Versatile multi-transgene expression using improved BAC TG-EMBED toolkit, novel BAC episomes, and BAC-MAGIC

Binhui Zhao[1†], Pankaj Chaturvedi[1†], David L. Zimmerman[1] and Andrew S. Belmont[1,*]

[1] Department of Cell and Developmental Biology, University of Illinois, Urbana, Illinois, 61801, United States of America

* To whom correspondence should be addressed. Tel: 217-244-2311; Email: asbel@illinois.edu

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Present Address: David L. Zimmerman, Biology Department, College of the Ozarks, Point Lookout, Missouri, 65726, United States of America

ABSTRACT

Achieving reproducible, stable, and high-level transgene expression in mammalian cells remains problematic. Previously, we attained copy-number-dependent, chromosome-position-independent expression of reporter minigenes by embedding them within a BAC containing the mouse *Msh3-Dhfr* locus (DHFR BAC). Here we extend this "BAC TG-EMBED" approach. First, we report a toolkit of endogenous promoters capable of driving transgene expression over a 0.01-5 fold expression range relative to the CMV promoter, allowing fine-tuning of relative expression levels of multiple reporter genes expressed on a single BAC. Second, we show small variability in both the expression level and long-term expression stability of a reporter gene embedded in BACs containing either transcriptionally active or inactive genomic regions, making choice of BACs more flexible. Third, we describe an intriguing phenomenon in which BAC transgenes are maintained as episomes in a large fraction of stably selected clones. Finally, we demonstrate the utility of BAC TG-EMBED by simultaneously labeling three nuclear compartments in 94% of stable clones using a multi-reporter DHFR BAC, constructed with a combination of synthetic biology and BAC recombineering tools. Our extended BAC TG-EMBED method provides a versatile platform for achieving reproducible, stable simultaneous expression of multiple transgenes maintained either as episomes or stably integrated copies.

INTRODUCTION

Transgene expression has been widely used in both basic research and biotechnology. Applications of transgene expression range from the elucidation of gene function by ectopic expression of selected transgenes, to the expression of transgenes for gene therapy, and to the overexpression of genes for production of biopharmaceuticals (1–5). Examples of such applications include the expression of multiple fluorescent proteins for live-cell imaging (6), the expression of the four or more Yamanaka transcription factors for efficient generation of induced pluripotent stem (iPS) cells (7), and the expression of multiple proteins for reconstitution of protein complexes (8).

Despite the currently widespread use of transgene expression, most transgene expression systems still suffer from serious experimental limitations. Plasmid-, lentivirus- and transposon- based systems, all still show varying degrees of chromosome position effects (9, 10) and position effect variegation (PEV) (11–15). Moreover, foreign sequences by themselves are targets for epigenetic silencing (16–19), and transgene concatamers can induce the formation of heterochromatin (20, 21). Together these transgene silencing mechanisms result in unpredictable transgene expression levels that do not correlate with copy number and are unstable with long-term culture or changes in the cell physiological or differentiated state (22–24).

Such limitations are compounded when the simultaneous and reproducible expression of multiple transgenes is required. For example, a common application in the emerging field of synthetic biology is the design of

61 novel gene circuits, involving the expression of multiple proteins, in many cases
62 at precise relative levels (25). While this approach has worked well in
63 prokaryotes and yeast, it has been difficult to implement in mammalian cells due
64 to the lack of suitable multi-transgene expression methods which overcome
65 chromosome position effects and allow expression of different transgenes at
66 reproducible relative levels.

67 A commonly used approach to countering transgene silencing and
68 variegation has been through the inclusion of *cis*-elements. These include
69 insulators (26, 27), locus control regions (LCRs) (28, 29), scaffold/matrix
70 attachment regions (S/MARs) (30, 31), ubiquitous chromatin opening elements
71 (UCOEs) (32, 33) and anti-repressors (34); some of these regulatory elements
72 have context-dependent and/or vector dependent activity. While these *cis*-
73 elements improve transgene expression to varying degrees, they are insufficient
74 for chromosome-position independent, copy-number-dependent transgene
75 expression (29, 35–37).

76 Additionally, in some transgene expression applications the ability to avoid
77 transgene chromosomal integration and eventually eliminate these transgenes
78 from the cells is highly desirable. Both viral-sequence based and non-viral, pEPI
79 based episomal vectors have been developed (38–41). Viral-based vectors have
80 the potential of causing transformation of the transfected cells (42), while pEPI-
81 like vectors, containing a S/MAR sequence immediately downstream of an active
82 transcription unit, are mitotically stable without selection (43–47), and thus
83 cannot be removed from the cells. Moreover, transgenes on these episomal

vectors are still subject to silencing (48), possibly due to the prokaryotic or viral sequences on these vectors (49, 50).

Bacterial artificial chromosomes (BACs) carrying ~100-200 kb mammalian genomic DNA inserts harbor most of the *cis*-regulatory sequences required for expression of the endogenous genes contained within these genomic inserts. Previously we demonstrated how embedding minigene constructs at different locations within the DHFR BAC provided reproducible expression of single or multiple reporter genes independent of the chromosome integration site (51). Similar approaches were used by other labs for high-level recombinant protein production (52, 53). Recently, our lab demonstrated stable transgene expression after cell-cycle arrest or after terminal cell differentiation, using the BAC-TG EMBED approach (54). All of these studies tested only BACs containing actively transcribed regions, based on the hypothesis that the expression level of the transgenes inserted into the BACs was determined by the chromatin environments reconstituted by the genomic inserts within the BACs. Indeed, because of this assumption, previous studies have specifically targeted the inserted transgenes to transcription units and even exons (51–53).

However, this hypothesis has not been tested. Moreover, overexpression from the genes on the BAC genomic inserts might change the properties of the transfected cells, or interfere with other assays of a study. Thus, BACs with no transcription units would be more desirable. Another improvement over our previous BAC TG-EMBED system (51, 54) would be a toolkit of endogenous promoters capable of driving transgene expression over a wide range of defined,

107 relative expression levels. Viral promoters, including the CMV promoter we used
108 previously, are known to be prone to epigenetic silencing (55, 56), while most
109 previously used endogenous and synthetic promoters were selected for their
110 strength (53, 57–60). While high-level transgene expression is preferable in
111 applications calling for overexpression, a low or near-physiological expression is
112 important for many other applications, including gene therapy. Additionally,
113 multiple transgenes may need to be expressed simultaneously but at
114 reproducible differential levels.

115 Here we describe further extensions to the BAC TG-EMBED method that
116 together provide a more versatile BC TG-EMBED toolkit for a range of future
117 potential applications. First, we describe a toolkit of endogenous promoters, for
118 which we have measured relative promoter strength, that can drive transgene
119 expression at reproducible relative levels over a 500-fold range. Second, we
120 show that multiple BAC scaffolds can be used to drive sustained high-level
121 transgene expression driven by the UBC promoter without selection for up to 12
122 weeks, including BAC scaffolds containing no active transcription units. Third,
123 we describe an episomal version of BAC TG-EMBED, where BAC transgenes
124 form circular, ~1 Mb episomes and can be eliminated from the cells by removing
125 selection. Fourth, we developed a "BAC-MAGIC" (**BAC**-**M**odular **A**ssembly of
126 **G**enomic loci **I**nterspersed **C**assettes) to more rapidly assemble BACs containing
127 multiple transgene expression cassettes. Finally, as a proof-of-principle
128 demonstration of our new, more versatile BAC TG-EMBED toolkit, we
129 demonstrate simultaneous expression of fluorescently tagged proteins labeling

130 three different nuclear compartments, achieving >90% optimally labeled cell

131 clones after a single, stable transfection.

132

133 **MATERIALS AND METHODS**

134 **PCR amplification of endogenous promoters**

135 Primers (Supplementary Table S1) were designed using Primer3 (61) or
136 NCBI primer blast (62) to amplify 1-3 kb promoter regions which included either
137 the entire or part of the 5' UTRs upstream of the first exons of target genes. We
138 used human genomic DNA extracted from BJ-hTERT cells as the template for
139 PCR. However, the UBC promoter, including a partially synthetic intron, was
140 amplified from plasmid pUGG (54).

141

142 **Construction of dual reporter DHFR BACs**

143 The original dual reporter BAC, DHFR-HB1-GN-HB2-RZ (51), was derived
144 from the CITB-057L22 BAC (DHFR BAC) containing mouse chr13:92992156-
145 93161185 (mm9). DHFR-HB1-GN-HB2-RZ has an EGFP expression cassette
146 inserted 26 kb downstream of the *Msh3* transcription start site, and a mRFP
147 expression cassette inserted at 121 kb downstream of the *Msh3* transcription
148 start site. The EGFP expression cassette contains a CMV promoter-driven
149 EGFP gene and a SV40 promoter-driven Kanamycin/Neomycin resistance gene,
150 while the mRFP expression cassette has a CMV promoter-driven mRFP gene
151 and a SV40-driven Zeocin resistance gene. New dual reporter DHFR BACs
152 were created using a similar strategy to that used to create DHFR-HB1-GN-HB2-
153 RZ, except that new mRFP expression cassettes were used, where the CMV
154 promoter was replaced with alternative, human endogenous promoters. The
155 intermediate DHFR BAC containing only the EGFP expression cassette, DHFR-

156 HB1-GN (51), was used to insert these new mRFP expression cassettes using λ
157 Red-mediated homologous recombination (63, 64).
158 Plasmid p[MOD-HB2-CRZ] (51) contains a CMV driven mRFP and a SV40
159 driven Zeocin resistance gene, flanked by two ~500 bp regions homologous to
160 the DHFR BAC target site. Plasmid p[MOD-HB2-RCS-Zeo] was created by
161 replacing the CMV-mRFP fragment between NotI and NheI sites of p[MOD-HB2-
162 CRZ] with a synthetic DNA fragment "RCS" containing multiple rare restriction
163 sites (Supplementary Table S2). The mRFP fragment generated by digesting
164 p[MOD-HB2-CRZ] with NheI was then inserted into the NheI site of p[MOD-HB2-
165 RCS-Zeo], yielding plasmid p[MOD-HB2-RCS-RZ]. The PCR-amplified
166 endogenous promoters were then inserted into the RCS, generating plasmids
167 p[MOD-HB2-promoter name-RZ]. Promoter functionality was tested by transient
168 transfection of NIH 3T3 cells with these plasmids.
169 To insert the new mRFP expression cassettes into the DHFR-HB1-GN
170 BAC, one round of λ Red-mediated recombination, using Zeocin resistance as
171 positive selection, was performed according to a published protocol (63). DNA
172 fragments containing the new mRFP expression cassettes with a given promoter
173 with flanking homologous arms were excised from p[MOD-HB2-promoter name-
174 RZ] plasmids by PmeI. SW102, a derivative strain of *Escherichia coli* (*E. coli*),
175 was used for recombination. Recombinants were selected on low-salt LB plates
176 containing 25 μg/ml Zeocin and 12.5 μg/ml Kanamycin at 32°C for ~20 hours.
177 Recombinant colonies were screened by PCR amplification of sequences
178 flanking the site of insertion (primers listed in Supplementary Table S1). The

179 integrity of BAC constructs was verified by restriction enzyme fingerprinting,
180 where observed band patterns on agarose gels were compared with predicted
181 ones.
182
183 **Construction of BACs containing the UBC-GFP-ZeoR cassette**
184 Construction of pUGG containing the UBC-GFP-ZeoR-FRT-GalK-FRT
185 cassette was described previously (54). Human BACs RP11-138I1 (UBB BAC),
186 CTD-2643I7 (HBB BAC), CTD-2207K13 (2207K13 BAC) and mouse BAC RP23-
187 401D9 (ROSA BAC) were obtained from Thermo Fisher Scientific. Mouse BAC
188 CITB-057L22 (DHFR BAC) was a gift from Edith Heard (Curie Institute, Paris,
189 France).
190 The UBC-GFP-ZeoR reporter gene insertion positions (mm9 or hg19) are
191 chr17:16,301,887-16,301,888 in the UBB BAC, chr6:113,043,332-113,043,333 in
192 the ROSA BAC, chr13:93,099,101-93,099,102 in the DHFR BAC,
193 chr1:79,224,725-79,224,726 in the 2207K13 BAC, and chr11:5,390,233-
194 5,390,244 in the HBB BAC.
195 λ Red-mediated BAC recombineering (63, 64) using a *galK*-based dual-
196 selection scheme was used to introduce the UBC-GFP-ZeoR reporter cassette
197 onto the BACs according to published protocols (63). DNA fragments with
198 homology ends for recombineering were prepared by PCR using primers
199 (Supplementary Table S1) with 74-bp homology sequences plus 16-bp
200 sequences (forward, 5'- acagcagagatccagt-3'; reverse, 5'-tgttggctagtgcgt-3') that
201 amplify the UBC-GFP-ZeoR-FRT-GalK-FRT cassette from plasmid pUGG. *E.*

202 *coli* strain SW105 was used for BAC recombineering. Recombinants containing
203 the UBC-GFP-ZeoR-FRT-GalK-FRT cassette were selected for *galK* insertion at
204 32°C on minimal medium in which D-galactose was supplied as the only carbon
205 source. Recombinant colonies were screened using PCR with BAC specific
206 primers flanking the target regions (Supplementary Table S1). Subsequently,
207 FLP recombinase-mediated removal of *galK* from selected recombinant clones
208 was done by inducing actively growing SW105 cells with 0.1% (w/v) L-arabinose.
209 Negative selection against *galK* used minimal medium containing 2-deoxy-
210 galactose; deletion of *galK* in recombinants was again verified using BAC specific
211 primers (Supplementary Table S1). The integrity of BAC constructs was verified
212 by restriction enzyme fingerprinting.
213 The UBB, HBB, 2207K13, ROSA, DHFR BACs with the UBC-GFP-ZeoR
214 reporter gene inserted were named UBB-UG, HBB-UG, 2207K13-UG, ROSA-UG
215 and DHFR-UG, respectively.
216
217 **Cell culture and establishment of BAC cell lines**
218 Mouse NIH 3T3 fibroblasts (ATCC CRL-1658™) were grown in Dulbecco's
219 modified Eagle medium (DMEM, with 4.5 g/l D-glucose, 4 mM L-glutamine, 1 mM
220 sodium pyruvate and 3.7 g/l $NaHCO_3$) supplemented with 10% HyClone Bovine
221 Growth Serum (GE Healthcare Life Sciences, Cat. # SH30541.03). Human
222 HCT116 cells (ATCC CCL-247™) were grown in McCoy's 5A medium
223 supplemented with 10% Fetal Bovine Serum (Seradigm, Cat. # 1500-500H).

224 BAC DNA for transfection of mammalian cells was prepared with the
225 QIAGEN Large Construct Kit (QIAGEN, Cat. # 12462) as per the manufacturer's
226 instructions. All BACs except DHFR BAC derived BACs were linearized before
227 transfection: 2207K13-UG BAC with SgrAI (New England Biolabs, Cat. #
228 R0603S), HBB-UG BAC with NotI (New England Biolabs, Cat. # R3189S) and all
229 other BACs with the PI-SceI (New England Biolabs, Cat. # R0696S).
230 Lipofectamine 2000 (Thermo Fisher Scientific, Cat. # 11668019) was used to
231 transfect the cells with the BACs according to the manufacturer's directions. The
232 dual reporter DHFR BACs and the BACs containing the UBC-GFP-ZeoR reporter
233 gene were transfected into NIH 3T3. The 2207K13-UG BAC was also transfected
234 into HCT116. The DHFR BACs containing the Lac operator repeats were
235 transfected into an NIH 3T3 cell clone 3T3_LG_C29 stably expressing the EGFP-
236 dimer LacI-NLS fusion protein (EGFP-LacI) (65). Mixed clonal populations of
237 stable transformants were obtained after ~2 weeks of selection (75 µg/ml Zeocin
238 and 500 µg/ml G418 for NIH 3T3 cells transfected with the dual reporter DHFR
239 BACs; 75 µg/ml or 200 µg/ml Zeocin for NIH 3T3 or HCT116 cells, respectively,
240 transfected with the BACs containing the UBC-GFP-ZeoR reporter gene; 75
241 µg/ml Zeocin and 200 µg/ml Hygromycin B for 3T3_LG_C29 transfected with the
242 DHFR BACs); individual cell clones were obtained by serial dilution or colony
243 picking using filter discs (66).
244 To analyze the stability of reporter gene expression in NIH 3T3 cells,
245 individual cell clones were grown continuously with or without Zeocin (75 µg/ml)
246 selection for 96 days. We used the following clones (Figure 4 and Supplementary

247 Figure S1): DHFR-UG BAC- f1-7, f3-13, f3-15 (uniform), f1-6, f2-1, f2-3
248 (heterogeneous); ROSA-UG BAC- 2D6- 3C11, 3D7 (uniform), 2C12, 3A1
249 (heterogeneous); UBB-UG BAC- 1C2, 1F1, 1F12, 2F5, 2G4, 4D3, 5C1, 5C7
250 (uniform), 1A8, 1D5, 6H2 (heterogeneous); 2207K13-UG BAC- 3E3, 5C8, 5E1,
251 6B9, 6E12, 6F4, 7B2 (uniform), 1E3, 6A2, 6C10, 7B9 (heterogeneous).
252

253 **Flow cytometry**

254 For analysis of reporter gene expression, cells were grown to ~40%-80%
255 confluence, trypsinized, and resuspended in growth media at ~0.5-1 million/ml.
256 For analysis of the expression of mRFP and EGFP, or mRFP alone, cell
257 suspensions were run on a BD FACS AriaII (BD Biosciences) or a BD LSR
258 Fortessa (BD Biosciences), using the PE channel (561 nm laser and 582/15 nm
259 bandpass filter) for mRFP, and the FITC channel (488 nm laser, 505 longpass
260 dichroic mirror and 530/30 nm bandpass filter) for EGFP. For analysis of GFP
261 expression alone, the cell suspensions were run on a BD FACS Canto II Flow
262 Cytometry Analyzer (BD Biosciences), using the FITC/Alexa Fluor-488 channel
263 (488nm laser, 502 longpass dichroic mirror and 530/30 bandpass filter).
264 Rainbow fluorescent beads (Spherotech, Cat. # RFP-30-5A) were used as
265 fluorescence intensity standards. Each sample was run for 1-2 min or until the
266 number of events after gating reached 10-20 thousand.
267 For cell sorting, cells were resuspended at ~10 million/ml in growth media
268 and run on a BD FACS AriaII for up to 30-40 minutes. Sorting windows are
269 shown in the main and supplementary figures.

270

271 **Estimation of relative promoter strength**

272 The red and green fluorescence of the mixed-clonal populations stably
273 transfected with the dual-reporter DHFR BACs was measured by flow cytometry.
274 The mean florescence values of all gated cells were divided by the bead intensity
275 values for normalization. The ratio of normalized mRFP to normalized EGFP
276 was calculated as a measure of promoter strength (Equation 1). All promoter
277 strengths were then normalized with the CMV promoter strength (comparing the
278 CMV-driven mRFP to the CMV-driven EGFP expression) to calculate the relative
279 promoter strength (Equation 2) using the CMV promoter as the reference.

280

$$\text{promoter strength} = \frac{median(\text{PE}_{\text{cells}})/median(\text{PE}_{\text{beads}})}{median(\text{FITC}_{\text{cells}})/median(\text{PE}_{\text{beads}})} \quad 1$$

$$\text{relative promoter strength} = \frac{\text{promoter strength}_x}{\text{promoter strength}_{\text{CMV}}} \quad 2$$

281

282 **Genomic DNA extraction**

283 Genomic DNA was isolated by phenol/chloroform extraction (67). Cultured
284 cells were harvested and washed with 1x Cell Culture Phosphate Buffered Saline
285 (PBS, Corning, Cat. # 21040CV). Sorted cells were pelleted. Up to ~2 million cells
286 were resuspended in 100 µl High-TE buffer (10 mM Tris-Cl, pH 8, 10 mM EDTA,
287 25-100 µg/ml RNase A (QIAGEN, Cat. # 19101)) and lysed by adding 2.5 µl 20%
288 SDS. After incubation at 37°C for several hours, the lysate was digested by ~0.2
289 mg/ml Proteinase K (New England Biolabs, Cat. # P8102 or P8107S) at 55°C for

290 ~1 day. 1 M Tris-Cl (pH 8.0), 5 M NaCl and nuclease free water were added to
291 the lysate to bring up the total volume to ~600 µl and final concentrations of Tris-
292 Cl to ~0.1 M and NaCl to ~0.2 M. The lysate was then extracted once with an
293 equal volume of phenol/chloroform/isoamyl alcohol (25:24:1 mixture, Fisher
294 Scientific, Cat. # BP1752I-400) and once with an equal volume of
295 chloroform/isoamyl alcohol (24:1 mixture, MilliporeSigma, Cat. # C0549). DNA
296 was precipitated by adding 2.5 volumes of 100% ethanol, washed with 70%
297 ethanol and resuspended in EB (10mM Tris-Cl, pH 8.5).
298

299 **Estimation of transgene copy number**

300 BAC or plasmid transgene copy number within individual cell clones or
301 sorted cells was measured by real-time quantitative PCR (qPCR), using purified
302 genomic DNA, iTaq universal SYBR Green Supermix (Bio-Rad Laboratories, Cat.
303 # 1725121) and a StepOnePlus (Applied Biosystems). Relative quantitation
304 methods were used for copy number calculation. Primers used for qPCR are
305 listed in Supplementary Table S1. Mouse genes *Sgk1* and *Hprt1* were used as
306 endogenous controls, assuming four copies of each gene per cell in NIH 3T3.
307 For Figure 3d and Figure 5c, a primer pair (Zeo-GFP2for/rev) that binds to the
308 UBC-GFP-ZeoR region was used to estimate transgene copy number. For Table
309 2, Table 3 and Supplementary Figure S5, in addition to Zeo-GFP2for/rev, 4
310 primer pairs binding to the DHFR BAC or 6 primer pairs binding to the HBB BAC
311 were used to estimate the copy number of DHFR-UG or HBB-UG BAC,
312 respectively. The ΔC_T method (Equations 3 and 5) was used to estimate the copy

313 numbers of the PCR amplification regions on the UBC-GFP-ZeoR reporter gene
314 or on the HBB BAC, and $\Delta\Delta C_T$ method (Equations 4 and 6) was used to estimate
315 the copy numbers of the PCR amplification regions on the DHFR BAC. When
316 multiple primer pairs were used for a region, the mean copy number of all PCR
317 amplification regions was calculated as the copy number of that region.
318 Equations 3 and 7 were used to calculate the fold increase of BAC copy numbers
319 in H1 and H2 samples relative to L.
320

$$\Delta C_T = C_{T\,\text{test region}} - (C_{T\,Sgk1} + C_{T\,Hprt1})/2 \quad (3)$$

$$\Delta\Delta C_T = \Delta C_{T\,\text{transgene clone}} - \Delta C_{T\,\text{NIH 3T3}} \quad (4)$$

$$\text{copy number}_{\Delta C_T} = 4 \times 1.95^{-\Delta C_T} \quad (5)$$

$$\text{copy number}_{\Delta\Delta C_T} = 4 \times 1.95^{-\Delta\Delta C_T} \quad (6)$$

$$\text{BAC fold increase} = 1.95^{\Delta C_{T\,\text{L}} - \Delta C_{T\,\text{H1|H2}}} \quad (7)$$

