Whole-genome epigenetic function annotation through sgRNA libraries synthesized by controlled template-dependent elongation

Junling Jia (junling.jia@alumni.bcm.edu)
The First Affiliated Hospital, Zhejiang University
Chen Pan
ZheJiang Univeristy
Ran Li
life sciences institute zhejiang university
Liyan Shui
Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital
Yali Wang
life sciences institute zhejiang university
Zhengyun Xiao
zhejiang university
Jing Zhu
Mingtian Genetics
Chao Wu
Zhe Jiang Univerisity https://orcid.org/0000-0002-6193-4398
Min Zheng
ZheJiang Univeristy

Article

Keywords: Genetic screen, CRISPR/Cas9, sgRNA library, H3K4me3, CTCF, mESCs, Liver cancer, LincRNA, CDC42

Posted Date: November 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-100657/v1

Whole-genome epigenetic function annotation through sgRNA libraries synthesized by controlled template-dependent elongation

Chen Pan[1,2,#], Ran Li[1,2,#], Liyan Shui[2,#], Yali Wang[1], Zhengyun Xiao[1], Jing Zhu[3], Chao Wu[2,*], Min Zheng [2,*] and Junling Jia[2,4,*]

1. Life Sciences Institute, Zhejiang University, Hangzhou, Zhejiang, 310058, PRC
2. Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, The First Affiliated Hospital, Zhejiang University, Hangzhou, Zhejiang 310003, PRC Zhejiang University, Hangzhou, Zhejiang 310003, PRC
3. Department of Anesthesiology & Center for Shock, Trauma and Anesthesiology Research, University of Maryland School of Medicine, Baltimore, MD, 21201, USA
4. Innovation Center for Precision Medicine, Zhongtong-Lanbo Diagnostic LTD, Beijing, 100166, PRC

*. Correspondence: minzheng@zju.edu.cn (M.Z.), wuchao1984@zju.edu.cn (C.W.), Junling.jia@alumni.bcm.edu (J.J.)

# These authors contributed equally

Keywords: Genetic screen, CRISPR/Cas9, sgRNA library, H3K4me3, CTCF, mESCs, Liver cancer, LincRNA, CDC42

Abstract

Epigenome is the set of DNA-associated proteins or chemical modifications to DNA, which regulates gene expression in processes of development and disease. While current advances have allowed researchers to routinely profile epigenomes from given samples, our understandings of the functions of epigenetic hallmarks are nonspecific at best. Applying CRISPR-screening to genome-widely interrogate the function of individual epigenetic hallmarks demands massive sgRNA libraries which are unaffordable via commercial syntheses. Our development consists of a high throughput and cost-effective controlled template-dependent elongation (CTDE) approach which converts source DNA to sgRNA templates. Affiliated screenings encompass 3.8M sgRNAs generated by CTDE targeting all major H3K4me3 and CTCF hallmarks in mESCs and HepG2 and identified 20K essential epigenetic hallmarks, which render the first batch of functional epigenome annotation of H3K4me3 and CTCF hallmarks in mammals. As an application example, we show that a H3K4me3 hallmark orchestrates CDC42 level and cell-cycle progression through promoting LINC00339 expression in HepG2.

Introduction

Humans have over 20,000 protein-coding genes, which account for about 2% of the overall genome DNA[1]. The expression of those genes are modulated by elements in the remaining portions of the genome, which are regarded as the regulome - an assortment of chromatin components, such as promoters, transcriptional regulatory regions, and high dimensional chromatin structures[1,2]. Being a major component of regulome, epigenome contains the set of DNA-associated proteins or chemical modifications to DNA[3]. Systematic discovery and location annotation of epigenome are accomplished primarily by Consortiums such as ENCODE and Roadmap Epigenomics Mapping through assays such as DNase hypersensitivity assays, DNA methylation assays and chromatin immunoprecipitation sequencing (ChIP-seq) assays[4-8]. These refined techniques have allowed most labs to routinely profile their interested epigenomic information under deliberate experimental conditions.

The function of many epigenetic modifications is generally known. For example, H3K4me3 (tri-methylation at the 4th lysine residue of the DNA packaging protein Histone H3) is involved in the positive regulation of the nearby gene transcription[9,10]. H3K4me3 plays a significant role regarding the regulation of stem cell lineage potency[11] as well. CTCF hallmarks primarily have a CCCTC-like motif and are bound by CTCF protein, a 11-zinc finger protein[12]. Moreover, CTCF hallmarks are involved in various cellular processes such as transcriptional regulation, insulator activity, V(D)J recombination and the regulation of chromatin architecture[13-17].

Mouse embryonic stem cells (mESCs) have around 42K (23.8Mb) H3K4me3 hallmarks and 37K (10.6Mb) CTCF hallmarks. Human liver cancer cells (HepG2) have around 41K (46.9Mb) H3K4me3 hallmarks and 28K (5.2 Mb) CTCF hallmarks. However, there is still a scarcity of a functional epigenome annotation of these hallmarks considering genetic screens have prioritized protein-coding genes or expressed non-coding loci[18,19].

Numerous reports have depicted pooled CRISPR-screenings that interrogate the function of regulatory elements using designed dense tiling sgRNA libraries targeting limited genomic loci[8,20,21]. While these studies have provided proof of concept for the application of pooled CRISPR-screening in the functional characterization of regulatory elements, the high-cost of synthesizing a dense tiling sgRNA library to cover an epigenetic hallmark genome-widely (around half-million USD for H3K4me3 or CTCF) is unrealistic and a major hurdle to further functional epigenome studies.

We have developed a simple- and cost-effective controlled template-dependent elongation (CTDE) approach that can convert any DNA sample to a sgRNAs library, which covers 98.47% of the effective CRISPR/Cas9 targeting sites within the source DNA. Significantly, over 99% CTDE-sgRNAs targeting sequences have a protospacer adjacent motif (PAM)[22-24]. We have generated sgRNA libraries targeting all H3K4me3 and CTCF hallmarks in mESCs and HepG2. In total, we have screened 3.8 M sgRNAs and identified 14K (14265) H3K4me3 and 6K (6235) CTCF essential hallmarks for the proliferation of mESCs and HepG2. mESCs CTCF dataset shows that mESCs

maintains a high proportion of non-essential cell-type specific CTCF hallmarks, which may be important for the implement of pluripotency. Importantly, the HepG2 H3K4me3 dataset helps confirm that an essential H3K4me3 hallmark inside the intron of LINC00339 orchestrates the cell-cycle progression and the expression of CDC42, a pivotal factor for the proliferation and invasion of cancer cells, through promoting LINC00339 expression.

Our studies have brought us to develop an efficient and budget-friendly approach (CTDE) to convert DNA to a sgRNA library. Several breakthroughs have ensued, one of them being the first batch of functional epigenome annotation of H3K4me3 and CTCF hallmarks in mammals through CTDE library coupled CRISPR-screening. Various other findings through characterizing an essential H3K4me3 hallmark in HepG2 have shown the significance of functional epigenome annotation in cancer research.

Result

Convert DNA to a sgRNA library through synthesized by controlled template-dependent elongation (CTDE)

The CRISPR/Cas9 system has been developed into genome mutating tools with wide-ranging applications[25]. The single guide RNA (sgRNA) binds to the Cas9 enzyme and guides the complex to a target via pairing the complementary DNA sequences followed by a protospacer adjacent motif (PAM), where Cas9 performs its endonuclease activity immediately[22,25]. Currently, sgRNA templates are designed as 17-22bp DNA sequences and commercially synthesized *in vitro* as ready-to-clone fragments[25-27].

To directly generate large-scale sgRNA templates from source DNA, we fragmented the DNA template within a 1kb length (Figure 1a). We then ligated the DNA fragments with A1 adaptors (red), immobilizing them on streptavidin beads and washing away their positive strands under a denaturing condition (Figure 1a). Next, we annealed the priming primer (the positive strand of A1 adaptor) onto the immobilized minus strand, then extending the primer using DNA polymerase coupled with reversible terminator (RT) nucleotides (3'-O-N_3-dNTP) that allow singular nucleotide incorporation before the restoration of their 3′-hydroxy groups (Figure 1a)[28]. We restored the 3′hydroxyl group via tris (2-carboxyethyl) phosphine (TCEP) treatment and repeated another round of nucleotide incorporation following the previous extension (Figure 1a)[28]. After 23 rounds of cycling, we blunt the 3′ terminus to get 23bp DNA fragments (not including the adapter) (Figure 1a). The most used Cas9 from

126 *Streptococcus pyogenes* recognizes the 5′-NGG-3′ (where "N" can be any nucleotide
127 base) PAM sequence. If the last two nucleotides of 3′end of the 23bp DNA are GG, it
128 will compose an AscI cutting site after the A2 adaptor ligation, which is used to select
129 the DNA fragment with 5′-NGG-3′ PAM (Figure 1a). Following PAM selection, we
130 remove the NGG triplet using a type II restrict endonuclease (BbsI) and put an A4
131 adaptor onto the 3′ terminus for further Gibson assembly into a sgRNA expressing
132 vector (Figure 1a)[29]. Since only two rare-cutting endonucleases are employed, the
133 dropout rate of sgRNA template caused by endonuclease cutting is very low (0.86%
134 per mouse genome and 0.82% per human genome).

135 To test the efficiency of this technique, we used a 14.9 kb plasmid (lentiCRISPR-
136 v2) modeling a DNA template. After implementing the CTDE steps described above
137 (Figure 1a and S1a), we generated sgRNAs targeting 98.47% of the sites with 5′-NGG-
138 3′ PAM sequence (Figure 1b and S1b). As expected, few sgRNAs can be generated
139 from AT rich region because of the low complimentary binding affinity between AT
140 rich sequences (Figure S1b). We also checked the capability of CTDE to enrichment
141 the DNA fragments with 3′ NGG triplet. There are 6.23% input DNA fragments that
142 are adjacent to a 3′ NGG triplet, and after enrichment, the rate is raised to 99.7% (15.98
143 folds) (Figure 1c). As expected, the lengths of the sgRNA templates are predominantly
144 20bp (83.93%) and the functional sgRNA templates (17-22bp) are 98.66% (Figure
145 1d)[26,27].

An evaluation of the faithful of the CTDE library ensued the CTDE library was compared with the popular sgRNA libraries (GeCKO 2.0 human/mouse) generated by Chip Based DNA Synthesis (CBDS). Typically, 96.4% sgRNAs of CTDE library can perfectly match a position in their template, and 1.6% sgRNAs of CTDE library carry one mismatch (Figure 1e). The error rate of the CTDE procedure is around 0.97 bases per 1000 bases. Around 90% sgRNAs of CBDS can perfectly match their targeting positions and around 7.5%/9.5% sgRNAs carry one mismatch (Figure 1e). The CBDS library, having an error rate of around 4.5-5.7 bases per 1000 bases, assimilates itself with the CTDE library. We also compared the amplification bias between CTDE library and CBDS library. The sgRNA abundance of CBDS library and CTDE library are both similarly low (Figure 1f). Treating of the abovesaid data, CTDE, consistent in the conversion of source DNA to sgRNA library, produces the same qualitative results as the library generated via Chip Based DNA Synthesis (CBDS).

Validations of CTDE sgRNA library screening

To implement mega-level screenings, we planned to infect cells with the lenti-viral CTDE library expressing sgRNAs along with Cas9 protein (around 40K sgRNAs per batch) (Figure 2a). Then, we sequence the abundance of each sgRNA template at two time points (3 days selected in puromycin media as P1 and 20 days expanded after puromycin selection as P10) to calculate the abundance change from P1 to P10 of each sgRNA template (Figure 2a; details in method).

Current bioinformatics efforts for the analysis of pooled CRISPR screens are devoted to identifying genes rather than non-coding genomic loci[30]. A coding gene is targeted usually by multiple sgRNAs with similar mutational abilities in traditional CRISPR screens; and the essentiality of genes are evaluated through the integrative analysis abundance change of these sgRNAs in frequented algorithms, such as MAGeCK[31]. For the purpose, however, of screening non-coding regulatory loci inside epigenetic hallmarks using CTDE library, the objective becomes the identification of narrow essential genomic sites, the majority of said can only be efficiently targeted by one sgRNA. Traditional calling algorithms are unreliable in this scenario as it will report numerous false positive significantly changed sgRNA (ssgRNA) (Figure 2b).

During the screening, many minor uncertain factors bias the abundance of each sgRNA randomly and cause the abundance change of the sgRNAs following a normal distribution in the scenario of non-selection pressure[32]. Some sgRNA disruptions will cause a negative selection pressure, bias their abundance in P10 systematically (<20%), and result in their abundance change following another normal distribution. While the distance of above two normal distribution is large enough (Figure S2a), we can efficiently identify the sgRNAs (FDR<0.1) that leads to negative selections using a self-developed straight-forward approach (NSgRNAShot; details in method), which has 96.42% precision rate and 96.27% recall rate on a simulation dataset (Figure S2a; Table S1). Most importantly, on identical testing datasets, NSgRNAShot has significantly lower false positive rates than MAGeCK (Figure 2b). Additionally, we compared the

187 ability of NSgRNAShot to call true positive ssgRNA with MAGeCK using two
188 published essential gene screen datasets. Due to significant false positive rates,
189 MAGeCK reports a noticeably higher amount of ssgRNAs than NSgRNAShot, most of
190 which target non-essential genes (Figure 2b-c), whereas major portions of ssgRNA
191 identified by NSgRNAShot target essential genes and overlap (100%) with ssgRNAs
192 targeting essential genes reported by MAGeCK (Figure 2c).

193 Following this, through targeting a drug resistance gene (Neo) in mESCs (Figure
194 S2b), we compared the efficiency of CTDE library with the library designed via
195 standard CBDS methods (details in method). The CBDS library has 115 sgRNAs
196 targeting all possible sites (NGG PAM) in Neo gene, along with 20 non-targeting
197 control gRNAs. The CDTE library was generated from the abovesaid Neo gene
198 fragments with additions of the same 20 non-targeting control gRNAs (Figure S2b).
199 The mESCs (expressing Neo) were infected with above lenti-viral libraries and were
200 cultured in medium with Neomycin, and the screening was performed according to the
201 abovesaid procedure. As expected, the abundance of Neo targeting sgRNAs of both
202 libraries exhibit significant decrease and can be identified as ssgRNAs (Figure 2d).
203 Most importantly, the ssgRNAs from both libraries are highly overlapped (97%)
204 (Figure 2d), exhibiting implications that sgRNA library generated by CTDE performs
205 similarly, if not better as the library produced by standard method such as CBDS. Taken
206 together, multiple validations have shown that our CTDE sgRNA library screening
207 approach is both practicable and efficient.

208

209 **Annotation of the essential H3K4me3 hallmarks for mESCs self-renewal**

210 Tri-methylation of Histone H3 at Lysine 4 (H3K4me3) hallmarks functions in
211 transcriptional activation and maintaining bivalent chromatin of mESCs[11]. Previous
212 small-scale CRISPR/Cas9-based screens of *Pou5f1* promoter in human ES cells show
213 that numerous critical regulatory elements exist inside H3K4me3 hallmarks and that
214 they are of great important regarding our understanding of gene regulation mechanisms
215 under a biological context[21,33].

216 To genome-widely interrogate essential regulatory elements within H3K4me3
217 hallmarks during mESCs self-renewal, we first acquired the H3K4me3 labeled DNA
218 fragments through chromatin immunoprecipitation (ChIP) and converted them to
219 sgRNA libraries through CTDE (Figure 1a). To implement the screening, we infected
220 the mESCs with the lenti-viral library expressing sgRNAs along with Cas9 protein
221 (around 40K sgRNAs per batch). Then, we performed the screening as described above
222 (Figure 2a).

223 In total, we screened 926K sgRNAs targeting 82.99% of the H3K4me3 enriched
224 regions (Figure 3a-d; Table S2-3). In H3K4me3 highly enriched regions (top 100), the
225 sgRNA density is high and reaches to 59 sgRNA per kb. The abundance distribution of
226 sgRNAs inside H3K4me3 hallmarks observes the pattern of their template DNAs
227 (Figure 3a-b and S3a). We identified 24189 sgRNAs causing significant negative
228 selection (called ssgRNA from here), which indicates that their targeting sites in

H3K4me3 hallmarks are essential regulatory elements for mESCs self-renewal (Figure 3a-d; Table S4). We verified three randomly picked ssgRNAs in non-coding regions. As expected, all three ssgRNAs can significantly inhibit the proliferation of mESCs (Figure 3e and S3b-c).

H3K4me3 ssgRNAs distribute evenly on most chromosomes, regions on chromosome 6, 9 and X being an exception (Figure 3a). Given these biases are not the results of lacking matrix H3K4me3 hallmarks or sgRNAs (Figure 3a), we reason that these regions contain fewer essential genes regulated by H3K4me3 hallmarks. ssgRNAs appear in regions, arrayed from weak to strong H3K4me3 elements, the majority of which are among strong elements regions (Figure 3b). Among all essential H3K4me3 elements, 63.24% of the elements are targeted by 1 ssgRNA, and the remaining 37.76% are targeted by multiple ssgRNAs (Table S3). Detailed positions of ssgRNAs inside their H3K4me3 elements can be found in the Supplemental table 4. Given that the H3K4me3 elements are generally wide, the location of these ssgRNAs should reflect the core regulatory sites within the elements.

Many H3K4me3 hallmarks locate on exons (Figure 3c; Table S4). Exon mutations will inactivate their genes and cause stronger phenotypes rather than disrupting H3K4me3's regulatory elements. Consistent with this supposition, the percentage of ssgRNAs on exons is significantly greater than the percentage of their matrix sgRNAs, whereas the percentage of the ssgRNAs on other regions are similar as the percentage of their matrix sgRNAs (Figure 3c). Because the sgRNAs targeting known essential

genes should be efficiently identified as ssgRNAs, the exons targeting sgRNAs can be used as spike-ins, providing further validation to the CTDE approach under an authentic mega-library screening condition. In sum, exons of 8165 genes are targeted, while 659 known essential genes are included (Figure 3f). As expected, our studies exhibit the ssgRNAs targeting 534 essential genes from the said 659, strongly accrediting the success of our screening (Figure 3f).

The enriched Gene Ontology (GO) terms of ssgRNAs targeting exons are related to essential biological processes such as ncRNA metabolic and ribonucleoprotein biogenesis (Figure S3d; Table S5). The corresponding genes of ssgRNAs targeting proximal promoters and UTRs can also be confidently identified (Details in method), and their enriched GO terms are related to the essential biological processes of cell survival as well (Figure S3e-f; Table S5).

Thus, we have successfully performed a genome-wide CRISPR-screening to interrogate the essential H3K4me3 regulatory elements for mESCs self-renewal.

Annotation of the essential CTCF hallmarks for mESCs self-renewal

Gene expressions are orchestrated by regulatory elements at local, long-range and high-dimensional levels[34-36]. CTCF stabilizes chromosomal architecture and coordinates the genome spatial positioning, which functions as a transcriptional activator or repressor[16]. In addition to local level regulatory elements (H3K4me3),

interrogating CTCF hallmarks provide an important understanding for gene regulation mechanisms in a biological context.

Mouse epigenome has around 55,000-65,000 CTCF hallmarks[37]. ~50% of them are intergenic and ~35% of them are intragenic[37]. To genome-widely interrogate essential CTCF hallmarks during mESCs self-renewal, we converted CTCF ChIPed DNA fragments to sgRNA libraries via CTDE (Figure 1a) and performed the CRISPR-screening as described above (Figure 2a). In total, we screened 848K sgRNAs which targets 64.12% of the CTCF hallmarks in mESCs (Figure 4a-c and S4a-b; Table S3 and S6). The sgRNA density is 24 sgRNAs per kb in CTCF strong binding sites (top 100). As CTCF elements display a consistent size, whilst maintaining a moderate diversity of the input DNA amount, that of which is significantly smaller than the input amount of H3K4me3, the abundance distribution of sgRNAs inside CTCF hallmarks is notably more even than that of H3K4me3, although the pattern of their template DNAs is still observed (Figure 4b and S4a). We identified 3038 CTCF ssgRNAs, which indicates that the corresponding CTCF hallmarks (47.02 % in intergenic regions) are essential for mESCs self-renewal (Figure 4a-c and S4b; Table S4). ssgRNAs appear in regions, arrayed from weak to strong CTCF elements. Among all essential CTCF elements, 87.07% elements are targeted by 1 ssgRNA whilst 12.93% elements are targeted by multiple ssgRNAs (Table S3). Detailed positions of ssgRNAs inside CTCF elements can be found in the Supplemental table 4. CTCF elements are narrow (83% < 400 bp) and have clear binding motif, and the ssgRNAs are rather important in indication of the

essentiality of their belonging CTCF elements than their targeting location inside. We verified three randomly picked intergenic CTCF ssgRNAs. As expected, all can significantly compromise mESCs proliferation (Figure 4d and S4c-d).

