for *rpoD* consensus content
183 (as *rpoD* consensus sequences are AT-rich), but GC content itself may not be an accurate predictor
184 of library presence/abundance; indeed, in some cases, a genome may have a moderate or relatively
185 high percent GC but also possess an unusually high *rpoD* consensus content, leading to an under-
186 representation in the cosmid library that could not have been predicted from GC content alone
187 (**Figure 6**). In our view, these results are also consistent with the previous observation that library
188 bias was more obvious among organisms with low GC content [2] because AT-rich genomes would
189 have an increased number of promoter-like sequences simply by chance [25].

190 **Examining the published literature: evidence for transcriptional activity of cloned AT-rich**
191 **DNA interfering with stability of circular vectors**

192 In this report, we have presented analysis concerning metagenomic DNA. However, if *rpoD*
193 consensus-like sequences are interfering with the maintenance of foreign DNA in *E. coli*, then the
194 scope of the problem extends beyond metagenomics applications. Curious about the extent of the
195 problem, we performed literature searches to find reports of experienced difficulties cloning AT-rich
196 DNA and/or investigations of possible mechanisms for those difficulties. Our search was fruitful,
197 leading us to literature that spans the past three decades.

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

198 It was reported that there are difficulties associated with cosmid-cloning of very AT-rich
199 genomic DNA [26, 27], and even when genomic libraries can be constructed, cosmid clones may be
200 unstable [28–31], which simply means that foreign DNA fragments are not able to be maintained in
201 the *E. coli* library host. Thus, if selection is applied for a marker present on the vector, then *in vivo*
202 events may lead to insert deletion, which has been observed by us as well as others, despite using a
203 host that is a *recA* mutant [31]. This is particularly evident when the library is constructed using a
204 high-copy number vector (e.g., one containing a ColE1-type origin of replication), which has been
205 experienced by us and others [32] and is in agreement with the observation that F-based, single-
206 copy fosmids perform better than multi-copy cosmids at stably maintaining insert DNA [33]. Loss
207 of cloned sequence is even more widespread for inserts that have repetitive DNA sequences [34], as
208 such sequences may be conducive to recombination. One way to combat insert loss is by
209 minimizing outgrowth of the library-containing cells as much as possible [31], though this is not
210 always feasible for shared cosmid libraries such as our Canadian MetaMicroBiome Library
211 collection [41].
212 But what is the mechanism for plasmid instability? It was previously shown that transcriptional
213 activity from a cloned strong promoter could affect plasmid stability by (1) interfering with the
214 origin of replication via transcription read-through into the vector, as well as (2) changing the
215 abundance of protein products involved in plasmid copy number. Furthermore, plasmid instability
216 was alleviated by placing transcriptional terminator sequences that flank the multiple cloning site
217 [36]. It was also observed that strong phage promoters could only be cloned into plasmids that
218 possess a downstream termination signal [37, 38]. Similarly, AT-rich pneumococcal DNA was found
219 to contain a high incidence of *E. coli* strong promoter sequences, and that cloning of the DNA was
220 improved by using a vector with efficient transcriptional terminators [3, 32, 39], although analysis
221 of a set of pneumococcal promoter-containing sequences indicated that transcription strong enough
222 to interfere with plasmid stability may be relatively rare and that other factors could be contributing

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

223 to cloning difficulty [40].

224 Another consideration is that efficient transcription of poly-dT (as well as poly-dG) DNA tracts
225 may cause the DNA to form a stable complex with its own accumulated transcription products,
226 leading to transcriptional stalling that may interfere with the replication fork [41–43]. One
227 particularly interesting observation that has surprisingly not attracted more interest, is that linear
228 cloning vectors with transcriptional terminators provide even more stability than circular vectors
229 with transcriptional terminators [26, 44]. The advantage of these vectors is due to their linear
230 conformation, but intriguingly, the mechanism remains unclear, although DNA supercoiling of
231 plasmids is thought to play a role (Ronald Godiska, personal communication). These findings along
232 with the aforementioned facts suggest that multiple, distinct mechanisms may be at play to cause
233 cloning bias in *E. coli*, but that there is evidence that transcriptional activity of cloned DNA may be
234 a cause of sequence bias observed in metagenomic libraries. It is often assumed that toxicity of gene
235 products may influence the stable maintenance or “clonability” of DNA in *E. coli* [4, 5, 45] but it is
236 currently unclear whether gene product toxicity is a major factor in the bias of typical clone libraries
237 constructed using circular vectors. It is interesting to consider that cloning bias could be due
238 primarily to purely transcriptional activity rather than the often-blamed protein toxicity.

