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

Summary

Selection of the target site is an inherent question for any project aiming for directed transgene integration. Genomic safe harbour (GSH) loci have been proposed as safe sites in the human genome for transgene integration. Although several sites have been characterised for transgene integration in the literature, most of these do not meet criteria set out for a GSH and the limited set that do have not been characterised extensively. Here, we conducted a computational analysis using publicly available data to identify 25 unique putative GSH loci that reside in active chromosomal compartments. We validated stable transgene expression and minimal disruption of the native transcriptome 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.

Introduction

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 (Mitchell et al., 2004). The randomly integrated transgenes may undergo silencing (Ellis, 2005; Mok et al., 2007) 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 (Hacein-Bey-Abina et al., 2003; Howe et al., 2008) in gene therapy trials. GSH loci have been previously suggested as safe sites for transgene integration. Criteria proposed for a putative GSH include; a set distance from coding and non-coding genes; with added separation from known oncogenes and miRNAs, and no disruption of transcriptional units or ultraconserved regions (Papapetrou and Schambach, 2016; Papapetrou et al., 2011; Sadelain et al., 2012). 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 (Papapetrou and Schambach, 2016). Here, we conducted a computational analysis to filter sites that meet criteria for GSH loci. In addition to the safety criteria, we identified regions that reside in active chromosomal compartments in many human cell and tissue types. Our analysis yielded a final list of 25 unique putative GSH that are predicted to be accessible in multiple cell types. We used human embryonic stem cells (hESCs) and their differentiated progeny to validate stable transgene expression in three of the putative GSH sites *in vitro*. Furthermore, to enable easy targeted transgene expression, we generated three hESC lines with constitutive landing pad expression constructs targeted into the three validated GSH.

Results

Computational filtering for safe and accessible loci

A list of criteria has previously been suggested for a given locus to qualify as putative GSH (Papapetrou and Schambach, 2016; Sadelain et al., 2012). 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 (Meuleman et al., 2020). 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 (Schmitt et al., 2016). To extend the analysis beyond the limited set of samples, we utilised RNA-seq data of all available tissue types from the GTEx portal (Carithers et al., 2015). 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 1). 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 1) using BLAT to generate a final shortlist of 25 unique putative universal GSH sites in the human genome (Fig. 1 A-B, Table 1).

