Computationally defined and *in vitro* validated putative genomic safe harbour loci for transgene expression in human cells

Authors
Matias I. Autio[1, 2*], Efthymios Motakis[2§†], Arnaud Perrin[1, 2§], Talal Bin Amin[3], Zenia Tiang[1,2], Dang Vinh Do[1,2], Jiaxu Wang[4], Joanna Tan[5], Wei Xuan Tan[6,7], Adrian Kee Keong Teo[6,7,8] and Roger S.-Y. Foo[1, 2*]

§ equal contribution
* co-corresponding authors

Affiliations

1. Laboratory of Molecular Epigenomics and Chromatin Organization, Genome Institute of Singapore, A*STAR, 138672, Singapore
2. Cardiovascular Disease Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, 117599, Singapore
3. Laboratory of Systems Biology and Data Analytics, Genome Institute of Singapore, A*STAR, 138672, Singapore
4. Laboratory of RNA Genomics and Structure, Genome Institute of Singapore, A*STAR, 138672, Singapore
5. Center for Genome Diagnostics, Genome Institute of Singapore, A*STAR, 138672, Singapore
6. Stem Cells and Diabetes Laboratory, Institute of Molecular and Cell Biology, A*STAR, 138673, Singapore
7. Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, 119228, Singapore
8. Precision Medicine Translational Research Programme, Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 117596, Singapore

† present address: The Jackson Laboratory for Genomic Medicine, Computational Sciences, Farmington, CT, USA

Abstract
Stable expression of transgenes is essential in both therapeutic and research applications. Traditionally, transgene integration has been accomplished via viral vectors in a semi-random fashion, but with inherent integration site biases linked to the type of virus used[1]. The randomly integrated transgenes may undergo silencing[2,3] and more concerningly, can also lead to dysregulation of endogenous genes. Gene dysregulation can lead to malignant transformation of cells and has unfortunately given rise to cases of leukaemia[4,5] in gene therapy trials. Genomic safe harbour (GSH) loci have been proposed as safe sites for transgene integration [6–8]. To date, a number of sites in the human genome have been used for directed integration; however none of these pass scrutiny as *bona fide* GSH[6]. Here, we conducted a computational analysis to identify 25 putative GSH loci that reside in active chromosomal compartments. We validated stable transgene expression in three GSH sites *in vitro* using human embryonic stem cells (hESCs) and their differentiated progeny. Furthermore, for easy targeted transgene expression, we have engineered constitutive landing pad expression constructs into the three validated GSH in hESCs.

Results
A list of criteria has previously been suggested for a given locus to qualify as putative GSH[6,8]. These criteria state that GSH: must not be in proximity to genes coding or non-coding, with added distance from known oncogenes and miRNAs, and must not disrupt transcriptional units or ultra-conserved genomic regions. To shortlist putative GSH we conducted a computational search of the human genome using publicly available data (Fig. 1 A). We included the previously published safety criteria and added a further filter to exclude any regions of DNaseI hypersensitivity, as these regions are likely enriched in transcription factor binding and regulatory elements[9]. A total of 12,766 sites, ranging from 1 b to approximately 30 Mb, passed the filters used (Fig. 1 A-B, Supp. Table 1). For a universal GSH site to be useful, it needs not only to be safe, but also enable stable expression of a transgene in any tissue type. We filtered the human genome for regions consistently in the active chromatin compartment based on 21 different human cell and tissue types[10]. To extend the analysis beyond the limited set of samples, we utilised RNA-seq data of all available tissue types from the GTEx portal[11]. We selected an empirical set of ubiquitously expressed genes with low variance. We then cross referenced the chromosomal locations of these genes with the consistently active chromatin regions. This analysis yielded 399 1Mb active regions that overlapped with a ubiquitously expressed gene (Fig. 1 A-B, Supp. Table 2). By overlapping the two datasets, we found 49 safe sites within the active regions. We further filtered the 49 sites (Fig. 1 A-B, Supp. Table 3) using BLAT to generate a final shortlist of 25 unique putative universal GSH sites in the human genome (Fig. 1 A-B, Supp. Table 4, Supp. Fig. 1).

To validate our candidate GSH, we selected seven of the 25 sites for *in vitro* experiments (Fig. 1 C). None of the selected seven sites lie at or immediately adjacent to borders of topologically associated domains (TADs) (Supp. Fig. 2). We targeted H1 hESC using CRISPR/Cas9 and a donor landing pad construct (Fig. 2 A) at each of the seven candidate sites (Methods and Supp. Table 5). To exclude potential off-target effects of CRISPR/Cas9 targeting, we used a version of Cas9 with enhanced specificity[12] and effective guides with highest predicted specificity available. Following antibiotic selection, single clones were expanded and screened for successful homology directed repair driven integration of the expression construct with junction- and digital-PCR (Fig. 2 B, Supp. Table 6, and Supp. Fig. 3). Successful heterozygous targeting of the donor construct was confirmed at three candidate GSH sites on chromosomes 1, 18 and 19. No evidence of off-target activity was observed following PCR amplification and Sanger sequencing of the top 5 predicted off-target sites for each

of the targeted clones (Supp. Fig. 4). We named the successfully targeted safe harbours after real world harbours, designating them Pansio-1, Olônne-18, and Keppel-19.

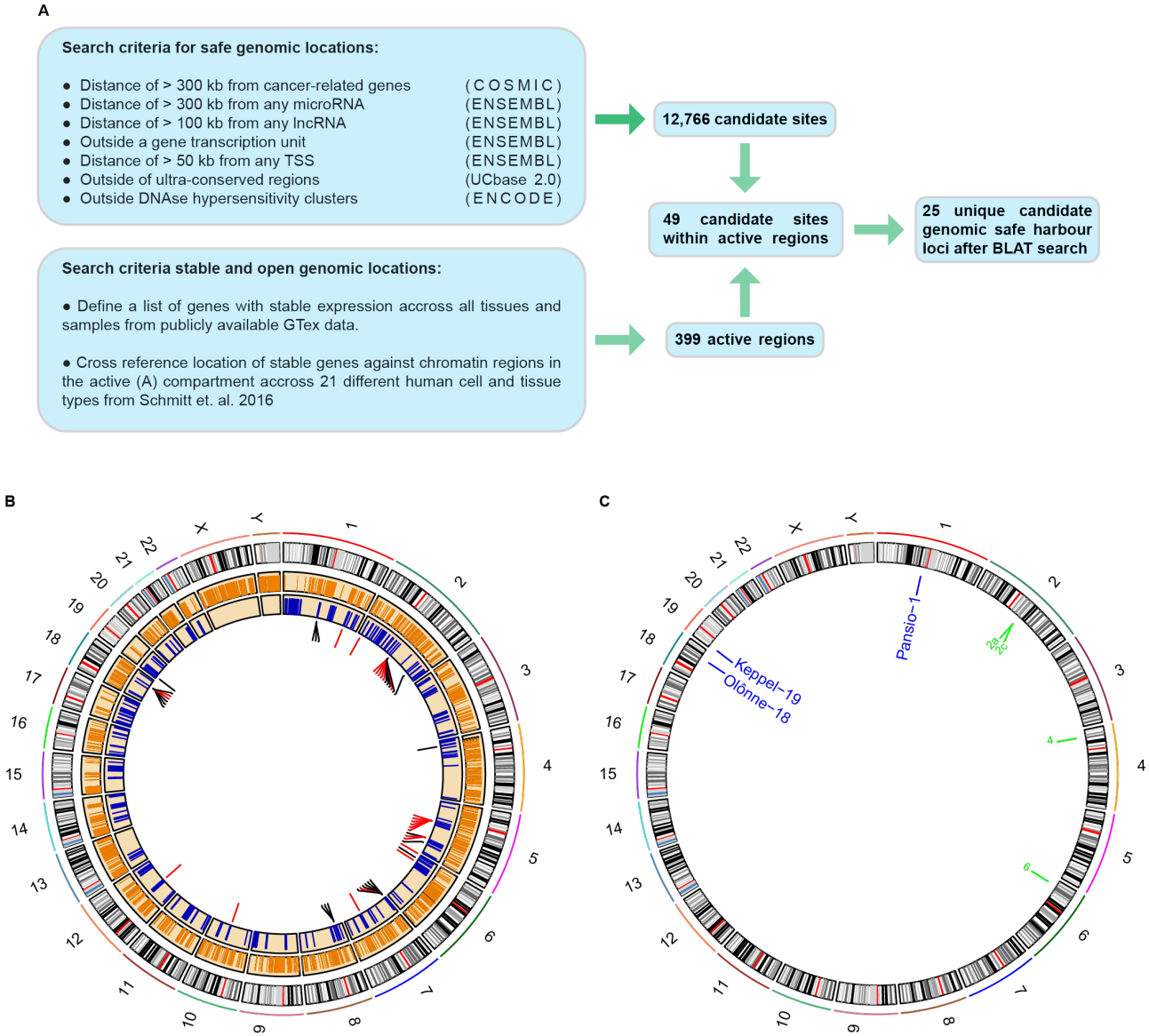


Figure 1. A) Schematic representation of the computational workflow for defining candidate GSH. **B)** CIRCOS plot summarising computational search results. Ring 1: chromosome ideograms; ring 2: orange bars indicating safe sites; ring 3: blue bars indicating active regions; ring 4: candidate sites within active regions, red bars site failed BLAT screening, black bars site passed BLAT screening. **C)** Locations of candidate GSH targeted *in vitro*. Blue labels: targeted clone established; green labels: no clone established.