Because CTCF hallmarks stabilize high-dimensional architecture of chromosome at multiple levels, disrupting CTCF hallmarks will disrupt the topological structure of genomes at different levels and can cause different level cell stresses. Thus, the distribution of CTCF ssgRNA on most chromosomes is not even as the distribution of their matrix sgRNAs (Figure 4a-b).

Distal promoters, introns and proximal promoters are major parts on which CTCF hallmarks locate (Figure S4b; Table S4). The GO terms of genes close to ssgRNAs targeting introns and proximal promoters are essential biological processes and tissue developments (Figure 4e and S4e; Table S7). Because the expression of differentiation and development related genes generally antagonizes mESCs pluripotency and self-renewal, we believe that these essential CTCF hallmarks should inhibit their expression. Unlike H3K4me3 ssgRNAs, only a minor part of CTCF ssgRNAs target exons (Figure S4b; Table S4). Although most GO terms of expressed genes of these exons are also related to essential biological processes, their significance is much lower than that of H3K4me3 ssgRNA (Figure S4e and S3d; Table S5 and S7). We reason that major functions of these essential CTCF hallmarks are beyond promoting the expression of their sitting genes.

311 mESCs differentiate into various cell types during embryonic development.
312 Previous studies have shown that the chromosome spatial structure will rearrange
313 accordingly to fit the change of gene expression patterns during differentiation[38]. We
314 compared the CTCF hallmarks with 16 mouse cell types/tissues and found that 59.63%
315 CTCF hallmarks in mESCs are cell-type specific and 40.37% are common (Figure 4f;
316 Table S8)[37,39]. The common CTCF hallmarks should help maintain the universal spatial
317 structure of chromosome, while the cell-type specific CTCF hallmarks should be either
318 mESCs specific or pre-loaded hallmarks for further differentiated cells. Consistent with
319 this supposition, the percentage of the cell-type specific essential CTCF hallmarks
320 (28.85%) of mESCs is significantly smaller than the percentage of the cell-type specific
321 CTCF hallmarks (59.63%) (Figure 4f; Table S8).
322
323 **Annotation of the essential H3K4me3 hallmarks in human liver cancer cells**
324 Whole-genome sequencing has surveyed large sets of cancer genomes and studied
325 the role and extent of single-nucleotide variants (SNVs), small insertions/deletions
326 (indels) and larger structural variants in cancers[40,41]. While the initial focus on the
327 genetic variations in protein-coding regions has dramatically expanded our knowledge
328 of cancer genetics, the remaining (>90%) non-coding part of the genetic variations are
329 much more difficult to understand and have remained largely unexplored[42], which is
330 due to a lack of functional annotation of regulatory elements inside.

To genome-widely interrogate essential activating regulatory elements in human liver cancer cells (HepG2) (Figure S5a-b), we performed a H3K4me3 CRISPR-screening as described above (Figure 1a and 2a). In total, we screened 1.19M sgRNAs targeting 80.91% of the H3K4me3 hallmarks in HepG2 (Figure 5a-c and S5c-d; Table S3 and S9). In H3K4me3 highly enriched regions (top100), the sgRNA density is 43sgRNAs per kb. The abundance distribution of sgRNAs inside H3K4me3 hallmarks observes the pattern of their template DNAs (Figure 5b and S5c). We have identified 14540 ssgRNAs (75.82% are inside non-coding regions), which represent 6475 essential regulatory elements in HepG2 (Figure 5a-c and S5d; Table S4). ssgRNAs appear in regions from weak to strong H3K4me3 elements, and the majority are in strong elements regions (Figure 5b). Among all essential H3K4me3 elements 62.89% elements are targeted by 1 ssgRNA while 37.11% elements are targeted by multiple ssgRNAs (Table S3). Detailed positions of ssgRNAs inside H3K4me3 elements can be found in the Supplemental table 4, indicating the core regulatory sites of these essential H3K4me3 elements. We also verified three randomly picked ssgRNAs targeting non-coding regions, and all can significantly inhibit HepG2 growth (Figure 5d and S5e-f).

The H3K4me3 ssgRNAs evenly distribute on most chromosomes, with regions on chromosome 5, 8 and 13 being exceptions, in which there is no shortage of H3K4me3 hallmarks and sgRNAs (Figure 5a). This indicates that fewer essential genes exist in these regions.

Most H3K4me3 ssgRNAs locate on exons, proximal promoters, and introns (Figure S5d; Table S4). Normally, genes are more efficiently inactivated by mutations on their exons than their regulatory regions. Hence, the percentage of the ssgRNAs on exons is significantly greater than the percentage of the ssgRNAs on other regions (Figure S5d). The enriched GO terms of ssgRNA targeting exons are related to essential biological processes such as ncRNA metabolic and mitochondrial function, which play central roles in malignancy through macromolecular synthesis and energy production[43,44] (Figure S5g; Table S10), while the enriched GO terms of ssgRNAs targeting proximal promoters and UTRs are also related to the essential biological processes of cell survival (Figure 5e and S5g; Table S10).

Annotation of the essential CTCF hallmarks in human liver cancer cells

CTCF hallmarks stabilize mammalian genomes into discrete structural and regulatory domains, those of which can either prevent or facilitate the interactions of promoters and enhancers across their boundaries[15]. Although previous works substantiate that CTCF hallmarks can evolve in human cancers[45], their underlying mechanisms are principally more difficult to define as a result of missing functional annotation of CTCF hallmarks in human.

To interrogate essential CTCF hallmarks in HepG2, we generated CTCF sgRNA libraries and performed the CRISPR-screening as described above (Figure 1a and 2a). Overall, we screened 1.06M sgRNAs, targeting 79.75% of the CTCF hallmarks in

HepG2 (Figure 6a-c and S6a-b; Table S3 and S11). In strong CTCF binding regions (top100), the sgRNA density is 63sgRNAs per kb. As expected, the abundance distribution of sgRNAs inside CTCF hallmarks is notably more even than that of H3K4me3, and the patterns of their template DNAs are observed all the while (Figure 6b and 5b). We identified 4628 CTCF ssgRNAs which represent 3583 (44.63 % inside intergenic regions) essential CTCF hallmarks for HepG2 growth (Figure S6b; Table S4). ssgRNAs appear in regions from weak to strong CTCF elements (Figure 6b). Among all HepG2 essential CTCF elements, 78.76% elements are targeted by 1 ssgRNA while 21.24% elements are targeted by multiple ssgRNAs (Table S3). Detailed positions of ssgRNAs inside CTCF elements can be found in the Supplemental table 4. We verified three randomly picked intergenic CTCF ssgRNAs. As expected, all can significantly inhibit HepG2 growth (Figure 6d and S6c-d).

CTCF ssgRNAs on chromosomes roughly follow the distribution of their matrix hallmarks and sgRNAs (Figure 6a). As we known that CTCF hallmarks stabilize genome from smaller loops into huge megabase-sized loops called topologically associated domains (TADs)[46,47], CTCF ssgRNAs will disrupt the topological structure at different levels and result in their uneven distribution on many regions (Figure 6a).

Major CTCF hallmarks exist on distal promoters, introns, and proximal promoters in HepG2 (Figure S6b; Table S4). CTCF ssgRNA targeting introns and proximal promoters have significant GO terms including signaling transduction and cell morphogenesis, all of which are crucial for cancers (Figure 6e and S6e; Table S12).

Only a minor portion of CTCF ssgRNAs target exons (Figure S6b). The majority of GO terms of expressed genes whose exons are targeted by ssgRNAs are related to essential biological processes (Figure S6f; Table S12). Because major functions of CTCF hallmarks on exons are beyond promoting their sitting genes expression[15], the GO term significance is lesser than that of H3K4me3 ssgRNA targeting exons (Figure S6f and S5g; Table S12 and S10).

The majority of CTCF hallmarks in HepG2 are common (79.79%) among 55 human cell types (Figure 6f; Table S13). As a tissue-specific cell line, the spatial chromosome structure of HepG2 has been adapted to the requirements of liver functions, and it is not necessary to keep so many spatial chromosome structures specific for other cell types. Unlike mESCs, the percentage of the cell-type specific essential CTCF hallmarks corresponds with the percentage of the cell-type specific CTCF hallmarks in HepG2 (Figure 6f and 4f).

H3K4me3 hallmark-LINC00339-CDC42 axis maintaining HepG2 growth

We believe that functional epigenome annotation can facilitate uncovering novel regulation mechanisms and biomarkers of cancer cells. We focused on a H3K4me3 ssgRNA (chr1-22352881), which targets the intron of LINC00339 that is highly expressed in multiple cancer cell lines (Figure 7a). Because H3K4me3 activates local gene expression, we checked if ssgRNA (chr1-22352881) would disrupt LINC00339 expression. As predicted, the LINC00339 level is significantly decreased after ssgRNA

414 (chr1-22352881) disruption (Figure 7b). Ensuing, we knocked down the expression of
415 LINC00339, and discovered that the proliferation of HepG2 is significantly inhibited
416 into the same level as ssgRNA (chr1-22352881) disruption (Figure 7c-e). Cell-cycle
417 analysis shows that both ssgRNA (chr1-22352881) disruption and knocking down
418 LINC00339 blocks the S-phase entry of HepG2 but not hESCs (Figure 7f). Altogether,
419 the data suggests that ssgRNA (chr1-22352881) disruption compromises HepG2
420 proliferation through downregulating the expression of LINC00339.

421 It has recently became apparent that long non-coding RNAs (lncRNAs) can
422 function as transcriptional activators[48]. Through binding to histone-modifying
423 complexes, transcription factors and RNA polymerase II, lncRNAs can promote gene
424 expression in *cis* or in *trans*[48]. Located at the immediate downstream of LINC00339,
425 CDC42 functions in cell-cycle and anchorage-independent growth (Figure S5f)[49,50].
426 CDC42 also transduces growth and adhesion signals to drives cell-cycle progress from
427 G1 to S phase[49,50]. Therefore, LINC00339 may promote CDC42 expression in HepG2.
428 As expected, we found that both ssgRNA (chr1-22352881) disruption and knocking-
429 down LINC00339 significantly decreases the expression of CDC42 (Figure 7g-h). Our
430 data establishes that the H3K4me3 hallmark targeted by ssgRNA (chr1-22352881)
431 orchestrates the activity of LINC00339-CDC42 axis to promote HepG2 growth.
432

Discussions

A vast proportion of mammalian genomes are non-coding and contain vital regulatory roles. Unlike coding-regions, functional analyses of non-coding regions through CRISPR-screening requires high density sgRNA coverage. The commercial synthesis of a sgRNA library covering an epigenetic hallmark (H3K4me3 for example: 200-300K USD) is unaffordable for most labs. As a result, molecular biology techniques that directly convert the source DNA into sgRNA library are urgently needed to unleash the full capabilities of CRISPR-screening for decoding non-coding-genomes. To this end, several approaches have been developed. One of which, named CORALINA (comprehensive gRNA library generation through controlled nuclease activity), employs MNase to break down the source DNA[51]. CORALINA incorporates fragmented DNA into library vector without a step to enrich the fragment flanking a Protospacer Adjacent Motif (PAM)[51]. Without PAM, around 70-77% sgRNAs generated by CORALINA are ineffective. Hiroshi Arakawa developed another approach that employs six type IIs or type III restriction enzymes (EcoP15I, Bgl II, Acu I, XbaI, Bsm BI and AatI) digestions and PAGE-Gel purifications to convert mRNA into sgRNAs with flanking PAM[52]. Because six restriction enzymes digestions will destroy 6.2% potential sgRNAs and multiple PAGE-Gel purification significantly increases DNA loss, Arakawa's approach is not practicable for large-scale CRISPR-screening. Our CTDE approach overcomes pervious limitations by generating sgRNA libraries from source DNAs with a simple- and cost-effective procedure.

Because the CTDE sgRNA library is directly from a DNA source, the abundance distribution of sgRNAs follows the abundance distribution of their input DNA. Benefited by this characteristic, our screen can focus more on the highly enriched H3K4me3 and CTCF regions which normally would have more biological significance. However, the biological significance of RNAs do not always correlate with their abundance, which normally varies up to hundreds of folds inside cells. Thus, CTDE sgRNA library is not practicable for whole transcriptome screening.

The approach utilizing paired-guide RNA CRISPR/Cas9 library(CREST-seq) has been shown efficient discovery and functional characterization of regulatory elements[21]. This type of approach needs more than three paired-guide RNAs for each hallmark, and the library cost to cover all CTCF and H3K4me3 hallmarks (79K in mESCs and 69K in HepG2) proves to be too costly. In addition, once asynchronous cutting happens, the sequence of the first cutting site will change[53] and the hallmark cannot be deleted by the paired-guide RNAs anymore. So, the paired-guide RNA approach needs CRISPR/Cas9 cut both sites simultaneously, which leads to a lower efficiency than single sgRNA system. Thus, the paired-guide RNA library-based screening requires a more sensitive readout than the survival readout used in this work.

The essential hallmarks datasets of H3K4me3 and CTCF in mESCs and HepG2 can provide potential benefits for multiple fields in the future. For instance, Induced Pluripotent Stem Cells (iPSCs) researchers should evaluate the somatic mutations inside the essential hallmarks of mESCs, which can significantly compromise the

475 reprogramming efficiency. Additionally, the H3K4me3 essential hallmarks related to
476 core transcriptional factors can be utilized to boost the reprogramming efficiency
477 through CRISPR mediated activation[54]. The HepG2 datasets permits cancer researchers
478 to direct their focus on H3K4me3 or/and CTCF essential hallmarks, which are specific
479 in liver cancer cells. Through reversible or permanent inactivation of these hallmarks
480 via CRISPR mediated approaches, potential novel liver cancer treatment strategies can
481 be developed[55].

482 CTDE provides a simple, time- and cost-effective procedure to convert source
483 DNA to sgRNA library, which could be more broadly applied to functional epigenome
484 annotation in a biological context.

485
486

References

1. TA., B. Genomes. 2nd edition.
2. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat Rev Genet* **17**, 772 (2016).
3. Bernstein, B.E., Meissner, A. & Lander, E.S. The mammalian epigenome. *Cell* **128**, 669-681 (2007).
4. Abbott, A. Project set to map marks on genome. *Nature* **463**, 596-597 (2010).
5. Consortium, E.P. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
6. Boyle, A.P. et al. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322 (2008).
7. Kurdyukov, S. & Bullock, M. DNA Methylation Analysis: Choosing the Right Method. *Biology (Basel)* **5** (2016).
8. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).
9. Sims, R.J., 3rd, Nishioka, K. & Reinberg, D. Histone lysine methylation: a signature for chromatin function. *Trends Genet* **19**, 629-639 (2003).
10. Wysocka, J. et al. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature* **442**, 86-90 (2006).
11. Bernstein, B.E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326 (2006).
12. Phillips, J.E. & Corces, V.G. CTCF: master weaver of the genome. *Cell* **137**, 1194-1211 (2009).
13. Chaumeil, J. & Skok, J.A. The role of CTCF in regulating V(D)J recombination. *Curr Opin Immunol* **24**, 153-159 (2012).
14. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951 (2008).
15. Kim, S., Yu, N.K. & Kaang, B.K. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp Mol Med* **47**, e166 (2015).
16. Ong, C.T. & Corces, V.G. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* **15**, 234-246 (2014).
17. Khoury, A. et al. Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. *Nat Commun* **11**, 54 (2020).
18. Joung, J. et al. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nat Protoc* **12**, 828-863 (2017).
19. Shalem, O., Sanjana, N.E. & Zhang, F. High-throughput functional genomics using CRISPR-Cas9. *Nat Rev Genet* **16**, 299-311 (2015).
20. Fulco, C.P. et al. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769-773 (2016).
21. Diao, Y. et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods* **14**, 629-635 (2017).
22. Shah, S.A., Erdmann, S., Mojica, F.J. & Garrett, R.A. Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol* **10**, 891-899 (2013).

23. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).
24. Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C. & Doudna, J.A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62-67 (2014).
25. Hsu, P.D., Lander, E.S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262-1278 (2014).
26. Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M. & Joung, J.K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* **32**, 279-284 (2014).
27. Ran, F.A. et al. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380-1389 (2013).
28. Metzker, M.L. et al. Termination of DNA synthesis by novel 3'-modified-deoxyribonucleoside 5'-triphosphates. *Nucleic Acids Res* **22**, 4259-4267 (1994).
29. Gibson, D.G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343-345 (2009).
30. Bodapati, S., Daley, T.P., Lin, X., Zou, J. & Qi, L.S. A benchmark of algorithms for the analysis of pooled CRISPR screens. *Genome Biol* **21**, 62 (2020).
31. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol* **15**, 554 (2014).
32. Bracewell, R. The Fourier Transform & Its Applications 3rd Edition. *Book*.
33. Diao, Y. et al. A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. *Genome Res* **26**, 397-405 (2016).
34. Weake, V.M. & Workman, J.L. Inducible gene expression: diverse regulatory mechanisms. *Nat Rev Genet* **11**, 426-437 (2010).
35. Malik, S. & Roeder, R.G. The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. *Nat Rev Genet* **11**, 761-772 (2010).
36. Ong, C.T. & Corces, V.G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12**, 283-293 (2011).
37. Kim, T.H. et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231-1245 (2007).
38. Dixon, J.R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).
39. Cuddapah, S. et al. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19**, 24-32 (2009).
40. Blum, A., Wang, P. & Zenklusen, J.C. SnapShot: TCGA-Analyzed Tumors. *Cell* **173**, 530 (2018).
41. Consortium, I.T.P.-C.A.o.W.G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).
42. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102-111 (2020).
43. Zong, W.X., Rabinowitz, J.D. & White, E. Mitochondria and Cancer. *Mol Cell* **61**, 667-676 (2016).
44. Anastasiadou, E., Jacob, L.S. & Slack, F.J. Non-coding RNA networks in cancer. *Nat Rev Cancer* **18**, 5-18 (2018).
45. Song, S.H. & Kim, T.Y. CTCF, Cohesin, and Chromatin in Human Cancer. *Genomics Inform* **15**, 114-122 (2017).

46. Yu, M. & Ren, B. The Three-Dimensional Organization of Mammalian Genomes. *Annu Rev Cell Dev Biol* **33**, 265-289 (2017).
47. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol* **16**, 245-257 (2015).
48. Long, Y., Wang, X., Youmans, D.T. & Cech, T.R. How do lncRNAs regulate transcription? *Sci Adv* **3**, eaao2110 (2017).
49. Chou, M.M., Masuda-Robens, J.M. & Gupta, M.L. Cdc42 promotes G1 progression through p70 S6 kinase-mediated induction of cyclin E expression. *J Biol Chem* **278**, 35241-35247 (2003).
50. Qadir, M.I., Parveen, A. & Ali, M. Cdc42: Role in Cancer Management. *Chem Biol Drug Des* **86**, 432-439 (2015).
51. Koferle, A. et al. CORALINA: a universal method for the generation of gRNA libraries for CRISPR-based screening. *BMC Genomics* **17**, 917 (2016).
52. Arakawa, H. A method to convert mRNA into a gRNA library for CRISPR/Cas9 editing of any organism. *Sci Adv* **2**, e1600699 (2016).
53. Sander, J.D. & Joung, J.K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol* **32**, 347-355 (2014).
54. Konermann, S. et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583-588 (2015).
55. Qi, L.S. et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173-1183 (2013).

