Defining the molecular state of human cancer

Biswajyoti Sahu[1,2,3], Päivi Pihlajamaa[1,3], Kaiyang Zhang[4], Kimmo Palin[1,2], Saija Ahonen[1,2], Alejandra Cervera [4], Lauri A. Aaltonen[1,2], Sampsa Hautaniemi[4] and Jussi Taipale[1,3,5]*

1. *Applied Tumor Genomics Research Program, Faculty of Medicine, University of Helsinki, Finland*
2. *Medicum, Faculty of Medicine, University of Helsinki, Finland.*
3. *Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom*
4. *Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Finland*
5. *Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden*

* *Corresponding author: Jussi Taipale (ajt208@cam.ac.uk)*

Cancer is the most complex genetic disease known, with mutations in more than 250 genes contributing to different forms of the disease[1,2,3]. Most human driver mutations are specific to particular types of cancer, at least in part due to differences in expression pattern between cell types, and the diversity of mutational mechanisms across different human tissues[4]. However, the fact that many apparently oncogenic mutations fail to transform fibroblastic cells in culture suggests that different cell types could be susceptible to transformation by different sets of oncogenes. Here we show that reprogramming human fibroblasts to induced hepatocytes (iHeps) makes the cells sensitive to transformation by a combination of oncogenes that is characteristic of liver cancer (CTNNB1, TERT and MYC). The transformed iHeps are highly proliferative, tumorigenic in nude mice, and bear gene expression signatures of liver cancer. Temporal analysis of the tumorigenic program using single-cell RNA-seq and RNA velocity analysis revealed that the cells progress along a common path to transformation, invariably acquiring liver cell identity prior to expressing markers characteristic of liver tumor cells. These results, together with analysis of chromatin accessibility using ATAC-seq and NaNoMe-seq indicate that lineage-determining factors act by defining a chromatin state that is permissive for transformation. Taken together, our results indicate that cell identity is a key determinant in transformation, and establish a paradigm for defining the molecular states of distinct types of human cancer.

Cancer genetics and genomics have identified a large number of genes implicated in human cancer[1,2,3]. Although some genes such as *p53* and *PTEN* are commonly mutated in many different types of cancer, most cancer genes are lineage-specific. It is well established that human cells are harder to transform than rodent cells[5,6,7,8,9,10,11], which can be transformed using only MYC and RAS oncogenes[12,13,14]. Seminal experiments by Hahn and Weinberg established already 20 years ago that different human cell types can be transformed

using a set of oncogenes that includes the powerful viral large-T and small-T oncoproteins from the SV40 virus[15]. Despite this early major advance, determining which specific mutations found in human patients lead to tumorigenesis has proven to be exceptionally difficult. This is because although viral oncoproteins are linked to several cancer types[16], most major forms of human cancer result from mutations affecting tumor-type specific sets of endogenous proto-oncogenes and tumor-suppressors. The fact that the combinations of oncogenes are distinct between tumor types suggests that cell lineage-specific factors could somehow interact with oncogenes to drive most cases of human cancer, confounding mechanistic studies utilizing simple model cell types. This prompted us to systematically investigate the factors required for transformation of human cells using a combination of cell fate conversion and oncogene activation.

Many human cell types can be converted to other cell types via a pluripotent state[17]. However, as pluripotent cells are tumorigenic in nude mice, we chose to use direct lineage conversion[18,19,20] in combination with oncogene expression to identify the set of factors that define a particular type of human cancer cell. For this purpose, we developed a cellular transformation assay protocol, in which human fibroblasts (HF) are converted to induced hepatocytes (iHeps) using lentiviral overexpression of a combination of lineage-specific transcription factors (TF), followed by ectopic expression of liver cancer-specific oncogenes (**Fig. 1a**). Transdifferentiation of fibroblasts to iHeps has previously been reported by several groups[20,21,22,23]. To identify an optimal protocol for generating iHeps from HFs (from human foreskin), we tested the previously reported combinations of TFs in parallel transdifferentiation experiments and analyzed the efficiency of iHep conversion by measuring the mRNA levels for liver markers[21,22,23] such as *ALBUMIN, TRANSFERRIN,* and *SERPINA1* at different time points (**Fig. 1b**, **Extended Data Fig. 1**). The combination of three TFs, HNF1A, HNF4A and FOXA3[22] resulted in the most efficient iHep generation, based on the observation that out of all combinations tested, this combination resulted in the

highest expression level of liver-specific genes at two, three, and four weeks after iHep induction (**Fig. 1b**). This protocol also resulted in most efficient lineage conversion based on the analysis of cell morphology; by two weeks after iHep induction, the cells lost their fibroblast phenotype and formed iHep colonies, from which the iHeps migrated and matured by six to seven weeks after induction (**Fig. 1c** and **Extended Data Fig. 2**).

To determine whether the iHeps could be transformed to liver cancer-like cells, we first plated the iHeps on collagen-coated dishes and maintained them in hepatocyte culture media (HCM). The proliferation of iHeps under such conditions is arrested[22] and the cells undergo apoptosis after two to three passages (**Fig. 1d**). To confer iHeps with unlimited proliferation potential and to drive them towards tumorigenesis, we transduced iHeps with a set of the most common driver genes for liver cancer using lentiviral constructs. For this purpose, we chose the five oncogenic drivers with the highest number of recurrent genetic alterations reported for liver cancer or hepatocellular carcinoma (HCC; from COSMIC, https://cancer.sanger.ac.uk/cosmic); these included four oncogenes, telomerase (*TERT*), β-catenin (*CTNNB1*), PI3 kinase (*PIK3CA*), and the transcription factor NRF2 (*NFE2L2*), as well as one tumor suppressor, p53 (*TP53*). In addition, we included the oncogene *MYC*, which is under tight control in normal cells[24], but overexpressed in many cancer types, including HCC[25]. Lentiviral expression of the fluorescent reporter mCherry with the oncogenic drivers in different combinations revealed that the pool of three oncogenes, *i.e.* constitutively active β-catenin (CTNNB1^{T41A}), MYC and TERT, together with *TP53* inactivation by CRISPR-Cas9 (CMT+sg*TP53*) resulted in highly proliferative iHeps with apparently unlimited proliferative potential (> 50 passages over more than one year; **Fig. 1d**). Importantly, expression of the three oncogenes CTNNB1^{T41A}, MYC and TERT (CMT) alone also resulted in similar iHeps with long-term proliferative potential (**Fig. 1d**). By contrast, ectopic expression of these oncogenic drivers in HFs failed to yield transformed, proliferating fibroblasts (**Fig. 1d**). This is the first instance to our knowledge where HFs can be directly

transformed using this minimal combination of defined factors, indicating that lineage-specific TFs are the missing link for human cellular transformation using oncogenic drivers.

To test for the tumorigenicity of the proliferative iHeps, we performed xenograft experiments. Subcutaneous injection of the CMT+sg*TP53* transformed iHeps into nude mice resulted in tumor formation (**Fig. 2a**). The process was reproducible in subsequent experiments; in addition, the effect was not specific to the fibroblast line used, as we also successfully reprogrammed another HF cell line (human fetal lung fibroblast) using the same lineage-specific TFs and oncogenic drivers. The xenograft tumors from the CMT+sg*TP53* transformed iHeps derived from either fibroblast line can be detected by *in vivo* fluorescent imaging as early as 11-12 weeks (**Fig. 2b**). Similarly, the CMT-transformed iHeps without *TP53* inactivation also resulted in tumor formation in nude mice 12 weeks post-injection (**Fig. 2b**). These results demonstrate that both CMT and CMT+sg*TP53* transformed iHeps are tumorigenic, and indicate that ectopic expression of defined lineage-specific TFs and oncogenes can reprogram and transform HFs into cells that can robustly initiate tumors in nude mice.

