Multi-omics analyses identify transcription factor interplay in corneal epithelial fate determination and disease

Jos GA Smits[1], Dulce Lima Cunha[1], Jieqiong Qu[1,2], Nicholas Owen [3], Lorenz Latta[4,5], Nora Szentmary[4], Berthold Seitz[5], Lauriane N Roux[6], Mariya Moosajee [3,7], Daniel Aberdam[6,8], Simon J. van Heeringen[1], Huiqing Zhou[1,9]**

*Corresponding author

[1] Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University, Nijmegen, The Netherlands,

[2] Department of Medical Microbiology, Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, Nijmegen, the Netherlands

[3] Development, Ageing and Disease, UCL Institute of Ophthalmology, London, United Kingdom

[4] Dr. Rolf M. Schwiete Center for Limbal Stem Cell and Aniridia Research, Saarland University, Homburg/Saar, Germany

[5] Department of Ophthalmology, Saarland University Medical Center, UKS, Homburg/Saar, Germany

[6] INSERM U976, Paris, France

[7] Department of Genetics, Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom

[8] Université de Paris, INSERM U1138, Centre des Cordeliers, Paris, France

[9] Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands

Running title: Cell fate regulation of limbal stem cells and epidermal keratinocytes

Keywords: Cell identity, Transcription factors, Gene Regulating Network, Limbal Stem Cells, Epidermal Keratinocytes, multi-omics

Abstract

The homeostasis of the transparent corneal epithelium in the eye is maintained by limbal stem cells with proper cell fates. A potential disease mechanism underlying corneal opacity has been proposed to be limbal stem cells acquiring characteristics of keratinocytes of the non-transparent epidermis. The precise cell fate differences between these two epithelial cells are however unknown. We performed a multi-omics analysis of human limbal stem cells derived from the cornea and keratinocytes from the epidermis, and characterized their similar yet distinct molecular signatures. With gene regulatory network analyses, we identified cell fate defining transcription factors and their regulatory hierarchy that are shared but also distinct for specific epithelial programs. Our findings indicate that shared transcription factors such as p63, FOXC1 and FOSL2 often regulate limbal stem cell-specific transcription factors such as PAX6, SMAD3 and OTX1. Single-cell RNA-seq analysis confirms the shared and specific transcription factors controlling the stem cell fates of the cornea and the epidermis. Importantly, genes associated with corneal opacity can cooperatively be targeted by the shared and limbal stem cell-specific transcription factors. Finally, by leveraging these key transcription factors, we identified *FOSL2* as a novel candidate associated with corneal opacity. By characterizing molecular signatures, our study uncovers the distinct regulatory circuitry controlling limbal stem cell fates and corneal opacity.

Introduction

Cell fate determination is a complex process essential for normal development and homeostasis. The key role of transcription factors (TFs) in cell fate determination has been demonstrated by a plethora of seminal studies where cell conversions can be achieved by forced expression of specific sets of TFs, e.g., generation of induced pluripotent stem cells[1,2]. TFs control cell fate determination by regulating the transcriptional program, through binding to cis-regulatory elements (CREs) of the targets on DNA and by modifying chromatin environments[3,4]. The cell identity changes are therefore concomitant with rewiring the chromatin and epigenetic landscape of the cells, highlighting the importance of chromatin modulation in cell fate control[3,4]. In somatic tissues, the precise control of the corresponding somatic stem cell fates is essential for tissue integrity and tissue-specific function, which is often deregulated in pathological conditions.

The epithelium of the cornea in the eye and the epidermis of the skin are two types of stratified epithelia with multiple layers of epithelial cells (Figure 1A), both derived from the surface ectoderm during embryonic development. The human corneal epithelium is the outermost layer of the cornea, supported by underlying stroma and endothelium, protecting the eye from the outside environment[5–7]. It is avascular and transparent, which allows the light into the eye. The proper structure and function of the corneal epithelium are maintained by stem cells in the limbus, limbal stem cells (LSCs), which are located at the rim of the cornea. Differentiating LSCs move centrally to form basal epithelial cells, and to stratify to differentiated epithelial layers[8]. Similar to the corneal epithelium in the barrier function, the epidermis of the skin, on the other hand, is non-transparent. The homeostasis of the epidermis is controlled by keratinocytes (KCs) in the basal layer, residing on the basal membrane between the epidermis and the dermis. Basal KCs differentiate vertically and migrate upwards to form different strata of the epidermis[6]. The somatic stem cells, LSCs of the cornea and KCs in the epidermis, are seemingly similar in their cellular morphology, indistinguishable in culture, and share expression of many basal epithelial genes such as *KRT5* and *KRT14*. Nevertheless, cell fates of LSCs and KCs are intrinsically distinct, as they initiate and maintain specific epithelial differentiation programs that give rise to the transparent corneal epithelium and non-transparent epidermis, respectively. Insights into the similarities and differences between cell fates of LSCs and KCs will shed light on the control mechanism of their cellular function and related pathological conditions, e.g., corneal opacity. So far, however, the cell fate similarities and

differences between KCs and LSCs controlled by TFs and their associated epigenetic mechanisms, are not yet understood.

In KCs and the epidermis, key TFs have been studied extensively, both *in vitro* and *in vivo* [9–12]. Key TFs including p63, GRHL family proteins, KLF4 and ZNF750 regulate transcriptional programs important for KC proliferation and differentiation[10,12–14]. Many of these TFs, sometimes cooperating with each other, are also known to modulate the chromatin landscape through enhancers[10,12,14]. The TF p63 encoded by *TP63* is a key regulator of stratified epithelia and is important for commitment, proliferation and differentiation of KCs[12]. It binds mainly to enhancers and maintains the epigenetic landscape for the proper epidermal cell identity[14–16]. Mutations in *TP63* are associated with developmental disorders like ectrodactyly, ectodermal dysplasia, and cleft lip/palate (EEC) syndrome (OMIM 604292), where patients present with defects in ectodermal derivatives, e.g., epidermis, hair follicles and nails, but also in other epithelium-lined tissues such as the cornea[17–19]. The disease phenotypes of *TP63* mutation associated disorders are consistent with p63 expression in stratified epithelia[17–19]. It has been shown that loss of the typical epidermal identity due to rewired epigenetic circuitry is characteristic of KCs carrying *TP63* mutations associated with EEC[14].

As compared to the wealth of molecular insights of TFs in KCs, the control mechanism of TFs in the corneal epithelium and LSCs is less understood. One of the better studied TFs is the eye master regulator PAX6. PAX6 is essential for specification and determination of different parts of the eye, including retina, iris, lens and the cornea[20–22]. In retina and lens, PAX6 interacts with chromatin modifier such as EZH2, cooperates with and regulates other TFs to define cell fates[23–25]. In LSCs of the cornea, PAX6 binds to enhancers, together with TFs such as RUNX1 and SMAD3, important for controlling the LSC identity[21,26–30]. Mutations and deregulation of *PAX6* are associated with aniridia (OMIM 106210), a disorder initially characterized by an absent or underdeveloped iris, among other phenotypes such as defects in the retina, pancreas and neurological systems[31], which is consistent with PAX6 expression in these tissues and organs[22]. Relevant to the cornea, up to 90% of aniridia patients show progressive limbal stem cell deficiency (LSCD) leading to corneal opacities[29,32]. Interestingly, is present in over 60% of patients with *TP63* mutation associated EEC syndrome[33,34]. In addition to PAX6 and p63, another TF that has been associated with corneal abnormalities is FOXC1, of which mutations are involved in the spectrum of anterior segment dysgenesis, including Peters anomaly and Axenfeld-Rieger syndrom (OMIM 602482)[35]. FOXC1 is expressed in the epithelium, stromal and

endothelial cells of the cornea, and is shown to be upstream and regulating PAX6[36,37]. Further to genetic causes, LSCD can also be caused by autoimmune and/or inflammatory disorders like Stevens-Johnson syndrome or ocular cicatricial pemphigoid, as well as infections or chemical/thermal burns[38]. Recently, two reports suggested that loss of *PAX6* or *FOXC1* in LSCs gives rise to loss of the LSC identity, and these *PAX6* or *FOXC1* deficient LSCs acquire a KC-like cell signature, indicated by upregulated expression of suprabasal epidermal genes[28,37,39].

These observations postulate an intriguing hypothesis that the cell fate conversion from LSCs to non-transparent KC-like cells may represent the pathomechanisms of LSCD and corneal opacities. However, as a general pathomechanism for LSCD and corneal opacities, this hypothesis is problematic, because many of the identified key TFs defining the LSC fate, such as p63, RUNX1, and SMAD3[14,30], are also expressed in KCs. How TFs like PAX6, p63 and FOXC1 regulate their target genes in LSCs and how their mutations give rise to LSCD and corneal opacities are not yet fully understood. Therefore, a comprehensive characterization and comparison of molecular signatures between LSCs and KCs will not only identify shared and tissue-specific TFs controlling cell fates but also provide insights into the pathomechanisms of LSCD and other corneal opacity disease mechanisms.

In this study, we performed in-depth analyses of the transcriptome and the epigenome of human LSCs and KCs cultured *in vitro*, and characterized differentially expressed genes and regulatory regions between the two cell types. Subsequently, using a gene regulatory network-based method, we identified key TFs and their hierarchy controlling epithelial programs that are shared by KCs and LSCs, and those that are distinct for each cell type. Expression pattens of the key TFs were further validated with *in vivo* single-cell RNA-seq data from the cornea and the epidermis. Importantly, we showed that the key TFs and their target genes that drive the specific LSC epithelial program are associated with corneal diseases, and identified novel disease gene *FOSL2* associated with corneal opacity.

Results

Distinct epithelial gene expression patterns define cell fate differences of skin keratinocytes and cornea limbal stem cells

To characterize gene expression patterns that define the cell fate difference between human cornea limbal stem cells (LSCs) and human skin keratinocytes (KCs) (figure 1A), we used LSCs established from limbal biopsies taken from post-mortem cornea and basal KCs from skin donors. Both cultured cells have the capacity to re-generate stratified epithelial tissues *in vitro*[40,41], and have high p63 expression (supplementary figure 1F), thus exhibiting the progenitor cell state. We performed RNA-seq analyses on. These analyses include several pseudobulk RNA-seq data aggregated from single-cell RNA-seq (scRNA-seq) experiments performed with cultured LSCs. This was because no measurable heterogeneity was detected in these cultured LSCs, except cell cycle differences (supplementary figure 1). To measure gene expression differences between the two cell types more robust, we incorporated our data with publicly available RNA-seq data (supplementary Table 1)[14,30].

Using these combined datasets in the pair-wise comparison, we identified 1251 differential expressed genes between LSCs and KCs. Among them, 793 genes had higher expression in LSCs (LSC-high genes), , while 459 differential genes were more highly expressed in KCs (referred to as KC-high genes). This analysis resulted in typical genes for both epithelial cell types: LSC-high genes contained limbal and corneal epithelial genes including *KRT19*, *KRT12,* and the eye master regulator *PAX6* (figure 1B, supplementary figure 1F), whereas KC-high genes contained epidermal markers such as *KRT1*, *KRT10*, *LCE3D* & *LCE3C*. Although some of these detected genes are associated with epithelial stratification, e.g., *KRT3* and *KRT12* with corneal stratification and *KRT1* and *KRT10* with epidermal stratification, their expression was much lower than their corresponding stratified epithelial cells[14,37](see supplementary figure 1G). This indicates that these cultured cells are limbal and epidermal progenitor cells. It should be noted that *TP63* is highly expressed in both LSCs and KCs (supplementary figure 1F), and therefore is not identified as differential.

Gene Ontology (GO) enrichment analysis[42,43] of KC-high genes identified enrichment of GO terms related to epidermis and skin development (figure 1C). GO terms associated with virus response, e.g., “response to virus” and “defense response to virus”, were also enriched due to detected immune and interferon related genes among KC-high genes.

This observation was consistent with the gene set enrichment analysis (GSEA) using the hallmark gene set of the MsigDB collection[44,45] that identified enrichment of the interferon-alpha and gamma response. Finally, PROGENy pathway target gene analysis[46] identified higher VEGF signaling in KC-high genes, such as *ALDH1A1* and *CPA6*. This is in line with the avascularized state of the cornea, and the vascularization-related genes are completely repressed. GO annotation of LSC-high genes, on the other hand, was enriched for terms such as “positive regulation of locomotion and cell motility” and ”angiogenesis” due to the presence of genes such as *MMP14*, *VEGFC*, *CXCL17,* and *IL1β*. Furthermore, eye and neural-related GO function terms such as “visual system development” and “regulation of nervous system development” were detected in LSC-high genes including *KRT12*, *PAX6*, *WNT7A,* and *ALDH1A3* (figure 1D). PROGENy analysis identified the TNF-α and NFKβ pathways associated with LSC-high genes (figure 1E), such as *CXCL1,3,5,6* and *TNF-α*. Consistently, TNF-α and NF-kB signaling was also identified by KEGG pathway[47], GO and GSEA analyses using the hallmark gene set of the MsigDB collection (supplementary figure 2). Finally, GSEA enrichment using the C8 dataset that contains single cell datasets within the MsigDB collection also identified enrichment for genes in the “Descartes fetal eye corneal and conjunctival epithelial cells. All enrichment results are summarized in supplementary table 4.

Next, we asked the question whether LSCD-associated LSCs acquire a KC-like cell fate. To test this, we examined the cell fate of LSCs from patients with aniridia. We performed RNA-seq analysis of primary LSCs of two aniridia patients and of controls. These data were integrated with other aniridia RNA-seq data published previously[48] (supplementary Table 1). We obtained 73 differentially downregulated genes (aniridia low genes) and 22 upregulated genes (aniridia high genes) in aniridia patient LSCs, as compared to LSCs from healthy controls (supplementary figure 3). Aniridia low genes included the *PAX6* target gene *KRT12* and other corneal and epithelial genes such as *TGFBI*, *CLND1*, *GJB6*, *IL36G*, *LAYN*, *NMU,* and *TMEM47.* Many of these epithelial genes are potential PAX6 target genes reported in an immortalized LSC model where one allele of *PAX6* was deleted[49]. We then applied GSEA to compare them to LSC and KC gene expression signatures, to investigate whether these deregulated genes due to *PAX6* haploinsufficiency represent the changed cell fate of these aniridia LSCs. Indeed, as expected, aniridia low genes were enriched among genes expressed highly in LSCs (P-value 2.8E-05) (figure 1F), indicating a loss of LSC cell fate in aniridia LSCs. Among aniridia high genes, we found GATA3 present in genes expressed highly in KCs

(supplementary figure 3). Nevertheless, there was no detected enrichment of aniridia high genes among genes expressed highly in KCs representing the KC cell fate, arguing against the postulated model thatPAX6-deficienct LSCs acquire a KC-like cell fate at the transcriptome level. These data suggest that additional mechanisms, such as TFs other than *PAX6*, contribute to the cell fate difference between KCs and LSCs.

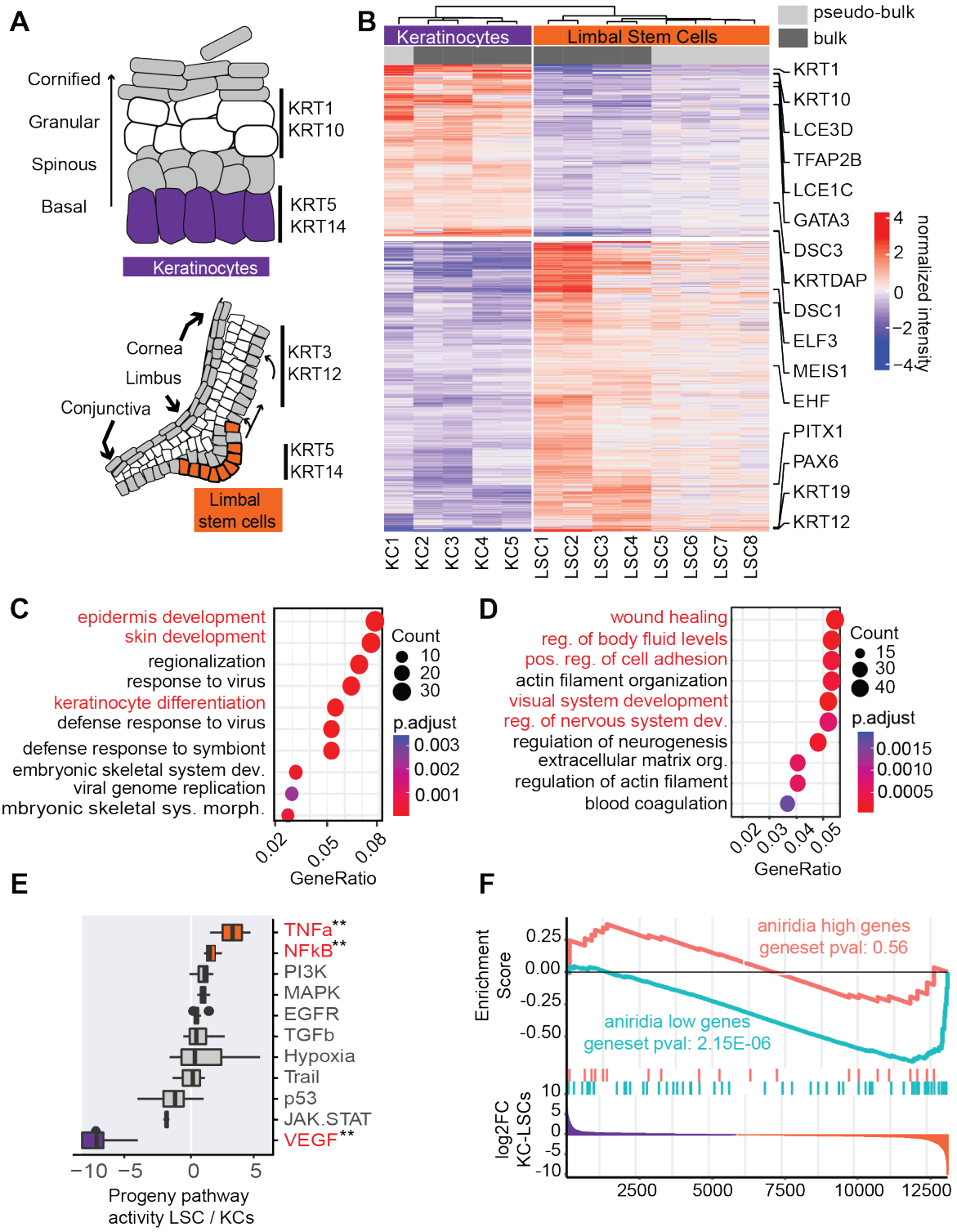


Figure 1 RNA-seq analysis of LSCs and KCs. A) Schematic picture of the epidermis and the limbus. **B)** Heatmap of normalized expression of differentially expressed genes between LSCs and KCs (adjusted pval < 0.01, log2 FC > 1.5). Differentially expressed genes are clustered using k-means clustering with 2 clusters. **C)** GO-term enrichment of KC-high genes. **D)** GO-term enrichment of LSC-high genes. **E)** PROGENy pathway activity analysis, with scores sorted based on LSC/KC ratio. Pathways depicted in red are differential color is grey if non-differential, orange if higher in LSC and purple if higher in KCs. **F)** Gene Set Enrichment Analysis of differentially expressed genes identified in aniridia patient LSCs, as compared to controls, Up- and down-regulated genes (aniridia high and low, respectively) were tested for enrichments against KC-high and LSC-high genes, respectively.