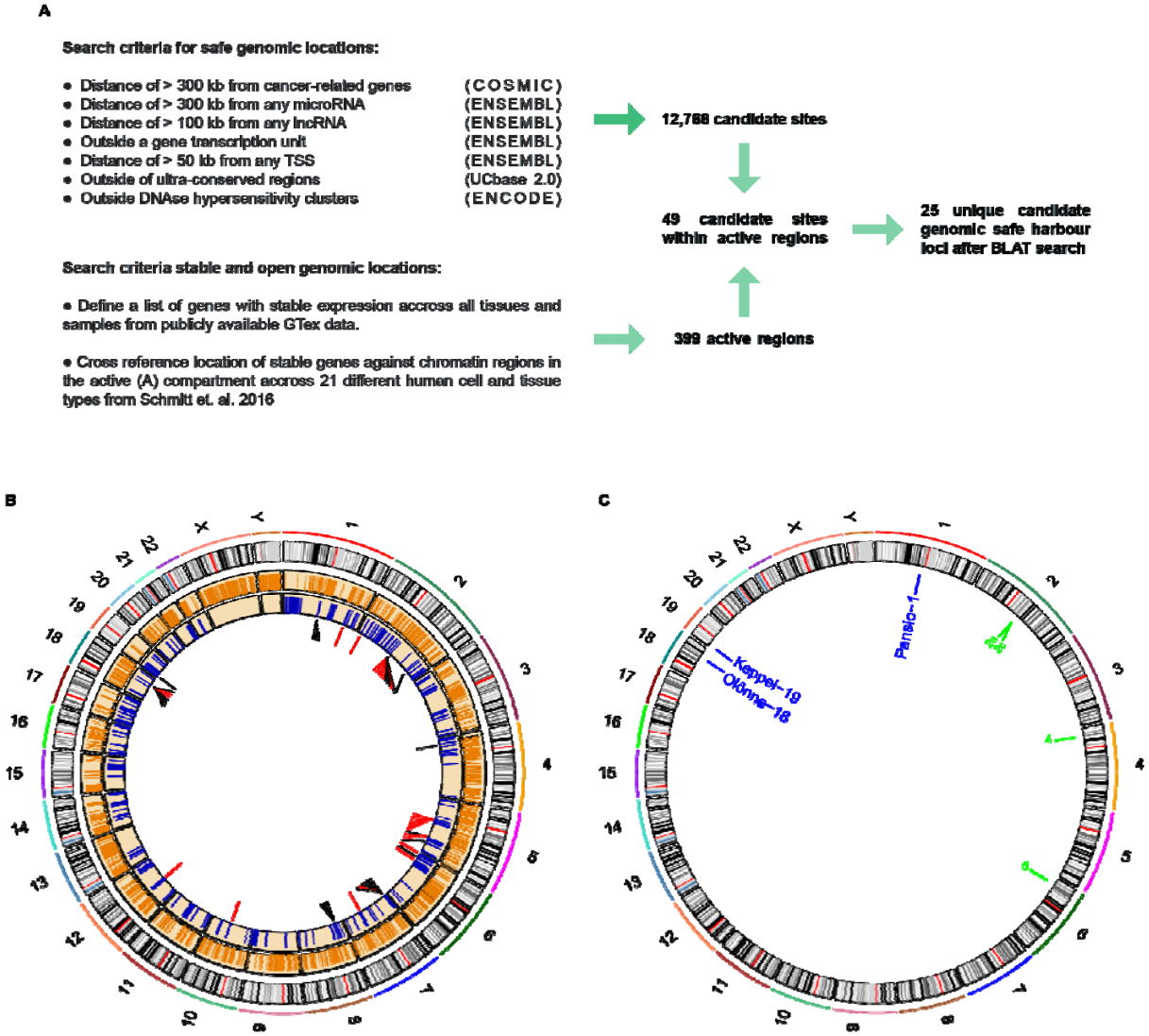


Figure 1. A) Schematic representation of the computational workflow for defining candidate GSH. **B)** CIRCOS plot summarising computational search results. Ring 1: chromosome ideograms; ring 2: orange bars indicating safe sites; ring 3: blue bars indicating active regions; ring 4: candidate sites within 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.

Table 1: Coordinates of candidate GSH, their associated active chromosome regions & housekeeping gene, and BLAT score against the most similar region

Chromosome	Start	End	Width	Active region start	Active region end	Housekeeping gene	BLAT-score
1	113289036	113289342	307	113000001	114000000	HIPK1	0.21
1	113314841	113318369	3529	113000001	114000000	HIPK1	0.18
1 *	113339961	113340514	554	113000001	114000000	HIPK1	0.27
2 *	128912721	128914814	2094	128000001	129000000	UGGT1	0.08
2	128918961	128919839	879	128000001	129000000	UGGT1	0.16
2 *	128932307	128935799	3493	128000001	129000000	UGGT1	0.15
2	128963272	128965759	2488	128000001	129000000	UGGT1	0.44
2	208992998	208997459	4462	208000001	209000000	PIKFYVE	0.28
4 *	17373361	17374159	799	17000001	18000000	MED28	0.04
5	131058585	131058947	363	131000001	132000000	FNIP1	0.41
5	148753741	148757219	3479	148000001	149000000	FBXO38	0.10
6 *	15727241	15727490	250	15000001	16000000	DTNBP1	0.47
7	4314741	4315279	539	4000001	5000000	FOXK1	0.15
7	4321017	4323839	2823	4000001	5000000	FOXK1	0.09
7	4328040	4329659	1620	4000001	5000000	FOXK1	0.21
7	4353504	4354219	716	4000001	5000000	FOXK1	0.32
7	4454808	4456201	1394	4000001	5000000	FOXK1	0.17
8	23945241	23945819	579	23000001	24000000	R3HCC1	0.34
8	23986981	23988319	1339	23000001	24000000	R3HCC1	0.06
8	23999628	24001194	1567	23000001	24000000	R3HCC1	0.02
18	56339813	56340245	433	56000001	57000000	TXNL1	0.06
18	56396821	56397319	499	56000001	57000000	TXNL1	0.07
18	56410681	56411039	359	56000001	57000000	TXNL1	0.14
18 *	56534775	56536439	1665	56000001	57000000	TXNL1	0.16
19 *	5400761	5402139	1379	5000001	6000000	SAFB	0.18

* = GSH shortlisted for *in vitro* validation

Targeted knock-in at putative GSH with CRISPR/Cas9

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. 1). 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 2). To exclude potential off-target effects of CRISPR/Cas9 targeting, we used a version of Cas9 with enhanced specificity (Slaymaker et al., 2016) 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 3, and Supp. Fig. 2). 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. 3). We named the successfully targeted safe harbours after real world harbours, designating them Pansio-1, Olônne-18, and Keppel-19.