239 CONCLUSIONS

240 Our own experiences in the lab, the results presented in this report, and what was already known
241 from the literature altogether support the hypothesis that GC bias in typical clone libraries (that is,
242 using circular vectors) is related to promoter activity of the insert in *E. coli*, although DNA topology
243 as well as toxic protein effects may also influence insert and plasmid maintenance. In our analyses,
244 we have focused only on would-be strong constitutive promoters in *E. coli* (sigma 70 consensus
245 sequences) because there is evidence that high level transcription may have negative effects. It is
246 important to acknowledge, however, that functional metagenomic approaches rely on *E. coli* (or

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

247 other hosts) being able to transcribe and translate foreign DNA, in order to identify fragments
248 encoding functions of interest. This ability of *E. coli* to initiate low-level transcription from diverse
249 sources [46] and to be able to produce foreign proteins, has been immensely advantageous for
250 functional metagenomics, and likely has contributed to the general assumption that *E. coli* is
251 tolerant of foreign DNA, whether it expresses it or not. Our work, however, suggests that more
252 careful consideration of cloning strategies may be required.

253 Currently, there are three outstanding questions: (1) to what extent does transcription contribute
254 to metagenomic library bias, (2) what factors affect whether transcription will be problematic, and
255 (3) how can transcriptional effects be minimized so that DNA can be faithfully maintained in *E.*
256 *coli*. An important consideration may be the likelihood of an *rpoD* consensus sequence being cloned
257 on any given fragment from a genome or metagenome. As an example, let us consider
258 *Ruminococcus bromii*, which was one of the most highly abundant species in the original sample
259 but became nearly absent in the cosmid library according to our analyses (~7% vs. ~0.01%,
260 respectively; see **Additonal file 4: Table S4**). *R. bromii* has a genome size of 2.25 Mb;
261 theoretically, its genome can be represented in ~80 fragments if we consider that the average
262 fragment in the particular cosmid library discussed here is ~28 kb (data not shown). Given that
263 there were 77 *rpoD* consensus sequences identified in its genome (**Additonal file 6: Table S6**),
264 potentially many fragments could include a sequence that behaves as a strong, constitutive promoter
265 in *E. coli*.

266 In general, it may be helpful to use cloning vectors that include transcriptional terminators
267 flanking the cloning site. We are currently investigating the extent to which transcriptional
268 terminators alleviate the cosmid library sequence bias, which should help tease apart the issue of
269 transcription from that of gene product toxicity. While it is generally recognized that different host
270 backgrounds are needed for functional screening [45, 47–52], it is not as widely acknowledged that
271 the *E. coli* library host itself may be quite limiting. It is interesting that despite decades of using *E.*

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

272 *coli* as "the workhorse of molecular biology", there is still much left to discover about how it
273 tolerates exogenous DNA, which should serve as a reminder to us of how necessary it is to
274 continually re-evaluate even our most basic methodological assumptions, particularly when they
275 concern the inner workings of the cell.

276 **METHODS**

277 **Sampling of DNA during steps of metagenomic cosmid library construction**

278 Methods for the construction of cosmid libraries, including the specific human gut metagenomic
279 library discussed here (NCBI BioSample ID SAMN02324081), have been previously described in
280 detail [9]. Briefly, DNA was extracted from pooled human fecal samples using freeze-grinding with
281 liquid nitrogen followed by gentle lysis. Crude extracted DNA was then size-selected by pulsed
282 field gel electrophoresis using a CHEF MAPPER Pulsed Field Gel Electrophoresis System (Bio-
283 Rad), followed by electroelution, retaining fragments between approximately 40 to 70 kb. The size-
284 selected DNA was end-repaired, purified, and ligated into the Eco72I site of linearized
285 dephosphorylated pJC8 vector DNA (Genbank accession KC149513). The ligation product was
286 packaged into lambda phage heads using Gigapack III XL Packaging Extract (Stratagene), followed
287 by transduction of *E. coli* HB101. Transductants were recovered on LB agar supplemented with
288 tetracycline (20 µg/ml), and incubated overnight at 37ºC. Resulting colonies were enumerated to
289 estimate library size (~42,000 clones), and colonies were resuspended, pooled, and frozen at -80ºC
290 to form the cosmid library stock.

