CELL BIOLOGY

Ancestry-dependent gene expression correlates with reprogramming to pluripotency and multiple dynamic biological processes

Laura S. Bisogno[1], Jun Yang[1], Brian D. Bennett[2], James M. Ward[2], Lantz C. Mackey[1], Lois A. Annab[1], Pierre R. Bushel[3], Sandeep Singhal[4], Shepherd H. Schurman[5], Jung S. Byun[6], Anna María Nápoles[6], Eliseo J. Pérez-Stable[6,7], David C. Fargo[8], Kevin Gardner[6,9], Trevor K. Archer[1]*

Induced pluripotent stem cells (iPSCs) can be derived from differentiated cells, enabling the generation of personalized disease models by differentiating patient-derived iPSCs into disease-relevant cell lines. While genetic variability between different iPSC lines affects differentiation potential, how this variability in somatic cells affects pluripotent potential is less understood. We generated and compared transcriptomic data from 72 dermal fibroblast–iPSC pairs with consistent variation in reprogramming efficiency. By considering equal numbers of samples from self-reported African Americans and White Americans, we identified both ancestry-dependent and ancestry-independent transcripts associated with reprogramming efficiency, suggesting that transcriptomic heterogeneity can substantially affect reprogramming. Moreover, reprogramming efficiency–associated genes are involved in diverse dynamic biological processes, including cancer and wound healing, and are predictive of 5-year breast cancer survival in an independent cohort. Candidate genes may provide insight into mechanisms of ancestry-dependent regulation of cell fate transitions and motivate additional studies for improvement of reprogramming.

INTRODUCTION

Reprogramming of adult somatic cells to a pluripotent state has the potential to transform disease modeling and regenerative medicine. It is now well established that induced pluripotent stem cells (iPSCs) can be generated from human fibroblasts by expressing four transcription factors: Oct4, Sox2, Klf4, and c-Myc (*1*). Consequently, unique and personalized disease models can be generated for any individual by differentiating patient-derived iPSCs into disease-relevant cell lines and organoids, which could greatly enhance precision and regenerative medicine.

Human genomes and epigenomes are heterogenous, and individual genetic variation can confound our understanding of biological mechanisms of disease. There is also substantial variation across major ancestral groups. Despite this, efforts in racial diversity have not maintained pace with advances in genomic technologies, and as a result, non-European ancestry samples are underrepresented in genomic databases, including The Cancer Genome Atlas (TCGA) and genome-wide association studies (GWAS) (*2*, *3*). Furthermore, with a few notable exceptions, most cell culture models currently in use are derived from individuals of European ancestry (*4*, *5*). iPSC models are also limited in the availability of non-European lines. A few recent studies have recognized this and focused on creating diverse human pluripotent cell lines (*6*–*9*). Nonetheless, availability of iPSC lines and data derived from these cells have not kept pace with iPSC technologies. This lack of diversity has contributed to large knowledge gaps in population-specific genetic variants that can cause disease susceptibility or altered drug responses, leading to imbalanced outcomes across ethnicities in genomics-based medicine initiatives (*2*, *3*, *7*).

Several studies have reported that genetic background affects the lineage commitment potential of iPSCs more so than most factors, including cell source (*10*, *11*). While the extent to which genetic variability in cells of origin affects dedifferentiation potential is less understood, work in inbred mouse models suggests that genetic variability plays an important role in the efficiency of iPSC induction (*12*). The importance of incorporating genetic heterogeneity into research models has been recognized and recently begun to be addressed with models such as the Diversity Outbred mice and the Collaborative Cross inbred strains (*13*, *14*). Diversifying both animal and cell culture research models will provide powerful tools for better understanding how genetic heterogeneity translates to important biological processes. Specifically, understanding how human donor–specific genetic variability affects iPSC generation is critical to establishing optimized protocols for iPSC-based disease, therapeutic, and toxicity models.

We recently generated primary dermal fibroblasts (DFs) and matched iPSCs from healthy self-reported African American and White American donors (*15*). We observed reproducible variability in the efficiency of reprogramming to iPSCs, with some samples having consistently low and others having consistently high efficiency. Notably, fibroblasts derived from African Americans reprogrammed significantly better as a group than those derived from White Americans. We found that individual variability in SWI/SNF complex subunits in iPSCs was significantly correlated with reprogramming efficiency. In the present study, we aimed to define genes

[1]Chromatin and Gene Expression Section, Epigenetics and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA. [2]Integrative Bioinformatics, Epigenetics and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA. [3]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA. [4]Department of Pathology, Department of Computer Science, University of North Dakota, Grand Forks, ND, USA. [5]Clinical Research Unit, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA. [6]Division of Intramural Research, Office of the Scientific Director, National Institute on Minority Health and Health Disparities, Bethesda, MD, USA. [7]Division of Intramural Research, National Heart, Lung and Blood Institute, Bethesda, MD, USA. [8]Office of Scientific Computing, National Institute of Environmental Health Sciences, Research Triangle Park, NC, USA. [9]Department of Pathology and Cell Biology, Columbia University Medical Center, Columbia University, New York, NY, USA.
*Corresponding author. Email: archer1@niehs.nih.gov

that contribute to the efficiency at which DFs reprogram to iPSCs, as well as investigate ancestry-dependent gene expression involved in reprogramming efficiency on a transcriptome-wide scale. We identified both ancestry-dependent and ancestry-independent gene expression signatures that can predict and explain interpersonal differences in reprogramming efficiency. Genes associated with reprogramming efficiency were enriched for involvement in wound healing and cancer. Thus, we identified both ancestry-dependent and ancestry-independent gene expression signatures that are not only predictive of reprogramming to pluripotency but also involved in other important cell fate transitions and biological processes.

RESULTS

Gene expression changes during iPSC induction are sex and race independent

To investigate individual differences in gene expression during reprogramming, we measured the transcriptome with RNA sequencing of DFs and matched iPSCs obtained from 36 African American and 36 White American individuals from our previously described cohort (table S1) (*15*). We detected extensive changes in RNA abundance, with more than 6000 transcripts changing $\log_2$ fold ≥ 4 [false discovery rate (FDR) ≤ 0.05] (Fig. 1A), including expected canonical fibroblast and stem cell markers (Fig. 1B). DFs were distinguishable from iPSCs by RNA sequencing, but race and sex were not (Fig. 1C). In line with this, gene expression differences during reprogramming between the sex and race subcohorts were highly correlated, as expected (fig. S1), and plotting the pairwise difference in expression for the top 5760 up-regulated genes ($\log_2$ fold change > 4 and FDR ≤ 0.05) between two samples being compared showed similar distributions for each comparison (Fig. 1D).

To formally test whether the profound gene expression changes observed during the reprogramming of DFs to iPSCs were common of both sex and race groups, we tested for interactions between cell type and sex or ancestry. Briefly, we ran DESeq2 analysis models adding in cell type/sex or cell type/race interaction terms. Very few genes reached significance, with only 84 expressed genes reaching significance of FDR ≤ 0.05 for the cell type by sex interaction term and only 160 expressed genes reaching significance for the cell type by race interaction term (table S2). Thus, a relatively small number of genes exhibit cell type changes dependent on these known factors compared to the large number of changes occurring during reprogramming.

Ancestry-dependent and ancestry-independent genes are associated with reprogramming efficiency

To identify genes driving differences in reprogramming efficiency between individuals, we incorporated the experimentally determined reprogramming efficiency values as a continuous feature into our RNA-sequencing data analysis model and calculated Spearman and Pearson correlation coefficients for gene expression and reprogramming efficiency in the DFs (table S2). We calculated these metrics for the total combined cohort ($n = 72$), African Americans only ($n = 36$), and White Americans only ($n = 36$). Spearman and Pearson correlations were highly congruent, and genes identified as significant by the RNA-sequencing model were generally identified as significant by both Pearson and Spearman correlations (fig. S2A). Because of the presence of outliers, we focused our analyses on the Spearman's rank correlations.

To determine an appropriate Spearman cutoff, and to see whether our metric identified more associated genes than expected by random chance, we performed 10,000 permutations randomizing reprogramming efficiency values while maintaining gene expression values in each fibroblast line. We found that the number of experimentally determined genes associated with efficiency was higher than the median number of associated genes from randomized permutations. The difference between experimental and randomized gene sets reached significance $P \leq 0.05$ at a Spearman cutoff of $P \leq 0.01$ (fig. S2B). For this reason, we chose to focus on genes that reached a Spearman significance threshold of 0.01 in subsequent analyses.

At the Spearman $P \leq 0.01$ threshold, we identified 1764 genes associated with reprogramming efficiency in the combined cohort, 845 positively associated and 919 negatively associated (Fig. 2A and table S2). While we did not observe a difference in reprogramming efficiency between different sexes (*15*), we nonetheless tested to see the effect of this additional variable on the genes associated with reprogramming efficiency. To do this, we stratified the cohort by sex and reran our RNA analysis model with reprogramming efficiency as a continuous feature. This analysis resulted in a relatively small number of significantly associated genes, with 192 expressed genes associated with reprogramming efficiency in females and 18 expressed genes associated with reprogramming efficiency in males. Thus, a small number of genes are likely dependent on two univariate results, and therefore, these are likely not limiting our interpretation (table S2).

Genes positively associated with reprogramming efficiency had higher RNA expression in high reprogramming samples, while genes negatively associated with reprogramming efficiency had lower RNA expression in high reprogramming samples (Fig. 2, B to C, and fig. S3, A to D). Unexpectedly, when we compared genes associated with reprogramming efficiency (Spearman $P \leq 0.01$) in the total cohort and across the two ancestries, we found very little overlap between the genes called significantly associated with efficiency, with only 31 genes (11 positively and 20 negatively associated) reaching significance in both ancestries and the combined cohort (Fig. 2A). This indicates that there were ancestry-dependent genes associated with reprogramming efficiency.