321

322 **Correlation of reporter gene expression and reporter gene copy number**

323 Mean fluorescence intensity (in arbitrary units) of individual clones were
324 measured by flow cytometry and normalized by fluorescent bead intensity to be
325 used as a measure of reporter gene expression. To ensure uniform
326 normalization for all samples, fluorescent beads from the same batch were used
327 for all measurements. Untransfected cells were used to establish background
328 fluorescence levels. Linear correlations of GFP expression level versus
329 transgene copy number for each group of cell clones were calculated using the

330 linear trend line tool in Microsoft Excel with the y-intercept fixed to 0
331 (autofluorescence normalized by beads was almost 0).
332
333 **DNA FISH probes**
334 Biotin or digoxigenin labeled DNA FISH probes were made from BAC
335 DNA, using a published protocol (68), with the following reagents: AluI, DpnI,
336 HaeIII, MseI, MspI, RsaI (New England Biolabs, Cat. # R0137S, R0176S,
337 R0108S, R0525S, R0106S, R0167S, respectively) and CutSmart Buffer (New
338 England Biolabs); Terminal Deoxynucleotidyl Transferase and reaction buffer
339 (Thermo Fisher Scientific, Cat. # EP0161); dATP (New England Biolabs, Cat. #
340 N0446S) and Biotin-14-dATP (Thermo Fisher Scientific, Cat. # 19524016) for
341 biotin labelling, or dTTP (New England Biolabs, Cat. # N0446S) and Digoxigenin-
342 11-dUTP (MilliporeSigma, Cat. # 11093088910) for digoxigenin labelling.
343
344 **3D DNA FISH**
345 DNA FISH of interphase nuclei used published protocols (69, 70) with
346 small modifications. Cells grown on coverslips (12 mm diameter) were fixed with
347 3-4% paraformaldehyde in Dulbecco's phosphate buffered saline (DPBS, 8 g/l
348 NaCl, 0.2 g/l KCl, 2.16 g/l Na_2HPO_4-$7H_2O$, 0.2 g/l KH_2PO_4) for 10 min, followed
349 by permeabilization with 0.5% Triton X-100 (Thermo Fisher Scientific, Cat. #
350 28314) in DPBS for 10-15 min. Cells were subjected to six freeze-thaw cycles
351 using liquid nitrogen, immersed in 0.1M HCl for 10-15 min, and then washed 3x
352 with 2x saline-sodium citrate (SSC). Freeze-thaw cycles sometimes were

353 skipped with no noticeable difference in FISH signals. Cells were incubated in
354 50% deionized formamide (MilliporeSigma, Cat. # S4117)/2x SSC for 30 min at
355 room temperature (RT), and stored for up to 1 month at 4°C. Each coverslip
356 used ~4 µl hybridization mixture, consisted of 5-20 ng/µl probes, 10x of mouse
357 (for NIH 3T3 cells) or human (for HCT116 cells) Cot-1 DNA (Thermo Fisher
358 Scientific Cat. # 18440016 or 15279011,) per ng probe, 50% deionized
359 formamide, 10% dextran sulfate (MilliporeSigma, Cat. # D8906) and 2x SSC.
360 Cells and probes were denatured together on a heat block at ~76°C for 2-3 min
361 and hybridized at 37°C for 16 hrs-3 days. After hybridization, cells were washed
362 3 x 5 min in 2x SSC at RT, and for 3 x 5 min in 0.1x SSC at 60°C, and then
363 rinsed with SSCT (4x SSC with 0.2% TWEEN 20) at RT. FISH signals were
364 detected by incubation with Alexa Fluor 647 conjugated Strepavidin (1:200;
365 Jackson ImmunoResearch, Cat. # 016-600-084) or Alexa 594 conjugated
366 Strepavidin (1:200; Life Technology, Cat. # S11227) for biotin-labeled probes, or
367 Alexa Fluor 647 conjugated IgG fraction monoclonal mouse anti-digoxin (1:200;
368 Jackson ImmunoResearch, Cat. # 200-602-156) for digoxigenin labeled probes,
369 diluted in SSCT with 1% Bovine Serum Albumin (MilliporeSigma, Cat. # A7906),
370 for 40 min-2 hrs at RT. Coverslips were washed in SSCT for 4 × 5 min, rinsed
371 with 4x SSC and mounted.
372
373 **Mitotic FISH**
374 Metaphase spreads were prepared according to a published protocol (71)
375 with small modifications. Cells grown to 70-80% confluence were incubated with

376 0.1 µg/ml Colcemid (Thermo Fisher Scientific, Cat. # 15212012) in growth media
377 for ~1 hr. Cells were then harvested and swollen by incubation in 0.075 M KCl
378 for 10-20 min at 37°C, followed by fixation with freshly prepared Carnoy's fixative
379 (3:1 v/v ratio of methanol/acetic acid). Chromosomal spreads were made by
380 dropping the fixed swollen cells onto cold wet glass slides. DNA FISH of mitotic
381 spreads was performed using a published protocol (71).
382
383 **Microscopy and image analysis**
384 For examining EGFP-LacI signals cells were grown on coverslips and
385 fixed with 3-4% paraformaldehyde in DPBS before mounting. For examining the
386 expression of the three reporter minigenes, SNAP tagged-Lamin B1, SNAP-
387 tagged Fibrillarin and mCherry-Magoh, the cells were first labeled with cell-
388 permeable substrate SNAP-Cell Fluorescein (New England Biolabs, Cat. #
389 S9107S) overnight at 240 nM concentrations. To reduce background of
390 unreacted SNAP-tag substrate, cells were incubated 3x 30 mins with media in
391 the incubator, washed with PBS, and fixed with freshly prepared 4%
392 paraformaldehyde in PBS for 15 min at RT. All samples- including fixed cells
393 expressing fluorescently tagged transgenes, 3D DNA FISH, and mitotic FISH
394 sample- were mounted with a Mowiol-DABCO anti-fade medium (72) containing
395 ~3 µg/ml DAPI (MilliporeSigma, Cat. # D9542).
396 3D z-stack images were acquired using a Deltavision wide-field
397 microscope (GE Healthcare), equipped with a Xenon lamp, 60X, 1.4 NA oil
398 immersion objective (Olympus) and CoolSNAP HQ CCD camera (Roper

399 Scientific) or a V4 OMX (GE healthcare) microscope, equipped with a 100X, 1.4
400 NA oil immersion objective (Olympus) and two Evolve EMCCDs (Photometrics).
401 Images were deconvolved using the deconvolution algorithm (72) provided by the
402 *softWoRx* software (GE Healthcare). Gamma = 0.5 was applied to green
403 channels in Figure 5e, Supplementary Figure S8 and Supplementary Figure S10
404 for proper display of spots with relatively low signals. All image analysis and
405 preparation were done using Fiji (73). Images were assembled using Illustrator
406 (Adobe), Photoshop (Adobe), or GIMP.
407 For estimation of episome size, the z-sections containing focused episome
408 images for the DAPI and FISH channels were selected manually from the
409 deconvolved z-stack image. Chromosomes and FISH spots were segmented by
410 applying the k-mean clustering algorithm (number of clusters = 3, cluster center
411 tolerance = 0.0001, randomization seed = 48) from the IJ Plugins Toolkit (http://ij-
412 plugins.sourceforge.net/plugins/toolkit.html). The smallest chromosome was
413 identified by manually searching for the chromosome with the smallest area.
414 Segmented FISH spots overlapping or touching chromosomes were removed
415 manually. Integrated DAPI intensities of the smallest chromosome and of the
416 FISH spots not overlapping or touching chromosomes were calculated by
417 Equation 8 (Mean gray value and Area were measured by Fiji). Average
418 episome size was calculated by Equation 9 (n is the number of FISH spots, chro
419 is the smallest chromosome found in the field, 61.4 Mb is the size of chr19 in
420 mm10).
421

$$\text{integrated density} = (\text{Mean gray value} - 200) \times \text{Area} \quad (8)$$

$$\text{episome size} = \frac{\text{integrated density}_{\text{FISH}}}{n \times \text{integrated density}_{\text{chro}}} \times 61.4 \text{ Mb} \quad (9)$$

422

423 Comparison of reporter gene expression levels in for NIH 3T3 cell clones
424 (Figure 7c) was done by projecting deconvolved images stacks and then
425 measuring the integrated intensity within individual nuclei after subtracting
426 background intensity levels measured in the cytoplasm. Regions of interest
427 circumscribing individual nuclei were drawn manually based on the SNAP-lamin
428 B1 signal. Linear correlations of the integrated intensities of the nuclear SNAP-
429 tag and mCherry signals were calculated using Microsoft Excel with the y-
430 intercept fixed to 0.

431 A non-linear Gamma correction (0.7) to reduce the grey-scale dynamic
432 range followed by a maximum intensity projection of 3-4 z-sections was used to
433 better visualize both lamin and nucleolar staining simultaneously (Figure 7d).

434

435 **Agarose embedded DNA preparation and S1 Nuclease digestion**

436 Agarose embedded DNA was prepared according to published protocols
437 (74, 75) with modifications. To prepare mammalian cell suspensions, cells were
438 grown without selection for 3-4 days after passaging, reaching 80%-90%
439 confluence. Cells were trypsinized, resuspended in cell media, washed with
440 PBS, and resuspended in PBS at a concentration of ~8 x 10^6 cells / 100 μl. To
441 prepare *E. coli* cell suspensions, ~0.1 ml of overnight culture was diluted in 15 ml
442 fresh LB and grown to an OD_{600} of ~1. Cells were washed with L Buffer (10 mM

443 Tris-Cl pH7.6, 20 mM NaCl, 100 mM EDTA) once and resuspended in L Buffer at
444 a concentration of ~10^9/100 μl, assuming a cell concentration of ~8 x 10^7/100 μl
445 at an OD_{600} of 1.
446 2% certified low melt agarose (Bio-Rad Laboratories, Cat. # 1613111) was
447 prepared with L Buffer and kept at 75°C. Equal volumes of the cell suspension
448 (RT) and the agarose solution (75°C) were mixed and immediately transferred to
449 plug molds (Bio-Rad Laboratories, Cat. # 1703713), ~100 μl mixture per plug.
450 The agarose plugs were incubated in L Buffer with 1% Sarcosyl (MilliporeSigma,
451 Cat. # L5125) and 0.5 mg/ml proteinase K at 55°C for 1-2 days. The agarose
452 plugs were washed with W Buffer (20 mM Tris-Cl, pH7.6, 50 mM EDTA) for 2 x
453 15 min, incubated in 1 mM PMSF in W Buffer for 30 min, and washed with W
454 Buffer again. Prepared agarose plugs were stored in 0.5 M EDTA at 4°C before
455 use.
456 For S1 Nuclease (Promega, Cat. # M5761) digestion, agarose plugs were
457 first washed in TE (10 mM Tris-Cl, 1 mM EDTA, pH 7.6) for 3 x 10 min and in 1x
458 S1 Nuclease Buffer for 20 min on ice. The agarose plugs were then digested
459 with 1-16 U/0.4 ml S1 Nuclease in 1x S1 Nuclease Buffer at 37°C for 45 min.
460 The reaction was stopped by washing the agarose plugs with 0.5 M EDTA or W
461 Buffer.
462
463 **Pulsed Field Gel Electrophoresis (PFGE)**
464 PFGE was performed using a CHEF-DR III (Bio-Rad Laboratories)
465 according to the manufacturer's manual using a 1% certified megabase agarose

466 (Bio-Rad Laboratories, Cat. # 1613108) gel in 0.5x Tris-borate-EDTA buffer
467 (TBE), a 0.5x TBE running buffer, and the following parameters: voltage = 6
468 V/cm, angle = 120°, pulse = 60-120 sec, temperature = 14 °C, run time = 20 or
469 24 hrs (stopped at 18-20 hrs). Yeast chromosomes (Bio-Rad Laboratories, Cat.
470 # 170-3605) were used as DNA size markers.
471
472 **Southern hybridization probes**
473 Southern hybridization probes were created and labeled with digoxigenin
474 by PCR using primers listed in Supplementary Table S1. Set 1 contains a 620bp
475 and a 615 bp fragment amplified from the GFP-ZeoR region; Set 2 contains 525
476 bp, 534 bp, and 504bp fragments amplified from the BAC vector region; Set 3
477 contains 446 bp, 681 bp, and 424 bp fragments amplified from the HBB BAC.
478 Pooled Set 1 and Set 2 fragments were used for detecting the DHFR BAC, and
479 pooled Set 1 and Set 3 for detecting the HBB BAC. PCR was done using *Taq*
480 DNA polymerase (New England Biolabs, Cat. # M0267L) with the following
481 recipe: 1x ThermoPol Buffer, 0.2 mM dATP/dCTP/dGTP (New England Biolabs,
482 Cat. # N0446S), 0.165 mM dTTP (New England Biolabs, Cat. # N0446S), 0.035
483 mM Digoxigenin-11-dUTP, 0.5 ng HBB BAC, 1.25 U *Taq* DNA polymerase, 0.5
484 μM forward/reverse primers, 50 μl total reaction volume. PCR products were
485 column (QIAGEN, Cat. # 28104) purified. Pooled probes were denatured in
486 nuclease free water, at ~100°C for ~10 min and snap-chilled on ice before use.
487
488 **Southern hybridization**

489 Southern blotting used a published protocol (76) with modifications. After
490 ethidium bromide staining and imaging, the gel was depurinated in 0.25 M HCl
491 for 2x 30 min, denatured in 0.4 M NaOH for 2x 25 min, neutralized in 0.5 M Tris-
492 Cl/1.5 M NaCl (pH 7.6) for 2x 20 min and washed in 2x SSC for 2x 20 min. DNA
493 was transferred to Zeta-Probe membranes (Bio-Rad Laboratories, Cat. #
494 1620165) using a Model 785 Vacuum Blotter (Bio-Rad Laboratories), with 2x
495 SSC as transfer buffer, ~5 inches Hg pressure, and ~16 hrs transfer time. A
496 Stratalinker (Strategene) was used to cross-link DNA to the membrane.
497 Hybridization used a standard protocol (77) with modifications. The
498 hybridization buffer was composed of 1:1 volumes of 1 M Na_2HPO_4 (pH 7.2) and
499 14% (w/v) SDS. Total concentration of pooled probes was ~100 ng/ml.
500 Hybridization was carried out at 65°C for ~16 hrs. After hybridization, the
501 membrane was washed with 2x SSC/0.1% SDS for 2 x 5 min at room
502 temperature, and with 1x SSC/0.1% SDS for 2 x 10 min at 65°C and rinsed with
503 2x SSC. Signals were detected using the DIG Nucleic Acid Detection Kit
504 (MilliporeSigma, Cat. # 000000011175041910) according to the manufacturer'
505 manual, except that in the final step, CDP-*Star* (MilliporeSigma, Cat. #
506 11685627001) was used instead of NBT/BCIP, and the membrane was imaged
507 by an iBright system (Thermo Fisher Scientific).
508
509 **Estimation of average BAC DNA content per episome**
510 To estimate the average BAC DNA content per episome of clone DHFR-
511 UG-s3 and clone HBB-UG-100d3, cells at the same passage were seeded on

512 glass coverslips for DNA FISH using BAC probes, and in different plates for
513 genomic DNA extraction followed by qPCR. The mean number of FISH spots
514 per nucleus, counted from z-stack projected images, provided the average
515 episome copy number per cell. For the DHFR-UG clone, 3 was subtracted from
516 the mean number of FISH spots, as the parental NIH 3T3 cells had ~3 FISH
517 spots, corresponding to the endogenous DHFR loci, using FISH probes prepared
518 from the DHFR BAC. qPCR estimation of BAC copy number per cell was
519 described in section "Estimation of transgene copy number". BAC DNA content
520 per episome was calculated using equation 10.
521

$$\text{BAC content per episome} = \frac{\text{BAC copy number per cell}}{\text{episome copy number per cell}} \times \text{BAC size} \qquad 10$$

522

523 **Whole genome sequencing**

524 Clone DHFR-UG-s3 and clone HBB-UG-100d3 were sorted by flow
525 cytometry using the H1, H2 and L sorting windows shown in Figure 6c and
526 Supplementary Figure S4a. Cells from the H2 and L regions were sorted in the
527 same experiment, while cells from the H1 regions were sorted in another
528 experiment. 100-200 thousand cells were collected from each window. Genomic
529 DNA from sorted cells was isolated by phenol-chloroform extraction. To prepare
530 sequencing libraries, genomic DNA was first fragmented to 100-500 bp by
531 sonication using a Bioruptor Pico (Diagenode), with the following conditions: 4
532 ng/μl DNA in 120 μl EB, 1.5 ml tube, 10-11 cycles of 30 secs on and 30 secs off.
533 Next, indexed adaptors was attached to the fragmented DNA using True-Seq

534 ChIP Sample Preparation kit (Illumina, Cat. # IP-202-1012) according to the
535 manufacturer's instructions with the following modifications: after the fragmented
536 DNA was end repaired, 3' end adenylated, and ligated to indexed adaptors
537 without size selection, the ligation products were PCR amplified for 7~9 cycles.
538 Libraries were quality checked on a Fragment Analyzer (Agilent) and quantitated
539 by qPCR. Every 6 libraries were pooled at equal molar ratios and sequenced on
540 one lane using a HiSeq 4000 for 101 cycles from one end of the fragments using
541 a HiSeq 4000 sequencing kit version 1. Fastq files were generated and de-
542 multiplexed with the bcl2fastq v2.20 Conversion Software (Illumina). Library
543 quality checking, quantitation and sequencing, and fastq file generation and de-
544 multiplexing were done by the DNA services lab, Roy J. Carver Biotechnology
545 Center, UIUC. 59-65 million reads with quality score >30 were obtained for each
546 library.
547
548 **Sequencing reads processing and copy number variation analysis**
549 Low quality bases and adaptor sequences were trimmed from raw reads
550 using cutadapt 1.14 with Python 2.7.13 with the following parameters: -a
551 AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -q 20,20 -m 20, resulting
552 in ~0.2% bp being trimmed. Reads were then aligned to a reference genome
553 (mm10 plus HBB BAC (CTD-2643I7, sequence from hg38), the BAC vector
554 (pBelo11, GenBank Accession #: U51113) and UBC-GFP-ZeoR, each as an
555 individual chromosome) using Bowtie2 (version 2.3.2) with default parameters.
556 Overall alignment rate of each sample was ~98%-99%. Finally, PCR duplicates

557 were removed by SAMtools rmdup (version 1.7) with default parameters,
558 resulting in 42-48 million total mapped reads in each sample.
559 For reads binning, each chromosome of the reference genome was
560 divided into non-overlapping 3 kb or 30 kb bins; the number of alignments with
561 centers falling into each bin (binned reads) was counted and then divided by the
562 mean read count (Equation 11), generating normalized binned reads (normalized
563 reads, Equation 12), and finally the normalized binned reads of the test sample
564 (H1 or H2 cells) were divided by that of the reference sample (L cells), generating
565 the ratio of normalized binned reads (ratio, Equation 13). The mean read count
566 was ~50 or ~500 for 3 kb or 30 kb bin size, respectively. To reduce noise caused
567 by extremely low read counts, a threshold for determining outliers was calculated
568 based on the quantile range (Equation 14). Bins with $\log_2$(reads) smaller than
569 the threshold in the test sample were removed from further analysis. The
570 excluded bins took up ~6.0% of total bins for both 3 kb and 30 kb bin sizes,
571 including zero read count bins, which took up ~5.5% or ~3.5% of total bins for 3
572 kb or 30 kb bin size, respectively. The maximum number of reads of the
573 excluded bins were ~7 or ~108 for 3 kb or 30 kb bin size, respectively.
574

$$\text{mean read count} = \frac{\text{total mapped reads} \times \text{bin size}}{\text{reference genome size}} \tag{11}$$

$$\log_2(\text{normalized reads}) = \begin{cases} \log_2 \dfrac{\text{binned reads}}{\text{mean read count}}, & \text{binned reads} > 0 \\ \log_2 \dfrac{0.1}{\text{mean read count}}, & \text{binned reads} = 0 \end{cases} \tag{12}$$

$$\log_2(\text{ratio}) = \log_2 \frac{\text{reads}_{\text{H1|H2}}}{\text{reads}_{\text{L}}} \quad (13)$$

$$\text{outlier threshold} = 25\% \text{ quantile} - 4 \times (75\% \text{ quantile} - 25\% \text{ quantile}) \quad (14)$$

A circular binary segmentation algorithm (78, 79) from the R-package DNAcopy (version 1.52.0) was used to merge bins with similar $\log_2$(ratio) into segments, with the following parameters for the segment function: verbose = 1, undo.splits="sdundo", undo.SD=1. The mean $\log_2$(ratio) of each segment was calculated for identifying episome-localizing regions.