The epigenetic states of cis-regulatory elements correlate with gene expression patterns

To understand the mechanisms underlying different cell fate controls of LSCs and KCs, we identified cis-regulatory elements (CREs) and their epigenetic states that drive gene expression differences. We generated an extensive multi-omics dataset of LSCs and KCs, and integrated these with other published data (supplementary figure 4a, supplementary Table 2, and Table 3). The complete dataset included ATAC-seq for open chromatin regions representing CREs[30] and ChIP-seq of histone modifications, H3K27ac and H3K4me3 marking active CREs, and H3K27me3 that marks repressed chromatin regions (figure 2A)[15,30]. Using ATAC-seq analysis, we identified 124,062 CREs in the two cell types. Approximately 80% of these CREs were accessible in both cell types (supplementary figure 5). To examine differential epigenetic states of these CREs in LSCs and KCs, we quantified ATAC-seq and histone modification signals in windows covering these CREs (figure 2A). This resulted in 35,348 CREs with differential epigenetic signals, about 28.5% of CREs (supplementary figure 4). To assess the correlation between these differential CREs and expression of their nearby genes, we considered both CREs at the promoter regions (promoter CREs) and enhancer CREs located within 50kb-distance from the genes (enhancer CREs) figure 2B, supplementary figure 5B). As expected, high ATAC, H3K4me3, and H3K27ac signals correlated with high gene expression, while the strong signals of repressive H3K27me3 correlated well with lowly expressed genes or genes with undetectable expression in the corresponding cell types, e.g., the loci of *PAX6*, *GATA3*, *HOXA9* and *TNF-α* (figure 2B, 2D, supplementary figure 5D).

Next, we performed GO analysis on genes that are close to differential CREs (figure 2C). For H3K27ac and H3K4me3 that mark active CREs, GO terms such as “epidermis” and “skin development” were identified for CREs with strong signals in KCs, whereas “positive regulation of cell adhesion” and “extracellular matrix organization” terms were found for CREs with strong H3K27ac and H3K4me3 signals in LSCs, consistent with identified differentially expressed genes in each corresponding cell type. In contrast, the repressive mark H3K27me3 anti-correlated with gene expression; GO terms of CREs with high H3K27me3 signals in KCs correlated with those of LSC-high genes, such as “positive regulation of cell adhesion”, and GO terms of CREs with strong H3K27me3 signals in LSCs correlated with those of KC-high genes, such as “epidermal development” (figure 2C). Intriguingly, some of the CREs close to KC stratification genes such as kallikrein gene family (KLK4,5,6,7,8) were already occupied by high levels of H3K4me3 and H3K27ac in KCs, even though these genes were not differentially expressed (see suplementary figure 5E).

Furthermore, in line with the enrichment of *TNF-α* and *NF-kB* signaling pathways in LSCs identified by PROGENy analysis of differentially expressed genes, higher H3K27ac and H3K4me3 signals were present in the loci of *TNF-α* and *NF-kB* target genes in LSCs, as compared to in KCs (supplementary figure 4C), while these loci in KCs were repressed by H3K27me3. The strong repression signals marked by H3K27me3 in KCs at genes that are key for the LSC fate such as PAX6 suggest a repression mechanism in KCs to prevent inappropriate gene expression that defines the LSC fate.

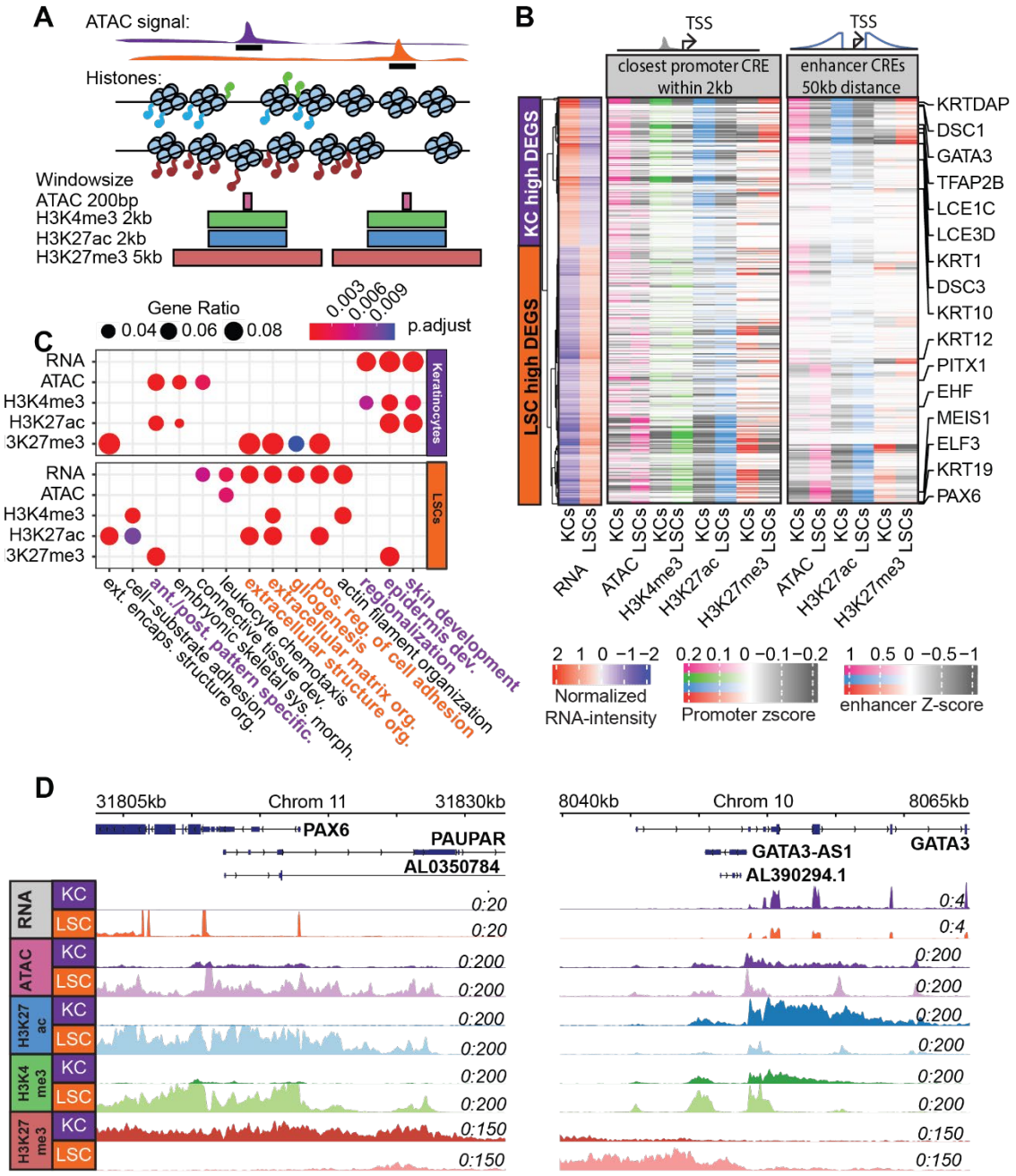


Figure 2 Cis Regulatory Element (CRE) analysis. A) Schematic overview of CRE identification and quantification. Signals of each analysis were quantified by different window sizes covering the ATAC-seq peak submit. **B)** Heatmap of the Z-scores of the quantile normalized ATAC-seq and histone mark signals near LSC- and KC-high genes. For promoter CREs, the closest CRE within 20kb to the transcription start site (TSS) was used. For enhancer CREs, the signals of all CREs within a 100kb window near a TSS were quantified, distance weighted, and finally summed **C)** GO-term enrichment of LSC- and KC-high genes and genes close (within 20kb) to differential CREs. **D)** UCSC genome browser screenshots showing signals of RNA-seq, ATAC-seq, ChIP-seq of H3K27ac, H3K4me3, and H3K27me3 om KCs and LSCs at the loci of *PAX6* and *GATA3*.

Gene regulatory network analysis identifies transcription factors controlling distinct epithelial cell fates and their hierarchy

Using the identified differential CREs, we set out to identify key TFs driving the cell fate differences between LSCs and KCs. TF binding motif enrichment was performed using Gimme Motifs[50] in all differential CREs marked by ATAC, H3K4me3, H3K27ac, and/or H3K27me3 signals. In general, TF motifs enriched in CREs with active marks, ATAC, H3K4me3, or H3K27ac, in one cell type, were also enriched within regions with the repressive H3K27me3 mark in the other cell type (figure 3A, supplementary figure 6A). For example, TF motifs that are linked to FOXC1, TEAD1, JUN, PAX6, FOS, RUNX2, OTX1, ELF3, SOX9 and REL were detected in differential CREs marked by high active mark signals in LSCs but also marked by high H3K27me3 in KCs. Consistent with our expectation, the enrichment of the PAX6 motif in differential CREs with higher active mark signals was detected in LSCs, as PAX6 is specific for LSCs but not for KCs. As REL is a TF involved in TNF-α and NF-kB pathways, the detection of the REL motif is consistent with the enrichment of TNF-α and NFKβ signaling genes among LSC-high genes. Notably, the FOS motif that is associated with FOS, FOSL1, FOSL2, JUN, etc.(supplementary figure 6A) was present in approximately 10% of all variable CREs in LSCs, the highest among all motifs (figure 3A). Motifs enriched in KC active CREs included those linked to KLF6, GRHL1, HOXC10, GATA3, NFIA, CEBPA, and CTCF. These motifs were also enriched in CREs marked by high H3K27me3 signals in LSCs. Enriched motifs could mostly be linked to TFs with high expression differences between LSCs and KCs, e.g., FOXC1, PAX6 and FOS are highly expressed in LSCs, while HOXC10, GATA3, and CEBPA are highly expressed in KCs (figure 3A). However, this is not the case for all identified motifs. For example, the motifs of RUNX2, KLF6, GRHL1, and CTCF were highly enriched in active CREs in KC, but no significant gene expression difference was detected between LSCs and KCs, suggesting that these TFs have a role shared in both LSCs and KCs but may control different target genes in the two cell types.

As the motif prediction approach using Gimme Motifs does not consider the expression of TFs and their targets, we applied ANANSE[51], a gene regulatory network method to identify key TFs for cell identities and cell fate conversions. ANANSE integrates CRE activities and TF motif predictions with the expression of TFs and their target genes, to generate a gene regulatory network of the specific cell type. Subsequently, a pairwise comparison of gene regulatory networks from two cell types is performed to identify the most influential TFs

that differentiate between the two cell types. The overall importance of identified TFs is represented by their influence score, scaled from 0-1 (see Material and Method)[51].

Because many key TFs for LSC and KC fates shared in the two epithelial cell types such as p63 showed similar gene expression levels, partially due to the common ectodermal origin of these cells (supplementary figure 1F), we could not identify these TFs using the differential gene regulatory network implemented in ANANSE. Therefore, we decided to include differential genes and CREs of embryonic stem cells (ESCs), as compared to either LSCs or KCs. This enabled us to not only identify distinct but also shared TFs for LSC and KC fates. Using RNA-seq, ATAC-seq and H3K27ac data from ESCs[52], we performed a pairwise comparison of gene regulatory networks between ESCs versus KCs or LSCs. When predicting TFs driving the ESC cell fate from both LSCs and KCs using ANANSE, we detected ESC specific TFs, such as NANOG, ZIC2, ZIC3, POU5F1 and SOX2 that are known to induce pluripotency from somatic cell types (supplementary figure 6E, F), demonstrating the effectiveness of ANANSE in predicting cell fate driving TFs. When predicting the TFs driving the LSC or KC fates from ESCs, ANANSE resulted in 70 epithelial TFs that had influence scores above 0.5 in both ESC-LSC and ESC-KC pairwise differential network analysis. Many of these shared TFs are known to be important for epithelial cell function, such as TP63, EHF, TFAP2A, TFAP2C, FOSL2, the KLF family (3,4,5,6,7), JUNB, CEBPD, CEBPB, and RUNX1. We classified these TFs as shared epithelial TFs (figure 3B, supplementary figure 7). Intriguingly, the prediction of one TF regulating other TFs represented by the outdegree analysis of the top 20 TFs detected that FOSL2, JUN, TP63 and TFAP2A are most likely to regulate other TFs in both LSCs and KCs (supplementary figure 7B, C), which is in line with the high percentage of detected FOS motif (figure 3).

Importantly, in the ESC-LSC and ESC-KC differential network analysis, TFs with high influence scores in LSCs but with undetectable (PAX6, ELF3, OTX1, PPARD) or low (FOSL1 and SMAD3) influence scores in KCs were considered as LSC specific TFs (figure 3B). Consistent with these findings, in the pairwise differential network analysis between LSCs and KCs (figure 3C), the predicted LSC specific TFs from this comparison were largely consistent with LSC specific TFs predicted from the ESC-LSC and ESC-KC differential network analysis (figure 3B). Interestingly, FOXC1 was annotated as a shared epithelial TFs in the approach of using ESC-LSC and ESC-KC comparisons, whereas in the KC-LSC comparison it was identified as a specific TF for the LSC fate. This is probably due to the higher expression of FOXC1 in LSCs. For KC specific TFs, we only identified

HOXA9 in the ESC-LSC and ESC-KC differential network analysis (figure 3B), This was confirmed by the pairwise comparison between KC-LSC. Next to HOXA9, the pairwise comparison of KC-LSC identified other TFs such as GATA3, IRX4, and CEBPA with high influence scores in KCs (figure 3D), indicating that KC-LSC pairwise comparison is more sensitive for detecting KC specific genes.

Finally, we set out to dissect the TF regulatory hierarchy for the cell identity differences between LSCs and KCs, by identifying potential target TFs of the shared and specific TFs. For this analysis, we did not consider the expression level of TFs themselves, because we considered that the potential binding of a TF to its target loci, represented by a binding score, is more important in this prediction. If a target TF is regulated by a TF with similar binding scores compared to ESCs in both LSCs and KCs, this regulation is annotated as a 'shared regulation'; if the binding score is significantly higher in one cell type than in the other (see Material and Method), the regulation of the TF-target TF pair is annotated as 'cell type specific regulation'. We included the top shared and specific TFs, 15 shared epithelial TFs, and six LSC specific TFs (PAX6, ELF3, OTX1, PPARD, SMAD3, and FOSL1), and the only KC specific TF HOXA9 (figure 3B, 3E, supplementary figure 6B, C, D).

As expected, there are many shared TFs regulating each other via 'shared regulation' (figure 3E, grey arrows). Consistently, cell type specific TFs regulate their target TFs via 'cell type specific regulation' (figure 3E, orange arrows), e.g., *PAX6* is predicted to be regulated by SMAD3 and PPARD. Furthermore, many autoregulation loops were also detected, e.g., PAX6 in LSCs and HOXA9 in KCs. Strikingly, we also found that shared TFs may regulate cell type specific TFs in via 'cell type specific regulation'. For example, p63, FOXC1, and TFAP2A were identified as shared TFs between KCs and LSCs, but they were predicted to regulate *PAX6* in LSCs.

In summary, our molecular characterization using KCs and LSCs cultured *in vitro* identified shared and cell type specific TFs for the LSC and KC fates. p63, FOSL2, EHF, TFAP2A, KLF5, RUNX1, CEBPD, and FOXC1 are among the shared epithelial TFs for both LSCs and KCs. PAX6, SMAD3, OTX1, ELF3, and PPARD are LSC specific TFs for the LSC fate, and HOXA9, IRX4, CEBPA, and GATA3 were identified as KC specific TFs. Furthermore, LSC and KC fates are defined by cooperative regulation of both shared and cell type specific TFs.

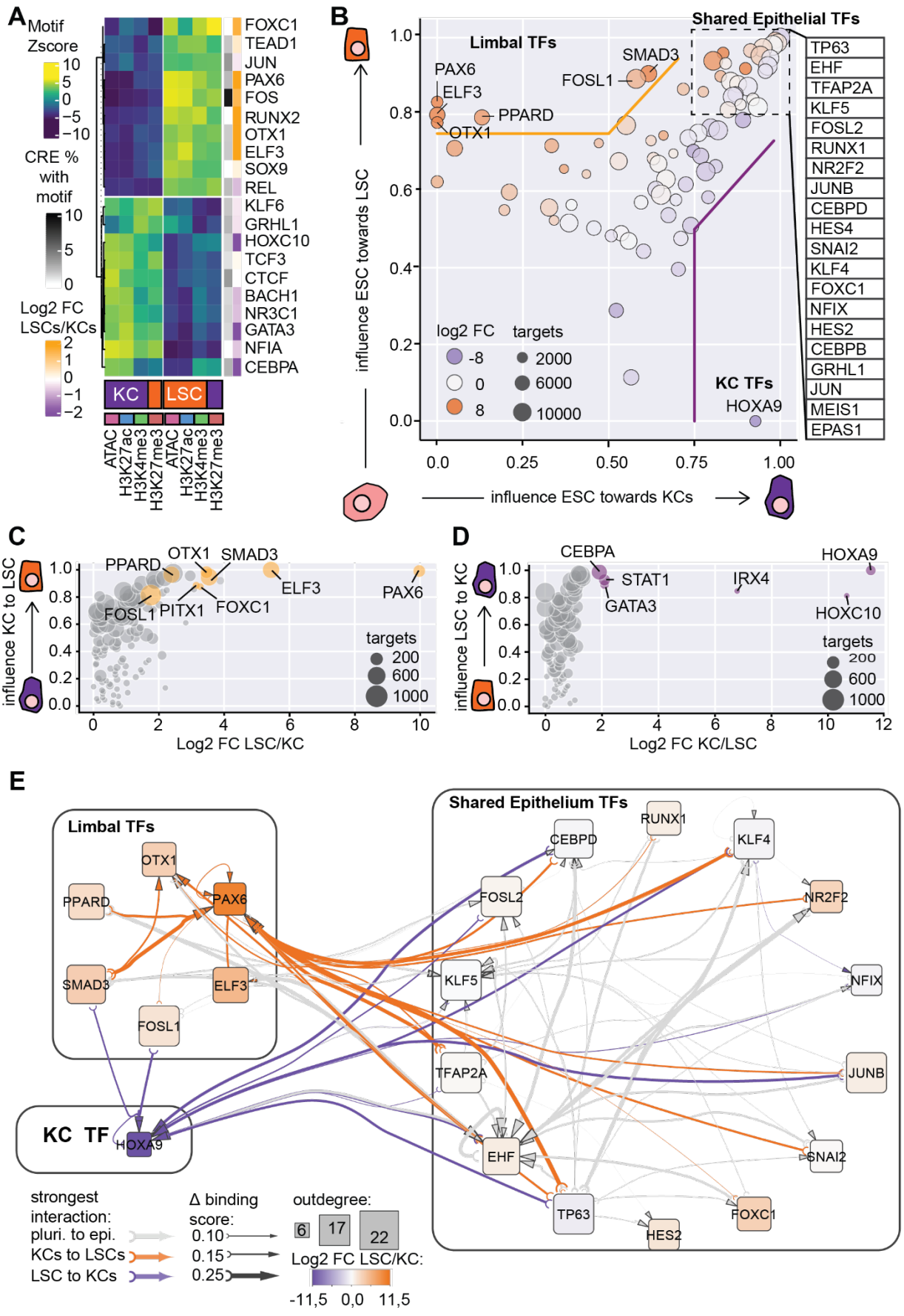


Figure 3 TFs and TF hierarchy controlling distinct epithelial cell identity. A) Heatmap of TF motif enrichment Z-scores detected in variable CREs and the corresponding TFs. The percentage of CREs containing the motifs and the expression ratio of TFs in LSCs and KCs are indicated. **B)** ANANSE influence score plot of TFs identified in ESC-KC (x-axis) and ESC-LSC (y-axis) comparison. Circle size represents the maximum number of target genes of a TF. The color represents log2FC between LSC/KC (orange LSC high; purple KC high). **C)** ANANSE influence score plot of TFs identified in KC-LSC comparison. **D)** ANANSE influence score plot of TFs identified in LSC-KC comparison. **E)** TF hierarchy is indicated by the binding score of a TF to its target TF locus, and the cell type specific regulation is indicated by the binding score difference of the TF at the target TF locus between cell types. When a binding score difference in KC-LSC comparison is greater than the mean of the difference in ESC-KC and ESC-LSC comparison, this TF regulation of the target TF is annotated as either KC- (purple arrows) or LSC specific (orange arrows) regulation. Otherwise, the regulation is annotated as 'shared regulation' for both cell types (grey arrows). The degree of binding score difference binding score is indicated by the thickness of the arrows. Outdegree node size represents the number of target genes. Fold change of TF gene expression in LSC and KCs is represented by orange (LSC-high) and purple (KC-high) colors and TFs with light colors are shared TFs.