291 During construction of the cosmid library, DNA was sampled from three steps: (1) the crude
292 extract DNA, (2) the size-selected DNA, and (3) the final cosmid library DNA, prepared from the
293 frozen stock using a GeneJET Plasmid Miniprep Kit (Thermo Scientific).

294 **Purification, quantification, and Illumina sequencing of DNA**

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

295 Two of the three DNA samples, the cosmid library DNA and the size-selected DNA, were
296 sufficiently pure for Illumina sequencing, as gauged by 260/280 nm and 260/230 nm ratios
297 (Nanodrop ND-1000 Spectrophotometer); however, the crude extract DNA required further
298 purification. Crude extract DNA concentration was estimated by gel electrophoresis, using
299 bacteriophage lambda DNA as a standard; ~150 μg in 1 ml was purified and concentrated on the
300 synchronous coefficient of drag alteration (SCODA) instrument (Boreal Genomics), using an
301 established protocol [53].

302 All samples were re-quantified by gel electrophoresis, using bacteriophage lambda DNA as a
303 standard, and >2 μg of each sample was sent to the Beijing Genomics Institute (BGI, Hong Kong)
304 for 90-base paired-end sequencing on the Illumina HiSeq 2000 platform, using their in-house
305 protocols and reagents for 350 bp fragment library construction. ~6.7 million reads were obtained in
306 both the forward and the reverse direction, generating ~1.2 Gb of sequence data per sample. All
307 sequence data have been made publicly available (see **Data** section).

308 Subtraction of *E. coli* genome and cosmid vector contamination

309 The cosmid library sequence data were expected to have substantial contamination with *E. coli*
310 genomic DNA and pJC8 vector sequences. Sequence data were cleaned of contaminating *E. coli*
311 genomic DNA and vector DNA, using BLAT [54] with a conservative criterion of 100% identity. To
312 remove *E. coli* contamination, we used the genome of *E. coli* K12 MG1655 (Genbank accession
313 U00096.3), which to our knowledge is currently the closest sequenced relative of HB101, the
314 library host strain; to remove vector contamination, we used the sequence of pJC8 (Genbank
315 accession KC149513), formatted to simulate Eco72I-cut, cloning-ready vector by removing the 0.8-
316 kb gentamicin resistance gene stuffer present between the two Eco72I sites.

317 Taxonomic analysis

318 To examine taxonomy based on only the 16S rRNA gene sequences present in the data, we

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

319 identified 16S-containing reads using Infernal version 1.1 [55] and classified them using the RDP
320 Classifier version 2.8 [56]. The classifier output was visualized using the MEtaGenome ANalyzer
321 (MEGAN) version 5.6 [57]. To examine taxonomy using all sequence reads (i.e., not only those
322 identified as 16S reads), we used the Metagenome Phylogenetic Analysis tool (MetaPhlAn) version
323 2.0, along with its built-in scripts for visualization [13].

324 **Promoter analysis**

325 To estimate promoter content in the data, we searched for known sigma factor consensus sequences
326 for the *E. coli* sigma factors, *rpoD*/σ^{70} (TTGACAN$_{15\text{-}19}$TATAAT), *rpoE*/σ^{24} (GGAACTTN$_{15\text{-}}$
327 $_{19}$TCAAA), *rpoH*/σ^{32} (TTG[A/T][A/T][A/T]N$_{13\text{-}14}$CCCCAT[A/T]T), *rpoN*/σ^{54} (TGGCAN$_7$TGC), as
328 well as for the *Bacteroides* primary sigma factor, σ^{ABfr} (TTTGN$_{19\text{-}21}$TAN$_2$TTTG). To do this, we used
329 regular expression pattern matching with Python version 2.7.3; consensus promoter sequences,
330 literature references, and regular expressions are provided (**Additonal file 1: Table S1)**.

331 **Analysis of reference genomes**

332 Genome sequences were downloaded from the NCBI Genbank database as complete genomes, draft
333 genomes, or from whole genome shotgun sequencing projects. Organism names and accession
334 numbers, as well as other relevant information, are provided (**Additonal file 2: Table S2)**.