To identify possible episome-localizing regions, we first measured BAC transgene copy numbers in the H1, H2 and L samples by qPCR and then calculated the theoretical episome copy numbers using the estimated BAC copy number per episome of unsorted cells (Table 2). The minimum copy number increase of episome-localizing host DNA (minimum increase) was then calculated assuming NIH 3T3 to be tetraploid and each episome to have the same host DNA sequence (Equation 15). Segments with mean $\log_2$(ratio) equal to or greater than $\log_2$(minimum increase) in both H1 and H2 samples were selected as candidate episome-localizing regions.

$$\text{minimum increase}_{\text{H1|H2}} = \frac{\text{episome copy number}_{\text{H1|H2}} + 4}{\text{episome copy number}_{\text{L}} + 4} \quad (15)$$

Construction of DHFR BAC deletions

593 We tested several DHFR BAC deletions- made for other purposes- for their
594 ability to produce episomes. The DHFR-c27 BAC (51) containing a 256-mer Lac
595 operator (LacO) repeats and a CMV-mRFP-SV40-ZeoR expression cassette was
596 derived from the DHFR BAC, and was used for making the DHFR BAC deletions.
597 DHFR-c27d2 contains a ~70 kb deletion of the 3' part of the *Msh3* gene. DHFR-
598 c27d3-crz contains a ~80 kb deletion of the whole *Dhfr* gene and the 5' part of the
599 *Msh3* gene, including the CMV-mRFP-SV40-ZeoR expression cassette inserted in
600 the *Msh3* gene, and contains a new CMV-mRFP-SV40-ZeoR expression cassette
601 introduced at the remaining part of *Msh3* gene. DHFR-c27d4 contains a ~20 kb
602 deletion around the divergent promoter region. λ Red-mediated BAC
603 recombineering with a *galK*-based dual-selection scheme was used to create the
604 deletion BACs from the DHFR-c27 BAC, as described in "Construction of dual
605 reporter DHFR BACs" and "Construction of BACs containing the UBC-GFP-ZeoR
606 cassette". DNA fragments containing either GalK or FRT-GalK-FRT and homology
607 ends were produced by PCR using either pGalK or pUGG as templates. For
608 DHFR-c27d2, the GalK cassette was introduced by the first round of recombination
609 and was subsequently removed by another round of recombination using a DNA
610 fragment created by a pair of partially overlapping primers. For DHFR-c27d3 and
611 DHFR-c27d4, the FRT-GalK-FRT cassette was introduced in the first round
612 instead and was subsequently removed by inducing FLP recombinase as
613 described in "Construction of BACs containing the UBC-GFP-ZeoR cassette". To
614 create DHFR-c27d3crz, the CMV-mRFP-SV40-ZeoR cassette was introduced into
615 DHFR-c27d3 by one round of recombination using Zeocin resistance as positive

616 selection as described in “Construction of dual reporter DHFR BACs”. Each round
617 of recombination was validated by PCR and restriction enzyme fingerprinting. All
618 primers are listed in Supplementary Table S1.
619
620
621 **Construction of multi-reporter DHFR BAC by BAC-MAGIC**
622 **Overview:** Construction of the 3-reporter BAC was done by serially
623 inserting ~10-15 kb DNA cassettes into the DHFR BAC scaffold by BAC
624 recombineering. These DNA cassettes were constructed from two different DNA
625 plasmid module types: reporter modules and intervening DHFR sequence
626 modules. DNA cassettes were inserted sequentially into the DHFR BAC using
627 multiple rounds of BAC recombineering and positive selection with one of two
628 different positive selectable markers. After insertion of the first DNA cassette,
629 each subsequent insertion of the next DNA cassette removed the preceding
630 positive selectable located at the 3’ end of the preceding cassette while inserting
631 the alternative selectable marker located at the 3’ end of the new cassette.
632 Three reporter gene modules (Rep Mod 01, 02, 03) plus three intervening DHFR
633 sequence modules (DHFR 02, 03, 04) were constructed and then inserted into
634 the DHFR BAC using 6 sequential rounds of BAC recombineering. In this way,
635 45 kb of the original DHFR BAC effectively was reconstructed such that the
636 original DHFR sequences were retained but the 3 reporter mini-genes were
637 inserted into this BAC region with each reporter minigene spaced by ~10 kb of

638 DHFR sequence. We call this overall construction approach BAC-MAGIC (**BAC**-
639 **M**odular **A**ssembly of **G**enomic loci **I**nterspersed **C**assettes).
640 Each DNA cassette was constructed using traditional cloning methods,
641 Gibson assembly (80), and/or DNA Assembler (81, 82). Three reporter recipient
642 modules (pRM01-Spec, pRM02-Spec, and pRM03-Spec) were designed to
643 incorporate a rare AgeI restriction site for insertion of reporter expression
644 cassettes of choice, in order to create the final reporter modules for BAC
645 recombineering. Unless mentioned specifically all the enzymes were procured
646 from New England Biolabs. All primers and oligos are listed in Supplementary
647 Table S1. Gibson assembly used Gibson assembly cloning kit (New England
648 Biolabs, Cat. # E5510S) as per the manufacturer's instructions.
649 DNA Assembler used *Saccharomyces cerevisiae (S. cerevisiae)* strain
650 VL6-48N (MATα, his3-Δ200, trp1-Δ1, ura3-Δ1, lys2, ade2-101, met14, cir°),
651 transformed with 43 fmol pRS413 vector backbone and 130 fmol of all other
652 fragments using the LiAc/SS carrier DNA/PEG method (83). The *S. cerevisiae*
653 single-copy shuttle vector pRS413 contains CEN6/ARS autonomously replicating
654 sequence, auxotrophic selection marker *HIS3* for propagation in yeast, and
655 pMB1 origin of replication and *bla* (Ap^R) marker for selection with ampicillin in *E.*
656 *coli*. The 3.8 kb pRS413 vector backbone was PCR amplified from plasmid
657 pRS413 (New England Biolabs) using primer pair RS413-Fw/RS413-Rev for all
658 yeast assembly reactions. The vector backbone and all other fragments made by
659 PCR were digested with DpnI to remove template DNA. Transformants were
660 selected on SC selection media plates lacking histidine [0.17% Bacto-yeast

661 nitrogen base without amino acids (MilliporeSigma, Cat. # Y1251-100G), 0.5%
662 ammonium sulfate, 2% D-glucose, 0.2% Dropout mix (MilliporeSigma, Cat. #
663 Y2001-20G), 2% agar, 80 mg/l uracil, 80 mg/l L-tryptophan, and 240 mg/l L-
664 leucine] at 30°C for 3-4 days. Plasmid DNA were prepared using QIAprep Spin
665 Miniprep Kit (Qiagen, Cat. # 27104) and screened by restriction enzyme
666 fingerprinting. Plasmid DNA from selected yeast colonies was introduced into *E.*
667 *coli* strain DH5α and isolated plasmid DNA then further validated by additional
668 restriction enzyme fingerprinting.

669 Below we describe construction of each reporter and intervening spacer
670 modules and BAC recombineering assembly of these modules to create the 3-
671 reporter BAC. ApE (M. Wayne Davis, University of Utah,
672 http://biologylabs.utah.edu/jorgensen/wayned/ape/) and SnapGene (from GSL
673 Biotech; available at snapgene.com) programs were used to analyze sequence
674 data, design primers, and design cloning strategies.

675 **Construction of plasmid pRM01-RSLB1-Spec (Reporter module 01):**
676 Plasmid pRM01 was made by sequential addition of two DHFR homology
677 regions to plasmid pEGFP-C1 (Clontech). First, the 2.1 kb DHFR homology
678 region (M1F4) was PCR amplified from the DHFR BAC using primer pair M1F4-
679 BamHIfor/M1F4-AgeIrev, double digested with BamHI/AgeI, and ligated with the
680 BamHI/AgeI digested pEGFP-C1 to generate intermediate plasmid pEG-Rep-
681 Module-1a. Next, the 2.0 kb DHFR homology region (M2F12) was PCR
682 amplified from the DHFR BAC using primer pair M2F12-AgeIFor/M2F12-PshRev,

683 double digested with AgeI/PshAI and ligated with the AgeI/SnaBI digested
684 plasmid pEG-Rep-Module-1a to produce plasmid pRM01.
685 To create plasmid pRM01-Spec (Reporter recipient module 01), a 1.6 kb
686 Spectinomycin resistance gene expression cassette (SpecR), derived from
687 plasmid pYES1L (Thermo Fisher Scientific), was inserted into pRM01, 400 bp
688 upstream of the 3' end of the M2F12 DHFR homology region by two-fragment
689 Gibson Assembly (80). The two fragments for Gibson assembly were PCR
690 amplified from pRM01 using primer pair GA-RM01-Spec-For/ GA-RM01-Spec-
691 Rev (PCR product size: 7.8 kb), or from pYES1L using primer pair
692 Specfor/SpecRev (PCR product size: 1.6 kb) respectively.
693 The pRSLB1 (hRPL32-SNAP-Lamin B1) plasmid harboring SNAP-tagged
694 Lamin B1 reporter expression cassette (RSLB1) was constructed by three-
695 fragment Gibson assembly. pEGFP-Lamin B1 plasmid vector backbone 5.3 kb
696 fragment was prepared by AseI/BsrGI double digestion. The hRPL32 promoter
697 (2.2 kb) and SNAP tag (561 bp) fragments were PCR amplified using primer
698 pairs GA-hRPL32-fwd/GA-hRPL32-rev (template plasmid pMOD-HB2-hRPL32-
699 RZ, made in this study), and GA-SNAP-fwd/GA-SNAP-rev (template plasmid
700 pSNAPf, New England Biolabs).
701 pRM01-Spec was linearized by AgeI and simultaneously
702 dephosphorylated by Shrimp Alkaline Phosphatase (New England Biolabs, Cat. #
703 M0371S). The RSLB1 expression cassette was PCR amplified from plasmid
704 pRSLB1 using primer pair R32CerLBAgeIfor/newPCFAgeIrev (PCR product size:
705 4.9 kb) and double digested with DpnI/AgeI. The linearized pRM01-Spec and the

706 digested RSLB1 PCR product were ligated to produce plasmid pRM01-RSLB1-
707 Spec, which was digested with AseI to produce the final BAC recombineering
708 10.3 kb targeting construct.

709 **Construction of plasmid pRM02-PSF-Spec (Reporter module 02):**
710 Plasmid pRM02 was made using similar cloning steps used to produce pRM01
711 except two different DHFR homology regions were added to pEGFP-C1: 2.0 kb
712 PCR product M2F4 (primer pair M2F4-BamHIfor/M2F4-AgeIrev) replaced M1F4
713 and 2.0 kb PCR product M3F1 (primer pair M3F1-AgeIFor/M3F1-PshRev)
714 replaced M2F12. Plasmid pRM02-Spec was made the same way as pRM01-
715 Spec except that fragment 1 for Gibson assembly was PCR amplified from
716 plasmid pRM02 using primer pair GA-RM02-Spec-For/GA-RM02-Spec-Rev
717 (PCR product size: 7.8 kb). The final plasmid pRM02-Spec (pRep-module 02-
718 Spec) is Reporter recipient module 02 for the SNAP-tagged Fibrillarin reporter
719 expression cassette (PSF).

720 To create plasmid pPSF (pPPIA-SNAP-Fibrillarin), the GFP cassette
721 between KpnI/HpaI restriction sites of plasmid GFP-Fibrillarin was replaced with
722 a 730 bp Cerulean cassette PCR amplified from plasmid pCerulean-N1 (New
723 England Biolabs) using primer pair ForCerFib/RevCerFib, resulting in an
724 intermediate plasmid pPCF. Next, the CMV promoter between SnaBI/HindIII
725 sites of pPCF was replaced with the 2.8 kb PPIA promoter PCR amplified from
726 plasmid p[MOD-HB2-PPIA-RZ] (made in this study) using primer pair
727 PPIACerFibFor/ PPIACerFibRev, resulting in plasmid pPPIA-Cer-Fib. Finally, the
728 720 bp Cerulean cassette between the AgeI/HpaI sites of pPPIA-Cer-Fib was

729 replaced with a 560 bp SNAP tag fragment PCR amplified from plasmid pSNAPf
730 (New England Biolabs) using primer pair Snap-XmaI-For/Snap-HpaI-Fib-Rev and
731 double digested with XmaI/HpaI, producing pPSF.
732 pRM02-Spec was linearized by AgeI and simultaneously
733 dephosphorylated by Shrimp Alkaline Phosphatase (New England Biolabs, Cat. #
734 M0371S). The 4.6 kb PSF expression cassette was PCR amplified from pPSF
735 using primer pair PSF-AgeI-For/ PSF-AgeI-Rev and double digested with
736 DpnI/AgeI. Their ligation produced plasmid pRM02-PSF-Spec, which provided
737 the 10.4 kb BAC recombineering targeting construct after BamHI/AatII/RsrII triple
738 digestion of pRM02-PSF-Spec.
739 **Construction of plasmid pRM03-PCM-Spec (Reporter module 03)**:
740 Plasmid pRM03 was made using similar cloning steps used to produce pRM01
741 except two different DHFR homology regions were added to pEGFP-C1: 2.1 kb
742 PCR fragment M3F4 (primer pair M3F4-BamHIfor/M3F4-AgeIrev) replaced M1F4
743 and 2.1 PCR fragment M4F1 (primer pair M4F1-AgeIFor/M4F1-PshRev)
744 replaced M2F12. Plasmid pRM03-Spec was made the same way as pRM01-
745 Spec except that fragment 1 for Gibson assembly was PCR amplified from
746 plasmid pRM03 using using primer pair GA-RM03-Spec-For/ GA-RM03-Spec-
747 Rev (PCR product size: 7.8 kb). The final plasmid pRM03-Spec (pRep-module
748 03-Spec) is Reporter recipient module 03 for the mCherry-tagged Magoh reporter
749 expression cassette (PCM).
750 Plasmid pPCM (pPPIA-mCherry-Magoh) was created in two steps. First,
751 the CMV promoter between the NdeI/NheI sites of plasmid pmRFP-Magoh was

752 replaced with the PPIA promoter (2.8 kb), PCR amplified from plasmid pMOD-
753 HB2-PPIA-RZ using primer pair PPIA-Magohfor/ PPIA-MagohRev and double
754 digested with NdeI/NheI, resulting in intermediate plasmid pPMM. Next, the
755 mRFP tag between the NheI/HindIII sites of pPMM was replaced with a 720 bp
756 mCherry tag PCR amplified from plasmid pQCXIN-TetR-mCherry using primer
757 pair mCherry-NheI-Magoh-For/mCherry-H3-Magoh-Rev, resulting in plasmid
758 pPCM.

759 To create plasmid pRM03-PCM-Spec (Reporter module 03), plasmid
760 pRM03-Spec was linearized by AgeI and simultaneously dephosphorylated by
761 Shrimp Alkaline Phosphatase (New England Biolabs, Cat. # M0371S). A 4.2 kb
762 PCM expression cassette was PCR amplified form plasmid pPCM using primer
763 pair MMorCF-AgeIfor/newPCFAgeIrev and double digested with DpnI/AgeI.
764 pRM03-Spec and the PCM PCR product were ligated, producing plasmid
765 pRM03-PCM-Spec, which was used as a template for PCR amplification using
766 primer pair M3F4-PCR-Fw/M4F1-PCR-Rev to produce the 9.9 kb BAC
767 recombineering target. After PCR, any remaining template plasmid was digested
768 with DpnI.

769 **Construction of plasmid pRS413-DHFR-Mod-02-Kan (Intervening**
770 **DHFR module 02):** Plasmid pRS413-DHFR-Mod-02 was made by assembling
771 the vector backbone with four additional fragments using the DNA assembler
772 method (81, 82). Fragment 5'-DHM2 (4.3 kb) and fragment 3'-DHM2 (6.3 kb)
773 with an overlap of 659 bp and were both PCR amplified from the DHFR BAC,
774 using primer pair M2F12-AgeIfor/M2F1rev or DHM2-Seq2/M2F4-AgeIrev,

775 respectively. Two bridging oligomers, with a 125 bp homology to the pRS413
776 vector backbone, and a 125 bp homology to fragment 5'-DHM2 (oligo M2F1-
777 pRS413) or fragment 3'-DHM2 (oligo M2F4-pRS413) were synthesized at
778 Integrated DNA Technologies, Inc. The final Intervening DHFR module 02,
779 plasmid pRS413-DHFR-Mod-02-Kan, was created by ligating a 2.4 kb Kan/NeoR
780 cassette derived from DraI digestion of plasmid pEGFP-C1, with the plasmid
781 pRS413-DHFR-Module-02 linearized by DraIII and blunted by DNA Polymerase
782 I, Large (Klenow) Fragment.
783 For BAC recombineering an 11.7 kb of targeting construct was amplified
784 from plasmid pRS413-DHFR-Mod-02-Kan using primer pair M2F12-
785 AgeIFor/DH2-4rev and purified by gel extraction after DpnI digestion of the
786 template plasmid.
787 **Construction of plasmid pRS413-DHFR-Mod-03-Kan (Intervening**
788 **DHFR module 03)**: Plasmid pRS413-DHFR-Mod-03 was made by assembling
789 the vector backbone with four additional fragments using the yeast DNA
790 assembler method. Fragment 5'-DHM3 (6.5 kb) and fragment 3'-DHM3 (5.0 kb)
791 with an overlap of 1553 bp were both PCR amplified from the DHFR BAC using
792 primer pair M3F1-AgeIFor/M3F3-BamHIrev or M3-F3For/M3F4-AgeIRev,
793 respectively. Two bridging oligomers, with a 125 bp homology to the pRS413
794 vector backbone, and a 125 bp homology to fragment 5'-DHM3 (oligo M3F1-
795 pRS413) or to fragment 3'-DHM3 (oligo M3F4-pRS413), respectively, were
796 synthesized at Integrated DNA Technologies, Inc. The final Intervening DHFR
797 module 03, plasmid pRS413-DHFR-Mod-03-Kan, was created by ligating a 2.4

798 kb Kan/NeoR cassette derived from DraI digestion of plasmid pEGFP-C1, with
799 the plasmid pRS413-DHFR-Mod-03 linearized by SmaI.
800 For BAC recombineering a 12.2 kb targeting construct was amplified from
801 plasmid pRS413-DHFR-Mod-03-Kan using primer pair DH3-1for/DH3-4rev and
802 purified by gel extraction after DpnI digestion of the template plasmid.
803 **Construction of plasmid pRS413-DHFR-Mod-04-Zeo (Intervening**
804 **DHFR module 04):** Plasmid pRS413-DHFR-Mod-04 was made by assembling
805 the vector backbone plus 5 additional fragments using the yeast DNA assembler
806 method (4). Fragment 5'-DHM4 (4.9 kb), fragment Mid-DHM4 (5.2 kb) and
807 fragment 3'-DHM4 (5.2 kb) with an overlap of 2663 bp in between 5'-DHM4 and
808 Mid-DHM4, and an overlap of 2542 bp in between Mid-DHM4 and 3'-DHM4,
809 were PCR amplified from the DHFR BAC using primer pair M4F1-
810 AgeIfor/DHM4F2-R, DHM4F2-Fw/DHM4F3-R, or Fw-M4F2-BamHI/RevM4F5-
811 MluI, respectively. Two bridging oligomers, with a 125 bp homology to pRS413
812 vector backbone, and a 125 bp homology to fragment 5'-DHM4 (oligo M4F1-
813 pRS413), or to fragment 3'-DHM4 (oligo M4F5-pRS413), were synthesized at
814 Integrated DNA Technologies, Inc. The final Intervening DHFR module 04,
815 plasmid pRS413-DHFR-Mod-04-Zeo, was created by ligating a 1.1 kb ZeoR
816 expression cassette PCR amplified from plasmid pSV40/Zeo2 (ThermoFisher
817 Scientific) using 5' phosphorylated primer pair ZeoMluIFor/ZeoMluIRev, with the
818 plasmid pRS413-DHFR-Module-04 linearized by BmgBI.

819 For BAC recombineering an 11.6 kb targeting construct was excised out
820 from plasmid pRS413-DHFR-Mod-04-Zeo using KpnI/DrdI restriction enzymes
821 and gel purified.
822 **Assembly of modules to create multi-reporter DHFR BAC:** The six
823 targeting constructs derived from the three reporter modules and the three
824 intervening DHFR modules were incorporated into the DHFR BAC by BAC
825 recombineering, with the following order: Reporter module 01, Intervening DHFR
826 module 02, Reporter module 02, Intervening DHFR module 03, Reporter module
827 03 and Intervening DHFR module 04. *E. coli* strain SW102 was used for BAC
828 recombineering. Each round of BAC recombineering used a corresponding
829 antibiotic (50 μg/ml Kanamycin, 50 μg/ml Spectinomycin, or 25 μg/ml Zeocin) as
830 positive selection for incorporation of the current targeting construct as described
831 in section "Construction of dual reporter DHFR BACs". In the second to the last
832 round of BAC recombineering, colonies were further screened for loss of the
833 antibiotic resistance gene incorporated in the previous round of BAC
834 recombineering by streaking colonies onto a plate containing the corresponding
835 antibiotic. Each round of recombination was validated by restriction enzyme
836 fingerprinting.
837
838

839 **RESULTS**

840 **Overview of BAC TG-EMBED toolkit development:**

841 We previously demonstrated the feasibility of the BAC TG-EMBED
842 approach using both the DHFR BAC (51) and a BAC containing the human
843 GAPDH gene locus (GAPDH BAC) (54). We set out to extend this BAC TG-
844 EMBED methodology in two new directions (Figure 1).

845 First, to better control transgene expression and to be able to express
846 multiple transgenes at reproducible expression ratios, we explored a set of
847 constitutive promoters with various strengths for transgene expression. A
848 previous similar survey of promoters within BAC scaffolds focused only on strong
849 promoters (53). Moreover this survey compared average expression in pools of
850 cell colonies containing different copy-number BAC insertions (53). Here we
851 used a two-reporter, single-cell ratio assay and also examined promoters with a
852 wide range of promoter strengths. Testing each promoter with each BAC
853 scaffold would have generated too large a number of possible combinations. We
854 therefore decided to test a number of different promoters with the original DHFR
855 BAC.

