# Establishment of quantitative RNAi-based forward genetics in *Entamoeba histolytica* and identification of genes required for growth

Akhila Bettadapur[1], Samuel S. Hunter[2], Charles G. Barbieri[3], Matthew L. Settles[2], and Katherine S. Ralston[1*]

[1]Department of Microbiology and Molecular Genetics, University of California, Davis, USA

[2]Genome Center, University of California, Davis, USA

[3]SeqMatic, LLC, Fremont, California, USA

Running Head: Forward genetics and identification of genes required for *Entamoeba histolytica* growth

[*]Address correspondence to Katherine S. Ralston, ksralston@ucdavis.edu

**Abstract**

*Entamoeba histolytica* is a globally important pathogen that is dramatically understudied. Its challenging genome has limited tractability. To enable forward genetics, we constructed the first genome-wide *E. histolytica* RNAi knockdown mutant library. The library is designed to enable deep sequencing analysis for quantitative identification of mutants after selection. We developed a novel analysis pipeline to precisely define and quantify full-length gene fragments inferred from read mapping. We performed the first *E. histolytica* RNAi screen and identified slow growth mutants. Growth phenotypes were reproducible in independently generated mutants. Some of the genes targeted in slow growth mutants had annotated functions consistent with roles in growth or metabolism. Some targeted genes lacked annotation, supporting the power of forward genetics in uncovering gene function. This work opens up the possibility of applying genetics to improve understanding of this important pathogen. Moreover, the strategies behind this RNAi library, and its analysis, are novel, and can be applied to other organisms.

29 **Introduction**

30 *Entamoeba histolytica* is the causative agent of amoebiasis in humans. *E. histolytica*
31 is common in developing nations and is responsible for ~50 million infections/year [1].
32 Children under five are particularly vulnerable to diarrheal infections including amoebiasis,
33 which are responsible for close to half a million deaths annually [2]. In the first two years of
34 life, approximately 80% of infants in an urban slum in Dhaka, Bangladesh become infected
35 with *E. histolytica* [3]. Malnutrition and stunting are associated with childhood *E. histolytica*
36 infections [4]. Among children aged 12 – 23 months enrolled in the in the Global Enteric
37 Multicenter Study, *E. histolytica* infection was associated with the highest risk of death (5%)
38 between enrollment and follow up [5]. Invasive infections result in an estimated 15,500
39 deaths/year in children and 67,900 deaths/year among people of all ages [2]. There is no
40 vaccine and therapy is limited to nitroimidazoles [6]. The continued morbidity and mortality
41 indicate that current treatment and prevention approaches are insufficient.

42 Infection can be asymptomatic, or can lead to diarrhea, bloody diarrhea with
43 intestinal ulceration, and fatal abscesses outside of the intestine. Fitting with its capacity to
44 cause ulcers and abscesses, *E. histolytica* was named for its ability to lyse tissue (*histo-:*
45 tissue; *lytic-:* dissolving). Despite the name-giving property, it is not known how amoebic
46 trophozoites ("amoebae") damage tissue [7]. The major known virulence factors are the
47 surface Gal/GalNAc lectin [8, 9], which mediates attachment to human cells, and the
48 cysteine proteases, which cleave human substrates [10, 11]. It was recently shown that
49 amoebae kill human cells by "nibbling" them, in a process termed trogocytosis [12].
50 Amoebic trogocytosis contributes to tissue damage, and warrants further study for its

51 potential relevance to invasive disease [12]. However, trogocytosis is not yet well-defined at
52 the molecular level [13].

53 Despite its impact on human health, *E. histolytica* is dramatically understudied
54 relative to other parasites. Given the limitations of current therapies and the limitations of
55 our current understanding of pathogenesis, new tools are needed to study this infection.
56 Although the genome has been sequenced, it is difficult to identify genes relevant to
57 pathogenesis from sequence information alone. The genome is relatively difficult to
58 manipulate experimentally. *E. histolytica* is thought to be tetraploid [14]. The genome
59 contains a very large number of expanded gene families, with 897 protein families that
60 contain 4,564 genes, and represent 56% of the proteome [15, 16]. The nucleotide
61 composition of the *E. histolytica* genome is ~75% A + T, which creates challenges for
62 genomic approaches and has contributed to the relatively poor draft genome quality. The
63 genome is ~20 Mb with ~8,000 genes [15, 17, 18]. For one third of genes, no homologues
64 have been identified [17] and less than half of genes have been assigned gene ontology
65 (GO) terms [15]. Thus, with limited sequence homology information and a difficult genome,
66 functional approaches such as forward genetics are needed to identify genes that are
67 relevant to pathogenesis.

68 Historically, the genetic tractability of *E. histolytica* has been limited. CRISPR/Cas9
69 has not yet been applied for use in *E. histolytica.* Previous efforts to integrate sequences
70 into the genome or to genetically manipulate endogenous loci have been unsuccessful.
71 Stable transfection with episomal plasmids has been used in order to overexpress tagged
72 proteins or dominant negative mutants [19, 20]. Forward genetics have not yet been

73 established in *E. histolytica*, outside of two studies that used a plasmid-based
74 overexpression library to screen for phenotypes [21, 22]. Relatively recently, the
75 endogenous RNAi pathway in *E. histolytica* has been characterized, and exploited as a useful
76 tool for targeted gene knockdown.

77 The existence of an RNAi pathway in *E. histolytica* that could be exploited for gene
78 knockdown was implied by studies that successfully knocked down expression of the
79 amoebapore A gene [23]. Amoebae were stably transfected with a plasmid containing a
80 fragment of the amoebapore A promoter together with a segment of a retrotransposon and
81 a short T-rich stretch [23]. While this led to permanent silencing of amoebapore A, the
82 mechanism of action was unclear, and the expression of off-target genes was affected [24,
83 25]. It was possible to silence other genes, but only in amoebae that already contained the
84 amoebapore A silencing plasmid. A clonal line of stable transfectants carrying the
85 amoebapore A plasmid was named the "G3" strain, and was subsequently used to
86 knockdown additional genes in the context of an amoebapore A knockdown mutant
87 background [25]. Other strategies for gene knockdown have included expression of genes in
88 the antisense orientation [26], expression of long double stranded RNA [27], phagocytosis
89 of bacteria expressing dsRNA [28], and expression of short hairpin RNA [29]. These
90 techniques have been useful for reverse genetic studies, but not robust enough to be
91 effective for every gene of interest.

92 After it had already been exploited for targeted gene knockdown, the *E. histolytica*
93 RNAi pathway was characterized in detail. Small RNAs were found to associate with a Piwi-
94 related protein [30]. Small RNAs were 27nt in length, 5' polyphosphorylated, and mapped

95 antisense to genes [30]. Thus far, 5' polyphosphorylated siRNAs have only been seen in *E.*
96 *histolytica* and *Caenorhabditis elegans* [31]. The level of *E. histolytica* small RNAs was
97 inversely correlated with gene expression [30]. Supporting that these small RNAs are
98 relevant to previous approaches that exploited gene silencing in *E. histolytica,* in the G3
99 strain, 27nt small RNAs mapped to amoebapore A, and when additional genes were
100 silenced in this background, new small RNAs mapped to the newly silenced gene [32]. There
101 is no obvious Dicer homologue, but a potential noncanonical Dicer and Argonaute
102 homologues have been characterized [33, 34].

103 Because endogenous small RNAs are polyphosphorylated, it was initially a challenge
104 to determine how to effectively exploit this pathway for gene knockdown. A major
105 breakthrough was the discovery that RNAi knockdown could be induced by spreading of
106 silencing [35]. A new "trigger" plasmid approach was established, which relies on
107 endogenous small RNAs directed to a 132 bp fragment of an endogenously-silenced gene
108 (called the trigger) to drive spreading of silencing to a fragment of gene of interest that is
109 adjacent to the trigger [35]. This leads to long-term silencing of the gene of interest, even
110 after the trigger plasmid is removed, which may be due to the deposition of the repressive
111 histone mark H3K27 at the chromosomal copy of the silenced gene [36]. This trigger
112 plasmid approach has been successfully applied to knockdown many genes, and has been
113 adapted for use in *E. invadens* [37-40].

114 In other parasites, RNAi knockdown has been successfully adapted for forward
115 genetic approaches, and this has been very effective in uncovering genes relevant to
116 parasite metabolism [41, 42], DNA repair [43], virulence [44], transmission [45, 46], and

117 drug therapies [47-50]. Initially, these approaches relied on the generation of clonal lines
118 after selection, and Sanger sequencing analysis of individual clones to identify the knocked
119 down gene [41, 42]. Later approaches have used next generation sequencing on non-clonal
120 mutant populations in order to identify mutants, and their relative abundance [43-51]. The
121 latter approaches are useful because they can quantitatively identify mutants that are
122 enriched by selection, and mutants that are less abundant or are lost after selection.

123 To enable forward genetics, we constructed and validated the first genome-wide *E.*
124 *histolytica* RNAi knockdown mutant plasmid library. Its design allows for Illumina deep
125 sequencing, and our novel analysis pipeline allows for quantitative identification of plasmids
126 that are enriched or depleted after selection. Using a pilot version of the library, we
127 performed the first RNAi screen in *E. histolytica* and identified 12 slow growth mutants.
128 Genes that were knocked down in slow growing mutants included genes for which existing
129 annotation provided no clues on function. Independent knockdown mutants phenocopied
130 the growth defects seen in the library mutants. In the course of this work, new details of the
131 *E. histolytica* RNAi pathway were revealed, as we found that both frame and directionality
132 of the gene of interest are dispensable for silencing. These studies establish the first RNAi
133 screen in *E. histolytica*, thereby dramatically improving the tractability of this pathogen.

134 **Results**

135 **Approach for quantitative RNAi-based forward genetics**

136 We devised an overall strategy for RNAi-based forward genetics in *E. histolytica* (Fig.
137 1). For this approach, a library of genes would be introduced into the *E. histolytica* "trigger"
138 RNAi plasmid (pTrigger) [35], to create an RNAi library (Fig. 1a). Amoebae would then be
139 stably-transfected with the library, and exposed to a selective pressure or screening
140 strategy. Illumina deep sequencing analysis would then be used to sequence the inserts and
141 to compare the selected mutants to control, non-selected amoebae, in order to identify
142 mutants that are quantitatively enriched or depleted. We empirically determined that the
143 stable transfection efficiency with the backbone of pTrigger is ~1/1000 amoebae. This
144 relatively high stable transfection efficiency suggested that transfecting enough amoebae to
145 achieve genome-wide RNAi knockdown would be feasible.

146 The Illumina priming sequences that are necessary for sequencing are typically
147 added to samples by using a nested strategy with two sequential PCR reactions. We
148 modified the trigger plasmid in order to flank the insert with Illumina adaptor sequences,
149 creating pTriggerAdaptor, so that after extracting amoebic genomic DNA (gDNA), only a
150 single PCR step would be needed to generate DNA ready for sequencing (Fig. 1b). Using
151 fewer PCR reactions is expected to reduce the potential for PCR bias that can change the
152 representation of sequences. An additional benefit of this modification is that the original
153 sequences flanking the insert were A/T rich and were not optimal for primer design. The
154 insertion of these Illumina sequences into the knockdown plasmid did not change
155 knockdown efficacy (Fig. 1c).

156 **Construction of a plasmid RNAi library**

157 Building on these findings, we devised a strategy to construct a genome-wide
158 plasmid RNAi library (Fig. 2a). We reasoned that since ~50% of the genome is coding,
159 constructing a gDNA library would be a practical approach. To generate the plasmid library,
160 random fragments of gDNA were inserted into the pTriggerAdaptor plasmid. Library
161 construction was initially done on a small scale, to create a "pilot" library, where plasmids
162 were harvested from a limited number of bacterial colonies. Building on the results of the
163 pilot, a second library was constructed on a larger scale, to make a "final" library that was
164 expected to have genome-wide coverage.

165 Previous studies defined that an insert size ≥ 500 bp is sufficient for knockdown [35,
166 52]. Larger insert sizes are likely to lead to more efficient knockdown, but fragments > 1 kb
167 are problematic for Illumina sequencing. The pilot library was constructed using gDNA
168 fragments between ~400 – 600 bp. Building on the robust Illumina sequencing results from
169 the pilot library (described in detail below), we constructed the final library, where gDNA
170 fragments were larger (between ~500 – 900 bp) in order to maximize the efficacy of RNAi
171 knockdown.