Figure 1

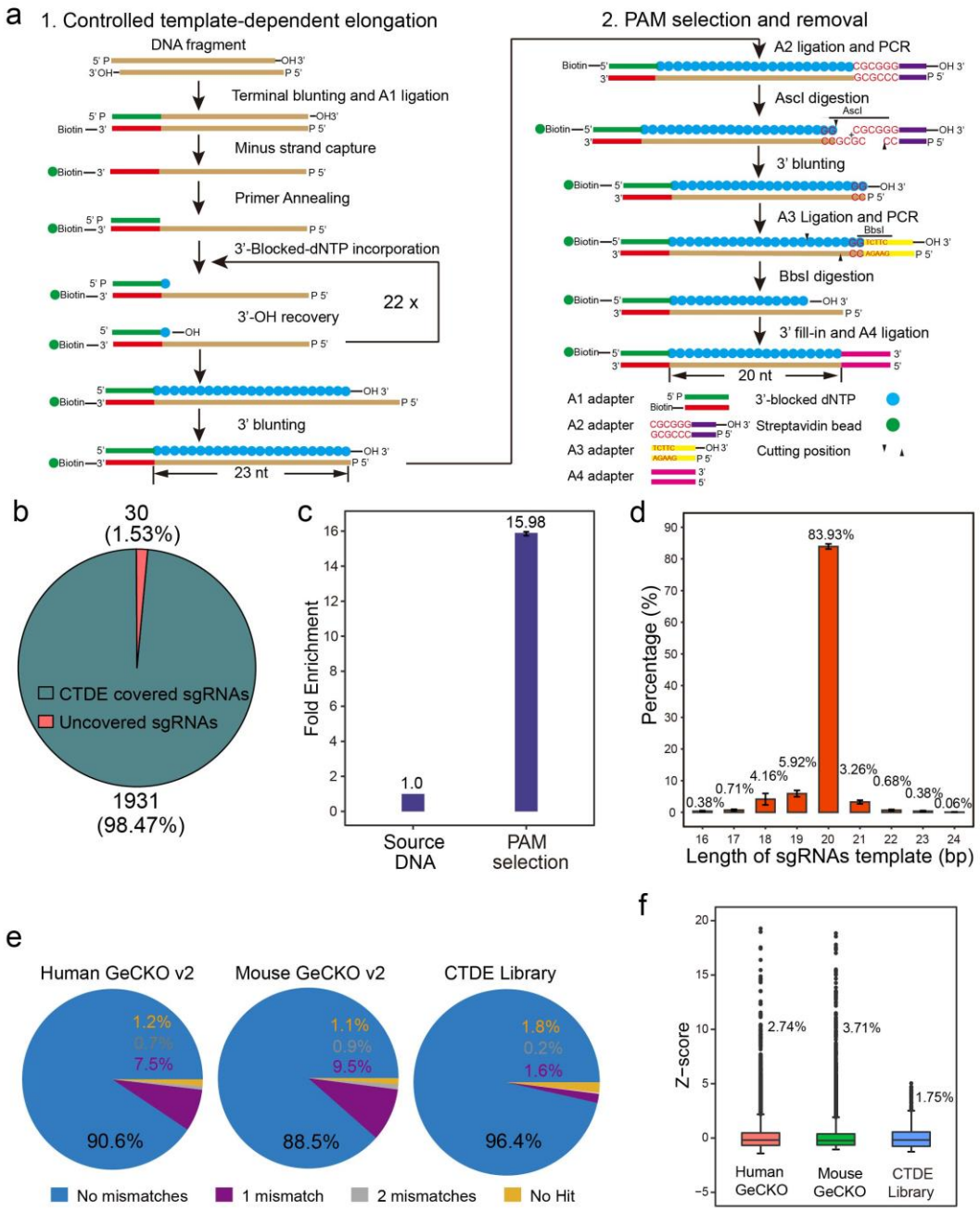

Figure 1 Controlled template-dependent elongation (CTDE) converts source DNA to a sgRNA library

(a) Detailed schematics of the CTDE procedure. Briefly, sgRNA templates (23bp) are generated from fragmented input DNA by controlled template-dependent elongation; sgRNA templates with NGG triplet in 3′ terminus are enriched and NGG triplets are removed before sgRNA templates are cloned into a vector.

(b) The coverage rate of the potential sgRNA templates on source DNA (lentiCRISPR-V2) by CTDE.

(c) Fold enrichment of the sgRNA templates after NGG-PAM selection step. Error bars SD of triplicates.

(d) Length distribution of the sgRNA templates from source DNA (lentiCRISPR-V2). Error bars SD of triplicates.
(e) The fidelity of sgRNAs generated by CBDS (GeCKO libraries) and CTDE approaches; the mapping quality of the sgRNAs against their targeting sites is used to show the library fidelity.
(f) The abundance bias of sgRNAs of CBDS (GeCKO libraries) and CTDE approaches; distribution of sgRNA counts z-score of two GeCKO libraries (red and green) and CTDE library (blue) are displayed.

Figure 2

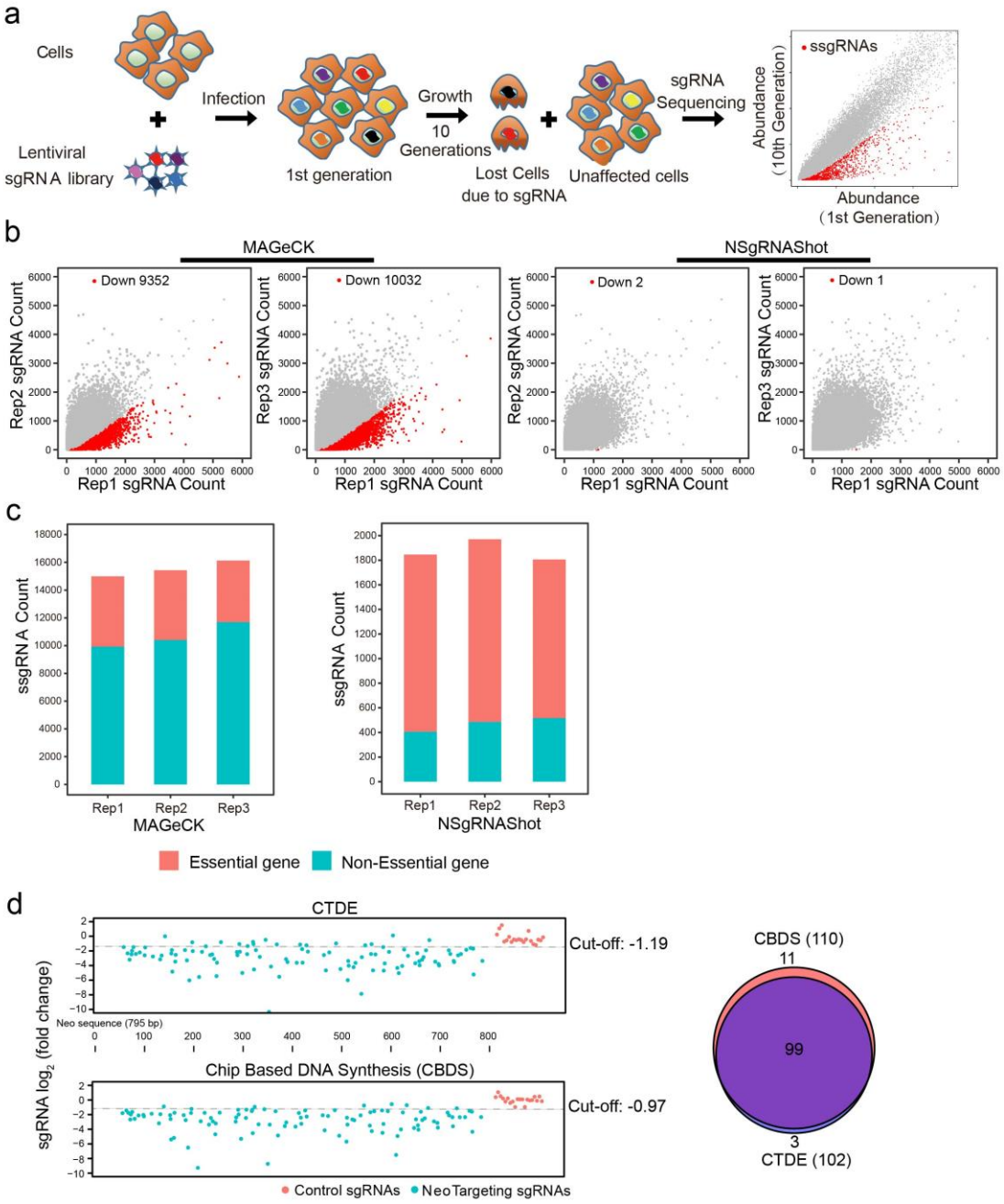

Figure 2 Validations of CTDE sgRNA library screening

(a) The schematic of CRISPR/Cas9 dropout screen: Cells are infected by library virus at low MOI; infected cells are cultured 10 generations; by comparing the abundance of sgRNA templates between the 1st and 10th generation, the sgRNAs which limit cell proliferation are identified by NSgRNAShot as significant sgRNAs(ssgRNAs).

(b) Comparison of false-positive rate of significantly changed sgRNA (ssgRNA) calling between MAGeCK and NSgRNAShot. Datasets of three biological replicates from an CRISPR dropout screen project (Tzelepis et al., 2016) are used (details in method). Scatter-plots exhibit sgRNA abundance at same time point between two biological replicate datasets. MAGeCK identifies

noticeably more false-positive ssgRNAs (red dots); NSgRNAShot identifies few false-positive ssgRNAs.

(c) Bar plots display the number and distribution (essential genes and non-essential genes) of ssgRNAs identified by MAGeCK and NSgRNAShot from above (b) CRISPR dropout screen project.

(d) Left panel: $\log_2$ fold change of the sgRNA abundance of Neo gene targeting libraries generated by CBDS and CTDE during validation screening (details in method). Red dots are non-targeting control sgRNAs; green dots are Neo targeting sgRNAs. Right panel: the overlap of significantly changed sgRNA (ssgRNA) from CBDS library and CTDE library.

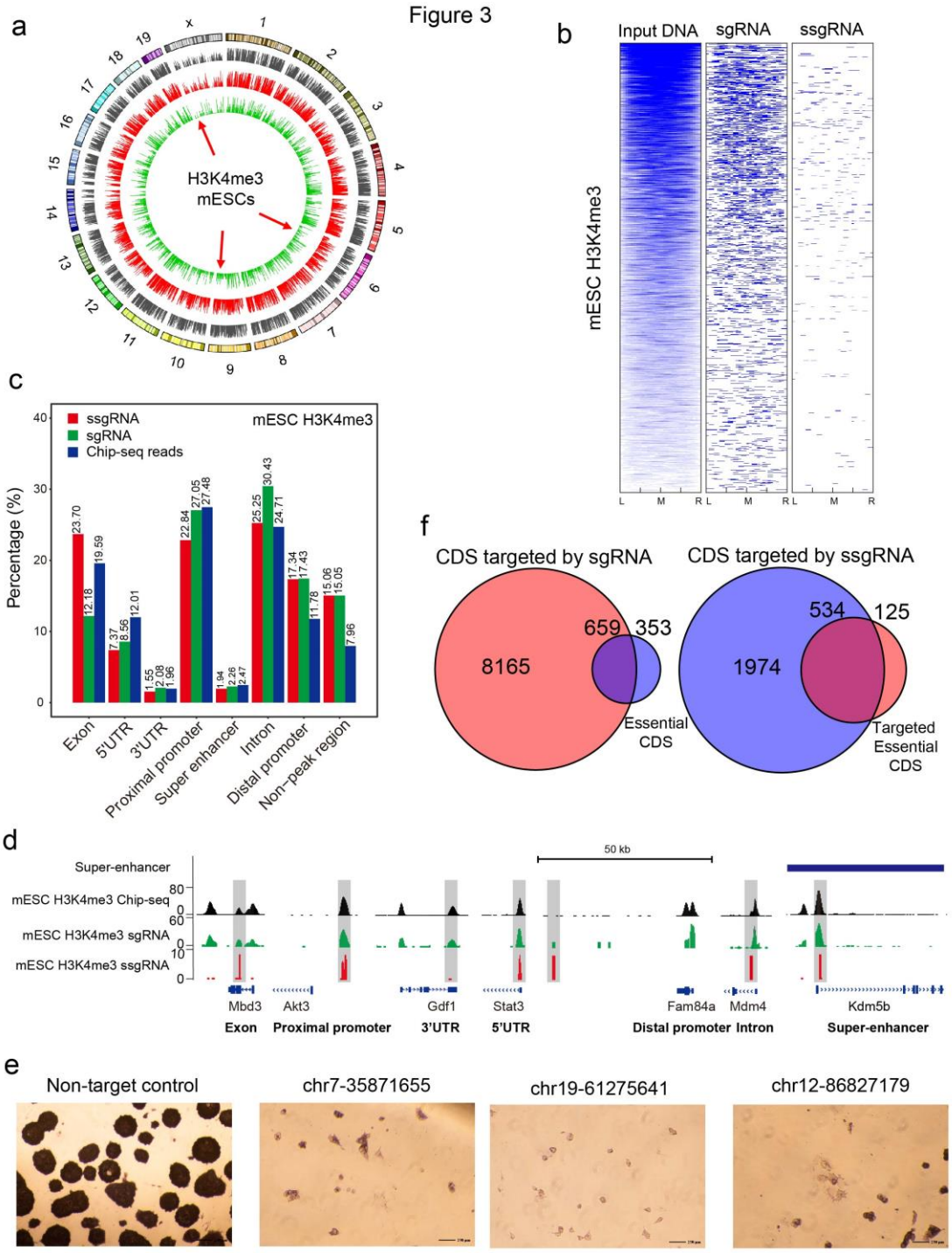

Figure 3 Annotation of the essential H3K4me3 hallmarks in mESCs self-renewal

(a) The circos plot of H3K4me3 ChIP-seq, H3K4me3 sgRNAs and H3K4me3 ssgRNAs profiles on mouse genome. The circles, from exterior to interior, illustrate mouse genome (mm9), ChIP-Seq reads density (black), H3K4me3 sgRNAs density (red) and H3K4me3 ssgRNAs density (green). Arrows show regions with biased ssgRNAs distribution.

(b) Heat-map of input DNA, sgRNA and ssgRNA enrichment inside mESCs H3K4me3 elements. All H3K4me3 elements are displayed in relative scale. L: Left boundary; M: Middle; R: Right boundary. The heat-maps are rank-ordered according to the enrichment of input DNA (blue,

enriched; white, not enrichment).
(c) The distribution of H3K4me3 ChIP-seq reads, H3K4me3 sgRNAs and H3K4me3 ssgRNAs on major regulatory regions of mouse genome.
(d) Plots of H3K4me3 ChIP-seq reads, H3K4me3 sgRNAs and H3K4me3 ssgRNAs at seven genomic loci. Y-axes, RPKM. Genomic regions with enriched H3K4me3 ssgRNA are shaded grey.
(e) Validation of three H3K4me3 ssgRNA in mESCs self-renewal. Alkaline phosphatase (AP) staining illustrates that three randomly selected H3K4me3 ssgRNAs significantly inhibit mESCs self-renewal. ssgRNA is named by three factors (chromosome number; targeting strain + or -; mapping position). Scale bar, 250 μm.
(f) Left panel: CDS targeted by mESCs H3K4me3 sgRNAs includes 659 well-recognized essential genes (Shohat and Shifman, 2019; Tzelepis et al., 2016). Right panel: CDS targeted by mESCs H3K4me3 ssgRNAs includes 534 genes from above 659 essential genes.

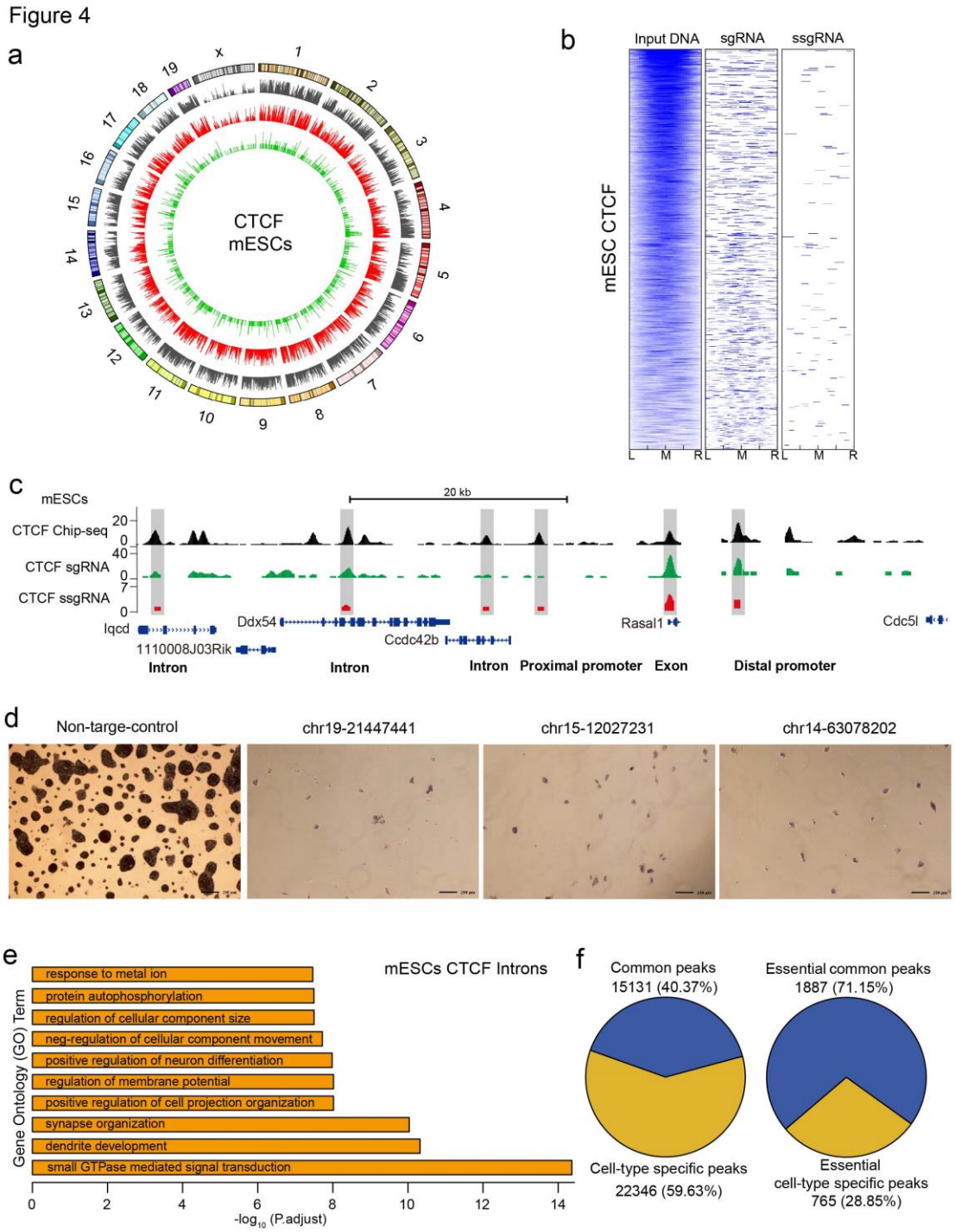

Figure 4 Annotation of the essential CTCF hallmarks in mESCs self-renewal

(a) The circos plot of CTCF ChIP-seq, CTCF sgRNAs and CTCF ssgRNAs profiles on mouse genome. The circles, from exterior to interior, illustrate mouse genome (mm9), CTCF ChIP-Seq reads density (black), CTCF sgRNAs density(red) and CTCF ssgRNAs density (green).

(b) Heat-map of input DNA, sgRNA and ssgRNA enrichment inside mESCs CTCF elements. All CTCF elements are displayed in relative scale. L: Left boundary; M: Middle; R: Right boundary. The heat-maps are rank-ordered based on the enrichment of input DNA as follows: blue, enriched; white, not enrichment.