335 **Data**

336 Raw Illumina sequence data are available at the NCBI Sequence Read Archive under Study
337 SRP031898. Accession numbers for SRA Experiments are: [NCBI:SRX683591] for the crude
338 extract, [NCBI:SRX683589] for the size-selected, and [NCBI:SRX683586] for the cosmid library.
339 In addition, raw data and other relevant data for this study may be accessed online through our
340 website [58].

341 **LIST OF ABBREVIATIONS**

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

342 CE, crude extract; SS, size-selected; CL, cosmid library; F, forward reads; R, reverse reads.

343 AUTHORS' CONTRIBUTIONS

344 TCC and KNL designed the experiments. KNL performed the experiments, structured and
345 performed the data analysis, and wrote the paper. TCC provided constructive criticism, revised the
346 manuscript, and provided reagents and materials.

347 COMPETING INTERESTS

348 The authors declare that they have no competing interests.

349 ACKNOWLEDGMENTS

350 We thank the seven anonymous individuals who donated fecal samples for the construction of the
351 CLGM1 cosmid library. We are grateful to the Shared Hierarchical Academic Research Computing
352 Network (SHARCNET: www.sharcnet.ca) and Compute/Calcul Canada for computing resources.
353 Research funding was provided by a Strategic Projects Grant (381646-09) from the Natural
354 Sciences and Engineering Research Council of Canada, by Genome Canada for the project
355 ‘‘Microbial Genomics for Biofuels and Co-Products from Biorefining Processes’’ (MGCB^2), and by
356 a University of Waterloo CIHR Research Incentive Fund. KNL was supported by a CGS-D
357 scholarship from the Canadian Institutes of Health Research.

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

REFERENCES

1. Temperton B, Field D, Oliver A, Tiwari B, Mühling M, Joint I, Gilbert JA: **Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing**. *ISME J* 2009, **3**:792–796.

2. Danhorn T, Young CR, DeLong EF: **Comparison of large-insert, small-insert and pyrosequencing libraries for metagenomic analysis**. *ISME J* 2012:1–11.

3. Chen J-D, Morrison DA: **Construction and properties of a new insertion vector, pJDC9, that is protected by transcriptional terminators and useful for cloning of DNA from Streptococcus pneumoniae**. *Gene* 1988, **64**:155–164.

4. Sorek R, Zhu Y, Creevey CJ, Francino MP, Bork P, Rubin EM: **Genome-wide experimental determination of barriers to horizontal gene transfer**. *Science (80-)* 2007, **318**:1449–1452.

5. Feingersch R, Béjà O: **Bias in assessments of marine SAR11 biodiversity in environmental fosmid and BAC libraries?**. *ISME J* 2009, **3**:1117–1119.

6. Lombard N, Prestat E, van Elsas JD, Simonet P: **Soil-specific limitations for access and analysis of soil microbial communities by metagenomics**. *FEMS Microbiol Ecol* 2011, **78**:31–49.

7. Gori F, Tringe SG, Folino G, van Hijum SAFT, Op den Camp HJM, Jetten MSM, Marchiori E: **Differences in sequencing technologies improve the retrieval of anammox bacterial genome from metagenomes**. *BMC Genomics* 2013, **14**:1–9.

8. Rodriguez-R LM, Konstantinidis KT: **Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets**. *Bioinformatics* 2014, **30**:629–635.

9. Lam KN, Hall MW, Engel K, Vey G, Cheng J, Neufeld JD, Charles TC: **Evaluation of a pooled strategy for high-throughput sequencing of cosmid clones from metagenomic libraries**. *PLoS One* 2014, **9**:e98968.

10. Meinicke P, Asshauer KP, Lingner T: **Mixture models for analysis of the taxonomic composition of metagenomes**. *Bioinformatics* 2011, **27**:1618–1624.

11. Klingenberg H, Aßhauer KP, Lingner T, Meinicke P: **Protein signature-based estimation of metagenomic abundances including all domains of life and viruses**. *Bioinformatics* 2013, **29**:973–980.

12. Lingner T, Asshauer KP, Schreiber F, Meinicke P: **CoMet — a web server for comparative functional profiling of metagenomes**. *Nucleic Acids Res* 2011, **39**:W518–W523.

13. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C: **Metagenomic microbial community profiling using unique clade-specific marker genes**. *Nat Methods* 2012, **9**:811–814.

14. El Kaoutari A, Armougom F, Gordon JI, Raoult D, Henrissat B: **The abundance and variety of carbohydrate-active enzymes in the human gut microbiota**. *Nat Rev Microbiol* 2013, **11**:497–504.

15. Spor A, Koren O, Ley R: **Unravelling the effects of the environment and host genotype on the gut microbiome**. *Nat Rev Microbiol* 2011, **9**:279–290.

16. Marchesi JR: **Human distal gut microbiome**. *Environ Microbiol* 2011, **13**:3088–3102.

17. Djordjevic G, Bojovic B, Banina A, Topisirovic L: **Cloning of promoter-like sequences from Lactobacillus paracasei subsp . paracasei C G l l and their expression in Escherichia coli , Lactococcus lactis , and Lactobacillus reuteri**. *Can J Microbiol* 1994, **40**:1043–1050.

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

18. Morrison DA, Jaurin B: **Streptococcus pneumoniae possesses canonical Escherichia coli (sigma 70) promoters**. *Mol Microbiol* 1990, **4**:1143–1152.

19. Shimada T, Yamazaki Y, Tanaka K, Ishihama A: **The whole set of constitutive promoters recognized by RNA polymerase RpoD holoenzyme of Escherichia coli**. *PLoS One* 2014, **9**:e90447.

20. Rhodius VA, Suh WC, Nonaka G, West J, Gross CA: **Conserved and variable functions of the sigmaE stress response in related genomes.** *PLoS Biol* 2006, **4**:e2.

21. Nonaka G, Blankschien M, Herman C, Gross CA, Rhodius VA: **Regulon and promoter analysis of the E. coli heat-shock factor, sigma32, reveals a multifaceted cellular response to heat stress**. *Genes Dev* 2006, **20**:1776–1789.

22. Zhao K, Liu M, Burgess RR: **Promoter and regulon analysis of nitrogen assimilation factor, sigma54, reveal alternative strategy for E. coli MG1655 flagellar biosynthesis**. *Nucleic Acids Res* 2010, **38**:1273–1283.

23. Bayley DP, Rocha ER, Smith CJ: **Analysis of cepA and other Bacteroides fragilis genes reveals a unique promoter structure**. *FEMS Microbiol Lett* 2000, **193**:149–154.

24. Mastropaolo MD, Thorson ML, Stevens AM: **Comparison of Bacteroides thetaiotaomicron and Escherichia coli 16S rRNA gene expression signals**. *Microbiology* 2009, **155**:2683–2693.

25. Mulligan ME, Mcclure WR: **Analysis of the occurrence of promoter-sites in DNA**. *Nucleic Acids Res* 1986, **14**:109–126.

26. Godiska R, Mead D, Dhodda V, Wu C, Hochstein R, Karsi A, Usdin K, Entezam A, Ravin N: **Linear plasmid vector for cloning of repetitive or unstable sequences in Escherichia coli**. *Nucleic Acids Res* 2010, **38**:e88.

27. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan M-S, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DMA, et al.: **Genome sequence of the human malaria parasite Plasmodium falciparum**. *Nature* 2002, **419**:498–511.

28. Hattori Y, Omori H, Hanyu M, Kaseda N, Mishima E, Tabata S, Saeki K: **Ordered Cosmid Library of the Mesorhizobium loti MAFF303099 Genome for Systematic Gene Disruption and Complementation Analysis**. *Plant Cell Physiol* 2002, **43**:1542–1557.

29. Brayton KA, Villiers EP DE, Fehrsen J, Nxomani C, Collins NE, Allsopp BA: **Cowdria ruminantium DNA is unstable in a SuperCos1 library**. *Onderstepoort J Vet Res* 1999, **117**:111–117.

30. Rhee D-K: **Instability of Pneumococcus Library in pHC79 and pACYC184**. *Arch Pharm Res* 1995, **18**:31–37.

31. Shoemaker NB, Barber RD, Salyers AA: **Cloning and characterization of a Bacteroides conjugal tetracycline-erythromycin resistance element by using a shuttle cosmid vector**. *J Bacteriol* 1989, **171**:1294–1302.