856 Second, we used one specific reporter gene construct to survey the effect
857 of different BAC scaffolds on reporter gene expression. Previous similar
858 applications used BAC scaffolds containing multiple endogenous genes which
859 would also be expressed in addition to added transgenes (51–53). Moreover, in
860 a previous, similar application, different strong promoters were tested by insertion
861 into the exon of an active BAC gene (53). Here we compared BAC scaffolds

862 containing expressed genes with BAC scaffolds from gene deserts or regions
863 containing silenced genes. We assayed the level, stability, and reproducibility of
864 the embedded reporter gene expression when inserted into different BAC
865 scaffolds to identify optimal BAC scaffolds for the BAC TG-EMBED system.
866
867 **A toolset of 7 endogenous promoters for tuning relative transgene**
868 **expression levels**
869 We selected 7 endogenous promoters to test, either because of their
870 known ability and use to drive transgene expression in a range of cell types
871 (EEF1α, UBC) (60, 84–86), or because these promoters were from
872 housekeeping genes (RPL32, PPIA, B2M, RPS3A, GUSB) known to be
873 expressed uniformly across a wide range of tissue types (87–90). We amplified
874 1-3 kb of regulatory regions upstream of the transcription start sites of these
875 genes using either human genomic DNA as a template or, for the UBC promoter,
876 using the pUGG plasmid (54).
877 To assay relative promoter strength, we used the two-minigene reporter
878 system developed in our previous study in which we compared expression of
879 CMV-driven EGFP and mRFP minigenes inserted in the same mouse DHFR
880 BAC scaffold (51). We previously showed that the mRFP minigene reporter
881 expression varied less than or equal to 2.4-fold when the mRFP reporter was
882 inserted at 6 different positions ranging 3-80 kb away from the EGFP reporter
883 gene location on the same BAC (51). To compare relative promoter strengths,
884 we fixed the insertion positions of mRFP and EGFP, and measured the relative

885 fluorescence levels of mRFP and EGFP when they were both driven by the CMV
886 promoter versus when the mRFP reporter was driven by an endogenous
887 promoter (Figure 1). Thus our assay measured the strength of different
888 endogenous promoters relative to the viral CMV promoter, while also measuring
889 the variation in this relative strength in different cells of a mixed clonal population.
890 For this assay, the EGFP reporter minigene was inserted 26kb
891 downstream of the Msh3 transcription start site (51) (Figure 2a). PCR-amplified
892 promoters from 7 different housekeeping genes were cloned upstream of the
893 mRFP expression cassette (Figure 2b), and then this mRFP expression cassette
894 was introduced 121 kb downstream of the Msh3 transcription start site by BAC
895 recombineering (Figure 2a), generating the dual reporter DHFR BAC. As a
896 control, we used the dual reporter BAC previously constructed (51) in which the
897 same mRFP cassette driven by the CMV promoter was inserted at this same
898 location 121 kb downstream of the Msh3 start site.
899 Mouse NIH 3T3 fibroblasts were then stably transfected with these
900 modified BAC constructs. After dual selection with G418 and Zeocin for two
901 weeks, mixed populations of stable clones carrying the BAC transgenes were
902 analyzed by flow cytometry to measure the relative expression ratio of mRFP and
903 EGFP (Figure 2c). Fluorescent beads were used as an invariant fluorescence
904 standard to calibrate the flow cytometer intensity outputs. The ratio of mRFP to
905 EGFP expression was then normalized by the ratio observed with the original
906 dual-reporter BAC construct in which both reporters were driven by CMV

907 promoter, providing the endogenous promoter strength relative to the CMV
908 promoter.
909 We observed an overall variation in promoter strength of over 500-fold,
910 ranging from the 4-5 fold relative promoter strength of the RPL32 and EEF1α
911 promoters to the 0.01-fold relative promoter strength for the GUSB promoter as
912 compared to the CMV promoter (Figure 2d). This expression ratio appeared to be
913 similar across the cell population.
914
915 **Reporter gene expression as a function of transcriptionally active and**
916 **inactive BAC scaffolds**
917 To find the best BAC scaffold for the BAC TG-EMBED system, we tested
918 BAC scaffolds from both actively transcribed regions and regions containing
919 silenced genes or no genes. Specifically, we measured the expression as a
920 function of copy number of one specific reporter gene construct inserted into
921 these BAC scaffolds. Previous applications of BAC TG-EMBED showed a linear
922 relationship between copy number and expression level, largely independent of
923 the chromosome integration site, demonstrating copy-number dependent,
924 position independent transgene expression (51, 54). For active chromosomal
925 regions, we chose the RP11-138I1 BAC containing the human ubiquitin B gene
926 locus (UBB BAC), the RP23-401D9 BAC containing the "safe-haven" mouse
927 *Rosa26* genetrap locus (ROSA BAC) (91), and the CITB-057L22 BAC carrying
928 the mouse Dhfr gene locus (DHFR BAC). For inactive chromosomal regions, we
929 chose the CTD-2207K13 BAC (2207K13 BAC) that contains no known gene or

930 regulatory element from a gene-desert region from the human genome, and the
931 CTD-2643I7 (HBB BAC) containing the human HBB gene locus and multiple
932 olfactory genes, all of which are transcriptionally silenced in fibroblasts (92).
933 We selected the UBC promoter for this reporter gene cassette as this
934 promoter had previously been shown to drive high expression across multiple cell
935 types (86); in our dual reporter system the UBC promoter was 2.6-fold stronger
936 than the CMV promoter (Figure 2d). Moreover, to eliminate any possible
937 transcriptional interference from closely spaced reporter and selectable marker
938 minigenes and to minimize any epigenetic silencing arising from DNA
939 methylation of this reporter gene-selectable marker construct, we used a
940 commercially available GFP-ZeoR fusion protein gene construct in which all CpG
941 dinucleotides had been removed and replaced by synonymous codons (Figure
942 3a).
943 We inserted this UBC-GFP-ZeoR reporter gene construct into different
944 BAC scaffolds by BAC recombineering, using *galK* for positive/negative selection
945 (63, 64). To eliminate potential artifacts caused by proximity to active promoters,
946 transcriptional start sites (TSS), or miRNA sequences, we chose insertion sites
947 flanked on both sides by at least 5 kb free of such sequence elements (Figure
948 3b). The UBB, HBB, 2207K13, ROSA, DHFR BACs with the UBC-GFP-ZeoR
949 reporter gene insertion were named as UBB-UG, HBB-UG, 2207K13-UG, ROSA-
950 UG and DHFR-UG.
951 After transfection, multiple cell clones (n=20-40) carrying stably integrated
952 BAC arrays were selected for Zeocin resistance and analyzed for reporter gene

953 expression by flow cytometry, using untransfected NIH 3T3 cells to determine
954 background, autofluorescence levels. For each cell clone, we used flow
955 cytometry to measure the mean GFP reporter expression and qPCR to measure
956 reporter gene copy number. These cell clones showed GFP fluorescence mean
957 levels ranging from 10-1000 fold higher than the background autofluorescence.
958 Our original working hypothesis predicted that the BAC TG-EMBED
959 reporter expression should be uniform in all cells of the same clone. Also, we
960 expected to see a linear relationship between mean reporter gene fluorescence
961 and number of BAC copies, signifying a copy-number-dependent, position
962 independent expression. Furthermore, we expected that the slope of this linear
963 relationship would be higher for BAC scaffolds expected to reconstitute an active
964 chromatin environment permissive for transgene expression as compared to
965 BAC scaffolds expected to reconstitute a more condensed, inactive chromatin
966 environment (Figure 1). In contrast, we expected that the reporter gene cassette
967 transfected without any BAC scaffold would show clonal expression levels that
968 poorly correlated with reporter gene copy number (copy-number-independent
969 expression).
970 Unexpectedly, the stable cell clones we isolated showed two distinct types
971 of population expression profiles- uniform versus heterogeneous. Uniform clones
972 showed single, relatively narrow expression peaks in the flow cytometry
973 histograms, with more than 90% of the cells showing GFP fluorescence varying
974 only over a 10-fold intensity range (Figure 3c, left). Heterogeneous clones
975 instead showed two peaks with a range of GFP expression varying ~1000-fold,

976 with the lower GFP intensity peak overlapping with the autofluorescence
977 distribution of control cells (Figure 3c, right). We had not previously observed
978 such heterogeneous expression profile using our original DHFR BAC containing
979 the CMV-driven mRFP alone or both the CMV-driven EGFP and CMV-driven
980 mRFP reporter genes (51). However, we had observed ~80% uniform clones for
981 a GAPDH BAC scaffold with the UBC-GFP-ZeoR reporter gene inserted (54).
982 The percentage of clones showing such heterogeneous expression varied from
983 58% to 83% for the 5 BAC scaffolds surveyed here (Table 1). No similar
984 heterogeneous expression profile was observed when the reporter gene
985 construct was transfected by itself (Table 1).
986 As expected, the control transfection of the reporter gene cassette by itself
987 resulted in copy-number-independent expression of the reporter gene (Figure 3d,
988 R^2=0.09), while the reporter gene embedded within the BACs yielded a linear
989 relationship between reporter gene fluorescence for both uniform (black) and
990 heterogeneous (red) BAC transgene clones (Figure 3d, R^2=0.561 to 0.914).
991 Surprisingly, we observed no more than a 4-fold variation in expression
992 per copy number among the 5 different BAC scaffolds tested, with no obvious
993 relationship between the observed slope and the type of BAC scaffold (Figure
994 3d). Although the transcriptionally active DHFR BAC produced the highest slope,
995 the transcriptionally inactive HBB BAC and the 2207K13 BAC containing DNA
996 from a gene desert produced the second and third highest slopes, while the BAC
997 containing DNA from the “safe haven” mouse *Rosa26* locus produced the lowest
998 slope.

999 Overall, these results show that for this UBC-GFP-ZeoR reporter gene,
1000 high-level, copy-number-dependent transgene expression using the BAC TG-
1001 EMBED method does not require BACs containing active, housekeeping
1002 genomic regions, but can also be obtained from a wide range of BAC genomic
1003 DNA inserts, including gene-desert regions. This means BAC TG-EMBED can
1004 be used to drive expression of only the transgenes added to the BAC scaffold,
1005 without overexpression of the genes contained within the BAC scaffold.
1006

1007 **Temporal stability of BAC-embedded reporter gene expression in uniform**
1008 **cell clones**

1009 We previously showed that the BAC TG-EMBED method provided long-
1010 term stability of transgene expression in the presence of continued drug selection
1011 (51). However, in the absence of drug selection we observed a 30-80% drop in
1012 expression over several months of cell passaging without any apparent drop in
1013 the integrated BAC copy number (51).

1014 Here we determined the long-term stability of the UBC-GFP-ZeoR reporter
1015 gene expression for both uniform and heterogeneous clones for four different
1016 BAC scaffolds. Individual clones for each BAC scaffold (3 uniform and 2
1017 heterogeneous for ROSA-UG BAC, 7 uniform and 4 heterogeneous for 2207K13-
1018 UG BAC, 8 uniform and 3 heterogeneous for UBB-UG BAC, and 3 uniform and 3
1019 heterogeneous for DHFR-UG BAC) were passaged up to three months in the
1020 absence or presence of drug selection and analyzed for reporter gene
1021 fluorescence at regular intervals after removal of drug selection.

1022 With the exception of a small number of apparent fluctuations possibly
1023 related to transient changes in culture conditions, clones with uniform reporter
1024 gene expression showed no significant change either in the mean fluorescence
1025 values (Figure 4a) or in the distribution of fluorescence among the same clones
1026 (Figure 4b and Supplementary Figures S1) over time in the absence of selection
1027 for all four BAC scaffolds tested. In the presence of continued selection, uniform
1028 clones containing DHFR-UG or ROSA-UG BACs showed no significant reporter
1029 gene expression change, while an ~50% or 100% increase was observed for the
1030 UBB-UG or 2207K13-UG BAC clones, respectively (Figure 4a). No changes in
1031 estimated BAC copy number based on qPCR measurement were observed for
1032 any of these clones during this time series. This suggests that epigenetic
1033 changes driven by selection pressure may be responsible for these small
1034 increases in reporter gene expression.
1035 Notably, in the absence of selection, heterogeneous clones for all tested
1036 BAC scaffolds showed a significant and progressive loss of reporter gene
1037 expression over time. This led to a significant fraction of cells showing
1038 autofluorescence levels of fluorescence by the end of the experiment (Figure 4a).
1039 Reporter gene expression-level became progressively more homogenous, but at
1040 lower fluorescence levels (Figure 4b and Supplementary Figure S1). With
1041 selection, UBB-UG and DHFR-UG BAC heterogeneous clones showed a 1.6 to
1042 3-fold increase in reporter gene expression, respectively, while the other BAC
1043 scaffold heterogeneous clones showed no significant changes (Figure 4a).
1044

1045 **BAC transgenes are maintained as episomes in heterogeneous clones**

1046 In our previous work, all stable cell clones obtained after BAC transfection
1047 and drug selection contained single BAC copies or multi-copy BAC arrays that
1048 had integrated into endogenous chromosomes (51, 65, 93–95) consistent with
1049 similar results from numerous laboratories. Thus, we initially assumed that the
1050 broad distribution of reporter gene fluorescence observed in heterogeneous cell
1051 clones was due to position effect variegation (PEV) of the BAC TG-EMBED
1052 reporter genes. We hypothesized that integrations into some chromosome
1053 integration sites led to uniformly-expressing clones, while integration into other
1054 chromosome sites prone to PEV led to heterogeneous clones with variegated
1055 transgene expression.

1056 However, the observation of a progressive loss over time of reporter gene
1057 expression for all heterogeneous clones led us to question the genome stability
1058 of the BAC transgenes in these clones. To test the relationship between
1059 changes in reporter gene expression and BAC copy number, we first sorted cells
1060 from the heterogeneous DHFR-UG-s3 cell clone by fluorescence-activated cell
1061 sorting (FACS), using a narrow sorting-window centered around the GFP peak
1062 fluorescence level (Figure 5a). After cell-sorting, with drug selection the original
1063 heterogeneous reporter gene expression distribution reestablished itself within
1064 one week of culture (Figure 5b). We then resorted cells showing different levels
1065 of GFP fluorescence using four narrow fluorescence windows P1, P2, P3, and P4
1066 (Figure 5b), and then used qPCR to measure the BAC copy number in cells from
1067 each of these sorting windows. Plotting mean cell fluorescence intensity levels

1068 versus copy number for these clonal subpopulations yielded a strikingly linear
1069 relationship (R^2=0.99) (Figure 5c). Thus, the variable reporter gene expression
1070 level in this heterogeneous cell clone is the result of loss of BAC transgenes
1071 rather than chromosome PEV.

1072 To identify the source of this BAC copy-number instability, we next used
1073 DNA FISH to visualize BAC transgenes within interphase nuclei and mitotic
1074 chromosome spreads. We compared the distribution of BAC transgenes within
1075 the heterogeneous clone, DHFR-UG-s3, versus a uniform clone, DHFR-UG-f3-
1076 15.

1077 DNA FISH suggested that whereas the uniform clone contained cells with
1078 an integrated DHFR BAC array, the heterogeneous clone contained cells in
1079 which the DHFR BAC was present as episomes. Specifically, interphase FISH
1080 against the DHFR BAC in the heterogeneous clone revealed multiple,
1081 noncontiguous, small spots distributed randomly throughout the nuclei (Figure
1082 5d). The number of these spots was highly variable in different cell nuclei,
1083 suggesting unequal segregation of BAC transgenes. In contrast, most cells from
1084 the uniform clone showed just one large, fiber-like FISH spot per nucleus (Figure
1085 5e). Moreover, FISH spots in mitotic spreads from the heterogeneous clone
1086 were either touching or spatially separated from the chromosomes (Figure 5f),
1087 whereas FISH spots in mitotic spreads from the uniform clone were always
1088 located within the chromosome (Figure 5g). The number of FISH spots per
1089 mitotic spread was highly variable in the heterogeneous clone, with each spot
1090 much smaller than the single FISH spot visualized within the mitotic chromosome

1091 from the uniform clone. Interestingly, the FISH spots in the heterogeneous clone
1092 mitotic spreads had weak DAPI staining, varying from slightly elevated over
1093 background to no difference from background (Figure 5h), suggesting these
1094 structures are much smaller than previously described double minute
1095 chromosomes (DMs) generated by gene amplification (96–98).
1096 Using DNA FISH of both interphase nuclei and mitotic spreads, we
1097 confirmed this finding of integrated BACs in all uniformly expressing clones
1098 versus episomal BACs in all heterogeneously expressing clones in additional cell
1099 clones carrying BAC transgenes based on three different BAC scaffolds
1100 (Supplementary Figure S2 and data not shown). Specifically, this includes 4
1101 heterogeneous and 4 uniform DHFR-UG BAC clones, 3 heterogeneous and 6
1102 uniform HBB-UG BAC clones, and one heterogenous and 2 uniform COL1A1-UG
1103 BAC clones.
1104 Unequal segregation of these BAC episomes during cell division would
1105 explain the heterogeneity of BAC transgene copy number in the cell population of
1106 heterogeneous clones, leading to variability of reporter gene expression. Indeed,
1107 telophase cells from heterogeneous clones showed unequal numbers of FISH
1108 spots in the two daughter nuclei (Figure 5i). In the absence of continued drug
1109 selection, we would expect cells that have lost BAC transgenes will accumulate if
1110 there is any selective growth advantage for cells with fewer BAC copies.
1111
1112 **BAC episomes are circular and ~1 Mb in size**

1113 We analyzed the average amount of DNA per BAC episome, using two
1114 independent methods- light microscopy and pulsed-field gel electrophoresis
1115 (PFGE). Both methods produced a similar estimate of ~800-1000 kb per BAC
1116 episome.

1117 Using light microscopy, we measured the average DHFR BAC episome
1118 DAPI integrated staining intensity in mitotic spreads from cell clone DHFR-UG-s3
1119 relative to the smallest mouse chromosome (chr19) with known DNA content of
1120 61.4 Mbp (Figure 6a). This comparison produced an estimated mean episome
1121 size of 770 kb in this DHFR-UG-s3 clone.

1122 Using PFGE, we observed that the BAC episomes were circular and
1123 estimated the modal BAC episome size to be ~900 kb and 1 Mbp for DHFR-UG
1124 and HBB-UG BAC episomes in cell clones DHFR-UG-s3 and HBB-UG-100d3,
1125 respectively. Two different cell clones, DHFR-UG-f3-1 and HBB-UG-fD2, were
1126 used as negative controls as they contained the same DHFR-UG or HBB-UG
1127 BAC DNA as the cell clones with episomes but the BAC DNA was integrated
1128 within endogenous mouse chromosomes. *E. coli* strains containing the DHFR or
1129 HBB BACs were used as positive controls for detection of circular episomes.

1130 Pulsed-field gels were analyzed by Southern blotting using pooled BAC
1131 DNA PCR products as the hybridization probes. Linear but not circular DNAs
1132 migrate in pulsed-field gels. Similar to the *E. coli* controls containing circular
1133 BACs, the Southern blot signals for the BAC DNA from the two clones containing
1134 episomes did not migrate out of the wells (Figure 6b and Supplementary Figure
1135 S3a-b), consistent with circular rather than linear BAC episomes.

1136 To validate that the BAC episomes are really circular, and to estimate their
1137 size, the agarose-embedded DNA was digested using the ssDNA specific
1138 Nuclease S1 prior to PFGE and Southern blot hybridization. After removal of
1139 proteins, circular DNA episomes in both bacteria and mammalian cells are
1140 typically negatively supercoiled. This supercoiling generates torsional stress
1141 which is relieved by local formation of single-stranded regions. Thus, S1
1142 nuclease has been used to cut these single-stranded regions and linearize
1143 circular DNA episomes (99–101). After S1 digestion, DNA from the cell clones
1144 carrying BAC episomes now showed DNA smears with peak intensities of ~900
1145 kb and 1 Mb for the DHFR-UG-s3 and HBB-UG-100d3 cell lines, respectively
1146 (Figure 6b and Supplementary Figure S3a-b). In contrast, after S1 nuclease
1147 digestion, DNA from the integrated BAC clones showed signals within the wells
1148 and above 2 Mb, overlapping the smears of fragmented genomic DNA (Figure 6b
1149 and Supplementary Figure S3a-b). DNA of *E. coli* containing DHFR BAC and
1150 HBB BAC episomes produced bands at ~200-300 kb, in addition to signals in the
1151 wells (Figure 6b and Supplementary Figure S3a-b) after S1 nuclease digestion.
1152 These estimated BAC sizes measured slightly larger than the actual BAC sizes
1153 (~200 kb), indicating there might be a slight overestimation of episome sizes
1154 using our PFGE running conditions.
1155

1156 **BAC episomes contain no detectable host DNA as revealed by CNV analysis**

1157 The propagation of BAC transgenes as episomes was unexpected. A
1158 major question is whether these episomes consist solely of BAC DNA, or

1159 whether host DNA is also included and possibly required for episome
1160 propagation.
1161 We first compared the estimated episome DNA content size with
1162 estimates of BAC copies per episome. BAC episome sizes estimated by either
1163 light microscopy or PFGE were approximately twice as large as predicted from
1164 qPCR BAC copy number estimates (Table 2). The estimated average BAC
1165 content per episome was 445 kb in DHFR-UG-s3 and 716 kb in HBB-UG-100d3.
1166 The difference in the estimated BAC DNA content per episome and the
1167 average episome size is at most a few hundred kb, and may be accounted for by
1168 inaccuracies of the qPCR copy number, PFGE size estimation, and possible
1169 variation in sequence representation within the BAC episomes due to shearing of
1170 DNA and/or recombination during the transfection and creation of the BAC
1171 episomes.
1172 Alternatively, this difference in episome size versus qPCR estimation of
1173 BAC copies per episome could also be caused by presence of host cell genomic
1174 DNA on the episomes. To search with higher sensitivity for the possible presence
1175 of host DNA within the episome, we performed Whole Genome Sequencing
1176 (WGS) based copy number variation (CNV) analysis of the two clones, DHFR-
1177 UG-s3 and HBB-UG-100d3. Genomic regions present on the episomes would
1178 appear amplified in cells containing episomes (test sample), comparing to cells
1179 with no episomes (reference sample). Thus the ratio in the number of reads for a
1180 given bin between the test sample and the reference sample was calculated. To
1181 reduce noise, bins were merged into segments based on the $\log_2$(ratio), using a

1182 circular binary segmentation (CBS) algorithm (78, 79). The mean $\log_2$(ratio) of
1183 each segment was used to estimate the CNV of this segment in the test sample
1184 relative to the reference sample.

1185 Mouse 3T3 cells show genomic instability; therefore we anticipated CNV
1186 between the parental cell line and individual clones. To reduce false-positives
1187 derived from CNV between different 3T3 clones, independent of episome
1188 content, we used cells with low reporter gene fluorescence sorted from the cell
1189 clone containing the episomal BAC transgenes (region L, Figure 6c-d, and
1190 Supplementary Figure S4a) as the reference sample. To further reduce false
1191 positives, we also imposed constraints that copy number increase for true
1192 positive regions should be reproducible between experimental replicates and
1193 correlate with episomal copy number. We calculated the estimated CNV in cells
1194 sorted with high (H2) reporter gene expression, using sorted cells with low (L)
1195 expression as the reference sample (H2, L, Figure 6c-d, and Supplementary
1196 Figure S4a). We also compared the estimated CNV in cells sorted with high
1197 reporter gene expression in an independent experiment (H1, Figure 6c-d, and
1198 Supplementary Figure S4a) with the estimated CNV from the first experiment. All
1199 samples were sequenced to ~2x coverage.

1200 We used 3 and 30 kb bin sizes for analysis. To reduce noise, we excluded
1201 all bins in the test sample with zero reads (5.5% of total bins for 3 kb bin analysis
1202 and 3.5% for 30 kb bin analysis) plus extreme outlier bins, defined by the lower
1203 quantile minus 4 times the interquantile distance, with unusually low read count

1204 (~0.5% of total bins for 3 kb bin analysis and 2.5% for 30 kb bin analysis) in the
1205 test sample before calculating ratios.

1206 As a test of our analysis method, we compared the mean segment
1207 $\log_2$(ratio) of the BAC regions in H1 and H2, generated by the above analysis
1208 method, to the fold increase of BAC regions in H1 and H2 relative to L measured
1209 by qPCR. As expected, the results from the CNV and qPCR analysis were very
1210 similar (Supplementary Figure S5).

1211 We were interested in asking whether a specific host DNA element was
1212 present on each episome copy present within a cell clone. We estimated that on
1213 average the sorted cells with high reporter gene expression had 15-20 episome
1214 copies per cell, depending on the cell clone, based on qPCR of BAC DNA
1215 sequences and the estimated number of BACs per episome (Table 3).

1216 We estimated theoretical minimum copy number increase for episome-
1217 localizing host DNA (minimum increase) in the H1 and H2, based on BAC copy
1218 number measured by qPCR, and assuming the NIH 3T3 to be tetraploid and
1219 each episome to have the same host cell genomic DNA (Table 3). Segments
1220 with mean $\log_2$(ratio) equal to or greater than $\log_2$(minimum increase) in both H1
1221 and H2 samples were selected as candidates for being on the episomes (Figure
1222 6e).

1223 Interestingly, all candidate segments identified belonged to the BAC
1224 regions, including the UBC-GFP-ZeoR and the BAC vector (Figure 6f-g,
1225 Supplementary Figure S4b-c and Supplementary Figure S6), and no other
1226 mouse genomic sequence satisfied all of the above conditions.