172 To make the RNAi library, high quality, RNA-free, gDNA (Fig. S1a – b) was sonicated
173 and size-selected (Fig. S1c). Y-shaped adaptors were ligated to the fragmented DNA to
174 ensure that each fragment was flanked by sequences matching the trigger plasmid
175 upstream and downstream of the cloning site, in order to enable subsequent Gibson
176 cloning. Adaptor ligation was successful, with an appropriate size shift seen using
177 bioanalyzer analysis (Fig. 2b – c). Gibson cloning and plasmid preps were done *en masse*

178 with pooled plasmids. To check cloning efficiency, individual colonies were obtained and
179 analyzed. Restriction analysis demonstrated that the ligation efficiency was very high, since
180 63 out of 67 plasmids examined contained an insert (Fig. 2d).

181 **The RNAi library has genome-wide coverage**

182 Illumina sequencing was used to analyze the pilot and final libraries, and the
183 fragmented gDNA that was used to construct each library. Notably, our approach for
184 analysis of the sequencing data is novel and differs from previous approaches for analysis of
185 RNAi screens [43-51], since all results presented here are based on full-length DNA-
186 fragments inferred from read mapping results (Fig. S2), rather than on more traditional read
187 mapping coverage or CDS overlap counting strategies. Our approach was designed to better
188 represent the actual DNA fragments present in the plasmid library, instead of just the short
189 ends of these fragments that were sequenced on the Illumina platform. This strategy
190 allowed us to precisely define whether a fragment might target one or multiple genes, and
191 enabled a clearer view of fragment based genome coverage. This approach will also allow
192 us to readily quantify the representation of gene fragments in future RNAi screen
193 experiments, in order to identify specific fragments that are particularly impactful in a
194 particular screen.

195 There did not appear to be bias in gDNA fragmentation or construction of the
196 library, since the nucleotide and dinucleotide frequencies of the fragmented gDNA and the
197 plasmids were the same as the genome (Fig. 3a and S3a). Rarefaction curves and read
198 fragment analysis showed that sequencing was performed to a sufficient depth to identify
199 all unique fragments within the plasmid samples, but the same depth of sequencing was not

200 deep enough to identify all unique fragments within the fragmented gDNA samples (Fig. S3b
201 – c).
202 Illumina sequencing analysis was performed on separate batches of the plasmid
203 library that had been cloned on different days, and on technical sequencing replicates.
204 Similar read counts per fragment were obtained from each batch and each sequencing
205 replicate, showing that cloning and sequencing were reproducible and there was no
206 evidence for bias within these steps (Fig. S4a – b, Table S1). Analysis of the sequencing
207 results showed that the gDNA fragments that were used to create the pilot library had a
208 mean size of 421 bp, and the fragments in the pilot plasmid library had a mean size of 427
209 bp (Fig. 3b). The gDNA fragments that were used to create the final library had a mean size
210 of 540 bp, and the fragments in the pilot plasmid library had a mean size of 555 bp (Fig. 3b).
211 The nearly identical size ranges of the gDNA fragments and cloned fragments further
212 support that there was no bias during library construction.
213 Genome coverage plots of the final plasmid library over the 5 largest contigs in the
214 genome visually showed that coverage of these contigs was complete, and the depth of
215 coverage was consistent (Fig. 3c). The breadth and depth of coverage over a zoomed in
216 region of the largest contig is represented in Fig. 3d. On a genome-wide level, the genome
217 was well covered by fragments that were cloned into the library in both the forward and
218 reverse orientation relative to the orientation of coding sequences (Fig. 4, Table 1).
219 Supporting that the final library contained genome-wide coverage, 97.24% of the genome
220 was represented (Table 1). 8307 of 8333 annotated genes were represented, and the
221 majority (8301 genes) were represented by more than one unique fragment. Of the 26

222 genes that are missing, 22 are likely to be misannotated in the genome, since they were not
223 seen in any sample, including fragmented gDNA (Fig. S5). Additionally, for the majority of
224 these genes, there was no empirical evidence for gene expression in prior RNAseq or
225 proteomic studies (Tables S2 and S3). From these findings, we concluded that the final
226 plasmid library has genome-wide coverage. We named the final library pEhKDL, since it
227 represents an *E. histolytica* knockdown library.

228 **Forward genetic isolation of slow growing mutants**

229 While the final library was under construction, *E. histolytica* was stably transfected
230 with the pilot library. To determine if the RNAi library could be used to identify genes that
231 contribute to a phenotype, we screened for mutants with abnormal growth. To first ask
232 how many different plasmids were present per stably transfected amoeba, clonal lines were
233 obtained using limiting dilution. Sanger sequencing analysis of clonal lines was consistent
234 with only one unique plasmid (*i.e.,* one unique insert) per transfected amoeba (Fig. S6).

235 To screen for growth phenotypes, growth rates were measured for a selected
236 number of clonal lines. Slow Growth mutants 1 (SG1) and 2 (SG2) were isolated first and
237 chosen for further characterization. These mutants had inserts corresponding to
238 EHI_194310 and EHI_007010, which are annotated as a putative amino acid transporter and
239 a hypothetical protein, respectively (Table 2). These mutants had a statistically significant
240 growth defect (Fig. 5a). We have previously knocked down superoxide dismutase
241 (EHI_159160) and determined that this results in a growth defect; thus, superoxide
242 dismutase RNAi was used as a control. RT-PCR analysis showed that EHI_194310 and
243 EHI_007010 were knocked down in the corresponding mutants (Fig. 5b).

244 To determine if knockdown mutant phenotypes were reproducible, Gibson cloning
245 was used to generate independent knockdown plasmids that contained a fragment of either
246 EHI_194310 or EHI_007010, which differed from the fragment in the library mutants. *E.*
247 *histolytica* was stably transfected with each of these new plasmids to create the
248 independent knockdown mutants, SG1i and SG2i. Growth of heterogeneous, stable
249 transfectants was measured, and these mutants had statistically significant growth defects
250 (Fig. 5c). RT-PCR analysis demonstrated that the corresponding target genes were knocked
251 down (Fig. 5d). The library mutants and independently generated mutants each exhibited
252 statistically significant growth defects (Fig. 5e). Thus, the phenotypes seen in library
253 mutants were reproduced in independently generated mutants.

254 **Bioinformatic analysis of slow growing mutants**

255 From the pilot RNAi library transfection, ~64 additional clonal lines were obtained
256 and 10 additional clonal lines with statistically significant growth defects were identified
257 (Fig. S7; Table 2). The insert present in each clonal line was identified by Sanger sequencing,
258 and the results were again consistent with one unique plasmid per transfected amoeba.
259 Inserts ranged in size from 322 to 555 bp. In previous studies that used pTrigger for gene
260 silencing, the gene of interest for knockdown was in the forward orientation and in frame
261 with the trigger [35, 37-40, 52]. In the slow growth mutants, inserts were in the forward or
262 reverse orientation, and in and out of frame (Table 2). Inserts that were in the reverse
263 orientation and out of frame, like SG1, nonetheless led to measurable knockdown (Fig. 5b).
264 Both the directionality of the gene of interest, and the frame, are therefore dispensable for
265 knockdown.

266 Nine of the twelve slow growth mutants contained an insert that mapped to a gene
267 (Table 2). The three mutants with inserts that did not clearly map to an annotated gene
268 (SG3, SG9 and SG12) did map to potentially unannotated ORFs, and also partially mapped
269 (*i.e.,* contained stretches of homology) to annotated genes. BLAST was used to search for
270 homologs of the ORFs that mapped to SG3, SG9, and SG12, and homologs were found in *E.*
271 *dispar* for each ORF (Expect values of 4e -31, 1e -10, 1e -9, respectively), suggesting that
272 these ORFs are genes that were missed in the annotation of the *E. histolytica* genome. The
273 *E. histolytica* genome contains a very large number of gene families [15, 16], evidencing a
274 high level of functional redundancy. One mutant, SG5, contained an insert that mapped to
275 more than one gene (Table 2). Due to the level of identity between the multiple hits,
276 multiple related genes may be knocked down in this mutant.
277 Several mutants contained inserts that mapped to genes with annotated functions
278 consistent with roles necessary for cellular growth, such as an amino acid transporter, a
279 major facilitator transporter, and a splicing factor (Table 2). Three mutants contained
280 inserts with predicted roles in metabolic pathways, based on data from KEGG Metabolic
281 Pathways (https://www.kegg.jp), the MetaCyc Metabolic Pathway Database [53], and
282 Chemical Entities of Biological Interest (ChEBI) database [54] (Table 2). Several mutants
283 contained inserts that mapped to genes annotated as hypothetical, or that lacked
284 annotated domains (Table 2). Thus, in addition to validating the RNAi library approach, the
285 12 slow growth mutants that were uncovered shed light on genes that are required for *E.*
286 *histolytica* growth and emphasize how forward genetics can be used to assign functional
287 information to hypothetical proteins.

288 **Discussion**

289 We constructed and validated the first genome-wide *E. histolytica* RNAi knockdown
290 mutant library. We performed a screen for slow growth mutants and identified 12 new
291 mutants, providing a proof-of-concept that demonstrates the library can be used to identify
292 mutant phenotypes. Independent knockdown mutants phenocopied the growth defects
293 seen in the library mutants, showing that library mutant phenotypes are reproducible. Thus,
294 our studies both identify genes that are critical for growth, and establish RNAi-based
295 forward genetics in *E. histolytica* for the first time. The library represents a valuable new
296 tool for *E. histolytica*, that can now be applied towards the study of many processes.

297 Slow growth mutants were identified among clonal knockdown mutant lines. 12
298 mutants were identified among ~80 total clonal lines. Among the identified genes targeted
299 in these mutants, many genes had annotated functions consistent with roles in cellular
300 growth, and/or had predicted roles in metabolic pathways, based on available metabolic
301 pathway databases (https://www.kegg.jp/) [53, 54]. No genes were identified for which a
302 role in cellular growth would be implausible. Several of the identified genes were annotated
303 as hypothetical, or lacked annotated domains. All of the gene hits in our screen had
304 previous evidence of gene expression in RNAseq analyses [55, 56]. Thus, the slow growth
305 phenotype begins to provide functional information to genes for which existing annotation
306 does not inform understanding of protein function. All of the annotated gene hits in our
307 screen had homologues in the related human-infectious *Entamoeba* species *E. dispar, E.*
308 *nuttali* and *E. moshkovskii.* Thus, this type of screen may be informative in identifying
309 potential drug targets that are conserved among human-infectious *Entamoeba* species.

310 The *E. histolytica* genome contains numerous gene families [15, 16]. An advantage of
311 RNAi *vs.* alternative approaches (*e.g.*, gene knockouts) is the potential to simultaneously
312 interfere with multiple gene family members. RNAi is similarly advantageous for genomes
313 that are polyploid, which is also relevant to the potentially tetraploid [14] *E. histolytica*
314 genome. We identified one mutant that contained an insert that mapped to more than one
315 gene. With the high levels of sequence identity, it is likely that multiple genes are knocked
316 down in this mutant. We were able to identify a slow growth phenotype in this mutant,
317 which suggests that the RNAi library may be effective in identifying mutant phenotypes
318 even in the context of high levels of functional redundancy.
319 We have used pTrigger to knockdown many genes in our laboratory, and, in general,
320 have been able to recover knockdown mutants for essentially every targeted gene.
321 Compared to clonal lines, heterogeneous transfectants have typically had more variable
322 levels of knockdown, and sometimes appear to have lower levels of knockdown
323 (unpublished data). Among clonal lines obtained from heterogeneous transfectants, we
324 have seen wide variation, including clones with high levels of knockdown and clones that
325 lack knockdown. The SG1 and SG2 clonal lines initially obtained from the library appeared to
326 have an equal or higher level of knockdown than the heterogeneous independent SG1i and
327 SG2i mutants that targeted the same gene. The heterogeneous independent knockdown
328 mutant SG1i had a less severe growth phenotype than the clonal library mutant SG1, which
329 may be due to variation in knockdown in the heterogeneous population. Loss of gene
330 silencing (and loss of a growth phenotype) over time was noted in the heterogeneous SG2i
331 mutants (data not shown), consistent with the selective disadvantage of the growth

332 phenotype. Related to this point, when a selection is applied to a library mutant population,
333 it will be important to compare selected and non-selected controls, in order to differentiate
334 mutants that are lost due to a growth disadvantage *vs.* mutants that are lost due to
335 selection.