Cancer genomes harbor large-scale chromosomal aberrations and are characterized by aneuploidy[26,27]. To understand the gross chromosomal aberrations in the transformed tumorigenic CMT and CMT+sg*TP53* iHeps compared to normal HFs, we performed spectral karyotyping, which showed a normal diploid male (46, XY) in HFs and aneuploid karyotypes in transformed iHeps (**Fig. 2c**). The aneuploid transformed iHeps with CMT+sg*TP53* at early passage were characterized by two different populations with two distinct modal chromosome numbers (**Fig. 2c**). The modal chromosome number of the first population was 45, XY, whereas the second population was pseudotetraploid, with a modal chromosome number between 67-92, XY; this pseudotetraploid state was consistently observed in late passage transformed iHeps. The major chromosomal aberrations that were similar between the two populations were missing copies of chromosomes 4 and 13, a derivative of

chromosome 19 containing a small portion of chromosome 3 [t3:19], an extra copy of Y and a loss of most of the p arm of chromosome 2. In comparison, the most common chromosomal aberrations reported in HCC are the gains of 1q (suggested target genes include *WNT14, FASL*) and 8q (*MYC, WISP1*) and the loss of 17p (*TP53, HIC1*), followed by losses of 4q (*LEF1, CCNA*) and 13q (*RB1, BRCA3*)[28,29] (**Fig. 2d**). The first three chromosomal aberrations are expected not to be present in our case, as the transformation protocol leads to activation of the Wnt pathway and MYC expression, and loss of p53. Consistently with this, we did not observe lesions in 1q, 8q or 17p in our cells. However, other common aberrations found in HCC cells, loss of chromosomes 4 and 13 were detected in our transformed CMT+sg*TP53* iHep cells (**Fig. 2c-d**). However, these chromosomal aberrations appeared not to be necessary for formation of tumors, as in the absence of targeted loss of p53 in CMT iHep cells, we did not observe these lesions (**Fig. 2c**). However, both CMT+sg*TP53* and CMT iHeps displayed pseudotetraploidy, similar to what is commonly observed in HCC (**Fig. 2c-d**). These results indicate that the transformed iHeps have similar chromosomal aberrations to those reported earlier in liver cancer, consistent with their identity as HCC-like cells.

To understand the gene expression dynamics and to map the early events of lineage conversion and oncogenic transformation, we performed single cell RNA-sequencing (scRNA-seq) of HFs, iHeps after one, two, and three weeks after induction, and from CMT-iHeps (one-week iHeps transduced with CMT and harvested two weeks later). The cells were clustered according to their expression profiles using Seurat[30] (version 2.3.4); a total of ten separate clusters of cells were identified during the course of the transdifferentiation and reprogramming and visualized by t-distributed stochastic neighbor embedding (t-SNE) plots[31] (**Fig. 3a-b**). Importantly, the scRNA-seq indicated that the CMT-transformed iHeps are a clearly distinct population of cells compared to the iHeps (**Fig. 3b**).

To determine the trajectory of differentiation of the cells, we performed RNA velocity analysis[32], which determines the direction of differentiation of individual cells based on

comparison of levels of spliced mRNAs (current state) with nascent unspliced mRNAs (representative of future state). This analysis confirmed that the cell populations analyzed were differentiating along the fibroblasts–iHep–transformed iHep axis (**Fig. 3c**). We next identified marker genes for each cell cluster (see **Methods**). This analysis revealed that CMT-iHeps have a distinct gene expression signature and that they have lost the fibroblast gene expression program during the course of the reprogramming. These results indicate that the iHep conversion and transformation have led to generation of liver-cell like transformed cells (**Fig. 3d**).

To further analyze gene expression changes during reprogramming and transformation, we performed pseudo-temporal ordering analysis of the scRNA-seq. Consistently with the RNA velocity analysis, the pseudotime analysis showed transition from fibroblasts to iHeps and subsequently to CMT-transformed iHeps (**Extended Data Fig. 3**). The scRNA-seq analyses allow detection of the precise early events that occur during iHep formation and the origin of HCC by mapping the gene expression changes in the cells across the pseudotime. During iHep differentiation, the expression of non-canonical Wnt pathway components, including Wnt5a ligand and the Frizzled 5 receptor, are upregulated (**Fig. 3e**). By contrast, during transformation, the exogenous CTNNB1^{T41A} activates the canonical Wnt pathway, suppressing expression of the non-canonical ligand Wnt5a. We also observe activation of the NOTCH pathway early during tumorigenesis; expression of *NOTCH1*, *NOTCH3* and their ligand *JAG1* (**Fig. 3e**, top) are strongly upregulated, together with the canonical NOTCH target gene *HES1*[33] and the liver specific target *NR4A2*[34]. These results are consistent with the proposed role of the NOTCH pathway in liver tumorigenesis[34,35].

To determine whether the gene expression signatures observed in transformed iHeps were similar to those observed in human liver tumors, we compared the scRNA-seq results to the published liver cancer data sets [28]. Majority of the CMT-iHep-specific marker genes (**Fig. 3d**) overlapped with the genes with genetic alterations in the TCGA HCC pan-cancer dataset

(74% of 372 cancer cases) and showed larger overlap with HCC when compared to cancers of pancreas and prostate, suggesting the specificity of this set of genes for liver tumorigenesis (**Extended Data Fig. 4**). We also analyzed the expression of the CMT-iHep marker genes that show genetic alterations in TCGA liver cancer data across the pseudotime in our scRNA-seq data. The expression of this subset of the CMT-iHep marker genes was also clearly increased, lending further credence to the fact that upregulation of this set of genes is an early event in liver tumorigenesis (**Extended Data Fig. 5**).

To determine the changes in gene expression and chromatin accessibility in the proliferative iHeps, we first performed bulk RNA-seq analysis from the tumorigenic CMT and CMT+sg*TP53* iHeps that were used for the xenograft implantation, as well as cells derived from the resulting tumors. Importantly, the genes that were differentially expressed in both CMT- and CMT+sg*TP53*-transformed iHeps compared to fibroblasts showed a clear and significant positive enrichment for the previously reported "subclass 2" liver cancer signature[36], associated with proliferation and activation of the MYC and AKT signaling pathways (**Fig. 3f**). The effect was specific to liver cancer, as we did not observe significant enrichment of gene expression signatures of other cancer types (**Extended Data Fig. 6**). During the reprogramming, we observed a clear up-regulation of common liver marker genes such as *ALB*, *APOA2*, *SERPINA1*, and *TF*, and down-regulation of fibroblast markers such as *MMP3*, *FGF7*, *THY1*, and *FAP*, in proliferative and tumorigenic iHeps. Importantly, the xenograft tumor from the CMT+sg*TP53* cells retained similar liver-specific gene expression profile (**Fig. 4a**). We also detected a clear up-regulation of several liver cancer marker genes such as *AFP, GPC3, SAA1,* and *VIL1* in transformed iHeps and in CMT+sg*TP53* tumors compared to control fibroblasts (**Fig. 4a**); *AFP* was also found among the most enriched genes (**Extended Data Fig. 7**) in both CMT+sg*TP53*- and CMT-transformed iHeps. Furthermore, we observed a negative correlation between the CMT+sg*TP53* and CMT iHep specific genes and the genes positively associated with liver cancer survival (**Extended Data**

Fig. 8), lending further credence to liver cancer-identity of the CMT+sg*TP53* and CMT transformed iHeps.