Many associated genes did not reach a significant association threshold when we split our cohort up by race (Fig. 2A and fig. S3B), or only reached significance in one race (fig. S3, C and D), suggesting that our detection of these genes was limited by power. However, the converse was also true, where other genes were no longer significant when we combined the cohort (Fig. 2, B and C). Genes differentially associated with reprogramming by ancestry had expression patterns that were so different between the ancestries that the association was lost when combined. For example, *GAS2* and *PLCE1* were associated with reprogramming efficiency in African Americans, but not White Americans, and therefore did not reach significance when the two ancestries were combined (Fig. 2B). Conversely, *FAM69B* and *WASHC2C* were associated with reprogramming efficiency in White Americans only (Fig. 2C). In total, 177 positively associated and 187 negatively associated genes were strictly African American dependent, and 223 positively associated and 146 negatively associated genes were White American dependent (Fig. 2A). Ancestry-dependent associated genes were not differentially expressed (fig. S3E), indicating that ancestry-dependent differences in the association were not driven by underlying gene expression differences.

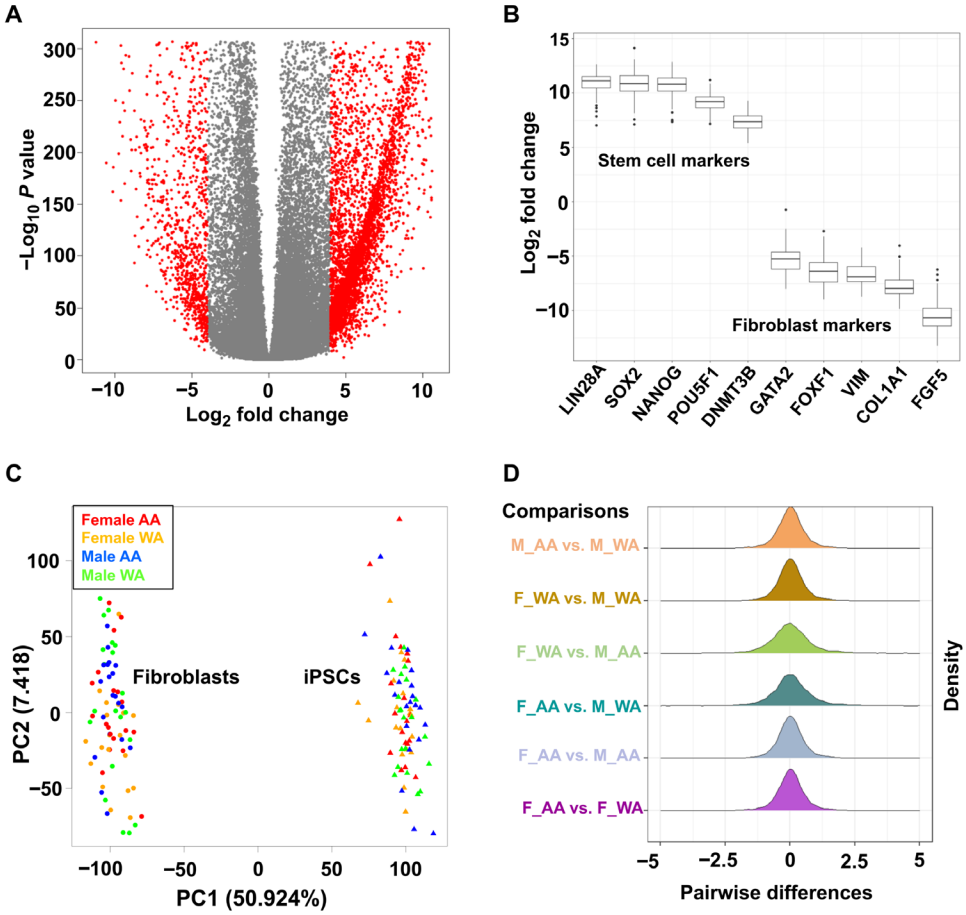

Fig. 1. Substantial gene expression changes during iPSC induction are sex and race independent. (**A**) Volcano plot displays the $\log_2$ fold change in expression during reprogramming for all detected RNAs. Transcripts with $\log_2$ fold change ≥4 or ≤−4 (FDR ≤ 0.05) are highlighted in red. (**B**) Box plots show the $\log_2$ fold change during reprogramming in each matched DF-iPSC pair for a subset of canonical stem cell and fibroblast markers. (**C**) Principal components analysis (PCA) plot showing clusters of samples based on similarity. The first two components (PC1 and PC2) of gene expression variance are displayed. Each dot represents a sample color coded by both cell type and demographic. (**D**) Ridgeline plots compare the pairwise differences of expression data for each gene up-regulated during reprogramming. AA, African American; WA, White American; M, male; F, female.

This further emphasizes the lack of overall expression differences between the ancestries shown in Fig. 1 and fig. S1.

Gene Ontology (GO) analysis indicated significant enrichment of categories representative of cell fate commitment and development, intracellular transport/protein localization, and cytoskeletal organization in both the total analysis and the race-specific analyses (Fig. 2D, fig. S3F, and table S3), suggesting that associated genes function in cell state organization and dynamics. This is in agreement with the notion that cellular properties change drastically with cell state dynamics, and proteins functioning in transport and cytoskeletal structure are significantly more abundant in fibroblasts compared to stem cells (*16*). However, there were categories that were unique to the individual ancestries (Fig. 2D), including apoptosis and hypoxic response, both of which have previously been linked to reprogramming (*1*, *17*). There were also enriched categories in the total that were enriched in one race, including WNT signaling and endoplasmic reticulum (ER) stress response, which have also been shown to regulate reprogramming efficiency (*18*, *19*). Our work suggests that transcripts encoding proteins functioning in pathways known to regulate reprogramming to pluripotency vary in primary donor-derived fibroblasts, and this variability might determine whether a fibroblast will reprogram efficiently.

Most genes associated with reprogramming efficiency are primed at the RNA level

While examining expression patterns of associated genes, we noticed a common pattern where fibroblasts that reprogrammed at higher efficiencies had expression levels more like iPSCs when compared to lines that reprogrammed at lower efficiencies (Fig. 2, B and C, and fig. S3, A to D). We quantified the number of associated genes with this "primed" expression pattern, and it was more common than nonprimed (Fig. 3A). For genes with this expression pattern, positively associated genes tended to increase expression and negatively associated genes tended to decrease expression during reprogramming. Therefore, fibroblast lines that reprogrammed with higher efficiencies generally had smaller expression changes for associated genes during reprogramming, as indicated by a $\log_2$ fold change closer to 0 in high reprogramming samples (Fig. 3, B and C). This suggests that DFs that reprogrammed with high efficiency were primed at the RNA level.

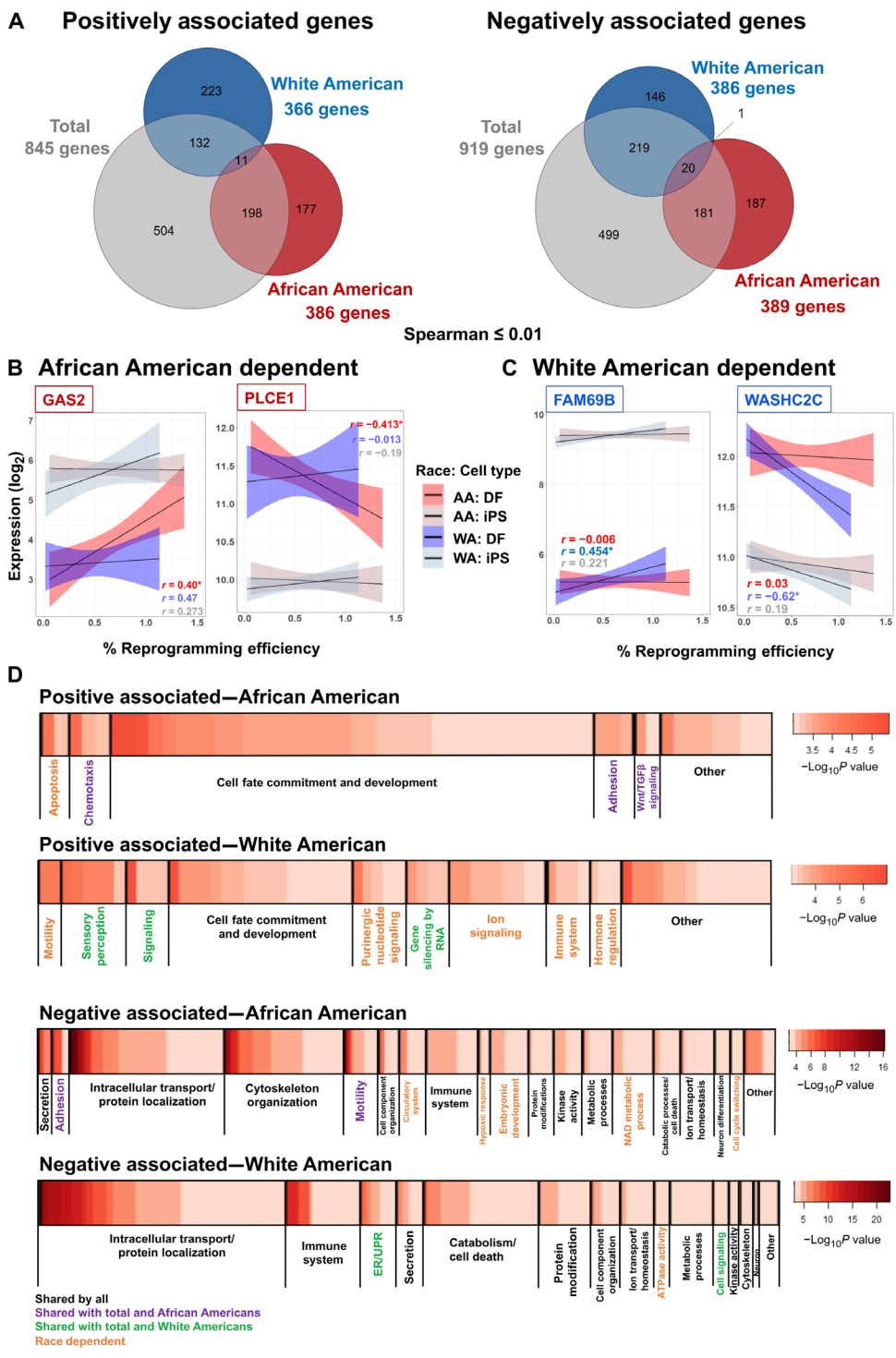

Fig. 2. Ancestry-dependent and ancestry-independent genes are associated with reprogramming efficiency. (**A**) Overlap of genes significantly associated with reprogramming efficiency (Spearman, $P \leq 0.01$) in the full cohort and in the ancestries independently separated into positively and negatively associated genes. Spearman values for a $P \leq 0.01$ cutoff are 0.388 for $n = 36$ (race cohorts) and 0.274 for $n = 72$ (total cohort). (**B** and **C**) Line plots with 95% confidence intervals of RNA expression (log_2) in DFs (dark colors) or iPSCs (iPS; matching light colors) for each individual sample in the cohort plotted against reprogramming efficiency, separated out by African American (AA; red) and White American (WA; blue). (B) Examples of genes uniquely associated in African Americans only (*GAS2* and *PLCE1*). (C) Examples of genes uniquely associated in White Americans only (*FAM69B* and *WASHC2C*). Spearman correlations calculated in the total population are in gray, African American in red, and White American in blue. Spearman correlations that reach a significance of $P \leq 0.01$ are denoted by * in individual plots. (**D**) GO analysis for genes associated with reprogramming efficiency in DFs. Heatmaps show the $-\log_{10} P$ value for enriched GO categories, grouped into related broader groupings, with ancestry-dependent functional categories in orange and functional categories identified in the combined analysis and only one race in purple (AA) and green (WA).