669 (c) Plots of CTCF ChIP-seq reads, CTCF sgRNAs and CTCF ssgRNAs at six genomic loci. Y-axes,
670 RPKM. Genomic regions with CTCF ssgRNA are shaded grey.
671 (d) Validation of three CTCF ssgRNA in mESCs self-renewal. Alkaline phosphatase (AP) staining
672 shows that three randomly selected CTCF ssgRNAs significantly inhibit mESCs self-renewal.
673 ssgRNA is named by three factors (chromosome number; targeting strain + or -; mapping position).
674 Scale bar, 250 μm.
675 (e) Top 10 GO terms of the genes whose introns are targeted by CTCF ssgRNAs.
676 (f) Cell-type specific analysis of essential CTCF hallmark (targeted by ssgRNA) in mESCs self-
677 renewal. Major part (71.15%) of essential CTCF hallmark in mESCs self-renewal are common VS
678 major part (59.63%) of CTCF hallmarks in mESCs are cell-type specific.
679

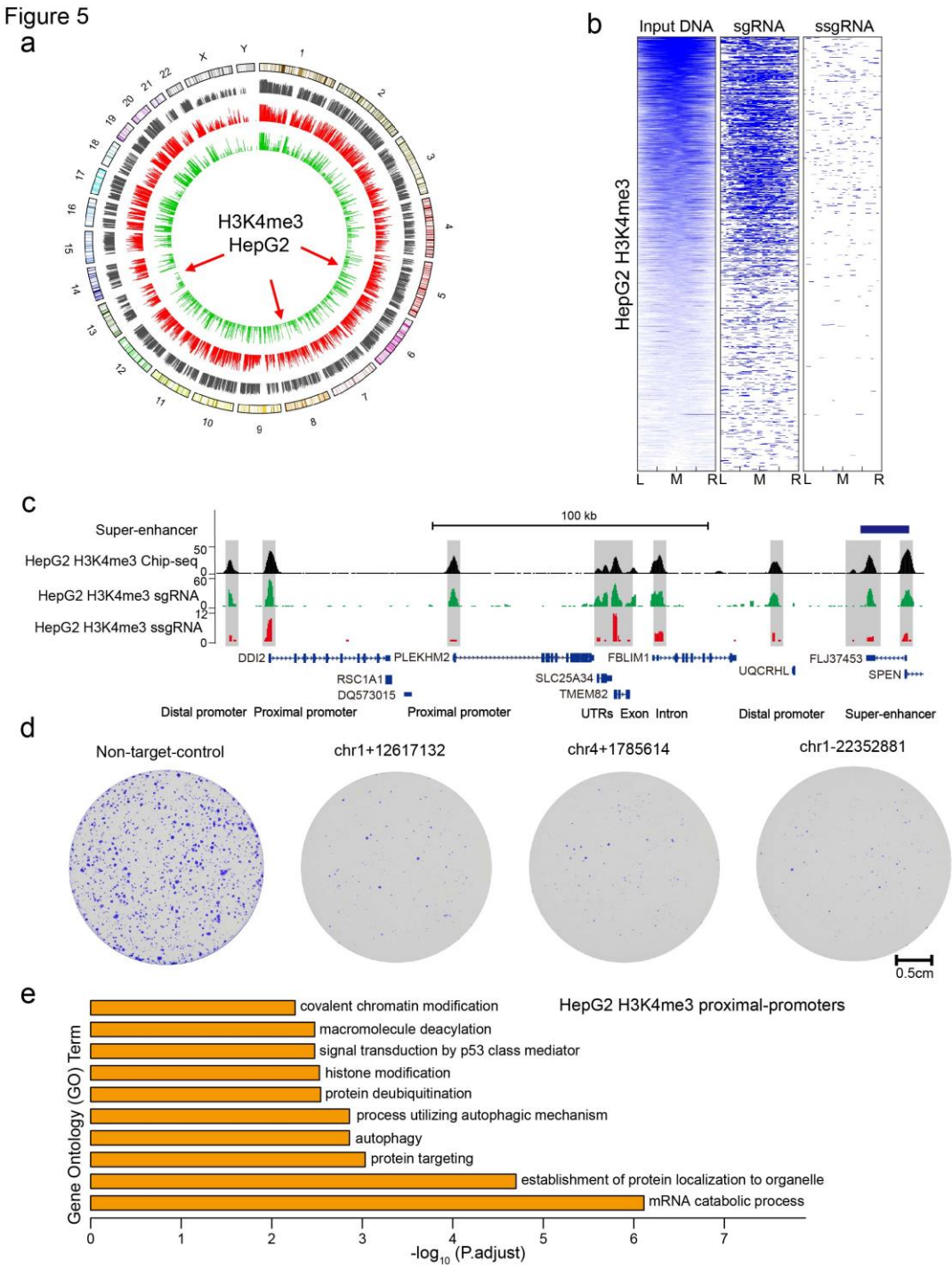

Figure 5 Annotation of the essential H3K4me3 hallmarks in HepG2

(a) The circos plot of H3K4me3 ChIP-seq, H3K4me3 sgRNAs and H3K4me3 ssgRNAs profiles on human genome. The circles, from exterior to interior, illustrate human genome (hg19), ChIP-Seq reads density (black), H3K4me3 sgRNAs density(red) and H3K4me3 ssgRNAs density (green). Arrows show regions with biased ssgRNAs distribution.

(b) Heat-map of input DNA, sgRNA and ssgRNA enrichment inside HepG2 H3K4me3 elements. All H3K4me3 elements are displayed in relative scale. L: Left boundary; M: Middle; R: Right boundary. The heat-maps are rank-ordered based on the enrichment of input DNA as follows: blue,

enriched; white, not enrichment.
(c) Plots of H3K4me3 ChIP-seq reads, H3K4me3 sgRNAs and H3K4me3 ssgRNAs at eight genomic loci. Y-axes, RPKM. Genomic regions with enriched H3K4me3 ssgRNA are shaded grey.
(d) Validation of three H3K4me3 ssgRNAs in HepG2. Crystal violet staining shows that three randomly selected H3K4me3 ssgRNAs significantly inhibit HepG2 proliferation. ssgRNA is named by three factors (chromosome number; targeting strain + or -; mapping position). Scale bar, 500 μm
(e) Top 10 GO terms of the genes whose proximal promotors are targeted by H3K4me3 ssgRNAs.

Figure 6

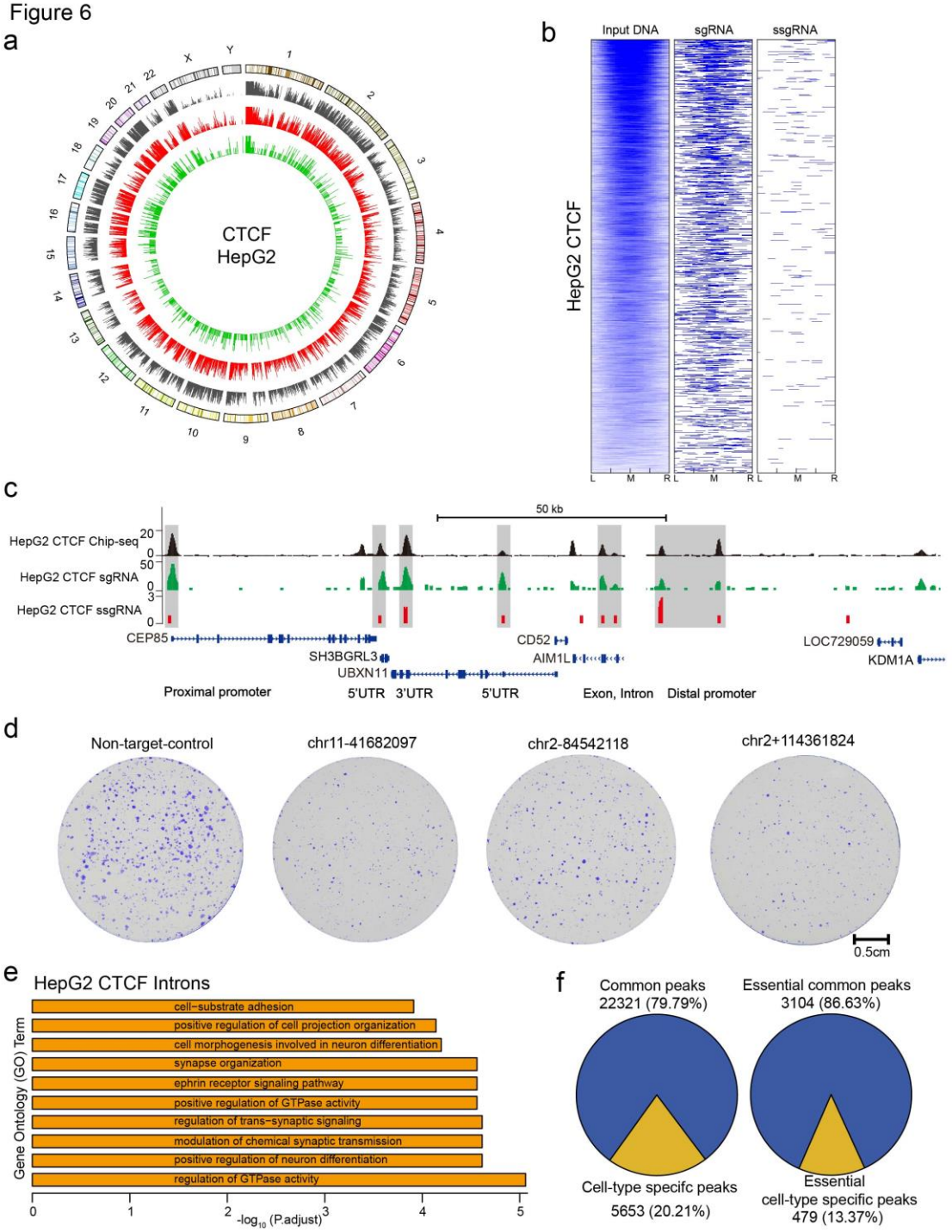

Figure 6 Annotation of the essential CTCF hallmarks in HepG2

(a) The circos plot of CTCF ChIP-seq, CTCF sgRNAs and CTCF ssgRNAs profiles on human genome. The circles, from outside to inside, illustrate human genome (hg19), CTCF ChIP-Seq reads density (black), CTCF sgRNAs density(red) and CTCF ssgRNAs density (green).

(b) Heat-map of input DNA, sgRNA and ssgRNA enrichment inside HepG2 CTCF elements. All CTCF elements are displayed in relative scale. L: Left boundary; M: Middle; R: Right boundary. The heat-maps are rank-ordered based on the enrichment of input DNA as follows: blue, enriched; white, not enrichment.

(c) Plots of CTCF ChIP-seq reads, CTCF sgRNAs and CTCF ssgRNAs at seven genomic loci. Y-axes, RPKM. Genomic regions with CTCF ssgRNA are shaded grey.
(d) Validation of three CTCF ssgRNAs in HepG2 proliferation. Crystal violet staining illustrates that three randomly selected CTCF ssgRNAs significantly inhibit HepG2 proliferation. ssgRNA is named by three factors (chromesome number; targeting strain + or -; mapping position). Scale bar, 500 μm
(e) Top 10 GO terms of the genes whose intron are targeted by CTCF ssgRNAs.
(f) Cell-type specific analysis of essential CTCF hallmark (targeted by ssgRNA) in HepG2 proliferation. Major part of both CTCF hallmarks (79.79%) and essential CTCF hallmarks (86.63%) are common.

Figure 7

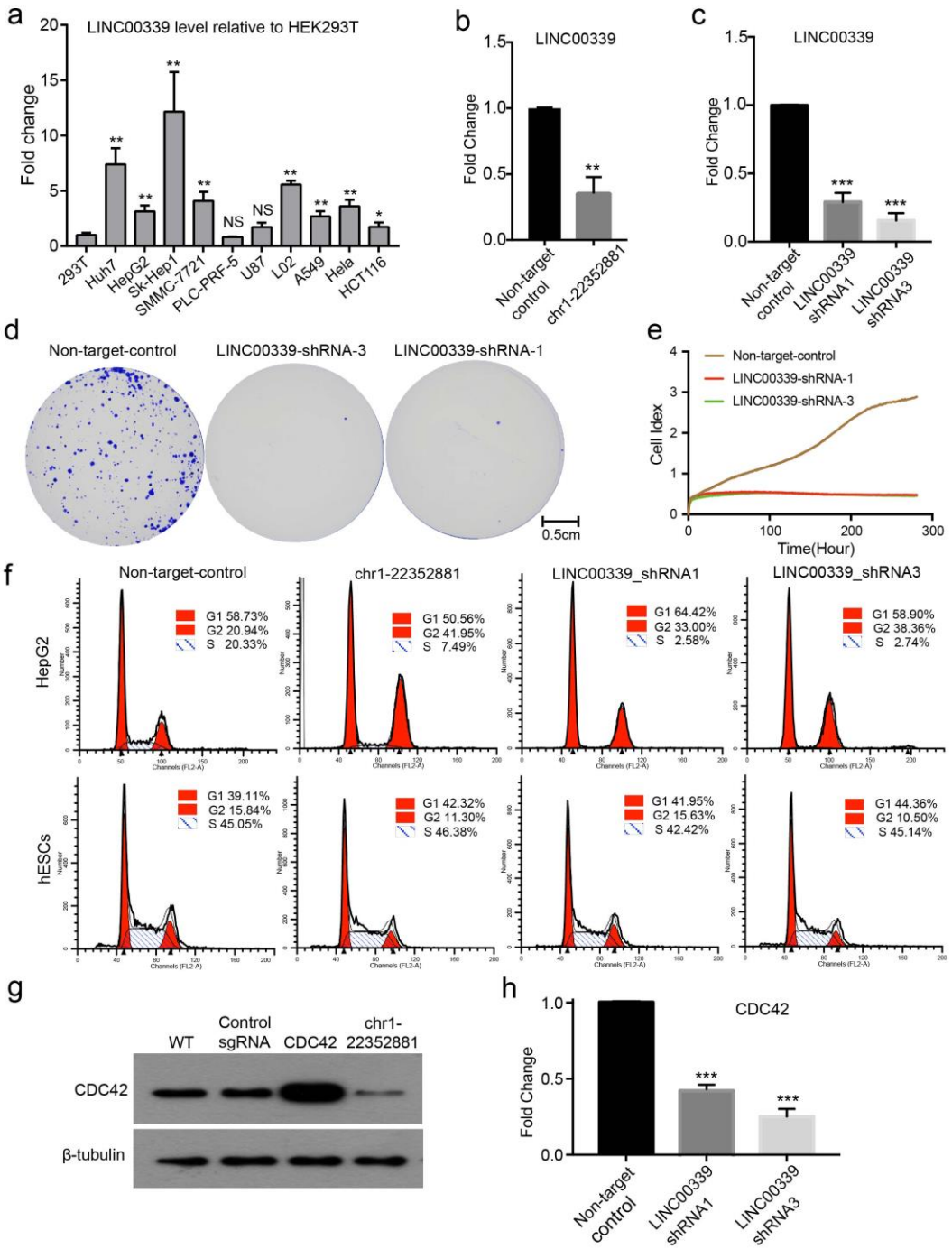

Figure 7 H3K4me3 hallmark-LINC00339-cdc42 axis maintaining HepG2 growth

(a) LINC00339 expression in ten human cancer cell lines by RT-PCR. Most human cancer cell lines have more LINC00339 than HEK293T. Student's T-test, ***p<0.001, **p < 0.01. Error bars SD of triplicates.

(b) Quantification of LINC00339 RNA level by RT-PCR after ssgRNA (chr1-22352881) disruption. ssgRNA (chr1-22352881) disruption significantly decreases LINC00339 level in HepG2. Student's T-test, ***p<0.01. Error bars SD of triplicates.

(c) Quantification of LINC00339 RNA level by RT-PCR after shRNA knockdown. Two shRNAs

734 can significantly decrease LINC00339 level in HepG2. Student's T-test, ***p<0.001. Error bars SD
735 of triplicates.
736 (d) Proliferation assay of HepG2 after knockdown LINC00339. Knockdown LINC00339
737 significantly inhibits HepG2 proliferation. 14 days culture after LINC00339 knockdown, cells were
738 stained by crystal violet.
739 (e) Knockdown LINC00339 significantly inhibits HepG2 proliferation. Real-time cell numbers are
740 plotted.
741 (f) Flow cytometry analysis of HepG2 and hESCs after ssgRNA (chr1-22352881) disruptio and
742 LINC00339 knockdown. ssgRNA (chr1-22352881) disruption and LINC00339 knockdown only
743 block S-phase entry in HepG2.
744 (g) Western blot analysis of CDC42 after ssgRNA (chr1-22352881) disruption in HepG2. CDC42
745 over-expression and a scramble sgRNA as controls. Loading control, tubulin.
746 (h) Quantification of CDC42 mRNA by RT-PCR after LINC00339 knockdown. Student's T-test,
747 ***p<0.001. Error bars SD of triplicates.
748
749
750

Figure S1

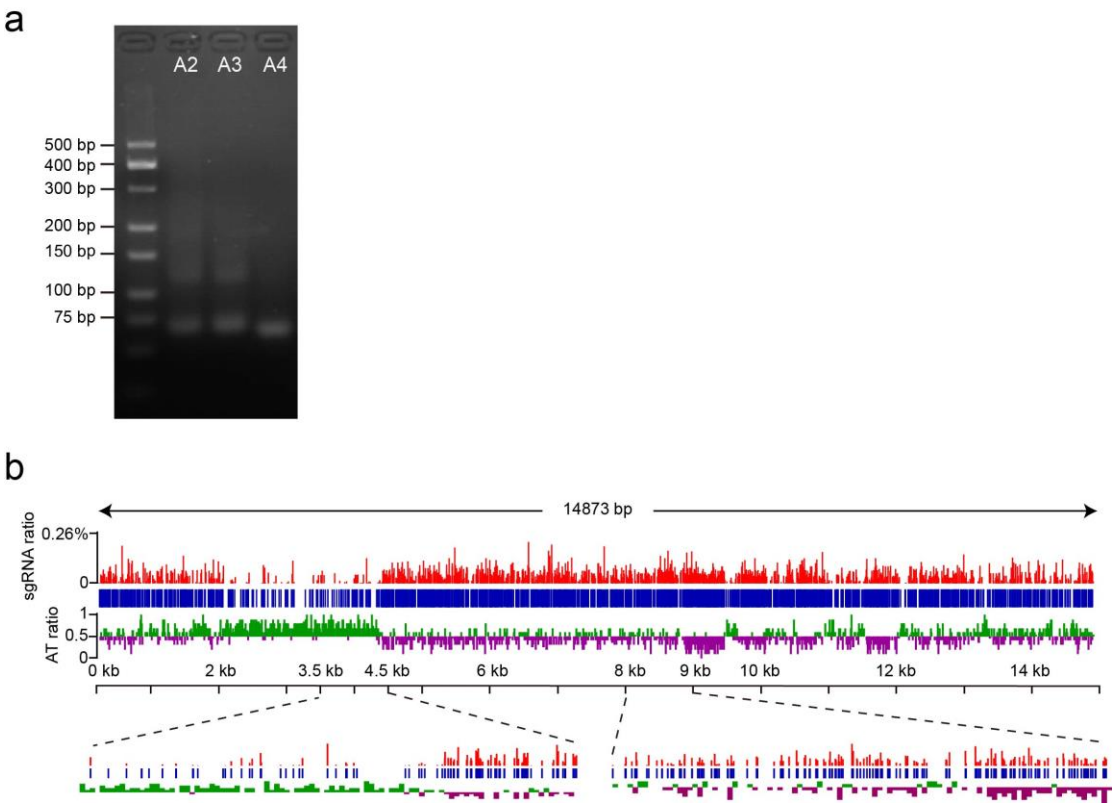

Figure S1

(a) Agarose gel (3.5%) image of PCR product after adapter (A2, A3 and A4) ligation.

(b) The abundance distribution of sgRNA templates generated by CTDE on source DNA (LentiCRISPR-V2). Up panel: Red displays the relative abundance of sgRNA templates on each PAM (NGG) site. Middle: blue represents is potential sgRNA sites (NGG PAM) on source DNA. Down panel: The green and purple are the relative abundance of A/T in 10bp bins (details in method). A A/T-rich region and a balanced nucleotide region are zoomed in.

Figure S2

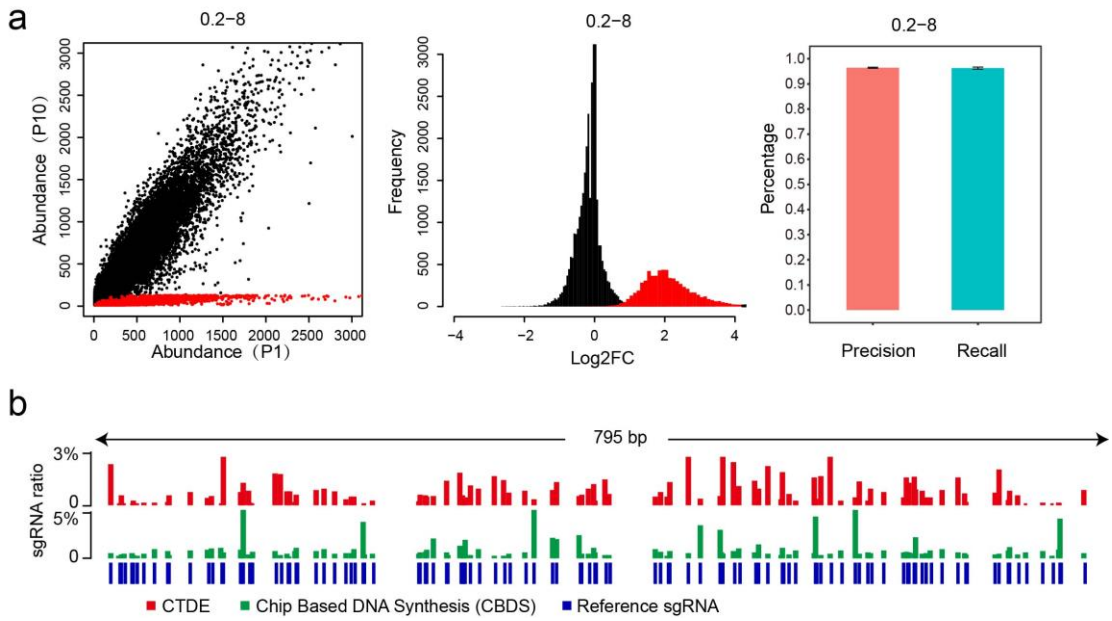

Figure S2

(a) NSgRNAShot can efficiently identify ssgRNA from a typical (α=0.2; β=8; details in method) simulated dataset. Left picture: Scatter-plot of normalized read counts of sgRNAs at P1 and P10 of the simulated dataset. Middle picture: the abundance fold-change (FC) distribution of the sgRNAs. Right picture: the precision rate and the recall rate of NsgRNAShot in simulated dataset. The black represents the sgRNAs without suffering dropout. The red represents the sgRNAs suffering dropout. (b) The abundance distribution of sgRNA templates generated by CTDE and CBDS on Neo gene. Red displays the relative abundance of sgRNA templates (CTDE) on each PAM (NGG) site; green is the relative abundance of sgRNA templates (CBDS) on each PAM (NGG) site; blue shows potential sgRNA sites (NGG PAM) on Neo gene.

