Cancer-Related Mutations are Not Enriched in Naïve Human Pluripotent Stem Cells

Giuliano Giuseppe Stirparo[1,2], Austin Smith[1,2,3*] and Ge Guo[1,2,4*]

[1] Wellcome-MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge, CB2 0AW, United Kingdom
[2] Living Systems Institute, University of Exeter, Exeter, EX4 4QD, United Kingdom
[3] Department of Biochemistry, University of Cambridge, Cambridge, CB2 1QR, United Kingdom
[4] Lead contact

* Correspondence: austin.smith@exeter.ac.uk; g.guo@exeter.ac.uk

Summary

Previous analysis of RNA-seq data from human naïve pluripotent stem cells reported multiple point mutations in cancer-related genes and implicated selective culture conditions. We observed, however, that those "mutations" were only present in co-cultures with mouse feeder cells. Inspection of reads containing the polymorphisms revealed complete identity to mouse reference genome. After filtering to remove sequences of mouse origin, the actual incidence of oncogenic polymorphisms arising in naïve pluripotent stem cells is close to zero.

Keywords

Pluripotent stem cell; single nucleotide polymorphism (SNP); naïve pluripotency; sequencing informatics; cancer-related mutations; TP53

Introduction

An important consideration for the use of human pluripotent stem cells in biomedical research and regenerative medicine is the acquisition of mutations, in particular in genes associated with cancer. This issue was highlighted in a recent study that reported point mutations in many cancer-related genes in one third of hPSC lines (Avior et al., 2019). Using RNA-seq data from a large panel of primed and naïve hPSCs, Avior et al. (2019) discovered recurrent non-synonymous single nucleotide polymorphisms (SNPs) in multiple Tier 1 cancer genes. Of particular note, the authors highlighted a 4-fold higher incidence of these mutations in naïve hPSCs compared with primed hPSCs. Naïve cells are maintained via chemical inhibition of several signalling pathways (Dong et al., 2019) and Avior et al. (2019) proposed that oncogenic mutations are selected for because they confer a growth advantage in the presence of the inhibitors. The finding of mutations in genes linked to growth and cancer raises potentially grave concerns about consequences for in vitro phenotypes and in vivo tumorigenicity.

The study by Avior et al (2019) included analysis of some samples from a dataset deposited by our laboratory (Guo et al., 2017). They reported detection of mutations in *TP53* and other genes in the naïve cell line cR-S6EOS. In our initial characterisation of cR-S6EOS, we did not observe the four functionally validated dominant-negative mutations in *TP53* that had previously been detected in a number of conventional hPSCs (Merkle et al., 2017). To clarify the prevalence of cancer-related mutations in naïve hPSCs, we re-examined RNA-seq data from different cultures of cR-S6EOS and other naïve cell lines.

Results

We first inspected the existence of the cancer-related mutations reported by Avior (2019) in our cR-S6EOS dataset (Guo et al., 2017). We applied the established GATK pipeline for calling single nucleotide polymorphisms (SNPs) from RNA-seq data (McKenna et al., 2010) (Figure S1A) and detected an average of ~14000 SNPs. However, the mutations reported by Avior (2019) were not present (Table S1). We reasoned that failure to detect these point mutations may have been attributable to our use of the optional Variants hard-filtering step, designed to increase the stringency of SNP calls. Indeed, when we omitted the hard-filtering, we detected a similar number of cancer-related mutations as reported by Avior et al (2019). We identified a total of 17 of the Avior SNPs across all the replicates of cR-S6EOS at two different passage numbers (Table S1). We therefore applied the pipeline without hard-filtering to analyse additional samples in our previously deposited dataset.

The data are from naïve cells in two culture conditions: (i) maintained on feeder layers of mouse embryo fibroblasts (MEF); (ii) transferred from MEF onto laminin for more than three passages. Cultures were of similar total passage number and libraries were prepared and sequenced in parallel (Guo et al., 2017). Remarkably, however, in cR-S6EOS cultures on laminin we did not detect any of the cancer-related SNPs identified in the MEF co-cultures (Figure 1A). We examined coverage per base of three SNPs identified by Avior in *TP53*, *FAT1* and *SMARCA4*. The SNPs were present in a fraction of reads from MEF cultures but completely absent from laminin samples (Figure 1B). Strikingly, in addition to the non-synonymous SNPs highlighted by Avior (2019) we noted multiple nearby SNPs in samples from cultures on MEFs that were likewise completely absent in the laminin cultures.

These observations are counter-intuitive, particularly as transition to feeder-free culture would be expected to impose stress and increase selective pressure. Moreover, collective presence or absence of multiple SNPs in multiple genes in the same cells is not consistent with natural selection. We repeated the analysis for the embryo-derived naïve cell line HNES1 (Guo et al., 2016) and again found that the cancer related mutations reported by Avior (2019) are detected only in MEF cultures and not in feeder-free conditions (Figure S1B). We were further intrigued by a significant overlap in the cancer-related SNPs identified in MEF cultures between two

entirely independent naïve cell lines, one generated by resetting and the other embryo-derived (Guo et al., 2017; Guo et al., 2016) (Figure S1C). Each of the Avior SNPs identified in HNES1 is also present in cR-S6EOS. It seems improbable that cell lines of independent genetic origins would show such a high number of identical mutations, and that these would only be present in co-cultures with MEF.

These observations prompted us to investigate whether contaminating MEF-derived sequences may contribute to SNP calls. We retrieved sequence reads harbouring SNPs reported by Avior that are detectable in cR-S6EOS MEF samples. These comprise 17 non-synonymous SNPs in 14 genes (Figure 1C). Alignment with the reference human and mouse gene sequences respectively revealed that these reads have an average of >99% identity with mouse, and less to human. In all cases the Avior SNP matches mouse gene sequence. Notably, numerous additional mismatches with human correspond to mouse nucleotide substitutions (Figure 1D).

In light of these findings we investigated systematically the contribution of contaminating MEF-derived sequences to SNP calls. We mapped a similar number of reads as Avior et al (2019) across all the studies (Figure S1D). We then applied XenofilteR, a tool previously developed for analysis of human xenografts in mice (Kluin et al., 2018). XenofilteR identifies and removes reads that map with higher efficiency to mouse than to human reference genome (Figure S1E). Direct comparison of samples of the same cell lines cultured with and without MEFs showed that XenofilteR detected and removed a high number of reads from co-cultures (Figure 1E). The fraction of reads removed by XenofilteR was significantly larger for naïve than primed hPSC samples (Figure 1F). An independent analysis using the metagenomic tool Sequence Expression AnaLyzer (SEAL) to classify human or mouse sequences yielded similar results (Table S1). Naïve cells are typically maintained at lower density than primed hPSCs, which will result in a higher contribution of MEFs in RNA-seq libraries. Variability in the representation of MEF sequences between samples likely relates to differences between cultures and laboratories in MEF preparation, relative density of hPSCs at time of harvesting, and extent to which measures are taken to deplete MEFs prior to RNA preparation. Application of XenofilteR did not significantly alter quantification of expression of the cancer associated genes (Figure S2A). We also investigated the impact on the global transcriptome by performing principal component analysis (PCA) for all expressed protein coding genes. This analysis (Figure S2B) shows no change in the separation of naïve and primed cells on PC1 with minor shifts in distribution on PC2.

We applied the GATK for RNA-seq pipeline to all the samples, with or without application of XenofilteR (Figure S1E). We initially focussed on the cancer-related SNPs identified by Avior et al. (2019). Remarkably, after depletion of mouse sequences the number of Avior SNPs fell to zero in most cases (Figure 2A, Figure S2C; Table S2). We also noticed that the number of those SNPs detectable before XenofilteR reflects the total number of mouse reads identified in each dataset (Figure 2B). A similar positive correlation (r=0.81) was identified between the number of cancer-related SNPs identified in naïve samples and the percentage of mouse reads assigned by SEAL.