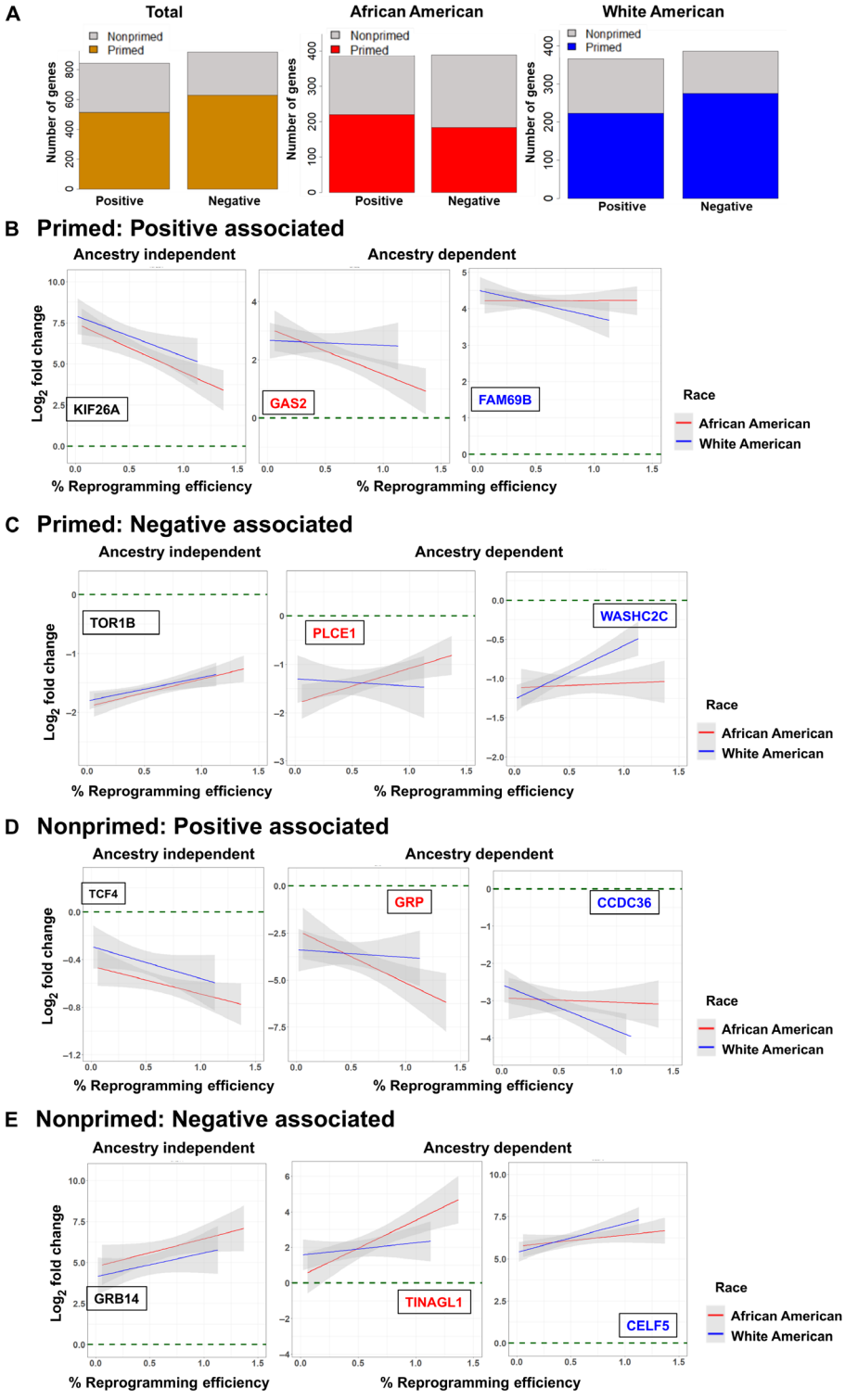

Fig. 3. Majority of genes associated with reprogramming efficiency are primed. (**A**) Number of associated genes (Spearman ≤ 0.01) considered to have a primed (orange, red, or blue) or a nonprimed (gray) gene expression pattern. (**B** to **E**) Log_2 fold change between each individual DF line and the mean iPSC expression for example primed genes that are (B) primed and positively associated with reprogramming efficiency in the combined ancestries (*KIF26A*), African Americans only (*GAS2*), and White Americans only (*FAM69B*); (C) primed and negatively associated with reprogramming efficiency in the combined ancestries (*TOR1B*), African Americans only (*PLCE1*), and White Americans only (*WASHC2C*); (D) nonprimed and positively associated with reprogramming efficiency in the combined ancestries (*TCF4*), African Americans only (*GRP*), and White Americans only (*CCDC36*); and (E) nonprimed and negatively associated with reprogramming efficiency in the combined ancestries (*GRB14*), African Americans only (*TINAGL1*), and White Americans only (*CELF5*). Note that the samples that reprogram with higher efficiency have log_2 fold changes closer to 0 (dashed green line) in primed genes and farther from 0 in nonprimed genes.

Although most genes associated with reprogramming efficiency exhibited this primed RNA expression pattern, there were several examples, both ancestry dependent and independent, of associated genes that did not exhibit this pattern (Fig. 3A). A subset of the non-primed genes was anti-primed, where samples that reprogrammed with higher efficiency were less like iPSCs (Fig. 3, D and E). Some genes that fit this expression pattern, including *TINAGL1* and *TCF4*, may both enhance or activate pluripotency in a time-dependent manner (*20*–*22*). Consequently, the anti-primed gene expression patterns may also be informative in the context of pluripotency.

Ancestry-dependent gene signatures effectively rank samples by efficiency in the corresponding cohort

Given the large number of genes associated with efficiency, we wanted to see how the summation of gene expression contributed to relative reprogramming efficiencies. We reasoned that, on the basis of the common occurrence of the primed gene expression pattern, samples that reprogrammed with higher efficiency were more likely to have higher expression levels of positively associated genes and lower expression levels of negatively associated genes. However, due to expression variability, the extent of primed gene expression for the sum of large gene sets in any given individual is unknown. We scored each fibroblast line using binary indicators: If an individual line had expression levels of a positively associated gene above the cohort mean, or expression levels of a negatively associated gene below the cohort mean, then the score for the individual line for that gene was "1." If expression levels did not meet these criteria, then a value of "0" was assigned (Fig. 4A). We summed the scores for positive and negative associated genes and used scores to rank samples. We then determined the correlation between these ranks and reprogramming efficiency.

To see how our associated gene sets ranked samples relative to other combinations of expressed genes, we applied our scoring system to 10,000 random sets of 750 genes in both the total cohort and race subcohorts (Fig. 4B). When we ranked samples according to the top 750 associated genes in the total analysis, the rankings were significantly better than random in the total analysis but right in the middle of the random distribution for the race cohorts individually (Fig. 4B, gray arrows). However, when we ranked with the top 750 associated genes in the race cohorts, the rankings were significantly better than random in the corresponding cohort but significantly worse than random in the opposite cohort. Therefore, considering the ancestries independently significantly improved our ability to determine relative reprogramming efficiencies based on associated gene expression.

Genes associated with reprogramming efficiency are involved in other dynamic biological processes and cell fate transitions

During reprogramming to pluripotency, cells undergo a dramatic change in cell identity, and our GO analysis suggested enrichment of genes functioning in categories related to cell state dynamics. To further investigate whether genes associated with reprogramming efficiency could be more generally related to how readily a cell changes state, we performed gene set enrichment analysis (GSEA) using all genes preranked on Pearson correlation coefficients for gene expression and reprogramming efficiency. We used Pearson correlations instead of Spearman correlations for this analysis because, given that Spearman correlation uses rank, there were duplicated correlation values for genes of interest. We first tested gene sets in the Canonical Pathways curated gene set (c2.cp) collection from the Molecular Signatures Database (MSigDB). While enriched gene sets unsurprisingly overlapped many of the enriched GO categories (Fig. 2D and fig. S3F), we also observed significant positive enrichment of gene sets related to transcriptional regulation, epigenetics (including DNA methylation), and chromosome maintenance (Fig. 5A and table S4), all of which function in the regulation of cell plasticity (*23*). In addition, we observed negative enrichment of many cancer-related gene sets (Fig. 5A and table S4). Cancer initiation and progression are highly dynamic, and metastatic dissemination is often promoted by an epithelial to mesenchymal transition (EMT), a cellular transformation in which epithelial cells gain mesenchymal characteristics. Reprogramming to iPSCs requires the opposite of an EMT, a mesenchymal to epithelial transition (MET) (*24*). We therefore tested an EMT gene set and, as expected, saw a negative enrichment score (Fig. 5A). Similarly, wound healing involves cell state transitions and an EMT is activated during this process to promote epithelial cell migration, and wound healing–related gene sets were significantly negatively enriched in our ranked gene set (Fig. 5A). This agrees with a recent single-cell RNA-sequencing analysis showing that genes involved in the wounding response pathway are enriched in mouse embryonic fibroblasts that failed to reprogram (*25*). These analyses suggested that genes associated with reprogramming efficiency were indicative of other processes requiring dynamic cell transition states.

