ARTICLE

The in vivo genetic program of murine primordial lung epithelial progenitors

Laertis Ikonomou1,2*, Michael J. Herriges1,2, Sara L. Lewandowski1,2, Robert Marsland III3, Carlos Villacorta-Martin1, Ignacio S. Caballero1, David B. Frank4, Reeti M. Sanghrajka1,2, Keri Dame1,2, Maciej M. Kańduła5,6, Julia Hicks-Berthet7, Matthew L. Lawton1,2, Constantina Christodoulou2, Attila J. Fabian8, Eric Kolaczyk5, Xaralabos Varelas7, Edward E. Morrisey9, John M. Shannon10, Pankaj Mehta3 & Darrell N. Kotton1,2*

Multipotent Nkx2-1-positive lung epithelial primordial progenitors of the foregut endoderm are thought to be the developmental precursors to all adult lung epithelial lineages. However, little is known about the global transcriptomic programs or gene networks that regulate these gateway progenitors in vivo. Here we use bulk RNA-sequencing to describe the unique genetic program of in vivo murine lung primordial progenitors and computationally identify signaling pathways, such as Wnt and Tgf-β superfamily pathways, that are involved in their cell-fate determination from pre-specified embryonic foregut. We integrate this information in computational models to generate in vitro engineered lung primordial progenitors from mouse pluripotent stem cells, improving the fidelity of the resulting cells through unbiased, easy-to-interpret similarity scores and modulation of cell culture conditions, including sub-stratum elastic modulus and extracellular matrix composition. The methodology proposed here can have wide applicability to the in vitro derivation of bona fide tissue progenitors of all germ layers.
 pluripotent stem cell (PSC)-based systems offer the possibility of de novo somatic cell derivation through directed differentiation, a multitask process recapitulating developmental milestones\(^1\). This methodology relies heavily on prior knowledge of developmental pathways and processes within the tissue/organ of interest, and incomplete developmental knowledge can be a significant impediment to the establishment of efficient directed differentiation protocols\(^2\).

In the case of lung specification, elucidation of the role of Bmp and Wnt signals through loss-of-function studies in murine development\(^3\–\(^5\) has led to the development of growth factor cocktails for derivation of lung progenitors from mouse and human PSCs\(^6\–\(^9\)\). Similarly, the in vitro derivation of ciliated cells\(^10\) as well as proximal and distal lung progenitors\(^12\,\(^13\) were made possible by in vivo studies of Notch signaling in ciliated cell differentiation\(^14\) and Wnt signaling in lung proximal-distal patterning\(^15\–\(^17\), respectively.

An important intermediate population in lung lineage derivation from PSCs is the lung epithelial primordial progenitor population, i.e., the first cells within the foregut that are specified to a lung epithelial cell fate, but do not yet express markers of lung epithelial differentiation, such as surfactant genes. Nkx2-1, the earliest marker of lung fate, is first expressed in the prospective lung domain of the gut tube at around embryonic (E) day 9.0 during mouse embryogenesis\(^18\). It is also expressed in the prospective thyroid domain and in the developing forebrain\(^19\). Previous work with mouse and human PSC lines has therefore used in vitro Nkx2-1 expression as a marker of lung progenitors\(^6\–\(^9\),\(^12\,\(^13\),\(^20\) with both distal alveolar and proximal airway epithelial competence. In addition, we still need to study early thyroid and lung patterning to use microdissection techniques that inevitably lead to isolation of mixed populations\(^22\). Furthermore, the equivalent stage in human development is not readily accessible and in vivo information on the lung primordium, even from other species, would be highly informative for human PSC-based lung differentiation studies.

Here, we report the identification, purification, and global transcriptomic analysis of the earliest detectable in vivo mouse lung progenitors. We compare the genetic program of these developing lung progenitors to foregut endodermal precursors purified just prior to the onset of expression of the Nkx2-1 program and contrast these programs to those of the Nkx2-1\(^+\) lineages that compose the developing forebrain and thyroid. These comparisons delineate the unique genetic program of the lung primordium at the moment of lung lineage specification in vivo and suggest pathways that regulate the emergence of lung fate within the developing foregut endoderm. We then mine the global transcriptome of the in vivo lung primordium to improve the critical stage of in vitro lung specification in a mouse PSC-based system. We show that modification of cell culture conditions, including biomechanical properties of the culture substrate, can enhance the epithelial character of PSC-derived mouse lung progenitors. By applying computational methods for classifying cell fate consisting of Linear Algebra-based projections, we demonstrate the enhanced fidelity of progenitors derived under improved conditions as reflected by increased projection scores onto the in vivo lung epithelial primordial population. Thus, our approach delineates the unique genetic program of the earliest primordial lung progenitors and underscores the importance of applying such an in vivo roadmap to guide the specification of primordial organ progenitors that may be engineered in vitro from alternate sources, such as PSCs.

### Results

**Lung epithelium is derived from Nkx2-1\(^+\) primordial progenitors.** We first tested whether lung epithelial lineages are derived from lung primordial progenitors in mouse development. Prior reports have characterized the lung epithelia that express Nkx2-1\(^23\) and demonstrate that Nkx2-1 lineage labeling of specific lung cell types such as pulmonary neuroendocrine cells\(^24\) and type II alveolar epithelial cells (AECs)\(^25\) is evident when using a transgenic Nkx2-1\(^\text{Cre}\) driver mouse line\(^26\) (Supplementary Fig. 1A). In our hands, this line labeled both proximal and distal lung epithelia (Supplementary Fig. 1B), as well as thyroid and forebrain cells. However, labeling in the lung did not occur until E11.5 (Supplementary Fig. 1C) making this line unsuitable for marking the progeny of E9.0 lung primordial progenitors. Next, we employed a mouse knock-in line, Nkx2-1\(^\text{CreERT2}\)\(^27\), and crossed it with the R26\(^\text{RtTA}\)\(^28\) line (Fig. 1A). In the resulting inducible mouse line (Nkx2-1\(^\text{Cre}\)), Nkx2-1-expressing cells and their progeny should indelibly be marked by nuclear enhanced green fluorescence protein (hereafter nGFP) upon tamoxifen administration. Indeed, when timed-pregnant mice were exposed to tamoxifen before E9.0 (Fig. 1B; Supplementary Fig. 1D, E), nGFP-expressing cells throughout the developing lung epithelium were observed at E14.5 and E18.5. GFP expression was strictly localized to the epithelial compartment, as shown by co-staining with the pan-epithelial marker EPCAM (Supplementary Fig. 1E) and NKKX2-1 (Fig. 1C).

### Faithful and specific lung expression of the Nkx2-1\(^\text{GFP}\) reporter

Having established that proximal and distal lung epithelia are derived through an Nkx2-1\(^+\) lung epithelial progenitor intermediate, we then sought to develop an approach for the identification and isolation of Nkx2-1\(^+\) primordial progenitors as they emerge from foregut endodermal precursors. We previously reported the generation of a mouse line carrying a cytoplasmic GFP reporter construct targeted to the Nkx2-1 locus (hereafter Nkx2-1\(^\text{GFP}\)) (Fig. 1D)\(^7\), and we analyzed the kinetics of Nkx2-1\(^\text{GFP}\) expression in this mouse from adulthood back to endodermal gut tube formation (Fig. 1E, F). Nkx2-1\(^\text{GFP}\) expression in the prospective lung domain was first detectable at E9.0 (18–23 somites), consistent with the time of lung lineage specification. Prior to E9.0, GFP was expressed only in the thyroid and forebrain domains, consistent with earlier initiation of the specification of these Nkx2-1\(^+\) lineages\(^29\). GFP reporter expression recapitulated endogenous Nkx2-1 transcript expression in all three embryonic domains (forebrain, thyroid, and lung) (Fig. 1D).

Within the lung, GFP expression continued exclusively in the epithelium throughout development, consistent with previously described Nkx2-1 expression patterns\(^30\). Postnatally, GFP-expressing nuclear NKX2-1 protein and markers of either club cells (SCGB1A1), basal cells (P63), or type II AECs (SFTPC) (Fig. 1F). GFP expression in Type I AECs (PDPN) and ciliated...
cells (acetylated α-tubulin) was more variable, with low to non-detectable GFP in these cell types (Fig. 1f). The immunostaining data were corroborated by the enrichment of club (Scgb1a1, Scgb3a2) and Type II AEC marker (Sftpa, Sftpb, Sftpc, and Abca3) gene expression in sorted NKX2-1GFP+ EPCAM+ cells (Fig. 1g, h). These data correlated well with prior reports that have established robust expression of NKX2-1 in postnatal club cells and Type II AECs and studies that demonstrate NKX2-1 expression in both early developing P63+ lung epithelia as well as postnatal mature P63+ airway basal cells31,32. Interestingly, a recent report indicated that NKX2-1 expression is required for both Type I AEC development and cell-fate maintenance33. Taken together, these data indicate that GFP expression recapitulates endogenous Nkx2-1 expression faithfully and specifically in the Nkx2-1GFP mouse. Therefore, this mouse can be used to sort to purity lung epithelial progenitors at the moment of lung specification, E9.0.