ATAC-seq analysis of the fibroblasts and CMT+sg*TP53* cells revealed that the changes in marker gene expression were accompanied with robust changes in chromatin accessibility at the corresponding loci (**Fig. 4b**). To assess chromatin accessibility and DNA methylation at a single-allele level, we performed NaNoMe-seq (see **Methods**), where accessible chromatin is methylated at GpC dinucleotides using the bacterial methylase M.CviPI[37]. Sequencing of the genome of the treated cells using single-molecule Nanopore sequencer then allows both detection of chromatin accessibility (based on the presence of methylated cytosines at GC dinucleotides) and DNA methylation at CG dinucleotides. This analysis confirmed the changes in DNA accessibility detected using ATAC-seq (**Fig. 4c**). Changes in DNA methylation at promoters of the differentially expressed genes were relatively minor (**Fig. 4c**), suggesting that the mechanism of reprogramming does not critically depend on changes in CpG methylation at the marker loci. Taken together, these results indicate that our novel cell transformation assay using lineage-specific TFs and cancer-specific oncogenes can reprogram fibroblasts to lineage-specific cancer that bears a gene expression signature similar to that observed in HCC.

To identify the necessary and sufficient factors that define lineage-specific cancer types we have here developed a novel cellular transformation protocol, and, for the first time, report direct conversion of HFs to liver cancer cells. First, lentiviral overexpression of three lineage-specific TFs reprograms HFs to iHeps, and subsequent ectopic expression of liver cancer-specific oncogenic factors transforms iHeps to a highly proliferative and tumorigenic phenotype with chromosomal aberrations and gene expression signature patterns similar to HCC. Importantly, lineage-conversion by specific TFs is required for the transformation process since the same oncogenic drivers alone do not transform HFs (**Fig. 4d**). After lineage conversion by the defined TFs, oncogenes alone (MYC, CTNNB1 and TERT) are sufficient

to drive the transformation with or without inactivation of the tumor suppressor *TP53*. These results establish a paradigm for testing the tumorigenicity of combinations of cancer genes, and their interactions with cellular lineage (**Fig 4d**). In addition, reprogramming normal cells to cancer cells allow "live" analysis of the early stages of the tumorigenic program, facilitating approaches towards early molecular detection and prevention of cancer.

In the past half-century, a very large number of genetic and genomic studies have been conducted using increasingly powerful technologies, resulting in identification of more than 250 genes that are recurrently mutated in cancer. However, in most cases, the evidence that the mutations in the genes actually cause cancer is correlative in nature, and requires assumptions about background mutation frequency and rates of clonal selection in normal tissues[38]. Furthermore, cancer genes are known to act in combination, and determining candidate sets of genes that are sufficient to cause cancer using genetic data alone would require astronomical sample sizes. Mechanistic studies are thus critical for conclusively determining that a particular gene is essential for cancer formation, and for identification of sets of genes that are sufficient for tumorigenesis. In principle, individual driver genes and their combinations could be identified and validated using particular primary cell types. Previously using primary cells, particular combinations of oncogenes that can transform specific types of human cells including colon, pancreatic, prostate and lung epithelial cells have been identified[39,40,41]. Our approach allows more precise control of cell identity, facilitating analysis of interactions between lineage-determining factors and oncogenes. In addition, approaches using primary cells are severely limited by the fact that for most tissues, sufficient amounts of live human tissue material are hard to obtain. Furthermore, the cell type of origin for most cancer types is not known, and it is commonly assumed that tumors originate from rare and hard-to-isolate subpopulations of cells (e.g. stem cells, or transient progenitor cells in the case of pediatric tumors). Our results using the novel cellular transformation assay show that HFs can be directly converted to lineage-specific cancer.

Using this assay, we were able to determine the minimum events necessary for making human liver cancer in culture. By using lineage-specific TFs to generate the cell type of interest for transformation studies, our molecular approach can be generalized for identifying minimal determinants of any cancer type, paving the way towards elucidating the exact molecular mechanisms by which specific combinations of mutations cause particular types of human cancer.

Methods

Plasmids and lentiviral production

Full-length coding sequences for the TFs and oncogenes were obtained from GenScript and cloned into the lentiviral expression vector pLenti6/V5-DEST using the Gateway recombination system (Thermo Fisher Scientific). Expression construct for mCherry (#36084), lentiviral Cas9 expression construct LentiCas9-Blast (#52962) and a cloning backbone lentiGuide-Puro (#52963) were obtained from Addgene, and the six pairs of single-stranded oligos corresponding to the guide sequences targeting the *TP53* gene in the GeCKO library were ordered from IDT, annealed, and ligated into lentiGuide-Puro backbone[42]. For virus production, the plasmids were co-transfected with the packaging plasmids psPAX2 and pMD2.G (Addgene #12260 and #12259, respectively) into 293FT cells (Thermo Fisher Scientific) with Lipofectamine 2000 (Thermo Fisher Scientific). Fresh culture medium was replenished on the following day, and the virus-containing medium was collected after 48 h. The lentiviral stocks were concentrated using Lenti-X concentrator (Clontech) and stored as single-use aliquots.

Cell lines and generation of iHeps

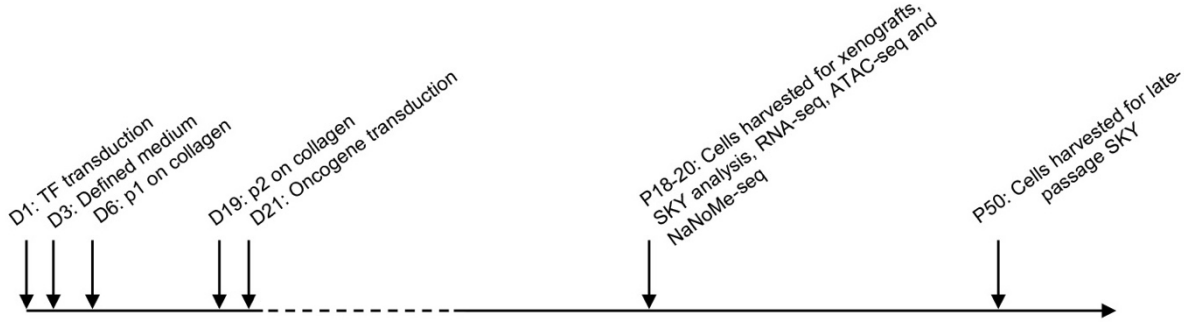
Human foreskin fibroblasts (HFF, CCD-1112Sk) and human fetal lung (HFL) fibroblasts were obtained from ATCC (#CRL-2429 and #CCL-153, respectively) and

cultured in fibroblast medium (DMEM supplemented with 10% FBS and antibiotics, Thermo Fisher Scientific). LentiCas9-Blast virus was transduced to early-passage fibroblasts (MOI = 1) with 8 µg/ml polybrene. Blasticidin selection 4 µg/ml was started two days after transduction and continued for two weeks. Early passage blasticidin-resistant cells were used in the reprogramming experiments by transducing cells with constructs for TF expression in combinations reported earlier by Morris et al. (FOXA1, HNF4A, KLF5)[23], Du et al. (HNF4A, HNF1A, HNF6, ATF5, PROX1, CEBPA)[21] and Huang et al. (FOXA3, HNF4A, HNF1A)[22] with MOI = 0.5 for each factor and 8 µg/ml polybrene (day 1). The medium was changed to fresh fibroblast medium containing β-mercaptoethanol on day 2 and to a defined hepatocyte growth medium (HCM, Lonza) on day 3. On day 6, the cells were passaged on plates coated with type I collagen (Sigma) in several technical replicates, and thereafter, the HCM was replenished every two–three days.