To investigate the safety of our targeted GSH, we first checked the mRNA expression levels of the nearest genes *MAGI3*, *TXNL1* and *ZNRF4* to Pansio-1, Olônne-18 and Keppel-19 respectively using qPCR. When compared to un-targeted H1 hESCs, the 95% confidence intervals of log2 fold-change (log2-FC) for each gene in the respective GSH line overlapped with zero, indicating that the data do not show evidence of statistically significant change in mRNA expression levels of the nearest genes (Fig. 2 C). We then conducted RNA-seq analysis to look for gene expression changes on a global scale. Our GSH targeted clones yielded very low numbers of differentially expressed (DE) genes; 30, 29 and 40 respectively for Pansio-1, Olônne-18, and Keppel-19 (Fig2 D and Supp. Table 7). Notably *CASP9*, the suicide gene included in our targeting construct, was the gene with lowest false discovery rate (FDR) in Pansio-1 and Olônne-18 and second lowest in Keppel-19 (Fig. 2 D and Supp. Table 7). A high proportion of the DE genes found in our GSH targeted lines were shared with control

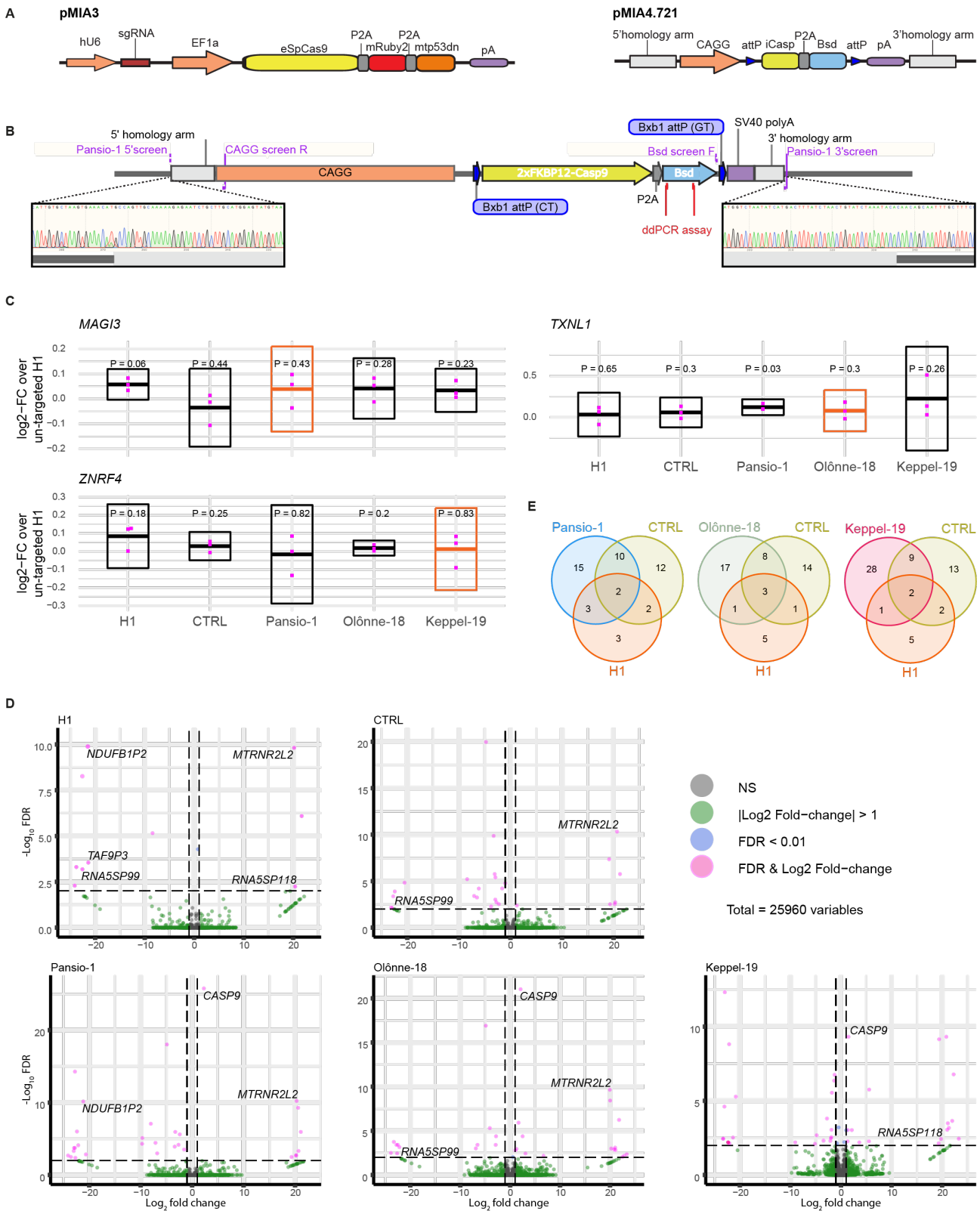


Figure 2. A) Schematic representation of CRISPR/Cas9 plasmid (pMIA3) and homology directed repair donor (pMIA4.721) used for targeting with functional components annotated. **B)** Schematic of integrated landing pad expression construct. Positions of primers for junction-PCR as well as of ddPCR assay are indicated. Representative junction-PCR Sanger sequencing reads from Pansio-1 targeted clones shown in expanded view. **C)** Log2-FC of mRNA expression levels against un-targeted H1 hESC samples for the nearest genes of Pansio-1, Olônne-18, and Keppel-19 candidate GSH. Evaluated samples: H1 = un-targeted hESC, CTRL = H1 cells non-GSH targeted, Pansio-1 = landing pad construct integrated to Pansio-1 GSH in H1 hESC, Olônne-18 = landing pad construct integrated to Olônne-18 GSH in H1 hESC and Keppel-19 = landing pad construct

integrated to Keppel-19 GSH in H1 hESC. Box plots representing 95% confidence intervals of mean log2-FC. Nearest gene for each GSH indicated in orange. Individual data points shown in pink with P-value for each comparison shown above. **D)** Volcano plots of RNA-seq analysis against un-targeted H1 hESC. Samples analysed as in **C)**. Differentially expressed (DE) genes with FDR ≤ 0.01 and |logFC| ≥ 1 in pink, genes with |logFC| ≥ 1 in green, genes with FDR ≤ 0.01 in blue, others in grey. **E)** Venn-diagrams illustrating the overlap of DE genes between un-targeted H1 hESC, non-GSH targeted H1 hESC and the three GSH targeted H1 hESC lines.

targeted H1 cells and/or untargeted wild-type H1 cells (Fig. 2 E and Supp. Table 7), suggesting that the observed changes were unrelated to the GSH targeting and transgene expression. The closest DE gene found from our analysis lies >42 Mb away for Pansio-1 and >47 Mb for Keppel-19. For Olônne-18 there were no DE genes on the same chromosome (Supp. Table 7). Functional enrichment analysis of the DE genes revealed relatively few terms; 9, 8 and 15 for Pansio-1, Olônne-18, and Keppel-19 respectively (Supp. Table 8). In the case of Pansio-1 all the terms were shared with the control targeted H1 cells except for one, which was due to the increased level of *CASP9* (Supp. Table 8). As a further safety check, we also conducted karyotyping of the three targeted clones and observed no abnormalities (Supp. Fig. 5). Taken together, these results suggest minimal disruption of the native genome and no karyotypic abnormalities following transgene integration to our three GSH.

