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

Akhila Bettadapur[1], Samuel S. Hunter[2*], Rene L. Suleiman[1*], Maura C. Ruyechan[1], Wesley Huang[1], 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

[*]These authors contributed equally to this work

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

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

16 **Author Contributions**

17 A.B. designed and performed the experiments, except fragmentation of gDNA,
18 immunofluorescence, and growth curves with expression of full-length or truncated
19 proteins. R.L.S. performed immunofluorescence experiments. M.C.R. performed growth
20 curves with expression of full-length or truncated proteins. W.H. generated transfectants
21 expressing full-length SG2. C.G.B. fragmented gDNA and processed the fragments to enable
22 subsequent plasmid library construction. S.S.H. designed the analysis strategy for Illumina
23 sequencing data, and analyzed all Illumina sequencing data. M.L.S. oversaw the Illumina
24 analysis approach. K.S.R. performed growth curves, conceived of the overall approach and
25 oversaw the design and analysis of the experiments. A.B., S.S.H., and K.S.R. wrote the
26 manuscript.

Abstract

While *Entamoeba histolytica* remains a globally important pathogen, it is dramatically understudied. The tractability of *E. histolytica* has historically been limited, which is largely due to challenging features of its genome. To enable forward genetics, we constructed and validated the first genome-wide *E. histolytica* RNAi knockdown mutant library. This library allows for Illumina deep sequencing analysis for quantitative identification of mutants that are enriched or depleted after selection. We developed a novel analysis pipeline to precisely define and quantify gene fragments. We used the library to perform the first RNAi screen in *E. histolytica* and identified slow growth (SG) mutants. Among genes targeted in SG mutants, many had annotated functions consistent with roles in cellular growth or metabolic pathways. Some targeted genes were annotated as hypothetical or lacked annotated domains, supporting the power of forward genetics in uncovering functional information that cannot be gleaned from databases. While the localization of neither of the proteins targeted in SG1 nor SG2 mutants could be predicted by sequence analysis, we showed experimentally that SG1 localized to the cytoplasm and cell surface, while SG2 localized to the cytoplasm. Overexpression of SG1 led to increased growth, while expression of a truncation mutant did not lead to increased growth, and thus aided in defining functional domains in this protein. Finally, in addition to establishing forward genetics, we uncovered new details of the unusual *E. histolytica* RNAi pathway. These studies dramatically improve the tractability of *E. histolytica* and open up the possibility of applying genetics to improve understanding of this important pathogen.

Author Summary

Entamoeba histolytica is a globally important pathogen that is dramatically understudied. One of the major limitations of this organism is its challenging genome. RNAi is the state-of-the-art tool for genetic manipulation in *E. histolytica,* though the RNAi pathway has several noncanonical features. Here, we harnessed the RNAi pathway to enable forward genetics for the first time in this organism. We validated the RNAi library by performing the first *E. histolytica* RNAi screen and identified slow growth mutants. We showed that independently-generated mutants also exhibited slow growth phenotypes, and we characterized protein localization and domains for some of the identified slow growth genes. The RNAi library that we constructed enables modern, quantitative Illumina deep sequencing analysis to identify mutants that are enriched or depleted after selection. We developed a novel analysis pipeline to precisely define and quantify full-length gene fragments inferred from read mapping. Our approach differs from previous approaches for analysis of RNAi screens, and it better represents the actual DNA fragments and their quantities. This study dramatically improves the tractability of this important pathogen. Moreover, the strategies behind this RNAi library, and its analysis, are novel, and can be applied to other organisms.

65 **Introduction**

66 *Entamoeba histolytica* is the causative agent of amoebiasis in humans. *E. histolytica*
67 is common in developing nations and is responsible for ~50 million infections/year [1]. In an
68 endemic area, approximately 80% of infants are infected with *E. histolytica* [2]. Malnutrition
69 and stunting are associated with childhood *E. histolytica* infections [3]. In children enrolled
70 in the Global Enteric Multicenter Study, *E. histolytica* infection was associated with the
71 highest risk of death between enrollment and follow up [4]. Amoebiasis results in an
72 estimated 15,500 deaths/year in children and 67,900 deaths/year in people of all ages [5].
73 There is no vaccine and therapy is limited to nitroimidazoles [6]. The continued morbidity
74 and mortality indicate that current treatment and prevention approaches are insufficient.

75 Despite its impact on human health, *E. histolytica* is dramatically understudied
76 relative to other parasites. The genome is relatively difficult to manipulate. *E. histolytica* is
77 tetraploid [7]. The genome contains many expanded gene families, which comprise 56% of
78 the proteome [8, 9]. The genome is ~75% A + T, which creates challenges for genomic
79 approaches and has contributed to the relatively poor draft genome quality. The genome is
80 ~20 Mb with ~8,000 genes [8, 10]. For one third of genes, no homologues have been
81 identified [10] and less than half of genes have gene ontology (GO) terms [8]. Thus, with
82 limited sequence homology information and a difficult genome, functional approaches such
83 as forward genetics are needed to identify genes that are relevant to pathogenesis.

84 Historically, the genetic tractability of *E. histolytica* has been limited. Forward
85 genetics have not yet been established in *E. histolytica*, outside of two studies that used a
86 plasmid-based overexpression library to screen for phenotypes [11, 12]. More recently, the

87 endogenous RNAi pathway in *E. histolytica* has been characterized, and exploited as a useful
88 tool for targeted gene knockdown.

89 The existence of an RNAi pathway in *E. histolytica* that could be exploited for gene
90 knockdown was implied by an approach that knocked down expression of amoebapore A
91 [13]. However, the mechanism was unclear, and the expression of off-target genes was
92 affected [14, 15]. A clonal line of stable transfectants carrying the amoebapore A plasmid
93 was named the "G3" strain, and was subsequently used to knockdown additional genes in
94 the context of an amoebapore A knockdown mutant background [14]. Other strategies for
95 gene knockdown have included expression of genes in the antisense orientation [16],
96 expression of long double stranded RNA [17], phagocytosis of bacteria expressing dsRNA
97 [18], and expression of short hairpin RNA [19]. These techniques have been useful for
98 reverse genetic studies, but not robust enough to be effective for every gene of interest.

99 In later studies, the *E. histolytica* RNAi pathway was characterized in detail. Small
100 RNAs were found to be 27nt in length and 5' polyphosphorylated [20]. 5'
101 polyphosphorylated siRNAs have only been seen in *E. histolytica* and *Caenorhabditis elegans*
102 [21]. Supporting that these small RNAs are relevant to previous approaches that exploited
103 gene silencing in *E. histolytica,* in the G3 strain, 27nt small RNAs mapped to amoebapore A,
104 and when additional genes were silenced in this background, new small RNAs mapped to
105 the newly silenced gene [22]. There is no obvious Dicer homologue, but a potential
106 noncanonical Dicer and Argonaute homologues have been characterized [23, 24].

107 Because endogenous small RNAs are polyphosphorylated, it was initially a challenge
108 to exploit this pathway for gene knockdown. A major breakthrough was the discovery that

109 RNAi knockdown could be induced by spreading of silencing [25]. A new “trigger” plasmid
110 approach was established, which relies on endogenous small RNAs directed to short
111 fragment of an endogenously-silenced gene (called the trigger) to drive spreading of
112 silencing to a fragment of gene of interest adjacent to the trigger [25]. This leads to long-
113 term silencing of the gene of interest and the deposition of a repressive histone mark [26].
114 This approach has been successfully applied in many studies, and in *E. invadens* [27-29].
115 In other parasites, RNAi knockdown has been successfully adapted for forward
116 genetic approaches, and this has been very effective in uncovering genes relevant to
117 parasite metabolism [30, 31], DNA repair [32], virulence [33], transmission [34, 35], and
118 drug therapies [36-39]. Initially, these approaches relied on the generation of clonal lines
119 after selection, and Sanger sequencing analysis of individual clones to identify the knocked
120 down gene [30, 31]. More recent studies used next generation sequencing on non-clonal
121 mutant populations in order to identify mutants and quantify their relative abundance [32-
122 39]. The latter approaches are useful because they can quantitatively identify mutants that
123 are enriched by selection, and mutants that are less abundant or are lost after selection.
124 To enable forward genetics, we constructed and validated the first genome-wide *E.*
125 *histolytica* RNAi knockdown mutant plasmid library. Its design allows for Illumina deep
126 sequencing, and our novel analysis pipeline allows for quantitative identification of plasmids
127 that are enriched or depleted after selection. Using a pilot version of the library, we
128 performed the first RNAi screen in *E. histolytica* and identified slow growth mutants. These
129 studies establish the first RNAi screen in *E. histolytica*, thereby dramatically improving the
130 tractability of this pathogen.

131 **Results**

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

133 We devised an overall strategy for RNAi-based forward genetics in *E. histolytica* (Fig.
134 1). For this approach, a library of genes would be introduced into the *E. histolytica* "trigger"
135 RNAi plasmid (pTrigger) [25], to create an RNAi library (Fig. 1a). Amoebae would then be
136 stably-transfected with the library, and exposed to a selective pressure or screening
137 strategy. Illumina deep sequencing analysis would then be used to sequence the inserts and
138 to compare the selected mutants to control, non-selected amoebae, in order to identify
139 mutants that are quantitatively enriched or depleted. We empirically determined that the
140 stable transfection efficiency with the backbone of pTrigger is ~1/1000 amoebae. This
141 relatively high stable transfection efficiency suggested that transfecting enough amoebae to
142 achieve genome-wide RNAi knockdown would be feasible.

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

153 **Construction of a plasmid RNAi library**

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

162 Previous studies defined that an insert size ≥ 500 bp is sufficient for knockdown [25,
163 40]. Larger insert sizes are likely to lead to more efficient knockdown, but fragments > 1 kb
164 are problematic for Illumina sequencing. The pilot library was constructed using gDNA
165 fragments between ~400 – 600 bp. Building on the robust Illumina sequencing results from
166 the pilot library (described in detail below), we constructed the final library, where gDNA
167 fragments were larger (between ~500 – 900 bp) in order to maximize the efficacy of RNAi
168 knockdown.

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

175 with pooled plasmids. To check cloning efficiency, individual colonies were obtained and
176 analyzed. Restriction analysis demonstrated that the ligation efficiency was very high, since
177 63 out of 67 plasmids examined contained an insert (Fig. 2d).

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

179 Illumina sequencing was used to analyze the pilot and final libraries, and the
180 fragmented gDNA that was used to construct each library. Notably, our approach for
181 analysis of the sequencing data is novel and differs from previous approaches for analysis of
182 RNAi screens [32-39] since all results presented here are based on full-length DNA-
183 fragments inferred from read mapping results (Fig. S2), rather than on more traditional read
184 mapping coverage or CDS overlap counting strategies. Our approach was designed to better
185 represent the actual DNA fragments present in the plasmid library, instead of just the short
186 ends of these fragments that were sequenced on the Illumina platform. This strategy
187 allowed us to precisely define whether a fragment might target one or multiple genes, and
188 enabled a clearer view of fragment based genome coverage. This approach will also allow
189 us to readily quantify the representation of gene fragments in future RNAi screen
190 experiments, in order to identify specific fragments that are particularly impactful in a
191 particular screen.

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

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

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

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

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

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

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

251 **Localization and domain analysis of proteins targeted in slow growth mutants**

252 To further characterize SG1 and SG2, protein localization was examined. SG1 is
253 annotated as a putative amino acid transporter (Table 2), with predicted transmembrane
254 domains, but no predicted signal peptide sequence. SG2 is annotated as a hypothetical
255 protein (Table 2), with no annotated domains. SG2 also lacks predicted transmembrane
256 domains, predicted lipid modification sites, and a predicted signal peptide sequence. Thus,
257 the localization of neither protein could be definitively inferred from sequence analysis.

258 We found that SG1 was localized to the amoeba surface and was also detected
259 intracellularly in the cytoplasm, and intracellularly with a punctate distribution (Fig. 6a – 6b,
260 S8). Since these transfectants were heterogeneous, there was heterogeneity in the intensity
261 of SG1 staining. Interestingly, amoebae overexpressing full-length SG1 consistently
262 exhibited brighter DAPI staining (Fig. 6c) and/or more dividing cells with two nuclei,

263 compared to wild-type amoebae (Fig. 6a – 6b), suggesting that the growth rate was higher
264 in these amoebae. Growth analysis showed that amoebae overexpressing full-length SG1
265 grew faster than vector control transfectants (Fig. 6d). To gain a deeper understanding of
266 domains required for SG1 function, a C-terminal truncation mutant was overexpressed, and
267 since it did not lead to increased growth, we concluded that this truncation interfered with
268 protein function (Fig. 6d).

269 Unlike SG1, SG2 was not detectable on the amoeba surface. Instead, SG2 exhibited
270 staining throughout the amoeba cytoplasm (Fig. 7a, S8). Since these transfectants were
271 heterogeneous, there was heterogeneity in the intensity of SG2 staining. Overexpression of
272 full-length SG2 was not associated with increased growth, and no difference was detected
273 in the intensity of DAPI staining between SG2 overexpressors and wild-type cells (Fig. 7b).
274 Unlike the rabbit antibody used to localize SG1 (Fig. 6a – 6b), the mouse antibody used to
275 localize SG2 was associated with some background nuclear staining in wild-type cells (Fig.
276 7b, 7c, S8). By contrast, nuclear staining was not pronounced in SG2 overexpressors (Fig. 7b,
277 7c, S8).

278 **Bioinformatic analysis of slow growing mutants**

279 From the pilot RNAi library transfection, ~64 additional clonal lines were obtained
280 and 10 additional clonal lines with statistically significant growth defects were identified
281 (Fig. S7; Table 2). The insert present in each clonal line was identified by Sanger sequencing,
282 and the results were again consistent with one unique plasmid per transfected amoeba.
283 Inserts ranged in size from 322 to 555 bp. In previous studies that used pTrigger for gene
284 silencing, the gene of interest for knockdown was in the forward orientation and in frame

285 with the trigger [25, 27-29, 40, 41]. In the slow growth mutants, inserts were in the forward
286 or reverse orientation, and in and out of frame (Table 2). Inserts that were in the reverse
287 orientation and out of frame, like SG1, nonetheless led to measurable knockdown (Fig. 5b).
288 Both the directionality of the gene of interest, and the frame, are therefore dispensable for
289 knockdown.

290 Nine of the twelve slow growth mutants contained an insert that mapped to a gene
291 (Table 2). The three mutants with inserts that did not clearly map to an annotated gene
292 (SG3, SG9 and SG12) did map to potentially unannotated ORFs, and also partially mapped
293 (*i.e.,* contained stretches of homology) to annotated genes. BLAST was used to search for
294 homologs of the ORFs that mapped to SG3, SG9, and SG12, and homologs were found in *E.*
295 *dispar* for each ORF (Expect values of 4e -31, 1e -10, 1e -9, respectively), suggesting that
296 these ORFs are genes that were missed in the annotation of the *E. histolytica* genome. The
297 *E. histolytica* genome contains a very large number of gene families [8, 9], evidencing a high
298 level of functional redundancy. One mutant, SG5, contained an insert that mapped to more
299 than one gene (Table 2). Due to the level of identity between the multiple hits, multiple
300 related genes may be knocked down in this mutant.

301 Several mutants contained inserts that mapped to genes with annotated functions
302 consistent with roles necessary for cellular growth, such as an amino acid transporter, a
303 major facilitator transporter, and a splicing factor (Table 2). Three mutants contained
304 inserts with predicted roles in metabolic pathways, based on data from KEGG Metabolic
305 Pathways (https://www.kegg.jp), the MetaCyc Metabolic Pathway Database [42], and
306 Chemical Entities of Biological Interest (ChEBI) database [43] (Table 2). Several mutants

307 contained inserts that mapped to genes annotated as hypothetical, or that lacked

308 annotated domains (Table 2). Thus, in addition to validating the RNAi library approach, the

309 12 slow growth mutants that were uncovered shed light on genes that are required for *E.*

310 *histolytica* growth and emphasize how forward genetics can be used to assign functional

311 information to hypothetical proteins.

312 **Discussion**

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

321 Slow growth mutants were identified among clonal knockdown mutant lines. 12
322 mutants were identified among ~80 total clonal lines. Among the identified genes targeted
323 in these mutants, many genes had annotated functions consistent with roles in cellular
324 growth, and/or had predicted roles in metabolic pathways, based on available metabolic
325 pathway databases (https://www.kegg.jp/) [42, 43]. No genes were identified for which a
326 role in cellular growth would be implausible. Several of the identified genes were annotated
327 as hypothetical, or lacked annotated domains. All of the gene hits in our screen had
328 previous evidence of gene expression in RNAseq analyses [44, 45]. Thus, the slow growth
329 phenotype begins to provide functional information to genes for which existing annotation
330 does not inform understanding of protein function. All of the annotated gene hits in our
331 screen had homologues in the related human-infectious *Entamoeba* species *E. dispar, E.*
332 *nuttali* and *E. moshkovskii.* Thus, this type of screen may be informative in identifying
333 potential drug targets that are conserved among human-infectious *Entamoeba* species.