32. Chen J-D, Morrison DA: **Cloning of Streptococcus pneumoniae DNA fragments in Escherichia coli requires vectors protected by strong transcriptional terminators**. *Gene* 1987, **55**:179–187.

33. Kim U-J, Shizuya H, de Jong PJ, Birren B, Simon MI: **Stable propagation of cosmid sized human DNA inserts in an F factor based vector**. *Nucleic Acids Res* 1992, **20**:1083–1085.

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

34. Case ST: **Selective deletion of large segments of Balbiani ring DNA during molecular cloning**. *Gene* 1982, **20**:169–176.

35. Neufeld JD, Engel K, Cheng J, Moreno-Hagelsieb G, Rose DR, Charles TC: **Open resource metagenomics: a model for sharing metagenomic libraries**. *Stand Genomic Sci* 2011, **5**:1–12.

36. Stueber D, Bujard H: **Transcription from efficient promoters can interfere with plasmid replication and diminish expression of plasmid specified genes**. *EMBO J* 1982, **1**:1399–1404.

37. Gentz R, Langner A, Changt ACY, Cohent SN, Bujard H: **Cloning and analysis of strong promoters is made possible by the downstream placement of a RNA termination signal**. *Proc Natl Acad Sci* 1981, **78**:4936–4940.

38. Kwon Y-S, Kim J, Kang C: **Viability of E. coli cells containing phage RNA polymerase and promoter: interference of plasmid replication by transcription**. *Genet Anal Biomol Eng* 1998, **14**:133–139.

39. Stassi DL, Lacks SA: **Effect of strong promoters on the cloning in Escherichia coli of DNA fragments from Streptococcus pneumoniae**. *Gene* 1982, **18**:319–328.

40. Dillard JP, Yother J: **Analysis of Streptococcus pneumoniae sequences cloned into Escherichia coli: effect of promoter strength and transcription terminators**. *J Bacteriol* 1991, **173**:5105–5109.

41. Kiyama R, Oishi M: **Instability of plasmid DNA maintenance caused by transcription of poly (dT)-containing sequences in Escherichia coli**. *Gene* 1994, **150**:57–61.

42. Kiyama R, Oishi M: **In vitro transcription of a poly (dA)· poly (dT)-containing sequence is inhibited by interaction between the template and its transcripts**. *Nucleic Acids Res* 1996, **24**:4577–4583.

43. Krasilnikova MM, Samadashwily GM, Krasilnikov AS, Mirkin SM: **Transcription through a simple DNA repeat blocks replication elongation**. *EMBO J* 1998, **17**:5095–5102.

44. Godiska R, Patterson M, Schoenfeld T, Mead DA: **Beyond pUC: Vectors for Cloning Unstable DNA**. In *Optim DNA Seq Process*. Edited by Kieleczawa J. Jones and Bartlett Publishers; 2005:55–75.

45. Taupp M, Mewis K, Hallam SJ: **The art and design of functional metagenomic screens**. *Curr Opin Biotechnol* 2011, **22**:1–8.

46. Miller WG, Simons RW: **DNA from diverse sources manifests cryptic low-level transcription in Escherichia coli**. *Mol Microbiol* 1990, **4**:881–893.

47. Craig JW, Chang F-Y, Kim JH, Obiajulu SC, Brady SF: **Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria**. *Appl Environ Microbiol* 2010, **76**:1633–1641.

48. Aakvik T, Degnes KF, Dahlsrud R, Schmidt F, Dam R, Yu L, Völker U, Ellingsen TE, Valla S: **A plasmid RK2-based broad-host-range cloning vector useful for transfer of metagenomic libraries to a variety of bacterial species.** *FEMS Microbiol Lett* 2009, **296**:149–158.

49. Li Y, Wexler M, Richardson DJ, Bond PL, Johnston AWB: **Screening a wide host-range, waste-water metagenomic library in tryptophan auxotrophs of Rhizobium leguminosarum and of Escherichia coli reveals different classes of cloned trp genes**. *Environ Microbiol* 2005, **7**:1927–1936.

50. Cheng J, Pinnell L, Engel K, Neufeld JD, Charles TC: **Versatile broad-host-range cosmids for construction of high quality metagenomic libraries**. *J Microbiol Methods* 2014, **99C**:27–34.