Distinct genetic program of lung primordial progenitors. Next, we sought to understand the genetic programs of the lung, forebrain, and thyroid progenitors during mouse development. We purified Nkx2-1GFP+ progenitors from the forebrain, lung, and thyroid by flow cytometry and analyzed their global...
transcriptomes by RNA-Seq (RNA-Seq) (Fig. 2a; Supplementary Fig. 2A). On average, we identified 100 GFP+primordial lung progenitors per embryo at E9.0 with higher numbers (~1300 cells) of GFP+forebrain cells accessible at this time point. Due to the low number of thyroid progenitors at E9.0 (~50 purified cells/embryo), we purified epithelial thyroid progenitors at a later time point (E13.5) in order to extract sufficient RNA for analysis (Fig. 2c). Highly enriched Nkx2-1 expression in sorted Nkx2-1-GFP+ cells at three time points confirmed the high purity of the sorts as well as the specificity of the reporter by both RT-qPCR and RNA-Seq (Supplementary Fig. 2B; Fig. 2d, respectively). On average, forebrain cells expressed higher levels of Nkx2-1 transcripts compared to E9.0 lung and E13.5 thyroid. As two alternative Nkx2-1 transcripts have been reported, one including all three exons and one including exons 2 and 3, we mapped Nkx2-1 sequencing reads on the Nkx2-1 locus (Fig. 2e). No obvious difference of transcript distribution was found between the three Nkx2-1-expressing populations.

Since Nkx2-1+ lineages in the lung or thyroid arise from foregut endodermal precursors and Nkx2-1+ forebrain originates from the ectodermal germ layer, we used established flow cytometry sorting algorithms 35 to prepare comparator RNA-Seq data sets from purified embryonic foregut endoderm (ENDM1+ EPCAM+) and ectodermal (ENDM1 EPCAM-) populations isolated from E8.25 embryos (Fig. 2a; Supplementary Fig. 2D). Enrichment of Sox17 expression in the ENDM1+ EPCAM+ fraction (Supplementary Fig. 2C), confirmed the specificity of the sort. All the populations used in RNA-Seq analysis are shown in Fig. 2b.

Lung primordial progenitors appear to possess a genetic program quite distinct from both thyroid and forebrain progenitor populations as well as foregut endoderm, their precursor population (Fig. 2f). To further understand the distinct genetic programs of embryonic Nkx2-1+ progenitors, we performed pairwise comparisons between these progenitors and their precursor populations (foregut endoderm for lung and thyroid, ectoderm for the forebrain). As shown in Fig. 3a, different sets of genes were differentially expressed between Nkx2-1+ progenitors and pre-specified populations (false discovery rate (FDR) < 0.05, average expression > 0 and |log2(fold change)| > 2).

Nine hundred eighty-five genes were differentially expressed between lung primordial progenitors and foregut endoderm. Those upregulated in the lung primordium included Nkx2-1, transcription factors (TFs) expressed in early lung epithelium (Foxp2, Irx2) 36,37 and specific Hox genes (Hoxa2, Hoxa4). Definitive endoderm, primitive streak, and node marker genes such as T, Eomes, Sox17, and Nodal were downregulated in the lung primordium, consistent with lineage specification. Of note, Dpp4 (Gsd2) that we have previously published as a negative selection marker of human PSC-derived lung progenitors 33 was significantly downregulated in the lung primordial cell population. Regarding the thyroid genetic program, 2340 genes were differentially expressed between E13.5 thyroid epithelium and foregut endoderm. Those upregulated in thyroid progenitors included the Tfs Nkx2-1, Pax8, Foxe1, and Hhex that form an integrated network controlling thyroid cell fate 38,39. Lastly, the genetic program of forebrain specification consisted of 1233 genes differentially expressed between E9.0 Nkx2-1+ forebrain progenitors and ectoderm. Among the genes upregulated in the former were Otx1/2, Dbx1, Six3, Lhx5, which have well-established roles in forebrain development 39,40. Interestingly, Foxa2 was also upregulated in the E9.0 forebrain, signifying that Nkx2-1/Foxa2 co-expression in vitro is not always indicative of lung endodermal progenitors, as has been widely assumed.

Pairwise comparison between the two Nkx2-1+ endodermal progenitor cell populations showed differential expression of 1257 genes (Fig. 3b). Genes upregulated in the thyroid progenitor population included Pax8, Hhex, and Foxe1 as well as genes including C444, Prlr, Cldn, Scar5, S100a5, and Ccak that are enriched in the E10.5 thyroid bud 41. Nkx2-1+ was also upregulated, consistent with higher relative expression levels in thyroid compared to lung epithelia (Fig. 2d; e; Supplementary Fig. 2B). Wt target genes (Left1 and Nkd2) were upregulated in the lung epithelial progenitors, reflecting the distinct role of Wnt signaling in lung vs. thyroid specification 9,21.

Having outlined the distinct temporal genetic programs of the lung, thyroid, and forebrain progenitors, we then tried to define tissue-specific transcriptomic signatures in an unbiased manner by comparing the three Nkx2-1+ populations (false discovery rate (FDR) < 0.05, average expression > 0 and |log2(fold change)| > 3) (Fig. 3c). We found 679 genes with significant gene expression changes, organized in three clusters, outlined in Fig. 3c. The “thyroid” cluster contained two of the key thyroid Tfs (Pax8, Hhex), markers of early thyroid epithelium (Prlr) 9,22, as well as putative thyroid markers described by us in mouse PSC-derived thyrocytes (Aqp1, Col25a1) 43. The “lung” cluster contained Gap43/4, Irx2, Hoxb8, and Cpm, a marker of human PSC-derived ventralized Nkx2-1+ foregut progenitors 44, Ankrd1, a canonical target of the Hippo pathway, was found in this cluster and also among the genes that distinguish the lung primordium from foregut endoderm (Fig. 3a, c). On the other hand, the “forebrain” cluster contained genes (Dbx1, Otx1/2, Lhx5, Foxd1, and Six3) that distinguish forebrain from pre-specified ectoderm
(Fig. 3a) and that have also been identified by us as part of the transcriptional signature of in vitro human PSC-derived NKK2–1GFP+ putative forebrain progenitors\(^{29}\).

This analysis demonstrates the distinct genetic programs of Nkk2–1+ developing lineages and how these programs temporally differ from cells within each relevant germ layer.
Fig. 3 Gene expression profiles of Nkx2-1+ endodermal and ectodermal populations. a Heatmaps containing row-normalized z-scores and representing unsupervised hierarchical clustering of samples analyzed in population RNA-Seq. Each heatmap depicts the pairwise comparison of specified endodermal or ectodermal Nkx2-1+ populations and their precursor stage (foregut endoderm and ectoderm, respectively). The number of genes used to create each heatmap appears underneath the heatmap. The cutoff criteria used define the number of genes for each heatmap are: false discovery rate (FDR) < 0.05, average expression > 0 and |log2(fold change)| > 2. b Heatmap of differentially expressed transcripts between the two Nkx2-1+ endodermal progenitor cell populations (lung and thyroid). The same cutoff criteria as in (a) were used. c Heatmap of differentially expressed transcripts across the three Nkx2-1+ cell populations (tissue-specific ‘signatures’). The same cutoff criteria as in (a) were used with the only difference of |log2(fold change)| > 3.
Wnt, Hippo, and Tgf-β pathways in in vivo lung specification. To delve into higher order biological processes that are unique to lung specification, we performed gene set enrichment analysis (GSEA) for all three Nkx2-1+ embryonic progenitors and their respective pre-specified populations (Fig. 4a; Supplementary Fig. 3A). Several pathways/processes appear to be common in either all three progenitor populations (e.g., “Circadian Entrainment”), or in two out of three populations (e.g., “Cytokine–Cytokine Receptor Interaction” in the lung primordium and thyroid). On the other hand, certain signaling pathways were progenitor-specific (seven for the lung, ten for thyroid, and one for forebrain, Supplementary Fig. 3B). Wnt, Hedgehog, Tgf-β superfamily, and Hippo pathways were upregulated exclusively in lung primordial progenitors upon
specification, whereas the PI3-kinase (PI3K) pathway was upregulated in specified thyroid progenitors (Fig. 4a; Supplementary Fig. 3C). Upregulation of the Hippo pathway in the lung domain was unlikely to be an artifact of tissue digestion leading to release of junctional Yap/Taz reserves as this pathway was not upregulated in either thyroid epithelium or forebrain, tissues that underwent the same processing. Moreover, several pathways that were downregulated upon lung specification were related to lipid metabolism (Fig. 4a; Supplementary Fig. 3A).

To gain further insights into these findings, we compared the expression of select Hippo-, Tgf-β superfamily-, and Wnt-related genes in the lung primordium and foregut endoderm (Fig. 4b). Bmp receptor (Bmpr1b, Bmpr2) expression was upregulated in the lung primordium along with several genes involved in inhibition of Tgf-β signaling or bioavailability (Dcn, Tgfbi, Ltbp1) raising the possibility that both Bmp signaling activation and Tgf-β signaling inhibition are involved in lung specification. The upregulation of several direct Wnt targets (Nkd2, Vcan, Nrcam, Twist1) and downregulation of Wnt inhibitors (Dkk1/3 45 and Sfrp5) point to activation of Wnt signaling within the lung primordium, but not the thyroid. Finally, Hedgehog signaling was also found to be differentially regulated (Supplementary Fig. 4A).