Generation of HCC-like cells

The iHeps generated using the three TFs (FOXA3, HNF4A, HNF1A) were passaged on type I collagen-coated plates on day 19 after iHep induction (p2) in HCM and transduced with different combinations of lentiviral constructs encoding the oncogenes (CTNNB1, MYC, TERT) on day 21 (MOI = 1 for each factor with 8 µg/ml polybrene). For CMT+sg*TP53* condition, the oncogenes were transduced along with a pool of six sgRNAs targeting the *TP53* gene. Fresh HCM was replenished on the day following the transduction, cells were maintained in HCM, and passaged when close to confluent. From fifth passaging onwards after oncogene induction, cells were maintained in HCM supplemented with 1% defined FBS (Thermo Fisher Scientific). For single-cell RNA-sequencing experiments, the iHeps were transduced with CMT oncogenes (MOI = 1 with 8 µg/ml polybrene) on day 8 with fresh HCM replenished on day 9, and the cells were harvested for single-cell RNA-

sequencing at the indicated time points from replicate culture wells. In all experiments, viral construct for mCherry expression was co-transduced with the oncogenes.



Xenografts

Oncogene-induced CMT and CMT+sg*TP53* cells were harvested at p20, 10^7 cells were resuspended in HCM supplemented with 1% defined FBS and mixed with equal volume of Matrigel (growth factor reduced basement membrane matrix, Corning #356231) and injected subcutaneously into the flank of a 6-week old immunodeficient BALB/c nude male mice (Scanbur). *In vivo* imaging of the tumors was performed for the mice under isoflurane anesthesia using the Lago system (Spectral Instruments Imaging). Photon counts from the mCherry were detected with fluorescence filters 570/630 nm and superimposed on a photographic image of the mice. Tumors were harvested 23-25 weeks after injection. All the experiments were performed according to the guidelines for animal experiments at the University of Helsinki and under license from appropriate Finnish Review Board for Animal Experiments.

SKY analysis

Spectral karyotype analysis was performed at Roswell Park Cancer Institute Pathology Resource Network. Cells were treated for 3 hours with 0.06 µg/ml of colcemid, harvested and fixed with 3:1 methanol and acetic acid. Metaphase spreads from fixed cells were hybridized with SKY probe (Applied Spectral Imaging) for 36 hours at 37 degrees

Celsius. Slides were prepared for imaging using CAD antibody kit (Applied Spectral Imaging) and counterstained with DAPI. Twenty metaphase spreads for each cell line were captured and analyzed using HiSKY software (Applied Spectral Imaging).

RNA isolation, qPCR and bulk RNA-sequencing

Total RNA was isolated from the control fibroblasts, iHeps harvested at day 5 and at weeks two, three, and four, CMT and CMT+sg*TP53* cells harvested at p20, and from tumor tissues stored in RNALater (Qiagen), using RNeasy Mini kit (Qiagen) with on-column DNase I treatment. For qRT-PCR analysis, cDNA synthesis from two biological replicates was performed using the Transcriptor High-fidelity cDNA synthesis kit (Roche) and real-time PCR using SYBR green (Roche) with primers specific for each transcript (**Extended Data Table 1**). The Ct values for the target genes were normalized to those of GAPDH, and the mean values of sample replicates were shown for different conditions at the indicated time points. RNA-sequencing was performed from three biological replicate samples for each condition, using 400 ng of total RNA from each sample for poly(A) mRNA capture followed by stranded mRNA-seq library construction using KAPA stranded mRNA-seq kit for Illumina (Roche) as per manufacturer's instruction. Final libraries with different sample indices were pooled in equimolar ratios based on quantification using KAPA library quantification kit for Illumina platforms (Roche) and size analysis on Fragment Analyzer (AATI) and sequenced on HiSeq 4000 (Illumina).

For preprocessing and analysis of the RNA-Seq reads the SePIA pipeline[43] based on the Anduril framework[44] was used. Quality metrics from the raw reads were estimated with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and Trimmomatic[45] clipped adaptors and low-quality bases. After trimming, reads shorter than 20bp were discarded. Kallisto (v0.44.0) with Ensembl v85[46] was used for quantification followed by tximport[47] and DESeq2[48] (v1.18.1) for differential expression calculating log2(fold change)

and standard error from triplicate samples. Gene set enrichment analysis[49] was performed using GSEAPY (version 0.9.8) by ranking differentially expressed genes based on their -log10(p-value)*sign(fold-change) as metric. The gene signatures analysed for enrichment were collected from Molecular Signatures Database (MSigDB, version 6.2).

Single-cell RNA-sequencing

For single cell RNA-sequencing (scRNA-seq), iHeps at different time points were harvested, washed with PBS containing 0.04% bovine serum albumin (BSA), resuspended in PBS containing 0.04% BSA at the cell density of 1000 cells / µl and passed through 35 µm cell strainer. Library preparation for Single Cell 3'RNA-seq run on Chromium platform (10x Genomics) for 4000 cells was performed according to manufacturer's instructions and the libraries were paired-end sequenced (R1:27, i7-index:8, R2:98) on HiSeq 4000 (Illumina). Preprocessing of scRNA-seq data, including demultiplexing, alignment, filtering, barcode counting, and unique molecular identifier (UMI) counting was performed using CellRanger.

To filter low quality cells, cells with fewer than 50,000 mapped reads, cells expressing fewer than 4000 genes or cells with greater than 6% UMI originating from mitochondrial genes were excluded. All genes that were not detected in at least 5 cells were discarded. From each sample, 500 cells were down-sampled for further analysis. The data was normalized and log-transformed using Seurat[30] (version 2.3.4). A cell cycle phase-specific score was generated for each cell, across five phases (G1/S, S, G2/M, M and M/G1) based on Macosko et al.[50] using averaged normalized expression levels of the markers for each phase. The cell cycle phase scores together with nUMI and percentage of UMIs mapping to mitochondrial genes per cell were regressed out using a negative binomial model. The graph-based method from Seurat was used to cluster the cells. The first 30 PCs were used in construction of SNN graph, and 10 clusters were detected with a resolution of 0.8. Markers specific to each cluster were identified using the "negbinom" model. Pseudotime

trajectories were constructed with URD[51] (version 1.0.2). The RNA velocity analysis was performed using velocyto[32] (version 0.17).

Oil-Red-O- and PAS-staining

Oil-Red-O and Periodic Acid-Schiff (PAS) staining were performed according to the manufacturer's recommendation (Sigma). Briefly, for Oil-Red-O-staining, cells were fixed with paraformaldehyde (4%) for 30 mins, washed with PBS, incubated with 60% isopropanol for 5 mins and Oil-Red-O working solution for 10 mins, and washed twice with 70% ethanol. For PAS-staining, cells were fixed with alcoholic formalin (3.7%) for 1 min, incubated with PAS solution for 5 mins and Schiff's reagent for 15 mins with several washes with water between each step, and counter-stained with hematoxylin.

ATAC-seq

Fibroblasts and CMT+sg*TP53* cells (p20) were harvested and 50,000 cells for each condition were processed for ATAC-seq libraries using previously reported protocol[52] and sequenced PE 2x75 NextSeq 500 (Illumina). The quality metrics of the FASTQ files were checked using FASTQC and the adapters were removed using trim_galore. The reads were aligned to human genome (hg19) using BWA, and the duplicate reads and the mitochondrial reads were removed using PICARD. The filtered and aligned read files were used for peak calling using MACS2 and for visualizing the traces using the IGV genome browser.