Figure S3

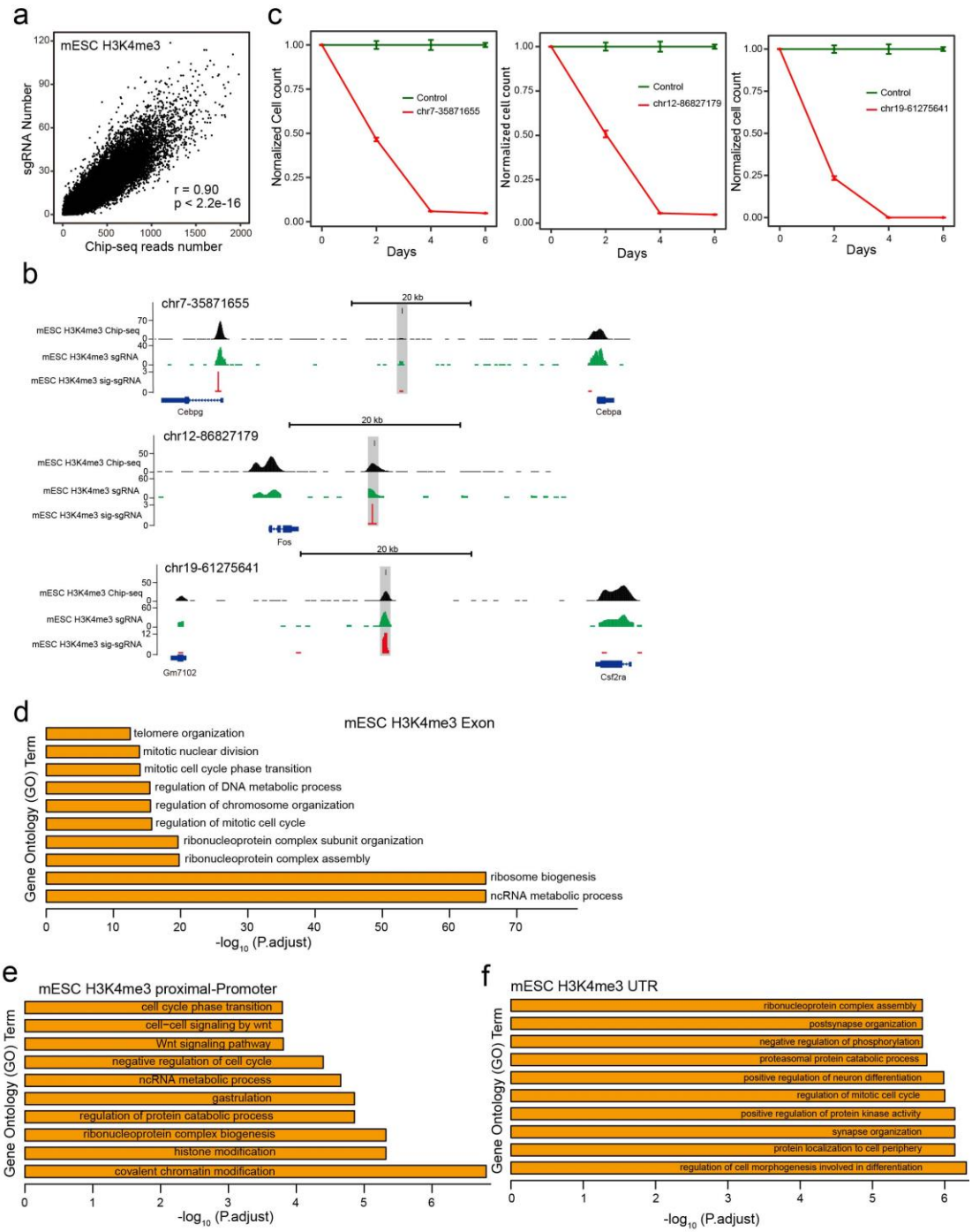

Figure S3

(a) Scatter-plot of the correlation between ChIP-seq reads number and sgRNA number in mESCs H3K4me3 elements. Spearson's correlation coefficient with two-tailed test was calculated.

(b) Plots of H3K4me3 ChIP-seq reads, H3K4me3 sgRNAs and H3K4me3 ssgRNAs at three genomic loci. Y-axes, RPKM. The genomic regions with H3K4me3 ssgRNAs validated in Figure

3e are shaded grey.
(c) Three H3K4me3 ssgRNAs validated in Figure 3e significantly inhibit mESCs proliferation. Cell numbers of three time points are plotted. Error bars the SD of triplicates
(d) Top 10 GO terms of the genes whose exons are targeted by H3K4me3 ssgRNAs.
(e) Top 10 GO terms of the genes whose proximal promoters are targeted by H3K4me3 ssgRNAs.
(f) Top 10 GO terms of the genes whose UTRs are targeted by H3K4me3 ssgRNAs.

Figure S4

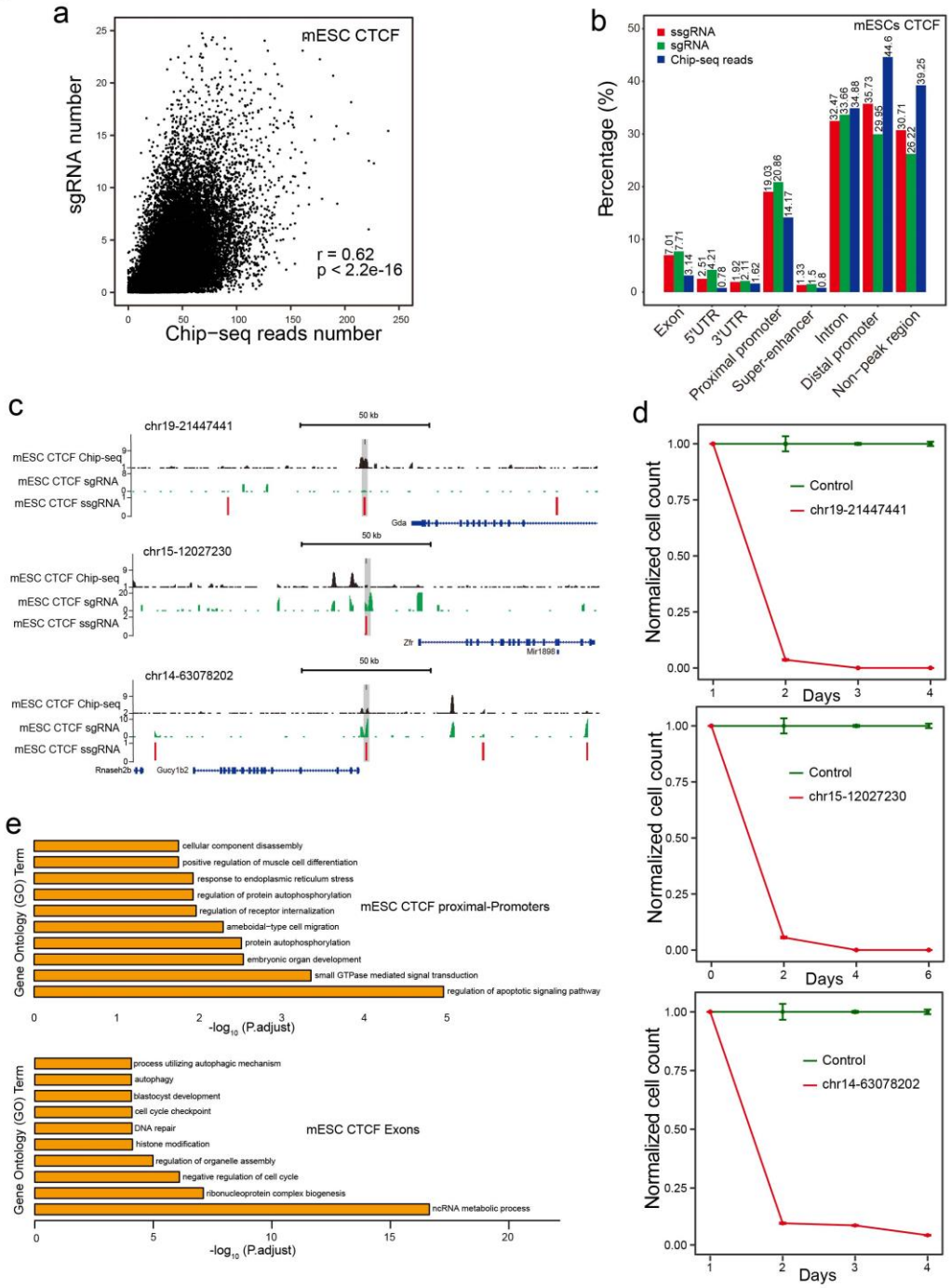

Figure S4

(a) Scatter-plot of the correlation between ChIP-seq reads number and sgRNA number in mESCs CTCF elements. Spearson's correlation coefficient with two-tailed test was calculated.

(b) The distribution of CTCF ChIP-seq reads, CTCF sgRNAs and CTCF ssgRNAs on major regulatory regions of mouse genome.

(c) Plots of CTCF ChIP-seq reads, CTCF sgRNAs and CTCF ssgRNAs at three genomic loci. Y-axes, RPKM. The genomic regions with CTCF ssgRNAs (validated in Figure 4d) are shaded grey.

(d) Three CTCF ssgRNA (validated in Figure 4d) significantly inhibits mESCs proliferation. Cell

numbers of three time-points are plotted. Error bars the SD of triplicates.
(e) Top 10 GO terms of the genes whose proximal promoters and exons are targeted by CTCF ssgRNAs.

Figure S5

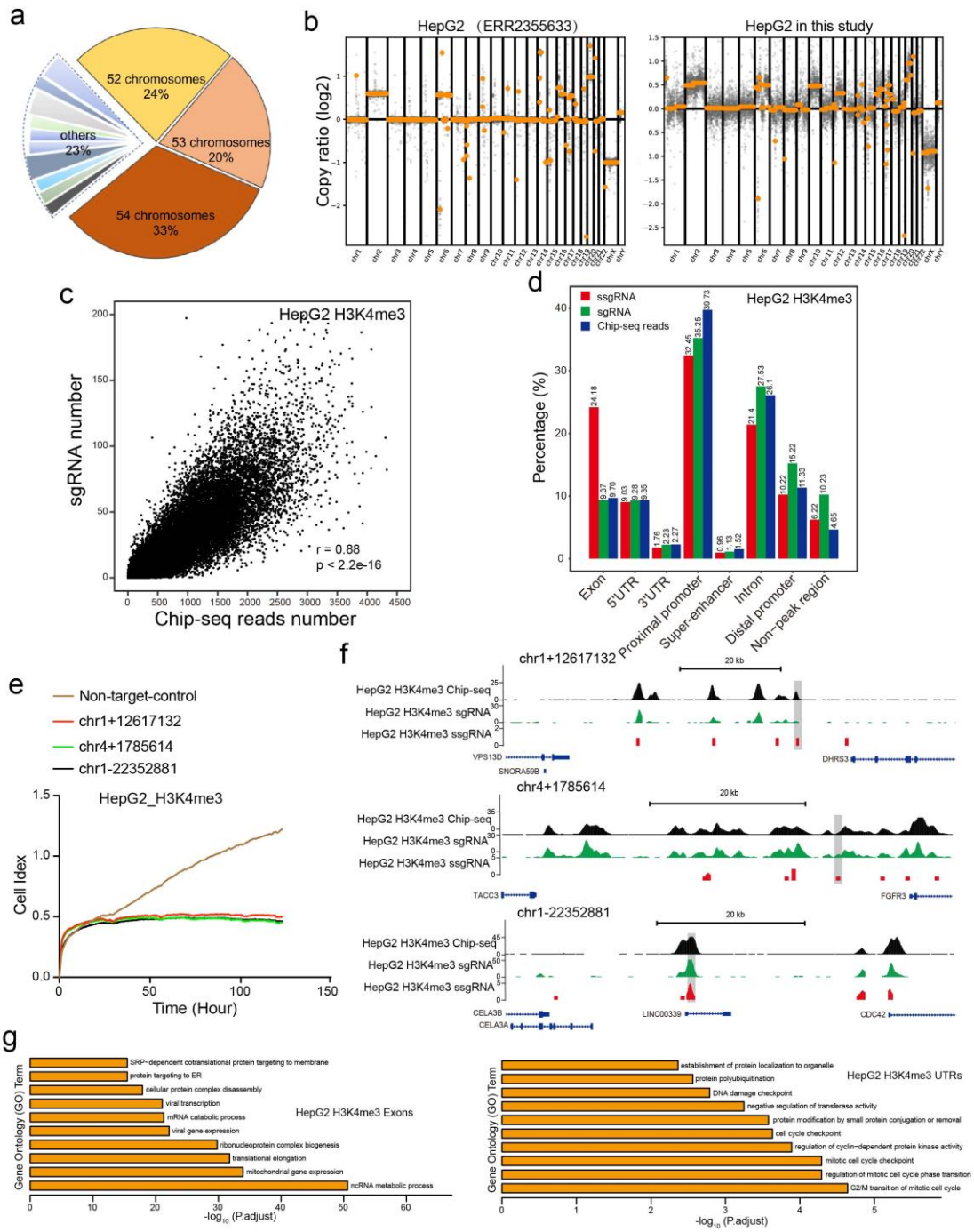

Figure S5

(a) Chromosome number of HepG2 used in this study based on 100 Giemsa-stained metaphases, which illustrates that the modal number was 52-54 chromosomes.

(b) Copy Number Variation (CNV) analysis of HepG2 in this study through low depth whole genome sequencing. The CNV distribution of HepG2 used in this study is similar with that of a standard HepG2 cell from Sequence Red Archive (SRA) database (ERR2355633).

(c) Scatter-plot of the correlation between ChIP-seq reads number and sgRNA number in HepG2 H3K4me3 elements. Spearson's correlation coefficient with two-tailed test was calculated.

819 (d) The distribution of H3K4me3 ChIP-seq reads, H3K4me3 sgRNAs and H3K4me3 ssgRNAs on
820 major regulatory regions of human genome.
821 (e) All three H3K4me3 ssgRNAs (validated in Figure 5d) can significantly inhibit HepG2
822 proliferation. Real-time cell numbers are plotted.
823 (f) Plots of H3K4me3 ChIP-seq reads, H3K4me3 sgRNAs and H3K4me3 ssgRNAs at three
824 genomic loci. Y-axes, RPKM. The genomic regions with H3K4me3 ssgRNAs (validated in Figure
825 5d) are shaded grey.
826 (g) Top 10 GO terms of the genes whose exons and UTRs are targeted by H3K4me3 ssgRNAs.
827

Figure S6

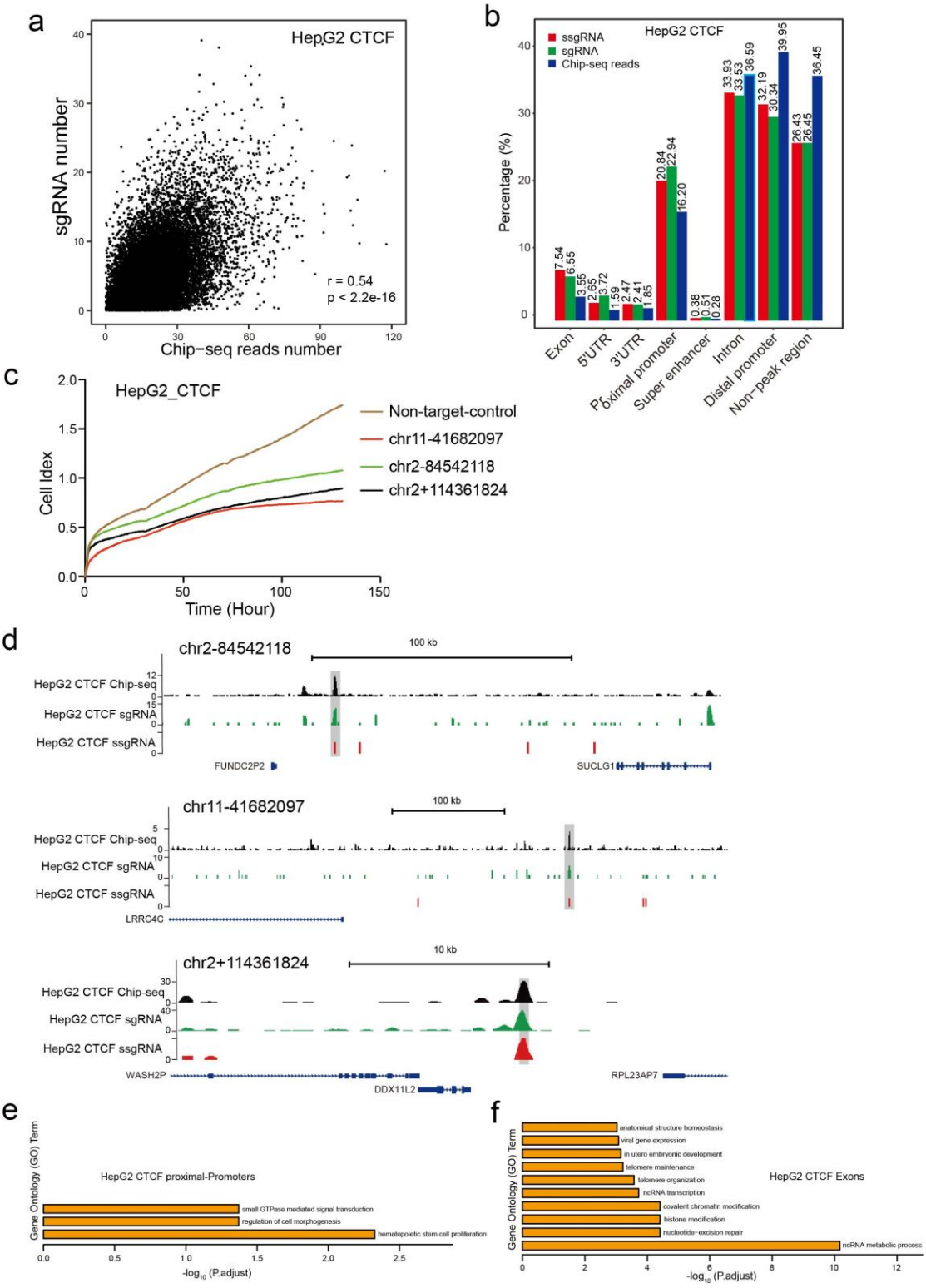

Figure S6

(a) Scatter-plot of the correlation between ChIP-seq reads number and sgRNA number in HepG2 CTCF elements. Spearson's correlation coefficient with two-tailed test was calculated.

(b) The distribution of CTCF ChIP-seq reads, CTCF sgRNAs and CTCF ssgRNAs on major

regulatory regions of human genome.
(c) All three CTCF ssgRNAs (validated in Figure 6d) can significantly inhibit HepG2 proliferation. Real-time cell numbers are plotted.
(d) Plots of CTCF ChIP-seq reads, CTCF sgRNAs and CTCF ssgRNAs at three genomic loci. Y-axes, RPKM. The genomic regions with CTCF ssgRNAs (validated in Figure 6d) are shaded grey.
(e) All three GO terms of the genes whose proximal promotors are targeted by CTCF ssgRNAs.
(f) Top 10 GO terms of the genes whose exons are targeted by CTCF ssgRNAs.

Methods and Materials

Part 1: Wet lab experiments

sgRNA library construction through CTDE (Controlled Template-dependent elongation)

1. 23bp DNA fragment generation

Source DNA is fragmented. We ligate them to the A1 adapter and capture them onto T1 streptavidin magnetic beads (Thermo Fisher, 65601). Then we denature DNA with 0.1M NaOH (Sigma, 79724) for 10min, and anneal primer for chain extension. We apply 2U Therminator™ DNA Polymerase (NEB, M0261) to incorporate one 3´ hydroxyl-reversible dNTP (Jena Bioscience, 3'-O-N_3-dNTPs). We restore the 3´ hydroxyl group by 100mM TCEP (Sigma, 646547) treatment and repeat the incorporation-and-reversion for 22 cycles. We blunt DNA with mung bean nuclease (NEB, M0250) for 30min, and then apply T4 PNK (NEB, M0201) to phosphorylate it for 30min.

2. NGG PAM selection

We ligate the A2 adapter and amplify the library for ten cycles. After the gel extraction of the amplified library, we digest the DNA with AscI (NEB, R0558). Then we capture the library onto streptavidin beads.