Avior et al. (2019) highlighted SNPs in genes associated with signalling pathways inhibited in naive stem cell culture (*CCND2, HIF1a, FAT1, APC, BCL9L*, *MYH9* and *CDKN1B*) and asserted that these were mutations conferring selective advantage. Every one of these SNPs was eliminated by applying XenofilteR (Table S2). Importantly, XenofilteR does not prevent detection of authentic human SNPs; >40,000 SNPs were still detected in cR-S6EOS and HNES1 samples (Figure 2C). Notably, for laminin cultures this number is not significantly changed before and after XenofilteR.

We examined the reads containing Avior SNPs that were removed by XenofilteR and also those for three SNPs that remained. We aligned the reads to human and mouse reference sequences. Reads with SNPs removed by XenofilteR matched to mouse reference and harboured, on average, more than 4 mismatches with human gene reference sequence

(Figure 2D). Conversely, reads containing the three SNPs that remained after XenofilteR exhibited more mismatches with mouse than human. These SNPs were in *TP53* (pR181H, pR248Q) and *CDK12* (pE131K) (Figure 2D). Both of the *TP53* SNPs were previously detected in some primed hPSCs (Merkle et al., 2017). In each of the two positive datasets in this analysis, the *TP53* SNP pre-existed in the primed hPSCs and was therefore inherited by naïve hPSCs (Table S2). The *CDK12* SNP was detected in only one of two technical replicates in a total of 7 samples from (Sahakyan et al., 2017) (Table S1 &S2).

The SNPs reported by Avior and eliminated by XenofilteR show a very high overlap across cell lines of different genetic backgrounds cultured in different conditions and laboratories (Table S2). Incidence of identical SNPs in these circumstances would be remarkable. This is readily explained, however, by shared contamination with MEFs. For example, without use of XenofilteR, examination of reads harbouring the *CCND2* SNP revealed more than 15 single nucleotide variants which are common between different data sets and each of which matches to mouse reference (Figure 2E).

Finally, we carried out a systematic analysis of naïve hPSCs cultured in our laboratory, either in our original media formulation (t2iLGö) (Takashima et al., 2014) or improved medium (PXGL) (Bredenkamp et al., 2019a; Bredenkamp et al., 2019b). In either medium, Avior SNPs were detected only in cultures on MEF and all were removed by XenofilteR (Figure 2F). We then broadened the investigation to search for any other potential SNPs in Tier 1 cancer genes. We uncovered only one recurrent polymorphism. A non-synonymous SNP in *ARID1A* (pI692V) was detected in HNES1 samples but was not present in any other naïve cell line. *ARID1A* is frequently mutated in colon cancer, with nonsense and out of frame mutations (Forbes et al., 2016). However, missense mutations have not been functionally annotated. HNES1 is an embryo derived cell line. We examined the earliest passage dataset available (Guo et al., 2016) and detected the *ARID1A* polymorphism at an allelic frequency of around 50%, as seen in the later passage samples. Notably, we did not detect this SNP in other embryo-derived cell lines, HNES2 and HNES3 (Guo et al., 2016).

Discussion

In summary, we find no evidence for prevalence of cancer-related point mutations in naïve hPSCs. Analysis of RNA-seq can be an effective method for identifying SNPs in hPSCs, as previously shown for certain *TP53* mutations (Merkle et al., 2017) and confirmed here. However, culture of hPSCs on MEF feeder layers results in presence of mouse gene sequences in hPSC RNA-seq datasets, which can lead to erroneous SNP calls. This is particularly relevant for naïve hPSCs, which in current protocols are predominantly cultured at relatively low density on MEFs. In general the impact of MEF sequences on gene expression is small because the majority are removed during genome alignment and reads per gene are normalised (Figure S2A,B). Nonetheless, unfiltered MEF sequences can distort measurement of genes that are lowly expressed in PSCs and highly expressed in MEF such as *CCND2*, or skew comparisons between hPSCs in the presence or absence of feeders. A filtration step such as XenofilteR is advisable in such cases, in particular for short read sequencing protocols with reduced quality of genome alignment.

Our analyses demonstrate that the reported detection of multiple cancer-related SNPs (Avior et al., 2019) in naïve hPSCs is attributable to contamination with MEF-derived sequences. Following our report, Avior et al have revised their methodology (Avior et al, 2020, this issue). It is essential to apply XenofilteR or an equivalent stringent quality measure to exclude mouse sequences from co-culture samples. Further analyses of naïve cells in t2iLGö or PXGL culture conditions, including additional independent cultures, did not detect recurrent SNPs in any Tier 1 cancer genes. Therefore, neither the generation of naïve hPSCs nor their propagation impose heightened susceptibility to point mutations in cancer-associated genes.

Acknowledgments
We are grateful to James Clarke for cell culture support and to Vicki Murry and Maike Paramor for generating sequencing libraries. This research was funded by the Medical Research Council (MRC) of the United Kingdom. The Wellcome-MRC Cambridge Stem Cell Institute receives core support from Wellcome and MRC. AS is a Medical Research Council Professor.

Author Contributions
Conceptualization, GG; Investigation, GGS; Methodology, GGS; Formal analysis, GGS; Writing; GGS, AS, GG; Supervision GG, AS

Declaration of Interests
AS and GG are inventors on a patent application relating to human naïve stem cells filed by the University of Cambridge.

FIGURE LEGENDS

Figure 1

(A) Numbers of cancer-associated SNPs from Avior et al. (2019) in cR-S6EOS samples cultured on mouse feeders (MEF) or on laminin (LN) detected by the GATK pipeline without VARIANTS hard- filtering

(B) Integrative Genome Browser screenshot of selected cancer-associated SNPs from Avior et al (2019) showing per base read coverage (0-100) in cR-S6EOS cultures on MEF or laminin. Dotted lines highlight the SNP reported by Avior (2019). Positions with alternative nucleotides are represented using different colors.

(C) Average mapping percentage of total reads from cR-S6EOS(MEF) samples harbouring the indicated SNPs reported by Avior et al (2019) when aligned against human or mouse reference. See also Table S1&S2

(D) Number of mismatches in reads as in (C) aligned against human or mouse reference.

(E) Boxplots of the number of mouse reads detected by XenofilteR in naïve cell samples from cultures on MEF or laminin.

(F) Boxplots of the number of mouse reads identified by XenofilteR in naïve and primed conditions across different datasets analysed in Avior et al. (2019).

Figure 2

(A) Number of cancer associated SNPs from Avior et al. (2019) in different datasets, as reported in Avior et al., 2019 (red), detected in this study without XenofilteR (Blue), and detected after removal of mouse reads using XenofilteR (Grey).

(B) Correlation between percentage of mouse reads and numbers of cancer-associated SNPs detected for all naïve hPSCs in this study.

(C) Total number of SNPs before and after removal of mouse reads in cR-S6EOS and HNES1 cultures on MEF or laminin.

(D) Numbers of mismatches in reads harbouring the cancer-related mutation aligned against human or mouse reference. Each bar represents average number of mismatches for all reads with SNPs reported by Avior et al (2019) in naïve hPSCs. N represents number of datasets with the indicated SNP.

(E) Integrative Genome Browser screenshot of *CCND2* transcripts showing the SNP reported by Avior et al (2019) in dashed box and nearby mismatches in reads across indicated human naïve hPSC datasets.

(F) Heatmap showing number of Avior SNPs detected in human naïve hPSCs cultured in t2ilGö medium or PXGL medium on MEF or on Laminin (LN) with or without application of XenofilteR. Samples from Bredenkamp (2019) are pooled data from cultures on laminin (LN) or Geltrex (GT).

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Ge Guo, g.guo@exeter.ac.uk

Materials Availability

This study did not generate new unique reagents.

Data and Code

RNA-seq data from this study are deposited in Gene Expression Omnibus with accession number GSE150933.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell culture

Research use of hPSCs is approved by the United Kingdom Stem Cell Steering Committee.

Naïve hPSCs were cultured in 5% O_2, 7% CO_2 in a humidified incubator at 37°C. Cell lines were maintained without antibiotics and confirmed free of mycoplasma contamination by periodic in-house PCR assay.