334 The *E. histolytica* genome contains numerous gene families [8, 9]. An advantage of
335 RNAi *vs.* alternative approaches (*e.g.,* gene knockouts) is the potential to simultaneously
336 interfere with multiple gene family members. RNAi is similarly advantageous for genomes
337 that are polyploid, which is also relevant to the tetraploid [7] *E. histolytica* genome. We
338 identified one mutant that contained an insert that mapped to more than one gene. With
339 the high levels of sequence identity, it is likely that multiple genes are knocked down in this
340 mutant. We were able to identify a slow growth phenotype in this mutant, which suggests
341 that the RNAi library may be effective in identifying mutant phenotypes even in the context
342 of high levels of functional redundancy.
343 We have used pTrigger to knockdown many genes in our laboratory, and, in general,
344 have been able to recover knockdown mutants for essentially every targeted gene.
345 Compared to clonal lines, heterogeneous transfectants have typically had more variable
346 levels of knockdown, and sometimes appear to have lower levels of knockdown [46]. We
347 obtain clonal lines by limiting dilution, but another effective available method is cloning by
348 isolating *E. histolytica* colonies on soft agar plates [47]. Among clonal lines obtained from
349 heterogeneous transfectants, we have seen wide variation, including clones with high levels
350 of knockdown and clones that lack knockdown. The SG1 and SG2 clonal lines initially
351 obtained from the library appeared to have an equal or higher level of knockdown than the
352 heterogeneous independent SG1i and SG2i mutants that targeted the same gene. The
353 heterogeneous independent knockdown mutant SG1i had a less severe growth phenotype
354 than the clonal library mutant SG1, which may be due to variation in knockdown in the
355 heterogeneous population. Loss of gene silencing (and loss of a growth phenotype) over

356 time was noted in the heterogeneous SG2i mutants (data not shown), consistent with the
357 selective disadvantage of the growth phenotype. Related to this point, when a selection is
358 applied to a library mutant population, it will be important to compare selected and non-
359 selected controls, in order to differentiate mutants that are lost due to a growth
360 disadvantage *vs.* mutants that are lost due to selection.

361 The library was designed to facilitate Illumina deep sequencing analysis in order to
362 quantitatively identify mutants that are enriched or lost after selection. We constructed
363 pilot and final versions of the library. Multiple lines of evidence supported that there was no
364 bias in library construction, including nucleotide and dinucleotide frequencies, and size
365 comparisons between the gDNA fragments and fragments in the final plasmids. Libraries
366 were prepared in batches that were cloned on different days. Similar read counts per
367 fragment were obtained from batches, again supporting that there was no bias in cloning or
368 sequencing analysis. The majority of the genome (97.24%) was covered in the final plasmid
369 library, and the majority of genes were represented (8307 out of 8333 annotated genes). 26
370 annotated genes are not represented. Of these, 22 genes were not seen in any sample,
371 including gDNA, suggesting that they represent errors in annotation. Consistent with this,
372 there was no published evidence for gene expression for the majority of these genes. The 4
373 remaining genes that were not seen in the final library had some of the lowest sequencing
374 coverage in the fragmented gDNA sample, suggesting that they were missed due to lack of
375 abundance in the fragmented gDNA that was used for library construction. Thus, with 8307
376 out of 8311 correctly annotated genes represented, and 97.24% of the genome, we
377 concluded that the library represented full genome coverage.

378 A future variation on this approach would be to make a library using fragmented
379 cDNA instead of gDNA. We reasoned that since ~50% of the genome is coding, gDNA is a
380 reasonable approach, since ~50% of the plasmids in the library would lead to gene silencing.
381 A limitation of a gDNA library is that some sequences such as rRNA and tRNA, which are
382 abundant in the genome, are represented in the library, and are not useful for identification
383 of mutant phenotypes. A cDNA library would remove these sequences, but would carry
384 other limitations. The most highly expressed genes would be overrepresented in a cDNA
385 library, and genes that are not expressed in the conditions in which cDNA was obtained
386 would not be represented in the library. The lower abundance of material for fragmentation
387 and cloning is also a consideration. mRNA is a very small proportion of total RNA, and in our
388 initial attempts to prepare cDNA for library construction, insufficient material was
389 recovered. With more initial total RNA, a cDNA library could be constructed in the future,
390 and would be complementary to the gDNA library that was constructed here.
391 We asked whether the frame or the directionality of the insert was critical for
392 knockdown using the trigger plasmid. In previous studies, the gene of interest for
393 knockdown was in the forward orientation and in frame with the trigger [25, 27-29, 40, 41];
394 thus, the frame and orientation requirements had not been tested experimentally. Among
395 the slow growth mutants that we identified, inserts were in the forward or reverse
396 orientation, and in and out of frame. Note that the adaptor sequence that was inserted
397 between the trigger and the insert does not change the frame. Inserts that were in the
398 reverse orientation and out of frame still led to knockdown. Both the directionality of the
399 gene of interest, and the frame, are therefore dispensable for knockdown. Since the RNAi

400 pathway in *E. histolytica* is unusual and mechanistic details are still coming into focus, our
401 results add to understanding of this pathway. The lack of orientation and frame
402 requirements for knockdown is useful, since gDNA fragments are random and thus are
403 ligated into the library plasmid in either orientation and in any frame. Previous studies had
404 defined that a gene fragment ≥ 500 bp is sufficient for knockdown [25, 40]. We found that
405 fragments as small as 322 bp were sufficient to result in slow growth phenotypes,
406 suggesting that efficient knockdown is possible with smaller fragments.
407 *E. histolytica* is dramatically understudied relative to other parasites, even though it
408 heavily impacts human health, causing significant morbidity and mortality. The RNAi library
409 that we developed dramatically improves the tractability of *E. histolytica.* In other parasites,
410 RNAi screens have uncovered genes relevant to molecular and cellular biology, virulence
411 and therapeutics. We look forward to the new information about *E. histolytica* that can be
412 gleaned by applying RNAi screens in the future. Since the design of this RNAi library allows
413 for quantitative identification of mutants that are enriched or depleted after selection, we
414 expect that this library can be used to identify genes involved in a wide variety of processes,
415 including essential genes that are relevant to the development of new therapies.

416 **Materials and Methods**

417 *Cell culture*

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

423 *gDNA Extraction*

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

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

443 *Plasmids and cloning*

444 The trigger plasmid [25] contains 132 base pairs of the trigger gene (EHI_048600) in the
445 pKT3M expression plasmid backbone [48]. Q5 Site Directed Mutagenesis (NEB) was used to
446 insert a SmaI restriction enzyme site after the trigger for cloning purposes, to create "pTrigger".
447 To create knockdown plasmids to target individual genes, pTrigger was linearized using SmaI
448 (NEB) and dephosphorylated using rSAP (NEB). A fragment of the gene of interest was amplified
449 from gDNA using PCR primers that contained homology to pTrigger, and the PCR product was
450 then cloned into the backbone using the Gibson Assembly Ultra kit (Synthetic Genomics Inc.),
451 following the manufacturer's instructions. See Table S4 for the sequences of primers that were
452 used during plasmid construction.

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

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

464 Superoxide dismutase knockdown mutants were generated by inserting 597 bp of
465 EHI_159160 (1 – 597) into pTrigger. Independent knockdown mutants (SG1i) of SG1 were
466 generated by inserting 525 bp of EHI_194310 (32 – 556) into pTrigger. Independent knockdown
467 mutants (SG2i) of SG2 were generated by inserting 487 bp of EHI_007010 (59 – 545) into
468 pTrigger.

469 To create the N terminal FLAG-tag constructs, Q5 Site Directed Mutagenesis (NEB) was
470 used to insert a 3X FLAG epitope tag sequence preceding an XhoI restriction site in the pEhEx
471 expression plasmid backbone [48], to create pEhEx-FLAG. For localization of SG1 (EHI_194310)
472 and SG2 (EHI_007010), each individual full-length gene was cloned into pEhEx-FLAG, to create
473 pEhEx-FLAG-SG1 and pEhEx-FLAG-SG2, respectively. For expression of truncated SG1, nt 1010-
474 1527 were omitted, and the resulting truncated gene was cloned into pEhEx-FLAG, with a stop
475 codon added at the 3' end of the truncated sequence.

476 *Plasmid library construction*

477 *E. histolytica* gDNA was harvested using the Wizard Genomic DNA Purification kit
478 (Promega), following the manufacturer's instructions for the tissue culture cells protocol.
479 RNAse treatment is included in this protocol. Quality and size of gDNA was verified by Qubit
480 and gel electrophoresis. gDNA was further purified and concentrated using TailorMag
481 Purification beads (SeqMatic), and then sheared using a Covaris M220 sonicator. For size

482 selection, sheared gDNA was loaded into three lanes on a 6% TBE gel. The gel was stained using
483 SYBR Safe DNA Gel Stain (Invitrogen), and three bands, covering from ~400 – ~600 bp (for the
484 pilot library) and ~500 – ~900 bp (for the final library) were extracted from the gel using the
485 TailorCut Gel Cutter and Gel Breaker system (SeqMatic). gDNA was purified using TailorMag
486 Purification beads. The sizing and concentration of the purified material was determined by
487 TapeStation (Agilent). Purified gDNA corresponding to the second and third gel slices from each
488 lane was combined. gDNA was further purified and concentrated using TailorMag Purification
489 beads, and then processed using a TruSeq XT library prep kit (Illumina). All steps followed the
490 manufacturer's protocol except: 1. Ligation time was increased to 180 minutes, 2. Custom
491 adaptors were used instead of Illumina TruSeq adaptors. The adaptors that were ligated had
492 the following sequences: Adaptor 1: ACACTCTTTCCCTACACGACGCTCTTCCGATCCCC*T and
493 Adaptor 2: /5'Phos/GGGGATCGGGAGAGCACACGTCTGAACTCTAGTCAC. After adaptor ligation,
494 the gDNA fragments were amplified using 5 rounds of PCR with KAPA HiFi 2X master mix (KAPA
495 Biosystems). See Table S4 for primer sequences. After PCR, fragmented gDNA was purified and
496 concentrated using TailorMag Purification beads.
497 pTriggerAdaptor was linearized using SmaI (NEB) and dephosphorylated using rSAP
498 (NEB). Linearization was verified using gel electrophoresis. 25 ng of fragmented gDNA with
499 ligated adaptors was combined with 25 ng of linearized pTriggerAdaptor. The Gibson Assembly
500 Ultra kit (Synthetic Genomics Inc.) was used according to the manufacturer's instructions. High
501 Efficiency 5-alpha Competent *E. coli* (NEB) were transformed with 3 μl of the Gibson mix and
502 plated on four LB plates containing ampicillin (Sigma). We determined that the Gibson cloning
503 efficiency with pTriggerAdaptor was ~1/250, which was used to calculate the amount of insert

504 needed to generate a library at a given target level of coverage. With an average insert size of
505 ~550 bp, a minimum of ~36,354 plasmids would be needed to fully cover the genome at 1X
506 coverage. For the pilot library, one *E. coli* transformation reaction was performed and 4 plates
507 of bacteria were harvested into 200 ml of LB liquid broth containing ampicillin, and incubated at
508 37°C for 12-18 hours, shaking. For the final library, Gibson cloning, *E. coli* transformation, and
509 plasmid preparations were performed over four separate days to create four batches of final
510 library plasmids. The final plasmid library represents the combined total of the four batches.
511 For the final library, 25 *E. coli* transformation reactions were performed and 100 plates of
512 bacteria were harvested into LB liquid broth containing ampicillin, at a concentration of ~1,000
513 colonies per 200 ml. Plasmids were harvested using the GenElute HP plasmid maxiprep kit
514 (Sigma), using 200 ml of culture per plasmid purification column, and ethanol precipitated
515 according to the manufacturer's instructions.

516 *Illumina sequencing of gDNA fragments and plasmid samples*

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

526 *Illumina sequencing analysis*

527 Illumina sequencing data was demultiplexed using barcode sequences added during
528 library preparation with Illumina demultiplexing software. All reads were sequenced as PE
529 151bp runs. The resulting paired reads were then cleaned using HTStream
530 (https://github.com/ibest/HTStream) to 1) remove reads containing PhiX Control Library
531 sequences, 2) trim Illumina sequencing adaptors, and 3) discard primer dimers and other reads
532 where adaptor trimming was unsuccessful.

533 Cleaned reads were mapped against the *E. histolytica* release 46 genome assembly
534 (downloaded from https://amoebadb.org/common/downloads/release-
535 46/EhistolyticaHM1IMSS/fasta/data/AmoebaDB-46_EhistolyticaHM1IMSS_Genome.fasta) using
536 bwa mem v2.0pre1 [49]. Output from bwa was passed to samtools v1.10 [50] for sorting and
537 indexing, and also to extract reads from for visualization of specific loci. Results from mapping
538 and subsequent analysis were visualized and reviewed in Geneious Prime 2020.0.3
539 (https://www.geneious.com). Identifying unique fragments and counting their frequency was
540 done using a custom Python script. Briefly, the script iterates over read-sorted BAM records.
541 Read pairs that were mapped to the same contig with proper orientation, a minimum of 146
542 matched bases, and no more than four bases soft clipped from the 5' end of either read pair
543 were used to identify fragments. Read mapping orientation was used to infer fragment
544 orientation. A counter for each unique fragment was incremented for additional reads
545 supporting the fragment. Reads that were mapped in proper orientation but that did not meet
546 other criteria were compared against the fragment intervals with an incremented count

547 assigned to the highest scoring overlap. As a last step, the script generates summary statistics,
548 an interval table with counts for all samples, and SAF and gff annotation files for each sample.
549 Output from the Python script was then loaded into R v3.6.1 [51] for further analysis
550 and to generate figures. Briefly, data.table [52] was used to load the full set of intervals, order
551 them by genomic position, and to compute summaries by sample group. Tidyr [53] was used for
552 reformatting data, and most plots were generated with ggplot2 [54]. The vegan package [55]
553 was used for calculating saturation curve values. The GenomicRanges package [56] was used in
554 combination with gene annotation (downloaded from
555 https://amoebadb.org/common/downloads/release-
556 46/EhistolyticaHM1IMSS/gff/data/AmoebaDB-46_EhistolyticaHM1IMSS.gff) to determine and
557 count fragments targeting genes, all reported counts used a minimum overlap of 27bp.
558 Genome coverage plots were generated with karyoploteR [57] and heatmaps were plotted
559 using pheatmap [58]. Scripts and more details about the analysis are available at the GitHub
560 repository associated with this publication: https://github.com/samhunter/Bettadapur-et-al-
561 Entamoeba-histolytica-RNAi-library.
562 All sequencing data from this project have been deposited in the Sequence Read Archive
563 (SRA) database (https://www.ncbi.nlm.nih.gov/sra), under project number SUB7459667.
564 *Transfection and clonal lines*
565 Library plasmids were transfected into amoebae using Attractene transfection reagent
566 (Qiagen) [59, 60]. 50 ml M199 (Gibco medium M199 with Earle's salts, L-glutamine, and 2.2
567 g/liter sodium bicarbonate, without phenol red), was supplemented with 0.045 g L-Cysteine,
568 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

569 this medium, 7.5 ml of heat inactivated Adult Bovine Serum (Gemini bio-products) was added
570 to make transfection medium. Amoebae were resuspended in transfection medium at a
571 concentration of $2.5x10^5$ cells/ml, and then 1.8 ml of amoebae, 10 μg plasmid DNA (in 200 μl of
572 transfection medium), and 37.5 μl Attractene reagent (Qiagen) were combined into a 2.0 ml
573 screw-cap cryotube. Cryotubes were incubated horizontally at 35°C for 3 hours. Two cryotubes
574 per transfection were then added to a T25 flask containing 66 ml of fresh TYI media. 24 hours
575 later, 6 μg/ml geneticin (Invitrogen) was added. Stable transfectants were selected and
576 maintained with geneticin at 6 μg/ml.

577 The efficiency of stable transfection was determined by performing limiting dilution.
578 Three independent experiments were performed. Amoebae were transfected using the process
579 described above, except that 24 hours after transfection, amoebae were harvested and
580 resuspended in TYI at a concentration of $2x10^6$ cells/ml. Amoebae were then serially diluted
581 into 96-well plates, in the presence or absence of 6 μg/ml geneticin. Plates were incubated in
582 GasPak EZ pouch (BD Biosciences) and growth was monitored daily. The plates without
583 Geneticin were used to calculate the plating efficiency. In the plates without geneticin, growth
584 was seen in wells corresponding to approximately 2.5 cells on average. In the plates containing
585 geneticin, growth was seen in wells containing 781 to 3125 amoebae. Taken together, the
586 corrected stable transfection efficiency was ~1/1000 amoebae.

587 For generation of clonal mutant lines for phenotypic characterization, limiting dilution
588 was performed in 96-well plates in a GasPak EZ pouch (BD Biosciences). For identification of
589 mutants with growth defects, clonal lines that visually appeared to have slow growth when
590 examined using microscopy were selected for further analysis. RNA was extracted using the

591 Direct-zol RNA kit (Zymo). gDNA (which contains exogeneous plasmids) was extracted using the
592 Quick-DNA miniprep kit (Zymo). Sanger sequencing was used to identify the library insert
593 associated with each clonal line. Sequencing was carried out by Genewiz. To identify the gene
594 targeted for knockdown in each clonal line, the insert sequence was used to search the
595 reference genome. Genes with at least 50nt of continuous sequence identity with the insert
596 were considered hits. See Table S4 for sequencing primer information.

597 *Growth analysis*

598 For growth analysis of RNAi knockdown mutants, clonal stable transfectants (SG1 –
599 SG12) or heterogeneous stable transfectants (SG1i and SG2i) were resuspended to $8x10^3$
600 cells/ml and 1 ml of cells was added to three tubes containing 13 ml fresh TYI with geneticin.
601 Cells were incubated at 35°C. At days 2, 4, and 6, one tube of cells was harvested per day and
602 counted using a TC20 Automatic cell counter (BioRad). Vector control transfected amoebae
603 were used as a negative control, and superoxide dismutase heterogeneous knockdown mutants
604 were used as a positive control for slow growth. Two technical replicates were performed per
605 experiment and three independent growth curve experiments were performed for each
606 mutant.