To functionally evaluate our bioinformatics findings, we combined mouse foregut explant culture with pharmacological or genetic manipulation of specific pathways (Fig. 4c, d). Previous work both in vivo and in vitro has shown that Wnt is required for the lung, but not thyroid specification, Bmp signaling is involved in specification of both endodermal domains3–5,9,21, and Fgf-dependent PI3K signaling is required for thyroid, but not lung specification21. We first used a Shh driver to delete β-catenin (ShhCre, βcat^floxed/floxed), thereby abolishing β-catenin-dependent Wnt signaling within the majority of the anterior foregut domain (Fig. 4c). As expected, this resulted in elimination of the Nkx2-1\(^+\) lung epithelial domain, similar to prior findings using genetic Wnt loss-of-function mouse models43. Although the Nkx2-1\(^+\) thyroid domain was not affected, based on prior work profiling Shh kinetics in thyroid organogenesis46, it is likely that the ShhCre driver is not expressed in the thyroid primordial zone at the stage of thyroid lineage specification (Fig. 4c). Therefore, we applied the well-documented Wnt inhibitor, XAV-93947, to mouse foregut cultures at this developmental stage to test the effect of Wnt loss of function throughout the foregut and observed the same phenocopied effect of abrogated lung specification while retaining thyroid specification (Fig. 4c). Surprisingly, absence of Wnt signaling in vivo at the moment of lung specification caused expansion of the Pdx1\(^+\) domain (Supplementary Fig. 4B). We then asked whether Shh signaling was involved in lung specification as implied by our bioinformatics analysis. Use of cyclopamine to block Hedgehog signaling at 6–8 somite stage indeed resulted in lung agenesis, without affecting thyroid specification (Fig. 4d). We were able to rescue lung specification in the presence of cyclopamine by addition of CHIR99021, a specific GSK-3β inhibitor that activates Wnt signaling by stabilizing β-catenin, indicating that Wnt acted downstream of Shh signaling. These findings are in keeping with prior work that established that endodermally secreted Shh acts indirectly by inducing hedgehog signaling in adjacent foregut mesenchymal cells which then secrete canonical Wnt ligands, which in turn act on the foregut endoderm to promote lung epithelial lineage specification48. In contrast to a requirement for Wnt signaling, we found that addition of several FGFs (FGF2, FGF7, and FGF10) to cyclopamine-treated foregut explants did not rescue lung specification (Fig. 4d), validating our previous findings9 that FGF signaling is neither necessary nor sufficient for lung lineage specification in vitro.

Overall, our pathway analysis showed direct activation of the Bmp and Wnt pathways within the specified mouse lung primordium, a finding that is in keeping with both in vivo loss-of-function studies3–5 and in vitro minimal signaling requirements for PSC derivation of lung primordial progenitors8. As the Hippo pathway was found to be differentially regulated at this early stage of lung development, we performed YAP/TAZ staining in the foregut endoderm at around E9.0 (Supplementary Fig. 4C). YAP/TAZ was cytoplasmic in most Nkx2-1\(^+\) cells but nuclear in the adjacent foregut, raising the possibility that transient YAP/TAZ activation immediately precedes lung specification though previous studies have suggested Yap/Taz dispensability for lung specification49. Lastly, our pathway analysis supports the apparent differential requirement for PI3K signaling in determination of thyroid fate and its dispensability for lung fate21.

**Similarity model reveals deficiencies in in vitro lung specification.** Having defined the genetic program of in vivo mouse primordial lung progenitors and the signaling pathways that are involved in lung specification in vivo, we next revisited our published mouse lung-directed differentiation protocol to assess the resemblance of our PSC-derived lung progenitors to in vivo lung primordial progenitors9.

To this end, we employed our own computational approach, namely Linear Algebra Projections (LAP), which we had successfully applied in our previous work43. As this method relies on global gene expression of in vitro populations and their comparison to a reference basis of in vivo cell types, we first updated the reference basis of 61 cell types used in Lang et al.50 to include our E8.25 ectoderm, E8.25 foregut endoderm, E13.5 Nkx2-1\(^+\), and E9.0 Nkx2-1\(^+\) lung and forebrain bulk RNA-Seq data as well as an E16.5 lung epithelial cell data set22 (six additional data sets). We also included single-cell data from E9.0 cells from mixed lineages captured separately from dissected thyroid, lung, and forebrain domains (eight clusters, see the Methods section) as well as single-cell RNA-Seq data from flow cytometry-enriched lung Nkx2-1\(^+\) cells at E13.5 (five clusters), E15.5 (five clusters) and E17.5 (six clusters)52 (Supplementary Fig. 6). Our analysis showed that the single-cell clusters from the E9.0 lung and thyroid domains were almost entirely Nkx2-1\(^+\) as expected given the rarity of GFP\(^+\) progenitors at this
developmental stage. Therefore, these comparators did not overlap with the Nkx2-1+ populations studied by bulk RNA-Seq. Thus, we created an updated reference basis of 91 in vivo cell populations, including mouse embryonic and adult populations from diverse stages, domains, and cell types.

As detailed in the methods, LAP scores are robust to comparisons across sequencing platforms (i.e., RNA-Seq versus microarrays). To further assess the robustness of the LAP method, following the inclusion of additional cell types and single-cell data, we re-analyzed the data from our previous work that had shown mouse ESC-derived thyroid progenitors have strong similarity to E13.5 mouse thyrocytes (Supplementary Fig. 7B). Day(D)14 thyroid progenitors projected strongly on E13.5 Nkx2-1+ thyrocytes, whereas D1 cells (cells exiting pluripotency) had a high undifferentiated-ESC similarity score.

We then proceeded with LAP analysis of the genome-wide microarray data of Nkx2-1+ cell population derived in our minimal lung specification media (Wnt3a/Bmp4/RA) on 2D gelatin (2D-Nkx2-1+) (Fig. 5a). This population expressed key genes of lung endoderm identity as previously published, but we surprisingly found that also have high projection scores for mesenchymal-like cell populations such as mouse embryonic fibroblasts (MEFs), and a low score when compared with the E9.0 Nkx2-1+ lung primordium program, indicating that the lung specification stage might be suboptimal (Fig. 5a).

We then proceeded with LAP analysis of the genome-wide microarray data of Nkx2-1+ cell population derived in our minimal lung specification media (Wnt3a/Bmp4/RA) on 2D gelatin (2D-Nkx2-1+) (Fig. 5a). This population expressed key genes of lung endoderm identity as previously published, but we surprisingly found that also have high projection scores for mesenchymal-like cell populations such as mouse embryonic fibroblasts (MEFs), and a low score when compared with the E9.0 Nkx2-1+ lung primordium program, indicating that the lung specification stage might be suboptimal (Fig. 5a).

We thus sought to troubleshoot our protocol for differentiating PSCs into lung primordial progenitors, focusing specifically on approaches that might reduce the aberrant mesenchymal program present in our in vitro engineered cells. As the “ADHERENS JUNCTION” pathway was involved in in vivo lung specification (Fig. 4a) and adherens junctions formation has been implicated in PSC cell fate decisions, we hypothesized that the fidelity of in vitro-derived lung progenitors could be improved by appropriate modulation of the substratum biomechanical properties, which could minimize the mesenchymal program and preserve an Epacm+ epithelial program. We further hypothesized this process might depend in part on the main epithelial adherens junction cadherin, CDH1.

To this end, we used the Nkx2-1mCherry mouse embryonic stem cell (ESC) line to specify Nkx2-1+ lung progenitors in Wnt3a/Bmp4/RA containing media on either gelatin-coated (2D Gelatin) or thick Matrigel-coated (3D Matrigel) condition (Fig. 5b). We noticed higher frequency of epithelial-like Nkx2-1mCherry+ colonies in the 3D vs. the 2D condition as early as D8 of the protocol (D2 of the specification stage) and this pattern was maintained by D14, which is the time point for lung primordial progenitor purification (Fig. 5c, d). Lung specification under 3D conditions brought about an increase in frequency of Nkx2-1mCherry+Epacm+ epithelial lung progenitors compared with 2D conditions and increased the yield of such progenitors per input undifferentiated (D0) ESC (Fig. 5e, f). When 3D Matrigel−Nkx2-1mCherry+Epacm+ cells were sorted to purity on D13–14 and embedded in Matrigel in the presence of lung differentiation media, they maintained high Epacm expression and gave rise to a variety of respiratory lineages, as shown by high expression levels of proximal (Sgb3a2, Sgb1a1, p63) and distal (Sfpf, Sftp8) lung markers and absence of thyroid markers (Pax8) (Fig. 5g). To investigate the implication of adherens junctions in the observed phenotype (Fig. 5d), we specified lung progenitors on three different substrata (2D-Gelatin, 2D-Matrigel (thin Matrigel layer), and 3D-Matrigel), and co-stained for the main epithelial adherens junction cadherin, E-cadherin (CDH1), and NKKX2-1 (Supplementary Fig. 5A). E-cadherin+ clusters contained the majority of NKKX2-1+ cells and their highest frequency was observed in the “3D Matrigel” conditions, indicating that both Matrigel and a low elastic modulus were probably required for the increased derivation of Nkx2-1mCherry+Epacm+ lung progenitors. When adherens junction formation was inhibited by repeated addition of an anti-CDH1 monoclonal antibody during lung specification on “3D Matrigel”, there was concomitant decrease in the Nkx2-1mCherry+Epacm+ percentage, implying that adherens junction formation was necessary for the specification of the putative lung primordial epithelial population (Supplementary Fig. 5B, C).