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

51. Troeschel SC, Drepper T, Leggewie C, Streit WR, Jaeger K-E: **Novel Tools for the Functional Expression of Metagenomic DNA**. *Metagenomics Methods Protoc* 2010, **668**:117–139. [*Methods in Molecular Biology*]

52. Uchiyama T, Miyazaki K: **Functional metagenomics for enzyme discovery: challenges to efficient screening**. *Curr Opin Biotechnol* 2009, **20**:616–622.

53. Engel K, Pinnell L, Cheng J, Charles TC, Neufeld JD: **Nonlinear electrophoresis for purification of soil DNA for metagenomics**. *J Microbiol Methods* 2012, **88**:35–40.

54. Kent WJ: **BLAT - The BLAST-Like Alignment Tool**. *Genome Res* 2002, **12**:656–664.

55. Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches**. *Bioinformatics* 2013, **29**:2933–2935.

56. Wang Q, Garrity GM, Tiedje JM, Cole JR: **Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy**. *Appl Environ Microbiol* 2007, **73**:5261–5267.

57. Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC: **Integrative analysis of environmental sequences using MEGAN4**. *Genome Res* 2011, **21**:1552–1560.

58. **Canadian MetaMicrobiome Library Publication Data** [http://www.cm2bl.org/~data]

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

FIGURE LEGENDS

Figure 1: Overview of the experimental design for this study. A pooled human fecal sample was used to construct a metagenomic cosmid library, during which DNA from three distinct steps was collected and sequenced in order to investigate possible sequence biases and at what steps the biases were introduced.

Figure 2: Gel electrophoresis of crude extract, size-selected, and cosmid library DNA samples. Diluted and undiluted amounts of each sample were gel electrophoresed for quality control check of DNA prior to Illumina sequencing.

Figure 3: Histogram of abundance of the top four phyla in crude extract, size-selected, and cosmid library samples. Abundance of the Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria phyla in each sample, as determined using MetaPhlAn.

Figure 4: Heatmap of 50 species with differential abundance across crude extract, size-selected, and cosmid library samples. Abundance in each sample of the top 50 species determined to be differentially abundant using MetaPhlAn. Abundance is depicted on a log scale.

Figure 5: Histogram of sigma factor consensus sequence content in crude extract, size-selected, and cosmid library samples. Bars indicate the number of consensus sequences in each sample, for select *E. coli* sigma factors and the *Bacteroides* primary sigma factor, normalized to the amount of sequence data for that sample. Consensus content is depicted on a log scale.

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 6: Bias in cosmid library relative to crude extract, against GC content or *rpoD* consensus content. Species abundance was obtained from MetaPhlAn analysis of the crude extract and cosmid library samples. Bias is calculated as fold change in percent abundance (cosmid library abundance / crude extract abundance) plotted against GC content (A) or *rpoD* consensus content (B). Change in abundance is depicted on a log scale; CE_0 values indicate zero abundance in the crude extract sample and CL_0 values indicate zero abundance in the cosmid library sample.

TABLES

Table 1: Percent GC of crude extract, size-selected, and cosmid library datasets. GC content was calculated after subtraction of *E. coli* and vector DNA from all samples.

Sample/Dataset	No. Reads	No. Mb	% GC
crude extract F	6 654 484	599	47.7
crude extract R	6 654 567	599	47.8
size-selected F	6 645306	598	46.9
size-selected R	6 645 817	598	46.9
cosmid library F	5 134 020	462	53.0
cosmid library R	5 191 538	467	53.1

bioRxiv preprint doi: https://doi.org/10.1101/013763; this version posted January 15, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ADDITIONAL FILES

Additional file 1: Table S1. Consensus sequences for the five sigma factors used, PMID number for the literature reference, and corresponding regular expressions used to search sequence data. (.txt)

Additional file 2: Table S2. NCBI Genbank accession numbers for genome sequences of the 46 species selected for GC content and *rpoD* consensus content analysis. (.txt)

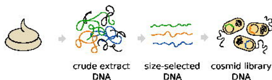
Additional file 3: Figure S3. Estimate of sample sequencing coverage using Nonpareil. (.pdf)

Additional file 4: Table S4. Taxa abundance output from MetaPhlAn for both forward and reverse datasets of each sample. (.txt)

Additional file 5: Figure S5. 16S rRNA analysis results using Infernal for identification of 16S-containing reads, RDP classifier to classify reads, and MEGAN for visualization of results. (.pdf)