607 For growth analysis of cells overexpressing full-length or truncated SG proteins,
608 geneticin concentration was doubled daily until it reached 48 μg/ml, in order to increase
609 plasmid copy number and lead to protein overexpression. Geneticin concentration was also
610 increased in the same manner in vector control transfectants. Heterogeneous stable
611 transfectants were resuspended to $5x10^4$ cells/ml and 1 ml of cells was added to five tubes
612 containing 13 ml fresh TYI with 48 μg/ml geneticin. Cells were incubated at 35°C. At days 1, 2, 3,

613 4 and 5, one tube of cells was harvested per day and counted using a TC20 Automatic cell
614 counter (BioRad). Vector control transfected amoebae were used as a negative control. Four
615 technical replicates were performed per experiment and three independent growth curve
616 experiments were performed for each cell line. Growth curves were also performed at 24 μg/ml
617 geneticin, which led to the same growth phenotypes seen at higher concentrations of geneticin.
618 *RT-PCR*
619 RT-PCR was used to assess knockdown in transfectants. RNA was extracted using the
620 Direct-zol RNA kit (Zymo). DNAseI (Turbo) treatment and SuperScript Reverse Transcription II
621 (Invitrogen) was performed as per manufacturer instructions to obtain cDNA. RT-PCR was
622 performed on cDNA harvested from transfectants containing the following constructs: BAR/SH3
623 domain-containing protein with and without Illumina adaptors, Superoxide Dismutase;
624 SG1/SG1 independent knock down; and SG2/SG2 independent knockdown. GAPDH
625 (EHI_187020) RT-PCR was used as a positive control. See Table S4 for RT-PCR primer sequences.
626 *Immunofluorescence assays*
627 For immunofluorescence studies, geneticin concentration was doubled daily until it
628 reached 24 μg/ml, in order to increase plasmid copy number and lead to protein
629 overexpression. Amoebae were washed in M199s and fixed with 4% paraformaldehyde in PBS
630 for 30 minutes at room temperature. Samples were permeabilized with 0.2% Triton X-100, or
631 were not permeabilized. For permeabilized samples, for SG1, all subsequent blocking, antibody
632 incubation and wash steps included 0.1% Triton X-100. For permeabilized samples, for SG2, all
633 subsequent blocking, antibody incubation and wash steps included 0.1% Tween 20. For non-
634 permeabilized samples, all subsequent blocking, antibody incubation and wash steps included

635 0.1% Tween 20. Samples were blocked with goat serum (Jackson ImmunoResearch
636 Laboratories, Inc.) and bovine serum albumin (Gemini). Antibody labeling was carried out in
637 solution. For SG1, a rabbit anti-FLAG primary antibody (Abcam Rabbit Anti-DDDDK tag, binds to
638 FLAG tag sequence, antibody ab1162) was used, followed by an alpaca anti-rabbit Alexa 647
639 secondary antibody (Jackson ImmunoResearch Laboratories, Inc.). For SG2, a mouse anti-FLAG
640 primary (Sigma Mouse Anti-FLAG M2 monoclonal antibody) was used, followed by an alpaca
641 anti-mouse Cy5 antibody (Jackson ImmunoResearch Laboratories). Samples were stained with
642 DAPI at 40 μg/ml. Data were collected using an ImageStreamX Mark II flow cytometer,
643 equipped with a 40X objective. 10,000 events were collected for each sample and data were
644 analyzed using Amnis IDEAS software (Fig. S8).

645 *Statistical analysis*

646 GraphPad Prism was used for student's unpaired *t* test statistical analysis and
647 normalization (baseline removal). Mean values and standard deviations are shown in the
648 figures, with *t* test values reported as follows: ns = $p > 0.05$, * = $p < 0.05$, ** = $p < 0.01$, *** = $p <$
649 0.001, **** = $p < 0.0001$.

650 **Acknowledgements**

651 We thank Hannah Miller, Rene Suleiman, Scott Dawson and Su-Ju Lin for helpful
652 discussions. We thank Lutz Froenicke for helpful advice on designing the library to enable
653 Illumina sequencing. DNA sequencing was carried out by the DNA Technologies and
654 Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared
655 Instrumentation Grant 1S10OD010786-01. A.B. was supported by a fellowship from the UC
656 Davis Training Program in Biomolecular Technology. M.C.R. was supported by the NIH T32
657 Comparative Medical Science Training Program, T32OD011147. This work was funded by
658 NIH grants 1R01AI146914, 5R21AI154163 and a Pew Scholarship awarded to K.S.R.

659 **Competing Interests**

660 The authors declare that they have no competing interests.

661 **References**

662 1. WHO/PAHO/UNESCO report. A consultation with experts on amoebiasis. Mexico City,
663 Mexico 28-29 January, 1997. Epidemiol Bull. 1997;18(1):13-4. Epub 1997/03/01. PubMed PMID:
664 9197085.
665 2. Gilchrist CA, Petri SE, Schneider BN, Reichman DJ, Jiang N, Begum S, et al. Role of the
666 Gut Microbiota of Children in Diarrhea Due to the Protozoan Parasite Entamoeba histolytica. J
667 Infect Dis. 2016;213(10):1579-85. Epub 2015/12/30. doi: 10.1093/infdis/jiv772. PubMed PMID:
668 26712950; PubMed Central PMCID: PMCPMC4837909.
669 3. Mondal D, Petri WA, Jr., Sack RB, Kirkpatrick BD, Haque R. Entamoeba histolytica-
670 associated diarrheal illness is negatively associated with the growth of preschool children:
671 evidence from a prospective study. Trans R Soc Trop Med Hyg. 2006;100(11):1032-8. Epub
672 2006/05/30. doi: S0035-9203(06)00066-6 [pii]
673 10.1016/j.trstmh.2005.12.012 [doi]. PubMed PMID: 16730764.
674 4. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, et al. Burden
675 and aetiology of diarrhoeal disease in infants and young children in developing countries (the
676 Global Enteric Multicenter Study, GEMS): a prospective, case-control study. Lancet.
677 2013;382(9888):209-22. doi: 10.1016/S0140-6736(13)60844-2. PubMed PMID: 23680352.
678 5. Mortality GBD, Causes of Death C. Global, regional, and national age-sex specific all-
679 cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for
680 the Global Burden of Disease Study 2013. Lancet. 2015;385(9963):117-71. Epub 2014/12/23.
681 doi: 10.1016/S0140-6736(14)61682-2. PubMed PMID: 25530442; PubMed Central PMCID:
682 PMCPMC4340604.
683 6. Gonzales MLM, Dans LF, Sio-Aguilar J. Antiamoebic drugs for treating amoebic colitis.
684 Cochrane Database Syst Rev. 2019;1:CD006085. Epub 2019/01/10. doi:
685 10.1002/14651858.CD006085.pub3. PubMed PMID: 30624763; PubMed Central PMCID:
686 PMCPMC6326239.
687 7. Willhoeft U, Tannich E. The electrophoretic karyotype of Entamoeba histolytica. Mol
688 Biochem Parasitol. 1999;99(1):41-53. Epub 1999/04/24. doi: 10.1016/s0166-6851(98)00178-9.
689 PubMed PMID: 10215023.
690 8. Lorenzi HA, Puiu D, Miller JR, Brinkac LM, Amedeo P, Hall N, et al. New assembly,
691 reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features
692 and protein content information. PLoS Negl Trop Dis. 2010;4(6):e716. Epub 2010/06/19. doi:
693 10.1371/journal.pntd.0000716. PubMed PMID: 20559563; PubMed Central PMCID:
694 PMCPMC2886108.
695 9. Wilson IW, Weedall GD, Lorenzi H, Howcroft T, Hon CC, Deloger M, et al. Genetic
696 Diversity and Gene Family Expansions in Members of the Genus Entamoeba. Genome Biol Evol.
697 2019;11(3):688-705. Epub 2019/01/23. doi: 10.1093/gbe/evz009. PubMed PMID: 30668670;
698 PubMed Central PMCID: PMCPMC6414313.
699 10. Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, Amedeo P, et al. The genome
700 of the protist parasite Entamoeba histolytica. Nature. 2005;433(7028):865-8. Epub 2005/02/25.
701 doi: nature03291 [pii]
702 10.1038/nature03291 [doi]. PubMed PMID: 15729342.

703 11. King AV, Welter BH, Koushik AB, Gordon LN, Temesvari LA. A genome-wide over-
704 expression screen identifies genes involved in phagocytosis in the human protozoan parasite,
705 Entamoeba histolytica. PLoS One. 2012;7(8):e43025. doi: 10.1371/journal.pone.0043025.
706 PubMed PMID: 22905196; PubMed Central PMCID: PMC3419234.
707 12. Koushik AB, Welter BH, Rock ML, Temesvari LA. A genomewide overexpression screen
708 identifies genes involved in the phosphatidylinositol 3-kinase pathway in the human protozoan
709 parasite Entamoeba histolytica. Eukaryot Cell. 2014;13(3):401-11. doi: 10.1128/EC.00329-13.
710 PubMed PMID: 24442890; PubMed Central PMCID: PMC3957588.
711 13. Bracha R, Nuchamowitz Y, Leippe M, Mirelman D. Antisense inhibition of amoebapore
712 expression in Entamoeba histolytica causes a decrease in amoebic virulence. Mol Microbiol.
713 1999;34(3):463-72. Epub 1999/11/17. doi: mmi1607 [pii]. PubMed PMID: 10564488.
714 14. Bracha R, Nuchamowitz Y, Anbar M, Mirelman D. Transcriptional silencing of multiple
715 genes in trophozoites of Entamoeba histolytica. PLoS Pathog. 2006;2(5):e48. Epub 2006/05/31.
716 doi: 10.1371/journal.ppat.0020048 [doi]. PubMed PMID: 16733544.
717 15. Bracha R, Nuchamowitz Y, Mirelman D. Transcriptional silencing of an amoebapore gene
718 in Entamoeba histolytica: molecular analysis and effect on pathogenicity. Eukaryot Cell.
719 2003;2(2):295-305. Epub 2003/04/10. PubMed PMID: 12684379; PubMed Central PMCID:
720 PMC154849.
721 16. Stock RP, Olvera A, Sanchez R, Saralegui A, Scarfi S, Sanchez-Lopez R, et al. Inhibition of
722 gene expression in Entamoeba histolytica with antisense peptide nucleic acid oligomers. Nat
723 Biotechnol. 2001;19(3):231-4. Epub 2001/03/07. doi: 10.1038/85671. PubMed PMID:
724 11231555.
725 17. Kaur G, Lohia A. Inhibition of gene expression with double strand RNA interference in
726 Entamoeba histolytica. Biochemical and biophysical research communications.
727 2004;320(4):1118-22. Epub 2004/07/14. doi: 10.1016/j.bbrc.2004.06.064. PubMed PMID:
728 15249205.
729 18. Solis CF, Santi-Rocca J, Perdomo D, Weber C, Guillen N. Use of bacterially expressed
730 dsRNA to downregulate Entamoeba histolytica gene expression. PLoS One. 2009;4(12):e8424.
731 Epub 2009/12/29. doi: 10.1371/journal.pone.0008424. PubMed PMID: 20037645; PubMed
732 Central PMCID: PMC2793006.
733 19. Linford AS, Moreno H, Good KR, Zhang H, Singh U, Petri WA, Jr. Short hairpin RNA-
734 mediated knockdown of protein expression in Entamoeba histolytica. BMC Microbiol.
735 2009;9(1):38. Epub 2009/02/19. doi: 1471-2180-9-38 [pii]
736 10.1186/1471-2180-9-38 [doi]. PubMed PMID: 19222852.
737 20. Zhang H, Ehrenkaufer GM, Pompey JM, Hackney JA, Singh U. Small RNAs with 5'-
738 polyphosphate termini associate with a Piwi-related protein and regulate gene expression in
739 the single-celled eukaryote Entamoeba histolytica. PLoS Pathog. 2008;4(11):e1000219. Epub
740 2008/12/02. doi: 10.1371/journal.ppat.1000219. PubMed PMID: 19043551; PubMed Central
741 PMCID: PMC2582682.
742 21. Sijen T, Steiner FA, Thijssen KL, Plasterk RH. Secondary siRNAs result from unprimed
743 RNA synthesis and form a distinct class. Science. 2007;315(5809):244-7. Epub 2006/12/13. doi:
744 10.1126/science.1136699. PubMed PMID: 17158288.
745 22. Zhang H, Alramini H, Tran V, Singh U. Nucleus-localized antisense small RNAs with 5'-
746 polyphosphate termini regulate long term transcriptional gene silencing in Entamoeba

histolytica G3 strain. J Biol Chem. 2011;286(52):44467-79. Epub 2011/11/04. doi: M111.278184 [pii]
10.1074/jbc.M111.278184. PubMed PMID: 22049083; PubMed Central PMCID: PMC3247957.
23. Pompey JM, Foda B, Singh U. A Single RNaseIII Domain Protein from Entamoeba histolytica Has dsRNA Cleavage Activity and Can Help Mediate RNAi Gene Silencing in a Heterologous System. PLoS One. 2015;10(7):e0133740. doi: 10.1371/journal.pone.0133740. PubMed PMID: 26230096; PubMed Central PMCID: PMC4521922.
24. Zhang H, Tran V, Manna D, Ehrenkaufer G, Singh U. Functional Characterization of Entamoeba histolytica Argonaute Proteins Reveals a Repetitive DR-Rich Motif Region That Controls Nuclear Localization. mSphere. 2019;4(5). Epub 2019/10/18. doi: 10.1128/mSphere.00580-19. PubMed PMID: 31619501; PubMed Central PMCID: PMCPMC6796981.
25. Morf L, Pearson RJ, Wang AS, Singh U. Robust gene silencing mediated by antisense small RNAs in the pathogenic protist Entamoeba histolytica. Nucleic Acids Res. 2013. doi: 10.1093/nar/gkt717. PubMed PMID: 23935116.
26. Foda BM, Singh U. Dimethylated H3K27 Is a Repressive Epigenetic Histone Mark in the Protist Entamoeba histolytica and Is Significantly Enriched in Genes Silenced via the RNAi Pathway. J Biol Chem. 2015;290(34):21114-30. doi: 10.1074/jbc.M115.647263. PubMed PMID: 26149683; PubMed Central PMCID: PMC4543668.
27. Pompey JM, Morf L, Singh U. RNAi pathway genes are resistant to small RNA mediated gene silencing in the protozoan parasite Entamoeba histolytica. PLoS One. 2014;9(9):e106477. Epub 2014/09/10. doi: 10.1371/journal.pone.0106477. PubMed PMID: 25198343; PubMed Central PMCID: PMCPMC4157801.
28. Manna D, Lentz CS, Ehrenkaufer GM, Suresh S, Bhat A, Singh U. An NAD(+)-dependent novel transcription factor controls stage conversion in Entamoeba. eLife. 2018;7. Epub 2018/10/31. doi: 10.7554/eLife.37912. PubMed PMID: 30375973; PubMed Central PMCID: PMCPMC6207428.
29. Suresh S, Ehrenkaufer G, Zhang H, Singh U. Development of RNA Interference Trigger-Mediated Gene Silencing in Entamoeba invadens. Infect Immun. 2016;84(4):964-75. Epub 2016/01/21. doi: 10.1128/IAI.01161-15. PubMed PMID: 26787723; PubMed Central PMCID: PMCPMC4807475.
30. Morris JC, Wang Z, Drew ME, Englund PT. Glycolysis modulates trypanosome glycoprotein expression as revealed by an RNAi library. EMBO J. 2002;21(17):4429-38. Epub 2002/08/29. PubMed PMID: 12198145; PubMed Central PMCID: PMC125414.
31. Drew ME, Morris JC, Wang Z, Wells L, Sanchez M, Landfear SM, et al. The adenosine analog tubercidin inhibits glycolysis in Trypanosoma brucei as revealed by an RNA interference library. J Biol Chem. 2003;278(47):46596-600. Epub 2003/09/16. doi: 10.1074/jbc.M309320200. PubMed PMID: 12972414.
32. Stortz JA, Serafim TD, Alsford S, Wilkes J, Fernandez-Cortes F, Hamilton G, et al. Genome-wide and protein kinase-focused RNAi screens reveal conserved and novel damage response pathways in Trypanosoma brucei. PLoS Pathog. 2017;13(7):e1006477. Epub 2017/07/26. doi: 10.1371/journal.ppat.1006477. PubMed PMID: 28742144; PubMed Central PMCID: PMCPMC5542689.