336 The library was designed to facilitate Illumina deep sequencing analysis in order to
337 quantitatively identify mutants that are enriched or lost after selection. We constructed
338 pilot and final versions of the library. Multiple lines of evidence supported that there was no
339 bias in library construction, including nucleotide and dinucleotide frequencies, and size
340 comparisons between the gDNA fragments and fragments in the final plasmids. Libraries
341 were prepared in batches that were cloned on different days. Similar read counts per
342 fragment were obtained from batches, again supporting that there was no bias in cloning or
343 sequencing analysis. The majority of the genome (97.24%) was covered in the final plasmid
344 library, and the majority of genes were represented (8307 out of 8333 annotated genes). 26
345 annotated genes are not represented. Of these, 22 genes were not seen in any sample,
346 including gDNA, suggesting that they represent errors in annotation. Consistent with this,
347 there was no published evidence for gene expression for the majority of these genes. The 4
348 remaining genes that were not seen in the final library had some of the lowest sequencing
349 coverage in the fragmented gDNA sample, suggesting that they were missed due to lack of
350 abundance in the fragmented gDNA that was used for library construction. Thus, with 8307
351 out of 8311 correctly annotated genes represented, and 97.24% of the genome, we
352 concluded that the library represented full genome coverage.

353 A future variation on this approach would be to make a library using fragmented

354 cDNA instead of gDNA. We reasoned that since ~50% of the genome is coding, gDNA is a
355 reasonable approach, since ~50% of the plasmids in the library would lead to gene silencing.
356 A limitation of a gDNA library is that some sequences such as rRNA and tRNA, which are
357 abundant in the genome, are represented in the library, and are not useful for identification
358 of mutant phenotypes. A cDNA library would remove these sequences, but would carry
359 other limitations. The most highly expressed genes would be overrepresented in a cDNA
360 library, and genes that are not expressed in the conditions in which cDNA was obtained
361 would not be represented in the library. The lower abundance of material for fragmentation
362 and cloning is also a consideration. mRNA is a very small proportion of total RNA, and in our
363 initial attempts to prepare cDNA for library construction, insufficient material was
364 recovered. With more initial total RNA, a cDNA library could be constructed in the future,
365 and would be complementary to the gDNA library that was constructed here.
366 We asked whether the frame or the directionality of the insert was critical for
367 knockdown using the trigger plasmid. In previous studies, the gene of interest for
368 knockdown was in the forward orientation and in frame with the trigger [35, 37-40, 52];
369 thus, the frame and orientation requirements had not been tested experimentally. Among
370 the slow growth mutants that we identified, inserts were in the forward or reverse
371 orientation, and in and out of frame. Note that the adaptor sequence that was inserted
372 between the trigger and the insert does not change the frame. Inserts that were in the
373 reverse orientation and out of frame still led to knockdown. Both the directionality of the
374 gene of interest, and the frame, are therefore dispensable for knockdown. Since the RNAi
375 pathway in *E. histolytica* is unusual and mechanistic details are still coming into focus, our

376 results add to understanding of this pathway. The lack of orientation and frame
377 requirements for knockdown is useful, since gDNA fragments are random and thus are
378 ligated into the library plasmid in either orientation and in any frame. Previous studies had
379 defined that a gene fragment ≥ 500 bp is sufficient for knockdown [35, 52]. We found that
380 fragments as small as 322 bp were sufficient to result in slow growth phenotypes,
381 suggesting that efficient knockdown is possible with smaller fragments.
382 *E. histolytica* is dramatically understudied relative to other parasites, even though it
383 heavily impacts human health, causing significant morbidity and mortality. The RNAi library
384 that we developed dramatically improves the tractability of *E. histolytica.* In other parasites,
385 RNAi screens have uncovered genes relevant to molecular and cellular biology, virulence
386 and therapeutics. We look forward to the new information about *E. histolytica* that can be
387 gleaned by applying RNAi screens in the future. Since the design of this RNAi library allows
388 for quantitative identification of mutants that are enriched or depleted after selection, we
389 expect that this library can be used to identify genes involved in a wide variety of processes,
390 including essential genes that are relevant to the development of new therapies.

391 **Materials and Methods**

392 *Cell culture*

393 *E. histolytica* HM1:IMSS (ATCC 30459) trophozoites were cultured in TYI-S-33 media,
394 supplemented with 15% heat inactivated Adult Bovine Serum (Gemini Bio Products), 80
395 units/ml penicillin and streptomycin (Gibco), and 2.3% Diamond Vitamin solution 80 Tween 40x
396 (Sigma Aldrich), at 35°C. Amoebae were harvested when flasks were approximately 80%
397 confluent.

398 *gDNA Extraction*

399 *E. histolytica* gDNA was harvested using the Quick-DNA miniprep kit (Zymo), cetyl
400 trimethylammonium bromide (CTAB) extraction, or the Wizard gDNA extraction kit (Promega).
401 The Quick-DNA miniprep kit (Zymo) was used according to manufacturer's instructions for cell
402 suspension samples. For CTAB extraction, cells were washed and then lysed at 55°C for 1 hour
403 in a solution of 0.1 M EDTA with 0.25% SDS and 0.1 mg/ml proteinase K. NaCl was added to a
404 final concentration of 0.7 M, pre-warmed CTAB was added to a final concentration of 1%, and
405 samples were incubated for 20 minutes at 65°C. An equal volume of chloroform was added,
406 and samples were centrifuged at 20,600 x *g* for 10 minutes. Supernatants were transferred into
407 fresh tubes, an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) was added, and
408 samples were centrifuged at 20,600 x *g* for 10 minutes. Supernatants were transferred into
409 fresh tubes, DNA was precipitated with 2.5 volumes of 100% ethanol at room temperature for 5
410 minutes, and samples were centrifuged at 20,600 x *g* for 10 minutes. DNA pellets were washed
411 in 70% ethanol. DNA was left to dry at room temperature and dissolved in water. For DNA
412 harvested using the Quick-DNA miniprep kit or CTAB extraction, RNAseIF (NEB) was used to

413 remove RNA contamination, according to manufacturer's instructions. The Wizard Genomic
414 DNA Purification kit (Promega) was used according to the manufacturer's instructions for the
415 tissue culture cells protocol. Because DNA harvested using the Wizard Genomic DNA
416 Purification kit was of higher quality, gDNA for plasmid library construction was prepared using
417 this kit.

418 *Plasmids and cloning*

419 The trigger plasmid [35] contains 132 base pairs of the trigger gene (EHI_048600) in the
420 pKT3M expression plasmid backbone [57]. Q5 Site Directed Mutagenesis (NEB) was used to
421 insert a SmaI restriction enzyme site after the trigger for cloning purposes, to create "pTrigger".
422 To create knockdown plasmids to target individual genes, pTrigger was linearized using SmaI
423 (NEB) and dephosphorylated using rSAP (NEB). A fragment of the gene of interest was amplified
424 from gDNA using PCR primers that contained homology to pTrigger, and the PCR product was
425 then cloned into the backbone using the Gibson Assembly Ultra kit (Synthetic Genomics Inc.),
426 following the manufacturer's instructions. See Table S4 for the sequences of primers that were
427 used during plasmid construction.

428 To test if the Illumina adaptor sequences interfere with knockdown, pTrigger plasmids
429 were created with and without adaptors. First, a pTrigger plasmid targeting a BAR/SH3 domain-
430 containing protein was created by inserting 812 bp of EHI_091500 (147 – 958). This plasmid was
431 further modified using Q5 Site Directed Mutagenesis (NEB) to add the custom adaptors
432 ACACTCTTTCCCTACACGACGCTCTTCCGATCT and GATCGGAAGAGCACACGTCTGAACTCCAGTCAC,
433 which are compatible with Illumina TruSeq sequencing primers, upstream and downstream of
434 the insert, respectively. After determining that the presence of the Illumina adaptors did not

435 interfere with knockdown, this construct was modified to remove the BAR/SH3 domain-
436 containing protein and replace it with a SmaI restriction site for further cloning purposes, to
437 create pTriggerAdaptor. pTriggerAdaptor was used as the backbone for plasmid library
438 generation.

439 Superoxide dismutase knockdown mutants were generated by inserting 597 bp of
440 EHI_159160 (1 – 597) into pTrigger. Independent knockdown mutants (SG1i) of SG1 were
441 generated by inserting 525 bp of EHI_194310 (32 – 556) into pTrigger. Independent knockdown
442 mutants (SG2i) of SG2 were generated by inserting 487 bp of EHI_007010 (59 – 545) into
443 pTrigger.

444 *Plasmid library construction*

445 *E. histolytica* gDNA was harvested using the Wizard Genomic DNA Purification kit
446 (Promega), following the manufacturer's instructions for the tissue culture cells protocol.
447 RNAse treatment is included in this protocol. Quality and size of gDNA was verified by Qubit
448 and gel electrophoresis. gDNA was further purified and concentrated using TailorMag
449 Purification beads (SeqMatic), and then sheared using a Covaris M220 sonicator. For size
450 selection, sheared gDNA was loaded into three lanes on a 6% TBE gel. The gel was stained using
451 SYBR Safe DNA Gel Stain (Invitrogen), and three bands, covering from ~400 – ~600 bp (for the
452 pilot library) and ~500 – ~900 bp (for the final library) were extracted from the gel using the
453 TailorCut Gel Cutter and Gel Breaker system (SeqMatic). gDNA was purified using TailorMag
454 Purification beads. The sizing and concentration of the purified material was determined by
455 TapeStation (Agilent). Purified gDNA corresponding to the second and third gel slices from each
456 lane was combined. gDNA was further purified and concentrated using TailorMag Purification

457 beads, and then processed using a TruSeq XT library prep kit (Illumina). All steps followed the
458 manufacturer's protocol except: 1. Ligation time was increased to 180 minutes, 2. Custom
459 adaptors were used instead of Illumina TruSeq adaptors. The adaptors that were ligated had
460 the following sequences: Adaptor 1: ACACTCTTTCCCTACACGACGCTCTTCCGATCCCC*T and
461 Adaptor 2: /5'Phos/GGGGATCGGGAGAGCACACGTCTGAACTCTAGTCAC. After adaptor ligation,
462 the gDNA fragments were amplified using 5 rounds of PCR with KAPA HiFi 2X master mix (KAPA
463 Biosystems). See Table S4 for primer sequences. After PCR, fragmented gDNA was purified and
464 concentrated using TailorMag Purification beads.

465 pTriggerAdaptor was linearized using SmaI (NEB) and dephosphorylated using rSAP
466 (NEB). Linearization was verified using gel electrophoresis. 25 ng of fragmented gDNA with
467 ligated adaptors was combined with 25 ng of linearized pTriggerAdaptor. The Gibson Assembly
468 Ultra kit (Synthetic Genomics Inc.) was used according to the manufacturer's instructions. High
469 Efficiency 5-alpha Competent *E. coli* (NEB) were transformed with 3 µl of the Gibson mix and
470 plated on four LB plates containing ampicillin (Sigma). We determined that the Gibson cloning
471 efficiency with pTriggerAdaptor was ~1/250, which was used to calculate the amount of insert
472 needed to generate a library at a given target level of coverage. With an average insert size of
473 ~550 bp, a minimum of ~36,354 plasmids would be needed to fully cover the genome at 1X
474 coverage. For the pilot library, one *E. coli* transformation reaction was performed and 4 plates
475 of bacteria were harvested into 200 ml of LB liquid broth containing ampicillin, and incubated at
476 37°C for 12-18 hours, shaking. For the final library, Gibson cloning, *E. coli* transformation, and
477 plasmid preparations were performed over four separate days to create four batches of final
478 library plasmids. The final plasmid library represents the combined total of the four batches.

479 For the final library, 25 *E. coli* transformation reactions were performed and 100 plates of
480 bacteria were harvested into LB liquid broth containing ampicillin, at a concentration of ~1,000
481 colonies per 200 ml. Plasmids were harvested using the GenElute HP plasmid maxiprep kit
482 (Sigma), using 200 ml of culture per plasmid purification column, and ethanol precipitated
483 according to the manufacturer's instructions.

484 *Illumina sequencing of gDNA fragments and plasmid samples*

485 Fragmented gDNA (with ligated custom adaptors; see 'plasmid library construction'), or
486 plasmid inserts were amplified with KAPA PCR (Kapa Biosystems) using full length TruSeq
487 Illumina primers, as follows: Primer 1:
488 GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[i7]ATCTCGTATGCCGTCTTCTGCTTG and Primer 2:
489 AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT. Three
490 technical replicates were sequenced for each sample. Samples were sequenced on an Illumina
491 MiSeq at either the DNA Technologies and Expression Analysis Core (University of California,
492 Davis), or on an Illumina MiSeq at SeqMatic, LLC. Illumina sequencing data analysis was
493 performed at the Bioinformatics Core (University of California, Davis).