Taken collectively, these data demonstrate that in vitro application of information gleaned from the in vivo characterization of lung primordial progenitors leads to identification of deficiencies in the derivation of lung multipotent epithelial progenitors from PSCs.

Fidelity of lung progenitors derived under improved conditions. We then asked whether the increased epithelialization of lung progenitors during in vitro lung specification resulted in an improved lung primordial progenitor signature.

To generate the in vitro transcriptome data required for LAP analysis, we performed bulk RNA-Seq of all epithelial cell populations from both 2D-Gelatin and 3D-Matrigel conditions, namely 2D-Nkx2-1+Epacm− (lung epithelial progenitors derived on 2D gelatin), 3D-Nkx2-1−Epacm+ (lung epithelial progenitors derived on 3D Matrigel), and 3D-Nkx2-1−Epacm+ (non-lung epithelial progenitors derived on 3D Matrigel) (Fig. 5h, i). Three populations were shown to be quite distinct both by hierarchical clustering (Fig. 5h) and principal component analysis (PCA) (Fig. 5i). Interestingly, the 3D-Nkx2-1+Epacm+ (non-lung epithelium) was clearly separated from the two Nkx2-1+ populations along the PCI axis (82% variance) in the PCA plot. For direct comparison between the old and new lung specification protocols (Fig. 6a), we included again the 2D-Nkx2-1− condition in the assessed conditions.

Enrichment of epithelial progenitors in our lung differentiation protocol correlated with improved primordial progenitor fidelity as reflected by increased similarity scores to the E9.0 in vivo Nkx2-1+ lung primordium of all Epacm+ populations (Fig. 6b). Nevertheless, the 3D-Nkx2-1−Epacm+ population had lower similarity scores to other lung epithelial cell populations and higher foregut endoderm and liver similarity scores compared with the Nkx2-1+Epacm+ populations (Fig. 6b, c). Notably, the mesenchymal (MEF) signature that was prominent in Nkx2-1+ lung progenitors derived under non-optimal biomechanical conditions all but disappeared in Epacm+ cells under improved culture conditions (Fig. 6d; Supplementary Fig. 6A, B). None of the four tested populations showed high similarity to neural or thyroid cell populations (Supplementary Fig. 7A, B). We then compared these data with the Nkx2-1mCherry+Epacm+ specificity of our in vitro specification and biology of lung progenitors derived in Epacm+ populations.

The two Nkx2-1−Epacm+ populations also had high projections on distinct lung embryonic epithelial populations, with 2D-Nkx2-1−Epacm+ having a higher similarity score to a proximal E13.5 subpopulation (enriched in Sgb3a2, Sox2, Sema3c, and Cldn10) and 3D-Nkx2-1−Epacm+ projecting more strongly on an E17.5 distal lung epithelial subpopulation (Supplementary Fig. 6A, B; Fig. 6b). This suggests the culture conditions of in vitro lung epithelial primordial progenitors may also affect downstream cell fate decisions.

To gain further confidence in the similarity scores described above, we applied an independent orthogonal signature projection method based on gene set analysis using the bulk RNA-Seq or microarray data sets in Fig. 6b–d and Supplementary Fig. 7A. As the overall patterns of radar plots in Supplementary Fig. 7C demonstrate, this method was able to replicate the LAP analysis for our control cell populations (D1,
D7, and D14 from Dame et al.\(^{(43)}\) with high projections on undifferentiated ESC and E8.25 ectoderm for D1 cells and high E13.5 thyroid projection for D14 cells. Similarly, a reduction of the MEF score and increases in E9.0 Lung Primordium and E8.25 Foregut Endoderm scores was observed in all EPCAM\(^{+}\) populations whereas the 3D-Nkx2-1\(^{+}\)EPCAM\(^{+}\) population also showed an increased liver score (Fig. 6c).

Thus, our unbiased, genome-wide similarity analysis allowed quantitative comparisons of PSC-derived lung and thyroid engineered progenitors to reference benchmarks including rare primordial Nkx2-1\(^{+}\) in vivo progenitors and demonstrated derivation of epithelial progenitors with bona fide lung primordial progenitor identity in our modified lung-directed differentiation protocol.
Fig. 5 ECM modulation during lung specification. a Projection strength scores for the PSC-derived lung progenitors described in Serra et al.9. b Schematic illustrating directed differentiation of mouse PSCs into lung epithelial lineages using 2D-gelatin-coated plates as well as 3D-Matrigel-coated plates. For these experiments, we used a mouse ESC line with an Nkx2-1mCherry reporter. The 3D condition leads to a low yield of Nkx2-1mCherry+/Epcam-+ cells and a dominant population of Nkx2-1mCherry+/Epcam-+ cells. c, d 3D Matrigel leads to increased numbers of Nkx2-1mCherry+ cells compared with 2D gelatin at both D8 and D14. Scale bar: 100 μm. e Bivariate flow cytometry plots for D14 lung progenitors showing expression of Nkx2-1mCherry and BV421-Epcam. f Quantification of Nkx2-1mCherry+/Epcam-+ populations shown in e as a percent of cells (left panel) and cell yield per starting (DO) undifferentiated ESC (right panel). P-value < 0.01 by two-tailed t tests, N = 3 independent experiments, error bars represent standard deviation. g RT-qPCR data from downstream differentiation of 3D-Matrigel Nkx2-1mCherry+/Epcam-+ sorted populations shown in d. Fold changes relative to undifferentiated cells, statistics from two-tailed, unpaired t tests, N = 3 independent experiments, error bars represent standard deviation. h D14 EPCAM+ cells from both 2D-gelatin-coated and 3D-Matrigel-coated plates were assessed by bulk RNA-seq. The resulting samples were then hierarchically clustered using Euclidean distances. i Principal Component Analysis (PCA) of bulk RNA-Seq samples. The top 500 most variable genes were used in this analysis and they are listed in Supplementary Data 2.

Discussion

Directed differentiation of PSCs has emerged as one of the most promising regenerative medicine platforms55. One of the remaining obstacles in the clinical use of PSCs for the treatment of lung disease is the partial understanding of the sequence of cell-fate decisions that lead from PSC-derived anterior foregut endoderm to functional, clinically relevant lung cell populations. In vivo, the formation of all lung epithelial lineages depends on the critical moment of lung specification. An incomplete understanding of this particular progenitor population can limit attempts to generate a similar population in vitro. By providing global transcriptomic signatures of the earliest in vivo Nkx2-1+ progenitors and quantitative algorithms for their application, our current work addresses deficiencies in the development of lung differentiation protocols by applying computational methods to improve the efficiency of in vitro lung progenitor derivation.

The underlying assumption of current lung-directed differentiation protocols is that lung epithelial lineages in vivo are derived through an Nkx2-1+ intermediate primordial progenitor6-8,20,56. We substantiated this statement by performing indelible marking of Nkx2-1+ lung progenitors around E9.0 using an inducible Nkx2-1 knock-in Cre driver and an nT/nG reporter mouse. Nuclear GFP-positive cells were exclusively found in the developing lung at E14.5 and in E18.5 cells expressing markers of lung epithelial lineages. These findings further support the wide use of Nkx2-1 as the earliest marker of lung epithelial fate and strongly imply that the use of Nkx2-1 fluorescent reporters in mouse and human PSCs7,20 in combination with lung specification cocktails leads to derivation and isolation of putative lung progenitors with broad differentiation competence.

Using our Nkx2-1GFP knock-in mouse, we sorted Nkx2-1+ lung epithelial progenitors at E9.0 and characterized their transcriptome by bulk RNA-Seq. The simultaneous purification and characterization of other Nkx2-1+ embryonic progenitors and E8.25 precursor populations (foregut endoderm and ectoderm) allowed us to define the genetic program of primordial lung progenitors and signaling pathways that appear to be peculiar to lung specification, relative to early endoderm and thyroid epithelium. Previous work in Xenopus and mouse has elucidated signals, such as RA, that impart lung competence to foregut endoderm48 as well as lung specification signals emanating from the mesenchyme, such as Bmp and Wnt5-5,9. Our pathway analysis does demonstrate activation of Wnt and Tgf-β superfamily pathways in lung primordial progenitors versus foregut endoderm vindicating the use of Wnt and Bmp activators in vitro lung specification cocktails for PSC-based systems.

Although the Shh pathway is also upregulated in primordial lung progenitors, our ex vivo foregut explant experiments demonstrate that it acts upstream of β-catenin-dependent Wnt signaling and may be redundant for in vitro PSC-derived foregut specification as long as Wnt activators are present. Interestingly, retinol metabolism is downregulated in lung progenitors which may imply that RA signaling is necessary for making foregut cells competent to adopt lung fate, but it is not required during lung specification as reported previously48. As far as pathways regulating thyroid versus lung endodermal fates are concerned, the P13K-mediated signaling features prominently in thyroid-related (3/10 pathways), but is absent from lung specification-related pathways. This is experimentally supported by the inability of FGF ligands to rescue lung ablation following inhibition of Shh signaling in foregut explants as well as previous work from our group9,21. More specifically, inhibition of P13K-dependent FGF signaling in either Xenopus or in mouse ESC-derived anterior foregut at the moment of thyroid specification completely abrogated thyroid fate21.

Intriguingly, the Hippo pathway is also differentially expressed during lung specification and various YAP/TAZ targets such as Ankrd1 and Cyr61 are upregulated, whereas YAP/TAZ negative regulators are downregulated in lung primordial progenitors relative to pre-specified foregut endoderm. Yet, the significance of these findings is, at present, unclear as ShhC-Cre driven Yap deletion within the foregut endoderm does not lead to lung agenesis49.