To further evaluate the role of associated genes in dynamic disease states, we tested whether genes associated with reprogramming efficiency could be predictive of survival and race in an independent cohort of 555 breast cancer patients: 262 self-reported African Americans and 293 self-reported White Americans (*26*). We tested the prediction ability for associated genes in a univariate fashion, by area under the curve (AUC) receiver operator characteristics (ROCs). To define genes that optimized prediction (AUC), genes were added one by one, according to their ranking (univariate, high to low), to the logistic model in Monte Carlo simulations to discriminate 5-year survival and self-reported race based on breast cancer gene expression. In addition to the logistic regression method, we used a 10-fold cross-validation to generate a logistic regression model and calculate the performance instead of using the Monte Carlo cross-validation (100-fold CV). Resulting gene sets had prediction power between 80 and 89% (Fig. 5, B and C, and table S5). Genes predictive of 5-year survival include those encoding proteins involved in cell adhesion and structure (*VCL*, *ACTB*, and *PDPN*), previously described tumor suppressive proteins (*UBE2QL1*, *RASSF3*, and *DAP*), as well as those not well described in the context of cancer (*LINC01521*, *SKIDA1*, and *RBMX2*) (table S5) (*27*–*31*).

There were many genes predictive of race of tumor origin (tables S5), including genes with both ancestry-dependent and ancestry-independent associations with reprogramming efficiency. Three of the genes identified in our analysis (*IL20RA*, *PPIL3*, and *MXRA7*) were previously found to be significantly differentially expressed in breast cancers from African Americans versus White Americans in TCGA, with *IL20RA* being one of the strongest predictors (*32*). Specifically, *IL20RA* and *PPIL3* are more up-regulated in individuals of self-reported Whites, whereas *MXRA7* is more up-regulated in those with self-reported Blacks. Of note, although extracellular matrix remodeling is known to be important in tumorigenesis, *MXRA7* (matrix remodeling–associated protein 7) has not been well studied in the context of cancer.

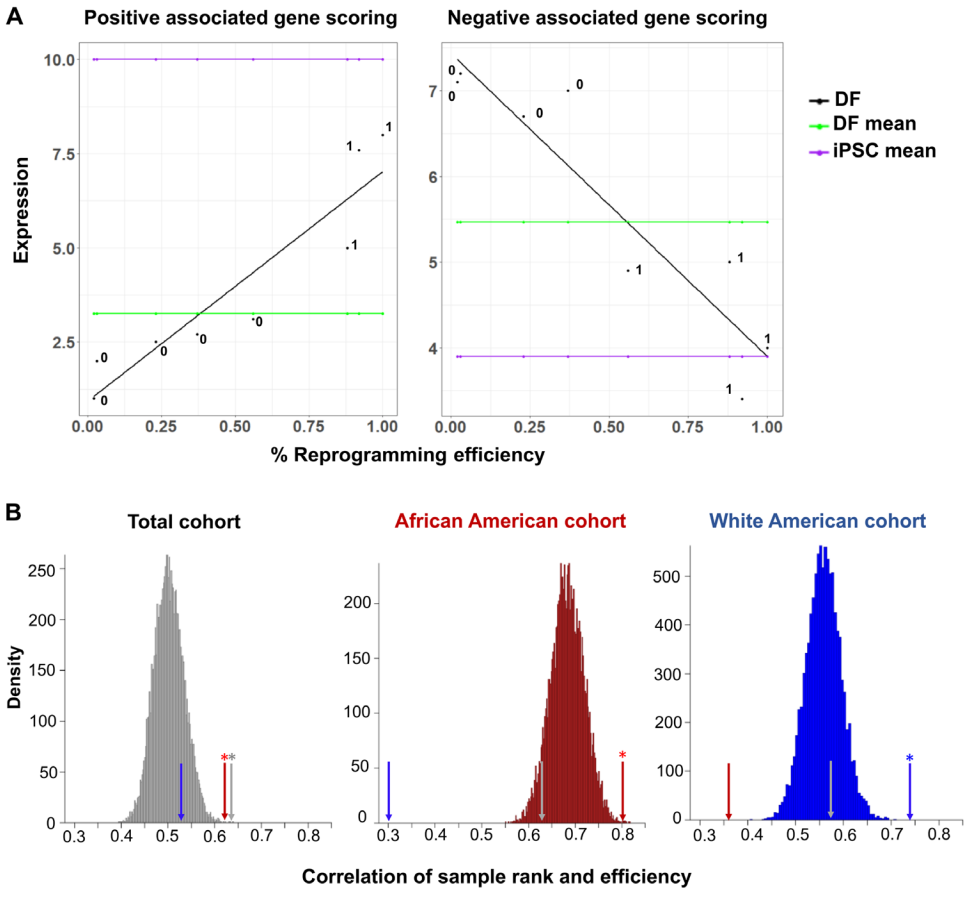

Fig. 4. Scoring by ancestry-dependent reprogramming efficiency improves rank score. (**A**) Schematic of binary scoring system used to rank samples on sum of expression of gene sets of interest. For a given fibroblast line (hypothetical examples shown in black) with expression levels of a positively associated gene above the cohort mean (green line) or expression levels of a negatively associated gene below the cohort mean, a score for that individual gene was 1. If expression levels did not meet these criteria, then a value of 0 was assigned. Total scores for all genes in a gene set of interest were summed and used to rank order samples. (**B**) Histograms of the distribution of the correlations between reprogramming efficiency and the relative sample ranks resulting from our scoring system when applied to 10,000 random sets of 750 genes, in the total cohort (gray) and race-specific subcohorts (red and blue). Arrows indicate the correlation values when the scoring system is applied to the top 750 associated genes (based on Spearman correlations) in the total cohort (gray), African American cohort (red), and White American cohort (blue).

To further test our ability to predict 5-year survival, we ran a GSEA analysis. We preranked genes expressed in the breast cancer cohort based on expression correlating with 5-year survival. We then ran GSEA to look for enrichment of reprogramming efficiency–associated genes. We found significant negative enrichment, indicating that reprogramming efficiency–associated genes are enriched for genes associated with 5-year survival (table S5). This further indicates that genes predictive of reprogramming efficiency are involved in other dynamic biological processes.

Prediction of new regulators of genes associated with reprogramming efficiency

Our previous work indicated that variations in expression of SWI/SNF chromatin remodeling complex subunits correlated with reprogramming efficiency; specifically, higher levels of mRNA encoding BRG1, BAF155, and BAF60a in iPSCs were associated with higher reprogramming (*15*). To determine whether SWI/SNF was an upstream regulator of reprogramming efficiency–associated genes, and to identify other potential upstream regulators that may enhance or block pluripotency, we used Ingenuity Pathway Analysis (IPA) using the set of genes for each ancestral group in our DF-iPSC cohort that met a Spearman significance cutoff $P \leq 0.01$ (table S6). SMARCA4 (BRG1) and SMARCD3 (BAF60c) were top predicted upstream regulators in associated genes in both the total and African American cohort (gray and red), but not the White American cohort (blue) (Fig. 6, A and B). However, SMARCA1, which encodes a SWI/SNF-related adenosine triphosphatase (ATPase), was predicted to be an upstream regulator exclusive to White Americans (Fig. 6, A and B).

We also found ancestry-dependent associated genes to differ in other potential upstream regulators. SYVN1 was one of the top upstream regulators that may be inhibited with higher reprogramming efficiency in African Americans, but not White Americans (Fig. 6, A and B). Conversely, there were several upstream regulators exclusive to White Americans, including NR3C1 and the 26S proteasome. We identified CST5 and miR-122 to also be upstream regulators of associated genes in the total cohort. Our laboratory recently demonstrated CST5 and miR-122 to be upstream regulators of genes down-regulated in response to proteasome inhibition (*33*). Combinatorial regulation by the 26S proteasome, CST5, and miR-122

