A single-cell atlas of the mouse and human prostate reveals heterogeneity and conservation of epithelial progenitors

Laura Crowley[1,2,3,4,7,10], Francesco Cambuli[1,2,3,4,7,10], Luis Aparicio[4,5,7,10], Maho Shibata[1,2,3,4,7,9], Brian D. Robinson[8], Shouhong Xuan[1,2,3,4,7], Weiping Li[1,2,3,4,7], Hanina Hibshoosh[6,7], Massimo Loda[8], Raul Rabadan[4,5,7,11], and Michael M. Shen[1,2,3,4,7,11]

[1]Department of Medicine, [2]Department of Genetics and Development, [3]Department of Urology, [4]Department of Systems Biology, [5]Department of Biomedical Informatics, [6]Department of Pathology and Cell Biology, [7]Herbert Irving Comprehensive Cancer Center, Columbia University Irving Medical Center, New York, NY 10032, USA

[8]Department of Pathology and Laboratory Medicine, Weill Medical College of Cornell University, New York, NY 10021, USA

[9]Present address: Department of Anatomy and Cell Biology, School of Medicine and Health Sciences, The George Washington University Cancer Center, The George Washington University, Washington, DC 20052, USA

[10]Authors made equal contributions

[11]Correspondence: rr2579@cumc.columbia.edu, mshen@columbia.edu

Summary

Understanding the cellular constituents of the prostate is essential for identifying the cell of origin for benign prostatic hyperplasia and prostate adenocarcinoma. Here we describe a comprehensive single-cell atlas of the adult mouse prostate epithelium, which demonstrates extensive heterogeneity. We observe distinct lobe-specific luminal epithelial populations (LumA, LumD, LumL, and LumV) in the distal region of the four prostate lobes, a proximally-enriched luminal population (LumP) that is not lobe-specific, as well as a periurethral population (PrU) that shares both basal and luminal features. Functional analyses suggest that LumP and PrU cells have multipotent progenitor activity in organoid formation and tissue reconstitution assays. Furthermore, we show that mouse distal and proximal luminal cells are most similar to human acinar and ductal populations, that a PrU-like population is conserved between species, and that the mouse lateral prostate is most similar to the human peripheral zone. Our findings elucidate new prostate epithelial progenitors, and help resolve long-standing questions about the anatomical relationships between the mouse and human prostate.

The significant anatomical differences between the mouse and human prostate have long hindered analyses of mouse models of prostate diseases. The mouse prostate can be separated into anterior (AP), dorsal (DP), lateral (LP), and ventral (VP) lobes; the mouse dorsal and lateral lobes are often combined as the dorsolateral prostate (DLP) (Cunha et al., 1987; Shappell et al., 2004; Shen and Abate-Shen, 2010). In contrast, the human prostate lacks defined lobes, and instead is divided into different histological zones (central, transition, and peripheral); the peripheral zone represents the predominant site of prostate adenocarcinoma, whereas benign prostatic hyperplasia (BPH) occurs in the transition zone (Cunha et al., 2018; Ittmann, 2018; Shappell et al., 2004). Moreover, unlike the mouse, the human prostate has distinct ductal and acinar regions. Although microarray gene expression profiling has suggested that the DLP is most similar to the human peripheral zone (Berquin et al., 2005), there is no consensus on the relationship between mouse lobes and human zones (Ittmann, 2018; Ittmann et al., 2013; Shappell et al., 2004).

The adult prostate epithelium is comprised of luminal, basal, and rare neuroendocrine cells (Shen and Abate-Shen, 2010; Toivanen and Shen, 2017), and cellular heterogeneity has been suggested within the luminal (Barros-Silva et al., 2018; Chua et al., 2014; Karthaus et al., 2020; Karthaus et al., 2014; Kwon et al., 2016; Liu et al., 2016) and basal compartments (Goldstein et al., 2008; Lawson et al., 2007; Wang et al., 2020). Lineage-tracing analyses have shown that the hormonally-intact adult prostate epithelium is maintained by unipotent progenitors within the basal and luminal epithelial compartments (Choi et al., 2012; Lu et al., 2013; Wang et al., 2013). However, following tissue dissociation, both basal and luminal cells can act as bipotent progenitors in organoid or tissue reconstitution assays (Chua et al., 2014; Karthaus et al., 2014). The progenitor properties of basal cells may reflect their ability to generate luminal progeny during tissue repair after wounding or inflammation (Kwon et al., 2014; Toivanen et al., 2016), but the role of presumptive luminal progenitors has been less clear. In particular, several studies have suggested increased progenitor potential in the proximal region of the prostate (nearest to the urethra) (Burger et al., 2005; Goto et al., 2006; Tsujimura et al., 2002), particularly for proximal luminal cells

(Karthaus et al., 2020; Kwon et al., 2016; Wei et al., 2019; Zhang et al., 2018). However, the nature and distribution of these epithelial populations have been poorly characterized.

Distinct luminal epithelial populations in the mouse prostate

To examine cellular heterogeneity, we performed single-cell RNA-sequencing of whole prostates from adult wild-type mice at 10 weeks of age. We microdissected the full proximal-distal extent of each prostate lobe down to its junction with the urethral epithelium (Figure 1—figure supplement 1A-C). We noted that the anterior (AP), dorsal (DP), and lateral (LP) lobes joined the urethra in close proximity on the dorsal side, whereas the ventral lobe (VP) had a distinct junction ventrally. As previously described (Cunha et al., 1987; Shappell et al., 2004; Shen and Abate-Shen, 2010), each lobe has a characteristic morphology, pattern of ductal branching, and histological appearance (Figure 1—figure supplement 1D-F).

To analyze these datasets, we applied *Randomly*, an algorithm that uses random matrix theory to reduce noise in single-cell datasets (Aparicio et al., 2020). Using the universality property of random matrix theory on eigenvalues and eigenvectors of sparse matrices, *Randomly* discriminates biological signals from noise and sparsity-induced confounding signals (which typically comprise approximately 98% of the data, based on a survey of published single-cell datasets) (Aparicio et al., 2020). Processing by *Randomly* facilitated the identification of cell populations with distinct transcriptional signatures (Figure 1—figure supplement 2). As visualized in an aggregated dataset composed of 5,288 cells from two whole prostates, we found distinct luminal, basal, and neuroendocrine populations that were annotated based on the expression of marker genes (Figure 1A). Notably, we could identify five different luminal epithelial populations, a single basal population, rare neuroendocrine cells, and a small population of epithelial cells that expresses both basal and luminal markers. We could also identify distinct stromal and immune components, corresponding to two different stromal subsets (Kwon et al., 2019), as well as immune cells (macrophages, T cells, B cells); some datasets also contained small populations of contaminating vas deferens and seminal vesicle cells.

To assess whether some of these epithelial populations might be lobe-specific, we next performed single-cell RNA-seq analyses of individual lobes (Figure 1B). We found that four of the luminal populations identified in the aggregated dataset were highly lobe-specific; hence, we named these populations LumA (AP-specific), LumD (DP-specific), LumL (LP-specific), and LumV (VP-specific). The remaining luminal population was observed in the datasets for all four lobes, and was highly enriched in the proximal portion of each lobe; thus, we termed this population LumP (proximal) (Figure 1C).

Spatial localization and morphology of epithelial populations

To examine the lobe-specificity and spatial distribution of these luminal populations, we identified candidate markers based on gene expression patterns in our single-cell datasets (Fig. 1d). If suitable antibodies were available, we tested these candidate markers for their specificity by immunofluorescence staining of prostate sections (Figure 1D,E; Figure 1—figure supplement 3). For example, we found that the LumA and LumD marker *Tgm4* was highly expressed by luminal cells in the distal region of the AP and DP, and that the LumL marker *Msmb* marked distal luminal cells in the LP. In contrast, the LumP marker *Ppp1r1b* was highly enriched in luminal cells that were primarily found in the proximal regions of all four lobes (Figure 1D,E). However, more general luminal markers such as androgen receptor (*Ar*) were expressed in all luminal populations (Figure 1—figure supplement 3).

Next, we investigated the spatial localization of these luminal populations along the proximal-distal axis in each lobe. We found that the LumP-containing proximal region extended from inside the rhabdosphincter to the first major ductal branch point in the AP, DP, and VP, but not the LP, whereas the bulk of the lobes corresponded to distal regions (Figure 1C). In the AP and DP, we found a discrete boundary in the medial region between the proximal LumP population and distal LumA or LumD populations, respectively (Figure 2A,C; Figure 2—figure supplement 1A). In contrast, the LP had a population between the proximal and distal regions that expressed low levels of LumL markers (Figure 2A; Figure 2—figure supplement 1A). Histological analyses

revealed that distal luminal cells of each lobe had a tall columnar appearance consistent with secretory function, whereas proximal LumP cells typically had a cuboidal morphology. Notably, this analysis also revealed heterogeneity in the distal region of each lobe, with rare clusters of 1-10 LumP cells observed in the distal AP, DP, and LP, but larger LumP clusters in the distal VP (Figure 2B).

To clarify the phenotypic differences between proximal and distal luminal populations, we performed scanning electron microscopy of an 8-week old anterior lobe (Figure 2C; Figure 2—figure supplement 1B). LumA cells displayed dense regions of rough endoplasmic reticulum throughout the cytoplasm, many free ribosomes, and abundant secretory vesicles on the apical surface, typical of secretory cells. In contrast, LumP cells displayed areas of high mitochondrial density, complex membrane interdigitation, and no vesicles. At the proximal-distal boundary, we observed an abrupt transition between cellular morphologies that took place within 1-2 cell diameters. These ultrastructural differences indicate that the LumA and LumP populations represent distinct cell types, rather than cell states.

Next, we investigated the remaining epithelial population, which shares basal and luminal features in our single-cell RNA-seq analysis (Figure 1A). We found that this small population was enriched in the most proximal region of all four lobes, residing inside the rhabdosphincter and adjacent to the urethral junction (Figure 2D; Figure 1—figure supplement 1C); hence, we termed this novel population PrU (periurethral). Although this PrU population co-expresses some markers with LumP (Figure 1D), it also expresses several urothelial markers such as *Ly6d* and *Aqp3* (Figure 2D).

The proximity of the PrU and LumP populations to the urethra and their co-expression of multiple markers led us to investigate their developmental origin. Consequently, we examined the expression of *Nkx3-1,* whose mRNA expression marks epithelial cells in ductal derivatives of the developing urogenital sinus, such as the prostate, but not the urothelium (Bhatia-Gaur et al., 1999); similarly, the *Nkx3-1*^{Cre} driver also marks early prostate bud cells but not the urogenital sinus

during development (Thomsen et al., 2008; Zhang et al., 2008). In the adult prostate, *Nkx3-1* is expressed by all four distal luminal populations (LumA, LumD, LumL, LumV), but is not expressed by LumP (Figure 1—figure supplement 3). However, we found the *Nkx3-1*^{Cre} driver lineage-marks most of the cells in all of these populations, including LumP and PrU, but not the urethra (Figure 2E). These data indicate that the LumP and PrU populations are derived from *Nkx3-1* expressing prostate epithelial cells, and are distinct from the urothelium.

Functional analysis of epithelial populations

We used an approach based on optimal transport theory (Wasserstein distance) to ascertain the relationships of these prostate epithelial populations (see *Methods*). The pair-wise comparisons (Figure 3A) can be captured by a neighbor-joining tree (Figure 3B), in which lower Wasserstein distance indicates greater similarity. We found that the distal populations grouped together, with the LumA and LumD populations being most closely related, followed by LumL and LumV. These distal populations were next most closely related to LumP, which in turn was most similar to PrU, followed by basal cells, suggesting a lineage relationship between LumP and distal luminal populations.

To investigate the functional properties of each epithelial population, we developed isolation strategies using microdissection and flow sorting (Figure 3C). We performed organoid formation assays with isolated cell populations using the defined ENR-based medium (Drost et al., 2016; Karthaus et al., 2014) as well as hepatocyte medium (HM), which has a more complex composition including serum (Chua et al., 2014). Despite differences in overall efficiency between media conditions, we consistently found that the PrU and basal populations were most efficient at forming organoids, followed by LumP, whereas the efficiency of distal LumA, LumD, LumL, and LumV was significantly lower (Figure 3D,E).

Next, we assessed the progenitor potential of isolated epithelial populations using *in vivo* tissue reconstitution assays. These assays involve recombination of dissociated epithelial cells with

rat embryonic urogenital mesenchyme followed by renal grafting, and have been extensively utilized for analysis of progenitor properties in the prostate (Lawson et al., 2007; Wang et al., 2013; Xin et al., 2003). We observed significant variation in the frequency of graft formation depending on the number and type of input epithelial cells (Figure 3F). Based on histological and immunostaining analyses, we found that each epithelial population typically gave rise to cells of the same type, but their ability to generate cells of other populations varied considerably (Figure 3G; Figure 3—figure supplement 1A). Notably, grafts using distal luminal cells required relatively large numbers of input cells (approximately 30,000 cells) (Figure 3F), and were mostly composed of the same population (*i.e.,* LumA cells generated LumA cells) together with a normal or reduced percentage of basal cells; additionally, these grafts could contain small patches of cells expressing LumP markers on their periphery (Figure 3G; Figure 3—figure supplement 1A). Interestingly, LumV cells had the lowest grafting efficiency and generally formed small ductal structures that lacked basal cells. In contrast, LumP, PrU, and basal cells could produce grafts with significantly lower input cell numbers (approximately 1,000 cells), which contained LumP cells together with multiple distal luminal populations as well as a normal ratio of basal cells. We excluded contribution of host cells and rat urogenital epithelium to grafts using control tissue reconstitution assays with GFP-expressing donor epithelial cells (Figure 3—figure supplement 1B).

Taken together, these results suggest a spectrum of progenitor potential among the different epithelial populations. Both PrU and basal cells possessed high progenitor activity in these assays, with LumP cells also displaying enhanced activity. In contrast, the distal luminal populations are much less efficient in these assays. These findings are consistent with the inferred relationships between these populations based on molecular (Figure 3A,B) as well as histological and ultrastructural analyses (Figure 2A,C).

Comparison of human and mouse prostate epithelial populations

To examine the conservation of epithelial populations between the mouse and human prostate, we performed single-cell RNA-seq analyses of tissue samples from human

196 prostatectomies (Figure 4—source data). We identified two distinct luminal populations as well as
197 a PrU-like population (Figure 4A-C), and examined their spatial distribution by immunostaining
198 of benign human prostate tissue (Figure 4—source data). Using specific markers (KRT7,
199 RARRES1), we found that one luminal population was primarily localized to ducts (Lum Ductal),
200 whereas the second one, positive for MSMB and MME, was predominantly acinar (Lum Acinar),
201 although some intermixing could be observed within ducts (Figure 4G). As in the mouse, the PrU-
202 like population expressed both basal and luminal genes. Similarly, some PrU-like markers were
203 shared with Lum Ductal, though these two populations are clearly distinguishable based on
204 anatomical and histological features.