Although pathway analyses of in vivo lung progenitors reveal that the minimal signals necessary and sufficient to induce lung fate in foregut progenitors have already been present in our lung specification media, the mesenchymal-like phenotype of in vitro-derived mouse lung progenitors implied that additional signals were required for proper preservation of the epithelial program during lung specification. The importance of biomechanical cues in cell-fate decisions of embryonic progenitors is increasingly recognized with a recent example of the role of the extracellular matrix in endocrine versus ductal fate decisions of bipotent pancreatic embryonic progenitors57. In this work, modulation of both the composition and the elastic modulus of the substratum resulted in increased derivation of EPCAM+ lung epithelial progenitors with broad differentiation potential. Future studies are required to decipher which of these variables is responsible for the improved yield of lung epithelial progenitors and in turn the signaling mechanisms regulated by these variables. Furthermore, it is still unclear whether this yield results from selection of an epithelial progenitor subset or inhibition of epithelial-to-mesenchymal drift in all cells.

Having transcriptionally defined E9.0 in vivo lung epithelial primordial progenitors as a benchmark for lung primordial progenitors, this work attempted then to define the fidelity of in vitro lung progenitors derived under improved biomechanical conditions. As the repertoire of in vitro engineered cell types...
Fig. 6 Similarity analysis of in vitro PSC-derived and in vivo lung progenitors. a Schematic depicting the four types of in vitro PSC-derived progenitors using a lung specification cocktail (in vitro-derived lineages) and derivation of similarity scores to in vivo populations (reference lineages) via two methods, namely linear projection analysis (LAP) and gene set analysis (GSA) (b–d). Projections strength scores from LAP analysis for the four populations shown in a, including projections to lung endodermal populations (b), non-lung endoderm (c), and mesenchymal-like populations (d). Scores shown in Fig. 5a are re-plotted here for comparison purposes. e Projection scores from gene set analysis using the reference cell populations from b–d and Supplementary Fig. 7A with bulk RNA-Seq data.
through directed differentiation, direct conversion, and forward programming is rapidly expanding, it is essential to define the correspondence of in vitro-derived and in vivo cell types. Several methods, including our own work, have been developed that apply network analysis to cell-fate conversions\(^\text{38}\), identify TFs that underlie cell identity\(^\text{39,40}\), and predict reprogramming TFs\(^\text{41,42}\). Previously, we successfully used LAP to interrogate the degree to which thyroid progenitors derived from PSCs were similar to E13.5 thyrocytes\(^\text{43}\). Application of the same method in the characterization of in vitro lung progenitors showed that enhancement of their epithelial properties resulted in increased resemblance to in vivo lung primordial progenitors and highly reduced mesenchymal cell signature. Nkx2-1\(^-\)EPCAM\(^+\) progenitors derived under 2D and 3D conditions had similar scores that were greatly improved compared to Nkx2-1\(^+\) progenitors without EPCAM enrichment. This implies that most of the lung competence in the 2D, suboptimal condition appears to be in the small (~1–2%) Nkx2-1\(^-\)+EPCAM\(^+\) population. Nevertheless, differential similarity scores to epithelial lung progenitors from later time points indicate that culture on different substrata may introduce fate biases, a tailoring possibility that merits further study. The increased similarity to in vivo lung primordium of even the Nkx2-1\(^-\)EPCAM\(^+\) population may seem perplexing at first glance. Yet, the fact that this population does not project on other lung epithelial lung populations, and has a higher liver signature suggests that Nkx2-1-negative cells are qualitatively different than Nkx2-1\(^+\) epithelial progenitors at the same stage of in vitro directed differentiation. It is also possible that the sorted Nkx2-1\(^-\)EPCAM\(^+\) population contains less differentiated cells that are about to adopt a lung fate, as suggested by the higher reported projection of this population.

In summary, this work defines the genetic program of in vivo primordial lung progenitors and uses similarity analysis to increase the correspondence of PSC-derived lung progenitors with their in vivo counterparts via manipulation of their biochemical environment. Our findings and the proposed methodology provide a framework for rational and systematic development and refinement of PSC-directed differentiation protocols and can propel future studies of lung primordial progenitors.

**Methods**

**Animal maintenance.** All mouse studies involving mice carrying GFP, CreERT2, Cre, and mT/mG transgenes were approved by the Institutional Animal Care and Use Committee of Cincinnati Children’s Hospital. We previously published the generation of Nkx2-1\(^+\)GFP mice (maintained on a C57BL/6J background)\(^\text{37}\). The Nkx2-1\(^+\)CreERT2 (Nkx2-1tm1can.1(creERT2)Zjh), R26R\(^\text{Tg}$$\text{Tg}$$B6.129S6-CAG-tdTomato*,-EGFP*$$\text{Tg}$$Zjh$$\text{Tg}$$Ees, Cat no.: 002624, 0.2 mg ml\(^{-1}\)) and Nkx2-1\(^-\)CreERT2 (Nkx2-1tm1.1(cre/ERT2)Zjh), R26R\(^\text{Tg}$$\text{Tg}$$B6.129S4-$$\text{Tg}$$ROSA26Sortm1Sor$$\text{Tg}$$Zjh$$\text{Tg}$$Ees, Cat no.: 002625) mouse strains were purchased from Jackson Laboratory.

**Characterization of the Nkx2-1\(^+\)GFP mouse.** To characterize the expression of Nkx2-1\(^+\) during mouse development, timed-pregnant females (2–6 months old) were euthanized by inhalation of isoflurane (purchased from IsoThesia, Henry Schein, Dublin, OH) and subsequent cervical dissection by embryonic (E) day 8.5 and E18.5, and embryos were delivered by cesarean using Dumont dissection forceps (Becton-Dickinson Science Tools, Foster City, CA). The lung domain was dissected out using Tungsten needles (Fine Science Tools) under an Olympus stereo microscope or cryo-embedded and sectioned. Cryosections were stained with an APC-anti-proliferating cell nuclear antigen (PCNA), rabbit anti-P63-alpha (1:100, Cell Signaling Technology, 3090s), rabbit anti-α-tubulin (1:800, Cell Signaling Technology, 3770s), rabbit anti-acetylated α-tubulin (1:800, Cell Signaling Technology, 3090s), or with an anti-hamster biotinylated antibody followed by incubation with an anti-biotin Cy3 antibody for the anti-PCNA antibody, counterstained with DAPI and imaged on a confocal Zeiss LSM 510 Meta NLO laser scanning inverted microscope using the LSM510 software or on a Zeiss LSM 700 Laser Scanning Confocal Microscope using the ZEN software (Carl Zeiss Microscopy, Thornwood, NY). Post-acquisition image processing was performed in ZEN 2.1 (black edition) software.

Lungs from adult Nkx2-1\(^+\)GFP mice were dissected out, minced and digested using Collagenase A (0.1%, Roche 103578) and Dispase II (2.4 U ml\(^{-1}\), Roche 29582S) supplemented with 2.5 mM CaCl\(_2\) for 60 min at 37 °C with periodic trituration. Monodisperse cells were suspended in HBSS with 2% FBS (HBSS + buffer), filtered through 30-μm FACS strainers (Miltenyi Biotec, Cat no.: 130-041-407) and spun with an anti-EPCAM antibody (APC/Cy7 anti-mouse CD326, clone G8.8, Biologend, Cat no.: 118218, 0.2 mg ml\(^{-1}\), 1:80) or an isotype control (APC/ Cy7 Rat IgG2a, isotype, clone RKT2758, Cat no.: 400524, 0.2 mg ml\(^{-1}\), 1:80) for 30 min on ice. Sorting was performed at the Boston University Flow Cytometry Core Facility using the MoFlo cell sorter and cells were stored at ~80 °C until further analysis.

**Real-time quantitative polymerase chain reaction.** RNA from sorted cells was extracted using QIAshredder columns (QIAGEN, Mansfield, MA, Cat no.: 79656) and the RNaseasy Plus Mini Kit (QIAGEN, Cat no.: 74136) according to the manufacturer’s instructions. Purified RNA was eluted in either 14 or 30 μl nucleic-free water, quantified on a NanoDrop ND-1000 microvolume spectrophotometer (ThermoFisher Scientific, Waltham, MA), and diluted as needed. cDNA was prepared using Taqman Reverse Transcription Reagents (Applied Biosystems, Waltham, MA, Cat no.: N808-0234). cDNA samples were prepared with Taqman Fast Universal Master Mix (Life Technologies, 4367846) and diluted to 6.25 ng per 25 μl reaction for real-time analysis on the StepOnePlus Real Time PCR System (Applied Biosystems). Relative expression analysis calculations were performed according to the 2-ΔΔCt method\(^\text{44}\) utilizing 18S rRNA as the reference gene and either es portraits or whole embryo cells as the reference sample. If expression was undetected based on the set threshold, the value was set to the maximum number of cycles (40) to allow fold change calculations. The data were graphed for visualization as shown in the figures using the GraphPad Prism 8 (version 8.0.2) software.

Taqman gene expression arrays were purchased from Applied Biosystems and are listed in Supplementary Table 1.