Single-cell RNA-seq analysis of the cornea and the epidermis validates expression of key transcription factors controlling cell fates

Since our multi-omics analysis was performed on LSCs and KCs cultured *in vitro,* we assessed single-cell RNA-seq datasets derived from the cornea and the epidermis, to confirm that the molecular signatures of LSCs and KCs in our study indeed represent those of somatic stem cells maintaining the corneal limbus and the epidermis[53,54]. By clustering single cells according to marker gene expression (supplementary figure 8), we selected the cell clusters corresponding to the stem cells as pseudobulk for further differential gene expression analysis. For the epidermis, we selected cells with high *KRT14*, *KRT5,* and low *KRT1* and *KRT10* expression as basal KCs, and for the cornea, cells with high *S100A2*, *PAX6* and *TP63* expression and without *CPVL* expression as LSCs, because *CPVL* has been proposed as a marker with neural crest origin[54].

Consistent with the *in vitro* findings, the *in vivo* LSCs expressed high levels of *PAX6*, *ELF3*, *FOXC1*, *FOSL1, OTX1,* and *SMAD3*, whereas the *in vivo* KCs expressed high levels of *HOXA9, CEBPA, and GATA3* (figure 4A). GO analysis identified similar functions of differentially expressed genes between the *in vivo* LSCs and KCs, as compared to those from *in vitro* cultured cells (supplementary figure 8E, F). Furthermore, PROGENy analysis of differentially expressed genes between *in vivo* LSCs and KCs showed that TNF-α and NF-kB pathway genes are significantly enriched in *in vivo* LSCs, e.g., *CXCL1,2,3,8,20* and *NFKB1* (figure 4B). GSEA analysis using the hallmark gene set of the MsigDB collection also identified enrichment for TNF-α signaling genes (see supplementary table 5).

As the data to analyze the epigenetic states of CREs of *in vivo* tissues were not available, we performed gene regulatory network analysis using the *in vitro* ATAC and H3K27ac datasets, together with the *in vivo* single-cell RNA-seq data that were aggregated as pseudobulk. Since the GRN analysis is largely driven by gene expression data, this analysis is meaningful to assess the influence of TFs on *in vivo* LSC and KC fate differences. Overall, the *in vivo* data identified similar cell type specific TFs, as compared to *in vitro* cultured cells (figure 4C and 4D). In *in vivo* LSCs, except PPARD that was not detected, PAX6, ELF3 FOXC1, and FOSL1 exhibited the highest influence scores, and TFs with the highest influence scores in *in vivo* basal KCs were HOX TFs and a few others such as CEBPA, GATA3, and IRX4.

Taken together, our analyses showed clear consistency between *in vivo* and *in vitro* derived data and identified key TFs driving the cell fate difference between LSCs and KCs.

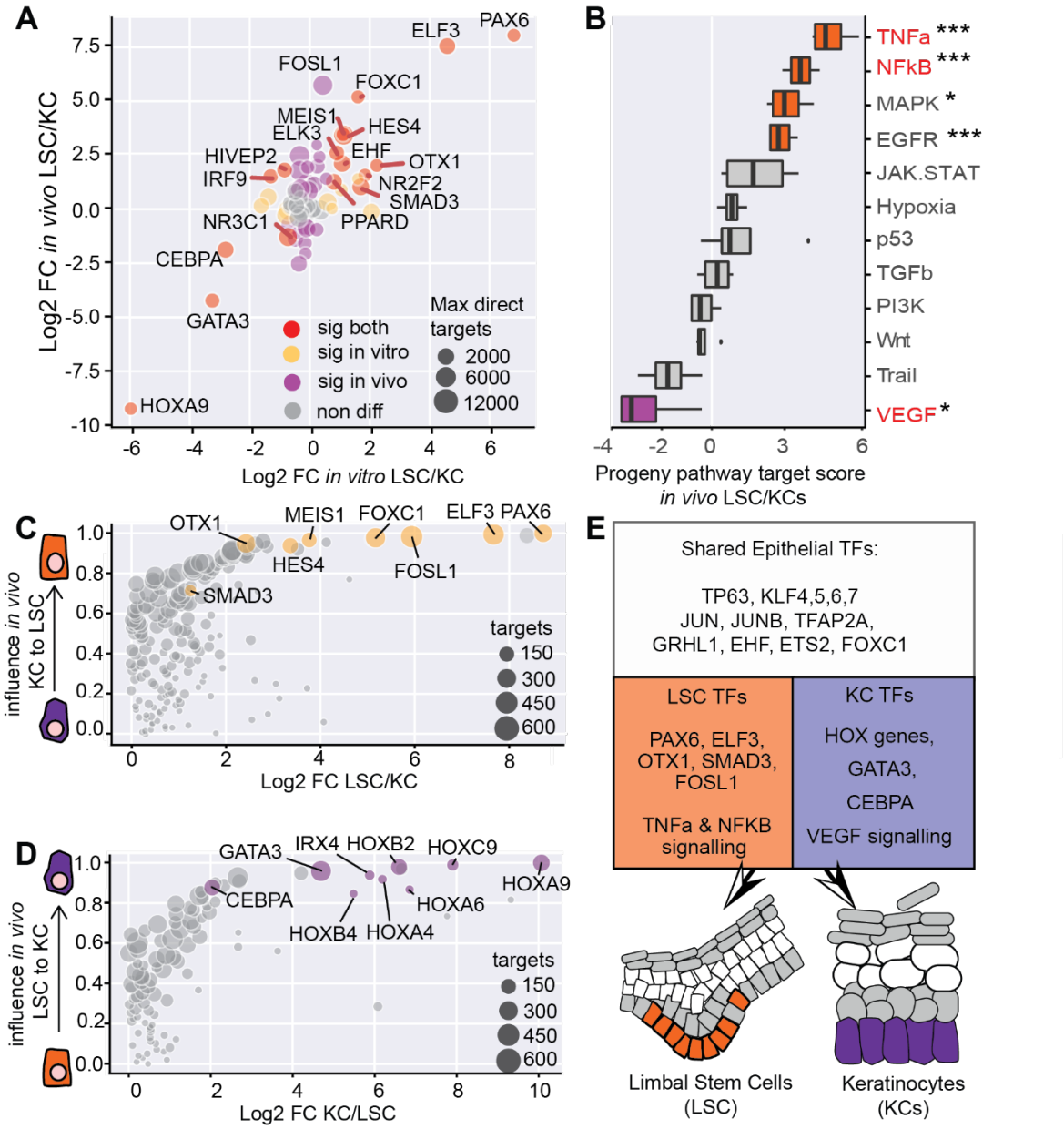


Figure 4 Validation of key TF expression using *in vivo* single-cell RNA-seq A) Fold change comparison of identified TFs using *in vivo* and *in vitro* data. **B)** PROGEny pathway analysis of *in vivo* LSCs and KCs **C)** ANANSE influence score plot of *in vivo* basal Cs to LSCs. **D)** ANANSE influence score plot of *in vivo* LSC to basal KCs. **E)** Summary of the identified shared and specific TFs.

Transcription factors controlling the LSC fate contribute to cornea diseases

As pathomechanisms associated with TF mutations in corneal opacity are not yet explored, we questioned whether key TFs defining the LSC fate and their target genes are relevant to cornea-related diseases. We used two approaches to collect genes that are associated with corneal phenotypes. First, we performed a literature search and constructed a gene list of 161 genes associated with LSCD and inherited corneal opacification diseases, which included epithelial but also stromal and endothelial cornea diseases (e.g., corneal dystrophies, keratoconus) and ocular as well as systemic syndromes with known corneal manifestations (Supplementary table 7). In parallel to this curated disease gene list, we used disease genes assembled in the EyeDiseases database[55] that includes genes associated with other eye diseases such as glaucoma and refractive error.

To interrogate whether the corneal disease genes can be regulated by the LSC defining TFs, we generated ChIP-seq of the p63 protein in LSCs and incorporated publicly available ChIP-seq data of PAX6, FOXC1, RUNX1, and SMAD3 in LSCs[28,30]. To assess whether binding of these TFs to corneal disease gene loci is more likely to occur than random, we examined TF binding signals by integrating TF ChIP-seq signals and the distance of the ChIP-seq peak to the disease genes (Supplementary figure 9A) and used a Mann Whitney U statistical test for significance. This analysis showed that genes in the curated corneal disease gene list are bound more often by FOXC1 and PAX6 with statistical significance, as compared to TF binding to all genes in the genome (figure 5A and 5B). Similarly, glaucoma genes are bound by p63, PAX6 and RUNX1, and p63, PAX6 and RUNX1 can bind to genes associated with Fuchs endothelial corneal dystrophy, refractive error and keratoconus, respectively (figure 5A and 5B). FOXC1, PAX6, and RUNX1 seemed to regulate most disease genes with high TF binding signals, especially those among the curated cornea disease genes, keratoconus and glaucoma genes (figure 5B). p63 regulates a small number of target disease genes that are probably expressed in epithelial cells. Many of these disease genes such as *TGFBI*, *JAG1* and *CAV1* are likely co-regulated by different TFs, as multiple binding sites of these TFs were observed at these gene loci (figure 5B and 5C). In line with these findings, alternative mapping of TF ChIP-seq binding sites to nearest genes resulted in similar statistical significance of disease genes being likely regulated by these TFs (Supplementary figure 9B,9C and 9D). As our results suggested that *TGFBI*, *JAG1*, and *CAV1* are regulated by PAX6, we examined whether their gene expression is affected in aniridia patient LSCs that have

PAX6 haploinsufficiency. Using RNA-seq data, indeed we observed that *TGFBI*, and to less extent also *JAG1* and *CAV1*, was downregulated in aniridia patient LSCs (figure 5D).

To explore the possible role of key TFs defining the LSC fate in contributing to the pathomechanism of corneal diseases, specifically corneal opacity, we leveraged the whole genome sequencing data in the 100,000 Genomes Project at Genomics England UK to identify variants of uncertain significance that may have functional consequences. To establish a suitable cohort, we identified a total number of 33 unsolved participants with human phenotype ontology (HPO) terms associated with corneal opacity (including HP:0007957, supplementary table 8). In a proband with band keratopathy (HP:0000585), we identified a *de novo* heterozygous missense variant in *FOSL2* (2:28412095:C:T, genome build GRCh38/hg38, NM_005253.4:c.628C>T), giving rise to a predicted damaging amino acid change (NP_005244.1:p.(Arg210Cys)), based on most major prediction tools (supplementary table 9).

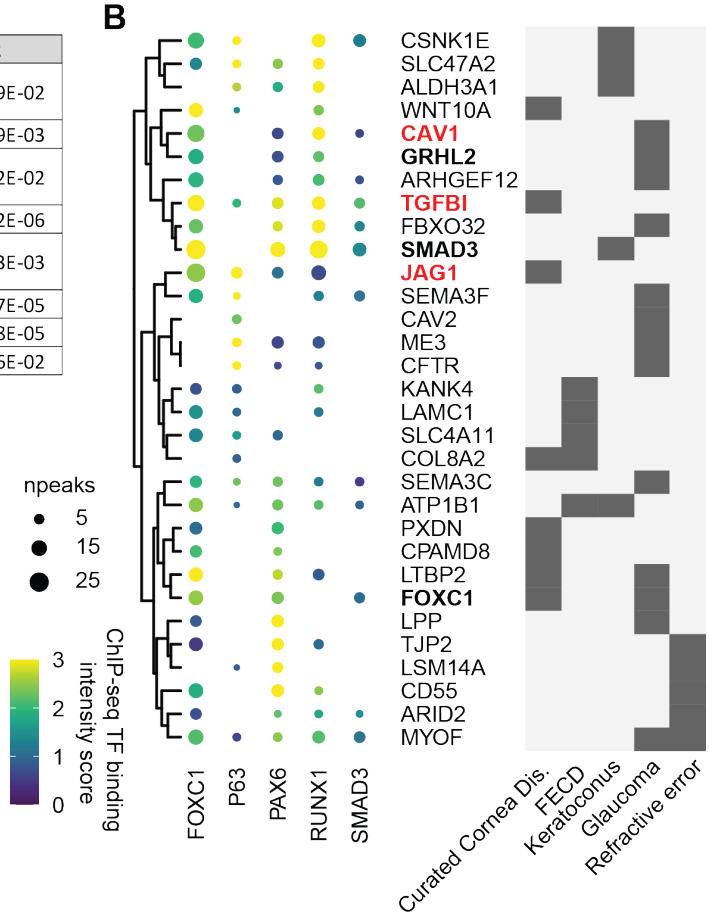
Since FOSL2 may be a novel gene associated with corneal opacity, we questioned whether FOSL2 could cooperate with the other three TFs that are known to associate with corneal opacity; PAX6 associated with aniridia and Peters Anomaly[31,56], FOXC1 associated with Axenfeld-Rieger Syndrome and Peters Anomaly[57,58], and p63 associated with EEC[17]. Using the curated corneal disease genes and those assembled in the EyeDiseases database, we constructed gene regulatory networks using predicted binding scores of PAX6, FOSL2, FOXC1 and p63 at the loci of corneal disease genes. This analysis showed that many eye disease genes, among which many are involved in corneal opacity, are potentially co-regulated by these four TFs (Supplementary figure 10). For example, genes associated with Peters Anomaly, such as *PXDN* and *CYP1B1,* are likely co-regulated by p63, FOXC1 and FOSL2. This analysis also revealed predicted regulations between these TFs, e.g., regulation of *PAX6* by p63 and FOSL2, auto-regulation of PAX6 and FOSL2, and coregulation between PAX6 and FOXC1 and between FOXC1 and FOSL2.

In summary, our results demonstrate that the identified shared and LSC specific TFs that define LSC fate contribute to corneal opacity and showed proof of principle that these key LSC TFs and their target genes can be leveraged as a resource for genetic studies of corneal opacity.

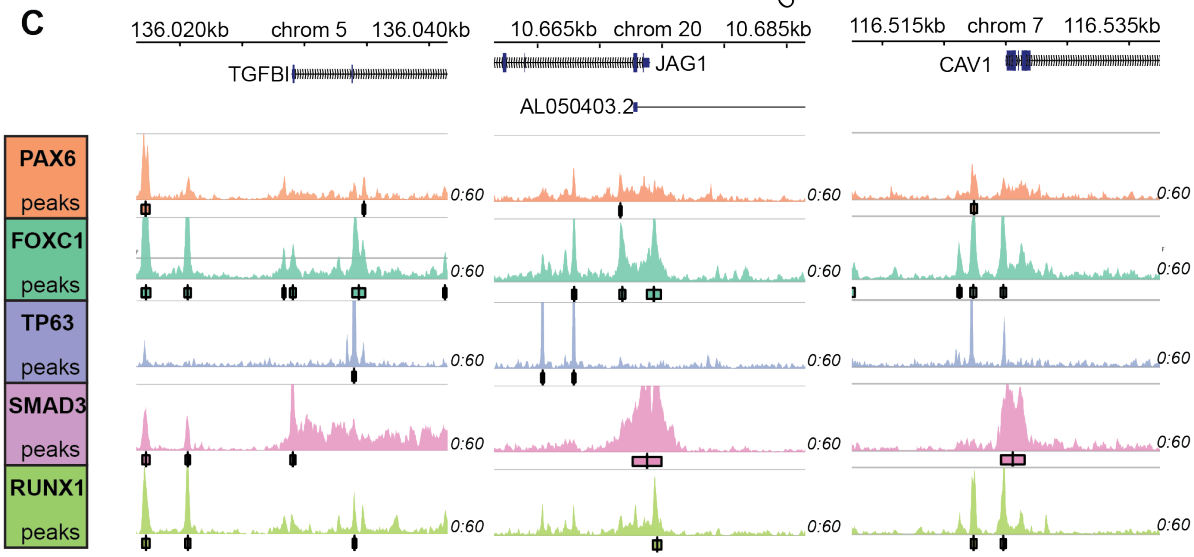
A

TF	disease	FDR
FOXC1	Curated cornea diseases	3.9E-02
P63	Glaucoma	4.9E-03
P63	Fuchs' endothelial corneal dystrophy	4.2E-02
PAX6	Glaucoma	6.2E-06
PAX6	Curated cornea diseases	1.3E-03
PAX6	Refractive error	8.7E-05
RUNX1	Glaucoma	6.8E-05
RUNX1	Keratoconus	3.6E-02

B



C



D

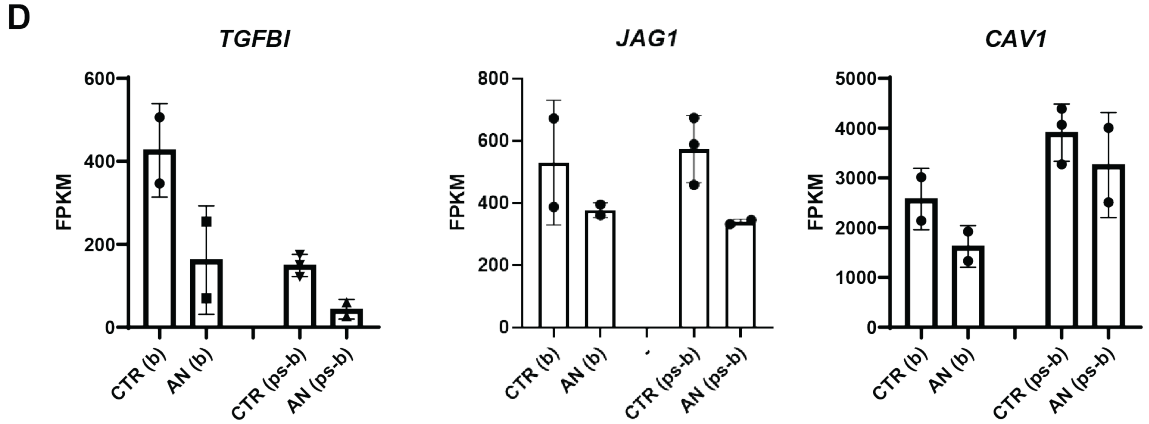


Figure 5 Disease-associated genes potentially targeted by TFs defining LSC fate. A) TFs that bind to gene loci associated with corneal abnormalities with significantly higher occurrence (FDR), as compared to binding to all genes in the whole genome. FDR was calculated with Mann-Whitney U one-sided test. **B**) Dot plot showing the ChIP-seq binding intensity (color bar) and the number of ChIP-seq peaks (npeaks, dot size) near the top 5 disease genes in disease gene lists that contain a significant number of potential TF targets. The number of peaks is peaks within 100kb of the transcription starting site (TSS); ChIP-seq intensity score is the weighted z score of the quantile log normalized ChIP-seq-peak intensities distance weighing. FECD, Fuchs Endothelial Corneal Dystrophy. Genes highlighted in red were further analyzed in D. **C)** UCSC genome browser screenshots showing ChIP-seq profile of PAX6, FOXC1, p63, and RUNX1 in LSCs at the loci of *TGFBI*, *JAG1*, and *CAV1,* highlighted in red in B. **D)** RNA-seq signals represented by Fragments per kilobase of transcript per million reads (FPKM) of *TGFBI*, *JAG1*, and *CAV1* in control (CTR) and in aniridia patient (AN) LSCs. RNA-seq was performed either in bulk (b) or in pseudobulk (ps-b), the latter aggregated from single-cell RNA-seq signals.