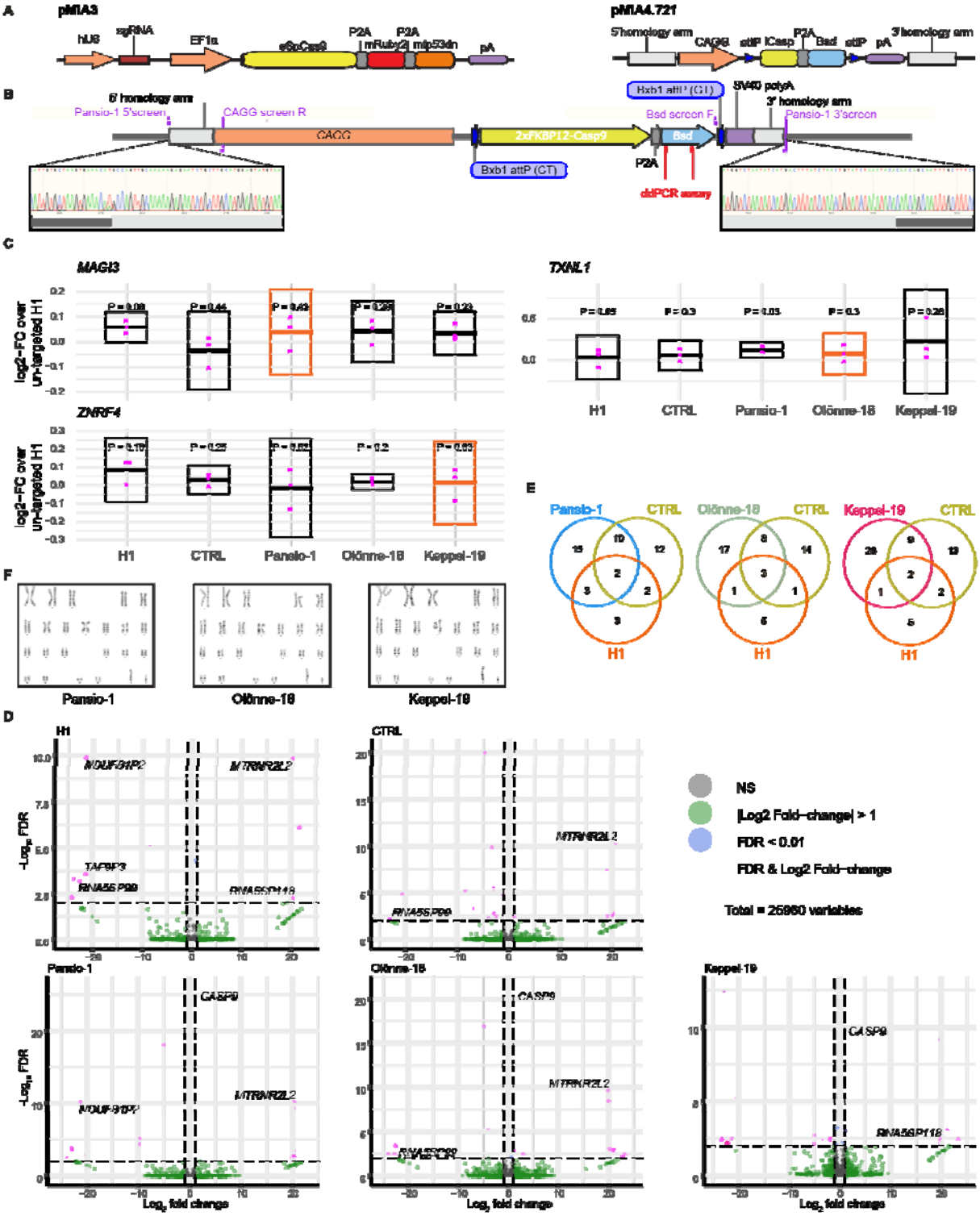


Figure 2. A) Schematic representation of CRISPR/Cas9 plasmid (pMIA3) and homology directed repair donor (pMIA4.721) used for targeting with functional components annotated. **B)** Schematic of integrated landing pad expression construct. Positions of primers for junction-PCR as well as of ddPCR assay 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. **F)** Representative images of metaphase spreads used for karyotyping the GSH targeted cell lines.

In vitro validation of targeted GSH in hESCs

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 (Fig. 2 D and Supp. Table 4). 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 4). A high proportion of the DE genes found in our GSH targeted lines were shared with control targeted H1 cells and/or untargeted wild-type H1 cells (Fig. 2 E and Supp. Table 4), 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 4). 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 4). 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 4). As a further safety check, we also conducted karyotyping of the three targeted clones and observed no abnormalities (Fig. 2 F). 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. 4) 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.