205 To extend this analysis, we computed the Wasserstein distances for each pair-wise
206 comparison between epithelial and major non-epithelial populations identified (Figure 4D-F). This
207 analysis showed that a PrU-like population was conserved between species. Furthermore, the
208 human Lum Ductal cells are most closely related to mouse LumP, whereas the Lum Acinar cells
209 are most closely related to LumL followed by LumV. Notably, these relationships were observed
210 in each dataset.
211

Discussion

We have generated a comprehensive cellular atlas of the prostate epithelium and have defined spatial, morphological, and functional properties of each epithelial population. Our analyses have revealed spatial and functional heterogeneity primarily in the luminal epithelial compartment, including distinct cell populations along the proximal-distal axis as well as lobe-specific identities. Notably, we have shown marked differences in progenitor potential between cell identities, which likely correspond to distinct cell types rather than cell states (Morris, 2019). In tissue reconstitution assays, the ability of LumP and PrU cells to generate luminal distal and basal cells suggests that both populations have properties of multipotent progenitors. In contrast to the luminal compartment, basal cells appear relatively homogeneous, suggesting that previously reported basal heterogeneity (Goldstein et al., 2008; Lawson et al., 2007; Wang et al., 2020) may be more limited. Notably, the PrU population is not readily classified in either compartment as it is comprised of cells with both basal and luminal features.

Our findings also provide a broader context for other reports of epithelial heterogeneity. Recent single-cell RNA-seq analysis of the mouse anterior prostate identified three distinct luminal populations (Karthaus et al., 2020), where L1 appears to correspond to our LumA and L2 to LumP; we also identified a population expressing L3 genes in both mouse and human datasets, but have annotated this as ductus/vas deferens based on marker expression (Figures 1A, 4C), relying on previous findings (Blomqvist et al., 2006). Based on patterns of gene expression, we suggest that the LumP population corresponds to Sca1^{high} luminal cells (Kwon et al., 2016) as well as Trop2 (Tacstd2) positive cells (Crowell et al., 2019), which have been described as proximal progenitor populations with scattered distal cells, and may also be responsible for the enhanced serial grafting efficiency of proximal prostate (Goto et al., 2006). In the human prostate, the progenitor activity of LumP and/or PrU-like cells may have been observed by retrospective lineage tracing using mitochondrial mutations (Moad et al., 2017). In addition, the LumP and PrU populations may share some similarities with the “hillock” and “club” cells originally described in a cellular atlas

238 of the mouse lung (Montoro et al., 2018) and subsequently reported in the human prostate (Joseph
239 et al., 2020), but the precise relationship of these populations is unclear.

240 Since LumP and PrU cells display multipotent progenitor activity in both organoid
241 formation and tissue reconstitution assays, their spatial distribution may reflect the ability of the
242 prostate to repair itself from a proximal to distal direction in response to extensive tissue damage
243 as well as from distal progenitors in response to more localized injury. Our findings suggest that
244 the novel PrU population in particular may play a role in prostate tissue repair and/or regeneration,
245 consistent with the previous identification of Ly6d-positive cells as a castration-resistant
246 progenitor (Barros-Silva et al., 2018). Notably, the ability of *Nkx3-1*^{Cre} to lineage-mark both PrU
247 and LumP, but not the urethra, suggests that these cell populations are distinct from the urothelium,
248 despite the molecular similarities between the proximal prostate and the urethra noted in a recent
249 report (Joseph et al., 2020). Nonetheless, our results imply that lineage relationships among the
250 tissues derived from the urogenital sinus (Georgas et al., 2015) require careful elucidation, since
251 they are of fundamental importance for understanding the genesis of congenital defects in the
252 urogenital system.

253 Our findings help resolve a long-standing question about the relationship of the mouse and
254 human prostates. Specifically, we speculate that mouse proximal and distal regions are most
255 related to human ductal and acinar regions, respectively, and that the mouse LP is most similar to
256 the human peripheral zone. Few if any studies have specifically assessed tumor phenotypes in the
257 LP, as it is small and usually combined with the DP as the dorsolateral prostate (DLP); however,
258 our analyses show that the DP differs significantly from the LP at the anatomical and molecular
259 levels. Consequently, we suggest that a re-evaluation of tumor phenotypes in genetically-
260 engineered mouse models may reveal a closer similarity to human prostate tumor histopathology
261 than previously appreciated.

262 Finally, the elucidation of prostate epithelial heterogeneity has potentially significant
263 implications for understanding the cell of origin for prostate adenocarcinoma. Previous studies

have suggested that luminal cells as well as basal cells can serve as the cell of origin for prostate cancer (Wang et al., 2013; Wang et al., 2014; Xin, 2019), yet known differences in human prostate cancer outcome (*e.g.,* (Zhao et al., 2017)) cannot be simply explained on this basis. Therefore, further analyses of epithelial heterogeneity and progenitor potential will likely lead to key insights into prostate tumor initiation and progression.

Key Resources Table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Strain, strain background (*Mus musculus*)	C57BL6/N (wild type)	Taconic	C57BL6/Ntac	8-10 week old males
Strain, strain background (*Mus musculus*)	SW (wild type)	Taconic	Tac:SW	8-10 week old males
Strain, strain background (*Mus musculus*)	UbC-GFP	Jackson Laboratory, JAX #004353	C57BL/6-Tg(UBC-GFP)30Scha/J	BL6 background, 8-13 week old males
Strain, strain background (*Mus musculus*)	R26r-YFP	Jackson Laboratory, JAX #007903	B6.Cg–Gt(ROSA)26Sortm3(CAG−EYFP)Hze/J	8-13 week old males
Strain, strain background (*Mus musculus*)	Nkx3-1Cre	Shen lab	Generated in our lab	BL6 background, 8-13 week old males
Strain, strain background (*Mus musculus*)	R2G2	Envigo	B6;129-Rag2tm1FwaIl2rgtm1Rsky/DwIHsd	8-15 week old males
Strain, strain background (*Mus musculus*)	NOD/SCID	Jackson Laboratory, JAX #001303	NOD.Cg-Prkdcscid/J	8-15 week old males
Strain, strain background (*Rattus norvegicus domestica*)	Sprague-Dawley embryos	Charles River #400	SAS Sprague Dawley	E18 embryos from pregnant females
Antibody	anti-mouse Cd66a (CEACAM1)-PE	Miltenyi	cat 130-106-209, lot 5190208411	(FACS 1:100uL)
Antibody	anti-mouse Tacstd2 (Trop2)-APC	R&D	cat FAB1122A, lot AAZB0117091	(FACS 1:100uL)
Antibody	anti-mouse Ly-6a/e Sca-1 PE-Vio770	Miltenyi	cat 130-106-258, lots 5190308494 & 5180615495	(FACS 1:100uL)
Antibody	anti-mouse Cd31 (Lin)-FITC	eBiosciences	cat 11-0311-82, lot 1978184, clone 390	(FACS 1:100uL)
Antibody	anti-mouse Cd45 (Lin)-FITC	eBiosciences	cat 11-0451-82, lot 2015744, clone 30-F11	(FACS 1:100uL)

Antibody	anti-mouse Ter119 (Lin)-FITC	eBiosciences	cat 11-5921-82, lot 2009756, clone TER-119	(FACS 1:100uL)
Antibody	anti-mouse Ly6d -APC-Vio770	Miltenyi	cat 130-115-315, lot 5190715088, clone REA A906	(FACS 1:100uL)
Antibody	mouse anti-mouse DARPP-32 (Ppp1r1b)	SCBT	cat sc-271111, lot C2719, clone H-3	(IF 1:50uL)
Antibody	rabbit anti-mouse DARPP-32 (Ppp1r1b)	Invitrogen	cat MA5-14968, lot VB2947074	(IF 1:400uL)
Antibody	rabbit anti-mouse Trpv6	alomone labs	cat ACC-036, lot ACC036AN1002	(IF 1:100uL)
Antibody	rabbit anti-mouse Lrrc26	alomone labs	cat APC-070, lot APC070AN0102	(IF 1:100uL)
Antibody	rabbit anti-mouse Msmb	Abclonal	cat A10092, lot 204440101	(IF 1:100uL)
Antibody	rabbit anti-mouse Cldn10	Invitrogen	cat 38-8400, lot UA279882	(IF 1:100uL)
Antibody	rabbit anti-mouse Mgll	Invitrogen	cat PA5-27915, lots TI2636340A & VB2935243A	(IF 1:250uL)
Antibody	rabbit anti-mouse Tgm4	Invitrogen	cat PA5-42106, lot uc2737144	(IF 1:100uL)
Antibody	rabbit anti-mouse Gsdma	abcam	cat ab230768, lot GR3212791-1	(IF 1:100uL)
Antibody	rabbit anti-mouse Krt7	abcam	cat ab68459, lot GR40294-6	(IF 1:250-500uL)
Antibody	rabbit anti-mouse Aqp3	Biorbyt	cat orb47955, lot B3440	(IF 1:500uL)
Antibody	chicken anti-mouse Krt5	Biolegend	cat 905901, lot B271562	(IF 1:500uL)
Antibody	rabbit anti-mouse p63	Biolegend	cat 619002, lot B262186	(IF 1:250uL)
Antibody	rat anti-mouse Krt8/18	DSHB	Troma-I NA	(1:250uL, lot specific)
Antibody	mouse anti-mouse Synaptophysin	BD Biosciences	cat BD611880, lot 8290534 2	(IF 1:500uL)

Antibody	rabbit anti-mouse Chromogranin A	abcam	cat ab15160, lot GR3205971-2	(IF 1:500uL)
Antibody	chicken anti-GFP	abcam	cat ab13970, lot GR3190550-30	(IF 1:1000uL)
Antibody	rabbit anti-mouse Nkx3.1	Athena Enzymes	cat "0315" 20316	(IF 1:100uL)
Antibody	rabbit anti-mouse ki67	abcam	cat ab15580, lot GR3198158	(IF 1:100uL)
Antibody	mouse anti-mouse Krt4	Invitrogen	cat MA1-35558, lot TB2524522	(IF 1:100uL)
Antibody	rabbit anti-mouse Clusterin	LS-Bio	cat LS-331486, lot 115142	(IF 1:100uL)
Antibody	rabbit anti-mouse Wfdc2	Invitrogen	cat PA5-80226, lot TK2671201	(IF 1:100uL)
Antibody	rabbit anti-mouse AR (Androgen receptor)	abcam	cat ab133273, lot GR3271456-1	(IF 1:100uL)
Antibody	mouse anti-human Krt7	Thermo Fisher	cat MA1-06316, lot OVTL12/30	(IF 1:200uL)
Antibody	mouse anti-human Rarres1	Thermo Fisher	cat MA5-26247, lot OTI1D2	(IF 1:200uL)
Antibody	mouse anti-human Mme (Cd10)	SCBT	cat sc-46656, clone F-4	(IF 1:100uL)
Antibody	rabbit anti-human Msmb	Abclonal	cat A10092	(IF 1:200uL)
Software, algorithm (code)	Random Matrix Theory	R. Rabadan Lab		https://rabadan.c2b2.columbia.edu/html/randomly/
Software, algorithm (code)	Python Optimal Transport	Rémi Flamary and Nicolas Courty, POT Python Optimal Transport library		https://github.com/rflamary/POT
Software, algorithm (code)	Phylogenetic tree analysis	Phangorn package		https://github.com/KlausVigo/phangorn
Software, algorithm (code)	Leiden algorithm	F. A. Wolf, P. Angerer, and F. J. Theis, Genome Biology (2018). "SCANPY: large-scale single-cell gene expression data analysis"		https://scanpy.readthedocs.io/en/stable

271

Materials and Methods

Mouse strains and genotyping

Wild type C57BL6/N (C57BL6/NTac, 8-10 weeks old) and Swiss-Webster (8-10 weeks) mice were purchased from Taconic. The *Ubc-GFP* (C57BL/6-Tg(UBC-GFP)30Scha/J, 8-13 weeks old; JAX #004353)(Schaefer et al., 2001) and *R26R–YFP* (B6.Cg–*Gt(ROSA)26Sor*$^{tm3(CAG–EYFP)Hze}$/J; JAX #007903) (Madisen et al., 2010) mice were obtained from the Jackson Laboratory. The *Nkx3-1*^{Cre} allele has been previously described (Lin et al., 2007; Thomsen et al., 2008). As hosts for renal grafts, R2G2 mice (B6;129-*Rag2*^{tm1Fwa}*Il2rg*^{tm1Rsky}/DwIHsd, 8-15 weeks old) were purchased from Envigo, and NOD/SCID mice (NOD.Cg-*Prkdc*^{scid}/J, 8-14 weeks old; JAX #001303) were purchased from the Jackson Laboratory. To obtain urogenital mesenchyme for tissue recombination and renal grafting, we used E18 Sprague-Dawley rat embryos from timed matings (Charles River #400). Animal studies were approved by and conducted according to standards set by the Columbia University Irving Medical Center (CUIMC) Institutional Animal Care and Use Committee (IACUC).

Isolation of mouse prostate tissue

The anterior (AP), dorsal (DP), lateral (LP), and ventral prostate (VP) lobes were dissected individually, using a transverse cut at the intersection of each lobe with the urethra to include the periurethral (PrU) region. For some analyses, we dissected PrU tissues in the most proximal regions (extending 0.5-2 mm from the connection of the lobes with the urethra), or the remainder of the proximal regions separately from the distal lobes. Paired lobes were collected from a single C57BL/6 mouse for each scRNA-seq experiment, or from 2-5 C57BL/6 mice for flow sorting, organoid culture, or tissue reconstitution experiments. For analyses of prostate anatomy, lobes were either dissected individually or a deep cut was made at the caudal end of the urethra for removal of entire urogenital apparatus.

296 **Acquisition and pathological assessment of human prostate tissue samples**

297 Human prostate tissue specimens were obtained from patients undergoing
298 cystoprostatectomy for bladder cancer or radical prostatectomy at Columbia University Irving
299 Medical Center or at Weill Cornell Medicine. Patients were aged 54-79 years old and gave
300 informed consent under Institutional Review Board-approved protocols. The clinical
301 characteristics of these patients are provided in Figure 4—source data. Following surgery, prostate
302 tissue was submitted for gross pathological annotation and sectioning, with ischemic time less than
303 one hour.

304 To acquire samples for single-cell RNA-sequencing, the prostate was transversely
305 sectioned perpendicular to the urethra in three main parts (apex, mid and base), which were further
306 divided based on laterality (left or right). Each part was cut in thick sections that included all three
307 prostatic zones (peripheral, transversal and central). Thick sections with low or no tumor burden
308 were selected for the study based on clinical findings and/or biopsies, and divided in three plates
309 by performing two parallel cuts. The upper flanking plate was flash-frozen, cryosectioned, and a
310 rapid review was performed by a board-certified surgical pathologist (H.H.) to provide preliminary
311 assessment on the presence of benign prostate tissue/absence of carcinoma. The middle flanking
312 plate was stored in RPMI medium with 5% FBS on ice, and immediately transferred to the research
313 facility for single-cell RNA sequencing. The lower flanking plate was processed by formalin
314 fixation and paraffin embedding, followed by sectioning and histological review to confirm
315 presence of benign prostate tissue/absence of carcinoma.