Discussion

Discussion

The corneal epithelium and the epidermis are both stratified epithelia, serving as barriers and the first-line defense against external insults. Nevertheless, they have distinct tissue-specific functions that are tightly controlled by the proliferation and differentiation program of their corresponding adult stem cells, LSCs for the cornea and basal KCs for the epidermis. In this study, we characterized molecular signatures defining the cell fates of these two cell types by integrating in house multi-omics data as well as publicly available datasets. Using motif and gene regulatory network analyses, we identified a collection of shared and specific epithelial TFs defining KCs and LSCs. Furthermore, we showed a proof-of-principle that this resource of LSC-defining TFs and their regulatory mechanisms can provide novel tools for dissecting pathomechanisms of corneal diseases.

In contrast to the well-studied TFs and their associated gene regulatory networks in KCs of the epidermis[9–12,59], TFs regulating LSCs of the cornea have only started to emerge recently. These TFs include PAX6, p63, SMAD3, RUNX1 and FOXC1 that are of importance for the proper LSC identity[28,30,37]. Except for PAX6, all other TFs are also expressed in KCs[9–12,59], which raises interesting questions whether these TFs are sufficient to determine cell fate differences between LSCs and KCs and how they control cell fate determination mechanisms. By specifically comparing LSCs to KCs, we identified PAX6, SMAD3, OTX1, FOSL1, and ELF3 as the TFs that determine LSC fate that is different from KCs. The identification of PAX6 as a LSC specific TF was expected. It is an eye development master regulator[28] and is associated with the disease aniridia where corneal opacity is one of the main manifestations[31]. Furthermore PAX6 has previously been shown to co-regulate target genes with RUNX1 and SMAD3[30]. Although SMAD3 is also expressed in KCs, it has clearly more influence in LSCs than in KCs in our study. OTX1 is an important TF for regulating the neural lineage[60,61]. In mice, both Otx1 and its ortholog Otx2 are vital for tissue specification during eye development, particularly of the retinal pigmented epithelium[62,63]. FOSL1 is known to have functions in epidermal cells. FOSL1, and FOSL2, together with other AP-1 TFs can form a complex[64], and this complex regulates many biological processes including epidermal stratification[65]. In our differential motif analysis, the FOS motif that can be bound by FOS, FOSL1 and FOSL2 is the most abundant motif enriched in LSCs, as compared to KCs, highlighting the potential role of FOSL1 in LSCs. ELF3 has previously been linked to *KRT12* and *KRT3* regulation[66], and is

one of the TFs identified to play a major role in LSC stratification[37]. Further studies on these LSC-specific TFs may provide insights into transdifferentiation strategies to regenerate LSCs from KCs.

We found that only a small number of TFs have been identified as KC specific, including GATA3 and CEBPA, and HOX genes. Among these TFs, GATA3 and CEBPA have been reported to have a role in the epidermis. CEBPA has been shown to regulate p63 expression[67], while GATA3 is regulated by p63 in the epidermis[68–70]. Interestingly, GATA3 was upregulated in aniridia patient LSCs, which is in line with the concept that loss of PAX6 in aniridia LSCs could lead to a KC-like signature. Nevertheless, our GSEA analyses did not show significant gene expression similarity between upregulated genes in aniridia LSCs and KCs, indicating that aniridia patient LSCs do not acquire a complete KC fate. Surprisingly, we detected HOXA9 as one of the KC specific TFs. HOX genes are well known in antero-posterior body patterning and segmentation where mesodermal genes are mainly involved[71,72], but little is known for their function in the epidermis. One plausible interpretation is that, as LSCs are from the eye and KCs are from the trunk of the body, the detection of HOX9 for the KC fate simply marks the positional information along antero-posterior axis. Nevertheless, the lower number of KC specific TFs, as compared to LSC specific TFs, indicates that repression of LSC specific TF expression is critical for the KC fate. This is in line with our observation that LSC specific TFs such as the *PAX6* locus is completely covered by H3K27me3, probably via polycomb repression. This may also indicate that the KC fate is similar to the cell fate of the surface ectoderm during embryonic development, and the LSC fate is later established during development via activation of LSC specific genes. It is also worth noting that the ANANSE prediction tool used in this study is unable to predict TFs with transcriptional repression functions[51], which limits the identification of TFs to repress LSC genes, if there is any in KCs. Consistent with this, ANANSE did not detect repressive TFs such as OVOL2[73] and SOX9[74] in LSCs.

We anticipated that TFs that are shared but important to both LSCs and KCs could be not identified through pair-wise gene regulatory network comparison between LSCs and KCs. For detecting these shared TFs, we compared both cell types to pluripotent stem cells in the gene regulatory network analysis. This approach indeed resulted in a significant number of shared epithelial TFs including p63, FOSL2, EHF, TFAP2A, KLF4/5, FOS, JUN, RUNX1 and FOXC1. Many of these common TFs have previously been linked to important functions in both epidermis and cornea[11,10,14,75,30,76–78]. Among them, p63 is a regulator of

stratified epithelial, and drives transdifferentiation of non-epithelial cells into both KCs and LSCs[12,79].

Our TF hierarchy analysis indicated that the shared TFs seem often to (co-)regulate cell type-specific TFs. For example, p63, KLF4 and TFAP2A can potentially co-regulate *PAX6* in LSCs. This is somewhat unexpected, because during mouse eye development, Pax6 expression was first observed in retina and lens, and subsequently in cornea at E12.5, prior to p63, whose expression was only detected in the cornea at E14.5[80]. However, the expression sequence of these genes *in vivo* may not be in conflict with the observed regulatory hierarchy. *TP63* is known to be expressed in the surface ectoderm where the corneal epithelium is also thought to be developed from[81]. In these ectodermal cells that later migrate to form the corneal epithelium, p63 may be an upstream regulator of *PAX6*. However, our data indicate that p63 expression itself is not sufficient to induce PAX6, and therefore other mechanisms, likely via chromatin and epigenetic modifications, must be involved to de-repress the *PAX6* locus.

In addition to identified TFs, NF-KB and TNF signaling pathway genes have been shown to be enriched for the LSC fate. The enrichment of the REL motif detected in LSCs provides further evidence for TNF and NF-KB pathway activation. In general, TNF signaling is thought to function mainly via immune cells, increasing inflammation of tissues[82] and leading to neovascularization and lymphangiogenesis[8384]. However, these studies do not directly address the function of TNF signaling in normal LSCs and corneal epithelium. Consistent with our findings, a single-cell RNA-seq study on the human cornea[54] reported TNF expression in LSCs in the cornea. Even more intriguingly, treatment of TNF on LSCs increased proliferation and expression of LSC markers such as p63 and GPHA2, supporting a role of TNF signaling in normal LSC function.

The key role of TFs in cell fate control is often demonstrated by their association with developmental diseases. PAX6, FOXC1 and p63 are known to be associated with corneal opacity[17,31,35], and FOSL2 is a novel corneal opacity candidate gene identified in this study. Except PAX6, all other TFs, p63, FOXC1 and FOSL2 are shared between LSCs and KCs. Based on our TF binding prediction, PAX6 is likely a downstream target of all these three TFs. PAX6 and FOXC1 seem to regulate most identified disease genes, which is consistent with their broad expression patterns in the eye and the phenotypic heterogeneity and overlap linked to *FOXC1* and *PAX6* mutations, e.g., iris and corneal

defects and higher prevalence of glaucoma[31,35], reinforcing a common regulatory network shared by the two TFs[85,86]. As for PAX6 and p63 associated disorders, although *PAX6* and *TP63* mutations are known to cause corneal opacity, other phenotypes are quite distinct, fully in line with their gene expression in different tissues, e.g., *PAX6* in the cornea epithelium, iris, retina, pancreas and parts of the central nervous system, and *TP63* in the cornea epithelium, skin epidermis and other stratified epithelia [17,31]. Although FOXC1 and FOSL2 were both annotated as shared TFs of LSCs and KCs in our study, their relevance for the skin and cornea might still be different. In our analysis, FOXC1 had a higher influence score in LSCs, probably due to its higher expression in LSCs. This is in line with skin phenotypes not being reported in anterior segment dysgenesis associated with *FOXC1* mutations. As FOSL2 often cooperates with FOSL1 that is a LSC specific TF[87], the importance of FOSL2 in LSCs can be envisaged. Consistent with our novel finding that a potentially damaging variant in *FOSL2* is associated with corneal opacity, decreased expression of FOSL1 in the cornea has been linked to keratoconus patients[88]. Further investigation of the role of FOSL2 and its variants in corneal opacity is required.

In summary, we identified shared and cell type specific epithelial TFs and signalling pathways that are important in determining cell fate of LSCs and KCs. This study also provides potential pathomechanisms for cornea and eye diseases. Furthermore, the identified TFs and signalling molecules can also be applied to develop transdifferentiation strategies to generate functional LSCs for corneal regeneration.

Data availability

All non-patient cell raw sequencing files generated in this study have been deposited in the GEO database with the accession number GSE206924. All aniridia-patient sequencing files have been deposited in the dbGaP database with controlled access. Publicly accessible data was downloaded from GEO using the accession codes provided in supplementary tables 1, 2 and 3. There are no restrictions on data availability. Source data are provided with this paper

All code used in this study is available at https://github.com/JGASmits/regulatory-networks-in-epidermal-and-corneal-epithelia

Abbreviations

Cis-regulatory element (CRE), Embryonic stem cells (ESCs), False Discovery Rate (FDR), Fuchs Endothelial Corneal Dystrophy (FECD), Gene Onthology (GO), Gene set enrichment analysis (GSEA), Keratinocytes (KCs), Limbal stem cell deficiency (LSCD), Limbal stem cells (LSCs), single-cell RNA-seq (scRNA-seq), Transcription factor (TF).

Author contribution

JGAS, SJH, and HZ designed the research; JGAS, JQ, LL, LNR, and DLC performed experiments; JGAS, SJH, and HZ analyzed multi-omics data; LL, LNR, NS, BS and DA contributed limbal stem cells and culturing expertise, NO and MM analyzed whole genome sequence data; JGAS, SJH and HZ wrote the original draft; all authors edited and contributed to the manuscript.

Acknowledgement

We thank M.P.A. Baltissen, L.A. Lamers and S. Rinzema for operating the Illumina analyzer and initial data demultiplexing, L. Wingens for support with the celseq library preparations, J. Arts, and W.N. Twilhaar for their pre-processing of the single cell RNAseq objects of the in vivo datasets. This research was supported by NWO-ALW (ALWOP 376, JGAS, HZ); EJP-RD (JPRD20-135, DLC, HZ); COST Consortium ANIRIDIA-NET (CA18116, DLC, HZ, LL, NS, DA). This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer

Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support.

Competing interests

The authors declare no competing interests.

Materials and Methods

KC and LSC cell culture in vitro

KCs were isolated and cultured as previously described[14]. Briefly, after isolation primary KCs were cultured in Keratinocyte Basal Medium supplemented with 100 U/mL Penicillin/Streptomycin, 0.1 mM ethanolamine, 0.1 mM O-phosphoethanolamine, 0.4% (vol/vol) bovine pituitary extract, 0.5 µg/mL hydrocortisone, 5 µg/mL insulin and 10 ng/mL epidermal growth factor. Medium was refreshed every other day until the cells were 90% confluent.

Limbal tissues were acquired as previously described[89]. Two aniridia Limbal tissue single biopsies were obtained from the superior limbus during penetrating keratoplasty from 2 patients with congenital aniridia as previously described [90]. Genetics of the aniridia patients were identified to be c.33delC p.Gly12Valfs*19 (NM_000280.2) for AN55 and c.990_993dup p.Met332Alafs*10 for AN40 (see supplementary table 1). Cell isolation was performed as previously described[89]. Briefly, limbal tissue was digested in collagenase A solution (4 mg/ml) in keratinocyte serum-free medium (KSFM) (Thermo Fisher Scientific; Waltham, MA) for 20 h at 37 °C. Cell suspensions were filtered through a use of Flowmi® micro strainer (SP Bel-Art; Wayne, NJ). LSC clusters were dissociated with trypsin-EDTA (0.05%) solution and cultivated in KSFM. Medium was refreshed every other day. Subconfluent (80–90%) limbal epithelial cells were harvested at passage 2.

Next to this approach, other LSC samples (LSC-Aberdam, see supplementary table 1) were isolated from postmortem donated peripheral corneal epithelium and cultured as previously described[49]. Briefly after isolation, they were expanded and cultured in KSFM (Gibco, Life Technologies) supplemented with 25 µg/ml Bovine Pituitary Extract (BPE; Gibco, Life Technologies), 0.2 ng/ml Epidermal Growth Factor (EGF, Peprotech, Neuilly-sur-Seine, France), 0.4 mM $CaCl_2$, 2 mM Glutamine (Gibco, Life Technologies) and 100 U/ml Penicillin/Streptomicin (Gibco, Life Technologies). Medium was refreshed every other day until the cells were 90% confluent.

Bulk RNA-seq

Total RNA was isolated using the Quick–RNA MicroPrep kit (Zymo Research), according to the manufacturer's protocol. RNA concentrations were measured using the the DeNovix DS-11FX spectrometer. 500 ng of RNA was prepared for sequencing using the KAPA

RNA HyperPrep Kit with RiboErase (Kapa Biosystems). Libraries were sequenced on the NextSeq 500 (Illumina), generating an average of 15–20 million reads per sample.

Single-cell RNA-seq library preparation and sequencing

A single-cell suspension was made using trypsin. After which cells were filtered using a 40uM filter to remove cell clumps. Cells were stained with 7-AAD. The live cells were selected for and FACS-sorted onto 384-well plates containing primers with unique molecular identifiers, according to the SORT-Seq protocol[91]. Plates were spun down (1200 × *g*, 1 min, 4 °C) and ERCC spike-in mix (1:50,000) was dispensed by a Nanodrop (BioNex Inc) into each well. 150 nl of the Reverse Transcription (RT) mix was dispensed into each well. Thermal cycling conditions were set at 4 °C 5 min; 25 °C 10 min; 42 °C 1 h; 70 °C 10 min. The library of each plate was pooled together and the cDNA was purified using AmpureXP (New England BioLabs) beads. Overnight in vitro transcription (Ambion MEGA-Script) was carried out at 16 °C, with the lid set at 70 °C. An exonuclease digestion step was performed thereafter for 20 min at 37 °C, followed by fragmentation of the RNA samples. After a beads cleanup, the samples were subjected to library RT and amplification to tag the RNA molecules with specific and unique sample indexes (Illumina), followed by a final beads cleanup (1:0.8, reaction mix: beads) and the sample cDNA libraries were eluted with DNAse free water. Libraries were quantified using the KAPPA quantification kit following manufacturers protocol after which the plates were sequenced on the NextSeq 500 (Illumina) for 25 million reads per plate.

ChIP-seq

Chromatin for ChIP was prepared as previously described (Kouwenhoven et al., 2010; Qu et al, 2018) with minor modifications. On average, 0.5M cells were used in each ChIP. Antibodies against H3K27ac (Diagenode #C15410174, 1.2 µg), H3K4me3 (Diagenode #C15410003, 1 µg), H3K27me3 (Diagenode #C15410069, 1.5 µg), p63 (Santa Cruz #H129, 1 µg, recognizing the C-terminal α tail of p63) were used in ChIP assay. Afterwards 5ng DNA fragments were pooled and proceeded on with library construction using KAPA Hyper Prep Kit (Kapa Biosystems #KK8504) according to the standard protocol. The prepared libraries were then sequenced using the NextSeq 500 (Illumina) according to standard Illumina protocols.

RNA-seq, ATAC-seq and ChIP-seq data preprocessing

Preprocessing of reads was done automatically with workflow tool seq2science v0.7.1[92]. Paired-end reads were trimmed with fastp v0.20.1[93] with default options. Genome assembly GRCh38.p13 was downloaded with genomepy 0.11.1[94]. Public samples were downloaded from the Sequence Read Archive[64] with help of the NCBI e-utilities and pysradb[95]. The effective genome size was estimated per sample by khmer v2.0[96]by calculating the number of unique kmers with k being the average read length per sample. scATAC fastq files were merged to pseudobulk by combining all fastq files from each plate using the bash command cat.

Reads of ChIP-seq and ATACseq were aligned with bwa-mem v0.7.17[97]with options '-M'. Reads of RNAseq samples were aligned with STAR v2.7.6a[98] with default options. Afterwards, duplicate reads were marked with Picard MarkDuplicates v2.23.8[99]. General alignment statistics were collected by samtools stats v1.14[100]. Mapped reads were removed if they did not have a minimum mapping quality of 30, were a (secondary) multimapper or aligned inside the ENCODE blacklist[101]. RNAseq sample counting and summarizing to gene-level was performed on filtered bam using HTSeq-count v0.12[102] . Sample sequencing strandness was inferred using RSeQC v4.0.0[103] in order to improve quantification accuracy.

ATAC samples were tn5 bias shifted by seq2science. ChiP and ATAC sample peaks were called with macs2 v2.2.7[104] with options '--shift -100 --extsize 200 --nomodel --keep-dup 1 --buffer-size 10000' in BAM mode. The effective genome size was estimated by taking the number of unique kmers in the assembly of the same length as the average read length for each sample. Narrowpeak files of ChiP-seq biological replicates belonging to the same condition were merged with the irreproducible discovery rate v2.0.4.2[105].