In addition to safety, a functional GSH also needs to allow for stable expression of a transgene. We took advantage of the landing-pad design of our targeting construct to swap in a sequence coding for Clover-fluorophore, by introducing a plasmid expressing BxbI-integrase as well as a donor construct into the three GSH lines (Fig. 3 A-B). Targeted cells were enriched with fluorescence activated cell sorting. Introduction of the payload transgene did not alter the expression of pluripotency markers OCT3/4 and SOX2 (Fig. 3C-D). We maintained the Clover targeted GSH lines in hESC state over 15 passages and consistently observed >98% Clover positive cells (Fig. 3E). Interestingly, the levels of Clover integrated to Olônne-18 seemed to be consistently lower than integrations to Pansio-1 or Keppel-19 (Fig. 3 C-E). To investigate the stability of our GSH in other cell types, we conducted directed differentiation of our Clover-integrated hESC lines into cell types from the three germ lineages. Clover expression remained consistent in neuronal, liver, and cardiac cells (Fig. 3F-H, Supp. Fig. 6) in Pansio-1, Olônne-18, and Keppel-19 targeted cells.

Overall, we have developed a computational pipeline to define GSH candidate sites from the human genome that fulfil criteria for safety as well as accessibility for transgene expression. Our pipeline defines 25 unique candidate GSH and we conducted *in vitro* validation experiments for three of them, Pansio-1, Olônne-18, and Keppel-19. Targeting and transgene expression in hESC at the three sites led to minimal or no change in the expression levels of the nearest native genes or the transcriptome overall and did not interfere with directed differentiation to the three germ lineages. Furthermore, we established landing pad expression lines in H1 hESC of Pansio-1, Olônne-18, and Keppel-19, which we hope will serve as useful research tools.

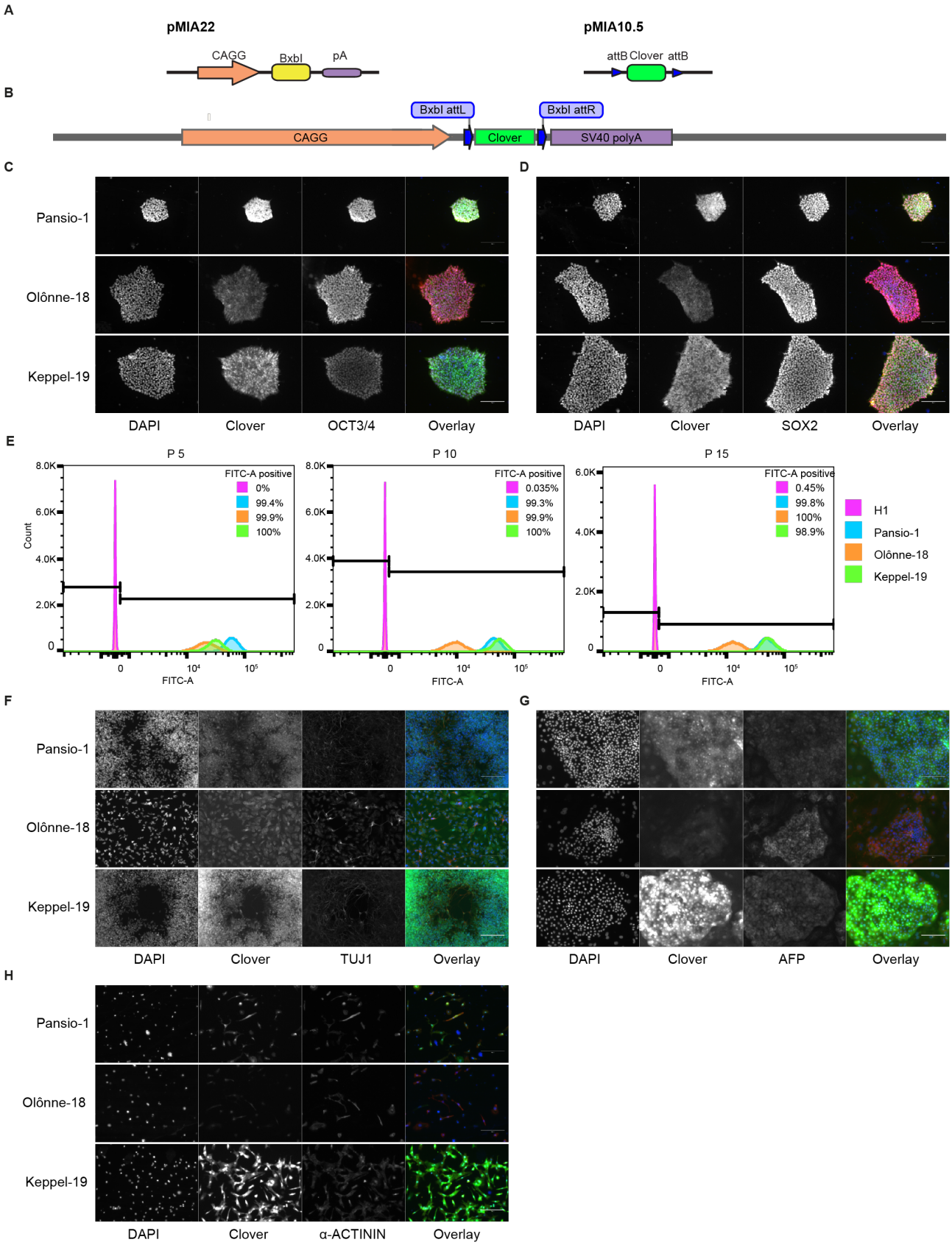


Figure 3. A) Schematic representation of integrase expression construct (pMIA22) and transposon donor construct (pMIA10.5). **B)** Schematic of landing pad construct with integrated Clover transgene. **C)** Representative immunofluorescence images of Clover-integrated GSH H1 cells. DAPI = nuclear staining with 4′,6-diamidino-2-phenylindole, Clover = fluorescence from Clover transgene, OCT3/4 = antibody staining against OCT3/4, Overlay = overlay of the three imaged channels. **D)** As in **C)** apart from antibody staining against SOX2. **E)** Histograms of flow cytometry analysis for FITC-A channel of un-targeted H1 hESC, and the

three GSH targeted hESC lines over 15 passages. Percentages of FITC-A positive cells according to the indicated gating. **F)** Representative immunofluorescence images of Clover-integrated GSH H1 cells differentiated to neuronal-like cells. Channels imaged as in **C)** apart from antibody staining against Tuj1. **G)** As in **F)** for cells differentiated to hepatocyte-like cells, antibody staining against AFP. **H)** As in **F)** for cells differentiated to cardiomyocyte-like cells, antibody staining against sarcomeric α-ACTININ. Scale bars for all immunofluorescence images equal to 150 µm.

A GSH site is an ideal location for transgene integration. To qualify as a GSH, a locus should be able to host transgenes enabling their stable expression as well as not interfere with the native genome[6]. A number of previous studies have reported discovery and usage of a handful of integration sites[7,13–16], which fulfil a subset of criteria previously suggested for GSH[6,8]. However, in contrast to the candidate sites we present (Fig. 1 A-B), the previously reported sites do not utilise criteria to avoid potential regulatory elements or criteria for universally stable and active genomic regions[7,13–17]. We utilised directed differentiation of our Pansio-1, Olônne-18 and Keppel-19 targeted hESC to show consistent expression in hESC and cells from all three germ lineages (Fig. 3 C-H), whereas majority of the previously reported sites remain studied in only a limited number of cell types[14–17].The integration sites that have been most heavily utilised in research, namely AAVS1, CCR5 and Rosa26 orthologous site[18–21], do not meet the criteria set out for GSH[6,8]. These sites reside in highly gene rich regions and in the case of AAVS1 actually within a gene transcription unit, furthermore all of these loci have known oncogenes in their proximity (<300kb)[8]. Thus, their utilisation in a clinical setting would require extensive further safety data. Furthermore variable transgene expression and silencing has been reported for AAVS1 in hepatocytes[22] and cardiomyocytes[23]. The three candidate sites we tested, Pansio-1, Olônne-18, and Keppel-19, demonstrated stable expression of the Clover transgene in both hepatocyte- and cardiomyocyte-like cells after differentiation from hESC (Fig. 3 G-H). The transcriptome analysis of our GSH targeted hESC lines showed very low number of differentially expressed genes (Fig. 2 D) when compared to untargeted H1 hESC. Furthermore, many of the DE genes we observed were shared with independent wild type H1 hESC as well as control targeted H1 hESC (Fig. 2 E). The nearest observed DE genes to each targeted candidate GSH were located >42 Mb away from the targeted site, which is far beyond the distance generally suggested for enhancer-promoter interactions and TADs[24].