316 For immunostaining analysis of prostate tissue sections, blocks previously assessed as
317 containing benign prostate histology were selected by a surgical pathologist (B.D.R.). Paraffin
318 sections were immunostained for markers of interest as well as CK5 to confirm the presence of
319 basal cells, and adjacent sections were stained by H&E. H&E sections were then reviewed to
320 confirm benign pathology.

Dissociation of mouse and human prostate tissue

Prostate tissues were minced with scissors and then incubated in papain (20 units/ml) with 0.1 mg/ml DNase I (Worthington LK003150) at 37°C with gentle agitation. After 45 minutes, samples were gently triturated, then incubated for another 20-45 minutes in papain as needed. Samples were gently triturated again, followed by quenching of the enzyme using 1 mg/ml ovomucoid/bovine serum albumin solution with 0.1 mg/ml DNase I (Worthington LK003150). Cells were passed through a 70 µm strainer (PluriSelect 43-10070-70) and washed with PBS-EDTA with 0.1 mg/ml DNase I (Corning MT-46034CI). If needed, the samples were additionally digested in TrypLE Express (Invitrogen 12605-036) for 3-5 minutes at 37°C with gentle agitation. The samples were gently triturated and the TrypLE was inactivated by addition of HBSS with 10% FBS and 0.1 mg/ml DNase I. Samples were passed through a 40 µm strainer (PluriSelect 43-10040-70), washed in 1x PBS, and resuspended in appropriate buffers for downstream analyses.

Single-cell RNA-sequencing

Dissociated cells were washed twice in 1x PBS, passed twice through 20 µm strainers (Pluriselect 43-10020-70), and counted using the Countess™ II FL Automated Cell Counter (ThermoFisher). If the viability of samples was >80% and the single-cell fraction was >95%, the cells were resuspended in 1x PBS with 0.04% BSA at approximately 1 x 10^6 cells/ml. Samples were submitted to the Columbia JP Sulzberger Genome Center for single-cell RNA-sequencing on the 10x Genomics Chromium platform. Libraries were generated using the Chromium Single Cell 3' Reagent Kit v2, with 12 cycles for cDNA amplification and 12 cycles for library construction. Samples were sequenced on a NovaSeq 6000 (r1 = 26, i1 = 8, r2 = 91). Sequencing data were aligned and quantified using the Cell Ranger Single-Cell Software Suite (v.2.1.1) using either the GRCm38 mouse or the GRCh38 human reference genomes.

For the mouse prostate, two independent biological replicate samples of whole prostate were submitted for scRNA-seq, with 2,361 and 2,927 cells sequenced. Two separate biological

346 replicates for individual lobes were also used for scRNA-seq, with 1,581-2,735 cells sequenced
347 for sample. For human prostate (Figure 4—source data), three independent samples (#1-3) that
348 corresponded to regions of benign histology were submitted for scRNA-seq, with 1,600, 2,303,
349 and 2,825 cells sequenced, respectively. Single-cell datasets have been deposited in GEO under
350 accession number GSE150692.

351 **Flow cytometry**

352 Dissociated cell suspensions were counted and resuspended in FACS buffer (1-3% fetal
353 bovine serum in 1x PBS or HBSS, 1 mM EDTA, and 0.1 mg/ml DNase I). Pre-incubation was
354 performed with Fc receptor blocking antibodies for some experiments. Primary antibodies were
355 added at a final concentration of 1:100 and incubated at 4°C for 20-30 minutes. Samples were
356 incubated with 1 µM propidium iodide for 15 minutes before sorting for dead cell exclusion. Cells
357 were sorted using a BD Influx, using the widest nozzles and lowest pressure settings (140 µm
358 nozzle and 7 psi for the Influx), and collected in low-binding tubes (Eppendorf 0030-108-116).
359 Data analysis was conducted using FCS Express 7 software. A slightly modified strategy was used
360 to sort prostate cells from *Ubc-GFP* mice.

361 **Organoid culture**

362 For ENR conditions (Karthaus et al., 2014), sorted cells from *Ubc-GFP* mice were
363 resuspended at a final concentration of 1,000 cells per 30 µl droplet of Matrigel (Corning 354234),
364 and placed in individual wells of a 24-well plate. The Matrigel was covered with 500 µl of ENR
365 medium, supplemented with 10 nM dihydrotestosterone (DHT) and 10 µM Y-27632. Media were
366 replenished at day 5 and organoids were imaged at day 10. Organoid measurements were
367 performed using the Fiji Particle Analysis Plugin (Rueden et al., 2017; Schindelin et al., 2012),
368 excluding particles with area < 2,000 μm^2 and roundness value < 0.5. If needed, Watershed was

applied to separate overlapping organoids/particles, with wells from at least 4 independent experiments analyzed.

For hepatocyte media (HM) conditions (Chua et al., 2014), sorted cells from wild-type C57BL/6 mice were plated at 2,000 and 5,000 cells per well in 5% Matrigel, 10 nM DHT, and 10 μM Y-27632 using ultra-low attachment 96-well plates (Corning 3474) and grown in hepatocyte media with 5% Matrigel, with media replenished every 5 days. Organoid formation efficiencies were calculated on day 12-13 of culture. Since LumV cells tended to form small structures containing 1-4 cells, we used a required minimum cut-off size; organoid images were analyzed using ImageJ. Data were collected from 3 biological replicate experiments, with a minimum of 2-3 technical replicates for each population in each experiment.

Renal grafting

For tissue reconstitution experiments, urogenital mesenchyme (UGM) cells were collected from embryonic day 18.5 rat embryos as described (Chua et al., 2018) and passed through a 100 μM filter (Pluriselect) before use. Sorted mouse epithelial cells were used at ranges from 250 to 60,000 cells depending on the specific population; since basal cells have been previously examined in graft experiments (*e.g.,* (Wang et al., 2013)), we did not investigate the minimum number of basal cells required for graft growth in these experiments. Rat UGM cells and sorted mouse epithelial cells were combined at pre-determined ratios (*e.g.,* 250,000 UGM:5,000 LumP cells) and resuspended in 10-15 μl buttons composed of 9:1 collagen 1 (Corning 354249):setting solution (10x EBSS, 0.2 M $NaHCO_3$, and 50 mM NaOH). After solidification of the collagen, the buttons were incubated in DMEM media with 10% FBS and 10 nM DHT overnight, followed by grafting under the kidney capsule of host immunodeficient mice on the next day. At the time of surgery, a slow-release testosterone pellet (12.5 mg testosterone, 90 day release; Innovative Research of America NA-151) was inserted subcutaneously in each host mouse. Grafts were analyzed 8-12 weeks after surgery.

Histology, immunostaining, and image analysis

For generation of paraffin blocks, prostate tissues were dissected in ice-cold HBSS and fixed in 10% formalin overnight, followed by processing through an ethanol gradient and embedding. For generation of frozen blocks, dissected tissues were fixed in 4% paraformaldehyde, immersed in sucrose overnight (30% in 1x PBS), embedded in OCT (Tissue-Tek 25608-930), and stored at -80°C. Alternatively, samples were flash-frozen in 2-methyl-butane (Sigma-Aldrich M32631) at -150°C for 1 hour, then stored at -80°C. Paraffin-embedded tissues were sectioned using a MICROM HM 325 microtome, and cryo-preserved tissues were sectioned using a Leica CM 1900 cryostat, at thicknesses of 5-13 μm.

For histological analyses, hematoxylin-eosin (H&E) staining was performed on paraffin sections using standard procedures. For immunofluorescence staining of paraffin sections, antigen retrieval was performed by boiling slides in citrate-based or Tris-based antigen unmasking buffer (Vector Labs H3300 and H3301) for 45 minutes. For immunofluorescence of cryosections, slides were rapidly fixed in either 4% paraformaldehyde or 10% NBF for 5 minutes after sectioning. Slides were washed, blocked in 5% animal serum for 1 hour, and incubated with primary antibodies overnight at 4°C. Slides were washed and incubated with Alexa Fluor secondary antibodies (Life Technologies) for one hour. Sections were stained with DAPI, and mounted (Vector Labs H-1200). Fluorescent images were acquired using a Leica TCS SP2, a Leica TCS SP5, or a Nikon Ti Eclipse inverted confocal microscope.

Electron microscopy

Prostate tissue from a C57BL/6 mouse at 8 weeks of age was dissected and fixed for 2 hours in 0.1 M sodium cacodylate buffer (pH 7.2) containing 2.5% glutaraldehyde and 2% paraformaldehyde. A portion of the AP lobe at the proximal-distal boundary was micro-dissected and post-fixed for two hours with 1% osmium tetroxide, contrasted with 1% aqueous uranyl acetate, dehydrated using an ethanol gradient and embedded in EMbed 812 (Electron Microscopy

Sciences, Hatfield, PA). 70 nm ultrathin sections were cut, mounted on formvar coated slot copper grids, and stained with uranyl acetate and lead citrate. Stained grids were imaged with a Zeiss Gemini300 scanning electron microscope using the STEM detector.

Statistical analysis

Prism v.8 was used for statistical analyses of functional data and for plot generation. For analyses of organoid formation efficiencies, the data passed the Shapiro-Wilk test for normal distribution ($p>alpha=0.05$), but did not pass the Bartlett's test for equal variance ($p<0.05$). We therefore used the Brown-Forsythe and Welch One-way Analysis of Variance (ANOVA) to confirm statistically significant differences between organoid populations ($p<0.0001$ for both HM and ENR conditions). Since all populations have fewer than 50 data points per sample, we used the Dunnett's T3 multiple comparisons test to determine which populations significantly differ ($p<0.05$). The p-values on the graphs indicate the least significant difference observed between compared populations.

For analyses of graft efficiency, the data did not pass the Shapiro-Wilk test for normal distribution ($p<a=0.05$), or Bartlett's test for equal variance ($p<0.05$). We therefore used a non-parametric Kruskal-Wallis test to confirm statistically significant differences between graft input cell numbers by epithelial population ($p<0.0001$), followed by the two-stage linear step-up procedure of Benjamini, Krieger and Yekutieli ($p<0.05$). The p-value on the graph indicates the average significant difference between the up to 10 lowest input cell numbers for each distal luminal population compared to the LumP population.

Bioinformatic analysis of single-cell RNA-seq data

Filtering the expression matrix

The starting pool of cells in the mouse prostate analysis is 13,429 cells, which is composed of two whole prostate samples of 2,361 and 2,927 cells, and 4 samples corresponding to each of

the lobes at 2,735 (AP), 1,781 (DP), 2,044 (LP), and 1,581 (VP) cells. The starting pool of cells for human prostate analyses is 6,728 cells coming from three independent samples of 2,303, 1,600 and 2,825 cells each. When filtering the data, we removed cells with less than 500 genes detected and cells with >10% of total transcripts derived from mitochondrial-encoded genes. The expression matrices are normalized by $log_2(1 + TPM)$, where TPM denotes transcripts per million.

Batch effect correction

For Figure 1A, we have aggregated the two samples corresponding to the whole mouse prostate. As a first step to remove the batch effect we have used the algorithm described in (Stuart et al., 2019), using default parameters.

Random Matrix Theory application to single-cell transcriptomics

Random Matrix Theory (RMT) is a field with many applications in different branches of mathematics and physics. The mathematical foundations of RMT were developed by the theoretical physicist Dyson in the 1960's when he described heavy atomic nuclei energy levels. One of the deepest properties of RMT is universality, *i.e.,* the insensitivity of certain statistical properties to variations of the probability distribution used to generate the random matrix. This property provides a unified and universal way to analyze single-cell data, where the gene and cell expression distributions are different for each cell. By using RMT universality, one can address the specific sparsity and noise of each single-cell dataset.

In this work we have used *Randomly,* an RMT-based algorithm (Aparicio et al., 2020). The idea of this algorithm is based on the fact that a single-cell dataset shows a threefold structure: a random matrix, a sparsity-induced (fake) signal and a biological signal. Indeed, 95% or more of the single-cell expression matrix is compatible with being a random matrix and hence, in such a case, with being pure noise (Aparicio et al., 2020). In order to detect the part of the expression matrix compatible with noise, *Randomly* uses the universality properties of RMT. More

specifically, let us suppose a $N \times P$ expression matrix X, where N is the number of cells and P is the number of genes, and where each column is independently drawn from a distribution with mean zero and variance σ, the corresponding Wishart matrix is defined as an $N \times N$ matrix:

$$W = \frac{1}{P} XX^T$$

The eigenvalues λ_i and normalized eigenvectors ψ_i of the Wishart matrix where $i = 1, 2, \ldots N$ are given by the following relation:

$$W\psi_i = \lambda_i \psi_i$$

If X happens to be a random matrix (a matrix whose entries x_{ij} are randomly sampled from a given distribution), then W becomes a random covariance matrix and the properties of its eigenvalues and eigenvectors are described by Random Matrix Theory. Universality properties of RMT arise in the limit $N \to \infty,\ P \to \infty,\ \gamma = \frac{N}{P}$ fixed. One of the consequences of universality at the level of eigenvalues λ_i, is that empirical density of states converges to the so-called Marchenko-Pastur (MP) distribution:

$$\rho_{MP}(\lambda) = \frac{1}{2\pi\gamma\sigma^2} \frac{\sqrt{(a_+ - \lambda)(\lambda - a_-)}}{\lambda} \mathbb{I}_{[a_-, a_+]}$$

where

$$a_\pm = \sigma^2(1 \pm \sqrt{\gamma})^2$$

and σ represents the variance of the probability distribution that generates each element in the random matrix ensemble. Any deviation of the eigenvalues from MP distribution would imply that the expression matrix X is not completely random, and therefore contains a signal that could be further analyzed.

One of the main novelties of the *Randomly* algorithm is the study of eigenvectors. At the level of eigenvectors, RMT universality is manifested through the so-called eigenvector delocalization, which implies that the norm of the eigenvectors ψ_i is equally distributed among all their components α:

492 $$\left|\psi_i^{(\alpha)}\right| \sim \frac{1}{\sqrt{N}}$$

493 Interestingly, the distribution of components for delocalized eigenvectors at large N approximates
494 a Gaussian distribution with mean zero and 1/N variance

495 $$f(\psi) \sim \frac{N}{\sqrt{2\pi}} e^{\left(\frac{-N\psi^2}{2}\right)}$$

496 The presence of any localized (non-delocalized) eigenvector implies that expression matrix X is
497 not completely random, and hence the existence of a signal that carries information. However, in
498 single-cell datasets, there is a very important subtlety due to the sparsity, which can generate a
499 fake signal (Aparicio et al., 2020). At the single-cell analysis level, the presence of localized
500 eigenvectors related with sparsity implies the existence of an undesired (fake) signal. The
501 *Randomly* algorithm is able to eliminate the sparsity-induced (fake) signal and isolate the
502 biological signal.