NaNoME-seq (NOME-seq using Nanopore sequencing)

To profile chromatin accessibility using GC methylase using NOME-seq protocol[37] and ability of Nanopore sequencing to detect CpG methylation without bisulfite conversion and PCR, we adapted the NOME-seq protocol for Nanopore sequencing on Promethion (NaNoME-seq). The nuclei isolation and treatment with GC methylase (M.CviPI) was

performed as described earlier[37]. The DNA was isolated from GC methylase treated nuclei by phenol chloroform followed by ethanol precipitation. The sequencing library for Promethion was prepared using the 1D genomic DNA by ligation kit (SQK-LSK109) as per manufacturer's recommendation and we loaded 50 fmol of final adapter-ligated high molecular weight genomic DNA to the flow cells for sequencing. After sequencing and basecalling, the Nanopore reads were aligned to GRCh37 reference genome with minimap2[53]. Nanopolish[54] was modified to call methylation in GC context. In total, 11Gbp of aligned read data from PCR amplified and GC methylated sequencing run was used to learn emission model for methylated GC sites. The learning process followed https://github.com/jts/methylation-analysis/blob/master/pipeline.make with adjustments for using human genome data and minimap2. For nuclear extract NaNoMe samples, methylation status was separately called for GC and CG sites. Similar independent method was recently described in a preprint by Lee et al (https://www.biorxiv.org/content/10.1101/504993v2). Reads with consecutive stretch of at least 80 GC sites with at least 75% methylated were filtered out due to expected cell free DNA contamination during library preparation as in Shipony et al. (https://www.biorxiv.org/content/10.1101/504662v1). The per site methylation levels in **Fig. 4c** are mean smoothed with triangular kernel 5 sites wide. Fibroblast and CMT+sg*TP53* NaNoMe analyses used 20.3Gbp and 24.8Gbp of aligned data, respectively.

References

1 Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777-D783 (2017).

2 Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17-37 (2013).

3 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558 (2013).

4 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).

5 Boehm, J. S., Hession, M. T., Bulmer, S. E. & Hahn, W. C. Transformation of human and murine fibroblasts without viral oncoproteins. *Mol. Cell. Biol.* **25**, 6464-6474 (2005).

6 Chaffer, C. L. & Weinberg, R. A. How does multistep tumorigenesis really proceed? *Cancer Discov.* **5**, 22-24 (2015).

7 Kamijo, T. *et al.* Tumor suppression at the mouse INK4a locus mediated by the alternative reading frame product p19ARF. *Cell* **91**, 649-659 (1997).

8 Metz, T., Harris, A. W. & Adams, J. M. Absence of p53 allows direct immortalization of hematopoietic cells by the myc and raf oncogenes. *Cell* **82**, 29-36 (1995).

9 Rangarajan, A., Hong, S. J., Gifford, A. & Weinberg, R. A. Species- and cell type-specific requirements for cellular transformation. *Cancer Cell* **6**, 171-183 (2004).

10 Ruley, H. E. Adenovirus early region 1A enables viral and cellular transforming genes to transform primary cells in culture. *Nature* **304**, 602-606 (1983).

11 Stevenson, M. & Volsky, D. J. Activated v-myc and v-ras oncogenes do not transform normal human lymphocytes. *Mol. Cell. Biol.* **6**, 3410-3417 (1986).

12 Land, H., Parada, L. F. & Weinberg, R. A. Tumorigenic conversion of primary embryo fibroblasts requires at least two cooperating oncogenes. *Nature* **304**, 596-602 (1983).

13 Shih, C., Padhy, L. C., Murray, M. & Weinberg, R. A. Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature* **290**, 261-264 (1981).

14 Sinn, E. *et al.* Coexpression of MMTV/v-Ha-ras and MMTV/c-myc genes in transgenic mice: synergistic action of oncogenes in vivo. *Cell* **49**, 465-475 (1987).

15 Hahn, W. C. *et al.* Creation of human tumour cells with defined genetic elements. *Nature* **400**, 464-468 (1999).

16 Moore, P. S. & Chang, Y. Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nat. Rev. Cancer* **10**, 878-889 (2010).

17 Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861-872 (2007).

18 Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987-1000 (1987).

19 Pang, Z. P. *et al.* Induction of human neuronal cells by defined transcription factors. *Nature* **476**, 220-223 (2011).

20 Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475**, 390-393 (2011).

21 Du, Y. *et al.* Human hepatocytes with drug metabolic function induced from fibroblasts by lineage reprogramming. *Cell Stem Cell* **14**, 394-403 (2014).

22 Huang, P. *et al.* Direct reprogramming of human fibroblasts to functional and expandable hepatocytes. *Cell Stem Cell* **14**, 370-384 (2014).

23 Morris, S. A. *et al.* Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158**, 889-902 (2014).

24 Lowe, S. W., Cepero, E. & Evan, G. Intrinsic tumour suppression. *Nature* **432**, 307-315 (2004).

25 Kalkat, M. *et al.* MYC Deregulation in Primary Human Cancers. *Genes (Basel)* **8**, E151 (2017).

26 Palin, K. *et al.* Contribution of allelic imbalance to colorectal cancer. *Nat. Commun.* **9**, 3664 (2018).

27 Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* **33**, 676-689 (2018).

28 Cancer Genome Atlas Research Network. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**, 1327-1341 (2017).

29 Moinzadeh, P., Breuhahn, K., Stutzer, H. & Schirmacher, P. Chromosome alterations in human hepatocellular carcinomas correlate with aetiology and histological grade--results of an explorative CGH meta-analysis. *Br. J. Cancer* **92**, 935-941 (2005).

30 Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495-502 (2015).

31 van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579-2605 (2008).

32 La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494-498 (2018).

33 Borggrefe, T. & Oswald, F. The Notch signaling pathway: transcriptional regulation at Notch target genes. *Cell. Mol. Life Sci.* **66**, 1631-1646 (2009).

34 Zhu, B. *et al.* Activated Notch signaling augments cell growth in hepatocellular carcinoma via up-regulating the nuclear receptor NR4A2. *Oncotarget* **8**, 23289-23302 (2017).

35 Villanueva, A. *et al.* Notch signaling is activated in human hepatocellular carcinoma and induces tumor formation in mice. *Gastroenterology* **143**, 1660-1669 e1667 (2012).

36 Hoshida, Y. *et al.* Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res.* **69**, 7385-7392 (2009).

37 Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22**, 2497–2506 (2012).

38 Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911-917 (2018).

39 Drost, J. *et al.* Sequential cancer mutations in cultured human intestinal stem cells. *Nature* **521**, 43-47 (2015).

40 Matano, M. *et al.* Modeling colorectal cancer using CRISPR-Cas9-mediated engineering of human intestinal organoids. *Nat. Med.* **21**, 256-262 (2015).

41 Park, J. W. *et al.* Reprogramming normal human epithelial tissues to a common, lethal neuroendocrine cancer lineage. *Science* **362**, 91-95 (2018).

42 Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014).

43 Icay, K. *et al.* SePIA: RNA and small RNA sequence processing, integration, and analysis. *BioData Min.* **9**, 20 (2016).

44 Ovaska, K. *et al.* Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med.* **2**, 65 (2010).

45 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

46 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754-D761 (2018).

47 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res.* **4**, 1521 (2015).

48 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).

49 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U S A* **102**, 15545-15550 (2005).

50 Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).

51 Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360** (2018).

52 Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959-962 (2017).