Our data suggest that the three candidate GSH, Pansio-1, Olônne-18, and Keppel-19, are able to support stable transgene expression in different cell types, and integration at these sites shows minimal perturbation of the native genome. As hESC currently offer the best available model to test the stability of transgene expression from a GSH in the human genome, we generated landing-pad H1 hESC lines for all three candidate sites. These cell lines will allow easy integration of various transgenes to the candidate GSH for research applications. Advances in genome engineering technologies using e.g. novel integrases or CRISPR RNA-guided transposons[25–28] will likely lead to easier integration of more complex transgene constructs. We hope that our list of 25 candidate GSH and the three *in vitro* validated sites will serve as a resource for research applications in many different cell types. After further experimentation and reproducibility validations, translation to clinical applications can be envisaged.

References

1. Mitchell, R. S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2**, (2004).
2. Ellis, J. Silencing and Variegation of Gammaretrovirus and Lentivirus Vectors. *Hum. Gene Ther.* **16**, 1241–1246 (2005).
3. Mok, H. P., Javed, S. & Lever, A. Stable gene expression occurs from a minority of integrated HIV-1-based vectors: transcriptional silencing is present in the majority. *Gene Ther.* **14**, 741–751 (2007).
4. Howe, S. J. *et al.* Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *J. Clin. Invest.* **118**, 3143–3150 (2008).
5. Hacein-Bey-Abina, S. *et al.* LMO2-Associated Clonal T Cell Proliferation in Two Patients after Gene Therapy for SCID-X1. *Science (80-.).* **302**, 415–419 (2003).
6. Papapetrou, E. P. & Schambach, A. Gene insertion into genomic safe harbors for human gene therapy. *Mol. Ther.* **24**, 678–684 (2016).
7. Papapetrou, E. P. *et al.* Genomic safe harbors permit high β-globin transgene expression in thalassemia induced pluripotent stem cells. *Nat. Biotechnol.* **29**, 73–78 (2011).
8. Sadelain, M., Papapetrou, E. P. & Bushman, F. D. Safe harbours for the integration of new DNA in the human genome. *Nat. Rev. Cancer* **12**, 51–8 (2012).
9. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
10. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.* **17**, 2042–2059 (2016).
11. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv. Biobank.* **13**, 311–319 (2015).
12. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science (80-.).* **351**, 84–88 (2016).
13. Costa, M. *et al.* The hESC line Envy expresses high levels of GFP in all differentiated progeny. *Nat. Methods* **2**, 259–260 (2005).
14. Pellenz, S. *et al.* New Human Chromosomal Sites with “Safe Harbor” Potential for Targeted Transgene Insertion. *Hum. Gene Ther.* **X**, hum.2018.169 (2019).
15. Rodriguez-Fornes, F. *et al.* Targeted gene therapy into a safe harbor site in human hematopoietic progenitor cells. *Gene Ther.* (2020) doi:10.1038/s41434-020-0144-x.
16. Eyquem, J., Poirot, L., Galetto, R., Scharenberg, A. M. & Smith, J. Characterization of three loci for homologous gene targeting and transgene expression. *Biotechnol. Bioeng.* **110**, 2225–2235 (2013).
17. Aznauryan, E. *et al.* Discovery and validation of novel human genomic safe harbor sites for gene and cell therapies. *bioRxiv* 2021.03.04.433856 (2021) doi:10.1101/2021.03.04.433856.
18. Irion, S. *et al.* Identification and targeting of the ROSA26 locus in human embryonic stem cells. *Nat. Biotechnol.* **25**, 1477–82 (2007).
19. Liu, R. *et al.* Homozygous Defect in HIV-1 Coreceptor Accounts for Resistance of Some Multiply-Exposed Individuals to HIV-1 Infection. *Cell* **86**, 367–377 (1996).
20. Kotin, R. M., Linden, R. M. & Berns, K. I. Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination. *EMBO J.* **11**, 5071–5078 (1992).
21. Perez, E. E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using

zinc-finger nucleases. *Nat. Biotechnol.* **26**, 808–816 (2008).
22. Ordovás, L. *et al.* Efficient recombinase-mediated cassette exchange in hPSCs to study the hepatocyte lineage reveals AAVS1 locus-mediated transgene inhibition. *Stem Cell Reports* **5**, 918–931 (2015).
23. Bhagwan, J. R. *et al.* Variable expression and silencing of CRISPR-Cas9 targeted transgenes identifies the AAVS1 locus as not an entirely safe harbour. *F1000Research* **8**, 1911 (2019).
24. Jerkovic´, I. & Cavalli, G. Understanding 3D genome organization by multidisciplinary methods. *Nat. Rev. Mol. Cell Biol.* **22**, 511–528 (2021).
25. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* (2019) doi:10.1038/s41586-019-1323-z.
26. Ioannidi, E. I. *et al.* Drag-and-drop genome insertion without DNA cleavage with CRISPR-directed integrases. *bioRxiv* (2021) doi:10.1101/2021.11.01.466786.
27. Anzalone, A. V *et al.* Programmable large DNA deletion, replacement, integration, and inversion with twin prime editing and site-specific recombinases. *bioRxiv* (2021) doi:10.1101/2021.11.01.466790.
28. Durrant, M. G. *et al.* Large-scale discovery of recombinases for integrating DNA into the human genome. *bioRxiv* (2021) doi:10.1101/2021.11.05.467528.

A

Search criteria for safe genomic locations:

- Distance of > 300 kb from cancer-related genes (COSMIC)
- Distance of > 300 kb from any microRNA (ENSEMBL)
- Distance of > 100 kb from any lncRNA (ENSEMBL)
- Outside a gene transcription unit (ENSEMBL)
- Distance of > 50 kb from any TSS (ENSEMBL)
- Outside of ultra-conserved regions (UCbase 2.0)
- Outside DNAse hypersensitivity clusters (ENCODE)

Search criteria stable and open genomic locations:

- Define a list of genes with stable expression accross all tissues and samples from publicly available GTex data.

- Cross reference location of stable genes against chromatin regions in the active (A) compartment accross 21 different human cell and tissue types from Schmitt et. al. 2016

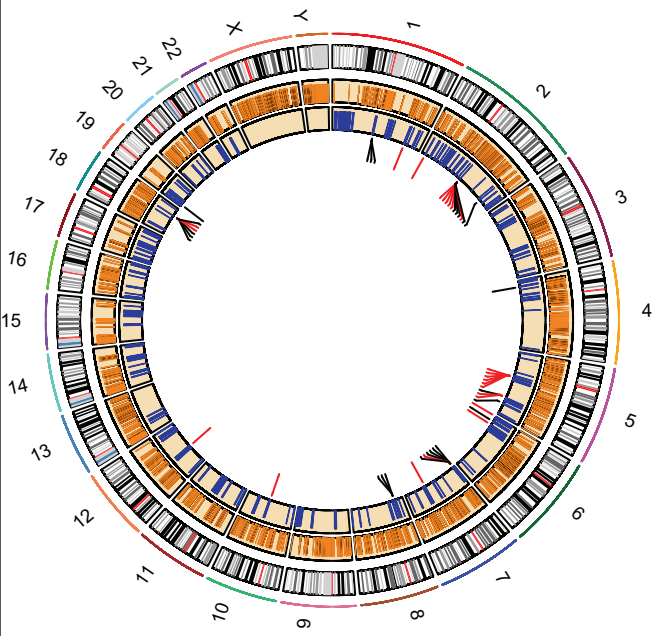
12,766 candidate sites

49 candidate sites within active regions

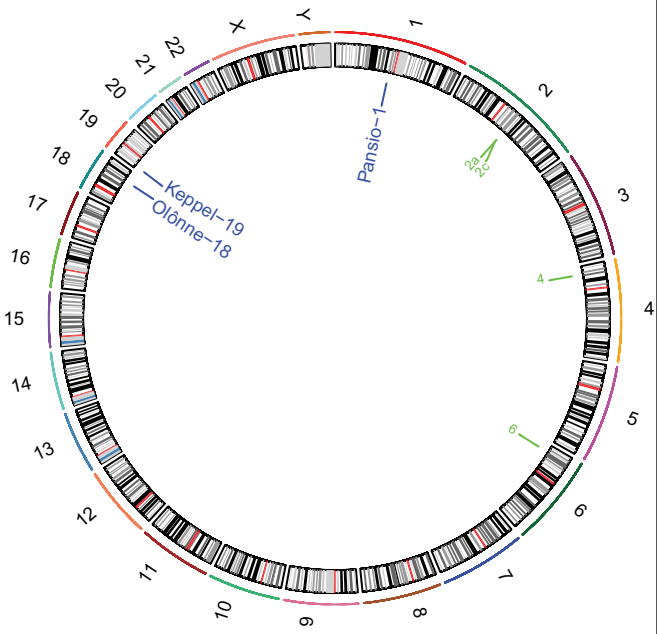
25 unique candidate genomic safe harbour loci after BLAT search