Discussion

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 (Papapetrou and Schambach, 2016). A number of previous studies have reported discovery and usage of a handful of integration sites (Costa et al., 2005; Eyquem et al., 2013; Papapetrou et al., 2011; Pellenz et al., 2019; Rodriguez-Fornes et al., 2020), which fulfil a subset of criteria previously suggested for GSH (Papapetrou and Schambach, 2016; Sadelain et al., 2012). 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 (Aznauryan et al., 2021; Costa et al., 2005; Eyquem et al., 2013; Papapetrou et al., 2011; Pellenz et al., 2019; Rodriguez-Fornes et al., 2020). 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 (Aznauryan et al., 2021; Eyquem et al., 2013; Pellenz et al., 2019; Rodriguez-Fornes et al., 2020).

The integration sites that have been most heavily utilised in research, namely AAVS1, CCR5 and Rosa26 orthologous site (Irion et al., 2007; Kotin et al., 1992; Liu et al., 1996; Perez et al., 2008), do not meet the criteria set out for GSH (Papapetrou and Schambach, 2016; Sadelain et al., 2012). 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) (Sadelain et al., 2012). 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 (Ordovás et al., 2015) and cardiomyocytes (Bhagwan et al., 2019). 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 (Jerkovic´ and Cavalli, 2021).

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 (Anzalone et al., 2021; Durrant et al., 2021; Ioannidi et al., 2021; Klompe et al., 2019) 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.