494 *Illumina sequencing analysis*

495 Illumina sequencing data was demultiplexed using barcode sequences added during
496 library preparation with Illumina demultiplexing software. All reads were sequenced as PE
497 151bp runs. The resulting paired reads were then cleaned using HTStream
498 (https://github.com/ibest/HTStream) to 1) remove reads containing PhiX Control Library
499 sequences, 2) trim Illumina sequencing adaptors, and 3) discard primer dimers and other reads
500 where adaptor trimming was unsuccessful.

501 Cleaned reads were mapped against the *E. histolytica* release 46 genome assembly
502 (downloaded from https://amoebadb.org/common/downloads/release-
503 46/EhistolyticaHM1IMSS/fasta/data/AmoebaDB-46_EhistolyticaHM1IMSS_Genome.fasta) using
504 bwa mem v2.0pre1 [58]. Output from bwa was passed to samtools v1.10 [59] for sorting and
505 indexing, and also to extract reads from for visualization of specific loci. Results from mapping
506 and subsequent analysis were visualized and reviewed in Geneious Prime 2020.0.3
507 (https://www.geneious.com). Identifying unique fragments and counting their frequency was
508 done using a custom Python script. Briefly, the script iterates over read-sorted BAM records.
509 Read pairs that were mapped to the same contig with proper orientation, a minimum of 146
510 matched bases, and no more than four bases soft clipped from the 5' end of either read pair
511 were used to identify fragments. Read mapping orientation was used to infer fragment
512 orientation. A counter for each unique fragment was incremented for additional reads
513 supporting the fragment. Reads that were mapped in proper orientation but that did not meet
514 other criteria were compared against the fragment intervals with an incremented count
515 assigned to the highest scoring overlap. As a last step, the script generates summary statistics,
516 an interval table with counts for all samples, and SAF and gff annotation files for each sample.
517 Output from the Python script was then loaded into R v3.6.1 [60] for further analysis
518 and to generate figures. Briefly, data.table [61] was used to load the full set of intervals, order
519 them by genomic position, and to compute summaries by sample group. Tidyr [62]was used for
520 reformatting data, and most plots were generated with ggplot2 [63]. The vegan package [64]
521 was used for calculating saturation curve values. The GenomicRanges package [65] was used in
522 combination with gene annotation (downloaded from

523 https://amoebadb.org/common/downloads/release-

524 46/EhistolyticaHM1IMSS/gff/data/AmoebaDB-46_EhistolyticaHM1IMSS.gff) to determine and

525 count fragments targeting genes, all reported counts used a minimum overlap of 27bp.

526 Genome coverage plots were generated with karyoploteR [66] and heatmaps were plotted

527 using pheatmap [67]. Scripts and more details about the analysis are available at the GitHub

528 repository associated with this publication: https://github.com/samhunter/Bettadapur-et-al-

529 Entamoeba-histolytica-RNAi-library.

530 All sequencing data from this project have been deposited in the Sequence Read Archive

531 (SRA) database (https://www.ncbi.nlm.nih.gov/sra), under project number SUB7459667.

532 *Transfection and clonal lines*

533 Library plasmids were transfected into amoebae using Attractene transfection reagent

534 (Qiagen) [12, 68]. 50 ml M199 (Gibco medium M199 with Earle's salts, L-glutamine, and 2.2

535 g/liter sodium bicarbonate, without phenol red), was supplemented with 0.045 g L-Cysteine,

536 0.3 g HEPES, 0.005 g ascorbic acid (all from Sigma) and the pH was adjusted to 6.8. To 42.5 ml of

537 this medium, 7.5 ml of heat inactivated Adult Bovine Serum (Gemini bio-products) was added

538 to make transfection medium. Amoebae were resuspended in transfection medium at a

539 concentration of $2.5x10^5$ cells/ml, and then 1.8 ml of amoebae, 10 µg plasmid DNA (in 200 µl of

540 transfection medium), and 37.5 µl Attractene reagent (Qiagen) were combined into a 2.0 ml

541 screw-cap cryotube. Cryotubes were incubated horizontally at 35°C for 3 hours. Two cryotubes

542 per transfection were then added to a T25 flask containing 66 ml of fresh TYI media. 24 hours

543 later, 6 µg/ml geneticin (Invitrogen) was added. Stable transfectants were selected and

544 maintained with geneticin at 6 µg/ml.

545 The efficiency of stable transfection was determined by performing limiting dilution.
546 Three independent experiments were performed. Amoebae were transfected using the process
547 described above, except that 24 hours after transfection, amoebae were harvested and
548 resuspended in TYI at a concentration of $2x10^6$ cells/ml. Amoebae were then serially diluted
549 into 96-well plates, in the presence or absence of 6 μg/ml geneticin. Plates were incubated in
550 GasPak EZ pouch (BD Biosciences) and growth was monitored daily. The plates without
551 Geneticin were used to calculate the plating efficiency. In the plates without geneticin, growth
552 was seen in wells corresponding to approximately 2.5 cells on average. In the plates containing
553 geneticin, growth was seen in wells containing 781 to 3125 amoebae. Taken together, the
554 corrected stable transfection efficiency was ~1/1000 amoebae.
555 For generation of clonal mutant lines for phenotypic characterization, limiting dilution
556 was performed in 96-well plates in a GasPak EZ pouch (BD Biosciences). For identification of
557 mutants with growth defects, clonal lines that visually appeared to have slow growth were
558 selected for further analysis. RNA was extracted using the Direct-zol RNA kit (Zymo). gDNA
559 (which contains exogeneous plasmids) was extracted using the Quick-DNA miniprep kit (Zymo).
560 Sanger sequencing was used to identify the library insert associated with each clonal line.
561 Sequencing was carried out by Genewiz. To identify the gene targeted for knockdown in each
562 clonal line, the insert sequence was used to search the reference genome. Genes with at least
563 50nt of continuous sequence identity with the insert were considered hits. See Table S4 for
564 sequencing primer information.
565 *Growth analysis*

566 Clonal stable transfectants were resuspended to $8x10^3$ cells/ml and 1 ml of cells was
567 added to three tubes containing 13 ml fresh TYI with geneticin. Cells were incubated at 35°C. At
568 days 2, 4, and 6, one tube of cells was harvested per day and counted using a TC20 Automatic
569 cell counter (BioRad). Vector control transfected amoebae (containing pTriggerAdaptor) were
570 used as a negative control, and superoxide dismutase heterogeneous knockdown mutants were
571 used as a positive control for slow growth. Two technical replicates were performed per
572 experiment and three independent growth curve experiments were performed for each
573 mutant.

574 *RT-PCR*

575 RT-PCR was used to assess knockdown in transfectants. RNA was extracted using the
576 Direct-zol RNA kit (Zymo). DNAseI (Turbo) treatment and SuperScript Reverse Transcription II
577 (Invitrogen) was performed as per manufacturer instructions to obtain cDNA. RT-PCR was
578 performed on cDNA harvested from transfectants containing the following constructs: BAR/SH3
579 domain-containing protein with and without Illumina adaptors, Superoxide Dismutase;
580 SG1/SG1 independent knock down; and SG2/SG2 independent knockdown. GAPDH
581 (EHI_187020) RT-PCR was used as a positive control. See Table S4 for RT-PCR primer sequences.

582 *Statistical analysis*

583 GraphPad Prism was used for student's unpaired *t* test statistical analysis and
584 normalization (baseline removal). Mean values and standard deviations are shown in the
585 figures, with *t* test values reported as follows: ns = $p > 0.05$, * = $p < 0.05$, ** = $p < 0.01$, *** = $p <$
586 $0.001$, **** = $p < 0.0001$.

587 **Acknowledgements**

588 We thank Hannah Miller, Rene Suleiman, Scott Dawson and Su-Ju Lin for helpful
589 discussions. We thank Lutz Froenicke for helpful advice on designing the library to enable
590 Illumina sequencing. DNA sequencing was carried out by the DNA Technologies and
591 Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared
592 Instrumentation Grant 1S10OD010786-01. A.B. was supported by a fellowship from the UC
593 Davis Training Program in Biomolecular Technology. This work was funded by NIH grant
594 1R01AI146914 and a Pew Scholarship awarded to K.S.R.

595 **Author Contributions**

596 A.B. designed and performed all of the experiments, except fragmentation of gDNA.
597 C.G.B. fragmented gDNA and processed the fragments to enable subsequent plasmid library
598 construction. S.S.H. designed the analysis strategy for Illumina sequencing data, and
599 analyzed all Illumina sequencing data. M.L.S. contributed to the overall strategy for
600 constructing, sequencing and analyzing the plasmid library, and oversaw the Illumina
601 analysis approach. K.S.R. conceived of the overall approach and oversaw the design and
602 analysis of the experiments. A.B., S.S.H., and K.S.R. wrote the manuscript.

603 **Competing Interests**

604 The authors declare that they have no competing interests.

605 **References**


606 1. *WHO/PAHO/UNESCO report. A consultation with experts on amoebiasis. Mexico City,*
607 *Mexico 28-29 January, 1997.* Epidemiol Bull, 1997. **18**(1): p. 13-4.
608 2. *GBD 2015 Mortality and Causes of Death Collaborators. (2016). Global, regional, and*
609 *national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes*
610 *of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015,*
611 in *Lancet (London, England).* Elsevier. p. 1459-1544.
612 3. Gilchrist, C.A., et al., *Role of the Gut Microbiota of Children in Diarrhea Due to the*
613 *Protozoan Parasite Entamoeba histolytica.* The Journal of infectious diseases, 2016.
614 **213**(10): p. 1579-1585.
615 4. Mondal, D., et al., *Entamoeba histolytica-associated diarrheal illness is negatively*
616 *associated with the growth of preschool children: evidence from a prospective study.*
617 Trans R Soc Trop Med Hyg, 2006. **100**(11): p. 1032-8.
618 5. Kotloff, K.L., et al., *Burden and aetiology of diarrhoeal disease in infants and young*
619 *children in developing countries (the Global Enteric Multicenter Study, GEMS): a*
620 *prospective, case-control study.* The Lancet, 2013. **382**(9888): p. 209-222.
621 6. Gonzales, M.L.M., L.F. Dans, and J. Sio-Aguilar, *Antiamoebic drugs for treating amoebic*
622 *colitis.* The Cochrane database of systematic reviews, 2019. **1**(1): p. CD006085-
623 CD006085.
624 7. Ralston, K.S. and W.A. Petri, Jr *Tissue destruction and invasion by Entamoeba histolytica.*
625 Trends Parasitol, 2011. **27**(6): p. 254-63.
626 8. Ravdin, J.I. and R.L. Guerrant, *Role of adherence in cytopathogenic mechanisms of*
627 *Entamoeba histolytica. Study with mammalian tissue culture cells and human*
628 *erythrocytes.* J Clin Invest, 1981. **68**(5): p. 1305-13.
629 9. Saffer, L.D. and W.A. Petri, Jr., *Role of the galactose lectin of Entamoeba histolytica in*
630 *adherence-dependent killing of mammalian cells.* Infect Immun, 1991. **59**(12): p. 4681-3.
631 10. Keene, W.E., et al., *The major neutral proteinase of Entamoeba histolytica.* J Exp Med,
632 1986. **163**(3): p. 536-49.
633 11. Lidell, M.E., et al., *Entamoeba histolytica cysteine proteases cleave the MUC2 mucin in*
634 *its C-terminal domain and dissolve the protective colonic mucus gel.* Proc Natl Acad Sci U
635 S A, 2006. **103**(24): p. 9298-303.
636 12. Ralston, K.S., et al., *Trogocytosis by Entamoeba histolytica contributes to cell killing and*
637 *tissue invasion.* Nature, 2014. **508**(7497): p. 526-30.
638 13. Bettadapur, A., H.W. Miller, and K.S. Ralston, *Biting off what can be chewed:*
639 *Trogocytosis in health, infection and disease.* Infection and Immunity., 2020. **In Press**.
640 14. Willhoeft, U. and E. Tannich, *The electrophoretic karyotype of Entamoeba histolytica.*
641 Molecular and Biochemical Parasitology, 1999. **99**(1).
642 15. Lorenzi, H.A., et al., *New Assembly, Reannotation and Analysis of the Entamoeba*
643 *histolytica Genome Reveal New Genomic Features and Protein Content Information.*
644 PLOS Neglected Tropical Diseases, 2010. **4**(6): p. e716.
645 16. Wilson, I.W., et al., *Genetic Diversity and Gene Family Expansions in Members of the*
646 *Genus Entamoeba.* Genome biology and evolution, 2019. **11**(3): p. 688-705.