Additional file 6: Table S6. Length, GC content, and *rpoD* consensus content of the 46 genomes selected for analysis. (.txt)

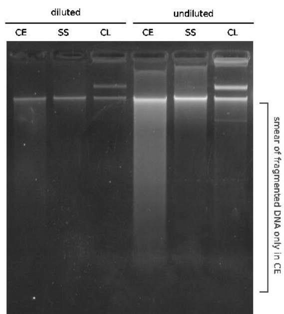
Construct metagenomic cosmid library

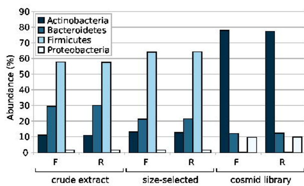


Save sample of DNA at each major step

Illumina sequence sample to ~1.2 Gb

Perform comparative analyses on data

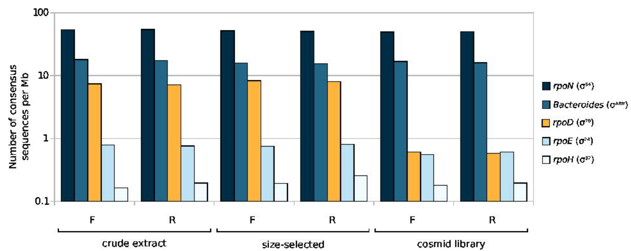

diluted
undiluted
CE
SS
CL
CE
SS
CL
smear of fragmented DNA only in CE

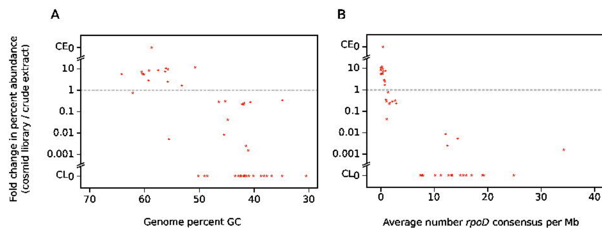

90
80
70
60
50
40
30
20
10
0
Abundance (%)
Actinobacteria
Bacteroidetes
Firmicutes
Proteobacteria
F
R
F
R
F
R
crude extract
size-selected
cosmid library

Megamonas hypermegale
Eubacterium eligens
Roseburia inulinivorans
Prevotella copri
Faecalibacterium unclassified
Ruminococcus torques
Bacteroides vulgatus
Parabacteroides merdae
Bacteroides uniformis
Bacteroides thetaiotaomicron
Clostridium bolteae
Faecalibacterium prausnitzii
Bacteroides unclassified
Eubacterium ventriosum
Bacteroides cellulosilyticus
Faecalibacterium cf
Bacteroides ovatus
Bacteroides xylanisolvens
Dorea formicigenerans
Ruminococcus obeum
Lactobacillus ruminis
Coprococcus comes
Catenibacterium mitsuokai
Dialister invisus
Dorea longicatena
Eubacterium rectale
Clostridium leptum
Streptococcus parasanguinis
Ruminococcus gnavus
Eubacterium hallii
Ruminococcus bromii
Holdemania filiformis
Clostridium nexile
Coprococcus catus
Desulfovibrio piger
Parabacteroides unclassified
Alistipes putredinis
Bifidobacterium breve
Bifidobacterium catenulatum
Escherichia coli
Bifidobacterium adolescentis
Bifidobacterium longum
Bifidobacterium pseudocatenulatum
Collinsella aerofaciens
Klebsiella pneumoniae
Akkermansia muciniphila
Eggerthella lenta
Bilophila wadsworthia
Alistipes shahii
Gordonibacter pamelaeae

crude extract R
crude extract F
size selected R
size selected F
cosmid library R
cosmid library F

10^{-2} 10^{-1} 10^{0} 10^{1}


100
10
1
0.1
Number of consensus sequences per Mb
rpoN (σ^{54})
Bacteroides (σ^{ABfr})
rpoD (σ^{70})
rpoE (σ^{24})
rpoH (σ^{32})
F
R
F
R
F
R
crude extract
size-selected
cosmid library


A
B
Fold change in percent abundance
(cosmid library / crude extract)
CE0
10
1
0.1
0.01
0.001
CL0
70
60
50
40
30
Genome percent GC
0
10
20
30
40
Average number rpoD consensus per Mb