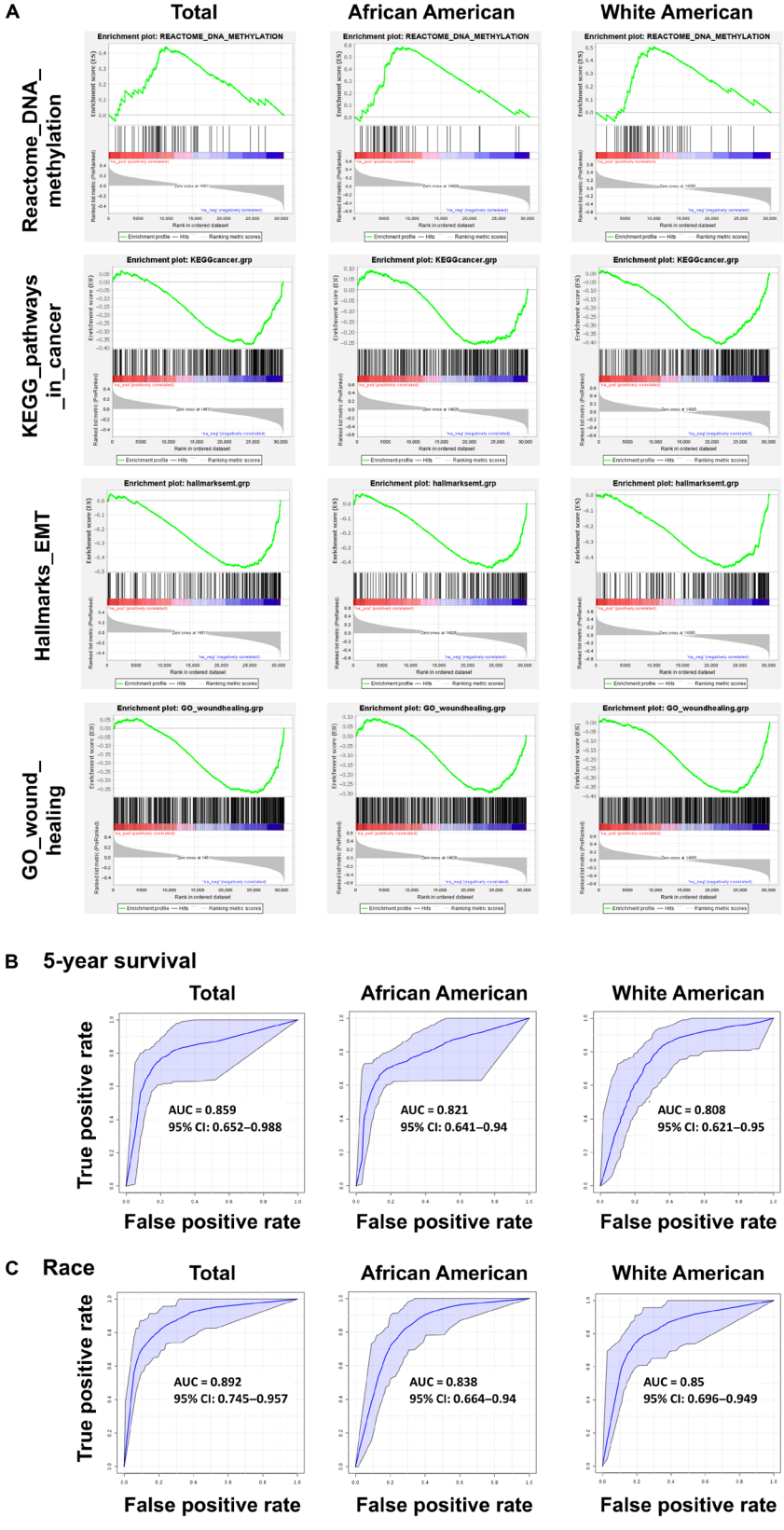

Fig. 5. Reprogramming efficiency–associated genes are involved in multiple dynamic processes. (**A**) Enrichment plots generated using preranked (Pearson) GSEA analysis for gene sets of interest. (**B**) Prediction ability for 5-year breast cancer survival or (**C**) race and reprogramming efficiency in univariate fashion, by AUC ROCs. The 95% confidence intervals (CIs) are shown.

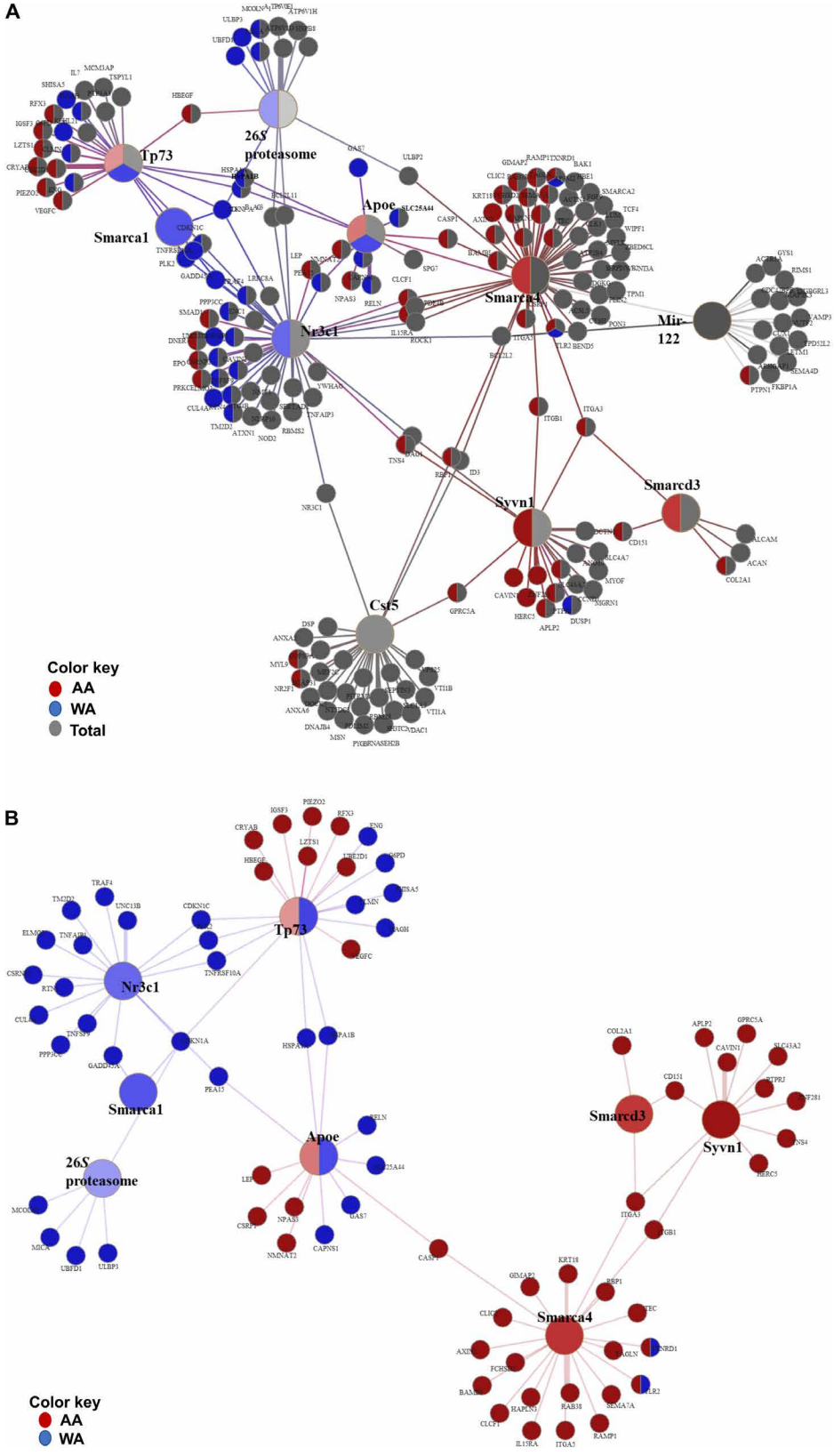

Fig. 6. Prediction of previously unknown regulators of genes associated with reprogramming efficiency. Concept network (Cnet) plots showing exemplar upstream regulators of efficiency-associated genes (full results are provided in table S5). (**A**) Regulators and target genes are color-coded by whether they are a hit/associated in the total cohort (gray), African American cohort (red), or White American cohort (blue). (**B**) Same as in (A) but only including information for the ancestry-dependent analyses.

may be an important determinant of pluripotent potential, and there may be some ancestry-dependent mechanisms of regulation.

APOE and TP73 were the only two upstream regulators predicted in all three comparisons: African American, White American, and total cohort. Targeted genes did not overlap between the two ancestries. However, with some exceptions (e.g., *CDKN1A*, *SHISA5*, *HSPA1B*, and *UBE2D1* were exclusively associated in one race), most targets approached significance in each race. This suggests that TP73 and APOE may be regulating both overlapping and unique gene sets depending on race.

DISCUSSION

Here, we present a comprehensive comparison of transcriptomic data from fibroblast-iPSC pairs obtained from 72 self-reported African American and White American individuals. Our study revealed both ancestry-dependent and ancestry-independent genes associated with the efficiency at which fibroblasts could be reprogrammed to iPSCs, suggesting that individual genetic heterogeneity can substantially affect reprogramming. Many genes associated with reprogramming efficiency would not have been identified without samples from different racial groups, indicative of the value in diversifying research models and considering genetic background while identifying factors that affect reprogramming efficiency. Our previous observation that African American samples as a group reprogrammed better overall compared to White Americans as a group may be explained by these ancestry-dependent differences (*15*), and our work lays the foundation for future investigation into ancestry-dependent mechanisms. Moreover, we add 36 transcriptome profiles of matched fibroblast-iPSCs derived from African American individuals to publicly available datasets, serving as a resource that can help address the lack of genetic data from individuals of non-European ancestry. iPSCs used to generate these datasets are available as models for future studies of disparate disease.

Our study provides many genes that may be involved in determining reprogramming efficiency. These include some of the strongest ancestry-independent associated genes, *KIF26A*, *TOR1B*, and *NSUN7*, all of which have demonstrated roles in development but have not been thoroughly studied in the context of pluripotency (*34*–*36*). Future studies using mouse or cell line models may be able to further clarify whether these play roles in stemness. While many of the top associated genes were independent of ancestry, many were ancestry dependent. For example, *GAS2*, *PLCE1*, and *GPRC5A* were uniquely correlated in African Americans, and none of these have known roles in regulating stem cells or pluripotency. We also identified many genes that have previous suggested roles in pluripotency, although we find that they are ancestry-dependent associations. For example, *NOMO2* was the strongest negatively associated gene in White Americans. *NOMO2* encodes a protein belonging to the nodal subfamily of transforming growth factor β (TGFβ) proteins and regulates crucial processes during embryonic development. Nodal signaling enhances pluripotency of ES (embryonic stem cells) and iPSCs, and NOMO is a nodal signaling antagonist (*37*, *38*). *WASHC2C* was the second highest negatively associated gene in White Americans, with no association in African Americans. *WASHC2C* encodes a WASH complex subunit, which has been demonstrated to be required for the differentiation of hematopoietic stem cells (*39*). While its requirement in stem cell differentiation has previously been established, we find a negative association with efficiency in samples from White Americans but not African Americans, although the average RNA expression decreases during reprogramming in both ancestries. This further demonstrates the need to consider ancestry, as not all research findings will apply to all backgrounds.

Our data suggest that genes involved in the regulation of intracellular transport, protein localization, and cytoskeletal organization are among the most significant transcripts negatively associated with reprogramming efficiency. In addition, genes involved in processes that require dynamic protein localization and restructuring of the cytoskeleton, namely, wound healing and cancer, were negatively enriched in our analysis. It is possible that this dynamic restructuring is indicative of the necessity of a MET during reprogramming versus an EMT during wound healing and tumor progression. While EMT has a recognized role in metastatic dissemination during tumorigenesis, it is a highly dynamic process, and the reverse MET is thought to be necessary for successful metastatic colonization (*24*). While the discovery that TWIST1 can induce both EMT and stem-like properties has suggested a link between the two concepts, other data suggest that cells with stem-like properties (i.e., "cancer stem cells") increase during the MET stage of metastasis (*24*). Differential cancer stem cell activity may explain racial disparities in health outcomes, and it was recently observed that individuals from Egypt had higher levels of breast cancer stem cells compared to those from England (*40*). These stem cells differed in both genetic content and drug resistance mechanisms. We speculate that diverse donor-derived iPSCs may serve as a model to better understand ancestral differences in cancer stem cells, and our identified ancestry-dependent genes associated with reprogramming efficiency may provide information on ancestry-dependent mechanisms of cancer stem cell formation.