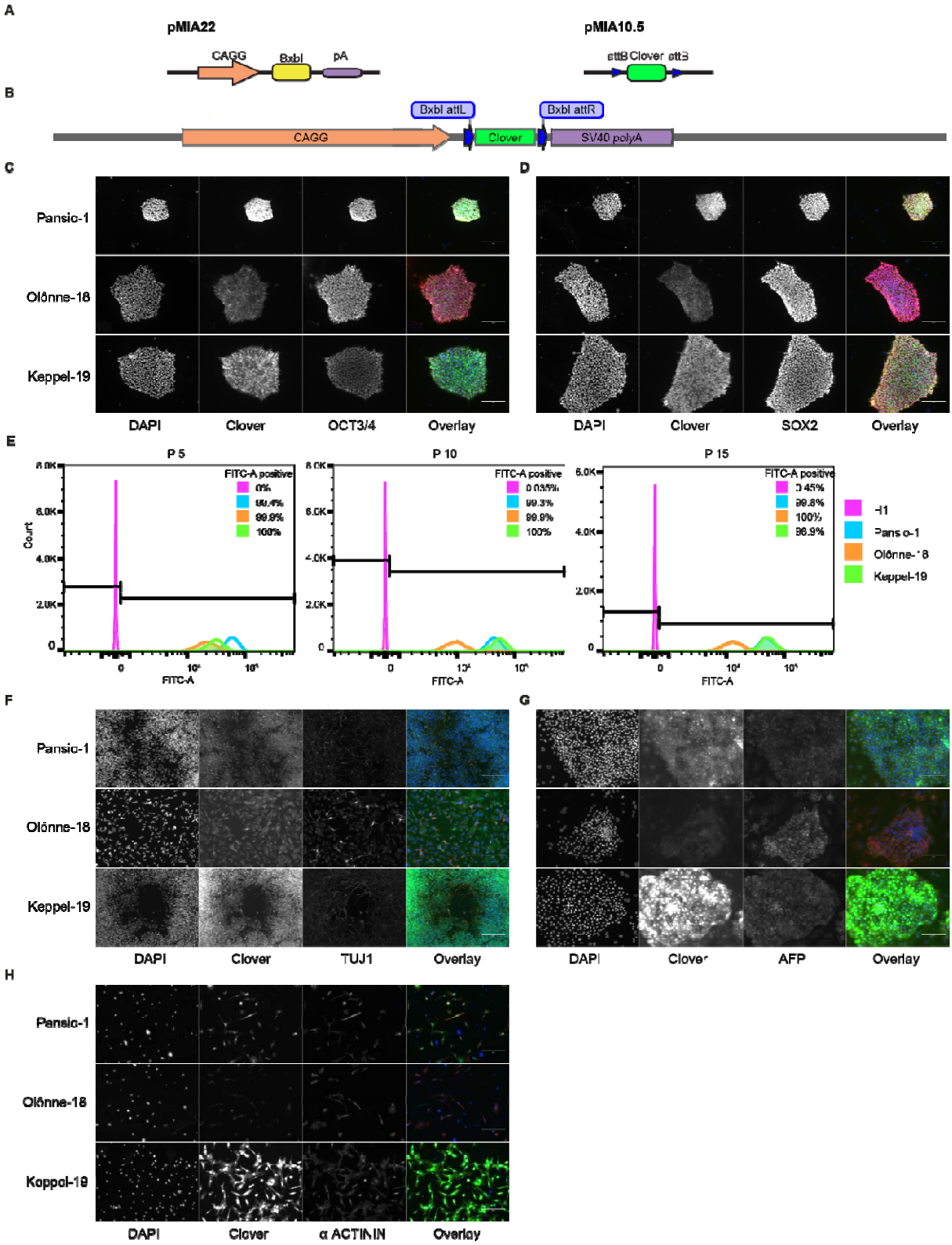


Figure 3. A) Schematic representation of integrase expression construct (pMIA22) and transposon donor construct (pMIA10.5). **B)** Schematic of landing pad construct with integrated Clover transgene. **C)** Representative immunofluorescence images of Clover-integrated GSH H1 cells. DAPI = nuclear staining 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.

References

Anzalone, A. V, Gao, X.D., Podracky, C.J., Nelson, A.T., Koblan, L.W., Raguram, A., Levy, J.M., Mercer, J.A.M., and Liu, D.R. (2021). Programmable large DNA deletion, replacement, integration, and inversion with twin prime editing and site-specific recombinases. BioRxiv.

Aznauryan, E., Yermanos, A., Kinzina, E., Kapetanovic, E., Milanova, D., Church, G.M., and Reddy, S.T. (2021). Discovery and validation of novel human genomic safe harbor sites for gene and cell therapies. BioRxiv 2021.03.04.433856.

Bhagwan, J.R., Collins, E., Mosqueira, D., Bakar, M., Johnson, B.B., Thompson, A., Smith, J.G.W., and Denning, C. (2019). Variable expression and silencing of CRISPR-Cas9 targeted transgenes identifies the AAVS1 locus as not an entirely safe harbour. F1000Research *8*, 1911.

Carithers, L.J., Ardlie, K., Barcus, M., Branton, P.A., Britton, A., Buia, S.A., Compton, C.C., DeLuca, D.S., Peter-Demchok, J., Gelfand, E.T., et al. (2015). A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreserv. Biobank. *13*, 311–319.

Costa, M., Dottori, M., Ng, E., Hawes, S.M., Sourris, K., Jamshidi, P., Pera, M.F., Elefanty, A.G., and Stanley, E.G. (2005). The hESC line Envy expresses high levels of GFP in all differentiated progeny. Nat. Methods *2*, 259–260.

Durrant, M.G., Fanton, A., Tycko, J., Hinks, M., Chandrasekaran, S., Perry, N.T., Schaepe, J., Du, P.P., Bintu, L., Bassik, M.C., et al. (2021). Large-scale discovery of recombinases for integrating DNA into the human genome. BioRxiv.

Ellis, J. (2005). Silencing and Variegation of Gammaretrovirus and Lentivirus Vectors. Hum. Gene Ther. *16*, 1241–1246.

Eyquem, J., Poirot, L., Galetto, R., Scharenberg, A.M., and Smith, J. (2013). Characterization of three loci for homologous gene targeting and transgene expression. Biotechnol. Bioeng. *110*, 2225–2235.

Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M.P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C.S., Pawliuk, R., Morillon, E., et al. (2003). LMO2-Associated Clonal T Cell Proliferation in Two Patients after Gene Therapy for SCID-X1. Science (80-.). *302*, 415–419.

Howe, S.J., Mansour, M.R., Schwarzwaelder, K., Bartholomae, C., Hubank, M., Kempski, H., Brugman, M.H., Pike-Overzet, K., Chatters, S.J., de Ridder, D., et al. (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. J. Clin. Invest. *118*, 3143–3150.

Ioannidi, E.I., Yarnall, M.T.N., Schmitt-ulms, C., Krajeski, R.N., Lim, J., Villiger, L., Zhou, W., Jiang, K., Roberts, N., Zhang, L., et al. (2021). Drag-and-drop genome insertion without DNA cleavage with CRISPR-directed integrases. BioRxiv.

Irion, S., Luche, H., Gadue, P., Fehling, H.J., Kennedy, M., and Keller, G. (2007). Identification and targeting of the ROSA26 locus in human embryonic stem cells. Nat. Biotechnol. *25*, 1477–1482.

Jerkovic´, I., and Cavalli, G. (2021). Understanding 3D genome organization by multidisciplinary methods. Nat. Rev. Mol. Cell Biol. *22*, 511–528.

Klompe, S.E., Vo, P.L.H., Halpin-Healy, T.S., and Sternberg, S.H. (2019). Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. Nature.