503 In Figure 1—figure supplement 2, we show one example of the performance of the
504 *Randomly* algorithm performance. Figure 1—figure supplement 2E shows a comparison of the
505 eigenvector localization with and without sparsity-induced signal in one of the single-cell mouse
506 datasets. The number of eigenvectors that carries signal is larger for the case with sparsity-induced
507 signal. After removal of the fake signal due to sparsity, Figure 1—figure supplement 2F shows the
508 distribution of eigenvalues and the fraction behaving in agreement with the MP distribution. More
509 than 95% of the expression matrix is compatible with a random matrix and therefore is equivalent
510 to random noise. In Figure 1—figure supplement 2G and 2H, the algorithm projects the original
511 expression matrix into the signal-like eigenvectors and the noise-like eigenvectors and performs a
512 chi-squared test for the variance (normalized sample variance), which allows identification of the
513 signal-like genes based on a false discovery rate.

514 The algorithm *Randomly* is a public Python package and can be found in
515 https://rabadan.c2b2.columbia.edu/html/randomly/.

Clustering

Clustering has been performed using the Leiden algorithm as is implemented in (Wolf et al., 2018). The selection of the number of clusters is based on the mean silhouette score. More specifically, we performed a set of clustering performances for different Leiden resolution parameters and compute the mean silhouette score for each case. The silhouette coefficient for a specific cell is given by:

$$s = \frac{b - a}{\max(a, b)}$$

where the a is the mean distance between a cell and all the other cells of the same class, and parameter b is the mean distance between a cell and all other cells in the next nearest cluster. Figure 1—figure supplement 2I shows the mean silhouette score as a function of the Leiden resolution parameter and the number of clusters for each case. The strategy we follow is selecting the number of clusters that maximizes this correlation. In some cases, it could be also useful to sub-cluster some of the clusters, repeating the strategy just described for one specific cluster. The sub-clustering has been used to disentangle immune populations or the vas deferens and the seminal vesicle populations.

t-SNE representations and gene visualizations

In order to visualize single-cell clusters, we performed a further dimensional reduction to two dimensions using t-distributed Stochastic Neighbor Embedding (t-SNE) representation. We used the default parameters, which are: Learning rate = 1000, Perplexity = 30 and Early exaggeration = 12. The tSNE, dot-plots, and ridge-plots were generated using the visualization functions of the *Randomly* package (Aparicio et al., 2020).

Comparison of RMT with traditional pipelines based on PCA dimensional reduction

To show a comparison with traditional approaches based on PCA, we have followed the pipeline in a public tool (Wolf et al., 2018) often used for single-cell analysis. We have performed

a PCA reduction, selecting principal components (PCs) through accumulated variance changes across the different PCs (Figure 1—figure supplement 2A). In this case, only 10 PCs are selected following this approach. After the dimensional reduction, we performed a clustering following the strategy described in the previous section (Figure 1—figure supplement 2B), selecting the number of clusters that maximize the mean silhouette score. Comparing Figure 1—figure supplement 2B and 2I, it is clear that the RMT generates a better clustering performance: the maximum of the silhouette score curve is larger than that generated by the traditional PCA approach, and one of these clusters is able to capture the periurethral (PrU) population (Figure 1—figure supplement 2J). On the other hand, the method based on traditional PCA is not able to capture the PrU population even if we allow for larger Leiden resolutions (Figure 1—figure supplement 2C and 2D).

Differential expression analysis

To test for differentially expressed genes among the different populations of prostate luminal cells, we have used a t-test on the datasets after de-noising them with *Randomly*. The p-value have been corrected for multiple hypothesis using Benjamini-Hochberg. We have used the implementation of (Wolf et al., 2018) with overestimation of the variance and comparison with a Wilcoxon test. Based on this analysis, we have selected genes with a corrected p-value smaller than 0.001.

Mouse population similarity

To calculate the phenotypic similitude/distance between epithelial populations in the mouse prostate, we have performed an analysis based on Optimal Transport (OT) (Kolouri et al., 2017; Villani, 2003). More specifically, we have used the Wasserstein-1 distance as a measure for phenotypic distance between cell populations, *i.e.,* among clusters in the latent space obtained after using *Randomly*. The Wasserstein-1 distance is defined as a distance function between probability

564 distributions in a given metric space. Assuming that the metric space is Euclidean, the Wasserstein-
565 1 distance (Kolouri et al., 2017; Villani, 2003) is defined as:

566 $$W_1(\mu,\nu) = min\left\{\int_{\mathbb{R}^d\times\mathbb{R}^d} \|y-x\|\,\gamma(dx,dy): \gamma \in couplings(\mu,\nu)\right\}$$

567 It is also known as the earth mover's distance, in which each probability distribution can be seen
568 as an amount of dirt piled in the metric space, with the Wasserstein distance corresponding to the
569 cost of turning one pile into the other. In our case, we would be evaluating the cost of transforming
570 one population into another.

571 The optimization of the Wasserstein calculation can be turned into an OT problem, based
572 on the Sinkhorn algorithm and the entropic regularization technique (Altschuler et al., 2017; Chizat
573 et al., 2018; Cuturi, 2013; Schmitzer, 2016). We have used the Python implementation of the
574 package POT Python Optimal Transport library (https://github.com/rflamary/POT), which solves
575 the entropic regularization OT problem and return the loss ($W_1(\mu,\nu)$). We have used as metric
576 cost matrix a Euclidean pairwise distance matrix and assumed that the cell populations correspond
577 to uniform probability distributions defined in the latent space obtained after using *Randomly*.

578 To calculate the distances between populations, we have constructed a matrix of
579 Wasserstein distances among the epithelial populations described in Figure 1 and visualized it
580 using a heatmap and a hierarchical clustering (Figure 3A). We have also generated a tree-like
581 visualization of all the information contained in the hierarchical clustering/heatmap using a
582 neighbor joining algorithm (Schliep, 2011). The length of the branches in the tree is measured in
583 units of the Wasserstein-1 distance (Figure 3B).

584 **Cross-species analysis**

585 We have performed a comparison between the epithelial populations in human and mouse
586 based on OT and Wasserstein distance. To harmonize the human and mouse datasets, we first
587 constructed a common latent space between the aggregated mouse data set and each of the three

588 human samples. To that end, we first looked for the mouse orthologous genes, and then normalized
589 mouse and human separately using $log_2(1 + TPM)$. We filtered out any gene which has an
590 average expression smaller than 0.1 for human or mouse, and merged the two corresponding
591 human and mouse datasets. Finally, we used *Randomly* to generate the common latent space.

592 We have used the Wasserstein distance to calculate the similitude among the clusters of
593 points previously identified with the different mouse and human populations in Figures 1 and 4.
594 We have visualized this with a set of nested heatmaps (Figure 4D-F) to make explicit which
595 populations have the minimum Wasserstein distance between each human population and mouse
596 populations.

597 We then validated the accuracy of this strategy. The first validation test is that the
598 conserved epithelial populations Basal and Lum P in human have a minimum in the Wasserstein
599 distance with their mouse equivalents. A second test of the robustness is to compare cell types that
600 are known to be well-conserved across species, such as immune cells. As with the conserved
601 epithelial cell types, the human immune cell populations have also a minimum Wasserstein
602 distance with respect to the corresponding mouse immune populations.
603

Data availability

Single-cell RNA-sequencing data from this study have been deposited in the Gene Expression Omnibus (GEO) under the accession number GSE150692. All other data are available from the authors upon reasonable request.

Acknowledgements

We would like to thank Reuben Akabas, Sarah Bergren, Mykola Bordyuh, Eva Leung, Bo Li, Maximilian Marhold, Agnieszka Pastula, Luis Pina, Roxanne Toivanen, and Sven Wenske for their assistance with this project, and Cory Abate-Shen and Jia Li for comments on the manuscript. We thank Erin Bush and Peter Sims for assistance with single-cell sequencing in the Columbia Genomics and High Throughput Screening Shared Resource of the Herbert Irving Comprehensive Cancer Center, as well as the resources of the Cancer Center Flow Core Facility, which are supported in part by the Cancer Center Support Grant P30CA013696. We also thank Michael Kissner and Daniel Troast for assistance with flow cytometry in the Columbia Stem Cell Initiative Flow Cytometry Core, which is supported in part by S10OD026845. For assistance with electron microscopy, we thank Alice Liang, Chris Petzold, and Joseph Sall at the New York University Langone Health DART Microscopy Lab, which is partially funded by NYU Cancer Center Support Grant P30CA016087 and by S10OD019974. These studies were supported by grants from the NIH R01CA238005 (M.M.S.), U54CA193313 (R.R. and M.M.S.), P50CA211024 (B.D.R., M.M.S., M.L.), K99CA194287 (M.S.), and by fellowships from the Department of Defense Prostate Cancer Research Program (W81XWH-18-1-0424; F.C.) and the National Science Foundation (L.C.).

627 ## Author contributions

628 Conceptualization (study design): R.R. and M.M.S.; Conceptualization (data interpretation
629 and presentation): L.C., F.C., L.A., and M.M.S.; Methodology (computational): L.A.;
630 Methodology (flow cytometry): F.C.; Investigation (single-cell analysis): L.A., L.C., F.C., and
631 M.S.; Investigation (computational analysis): L.A.; Investigation (urogenital anatomy): L.C. and
632 F.C.; Investigation (mouse prostate analysis): L.C., F.C., S.X., and W.L.; Investigation (flow
633 cytometry): L.C. and F.C.; Investigation (organoid assays): F.C. and L.C.; Investigation (renal
634 graft assays): L.C. and S.X.; Investigation (statistical analysis): L.C.; Investigation (human
635 prostate analysis): F.C. and B.D.R.; Resources: H.H. and M.L.; Writing (original draft): L.C., L.A.,
636 F.C. and M.M.S.; Writing (review and editing): L.C., F.C. L.A., M.S., S.X., R.R., and M.M.S.;
637 Supervision: R.R. and M.M.S.; Funding acquisition: L.C., F.C., M.S., M.L., R.R., and M.M.S.

638

639

640 ## Competing interests

641 The authors declare that they have no competing interests.

642

643 # References

644 Altschuler J, Niles-Weed J, Rigollet P. 2017. Near-linear time approximation algorithms for
645 optimal transport via Sinkhorn iteration. *Advances in Neural Information Processing*
646 *Systems* **30**.

647 Aparicio L, Bordyuh M, Blumberg AJ, Rabadan R. 2020. A random matrix theory approach to
648 denoise single-cell data. *Patterns* https://doi.org/10.1016/j.patter.2020.100035.

649 Barros-Silva JD, Linn DE, Steiner I, Guo G, Ali A, Pakula H, Ashton G, Peset I, Brown M, Clarke
650 NW, Bronson RT, Yuan GC, Orkin SH, Li Z, Baena E. 2018. Single-cell analysis identifies
651 LY6D as a marker linking castration-resistant prostate luminal cells to prostate progenitors
652 and cancer. *Cell Rep* **25**: 3504-3518 e3506. 10.1016/j.celrep.2018.11.069, PMID:
653 30566873

654 Berquin IM, Min Y, Wu R, Wu H, Chen YQ. 2005. Expression signature of the mouse prostate. *J*
655 *Biol Chem* **280**: 36442-36451. 10.1074/jbc.M504945200, PMID: 16055444

656 Bhatia-Gaur R, Donjacour AA, Sciavolino PJ, Kim M, Desai N, Young P, Norton CR, Gridley T,
657 Cardiff RD, Cunha GR, Abate-Shen C, Shen MM. 1999. Roles for *Nkx3.1* in prostate
658 development and cancer. *Genes Dev* **13**: 966-977. PMID: 10215624

659 Blomqvist SR, Vidarsson H, Soder O, Enerback S. 2006. Epididymal expression of the forkhead
660 transcription factor Foxi1 is required for male fertility. *EMBO J* **25**: 4131-4141.
661 10.1038/sj.emboj.7601272, PMID: 16932748

662 Burger PE, Xiong X, Coetzee S, Salm SN, Moscatelli D, Goto K, Wilson EL. 2005. Sca-1
663 expression identifies stem cells in the proximal region of prostatic ducts with high capacity
664 to reconstitute prostatic tissue. *Proc Natl Acad Sci USA* **102**: 7180-7185. PMID: 15899981

665 Chizat L, Peyre G, Schmitzer B, Vialard F-X. 2018. Scaling algorithms for unbalanced optimal
666 transport problems. *Mathematics of Computation* **87**: 2563-2609.

667 Choi N, Zhang B, Zhang L, Ittmann M, Xin L. 2012. Adult murine prostate basal and luminal cells
668 are self-sustained lineages that can both serve as targets for prostate cancer initiation.
669 *Cancer Cell* **21**: 253-265. 10.1016/j.ccr.2012.01.005, PMID: 22340597

670 Chua CW, Epsi NJ, Leung EY, Xuan S, Lei M, Li BI, Bergren SK, Hibshoosh H, Mitrofanova A,
671 Shen MM. 2018. Differential requirements of androgen receptor in luminal progenitors
672 during prostate regeneration and tumor initiation. *Elife* **7**. 10.7554/eLife.28768, PMID:
673 29334357

674 Chua CW, Shibata M, Lei M, Toivanen R, Barlow LJ, Bergren SK, Badani KK, McKiernan JM,
675 Benson MC, Hibshoosh H, Shen MM. 2014. Single luminal epithelial progenitors can
676 generate prostate organoids in culture. *Nat Cell Biol* **16**: 951-961. 10.1038/ncb3047,
677 PMID: 25241035

678 Crowell PD, Fox JJ, Hashimoto T, Diaz JA, Navarro HI, Henry GH, Feldmar BA, Lowe MG,
679 Garcia AJ, Wu YE, Sajed DP, Strand DW, Goldstein AS. 2019. Expansion of luminal
680 progenitor cells in the aging mouse and human prostate. *Cell Rep* **28**: 1499-1510 e1496.
681 10.1016/j.celrep.2019.07.007, PMID: 31390564

682 Cunha GR, Donjacour AA, Cooke PS, Mee S, Bigsby RM, Higgins SJ, Sugimura Y. 1987. The
683 endocrinology and developmental biology of the prostate. *Endocr Rev* **8**: 338-362.
684 10.1210/edrv-8-3-338, PMID: 3308446

685 Cunha GR, Vezina CM, Isaacson D, Ricke WA, Timms BG, Cao M, Franco O, Baskin LS. 2018.
686 Development of the human prostate. *Differentiation* **103**: 24-45.
687 10.1016/j.diff.2018.08.005, PMID: 30224091

688 Cuturi M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in*
689 *Neural Information Processing Systems* **26**.

690 Drost J, Karthaus WR, Gao D, Driehuis E, Sawyers CL, Chen Y, Clevers H. 2016. Organoid
691 culture systems for prostate epithelial and cancer tissue. *Nat Protoc* **11**: 347-358.
692 10.1038/nprot.2016.006, PMID: 26797458