790 33. Glover L, Hutchinson S, Alsford S, Horn D. VEX1 controls the allelic exclusion required for
791 antigenic variation in trypanosomes. Proc Natl Acad Sci U S A. 2016;113(26):7225-30. Epub
792 2016/05/27. doi: 10.1073/pnas.1600344113. PubMed PMID: 27226299; PubMed Central
793 PMCID: PMCPMC4932947.
794 34. Mony BM, MacGregor P, Ivens A, Rojas F, Cowton A, Young J, et al. Genome-wide
795 dissection of the quorum sensing signalling pathway in Trypanosoma brucei. Nature.
796 2014;505(7485):681-5. Epub 2013/12/18. doi: 10.1038/nature12864. PubMed PMID:
797 24336212; PubMed Central PMCID: PMCPMC3908871.
798 35. Rico E, Ivens A, Glover L, Horn D, Matthews KR. Genome-wide RNAi selection identifies a
799 regulator of transmission stage-enriched gene families and cell-type differentiation in
800 Trypanosoma brucei. PLoS Pathog. 2017;13(3):e1006279. Epub 2017/03/24. doi:
801 10.1371/journal.ppat.1006279. PubMed PMID: 28334017; PubMed Central PMCID:
802 PMCPMC5380359.
803 36. Alsford S, Eckert S, Baker N, Glover L, Sanchez-Flores A, Leung KF, et al. High-throughput
804 decoding of antitrypanosomal drug efficacy and resistance. Nature. 2012;482(7384):232-6.
805 Epub 2012/01/27. doi: 10.1038/nature10771. PubMed PMID: 22278056; PubMed Central
806 PMCID: PMCPMC3303116.
807 37. Collett CF, Kitson C, Baker N, Steele-Stallard HB, Santrot MV, Hutchinson S, et al.
808 Chemogenomic Profiling of Antileishmanial Efficacy and Resistance in the Related Kinetoplastid
809 Parasite Trypanosoma brucei. Antimicrob Agents Chemother. 2019;63(8). Epub 2019/06/05.
810 doi: 10.1128/AAC.00795-19. PubMed PMID: 31160283; PubMed Central PMCID:
811 PMCPMC6658743.
812 38. Trenaman A, Glover L, Hutchinson S, Horn D. A post-transcriptional respiratome regulon
813 in trypanosomes. Nucleic Acids Res. 2019;47(13):7063-77. Epub 2019/05/28. doi:
814 10.1093/nar/gkz455. PubMed PMID: 31127277; PubMed Central PMCID: PMCPMC6648352.
815 39. Zhang N, Zoltner M, Leung KF, Scullion P, Hutchinson S, Del Pino RC, et al. Host-parasite
816 co-metabolic activation of antitrypanosomal aminomethyl-benzoxaboroles. PLoS Pathog.
817 2018;14(2):e1006850. Epub 2018/02/10. doi: 10.1371/journal.ppat.1006850. PubMed PMID:
818 29425238; PubMed Central PMCID: PMCPMC5823473.
819 40. Khalil MI, Foda BM, Suresh S, Singh U. Technical advances in trigger-induced RNA
820 interference gene silencing in the parasite Entamoeba histolytica. Int J Parasitol.
821 2016;46(3):205-12. doi: 10.1016/j.ijpara.2015.11.004. PubMed PMID: 26747561; PubMed
822 Central PMCID: PMCPMC4767557.
823 41. Rastew E, Morf L, Singh U. Entamoeba histolytica rhomboid protease 1 has a role in
824 migration and motility as validated by two independent genetic approaches. Exp Parasitol.
825 2015. doi: 10.1016/j.exppara.2015.04.004. PubMed PMID: 25889553.
826 42. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc
827 database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome
828 databases. Nucleic Acids Res. 2016;44(D1):D471-80. Epub 2015/11/04. doi:
829 10.1093/nar/gkv1164. PubMed PMID: 26527732; PubMed Central PMCID: PMCPMC4702838.
830 43. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference
831 database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic
832 Acids Res. 2013;41(Database issue):D456-63. Epub 2012/11/28. doi: 10.1093/nar/gks1146.
833 PubMed PMID: 23180789; PubMed Central PMCID: PMCPMC3531142.

834 44. Shabardina V, Kischka T, Kmita H, Suzuki Y, Makalowski W. Environmental adaptation of
835 Acanthamoeba castellanii and Entamoeba histolytica at genome level as seen by comparative
836 genomic analysis. Int J Biol Sci. 2018;14(3):306-20. Epub 2018/03/22. doi: 10.7150/ijbs.23869.
837 PubMed PMID: 29559848; PubMed Central PMCID: PMCPMC5859476.
838 45. Hon CC, Weber C, Sismeiro O, Proux C, Koutero M, Deloger M, et al. Quantification of
839 stochastic noise of splicing and polyadenylation in Entamoeba histolytica. Nucleic Acids Res.
840 2013;41(3):1936-52. Epub 2012/12/22. doi: 10.1093/nar/gks1271. PubMed PMID: 23258700;
841 PubMed Central PMCID: PMCPMC3561952.
842 46. Suleiman RL, Ralston KS. Growth and genetic manipulation of Entamoeba histolytica.
843 Curr Protoc. 2021;Submitted.
844 47. Mueller DE, Petri WA, Jr. Clonal growth in Petri dishes of Entamoeba histolytica. Trans R
845 Soc Trop Med Hyg. 1995;89(1):123. Epub 1995/01/01. doi: 10.1016/0035-9203(95)90684-3.
846 PubMed PMID: 7747296.
847 48. Saito-Nakano Y, Yasuda T, Nakada-Tsukui K, Leippe M, Nozaki T. Rab5-associated
848 vacuoles play a unique role in phagocytosis of the enteric protozoan parasite Entamoeba
849 histolytica. J Biol Chem. 2004;279(47):49497-507. Epub 2004/09/07. doi:
850 10.1074/jbc.M403987200 [doi]
851 M403987200 [pii]. PubMed PMID: 15347665.
852 49. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
853 arXiv e-prints. 2013.
854 50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
855 Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. Epub 2009/06/10.
856 doi: 10.1093/bioinformatics/btp352. PubMed PMID: 19505943; PubMed Central PMCID:
857 PMCPMC2723002.
858 51. Team RC. R: A language and environment for statistical computing. R Foundation for
859 statistical computing. 2019.
860 52. Dowle M, Srinivasan A. data.table: Extension of `data.frame`. R package version 1.12.8.
861 2019.
862 53. Wickham H, Henry L. tidyr: Tidy Messy Data. R package version 1.0.2. 2020.
863 54. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.
864 2016.
865 55. Okansen J, Guillaume Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, et al.
866 vegan: Community Ecology Package. R package version 2.5-6. 2019.
867 56. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al. Software for
868 computing and annotating genomic ranges. PLoS Comput Biol. 2013;9(8):e1003118. Epub
869 2013/08/21. doi: 10.1371/journal.pcbi.1003118. PubMed PMID: 23950696; PubMed Central
870 PMCID: PMCPMC3738458.
871 57. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes
872 displaying arbitrary data. Bioinformatics. 2017;33(19):3088-90. Epub 2017/06/03. doi:
873 10.1093/bioinformatics/btx346. PubMed PMID: 28575171; PubMed Central PMCID:
874 PMCPMC5870550.
875 58. Kolde R. pheatmap: Pretty Heatmaps. R package version 1.0.12. 2019.
876 59. Ralston KS, Solga MD, Mackey-Lawrence NM, Somlata, Bhattacharya A, Petri WA, Jr.
877 Trogocytosis by Entamoeba histolytica contributes to cell killing and tissue invasion. Nature.

878 2014;508(7497):526-30. doi: 10.1038/nature13242. PubMed PMID: 24717428; PubMed Central
879 PMCID: PMC4006097.
880 60. Miller HW, Suleiman RL, Ralston KS. Trogocytosis by Entamoeba histolytica Mediates
881 Acquisition and Display of Human Cell Membrane Proteins and Evasion of Lysis by Human
882 Serum. MBio. 2019;10(2). doi: 10.1128/mBio.00068-19. PubMed PMID: 31040235; PubMed
883 Central PMCID: PMCPMC6495370.
884

885 **Figure Legends**

886 **Figure 1. RNAi library strategy.**

887 **a,** The plasmid RNAi library contains random fragments of gDNA cloned into the trigger
888 RNAi plasmid (pTriggerAdaptor). Different plasmid inserts are represented by different
889 colors in this schematic. A heterogeneous population of stably transfected RNAi knockdown
890 mutants can be exposed to selection or screening. To identify mutants that are represented
891 in the population after selection/screening, gDNA is extracted for analysis. gDNA extracted
892 from non-selected amoebae is used as a control. **b,** Schematic representation of the plasmid
893 RNAi library. The trigger plasmid is shown, with different inserts represented by different
894 colors. The "trigger" is a 132 bp fragment of EHI_048600. pTrigger was modified to create
895 pTriggerAdaptor, containing custom Illumina TruSeq adaptor sequences (IA) upstream and
896 downstream of the cloning site to facilitate amplicon sequencing. A single round of PCR is
897 needed to add the remaining Illumina sequences (shown in dark gray) and barcodes
898 necessary for Illumina deep sequencing analysis. **c,** An 812 bp fragment of a BAR/SH3
899 domain-containing protein (EHI_091530) was inserted into pTrigger with or without custom
900 Illumina TruSeq adaptors flanking the cloning site. The BAR/SH3 domain-containing protein
901 was in the forward orientation and in frame relative to the trigger sequence. Stably
902 transfected heterogeneous mutants were assessed for knockdown of the targeted gene
903 using RT-PCR, with (+) or without (-) reverse transcriptase. GAPDH (EHI_187020) was used
904 as a loading control.

905 **Fig. 2. RNAi library construction.**

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

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

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

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

929 pilot and final plasmid libraries. Size in nucleotides is plotted on the X axis, and probability
930 distribution density is plotted on the Y axis. **c,** Genome coverage plots of the final library,
931 over the five largest contigs in the reference genome. The positions of unique fragments
932 and the depth in fragments overlapping each genomic position are shown. The positions of
933 genes are shown with black and green bars. Fragments that are in the same orientation as
934 genes that are in the forward orientation in the reference genome are shown in black, and
935 fragments that are in the same orientation as genes that are in the reverse orientation in
936 the reference genome are shown in green. **d,** Visual representation of fragment coverage in
937 the final library, over 100 kb of contig DS571145.

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

939 Heat map showing the number of unique fragments per gene in the pilot and final gDNA
940 fragments, and the pilot and final plasmid libraries. The number of unique fragments (Log_2) for
941 each gene are plotted (color value, see legend). Stranded fragments are in the same orientation
942 as genes. Unstranded fragments are the total number of fragments, both in the same
943 orientation as genes and in the opposite orientation as genes.

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

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

951 vector control transfectants. **b,** RT-PCR analysis demonstrates that the genes corresponding
952 to the plasmid inserts in SG1 and SG2 (EHI_194310 and EHI_007010, respectively), were
953 knocked down. **c,** Knockdown mutants were independently generated by using fragments of
954 these genes that differed from the fragments in the library plasmids. These independent
955 mutants are referred to as SG1i and SG2i. Growth analysis was performed using
956 heterogeneous stable transfectants, and growth mutants exhibited a significant growth
957 defect when compared to vector control. **d,** RT-PCR analysis demonstrates that EHI_194310
958 and EHI_007010 were knocked down. **e,** Growth phenotypes in SG1, SG2, SG1i and SG2i.
959 Growth values for each mutant were normalized to vector control transfectants. Mutants
960 exhibited statistically significant growth defects relative to vector control transfectants.
961 **Figure 6. SG1 localizes to the cytoplasm and cell surface, and overexpression of SG1**
962 **enhances growth.**
963 Amoebae were stably transfected with a plasmid for expression of full-length FLAG-tagged
964 SG1. Immunofluorescence was used to determine the localization of SG1 in heterogeneous
965 transfectants and imaging flow cytometry was used for analysis. **a,** Localization of SG1 in
966 FLAG-SG1 transfectants (left panel) and control staining of wild-type cells (right panel).
967 Samples were permeabilized prior to antibody staining. Six random cells are shown for each
968 condition. Shown from left to right are bright field images (BF), rabbit anti-FLAG antibody
969 staining (FLAG, red), and DAPI staining (DAPI, blue). The numbers in the BF images indicate
970 the object/image number. **b,** Localization of SG1 in FLAG-SG1 transfectants (left panel) and
971 control staining of wild-type cells (right panel), as in panel A, except samples were not
972 permeabilized. **c,** Histograms showing the intensity of the DAPI staining in images collected

973 from FLAG-SG1 transfectants (left panel) or wild-type cells (right panel). The mean
974 fluorescence intensity (MFI) is indicated on each histogram. **d,** Growth analysis of
975 heterogeneous transfectants expressing full-length SG1 or a C-terminal SG1 truncation
976 mutant. Amoebae transfected with pEhEx (vector control) were used as control. Expression
977 of full-length SG1 was associated with an increase in growth compared to vector control
978 transfectants, that was statistically significant at days 3 ($p < .01$), 4 ($p < .01$), and 5 ($p < .05$),
979 as indicated. The growth of transfectants expressing truncated SG1 was associated with a
980 statistically significant decrease in growth compared to transfectants expressing full-length
981 SG1 at days 2 ($p < .05$), 3 ($p < .05$), 4 ($p < .01$), and 5 ($p < .01$). There was no significant
982 difference between the growth of transfectants expressing truncated SG1 and vector
983 control transfectants at any time point.

984 **Figure 7. SG2 localizes to the cytoplasm.**

985 Amoebae were stably transfected with a plasmid for expression of full-length FLAG-tagged
986 SG2. Immunofluorescence was used to determine the localization of SG2 in heterogeneous
987 transfectants and imaging flow cytometry was used for analysis. **a,** Localization of SG2 in
988 FLAG-SG2 transfectants (left panel) and control staining of wild-type cells (right panel).
989 Samples were permeabilized prior to antibody staining. Six random cells are shown for each
990 condition. Shown from left to right are bright field images (BF), mouse anti-FLAG antibody
991 staining (FLAG, red), and DAPI staining (DAPI, blue). The numbers in the BF images indicate
992 the object/image number. **b,** Histograms showing the intensity of the DAPI staining in
993 images collected from FLAG-SG2 transfectants (left panel) or wild-type cells (right panel).
994 The mean fluorescence intensity (MFI) is indicated on each histogram. **c,** Colocalization

995 analysis of FLAG and DAPI staining. The percentage of DAPI+/FLAG+ images with

996 colocalization of the two markers are indicated.

997 **Table Legends**

998 **Table 1. Library coverage analysis.**

999 Analysis of coverage in the pilot and final gDNA fragments, and the pilot and final plasmid
1000 libraries. Genomic coverage (reported as bp covered, forward, reverse, and unstranded)
1001 was calculated using fragments of the appropriate orientation, inferred from mapped reads
1002 and the *coverage* function in the GenomicRanges R/Bioconductor package. The total
1003 number of covered bp are indicated, and the corresponding percentage of the HM1:IMSS
1004 reference genome that this represents. Forward fragments are in the same orientation as
1005 genes that are in the forward orientation in the reference genome. Reverse fragments are
1006 in the same orientation as genes that are in the reverse orientation in the reference
1007 genome. Stranded fragments are in the same orientation as genes, and are the combined
1008 total of both forward and reverse fragments. Unstranded fragments are the total number of
1009 fragments, both in the same orientation as genes in the reference genome and in the
1010 opposite orientation as genes. Genes that are targeted are those with at least one unique
1011 fragment with greater than or equal to 27nt that overlaps the gene. Genes that are missed
1012 are those that do not fit the criteria for targeted genes. Genes that are targeted by multiple
1013 fragments are those with two or more unique fragments. Counts were calculated using the
1014 countOverlaps function in the GenomicRanges package and annotation from AmoebaDB.
1015 The number of unique fragments is the total number of unique fragments in each sample.

1016 **Table 2. Identified slow growth mutants.**

1017 Summary of growth phenotypes of SG1 – SG12 slow growth mutants, and analysis of the
1018 sequences present in each corresponding library plasmid. Growth values represent cell

1019 numbers normalized to vector control transfectants on day 4 of growth assays. The
1020 corresponding raw data growth assay data are found in Fig. 5a and Fig. S5a, c, e). Mutants
1021 exhibited statistically significant growth defects relative to vector control transfectants; *p*-
1022 values from day 4 are shown. The insert sizes are shown in bp. Insert directionality relative
1023 to the corresponding coding sequence in the annotated genome is shown. Insert frame
1024 relative to the trigger sequence in the pTriggerAdaptor is shown. The accession numbers for
1025 coding sequences that inserts correspond to are indicated, with available annotation
1026 information and GO terms for each gene (NA, not applicable). Predicted roles in metabolic
1027 pathways are shown; genes that are predicted to function in more than one metabolic
1028 pathway are indicated as "several." These predictions are based on data from KEGG
1029 Metabolic Pathways (https://www.kegg.jp), the MetaCyc Metabolic Pathway Database [42],
1030 and Chemical Entities of Biological Interest (ChEBI) database [43].

1031 **Supplementary Figure Legends**

1032 **Figure S1. Preparation of RNA-free gDNA from *E. histolytica.***

1033 Gel electrophoresis analysis of gDNA extracted using the Wizard gDNA extraction kit (Promega),
1034 Quick-DNA miniprep kit (Zymo), and CTAB extraction. **a,** gDNA isolated using the Wizard kit had
1035 no apparent RNA contamination, and the gDNA appeared to be of high quality. **b,** gDNA
1036 isolated using the Quick-DNA kit or CTAB was contaminated with RNA. RNA contamination was
1037 removed by treatment with RNAseIF and subsequent purification. **c,** Bioanalyzer traces of pilot
1038 and final gDNA samples, after fragmentation and size selection.

1039 **Figure S2. Fragment Mapping Analysis.**

1040 Example view of mapped genomic fragments ("intervals") represented in the final plasmid
1041 library. For comparison, the mapped paired end reads that give rise to each fragment are
1042 shown below each fragment. Shown are results from final plasmid batch two, sequencing
1043 replicate one, mapped to contig DS571153.