1227 In conclusion, we could not detect host cell DNA reproducibly present on
1228 all episomal copies using bin sizes of either 3 or 30 kb. We therefore conclude
1229 BAC DNA itself is sufficient for the creation and propagation of these BAC
1230 episomes. We cannot exclude the possibilities, however, that an unmappable,
1231 repetitive host DNA sequence is present on the episomes and confers their
1232 ability to propagate or that different host DNA sequences are present on each
1233 episome present within a single cell clone.
1234

1235 **Multiple promoters added to BACs support formation of episomal BAC**
1236 **transgenes but only in certain cell lines**

1237 Because we did not observe episomal BAC transgenes in our original
1238 BAC-TG EMBED work using the CMV-mRFP-SV40-ZeoR reporter gene (51), we
1239 hypothesized that addition of the UBC-GFP-ZeoR reporter gene might be
1240 responsible for BAC episome formation. Our dual-reporter assay showed that
1241 the UBC promoter was much stronger than the CMV promoter; therefore, we
1242 further hypothesized that promoter strength might correlate with the frequency of
1243 BAC episome formation.

1244 To test this hypothesis, we isolated clones stably transfected with the
1245 dual-reporter DHFR BAC transgenes and examined reporter gene expression
1246 patterns in these clones by flow cytometry (Supplementary Figure S7a). As
1247 expected, no heterogeneously GFP/RFP expressing clones where observed
1248 when the mRFP reporter gene was driven by CMV promoter (n=13) or B2M
1249 promoter (n=6). In contrast, we observed ~70% or ~30% heterogeneously

1250 GFP/RFP expressing clones when the mRFP was driven by the EEF1a promoter
1251 (12/18) or the RPL32 promoter (10/29), respectively (Supplementary Figure S7a-
1252 b). We confirmed that BAC transgenes in these heterogeneously expressing
1253 clones were episomal using DNA FISH (Supplementary Figure S7c).
1254 These results using human promoter sequences added to the BACs, did
1255 show a rough correlation of promoter strength with the frequency of clones
1256 containing episomes. However, when we examined a series of DHFR BAC
1257 constructs, we instead observed clones with episomes using BAC transgenes
1258 containing the dual reporter, selectable marker CMV-mRFP-SV40-ZeoR reporter
1259 cassette. This included the identical DHFR BAC construct used in our previous
1260 BAC-TG EMBED work (51), as well as various DHFR BAC deletions
1261 (Supplementary Figure S8a). All the DHFR BAC constructs contain LacO
1262 repeats, and a NIH 3T3 derived clone expressing EGFP-LacI was used for
1263 transfection, so that BAC transgenes could be observed directly in fixed cells.
1264 Although all clones showed a unimodal flow cytometry expression pattern,
1265 explaining why we did not observe this phenomenon previously, a large fraction
1266 (DHFR-c27: 1/2, DHFR-c27d2: 2/10, DHFR-c27d3crz: 2/16, DHFR-c27d4: 7/10)
1267 of clones showed episomal BAC transgenes (Supplementary Figure S8b-c).
1268 Thus, the promoter used to drive reporter and/or selectable markers
1269 appears to determine not whether episomal BAC transgenes are established but
1270 rather whether a unimodal versus bimodal distribution is observed in cells
1271 containing these episomal BAC transgenes. The presence of strong promoters
1272 (UBC, EEF1a and RPL32) appears to allow the formation of bimodal distributions

1273 of reporter gene expression, possibly related to the balance between the
1274 degradation rate of the initially high levels of selectable marker versus the rate of
1275 loss of BAC transgene episomes during cell division.

1276 Next, we tested whether BACs can form episomes in a different cell line
1277 other than mouse NIH 3T3 fibroblasts. Previously, we observed cell clones
1278 containing only integrated BAC transgenes in CHO (93, 95) and mouse ES cells
1279 (54, 94). Reasoning that cancer cells with some level of genomic instability might
1280 be more prone to formation of BAC episomes, we tested the human colorectal
1281 carcinoma epithelial cell line, HCT116, using the 2207K13-UG BAC which
1282 produced 79% episome clones in NIH 3T3 cells.

1283 Four out of 32 stable clones showed a heterogeneous GFP distribution
1284 similar to that observed in NIH 3T3 episome clones, with a broad high fluorescent
1285 peak and a tail/secondary peak near the auto-fluorescence level (Supplementary
1286 Figure S9). However, none of these four clones showed episomal BAC
1287 transgenes by DNA FISH (Supplementary Figure S10a). Instead, most cells in
1288 each clone showed the same number (one or two) of spots, but these spots
1289 varied in size from cell to cell. Therefore, it appears that the broad GFP peaks in
1290 these four clones are due to some form of genomic instability leading to CNV of
1291 integrated transgene arrays. Interestingly, one clone, HCT116-k13_06, out of the
1292 32, which had a single GFP peak, showed a small fraction of cells of with
1293 episomal BAC transgenes, in contrast to the vast majority of cells which
1294 contained integrated BAC transgenes (Supplementary Figure S10b). One out of
1295 24 subclones of this HCT116-k13_06 clone, HCT116-k13_06-10, showed a

1296 similar mixed population with either integrated BACs or episomal BACs, similar to
1297 the parent clone HCT116-k13_06 (Supplementary Figure S10c). The low
1298 frequency of clones with episomal BACs, the variable size of the integrated BAC
1299 transgene arrays, and the co-existence of integrated and episomal BAC
1300 transgenes in the same cells and from the same clone suggests these episomes
1301 might arise from the well-known phenomenon of gene amplification (96–98).
1302 Similarly, a small percentage of clones carrying the GAPDH BAC in stable
1303 mouse ES cell colonies showed broad GFP expression peaks by flow cytometry,
1304 but FISH revealed this was due to variable size, integrated BAC transgene arrays
1305 (Binhui Zhao, Ph.D thesis), due presumably to some type of CNV induced by
1306 genomic instability of these transgene arrays.
1307 In conclusion, the high frequency establishment of BAC transgene
1308 episomes seen in mouse 3T3 cells does not appear to occur in either HCT116 or
1309 mouse ES cells, or at detectable frequency in CHO cells (54, 93–95).
1310

1311 **Expression of multiple-reporters by BAC-MAGIC**

1312 As a proof-of-principle application of our improved toolkit for BAC TG-
1313 EMBED, we created a multi-transgene BAC to label simultaneously the nuclear
1314 lamina, nucleoli, and nuclear speckles with a single stable transfection. The
1315 original DHFR BAC was used for this multi-transgene expression. A SNAP-
1316 tagged Lamin B1 reporter mini-gene was used to label the nuclear lamina, a
1317 SNAP-tagged Fibrillarin the nucleoli, and an mCherry-Magoh the nuclear
1318 speckles. We used the RPL32 promoter to drive the expression of the SNAP-

1319 tagged Lamin B1, and a promoter of intermediate strength, PPIA, for the SNAP-
1320 tagged Fibrillarin and the mCherry-tagged Magoh, which are both abundant
1321 proteins.

1322 Previously, we used random Tn5 transposition to introduce expression
1323 cassettes into BAC scaffolds (50), but this approach is limited in the number of
1324 serial insertions that can be made due to the remobilization of existing
1325 transposons, its requirement for multiple selectable markers, and the
1326 randomness of the insertion sites. Alternatively, BAC recombineering using
1327 antibiotic resistance genes as positive selectable markers have been used to
1328 insert expression cassettes into precise locations on the BACs. However, like
1329 transposition, this method relies on the availability of multiple selectable markers
1330 and introduces unwanted selectable markers into the BACs. An alternative BAC
1331 recombineering scheme using cycles of *galK*-based positive selection to insert
1332 sequences followed by negative selection to remove *galK* have been used to
1333 make multiple BAC modifications without addition of unwanted selectable
1334 markers. However, the low efficiency of negative selection, due to a high
1335 background of competing, spontaneous deletions of mammalian DNA with its
1336 high repetitive DNA content, makes this approach quite time and labor intensive.
1337 Typically, one month is required for each cycle of insertion of DNA by positive
1338 selection, removal of the selectable marker by negative selection, and
1339 subsequent screening and testing of DNA from colonies that survive the negative
1340 selection to identify the small fraction of colonies containing the desired
1341 homology-driven, specific deletion of just the selectable marker.

1342 To accelerate creation of BACs containing multiple transgene, we created
1343 a new BAC assembly approach, BAC MAGIC (**BAC**-**M**odular **A**ssembly of
1344 **G**enomic loci **I**nterspersed **C**assettes). BAC MAGIC combines the DNA
1345 assembler method in yeast (81, 82) and/or Gibson assembly (80) with traditional
1346 cloning methods to create a number of BAC recombination modules followed by
1347 sequential rounds of BAC recombineering in which one fragment is inserted
1348 using one selectable marker followed by addition of a new fragment overlapping
1349 the previous fragment using a second positive selectable marker which replaces
1350 the first (102). Each round of fragment insertion only requires ~ 1 week for
1351 transformation and screening of clones. In this way, 45 kb of the DHFR BAC
1352 was effectively reconstructed such that DHFR sequences remained but 3
1353 fluorescent mini-gene expression cassettes were added, each spaced by ~10 kb
1354 of DHFR sequence (Figure 7a-b, Supplementary Figure S11). The large
1355 homologous sequences flanking each expression cassette reduces
1356 recombination between similar sequences in other expression cassettes already
1357 inserted into the BAC, increasing the efficiency of this overall approach.
1358 We began the process using a DHFR BAC. After six rounds of BAC
1359 recombineering, we had created a BAC with four expression cassettes (Figure
1360 7b): a SNAP-tagged Lamin B1 minigene, a SNAP-tagged Fibrillarin minigene, a
1361 mCherry-Magoh minigene, and a ZeoR selectable marker.
1362 We tested simultaneous expression of the three reporters in 17
1363 independent NIH 3T3 cell clones transfected with the multi-reporter BAC by
1364 examining fluorescence in fixed cells under a microscope (SNAP-tagged proteins

1365 were labeled with a Fluorescein conjugated SNAP tag substrate before fixation).
1366 We observed uniform expression of all the three reporters in 16/17 clones. The
1367 loss of SNAP-Lamin B1 expression in one of the clone (Cl#16) may be due to
1368 random breakage of the BAC during transfection, as PCR revealed the absence
1369 of this minigene from the cell clone. Similarly, 12/14 U2OS human osteosarcoma
1370 cell clones showed both SNAP-Lamin B1 and SNAP-Fibrillarin expression after
1371 transfection of a BAC containing only these two expression cassettes (data not
1372 shown).
1373 Within individual cells, a linear correlation was observed between the
1374 integrated fluorescence intensity per cell of SNAP-tagged proteins Lamin B1 and
1375 Fibrillarin versus mCherry-Magoh in 4/4 representative NIH 3T3 clones (04, 08,
1376 13 and 14, Figure 7c). Moreover, these fluorescently tagged proteins showed
1377 uniform rather than variegating expression in different cell nuclei of the same
1378 clone observed under the microscope (Figure 7d).
1379

1380 **DISCUSSION**

1381 We previously demonstrated the utility of the BAC TG-EMBED method to
1382 achieve position-independent, copy-number-dependent, one-step transgene
1383 expression in mammalian cells (51, 54). Here, we have extended the BAC TG-
1384 EMBED methodology through four new advances and provided a proof-of-
1385 principle demonstration of this new methodology by efficiently creating cell lines
1386 stably expressing uniform levels of three different fluorescently tagged proteins-
1387 Lamin B1, Fibrillarin, and Magoh in a single stable transfection.

1388 First, we describe a toolkit of endogenous promoters providing an ~500-
1389 fold range in promoter strength varying from ~5 fold higher to ~100-fold weaker
1390 than the commonly used viral CMV promoter. As these promoters are from
1391 human genes shown to be expressed in a wide range of cell lines and tissues
1392 (60, 84–90), we expect them to support transgene expression in most cell types
1393 and independent of cell proliferation or differentiation state. While most of the
1394 previous studies on transgene promoters focused on conventional, strong
1395 promoters (53, 57–60), including a similar approach that expressed mini-genes
1396 within BAC scaffold (53), we included moderate-strength and weak promoters in
1397 our survey. The weak promoters we identified, such as GUSB and RPS3A,
1398 could possibly replace the commonly used minimal promoters or inducible
1399 promoters where a sustained low-level of transgene expression is needed.
1400 Moreover, this wide range of promoter strengths allows reproducible expression
1401 of multiple transgenes over a wide range of relative expression levels from a
1402 single BAC scaffold, lending itself to such purposes, for example, as the design

1403 of synthetic gene circuits, which typically requires expression of different
1404 components at reproducible relative expression levels (25).
1405 Second, we show that with the UBC-GFP-ZeoR reporter gene, our BAC-
1406 TG EMBED system achieved stable reporter gene expression of integrated BAC
1407 transgenes for several months in the absence of drug selection. This is an
1408 improvement over the 30-80% drop in expression observed originally with the
1409 CMV-mRFP-SV40-ZeoR reporter gene (51). Both the UBC promoter and the
1410 CpG free GFP-ZeoR gene body could be contributing to this improvement.
1411 Third, we show that at least with the UBC-GFP-ZeoR expression cassette,
1412 our BAC TG-EMBED system is not dependent on BAC scaffolds containing
1413 active DNA genomic regions but also works with BAC scaffolds containing
1414 silenced DNA genomic regions as well as gene deserts. UBC may represent a
1415 member of a class of active, house-keeping gene promoters that is relatively
1416 insensitive to chromosome position effects. This allows choice of a BAC scaffold
1417 for the BAC TG-EMBED method that will not co-express any genes other than
1418 the introduced transgene cassettes. In contrast, both our previous BAC TG-
1419 EMBED studies (51, 54) and similar work from other laboratories (52, 53), used
1420 only BACs containing highly-transcribed house-keeping genes, due to the
1421 assumption that either an active chromatin region or active 5' *cis*-regulatory
1422 regions would be required for creating a transcriptionally permissive environment
1423 for transgene expression. Integration of the UBC-GFP-ZeoR reporter gene into
1424 the BAC was required for position-independent, copy-number dependent
1425 expression, as its expression was copy-number independent when the same

1426 UBC-GFP-ZeoR reporter gene was stably transfected by itself into cells. The
1427 expression levels of this UBC-GFP-ZeoR were similar, per copy number, in cell
1428 clones with episomal BAC transgenes to levels in clones with integrated BACs.
1429 Fourth, we describe an episome version of our BAC-TG EMBED system.
1430 In a single experiment, clones containing either stably integrated or
1431 extrachromosomally maintained BAC transgenes can be isolated. Most of the
1432 widely used episomal vectors are either based on viral sequences or derived
1433 from the non-viral pEPI plasmid (38–41). A notable feature of the episomes
1434 generated by our BAC TG-EMBED system is that they are lost rapidly in the
1435 absence of drug selection, whereas both of the other two systems show
1436 selection-independent mechanisms for stable episome maintenance (43–47,
1437 103). While episome stability is valuable for certain applications, in other cases
1438 one would like to be able to easily eliminate the episomes as needed. Moreover,
1439 in contrast to the low copy number of episomes per cell produced using the other
1440 two methods, the BAC TG-EMBED method yields tens of BAC copies per cell
1441 allowing for much higher transgene expression levels. Additionally, the sizes of
1442 the episomes generated by the BAC TG-EMBED method are much larger than
1443 those generated by the other two methods. In the two clones we examined, the
1444 episomes were ~1 Mb and containing several copies of the BACs per episome
1445 and no detectable host DNA.
1446 The high frequency creation and simple composition of these BAC
1447 episomes contrasts with human artificial chromosomes (HACs), which are
1448 special episomes, usually 1-10 Mb in size containing centromeric repeat

1449 sequences, mitotically stable, and maintained at low copy number (104–107).
1450 Capable of introducing large DNA sequences into recipient cells, HACs have
1451 shown great potential in a wide range of applications, such as recombinant
1452 protein production, drug selection and gene therapy (108–111). However, the
1453 construction of HACs remains non-trivial: it requires cloning of either telomere
1454 sequences and/or alphoid DNA, the formation of HACs occurs at very low
1455 frequency and only in certain cell lines (112), and the transfer of HACs from
1456 donor cells into recipient cells is difficult (113, 114). Moreover, the presence of
1457 large telomere sequences and/or alphoid DNA on the HACs, and the
1458 heterchromatic state associated with these repeats, increases the likelihood of
1459 transgene silencing.

1460 In contrast, with our BAC-TG EMBED system, 10s-100s of stable cell
1461 clones containing multiple copies of ~Mb-size episomes, likely containing only
1462 BAC DNA, can be obtained from a single transfection. Cells containing high
1463 copy numbers of these BAC episomes can be enriched by flow sorting, while
1464 cells from these clones containing no BAC episomes can be recovered after
1465 removal of drug selection and/or flow sorting. We anticipate that with additional
1466 engineering, these BAC episomes might possibly become a high-capacity
1467 episome system complementary to HACs, assuming they can be isolated from
1468 one cell line and then introduced and propagated in other cell lines.

1469 It remains unclear how these BAC episomes form in NIH 3T3 cells and
1470 why they do not do so in other cell lines. In the two clones we studied, the
1471 episomes were circular DNA and composed of several BAC copies.

1472 Interestingly, previous studies have shown that plasmids containing a MAR that
1473 is also a replication initiation region (IR) could initiate gene amplification in certain
1474 primary cancer cells, forming homogenously staining regions (HSRs), integrating
1475 into existing double minutes (DMs) or forming DMs *de novo* in cells without DMs
1476 (115, 116). It is believed that the IR/MAR plasmids are initially replicated as
1477 extrachromosomal circles, and then they multimerize into larger circular
1478 molecules. These amplified circles further multimerize to form DMs, recombine
1479 with pre-existing DMs or integrate into chromosomes and initiate HSR formation
1480 (117, 118). This model is very similar to the episome model of gene
1481 amplification, where instead of the IR/MAR plasmids, small extrachromosomal
1482 circular DNAs, which are several hundred kb in size and are possibly produced
1483 by small chromosome deletions, initiates DM and HSR formation (97, 98, 119).
1484 Given that both the MAR and IR sequences are ubiquitous in the
1485 mammalian genome, it is likely that the BACs used in this study also contain
1486 MAR and/or IR sequences. However, unlike the MAR/IR plasmids, these BACs
1487 did not generate typical HSRs when integrated into the chromosomes, and the
1488 episomes were much smaller than DMs in NIH 3T3 cells. One possible
1489 explanation is that the BACs undergo initial steps of gene amplification to form
1490 the episomes in NIH 3T3 cells, but the cells have mechanisms to stop the
1491 episomes from further multimerization or amplification. As gene amplification
1492 happens only in cancer cells, perhaps BACs can only form episomes in certain
1493 cell lines. As shown here, BAC transgene formed episomes in a small fraction of
1494 HCT116 cells, which could not be stably maintained even with drug selection.

1495 Further study of this BAC episome phenomenon may provide new insights into
1496 the process of gene amplification.
1497 Alternatively, the formation of BAC transgene episomes in NIH 3T3 cells
1498 might occur through a process completely unrelated to gene amplification.
1499 Future work will be needed to determine the actual mechanism of this BAC
1500 episomal formation in mouse 3T3 cells.
1501 To facilitate the assembly of BACs expressing multiple mini-genes, we
1502 developed BAC-MAGIC, allowing creation of a multi-transgene expressing BAC
1503 in several weeks, rather than the 4-5 months which would have been required by
1504 multiple rounds of DNA insertion using conventional BAC recombineering. Initial
1505 attempts to reassemble large, ~50kb regions of DHFR using yeast DNA
1506 assembly failed, apparently due to recombination between repetitive elements
1507 within the DHFR BAC sequence as well as the expression cassettes. In contrast,
1508 assembly of 10-15 kb modules from several DNA fragments using yeast DNA
1509 assembly worked with high efficiency. BAC-MAGIC exploits Gibson and yeast
1510 DNA assembly to build smaller modules with efficient serial BAC recombineering
1511 to reconstruct large BAC constructs containing multiple mini-gene expression
1512 cassettes. More generally, BAC-MAGIC should provide a tool for reconstruction
1513 of large eukaryotic DNA sequences containing high numbers of repetitive
1514 elements.
1515 Finally, as a demonstration of our new version of BAC TG-EMBED
1516 system, we created cell lines expressing three different fluorescently tagged
1517 proteins in a single stable transfection step requiring just several weeks to isolate

1518 and expand cell clones. Most cell clones expressed all three tagged proteins at
1519 uniform levels and at reproducible relative levels of expression. This contrasts
1520 with the 6-12 months we have devoted in previous studies to create similar cell
1521 lines expressing multiple tagged proteins (120) through a series of individual
1522 transfections followed by extensive screening of colonies to identify the small
1523 fraction expressing suitable levels of tagged proteins with minimal variegation
1524 and/or progressive long-term transgene silencing over time.
1525 We anticipate that our expanded BAC TG-EMBED toolkit similarly will
1526 facilitate a wide range of applications requiring simultaneous expression of
1527 multiple transgenes.
1528

1529 **Figure Legends**

1530 **Figure 1. Two-prong experimental approach.** Left: Identification of promoters
1531 of different strengths- We measured relative promoter strengths by embedding
1532 EGFP and mRFP reporter genes into the DHFR BAC, using the CMV promoter
1533 to drive EGFP expression and the test promoter to drive mRFP. The ratio of
1534 mRFP and GFP expression, normalized by this same ratio for a CMV test
1535 promoter, defines promoter strength relative to CMV. Right: Surveying reporter
1536 gene expression in different BAC scaffolds- (Top) The UBC-GFP-ZeoR reporter
1537 gene was inserted into BACs carrying DNA from mouse or human genomic
1538 regions corresponding to either transcriptionally active or inactive genomic
1539 regions. (Bottom) Plotting reporter gene expression (y-axis) versus reporter gene
1540 copy number (x-axis) for multiple cell clones stably expressing BAC transgenes:
1541 a linear correlation would indicate copy-number dependent, position independent
1542 expression, while the slope of this linear correlation would measure reporter gene
1543 expression per copy number.