693 Georgas KM, Armstrong J, Keast JR, Larkins CE, McHugh KM, Southard-Smith EM, Cohn MJ,
694 Batourina E, Dan H, Schneider K, Buehler DP, Wiese CB, Brennan J, Davies JA, Harding
695 SD, Baldock RA, Little MH, Vezina CM, Mendelsohn C. 2015. An illustrated anatomical
696 ontology of the developing mouse lower urogenital tract. *Development* **142**: 1893-1908.
697 10.1242/dev.117903, PMID: 25968320

698 Goldstein AS, Lawson DA, Cheng D, Sun W, Garraway IP, Witte ON. 2008. Trop2 identifies a
699 subpopulation of murine and human prostate basal cells with stem cell characteristics. *Proc*
700 *Natl Acad Sci USA* **105**: 20882-20887. 10.1073/pnas.0811411106, PMID: 19088204

701 Goto K, Salm SN, Coetzee S, Xiong X, Burger PE, Shapiro E, Lepor H, Moscatelli D, Wilson EL.
702 2006. Proximal prostatic stem cells are programmed to regenerate a proximal-distal ductal
703 axis. *Stem Cells* **24**: 1859-1868. PMID: 16644920

704 Ittmann M. 2018. Anatomy and histology of the human and murine prostate. *Cold Spring Harb*
705 *Perspect Med* **8**. 10.1101/cshperspect.a030346, PMID: 29038334

706 Ittmann M, Huang J, Radaelli E, Martin P, Signoretti S, Sullivan R, Simons BW, Ward JM,
707 Robinson BD, Chu GC, Loda M, Thomas G, Borowsky A, Cardiff RD. 2013. Animal
708 models of human prostate cancer: the consensus report of the New York meeting of the
709 Mouse Models of Human Cancers Consortium Prostate Pathology Committee. *Cancer Res*
710 **73**: 2718-2736. 10.1158/0008-5472.CAN-12-4213, PMID: 23610450

711 Joseph DB, Henry GH, Malewska A, Iqbal NS, Ruetten HM, Turco AE, Abler LL, Sandhu SK,
712 Cadena MT, Malladi VS, Reese JC, Mauck RJ, Gahan JC, Hutchinson RC, Roehrborn CG,
713 Baker LA, Vezina CM, Strand DW. 2020. Urethral luminal epithelia are castration-

714 insensitive progenitors of the proximal prostate. *bioRxiv*: 2020.2002.2019.937615.
715 10.1101/2020.02.19.937615,

716 Karthaus WR, Hofree M, Choi D, Linton EL, Turkekul M, Bejnood A, Carver B, Gopalan A,
717 Abida W, Laudone V, Biton M, Chaudhary O, Xu T, Masilionis I, Manova K, Mazutis L,
718 Pe'er D, Regev A, Sawyers CL. 2020. Regenerative potential of prostate luminal cells
719 revealed by single-cell analysis. *Science* **368**: 497-505. 10.1126/science.aay0267, PMID:
720 32355025

721 Karthaus WR, Iaquinta PJ, Drost J, Gracanin A, van Boxtel R, Wongvipat J, Dowling CM, Gao
722 D, Begthel H, Sachs N, Vries RG, Cuppen E, Chen Y, Sawyers CL, Clevers HC. 2014.
723 Identification of multipotent luminal progenitor cells in human prostate organoid cultures.
724 *Cell* **159**: 163-175. 10.1016/j.cell.2014.08.017, PMID: 25201529

725 Kolouri S, Park S, Thorpe M, Slepcev D, Rohde GK. 2017. Optimal Mass Transport: Signal
726 processing and machine-learning applications. *IEEE Signal Process Mag* **34**: 43-59.
727 10.1109/MSP.2017.2695801, PMID: 29962824

728 Kwon OJ, Zhang L, Ittmann MM, Xin L. 2014. Prostatic inflammation enhances basal-to-luminal
729 differentiation and accelerates initiation of prostate cancer with a basal cell origin. *Proc*
730 *Natl Acad Sci USA* **111**: E592-600. 10.1073/pnas.1318157111, PMID: 24367088

731 Kwon OJ, Zhang L, Xin L. 2016. Stem Cell Antigen-1 identifies a distinct androgen-independent
732 murine prostatic luminal cell lineage with bipotent potential. *Stem Cells* **34**: 191-202.
733 10.1002/stem.2217, PMID: 26418304

734 Kwon OJ, Zhang Y, Li Y, Wei X, Zhang L, Chen R, Creighton CJ, Xin L. 2019. Functional
735 Heterogeneity of Mouse Prostate Stromal Cells Revealed by Single-Cell RNA-Seq.
736 *iScience* **13**: 328-338. 10.1016/j.isci.2019.02.032, PMID: 30878879

737 Lawson DA, Xin L, Lukacs RU, Cheng D, Witte ON. 2007. Isolation and functional
738 characterization of murine prostate stem cells. *Proc Natl Acad Sci USA* **104**: 181-186.
739 10.1073/pnas.0609684104, PMID: 17185413

740 Lin Y, Liu G, Zhang Y, Hu YP, Yu K, Lin C, McKeehan K, Xuan JW, Ornitz DM, Shen MM,
741 Greenberg N, McKeehan WL, Wang F. 2007. Fibroblast growth factor receptor 2 tyrosine
742 kinase is required for prostatic morphogenesis and the acquisition of strict androgen
743 dependency for adult tissue homeostasis. *Development* **134**: 723-734. dev.02765, PMID:
744 17215304

745 Liu X, Grogan TR, Hieronymus H, Hashimoto T, Mottahedeh J, Cheng D, Zhang L, Huang K,
746 Stoyanova T, Park JW, Shkhyan RO, Nowroozizadeh B, Rettig MB, Sawyers CL, Elashoff
747 D, Horvath S, Huang J, Witte ON, Goldstein AS. 2016. Low CD38 identifies progenitor-
748 like inflammation-associated luminal cells that can initiate human prostate cancer and
749 predict poor outcome. *Cell Rep* **17**: 2596-2606. 10.1016/j.celrep.2016.11.010, PMID:
750 27926864

751 Lu TL, Huang YF, You LR, Chao NC, Su FY, Chang JL, Chen CM. 2013. Conditionally ablated
752 Pten in prostate basal cells promotes basal-to-luminal differentiation and causes invasive
753 prostate cancer in mice. *Am J Pathol* **182**: 975-991. 10.1016/j.ajpath.2012.11.025, PMID:
754 23313138

755 Madisen L, Zwingman TA, Sunkin SM, Oh SW, Zariwala HA, Gu H, Ng LL, Palmiter RD,
756 Hawrylycz MJ, Jones AR, Lein ES, Zeng H. 2010. A robust and high-throughput Cre
757 reporting and characterization system for the whole mouse brain. *Nat Neurosci* **13**: 133-
758 140. 10.1038/nn.2467, PMID: 20023653

759 Moad M, Hannezo E, Buczacki SJ, Wilson L, El-Sherif A, Sims D, Pickard R, Wright NA,
760 Williamson SC, Turnbull DM, Taylor RW, Greaves L, Robson CN, Simons BD, Heer R.
761 2017. Multipotent basal stem cells, maintained in localized proximal niches, support

762 directed long-ranging epithelial flows in human prostates. *Cell Rep* **20**: 1609-1622.
763 10.1016/j.celrep.2017.07.061, PMID: 28813673

764 Montoro DT, Haber AL, Biton M, Vinarsky V, Lin B, Birket SE, Yuan F, Chen S, Leung HM,
765 Villoria J, Rogel N, Burgin G, Tsankov AM, Waghray A, Slyper M, Waldman J, Nguyen
766 L, Dionne D, Rozenblatt-Rosen O, Tata PR, Mou H, Shivaraju M, Bihler H, Mense M,
767 Tearney GJ, Rowe SM, Engelhardt JF, Regev A, Rajagopal J. 2018. A revised airway
768 epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**: 319-324.
769 10.1038/s41586-018-0393-7, PMID: 30069044

770 Morris SA. 2019. The evolving concept of cell identity in the single cell era. *Development* **146**.
771 10.1242/dev.169748, PMID: 31249002

772 Rueden CT, Schindelin J, Hiner MC, DeZonia BE, Walter AE, Arena ET, Eliceiri KW. 2017.
773 ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* **18**:
774 529. 10.1186/s12859-017-1934-z, PMID: 29187165

775 Schaefer BC, Schaefer ML, Kappler JW, Marrack P, Kedl RM. 2001. Observation of antigen-
776 dependent CD8+ T-cell/ dendritic cell interactions in vivo. *Cell Immunol* **214**: 110-122.
777 10.1006/cimm.2001.1895, PMID: 12088410

778 Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden
779 C, Saalfeld S, Schmid B, Tinevez JY, White DJ, Hartenstein V, Eliceiri K, Tomancak P,
780 Cardona A. 2012. Fiji: an open-source platform for biological-image analysis. *Nat Methods*
781 **9**: 676-682. 10.1038/nmeth.2019, PMID: 22743772

782 Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**: 592-593.
783 10.1093/bioinformatics/btq706, PMID: 21169378

784 Schmitzer B. 2016. Stabilized sparse scaling algorithms for entropy regularized transport
785 problems. *arXiv*. 1610:06519,

786 Shappell SB, Thomas GV, Roberts RL, Herbert R, Ittmann MM, Rubin MA, Humphrey PA,
787 Sundberg JP, Rozengurt N, Barrios R, Ward JM, Cardiff RD. 2004. Prostate pathology of
788 genetically engineered mice: definitions and classification. The consensus report from the
789 Bar Harbor meeting of the Mouse Models of Human Cancer Consortium Prostate
790 Pathology Committee. *Cancer Res* **64**: 2270-2305. PMID: 15026373

791 Shen MM, Abate-Shen C. 2010. Molecular genetics of prostate cancer: new prospects for old
792 challenges. *Genes Dev* **24**: 1967-2000. 10.1101/gad.1965810, PMID: 20844012

793 Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M,
794 Smibert P, Satija R. 2019. Comprehensive integration of single-cell data. *Cell* **177**: 1888-
795 1902. 10.1016/j.cell.2019.05.031, PMID: 31178118

796 Thomsen MK, Butler CM, Shen MM, Swain A. 2008. Sox9 is required for prostate development.
797 *Dev Biol* **316**: 302-311. 10.1016/j.ydbio.2008.01.030, PMID: 18325490

798 Toivanen R, Mohan A, Shen MM. 2016. Basal progenitors contribute to repair of the prostate
799 epithelium following induced luminal anoikis. *Stem Cell Reports* **6**: 660-667.
800 10.1016/j.stemcr.2016.03.007, PMID: 27117783

801 Toivanen R, Shen MM. 2017. Prostate organogenesis: tissue induction, hormonal regulation and
802 cell type specification. *Development* **144**: 1382-1398. 10.1242/dev.148270, PMID:
803 28400434

804 Tsujimura A, Koikawa Y, Salm S, Takao T, Coetzee S, Moscatelli D, Shapiro E, Lepor H, Sun
805 TT, Wilson EL. 2002. Proximal location of mouse prostate epithelial stem cells: a model
806 of prostatic homeostasis. *J Cell Biol* **157**: 1257-1265. PMID: 12082083

807 Villani C. 2003. Topics in optimal transportation. *American Mathematical Soc* **58**.

808 Wang X, Xu H, Cheng C, Ji Z, Zhao H, Sheng Y, Li X, Wang J, Shu Y, He Y, Fan L, Dong B,
809 Xue W, Wai Chua C, Wu D, Gao WQ, He Zhu H. 2020. Identification of a Zeb1 expressing

basal stem cell subpopulation in the prostate. *Nat Commun* **11**: 706. 10.1038/s41467-020-14296-y, PMID: 32024836

Wang ZA, Mitrofanova A, Bergren SK, Abate-Shen C, Cardiff RD, Califano A, Shen MM. 2013. Lineage analysis of basal epithelial cells reveals their unexpected plasticity and supports a cell-of-origin model for prostate cancer heterogeneity. *Nat Cell Biol* **15**: 274-283. 10.1038/ncb2697, PMID: 23434823

Wang ZA, Toivanen R, Bergren SK, Chambon P, Shen MM. 2014. Luminal cells are favored as the cell of origin for prostate cancer. *Cell Rep* **8**: 1339-1346. 10.1016/j.celrep.2014.08.002, PMID: 25176651

Wei X, Zhang L, Zhou Z, Kwon OJ, Zhang Y, Nguyen H, Dumpit R, True L, Nelson P, Dong B, Xue W, Birchmeier W, Taketo MM, Xu F, Creighton CJ, Ittmann MM, Xin L. 2019. Spatially Restricted Stromal Wnt Signaling Restrains Prostate Epithelial Progenitor Growth through Direct and Indirect Mechanisms. *Cell Stem Cell* **24**: 753-768 e756. 10.1016/j.stem.2019.03.010, PMID: 30982770

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**: 15. 10.1186/s13059-017-1382-0, PMID: 29409532

Xin L. 2019. Cells of origin for prostate cancer. *Adv Exp Med Biol* **1210**: 67-86. 10.1007/978-3-030-32656-2_4, PMID: 31900905

Xin L, Ide H, Kim Y, Dubey P, Witte ON. 2003. In vivo regeneration of murine prostate from dissociated cell populations of postnatal epithelia and urogenital sinus mesenchyme. *Proc Natl Acad Sci U S A* **100 Suppl 1**: 11896-11903. PMID: 12909713

Zhang D, Jeter C, Gong S, Tracz A, Lu Y, Shen J, Tang DG. 2018. Histone 2B-GFP label-retaining prostate luminal cells possess progenitor cell properties and are intrinsically resistant to castration. *Stem Cell Reports* **10**: 228-242. 10.1016/j.stemcr.2017.11.016, PMID: 29276153

835 Zhang Y, Zhang J, Lin Y, Lan Y, Lin C, Xuan JW, Shen MM, McKeehan WL, Greenberg NM,
836 Wang F. 2008. Role of epithelial cell fibroblast growth factor receptor substrate 2alpha in
837 prostate development, regeneration and tumorigenesis. *Development* **135**: 775-784.
838 dev.009910 10.1242/dev.009910, PMID: 18184727

839 Zhao SG, Chang SL, Erho N, Yu M, Lehrer J, Alshalalfa M, Speers C, Cooperberg MR, Kim W,
840 Ryan CJ, Den RB, Freedland SJ, Posadas E, Sandler H, Klein EA, Black P, Seiler R,
841 Tomlins SA, Chinnaiyan AM, Jenkins RB, Davicioni E, Ross AE, Schaeffer EM, Nguyen
842 PL, Carroll PR, Karnes RJ, Spratt DE, Feng FY. 2017. Associations of luminal and basal
843 subtyping of prostate cancer with prognosis and response to androgen deprivation therapy.
844 *JAMA Oncol* **3**: 1663-1672. 10.1001/jamaoncol.2017.0751, PMID: 28494073

845

Figure Legends

Figure 1. Single-cell analysis identifies prostate luminal epithelial heterogeneity. (A) *t*-distributed stochastic neighbor embedding (tSNE) plot of 5,288 cells from an aggregated dataset of two normal mouse prostates, processed by *Randomly* and clustered using the Leiden algorithm. **(B)** tSNE representation of each prostate lobe (AP: 2,735 cells; DP: 1,781 cells; LP: 2,044 cells; VP: 1,581 cells). **(C)** Schematic model of prostate lobes with the urethral rhabosphincter partially removed, with the distribution of luminal epithelial populations indicated. **(D)** Dot plot of gene expression levels in each epithelial population for selected marker genes. **(E)** Ridge plots of marker genes showing expression in each population. **(F)** Hematoxylin-eosin (H&E) and immunofluorescence images of selected markers in serial sections; the periurethral/proximal region shown is from the AP and DP. Arrow in VP distal indicates distal cell with *Ppp1r1b* expression. Scale bars indicate 50 microns.