Single-cell RNA-seq data preprocessing

Single-cell libraries were pre-processed using the cellseq2 pipeline. Briefly, reads were aligned using star to the GRCh38.p13 genome. After which cells were quality controlled using Seurat, filtering cells on ERCC reads, genes measured and transcripts per cell. After visualization of the lack of heterogeneity by Umap, pseudobulk count data was generated by summing all the cells their UMI counts. Cellular heterogeneity was assessed using the analysis file Generate_scRNAseq_pseudobulk.Rmd . Finally single-cell and bulk gene count tables were merged for a combined bulk and pseudobulk analysis.

RNA-seq data analysis and normalization

The bulk and pseudobulk count tables were merged on gene names, keeping genes measured in either or both of the datasets. Due to potential sex differences between donors, genes located on chromosome X and Y were removed. Finally, genes with less than 10 counts per row were removed.

Variance visualization was performed using sample distance and PCA. Variance an distance was visualized before and after removing technical variation differences between bulk, pseudobulk, and in case of public available anirdia LSC datasets 3' end enriched RNAseq, using Limma[106]. This indicated that the driving difference between the samples after removing the technical difference was cell type (see supplementary figure 2A and supplementary figure 2A&B and supplementary figure 4B&C).

Rld normalization was used for normalizing gene intensities. Between all conditions differential genes were detected using Deseq2[107]. Non-batch corrected count tables were used for identifying the DEGs. Ashr log2 fold change shrinkage[108] was used to shrink the Log2 fold change values. DEG cutoffs were set as an adjusted *p* value of 0.01 or lower, and an absolute log2 FC of 0.58 and larger.

Complex Heatmap[86] with the circlize colour package[109] were used to visualize the DEGs. Subsequently Progeny enrichment[46] was performed to quantify signaling pathway target gene enrichment. Clusterprofiler[43] was run for GO-term enrichment on DEGS of each comparison. Finally foldchange of all genes were used to generate a gene list for GSEA enrichement of the the MySigDB collections[44]. Gene names were mapped to ENTREZID using AnnotationDbi[110] and these were used to run KEGG pathway enrichment. The enriched pathways were visualized using pathview[111].

Identification of CREs

In order to identify CREs, ATAC-seq was used. Bulk and scATAC data were merged from *in vitro* expanded KCs and LSCs. Next to the generated datasets, publicly available data was incorporated. To prevent a sequencing depth bias, the top 100.000 ATAC peaks from each cell type were combined, the overlapping peak summits were merged, and histone modifications in varying window sizes around these ATAC peaks were quantified using histone ChIP-seq datasets (figure 2A).

For ATAC signal quantification the ATAC intensity was quantified in 200bp around the peak summits, for the promoter mark H3K4me3 and the enhancer mark H3K27ac a 2kb window was used and finally for the repressive H3K27me3 mark a 5kb window was used

for quantification (figure 2A). This resulted in an extensive dataset containing cis regulatory elements and their respective histone modification signal intensity.

Differential CREs were identified for the ATAC-seq and H3K27ac reads by running DESEQ2 on the read counts within the defined windows and identified regions (adjusted p-value < 0.05). For H3K4me3 and H3K27me3 signals, Differential CREs were identified with two steps. First the histone mark distribution was plotted, and CREs with a low to no histone signal were disregarded. Next, high activity regions with variable signal were selected. (Supplementary Figure 5).

Variable cis-regulatory elements were linked to genes with two approaches

1. CREs were linked to all the TSS regions within a 100kb window using bedtool window[83] after which the CREs per TF were distance weighted and summed based on the ANANSE distance weighing approach including promoter peaks[51].
2. CREs were linked to the closest TSS region within 20kb using bedtool closest[112].

After linking the regions to genes in both approaches, intensity scores were printed to a CSV file. Heatmaps and go term enrichments were generated in R using clusterprofiler and complex heatmap.

Single-cell ATAC-seq

A single-cell suspension was made using trypsin. After which cells were filtered using a 40uM filter to remove cell clumps. The protocol on from Chen et all[113] was used to sequence single-cell ATAC. Briefly 50.000 cells were tagmented in bulk in 45 µl of tagmentation mix (20mM Tris pH 7.6, 10mM magnesium Chloride 20% Dimethylformamide), 5µl of tagmentation protein and 0.25 µl of Digitonin. Cells were tagmented for 30min at 37 °C and 800 rpm.
Tagmentation was stopped by adding 50 µl of tagmentation stop buffer (10mM Tris-HCL PH 7.8 and 20 mM EDTA). Cells were stained with DAPI and DAPI positive cells were FACs sorted in 384-well plates containing Nextera primers with unique molecular identifiers, NACL ProteinaseK and SDS page. Plates were spun down (1200 × *g*, 1 min, 4 °C) and were incubated for 15min at 65 °C.
4µl of Tween20 was added. 2µl of H20 was added and finally 10 µl of NEBNext High-Fidelity 2X PCR Master Mix was added to each well. Thermal cycling conditions were set at 72 °C 5 min; 98 °C 5 min; and then 20 repeats of 98°C for 10s, 63°C for 30s, 72°C for 20s.

Plate libraries were pooled and purified using a Qiagen PCR purification kit with adjusted buffer volumes according to Chen et all[113] . After the column cleanup a final beads cleanup was performed using AmpureXP (New England BioLabs) beads and the sample cDNA libraries were eluted with DNAse free water. Libraries were quantified using the KAPPA quantification kit following manufacturers protocol after which the plates were sequenced on the NextSeq 500 (Illumina) for 30 million reads per plate.

Motif analysis

The gimme motifs database was pre-filtered[50] to include only motifs linked to TFs which were expressed in either KC and/or LSCs (using a cutoff of at least 10 counts in total). When multiple motifs mapped to a TF, the most variable motif was used. In case multiple TFs mapped to a motif, the most differential TF on the transcriptome was annotated to the motif. All highly variable CREs their log10 quantile normalized values were used as an input for gimme maelstrom motif enrichment analysis.

ANANSE analysis

For the gene regulatory network analysis, all the called ATAC peaks were used, merging summits and excluding peaks on the chromosomes GL, Un, KI, MT, X, and Y due to potential donor sex differences. Next, ANANSE binding was ran using all the peaks as potential enhancer regions and using both ATAC and H3K27ac signals to predict potential TF binding. To select the TF binding model a Jaccard similarity score of min 0.2 was used, to minimalize the false-positive models used to predict TF binding. For ANANSE network, the ANANSE binding files were combined with the RNAseq TPM files. This included all bulk RNAseq samples of KCs, LSCs and ESCs (see Supplementary Table 1). In the case of the *in vivo* pseudobulk data, FPKM values were used based on the UMI tables.

Finally, ANANSE influence was ran using the top 500.000 differential edges between networks. Deseq2 was ran on the countfiles of each comparison to identify differential genes needed for ANANSE influence. To prevent missing values, for the final ESC-KC, ESC-LSC, KC-LSC and LSC-KC comparison all differential edges were taken from each comparison and used to reran each comparison with all these edges included. This prevented missing values in the differential networks while comparing different differential networks.

TF hierarchy was estimated using the TF-target TF binding score generated by the influence command running the –full-output flag. This represents the motif, ATAC&H3K27ac signal intensity in the target TF locus and is excluding the difference in expression. The Delta binding score was calculated by subtracting the score of a TF-gene interaction within one GRN with the score of the interaction within the other GRN.

The delta binding score of the ESC-KC and ESC-LSC comparisons were averaged. If this average was higher than the delta binding score of KC-LSC and LSC-KC the interaction was classified as 'shared epithelial', if the delta binding score was highest in LSC-KC it was classified as 'KC specific', if the delta binding score was highest in KC-LSC it was classified as 'LSC specific'.

Single-cell RNA-seq analysis of the epidermis and the cornea

The raw sequencing data was downloaded from GEO and split it into fastq files using seq2science. Cellranger count was run with Cellranger 6.0.1 to retrieve the matrix, barcodes, and features files necessary for Seurat[114] analysis in R. scRNA-seq cells were selected with a minimum count of 2000, a feature number above 1000, and a mitochondrial percentage below 30 percent. Cell cycle scoring was performed using Seurat CellCycleScoring() feature with the cell cycle genes from Tirosh et al[115]. All cells not in the G1 phase were removed. Leiden Clustering was performed and cell clusters were annotated based on described marker genes.

For the data of the epidermis, cell clusters were selected with high KRT14, KRT5, and low KRT1 and KRT10 expression as basal KCs. From the cornea dataset, cell clusters with high S100A2 with PAX6 and TP63 expression and without CPVL expression were selected as LSCs.
The *in vivo* vs *in vitro* fold change difference plot was generated by loading deseq2 result tables to identify the TF fold changes.

ChIP-seq data analysis

ChIP-seq peaks were called with MACs2 and validated by IDR (see preprocessing). Next for each peak summit reads were counted in 200bp windows across each summit. Values were log-transformed and quantile-based normalized. Peaks were linked to TSS regions in 100kb, using bedtools window [83]. After which they were distance weighted using the

ANANSE distance weighing approach. When genes did not have ChIP-seq peaks within a 100kb window, they got an intensity score of 0.

Disease gene lists were collected from the EyeDiseases database[40], including all disease gene lists of more than 20 genes. A one-sided Man-Whitney U test was performed to test the hypothesis that the disease genes have more TF binding than the other genes in the genome.

Of each significant hit, the top 5 of most bound gene loci were outputted to a list. And the final list was used to generate a dotplot in R (hipseq_intensity_npeak_dotplot.Rmd).

Alternatively, ChIP-seq peaks were mapped to the gene TSS start site using bedtool closest[112]. Next disease genes mapped vs non-mapped disease genes were compared to all genes mapped vs non-mapped. Using a Fisher exact test.

Cornea disease gene list

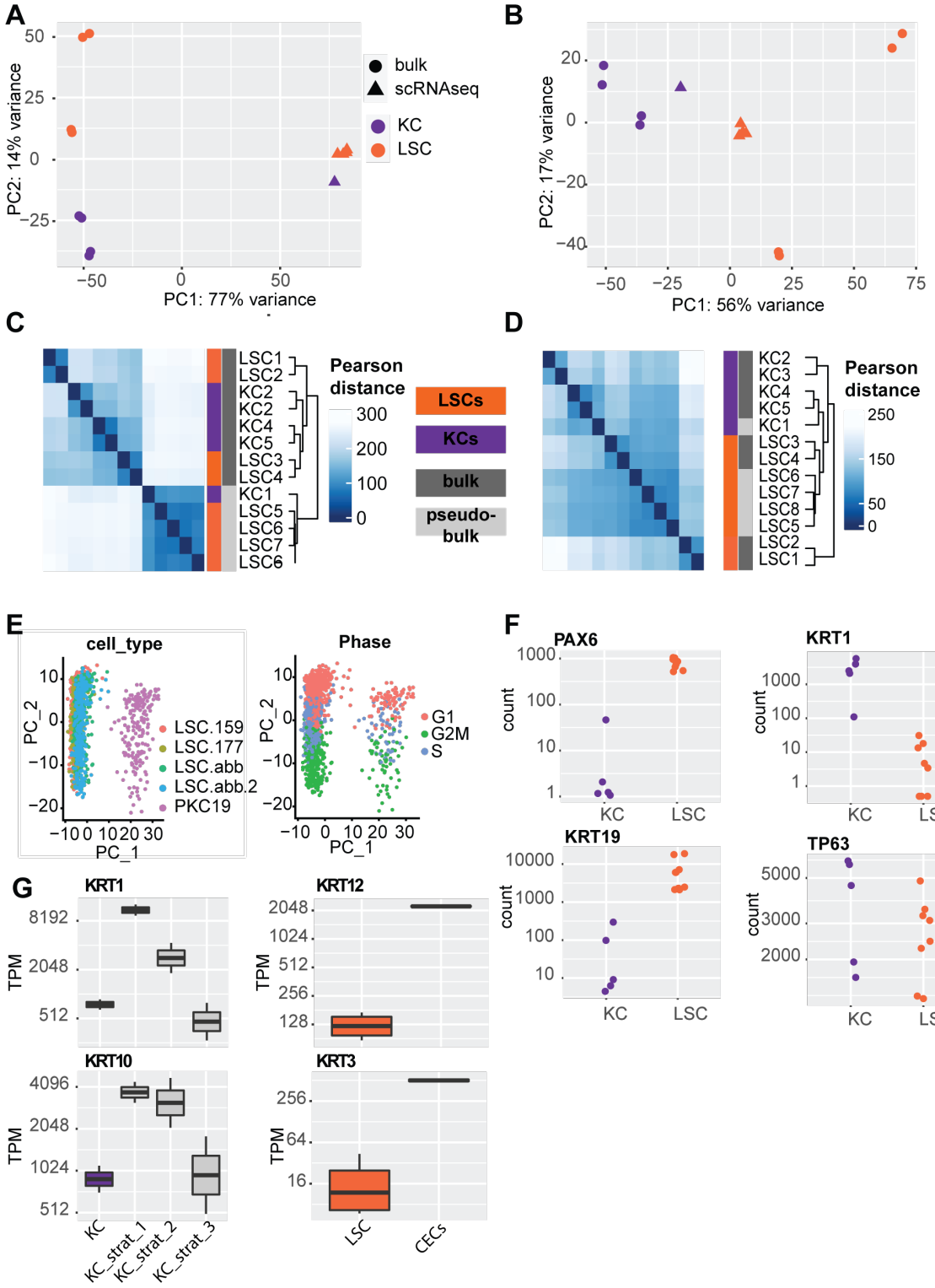
Curated cornea disease list was firstly compiled by retrieving all known genetic disorders affecting the cornea and respective affected genes from “Ophthalmic Genetic Diseases” [116] and then confirmed using available literature in Pubmed (https://pubmed.ncbi.nlm.nih.gov/) and online eye disease (https://gene.vision/) databases. Diseases were grouped as 1) corneal diseases; 2) systemic (or other ocular) disorders with corneal phenotypes, and 3) Diseases with secondary cornea involvement (due to exposure, or unclear involvement). Genes associated with multifactorial disease keratoconus were added based on literature search mainly of published genome wide association (GWAS) and linkage (GWLS) studies[117–119] Curated gene list is available as Supplementary Table 7.

Variant discovery

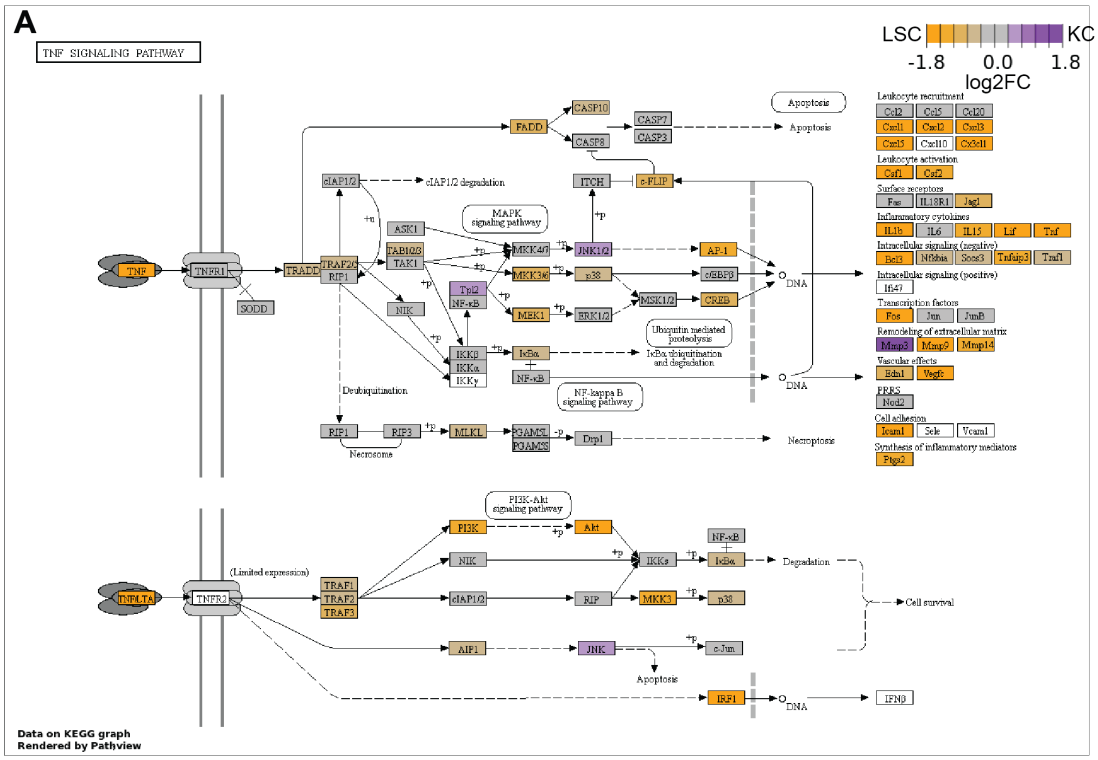
Participants of the 100,000 Genomes Project were identified for our analyses who had at least one of the following HPO terms or daughter terms present: corneal opacity (HP:0007957), corneal scarring (HP:0000559), Opacification of the corneal stroma (HP:0007759), central opacification of the cornea (HP:0011493), band keratopathy (HP:0000585), central posterior corneal opacity (HP:0008511), corneal crystals (HP:0000531), generalized opacification of the cornea (HP:0011494), peripheral opacification of the cornea (HP:0008011), punctate opacification of the cornea

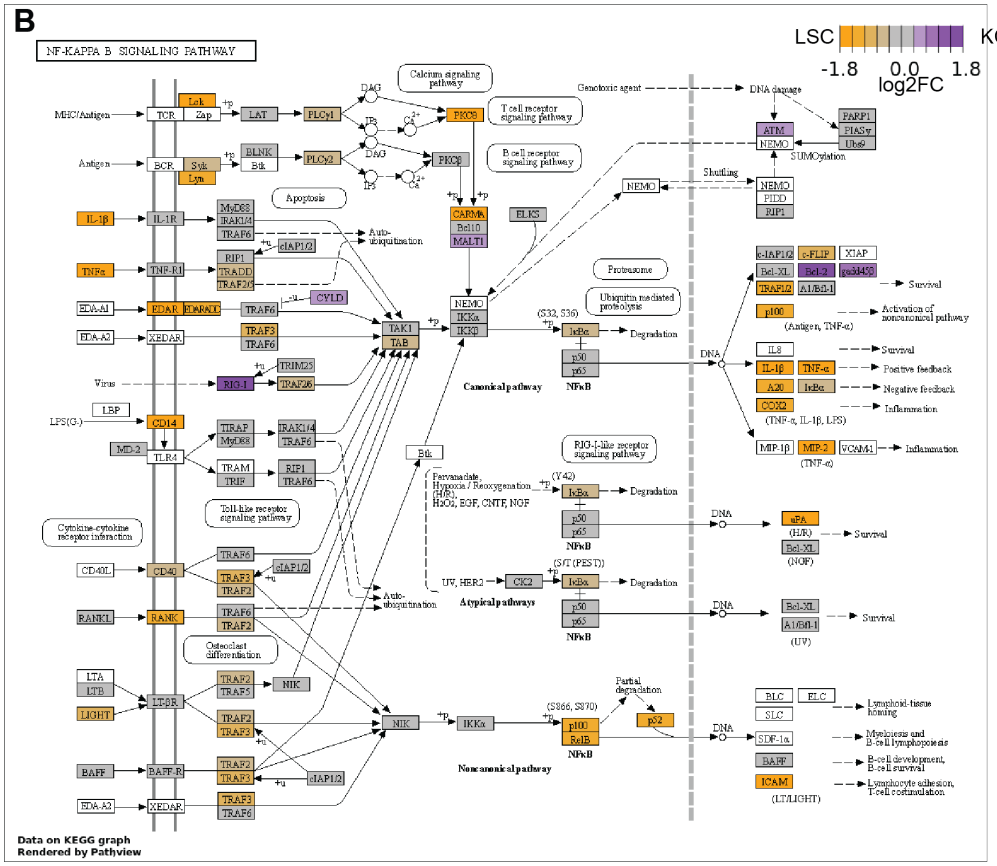
(HP:0007856) and sclerocornea (HP:0000647). A total of 33 probands were identified who remain genetically unsolved. The whole genome sequence data was interrogated for single-nucleotide variants (SNVs) and indels (insertions or deletions), copy number variants (CNVs) and structural variants as previously described (Owen et al 2022). Filtered variants were annotated using Ensembl Variant Effect Predictor (VEP v99) and prioritised identified variants using scores available from CADD, MutationTaster, Provean, Sift, polyphen2, MetaRNN, DANN, fathmm-MKL. Variant nomenclature was assessed using Variant Validator.