647 17. Loftus, B., et al., *The genome of the protist parasite Entamoeba histolytica.* Nature,
648 2005. **433**(7028): p. 865-8.
649 18. Lorenzi, H.A., et al., *New assembly, reannotation and analysis of the Entamoeba*
650 *histolytica genome reveal new genomic features and protein content information.* PLoS
651 neglected tropical diseases, 2010. **4**(6): p. e716-e716.
652 19. Hamann, L., R. Nickel, and E. Tannich, *Transfection and continuous expression of*
653 *heterologous genes in the protozoan parasite Entamoeba histolytica.* Proceedings of the
654 National Academy of Sciences, 1995. **92**(19): p. 8975-8979.
655 20. Hamann, L., H. Buß, and E. Tannich, *Tetracycline-controlled gene expression in*
656 *Entamoeba histolytica.* Molecular and Biochemical Parasitology, 1997. **84**(1): p. 83-91.
657 21. King, A.V., et al., *A genome-wide over-expression screen identifies genes involved in*
658 *phagocytosis in the human protozoan parasite, Entamoeba histolytica.* PloS one, 2012.
659 **7**(8): p. e43025-e43025.
660 22. Koushik, A.B., et al., *A genomewide overexpression screen identifies genes involved in*
661 *the phosphatidylinositol 3-kinase pathway in the human protozoan parasite Entamoeba*
662 *histolytica.* Eukaryotic cell, 2014. **13**(3): p. 401-411.
663 23. Bracha, R., et al., *Antisense inhibition of amoebapore expression in Entamoeba*
664 *histolytica causes a decrease in amoebic virulence.* Mol Microbiol, 1999. **34**(3): p. 463-
665 72.
666 24. Bracha, R., Y. Nuchamowitz, and D. Mirelman, *Transcriptional silencing of an*
667 *amoebapore gene in Entamoeba histolytica: molecular analysis and effect on*
668 *pathogenicity.* Eukaryotic cell, 2003. **2**(2): p. 295-305.
669 25. Bracha, R., et al., *Transcriptional silencing of multiple genes in trophozoites of*
670 *Entamoeba histolytica.* PLoS Pathog, 2006. **2**(5): p. e48.
671 26. Stock, R.P., et al., *Inhibition of gene expression in Entamoeba histolytica with antisense*
672 *peptide nucleic acid oligomers.* Nature Biotechnology, 2001. **19**(3): p. 231-234.
673 27. Kaur, G. and A. Lohia, *Inhibition of gene expression with double strand RNA interference*
674 *in Entamoeba histolytica.* Biochemical and Biophysical Research Communications, 2004.
675 **320**(4): p. 1118-1122.
676 28. Solis, C.F., et al., *Use of bacterially expressed dsRNA to downregulate Entamoeba*
677 *histolytica gene expression.* PloS one, 2009. **4**(12): p. e8424-e8424.
678 29. Linford, A.S., et al., *Short hairpin RNA-mediated knockdown of protein expression in*
679 *Entamoeba histolytica.* BMC microbiology, 2009. **9**: p. 38-38.
680 30. Zhang, H., et al., *Small RNAs with 5′-Polyphosphate Termini Associate with a Piwi-*
681 *Related Protein and Regulate Gene Expression in the Single-Celled Eukaryote Entamoeba*
682 *histolytica.* PLOS Pathogens, 2008. **4**(11): p. e1000219.
683 31. Sijen, T., et al., *Secondary siRNAs Result from Unprimed RNA Synthesis and Form a*
684 *Distinct Class.* Science, 2007. **315**(5809): p. 244.
685 32. Zhang, H., et al., *Nucleus-localized antisense small RNAs with 5'-polyphosphate termini*
686 *regulate long term transcriptional gene silencing in Entamoeba histolytica G3 strain.* J
687 Biol Chem, 2011. **286**(52): p. 44467-79.
688 33. Pompey, J.M., B. Foda, and U. Singh, *A Single RNaseIII Domain Protein from Entamoeba*
689 *histolytica Has dsRNA Cleavage Activity and Can Help Mediate RNAi Gene Silencing in a*
690 *Heterologous System.* PloS one, 2015. **10**(7): p. e0133740-e0133740.

691 34. Zhang, H., et al., *Functional Characterization of Entamoeba histolytica Argonaute*
692 *Proteins Reveals a Repetitive DR-Rich Motif Region That Controls Nuclear Localization.*
693 mSphere, 2019. **4**(5): p. e00580-19.
694 35. Morf, L., et al., *Robust gene silencing mediated by antisense small RNAs in the*
695 *pathogenic protist Entamoeba histolytica.* Nucleic acids research, 2013. **41**(20): p. 9424-
696 9437.
697 36. Foda, B.M. and U. Singh, *Dimethylated H3K27 Is a Repressive Epigenetic Histone Mark in*
698 *the Protist Entamoeba histolytica and Is Significantly Enriched in Genes Silenced via the*
699 *RNAi Pathway.* The Journal of biological chemistry, 2015. **290**(34): p. 21114-21130.
700 37. Pompey, J.M., L. Morf, and U. Singh, *RNAi pathway genes are resistant to small RNA*
701 *mediated gene silencing in the protozoan parasite Entamoeba histolytica.* PLoS One,
702 2014. **9**(9): p. e106477.
703 38. Manna, D., et al., *An NAD(+)-dependent novel transcription factor controls stage*
704 *conversion in Entamoeba.* Elife, 2018. **7**.
705 39. Rastew, E., L. Morf, and U. Singh, *Entamoeba histolytica rhomboid protease 1 has a role*
706 *in migration and motility as validated by two independent genetic approaches.* Exp
707 Parasitol, 2015. **154**: p. 33-42.
708 40. Suresh, S., et al., *Development of RNA Interference Trigger-Mediated Gene Silencing in*
709 *Entamoeba invadens.* Infect Immun, 2016. **84**(4): p. 964-975.
710 41. Morris, J.C., et al., *Glycolysis modulates trypanosome glycoprotein expression as*
711 *revealed by an RNAi library.* The EMBO journal, 2002. **21**(17): p. 4429-4438.
712 42. Drew, M.E., et al., *The adenosine analog tubercidin inhibits glycolysis in Trypanosoma*
713 *brucei as revealed by an RNA interference library.* J Biol Chem, 2003. **278**(47): p. 46596-
714 600.
715 43. Stortz, J.A., et al., *Genome-wide and protein kinase-focused RNAi screens reveal*
716 *conserved and novel damage response pathways in Trypanosoma brucei.* PLoS Pathog,
717 2017. **13**(7): p. e1006477.
718 44. Glover, L., et al., *VEX1 controls the allelic exclusion required for antigenic variation in*
719 *trypanosomes.* Proc Natl Acad Sci U S A, 2016. **113**(26): p. 7225-30.
720 45. Rico, E., et al., *Genome-wide RNAi selection identifies a regulator of transmission stage-*
721 *enriched gene families and cell-type differentiation in Trypanosoma brucei.* PLoS
722 pathogens, 2017. **13**(3): p. e1006279-e1006279.
723 46. Mony, B.M., et al., *Genome-wide dissection of the quorum sensing signalling pathway in*
724 *Trypanosoma brucei.* Nature, 2014. **505**(7485): p. 681-685.
725 47. Alsford, S., et al., *High-throughput decoding of antitrypanosomal drug efficacy and*
726 *resistance.* Nature, 2012. **482**(7384): p. 232-236.
727 48. Collett, C.F., et al., *Chemogenomic Profiling of Antileishmanial Efficacy and Resistance in*
728 *the Related Kinetoplastid Parasite Trypanosoma brucei.* Antimicrob Agents Chemother,
729 2019. **63**(8).
730 49. Trenaman, A., et al., *A post-transcriptional respiratome regulon in trypanosomes.*
731 Nucleic Acids Res, 2019. **47**(13): p. 7063-7077.
732 50. Zhang, N., et al., *Host-parasite co-metabolic activation of antitrypanosomal*
733 *aminomethyl-benzoxaboroles.* PLoS Pathog, 2018. **14**(2): p. e1006850.

734 51. Alsford, S., et al., *High-throughput phenotyping using parallel sequencing of RNA*
735 *interference targets in the African trypanosome.* Genome research, 2011. **21**(6): p. 915-
736 924.
737 52. Khalil, M.I., et al., *Technical advances in trigger-induced RNA interference gene silencing*
738 *in the parasite Entamoeba histolytica.* Int J Parasitol, 2016. **46**(3): p. 205-212.
739 53. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes and the*
740 *BioCyc collection of pathway/genome databases.* Nucleic Acids Res, 2016. **44**(D1): p.
741 D471-80.
742 54. Hastings, J., et al., *The ChEBI reference database and ontology for biologically relevant*
743 *chemistry: enhancements for 2013.* Nucleic Acids Res, 2013. **41**(Database issue): p.
744 D456-63.
745 55. Hon, C.C., et al., *Quantification of stochastic noise of splicing and polyadenylation in*
746 *Entamoeba histolytica.* Nucleic Acids Res, 2013. **41**(3): p. 1936-52.
747 56. Shabardina, V., et al., *Environmental adaptation of Acanthamoeba castellanii and*
748 *Entamoeba histolytica at genome level as seen by comparative genomic analysis.* Int J
749 Biol Sci, 2018. **14**(3): p. 306-320.
750 57. Saito-Nakano, Y., et al., *Rab5-associated vacuoles play a unique role in phagocytosis of*
751 *the enteric protozoan parasite Entamoeba histolytica.* J Biol Chem, 2004. **279**(47): p.
752 49497-507.
753 58. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.*
754 arXiv e-prints, 2013.
755 59. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009.
756 **25**(16): p. 2078-9.
757 60. team, R.c., *R: A language and environment for statistical computing. R Foundation for*
758 *Statistical Computing*. 2019: Vienna, Austria.
759 61. Dowle, M. and A. Srinivasan, *data.table: Extension of `data.frame`. R package version*
760 *1.12.8.* 2019.
761 62. Wickham, H. and L. Henry, *tidyr: Tidy Messy Data. R package version 1.0.2.* 2020.
762 63. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016: Springer-Verlag, New
763 York.
764 64. Oksanen, J., et al., *vegan: Community Ecology Package. R package version 2.5-6.* . 2019.
765 65. Lawrence, M., et al., *Software for Computing and Annotating Genomic Ranges.* PLOS
766 Computational Biology, 2013. **9**(8): p. e1003118.
767 66. Gel, B. and E. Serra, *karyoploteR: an R/Bioconductor package to plot customizable*
768 *genomes displaying arbitrary data.* Bioinformatics (Oxford, England), 2017. **33**(19): p.
769 3088-3090.
770 67. Kolde, R., *pheatmap: Pretty Heatmaps. R package version 1.0.12.* 2019.
771 68. Miller, H.W., R.L. Suleiman, and K.S. Ralston, *Trogocytosis by Entamoeba histolytica*
772 *Mediates Acquisition and Display of Human Cell Membrane Proteins and Evasion of Lysis*
773 *by Human Serum.* mBio, 2019. **10**(2).
774

775

776 **Figure Legends**

777 **Figure 1. RNAi library strategy.**

778 **a,** The plasmid RNAi library contains random fragments of gDNA cloned into the trigger
779 RNAi plasmid (pTriggerAdaptor). Different plasmid inserts are represented by different
780 colors in this schematic. A heterogeneous population of stably transfected RNAi knockdown
781 mutants can be exposed to selection or screening. To identify mutants that are represented
782 in the population after selection/screening, gDNA is extracted for analysis. gDNA extracted
783 from non-selected amoebae is used as a control. **b,** Schematic representation of the plasmid
784 RNAi library. The trigger plasmid is shown, with different inserts represented by different
785 colors. The "trigger" is a 132 bp fragment of EHI_048600. pTrigger was modified to create
786 pTriggerAdaptor, containing custom Illumina TruSeq adaptor sequences (IA) upstream and
787 downstream of the cloning site to facilitate amplicon sequencing. A single round of PCR is
788 needed to add the remaining Illumina sequences (shown in dark gray) and barcodes
789 necessary for Illumina deep sequencing analysis. **c,** An 812 bp fragment of a BAR/SH3
790 domain-containing protein (EHI_091530) was inserted into pTrigger with or without custom
791 Illumina TruSeq adaptors flanking the cloning site. The BAR/SH3 domain-containing protein
792 was in the forward orientation and in frame relative to the trigger sequence. Stably
793 transfected heterogeneous mutants were assessed for knockdown of the targeted gene
794 using RT-PCR, with (+) or without (-) reverse transcriptase. GAPDH (EHI_187020) was used
795 as a loading control.