Chemically reset (cR) (Guo et al., 2017), embryo-derived (HNES) (Guo et al., 2016) and reprogrammed (niPSC) (Bredenkamp et al., 2019b) naïve hPSCs were propagated in N2B27 with PXGL [1 μM PD0325901 (P), 2 μM XAV939 (X), 2 μM Gö6983 (G) and 10 ng/mL human LIF (L)] on irradiated MEF feeders as described (Bredenkamp et al., 2019a). ROCK inhibitor (Y-27632) and Geltrex (0.5μL per cm^2 surface area; hESC-Qualified, Thermo Fisher Scientific, A1413302,) were added to media during replating. Cultures were passaged by dissociation with Accutase (Biolegend, 423201) every 3-5 days.

METHOD DETAILS

Transcriptome sequencing

Total RNA was extracted from two biological replicate cultures of each cell line and time point using TRIzol/chloroform (Thermo Fisher Scientific, 15596018), and RNA integrity assessed by Qubit measurement and RNA nanochip Bioanalyzer. Ribosomal RNA was depleted from 1 μg of total RNA using Ribozero (Illumina kit). Sequencing libraries were prepared using the TruSeq RNA Sample Prep Kit (RS-122-2001, Illumina). Sequencing was performed on the Novaseq S1 or S2 platform (Illumina), according to the manufacturer's instructions.

QUANTIFICATION AND STATISTICAL ANALYSIS

Alignment was performed using the Genome build hg38 for human and Genome build mm10 for mouse. STAR (Dobin et al., 2013) was used for aligning reads. Ensembl release 96 was used to guide gene annotation in both species. Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used to remove adapter contamination, if present. Best practice for variant calling in RNA-seq pipeline was used (https://gatk.broadinstitute.org/hc/en-us) (FIG.S1A, FIG.S1E), together with dbSNP146 downloaded from GATK resource bundle repository (ftp.broadinstitute.org/bundle).

R package XenofilteR (Kluin et al., 2018) compared alignment quality between human and mouse mapped reads and filtered out sequences with higher mapping efficiency in mouse. We quantified alignments to gene loci with htseq-count (Anders et al., 2014) based on annotation from Ensembl 96. PCA were computed on FPKM/RPKM $\log_2$ normalized counts using all the expressed protein coding genes and R library FactoMineR (Lê et al., 2008). Integrative Genomics Viewer (IGV) was used to visualize aligned reads and coverage.

Cancer-related genes and SNP location was downloaded from Supplementary Table 2 in Avior et al., 2019.

Damaging and non-synonymous SNPs in coding regions were annotated using SNPnexus (SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine) and COSMIC database (https://cancer.sanger.ac.uk/cosmic)

Mapping between human and mouse

Reads harboring the mutations were retrieved with samtools (http://www.htslib.org/doc/samtools.html). The reads were subsequently aligned using Clustal Omega webtool (https://www.ebi.ac.uk/Tools/msa/clustalo/) against the human and mouse reference. Human reference was obtained by selecting the 50 bp before and after the mutations. This 100 bp fragment was then aligned to mouse using blastn (Altschul et al., 1990) in order to identify the syntenic mouse reference region.

During alignment of reads harboring the mutations to human and mouse reference, only aligned fragments longer than 45 bp were retained to compute number of mismatches and percentage of mapping. A seed of 8 bases was used. Sequence Expression Anayzer (SEAL) (https://jgi.doe.gov/data-and-tools/bbtools/) was used to quantify sequence abundance based on human and mouse reference genomes.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
RNA sequencing data from this study	Gene Expression Omnibus	GSE150933
Experimental Models: Cell Lines		
HNES1	Guo et al, 2016	N/A
cR-H9	Guo et al, 2017	N/A
EPC niPSC	Bredenkamp et al 2019	N/A
HDF16 niPSC	This study	N/A
Software and Algorithms		
STAR	Dobin A et al., 2013	
htseq-count	Anders S et al., 2014	
Samtools	Li et al., 2009	http://samtools.sourceforge.net/
XenofilteR	Kluin et al., 2018	https://github.com/PeeperLab/XenofilteR
R	R Core Team, 2017	https://www.R-project.org/
Genome and Genome annotation	GRCh38/mm10 Ensembl 96	http://apr2019.archive.ensembl.org/index.html
gplots	Gregory R. Warnes et al., 2019	https://cran.r-project.org/web/packages/gplots/index.html
IGV	(Robinson et al., 2011)	http://software.broadinstitute.org/software/igv/
GATK	McKenna et al., 2010	https://gatk.broadinstitute.org/hc/en-us

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq — A Python framework to work with high-throughput sequencing data. bioRxiv.

Avior, Y., Eggan, K., and Benvenisty, N. (2019). Retraction. Cell Stem Cell *28, this issue*, *bxs.

Avior, Y., Lezmi, E., Eggan, K., and Benvenisty, N (2020). **Cancer-Related Mutations Identified in Primed Human Pluripotent Stem Cells. Cell Stem Cell, this issue.**

Bredenkamp, N., Stirparo, G.G., Nichols, J., Smith, A., and Guo, G. (2019a). The Cell-Surface Marker Sushi Containing Domain 2 Facilitates Establishment of Human Naive Pluripotent Stem Cells. Stem Cell Reports *12*, 1212-1222.

Bredenkamp, N., Yang, J., Clarke, J., Stirparo, G.G., von Meyenn, F., Dietmann, S., Baker, D., Drummond, R., Ren, Y., Li, D.*, et al.* (2019b). Wnt Inhibition Facilitates RNA-Mediated Reprogramming of Human Somatic Cells to Naive Pluripotency. Stem Cell Reports.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Dong, C., Fischer, L., and Theunissen, T.W. (2019). Recent insights into the naive state of human pluripotency and its applications. Exp Cell Res, 111645.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L.*, et al.* (2016). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Research *45*, D777-D783.

Guo, G., von Meyenn, F., Rostovskaya, M., Clarke, J., Dietmann, S., Baker, D., Sahakyan, A., Myers, S., Bertone, P., Reik, W.*, et al.* (2017). Epigenetic resetting of human pluripotency. Development *144*, 2748-2763.

Guo, G., von Meyenn, F., Santos, F., Chen, Y., Reik, W., Bertone, P., Smith, A., and Nichols, J. (2016). Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass. Stem Cell Reports *6*, 437-446.

Kluin, R.J.C., Kemper, K., Kuilman, T., de Ruiter, J.R., Iyer, V., Forment, J.V., Cornelissen-Steijger, P., de Rink, I., ter Brugge, P., Song, J.-Y.*, et al.* (2018). XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data. BMC Bioinformatics *19*, 366.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: An R package for multivariate analysis. Journal of Statistical Software *25*, 1-18.