Figure 1—figure supplement 1. Anatomy and dissection of mouse prostate lobes. (A) Schematic of connections of prostate lobes to the urethra. Note that the AP, DP, and LP connect dorsally in close proximity, whereas the VP connects on the ventral side. **(B)** Whole-mount views of prostate lobe connections in *Ubc-GFP* mice. **(C)** H&E staining of transverse section through intact urogenital apparatus. The LP crosses the rhabdosphincter caudally (right), and the periurethral (PrU) region lies within the rhabdosphincter. **(D,E)** Bright-field and epifluorescence views of dissected prostate lobes from *Ubc-GFP* mouse. Proximal regions are oriented downwards; note that the LP is the smallest lobe and has a relatively long unbranched region. **(F)** H&E staining of sections from the indicated lobes. Scale bars in **(B-E)** indicate 2 mm, in **(F)** indicate 50 microns.

Figure 1—figure supplement 2. Random-matrix analysis of single-cell datasets. Comparison of dimensional reduction, clustering and visualization of 2,322 sequenced cells from the mouse anterior lobe, based on traditional PCA **(A-D),** and the *Randomly* algorithm **(E-J). (A)** "Elbow plot" describing the variance ratio of each principal component (PC) after a PCA reduction of the $\log_2(1+TPM)$ transformed count matrix. **(B)** Mean silhouette scores of the clusters obtained for different values of the Leiden resolution after performing a clustering in the first 11 PCs using the Leiden algorithm (as implemented in (Wolf et al., 2018)). **(C,D)** tSNE visualizations of the 11 PCs and clustering for Leiden resolutions of 0.2 (7 clusters) and 0.3 (8 clusters) respectively. The LumA* cells have similar expression profiles as the LumA population, but display altered expression of ribosomal and mitochondrial genes, which is consistent with cellular stress. **(E)** Normality test to detect eigenvector localization in the 2,322 cell-eigenvectors of the *z*-scored $\log_2(1+TPM)$ transformed count matrix. The red line corresponds to the sparse data before *Randomly* and the green line shows eigenvector behavior after elimination of the sparsity-induced signal. **(F)** Spectral distribution of the Wishart matrix for selected cells after elimination of the sparsity-induced signal (blue histogram) with a Marchenko-Pastur (MP) distribution fit (red line). Only 50 eigenvalues (~2% of the total) lie outside the MP distribution, and their corresponding eigenvectors carry true signal. The remaining ~98% of the data is comparable to a random matrix and is therefore noise. **(G)** Chi-squared test for the variance of the genes' projection into different sets of eigenvectors. Blue, the 50 signal-like eigenvectors; pink, the eigenvectors corresponding to the last 50 MP eigenvalues; green, the eigenvectors corresponding to the first 50 MP eigenvalues; brown, projection on 50 2,322-dimensional random vectors. **(H)** Selection of the genes that are mostly responsible for the signal in this dataset (purple line). The number of genes (orange line) is calculated with a false discovery rate (FDR) using the ratio of the blue and pink distributions in **(G)** Approximately 5000 genes are responsible for the signal using FDR $\leq$ 0.001. **(I)** Mean silhouette score of the clusters obtained for different values of the Leiden resolution after processing with *Randomly*. In comparison with **(B),** the score is much higher, indicating better clustering. **(J)** tSNE visualization of the latent space generated by *Randomly* and clustering

897 performed with the Leiden algorithm. For downstream analyses, we have combined the LumA*
898 cells with the LumA population. The number of clusters selected corresponds to the maximum of
899 the curve in **(I).** *Randomly* can assist in identification of populations (*e.g.,* PrU) by removing noise
900 and sparsity-induced signals, and by selecting genes responsible for the biological signal.

901

902 **Figure 1—figure supplement 3. Additional marker validation for epithelial populations.**
903 *(Above)* Ridge plots of marker genes show expression in each population. *(Below)*
904 Immunofluorescence staining of marker expression in sections; the periurethral/proximal region
905 shown is from the AP and DP. Scale bars indicate 50 microns.

906

907 **Figure 2. Luminal epithelial populations display spatial and morphological heterogeneity.**
908 **(A)** H&E and IF of serial sections from the DP and LP, showing expression of proximal (*Ppp1r1b*)
909 and distal (*Tgm4, Msmb*) markers; note apparent differences in the boundary regions of the two
910 lobes. **(B)** Detection of distally localized LumP cells (arrows) in all four lobes; these are most
911 abundant in the VP. **(C)** Scanning electron micrographs of the boundary region of the AP; central
912 low-power image is flanked by high-power images of boxed regions. Red arrow, mitochondria;
913 black arrow, membrane interdigitation; blue arrow, Golgi apparatus; green arrow, rough
914 endoplasmic reticulum. **(D)** Identification of the periurethral region. Cells in the periurethral region
915 express Ly6d, Ck7, Aqp3, and Ppp1r1b; notably, Cldn10-expressing LumP cells decrease
916 approaching the periurethral region **(E)** Lineage-marking in *Nkx3.1*$^{Cre/+}$*; R26R-YFP* mice (n=3)
917 shows widespread YFP expression in the periurethral, proximal, and distal AP; small patches
918 remain unrecombined and lack YFP (arrows). Scale bars in **(A,B,D,E)** indicate 50 microns; scale
919 bars in **(C)** indicate 2 microns.

920

Figure 2—figure supplement 1. Additional analysis of proximal-distal heterogeneity. (A) H&E and IF of serial sections from the AP and VP, showing expression of proximal (*Ppp1r1b*) and distal (*Tgm4, Trpv6*) markers. **(B)** Scanning electron micrographs of proximal and distal regions of the AP. *Left:* red arrows, mitochondria; black arrow, membrane interdigitation; *Right:* blue arrows, Golgi apparatus; green arrows, rough endoplasmic reticulum; yellow arrows, secretory vesicles near the apical surface. Scale bars in **(A)** indicate 50 microns; scale bars in **(B)** indicate 2 microns.

Figure 3. Functional analysis of epithelial populations in organoid and tissue reconstitution assays. (A) Heatmap visualization of the Wasserstein distance between epithelial populations with hierarchical clustering. **(B)** Tree visualization of Wasserstein distances. **(C)** Flow sorting of distinct epithelial populations from the AP lobe. **(D)** Organoid formation assays using sorted epithelial cells in two distinct culture conditions. ENR conditions: sorted cells from Ubc-GFP mice plated at 1,000 cells/well, imaged at day 10. Hepatocyte Media (HM) conditions: sorted cells from wild type C57BL/6 mice plated at 2,000-5,000 cells/well and imaged on day 12-13. Maximum p-values for each pair-wise comparison are indicated. **(E)** Organoid formation efficiency plots. **(F)** Grafting efficiency in tissue reconstitution assays. LumP, PrU, and Basal are significantly more efficient at generating grafts from smaller number of cells relative to distal luminal populations (average p-value shown). **(G)** H&E and IF of sections from fully-differentiated renal grafts; positive staining corresponds to results found in ≥3 independent grafts. Scale bars indicate 50 microns.

Figure 3—figure supplement 1. Additional marker analysis of renal grafts. (A) Immunofluorescence staining of grafts using the indicated markers; arrows indicate regions of patchy expression. The left-most two columns show fully-differentiated graft regions, whereas the right-most column shows less-differentiated regions that are relatively small, have less abundant

947 basal cells, are typically found on the periphery of the graft, and have fewer secretions. These less-
948 differentiated regions tend to express the LumP marker Ppp1r1b. **(B)** Immunofluorescence
949 detection of GFP in grafts from *Ubc-GFP* mice demonstrates donor origin of grafted epithelial
950 cells.

951

952 **Figure 4. Heterogeneity and conservation of luminal populations in the human prostate. (A-**
953 **C)** tSNE plot of scRNA-seq data (**A,** 1,600 cells; **B,** 2,303 cells; **C,** 2825 cells) from three
954 independent human prostatectomy samples. **(D-F)** Heatmap visualization of Wasserstein distances
955 between the human and mouse prostate populations for each dataset. **(G)** H&E and IF images of
956 serial sections from human prostate, showing regions of the prostatic utricle, central, transition,
957 and peripheral zones. Arrows indicate regions of ductal morphology. Scale bars indicate 50
958 microns.

959

960 **Figure 4—source data. Human prostate samples and corresponding clinical data.**