Supplementary figures

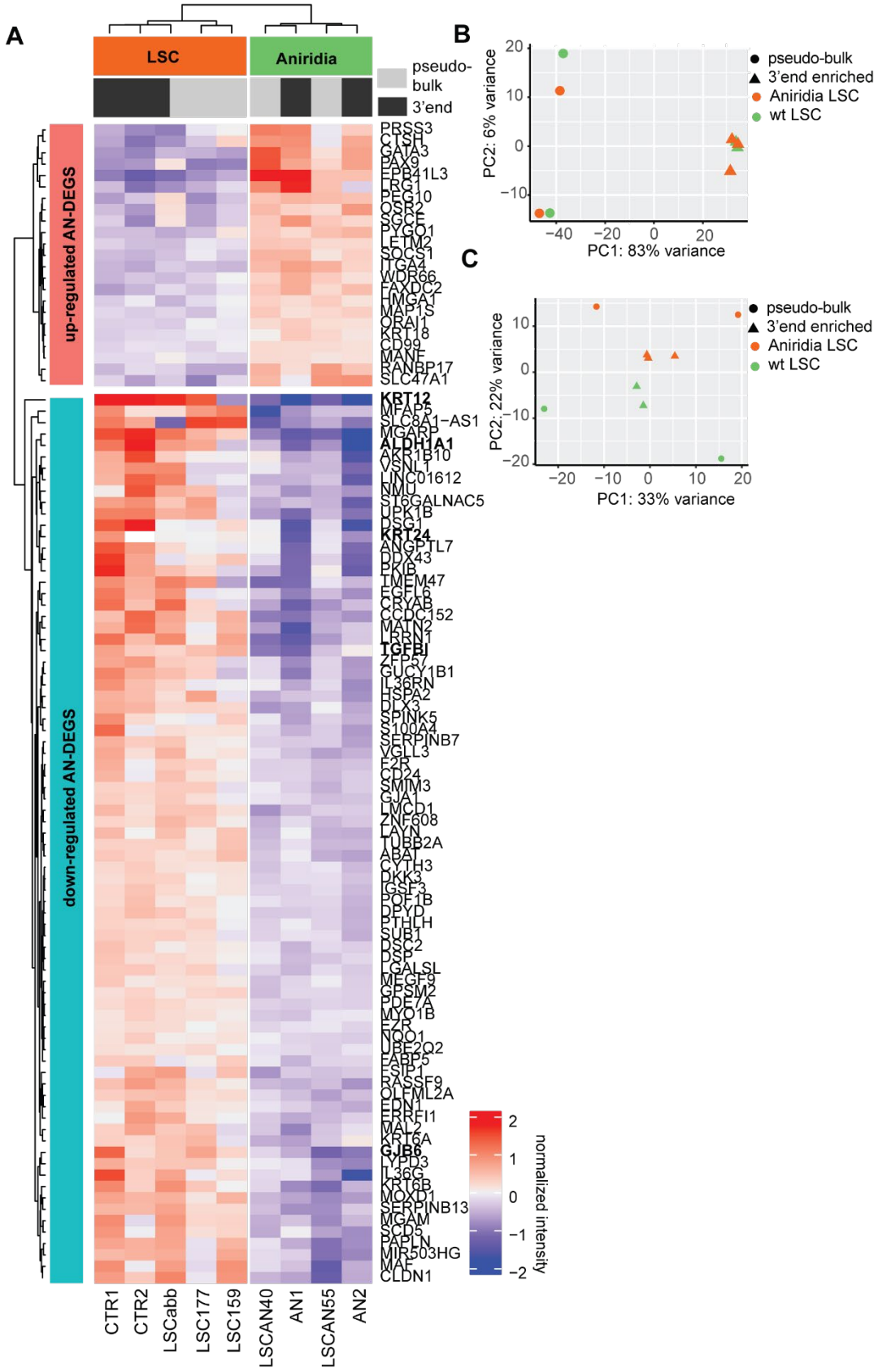


Supplementary figure 1 A) PCA plot of RNAseq samples before batch correction. **B)** PCA plot after batch correction. **C)** Pearson correlation matrix before batch correction. **D)** Pearson correlation matrix after batch correction. **E)** Umap dimensionality reduction of scRNA-seq data, visualizing the samples each cell is from on the left, and the cell cycle state on the right. **F)** Gene count plot for PAX6, KRT1, KRT19, and TP63 in all KC and LSC samples. **G)** TPM gene plots for KRT1, KRT10, KRT12 and KRT3 in KC, and various stratified KC samples, and in LSC and stratified Cornea epithelial cells.





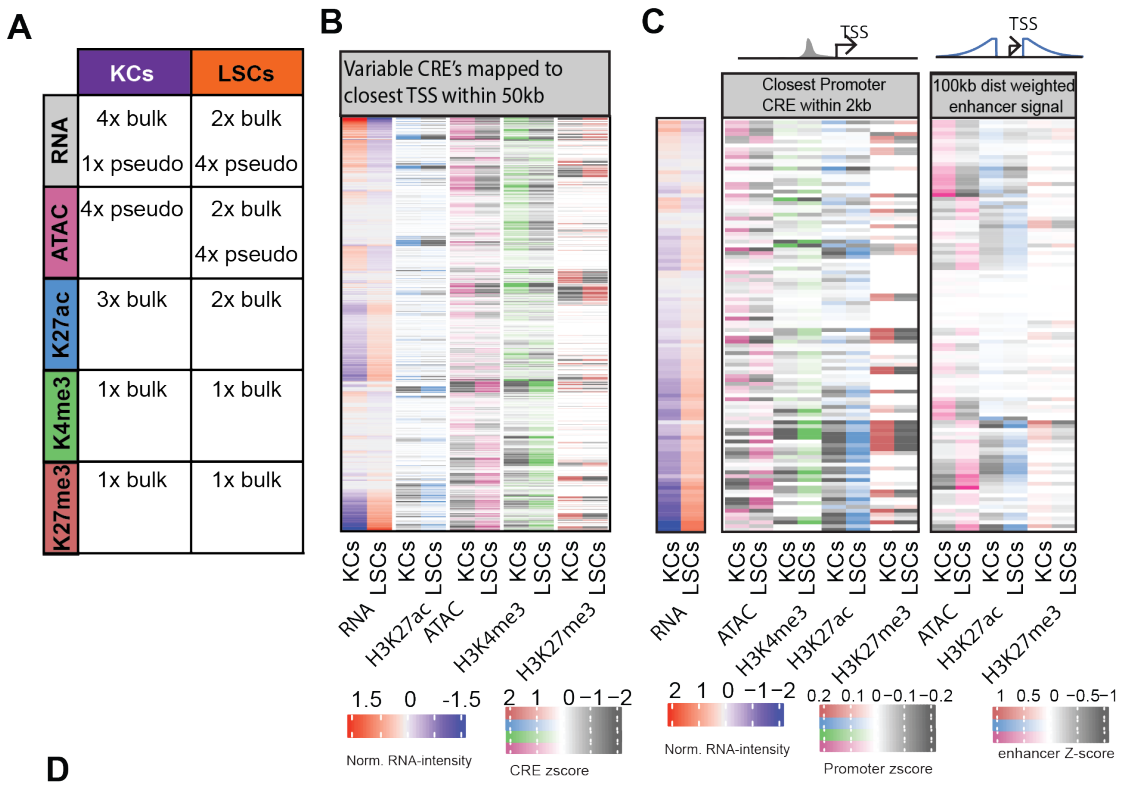
Supplementary figure 2 A) TNF signaling pathway component expression FC differences between KC and LSCs. **B)** NF-KAPPA B signaling pathway component expression FC differences between KC and LSCs.



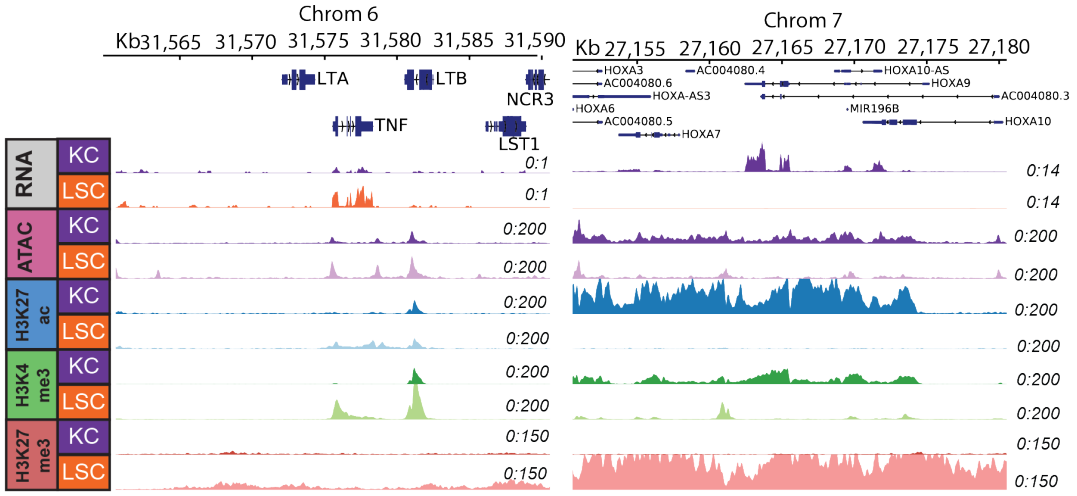
Supplementary figure 3 A) Heatmap of normalized DEG expression between control and aniridia patient LSCs (adjusted pval < 0.05), using k-means clustering with 2 clusters. **B)** PCA plot of RNA-seq samples before batch correction. **C)** PCA plot after batch correction.

A

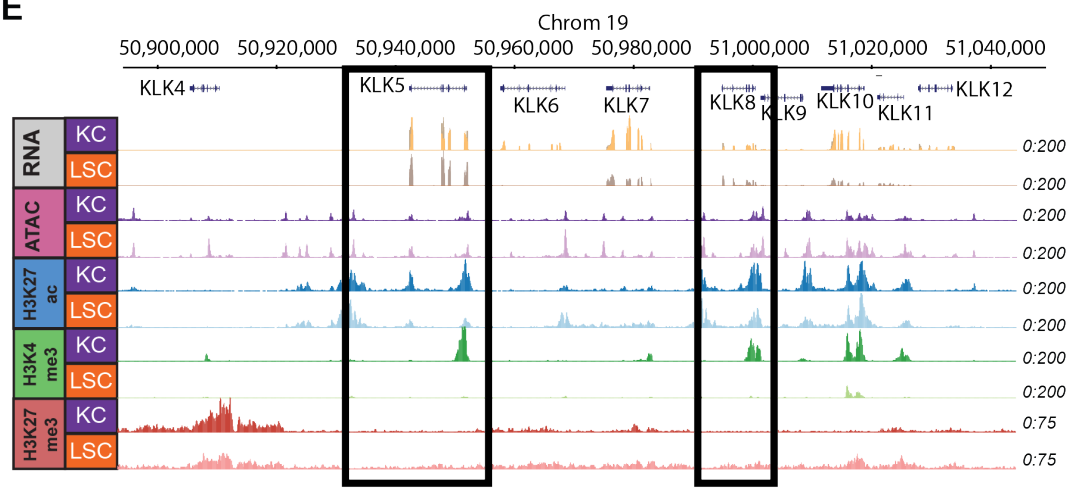
	KCs	LSCs
RNA	4x bulk 1x pseudo	2x bulk 4x pseudo
ATAC	4x pseudo	2x bulk 4x pseudo
K27ac	3x bulk	2x bulk
K4me3	1x bulk	1x bulk
K27me3	1x bulk	1x bulk