1544

1545 **Figure 2. Dual-reporter assay for promoter strength estimation.** (a) Dual
1546 reporter DHFR BAC showing the two genes on the BAC, *Dhfr* and *Msh3*, and the
1547 insertion sites of the two reporter expression cassettes. Longer vertical bars-
1548 exons; shorter vertical bars- UTRs; arrows- direction of transcription; green
1549 arrowhead- EGFP expression cassette insertion site; red arrowhead- mRFP
1550 expression cassette insertion site. (b) The two reporter gene/selectable marker
1551 cassettes used in the assay. The EGFP cassette (top) contains an EGFP

1552 minigene, driven by a CMV promoter, and a Kanamycin/Neomycin resistance
1553 gene (Kan/NeoR), driven by a SV40 promoter for expression in mammalian cells,
1554 or by a AmpR promoter for expression in bacteria. The mRFP cassette (bottom)
1555 contains a mRFP minigene and a Zeocin resistance gene (ZeoR). Different
1556 endogenous promoters were inserted immediately upstream of mRFP. ZeoR is
1557 driven by a SV40 promoter for expression in mammalian cells, or by a AmpR
1558 promoter for expression in bacteria. pA- poly(A) signal. (c) Scatter plots showing
1559 mRFP fluorescence (y-axis) vs EGFP fluorescence (x-axis) of cells from the
1560 mixed clonal populations stably transfected with dual reporter DHFR BACs.
1561 Promoters driving the mRFP and the ratio of mRFP/EGFP (promoter strength)
1562 are labeled in each plot. (d) Promoter strengths relative to CMV.
1563
1564 **Figure 3. Expression of reporter gene embedded in different BAC**
1565 **scaffolds**. (a) UBC-GFP-ZeoR-FRT-GalK-FRT cassette showing the GFP-ZeoR
1566 minigene driven by the UBC promoter and the *galK* positive/negative selection
1567 marker flanked by 34 bp flippase recognition target (FRT) sites (arrowheads). (b)
1568 Maps of the BACs used in the study. Longer vertical bars- exons; shorter vertical
1569 bars- UTRs; black arrows or arrowheads- direction of transcription; green arrow
1570 heads- UBC-GFP-ZeoR insertion site. (c) GFP fluorescence histograms obtained
1571 by flow-cytometry for “uniform” (left, green, clone DHFR-UG-f3-15) versus
1572 “heterogeneous” (right, green, clone DHFR-UG-f1-6) expressing NIH 3T3 clones
1573 carrying the DHFR-UG BAC. x-axis- fluorescence value, y-axis- cell number;
1574 gray- autofluorescence of untransfected cells. Fluorescence is measured in

1575 arbitrary units. (d) Scatter plots of mean normalized cellular GFP fluorescence
1576 (y-axis) vs reporter gene copy number (x-axis) for clonal populations transfected
1577 with the UBC-GFP-ZeoR cassette alone or with different BAC scaffolds carrying
1578 the UBC-GFP-ZeoR reporter gene. Linear regression fits (black lines, y-
1579 intercepts set to 0) are shown with corresponding R-squared values and
1580 equations. Red circles- heterogeneous clones; Black circles- uniform clones;
1581 Bottom right of plots: Number of clones analyzed.
1582
1583 **Figure 4. UBC-GFP-ZeoR reporter gene expression over time.** "Uniform"
1584 clones show stable expression with or without expression, while "heterogenous"
1585 clones show progressive loss of expression without selection. (a) Changes in
1586 GFP fluorescence of uniform versus heterogeneous clones, averaged over
1587 multiple clones (2-8), carrying indicated BAC transgenes during 96 days of
1588 continuous passaging with or without Zeocin selection. x-axis- number of days
1589 since removal of Zeocin; y-axis- mean fluorescence values of multiple clones
1590 divided by that at day zero; black- "uniform" expressing clones cultured with
1591 Zeocin; blue- "uniform" expressing clones cultured without Zeocin; red-
1592 "heterogeneous" expressing clones cultured with Zeocin; green-
1593 "heterogeneous" expressing clones cultured without Zeocin; (b) GFP
1594 fluorescence histogram of representative "uniform" and "heterogeneous"
1595 expressing NIH 3T3 clones at day 0, 24, 60 and 96 without selection. Gray-
1596 autofluorescence of untransfected cells; Green- GFP fluorescence of the
1597 indicated clones. x-axis- fluorescence; y-axis- cell number.

1598

1599 **Figure 5. BAC transgenes exist as episomes in heterogeneously**
1600 **expressing clones.** (a-c) BAC copy number analysis of sub-populations of a
1601 heterogeneous clone, DHFR-UG-s3, with different fluorescence levels. (a) GFP
1602 fluorescence histogram of DHFR-UG-s3 cells during first sorting (y-axis- cell
1603 number; x-axis- GFP fluorescence level). Cells within a narrow peak-window
1604 (dotted lines) were sorted by FACS. (b) GFP fluorescence histogram of sorted
1605 DHFR-UG-s3 cells after one week of cell growth. Cells within the four colored
1606 windows (P1-4) were sorted by FACS and used for BAC copy number estimation
1607 by qPCR. (c) Mean GFP fluorescence (y-axis) vs copy number (x-axis) of the
1608 four cell sub-populations and the original unsorted population shows linear
1609 correlation between fluorescence levels and copy number (R^2=0.99). (d-e) DNA
1610 FISH over interphase nuclei of the heterogeneous clone DHFR-UG-s3 (d) and a
1611 uniform clone DHFR-UG-f3-15 (e) to visualize the BAC transgenes. Maximum-
1612 intensity projections are shown. Gamma=0.5 was applied to FISH channel after
1613 projection to better display low intensity FISH spots. (f-g) DNA FISH over mitotic
1614 spreads of the heterogeneous clone DHFR-UG-s3 (f) and the uniform clone
1615 DHFR-UG-f3-15 (g). (h) DAPI intensity over an episome with strong FISH signal
1616 and one with weak FISH signal. Top: enlarged view of the white square area in
1617 (f); bottom: DAPI (red) and FISH signal (green) intensity profile along the white
1618 arrow in the top panel. (i) A pair of telophase nuclei of the heterogeneous clone,
1619 DHFR-UG-s3, showing unequal segregation of episomal BAC transgenes during
1620 mitosis. (d-i) Red- DNA DAPI stain; green- BAC FISH signal. Scale bars = 5 μm.

1621

1622 **Figure 6. BAC episome size estimation and CNV analysis.** (a) Estimation of
1623 average episome size in the DHFR-UG-s3 clone using mitotic FISH. Red- DNA
1624 DAPI stain; Green- BAC FISH signal; Red circles: regions of interest (ROIs) of
1625 FISH spots used for analysis; Yellow circles: ROI of the smallest chromosome in
1626 the field. Scale bars = 5 μm. This panel reuses the image in Figure 5f for
1627 analysis. (b) Southern hybridization using probes prepared from the DHFR BAC
1628 of cellular DNA without enzyme digestion, or digested with increasing amount of
1629 S1 Nuclease, separated by PFGE. Lane 1-4: uniform clone DHFR-UG-f3-1;
1630 Lane 5-8: heterogeneous clone DHFR-UG-s3; Lane 9-12: *E. coli* carrying the
1631 DHFR BAC. (c-g) CNV analysis of the DHFR-UG-s3 clone. (c) Flow chart of the
1632 CNV analysis. (d) Two FACS experiments for collecting cells with high (H1 and
1633 H2), and low (L) fluorescence subpopulation. x-axis- FITC channel intensity; y-
1634 axis- forward scatter; H1, H2, and L- sorting windows. (e) Episome-localizing
1635 genomic regions (pink highlighted regions) are expected to have mean $\log_2$(ratio)
1636 (red line) equal to or greater than $\log_2$(estimated minimum copy number
1637 increase) (blue dashed line). (f) $\log_2$(ratio) of individual bins (dark gray dots) and
1638 the segment mean $\log_2$(ratio) (red lines) around the *Dhfr-Msh3* locus belonging to
1639 the DHFR BAC (pink highlight) in the H1 and H2 subpopulations of the DHFR-
1640 UG-s3 clone. (g) Scatter plot of segment mean $\log_2$(ratio) vs segment mean
1641 $\log_2$(normalized reads) of all segments of the H1 and H2 subpopulations of the
1642 DHFR-UG-s3 clone. Pink dots- segments belonging to the DHFR BAC, including
1643 the *Dhfr-Msh3* locus, UBC-GFP-ZeoR and the BAC vector; Black dots- remaining

1644 segments in the genome. (f-g) Blue dashed line: $\log_2$(estimated minimum copy
1645 number increase).
1646
1647 **Figure 7. BAC-MAGIC and simultaneous multi-reporter expression.** (a-b)
1648 Construction of the multi-reporter DHFR BAC by BAC-MAGIC. (a) Modular
1649 design of BAC-MAGIC. Reporter module 01, 02 and 03 contain reporter gene
1650 expression cassettes (X), DHFR BAC homologous sequences (dark gray), and
1651 Spectinomycin resistance markers (SpecR, yellow) near the 3' ends for bacterial
1652 selection. Intervening DHFR module 02, 03 and 04 contain DHFR BAC
1653 homologous sequences (dark gray), and antibiotic resistance markers near the 3'
1654 ends (Kanamycin/Neomycin resistance marker (Kan/NeoR, blue) in module 02
1655 and 03 for bacterial selection, and Zeocin resistance marker (ZeoR, dark green)
1656 in module 04 for dual selection in bacterial or mammalian cells). The dotted lines
1657 mark homologous regions between the reporter modules and the intervening
1658 DHFR modules. (b) Six sequential steps of BAC recombineering introduce three
1659 reporter expression cassettes, RPL32-driven SNAP-tagged Lamin B1 (RSLB1),
1660 PPIA-driven SNAP-tagged Fibrillarin (PSF), and PPIA-driven mCherry-Magoh,
1661 onto the DHFR BAC (light gray) with ~10 kb of intervening DHFR BAC
1662 sequences (dark gray). Homologous regions are indicated by crossed lines. (c)
1663 Relative expression of the SNAP-tagged Lamin B1 and Fibrillarin to the mCherry-
1664 Magoh reporter in four representative NIH 3T3 cell clones (04, 08, 13 and 14)
1665 containing the multi-reporter BAC. Integrated fluorescence intensities per cell of
1666 SNAP--fluorescein (y-axis) and mCherry-Magoh (x-axis) are plotted. Linear

1667 regression lines (y-intercepts set to 0) are shown with corresponding R-squared
1668 values. Number of nuclei of each clone analyzed range from 18 to 27. Red-
1669 Clone 04; Blue- Clone 08, Black- Clone 13; Green- Clone 14. (d) Representative
1670 images (maximum intensity projections of 2-3 optical sections) from the four cell
1671 clones (Clone 04, 08, 13 and 14) showing expression of the three reporter genes.
1672 Nuclear lamina is labeled with SNAP-tagged Lamin B1 (green), nucleoli with
1673 SNAP-tagged Fibrillarin (green), and speckles with mCherry-Magoh (red). One
1674 magnified nucleus from each representative field (top panel) is shown in the
1675 bottom panel. Scale bars = 5 μm.
1676
1677
1678
1679 **Table 1.** Percentage of heterogeneously expressing clones transfected with the
1680 UBC-GFP-ZeoR cassette alone or with different BAC scaffolds carrying the UBC-
1681 GFP-ZeoR reporter gene.
1682
1683 **Table 2.** BAC copy number, episome copy number, and BAC DNA content per
1684 episome in clone DHFR-UG-s3 and clone HBB-UG-100d3.
1685
1686 **Table 3.** BAC copy number, estimated episome copy number, and estimated
1687 minimum copy number increase of episome-localizing DNA in H1 and H2
1688 subpopulations relative to L subpopulation of clone DHFR-UG-s3 and clone HBB-
1689 UG-100d3.

1692 **FUNDING**

1693 This work was supported by the National Institute of General Medical Sciences
1694 [GM098319 and, in part, GM58460 to A.S.B.] and by the National Institutes of
1695 Health Common Fund 4D Nucleome Program [U54 DK107965 to A.S.B]. The
1696 content is solely the responsibility of the authors and does not necessarily
1697 represent the official views of the National Institute of General Medical Sciences
1698 or the National Institutes of Health.

1699

1700

1701 **ACKNOWLEDGEMENTS**

1702 We thank Edith Heard (Curie Institute) for providing DHFR BAC (clone
1703 057L22 from CITB mouse library), Veena K Parnaik (CSIR-CCMB, Hyderabad,
1704 India) for GFP-Lamin B1 plasmid, Miroslav Dundr (Rosalind Franklin University of
1705 Medicine and Science) for GFP-Fibrillarin plasmid, Huimin Zhao (University of
1706 Illinois Urbana-Champaign) for pQCXIN-TetR-mCherry plasmid, Peter Adams
1707 (Sanford Burnham Prebys Medical Discovery Institute) for BJ-hTERT cells, and
1708 KV Prasanth (University of Illinois Urbana-Champaign) for mRFP-Magoh
1709 plasmid. Use of the BD FACS AriaII was assisted by the flow cytometry facility
1710 stuff at Roy J. Carver Biotechnology Center, University of Illinois at Urbana-
1711 Champaign (UIUC).

1712

1713

REFERENCES

1. Walsh,G. (2018) Biopharmaceutical benchmarks 2018. *Nat. Biotechnol.*, **36**, 1136–1145.
2. Prelich,G. (2012) Gene overexpression: uses, mechanisms, and interpretation. *Genetics*, **190**, 841–54.
3. Glover,D.J., Lipps,H.J. and Jans,D.A. (2005) Towards safe, non-viral therapeutic gene expression in humans. *Nat. Rev. Genet.*, **6**, 299–310.
4. Zhu,J. (2012) Mammalian cell protein expression for biopharmaceutical production. *Biotechnol. Adv.*, **30**, 1158–1170.
5. Wurm,F.M. (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.*, **22**, 1393–8.
6. Rizzo,M.A., Davidson,M.W. and Piston,D.W. (2009) Fluorescent Protein Tracking and Detection: Applications Using Fluorescent Proteins in Living Cells. *Cold Spring Harb. Protoc.*, **2009**, pdb.top64-pdb.top64.
7. Takahashi,K. and Yamanaka,S. (2006) Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, **126**, 663–676.
8. Machida,K., Masutani,M., Kobayashi,T., Mikami,S. and Nishino,Y. (2012) Reconstitution of the human chaperonin CCT by co-expression of the eight distinct subunits in mammalian cells. *PROTEIN Expr. Purif.*, **82**, 61–69.
9. Akhtar,W., de Jong,J., Pindyurin,A. V., Pagie,L., Meuleman,W., de Ridder,J., Berns,A., Wessels,L.F.A., van Lohuizen,M. and van Steensel,B. (2013) Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell*, **154**, 914–27.
10. Chen,M., Licon,K., Otsuka,R., Pillus,L. and Ideker,T. (2013) Decoupling Epigenetic and Genetic Effects through Systematic Analysis of Gene Position. *Cell Rep.*, **3**, 128–37.
11. Robertson,G., Garrick,D., Wu,W., Kearns,M., Martin,D. and Whitelaw,E. (1995) Position-dependent variegation of globin transgene expression in mice. *Proc. Natl. Acad. Sci. U. S. A.*, **92**, 5371–5.
12. Girton,J.R. and Johansen,K.M. (2008) Chromatin structure and the regulation of gene expression: the lessons of PEV in Drosophila. *Adv. Genet.*, **61**, 1–43.
13. Ramunas,J., Montgomery,H.J., Kelly,L., Sukonnik,T., Ellis,J. and Jervis,E.J. (2007) Real-time fluorescence tracking of dynamic transgene variegation in stem cells. *Mol. Ther.*, **15**, 810–7.
14. Tchasovnikarova,I.A., Timms,R.T., Matheson,N.J., Wals,K., Antrobus,R., Gottgens,B., Dougan,G., Dawson,M.A. and Lehner,P.J. (2015) Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells. *Science (80-.).*, **348**, 1481–1485.
15. Karpen,G.H. (1994) Position-effect variegation and the new biology of heterochromatin. *Curr. Opin. Genet. Dev.*, **4**, 281–91.
16. He,J., Yang,Q. and Chang,L.-J. (2005) Dynamic DNA Methylation and Histone Modifications Contribute to Lentiviral Transgene Silencing in Murine Embryonic Carcinoma Cells. *J. Virol.*, **79**, 13497–13508.
17. Suzuki,M., Kasai,K. and Saeki,Y. (2006) Plasmid DNA sequences present in

1760 conventional herpes simplex virus amplicon vectors cause rapid transgene
1761 silencing by forming inactive chromatin. *J. Virol.*, **80**, 3293–300.
1762 18. Scrable,H. and Stambrook,P.J. (1999) A genetic program for deletion of
1763 foreign DNA from the mammalian genome. *Mutat. Res.*, **429**, 225–37.
1764 19. Minoguchi,S. and Iba,H. (2008) Instability of retroviral DNA methylation in
1765 embryonic stem cells. *Stem Cells*, **26**, 1166–73.
1766 20. Dorer,D.R. and Henikoff,S. (1997) Transgene repeat arrays interact with
1767 distant heterochromatin and cause silencing in cis and trans. *Genetics*, **147**,
1768 1181–90.
1769 21. Garrick,D., Fiering,S., Martin,D.I. and Whitelaw,E. (1998) Repeat-induced
1770 gene silencing in mammals. *Nat. Genet.*, **18**, 56–9.
1771 22. Laker,C., Meyer,J., Schopen, a, Friel,J., Heberlein,C., Ostertag,W. and
1772 Stocking,C. (1998) Host cis-mediated extinction of a retrovirus permissive for
1773 expression in embryonal stem cells during differentiation. *J. Virol.*, **72**, 339–
1774 348.
1775 23. Herbst,F., Ball,C.R., Tuorto,F., Nowrouzi,A., Wang,W., Zavidij,O.,
1776 Dieter,S.M., Fessler,S., van der Hoeven,F., Kloz,U., *et al.* (2012) Extensive
1777 methylation of promoter sequences silences lentiviral transgene expression
1778 during stem cell differentiation in vivo. *Mol. Ther.*, **20**, 1014–21.
1779 24. Hotta,A. and Ellis,J. (2008) Retroviral vector silencing during iPS cell
1780 induction: an epigenetic beacon that signals distinct pluripotent states. *J.*
1781 *Cell. Biochem.*, **105**, 940–8.
1782 25. Brophy,J.A.N. and Voigt,C.A. (2014) Principles of genetic circuit design. *Nat.*
1783 *Methods*, **11**, 508.
1784 26. Pikaart,M.J., Recillas-Targa,F. and Felsenfeld,G. (1998) Loss of
1785 transcriptional activity of a transgene is accompanied by DNA methylation
1786 and histone deacetylation and is prevented by insulators. *Genes Dev.*, **12**,
1787 2852–2862.
1788 27. Emery,D.W., Yannaki,E., Tubb,J. and Stamatoyannopoulos,G. (2000) A
1789 chromatin insulator protects retrovirus vectors from chromosomal position
1790 effects. *Proc. Natl. Acad. Sci.*, **97**, 9150–9155.
1791 28. Grosveld,F., van Assendelft,G.B., Greaves,D.R. and Kollias,G. (1987)
1792 Position-independent, high-level expression of the human beta-globin gene
1793 in transgenic mice. *Cell*, **51**, 975–85.
1794 29. Guy,L.G., Kothary,R., DeRepentigny,Y., Delvoye,N., Ellis,J. and Wall,L.
1795 (1996) The beta-globin locus control region enhances transcription of but
1796 does not confer position-independent expression onto the lacZ gene in
1797 transgenic mice. *EMBO J.*, **15**, 3713–21.
1798 30. Phi-Van,L., von Kries,J.P., Ostertag,W. and Strätling,W.H. (1990) The
1799 chicken lysozyme 5' matrix attachment region increases transcription from a
1800 heterologous promoter in heterologous cells and dampens position effects
1801 on the expression of transfected genes. *Mol. Cell. Biol.*, **10**, 2302–7.
1802 31. Kim,J.M., Kim,J.S., Park,D.H., Kang,H.S., Yoon,J., Baek,K. and Yoon,Y.
1803 (2004) Improved recombinant gene expression in CHO cells using matrix
1804 attachment regions. *J. Biotechnol.*, **107**, 95–105.
1805 32. Williams,S., Mustoe,T., Mulcahy,T., Griffiths,M., Simpson,D., Antoniou,M.,

1806 Irvine,A., Mountain,A. and Crombie,R. (2005) CpG-island fragments from the
1807 HNRPA2B1/CBX3 genomic locus reduce silencing and enhance transgene
1808 expression from the hCMV promoter/enhancer in mammalian cells. *BMC*
1809 *Biotechnol.*, **5**, 17.
1810 33. Müller-Kuller,U., Ackermann,M., Kolodziej,S., Brendel,C., Fritsch,J.,
1811 Lachmann,N., Kunkel,H., Lausen,J., Schambach,A., Moritz,T., *et al.* (2015)
1812 A minimal ubiquitous chromatin opening element (UCOE) effectively
1813 prevents silencing of juxtaposed heterologous promoters by epigenetic
1814 remodeling in multipotent and pluripotent stem cells. *Nucleic Acids Res.*, **43**,
1815 1577–92.
1816 34. Kwaks,T.H.J., Barnett,P., Hemrika,W., Siersma,T., Sewalt,R.G.A.B.,
1817 Satijn,D.P.E., Brons,J.F., van Blokland,R., Kwakman,P., Kruckeberg,A.L., *et*
1818 *al.* (2003) Identification of anti-repressor elements that confer high and
1819 stable protein production in mammalian cells. *Nat. Biotechnol.*, **21**, 553–8.
1820 35. Grandchamp,N., Henriot,D., Philippe,S., Amar,L., Ursulet,S., Serguera,C.,
1821 Mallet,J. and Sarkis,C. (2011) Influence of insulators on transgene
1822 expression from integrating and non-integrating lentiviral vectors. *Genet.*
1823 *Vaccines Ther.*, **9**, 1.
1824 36. Truffinet,V., Guglielmi,L., Cogné,M. and Denizot,Y. (2005) The chicken beta-
1825 globin HS4 insulator is not a silver bullet to obtain copy-number dependent
1826 expression of transgenes in stable B cell transfectants. *Immunol. Lett.*, **96**,
1827 303–4.
1828 37. Bharadwaj,R.R., Trainor,C.D., Pasceri,P. and Ellis,J. (2003) LCR-regulated
1829 transgene expression levels depend on the Oct-1 site in the AT-rich region
1830 of beta -globin intron-2. *Blood*, **101**, 1603–10.
1831 38. Lufino,M.M.P., Edser,P.A.H. and Wade-Martins,R. (2008) Advances in high-
1832 capacity extrachromosomal vector technology: Episomal maintenance,
1833 vector delivery, and transgene expression. *Mol. Ther.*, **16**, 1525–1538.
1834 39. Ehrhardt,A., Haase,R., Schepers,A., Deutsch,M.J., Lipps,H.J. and Baiker,A.
1835 (2008) Episomal vectors for gene therapy. *Curr. Gene Ther.*, **8**, 147–61.
1836 40. Conese,M., Auriche,C. and Ascenzioni,F. (2004) Gene therapy progress and
1837 prospects: episomally maintained self-replicating systems. *Gene Ther.*, **11**,
1838 1735–41.
1839 41. Van Craenenbroeck,K., Vanhoenacker,P. and Haegeman,G. (2000)
1840 Episomal vectors for gene expression in mammalian cells. *Eur. J. Biochem.*,
1841 **267**, 5665–78.
1842 42. Frappier,L. (2012) Contributions of Epstein-Barr nuclear antigen 1 (EBNA1)
1843 to cell immortalization and survival. *Viruses*, **4**, 1537–47.
1844 43. Piechaczek,C., Fetzer,C., Baiker,A., Bode,J. and Lipps,H.J. (1999) A vector
1845 based on the SV40 origin of replication and chromosomal S/MARs replicates
1846 episomally in CHO cells. *Nucleic Acids Res.*, **27**, 426–428.
1847 44. Baiker,A., Maercker,C., Piechaczek,C., Schmidt,S.B., Bode,J., Benham,C.
1848 and Lipps,H.J. (2000) Mitotic stability of an episomal vector containing a
1849 human scaffold/matrix-attached region is provided by association with
1850 nuclear matrix. *Nat. Cell Biol.*, **2**, 182–4.
1851 45. Jenke,A.C.W., Stehle,I.M., Herrmann,F., Eisenberger,T., Baiker,A., Bode,J.,