**Statistical analysis.** Error bars in graphs represent SD or SEM as indicated in the figure legends. Biological sample replicates (N) are also indicated in the legends. Statistically significant differences between conditions (ΔCt values used for RT-qPCR calculations) were determined using two-tailed unpaired Student’s t tests or as specified in figure legends. Statistical significance is represented as P-values.

**Lineage trace.** To trace the contribution of lung primordial progenitor cells to lung epithelial cell lineages, Nkx2-1\(^+\)CreERT2 mice were crossed with R26R\(^\text{Tg}$$\text{Tg}$$B6.129S4-$$\text{Tg}$$ROSA26Sortm1Sor$$\text{Tg}$$Zjh$$\text{Tg}$$Ees mice. For these experiments, we used 5–8-month-old male mice and 3–6-month-old female mice. Tamoxifen (Sigma-Aldrich, St. Louis, MO, Cat no.: T5648) was dissolved in ethanol at room temperature, and then diluted 9:1 with corn oil to a final concentration of 10 mg ml\(^{-1}\) tamoxifen. Recombination was induced by two doses of tamoxifen (2.5 mg tamoxifen per 10 mg ml\(^{-1}\) of body weight) delivered by intraperitoneal injection at E7.5 and E8.0, and embryos were harvested at either E13.5 or E14.5. Lungs were either imaged using an Olympus stereo fluorescence imaging microscope or cryo-embedded and sectioned. Cryosections were stained with an APC- Cy7-conjugated anti-mouse rat EPCAM antibody (Biologend CD326, Cat no.: 118217) utilizing 18S rRNA as the reference gene and the NIS-Elements software (Nikon Instruments Inc., Melville, NY).

To counteract estrogenic effects of tamoxifen and prevent dystocia in pregnant animals at later time points, we modified our protocol by adding progesterone and switching the administration route to oral gavage. Tamoxifen and progesterone were dissolved in ethanol at room temperature, and then diluted 9:1 with corn oil to a final concentration of 20 mg ml\(^{-1}\) tamoxifen with 10 mg ml\(^{-1}\) progesterone. Recombination was induced by a single dose of tamoxifen–progesterone (2 mg tamoxifen and 1 mg progesterone per 10 g of body weight) delivered by oral gavage to pregnant females at E8.75. Control animals received either a single dose of tamoxifen–progesterone (2 mg tamoxifen and 1 mg progesterone per 10 g of body weight) at E7.5 or a gavage of 9:1 corn oil and ethanol (100 μl per 10 g of body weight) at E8.75.
Embryos receiving tamoxifen–progesterone or corn oil at E8.75 were harvested at E18.5, and lungs, brain, stomach, esophagus, and liver were dissected and evaluated for GFP expression by fluorescent microscopy. Embryos receiving tamoxifen–progesterone at E7.5 were harvested at E13.5 and sectioned intact to evaluate GFP expression in the lungs and throughout the embryo. Tail tissue was used to confirm the genotype of each embryo. Dissected organs or whole embryos were fixed overnight in 4% paraformaldehyde at 4 °C, rinsed in PBS, transferred to 30% sucrose, cryo-immobilized in OCT medium, and sectioned at 5 μm thickness. For immunohistochemistry, sections were rehydrated and then rinsed with 0.3% Triton-X, rinsed again, and incubated overnight at 4 °C with the following primary antibodies: rabbit anti-pro-SPC (1:1000, Seven Hills Bioreagents; Cat no. 9-ectoderm)35, Syt-1 (1:800, mouse28 E8.25 embryos, forebrain and lung cells were isolated from E8.25 embryos, foregut endoderm and ectoderm35. Briefly, embryos were dissected using tungsten needles with the help of an Olympus stereo fluorescence imaging microscope and kept in HBSS with 10% FBS. Tissues were digested using Collagenase A (0.1%, Roche 103578) and Dispase II (2.4 U ml−1, Roche 295825) supplemented with 2.5 mM CaCl2 for 60 min at 37 °C with periodic trituration. Monodispersed cells were resuspended in HBSS through 30 μm FACS strainers (Millipore, Temecula, CA) and stained with propidium iodide for dead cell exclusion (1:500) directly before sorting. Sorting was performed at the Boston University Flow Cytometry Core Facility using the MoFlo cell sorter and cells were stored at −80 °C until further analysis. For E9.0 and E13.5 embryos, the three Nkx2-1GFP-expressing domains (forebrain, endoderm, and ectoderm) were dissected using tungsten needles with the help of an Olympus stereo fluorescence imaging microscope and kept in HBSS with 10% FBS. Tissues were digested using Collagenase A (0.1%, Roche 103578) and Dispase II (2.4 U ml−1, Roche 295825) supplemented with 2.5 mM CaCl2 for 60 min at 37 °C with periodic trituration. Monodispersed cells were resuspended in HBSS through 30 μm FACS strainers (Millipore, Temecula, CA) and stained with propidium iodide for dead cell exclusion (1:500) directly before sorting. Sorting was performed at the Boston University Flow Cytometry Core Facility using the MoFlo cell sorter and cells were stored at −80 °C until further analysis. For E9.0 and E13.5 embryos, the three Nkx2-1GFP-expressing domains (forebrain, endoderm, and ectoderm) were dissected using tungsten needles with the help of an Olympus stereo fluorescence imaging microscope and kept in HBSS with 10% FBS. Tissues were digested using Collagenase A (0.1%, Roche 103578) and Dispase II (2.4 U ml−1, Roche 295825) supplemented with 2.5 mM CaCl2 for 60 min at 37 °C with periodic trituration. Monodispersed cells were resuspended in HBSS through 30 μm FACS strainers (Millipore, Temecula, CA) and stained with propidium iodide for dead cell exclusion (1:500) directly before sorting. Sorting was performed at the Boston University Flow Cytometry Core Facility using the MoFlo cell sorter and cells were stored at −80 °C until further analysis. Flow cytometry plots were generated and formatted using the FlowJo software.

Mouse foreleg explants. Whole forelegs were dissected from mouse embryos (6–8 somite pairs) in Hank’s balanced salt solution (HBSS) then explanted onto 8 μm pore size Whatman Nucleopore Track-Etch Membranes (Millipore). Explants were cultured for 2–3 days in a base medium [BGJb medium (Gibco) + 10% fetal bovine serum (FBS, Sigma) + 0.2 mg ml−1 ascorbic acid] containing either specific inhibitors or DMSO as a vehicle control. Whole-mount immunostaining was performed using a modification of the method of Ahnhelt-Ronne et al.30. The primary antibodies used were: guinea pig anti-Nkx-2.1 (Seven Hill Bioreagents; 1:300) and rat anti-E-cad (R&D Systems; 1:200). After staining, explants were post-fixed with 4% paraformaldehyde and imaged using a Nikon AI-Rsi inverted light confocal microscope. ImageJ software was used to analyze the images.

Single-cell RNA-Seq analysis. In order to enrich the reference basis for gene expression predictions (see the next section), we leveraged two previous experiments performed in our labs. The E13.5, E15.5, and E17.5 single-cell data were generated as described in ref. 32. Briefly, lungs from Nkx2-1GFP embryos of the respective embryonic day were harvested and processed into single-cell suspensions using an enzyme solution containing dispase (Collaborative Biosciences), collagenase I (Life Technologies), and RNAase-free DNase (Promega). Nkx2-1GFP, positive cells stained with DAPI and also DAPI to exclude the mouse ESCs68 were differentiated into de...
Genomics) to produce single-cell bar-coded droplets using ×10 Single Cell 3' v2 chemistry. Libraries generated were sequenced across at least one lane on the Illumina NextSeq 500 in 2×150 bp run. The NextSeq version was 2.17. The HiSeq Rapid SBS kit was used for the Illumina HiSeq 2500 instrument with the HiSeq Rapid SBS kit. The resulting reads were aligned and gene level unique molecular identifier (UMI) counts obtain using CellRanger version 2.0.1 (×10 Genomics).

For 9.0 single-cell RNA-Seq data, Nkx2-1 GPP expressing domains (forebrain, thyroid, and lung tissues containing both GPP+ and GPP− lineages) were isolated from E8.25 embryo EPCAM+ cells using mFACS strainers (Miltenyi Biotec, Cat no.: 130-041-047) and stained with Calcein Blue (1:500) directly before sorting viable cells. Given the rarity of GPP+ cells at this time point, GPP− cells were mixed together with GPP+ cells from that tissue for cell capture for single-cell RNA sequencing. This resulted in successful capture of 127.9 Nkx2-1 GPP− forebrain progenitors (out of a total of 620 cells), but for lung and thyroid progenitors the rarity of GPP progenitors resulted in <3 GPP+ Nkx2-1 cells captured (out of a total of 417 lung and 279 thyroid mixed non-epithelial lineages, respectively). Cells were sorted at the Boston University Flow Cytometry Core Facility using the MoFlo high speed cell sorter and collected in HBSS i-buffer. Freshly-sorted cells were spun briefly to concentrate the cells, and then cell concentration and viability of flow-sorted samples were measured using the Countess II instrument (ThermoFisher Scientific) and ranged from 3 to 8 × 10^4 cells ml^-1 with 80% viability. Cell samples were input into 10× Genomics Single Cell 3' v2 workflow and processed according to the manufacturer’s instructions for a targeted recovery of 1000 cells (×10 Genomics, USA). Resulting cDNA libraries were sequenced on an Illumina NextSeq 500 instrument with 1.95 pM input and 1% PhiX control library spike-in (Illumina, USA). Single-cell capture was conducted at the Boston University Single Cell Sequencing Core using a 10X Genomics Chromium System. The resulting reads were aligned and gene level unique molecular identifier (UMI) counts obtain using CellRanger version 2.0.1 (×10 Genomics). Four libraries derived from embryonic mouse (two "lung" domain, one "thyroid", and one "forebrain" domain totaling 1316 cells and a post-normalization mean coverage of 165238 reads per cell) were combined for downstream analysis.