961

Figure 1

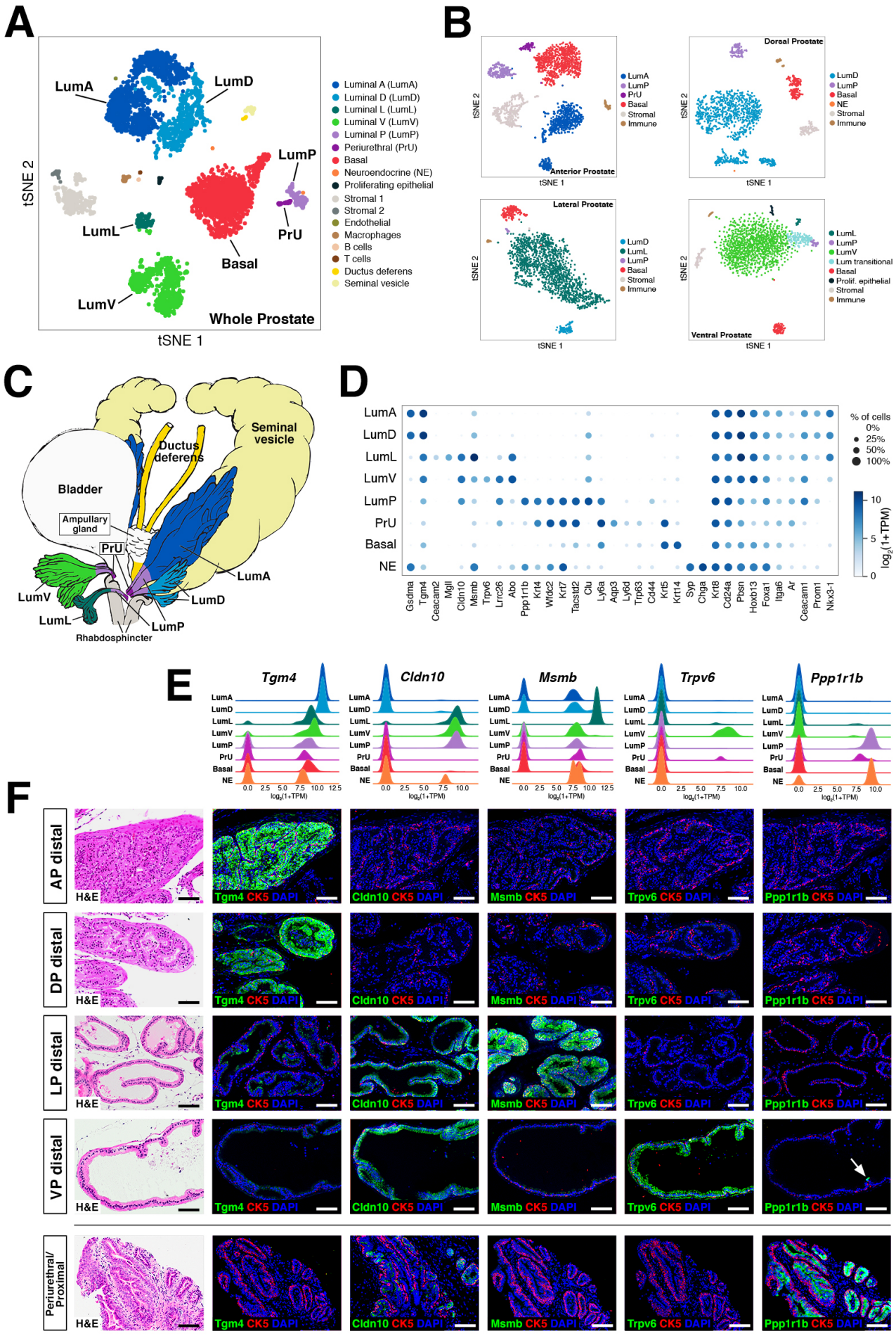

Figure 1—figure supplement 1

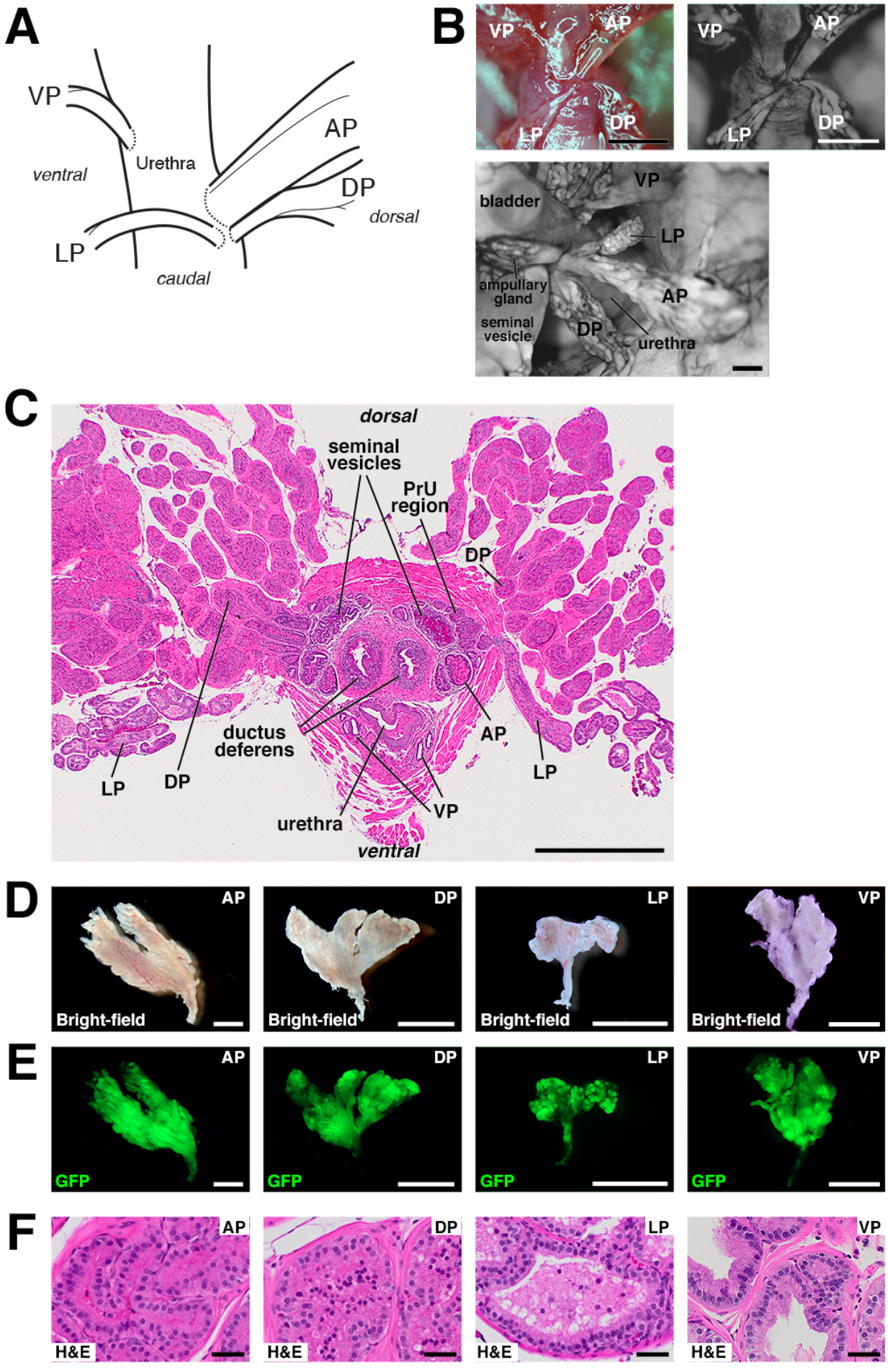

Figure 1—figure supplement 2

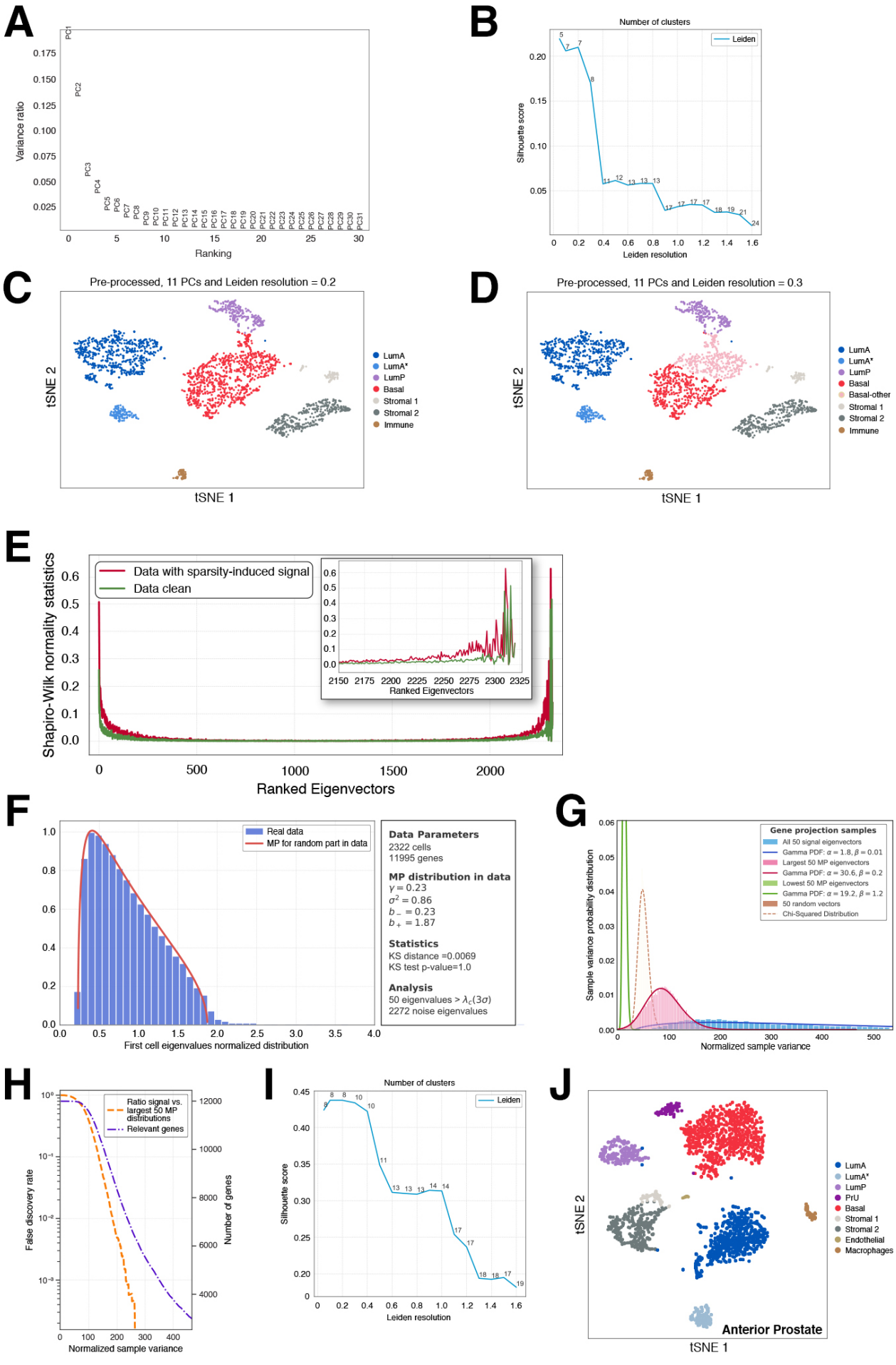

Figure 1—figure supplement 3

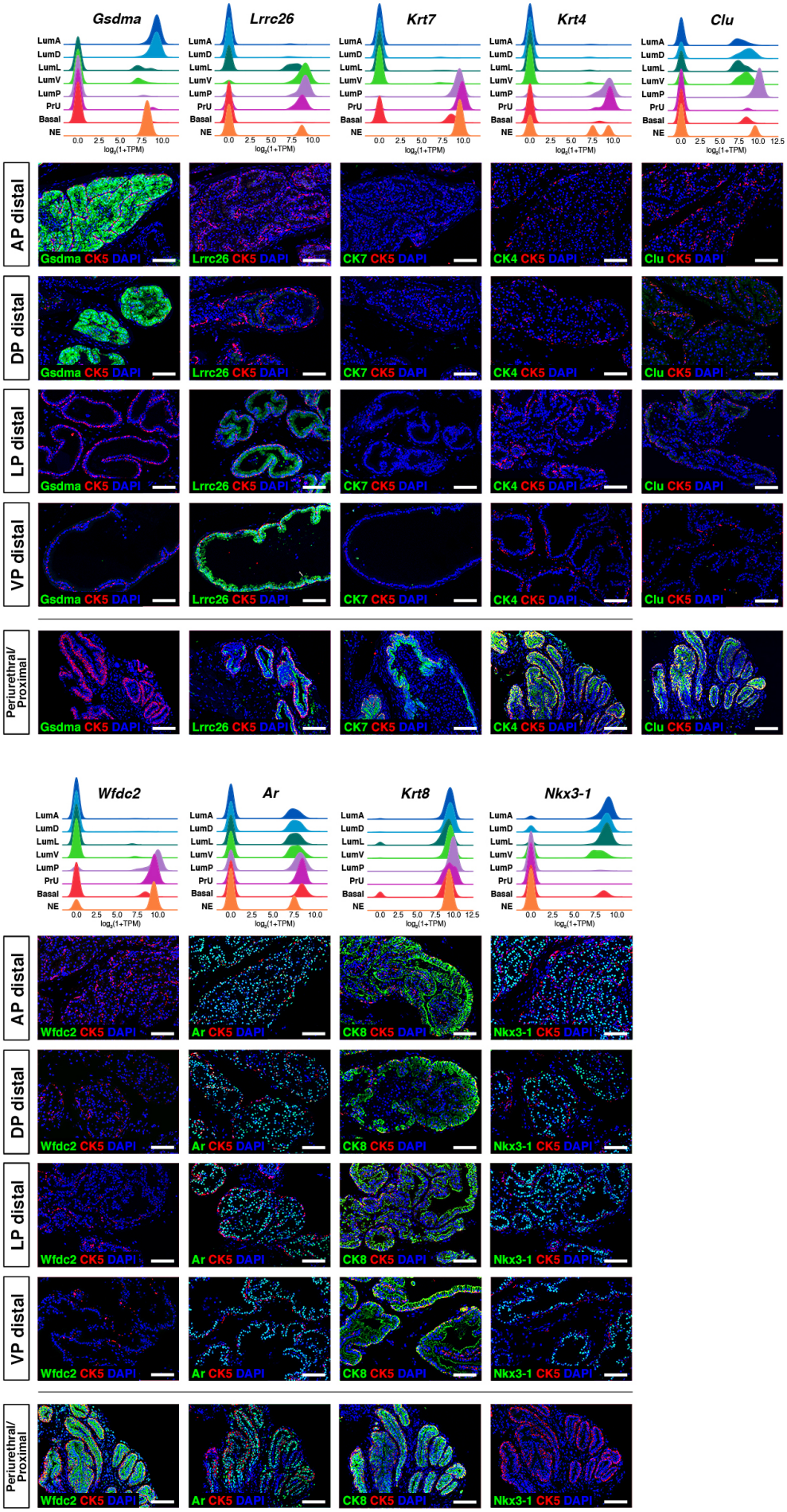

Figure 2

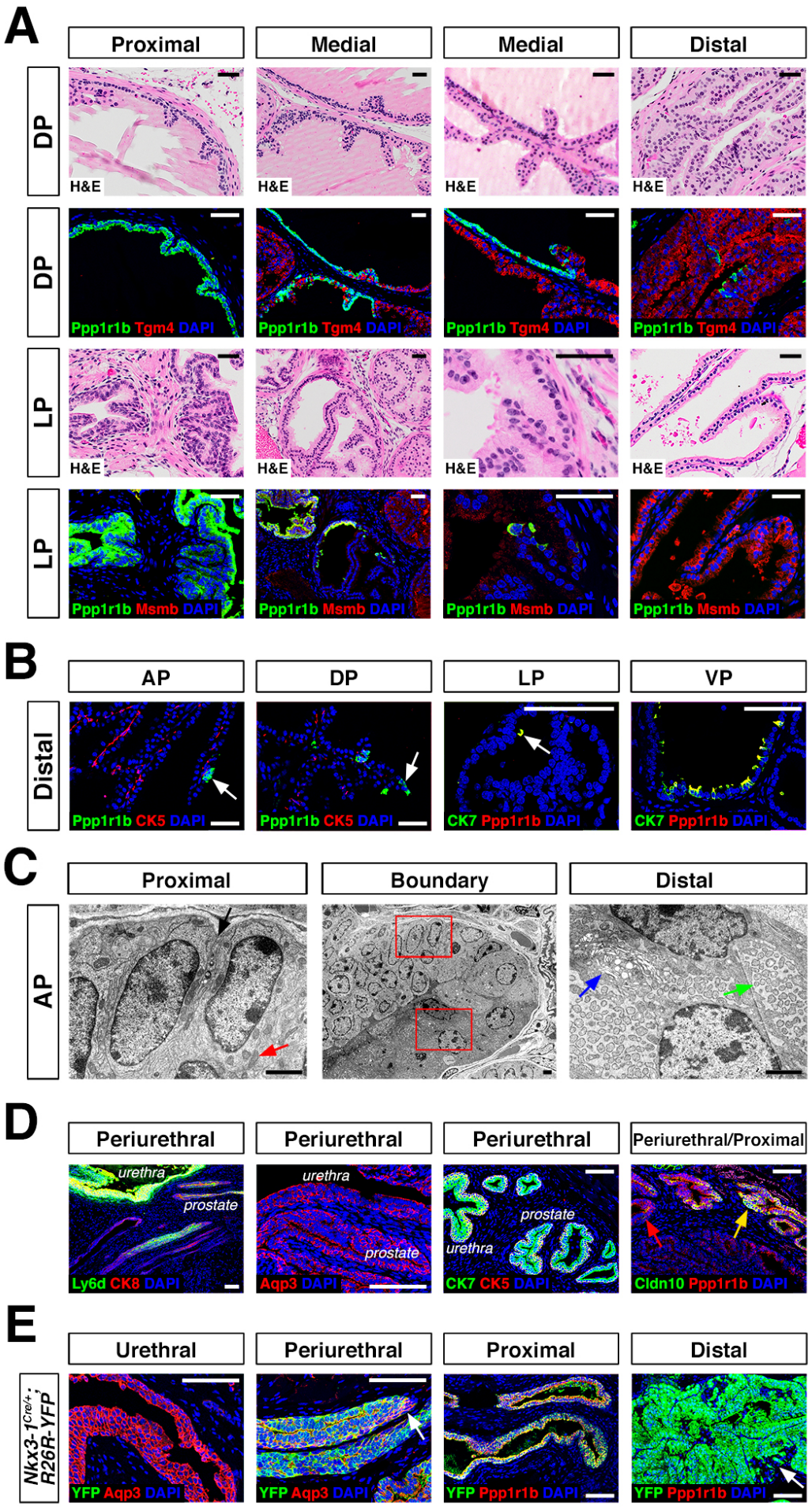

Figure 2—figure supplement 1

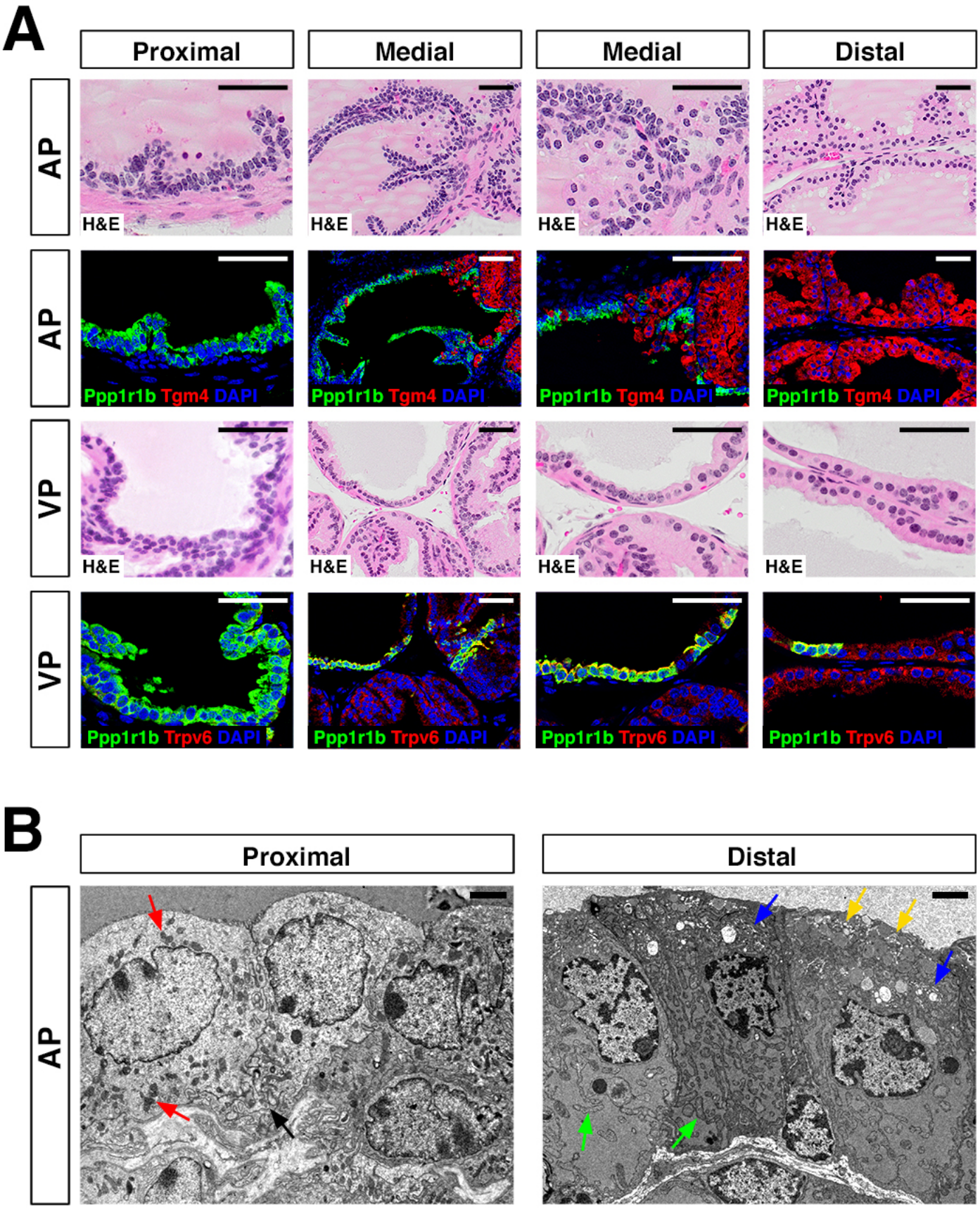

Figure 3

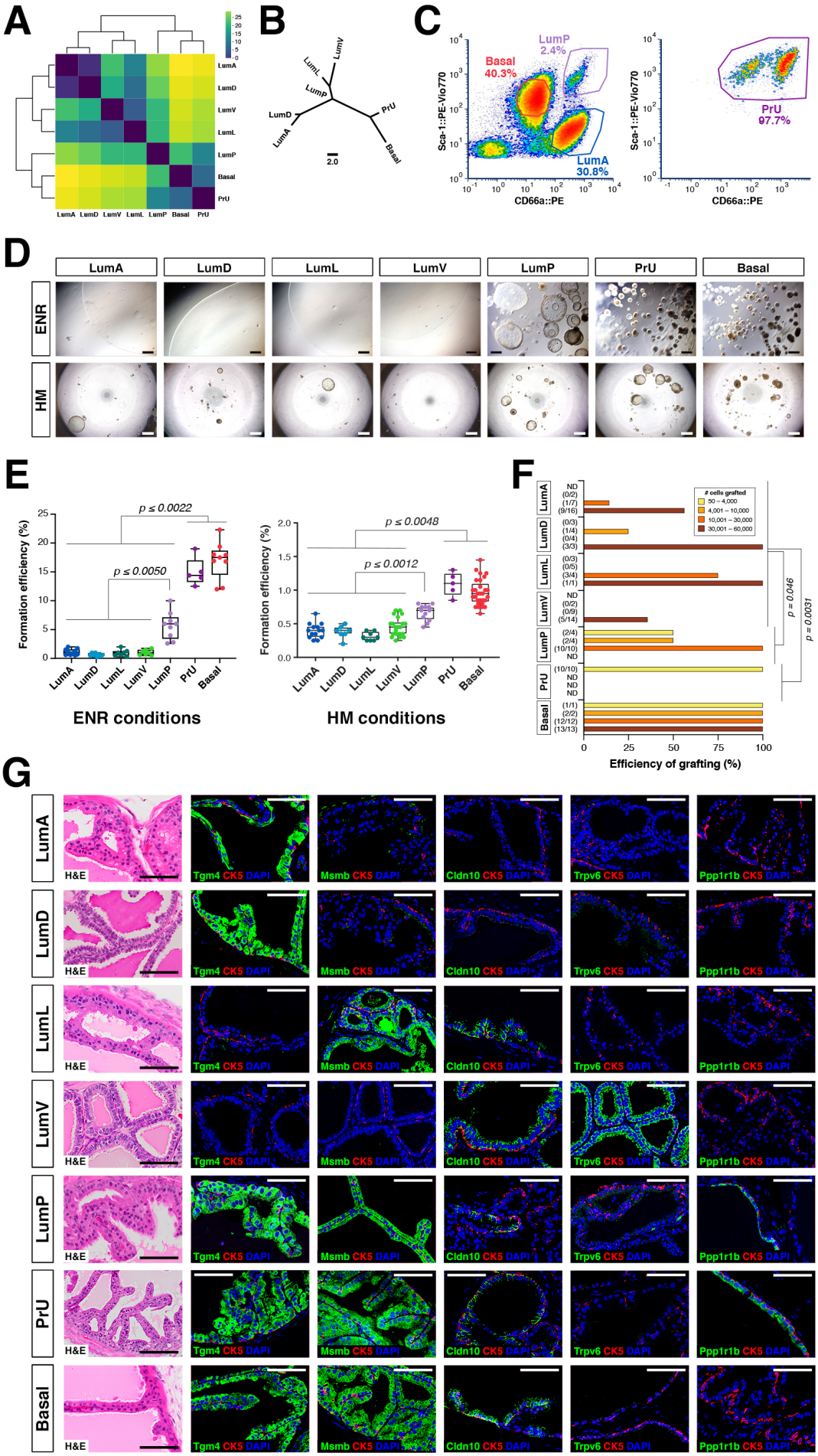

Figure 3—figure supplement 1

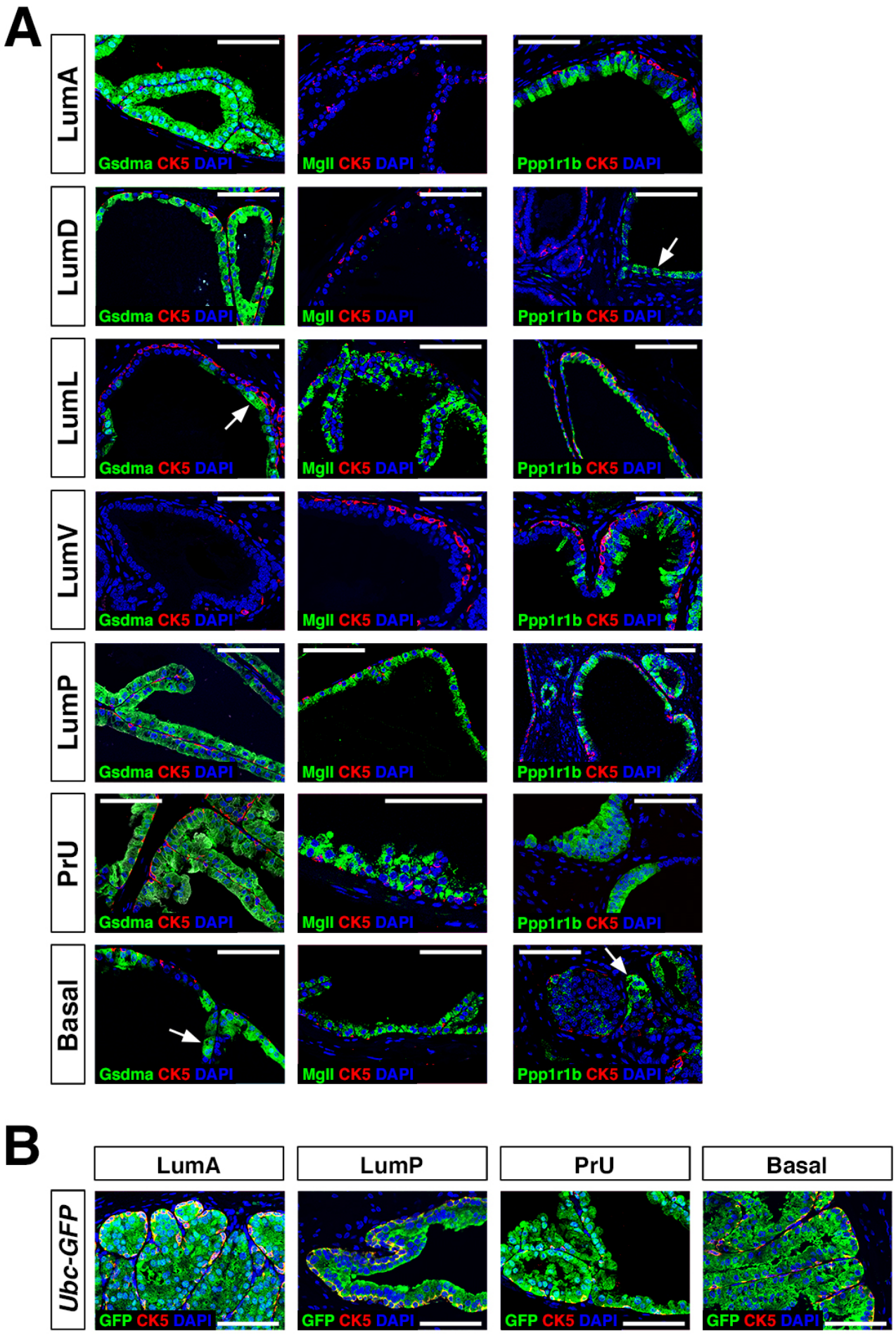

Figure 4

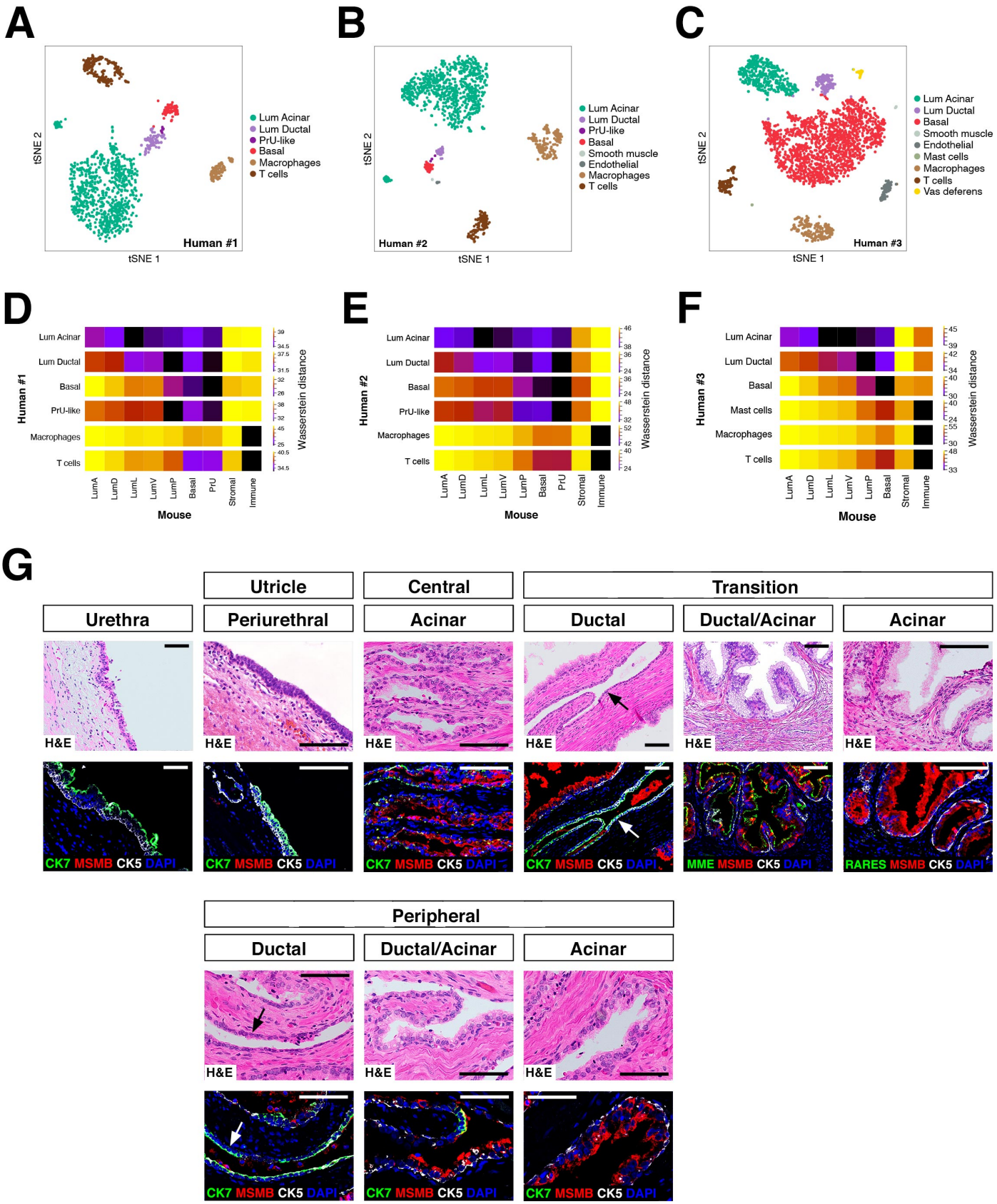

Figure 4—source data. Human prostate samples and corresponding clinical data

Patient	Site	Age	Procedure	Overall diagnosis	Pathology of sample analyzed	Comments
1	Columbia	70	Cystoprostatectomy	High grade urothelial carcinoma of bladder	Benign prostatic hyperplasia with granulomatous prostatitis	scRNA-seq dataset #1 (mm037)
2	Columbia	68	Cystoprostatectomy	High grade urothelial carcinoma of bladder	Benign prostatic hyperplasia with chronic inflammation	scRNA-seq dataset #2 (mm033)
3	Columbia	63	Radical prostatectomy	Prostate adenocarcinoma (Gleason 3+3=6, pT2 N0)	Benign prostate with inflammation	scRNA-seq dataset #3 (mf002)
4	Cornell	54	Radical prostatectomy	Prostatic adenocarcinoma (Gleason 4+3=7, pT2 N0)	Benign prostate	
5	Cornell	65	Radical prostatectomy	Prostatic adenocarcinoma (Gleason score 3+4=7, pT3a N1)	Benign prostate	
6	Cornell	79	Cystoprostatectomy	High grade urothelial carcinoma of bladder	Benign prostate	