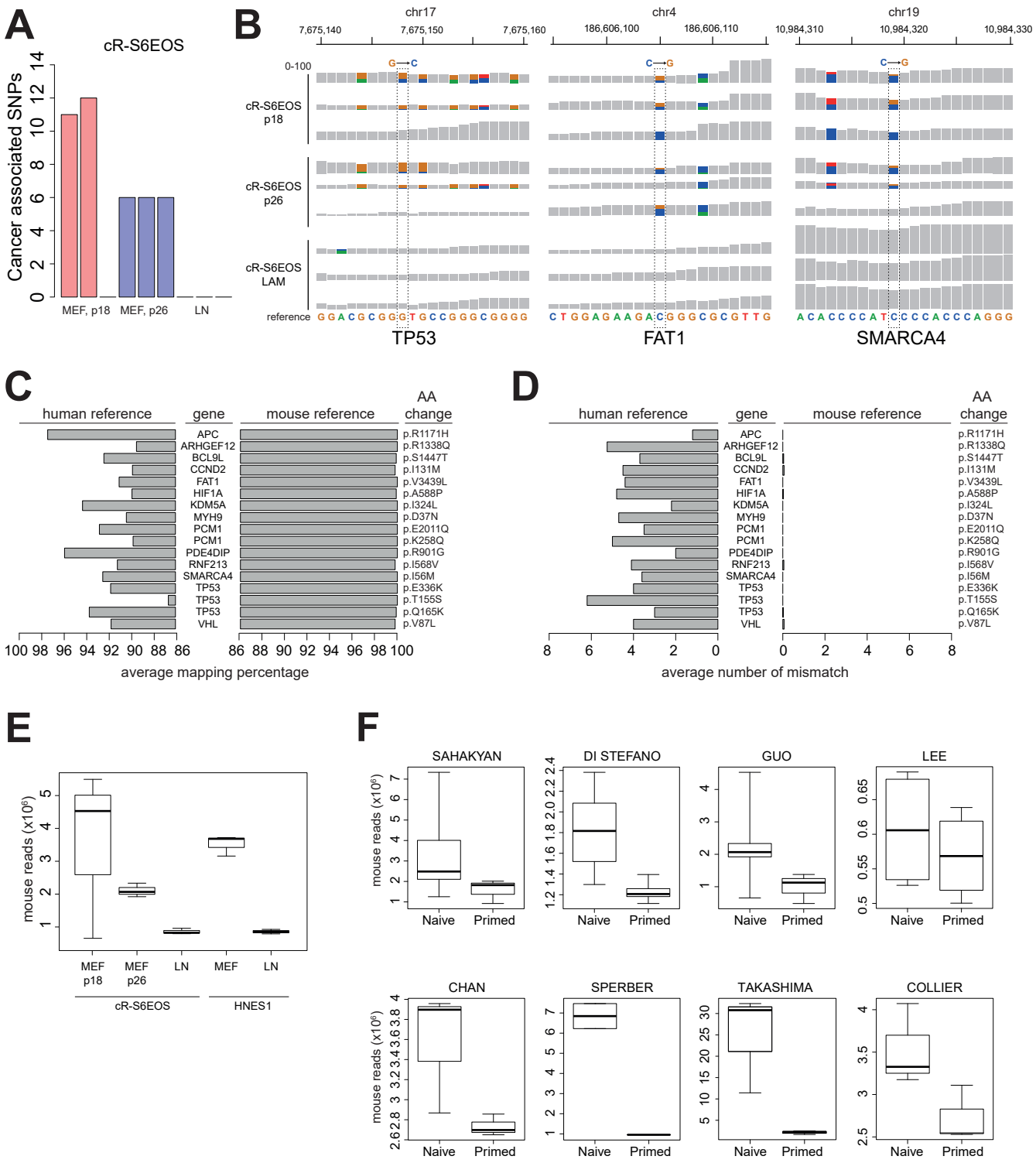
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297-1303.

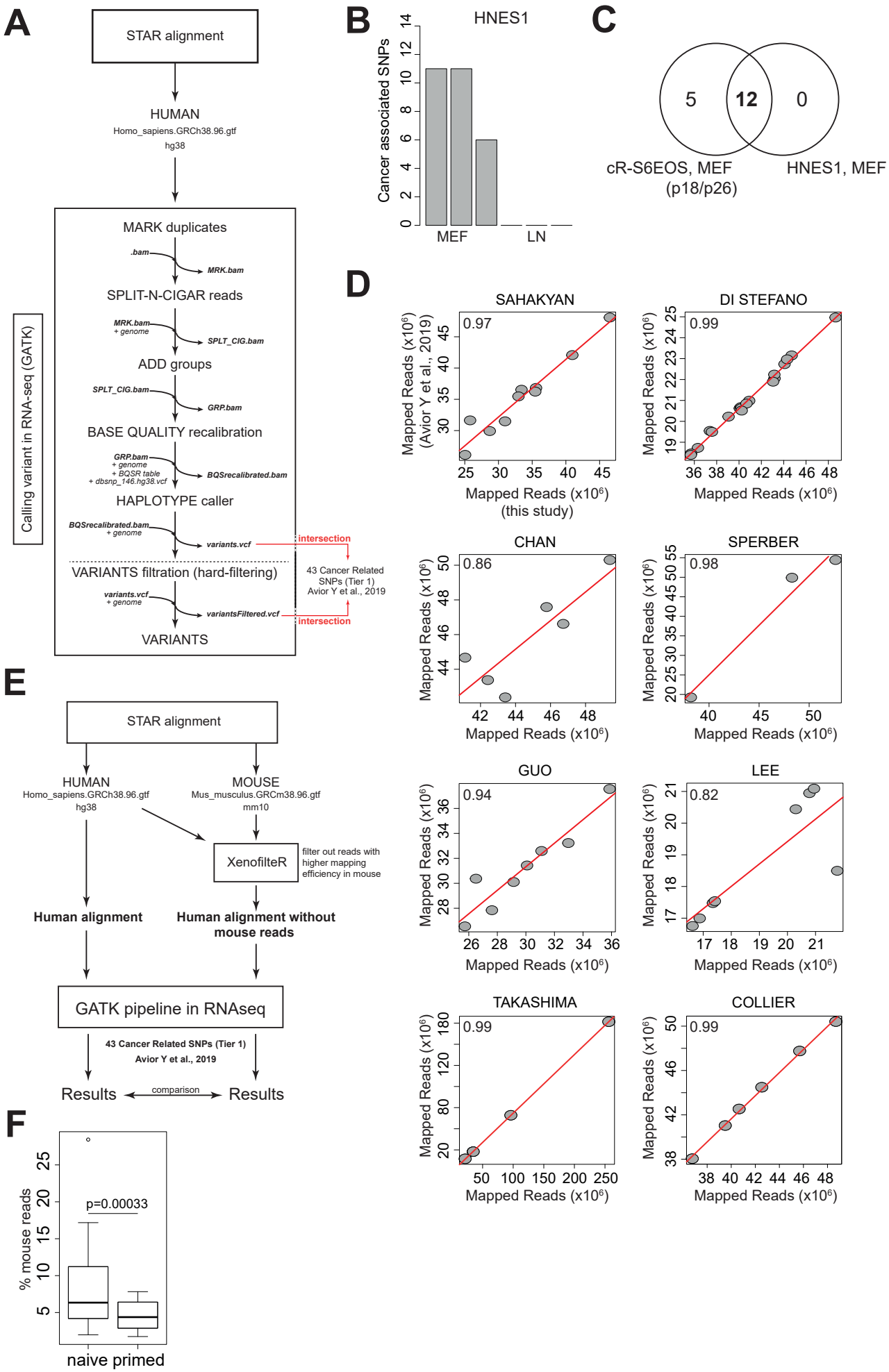
Merkle, F.T., Ghosh, S., Kamitaki, N., Mitchell, J., Avior, Y., Mello, C., Kashin, S., Mekhoubad, S., Ilic, D., Charlton, M.*, et al.* (2017). Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations. Nature *545*, 229-233.