53 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

54 Simpson, J. T. *et al*. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).

Acknowledgments: We thank Drs. Otto Kauko, Teemu Kivioja, and Minna Taipale for critical review of the manuscript, and Tomi Leung, Anu M. Luoto, and Kaisu Jussila for technical assistance. We also thank HiLIFE research infrastructures including Biomedicum Virus Core, Single-cell Analytics FIMM, Biomedicum Imaging Unit and Laboratory Animal Center.

Author contributions: JT conceived and supervised the study. BS designed and performed all the experiments with help from PP. BS performed the initial data processing of single cell and bulk RNA-seq data. KZ performed single cell RNA-seq analysis. BS performed the NaNoMe-seq and KP analyzed the data with inputs from SA and LA. AC performed bulk RNA-seq analysis with inputs from BS and PP. SH provided the bioinformatics support and inputs into the project. All authors contributed to the writing of the manuscript.

Author Information: Reprints and permissions information is available at

Competing interests: Authors declare no competing interests.

Corresponding author: Correspondence and requests for materials should be addressed to *ajt208@cam.ac.uk*

Funding: This work was supported by grants from Academy of Finland (Finnish Center of Excellence Program; 2012-2017, 250345 and 2018-2025, 312041, Post-doctoral fellowships; 274555, 288836 and Research Fellowships, 317807), Jane and Aatos Erkko Foundation, and the Finnish Cancer Foundation.

Data and materials availability: All sequence data will be made available under ENA.

Figure 1.

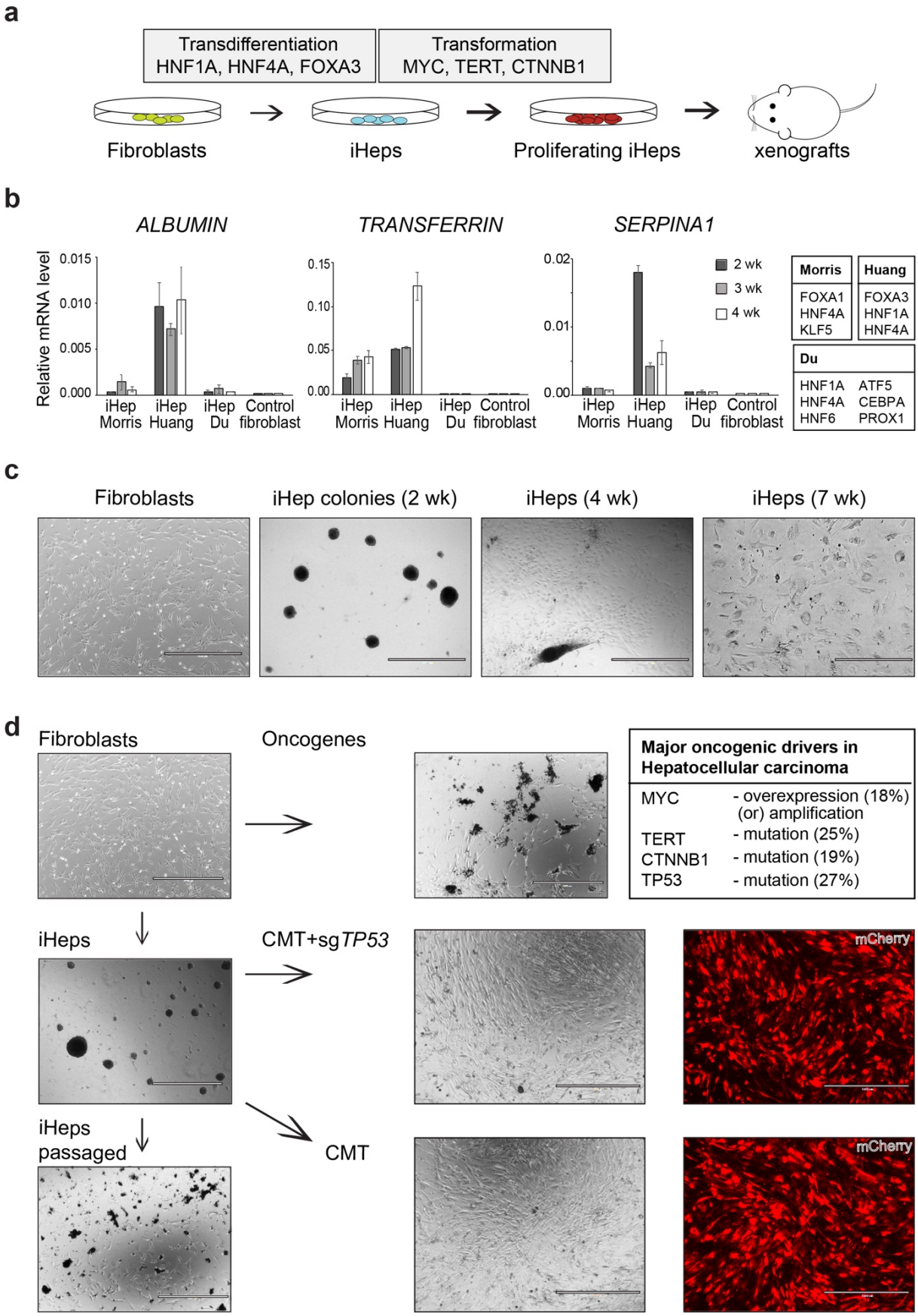


Figure 1: Generating proliferative induced hepatocytes using defined transcription factors and oncogenic drivers.

a, Schematic outline of the cell transformation assay for making lineage-specific cancer by lentiviral expression of three lineage-specific TFs to convert HFs to induced hepatocytes (iHep) and defined oncogenic drivers to transform iHeps to proliferating and tumorigenic cells.

b, Comparison of TF combinations[21,22,23] for converting human fibroblasts to iHeps by detecting transcript levels for liver marker genes (*ALBUMIN*, *TRANSFERRIN* and *SERPINA1/α-1-antitrypsin*) by qRT-PCR at different time points after iHep conversion, normalized to GAPDH levels (mean ± standard error).

c, Phase contrast microscope images showing the phenotype and morphology of the cells in the course of conversion of fibroblasts to iHeps at different times points after transduction of a cocktail of three TFs HNF1A, HNF4A and FOXA3[22].

d, Generation of highly proliferative iHep cells by transducing iHeps with two pools of liver cancer-specific oncogenic drivers. CMT pool contains three oncogenes CTNNB1^{T41A}, MYC, and TERT, and CMT+sg*TP53* pool contains the same oncogenes along with constructs for *TP53* inactivation by CRISPR-Cas9. Phase contrast microscope images showing the phenotype and morphology of the cells. Mutation rates of the oncogenic drivers as reported in the COSMIC database for HCC and MYC amplification as reported in[28]. Oncogenes are co-transduced with fluorescent reporter mCherry for detection of transduced cells. Oncogene transduction to fibroblasts fails to transform the cells, passaging of oncogene-expressing fibroblasts as well as iHeps without oncogenes results in apoptosis after few passages. Scale bar 1000 μm unless otherwise specified.

Figure 2.