3. NGG PAM removal

We ligate the A3 adapter and amplify the library for ten cycles, and then digest the library with BbsI (NEB, R3539). We fill in the gap with T4 DNA polymerase (NEB, M0203), ligate A4 adapter, and amplify the library with the KAPA HiFi polymerase

mix (Roche, KK2631) for ten cycles. We apply 20% TBE-PAGE to select the size of the library (61nt). After releasing DNA from PAGE, we amplify the library by PCR with KAPA HiFi polymerase and primer (the sequence is below) and size-selected via 2% agarose gel[1].

ArrayF

TAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGAC

GAAACACCG

ArrayR

ACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCT

AGCTCTAAAAC

A1-F GAAAGGACGAAACACCgT

A1-R cGGTGTTTCGTCCTTTCCACaagatAGATCGGAAGAGCGTC-Biotin

Anneal Primer

tgGACGCTCTTCCGATCTATCTTGTGGAAAGGACGAAACACCGT

A2-F

CGCGCCCACACGTCTGAACTCCAGTCACAGTCAACAATCTCGTATGCCGT

CTTCTGCT

A2-R

CAAGCAGAAGACGGCATACGAGATTGTTGACTGTGACTGGAGTTCAGACG

TGTGGGCGCG

A3-F

TCTTCAAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGT
TATCCGC

A3-R

AGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGCCAA
GCTTGAAGA

A4-F GTTTTAGAGCTAGAAATAGCAAGTTAAA

Chromatin Immunoprecipitation (ChIP)

The ChIP assay was performed according to manufacture of Millipore ChIP kit (Millipore, 20-153). Briefly, formalin was added to cells to crosslink protein and DNA and then cells (10^6-10^7) were lysed in SDS lysis buffer. Chromatin was fragmented via sonication to the size of 200bp-500bp and then immunoprecipitated with antibodies (5-15μg each sample). The enriched DNA was purified by QIAquick PCR purification kit (QIAGEN). These DNA samples were applied to CTDE protocol or illumina sequencing library construction. The antibodies are anti-H3K4me3 (Cell Signaling, 9751) and anti-CTCF (Millipore, 07-729).

Cell culturing

V6.5 ESCs were maintained in DMEM supplemented with 15% fetal bovine serum (Gibco), 0.1mM β-mercaptoethanol, 2mM L-glutamine, 0.1mM nonessential amino acid, 1000U/ml recombinant leukemia inhibitory factor (Millipore LIF, ESG1107), and

907 30U/ml penicillin/streptomycin. For feeder-free culturing, ESCs were grown on plates
908 coated with 0.1% gelatin (Millipore, ES-006-B). HepG2 cells were cultured in DMEM
909 supplemented with 10% FBS (Gibco) and 1% penicillin/streptomycin.
910
911 **sgRNA library cloning into lentiCRISPR v2**
912 The lentiCRISPR v2 vector was digested with BsmBI (Fermentas, FD0454), treated
913 with alkaline phosphatase (Fermentas, EF0654) at 37°C for 2 hours and gel-purified on
914 a 1% TAE-Agarose gel. A 20μl Gibson ligation reaction (NEB, E2611) was performed
915 using 30ng of inserts and 200ng of digested vector. After ligation, 10μl of the reaction
916 was transformed into 100μl of T1 Resistant Chemically competent cells (TransGen
917 Biotech, CD501-02) according to the manufacturer's protocol. The sgRNA diversity of
918 one reaction will be 40-60 thousand. For example one library, whose template is mESC
919 H3K4me3 ChIPed DNA, contains 0.3-0.4 million kinds of sgRNAs, we took about 20
920 assembly reactions so that we can get representative sub-pool. For template is not a
921 linear DNA, increasing assembly reaction will not significantly enhance representation
922 of this library. Plasmid DNA was extracted using Plasmid Preparation Kit (Axygen,
923 AP-MN-P-250G).
924
925 **Lentivirus production and transduction**
926 Lentivirus was produced through the co-transfection of the lentiviral vectors with
927 psPAX2 (Addgene, #12260) and pMD2.G (Addgene, #12259) into HEK293T cells

using PEI (Polysciences, 24765). Virus-containing supernatant was collected and filtered through a 0.45μm low protein-binding membrane (Millipore, SLHV033RB) 48 hours after transfection. We performed spin-infection in medium containing polybrene (Sigma-Aldrich, H9268) at 1800 rpm for 45min at room temperature. According to Poisson distribution[2], we ensured a multiplicity of infection (MOI) of less than 0.3 to get single-infected percentage of over 80%. For a 30000-40000 sgRNA diversity sub-pool, we used 1.2×10^7 cells for each transduction. Transduced cells were selected 24 hours under puromycin (Gibco, A1113803) for 3 days. 1μg/ml for mESCs and 3μg/ml for HepG2. After puromycin selection (3 days for mESC, 5 days for HepG2), we applied these cells for dropout screen.

CRISPR/Cas9 based dropout screen

After puromycin selection, 4 million cells were seeded into a 10-cm dish every generation. 10 million cells from the 1st generation and 10th generation were used for genomic DNA extraction and sgRNA-template targeted sequencing. For DNA extraction, we used ethanol-precipitated method to avoid DNA loss. Cell pellet was lysed in lysis buffer (1% SDS, 50mM Tris, 1mM EDTA, 20μg Protease K, 10μg RNase A) in 65℃ water bath overnight. Cell lysate was denatured by phenol:chloroform:isopentanol (25:24:1), and thoroughly mixed by shaking. After centrifugation, DNA was precipitated from supernatant by 2 volume 100% ethanol. We

washed DNA pellet twice with chilled 70% ethanol. After air drying, we dissolved DNA in ddH_2O.

sgRNA-template targeted sequencing

Two PCR steps were performed:

1. PCR from enough genomic DNA to preserve library complexity.

sgRNA template containing cassette was amplified using primers specific to the lentiCRISPR v2 vector (Round1-F and Round1-R)

Round1-F AATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCG

Round1-R CCAACTTCTCGGGGACTGTGGGCGATGTGCGCTCTGCCCACTGA

24 × 50µl PCR reactions were performed with 2µg genomic DNA in each tube to achieve over 100 × coverage. 20 cycles were used for minus bias.

2. Add illumina sequencing adapters.

10µl DNA from first round PCR product (combined all tubes together) was used as template for the second round PCR. Second round PCR product was purified by QIAquick PCR clean-up Kit and quantified by Agilent 2200 Bio-analyzer before illumina sequencing.

Round 2-F

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT

CCGATCTtcttgtggaaaggacgaaacaccg

Round2-R

CAAGCAGAAGACGGCATACGAGATxxxxxxGTGACTGGAGTTCAGACGTGT

GCTCTTCCGATCTcgctctgcccactgacgggcaccgg (xxxxxx is a barcode)

Anti-neomycin cell-line preparation

Anti-Neomycin mESCs cells were performed as described with modifications[3]. Briefly, the pEF1a-BirA-V5-neo (#100548, Addgene) plasmid was stably transfected into V6.5 ESCs via electroporation. Cells were selected by neomycin and were applied to screen experiment.

Commercial synthesis Neo sreen library contruction

According to 795bp neo fragment sequence, we pick all gRNA with NGG PAM, totally 115 gRNA. Also, we chose 20 gRNA from Gecko negative control (they can't target hg19, mm9 and neo reference sequence). We order these gRNA from oligo synthesis company (Shangya), and these oligoes were in uniform format: 5'-GTGGAAAGGACGAAACACCGNNNNNNNNNNNNNNNNNNNGTTTTAGA GCTAGAAATAGC-3', where N20 represent neo gRNA or negative control sequence. We combined 115 neo gRNA and 20 negative control equally together, and amplified 5µl library(10µM) with 2X HiFi DNA polymerase, Array-F and Array-R primer for 5 cycles. 140bp library was gel extracted and assembled into LentiCRISPR V2 plasmid. 20 negative control were separately prepared as above and assembled into LentiCRISPR v2 plasmid, and this library was spike-in of CTDE-neo library.

Neo fragment obtaining

We obtained neo fragment directly by PCR using pEF1a-BirA-V5-neo (Addgene, #100548) as template, and gel extract 795bp DNA as source of CTDE method to construct CTDE neo library.

Neo library Screen experiment

After puromycin selection, 1 million cells were seeded into a 10-cm dish，and 1 million cells were collected and labelled with Prev-selection. Anti-Neomycin mESCs infected by CTDE-Neo and Commercial synthesis Neo library were cultured with 150μg/ml Neomycin（Sangon）for 7 days. 7 days later, cells were collected and labelled with Post-selection. Prev- and Post-selection sample were extracted genomic DNA and illumina sequencing library were construct as above.

Plasmid construction

Oligoes (for sgRNA or shRNA) were annealed by gradient cooling process and were ligated to lentiCRISPR v2 (BsmBI digested) or PLKO.1 (digested by AgeI/EcoRI) using T4 DNA ligase. After transformation, colonies were picked. The confirmed colonies by sanger sequencing were preserved in glycerol and plasmid were applied to transfection. The CDC42 cDNA was obtained from ORF collection (Ultimate™ ORFs,

GeneID: 998). The entire open-reading frame was amplified by PCR and cloned into PLVX vector at BamHI and XhoI sites.

AP staining and survival assay

mESCs transduced with lenti-sgRNA were first selected under 1μg/ml puromycin for 3 days. AP staining was performed according to the manufacturer's instructions (Beyotime Biotechnology, C3206). HepG2 cells transduced with sgRNA or LINC00339-shRNA were plated 5 days after transduction at 1.8×10^5 cells per well in a 24-well plate. Then cells were plated in 6-well plates at a density of 3000 cells per well. The culture media with 1μg/ml puromycin was refreshed every 3 days for ~14 days. After PBS washing, colonies were stained with 0.1% crystal violet (Beyotime, C0121) for 10min at room temperature. Then the images of each well were recorded by a digital camera and the number of colonies was counted using ImageJ software.

Karyotype analysis

HepG2 (10^5) were trypsinized and resuspended in 5ml 0.56% KCl in water. After incubation at 37℃ for 10min, 1ml 25% freshly made fixative solution (methanol:glacial acetic acid = 3:1 by volume) in water was added. Cells were pelleted and resuspended in 1ml 100% fixative solution three times. To make cells spread, one drop of the cell suspension was dropped (from a height of 0.5m) onto a glass slide and allowed to air

dry before DAPI staining. The mitotic chromosome number was counted under the fluorescence microscopy.

Cell cycle analysis

Cells were transduced with lentivirus via spinfection in 6-well plates. After puromycin selection, cells were harvested and fixed in chilled 70% ethanol at -20℃ overnight, and stained with 40μg/ml propidium iodide (Sigma-Aldrich, P4170) containing 50μg/ml RNase (Sigma-Aldrich, R6513) at 37℃ for 15min in the dark, then analyzed cell cycle by FACSCalibur flow cytometer (BD Biosciences). Data were analyzed with ModFit LT software.

qPCR quantification

Total RNA was harvested using the RNeasy Plus Mini Kit (Qiagen, 74134) and 1μg RNA was used for reverse transcription with the PrimeScript™ RT Master Mix (Takara, RR036A). After reverse transcription, TB Green® Fast qPCR (Takara, RR430A) was performed with related primers (Table S14).

Western blot

Cells were harvested and lysed in SDS-PAGE sample buffer. Equal amounts of total protein were loaded in each lane. Proteins were resolved by SDS-PAGE, transferred to 0.45μm PVDF membranes, and probed with the indicated antibodies. The antibodies

used for western blotting are: anti-CDC42 (Cell Signaling Technologies, 2466), anti- β-tubulin (Cell Signaling Technologies, 2128).

Real-Time cell number analysis.

Experiments were performed using Real-Time Cell Analyzer (iCELLigence Analyzer, ACEA Biosciences Inc.) at 37℃ with 5% CO2. To measure the background, we placed 200μl DMEM medium in the E-plate L8 (ACEA Biosciences, 00300600840). Then, we added 400μl HepG2 (transduced with sgRNA or shRNA) cell suspension (8000 cells/well) with 10% FBS DMEM medium. The impedance (cell index), which was detected by sensor electrodes in E-plate L8 every 15min for 120 hours. The plot of the cell index was calculated automatically by the RTCA software package 1.1.1 (ACEA, Biosciences Inc.).

Part 2: Dry lab experiments and data analysis

Sequencing data pre-processing

The designed lentiCRISPR v2 sgRNA fragments were sequenced. The sequenced read comprises a 5'-adapter, a sgRNA, and a 3'-adapter, one by one. We trimmed the read, remove the pre-designed 5'-adapter and 3'-adapter (Table S14) to get the sequence of the sgRNA. Then we adhered "NGG" to the end of each sgRNA to recover its PAM sequence and prepared it for mapping on the reference genome.

The assessment of sgRNA library generated by CTDE from plasmid lentiCRISPR v2

1. We evaluated the detected sgRNAs from the designed reference of plasmid lentiCRISPR v2 (Addgene #52961, 14873bp) with the following steps:

Step 1, we built a mapping reference set from the plasmid lentiCRISPR v2 sequences using Bowtie build-index function.

Step 2, we prepared three sgRNA libraries with the CTDE approach. For each of the libraries, we sampled one million reads and then extracted 19-21 bp length sgRNAs from the "Sequencing data pre-processing" analysis. We merged three sgRNA libraries to get a sgRNA set.

Step 3, we aligned the sgRNAs set from step2 to the reference set by the bowtie aligner, allowing one base mismatch and reporting all alignments (-v 1 -a).

Step 4, we counted the number of mapped sgRNAs as x. We identified the "NGG" locus on either sense or antisense strands of the plasmid lentiCRISPR v2 sequences. For locus i, we counted the number of sgRNAs whose PAM tails mapped to the locus as y. Then we calculated the ratio of sgRNAs for locus i (Ratio_{locus_i}) as follows:

$$\mathrm{Ratio}_{locus\_i} = \frac{y}{x} \times 100$$

We repeatedly calculated the number and ratio of sgRNAs for each locus. Next we divided lentiCRISPR v2 sequences into equal-sized 10 bp bins and calculated the ratio of AT for each bin. We finally used Integrative Genomics Viewer (IGV) (v2.8.0) to

display the ratio of sgRNAs in each "NGG" locus and AT in each bin of the lentiCRISPR v2 sequences.

2. We calculated the coverage of the detected sgRNAs on the "NGG" locus in the designed reference with the following steps:

Step 1, we count all possible the "NGG" loci in the plasmid lentiCRISPR v2 sequences as N_{total}.

Step 2, we count the number of "NGG" loci covered by sgRNAs from the sgRNA set described above as $N_{covered}$. Then we calculated the coverage of the detected sgRNAs on the "NGG" locus in the designed reference (Cover_{SCTE}) as follows:

$$\mathrm{Cover}_{SCTE} = \frac{N_{covered}}{N_{total}}$$

3. We next analyzed the enrichment of the detected sgRNAs on the "NGG" locus with the following steps:

Step 1, we sequenced the plasmid lentiCRISPR v2, then we trimmed the sequenced reads to 23bp and aligned them to the reference set using bowtie, not allowing base mismatch (-v 0). The reads mapped on the plasmid lentiCRISPR v2 sequences and mapped on the "NGG" locus of the sequences were counted. Next, we calculated the percentage of reads on the "NGG" locus from the mapped reads as $P_{NGG-PAM}$.

Step 2, we sampled one million reads from each of three sgRNA libraries, then we got the sgRNAs from the "Sequencing data pre-processing" analysis and extracted the sgRNAs of 20bp length.

Step 3, we processed the sgRNAs in two different ways. In one way, we removed their "NGG" tails and mapped them to the reference set by bowtie, not allowing base mismatch (-v 0). Simultaneously, we mapped them to the reference set by bowtie, allowing one base mismatch (-v 1). After alignment, we counted the number of mapped sgRNAs as $N_{NGG-PAM}$. Then the fold enrichment of the detected sgRNAs on the "NGG" locus（$Enrich_{SCTE}$）was estimated as follows:

$$Enrich_{SCTE} = \frac{N_{NGG-PAM}}{N_{total} \times P_{NGG-PAM}}$$

4. We grouped the detected sgRNAs by their length and calculated the proportion of sgRNAs of different length with the following steps:

Step 1, for each of the three sgRNA libraries, we got sgRNAs from the "Sequencing data pre-processing" analysis and mapped them to reference set by bowtie, allowing one base mismatch (-v 1).

Step 2, we grouped sgRNAs by their length. For the sgRNAs of length i, we counted the number as N_i, and calculated their proportion as follows:

$$Proportion_i = \frac{N_i}{N_{all}}.$$

Here N_{all} is the number of all mapped sgRNAs. We repeatedly calculated the proportion for the sgRNAs of 16-24bp from the three sgRNA libraries.

5. To evaluate the how faithful sgRNA synthesis was, we mapped 16-24bp sgRNA to reference set by the following priority order: with no mismatch, with one mismatch, with two mismatches (not include mismatch in PAM site). For PAM site, we only tolerate mismatch at the first base. The error rate was calculated as total number of

mismatch base divided by total number of base of mapped sgRNA (not include PAM site).

Comparison between CTDE and commerical method library in Neo resistant screen

To demonstrate the performance of CTDE, we generated two libraries based on the Neo sequence using CTDE and commercial methods (Table S14). Twenty non-targeting control sgRNAs randomly selected from GeCKO v2 library were added to two libraries. After illumina sequencing and initial data processing, the number of reads were counted with an in-house script for all libraries. Then library sizes were normalized using total reads count of control sgRNA in pre- and post-selection library. After normalization, the fold change of each sgRNA between post- and pre-selection library was calculated. Targeting sgRNAs whose fold change less than the minmum fold change of non-targeting control sgRNAs were recognized as ssgRNA.

Comparison between NSgRNAShot and MAGeCK in identify ssgRNA

We downloaded one genome-wide CRISPR-Cas9 dropout screen data of mESCs as the benchmark data[4]. MAGeCK is a widely used method to identify of essential genes from genome-scale CRISPR/Cas9 knockout screens[5]. As described in the previous paper, since the CRISPR/Cas9 knockout system should show no difference in selection preference between control samples or between replicated treatment samples, a good

method should not detect many significantly selected sgRNAs and genes between these samples[5]. So we took the strategy to detect the possible false-positive ssgRNA with FDR < 0.1 between two replicates of ESC treatment samples with NSgRNAShot and MAGeCK based on benchmark data. When multiple sgRNAs are available for one gene, MAGeCK demonstrated the best performance to detect essential genes[6]. Therefore we evaluated the sensitivity of ssgRNA identified by our method with the following strategy: we took the essential genes reported by MAGeCK as the gold standard set, a good method should largely report the ssgRNA from essential genes but not from other genes.

ChIP-seq data analysis

For ChIP-seq of H3K4me3 and CTCF in mESC and HepG2 cell lines, libraries were sequenced using Illumina HiSeq X Ten and paired-end 150 bp long reads were obtained. Chip-seq reads were analyzed with the following steps:

Step 1, we trimmed Chip-seq reads to 100 bp (from 5' to 3'), and for redundant reads which have the same sequence, only one was retained. Then reads were mapped to the reference genome (mm9 for mESC or hg19 for HepG2) by bowtie, only uniquely mapped reads were reported (-m 1).

Step 2, Chip-seq peaks for H3K4me3 and CTCF were detected by MACS2 with default settings. For each mouse ES CTCF peak, we calculated its length ($peak\ length$) and

counted the number of reads mapped on the peak ($number\ of\ mapped\ reads$), then we calculated its average read depth ($average\ depth\ per\ peak$) as follows:

$$average\ depth\ per\ peak = \frac{reads\ length \times number\ of\ mapped\ reads}{peak\ length}$$

We filtered peaks whose average read depth less than ten.

Step 3, we manually identified regions which consistently contain significantly enriched reads in multiple input ChIP-Seq data provided by Encode project (Table S15) and regarded them as false positive peak regions. We filtered the peaks overlapped with these regions.

RNA-seq data analysis

1. We downloaded two total RNA-seq datasets of HepG2 from Gene Expression Omnibus (GEO) database (GEO:GSE88089), and two total RNA-Seq datasets of mESC from Beijing Institute of Genomics (BIG) Data Center (CRA001133).

2. For each RNA-seq dataset, raw reads were mapped to reference genome (mm9 for mESC or hg19 for HepG2) by bowtie, allowing one base mismatch and only uniquely mapped reads were retained (-v 1 -m 1). For redundant reads that fall on the same position on the genome, only one was retained.

3. Based on the UCSC knownGene annotations, we counted mapped reads in the exon regions of transcripts, and then we calculated transcript expression levels by normalizing the number of reads for each transcript to total mapped reads and mRNA length, namely reads per kilo-bases per million reads (RPKM).

4. We got the average RPKMs of transcripts for mESC and HepG2, respectively, and the largest RPKM of transcripts for each gene represents gene expression level. We selected the genes with RPKM higher than one as expressed genes.