1852 Fackelmayer,F.O. and Lipps,H.J. (2004) Nuclear scaffold/matrix attached
1853 region modules linked to a transcription unit are sufficient for replication and
1854 maintenance of a mammalian episome. *Proc. Natl. Acad. Sci.*, **101**, 11322–
1855 11327.
1856 46. Stehle,I.M., Postberg,J., Rupprecht,S., Cremer,T., Jackson,D.A. and
1857 Lipps,H.J. (2007) Establishment and mitotic stability of an extra-
1858 chromosomal mammalian replicon. *BMC Cell Biol.*, **8**, 1–12.
1859 47. Argyros,O., Wong,S.P., Niceta,M., Waddington,S.N., Howe,S.J., Coutelle,C.,
1860 Miller, a D. and Harbottle,R.P. (2008) Persistent episomal transgene
1861 expression in liver following delivery of a scaffold/matrix attachment region
1862 containing non-viral vector. *Gene Ther.*, **15**, 1593–1605.
1863 48. Tessadori,F., Zeng,K., Manders,E., Riool,M., Jackson,D. and van Driel,R.
1864 (2010) Stable S/MAR-based episomal vectors are regulated at the chromatin
1865 level. *Chromosome Res.*, **18**, 757–75.
1866 49. Chen,Z.Y., He,C.Y., Meuse,L. and Kay,M. a (2004) Silencing of episomal
1867 transgene expression by plasmid bacterial DNA elements in vivo. *Gene*
1868 *Ther.*, **11**, 856–864.
1869 50. Riu,E., Chen,Z.-Y., Xu,H., He,C.-Y. and Kay,M. a (2007) Histone
1870 modifications are associated with the persistence or silencing of vector-
1871 mediated transgene expression in vivo. *Mol. Ther.*, **15**, 1348–55.
1872 51. Bian,Q. and Belmont,A.S. (2010) BAC TG-EMBED: one-step method for
1873 high-level, copy-number-dependent, position-independent transgene
1874 expression. *Nucleic Acids Res.*, **38**, e127.
1875 52. Blaas,L., Musteanu,M., Eferl,R., Bauer,A. and Casanova,E. (2009) Bacterial
1876 artificial chromosomes improve recombinant protein production in
1877 mammalian cells. *BMC Biotechnol.*, **9**, 3.
1878 53. Zboray,K., Sommeregger,W., Bogner,E., Gili,A., Sterovsky,T., Fauland,K.,
1879 Grabner,B., Stiedl,P., Moll,H.P., Bauer,A., *et al.* (2015) Heterologous protein
1880 production using euchromatin-containing expression vectors in mammalian
1881 cells. *Nucleic Acids Res.*, **43**, e102.
1882 54. Chaturvedi,P., Zhao,B., Zimmerman,D.L. and Belmont,A.S. (2018) Stable
1883 and reproducible transgene expression independent of proliferative or
1884 differentiated state using BAC TG-EMBED. *Gene Ther.*, **25**, 376–391.
1885 55. Fitzsimons,H.L., Bland,R.J. and During,M.J. (2002) Promoters and regulatory
1886 elements that improve adeno-associated virus transgene expression in the
1887 brain. *Methods*, **28**, 227–36.
1888 56. Brooks,A.R., Harkins,R.N., Wang,P., Qian,H.S., Liu,P. and Rubanyi,G.M.
1889 (2004) Transcriptional silencing is associated with extensive methylation of
1890 the CMV promoter following adenoviral gene delivery to muscle. *J. Gene*
1891 *Med.*, **6**, 395–404.
1892 57. Hong,S., Hwang,D.-Y., Yoon,S., Isacson,O., Ramezani,A., Hawley,R.G. and
1893 Kim,K.-S. (2007) Functional analysis of various promoters in lentiviral
1894 vectors at different stages of in vitro differentiation of mouse embryonic stem
1895 cells. *Mol. Ther.*, **15**, 1630–9.
1896 58. Chen,C., Krohn,J., Bhattacharya,S. and Davies,B. (2011) A comparison of
1897 exogenous promoter activity at the ROSA26 locus using a PhiC31 integrase

1898 mediated cassette exchange approach in mouse ES cells. *PLoS One*, **6**,
1899 e23376.
1900 59. Qin,J.Y., Zhang,L., Clift,K.L., Hulur,I., Xiang,A.P., Ren,B.-Z. and Lahn,B.T.
1901 (2010) Systematic comparison of constitutive promoters and the
1902 doxycycline-inducible promoter. *PLoS One*, **5**, e10611.
1903 60. Hong,S., Hwang,D.D., Yoon,S., Isacson,O., Ramezani,A., Hawley,R.G. and
1904 Kim,K. (2007) Functional analysis of various promoters in lentiviral vectors at
1905 different stages of in vitro differentiation of mouse embryonic stem cells. *Mol.*
1906 *Ther.*, **15**, 1630–9.
1907 61. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users
1908 and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–86.
1909 62. Ye,J., Coulouris,G., Zaretskaya,I., Cutcutache,I., Rozen,S. and Madden,T.L.
1910 (2012) Primer-BLAST: A tool to design target-specific primers for
1911 polymerase chain reaction. *BMC Bioinformatics*, **13**, 134.
1912 63. Khanna,N., Bian,Q., Plutz,M. and Belmont,A.S. (2013) BAC manipulations for
1913 making BAC transgene arrays. *Methods Mol. Biol.*, **1042**, 197–210.
1914 64. Warming,S., Costantino,N., Court,D.L., Jenkins,N.A. and Copeland,N.G.
1915 (2005) Simple and highly efficient BAC recombineering using galK selection.
1916 *Nucleic Acids Res.*, **33**, e36.
1917 65. Bian,Q., Khanna,N., Alvikas,J. and Belmont,A.S. (2013) β-Globin cis-
1918 elements determine differential nuclear targeting through epigenetic
1919 modifications. *J. Cell Biol.*, **203**, 767–83.
1920 66. Strukov,Y.G. and Belmont, a. S. (2008) Development of Mammalian Cell
1921 Lines with lac Operator-Tagged Chromosomes. *Cold Spring Harb. Protoc.*,
1922 **2008**, pdb.prot4903-pdb.prot4903.
1923 67. Sambrook,J. and Russell,D.W. (2006) Purification of Nucleic Acids by
1924 Extraction with Phenol:Chloroform. *Cold Spring Harb. Protoc.*, **2006**,
1925 pdb.prot4455.
1926 68. Dernburg,A.F. (2011) Fragmentation and labeling of probe DNA for whole-
1927 mount FISH in Drosophila. *Cold Spring Harb. Protoc.*, **2011**, 1527–30.
1928 69. Solovei,I. and Cremer,M. (2010) 3D-FISH on cultured cells combined with
1929 immunostaining. *Methods Mol. Biol.*, **659**, 117–26.
1930 70. Cremer,M., Grasser,F., Lanctôt,C., Müller,S., Neusser,M., Zinner,R.,
1931 Solovei,I. and Cremer,T. (2008) Multicolor 3D fluorescence in situ
1932 hybridization for imaging interphase chromosomes. *Methods Mol. Biol.*, **463**,
1933 205–39.
1934 71. Beatty,B.G. and Scherer,S.W. (2002) Human chromosome mapping of single
1935 copy genes. *FISH A Pract. approach. B. Beatty, S. Mai, J. Squire, Ed.*
1936 *Oxford Univ. Press. Oxford*.
1937 72. Mowiol mounting medium (2006) *Cold Spring Harb. Protoc.*, **2006**,
1938 pdb.rec10255.
1939 73. Schindelin,J., Arganda-Carreras,I., Frise,E., Kaynig,V., Longair,M.,
1940 Pietzsch,T., Preibisch,S., Rueden,C., Saalfeld,S., Schmid,B., *et al.* (2012)
1941 Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, **9**,
1942 676–82.
1943 74. Khan,S.R. and Kuzminov,A. (2017) Degradation of RNA during lysis of

Escherichia coli cells in agarose plugs breaks the chromosome. *PLoS One*, **12**, e0190177.

75. Sambrook,J. and Russell,D.W. (2006) Preparation of DNA for Pulsed-field Gel Electrophoresis: Isolation of DNA from Mammalian Cells and Tissues. *Cold Spring Harb. Protoc.*, **2006**, pdb.prot4030.
76. Kimura,M., Stone,R.C., Hunt,S.C., Skurnick,J., Lu,X., Cao,X., Harley,C.B. and Aviv,A. (2010) Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths. *Nat. Protoc.*, **5**, 1596–607.
77. Sambrook,J. and Russell,D.W. (2006) Southern Hybridization of Radiolabeled Probes to Nucleic Acids Immobilized on Membranes. *Cold Spring Harb. Protoc.*, **2006**, pdb.prot4044.
78. Venkatraman,E.S. and Olshen,A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657–663.
79. Olshen,A.B., Venkatraman,E.S., Lucito,R. and Wigler,M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
80. Gibson,D.G., Young,L., Chuang,R.-Y., Venter,J.C., Hutchison,C.A. and Smith,H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343–5.
81. Shao,Z. and Zhao,H. (2012) DNA assembler: A synthetic biology tool for characterizing and engineering natural product gene clusters 1st ed. Elsevier Inc.
82. Shao,Z., Zhao,H.H. and Zhao,H.H. (2009) DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.*, **37**, e16.
83. Gietz,R.D. and Schiestl,R.H. (2007) Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.*, **2**, 38–41.
84. Schorpp,M., Jager,R., Schellander,K., Schenkel,J., Wagner,E.F., Weiher,H. and Angel,P. (1996) The Human Ubiquitin C Promoter Directs High Ubiquitous Expression of Transgenes in Mice. *Nucleic Acids Res.*, **24**, 1787–1788.
85. Mizushima,S. and Nagata,S. (1990) pEF-BOS, a powerful mammalian expression vector. *Nucleic Acids Res.*, **18**, 5322.
86. Lois,C., Hong,E.J., Pease,S., Brown,E.J. and Baltimore,D. (2002) Germline transmission and tissue-specific expression of transgenes delivered by lentiviral vectors. *Science*, **295**, 868–72.
87. Hong Cai,J., Deng,S., Kumpf,S., Lee,P., Zagouras,P., Ryan,A. and Gallagher,D. (2007) Validation of rat reference genes for improved quantitative gene expression analysis using low density arrays. *Biotechniques*, **42**, 503–512.
88. de Jonge,H.J.M., Fehrmann,R.S.N., de Bont,E.S.J.M., Hofstra,R.M.W., Gerbens,F., Kamps,W. a, de Vries,E.G.E., van der Zee,A.G.J., te Meerman,G.J. and ter Elst,A. (2007) Evidence based selection of housekeeping genes. *PLoS One*, **2**, e898.

89. Zhu,J., He,F., Song,S., Wang,J. and Yu,J. (2008) How many human genes can be defined as housekeeping with current expression data? *BMC Genomics*, **9**, 172.
90. She,X., Rohl,C. a, Castle,J.C., Kulkarni,A. V, Johnson,J.M. and Chen,R. (2009) Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics*, **10**, 269.
91. Zambrowicz,B.P., Imamoto, a, Fiering,S., Herzenberg,L. a, Kerr,W.G. and Soriano,P. (1997) Disruption of overlapping transcripts in the ROSA beta geo 26 gene trap strain leads to widespread expression of beta-galactosidase in mouse embryos and hematopoietic cells. *Proc. Natl. Acad. Sci. U. S. A.*, **94**, 3789–94.
92. Bertulat,B., de Bonis,M.L., Della Ragione,F., Lehmkuhl,A., Milden,M., Storm,C., Jost,K.L., Scala,S., Hendrich,B., D'Esposito,M., *et al.* (2012) MeCP2 Dependent Heterochromatin Reorganization during Neural Differentiation of a Novel Mecp2-Deficient Embryonic Stem Cell Reporter Line. *PLoS One*, **7**.
93. Hu,Y., Plutz,M. and Belmont,A.S. (2010) Hsp70 gene association with nuclear speckles is Hsp70 promoter specific. *J. Cell Biol.*, **191**, 711–9.
94. Sinclair,P., Bian,Q., Plutz,M., Heard,E. and Belmont,A.S. (2010) Dynamic plasticity of large-scale chromatin structure revealed by self-assembly of engineered chromosome regions. *J. Cell Biol.*, **190**, 761–76.
95. Hu,Y., Kireev,I., Plutz,M., Ashourian,N. and Belmont,A.S. (2009) Large-scale chromatin structure of inducible genes: transcription on a condensed, linear template. *J. Cell Biol.*, **185**, 87–100.
96. Schwab,M. and Amler,L.C. (1990) Amplification of cellular oncogenes: A predictor of clinical outcome in human cancer. *Genes, Chromosom. Cancer*, **1**, 181–193.
97. Carroll,S.M., DeRose,M.L., Gaudray,P., Moore,C.M., Needham-Vandevanter,D.R., Von Hoff,D.D. and Wahl,G.M. (1988) Double minute chromosomes can be produced from precursors derived from a chromosomal deletion. *Mol. Cell. Biol.*, **8**, 1525–33.
98. L'Abbate,A., Macchia,G., D'Addabbo,P., Lonoce,A., Tolomeo,D., Trombetta,D., Kok,K., Bartenhagen,C., Whelan,C.W., Palumbo,O., *et al.* (2014) Genomic organization and evolution of double minutes/homogeneously staining regions with MYC amplification in human cancer. *Nucleic Acids Res.*, **42**, 9131–45.
99. Marasini,D. and Fakhr,M. (2014) Exploring PFGE for Detecting Large Plasmids in Campylobacter jejuni and Campylobacter coli Isolated from Various Retail Meats. *Pathogens*, **3**, 833–844.
100. Barton,B.M., Harding,G.P. and Zuccarelli,A.J. (1995) A general method for detecting and sizing large plasmids. *Anal. Biochem.*, **226**, 235–40.
101. Walker,P.R., LeBlanc,J. and Sikorska,M. (1997) Evidence that DNA fragmentation in apoptosis is initiated and propagated by single-strand breaks. *Cell Death Differ.*, **4**, 506–515.
102. Richardson,S.M., Mitchell,L.A., Stracquadanio,G., Yang,K., Dymond,J.S., DiCarlo,J.E., Lee,D., Huang,C.L.V., Chandrasegaran,S., Cai,Y., *et al.* (2017)

2036 Design of a synthetic yeast genome. *Science*, **355**, 1040–1044.
2037 103. Nanbo,A., Sugden,A. and Sugden,B. (2007) The coupling of synthesis and
2038 partitioning of EBV's plasmid replicon is revealed in live cells. *EMBO J.*, **26**,
2039 4252–62.
2040 104. Kouprina,N., Earnshaw,W.C., Masumoto,H. and Larionov,V. (2013) A new
2041 generation of human artificial chromosomes for functional genomics and
2042 gene therapy. *Cell. Mol. Life Sci.*, **70**, 1135–48.
2043 105. Mills,W., Critcher,R., Lee,C. and Farr,C.J. (1999) Generation of an
2044 approximately 2.4 Mb human X centromere-based minichromosome by
2045 targeted telomere-associated chromosome fragmentation in DT40. *Hum.*
2046 *Mol. Genet.*, **8**, 751–61.
2047 106. Harrington,J.J., Van Bokkelen,G., Mays,R.W., Gustashaw,K. and
2048 Willard,H.F. (1997) Formation of de novo centromeres and construction of
2049 first-generation human artificial microchromosomes. *Nat. Genet.*, **15**, 345–
2050 55.
2051 107. Kazuki,Y. and Oshimura,M. (2011) Human artificial chromosomes for gene
2052 delivery and the development of animal models. *Mol. Ther.*, **19**, 1591–601.
2053 108. Kim,J.-H., Kononenko,A., Erliandri,I., Kim,T.-A., Nakano,M., Iida,Y.,
2054 Barrett,J.C., Oshimura,M., Masumoto,H., Earnshaw,W.C., *et al.* (2011)
2055 Human artificial chromosome (HAC) vector with a conditional centromere for
2056 correction of genetic deficiencies in human cells. *Proc. Natl. Acad. Sci. U. S.*
2057 *A.*, **108**, 20048–53.
2058 109. Hiratsuka,M., Uno,N., Ueda,K., Kurosaki,H., Imaoka,N., Kazuki,K., Ueno,E.,
2059 Akakura,Y., Katoh,M., Osaki,M., *et al.* (2011) Integration-free iPS cells
2060 engineered using human artificial chromosome vectors. *PLoS One*, **6**,
2061 e25961.
2062 110. Takahashi,Y., Tsuji,S., Kazuki,Y., Noguchi,M., Arifuku,I., Umebayashi,Y.,
2063 Nakanishi,T., Oshimura,M. and Sato,K. (2010) Development of evaluation
2064 system for bioactive substances using human artificial chromosome-
2065 mediated osteocalcin gene expression. *J. Biochem.*, **148**, 29–34.
2066 111. Kazuki,Y., Hiratsuka,M., Takiguchi,M., Osaki,M., Kajitani,N., Hoshiya,H.,
2067 Hiramatsu,K., Yoshino,T., Kazuki,K., Ishihara,C., *et al.* (2010) Complete
2068 genetic correction of ips cells from Duchenne muscular dystrophy. *Mol.*
2069 *Ther.*, **18**, 386–93.
2070 112. Larin,Z. and Mejía,J.E. (2002) Advances in human artificial chromosome
2071 technology. *Trends Genet.*, **18**, 313–319.
2072 113. Liskovykh,M., Lee,N.C., Larionov,V. and Kouprina,N. (2016) Moving toward
2073 a higher efficiency of microcell-mediated chromosome transfer. *Mol. Ther.*
2074 *Methods Clin. Dev.*, **3**, 16043.
2075 114. Fournier,R.E. and Ruddle,F.H. (1977) Microcell-mediated transfer of murine
2076 chromosomes into mouse, Chinese hamster, and human somatic cells.
2077 *Proc. Natl. Acad. Sci. U. S. A.*, **74**, 319–23.
2078 115. Shimizu,N., Miura,Y., Sakamoto,Y. and Tsutsui,K. (2001) Plasmids with a
2079 mammalian replication origin and a matrix attachment region initiate the
2080 event similar to gene amplification. *Cancer Res.*, **61**, 6987–6990.
2081 116. Shimizu,N., Hashizume,T., Shingaki,K. and Kawamoto,J.K. (2003)

Amplification of plasmids containing a mammalian replication initiation region is mediated by controllable conflict between replication and transcription. *Cancer Res.*, **63**, 5281–5290.

117. Shimizu,N., Shingaki,K. and Kaneko-sasaguri,Y. (2005) When , where and how the bridge breaks : anaphase bridge breakage plays a crucial role in gene amplification and HSR generation. **302**, 233–243.

118. Shimizu,N. (2009) Extrachromosomal double minutes and chromosomal homogeneously staining regions as probes for chromosome research. *Cytogenet. Genome Res.*, **124**, 312–326.

119. Schoenlein,P. V, Shen,D.W., Barrett,J.T., Pastan,I. and Gottesman,M.M. (1992) Double minute chromosomes carrying the human multidrug resistance 1 and 2 genes are generated from the dimerization of submicroscopic circular DNAs in colchicine-selected KB carcinoma cells. *Mol Biol Cell*, **3**, 507–520.

120. Khanna,N., Hu,Y. and Belmont,A.S.S. (2014) HSP70 Transgene Directed Motion to Nuclear Speckles Facilitates Heat Shock Activation. *Curr. Biol.*, **24**, 1138–1144.