Kotin, R.M., Linden, R.M., and Berns, K.I. (1992). 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.
Liu, R., Paxton, W.A., Choe, S., Ceradini, D., Martin, S.R., Horuk, R., MacDonald, M.E., Stuhlmann, H., Koup, R.A., and Landau, N.R. (1996). Homozygous Defect in HIV-1 Coreceptor Accounts for Resistance of Some Multiply-Exposed Individuals to HIV-1 Infection. Cell *86*, 367–377.
Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A., et al. (2020). Index and biological spectrum of human DNase I hypersensitive sites. Nature *584*, 244–251.
Mitchell, R.S., Beitzel, B.F., Schroder, A.R.W., Shinn, P., Chen, H., Berry, C.C., Ecker, J.R., and Bushman, F.D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol. *2*.
Mok, H.P., Javed, S., and Lever, A. (2007). 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.
Ordovás, L., Boon, R., Pistoni, M., Chen, Y., Wolfs, E., Guo, W., Sambathkumar, R., Bobis-Wozowicz, S., Helsen, N., Vanhove, J., et al. (2015). Efficient recombinase-mediated cassette exchange in hPSCs to study the hepatocyte lineage reveals AAVS1 locus-mediated transgene inhibition. Stem Cell Reports *5*, 918–931.
Papapetrou, E.P., and Schambach, A. (2016). Gene insertion into genomic safe harbors for human gene therapy. Mol. Ther. *24*, 678–684.
Papapetrou, E.P., Lee, G., Malani, N., Setty, M., Riviere, I., Tirunagari, L.M.S., Kadota, K., Roth, S.L., Giardina, P., Viale, A., et al. (2011). Genomic safe harbors permit high β-globin transgene expression in thalassemia induced pluripotent stem cells. Nat. Biotechnol. *29*, 73–78.
Pellenz, S., Phelps, M., Tang, W., Hovde, B.T., Sinit, R.B., Fu, W., Li, H., Chen, E., and Monnat, R.J. (2019). New Human Chromosomal Sites with "Safe Harbor" Potential for Targeted Transgene Insertion. Hum. Gene Ther. *X*, hum.2018.169.
Perez, E.E., Wang, J., Miller, J.C., Jouvenot, Y., Kim, K.A., Liu, O., Wang, N., Lee, G., Bartsevich, V. V., Lee, Y.-L., et al. (2008). Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. Nat. Biotechnol. *26*, 808–816.
Rodriguez-Fornes, F., Quintana-Bustamante, O., Lozano, M.L., Segovia, J.C., Bueren, J.A., and Guenechea, G. (2020). Targeted gene therapy into a safe harbor site in human hematopoietic progenitor cells. Gene Ther.
Sadelain, M., Papapetrou, E.P., and Bushman, F.D. (2012). Safe harbours for the integration of new DNA in the human genome. Nat. Rev. Cancer *12*, 51–58.
Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., et al. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep. *17*, 2042–2059.
Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. Science (80-.). *351*, 84–88.

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.

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. **F)** Representative images of metaphase spreads used for karyotyping the GSH targeted cell lines.