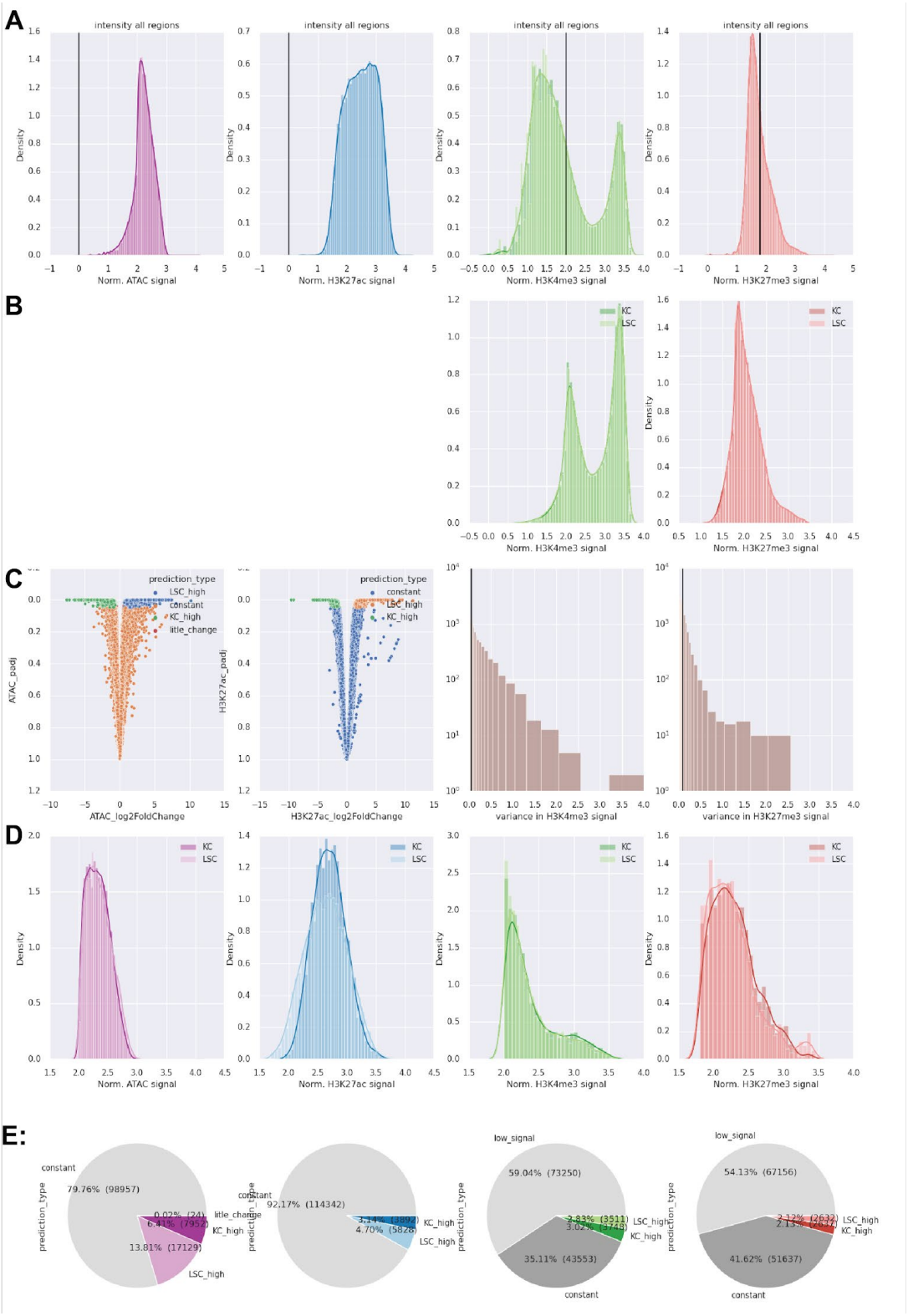
D



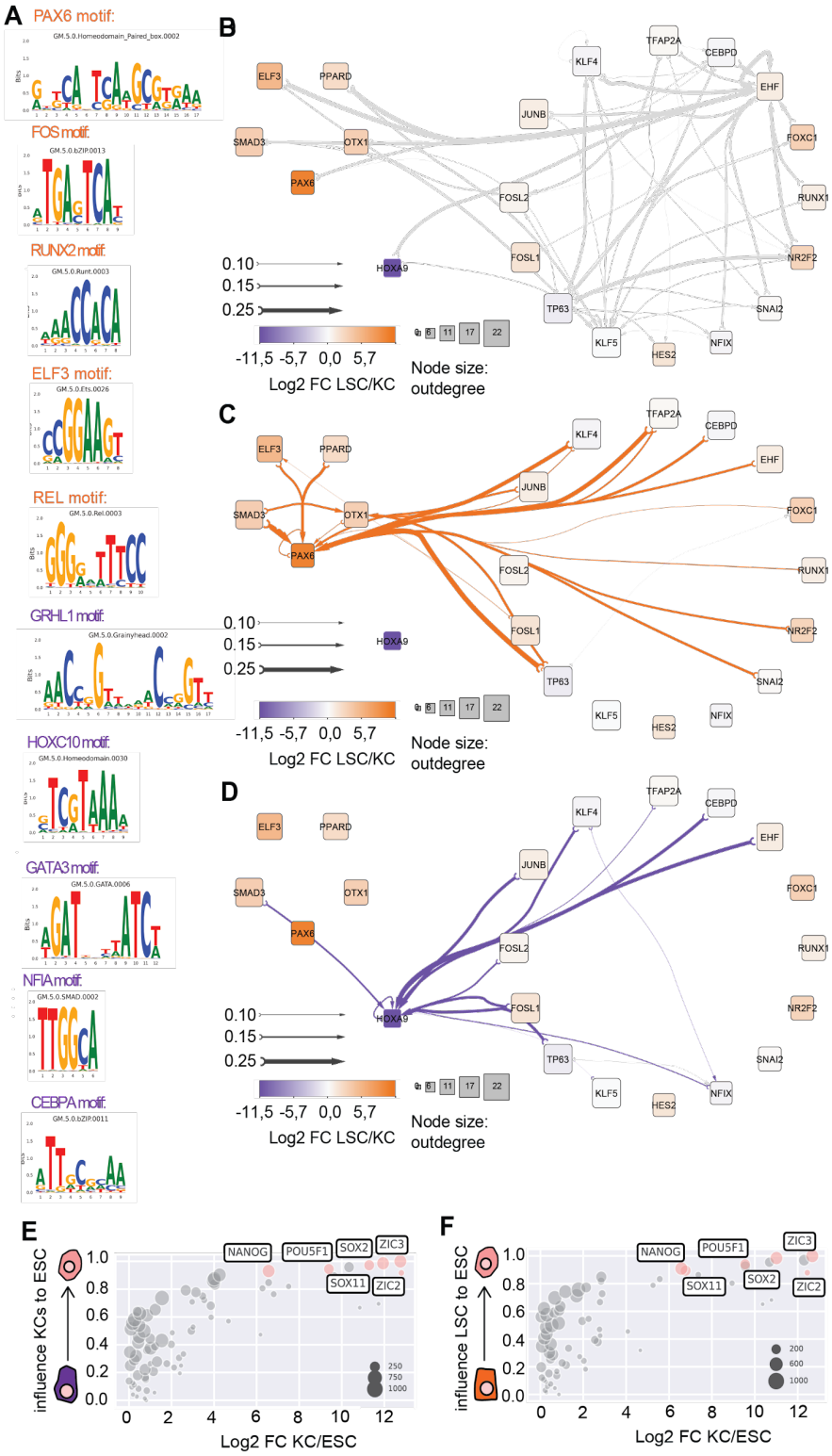
E



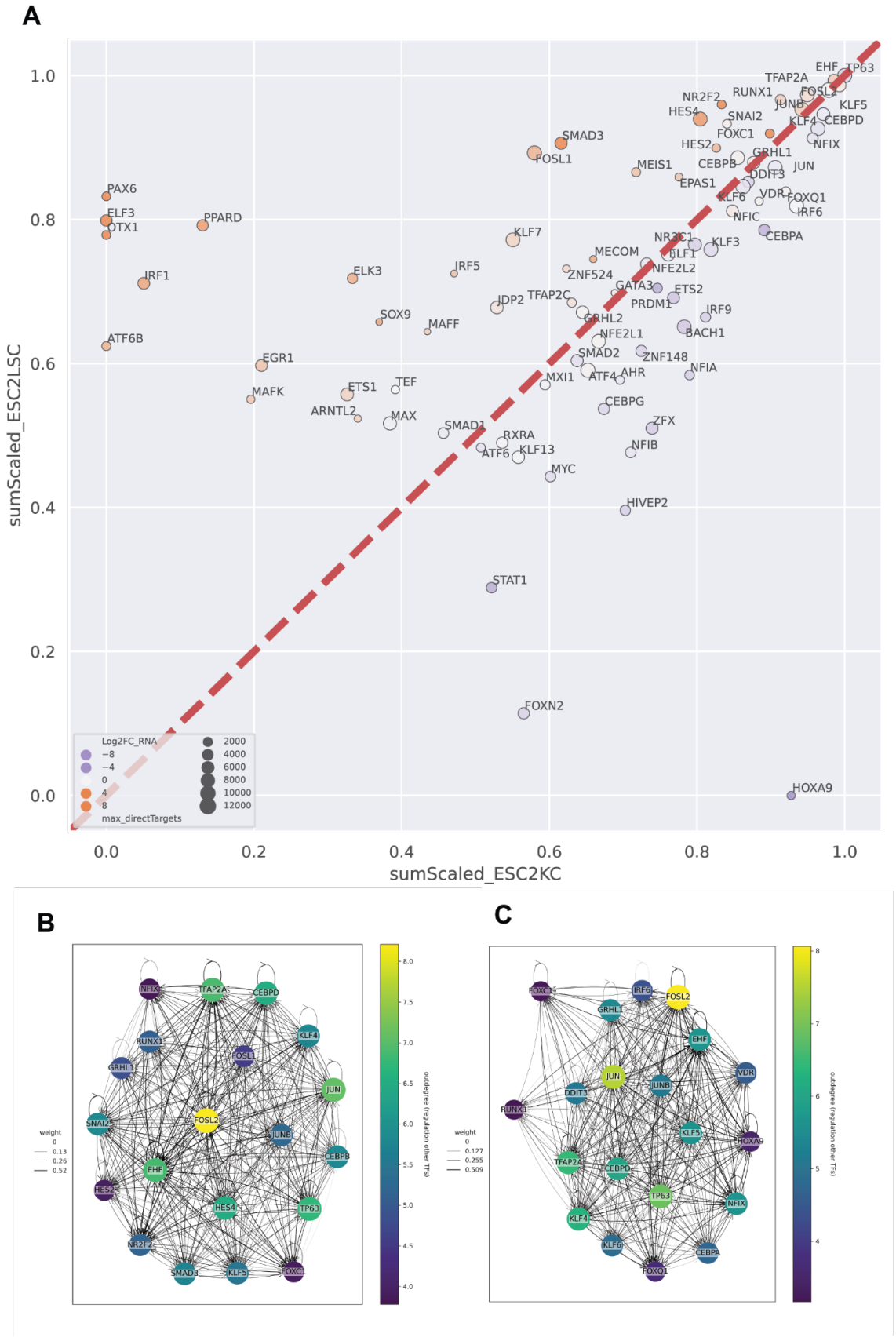
Supplementary figure 4 A) Overview of data types used in our analysis. **B)** Variable CREs mapped to the closest TSS within 50kb. Zscore normalized CRE signal intensities and normalized RNA-seq intensities. **C)** Heatmap of PROGENy TNF and NF-KB target genes and the Z-score of the quantile normalized histone intensity signal of the closest CRE and the distance weighted enhancer signal. The progeny weight attributed to each gene for TNF and NF-KB score is annotated right of the heatmap. **D)** Example loci of TNF and HOXA9.



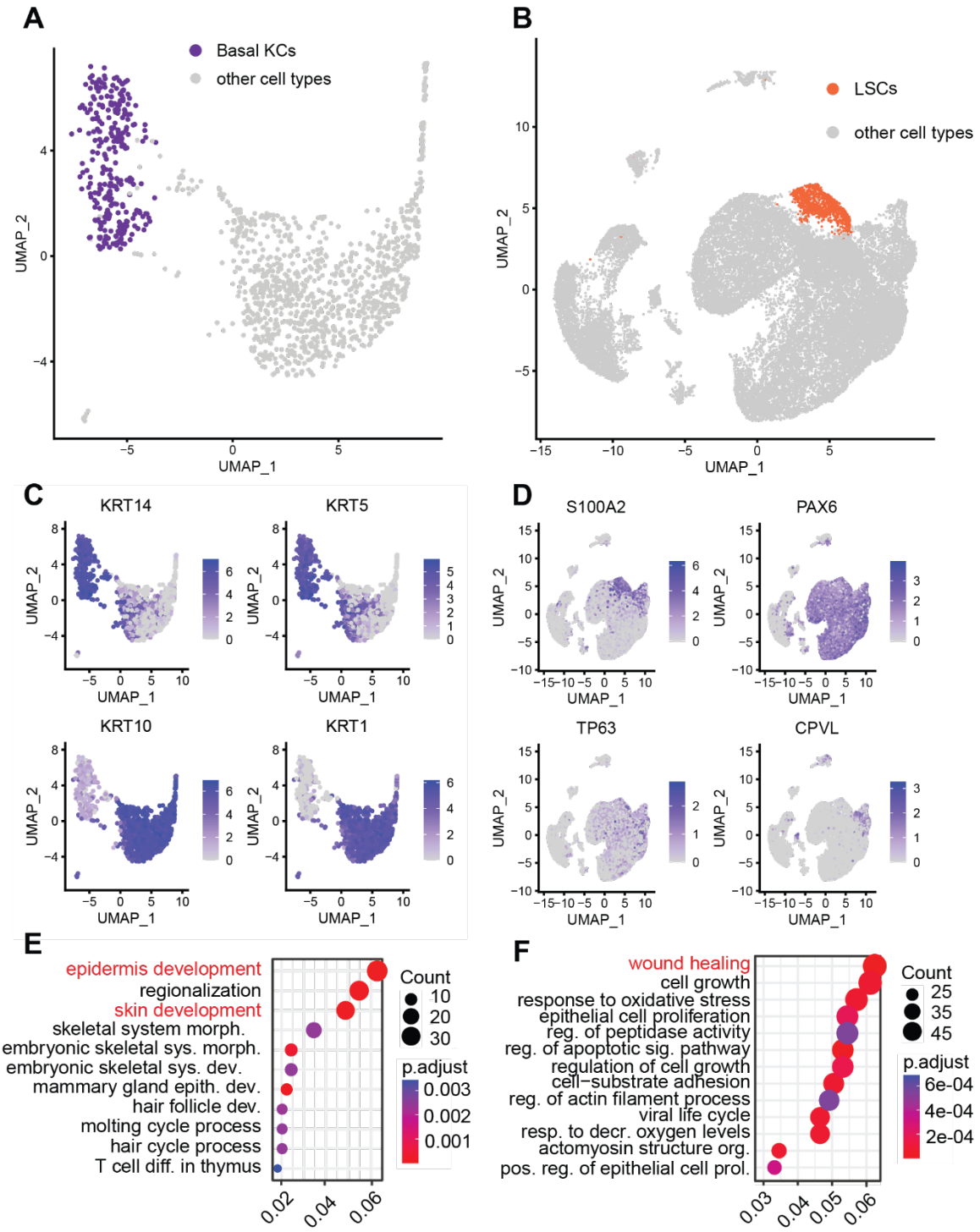
Supplementary figure 5 A) quantile normalized intensity score of all ATAC peaks for the varying histone datasets. Including cutoff value for H3K4me3 and H3K27me3 **B)** Resulting intensity score for H3K4me3 and H3k27me3 regions. **C)** Deseq2 volcano plot of all ATAC & H3K27ac regions. Variance with the variance cutoff for H3K4me3 and H3K27me3. **D)** resulting population of variable regions. **E)** Pie chart of region type distribution.



Supplementary figure 6 A) Enriched motifs linked to the various TFs **B)** General epithelial interactions between TFs. Edge Width corresponds with Ananse binding score predictions. Node color represents RNAseq fold change between LSC and KCs, while node size represents outdegree. **C)** Similar to B but with all the LSC specific interactions. **D)** Similar to B but with all the KC specific interactions. **E)** Ananse KC to ESC influence plot. **F)** Ananse LSC to ESC influence plot.

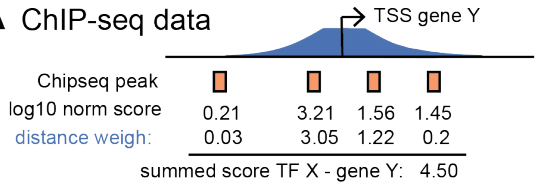


Supplementary figure 7 A) Ananse influence of ESC to KC (x-axis) and ESC to LSC (y-axis), circle size represents a maximum number of target genes in both comparisons. The circle color represents log2FC between LSC/KC. **B)** ESC-LSC top TF interaction network generated by ananse. **C)** ESC-KC top TF interaction network generated by ananse.

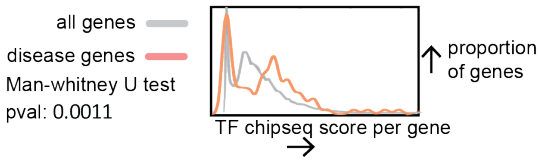


Supplementary figure 8 A) Umap of epidermal scRNAseq dataset of Atwood et al. Basal KC cluster used for validation is highlighted **B)** Umap of scCornea atlas of Collin et al, basal LSC cluster used for validation is highlighted **C)** Marker gene expression used to select the basal KCs cluster **D)** Marker gene expression used to select the basal LSC cluster. **E)** GO-term enrichment of the basal-KC high DEGS enriched vs the human genome as a background and simplified using simplify **F)**GO-term enrichment of the basal-LSCs high DEGS enriched vs the human genome as a background and simplified using simplify

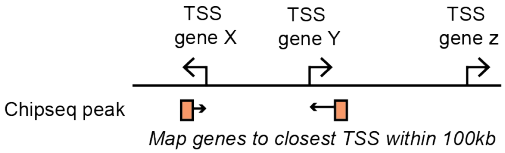
A ChIP-seq data



TF intensity score distributions:



B



Map genes to closest TSS within 100kb

	with peak	no peak
all genes	10045	8975
disease genes	77	31

fisher exact pval: 0.00013

FDR per TF: 0.013

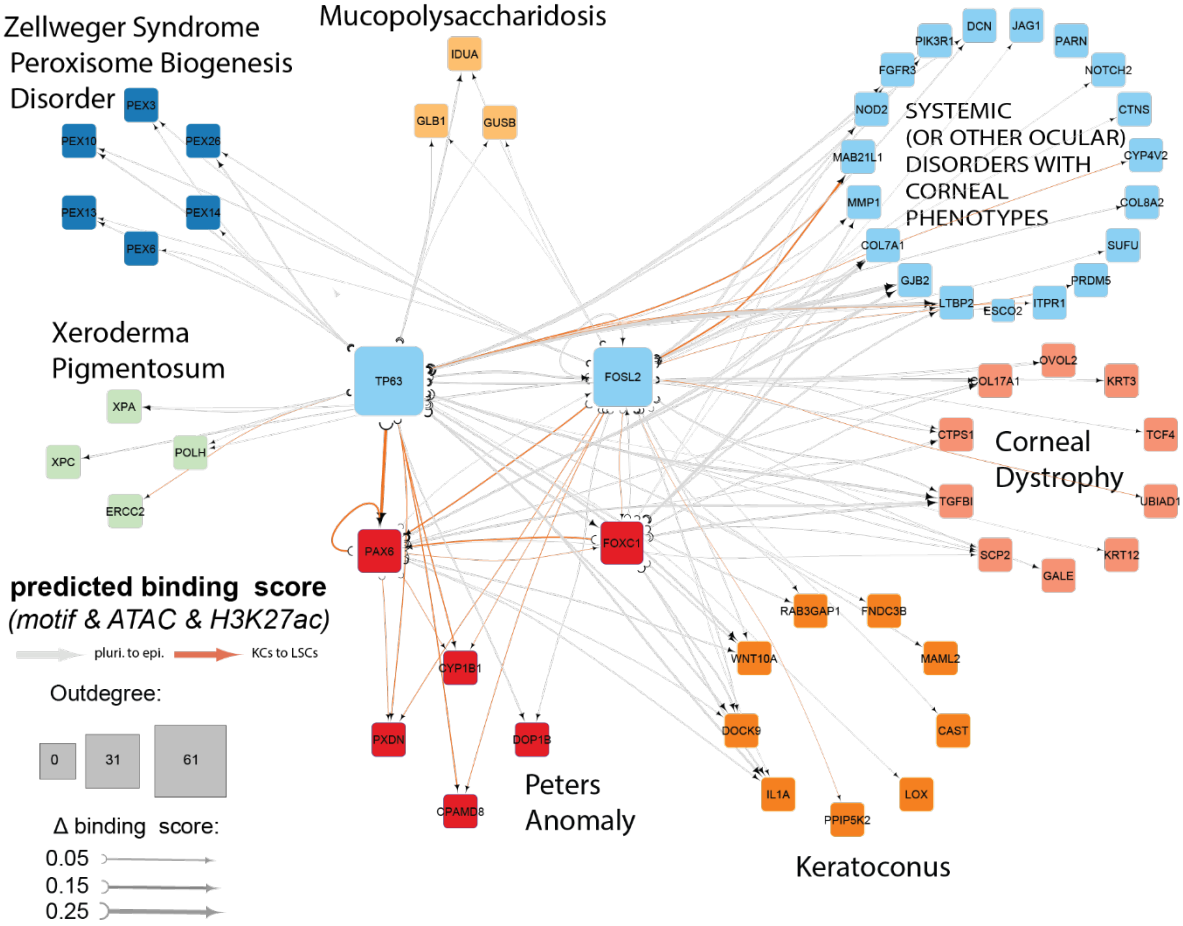
C

TF	disease	FDR
FOXC1	Curated cornea diseases	3.9E-02
P63	Glaucoma	4.9E-03
P63	Fuchs' endothelial corneal dystrophy	4.2E-02
P63	Myopia	4.2E-02
P63	Refractive error	4.2E-02
PAX6	Glaucoma	6.2E-06
PAX6	Refractive error	8.7E-05
PAX6	Curated cornea diseases	1.3E-03
RUNX1	Glaucoma	6.8E-05
RUNX1	Keratoconus	3.6E-02

D

TF	disease	FDR
FOXC1	Curated cornea diseases	2.5E-02
FOXC1	Glaucoma	2.5E-02
P63	Glaucoma	1.0E-02
P63	Fuchs' endothelial corneal dystrophy	1.0E-02
PAX6	Glaucoma	2.3E-06
PAX6	Curated cornea diseases	4.9E-03
PAX6	Refractive error	8.3E-03
RUNX1	Glaucoma	4.4E-04
RUNX1	Keratoconus	3.9E-02

Supplementary figure 9 A) Approach for distance weighing and merging of TF ChIP-seqs per TF. This resulted in a TF-disease gene score distribution that was compared to the distribution of all genes with a one-sided Mann-Whitney U test. **B)** Approach for linking the ChIP-seq peaks to the closest gene TSS, after which enrichment for disease genes was tested with a Fisher exact test. **C)** FDR values of the significant enriched TFs resulting from the ChIP-seq Mann-Whitney U tests. **D)** FDR values of the significant enriched TFs resulting from the Fisher exact test.



Supplementary figure 10 A) TF predicted binding is indicated by the binding score of a TF to its target TF locus, and the cell-type specific regulation is indicated by the binding score difference of the TF at the target TF locus between cell types. When a binding score difference in the KC-LSC comparison is greater than the mean of the difference in ESC-KC and ESC-LSC comparison, this TF regulation of the target TF is annotated as LSC specific (orange arrows). Otherwise, the regulation is annotated as 'common regulation' for both cell types (grey arrows). The degree of binding score difference Δ binding score is indicated by the thickness of the arrows. Outdegree node size represents the number of target genes. Target disease genes are colored based on the disease they were curated to.