Because of the sparse nature of single-cell data, with high dropout levels (undetected but expressed transcripts), we averaged the normalized expression profiles of cells in each cluster and used those averages for the projections. The clusters were identified following the usual procedure of de-processing and QC filtering, identifying highly variable genes, log-normalization and scaling, and dimensionality reduction with PCA and clustering using the Louvain method for community detection at a resolution of 0.5. Clusters were manually annotated based on their marker genes and subsequently averaged, with those averages later exported for downstream analysis.

Linear algebra projections. Here, we give a brief overview of the projection method plots shown in Fig. 6 and Supplementary Fig. 7, please see refs. 15,64 for complete mathematical details. The projection method allows one to compare experimentally acquired gene expression data versus a reference basis. A projection score is defined as the indicator function of the closest match of gene expression, while 0.0 indicates no match. The projection method has several key advantages relative to other techniques such as Principal Component Analysis (PCA). First, it provides a biologically interpretable numeric value of the similarity between the data and the reference basis. Second, it removes the inherent correlations of the reference basis by measuring the LOAPs for the projections onto the reference basis. Third, the projection score for each reference cell type has a projection of 1 with only itself and has a projection of 0 with all other reference cell types.

To create the reference basis cell types, we first started with the basis used in refs. 80,81 (61 cell types). Next, we added data from the following RNA-Seq samples: E8.25 Embryonic, E8.25 foregut endoderm, E9.0 Nkx2-1+ forebrain, E9.0 Nkx2-1+ lung, E13.5 Nkx2-1+ thyroid. We also included single-cell RNA-Seq samples from E9.0 (lung, thyroid, and forebrain domains), and E13.5, E15.5, and E17.5 (lung), averaged over the clusters described above (total of 24 clusters). Finally, we added the average over three E16.5 lung epithelium control samples (GSE57391, sample IDs Ezkh_01, Ezkh_02, Ezkh_03) described in Galvis et al.15. The basis thus consists of 91 cell types.

The data set consists of a mix of Affymetrix GeneChip Mouse Gene 1.0 ST and RNA-Seq data and was normalized as follows. First, raw microarrays and RNA-Seq were separately processed following standard techniques for each data type. Next, only genes with at least one read across all samples were kept, leading to N = 9035 genes. In order to make robust comparisons across platforms, the raw expression output was converted to a rank order. Next, we wanted to convert this rank order to the z-score of a log-normal distribution. We converted the rank to a percentile (for N genes, by dividing by N−1), and then this percentile into a normal z-score. For mathematical convenience, we normalized by N and not N−1 since then the Euclidean norm of each microarray gene expression was N. At this point, each sample was described by a Gaussian distribution with a Euclidean norm of N = 10990. Finally, we selected only the genes that were highly expressed (z > 1) in at least one cell type, keeping a total of N = 9035 genes for the subsequent analysis. A z score of 1 was set as all conditions are presented in Supplementary Data 3. As mentioned above, a projection score of 1.0 indicates a perfect match of gene expression, zero indicates no match, and −1.0 indicates perfectly anti-correlated genes. There are no direct equivalents of P-values to measure the significance of a projection. However, one can compare the expression of random signatures in the projection methodology described in ref. 72, as implemented in the GSVA package27.

Gene set analysis. In order to validate the scores obtained above, we applied an orthogonal signature projection method based on gene set analysis. In particular, we computed separate enrichment scores for each pairing of sample and gene set using the projection methodology described in ref. 72, as implemented in the GSVA package.

The procedure can be divided into the following steps: first, we filtered, log-transformed and scaled the data. Second, we selected the genes that constitute the molecular signatures for each of the conditions we want to score: "E8.25 Foregut Endoderm", "E8.25 Ectoderm", "E9.0 Forebrain", "Liver", "Npc" and "Mef" (for the experimental conditions) and "E8.25 Foregut Endoderm", "E8.25 Ectoderm", "E13.5 Thyroid", "Esc" (for the control conditions). Third, we computed the enrichment score of each of those gene sets for our experimental conditions ("2D-Nkx2-1","2D-Nkx2-1+PSCAM","3D-Nkx2-1+PSCAM" and "3D-Nkx2-1-1PSCAM") and our controls ("D1", "D7", and "D14").

For the first step, we filtered out genes that had a total log-normal expression <2 overall samples in our RNA-Seq data sets. Log transformation and normalization was done using counts per million and a prior count offset to avoid taking log of zero. Subsequently, the values were centered around 0, and the distributions were visualized to evaluate effective normalization and Gaussian shape. Finally, we minimized the effect of outliers by capping extreme values to 1.5 times the interquartile range for each sample.

For the second step, we extracted the 3000 genes with higher variance overall, to capture those with biological signal, and then, for each of the conditions we wanted to score, we extracted the 100 most expressed among those highly variable genes. These are the gene sets to be used for scoring our problem samples in the following step. In the third and final step, we computed enrichment scores by integrating the difference of the weighted Empirical Cumulative Distribution Function (ECDF) of the genes in the signature and the ECDF of the remaining genes as in the aforementioned publications26,27.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
All sequencing data that support the findings of this study have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) under following accession codes: (a) in vivo bulk RNA-Seq: E8.25 Foregut Endoderm and E8.25 Ectoderm, E8.25 foregut endoderm, E9.0 Nkx2-1+ lung and brain, E9.0 Nkx2-1+ lung and brain, E13.5 Nkx2-1+ Thyroid, E13.5 Nkx2-1+ Thyroid, GEO Series accession code GSE138903. (b) In vitro bulk RNA-Seq: 2D-Nkx2-1+PSCAM, 3D-Nkx2-1+PSCAM, 3D-Nkx2-1-1PSCAM; GEO Series accession code GSE138676. (c) In vivo E9.0 single-cell RNA-Seq: GEO Series accession code GSE138904. (d) E13.5 lung Nkx2-1 GPP+ single-cell data; GEO Series accession code GSE139186. The E13.5 and E17.5 lung Nkx2-1 GPP+ single-cell data have been previously deposited and are accessible under GEO Series accession code GSE113320. The microarray data containing the 2D-Nkx2-1+ condition have been previously deposited and are accessible under GEO Series accession code GSE92916. The bulk RNA-Seq for the thyroid directed differentiation (D1, D7, D14 conditions) have been previously deposited and are accessible under GEO Series accession code GSE29572. All other relevant data are available from the corresponding author upon reasonable request.

Code availability
All custom LAPS scripts can be made available at [https://github.com/Emergent-Behaviors-in-Biology/lung-primordium](https://github.com/Emergent-Behaviors-in-Biology/lung-primordium). Additional modified scripts can be accessed upon request.

Received: 20 May 2019; Accepted: 23 December 2019; Published online: 31 January 2020

References
1. Murry, C. E. & Keller, G. Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development. Cell 132, 661–680 (2008).

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-020-14348-3 | www.nature.com/naturecommunications
2. Ikonoumu, L. & Kotton, D. N. Derivation of endodermal progenitors from pluripotent stem cells. J. Cell Physiol. 230, 246–258 (2015).

3. Domyan, E. T. et al. Signaling through BMP receptors promotes respiratory identity in the foregut via repression of Sox2. Development 138, 971–981 (2011).

4. Goss, A. M. et al. Wnt2b/2b and beta-catenin signaling are necessary and sufficient to specify lung progenitors in the foregut. Dev. Cell 17, 290–298 (2009).

5. Harris-Johnson, K. S., Domyan, E. T., Vezina, C. M. & Sun, X. Beta-Catenin promotes respiratory progenitor identity in mouse foregut. Proc. Natl Acad. Sci. USA 106, 16287–16292 (2009).

6. Huang, S. X. L. et al. Efficient generation of lung and airway epithelial cells from human pluripotent stem cells. Nat. Biotechnol. 32, 84–91 (2014).

7. Loemij, T. A. et al. Efficient derivation of purified lung and thyroid progenitors from embryonic stem cells. Cell Stem Cell 10, 398–411 (2012).

8. Mou, H. M. et al. Generation of multipotent lung and airway progenitors from mouse ESCs and patient-specific cystic fibrosis iPSCs. Cell Stem Cell 10, 385–397 (2012).

9. Serra, M. et al. Pluripotent stem cell differentiation reveals distinct developmental pathways regulating lung–versus thyroid-lineage specification. Development 144, 3879–3893 (2017).

10. Firth, A. L. et al. Generation of multiciliated cells in functional airway epithelium from human induced pluripotent stem cells. Proc. Natl Acad. Sci. USA 111, E1723–E1730 (2014).

11. Konishi, S. et al. Directed induction of functional multi-ciliated cells in proximal airway epithelial spheroids from human pluripotent stem cells. Stem Cell Rep. 6, 18–25 (2016).

12. Jacob, A. et al. Differentiation of human pluripotent stem cells into functional lung alveolar epithelial cells. Cell Stem Cell 21, 472–488 (2017).