One limitation of our study is that we were unable to account for environmental and lifestyle factors. Incorporating this additional information in the future will allow us to identify gene expression differences that are dependent on external variables. An additional limitation is that we were unable to account for genetic heterogeneity within subpopulations of our cohort. African Americans are generally more genetically heterogeneous compared to White Americans, and accounting for admixed substructure, including local ancestry, may provide greater insight into ancestry-dependent gene expression. Population-based genomics studies have begun to use local ancestry–adjusted models, which can improve power in admixed populations (*7*, *41*). However, in the absence of accounting for population substructure, we still gained insights into ancestry-dependent contributions of associated genes, many of which may control other important cell fate transitions.

We found that a subset of genes strongly correlated with reprogramming efficiency were able to predict both self-reported race of tumor origin and 5-year survival in an independent breast cancer cohort. Genes that are both associated with reprogramming efficiency and predictive of the self-reported race of tumor origin may encode proteins that function in ancestry-dependent mechanisms of cell state transitions. In addition, 5-year survival prediction ability was improved when considering ancestry-dependent associated genes in one specific self-identified race. Compared to White Americans, African Americans have higher breast cancer mortality rates in the United States, even when accounting for lifestyle factors and access to healthcare (*26*). Recently, it was demonstrated that clinically homogenous tumors had ancestry-dependent variability of significance of master transcriptional regulators as biomarkers, implicating differences in downstream network activity (*26*). Focusing on genes

with differential prognostic value may help to elucidate differences in molecular mechanisms of cancer progression, which will be critical to overcoming cancer health disparities.

We also found ancestry-dependent associated genes to differ in potential upstream regulators. *SYVN1* was one of the top upstream regulators, inhibited with higher reprogramming efficiency in African Americans but not White Americans. *SYVN1* encodes a protein involved in ER-associated degradation and the unfolded protein response (ER/UPR). It has recently been shown that transient activation of the ER/UPR stress response is critical for acquiring pluripotency (*18*), but the involvement of *SYVN1* in pluripotency has not been studied. We suggest that *SYVN1* and the ER/UPR are interesting targets for future study in the context of stem cell maintenance and pluripotency. Conversely, the 26*S* proteasome was activated with higher reprogramming efficiency in White Americans but not African Americans. While many associated genes down-regulated by the 26*S* proteasome were clearly negatively associated in White Americans only (e.g., *MCOLN1* and *MICA*), others were ancestry independent (e.g., *HSPB8* and *BAG3*). Our group recently demonstrated that CST5 and miR-122 are upstream regulators of genes down-regulated in response to proteasome inhibition (*33*), and both CST5 and miR-122 were predicted upstream regulators in the total analysis, with targets of both being largely negatively associated and ancestry independent, although there were some examples of ancestry-dependent targets. Our data suggest that the regulation of the proteasome and proteasome responsive genes by CST5 and miR-122 may be important determinants of pluripotent potential and that there may be some ancestry-dependent mechanisms of downstream function.

Widespread chromatin remodeling takes place during iPSC generation, and reorganization of chromatin structure is regulated by several adenosine triphosphate (ATP)–dependent chromatin remodeling complexes (*42*). Previously, we showed that up-regulation of specific BAF complex subunits, BRG1, BAF155, and BAF60a, in iPSCs directly correlated with reprogramming efficiency (*15*), suggesting that cell lines that can more efficiently up-regulate this specific BAF complex generally reprogram with higher efficiency. In our current study, we found that *SMARCA4* (BRG1) and *SMARCD3* (BAF60c) were predicted upstream regulators of genes associated with reprogramming efficiency. Regulation by SWI/SNF appears to be stronger in African American associated genes. Our data suggest that *SMARCA4*/BRG1 may be activated with higher reprogramming efficiency, in line with our observed increase in *SMARCA4* expression during reprogramming. Conversely, our data indicate that *SMARCD3*/BAF60c may be inhibited with higher reprogramming efficiency. BAF60c and BAF60a are mutually exclusive SWI/SNF subunits, and although BAF60a is known to be part of a stem cell–specific BAF complex, its specific contribution to chromatin architecture and complex function is unknown. Given that BAF60c is a predicted upstream regulator of genes associated with reprogramming efficiency, with potential inhibition in fibroblasts that reprogram with higher efficiency, understanding the relative contributions of the BAF60 mutually exclusive subunits will be important to understanding complex function in the acquisition and maintenance of pluripotency.

In addition to chromatin-based remodeling, other epigenetic changes that alter gene expression, including DNA methylation and posttranslational histone modifications, occur during transitions to pluripotency. Several studies have demonstrated that DNA methylation patterns vary by ancestry (*43*). It will be interesting to see whether interindividual heterogeneity of methylation affects variability in reprogramming efficiency. Last, while we suggest that fibroblasts that reprogram with high efficiency are primed at the RNA level, one important caveat is that the proteome remains to be evaluated. Like most biological conditions, pluripotency has multiple levels of regulation, and protein levels often do not correlate with steady-state levels of mRNAs in the transcriptome due to changes in mRNA processing, translation, posttranslational modifications, and protein turnover (*44*). To fully understand regulatory mechanisms and functional outcomes of the observed priming, it will be important to evaluate protein expression and posttranslational modification.

Overall, our study provides a comprehensive investigation of how ancestry and transcriptome heterogeneity can affect reprogramming of fibroblasts to iPSCs, and we emphasize the value of considering ancestry to support diversifying research models. We report newly associated genes and upstream regulators that may be involved in the acquisition and maintenance of pluripotency, many of which are involved in other cell fate transitions and therefore may be applicable to other systems. Future study of candidate genes and regulators identified may provide insight into novel mechanisms of ancestry-dependent and ancestry-independent regulation of dynamic cell state transitions as well as motivate additional studies for improvement of reprogramming.

MATERIALS AND METHODS

Study design

DFs and iPSCs used in this study have been previously described (*15*), and details on samples used can be found in table S1. The cohort used in this study consisted of 36 African American (17 female and 19 male) and 36 White American (19 female and 17 male) individuals. We did not find a correlation with reprogramming efficiency and sex (*15*). The total cohort has a median age of 32 years, with a range of 20 to 64 years. The African American cohort has a median age of 34, with a range of 21 to 52 years. The White American cohort has a median age of 31, with a range of 20 to 64 years. The African American cohort had reprogramming efficiencies ranging from 0.06 to 1.37%, with a median of 0.655%. The White American cohort had reprogramming efficiencies ranging from 0.02 to 1.13%, with a median of 0.455%. Our goal was to define transcriptomic heterogeneity that could be contributing to differences in reprogramming efficiency between individuals and between groups.

RNA sequencing

Total RNA was isolated using the Qiagen RNeasy Kit and quantified using the Qubit RNA HS Assay (Thermo Fisher Scientific). RNA quality was assessed using a 2100 Bioanalyzer instrument and an Agilent 6000 RNA Pico kit (Agilent Technologies). Average RIN was 8.6. For each sample, 500 ng of total RNA was used as input for preparation of whole-transcriptome ribosomal RNA (rRNA)–depleted libraries. An adapter-ligated library was prepared with the KAPA HyperPrep Kit (Kapa Biosystems, Wilmington, MA) using Bioo Scientific NEXTflex DNA Barcoded Adapters (Bioo Scientific, Austin, TX, USA) according to the Kapa-provided protocol.

rRNA was depleted by incubating total RNA with probes complementary to rRNA sequences. Following hybridization, RNase H was used to enzymatically degrade rRNA. Cleanup and deoxyribonuclease (DNase) digestion were performed using Kapa Pure Beads and DNase according to Kapa protocol.

rRNA-depleted samples were fragmented at 85°C for 4.5 min in the presence of magnesium before first- and second-strand synthesis and A-tailing reactions. NEXTflex DNA Barcoded Adapters (1.5 μM) were ligated to A-tailed complementary DNA (cDNA) with a unique barcode for each sample. Products were purified with Kapa Pure Beads, and eight cycles of Library Amplification were performed. Following amplification, a final library cleanup was performed, and library quantification and quality control (QC) were assessed using Qubit DNA HS Assay (Thermo Fisher Scientific) and an Agilent DNA HS kit on a 2100 Bioanalyzer instrument.

The resulting multiplexed sequencing libraries were used in cluster formation on an Illumina cBOT (Illumina, San Diego, CA, USA), and sequencing was performed using an Illumina HiSeq 2500 following Illumina-provided protocols for 2 × 150 base pair (bp) paired-end sequencing. Each transcriptome was sequenced to a target depth of 125 million reads.

The following mean raw reads were obtained: African American DFs, 129,571,450; White American DFs, 131,939,505; African American iPSCs, 132,501,335; and White American iPSCs, 134,394,164. Raw reads were aligned to hg19 using the STAR alignment tool (https://github.com/alexdobin/STAR). The following mean aligned reads were obtained: African American DFs, 123,315,343; White American DFs, 125,178,035; African American iPSCs, 123,148,312; White American iPSCs, 123,886,992.

Gene Ontology

GO was analyzed using Gorilla (http://cbl-gorilla.cs.technion.ac.il/), and broader categorical groupings were assigned on the basis of overlapping semantics and function.

Gene set enrichment analysis

GSEA v4.0.3 was run using GSEAPreranked, default settings with no collapsing. Because Spearman correlation coefficients resulted in duplicate values, input data were all expressed genes ordered by Pearson correlation coefficients, using values calculated with the total cohort and separate ancestries. The gene sets tested were those in the collection c2.cp from the MSigDB, as well as specific sets: GO wound healing, KEGG cancer, and PTEN pathways downloaded from MSigDB. Enrichment plots were generated using GSEA to demonstrate positive or negative enrichment of the gene sets of interest.