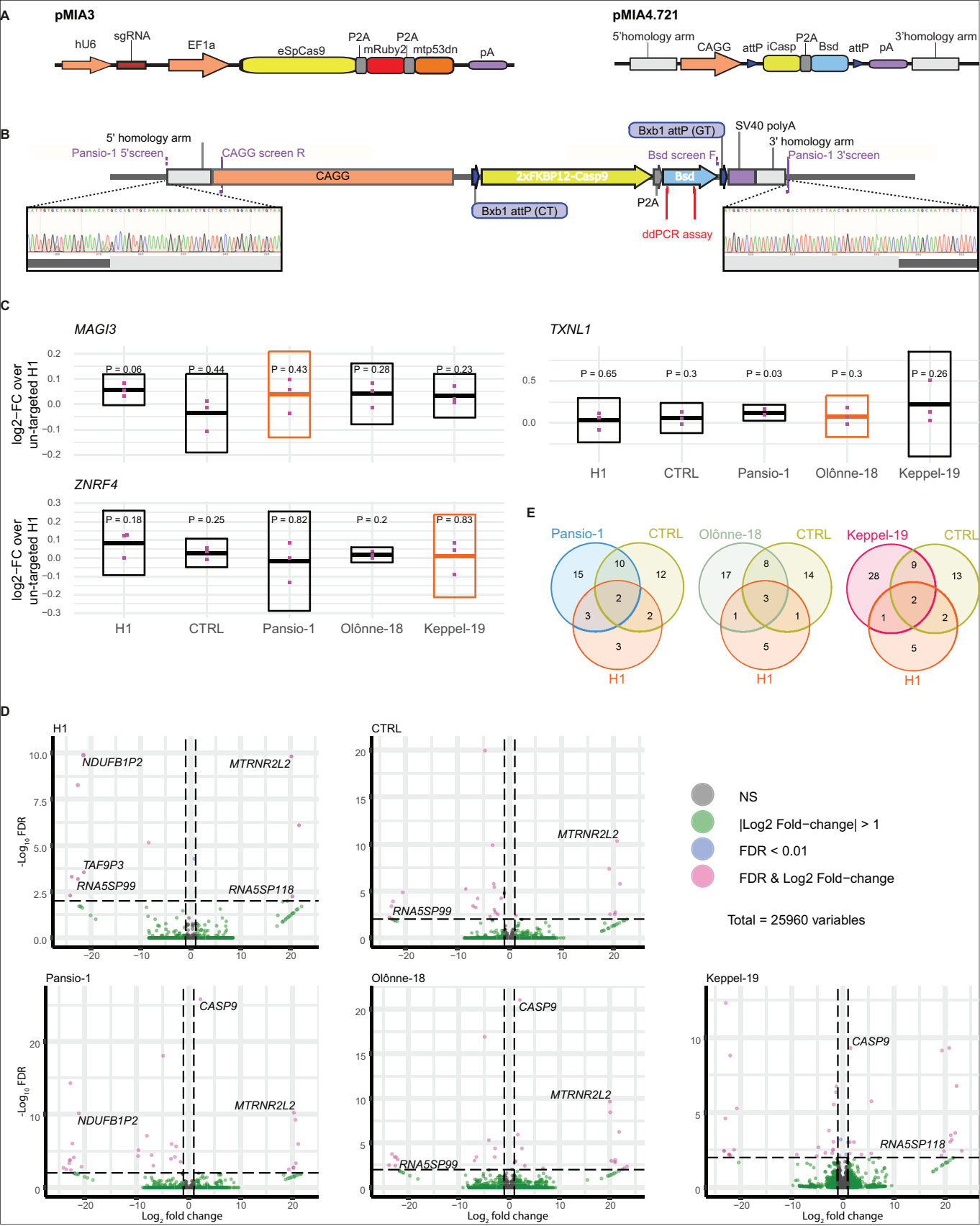
399 active regions

B

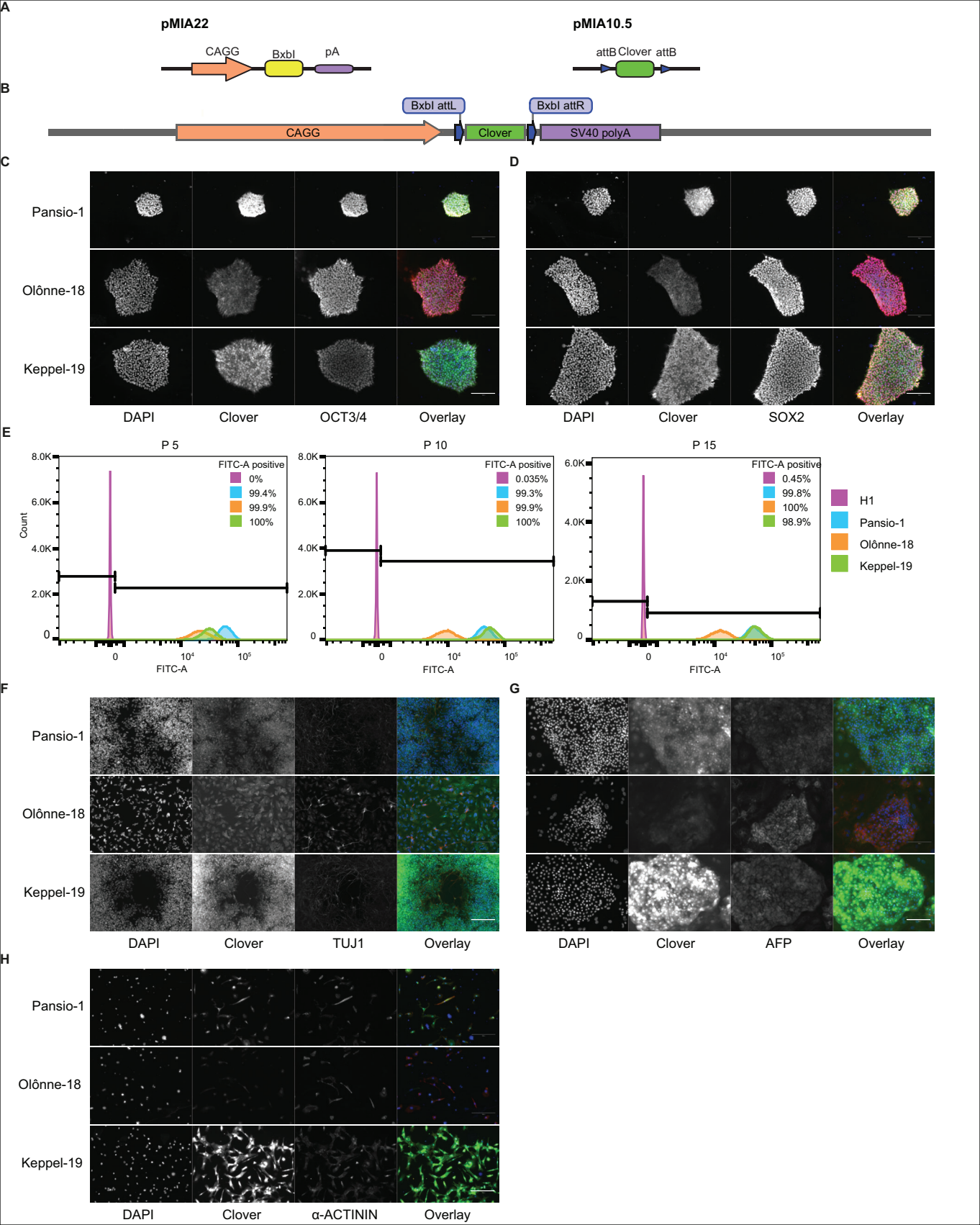


C





A
pMIA3
hU6
sgRNA
EF1a
eSpCas9
P2A
mRuby2
P2A
mtp53dn
pA
pMIA4.721
5'homology arm
CAGG
attP
iCasp
P2A
Bsd
attP
pA
3'homology arm
B
5' homology arm
Pansio-1 5'screen
CAGG screen R
CAGG
2xFKBP12-Casp9
Bxb1 attP (CT)
Bxb1 attP (GT)
Bsd screen F
Bsd
P2A
ddPCR assay
SV40 polyA
3' homology arm
Pansio-1 3'screen
C
MAGI3
log2−FC over un-targeted H1
P = 0.06
P = 0.44
P = 0.43
P = 0.28
P = 0.23
TXNL1
P = 0.65
P = 0.3
P = 0.03
P = 0.3
P = 0.26
H1
CTRL
Pansio-1
Olônne-18
Keppel-19
ZNRF4
P = 0.18
P = 0.25
P = 0.82
P = 0.2
P = 0.83
E
Pansio-1
CTRL
Olônne-18
CTRL
Keppel-19
CTRL
H1
D
H1
CTRL
Pansio-1
Olônne-18
Keppel-19
-Log10 FDR
Log2 fold change
NDUFB1P2
MTRNR2L2
TAF9P3
RNA5SP99
RNA5SP118
CASP9
NS
|Log2 Fold−change| > 1
FDR < 0.01
FDR & Log2 Fold−change
Total = 25960 variables