1044 **Figure S3. Appropriate sequencing depth and lack of bias during library construction.**

1045 **a,** Dinucleotide frequency analysis of the HM1:IMSS reference genome, the pilot and final gDNA
1046 fragments, and the pilot and final plasmid libraries. Stacked bar plots show the percentage of
1047 each dinucleotide. The dinucleotide composition of the gDNA fragments and the final plasmid
1048 library is the same as the genome. **b – c,** Rarefaction curves for pilot (panel b) and final (panel c)
1049 samples. gDNA fragment samples are shown in grey and plasmid samples are shown in orange.
1050 The number of fragments detected with increasing sequence reads shows if sequencing was
1051 performed to saturation. Three sequencing replicates are shown for pilot gDNA samples and
1052 three sequencing replicates are shown for pilot plasmid samples. Three sequencing replicates

1053 are shown for final gDNA samples and three sequencing replicates are shown for each of four
1054 batches of final plasmid samples.

1055 **Figure S4. Reproducible fragment and read counts from biological and technical replicates.**
1056 "Batches" of the final plasmid library were cloned on different days, and represent biological
1057 replicates. "Reps" correspond to independently prepared sequencing preparations, and
1058 represent technical replicates. **a,** Box plot of read counts per fragment in plasmid samples, with
1059 pilot plasmids in orange and final plasmids in grey. Log_2 read counts are shown; outliers
1060 (outside of 2nd and 98th percentile) are plotted as points, and the box represents the 25th, and
1061 75th percentile with median indicated as a bold line. Most fragments were represented by
1062 multiple sequenced reads. **b,** Stacked bar plots of the total number of fragments detected per
1063 sample shows that replicates are similar to each other in the number of fragments identified
1064 and that roughly equal numbers of forward and reverse fragments were produced consistently.
1065 Forward fragments are in the same orientation as genes that are in the forward orientation in
1066 the reference genome. Reverse fragments are in the same orientation as genes that are in the
1067 reverse orientation in the reference genome.

1068 **Figure S5. Annotated genes that were missed or were least covered by sequencing reads.**
1069 Heat map showing the number of unique fragments per gene, for the 50 least covered genes in
1070 the pilot and final gDNA fragments, and the pilot and final plasmid libraries. The color intensity
1071 indicates Log_2 fragment coverage, with dark blue indicating no coverage. Stranded fragments
1072 are in the same orientation as genes. Unstranded fragments are the total number of fragments,
1073 both in the same orientation as genes and in the opposite orientation as genes.

1074 **Figure S6. Clonal transfectants contain only one library plasmid.**

1075 SG1 gDNA was extracted using the Quick-DNA miniprep kit and PCR amplified using the CS5'
1076 and CS3' primers that sit in the pTriggerAdaptor backbone. Sanger sequencing was performed
1077 using the RNAi Sequencing Library F primer. Adaptor 2 starts at the "GGGG" series of
1078 nucleotides, and upstream of this is the SG1 (EHI_194310) sequence. Noncompeting peaks in
1079 the chromatogram suggest that there is one unique genome fragment present in this mutant.

1080 **Figure S7. Identification of slow growing RNAi knockdown mutants.**

1081 Growth analysis of mutants SG1 – SG12. **a, c, e,** Growth analysis demonstrates that selected
1082 clonal knockdown mutant lines have significant growth defects compared to vector control
1083 transfectants. **b, d, f**, Growth was normalized and compared to vector control transfectants
1084 on each day. Mutants exhibited statistically significant growth defects relative to vector
1085 control transfectants.

1086 **Figure S8. Gating scheme for imaging flow cytometry analysis.**

1087 Shown is the gating strategy that was used to analyze imaging flow cytometry data, with the
1088 percentage of gated events, and number of gated events, shown below each plot. This
1089 example illustrates the gating of the wild-type sample shown in Figure 7. 10,000 events
1090 were collected. **a,** In-focus events were gated using a gradient of brightfield (BF). **b,** Events
1091 gated in panel a were refined to remove images containing more than one amoeba, by
1092 examining the area and the aspect ratio of the object(s) masked in the brightfield image. **c,**
1093 Events gated in panel b were refined to remove images that contained saturation in the
1094 FLAG channel (channel 11), by examining the raw max pixel intensity (i.e., the intensity of
1095 the brightest pixel in the image). **d,** Events gated in panel c as "unsaturated" were further
1096 examined to identify images that contained both FLAG and DAPI signal, by examining the

1097 intensity of the object masked in the DAPI and FLAG channels. **e,** Events gated in panel d as

1098 "+DAPI/FLAG" were further examined to characterize the degree of colocalization (bright

1099 detail similarity) between the DAPI and FLAG signals.

1100 **Supplementary Table Legends**

1101 **Table S1. Sequencing and fragment count statistics.**

1102 Total number of paired reads mapping to the same contig (mappedPairs) and the number of
1103 fragments inferred from these mapped pairs for each sample. Reads were mapped with bwa
1104 mem, and processed from the resulting SAM formatted file using a custom Python script.
1105 Three sequencing replicates were performed for pilot gDNA samples and three sequencing
1106 replicates were performed for pilot plasmid samples. Three sequencing replicates were
1107 performed for final gDNA samples and three sequencing replicates were performed for each of
1108 four batches of final plasmid samples.

1109 **Table S2. Expression data for annotated genes that were missing in the final plasmid library.**

1110 Available empirical expression data for genes that were missing in the final plasmid library
1111 (genes that are represented in blue in Fig. S3). Accession numbers, annotation information, and
1112 the number of paralogs in the reference genome are indicated. Presence (+) or absence (-) in
1113 gDNA and plasmid datasets is shown. Gene expression from available RNAseq datasets [44, 45]
1114 is indicated, as the FPKM + 1 value. The highest value for FPKM + 1 for each gene is shown. The
1115 total number of spectra that have been identified in proteomics studies is indicated.

1116 **Table S3. Expression data for representative genes present in the final plasmid library.**

1117 To provide a frame of reference for the expression data in Table S2, shown are the available
1118 empirical expression data for 10 random, representative genes that were present in the final
1119 plasmid library. Accession numbers, annotation information, and the number of paralogs in the
1120 reference genome are indicated. Presence (+) or absence (-) in gDNA and plasmid datasets is
1121 shown. Gene expression from available RNAseq datasets [44, 45] is indicated, as the FPKM + 1

1122 value. The highest value for FPKM + 1 for each gene is shown. The total number of spectra that

1123 have been identified in proteomics studies is indicated.

1124 **Table S4. Primer sequences.**

1125 Sequences of the primers used to generate and sequence plasmids, perform RT-PCR, and

1126 create the plasmid RNAi library.

Figure 1

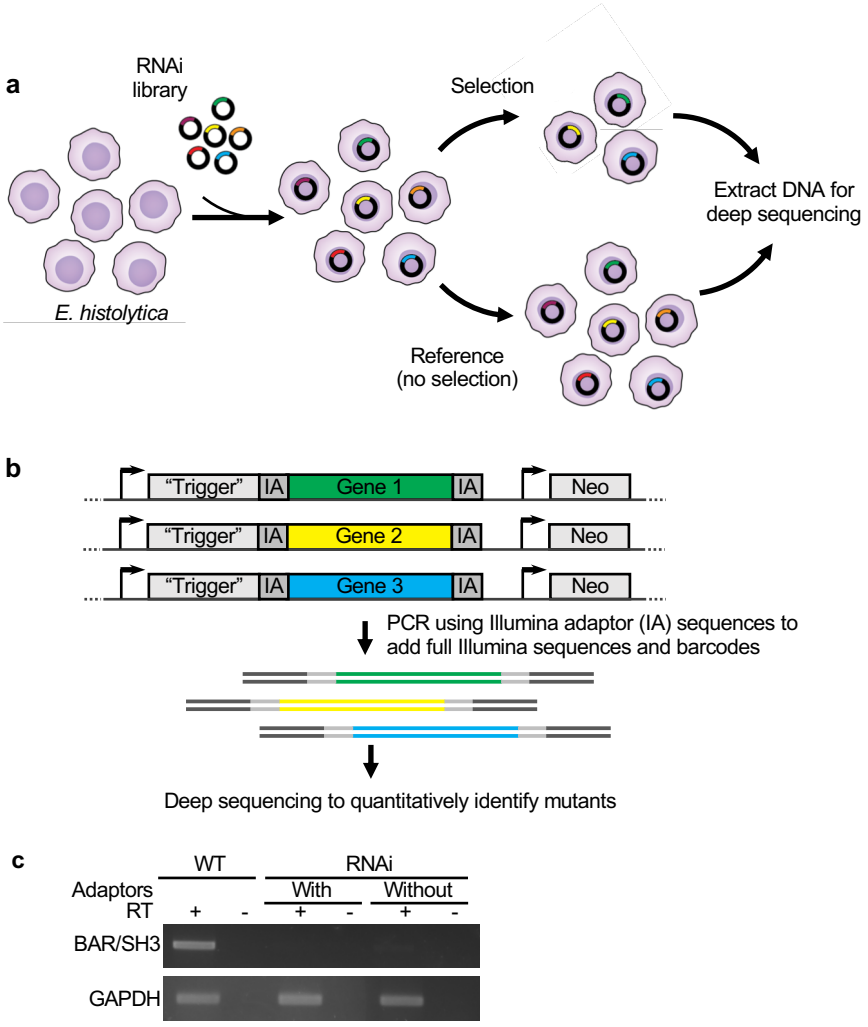

a
RNAi
library
E. histolytica
Selection
Reference
(no selection)
Extract DNA for
deep sequencing
b
"Trigger"
IA
Gene 1
Gene 2
Gene 3
Neo
PCR using Illumina adaptor (IA) sequences to
add full Illumina sequences and barcodes
Deep sequencing to quantitatively identify mutants
c
WT
RNAi
Adaptors
With
Without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

Figure 2

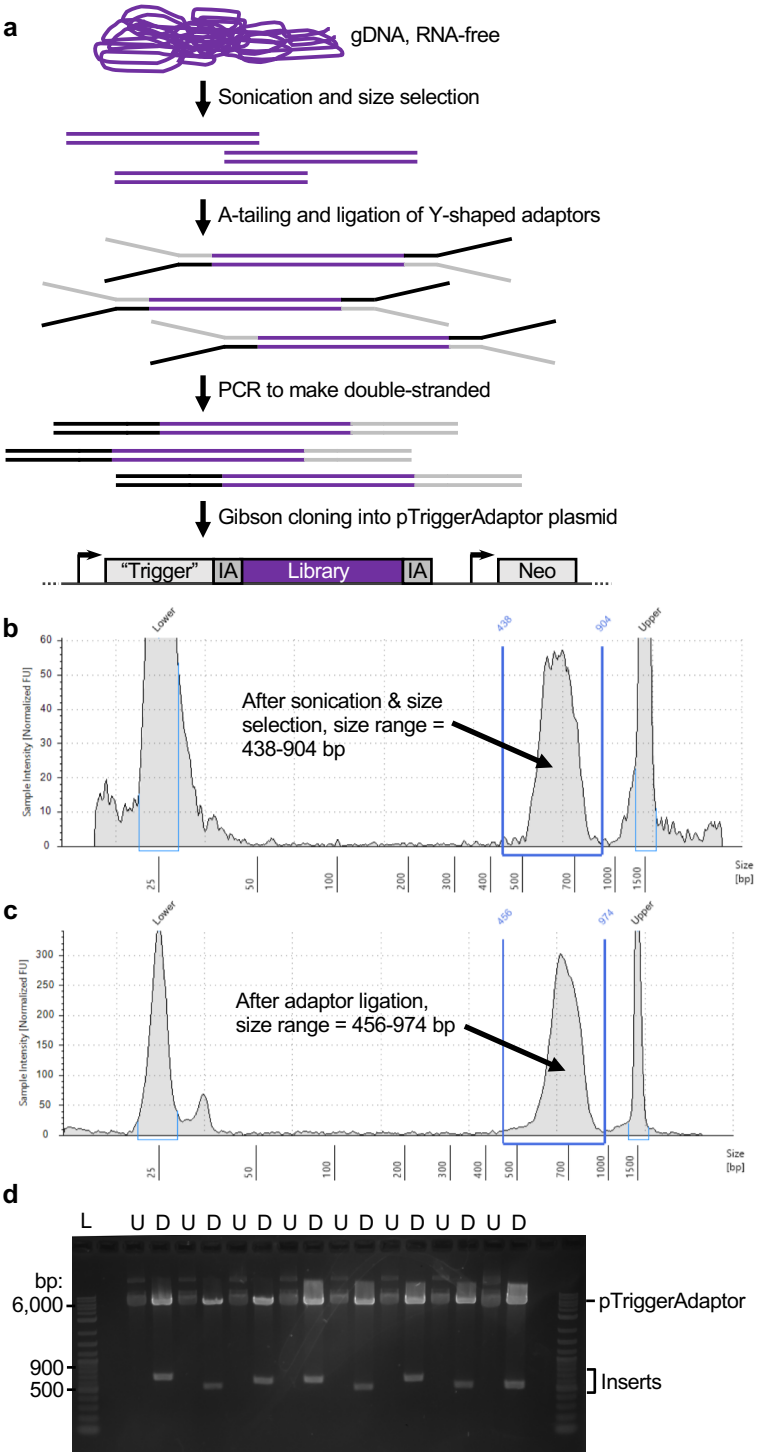

a
gDNA, RNA-free
Sonication and size selection
A-tailing and ligation of Y-shaped adaptors
PCR to make double-stranded
Gibson cloning into pTriggerAdaptor plasmid
"Trigger"
IA
Library
IA
Neo
b
Lower
438
904
Upper
After sonication & size selection, size range = 438-904 bp
Sample Intensity [Normalized FU]
Size [bp]
c
Lower
456
974
Upper
After adaptor ligation, size range = 456-974 bp
Sample Intensity [Normalized FU]
Size [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

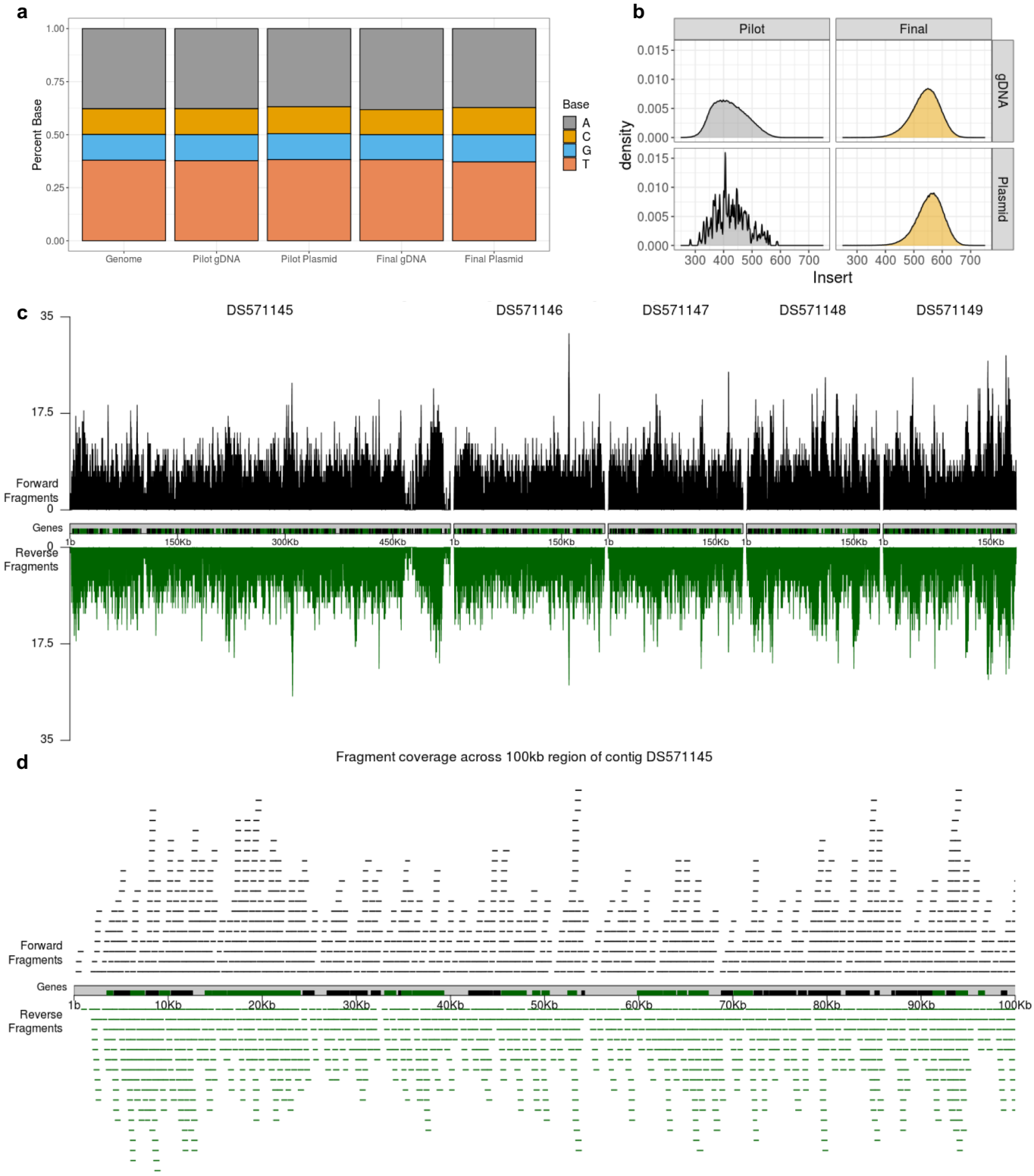

a
Percent Base
1.00
0.75
0.50
0.25
0.00
Genome
Pilot gDNA
Pilot Plasmid
Final gDNA
Final Plasmid
Base
A
C
G
T
b
Pilot
Final
gDNA
Plasmid
density
0.015
0.010
0.005
0.000
300 400 500 600 700
Insert
c
DS571145
DS571146
DS571147
DS571148
DS571149
35
17.5
0
Forward Fragments
Genes
Reverse Fragments
1b
150Kb
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