796 **Fig. 2. RNAi library construction.**

797 **a,** Steps to create the library. *E. histolytica* gDNA was extracted and treated with RNAse.

798 Sonication, followed by gel purification, was used to isolate gDNA fragments. Fragments
799 were end-repaired, A-tailed, and ligated to Y-shaped adaptors. The adaptors contain
800 sequences corresponding to the region upstream and downstream of the cloning site in
801 pTriggerAdaptor. The Y-shaped design ensures that, after PCR to make these sequences
802 double-stranded, each gDNA fragment will be flanked by both the upstream/IA plasmid
803 sequence (represented in black), and the downstream/IA sequence (represented in gray).
804 Finally, the adaptor-ligated gDNA fragments were cloned into pTriggerAdaptor using Gibson
805 cloning. **b – c,** Bioanalyzer traces show appropriate size shift after adaptor ligation,
806 supporting that Y-shaped adaptors were successfully ligated. Example data are shown from
807 optimization experiments that were performed to establish the protocols for library
808 construction. **d,** Gibson cloning and plasmid preps were done *en masse* with pooled
809 plasmids. To check cloning efficiency, individual colonies were obtained and analyzed.
810 Shown are plasmid samples and restriction analysis from the pilot plasmid library
811 preparation. Restriction analysis was performed with restriction enzymes that flank the
812 cloning site (U, uncut; D, digested). Essentially all plasmids that were analyzed contained an
813 insert of the appropriate size range.

814 **Figure 3. The final RNAi library has genome-wide coverage.**

815 **a,** Nucleotide frequency analysis of the HM1:IMSS reference genome, the pilot and final
816 gDNA fragments, and the pilot and final plasmid libraries. Stacked bar plots show the
817 percentage of each base. The nucleotide composition of the gDNA fragments and the final
818 plasmid library is the same as the genome. **b,** Histograms showing the distribution of
819 fragments sizes for the unique fragments in the pilot and final gDNA fragments, and the

820 pilot and final plasmid libraries. Size in nucleotides is plotted on the X axis, and probability
821 distribution density is plotted on the Y axis. **c,** Genome coverage plots of the final library,
822 over the five largest contigs in the reference genome. The positions of unique fragments
823 and the depth in fragments overlapping each genomic position are shown. The positions of
824 genes are shown with black and green bars. Fragments that are in the same orientation as
825 genes that are in the forward orientation in the reference genome are shown in black, and
826 fragments that are in the same orientation as genes that are in the reverse orientation in
827 the reference genome are shown in green. **d,** Visual representation of fragment coverage in
828 the final library, over 100 kb of contig DS571145.

829 **Figure 4. Fragment directionality and coverage analysis.**

830 Heat map showing the number of unique fragments per gene in the pilot and final gDNA
831 fragments, and the pilot and final plasmid libraries. The number of unique fragments ($Log_2$) for
832 each gene are plotted (color value, see legend). Stranded fragments are in the same orientation
833 as genes. Unstranded fragments are the total number of fragments, both in the same
834 orientation as genes and in the opposite orientation as genes.

835 **Figure 5. Identification of slow growth mutants, with phenotypes that are reproducible in**
836 **independently-generated knockdown mutants.**

837 **a,** Growth analysis of clonal knockdown mutant lines obtained from screening the
838 heterogeneous library transfectants for growth defects. Amoebae transfected with
839 pTriggerAdaptor (vector control) were used as a negative control, and superoxide dismutase
840 (EHI_159160) knockdown mutants were used as a positive control for slow growth. The
841 library mutants SG1 and SG2 each exhibited a significant growth defect when compared to

842 vector control transfectants. **b,** RT-PCR analysis demonstrates that the genes corresponding
843 to the plasmid inserts in SG1 and SG2 (EHI_194310 and EHI_007010, respectively), were
844 knocked down. **c,** Knockdown mutants were independently generated by using fragments of
845 these genes that differed from the fragments in the library plasmids. These independent
846 mutants are referred to as SG1i and SG2i. Growth analysis was performed using
847 heterogeneous stable transfectants, and growth mutants exhibited a significant growth
848 defect when compared to vector control. **d,** RT-PCR analysis demonstrates that EHI_194310
849 and EHI_007010 were knocked down. **e,** Comparison of growth phenotypes in SG1, SG2,
850 SG1i and SG2i. Values were normalized to vector control transfectants. Mutants exhibited
851 statistically significant growth defects relative to vector control transfectants.

852 **Table Legends**

853 **Table 1. Library coverage analysis.**

854 Analysis of coverage in the pilot and final gDNA fragments, and the pilot and final plasmid
855 libraries. Genomic coverage (reported as bp covered, forward, reverse, and unstranded)
856 was calculated using fragments of the appropriate orientation, inferred from mapped reads
857 and the *coverage* function in the GenomicRanges R/Bioconductor package. The total
858 number of covered bp are indicated, and the corresponding percentage of the HM1:IMSS
859 reference genome that this represents. Forward fragments are in the same orientation as
860 genes that are in the forward orientation in the reference genome. Reverse fragments are
861 in the same orientation as genes that are in the reverse orientation in the reference
862 genome. Stranded fragments are in the same orientation as genes, and are the combined
863 total of both forward and reverse fragments. Unstranded fragments are the total number of
864 fragments, both in the same orientation as genes in the reference genome and in the
865 opposite orientation as genes. Genes that are targeted are those with at least one unique
866 fragment with greater than or equal to 27nt that overlaps the gene. Genes that are missed
867 are those that do not fit the criteria for targeted genes. Genes that are targeted by multiple
868 fragments are those with two or more unique fragments. Counts were calculated using the
869 countOverlaps function in the GenomicRanges package and annotation from AmoebaDB.
870 The number of unique fragments is the total number of unique fragments in each sample.

871 **Table 2. Identified slow growth mutants.**

872 Summary of growth phenotypes of SG1 – SG12 slow growth mutants, and analysis of the
873 sequences present in each corresponding library plasmid. Growth values represent cell

874 numbers normalized to vector control transfectants on day 4 of growth assays. The
875 corresponding raw data growth assay data are found in Fig. 5a and Fig. S5a, c, e). Mutants
876 exhibited statistically significant growth defects relative to vector control transfectants; *p*-
877 values from day 4 are shown. The insert sizes are shown in bp. Insert directionality relative
878 to the corresponding coding sequence in the annotated genome is shown. Insert frame
879 relative to the trigger sequence in the pTriggerAdaptor is shown. The accession numbers for
880 coding sequences that inserts correspond to are indicated, with available annotation
881 information and GO terms for each gene (NA, not applicable). Predicted roles in metabolic
882 pathways are shown; genes that are predicted to function in more than one metabolic
883 pathway are indicated as "several." These predictions are based on data from KEGG
884 Metabolic Pathways (https://www.kegg.jp), the MetaCyc Metabolic Pathway Database [53],
885 and Chemical Entities of Biological Interest (ChEBI) database [54].

886 **Supplementary Figure Legends**

887 **Figure S1. Preparation of RNA-free gDNA from *E. histolytica.***

888 Gel electrophoresis analysis of gDNA extracted using the Wizard gDNA extraction kit (Promega),
889 Quick-DNA miniprep kit (Zymo), and CTAB extraction. **a,** gDNA isolated using the Wizard kit had
890 no apparent RNA contamination, and the gDNA appeared to be of high quality. **b,** gDNA
891 isolated using the Quick-DNA kit or CTAB was contaminated with RNA. RNA contamination was
892 removed by treatment with RNAseIF and subsequent purification. **c,** Bioanalyzer traces of pilot
893 and final gDNA samples, after fragmentation and size selection.

894 **Figure S2. Fragment Mapping Analysis.**

895 Example view of mapped genomic fragments ("intervals") represented in the final plasmid
896 library. For comparison, the mapped paired end reads that give rise to each fragment are
897 shown below each fragment. Shown are results from final plasmid batch two, sequencing
898 replicate one, mapped to contig DS571153.

899 **Figure S3. Appropriate sequencing depth and lack of bias during library construction.**

900 **a,** Dinucleotide frequency analysis of the HM1:IMSS reference genome, the pilot and final gDNA
901 fragments, and the pilot and final plasmid libraries. Stacked bar plots show the percentage of
902 each dinucleotide. The dinucleotide composition of the gDNA fragments and the final plasmid
903 library is the same as the genome. **b – c,** Rarefaction curves for pilot (panel b) and final (panel c)
904 samples. gDNA fragment samples are shown in grey and plasmid samples are shown in orange.
905 The number of fragments detected with increasing sequence reads shows if sequencing was
906 performed to saturation. Three sequencing replicates are shown for pilot gDNA samples and
907 three sequencing replicates are shown for pilot plasmid samples. Three sequencing replicates

908 are shown for final gDNA samples and three sequencing replicates are shown for each of four
909 batches of final plasmid samples.

910 **Figure S4. Reproducible fragment and read counts from biological and technical replicates.**
911 "Batches" of the final plasmid library were cloned on different days, and represent biological
912 replicates. "Reps" correspond to independently prepared sequencing preparations, and
913 represent technical replicates. **a,** Box plot of read counts per fragment in plasmid samples, with
914 pilot plasmids in orange and final plasmids in grey. $Log_2$ read counts are shown; outliers
915 (outside of $2^{nd}$ and $98^{th}$ percentile) are plotted as points, and the box represents the $25^{th}$, and
916 $75^{th}$ percentile with median indicated as a bold line. Most fragments were represented by
917 multiple sequenced reads. **b,** Stacked bar plots of the total number of fragments detected per
918 sample shows that replicates are similar to each other in the number of fragments identified
919 and that roughly equal numbers of forward and reverse fragments were produced consistently.
920 Forward fragments are in the same orientation as genes that are in the forward orientation in
921 the reference genome. Reverse fragments are in the same orientation as genes that are in the
922 reverse orientation in the reference genome.

923 **Figure S5. Annotated genes that were missed or were least covered by sequencing reads.**
924 Heat map showing the number of unique fragments per gene, for the 50 least covered genes in
925 the pilot and final gDNA fragments, and the pilot and final plasmid libraries. The color intensity
926 indicates $Log_2$ fragment coverage, with dark blue indicating no coverage. Stranded fragments
927 are in the same orientation as genes. Unstranded fragments are the total number of fragments,
928 both in the same orientation as genes and in the opposite orientation as genes.

929 **Figure S6. Clonal transfectants contain only one library plasmid.**

930 SG1 gDNA was extracted using the Quick-DNA miniprep kit and PCR amplified using the CS5'
931 and CS3' primers that sit in the pTriggerAdaptor backbone. Sanger sequencing was performed
932 using the RNAi Sequencing Library F primer. Adaptor 2 starts at the "GGGG" series of
933 nucleotides, and upstream of this is the SG1 (EHI_194310) sequence. Noncompeting peaks in
934 the chromatogram suggest that there is one unique genome fragment present in this mutant.

935 **Figure S7. Identification of slow growing RNAi knockdown mutants.**

936 Growth analysis of mutants SG1 – SG12. **a, c, e,** Growth analysis demonstrates that selected
937 clonal knockdown mutant lines have significant growth defects compared to vector control
938 transfectants. **b, d, f**, Growth was normalized and compared to vector control transfectants
939 on each day. Mutants exhibited statistically significant growth defects relative to vector
940 control transfectants.

941 **Supplementary Table Legends**

942 **Table S1. Sequencing and fragment count statistics.**

943 Total number of paired reads mapping to the same contig (mappedPairs) and the number of
944 fragments inferred from these mapped pairs for each sample. Reads were mapped with bwa
945 mem, and processed from the resulting SAM formatted file using a custom Python script.
946 Three sequencing replicates were performed for pilot gDNA samples and three sequencing
947 replicates were performed for pilot plasmid samples. Three sequencing replicates were
948 performed for final gDNA samples and three sequencing replicates were performed for each of
949 four batches of final plasmid samples.