A
pMIA22
CAGG
BxbI
pA
pMIA10.5
attB Clover attB
B
BxbI attL
BxbI attR
CAGG
Clover
SV40 polyA
C
Pansio-1
Olônne-18
Keppel-19
DAPI
Clover
OCT3/4
Overlay
D
DAPI
Clover
SOX2
Overlay
E
P 5
P 10
P 15
Count
FITC-A
FITC-A positive
0%
99.4%
99.9%
100%
0.035%
99.3%
99.9%
100%
0.45%
99.8%
100%
98.9%
H1
Pansio-1
Olônne-18
Keppel-19
F
Pansio-1
Olônne-18
Keppel-19
DAPI
Clover
TUJ1
Overlay
G
DAPI
Clover
AFP
Overlay
H
Pansio-1
Olônne-18
Keppel-19
DAPI
Clover
α-ACTININ
Overlay

Methods

Short-listing of putative safe harbour genomic regions

We applied a series of computational, whole-genome loci filtering criteria to pick a narrow list of high-confidence, putative safe harbour sites for experimental validation. In the first step, we selected genomic regions that satisfy simultaneously all of the following criteria: the loci should be located outside of ultra-conserved regions[1,2] (coordinates lifted over from hg19 to hg38 assembly), outside of DNase clusters +/- 2kb (ENCFF503GCK, ENCODE database https://www.encodeproject.org/[3]), more than 50 kb away from any transcription start site and outside a gene transcription unit (ENSEMBL Release 103, dataset *hsapiens_gene_ensembl*, http://www.ensembl.org/index.html), more than 300 kb away from cancer-related genes (Cancer Gene Census, GRCh38, COSMIC v92 database https://cancer.sanger.ac.uk/census), more than 300 kb away from any miRNA (ENSEMBL Release 103, dataset *hsapiens_gene_ensembl*, gene_biotype = miRNA, http://www.ensembl.org/index.html) and more than 100 kb away from any long non-coding RNA (ENSEMBL Release 103, dataset *hsapiens_gene_ensembl*, gene_biotype = lncRNA, http://www.ensembl.org/index.html). From the filtered loci we discarded the loci with high BLAT similarity to other sequences. On the RNA-seq level, we required the putative safe harbour sites to be associated with ubiquitously expressed, low variance genes. On the 3D chromosome organization level, they should belong to regions consistently located in active chromosomal compartments across multiple tissue types.

Ubiquitously expressed and low-variance genes

We downloaded the median gene-level TPMs by tissue type from GTEx (https://www.gtexportal.org/home/datasets) and identified an empirical set of low-variance housekeeping genes. To this extent, we estimated the mean and the variance of each gene across all available tissue types and, independently, selected the genes with the lowest, insignificant variability using the HVG function of scran R package that decomposes the total variance of each gene into its biological and technical components. We picked the genes whose expression levels do not change significantly across the tissue types ($FDR > 0.9$) and fit a mean vs variance non-parametric lowess regression model. We selected the genes with mean TPM ≥ 5 and variance below the average (smoothed) variance estimated from the lowess model.

Loci interaction via chromatin conformation capture

Using a set of publicly available Hi-C chromatin organization data[4] from human cell and tissue types, we shortlisted the genomic regions consistently located (at least 20/21 interrogated tissue types) in active (open chromatin) compartments[5].

BLAT analysis

We measured the uniqueness of target loci using BLAT (https://genome.ucsc.edu/cgi-bin/hgBlat) on the human genome GRCh38 with BLAT'S guess query type. BLAT takes the target DNA sequence as input and identifies similar ones in the whole human genome. Target sequences of more than 25,000 bps (BLAT's limit) were split into multiple smaller overlapping segments of length between 9,000 to 11,000 bps each (depending on the original target length) and tested separately. We reported the location and length of the matching sequences, their similarity to the target (BLAT score) and the number of the matching base

pairs. We filtered out the target loci with at least one matching sequence of more than 50% similarity.

TAD boundary check

We checked the locations of our *in vitro* targeted GSH candidates against the TAD borders using data from H1 human embryonic stem cells (hESCs)[6] on the 3D Genome Browser (http://3dgenome.org)[7]. Visual inspection of the candidate loci confirmed that all the candidate GSH are more than 80,000 bp away from TAD borders (Supp. Fig. 2).

Plasmid construction

All restriction enzymes were purchased from NEB. PCR reactions were conducted using Q5® Hot Start High-Fidelity 2X Master Mix (NEB, M0494L). Ligations were conducted using isothermal assembly with NEBuilder® HiFi DNA Assembly Master Mix (NEB, E2621L). All gBlocks, primers & oligos were ordered from Integrated DNA Technologies, Singapore. Plasmids used in this manuscript will be made available via Addgene and their sequences are listed in Supp. Fig. 7. Primers used for fragment amplification are listed in Supp. Table 5.

pMIA4.721: An in-house expression plasmid containing a CAGG promoter (pMIA4.9) was digested with BamHI & SphI. Two gBlocks (bxb-bsd and bxb-sv) were directly ligated into the digested plasmid. The resulting plasmid was digested with PmlI & KpnI. A fragment containing codon optimised iCasp9-2A-Bsd was amplified from a plasmid supplied by Genewiz (sequence of iCasp9 based on Straathof *et.al.* (2005)[8]) and ligated into the digested plasmid. This plasmid was subsequently digested with AgeI & SbfI. The SV40 polyA signal was amplified from pMAX-GFP and ligated to the digested plasmid to generate pMIA4.271.

To generate the HDR donors for each GSH candidate, the pMIA4.721 plasmid was digested with NheI for 5' homology arm and with SbfI for 3' homology arm. Homology arms ranging from 240bp to 769bp were amplified from H1 hESC gDNA. Ligation of homology arms was done in two sequential reactions. Order of ligation depended on the underlying sequence of homology arms for each target. Plasmid for control targeting was built by ligating 5' homology arm to pMIA4.9 digested by NheI and 3' by SbfI.

pMIA22: pMAX-GFP (Lonza) was digested with KpnI and SacI. A gBlock encoding a codon optimised BxbI-integrase[9] with a C-terminal bi-partite nuclear localisation signal[10] was amplified and ligated to the digested backbone.

pMIA10.5-Clover: An empty donor plasmid (pMIA10.5) containing a 5' BxbI attB (CT) and a 3'BxbI attB (GT) was ordered from Genewiz. Clover transgene was amplified and ligated into the plasmid after digestion with AgeI & KpnI.

Stem cell culture

Human ESC line H1 was maintained using mTeSR medium (Stemcell Technologies, 85850) on 1:200 Geltrex (Thermo fisher, A1413202) coated tissue culture plates and passaged regularly as cell aggregates every 4-5 days using ReLeSR (Stemcell Technologies, 05872).