Table 1

Construct	Heterogeneous clones%	Number of clones
UBC-GFP-ZeoR	0	58
DHFR-UG	60%	30
ROSA-UG	76%	38
UBB-UG	58%	41
2207K13-UG	69%	35
HBB-UG	83%	23

Table 2

Sample Name	BAC copy number per cell	Episome copy number per cell	BAC copy number per episome	BAC size (kb)	BAC content per episome (kb)
DHFR-UG-s3	15.4	6.2 (n=99)	2.5	178	445
HBB-UG-100d3	14.9	4.5 (n=100)	3.3	217	716

Table 3

Sample Name	BAC copy number per cell	Estimated episome copy number per cell	Minimum copy number increase of episome-localizing DNA relative to L
DHFR-UG-s3_H1	49.4	19.8	5.8
DHFR-UG-s3_H2	57.5	23.1	6.6
DHFR-UG-s3_L	0.3	0.1	/
HBB-UG-100d3_H1	48.5	14.7	4.6
HBB-UG-100d3_H2	51.2	15.5	4.8
HBB-UG-100d3_L	0.2	0.1	/

Figure 1

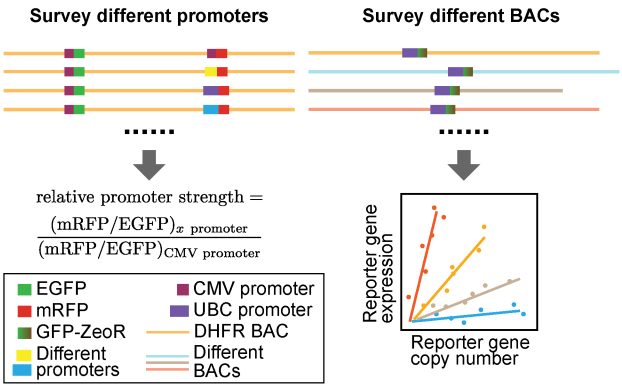


Figure 2

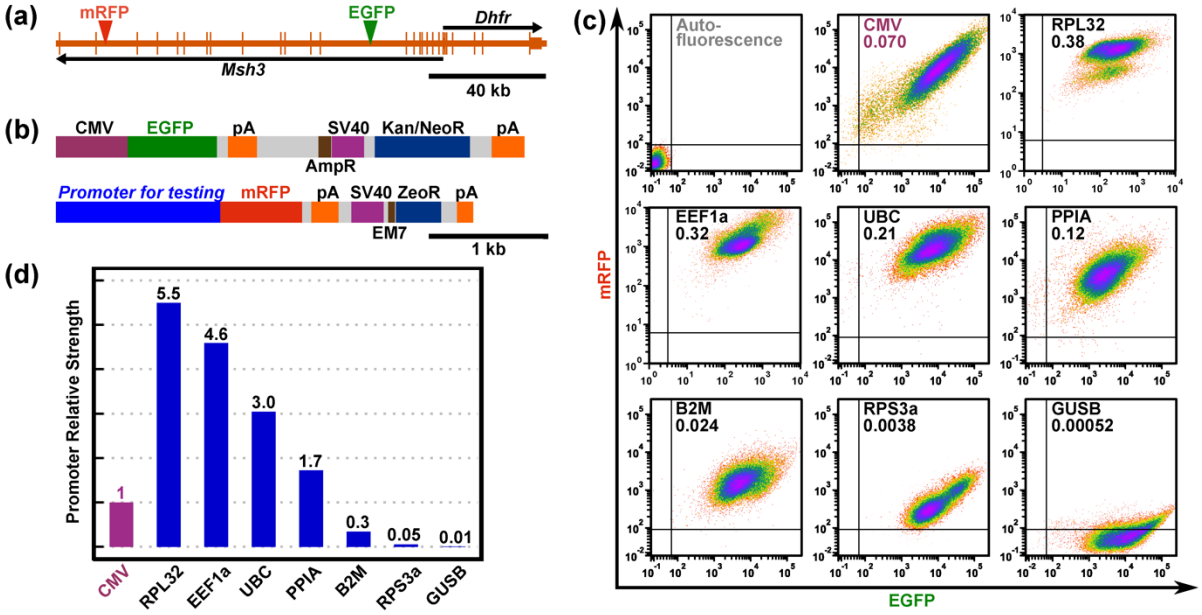


Figure 3

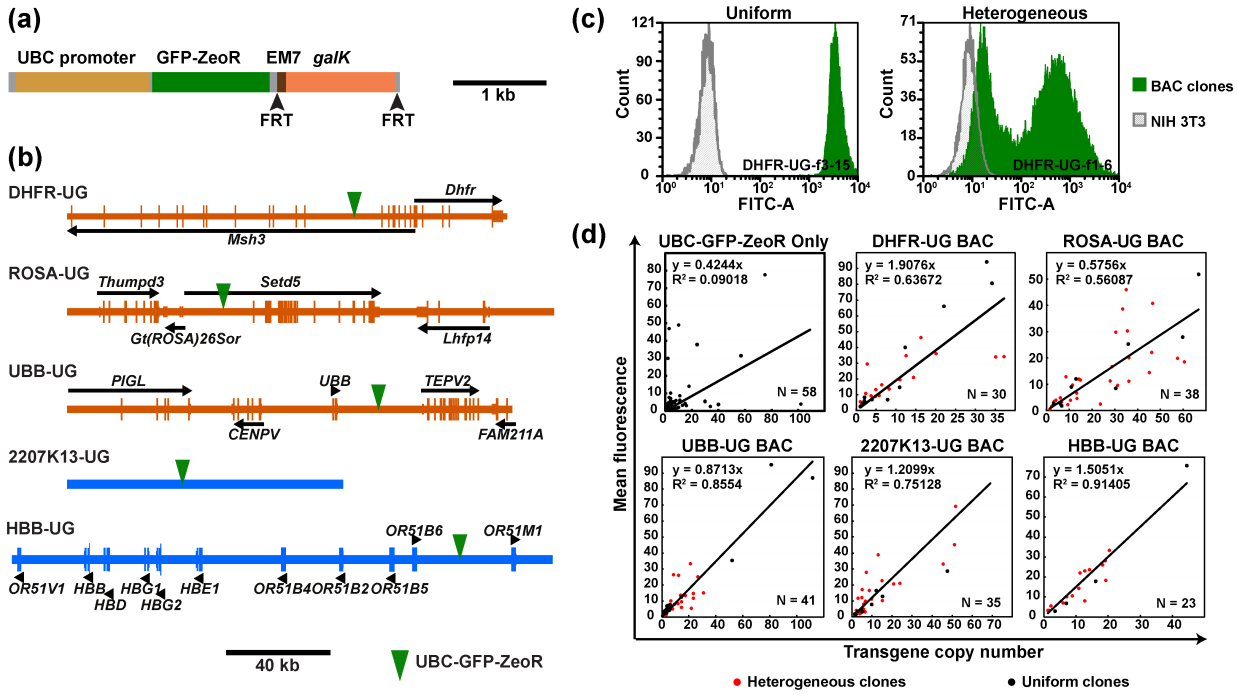

(a)
UBC promoter
GFP-ZeoR
EM7
galK
FRT
FRT
1 kb
(b)
DHFR-UG
Dhfr
Msh3
ROSA-UG
Thumpd3
Setd5
Gt(ROSA)26Sor
Lhfp14
UBB-UG
PIGL
UBB
TEPV2
CENPV
FAM211A
2207K13-UG
HBB-UG
OR51B6
OR51M1
OR51V1
HBB
HBG1
HBE1
OR51B4
OR51B2
OR51B5
HBD
HBG2
40 kb
UBC-GFP-ZeoR
(c)
Uniform
Heterogeneous
Count
FITC-A
DHFR-UG-f3-15
DHFR-UG-f1-6
BAC clones
NIH 3T3
(d)
UBC-GFP-ZeoR Only
y = 0.4244x
R² = 0.09018
N = 58
DHFR-UG BAC
y = 1.9076x
R² = 0.63672
N = 30
ROSA-UG BAC
y = 0.5756x
R² = 0.56087
N = 38
UBB-UG BAC
y = 0.8713x
R² = 0.8554
N = 41
2207K13-UG BAC
y = 1.2099x
R² = 0.75128
N = 35
HBB-UG BAC
y = 1.5051x
R² = 0.91405
N = 23
Mean fluorescence
Transgene copy number
Heterogeneous clones
Uniform clones

Figure 4

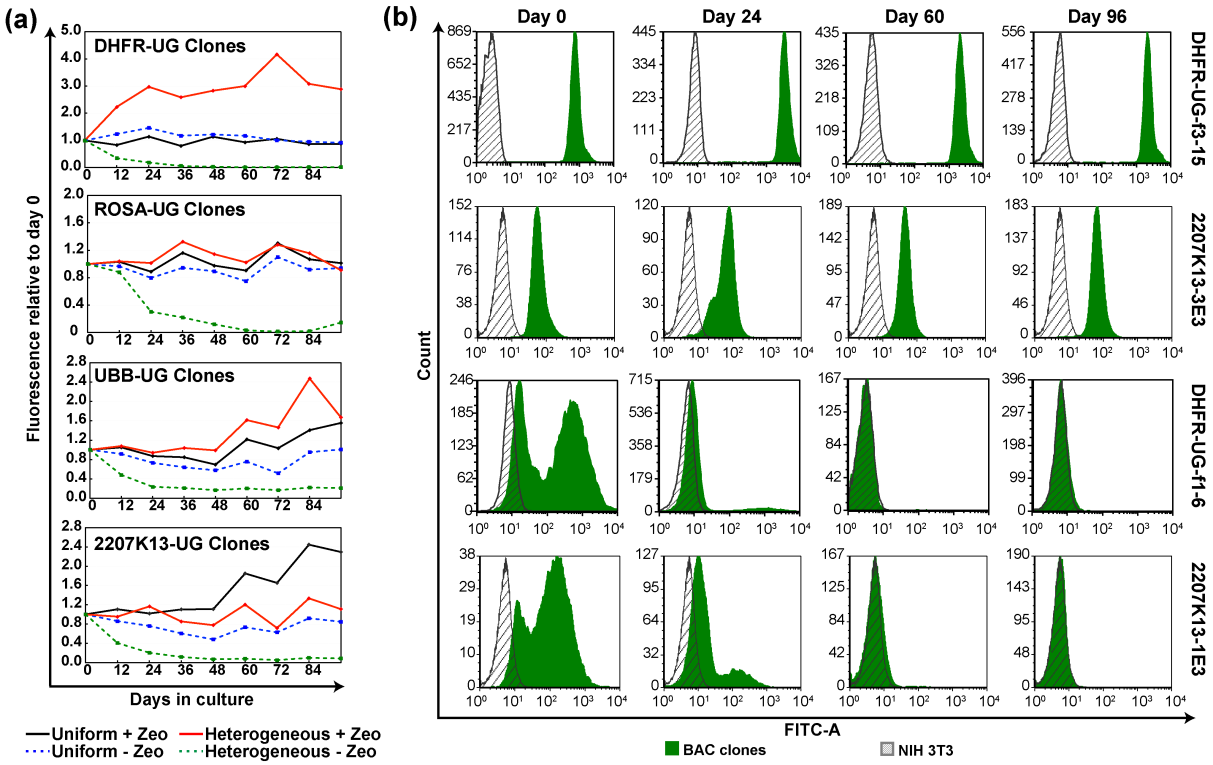

(a)
DHFR-UG Clones
ROSA-UG Clones
UBB-UG Clones
2207K13-UG Clones
Fluorescence relative to day 0
Days in culture
Uniform + Zeo
Heterogeneous + Zeo
Uniform - Zeo
Heterogeneous - Zeo
(b)
Day 0
Day 24
Day 60
Day 96
DHFR-UG-f3-15
2207K13-3E3
DHFR-UG-f1-6
2207K13-1E3
Count
FITC-A
BAC clones
NIH 3T3

Figure 5

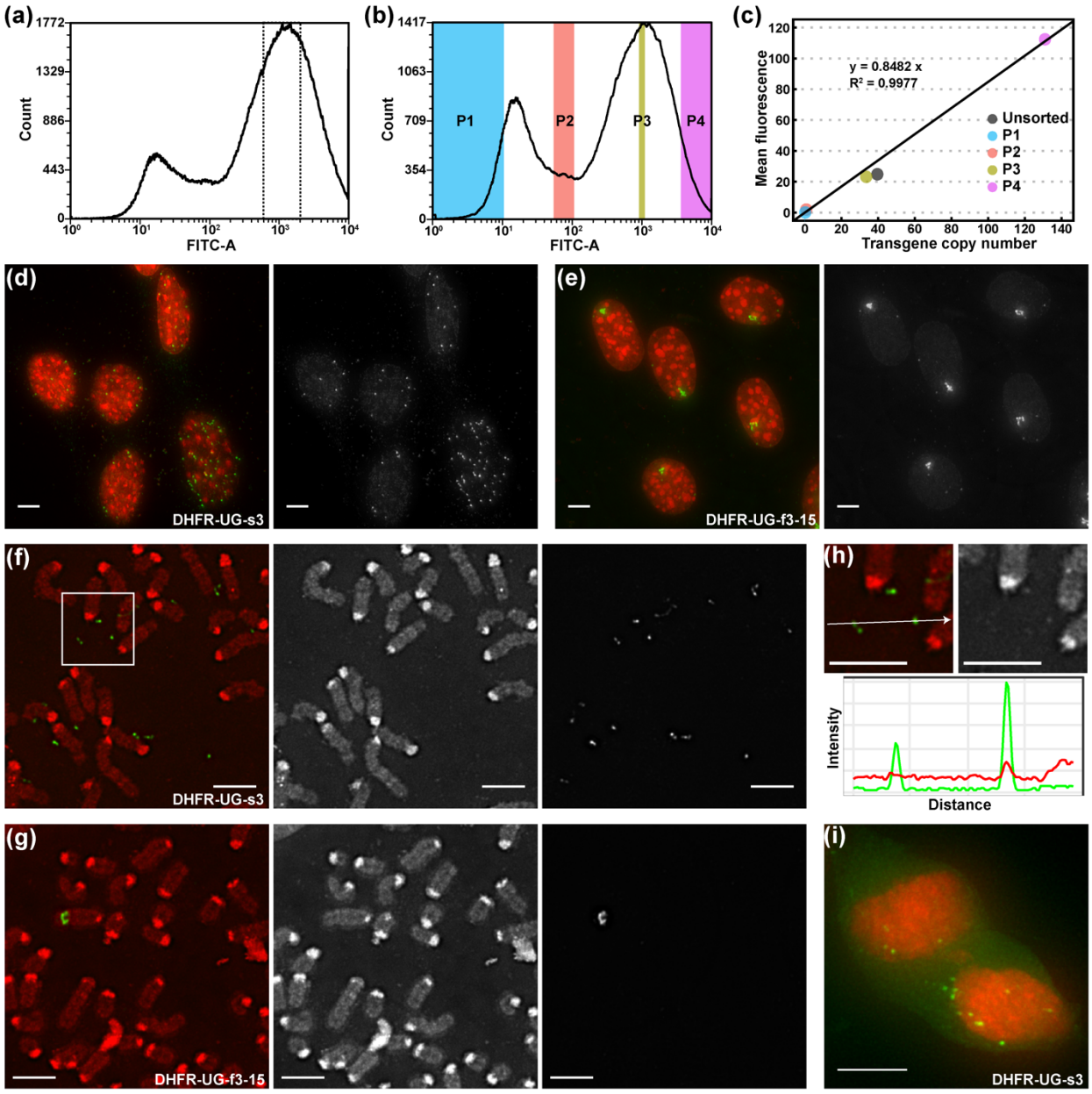

(a)
Count
FITC-A
(b)
P1
P2
P3
P4
Count
FITC-A
(c)
y = 0.8482 x
R² = 0.9977
Unsorted
P1
P2
P3
P4
Mean fluorescence
Transgene copy number
(d)
DHFR-UG-s3
(e)
DHFR-UG-f3-15
(f)
DHFR-UG-s3
(h)
Intensity
Distance
(g)
DHFR-UG-f3-15
(i)
DHFR-UG-s3

Figure 6

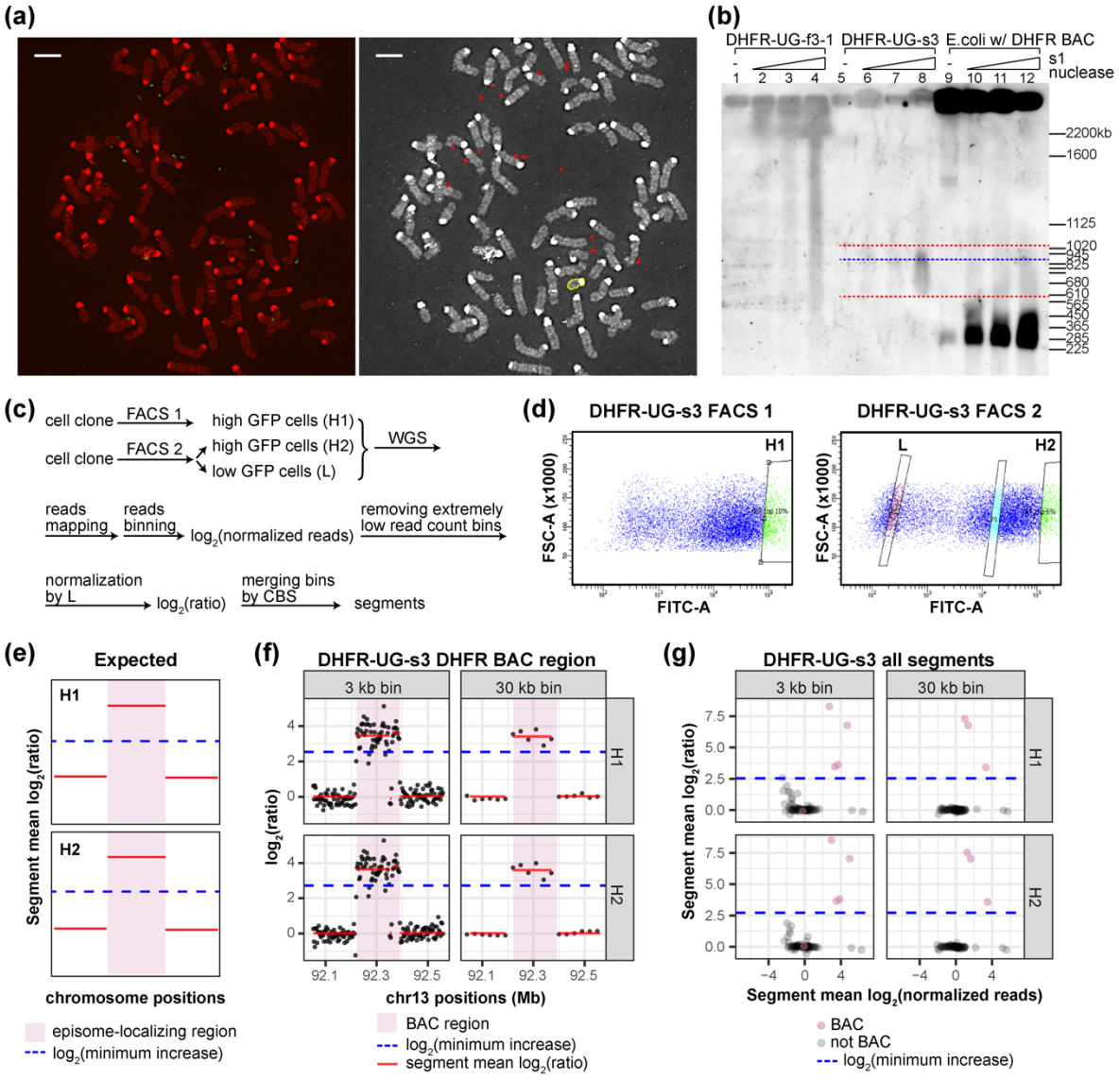
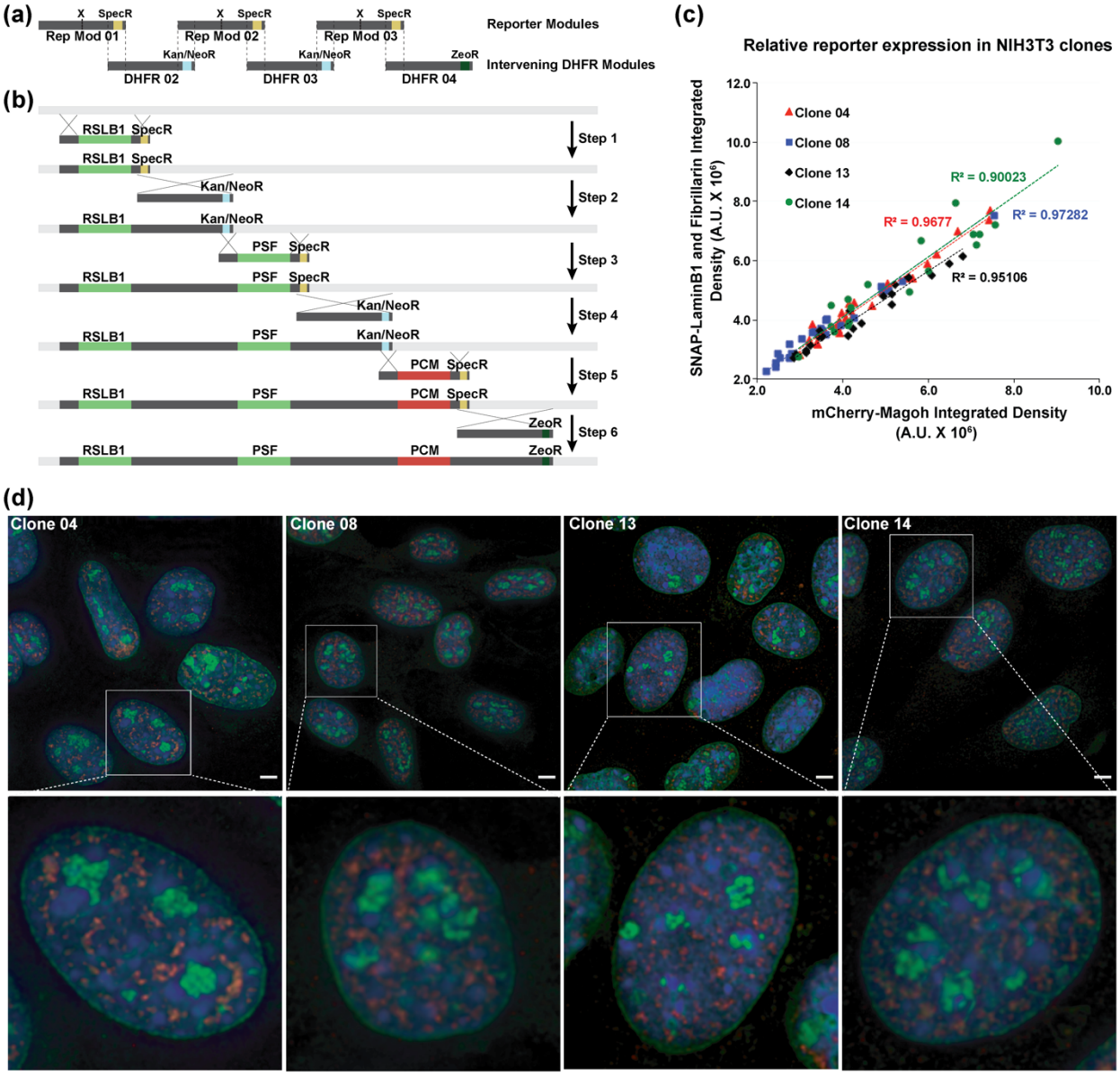

(a)
(b)
DHFR-UG-f3-1
DHFR-UG-s3
E.coli w/ DHFR BAC
s1 nuclease
1 2 3 4 5 6 7 8 9 10 11 12
2200kb
1600
1125
1020
945
825
680
610
565
450
365
285
225
(c)
cell clone
FACS 1
high GFP cells (H1)
cell clone
FACS 2
high GFP cells (H2)
low GFP cells (L)
WGS
reads mapping
reads binning
log2(normalized reads)
removing extremely low read count bins
normalization by L
log2(ratio)
merging bins by CBS
segments
(d)
DHFR-UG-s3 FACS 1
DHFR-UG-s3 FACS 2
H1
L
H2
FSC-A (x1000)
FITC-A
(e)
Expected
H1
H2
Segment mean log2(ratio)
chromosome positions
episome-localizing region
log2(minimum increase)
(f)
DHFR-UG-s3 DHFR BAC region
3 kb bin
30 kb bin
log2(ratio)
chr13 positions (Mb)
BAC region
log2(minimum increase)
segment mean log2(ratio)
(g)
DHFR-UG-s3 all segments
3 kb bin
30 kb bin
Segment mean log2(ratio)
Segment mean log2(normalized reads)
BAC
not BAC
log2(minimum increase)

Figure 7


(a)
X SpecR
Rep Mod 01
Rep Mod 02
Rep Mod 03
Reporter Modules
Kan/NeoR
ZeoR
DHFR 02
DHFR 03
DHFR 04
Intervening DHFR Modules
(b)
RSLB1 SpecR
Step 1
Kan/NeoR
Step 2
RSLB1
PSF SpecR
Step 3
Step 4
PCM SpecR
Step 5
ZeoR
Step 6
PCM
(c)
Relative reporter expression in NIH3T3 clones
Clone 04
Clone 08
Clone 13
Clone 14
R² = 0.90023
R² = 0.9677
R² = 0.97282
R² = 0.95106
SNAP-LaminB1 and Fibrillarin Integrated Density (A.U. X 10⁶)
mCherry-Magoh Integrated Density (A.U. X 10⁶)
(d)
Clone 04
Clone 08
Clone 13
Clone 14