Whole-genome sequencing data analysis

SNP (single nucleotide polymorphism) existed in the human and mouse population. To identified the sgRNAs from genomic regions with SNP, we built additional reference from the regions with the following steps:

1. We downloaded one whole-genome sequencing (WGS) dataset of HepG2 from the Sequence Read Archive (SRA) database (ERR2355633). Moreover, the WGS library of mESC was sequenced using Illumina HiSeq X Ten, and paired-end 150 bp long reads were obtained.

2. For each WGS dataset, paired-end reads were merged together. Then we trimmed adapter residues from reads by AdapterRemoval (Version 2.1.7), and reads shorter than 150 bp after trimming were filtered out.

3. We performed step-wise mapping process, step 1, reads were mapped to the reference genome (mm9 for mESC and hg19 for Hepg2) using bowtie2 with default settings. Step 2, mapped reads from step 1 were mapped to the reference genome using bowtie, not allowing base mismatch (-v 0). Unmapped reads from step 2 were regarded as derived from regions of mESC or HepG2 with SNP, we used these uniquely unmapped reads to build a mapping reference set using the Bowtie build-index function.

Identification of negatively selected sgRNAs

1. The sgRNA libraries generated by H3K4me3 or CTCF labeled DNA sequences were sequenced using Illumina HiSeq X Ten. For each sgRNA library, we got sgRNAs from "Sequencing data pre-processing" analysis.

2. We selected 18-21 bp length sgRNAs which have ability to direct Cas9-induced indels at target sites for further analysis. Then we performed step-wise mapping process. Step 1, sgRNAs were mapped to reference genome (mm9 for mESC and hg19 for HepG2) using bowtie, allowing one base mismatch (-v 1). Step 2, unmapped sgRNAs from step 1 were mapped to reference set from "Whole-genome sequencing data analysis" using bowtie, allowing one base mismatch (-v 1). We counted the number of each mapped sgRNA from above two steps.

3. The sgRNAs whose abundance had significantly reduced from P1 to P10 were identified using our designed NSgRNAShot algorithm. We used the term 'ssgRNA' to refer to negatively selected sgRNA.

Genome-wide distribution of chip-seq reads, sgRNAs and ssgRNAs

We analyzed the genomic distribution of chip-seq reads, sgRNAs, and ssgRNAs with following steps:

1. We got sgRNAs from “Identification of negatively selected sgRNAs”. Then we mapped sgRNAs to reference genome (mm9 for mESC and hg19 for HepG2) using bowtie, allowing one base mismatch (-v 1).

2. We got ssgRNAs from “Identification of negatively selected sgRNAs” and mapped them to reference genome using bowtie, allowing one base mismatch and refraining from reporting any alignments have more than three reportable alignments (-v 1 -a -m 3). For ssgRNAs having more than one reportable alignment, we only retained alignments located in peak region, otherwise, all alignments were retained.

3. We calculated the chip-seq reads, sgRNAs, and ssgRNAs density on each 100 bp window and sliding along the chromosomes with a step length of 20 bp. We used Circos (v0.69-6) tool to display the genome-wide density distribution of chip-seq reads, sgRNAs and ssgRNAs, and local density distribution was visualized by the UCSC Genome Browser.

4. To display the distribution of chip-seq reads, sgRNAs and ssgRNAs in functional elements of the genome, we assigned annotation for each nucleotide in the reference genome by using the following priority order: coding-exon > 5’-UTR > 3’-UTR > proximal promoter (10kb upstream of a TSS) > super-enhancer > intron > distal promoter (>10kb upstream of a TSS) = non-peak region. We assigned chip-seq read, sgRNAs and ssgRNAs to these categories based on location.

5. To demonstrate what the distribution of chip-seq reads, sgRNAs and ssgRNAs per peak look like, we split each peak into non-overlapping 30 bins. Then for each of these

non-overlapping bins, RPKM of chip-seq reads, sgRNAs and ssgRNAs was calculated for visualization separately. Peaks were ranked from highest to lowest based on median RPKM of chip-seq reads of bins.

Recall rate of essential genes from CTDE

We downloaded two gene sets described as essential for survival and proliferation of mESCs[4, 7]. The overlap of two gene sets was used as reference essential genes for mESCs. We mapped sgRNAs and ssgRNAs to reference genome (mm9) using bowtie (-v 1 -m 1), essential genes whose exons were targeted by more than 2 sgRNA were regarded as covered by CTDE. The ratio of essential genes, which are covered by CTDE, targeted by ssgRNA is the recall rate of essential genes.

Gene Ontology (GO) analysis

We assessed whether related genes of exon, intron, utr, and proximal promoters covered by ssRNAs were enriched for particular GO categories by calculating adjusted P-value using the Benjamini & Hochberg method. We performed GO analyses using R package clusterProfiler with default parameters.

1. For uniquely mapped ssgRNAs generated by H3K4me3 labeled DNA sequences, only related expressed genes (FPKM > 1) of an exon, utr, intron, and proximal promoters with ssgRNAs located were used to perform GO analysis.

2. For uniquely mapped ssgRNAs generated by CTCF labeled DNA sequences, all related genes of utr, intron and proximal promoters with ssgRNAs located were used for GO analysis, but only related expressed genes of exons with ssgRNAs located were used for GO analysis.

3. We sorted all GO categories according to adjusted P-values in an ascending order.

Analysis of CTCF hallmarks shared occupation in different cell types

1. We downloaded CTCF hallmarks of 16 mouse cell lines and tissues from ENCODE, the CTCF hallmarks of mESC were occupied in higher than 9 of 16 mouse cell lines and tissues were defined as common hallmarks.

2. We downloaded CTCF hallmarks of 55 human cell lines from ENCODE, the CTCF hallmarks of HepG2 were occupied in higher than 33 of 55 human cell lines were defined as common hallmarks.

3. We mapped ssgRNAs from “Identification of negatively selected sgRNAs” to reference genome (mm9 for mESC and hg19 for HepG2) using bowtie, allowing one base mismatch and reported uniquely mapped reads (-v 1 -m 1). Based on the position of CTCF hallmarks and ssgRNAs in the genome, we calculated the distribution of ssgRNAs in common and cell-type specific CTCF hallmarks. We used the term ‘essential common peaks’ to refer to common CTCF hallmarks with ssgRNAs located.

Negative-SgRNA-Shot (NSgRNAShot) method

We assumed that in the dish, during culture, 1) most sgRNAs do not change their abundance, which suggests they suffer no selection pressure; 2) no sgRNAs is undergoing positive selection and increases the abundance; 3) a few sgRNAs are suffering negative selection and decrease their abundance. Then we developed a method named Negative-SgRNA-Shot (NSgRNAShot) to detect significant-negatively-selected sgRNAs (ssgRNA). Here are the details.

Step 1. We got reads counts (r) profiles of sgRNAs at 1st generation (p1) and 10th generation (p10) in the dish from "Identification of negatively selected sgRNAs" analysis. Suppose we have detected n and m sgRNAs from p1 and p10, then we calculated the sequencing depth of sgRNAs profiles at p1(Sum_{p1}) and p10(Sum_{p10}) with the following formula,

$$\mathrm{Sum}_{p1} = \sum_{i=1}^{n} r_i,\ \mathrm{Sum}_{p10} = \sum_{j=1}^{m} r_j;$$

If $\mathrm{Sum}_{p1} \geq \mathrm{Sum}_{p10}$, we calculated the normalized reads counts (nr) of sgRNAs at p1 with the formula: $\mathrm{nr}_i = \mathrm{r}_i,\ where\ 1 \leq \mathrm{i} \leq \mathrm{n}$; and we calculated the normalized reads counts of sgRNAs at p10 with the formula: $\mathrm{nr}_j = \mathrm{r}_j * (\mathrm{Sum}_{p1}/\mathrm{Sum}_{p10}),\ where\ 1 \leq \mathrm{j} \leq \mathrm{m}$.

If $\mathrm{Sum}_{p1} < \mathrm{Sum}_{p10}$, we calculated the normalized reads counts of sgRNAs at p1 with the formula: $\mathrm{nr}_i = \mathrm{r}_i * (\mathrm{Sum}_{p10}/\mathrm{Sum}_{p1}),\ where\ 1 \leq \mathrm{j} \leq \mathrm{n}$; and we calculated the normalized reads counts of sgRNAs at p10 with the following formula: $\mathrm{nr}_j = \mathrm{r}_j,\ where\ 1 \leq \mathrm{j} \leq \mathrm{m}$.

Step 2. We calculated log2 fold change (log2fc) value of abundance of a sgRNA from p1 and p10 with the following formula:

$$\mathrm{log2fc} = \mathrm{log2}\left(\frac{\max\{\mathrm{nr}_{p1},10\}}{\max\{\mathrm{nr}_{p10},10\}}\right),$$

where nr_{p1} and nr_{p10} are the normalized reads count of the sgRNA at p1 and p10. However, if the sgRNA is not detected at either p1 or p10, we set $\mathrm{nr}_{p1} = 0$ or $\mathrm{nr}_{p10} = 0$, correspondingly.

With that, we calculated log2fc of all sgRNAs in the dish.

Step 3. At first, we made some deductions. In the dish, we assumed the number of the negatively-selected sgRNAs is no more than 20% of the non-selected sgRNAs. We assumed the log2fc of the former ones follows a normal distribution $\mathrm{N}(\mathrm{u}, \sigma^2)$, and the log2fc of the latter ones follows a normal distribution $\mathrm{N}(0, \sigma^2)$. Here $\mathrm{u} > 0$ and $\mathrm{u} > 4\sigma^2$. Then the log2fc of sgRNAs in the dish follows a mixture-normal distribution. The mode, which is the value occurring a maximum number of times in the distribution, will indicate the center of the normal distribution of the log2fc from the non-selected sgRNAs and should be zero.

Then, we conducted the analysis. We calculated the distribution of log2fc of all sgRNAs in the dish, find mode (m) of the distribution and correct log2fc of the sgRNAs with the following formula:

$$\mathrm{log2fc\_m} = \mathrm{log2fc} - \mathrm{m}.$$

With that, the log2fc_m of sgRNAs in the dish follows a mixture-normal distribution. The mode of the distribution indicates the center of the normal distribution of the log fold change from the non-selected sgRNAs and is zero.

Step 4. Again, we made some deductions at first. The previous deductions in step 3 demonstrated the log2fc_m of the negatively-selected sgRNAs follows the normal distribution $\mathrm{N}(\mathrm{u}, \sigma^2)$. We have assumed that $\mathrm{u} > 4\sigma^2$, thus the negatively-selected sgRNAs whose $\mathrm{log2fc\_m} < 0$ should occupy 4.5% of all negatively-selected sgRNAs. However, the percent of negatively-selected sgRNAs is less than 20% of non-selected sgRNAs in the dish. Thus the negatively selected sgRNAs whose $\mathrm{log2fc\_m} < 0$ should occupy less than 1% (0.2/(1+0.2)×0.045=0.0075) sgRNAs in the dish. Since this percent is relatively small, we assumed that the sgRNAs with $\mathrm{log2fc\_m} < 0$ are from the non-selected sgRNAs but not from the negatively-selected sgRNAs.

Then, we conducted the analysis. We randomly picked a sgRNA k with the positive log2fc_m ($\mathrm{log2fc\_m}_k$) from the dish. We took the sgRNAs in the dish whose $\mathrm{log2fc\_m} < -\mathrm{log2fc\_m}_k$ being from the non-selected sgRNAs, and recorded their number as x_1. We counted the ones whose $\mathrm{log2fc\_m} > \mathrm{log2fc\_m}_k$ in the dish as x_2. If $x_1 > 0$ and $x_2 > x_1$, then we can deduce that x_1 sgRNAs are from the non-selected sgRNAs and $x_2 - x_1$ sgRNAs are from the negatively-selected ones in all the x_2 sgRNAs. Then, taking $\mathrm{log2fc\_m}_k$ as the threshold, we can calculate the false discovery rate (fdr_k) of the negatively-selected sgRNAs as:

$$\mathrm{fdr}_k = x_1/x_2.$$

Step 5. We went through any sgRNA l in the dish with positive log2fc_m (log2fc_m$_l$), took the log2fc_m$_l$ as a threshold, recorded the number of sgRNAs whose log2fc_m $>$ log2fc_m$_l$ in the dish as x_{2_l}, and calculated the fdr (fdr$_l$) of the negatively-selected sgRNAs. Then we paired log2fc_m$_l$, fdr$_l$, x_{2_l}, recorded the pair of any sgRNA to get the pairs set. Next, we screened the pairs set and selected the pairs whose fdr$_l$ < 0.1, and further identified the one with the largest x_{2_l} in the selected pairs and recorded its log2fc_m$_l$. Finally, taking the log2fc_m$_l$ as the threshold, we identified the sgRNAs whose log2fc_m $>$ log2fc_m$_l$ in the dish as the ssgRNA. However, if we cannot find any pair whose fdr < 0.1 in pairs set, we claimed no ssgRNA in the dish.

Step 6. We can get the ssgRNAs by conducting the steps 1-5. However, some ssgRNAs may not be saturated sequenced and have small reads counts at either p1 or p10. The change of their abundance from p1 to p10 was suspected. With that, we modified step 2 to give weights to the log2 fold change (log2fc) of the sgRNAs in the dish. At first, we selected the larger one of its normalized reads counts (nr) at p1 and p10 for each sgRNA. Then we ranked all sgRNAs by their selected nr in descending order. Next, we calculated its percentile (perct) for each sgRNA in the rank list. At last, picking a sgRNA z with the percentile value perct$_z$, we calculated its log2fc (log2fc$_z$) with the following formula,

$$\text{log2fc}_z = \text{perct}_z * \text{log2}(\frac{\max\{\text{nr}_{p1},10\}}{\max\{\text{nr}_{p10},10\}}),$$

where nr_{p1} and nr_{p10} are the normalized reads count of the sgRNA at p1 and p10. However, if the sgRNA is not detected at either p1 or p10, we set $\mathrm{nr}_{p1} = 0$ or $\mathrm{nr}_{p10} = 0$, correspondingly.

We modified the log2fc of all sgRNAs in the dish by this approach.

Finally, we conducted step 1,6,3,4 and 5 to detect ssgRNAs in the dish.

Evaluation of Negative-SgRNA-Shot method

We designed a simulation experiment to test the performance of NSgRNAShot.

At first, we generated a series of simulated datasets.

Step 1. We picked a dish. We conducted steps 1 to 3 of NSgRNAShot to get the log2fc_m of all the sgRNAs in the dish. We also recorded the normalized reads count of the sgRNAs at p10 (nr_{p10}). For the sgRNAs who got no reads count at p10, we set their nr_{p10} as zero.

Step 2. We counted sgRNAs whose $\mathrm{log2fc\_m} < 0$ in the dish and recorded the number as r. We randomly picked a sgRNA p from the r sgRNAs and recorded its log2fc_m as $\mathrm{log2fc\_m}_p$. Then we tried to find a sgRNA q and recorded its log2fc_m as $\mathrm{log2fc\_m}_q$ where $|\mathrm{log2fc\_m}_p + \mathrm{log2fc\_m}_q| < 0.05$, thus we got a sgRNAs pair whose log2fc_m values are symmetrical about zero.

We repeated the process $r/2$ times, got $r/2$ sgRNAs pairs to form a sgRNAs set whose log2fc_m values are symmetrical about zero. Since we had recorded the nr_{p10}

of the sgRNAs in the set, we inferred their normalized reads count of the sgRNAs at p1 (nr_{p1}) with the following formula,

$$\mathrm{nr}_{p1} = \mathrm{nr}_{p10} * 2^{\mathrm{log2fc\_m}} \quad .$$

Here we got a sgRNAs set with the normalized reads count at both p1 and p10. We saw the set as the simulated non-selected sgRNAs set.

Step 3. We calculated variation (σ^2) of the simulated non-selected sgRNAs set. We randomly picked α ($0<\alpha<0.2$) percent of the non-selected sgRNAs, calculated their log2fc_m and modified their log2fc_m values to $\mathrm{log2fc\_m} + \beta\sigma^2$, where ($\beta \geq 4$). We had recorded the nr_{p10} of these sgRNAs, we inferred their normalized reads count of the sgRNAs at p1 (nr_{p1}) with the following formula,

$$\mathrm{nr}_{p1} = \mathrm{nr}_{p10} * 2^{\mathrm{log2fc\_m}+\beta\sigma^2} \quad .$$

Thus we got a sgRNAs set with the normalized reads count at both p1 and p10. We saw the set as the simulated negatively-selected sgRNAs set. Finally, we combined the simulated non-selected and negatively-selected sgRNAs sets to get a simulated sgRNAs dataset.

Step 4. We took (α, β) in step 3 as the parameter pair. We set the parameter pairs as (0.01,6), (0.01,8), (0.05,4), (0.05,6), (0.05,8), (0.1,4), (0.1,6), (0.1,8), (0.2,4), (0.2,6), and (0.2,8). We repeated steps 3 and 4 one hundred times with each parameter pair to generate the simulated datasets. We randomly took a dataset under each parameter pair. We plotted the normalized reads count of the non-selected and negatively selected sgRNAs in the dataset and calculated the distribution of the log2 fold change of these

sgRNAs. The scatter plot and distribution plot showed that the simulation is successful because they tend to fit the two plots in the real dataset (Figure S1a).

To this step, we had generated a series of simulated datasets. Then, we tested the performance of NSgRNAShot on detecting the negatively-selected sgRNAs at the simulated datasets. We used two indicators-precision ($prec$) and recall (rec), which are calculated with the following formula, to benchmark NSgRNAShot,

$$prec = \frac{TP}{TP+FP}, \text{ and } rec = \frac{TP}{TP+FN},$$

where TP is the number of negatively-selected sgRNA identified as ssgRNA; FP is the number of non-selected sgRNA identified as ssgRNA; FN is the number of negatively-selected sgRNA failed to be identified as ssgRNA.

We run NSgRNAShot on the simulated datasets to detect negatively-selected sgRNAs. We can see that in the datasets whose parameter pairs are (0.05,8), (0.1,6), (0.1,8), (0.2,6), and (0.2,8), the precision and recall indicators are both higher than 80% (Table S1). We believed NSgRNAShot could successfully identify the negatively-selected sgRNAs in the circumstances.

References

1. Shalem, O. et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014).
2. Chen, S. et al. Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* **160**, 1246-1260 (2015).
3. Kim, J., Cantor, A.B., Orkin, S.H. & Wang, J. Use of in vivo biotinylation to study protein-protein and protein-DNA interactions in mouse embryonic stem cells. *Nature protocols* **4**, 506-517 (2009).

4. Tzelepis, K. et al. A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep* **17**, 1193-1205 (2016).
5. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome biology* **15**, 554 (2014).
6. Bodapati, S., Daley, T.P., Lin, X., Zou, J. & Qi, L.S. A benchmark of algorithms for the analysis of pooled CRISPR screens. *Genome biology* **21**, 62 (2020).
7. Shohat, S. & Shifman, S. Genes essential for embryonic stem cells are associated with neurodevelopmental disorders. *Genome research* **29**, 1910-1918 (2019).

Figures

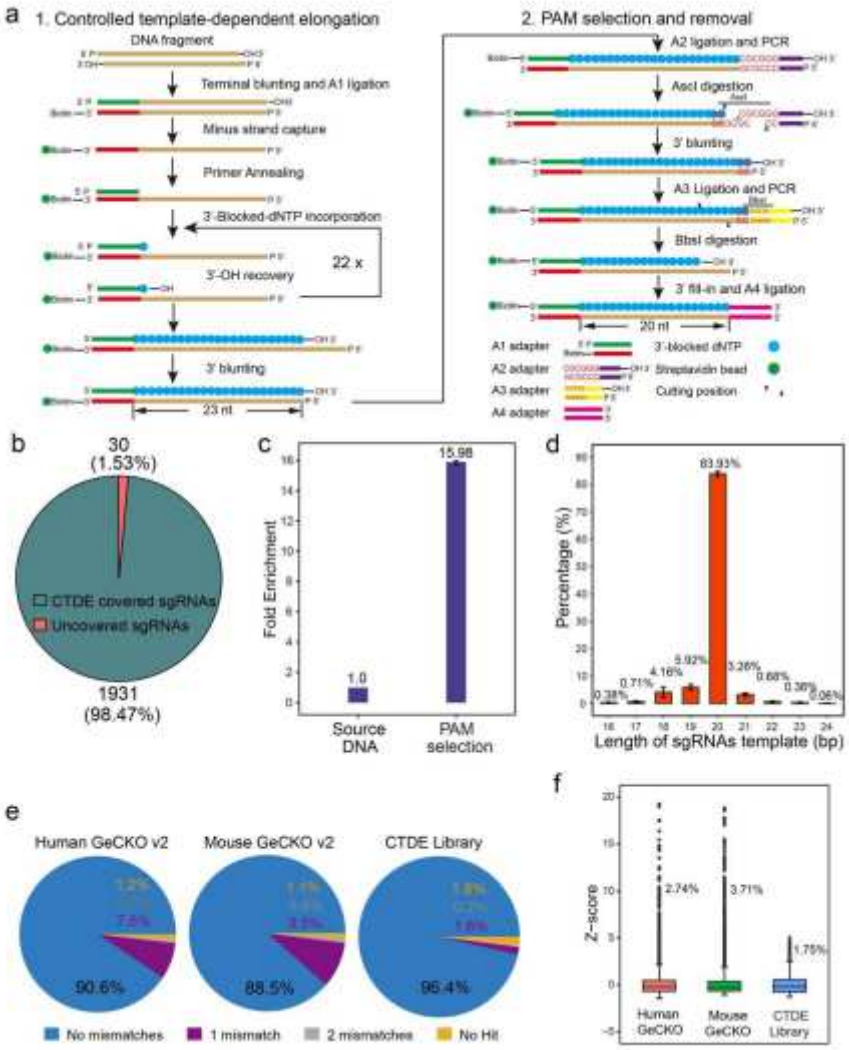

Figure 1

[Please see the manuscript file to view the figure caption.]