950 **Table S2. Expression data for annotated genes that were missing in the final plasmid library.**

951 Available empirical expression data for genes that were missing in the final plasmid library
952 (genes that are represented in blue in Fig. S3). Accession numbers, annotation information, and
953 the number of paralogs in the reference genome are indicated. Presence (+) or absence (-) in
954 gDNA and plasmid datasets is shown. Gene expression from available RNAseq datasets [55, 56]
955 is indicated, as the FPKM + 1 value. The highest value for FPKM + 1 for each gene is shown. The
956 total number of spectra that have been identified in proteomics studies is indicated.

957 **Table S3. Expression data for representative genes present in the final plasmid library.**

958 To provide a frame of reference for the expression data in Table S2, shown are the available
959 empirical expression data for 10 random, representative genes that were present in the final
960 plasmid library. Accession numbers, annotation information, and the number of paralogs in the
961 reference genome are indicated. Presence (+) or absence (-) in gDNA and plasmid datasets is
962 shown. Gene expression from available RNAseq datasets [55, 56] is indicated, as the FPKM + 1

963 value. The highest value for FPKM + 1 for each gene is shown. The total number of spectra that

964 have been identified in proteomics studies is indicated.

965 **Table S4. Primer sequences.**

966 Sequences of the primers used to generate and sequence plasmids, perform RT-PCR, and

967 create the plasmid RNAi library.

Figure 1

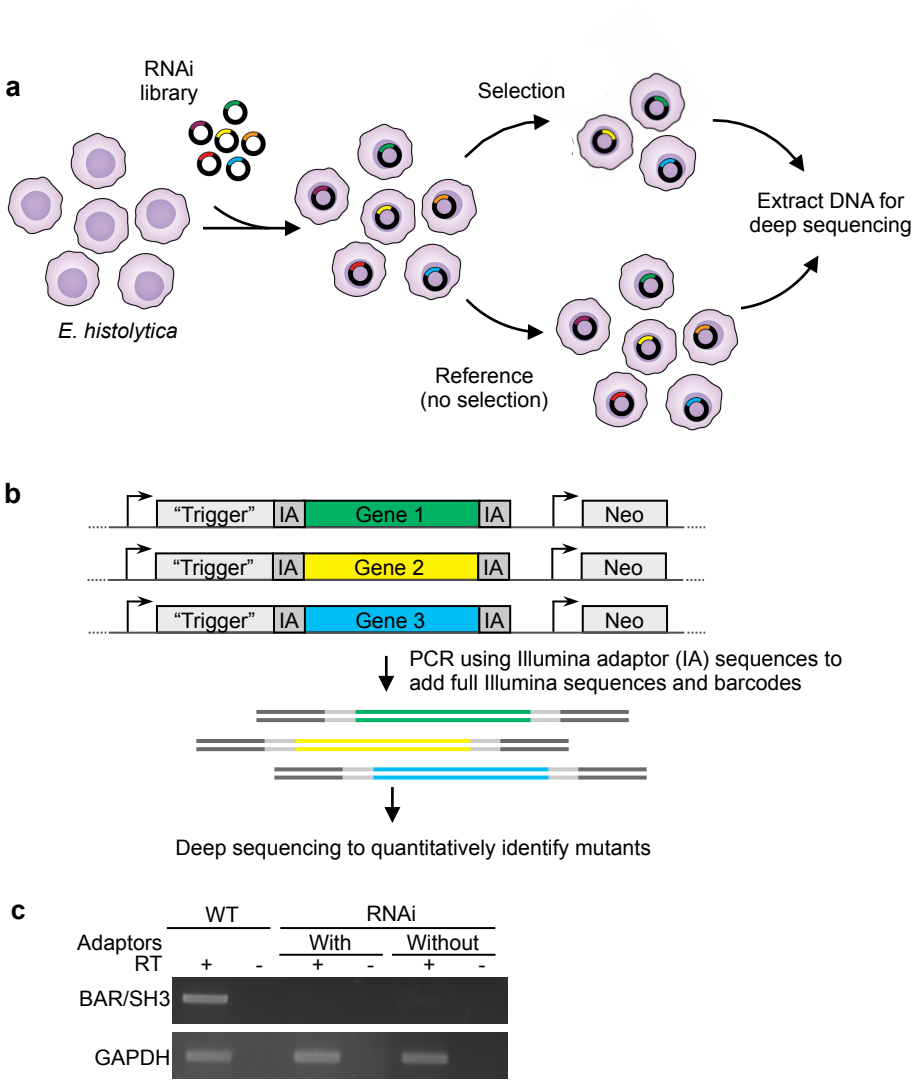
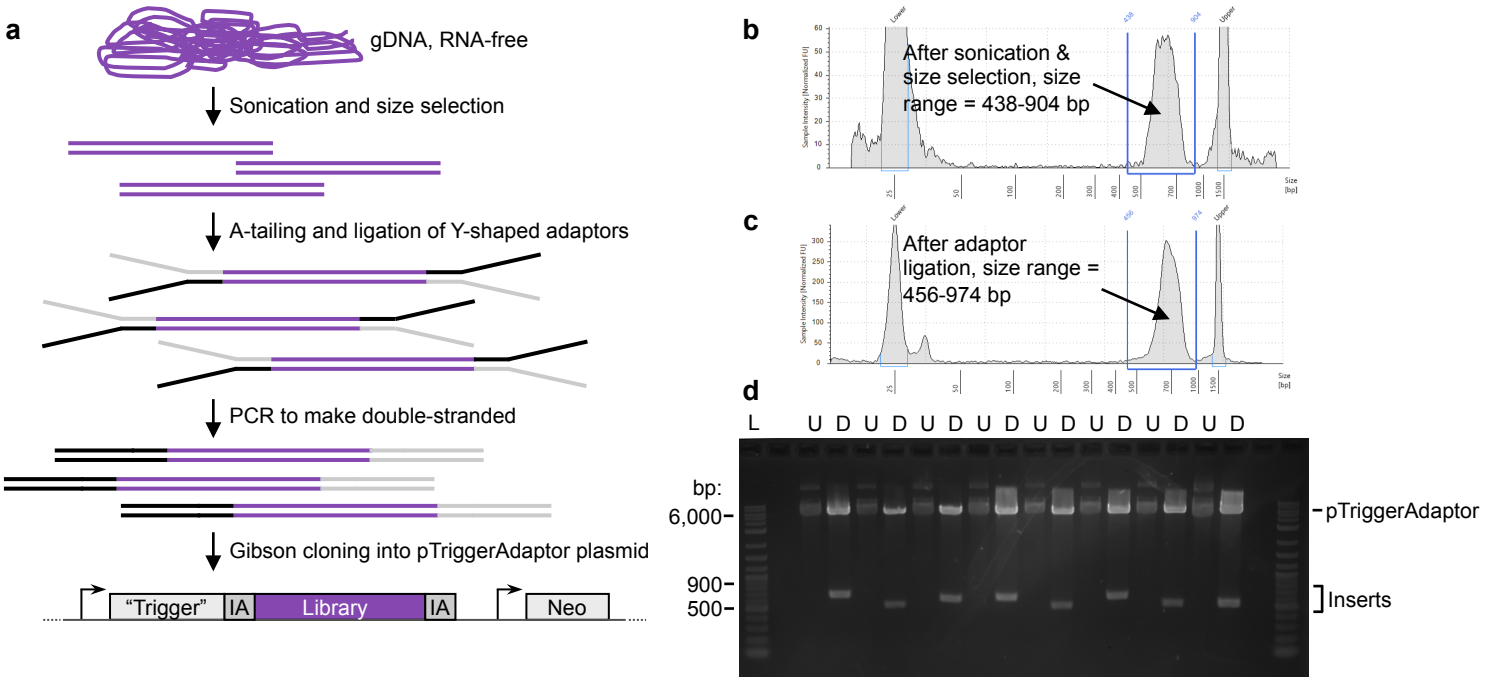

a
RNAi
library
E. histolytica
Selection
Reference
(no selection)
Extract DNA for
deep sequencing
b
"Trigger"
IA
Gene 1
Gene 2
Gene 3
Neo
PCR using Illumina adaptor (IA) sequences to
add full Illumina sequences and barcodes
Deep sequencing to quantitatively identify mutants
c
WT
RNAi
Adaptors
With
Without
RT
+
-
BAR/SH3
GAPDH

Figure 2


a
gDNA, RNA-free
Sonication and size selection
A-tailing and ligation of Y-shaped adaptors
PCR to make double-stranded
Gibson cloning into pTriggerAdaptor plasmid
"Trigger"
IA
Library
IA
Neo
b
After sonication & size selection, size range = 438-904 bp
c
After adaptor ligation, size range = 456-974 bp
d
L
U D U D U D U D U D U D U D U D
bp:
6,000
900
500
pTriggerAdaptor
Inserts

# Figure 3

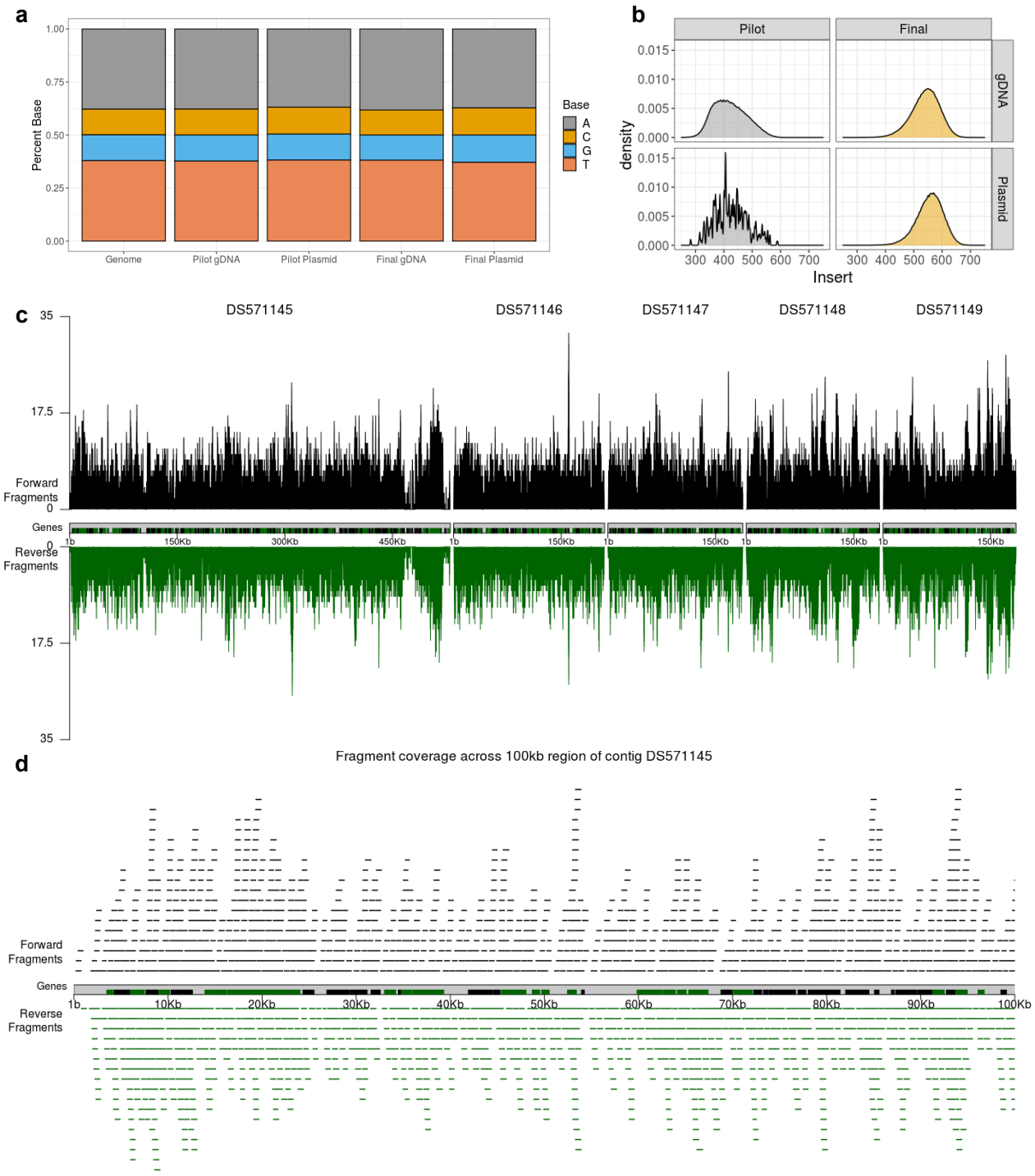
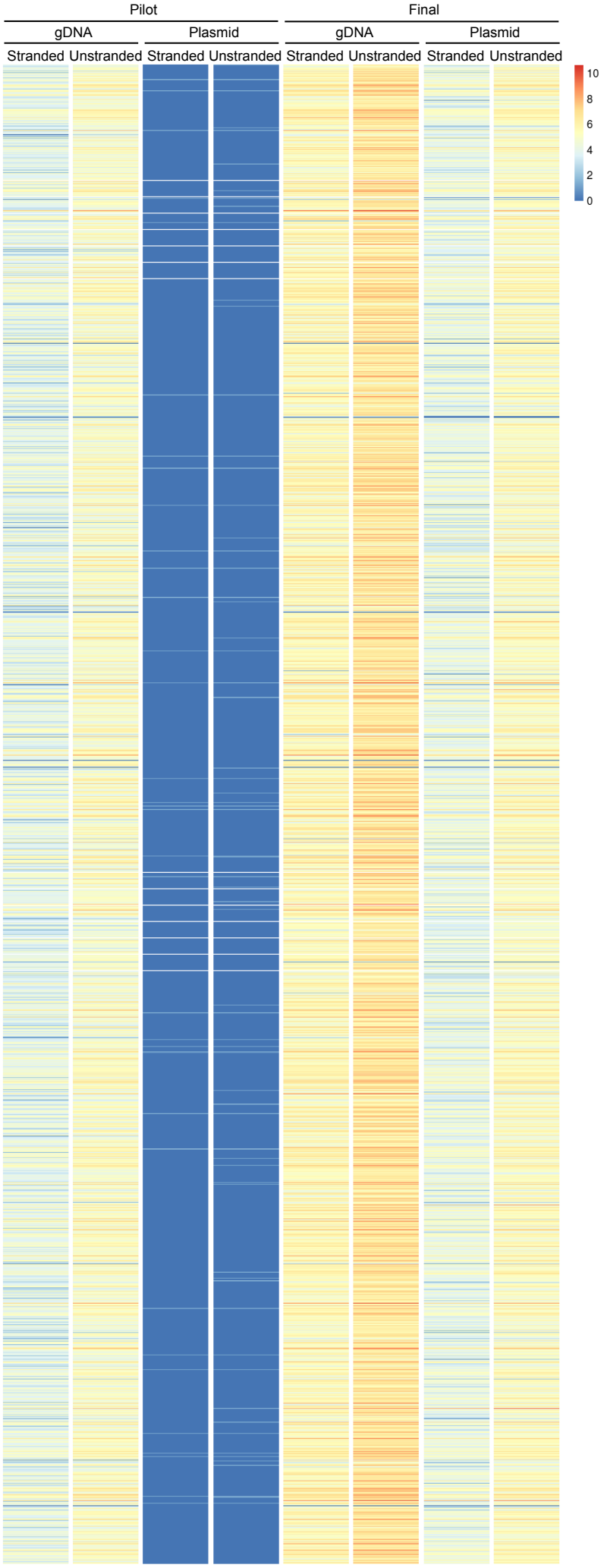