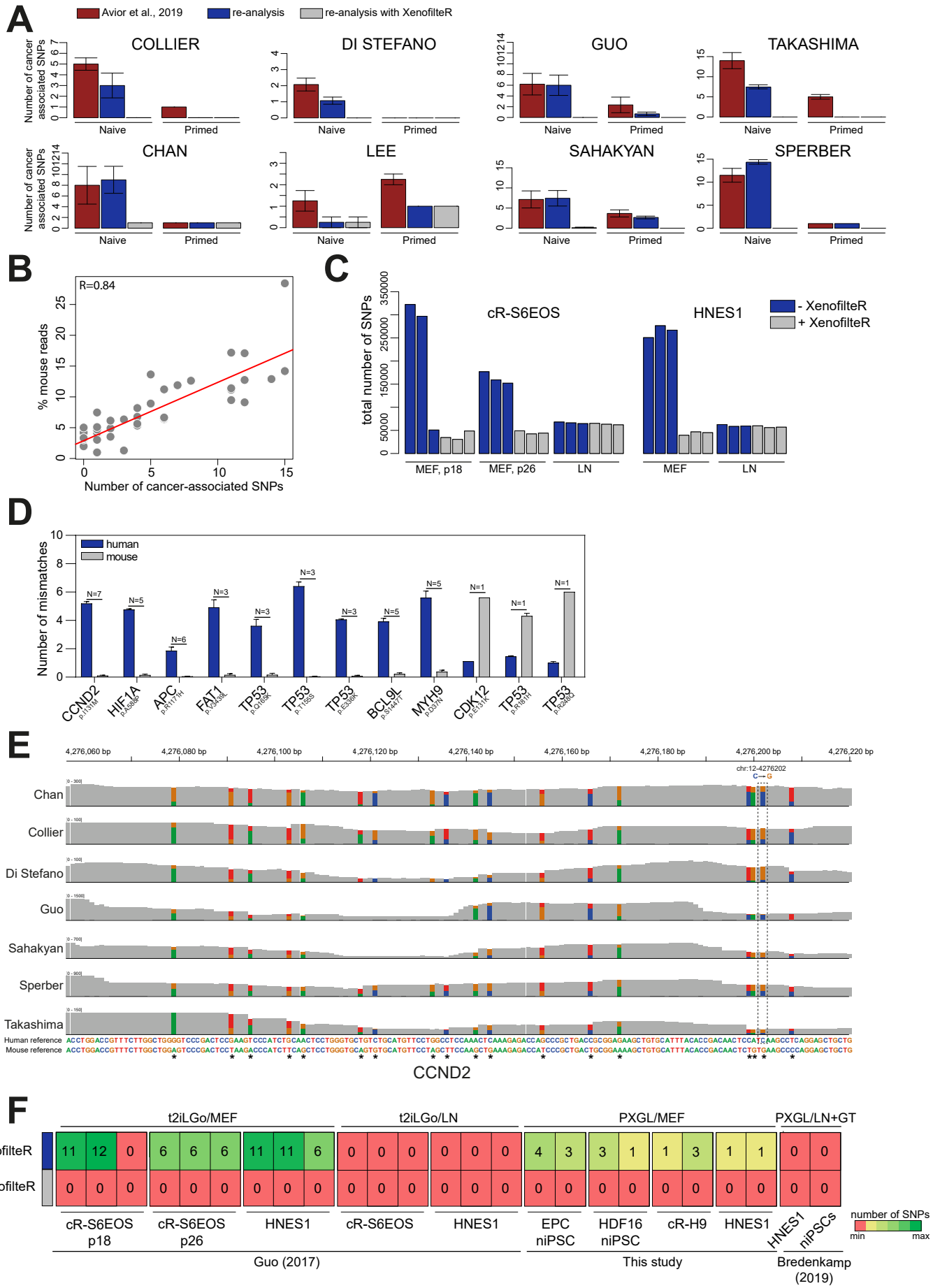
Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nature biotechnology *29*, 24-26.

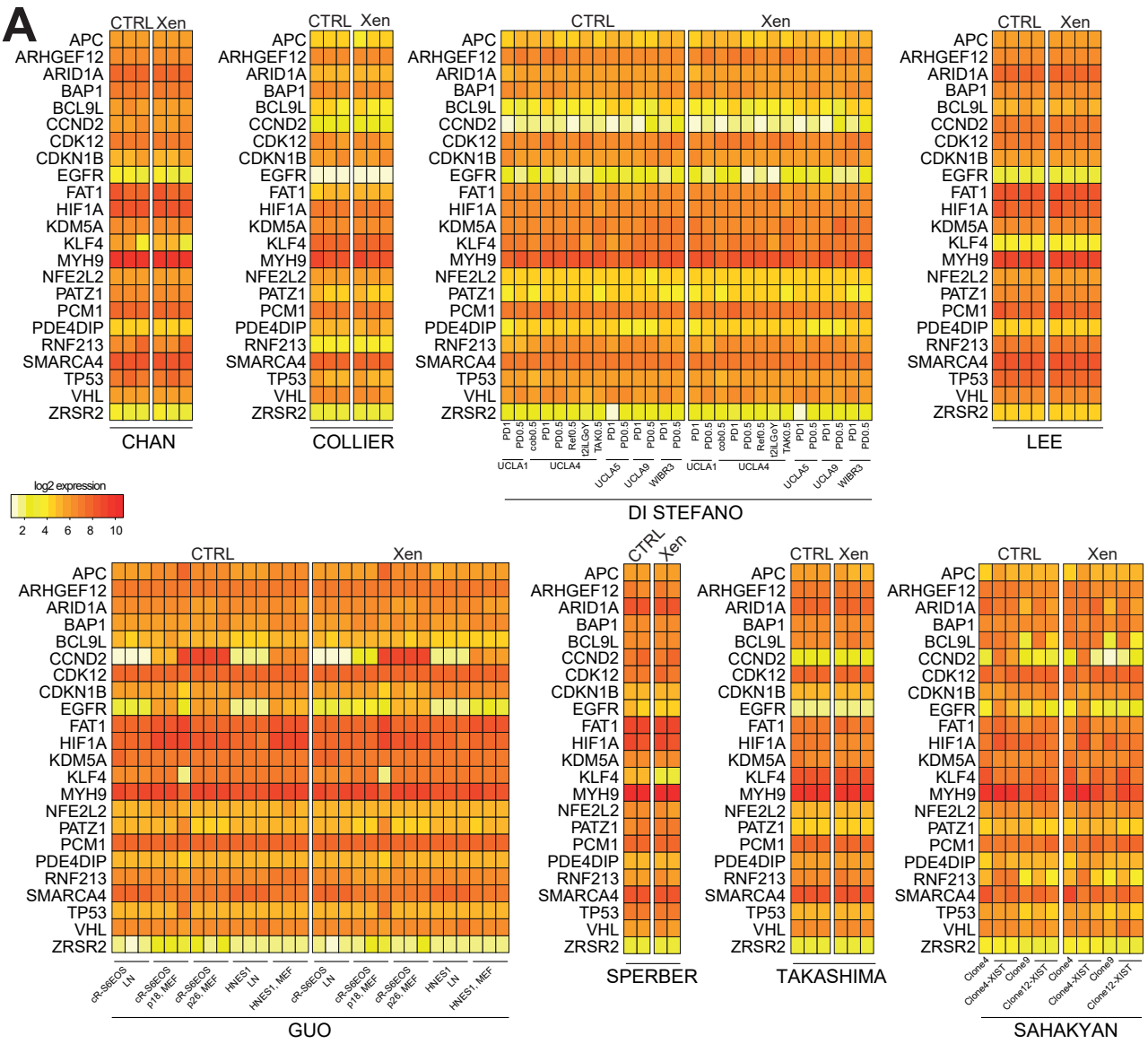
Sahakyan, A., Kim, R., Chronis, C., Sabri, S., Bonora, G., Theunissen, T.W., Kuoy, E., Langerman, J., Clark, A.T., Jaenisch, R.*, et al.* (2017). Human Naive Pluripotent Stem Cells Model X Chromosome Dampening and X Inactivation. Cell Stem Cell *20*, 87-101.

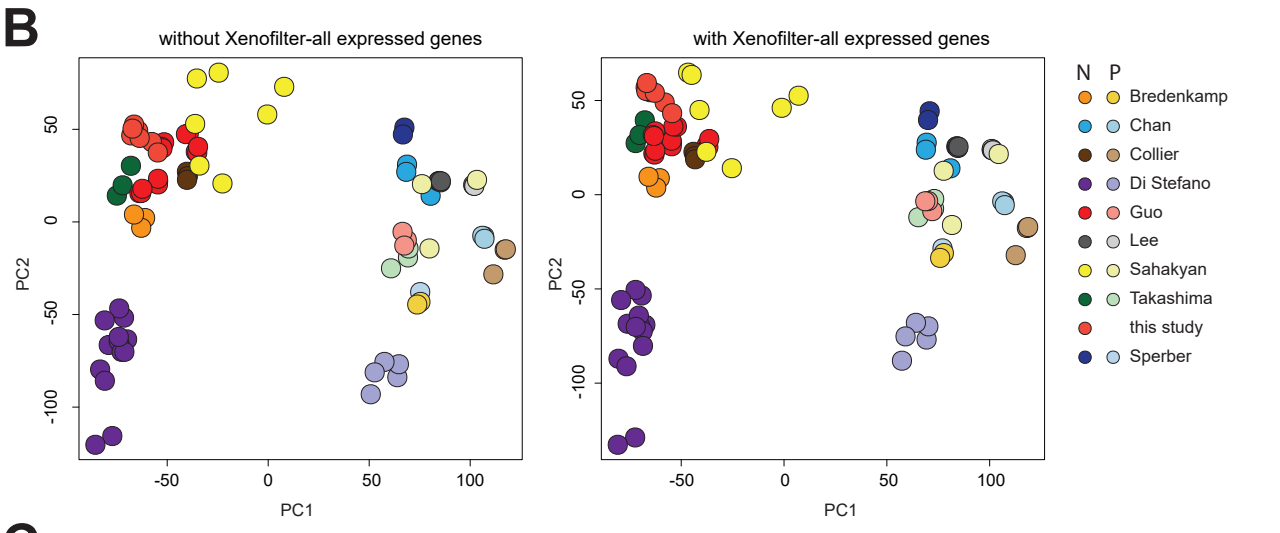
Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W.*, et al.* (2014). Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. Cell *158*, 1254-1269.