To look for enrichment of reprogramming efficiency–associated genes in the breast cancer cohort, input data were all expressed genes in the breast cancer cohort ordered by expression in samples with high 5-year survival (*26*). The gene sets tested were all reprogramming efficiency genes that were expressed in the breast cancer cohort, determined using the total iPSC cohort (942 expressed genes), the AA cohort (496 expressed genes), and the WA cohort (472 expressed genes).

Pathway enrichment analysis

IPA (Qiagen) was run using default settings for genes with Spearman correlations with $P \leq 0.01$ (0.388 for $n = 36$ and 0.274 for $n = 72$) for the total cohort combined and ancestries separately.

Multienrichment analysis was performed on IPA results using R version 3.6.2 and the multienrichjam package (version 17.900) (https://github.com/jmw86069/multienrichjam). Enriched pathways with an FDR-adjusted *P* value of less than 0.1 were used to generate a pathway-gene heatmap, which was clustered using Euclidian distance. Concept network (Cnet) plots were created using exemplar pathways from each cluster.

Statistical analysis

RNA sequencing

FeatureCounts from the Subread package was used to map reads to genes using the gene model GENCODE v27 (https://github.com/torkian/subread-1.6.1), and read counts were normalized using DESeq2 (https://github.com/mikelove/DESeq2). About 47,000 genes were detected in the DFs, and about 52,000 genes were detected in the iPSCs. A gene was considered expressed if there were at least 10 DESeq2 normalized counts in either DFs or iPSCs. This resulted in 30,494 genes expressed in the total cohort, 30,586 genes expressed in the African American cohort, and 30,365 genes expressed in the White American cohort. To test for cell type interactions with sex and ancestry, interaction terms were added to a DESeq2 model.

Principal components analysis was performed using R's prcomp function scale set to TRUE, and all other parameters were set to default. For pairwise comparisons, the average $\log_2$ fold change data from each by-group DESeq2 comparison were used. The union of the genes from the lists based on $\log_2$ fold change > 4.0 and FDR ≤ 0.05 cutoffs was considered. This gave 5760 genes. Pairwise differences were calculated for each of the resulting genes.

Reprogramming efficiency correlations

Pearson and Spearman correlations were calculated on the basis of DESeq2 normalized counts. DESeq2 was used with a model that had reprogramming efficiency as the response variable to quantify the effect of gene expression on reprogramming efficiency. Spearman values considered were 0.388 for $n = 36$ (race cohorts) and 0.274 for $n = 72$ (total cohort) (*45*).

Significance permutation tests were run by randomizing reprogramming efficiency across the cohort 10,000 times and determining the number of genes that met a given Spearman correlation significance threshold at each randomization. For this analysis, we only considered genes detected above noise in fibroblasts, defined by having at least 10 mean normalized counts, to avoid testing correlation on noise and to reduce effects of including nonresponsive genes during random permutations.

Primed genes were determined by first calculating the slope of the by-sample $\log_2$ fold change for each gene between DFs and iPSCs. Positively associated genes (positive Spearman) were considered primed if the cohort $\log_2$ fold change during reprogramming was positive (increased expression during reprogramming, adjusted $P \leq 0.05$), and the slope of the pairwise $\log_2$ fold change was negative (slope approaching 0). Negatively associated genes (negative Spearman) were considered primed if the cohort $\log_2$ fold change during reprogramming was negative and the slope of the pairwise $\log_2$ fold change was positive (slope approaching 0).

Rank scoring

Each fibroblast line was scored using binary indicators: If an individual line had expression levels of a positively associated gene above the cohort mean, or expression levels of a negatively associated gene below the cohort mean, then the score for the individual line for that gene was 1. If expression levels did not meet these criteria, then a value of 0 was assigned. Scores for positive and negative associated genes were summed and used scores to rank samples. The correlation between these ranks and reprogramming efficiency was then determined. Rank scores were determined with the top 750 associated genes in the total cohort, African American only cohort, and White American only cohort. To generate a correlation distribution of unassociated gene sets, samples were rank-scored as above using 10,000 random sets of 750 genes selected from all genes that met our

previously defined expression cutoff and that had Spearman correlations under $P \leq 0.05$ significance.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/47/eabc3851/DC1