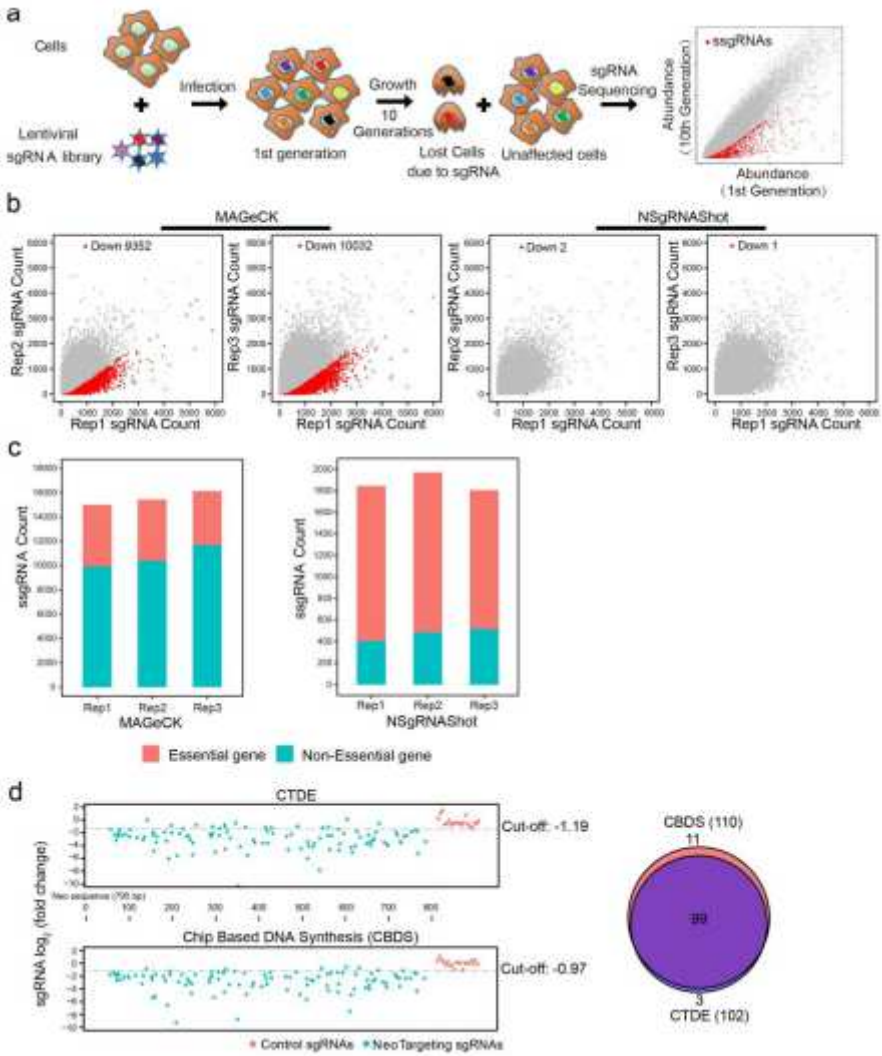

Figure 2

[Please see the manuscript file to view the figure caption.]

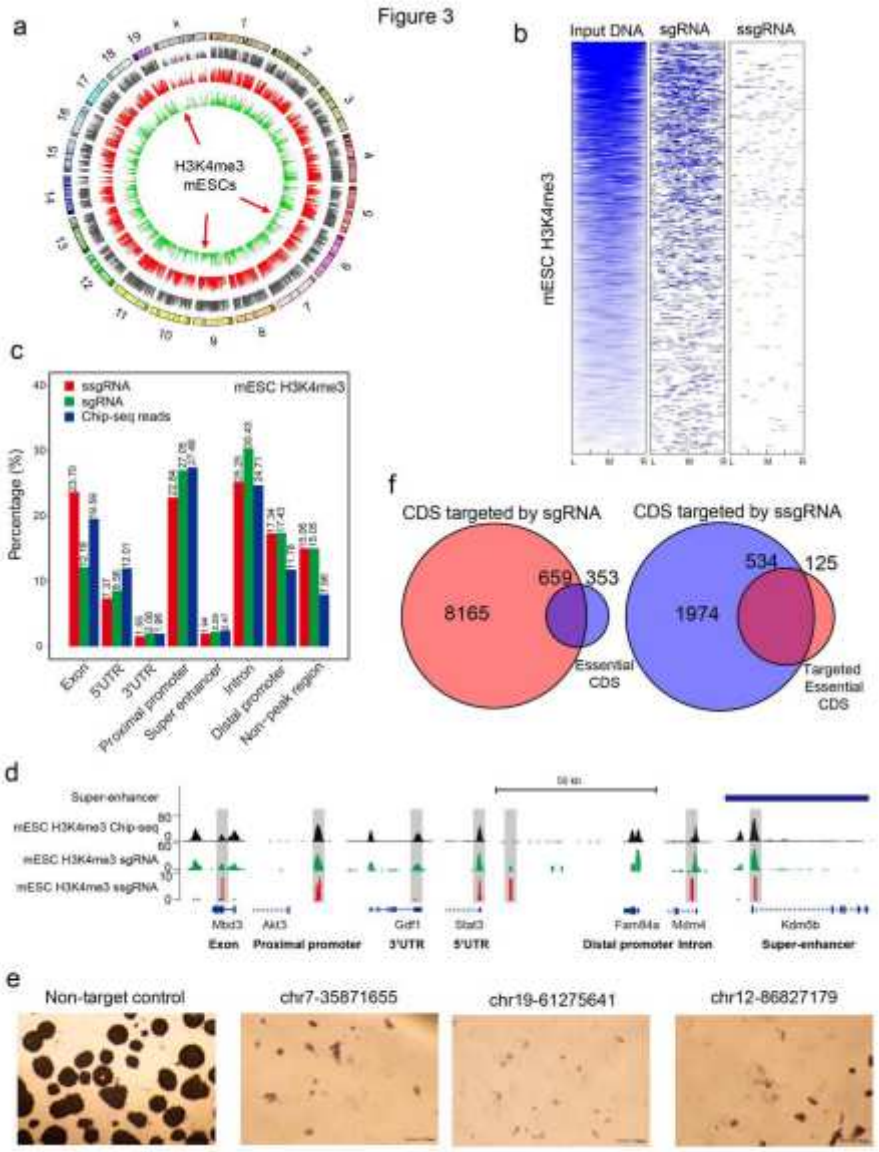

Figure 3

[Please see the manuscript file to view the figure caption.]

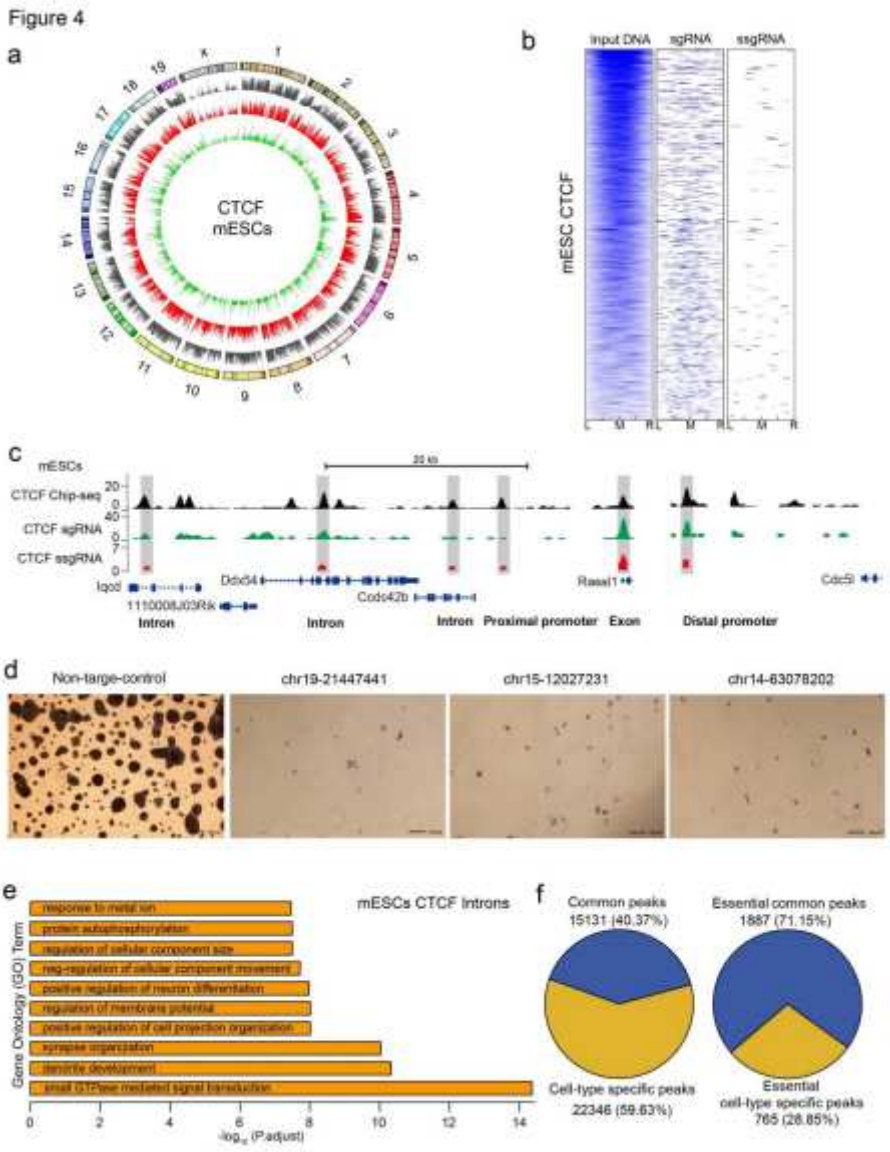

Figure 4

[Please see the manuscript file to view the figure caption.]

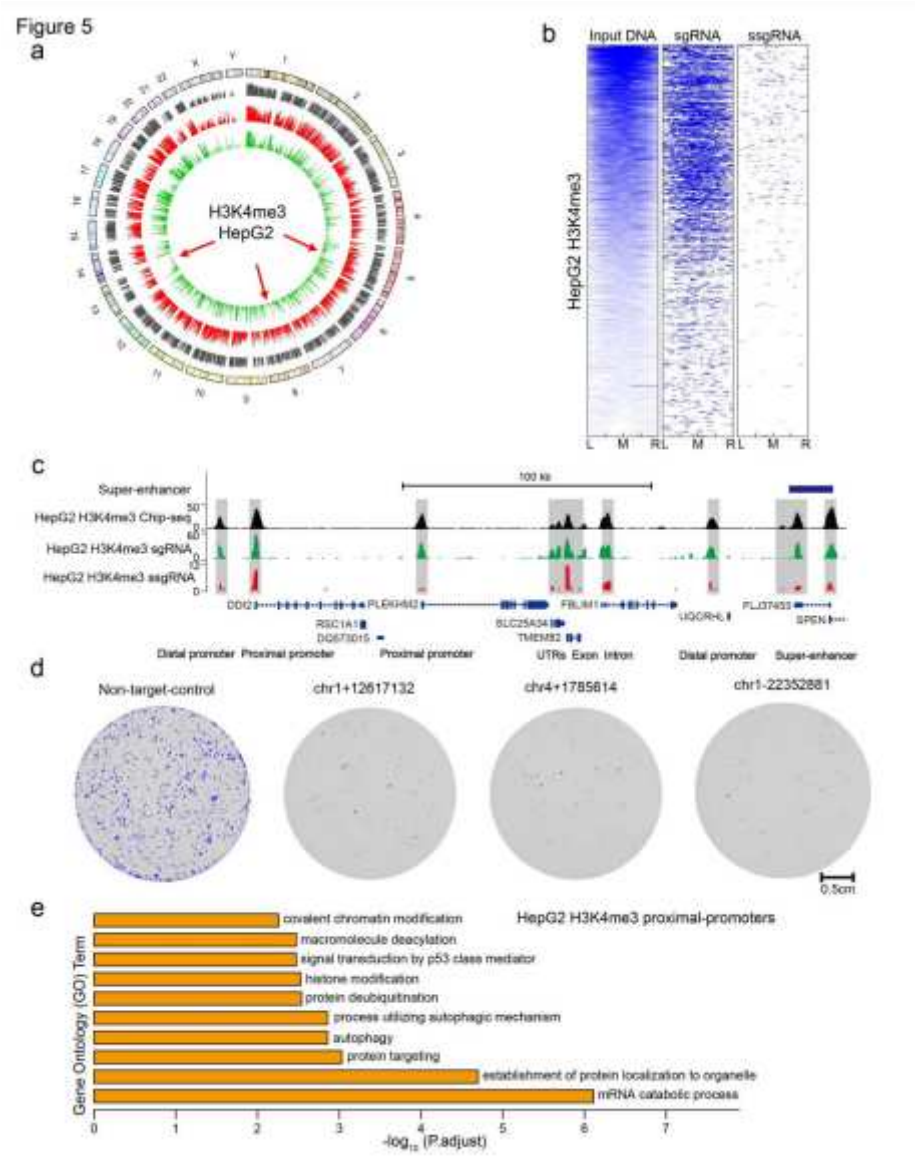

Figure 5

[Please see the manuscript file to view the figure caption.]

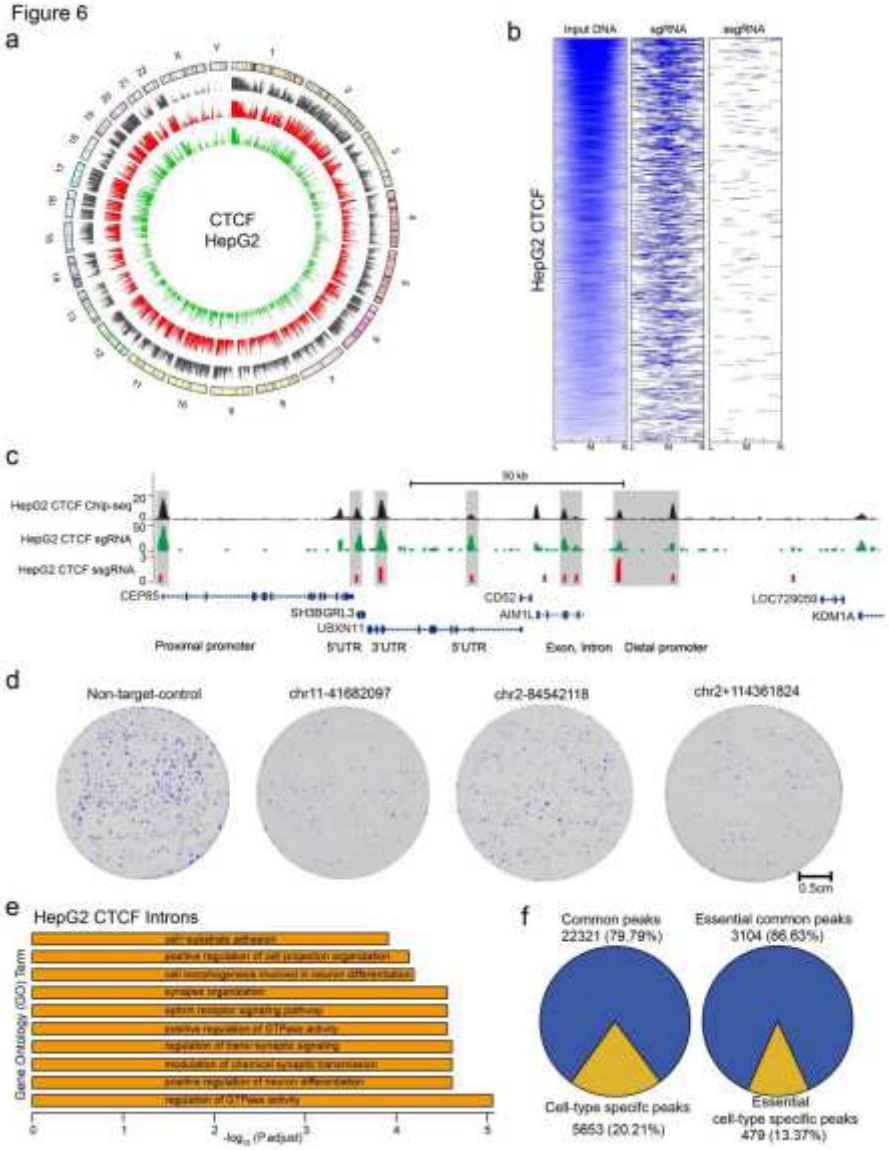

Figure 6

[Please see the manuscript file to view the figure caption.]

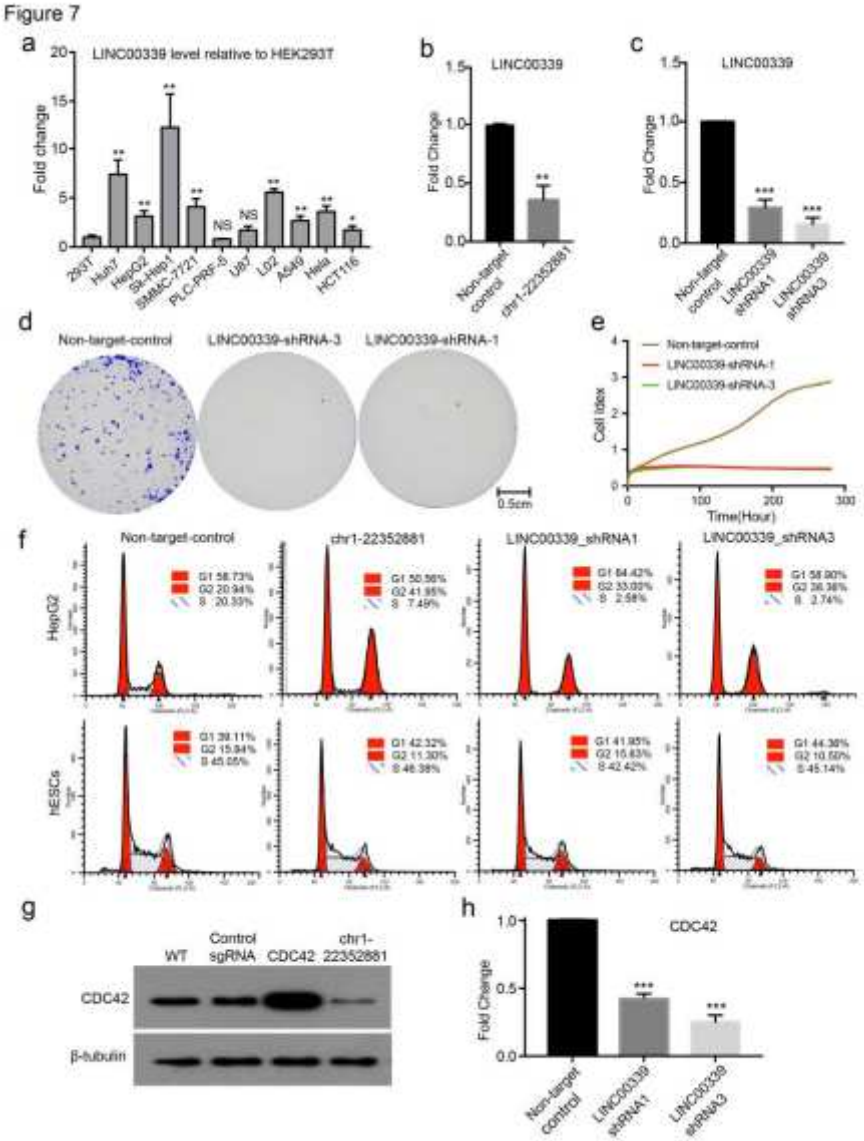

Figure 7

[Please see the manuscript file to view the figure caption.]

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalTable.zip