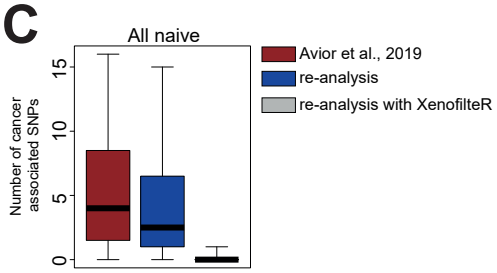

A
cR-S6EOS
Cancer associated SNPs
0 2 4 6 8 10 12 14
MEF, p18
MEF, p26
LN
B
chr17
7,675,140
7,675,150
7,675,160
chr4
186,606,100
186,606,110
chr19
10,984,310
10,984,320
10,984,330
0-100
cR-S6EOS p18
cR-S6EOS p26
cR-S6EOS LAM
reference
G G A C G C G G G T G C C G G G C G G G G
C T G G A G A A G A C G G G C G C G T T G
A C A C C C C A T C C C C A C C C A G G G
TP53
FAT1
SMARCA4
C
human reference
gene
mouse reference
AA change
APC p.R1171H
ARHGEF12 p.R1338Q
BCL9L p.S1447T
CCND2 p.I131M
FAT1 p.V3439L
HIF1A p.A588P
KDM5A p.I324L
MYH9 p.D37N
PCM1 p.E2011Q
PCM1 p.K258Q
PDE4DIP p.R901G
RNF213 p.I568V
SMARCA4 p.I56M
TP53 p.E336K
TP53 p.T155S
TP53 p.Q165K
VHL p.V87L
100 98 96 94 92 90 88 86
86 88 90 92 94 96 98 100
average mapping percentage
D
human reference
gene
mouse reference
AA change
8 6 4 2 0
0 2 4 6 8
average number of mismatch
E
mouse reads (x10^6)
1 2 3 4 5
MEF p18
MEF p26
LN
MEF
LN
cR-S6EOS
HNES1
F
SAHAKYAN
DI STEFANO
GUO
LEE
CHAN
SPERBER
TAKASHIMA
COLLIER
mouse reads (x10^6)
Naive
Primed


A
STAR alignment
HUMAN
Homo_sapiens.GRCh38.96.gtf
hg38
Calling variant in RNA-seq (GATK)
MARK duplicates
.bam
MRK.bam
SPLIT-N-CIGAR reads
MRK.bam
+ genome
SPLT_CIG.bam
ADD groups
SPLT_CIG.bam
GRP.bam
BASE QUALITY recalibration
GRP.bam
+ genome
+ BQSR table
+ dbsnp_146.hg38.vcf
BQSrecalibrated.bam
HAPLOTYPE caller
BQSrecalibrated.bam
+ genome
variants.vcf
intersection
VARIANTS filtration (hard-filtering)
43 Cancer Related SNPs (Tier 1) Avior Y et al., 2019
variants.vcf
+ genome
variantsFiltered.vcf
intersection
VARIANTS
B
HNES1
Cancer associated SNPs
0 2 4 6 8 10 12 14
MEF
LN
C
5
12
0
cR-S6EOS, MEF (p18/p26)
HNES1, MEF
D
SAHAKYAN
0.97
Mapped Reads (x10⁶) (Avior Y et al., 2019)
Mapped Reads (x10⁶) (this study)
DI STEFANO
0.99
Mapped Reads (x10⁶)
CHAN
0.86
SPERBER
0.98
GUO
0.94
LEE
0.82
TAKASHIMA
0.99
COLLIER
0.99
E
STAR alignment
HUMAN
Homo_sapiens.GRCh38.96.gtf
hg38
MOUSE
Mus_musculus.GRCm38.96.gtf
mm10
XenofilteR
filter out reads with higher mapping efficiency in mouse
Human alignment
Human alignment without mouse reads
GATK pipeline in RNAseq
43 Cancer Related SNPs (Tier 1)
Avior Y et al., 2019
Results
comparison
Results
F
% mouse reads
5 10 15 20 25
p=0.00033
naive primed


A
Avior et al., 2019
re-analysis
re-analysis with XenofilteR
COLLIER
DI STEFANO
GUO
TAKASHIMA
CHAN
LEE
SAHAKYAN
SPERBER
Number of cancer associated SNPs
Naive
Primed
B
R=0.84
% mouse reads
Number of cancer-associated SNPs
C
cR-S6EOS
HNES1
- XenofilteR
+ XenofilteR
total number of SNPs
MEF, p18
MEF, p26
LN
MEF
LN
D
human
mouse
Number of mismatches
CCND2 p.T151M
HIF1A p.A588P
APC p.A1171H
FAT1 p.G4434L
TP53 p.Q16K
TP53 p.T155S
TP53 p.E258K
BCL9L p.S1447T
MYH9 p.G37N
CDK12 p.E731K
TP53 p.R181H
TP53 p.R248Q
E
Chan
Collier
Di Stefano
Guo
Sahakyan
Sperber
Takashima
Human reference
Mouse reference
CCND2
F
t2iLGo/MEF
t2iLGo/LN
PXGL/MEF
PXGL/LN+GT
- XenofilteR
+ XenofilteR
cR-S6EOS p18
cR-S6EOS p26
HNES1
cR-S6EOS
HNES1
EPC niPSC
HDF16 niPSC
cR-H9
HNES1
HNES1
niPSCs
number of SNPs
min
max
Guo (2017)
This study
Bredenkamp (2019)


A
CTRL
Xen
APC
ARHGEF12
ARID1A
BAP1
BCL9L
CCND2
CDK12
CDKN1B
EGFR
FAT1
HIF1A
KDM5A
KLF4
MYH9
NFE2L2
PATZ1
PCM1
PDE4DIP
RNF213
SMARCA4
TP53
VHL
ZRSR2
CHAN
COLLIER
DI STEFANO
LEE
log2 expression
2 4 6 8 10
GUO
SPERBER
TAKASHIMA
SAHAKYAN


B
without Xenofilter-all expressed genes
with Xenofilter-all expressed genes
PC1
PC2
N P
Bredenkamp
Chan
Collier
Di Stefano
Guo
Lee
Sahakyan
Takashima
this study
Sperber


C
All naive
Number of cancer associated SNPs
Avior et al., 2019
re-analysis
re-analysis with XenofilteR

SUPPLEMENTAL LEGENDS

Supplement Figure 1

(A) Schematic of the pipeline used for the identification of SNPs from RNA-seq data.Variants were intersected with 43 cancer-related SNPs from Avior (2019) before and after application of hard-filtering.

(B) Total number of cancer-associated SNPs from Avior (2019) identified in HNES1 naïve cells on MEF or laminin substrates.

(C) Overlap between cancer associated SNPs from Avior et al (2019) identified in cR-S6EOS and HNES1 cells on MEF.

(D) Scatter plots of mapped reads in Avior et al. (2019) and this study.

(E) Schematic of the pipeline used for the identification of SNPs in RNA-seq data with and without removal of mouse reads by XenofilteR.