13. McGarry, K. B. et al. Efficient derivation of functional human airway epithelium from pluripotent stem cells via temporal regulation of Wnt signaling. Cell Stem Cell 20, 844–857 (2017).

14. Tsao, P. N. et al. Notch signaling controls the balance of ciliated and secretory cell fates in developing airways. Development 136, 2297–2307 (2009).

15. Mucenski, M. L. et al. beta-catenin is required for specification of proximal/distal cell fate during lung morphogenesis. J. Biol. Chem. 278, 40321–40328 (2003).

16. Shu, W. G. et al. Wnt/beta-catenin signaling acts upstream of N-myc, BMP4, and FGF signaling to regulate proximal-distal patterning in the lung. Dev. Biol. 283, 226–239 (2005).

17. Hashimoto, S. et al. beta-Catenin-SOX2 signaling regulates the fate of developing airway pathways. J. Cell Sci. 125, 932–942 (2012).

18. Desai, T. J., Malpel, S., Flentke, G. R., Smith, S. M. & Cardoso, W. V. Retinoic acid selectively regulates Fgfr10 expression and maintains cell identity in the prospective lung field of the developing foregut. Dev. Biol. 273, 402–415 (2004).

19. Lazzaro, D., Price, M., Defelice, M. & Dilauro, R. The transcription factor Tft-1 is expressed at the onset of thyroid and lung morphogenesis and in restricted regions of the fetal brain. Development 113, 1093–1104 (1991).

20. Hawkins, F. et al. Prospective isolation of NKX2-1-expressing human lung progenitors derived from pluripotent stem cells. J. Clin. Invest. 127, 2277–2294 (2017).

21. Kurmann, A. A. et al. Repression of Igf1 expression by Ezh2 prevents basal cell carcinoma. Proc. Natl Acad. Sci. USA 111, 13031–13036 (2016).

22. Dame, K. et al. Thyroid progenitors are robustly derived from embryonic stem cells through transient, developmental stage-specific overexpression of NKX2-1. Stem Cell Rep. 8, 216–225 (2017).

23. Gotoh, S. et al. Generation of alveolar epithelial spheroids via isolated progenitor cells from human pluripotent stem cells. Stem Cell Rep. 3, 394–403 (2014).

24. Niehrs, C. Function and biological roles of the Dickkopf family of Wnt antagonists. Dev. Biol. 385, 109–122 (2014).

25. Mahoney, I. E., Morii, M., Szymaniak, A. D., Vardelas, X. & Cardoso, W. V. The Hippo pathway effector Yap controls patterning and differentiation of airway epithelial progenitors. Dev. Cell 30, 137–150 (2014).

26. Lang, A. H., Li, H., Collins, J. J. & Mehta, P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. PLoS Comput. Biol. 10, e1003734 (2014).

27. Gokhan, L. et al. Repression of Igf1 expression by Ehx2 prevents basal cell differentiation in the developing lung. Development 142, 1458–1469 (2015).

28. Frank, D. B. et al. Nuclear double-catenin-mediated stabilization of E-cadherin is essential for primitive endoderm specification. PLoS Genet. 12, e1006243 (2016).

29. Ikonoumu, L., Wagner, D. E., Turner, L. & Weiss, D. J. Translating basic research into safe and effective cell-based treatments for respiratory diseases. Ann. Am. Thorac. Soc. 16, 657–668 (2019).

30. Pieters, T. et al. In vitro generation of human pluripotent stem cell derived lung organoids. eLife 4, e05098 (2015).

31. Mamidi, A. et al. Mechanosignalling via integrins directs fate decisions of pluripotent progenitors. Nature 564, 114–118 (2018).

32. Cahan, P. et al. CellNet: network biology applied to stem cell engineering. Cell 158, 903–915 (2014).

33. Gotoh, S. et al. Generation of alveolar epithelial spheroids via isolated progenitor cells from human pluripotent stem cells. Stem Cell Rep. 3, 394–403 (2014).

34. Niehrs, C. Function and biological roles of the Dickkopf family of Wnt antagonists. Dev. Biol. 385, 109–122 (2014).

35. Mahoney, I. E., Morii, M., Szymaniak, A. D., Vardelas, X. & Cardoso, W. V. The Hippo pathway effector Yap controls patterning and differentiation of airway epithelial progenitors. Dev. Cell 30, 137–150 (2014).

36. Lang, A. H., Li, H., Collins, J. J. & Mehta, P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. PLoS Comput. Biol. 10, e1003734 (2014).

37. Gokhan, L. et al. Repression of Igf1 expression by Ehx2 prevents basal cell differentiation in the developing lung. Development 142, 1458–1469 (2015).

38. Frank, D. B. et al. Early lineage specification defines alveolar epithelial ontogeny in the murine lung. Proc. Natl Acad. Sci. USA 116, 4362–4371 (2019).

39. Mahoney, I. E., Morii, M., Szymaniak, A. D., Vardelas, X. & Cardoso, W. V. The Hippo pathway effector Yap controls patterning and differentiation of airway epithelial progenitors. Dev. Cell 30, 137–150 (2014).

40. Lang, A. H., Li, H., Collins, J. J. & Mehta, P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. PLoS Comput. Biol. 10, e1003734 (2014).

41. Gokhan, L. et al. Repression of Igf1 expression by Ehx2 prevents basal cell differentiation in the developing lung. Development 142, 1458–1469 (2015).

42. Frank, D. B. et al. Early lineage specification defines alveolar epithelial ontogeny in the murine lung. Proc. Natl Acad. Sci. USA 116, 4362–4371 (2019).

43. Mahoney, I. E., Morii, M., Szymaniak, A. D., Vardelas, X. & Cardoso, W. V. The Hippo pathway effector Yap controls patterning and differentiation of airway epithelial progenitors. Dev. Cell 30, 137–150 (2014).

44. Lang, A. H., Li, H., Collins, J. J. & Mehta, P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. PLoS Comput. Biol. 10, e1003734 (2014).
59. D’Alessio, A. C. et al. A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Rep.* 5, 763–775 (2015).

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

61. Pusuluri, S. T., Lang, A. H., Mehta, P. & Castillo, H. E. Cellular reprogramming dynamics follow a simple 1D reaction coordinate. *Phys. Biol.* 15, 016001 (2018).

62. Soriano, P. Generalized lacZ expression with the ROSA26 Cre reporter strain. *Nat. Genet.* 21, 70–71 (1999).

63. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(ΔΔCt) method. *Methods* 25, 402–408 (2001).

64. Liu, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 41, 17 (2013).

65. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25 (2010).

66. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140 (2010).

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

68. Bilodeau, M., Shojaie, S., Ackerley, C., Post, M. & Rossant, J. Identification of a proximal progenitor population from murine fetal lungs with clonogenic and multilineage differentiation potential. *Stem Cell Rep.* 3, 634–649 (2014).

69. Dang, S. M., Gerecht-Nir, S., Chen, J., Itskovitz-Eldor, J. & Zandstra, P. W. Controlled, scalable embryonic stem cell differentiation culture. *Stem Cells* 22, 275–282 (2004).

70. Ahnfelt-Ronne, J. et al. An improved method for three-dimensional reconstruction of protein expression patterns in intact mouse and chicken embryos and organs. *J. Histochem. Cytochem.* 55, 925–930 (2007).

71. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214 (2015).

72. Barbie, D. A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462, 108–112 (2009).

73. Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 14, 7 (2013).

**Acknowledgements**

We thank Drs. Avi Spira and Gang Liu of the BU School of Medicine (BUSM) Computational Biomedicine Section for help with RNA-sequencing. We are indebted to Dr. Felicia Chen from BUSM Pulmonary Center for help with embryo staging and dissections. We thank Dr. Paul Gadue (UPenn) for the ENDM1 antibody gift. We are grateful to Brian R. Tilton and Dr. Patrick Autissier of the BUSM Flow Cytometry Core for assistance with epifluorescence microscopy. S.L.L. was supported by a Boston University Clinical and Translational Science Institute (CTSI) training grant (TL1TR001410). M.K.M. is a Marshall Plan Scholar supported by the Austrian Marshall Plan Foundation. R.M. and P.M. were supported by NIH NIGMS grant 1R35GM119461 and a Simons Investigator Award in the Mathematical Modeling of Living Systems (MMLS) to P.M. L.I. was supported by NIH grants (R01 HL111574, R01 HL124280), a chILD Foundation/American Thoracic Society award, an Evans Junior Faculty Merit Award and a CTSI Pilot Award (UL1TR001430). D.N.K. was supported by NIH grants U01HL134745, R01HL095993, R01DK105029, R01GM122096, R01HL122442, R01HL128172, and U01HL134766.

**Author contributions**

L.I. made the figures and L.I. and D.N.K. wrote the paper. L.I., M.J.H, S.L.L., R.M.S., K.D., J.H.-B., M.L.L., C.C., A.I.F. and J.M.S. performed experiments. I.S.C. and M.M.K. performed RNA-Seq and GSEA analyses. R.M., C.V.-M. and P.M. conceived and performed computational analyses. D.B.F. and E.E.M. contributed data sets. L.I., M.J.H., J.M.S. and D.N.K. conceived and designed experiments. L.I., E.K., X.V., P.M. and D.N.K. supervised research.

**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-14348-3.

**Correspondence** and requests for materials should be addressed to L.I. or D.N.K.

**Peer review information** Nature Communications thanks Tushar Desai and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2020