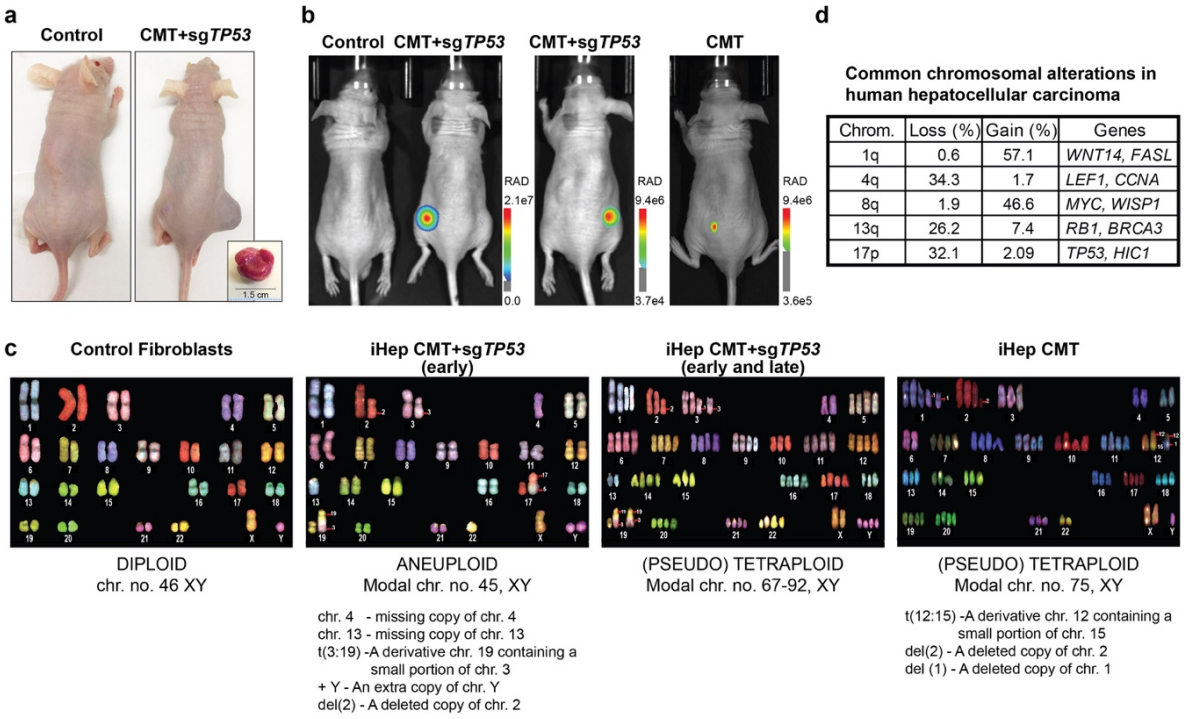


Figure 2: Tumorigenic properties of the transformed iHeps.
a, Subcutaneous injection of proliferating iHeps results in xenograft tumors in nude mice (tumor size of 1.5 cm ~ 23 weeks after xenotransplantation). Proliferative iHeps transduced with defined CMT oncogenes with *TP53* inactivation (CMT+sg*TP53*) or control iHeps without oncogenes were used in the injections.
b, *In vivo* imaging of xenograft tumors ~12 weeks after implantation. Two biological replicate experiments are shown for CMT+sg*TP53* cells with iHep conversion and oncogene transduction with *TP53* inactivation performed in two separate human fibroblast cell lines (foreskin fibroblast [*left*] and fetal lung fibroblast [*middle*]) as well as proliferative CMT iHeps without *TP53* inactivation (*right*). Fluorescence signal emitted by mCherry co-transduced with the oncogenes is detected *in vivo* using the Lago system (scale bar = radiance units). Control mice are injected with either fibroblasts or iHeps.
c, Analysis of chromosomal aberrations in the transformed iHeps by spectral karyotyping. CMT+sg*TP53* cells were analyzed at passage 18 (*early*) and p50 (*late*) and CMT cells at passage 18. Fibroblasts have normal diploid karyotype (46, XY, representative spectral image on *left*) and transformed iHeps show aneuploidies as indicated in the figure. Early passage CMT+sg*TP53* cells show two different populations with two distinct modal chromosome numbers (45, XY and 67-92, XY, representative spectral image for 45, XY on *middle-left*). Late passage CMT+sg*TP53* cells have modal chromosome number 67-92, XY (*middle-right*) and CMT cells 75, XY (*right*).
d, Frequencies of chromosomal alterations reported for human HCC samples (see[29]).

Figure 3.

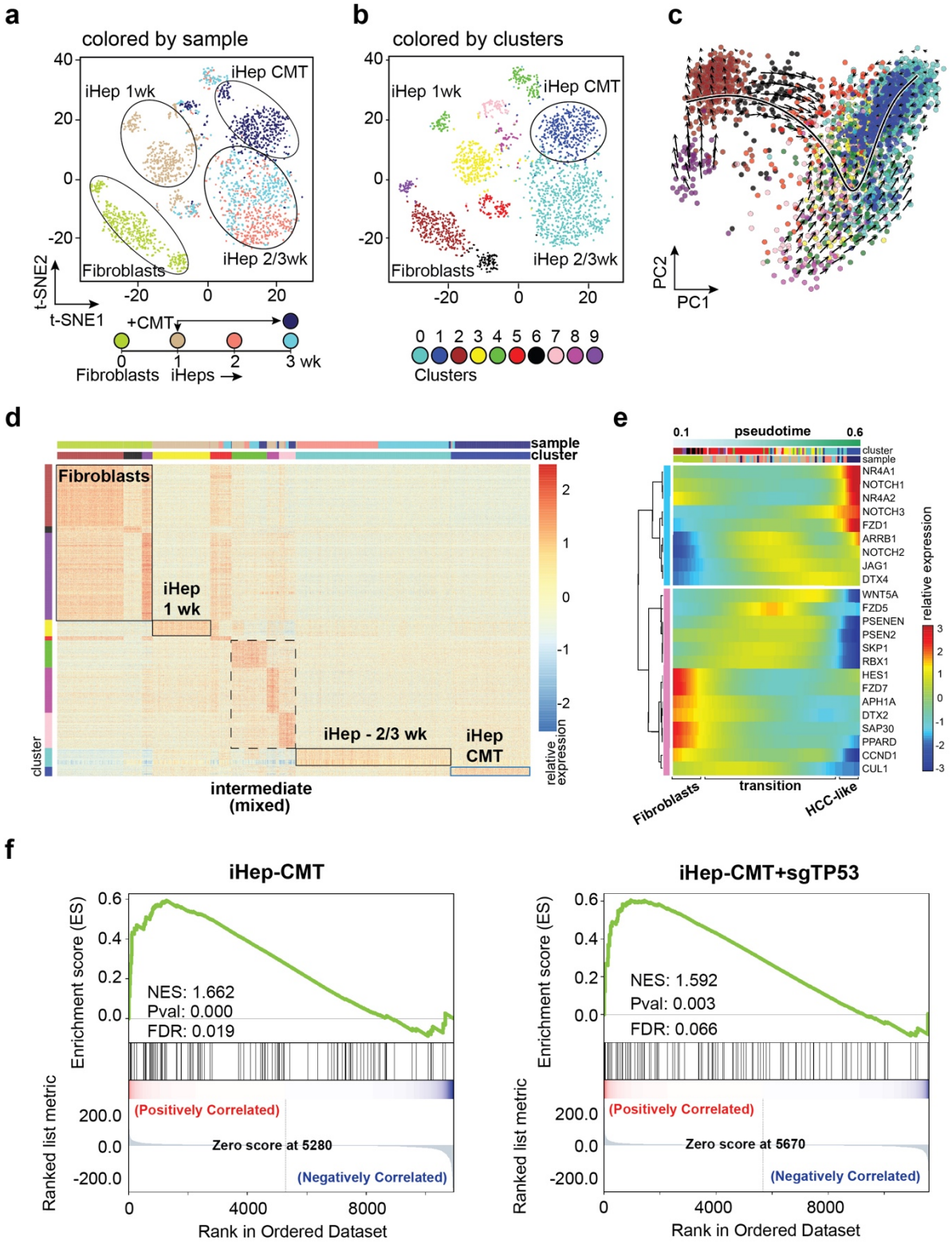


Figure 3: Transformed iHeps show gene expression profile similar to liver cancer.
a, t-SNE of single cells from fibroblasts, iHeps at one–three weeks after iHep induction, and iHeps transduced with CMT oncogenes at one week and harvested for scRNA-seq two weeks later. Cells are colored by sample, and sample collection timeline is indicated.