(F) Distribution of percentage of mouse reads for all naïve and primed hPSC samples.

Supplement Figure 2

(A) Heatmap with log2 expression value for cancer-associated genes in hPSCs before (CTRL) and after removal of mouse reads (XEN).

(B) PCA plots computed for all samples with all expressed protein coding genes. Left panel, before removal of mouse reads, right panel, after XenofilteR. N, naïve; P, primed as assigned by Avior et al. (2019). Purported naïve samples from Chan, Lee and Sperber align with conventional primed cells, as noted in previous analyses (Bredenkamp et al., 2019b; Takashima et al., 2014)

(C) Total number of cancer-associated SNPs detected in naïve hPSCs with different analyses.

Supplement Table 1

Summary of SNP analysis, showing datasets and samples analysed in this study and in Avior (2019). na denotes not analysed by Avior (2019). Note: cancer-related SNPs denote the 43 SNPs reported in Avior (2019)

Supplement Table 2

Table showing sample distribution of the 43 cancer-associated SNPs identified in Avior (2019) as determined in this study. The list of SNPs was downloaded from Supplemental Table 2 in Avior (2019) and include SNPs identified in hESC, iPSC or mesenchymal stromal cell (MSC) samples.

Supplemental Table 1

				SEAL		Without XenofilteR					With XenofilteR					Avior et al., 2019	
Dataset	Sample Name	Condition	Sequencing	Human	Mouse	Mapped	SNPs	SNPs-hardFiltering	Cancer-Related SNPs	Cancer-Related SNPs hardFiltering	Mapped	SNPs	SNPs-hardFiltering	Cancer-Related SNPs	Cancer-Related SNPs-hardFiltering	Mapped	SNPs
CHAN	hESC_3iL_1	Naïve	75-Single	44540426	5696945	41120847	303084	107987	11	7	37223900	59610	12194	1	1	44670511	11
	hESC_3iL_2	Naïve	75-Single	47357068	5740359	43414244	299886	106352	12	7	39456448	62156	13076	1	1	42380450	12
	hESC_3iL_3	Naïve	75-Single	47800369	1912424	42415874	133037	30271	4	1	39548240	61047	12672	1	1	43375823	1
	hESCs_1	Primed	75-Single	55996686	526523	49388963	80042	16573	1	1	46532612	69301	14593	1	1	50300583	1
	hESCs_2	Primed	75-Single	52813621	435231	46724102	77315	15775	1	1	44026776	66772	13902	1	1	46617640	1
	hESCs_3	Primed	75-Single	51898027	463113	45773732	78007	15659	1	1	43123745	66718	13687	1	1	47587086	1
COLLIER	Naive_Rep1	Naïve	150-Single	50630707	3442896	45676547	207768	55258	5	2	41604230	60085	13845	0	0	47756358	6
	Naive_Rep2	Naïve	150-Single	45191366	2852679	40647106	189211	47363	4	1	37320809	57819	13034	0	0	42530908	4
	Naive_Rep3	Naïve	150-Single	47458808	2386850	42531885	159923	38820	1	0	39356437	52220	12770	0	0	44489530	5
	Primed_Rep1	Primed	150-Single	41043477	208787	36754019	90169	16220	0	0	34223646	71285	13320	0	0	38047890	1
	Primed_Rep2	Primed	150-Single	54174033	304956	48686906	116882	20214	0	0	45577409	89762	16308	0	0	50411294	1
	Primed_Rep3	Primed	150-Single	44522799	201940	39493260	90261	15130	0	0	36945670	69550	12425	0	0	41035289	1
DI STEFANO	UCLA1_PD05_naive_1	Naïve	50-Paired	49206774	6196640	39063976	143393	23324	1	0	37235172	115968	20376	0	0	20225761	4
	UCLA1_PD1_naive_1	Naïve	50-Paired	53361716	4247006	40069814	134230	20676	0	0	38059340	117089	18945	0	0	20622182	0
	UCLA1_primed_1	Primed	50-Paired	60642694	1750370	48643956	161643	27093	0	0	47385046	160439	26934	0	0	24966532	0
	UCLA4_cob05_naive_1	Naïve	50-Paired	47394634	9163008	37391314	185056	26372	2	0	35093774	129548	17182	0	0	19541830	4
	UCLA4_PD05_naive_1	Naïve	50-Paired	45556554	4514046	36322056	172813	20138	1	0	34801292	152227	17942	0	0	18730172	2
	UCLA4_PD1_naive_1	Naïve	50-Paired	49176300	4789186	40172708	173746	23848	1	0	38873992	151135	21427	0	0	20681374	4
	UCLA4_primed_1	Primed	50-Paired	54683322	1791358	43179294	141579	23631	0	0	41784032	140464	23487	0	0	22052878	0
	UCLA4_Ref05_naive_1	Naïve	50-Paired	46027438	6253732	35708912	158866	21649	1	0	33889548	128893	17747	0	0	18462927	3
	UCLA4_t2iLGoY_naive_1	Naïve	50-Paired	47245652	6863726	37573798	184472	24519	3	1	35191578	136778	16859	0	0	19496963	2
	UCLA4_TAK05_naive_1	Naïve	50-Paired	54719410	7443720	44714646	226567	27842	2	0	42529038	181802	20592	0	0	23154031	4
	UCLA5_PD05_naive_1	Naïve	50-Paired	55956254	3480200	44074134	146379	20822	0	0	42676226	135410	19867	0	0	22735731	0
	UCLA5_PD1_naive_1	Naïve	50-Paired	45942896	4997092	35685660	129200	15772	1	0	34236474	116418	14815	0	0	18395337	1
	UCLA5_primed_1	Primed	50-Paired	51173974	1621878	40204430	147230	23616	0	0	39089110	146345	23450	0	0	20627247	0
	UCLA9_PD05_naive_1	Naïve	50-Paired	53207974	3665828	40910088	144065	21688	1	0	39202010	125237	19602	0	0	20985723	1
	UCLA9_PD1_naive_1	Naïve	50-Paired	53049840	3574380	40668666	149965	19819	1	0	39050742	134375	18236	0	0	20850033	2
	UCLA9_primed_1	Primed	50-Paired	51022394	1541090	40239150	151322	22341	0	0	39032464	150316	22179	0	0	20513429	0
	WIBR3_PD05_naive_1	Naïve	50-Paired	56916542	6913178	44310118	162888	25856	1	0	42225896	128486	21184	0	0	22966199	1
	WIBR3_PD1_naive_1	Naïve	50-Paired	54434762	4820978	43162864	161395	25437	0	0	41350338	131255	21796	0	0	22235035	1
	WIBR3_primed_1	Primed	50-Paired	54003506	1550922	43050376	156249	23829	0	0	41867940	154953	23678	0	0	21896548	0
GUO (2017)	cR_S6EOS_laminin_R1	Naïve	50-Single	70324204	1241532	59959551	68001	1111	0	0	59000074	65124	1047	0	0	na	na
	cR_S6EOS_laminin_R2	Naïve	50-Single	62529082	1082558	53527112	66178	1141	0	0	52702701	63163	1067	0	0	na	na
	cR_S6EOS_laminin_R3	Naïve	50-Single	60094062	1075376	51287521	64571	1255	0	0	50491942	61894	1175	0	0	na	na
	cR_S6EOS_p18_R1	Naïve	50-Single	32228691	23051237	31979065	322318	28444	11	1	26485700	34382	801	0	0	na	na
	cR_S6EOS_p18_R2	Naïve	50-Single	26478991	19035697	26503587	296891	21295	12	1	21975020	30500	557	0	0	30369314	12
	cR_S6EOS_p18_R3	Naïve	50-Single	37955244	795562	32974317	50715	926	0	0	32320187	48563	868	0	0	33229012	0