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

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

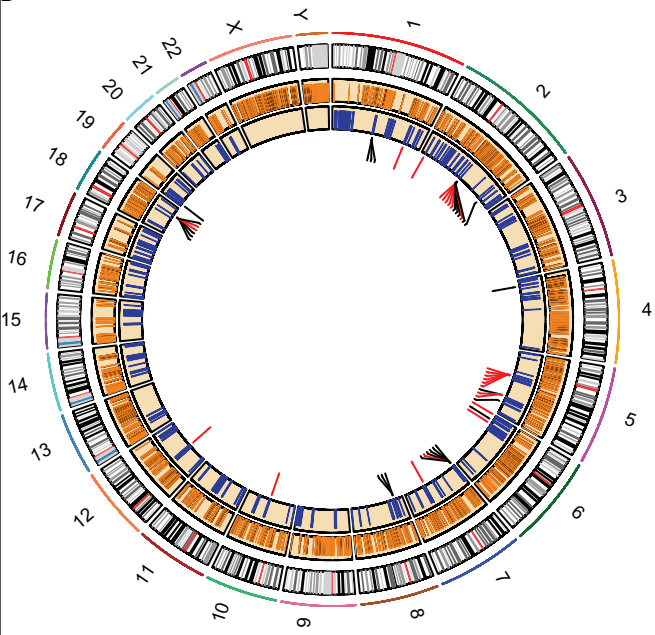
12,766 candidate sites

49 candidate sites within active regions

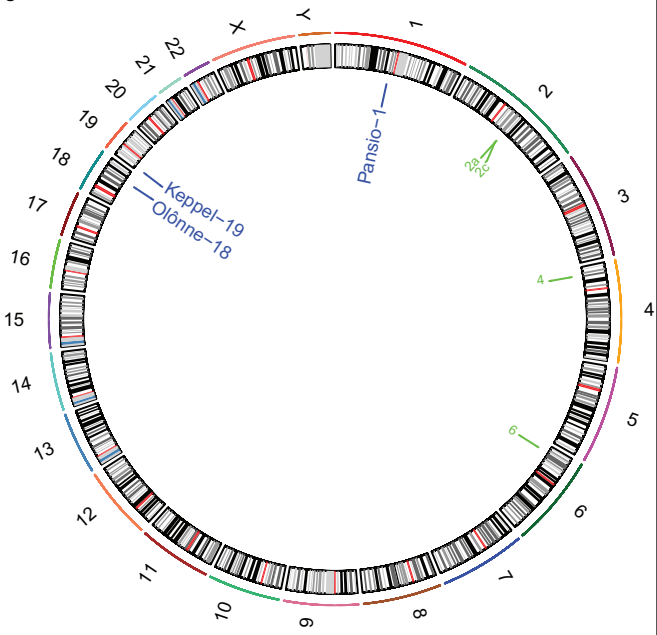
25 unique candidate genomic safe harbour loci after BLAT search