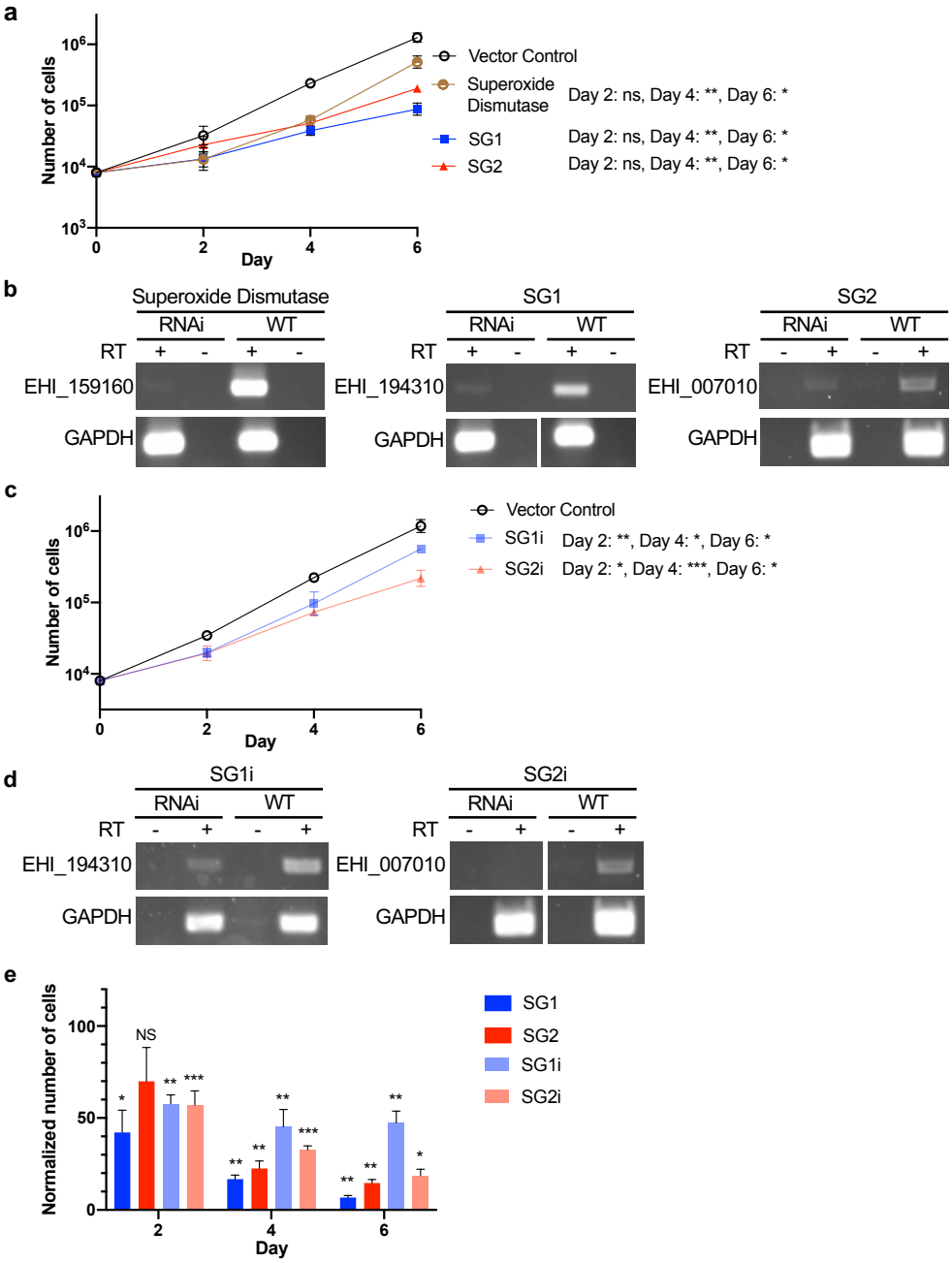

a
Number of cells
Day
Vector Control
Superoxide Dismutase
Day 2: ns, Day 4: **, Day 6: *
SG1
Day 2: ns, Day 4: **, Day 6: *
SG2
Day 2: ns, Day 4: **, Day 6: *
b
Superoxide Dismutase
SG1
SG2
RNAi
WT
RT
EHI_159160
EHI_194310
EHI_007010
GAPDH
c
Number of cells
Day
Vector Control
SG1i
Day 2: **, Day 4: *,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

Figure 6

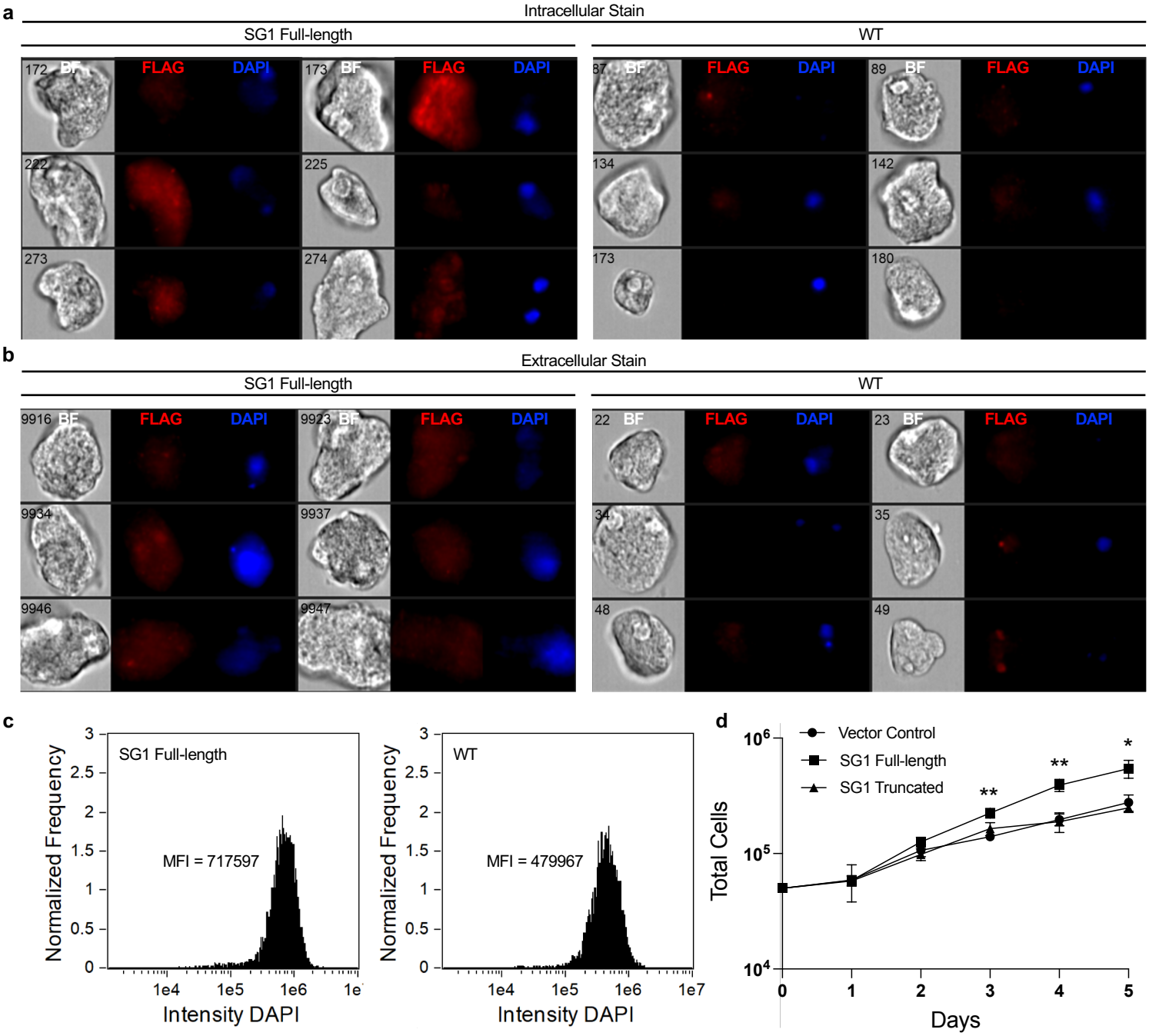

a
Intracellular Stain
SG1 Full-length
WT
BF
FLAG
DAPI
b
Extracellular Stain
SG1 Full-length
WT
c
SG1 Full-length
MFI = 717597
WT
MFI = 479967
Normalized Frequency
Intensity DAPI
d
Vector Control
SG1 Full-length
SG1 Truncated
Total Cells
Days

Figure 7

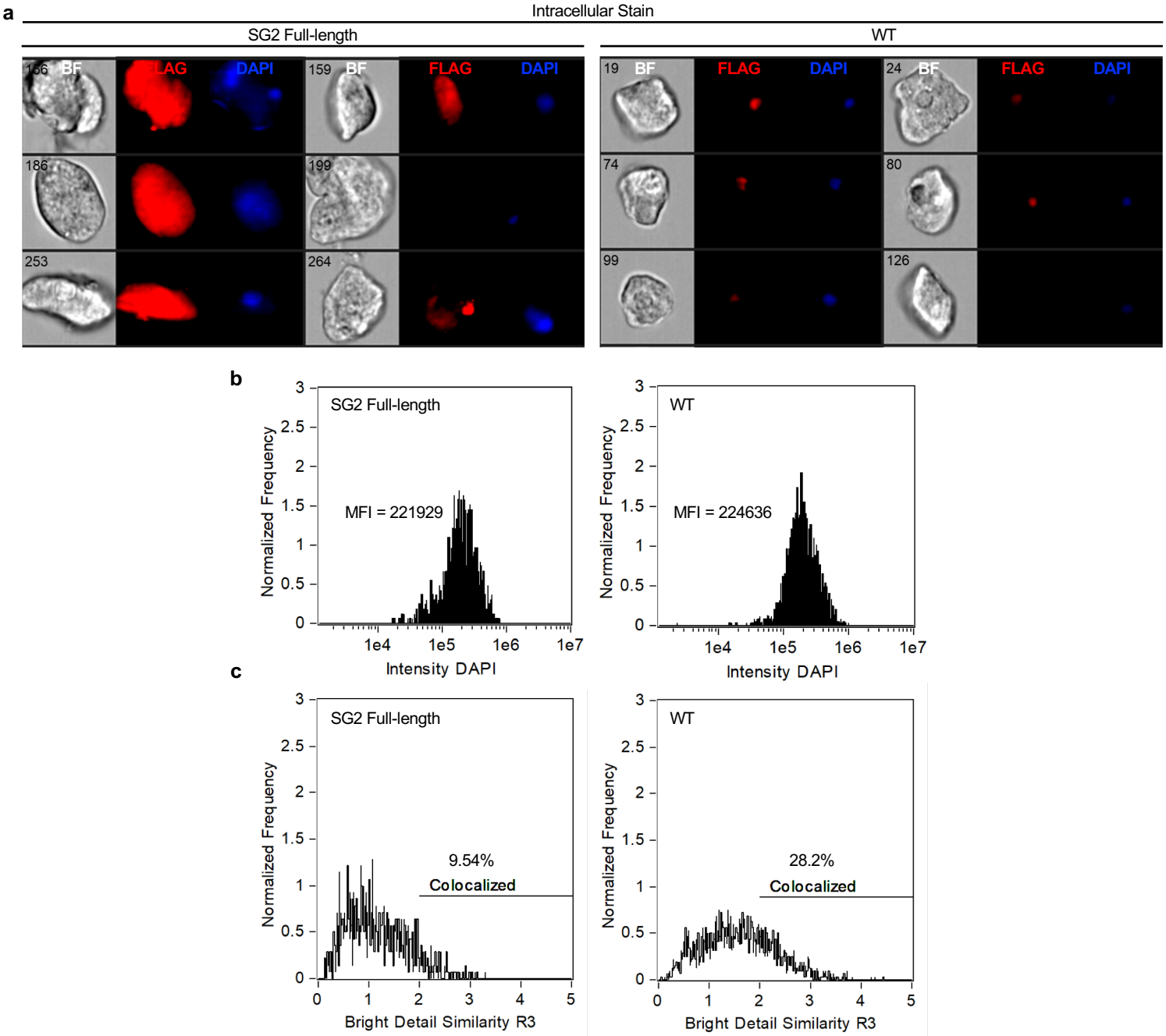

a
Intracellular Stain
SG2 Full-length
WT
BF
FLAG
DAPI
b
SG2 Full-length
MFI = 221929
WT
MFI = 224636
Normalized Frequency
Intensity DAPI
c
SG2 Full-length
9.54%
Colocalized
WT
28.2%
Colocalized
Bright Detail Similarity R3

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	Amino acid transporter, putative	Amino acid transporter, transmembrane domain	NA	NA
SG2	22.58 +/- 4.10	**	544	Forward	In	EHI_007010	Hypothetical protein	NA	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 transporter superfamily	NA	NA
SG5	14.50 +/- 0.91	***	463	Forward	In	EHI_024220	Hypothetical protein	Importin beta family	Protein import into nucleus	NA
				Forward	In	EHI_160920	Hypothetical protein	Importin beta family	Protein import into nucleus	NA
				Forward	In	EHI_171760	Hypothetical protein	Importin beta family	Obsolete protein import into nucleus, docking	NA
SG6	15.19 +/- 2.23	****	380	Reverse	In	EHI_093960	Pre-mRNA splicing factor, putative	Pre-mRNA-processing factor 6/Prp1/STA1	mRNA splicing, *via* spliceosome	Several
SG7	19.59 +/- 1.89	***	555	Reverse	In	EHI_009580	Tyrosine kinase, putative	Protein kinase-like domain superfamily	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	NA	Several
SG11	40.47 +/- 6.77	***	408	Reverse	Out	EHI_098820	Hypothetical protein	Dbl homology (DH) domain; Armadillo-type fold	Regulation of Rho protein signal transduction	NA
SG12	18.09 +/- 1.83	***	322	Reverse	In	DS571214	NA	NA	NA	NA

Figure S1

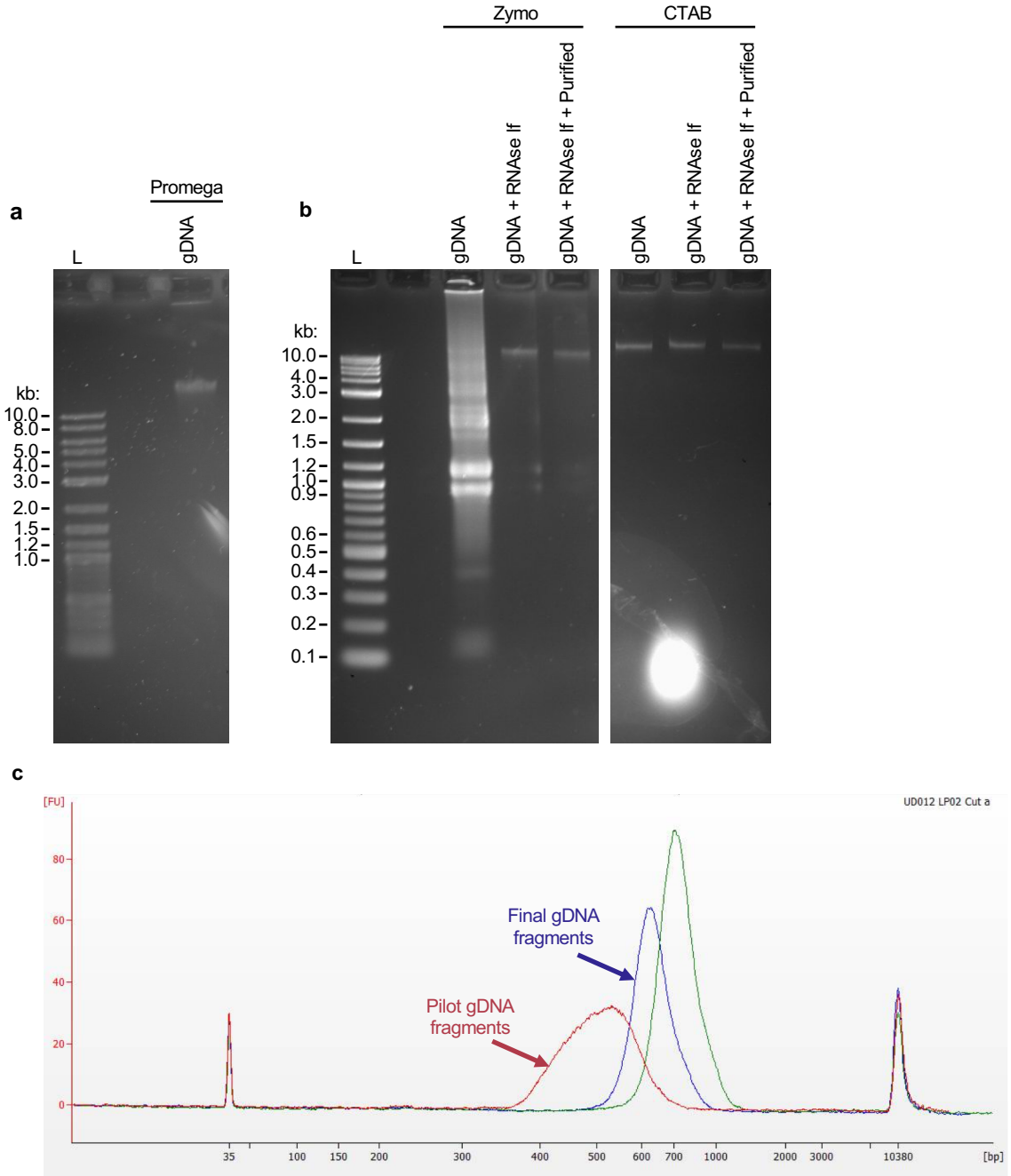

a
Promega
L
gDNA
kb:
10.0
8.0
5.0
4.0
3.0
2.0
1.5
1.2
1.0
b
Zymo
CTAB
L
gDNA
gDNA + RNAse If
gDNA + RNAse If + Purified
gDNA
gDNA + RNAse If
gDNA + RNAse If + Purified
kb:
10.0
4.0
3.0
2.0
1.5
1.2
1.0
0.9
0.6
0.5
0.4
0.3
0.2
0.1
c
[FU]
UD012 LP02 Cut a
80
60
40
20
0
Final gDNA fragments
Pilot gDNA fragments
35
100
150
200
300
400
500
600
700
1000
2000
3000
10380
[bp]