	cR_S6EOS_p26_R1	Naïve	50-Single	39279725	8748813	35864599	176854	13449	6	1	33532383	49038	1014	0	0	37558713	8
	cR_S6EOS_p26_R2	Naïve	50-Single	34067340	7756566	31095843	159176	10995	6	0	29028930	42290	795	0	0	32590948	7
	cR_S6EOS_p26_R3	Naïve	50-Single	33018807	6952144	30080969	152079	9747	6	1	28160544	43958	806	0	0	31428230	4
	HNES1_laminin_R1	Naïve	50-Single	68126457	1092303	57865742	62552	1353	0	0	56939236	59811	1274	0	0	na	na
	HNES1_laminin_R2	Naïve	50-Single	63958220	1216971	53378329	58523	1186	0	0	52263300	55882	1112	0	0	na	na
	HNES1_laminin_R3	Naïve	50-Single	57595383	1017371	48851058	59410	1177	0	0	48060941	56897	1129	0	0	na	na
	HNES1_R1	Naïve	50-Single	28884855	12962664	27538051	250592	16885	11	1	24379978	39369	680	0	0	na	na
	HNES1_R2	Naïve	50-Single	35010481	15327606	33306748	276404	21826	11	1	29587438	46661	865	0	0	na	na
	HNES1_R3	Naïve	50-Single	34674257	15254518	32829139	266788	18971	6	1	29146195	44967	758	0	0	na	na
	S6EOS_R1	Primed	50-Single	32949037	4507555	29139234	98924	5475	1	0	27761321	38661	790	0	0	30105365	5
	S6EOS_R2	Primed	50-Single	29159652	3642432	25729128	85619	4263	1	0	24602951	35225	674	0	0	26550319	2
	S6EOS_R3	Primed	50-Single	31714050	613262	27623031	46431	643	0	0	27140890	44382	605	0	0	27843468	0
LEE	hPSC_naive_rep1_1	Naïve	70-Single	23767439	366802	20783225	66818	6467	0	0	20093188	54392	5234	0	0	20946688	0
	hPSC_naive_rep1_2	Naïve	70-Single	24890455	385981	21772800	69358	6792	0	0	21103776	56043	5483	0	0	18494786	1
	hPSC_naive_rep2_1	Naïve	70-Single	19010754	294657	16620769	58174	5104	0	0	16078272	47594	4102	0	0	16752464	2
	hPSC_naive_rep2_2	Naïve	70-Single	19832433	308552	17346166	59777	5402	1	0	16819991	48758	4314	1	1	17476647	2
	hPSC_primed_rep1_1	Primed	70-Single	19018923	167446	16876074	50965	4922	1	1	16338350	46409	4372	1	1	16993418	2
	hPSC_primed_rep1_2	Primed	70-Single	19634443	172914	17416394	51656	5087	1	1	16916032	46946	4512	1	1	17535853	3
	hPSC_primed_rep2_1	Primed	70-Single	22921984	219104	20296540	56734	5859	1	0	19657901	51628	5252	1	1	20438502	2
	hPSC_primed_rep2_2	Primed	70-Single	23661220	226968	20946678	57684	6038	1	1	20347482	52386	5394	1	1	21088380	2
SAHAKYAN	UCLA1_Clone12_XIST_Pos_Rep1	Naïve	50-Single	35070242	12440999	33417203	303934	49375	12	1	29158857	37063	2496	0	0	36530621	9
	UCLA1_Clone12_XIST_Pos_Rep2	Naïve	50-Single	41722611	2794826	35570590	87185	7769	4	0	33569645	28638	2678	0	0	36817012	4
	UCLA1_Clone4_LatePassage_XIST_Pos	Naïve	50-Single	40275437	3095896	35458329	113497	8625	2	0	34204594	39367	2298	0	0	36229723	2
	UCLA1_Clone4_XIST_Neg_Rep1	Naïve	50-Single	33707418	10436850	32966915	274977	53940	11	2	29209331	33859	2380	1	0	35471787	13
	UCLA1_Clone4_XIST_Neg_Rep2	Naïve	50-Single	45734485	3278137	40956812	100975	11508	4	1	38748161	45106	3357	0	0	42070671	2
	UCLA1_Clone9_XIST_Pos_Rep1	Naïve	50-Single	22303036	26345413	25790607	481455	80217	15	2	18457920	27204	1354	0	0	31631824	16
	UCLA1_Clone9_XIST_Pos_Rep2	Naïve	50-Single	54748785	3063248	46521690	89320	8727	4	0	44037159	32167	3284	0	0	48078388	4
	UCLA1_Primed_Rep1	Primed	50-Single	27823436	3530342	25054152	114135	12239	3	0	23232549	29002	1543	0	0	26116987	4
	UCLA1_Primed_Rep2	Primed	50-Single	33550834	3762642	28708228	99136	13136	3	0	26694641	23298	1890	0	0	29923369	5
	UCLA1_Reprimed	Primed	50-Single	36035801	2076347	30979411	71001	5507	2	0	30057499	34792	1923	0	0	31479390	2
SPERBER	H1_4iLIF_rep1	Naïve	75-Single	52772165	15447557	52636975	480108	208721	15	8	45170260	52528	9329	0	0	54419471	13
	H1_4iLIF_rep2	Naïve	75-Single	48951468	14236265	48285856	460174	188272	14	9	42061030	47961	8656	0	0	49889697	10
	H1_rep1	Primed	100-Paired	52131568	610882	38181590	489824	36374	1	1	37226724	448121	34967	0	0	19252993	1
TAKASHIMA	H9_R1	Primed	100-Paired	42841158	5268024	34098988	346073	31714	0	0	31585186	299998	30200	0	0	18133411	6
	H9_R2	Primed	100-Paired	27655616	3658574	21435008	270694	19855	0	0	19757438	230394	18837	0	0	11539700	4
	H9_R3	Primed	100-Paired	42777274	3951036	34952954	407640	32989	0	0	32711134	353228	31264	0	0	18484742	5
	H9_reset_R1	Naïve	100-Paired	117642884	35033856	96061254	856660	152305	7	5	84634350	574314	67463	0	0	52754813	16
	H9_reset_R2	Naïve	100-Paired	343781150	86522164	256111320	1251189	249887	8	3	223784684	889372	138353	0	0	141321481	12
	H9_reset_R3	Naïve	100-Paired	281143414	84733026	225716650	1353177	278354	5	2	194933258	946944	154608	0	0	na	na
BREDENKAMP(2019)	HNES1(LN+GT)	Naïve	75-Paired	18596056	140036	16199188	93764	9204	0	0	14959196	96464	8762	0	0	na	na
	niPS_C2(LN+GT)	Naïve	75-Paired	19025232	566786	16460565	94479	8565	0	0	15180521	97900	8342	0	0	na	na
	niPS_C4(LN+GT)	Naïve	75-Paired	17110842	134776	14871279	87234	7907	0	0	13696599	90653	7742	0	0	na	na
GUO (this study)	EPC niPSC, PXGL/MEF, rep1	Naïve	100-Paired	44464406	4250786	41447192	332633	48281	4	3	40959938	286176	37680	0	0	na	na
	EPC niPSC, PXGL/MEF, rep2	Naïve	100-Paired	52263284	8147020	49070220	373803	64469	3	1	48234946	294278	40472	0	0	na	na
	HDF16 niPSC,PXGL/MEF, rep1	Naïve	100-Paired	46776824	4420114	43518900	333163	49916	3	1	43011184	283915	38343	0	0	na	na
	HDF16 niPSC,PXGL/MEF, rep2	Naïve	100-Paired	57913750	3991800	54064608	328668	52741	1	0	53499818	286861	43623	0	0	na	na
	cR-H9, PXGL/MEF, rep1	Naïve	100-Paired	89946474	1906762	83438534	414999	62161	1	0	82616854	385511	58498	0	0	na	na
	cR-H9, PXGL/MEF, rep2	Naïve	100-Paired	77241044	3807188	71905780	415901	62602	3	2	70973696	368571	53822	0	0	na	na
	HNES1, PXGL/MEF, rep1	Naïve	100-Paired	84562038	1767776	78661938	428639	60078	1	1	77877670	399487	56735	0	0	na	na
	HNES1, PXGL/MEF, rep2	Naïve	100-Paired	100986994	4431498	94863452	557358	82757	1	1	93901890	508903	74089	0	0	na	na