a
Percent Base
1.00
0.75
0.50
0.25
0.00
Genome
Pilot gDNA
Pilot Plasmid
Final gDNA
Final Plasmid
Base
A
C
G
T
b
Pilot
Final
gDNA
Plasmid
density
0.015
0.010
0.005
0.000
300 400 500 600 700
Insert
c
DS571145
DS571146
DS571147
DS571148
DS571149
35
17.5
0
Forward Fragments
Genes
Reverse Fragments
1b
150Kb
300Kb
450Kb
d
Fragment coverage across 100kb region of contig DS571145
Forward Fragments
Genes
Reverse Fragments
1b
10Kb
20Kb
30Kb
40Kb
50Kb
60Kb
70Kb
80Kb
90Kb
100Kb

Figure 4


Pilot
Final
gDNA
Plasmid
gDNA
Plasmid
Stranded
Unstranded
Stranded
Unstranded
Stranded
Unstranded
Stranded
Unstranded
10
8
6
4
2
0

# Figure 5

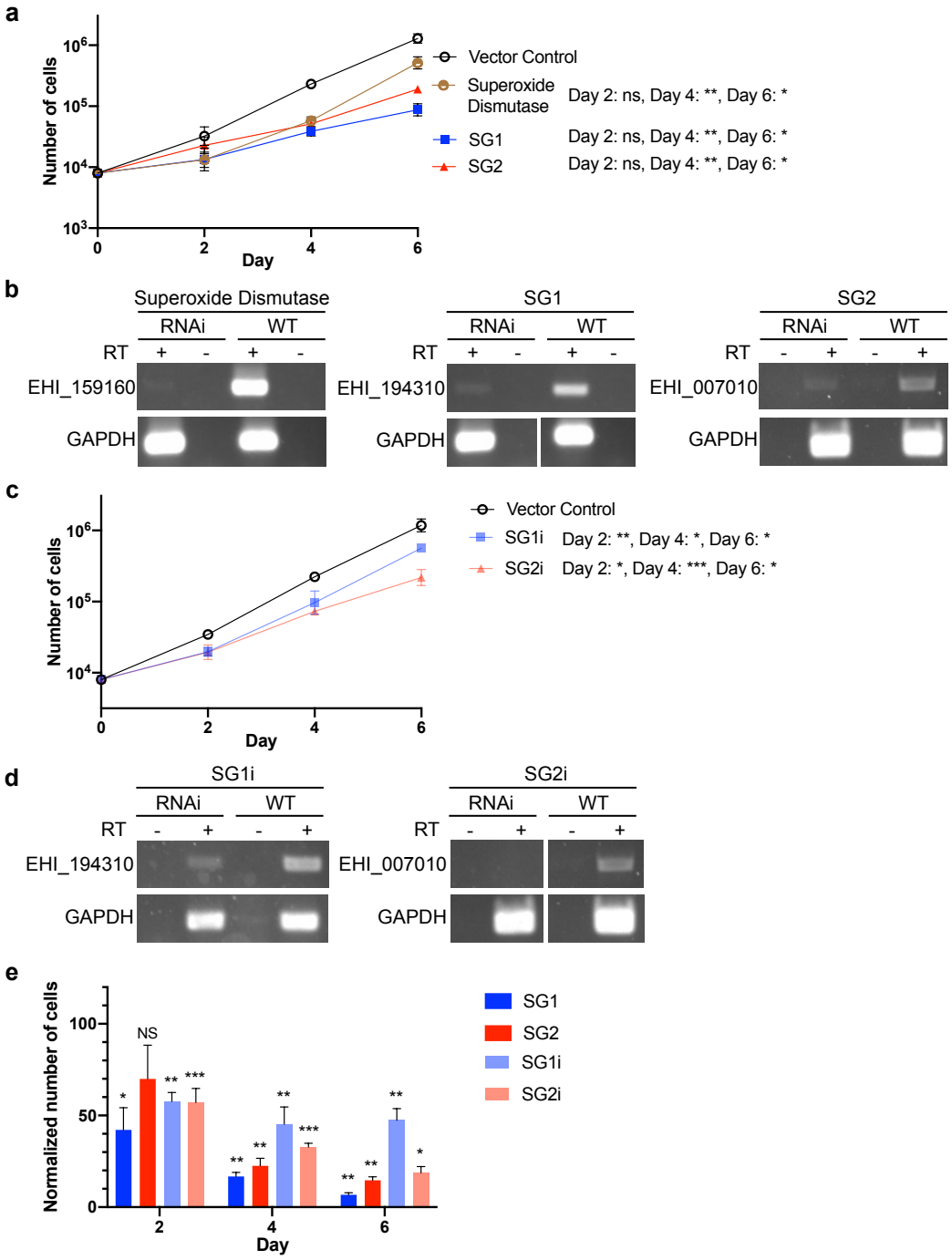

a
Number of cells
Day
Vector Control
Superoxide Dismutase
Day 2: ns, Day 4: **, Day 6: *
SG1
Day 2: ns, Day 4: **, Day 6: *
SG2
Day 2: ns, Day 4: **, Day 6: *
b
Superoxide Dismutase
SG1
SG2
RNAi
WT
RT
EHI_159160
EHI_194310
EHI_007010
GAPDH
c
Number of cells
Day
Vector Control
SG1i
Day 2: **, Day 4: *, Day 6: *
SG2i
Day 2: *, Day 4: ***, Day 6: *
d
SG1i
SG2i
RNAi
WT
RT
EHI_194310
EHI_007010
GAPDH
e
Normalized number of cells
Day
NS
SG1
SG2
SG1i
SG2i

Table 1

| | Pilot | | Final | |
|---|---|---|---|---|
| | gDNA | Plasmid | gDNA | Plasmid |
| **Covered bp, forward** | 19961090 | 93257 | 20500538 | 19730276 |
| **Covered %, forward** | 96.0 | 0.5 | 98.6 | 94.9 |
| **Covered bp, reverse** | 19949512 | 99274 | 20499865 | 19734222 |
| **Covered %, reverse** | 96.0 | 0.5 | 98.6 | 94.9 |
| **Covered bp, unstranded** | 20401392 | 191674 | 20571421 | 20225883 |
| **Covered %, unstranded** | 98.1 | 0.9 | 98.9 | 97.2 |
| **Genes targeted, stranded** | 8299 | 152 | 8310 | 8302 |
| **Genes missed, stranded** | 34 | 8181 | 23 | 31 |
| **Mean fragments per gene, stranded** | 21 | 0 | 59.2 | 23.3 |
| **Genes targeted by multiple fragments, stranded** | 8270 | 54 | 8308 | 8276 |
| **Genes targeted, unstranded** | 8310 | 305 | 8310 | 8307 |
| **Genes missed, unstranded** | 23 | 8028 | 23 | 26 |
| **Mean fragments per gene, unstranded** | 42.9 | 0.1 | 119.9 | 47.2 |
| **Genes targeted by multiple fragments, unstranded** | 8307 | 111 | 8310 | 8301 |
| **Number of unique fragments** | 537066 | 621 | 1509696 | 489981 |

# Table 2

| Name | Normalized growth | *p*-value | Insert size (bp) | Insert Direction | Insert Frame | Accession | Annotation | Annotated Domains | GO term | Metabolic Pathways |
|---|---|---|---|---|---|---|---|---|---|---|
| SG1 | 16.74 +/- 2.17 | ** | 451 | Reverse | Out | EHI_194310 | Putative amino acid transporter | Transmembrane amino acid transporter protein | NA | NA |
| SG2 | 22.58 +/- 4.10 | ** | 544 | Forward | In | EHI_007010 | Hypothetical protein | Prokaryotic membrane lipoprotein lipid attachment site profile | NA | NA |
| SG3 | 72.87 +/- 6.54 | * | 403 | Forward | Out | DS571580 | NA | NA | NA | NA |
| SG4 | 9.45 +/- 0.78 | *** | 351 | Reverse | Out | EHI_110390 | Transporter, major facilitator family | MFS general substrate transporter | NA | NA |
| SG5 | 14.50 +/- 0.91 | *** | 463 | Forward | In | EHI_024220 | Hypothetical protein | ARM repeat | binding | NA |
| | | | | Forward | In | EHI_160920 | Hypothetical protein | ARM repeat | binding | NA |
| | | | | Forward | In | EHI_171760 | Hypothetical protein | ARM repeat | Obsolete protein import into nucleus; Docking | NA |
| SG6 | 15.19 +/- 2.23 | **** | 380 | Reverse | In | EHI_093960 | Putative pre-mRNA splicing factor | PRP1 splicing factor, N-terminal | mRNA splicing, *via* spliceosome | Several |
| SG7 | 19.59 +/- 1.89 | *** | 555 | Reverse | In | EHI_009580 | Putative tyrosine kinase | Protein tyrosine kinase | Protein phosphorylation | Several |
| SG8 | 65.35 +/- 5.05 | *** | 370 | Forward | In | EHI_148960 | Hypothetical protein | NA | NA | NA |
| SG9 | 11.08 +/- 0.96 | *** | 352 | Forward | In | DS571952 | NA | NA | NA | NA |
| SG10 | 11.45 +/- 1.29 | *** | 425 | Forward | In | EHI_069320 | C2 domain containing protein | C2 domain (Calcium/lipid-binding domain, CaLB) | NA | Several |
| SG11 | 40.47 +/- 6.77 | *** | 408 | Reverse | Out | EHI_098820 | Hypothetical protein | RhoGEF domain; DBL homology domain | Regulation of Rho protein signal transduction | NA |
| SG12 | 18.09 +/- 1.83 | *** | 322 | Reverse | In | DS571214 | NA | NA | NA | NA |