CRISPR/Cas9 mediated targeted construct integration in hESC

H1 hESCs were targeted via nucleofection using an Amaxa-4D (Lonza) as described previously[11]. Briefly, gRNAs were designed using CRISPOR[12] (http://crispor.tefor.net/). Three gRNAs with the highest predicted off-target scores and containing a native G-base in the first position were selected for each GSH candidate (Supp. Table 5). The gRNAs were cloned into pMIA3 plasmid (Addgene #109399) digested with Esp3I and tested via a GFP reconstitution

assay in HEK239T cells. The target locus of each candidate GSH was amplified from H1 hESC gDNA (for primers see Supp. Table 5). The amplified target sequences, ranging from 232-974bp were cloned into the pCAG-EGxxFP plasmid[13] (Addgene # 50716, a kind gift from Dr Masahito Ikawa). The pMIA3 with the tested gRNA and the respective pCAG-EGxxFP target plasmid were transfected into HEK293T using lipofectamine3000 (ThermoFisher Scientific, L3000015), according to manufacturer's recommendations. For each candidate GSH the gRNA with the highest GFP signal at 48h post transfection (data not shown) was selected for use in hESC targeting.

Five micrograms of pMIA3 plasmid containing optimal gRNA for each candidate GSH and the respective pMIA4.721 HDR-donor plasmids were nucleofected into hESC using the P3 Primary Cell kit (V4XP-3024) and programme CA-137. 1.5x10^6 cells were used for each targeting and were plated onto geltrex coated wells on 6-well plates in mTeSR with CloneR (Stemcell Technologies, 05889) following nucleofection. After 24h media was changed to mTeSR, and cells were allowed to recover for another 24-48h. Once cells reached 70-80% confluency, Blasticidin (ThermoFisher Scientific, A1113903) was added to the culture media at 10 µg/ml. Individual colonies were manually picked from the wells after 7-14 days of selection and expanded further for screening.

Junction PCR

Genomic DNA samples for all the collected GSH clones was isolated using PureLink™ Genomic DNA Mini Kit (ThermoFisher Scientific, K182002) according to manufacturer's instructions. PCR reactions amplifying both 5' and 3' targeting HDR junctions as well as the wild type allele were set up using primers listed in Supp. Table 5. Samples were checked for the correct amplification size and alignment of the Sanger sequencing reads for each junction PCR and wild type allele.

Off-target analysis

The top five predicted off-targets were checked via PCR amplification and Sanger sequencing. PCR primers for respective off-targets for each gRNA are listed in Supp. Table 5. Sanger sequencing traces covering the off-target site for wildtype and the respective targeted clone are shown in Supp. Fig. 4.

Copy number analysis

We evaluated Blasticidin and RPP30 Copy Numbers using Droplet Digital™ Polymerase Chain Reaction (ddPCR) technology (Bio-Rad Technologies), according to manufacturer's specifications. Briefly and following fluorescence-based quantification (ThermoFisher Scientific, Qubit™), 2.5 ng double-stranded DNA was added to a reaction mix containing target-specific primers/probe mixes (900 nM primer/250 nM probe per FAM and HEX fluorophore; Bio-Rad, 10042958 Unique Assay ID: dCNS626289650 and 10031243 Unique Assay ID: dHsaCP2500350), 0.05 U HaeIII Restriction Enzyme (New England Biolabs, R0108S) and ddPCR-specific Supermix for Probes (no dUTP) (Bio-Rad, 1863024). This was randomly partitioned into at least 10,000 discrete oil droplets per reaction using microfluidics within the QX200™ Droplet Generator (Bio-Rad, 1864002; together with Droplet Generation Oil for Probes, 1863005), which were gently transferred using a multi-channel pipette into a semi-skirted 96-well plate before heat-sealing (Bio-Rad PX1™ PCR Heat Sealer, 1814000). Target amplification within each droplet was conducted in the C1000 Touch™ Thermal Cycler with

96-Deep Well Reaction Module (Bio-Rad, 1851197) through the following PCR protocol: 1) Enzyme Activation at 95°C for 10 minutes, 2) 40 cycles of Denaturation and annealing/extension at 94°C for 30 seconds and 55°C for 1 minute, respectively, 3) Enzyme Deactivation at 98°C for 10 minutes. The QX200 Droplet Reader (Bio-Rad, 1864003) then derived the number of target-containing droplets through assessing each droplet for elevated, target-specific fluorescence. Blasticidin (FAM)-positive droplet counts were normalised using its respective well-specific RPP30 (HEX)-positive counts prior to downstream analysis. All experiments were done in duplicates, with data visualised and assessed using the QuantaSoft software version 1.7.4.917 (Bio-Rad).

qPCR analysis

RNA was extracted from three biological replicates of the GSH targeted H1 hESC Pansio-1, Olônne-18 and Keppel-19, control targeted cells (cells that underwent CRISPR/Cas9-mediated targeting and HDR of an expression cassette at a non-GSH locus) and two independent cultures of untargeted cells using Direct-zol™ RNA Miniprep kit (Zymo Research, ZYR.R2052). 1µg of RNA was converted into cDNA with Superscript IV Vilo MM (ThermoFisher Scientific, 11766050). Quantitative PCR reactions using TaqMan gene expression assays and master mix were used to compare the expression levels of *MAGI3*, *TXNL1* and *ZNRF4* against reference genes *18S* and *GAPDH* (ThermoFisher Scientific, Hs00326365_m1; Hs00169455_m1; Hs00741333_s1; HS99999901_S1; Hs03929097_g1 and 4444557). Quantitative RT-PCR analysis was done as described previously[14] and resulting log2 fold change gene expression data was compared against reference H1 untargeted samples.

RNA-seq library prep

RNA samples described above for qPCR were also used for RNA-seq. RNA concentration and quality were checked with an Agilent 2100 RNA Pico Chip (Agilent, 5067-1513). RNA sequencing libraries were prepared using the TruSeq Stranded Total RNA Sample Prep Kit (Illumina, 20020596) including Ribo-Zero to remove abundant cytoplasmic rRNA. The remaining intact RNA was fragmented, followed by first- and second-strand cDNA synthesis using random hexamer primers. "End-repaired" fragments were ligated with a unique illumina adapters. All samples were multiplexed and pooled into a single library. Sequencing was done on a HiSeq 4000 to a minimum depth of 50 million 150 bp paired-end reads per biological sample.

RNA-seq quality Control

In all experiments, the raw paired-end reads in fastq format were initially processed with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for quality control at the base and sequence level. To remove the PCR duplicates we utilised the FastUniq algorithm[15]. The adaptor trimming was performed by Trimmomatic (version 0.39)[16]. We quantified the 229,649 annotated human transcripts of GENCODE v35 by Kallisto (version 0.46)[17] followed by conversion of transcript to raw and TPM-normalized gene counts by the tximport package in R[18]. In total, 40,198 genes were quantified. Subsequently, we performed QC at the raw gene counts, checking for bad quality samples having less than 100,000 reads or more than 10% reads mapped to mitochondrial RNA or less than 2,000 detected genes. All samples of the various experiments were of high quality and were retained for the main analysis.

RNA-seq Differential Expression Analysis

The differential expression analysis was conducted by DEseq2[19] evaluating all pairwise comparisons of the conditions of each experiment. In each comparison, we considered only the expressed genes, i.e. those with non-zero raw counts in at least one sample. The differentially expressed genes were those with |logFC| ≥ 1 and FDR ≤ 0.01.

Functional Enrichment Analysis

Functional enrichment analysis was performed on the differentially expressed genes using the g:GOSt R package for g:Profiler (version e104_eg51_p15_3922dba) with g:SCS multiple testing correction method applying significance threshold of 0.05[20].

Karyotyping

For each cell line, 20 GTL-banded metaphases were counted, of which a minimum of four have been analysed and karyotyped.