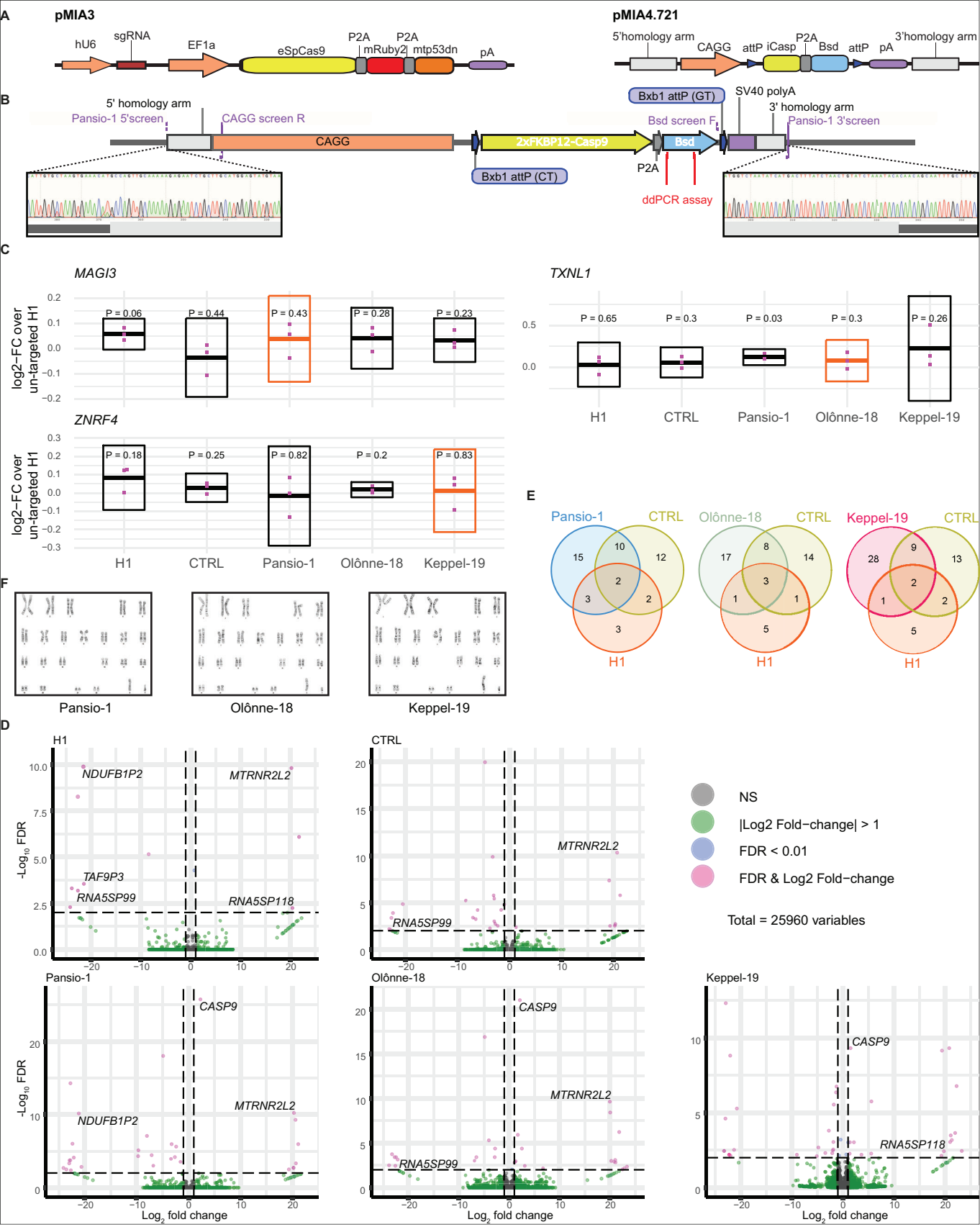
399 active regions

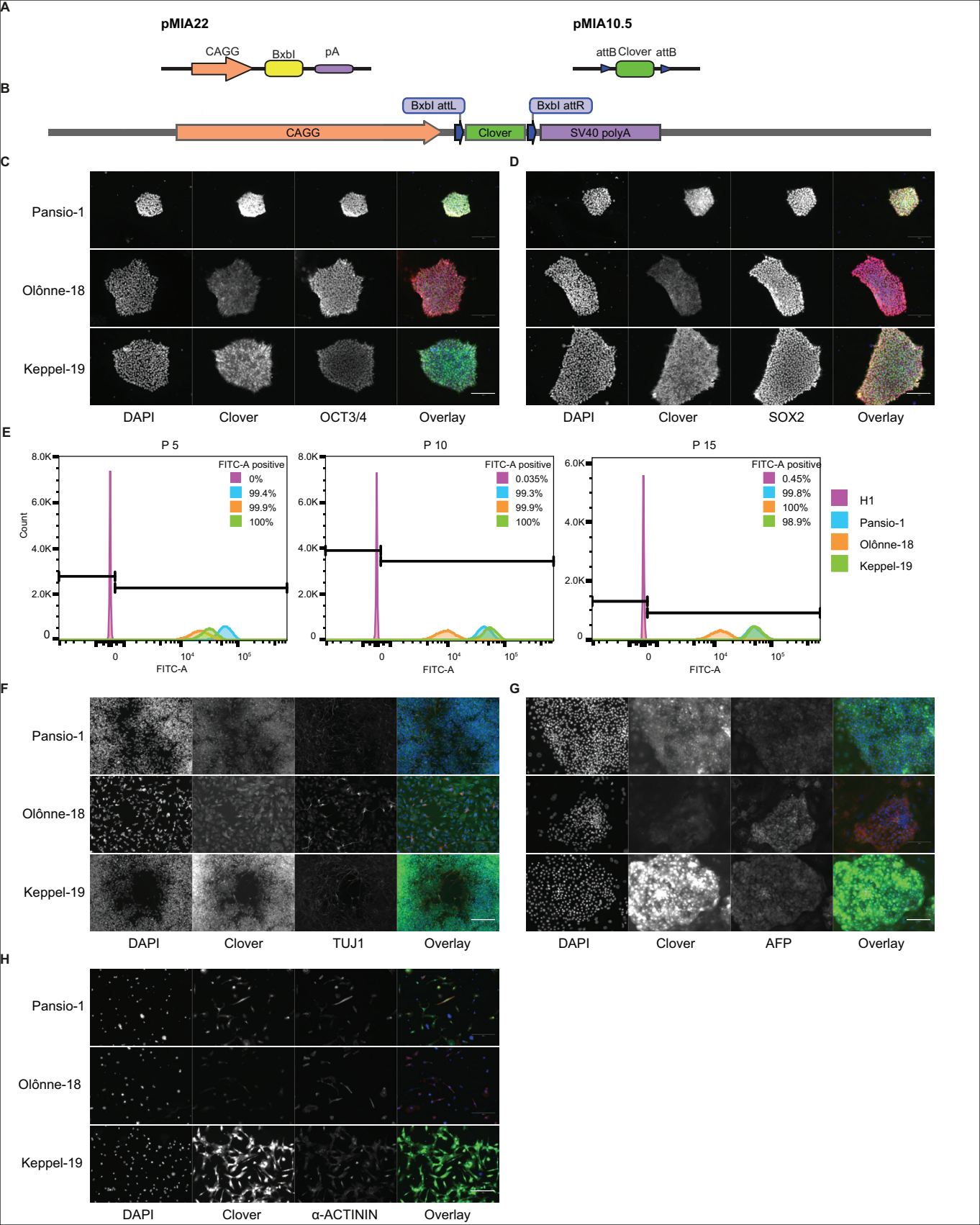
B



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)
SV40 polyA
3' homology arm
Bsd screen F
Pansio-1 3'screen
Bsd
P2A
ddPCR assay
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
ZNRF4
P = 0.18
P = 0.25
P = 0.82
P = 0.2
P = 0.83
H1
CTRL
Pansio-1
Olônne-18
Keppel-19
E
Pansio-1
CTRL
Olônne-18
CTRL
Keppel-19
CTRL
H1
H1
H1
F
Pansio-1
Olônne-18
Keppel-19
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