Figure S2

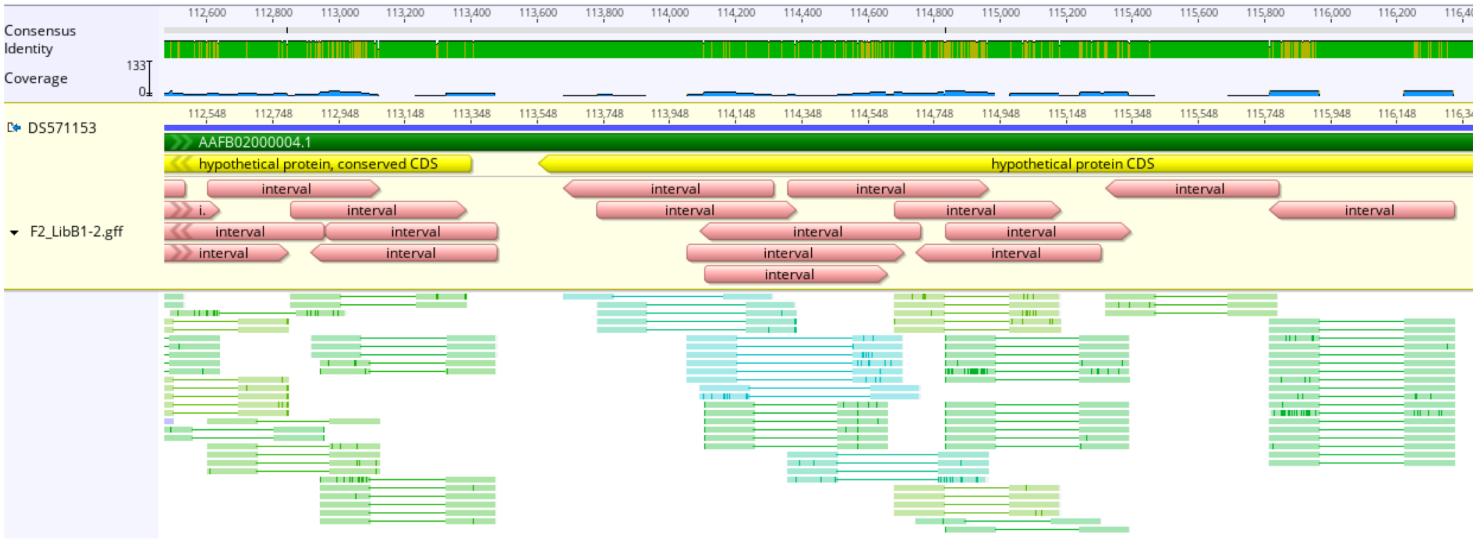

Consensus Identity
Coverage
133
0
DS571153
AAFB02000004.1
hypothetical protein, conserved CDS
hypothetical protein CDS
F2_LibB1-2.gff
interval

Figure S3

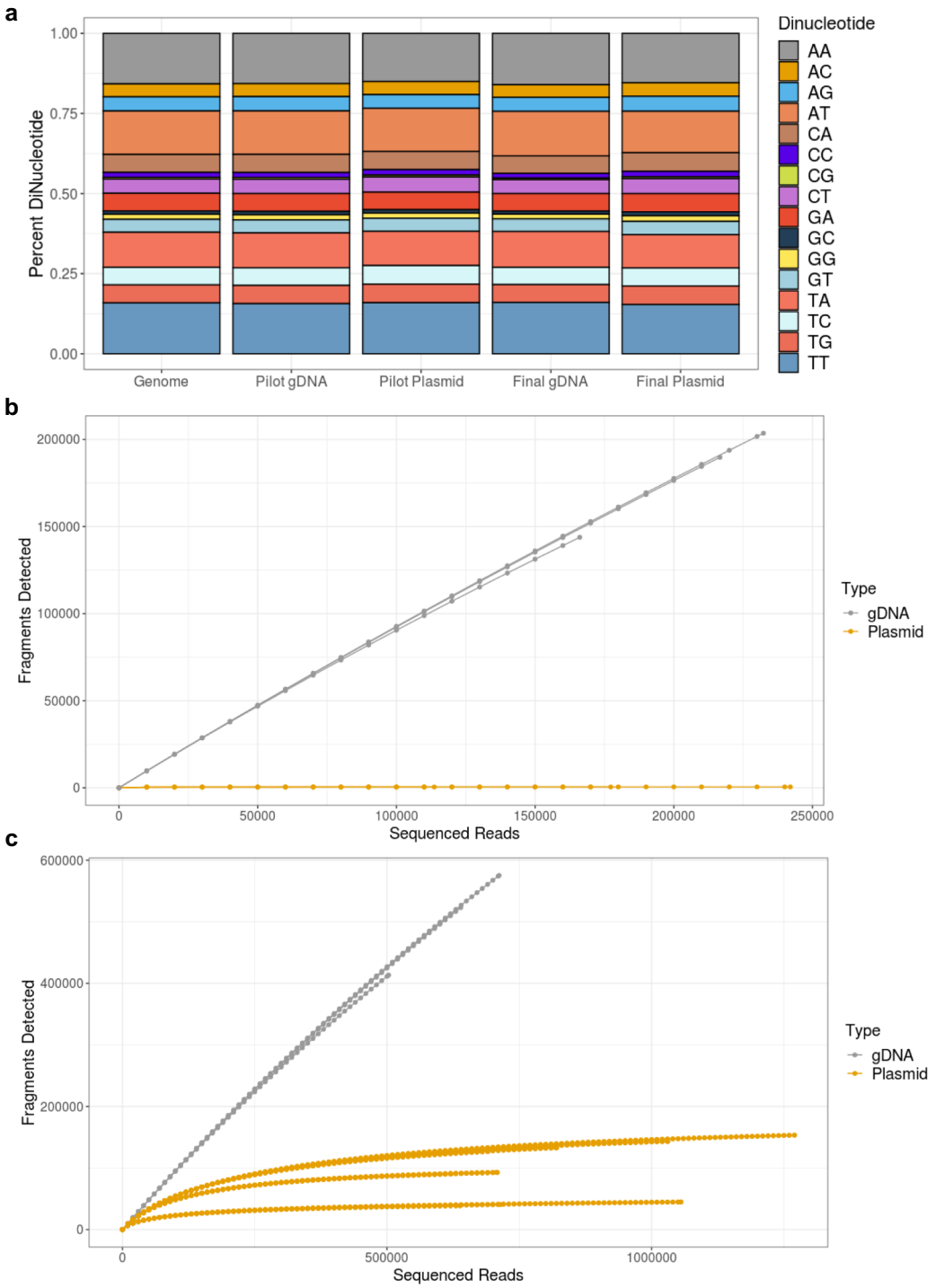

a
Percent DiNucleotide
1.00
0.75
0.50
0.25
0.00
Genome
Pilot gDNA
Pilot Plasmid
Final gDNA
Final Plasmid
Dinucleotide
AA
AC
AG
AT
CA
CC
CG
CT
GA
GC
GG
GT
TA
TC
TG
TT
b
Fragments Detected
200000
150000
100000
50000
0
0
50000
100000
150000
200000
250000
Sequenced Reads
Type
gDNA
Plasmid
c
Fragments Detected
600000
400000
200000
0
0
500000
1000000
Sequenced Reads
Type
gDNA
Plasmid

Figure S4

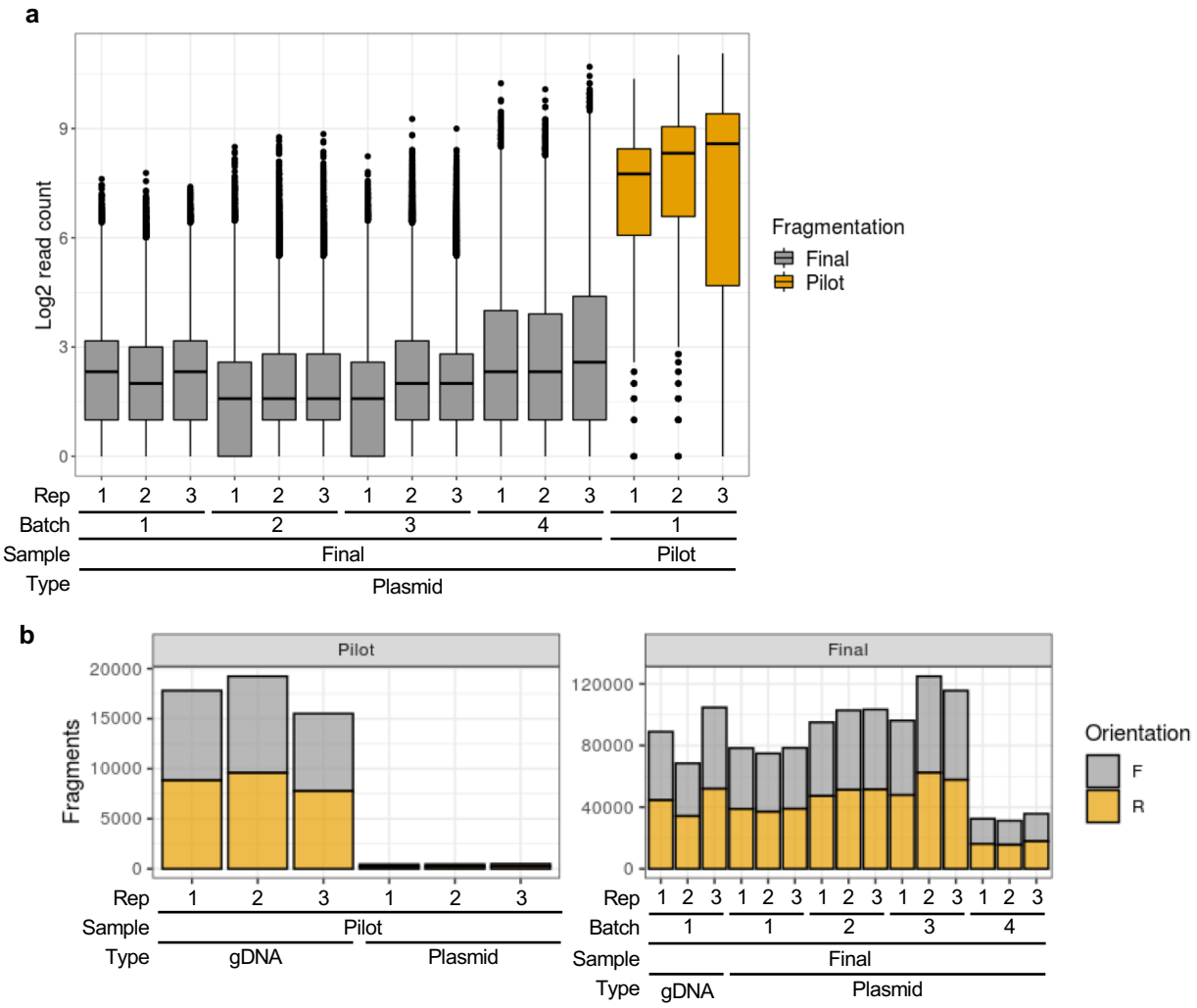

a
Log2 read count
0
3
6
9
Fragmentation
Final
Pilot
Rep 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
Batch 1 2 3 4 1
Sample Final Pilot
Type Plasmid
b
Pilot
Final
Fragments
20000
15000
10000
5000
0
120000
80000
40000
Orientation
F
R
Rep 1 2 3 1 2 3
Sample Pilot
Type gDNA Plasmid
Rep 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
Batch 1 1 2 3 4
Sample Final
Type gDNA Plasmid

Figure S5

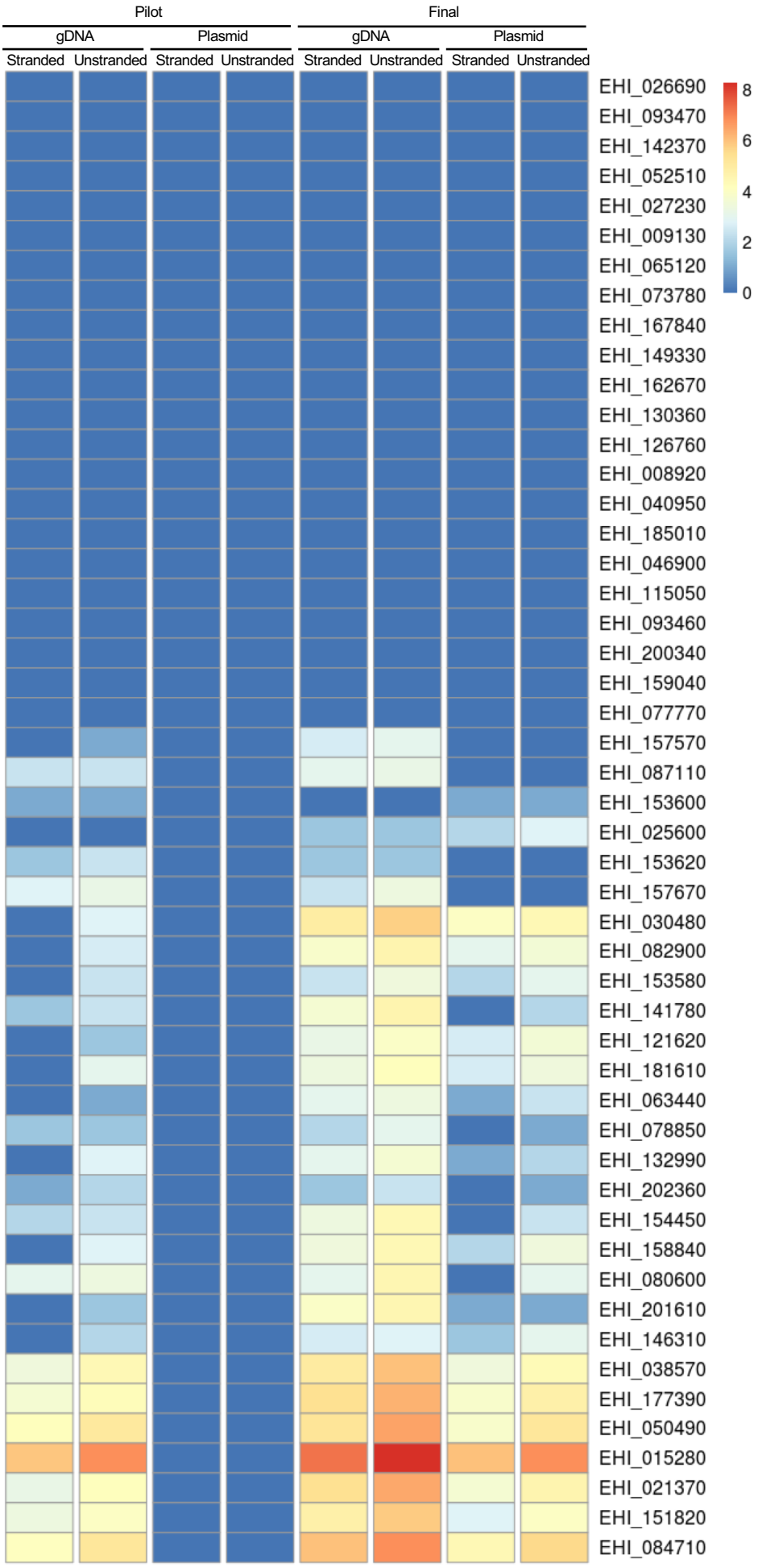

Pilot
Final
gDNA
Plasmid
gDNA
Plasmid
Stranded
Unstranded
Stranded
Unstranded
Stranded
Unstranded
Stranded
Unstranded
EHI_026690
EHI_093470
EHI_142370
EHI_052510
EHI_027230
EHI_009130
EHI_065120
EHI_073780
EHI_167840
EHI_149330
EHI_162670
EHI_130360
EHI_126760
EHI_008920
EHI_040950
EHI_185010
EHI_046900
EHI_115050
EHI_093460
EHI_200340
EHI_159040
EHI_077770
EHI_157570
EHI_087110
EHI_153600
EHI_025600
EHI_153620
EHI_157670
EHI_030480
EHI_082900
EHI_153580
EHI_141780
EHI_121620
EHI_181610
EHI_063440
EHI_078850
EHI_132990
EHI_202360
EHI_154450
EHI_158840
EHI_080600
EHI_201610
EHI_146310
EHI_038570
EHI_177390
EHI_050490
EHI_015280
EHI_021370
EHI_151820
EHI_084710
8
6
4
2
0