Flow cytometry

H1 wild type hESCs and Pansio-1, Olônne-18, and Keppel-19 H1 lines carrying Clover-transgene were disassociated with accutase (Stemcell Technologies, 07922) and resuspended in PBS. The single cells in PBS were analysed with a BD LSR Fortessa x-20 FACS Analyzer and FlowJo (v10.6.1).

hESC cardiac differentiation

Two days prior to starting differentiation, cells were dissociated using Accutase and seeded as single cells in Geltrex-coated 12-well plates at seeding density between 1 to 1.5x10^6 cells. Cardiac differentiation was performed following the published protocol by Lian et. al.[21], with modifications as follows. 6 µM of CHIR99021 (Stemcell Technologies, 72054) was added on day 0 and left for 24h followed by medium change. On day 3, 5µM IWP2 (Sigma Aldrich, I0536) was added using 50/50 mix of new fresh medium and conditioned medium collected from each well and left for 48h. Culture medium from day 0 until day 7 was RPMI1640 (HyClone, SH30027.01) plus B-27 serum-free supplement without insulin (Gibco, A1895601). From day 7 and onwards RPMI1640 with B-27 serum free supplement with insulin (Gibco, 17504044) was used and changed every 2-3 days.

hESC differentiation to hepatocyte-like cells

hESCs were differentiated into hepatocyte-like cells as described previously[22,23], with some modifications. Briefly, hESCs were dissociated into small clumps using ReLeSR and plated onto gelatin-coated coverslips in a 12-well plate with mTeSR. Two days later, hESCs were induced to differentiate into definitive endoderm (DE) cells in RPMI-1640 medium (Gibco) containing 2% B-27 (Invitrogen), 1% non-essential amino acids (Gibco), 1% GlutaMAX™ (Gibco) and 50µM 2-mercaptoethanol (Gibco) (basal differentiation medium), supplemented with 100ng/ml Activin A (R&D Systems), 3 µM CHIR99021 (Tocris) and 10 µM LY294002 (LC Labs) for the first three days (D0 to D3). From D3 to D6, cells were incubated in basal differentiation medium supplemented with 50 ng/ml Activin A to form foregut endoderm cells. From D6 to D10, cells were incubated in basal differentiation medium supplemented with 20 ng/ml BMP4 (Miltenyi Biotec) and 10ng/ml FGF10 (Miltenyi Biotec) to form hepatic endoderm cells. From D10 to D24, hepatic endoderm cells were incubated in HCM Bulletkit (Lonza) differentiation

media supplemented with 30 ng/ml Oncostatin M (Miltenyi Biotec) and 50 ng/ml HGF (Miltenyi Biotec). Differentiation medium was replaced every two or three days.

hESC neural induction

H1 cells cultured in mTeSR complete medium for 1-2 days were then used for neural induction as published[24,25]. Briefly, 20-30% confluent H1 cells were treated with CHIR99021, SB431542 and Compound E in neural induction media, changed every 2 days; 7 days later, the cells were split 1:3 by Accutase and seeded on matrigel-coated plates. ROCK inhibitor (1254, Tocris) was added (final concentration 10µM) to the suspension at passaging. Cells were then cultured in neural cell culture medium. These derived cells are neural precursor cells (NPC), which were used for further studies.

Neuronal differentiation

Spontaneous neuronal differentiation was performed as previously described[24]. Briefly, the derived 2 $X10^5$ NPCs were seeded on poly-l-lysine (P4707, Sigma) and laminin (L2020, Sigma) coated 6-well plates in neural cell culture medium. The next day, the cells were cultured in neuron differentiation medium: DMEM/F12 (11330-032), Neurobasal (21103-049), 1 X N2 (17502-048), 1 X B27 (17504-044), 300ng/ml cAMP (A9501), 0.2mM vitamin C (A4544-25), 10ng/ml BDNF (450-02), 10ng/ml GDNF (450-10) until day 30.

Immunofluorescence

Cells on coverslips were fixed in 4% paraformaldehyde (Wako) for 15 min at room temperature, before blocking in 5% donkey serum (EMD milipore) in PBS with 0.1% Triton X-100 for 1h at room temperature. Cells were stained with primary antibody overnight at 4°C (see Supp. Table 5 for antibodies used), or for control slides with blocking buffer. Secondary antibody staining was done with the appropriate AlexaFluor 594 for 1h at room temperature. Lastly, cells were stained with DAPI (Sigma-Aldrich, 1:5000) for 20 min at room temperature. Coverslips were mounted onto glass slides using Vectashield (Vector Laboratories). Images were taken using the EVOS M5000 microscope. Light intensity and gain were kept consistent across samples and controls with each antibody.

Author contributions

MIA conceived and designed the project, with help from **RSYF**. **EM**, **MIA** and **TBA** designed and conducted the computational search for GSH candidate loci. **MIA** and **AP** performed the experiments, with help from **VDD**, **JW**, **JT**, **ZT**, **WXT** and **AKKT**. **MIA** and **RSYF** analysed the data and wrote manuscript, with input from all authors. **RSYF** supervised the research.

Acknowledgements

This manuscript is dedicated to the memory of Seppo J. Autio. We would like to express our gratitude to Dr Alexander Lezhava for generously supporting our ddPCR experiments. We thank the late Albert Barillé for life-long inspiration. We would also like to thank Dr Steve Oh for insightful discussions and feedback on the manuscript, as well as Dr Lee Siggens for feedback and assistance on primer design. This research was funded by the BMRC YIG 2016 (1610851033) and A*Star CDF 2020 (202D8020) to MIA. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Supplementary Figure 1. CIRCOS plot annotating the 25 unique putative GSH. Sites used for testing highlighted in pink.

Supplementary Figure 2. Screenshots of Hi-C interaction matrices from H1 hESC for each of the shortlisted GSH candidate loci. TADs are indicated by the "pyramids" of high interaction observed in the Hi-C matrices. UCSC genome browser track annotating the candidate GSH (targeted GSH in pink), GENCODE v36 and H3K27Ac mark from ENCODE shown below.

Supplementary Figure 3. PCR gel images of junction PCR and wild type allele PCR reactions for screened clones.

Supplementary Figure 4. Sanger sequencing traces of the top five predicted gRNA off-target sites for un-targeted wild type H1 hESC and the respective GSH targeted clones.

Supplementary Figure 5. Representative images of metaphase spreads used for karyotyping the GSH targeted cell lines.

Supplementary Figure 6. Representative images of Pansio-1 line immunofluorescence staining with AF594 secondary antibody alone. DAPI = nuclear staining with 4',6-diamidino-2-phenylindole, Clover = fluorescence from Clover transgene, AF594 = staining with secondary antibody alone, Overlay = overlay of the three imaged channels.

Supplementary Figure 7. Sequence maps of the plasmids used in this study.

Supplementary references

1. Lomonaco, V. *et al.* UCbase 2.0: ultraconserved sequences database (2014 update). *Database* **2014**, bau062–bau062 (2014).
2. Taccioli, C. *et al.* UCbase & miRfunc: a database of ultraconserved sequences and microRNA function. *Nucleic Acids Res.* **37**, D41–D48 (2009).
3. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
4. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.* **17**, 2042–2059 (2016).
5. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science (80-.).* **326**, 289–293 (2009).
6. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
7. Wang, Y. *et al.* The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.* **19**, 1–12 (2018).
8. Straathof, K. C. *et al.* An inducible caspase 9 safety switch for T-cell therapy. *Blood* **105**, 4247–4254 (2005).
9. Ghosh, P., Pannunzio, N. R. & Hatfull, G. F. Synapsis in Phage Bxb1 Integration: Selection Mechanism for the Correct Pair of Recombination Sites. *J. Mol. Biol.* **349**, 331–348 (2005).
10. Wu, J., Corbett, A. H. & Berland, K. M. The Intracellular Mobility of Nuclear Import Receptors and NLS Cargoes. *Biophys. J.* **96**, 3840–3849 (2009).
11. Ang, L. T. *et al.* A Roadmap for Human Liver Differentiation from Pluripotent Stem Cells. *Cell Rep.* **22**, 2190–2205 (2018).
12. Concordet, J.-P. & Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* 1–4 (2018) doi:10.1093/nar/gky354.
13. Mashiko, D. *et al.* Generation of mutant mice by pronuclear injection of circular plasmid expressing Cas9 and single guided RNA. *Sci. Rep.* **3**, 3355 (2013).
14. Taylor, S. C. *et al.* The Ultimate qPCR Experiment: Producing Publication Quality, Reproducible Data the First Time. *Trends Biotechnol.* **37**, 761–774 (2019).

15. Xu, H. *et al.* FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS One* **7**, e52249 (2012).
16. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
17. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
18. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. *F1000Research* **4**, 1–19 (2016).
19. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
20. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
21. Lian, X. *et al.* Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/β-catenin signaling under fully defined conditions. *Nat. Protoc.* **8**, 162–175 (2013).
22. Hannan, N. R. F., Segeritz, C. P., Touboul, T. & Vallier, L. Production of hepatocyte-like cells from human pluripotent stem cells. *Nat. Protoc.* **8**, 430–437 (2013).
23. Ng, N. H. J. *et al.* HNF4A Haploinsufficiency in MODY1 Abrogates Liver and Pancreas Differentiation from Patient-Derived Induced Pluripotent Stem Cells. *iScience* **16**, 192–205 (2019).
24. Li, W. *et al.* Rapid induction and long-term self-renewal of primitive neural precursors from human embryonic stem cells by small molecule inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 8299–8304 (2011).
25. Wang, J. *et al.* Single-cell gene expression analysis reveals regulators of distinct cell subpopulations among developing human neurons. *Genome Res.* **27**, 1783–1794 (2017).