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. Y. Buganim, D. A. Faddah, R. Jaenisch, Mechanisms and models of somatic cell reprogramming. *Nat. Rev. Genet.* **14**, 427–439 (2013).
2. D. E. Spratt, T. Chan, L. Waldron, C. Speers, F. Y. Feng, O. O. Ogunwobi, J. R. Osborne, Racial/ethnic disparities in genomic sequencing. *JAMA Oncol.* **2**, 1070–1074 (2016).
3. G. Sirugo, S. M. Williams, S. A. Tishkoff, The missing diversity in human genetic studies. *Cell* **177**, 26–31 (2019).
4. J. Dutil, Z. Chen, A. N. Monteiro, J. K. Teer, S. A. Eschrich, An interactive resource to probe genetic diversity and estimated ancestry in cancer cell lines. *Cancer Res.* **79**, 1263–1273 (2019).
5. S. E. Hooker Jr., L. Woods-Burnham, M. Bathina, S. Lloyd, P. Gorjala, R. Mitra, L. Nonn, K. S. Kimbro, R. A. Kittles, Genetic ancestry analysis reveals misclassification of commonly used cancer cell lines. *Cancer Epidemiol. Biomarkers Prev.* **28**, 1003–1009 (2019).
6. E. A. Chang, M. L. Tomov, S. T. Suhr, J. Luo, Z. T. Olmsted, J. L. Paluh, J. Cibelli, Derivation of ethnically diverse human induced pluripotent stem cell lines. *Sci. Rep.* **5**, 15234 (2015).
7. F. A. Tofoli, M. Dasso, M. Morato-Marques, K. Nunes, L. A. Pereira, G. S. da Silva, S. A. S. Fonseca, R. M. Costas, H. C. Santos, A. da Costa Pereira, P. A. Lotufo, I. M. Bensenor, D. Meyer, L. V. Pereira, Increasing the genetic admixture of available lines of human pluripotent stem cells. *Sci. Rep.* **6**, 34699 (2016).
8. A. D. Panopoulos, M. D'Antonio, P. Benaglio, R. Williams, S. I. Hashem, B. M. Schuldt, C. DeBoever, A. D. Arias, M. Garcia, B. C. Nelson, O. Harismendy, D. A. Jakubosky, M. K. R. Donovan, W. W. Greenwald, K. Farnam, M. Cook, V. Borja, C. A. Miller, J. D. Grinstein, F. Drees, J. Okubo, K. E. Diffenderfer, Y. Hishida, V. Modesto, C. T. Dargitz, R. Feiring, C. Zhao, A. Aguirre, T. J. McGarry, H. Matsui, H. Li, J. Reyna, F. Rao, D. T. O'Connor, G. W. Yeo, S. M. Evans, N. C. Chi, K. Jepsen, N. Nariai, F.-J. Müller, L. S. B. Goldstein, J. C. Izpisua Belmonte, E. Adler, J. F. Loring, W. T. Berggren, A. D'Antonio-Chronowska, E. N. Smith, K. A. Frazer, iPSCORE: A resource of 222 iPSC lines enabling functional characterization of genetic variation across a variety of cell types. *Stem Cell Rep.* **8**, 1086–1100 (2017).
9. X. Gao, J. J. Yourick, R. L. Sprando, Generation of nine induced pluripotent stem cell lines as an ethnic diversity panel. *Stem Cell Res.* **31**, 193–196 (2018).
10. A. Kyttälä, R. Moraghebi, C. Valensisi, J. Kettunen, C. Andrus, K. K. Pasumarthy, M. Nakanishi, K. Nishimura, M. Ohtaka, J. Weltner, B. Van Handel, O. Parkkonen, J. Sinisalo, A. Jalanko, R. D. Hawkins, N.-B. Woods, T. Otonkoski, R. Trokovic, Genetic variability overrides the impact of parental cell type and determines iPSC differentiation potential. *Stem Cell Rep.* **6**, 200–212 (2016).
11. H. Kilpinen, A. Goncalves, A. Leha, V. Afzal, K. Alasoo, S. Ashford, S. Bala, D. Bensaddek, F. P. Casale, O. J. Culley, P. Danecek, A. Faulconbridge, P. W. Harrison, A. Kathuria, D. McCarthy, S. A. McCarthy, R. Meleckyte, Y. Memari, N. Moens, F. Soares, A. Mann, I. Streeter, C. A. Agu, A. Alderton, R. Nelson, S. Harper, M. Patel, A. White, S. R. Patel, L. Clarke, R. Halai, C. M. Kirton, A. Kolb-Kokocinski, P. Beales, E. Birney, D. Danovi, A. I. Lamond, W. H. Ouwehand, L. Vallier, F. M. Watt, R. Durbin, O. Stegle, D. J. Gaffney, Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
12. L. V. Schnabel, C. M. Abratte, J. C. Schimenti, T. L. Southard, L. A. Fortier, Genetic background affects induced pluripotent stem cell generation. *Stem Cell Res. Ther.* **3**, 30 (2012).
13. K. L. Svenson, D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, E. J. Chesler, A. A. Palmer, L. McMillan, G. A. Churchill, High-resolution genetic mapping using the Mouse Diversity outbred population. *Genetics* **190**, 437–447 (2012).
14. Collaborative Cross Consortium, The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics* **190**, 389–401 (2012).
15. L. C. Mackey, L. A. Annab, J. Yang, B. Rao, G. E. Kissling, S. H. Schurman, D. Dixon, T. K. Archer, Epigenetic enzymes, age, and ancestry regulate the efficiency of human iPSC reprogramming. *Stem Cells* **36**, 1697–1708 (2018).
16. L. C. Boraas, J. B. Guidry, E. T. Pineda, T. Ahsan, Cytoskeletal expression and remodeling in pluripotent stem cells. *PLOS ONE* **11**, e0145084 (2016).
17. K. Iida, T. Takeda-Kawaguchi, M. Hada, M. Yuriguchi, H. Aoki, N. Tamaoki, D. Hatakeyama, T. Kunisada, T. Shibata, K. Tezuka, Hypoxia-enhanced derivation of iPSCs from human dental pulp cells. *J. Dent. Res.* **92**, 905–910 (2013).
18. M. S. Simic, E. A. Moehle, R. T. Schinzel, F. K. Lorbeer, J. J. Halloran, K. Heydari, M. Sanchez, D. Jullié, D. Hockemeyer, A. Dillin, Transient activation of the UPR^{ER} is an essential step in the acquisition of pluripotency during reprogramming. *Sci. Adv.* **5**, eaaw0025 (2019).
19. S. E. Vidal, B. Amlani, T. Chen, A. Tsirigos, M. Stadtfeld, Combinatorial modulation of signaling pathways reveals cell-type-specific requirements for highly efficient and synchronous iPSC reprogramming. *Stem Cell Rep.* **3**, 574–584 (2014).
20. H. Neiswender, S. Navarre, D. J. Kozlowski, E. K. Lemosy, Early craniofacial defects in zebrafish that have reduced function of a Wnt-interacting extracellular matrix protein, Tinagl1. *Cleft Palate Craniofac. J.* **54**, 381–390 (2017).
21. L. Sun, Z. Dong, H. Gu, Z. Guo, Z. Yu, TINAGL1 promotes hepatocellular carcinogenesis through the activation of TGF-β signaling-medicated VEGF expression. *Cancer Manag. Res.* **11**, 767–775 (2019).
22. R. Ho, B. Papp, J. A. Hoffman, B. J. Merrill, K. Plath, Stage-specific regulation of reprogramming to induced pluripotent stem cells by Wnt signaling and T cell factor proteins. *Cell Rep.* **3**, 2113–2126 (2013).
23. A. Paksa, J. Rajagopal, The epigenetic basis of cellular plasticity. *Curr. Opin. Cell Biol.* **49**, 116–122 (2017).
24. T. Brabletz, EMT and MET in metastasis: Where are the cancer stem cells? *Cancer Cell* **22**, 699–701 (2012).
25. L. Guo, L. Lin, X. Wang, M. Gao, S. Cao, Y. Mai, F. Wu, J. Kuang, H. Liu, J. Yang, S. Chu, H. Song, D. Li, Y. Liu, K. Wu, J. Liu, J. Wang, G. Pan, A. P. Hutchins, J. Liu, D. Pei, J. Chen, Resolving cell fate decisions during somatic cell reprogramming by single-cell RNA-seq. *Mol. Cell* **73**, 815–829.e7 (2019).
26. J. S. Byun, S. Singhal, S. Park, D. I. Yi, T. Yan, A. Caban, A. Jones, P. Mukhopadhyay, S. M. Gil, S. M. Hewitt, L. Newman, M. B. Davis, B. D. Jenkins, J. L. Sepulveda, A. De Siervi, A. M. Napoles, N. A. Vohra, K. Gardner, Racial differences in the association between luminal master regulator gene expression levels and breast cancer survival. *Clin. Cancer Res.* **26**, 1905–1914 (2020).
27. M. Tsuneki, M. Yamazaki, S. Maruyama, J. Cheng, T. Saku, Podoplanin-mediated cell adhesion through extracellular matrix in oral squamous cell carcinoma. *Lab. Invest.* **93**, 921–932 (2013).
28. M. Maziveyi, S. K. Alahari, Cell matrix adhesions in cancer: The proteins that form the glue. *Oncotarget* **8**, 48471–48487 (2017).
29. N. C. Wake, C. J. Ricketts, M. R. Morris, E. Prigmore, S. M. Gribble, A.-B. Skytte, M. Brown, N. Clarke, R. E. Banks, S. Hodgson, A. S. Turnell, E. R. Maher, E. R. Woodward, UBE2QL1 is disrupted by a constitutional translocation associated with renal tumor predisposition and is a novel candidate renal tumor suppressor gene. *Hum. Mutat.* **34**, 1650–1661 (2013).
30. T. Kudo, M. Ikeda, M. Nishikawa, Z. Yang, K. Ohno, K. Nakagawa, Y. Hata, The RASSF3 candidate tumor suppressor induces apoptosis and G_1–S cell-cycle arrest via p53. *Cancer Res.* **72**, 2901–2911 (2012).
31. T. Raveh, G. Droguett, M. S. Horwitz, R. A. DePinho, A. Kimchi, DAP kinase activates a $p19^{ARF}$/p53-mediated apoptotic checkpoint to suppress oncogenic transformation. *Nat. Cell Biol.* **3**, 1–7 (2001).
32. D. Huo, H. Hu, S. K. Rhie, E. R. Gamazon, A. D. Cherniack, J. Liu, T. F. Yoshimatsu, J. J. Pitt, K. A. Hoadley, M. Troester, Y. Ru, T. Lichtenberg, L. A. Sturtz, C. S. Shelley, C. C. Benz, G. B. Mills, P. W. Laird, C. D. Shriver, C. M. Perou, O. I. Olopade, Comparison of breast cancer molecular features and survival by African and European ancestry in the Cancer Genome Atlas. *JAMA Oncol.* **3**, 1654–1662 (2017).
33. H. K. Kinyamu, B. D. Bennett, P. R. Bushel, T. K. Archer, Proteasome inhibition creates a chromatin landscape favorable to RNA Pol II processivity. *J. Biol. Chem.* **295**, 1271–1287 (2019).
34. L. Chi, P. Delgado-Olguín, Expression of NOL1/NOP2/sun domain (*Nsun*) RNA methyltransferase family genes in early mouse embryogenesis. *Gene Expr. Patterns* **13**, 319–327 (2013).
35. L. M. Tanabe, C.-C. Liang, W. T. Dauer, Neuronal nuclear membrane budding occurs during a developmental window modulated by torsin paralogs. *Cell Rep.* **16**, 3322–3333 (2016).
36. R. Zhou, S. Niwa, N. Homma, Y. Takei, N. Hirokawa, KIF26A is an unconventional kinesin and regulates GDNF-Ret signaling in enteric neuronal development. *Cell* **139**, 802–813 (2009).
37. K. E. Galvin, E. D. Travis, D. Yee, T. Magnuson, J. L. Vivian, Nodal signaling regulates the bone morphogenic protein pluripotency pathway in mouse embryonic stem cells. *J. Biol. Chem.* **285**, 19747–19756 (2010).
38. C. Haffner, M. Frauli, S. Topp, M. Irmler, K. Hofmann, J. T. Regula, L. Bally-Cuif, C. Haass, Nicalin and its binding partner Nomo are novel Nodal signaling antagonists. *EMBO J.* **23**, 3041–3050 (2004).
39. P. Xia, S. Wang, G. Huang, P. Zhu, M. Li, B. Ye, Y. Du, Z. Fan, WASH is required for the differentiation commitment of hematopoietic stem cells in a c-Myc–dependent manner. *J. Exp. Med.* **211**, 2119–2134 (2014).
40. M. Kamal, E. H. O. Nafie, S. Elsers, S. Alanwar, R. Ibrahim, F. Farag, M. Mlees, B. M. Simões, K. Spence, A. Santiago-Gómez, M. L. Salem, R. B. Clarke, Ethnicity influences breast cancer stem cells' drug resistance. *Breast J.* **24**, 701–703 (2018).
41. E. R. Martin, I. Tunc, Z. Liu, S. H. Slifer, A. H. Beecham, G. W. Beecham, Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. *Genet. Epidemiol.* **42**, 214–229 (2018).

42. P. C. Scacheri, P. J. Tesar, iPSC reprogramming is not just an open and shut case. *Cell Stem Cell* **21**, 711–712 (2017).
43. F. Kader, M. Ghai, DNA methylation-based variation between human populations. *Mol. Genet. Genomics* **292**, 5–35 (2017).
44. M. Li, J. C. Izpisua Belmonte, Deconstructing the pluripotency gene regulatory network. *Nat. Cell Biol.* **20**, 382–392 (2018).
45. P. H. Ramsey, Critical values for Spearman's rank order correlation. *J. Educ. Stat.* **14**, 245–253 (1989).

Acknowledgments: We would like to thank members of the Archer group, the NIEHS Epigenetics and Stem Cell Biology Laboratory, and the NIEHS Integrative Bioinformatics group for ongoing discussions and advice. We are also grateful to M. Shi, S. Mantooth, and L. Wyrick for helpful discussion and support. Last, we thank R. Jothi, G. Hu, S. London, H. Kinyamu, J. Hoffman, and K. Gunn for critical and thoughtful evaluation of this manuscript. **Funding:** This research was supported by the Intramural Research Program of the National Institute of Environmental Health Sciences (Z01 ES071006-20) and the National Institute on Minority Health and Health Disparities. S.S. was supported by the NIH (U54GM128729). K.G. was supported by the National Cancer Institute (1R01CA253368-01). **Author contributions:** Conceptualization: L.S.B. and T.K.A.; methodology: L.S.B., B.D.B., J.M.W., D.C.F., and T.K.A.; formal analysis: L.S.B., B.D.B., J.M.W., P.R.B., and S.S.; investigation: L.S.B., J.Y., L.C.M., and L.A.A.; resources: S.H.S., D.C.F., K.G., and J.S.B.; data curation: L.S.B. and B.D.B.; writing (original draft): L.S.B. and T.K.A.; writing (review and editing): L.S.B., J.Y., B.D.B., J.M.W., P.R.B., S.S., J.S.B., A.M.N., E.J.P.-S., D.C.F., and T.K.A.; visualization: L.S.B., B.D.B., J.M.W., P.R.B., and S.S.; supervision: T.K.A., D.C.F., and K.G.; funding acquisition: T.K.A., A.M.N., E.J.P.-S., and K.G. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. Original sequencing data will be deposited in dbGaP. Cell lines are available upon request.

Submitted 26 April 2020
Accepted 2 October 2020
Published 20 November 2020
10.1126/sciadv.abc3851

Citation: L. S. Bisogno, J. Yang, B. D. Bennett, J. M. Ward, L. C. Mackey, L. A. Annab, P. R. Bushel, S. Singhal, S. H. Schurman, J. S. Byun, A. M. Nápoles, E. J. Pérez-Stable, D. C. Fargo, K. Gardner, T. K. Archer, Ancestry-dependent gene expression correlates with reprogramming to pluripotency and multiple dynamic biological processes. *Sci. Adv.* **6**, eabc3851 (2020).