b, t-SNE of cells clustered according to their similar gene expression profiles. Each dot represents a data point for individual cell and the cells with similar gene expression profiles are colored according to the clusters.
c, Principal component analysis (PCA) projection of single cells shown with velocity field with the observed states of the cells shown as circles and the extrapolated future states shown with arrows for the first two principal components. Cells are colored by cluster identities corresponding to Fig. 3b.
d, Clustered heatmap showing the relative expression levels of cluster-specific marker genes (the expression of a gene in a particular cell relative to the average expression of that gene across all cells) from single cell RNA-seq analysis. Color code illustrating sample and cluster identities correspond to the colors in Fig. 3a and b, respectively.
e, Relative expression of the genes from the Notch signaling pathway across pseudotime in the single-cell RNA-seq data (the expression of a gene in a particular cell relative to the average expression of that gene across all cells). Color code illustrating sample and cluster identities correspond to the colors in Fig. 3a and b, respectively.
f, Gene set enrichment analysis (GSEA) results for CMT-iHeps and CMT+sg*TP53*-iHeps compared to control fibroblasts against liver cancer signature (Subclass 2[36]) from molecular signatures database (MSigDB). Positive normalized enrichment score (NES) reflects overrepresentation of liver cancer signature genes among the top ranked differentially expressed genes in CMT-iHep and CMT+sg*TP53*-iHep conditions compared to control fibroblasts.

Figure 4.

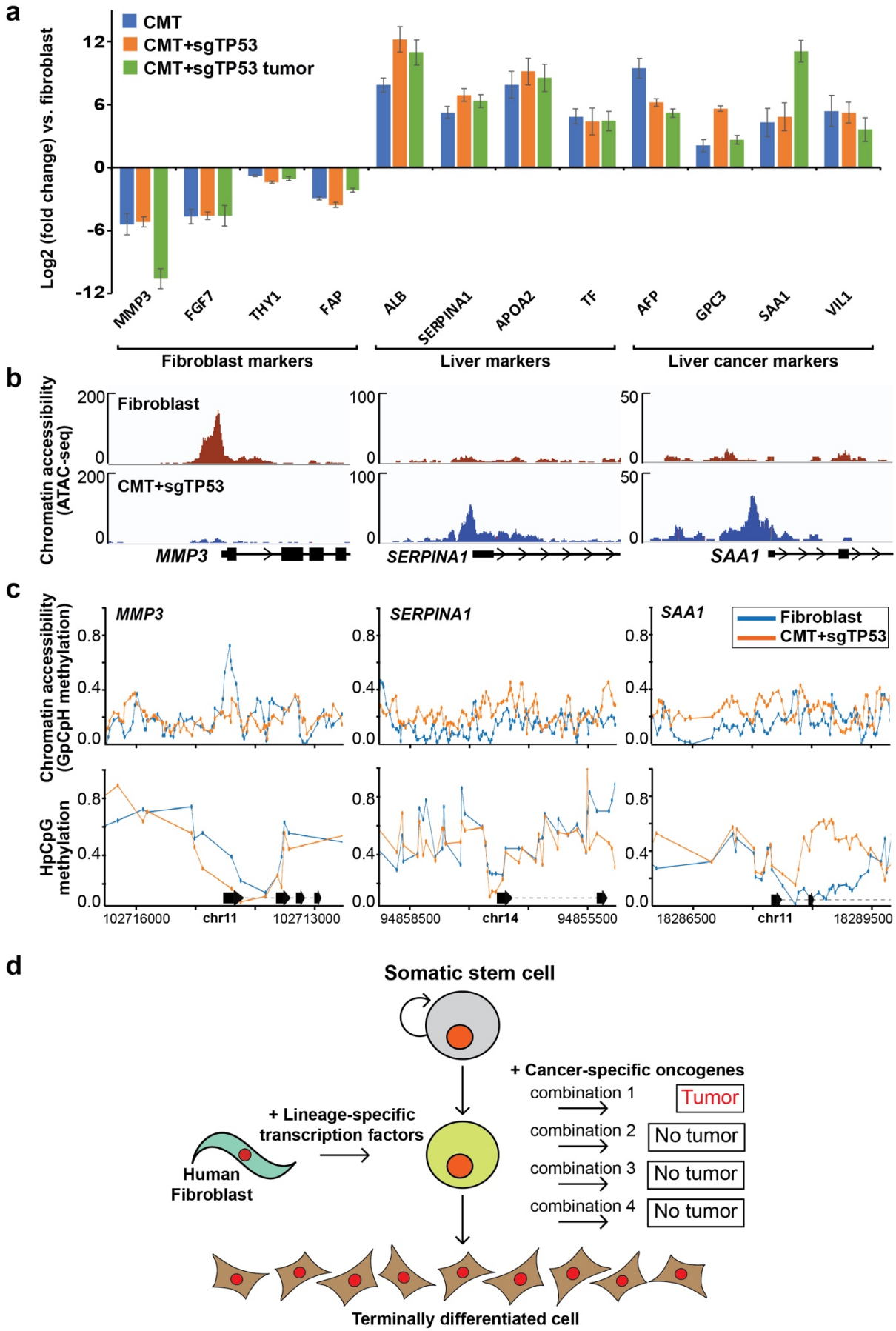


Figure 4: Direct conversion of human fibroblasts to liver cancer cells.
a, Differential expression levels [log2(fold change)] of marker genes for fibroblasts, hepatocytes, and liver cancer, and fibroblasts in bulk RNA-seq measurements from CMT+sg*TP53*-iHeps, CMT-iHeps and xenograft tumor from CMT+sg*TP53* against control fibroblasts (±standard error).
b, IGV snapshots for promoter regions of representative genes from fibroblast markers (*MMP3*), liver markers (*SERPINA1/α-1-antitrypsin*), and liver cancer markers (*SAA1*) showing ATAC-seq enrichment from fibroblast and CMT+sg*TP53*-iHeps.
c, Chromatin accessibility and CpG methylation of DNA measured using NaNoMe-seq. Cytosine methylation detected using Nanopore sequencing from CMT+sg*TP53*-iHeps and control fibroblasts is shown for promoter regions of representative genes from fibroblast markers (*MMP3*), liver markers (*SERPINA1/α-1-antitrypsin*), and liver cancer markers (*SAA1*) using a window of TSS ±1500 bp. GpCpH methylation (all GC sequences where the C is not part of a CG sequence also, top) reports on chromatin accessibility, whereas HpCpG methylation reports on endogenous methylation of cytosines in the CpG context.
d, Schematic presentation of the molecular approach for identifying minimal determinants of tumorigenesis in specific tissues. Lineage-specific transcription factors are used to reprogram human fibroblasts to precise cellular identity (left), whose transformation by specific combinations of oncogenes (right) can then be tested. This approach, combined with single-cell RNA-seq and RNA velocity analyses allows also analysis of which cell type along the stem cell to terminally differentiated cell axis (top to bottom) is susceptible for transformation.