Figure S6

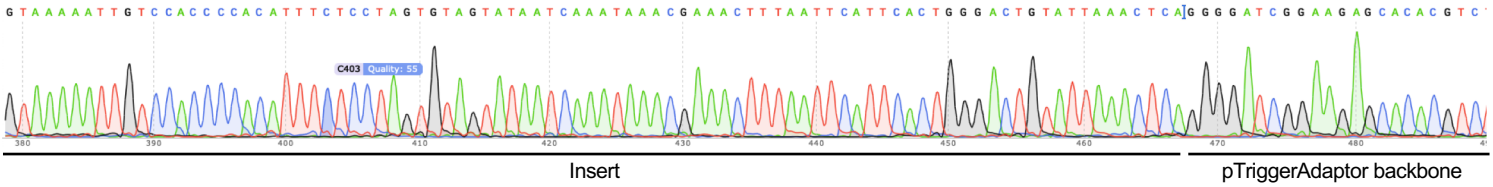


Figure S7

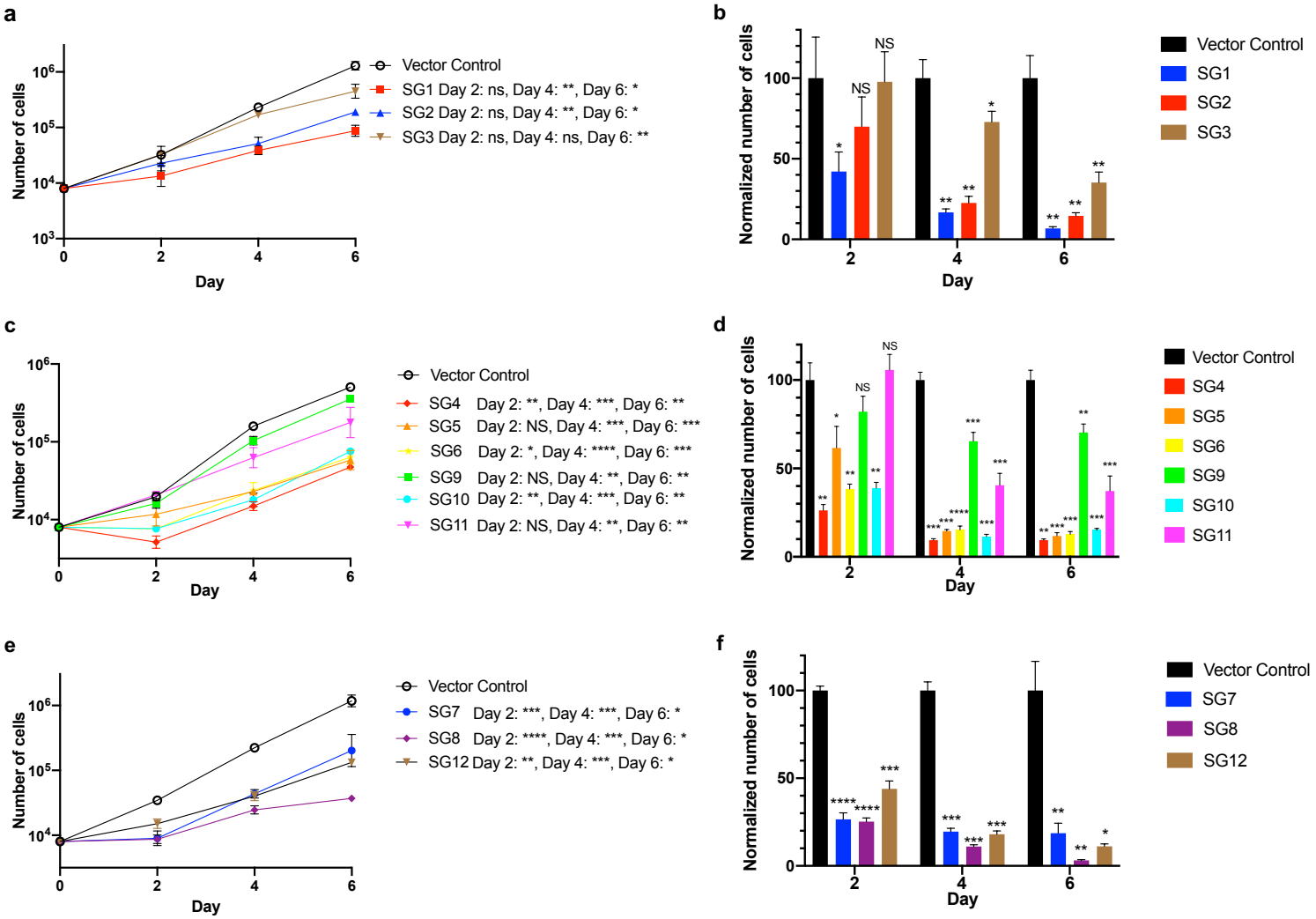
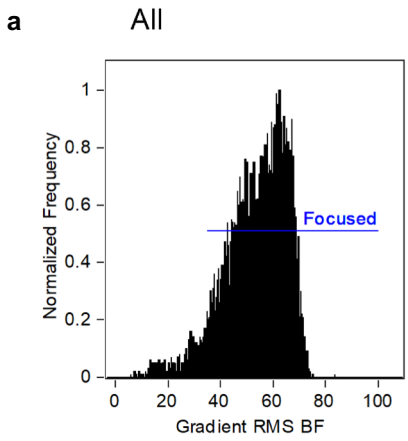
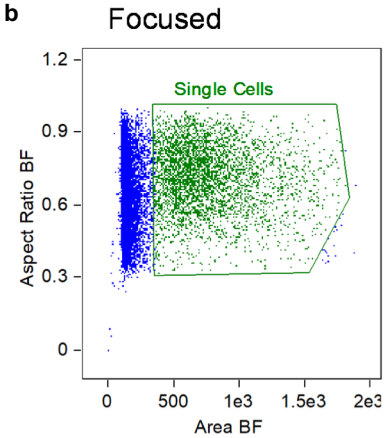

a
Number of cells
Day
Vector Control
SG1 Day 2: ns, Day 4: **, Day 6: *
SG2 Day 2: ns, Day 4: **, Day 6: *
SG3 Day 2: ns, Day 4: ns, Day 6: **
b
Normalized number of cells
Day
Vector Control
SG1
SG2
SG3
c
Number of cells
Day
Vector Control
SG4 Day 2: **, Day 4: ***, Day 6: **
SG5 Day 2: NS, Day 4: ***, Day 6: ***
SG6 Day 2: *, Day 4: ****, Day 6: ***
SG9 Day 2: NS, Day 4: **, Day 6: **
SG10 Day 2: **, Day 4: ***, Day 6: **
SG11 Day 2: NS, Day 4: **, Day 6: **
d
Normalized number of cells
Day
Vector Control
SG4
SG5
SG6
SG9
SG10
SG11
e
Number of cells
Day
Vector Control
SG7 Day 2: ***, Day 4: ***, Day 6: *
SG8 Day 2: ****, Day 4: ***, Day 6: *
SG12 Day 2: **, Day 4: ***, Day 6: *
f
Normalized number of cells
Day
Vector Control
SG7
SG8
SG12

Figure S8



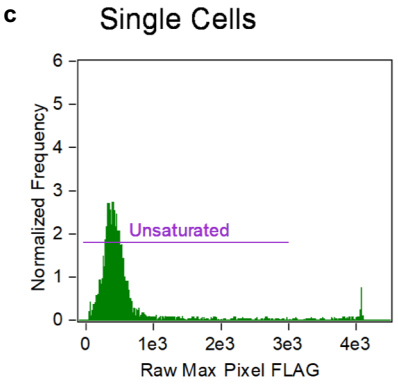
Gradient RMS BF

Population	Count	%Gated
All	10000	100
Focused	9312	93.1



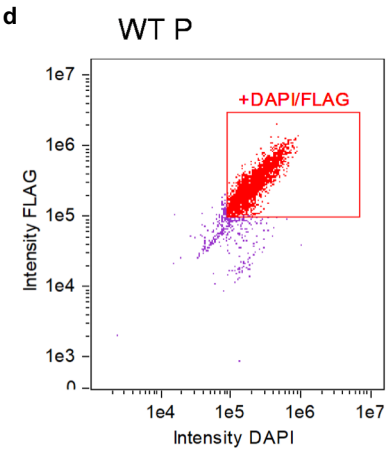
Area BF, Aspect Ratio BF

Population	Count	%Gated
Focused	9312	100
Single Cells	3634	39



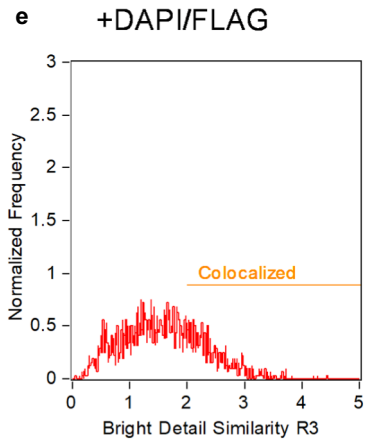
Raw Max Pixel FLAG

Population	Count	%Gated	Mean
Single Cells	3634	100	564.97
WT P	3534	97.2	473.1



Intensity DAPI, Intensity FLAG

Population	Count	%Gated
WT P	3534	100
+DAPI/FLAG	3198	90.5



Bright Detail Similarity R3

Population	Count	%Gated
+DAPI/FLAG	3198	100
Colocalized	901	28.2

Table S1

Sample	Mapped Pairs	Fragments
gDNA_Pilot_Rep1	447816	189741
gDNA_Pilot_Rep2	480134	203572
gDNA_Pilot_Rep3	342594	143817
Plasmid_Pilot_Rep1	276964	472
Plasmid_Pilot_Rep2	399114	481
Plasmid_Pilot_Rep3	573080	540
gDNA_Final_Rep1	1317646	522946
gDNA_Final_Rep2	1035946	413343
gDNA_Final_Rep3	1466722	575167
Plasmid_Final_Batch1_Rep1	1500958	92822
Plasmid_Final_Batch1_Rep2	1318560	90347
Plasmid_Final_Batch1_Rep3	1501272	93088
Plasmid_Final_Batch2_Rep1	1457526	126868
Plasmid_Final_Batch2_Rep2	1723194	134349
Plasmid_Final_Batch2_Rep3	1740592	132963
Plasmid_Final_Batch3_Rep1	1467720	129278
Plasmid_Final_Batch3_Rep2	2712002	153399
Plasmid_Final_Batch3_Rep3	2207970	143289
Plasmid_Final_Batch4_Rep1	1588276	40770
Plasmid_Final_Batch4_Rep2	1418664	38870
Plasmid_Final_Batch4_Rep3	2323702	44937

Table S2

Accession	Annotation	Final gDNA	Final Plasmid	Paralogs	RNAseq, FPKM	Proteomics, Spectra
EHI_026690	protein kinase domain containing protein	-	-	-	0.01	2
EHI_093470	hypothetical protein	-	-	-	0.01	-
EHI_142370	hypothetical protein, conserved	-	-	1	0.01	-
EHI_052510	hypothetical protein	-	-	5	0.01	-
EHI_027230	hypothetical protein, conserved	-	-	2	0.01	-
EHI_009130	hypothetical protein	-	-	58	0.01	-
EHI_065120	hypothetical protein, conserved	-	-	1	0.01	-
EHI_073780	DNAJ homolog subfamily A member 2, putative	-	-	2	0.01	-
EHI_167840	WH2 motif domain contaning protein	-	-	-	0.01	-
EHI_149330	hypothetical protein, conserved	-	-	2	0.01	-
EHI_162670	hypothetical protein	-	-	-	0.01	-
EHI_130360	modulator of drug activity B homolog, putative	-	-	-	0.01	-
EHI_126760	hypothetical protein	-	-	-	0.01	-
EHI_008920	hypothetical protein	-	-	4	0.01	-
EHI_040950	hypothetical protein	-	-	1	0.01	-
EHI_185010	ribosomal protein L10, putative	-	-	2	0.17	1
EHI_046900	4-alpha-glucanotransferase, putative	-	-	1	0.01	-
EHI_115050	protein kinase, putative	-	-	58	0.01	-
EHI_093460	hypothetical protein	-	-	-	0.01	-
EHI_200340	hypothetical protein	-	-	-	0.01	-
EHI_159040	plasma membrane calcium ion-transporting ATPase, putative	-	-	6	0.01	-
EHI_077770	UDP-glucose:glyco protein glucosyltransferase 1 precursor, putative	-	-	-	0.1	-
EHI_157570	tRNA-Asp (anticodon: GTC)	+	-	-	0.5	-
EHI_087110	hypothetical protein, conserved	+	-	3	432.1	-
EHI_153620	prolyl oligopeptidase family protein	+	-	2	0.05	-
EHI_157670	hypothetical protein	+	-	4	0.56	-

Table S3

Accession	Annotation	Final gDNA	Final Plasmid	Paralogs	RNAseq, FPKM	Proteomics, Spectra
EHI_148050	hypothetical protein, conserved	+	+	-	169.21	22
EHI_148280	p21-activated kinase	+	+	1	404.02	28
EHI_148450	hypothetical protein	+	+	-	17.41	-
EHI_148490	hypothetical protein, conserved	+	+	-	7.62	-
EHI_148550	protein tyrosine kinase domain-containing protein	+	+	13	3.55	-
EHI_148680	hypothetical protein, conserved	+	+	-	18.06	7
EHI_148770	hypothetical protein, conserved	+	+	-	7.66	4
EHI_148840	RhoGAP domain containing protein	+	+	-	16.71	-
EHI_148870	hypothetical protein	+	+	-	38.06	4
EHI_149020	hypothetical protein	+	+	-	3.46	-

Table S4

Primer	Purpose	Sequence
BAR_IFF_F	Cloning	TTAAAAACAAAAATTGCAGCATCAACGCCCACCAGATGCAGGAAGATCATCA
BAR_IFF_R	Cloning	GTTTAAAAAAGAAGAGTTCAACTCGAGCCCCTCCACTCCACCAATCACCT
BAR_RTPCR_F	RT-PCR	AGGCTTGTTGAGGCATGTACTA
BAR_RTPCR_R	RT-PCR	AAGTTCTCCACTCCACCAATCA
MSP1_IFF_F	Cloning	TTGCAGCATCAACGCCCGGGAATAATCAGAAACGAAATATTTTTGG
MSP1_OFF_F	Cloning	AAAGAAGAGTTCAACTCGAGACTAGATGAAAAACATCTTGAATCTT
MSP1_RTPCR_F	RT-PCR	GGTGAGACTCAAGCAAGGGG
MSP1_RTPCR_R	RT-PCR	GGCCATTCTTCTACTTCGCCA
Adaptor1primer_F	Library construction	ACACTCTTTCCCTACACGAC
Adaptor2primer_R	Library construction	GTGACTGGAGTTCAGACGTGT
SG0_F	Cloning	TTAAAAACAAAAATTGCAGCATCAACGCCCATGTCTTTCCAATTACCAC
SG0_R	Cloning	GTTTAAAAAAGAAGAGTTCAACTCGAGCCCTTGAGGGGAAAAGTGAA
SG1i_F	Cloning	TTAAAAACAAAAATTGCAGCATCAACGCCCGTGATGATACTTCTGTCTCTAAC
SG1i_R	Cloning	GTTTAAAAAAGAAGAGTTCAACTCGAGCCCCGATACAACGCTCAGTATCTC
SG2i_F	Cloning	TTAAAAACAAAAATTGCAGCATCAACGCCCCCATTACCACATTAAGTTCTTC
SG2i_R	Cloning	GTTTAAAAAAGAAGAGTTCAACTCGAGCCCGGATAAACAAGAACTAAGTTTTG
SG0_RTPCR_F	RT-PCR	GCTCTTGAGCCTCATATTAGT
SG0_RTPCR_R	RT-PCR	CAACGAACTTCCAGTTGACTAC
SG1_RTPCR_F	RT-PCR	GGAGATACTGAGCGTTGTATCG
SG1_RTPCR_R	RT-PCR	GCATACCACTGTGTTTCTCC
SG2_RTPCR_F	RT-PCR	CATGCCTCACCACATTCACT
SG2_RTPCR_R	RT-PCR	ACCAAACTCCCATCCTTCAA
GAPDH_RTPCR_F	RT-PCR	CGTCCACAGACAATTCGAAGGAAC
GAPDH_RTPCR_R	RT-PCR	AAGGCAGTTGGTTGTGCATGA
CS5'_F	Sanger sequencing	TCAGTCTTACCACGTCATAAAGT
CS3'_R	Sanger sequencing	TGCAAGAAGATGTTACAAAGCA
LibraryRNAiSeq_F	Sanger sequencing	ACACTCTTTCCCTACACGACG
LibraryRNAiSeq_R	Sanger sequencing	GTGACTGGAGTTCAGACGTGT