Supplementary tables

Supplementary table 1: RNAseq datasets

dataset	name:	datatype	GEO number
KC1	PKC19	pseudobulk	
KC2	PKC19	bulk	GSM2597280
KC3	Dombi23	bulk	GSM2597284
KC4	PKC19	bulk	
KC5	PKC19	bulk	
LSC1	LSC_ouyang	bulk	GSM4728059
LSC2	LSC_ouyang	bulk	GSM4728060
LSC3	LSCaberdam	bulk	
LSC4	LSCaberdam	bulk	
LSC5	LSC159	pseudobulk	
LSC6	LSC177	pseudobulk	
LSC7	LSCaberdam	pseudobulk	
LSC8	LSCaberdam	pseudobulk	
CTR1		3’ bulk	GSM3093490
CTR1		3’ bulk	GSM3093491
AN1		3’ bulk	GSM3093492
AN2		3’ bulk	GSM3093493
Aniridia3	AN40	pseudobulk	
Aniridia4	AN55	pseudobulk	
scCornea	Cornea tissue	Single-cell	GSE155683
scEpidermis	Epidermal tissue	Single-cell	GSE147482
ESC1	H1	3’ bulk	GSM915329
ESC2	H11	3’ bulk	GSM958733
KC_strat_1	PKC19	bulk	GSM2597281
KC_strat_1	PKC19	bulk	GSM2597285
KC_strat_2	PKC19	bulk	GSM2597282
KC_strat_2	PKC19	bulk	GSM2597286
KC_strat_3	PKC19	bulk	GSM2597283
KC_strat_3	PKC19	bulk	GSM2597287
CECs	LSC_ouyang	bulk	GSM2597287

CECs	LSC_ouyang	bulk	GSM4711911

Supplementary table 2: origin ATAC datasets

dataset	**name:**	**datatype**	**GEO number**
KC1	Dombi23	bulk	
KC2	Dombi23	bulk	
KC3	PKC19	pseudobulk	
KC4	PKC19	pseudobulk	
KC5	PKC19	pseudobulk	
KC6	PKC19	pseudobulk	
LSC1	LSCaberdam	pseudobulk	
LSC2	LSCaberdam	pseudobulk	
LSC3	LSC 159	pseudobulk	
LSC4	LSCaberdam	bulk	
LSC5	LSC_Ouyang	Bulk	GSM4728093
LSC6	LSC_Ouyang	Bulk	GSM4728094
ESC1	H1	Bulk	GSM2400260
ESC1	H1	bulk	GSM2400261

Supplementary table 3: origin ChIP-seq datasets

dataset	**Cell Type**	**name line:**	**datatype**	**origin cell line**	**GEO number**
KC1_H3K27ac	KC	HKC1	H3K27ac		GSM1446919
KC2_H3K27ac		Dombi23			
LSC1_H3K27ac	LSC	LSC JQ			
LSC2_H3K27ac		LSC_Ouyang			GSM4728063
LSC3_H3K27ac		LSC_Ouyang			GSM4728064
ESC1 H3K27ac	ESCs	H1			GSM466732
ESC2 H3K27ac		H1			GSM663427
ESC3 H3K27ac		H1			GSM733718
KC1_H3K27me3	KC	Dombi23	H3K27me3		GSM2597292
LSC1_H3K27me3	LSC	LSC_JQ			
LSC2_H3K27me3		LSC_Ouyang			GSM4728069

LSC3_H3K27me3		LSC_Ouyang			GSM4728070
KC1_H3K4me3	KC	Dombi23	H3K4me3		GSM2597288
LSC1_H3K4me3	LSC	LSC JQ			
LSC2_H3K4me3		LSC_Ouyang			GSM4728067
LSC3_H3K4me3		LSC_Ouyang			GSM4728068
LSC1 P63 1	LSC	LSC JQ	P63		
LSC2 P63 2		LSC JQ			
KC1 P63	KCs	HKC1			SRR1528616
KC2 P63		Dombi23			
LSC1 RUN1	LSCs	LSC_Ouyang	RUNX		GSM4728087
LSC2 RUNX1					GSM4728088
LSC1 PAX6			PAX6		GSM4728089
LSC2 PAX6					GSM4728090
LSC1 SMAD3			SMAD3		GSM4728091
LSC2 Smad3					GSM4728092
LSC1 FOXC1			FOXC1		GSM4711922
LSC2 FOXC1					GSM4711923

Supplementary table 4: RNAseq enrichment results in vitro cells

All_enrichments_KC_LSC.xlsx

Supplementary table 5: RNAseq enrichment results in vivo cells

all_invivo_enrichments

Supplementary table 6: CRE analsyis enrichment results

IDCRE_GO_clusters.xlsx

Supplementary table 7: Curated gene list

Cornea_genes_list.xlsx

Supplementary table 8: variant pathogenicity predictions

Included HP (daughter) terms of Corneal opacity		GEL probands
Corneal opacity	HP:0007957	12
Sclerocornea	HP:0000647	8
Corneal scarring	HP:0000559	3
Opacification of the corneal stroma	HP:0007759	3
Central posterior corneal opacity	HP:0008511	2
Generalized opacification of the cornea	HP:0011494	2

Central opacification of the cornea	HP:0011493	1
Band keratopathy	HP:0000585	1
Corneal crystals	HP:0000531	1
Kayser-Fleischer ring	HP:0200032	0
Opacification of the corneal epithelium	HP:0007727	0
Subepithelial corneal opacities	HP:0008039	0
Peripheral opacification of the cornea	HP:0008011	0
Punctate opacification of the cornea	HP:0007856	0
Total included:		33
Excluded terms		
Corneal arcus	HP:0001084	60
Central corneal dystrophy	HP:0007881	4
Mosaic central corneal dystrophy	HP:0100690	0
Nodular corneal dystrophy	HP:0007827	0
Crystalline corneal dystrophy	HP:0007760	1
Crystalline corneal dystrophy	HP:0007760	1
Total excluded:		66

Supplementary table 9: variant pathogenicity predictions

FOSL2(NM_005253.4):c.628C>T. (p.Arg210Cys)		
variant info	GenBank	NM_005253.4
	Uniprot	P15408
	genomicDNA-hg38	2-28412095-C-T
	cDNA	c.628C>T
	AminoAcid	210
	reference AA	R
	Alternative AA	c.628C>T
	gnomAD allele freq	0.0000922
ACMG/AMG guidelines[123] classification	variant of uncertain significance	
prediction tool	**prediction score**	**prediction**
SIFT	0.035	**Damaging**
SIFT4G	0.034	**Damaging**
PrimateAI	0.73	**Pathogenic**
MutationTaster	0.99999	**Disease Causing**
FATHMM-XF	0.90467	**Pathogenic**
List-S2	0.886611	**Damaging**
Meta-RNN	0.5091318	**Damaging**
M-CAP	0.096	Possibly Pathogenic
Poly-phen2	0.096	Probably Damaging

References

1. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
2. Chambers, S. M. & Studer, L. Cell fate plug and play: direct reprogramming and induced pluripotency. *Cell* **145**, 827–830 (2011).
3. Zaret, K. S. Pioneer Transcription Factors Initiating Gene Network Changes. *Annu. Rev. Genet.* **54**, 367–385 (2020).
4. Li, M. & Belmonte, J. C. I. Ground rules of the pluripotency gene regulatory network. *Nat. Rev. Genet.* **18**, 180–191 (2017).
5. Donati, G. & Watt, F. M. Stem cell heterogeneity and plasticity in epithelia. *Cell Stem Cell* **16**, 465–476 (2015).
6. Roberts, N. & Horsley, V. Developing stratified epithelia: lessons from the epidermis and thymus. *Wiley Interdiscip. Rev. Dev. Biol.* **3**, 389–402 (2014).
7. Bashir, H., Seykora, J. T. & Lee, V. Invisible Shield: Review of the Corneal Epithelium as a Barrier to UV Radiation, Pathogens, and Other Environmental Stimuli. *J. Ophthalmic Vis. Res.* **12**, 305–311 (2017).
8. Gonzalez, G., Sasamoto, Y., Ksander, B. R., Frank, M. H. & Frank, N. Y. Limbal Stem Cells: Identity, Developmental Origin and Therapeutic Potential. *Wiley Interdiscip. Rev. Dev. Biol.* **7**, (2018).
9. Bhaduri, A. *et al.* Network Analysis Identifies Mitochondrial Regulation of Epidermal Differentiation by MPZL3 and FDXR. *Dev. Cell* **35**, 444–457 (2015).
10. Rubin, A. J. *et al.* Lineage-specific dynamic and pre-established enhancer–promoter contacts cooperate in terminal differentiation. *Nat. Genet.* **49**, 1522–1528 (2017).
11. Li, L. *et al.* TFAP2C- and p63-Dependent Networks Sequentially Rearrange Chromatin Landscapes to Drive Human Epidermal Lineage Commitment. *Cell Stem Cell* **24**, 271-284.e8 (2019).
12. Soares, E. & Zhou, H. Master regulatory role of p63 in epidermal development and disease. *Cell. Mol. Life Sci.* **75**, 1179–1190 (2018).

13. Sen, G. L. *et al.* ZNF750 is a p63 target gene that induces KLF4 to drive terminal epidermal differentiation. *Dev. Cell* **22**, 669–677 (2012).

14. Qu, J. *et al.* Mutant p63 Affects Epidermal Cell Identity through Rewiring the Enhancer Landscape. *Cell Rep.* **25**, 3490-3503.e4 (2018).

15. Kouwenhoven, E. N. *et al.* Transcription factor p63 bookmarks and regulates dynamic enhancers during epidermal differentiation. *EMBO Rep.* **16**, 863–878 (2015).

16. Qu, J., Yi, G. & Zhou, H. p63 cooperates with CTCF to modulate chromatin architecture in skin keratinocytes. *Epigenetics Chromatin* **12**, 31 (2019).

17. Rinne, T., Hamel, B., van Bokhoven, H. & Brunner, H. G. Pattern of p63 mutations and their phenotypes--update. *Am. J. Med. Genet. A.* **140**, 1396–1406 (2006).

18. Rinne, T., Brunner, H. G. & van Bokhoven, H. p63-associated disorders. *Cell Cycle Georget. Tex* **6**, 262–268 (2007).

19. van Bokhoven, H. & Brunner, H. G. Splitting p63. *Am. J. Hum. Genet.* **71**, 1–13 (2002).

20. Hanson, I. & Van Heyningen, V. Pax6: more than meets the eye. *Trends Genet. TIG* **11**, 268–272 (1995).

21. Ashery-Padan, R. & Gruss, P. Pax6 lights-up the way for eye development. *Curr. Opin. Cell Biol.* **13**, 706–714 (2001).

22. Cvekl, A. & Callaerts, P. PAX6: 25th anniversary and more to learn. *Exp. Eye Res.* **156**, 10–21 (2017).

23. Xie, Q. & Cvekl, A. The orchestration of mammalian tissue morphogenesis through a series of coherent feed-forward loops. *J. Biol. Chem.* **286**, 43259–43271 (2011).

24. Cohen-Tayar, Y. *et al.* Pax6 regulation of Sox9 in the mouse retinal pigmented epithelium controls its timely differentiation and choroid vasculature development. *Dev. Camb. Engl.* **145**, dev163691 (2018).

25. Ypsilanti, A. R. *et al.* Transcriptional network orchestrating regional patterning of cortical progenitors. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2024795118 (2021).

26. van Heyningen, V. & Williamson, K. A. PAX6 in sensory development. *Hum. Mol. Genet.* **11**, 1161–1167 (2002).

27. Shaham, O., Menuchin, Y., Farhy, C. & Ashery-Padan, R. Pax6: a multi-level regulator of ocular development. *Prog. Retin. Eye Res.* **31**, 351–376 (2012).

28. Ouyang, H. *et al.* WNT7A and PAX6 define corneal epithelium homeostasis and pathogenesis. *Nature* **511**, 358–361 (2014).

29. Lagali, N. *et al.* Early phenotypic features of aniridia-associated keratopathy and association with PAX6 coding mutations. *Ocul. Surf.* **18**, 130–140 (2020).

30. Li, M. *et al.* Core transcription regulatory circuitry orchestrates corneal epithelial homeostasis. *Nat. Commun.* **12**, 420 (2021).

31. Lima Cunha, D., Arno, G., Corton, M. & Moosajee, M. The Spectrum of PAX6 Mutations and Genotype-Phenotype Correlations in the Eye. *Genes* **10**, (2019).

32. Kit, V., Cunha, D. L., Hagag, A. M. & Moosajee, M. Longitudinal genotype-phenotype analysis in 86 patients with PAX6-related aniridia. *JCI Insight* **6**, 148406 (2021).

33. Rinne, T. *et al.* A novel translation re-initiation mechanism for the p63 gene revealed by amino-terminal truncating mutations in Rapp-Hodgkin/Hay-Wells-like syndromes. *Hum. Mol. Genet.* **17**, 1968–1977 (2008).

34. Di Iorio, E. *et al.* Limbal stem cell deficiency and ocular phenotype in ectrodactyly-ectodermal dysplasia-clefting syndrome caused by p63 mutations. *Ophthalmology* **119**, 74–83 (2012).

35. Tümer, Z. & Bach-Holm, D. Axenfeld–Rieger syndrome and spectrum of PITX2 and FOXC1 mutations. *Eur. J. Hum. Genet.* **17**, 1527–1539 (2009).

36. Seo, S. *et al.* Forkhead box transcription factor FoxC1 preserves corneal transparency by regulating vascular growth. *Proc. Natl. Acad. Sci.* **109**, 2015–2020 (2012).

37. Li, M. *et al.* Loss of FOXC1 contributes to the corneal epithelial fate switch and pathogenesis. *Signal Transduct. Target. Ther.* **6**, 1–11 (2021).

38. Bonnet, C., Roberts, J. S. & Deng, S. X. Limbal stem cell diseases. *Exp. Eye Res.* **205**, 108437 (2021).

39. Kitazawa, K. *et al.* PAX6 regulates human corneal epithelium cell identity. *Exp. Eye Res.* **154**, 30–38 (2017).

40. van den Bogaard, E. H. *et al.* Rho kinase inhibitor Y-27632 prolongs the life span of adult human keratinocytes, enhances skin equivalent development, and facilitates lentiviral transduction. *Tissue Eng. Part A* **18**, 1827–1836 (2012).

41. Lužnik, Z. *et al.* Preservation of Ocular Epithelial Limbal Stem Cells: The New Frontier in Regenerative Medicine. *Adv. Exp. Med. Biol.* **951**, 179–189 (2016).

42. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–D261 (2004).

43. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innov. N. Y. N* **2**, 100141 (2021).

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

45. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).

46. Schubert, M. *et al.* Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* **9**, 20 (2018).

47. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

48. Latta, L. *et al.* Expression of retinoic acid signaling components ADH7 and ALDH1A1 is reduced in aniridia limbal epithelial cells and a siRNA primary cell based aniridia model. *Exp. Eye Res.* **179**, 8–17 (2019).

49. Roux, L. N. *et al.* Modeling of Aniridia-Related Keratopathy by CRISPR/Cas9 Genome Editing of Human Limbal Epithelial Cells and Rescue by Recombinant PAX6 Protein. *STEM CELLS* **36**, 1421–1429 (2018).

50. GimmeMotifs: an analysis framework for transcription factor motif analysis | bioRxiv. https://www.biorxiv.org/content/10.1101/474403v1.

51. Xu, Q. *et al.* ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination. *Nucleic Acids Res.* **49**, 7966–7985 (2021).

52. Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010).

53. Wang, S. *et al.* Single cell transcriptomics of human epidermis identifies basal stem cell transition states. *Nat. Commun.* **11**, 4239 (2020).

54. Collin, J. *et al.* A single cell atlas of human cornea that defines its development, limbal progenitor cells and their interactions with the immune cells. *Ocul. Surf.* **21**, 279–298 (2021).

55. Yuan, J. *et al.* EyeDiseases: an integrated resource for dedicating to genetic variants, gene expression and epigenetic factors of human eye diseases. *NAR Genomics Bioinforma.* **3**, lqab050 (2021).

56. Hanson, I. M. *et al.* Mutations at the PAX6 locus are found in heterogeneous anterior segment malformations including Peters' anomaly. *Nat. Genet.* **6**, 168–173 (1994).

57. Cella, W. *et al.* Structural assessment of PITX2, FOXC1, CYP1B1, and GJA1 genes in patients with Axenfeld-Rieger syndrome with developmental glaucoma. *Invest. Ophthalmol. Vis. Sci.* **47**, 1803–1809 (2006).

58. Honkanen, R. A. *et al.* A family with Axenfeld–Rieger syndrome and Peters Anomaly caused by a point mutation (Phe112Ser) in the FOXC1 gene. *Am. J. Ophthalmol.* **135**, 368–375 (2003).

59. Klein, R. H. *et al.* GRHL3 binding and enhancers rearrange as epidermal keratinocytes transition between functional states. *PLOS Genet.* **13**, e1006745 (2017).

60. Simeone, A. Otx1 and Otx2 in the development and evolution of the mammalian brain. *EMBO J.* **17**, 6790–6798 (1998).

61. Huang, B. *et al.* OTX1 regulates cell cycle progression of neural progenitors in the developing cerebral cortex. *J. Biol. Chem.* **293**, 2137–2148 (2018).

62. Martinez-Morales, J. R., Signore, M., Acampora, D., Simeone, A. & Bovolenta, P. Otx genes are required for tissue specification in the developing eye. *Development* **128**, 2019–2030 (2001).

63. Samuel, A., Housset, M., Fant, B. & Lamonerie, T. Otx2 ChIP-seq Reveals Unique and Redundant Functions in the Mature Mouse Retina. *PLOS ONE* **9**, e89110 (2014).

64. Karin, M., Liu, Z. g & Zandi, E. AP-1 function and regulation. *Curr. Opin. Cell Biol.* **9**, 240–246 (1997).

65. Eckert, R. L. *et al.* AP1 transcription factors in epidermal differentiation and skin cancer. *J. Skin Cancer* **2013**, 537028 (2013).

66. Yoshida, N., Yoshida, S., Araie, M., Handa, H. & Nabeshima, Y. Ets family transcription factor ESE-1 is expressed in corneal epithelial cells and is involved in their differentiation. *Mech. Dev.* **97**, 27–34 (2000).

67. Borrelli, S. *et al.* Reciprocal regulation of p63 by C/EBP delta in human keratinocytes. *BMC Mol. Biol.* **8**, 85 (2007).

68. Candi, E. *et al.* p63 is upstream of IKK alpha in epidermal development. *J. Cell Sci.* **119**, 4617–4622 (2006).

69. Chikh, A. *et al.* Expression of GATA-3 in epidermis and hair follicle: relationship to p63. *Biochem. Biophys. Res. Commun.* **361**, 1–6 (2007).

70. Zeitvogel, J. *et al.* GATA3 regulates FLG and FLG2 expression in human primary keratinocytes. *Sci. Rep.* **7**, 11847 (2017).

71. Rinn, J. L. *et al.* A dermal HOX transcriptional program regulates site-specific epidermal fate. *Genes Dev.* **22**, 303–307 (2008).

72. Gehring, W. J. The animal body plan, the prototypic body segment, and eye evolution. *Evol. Dev.* **14**, 34–46 (2012).

73. Kitazawa, K. *et al.* OVOL2 Maintains the Transcriptional Program of Human Corneal Epithelium by Suppressing Epithelial-to-Mesenchymal Transition. *Cell Rep.* **15**, 1359–1368 (2016).

74. Menzel-Severing, J. *et al.* Transcription factor profiling identifies Sox9 as regulator of proliferation and differentiation in corneal epithelial stem/progenitor cells. *Sci. Rep.* **8**, 10268 (2018).

75. McConnell, B. B., Ghaleb, A. M., Nandan, M. O. & Yang, V. W. The diverse functions of Krüppel-like factors 4 and 5 in epithelial biology and pathobiology. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **29**, 549–557 (2007).

76. Kenchegowda, D., Harvey, S. A. K., Swamynathan, S., Lathrop, K. L. & Swamynathan, S. K. Critical Role of Klf5 in Regulating Gene Expression during Post-Eyelid Opening Maturation of Mouse Corneas. *PLoS ONE* **7**, e44771 (2012).

77. Stephens, D. N. *et al.* The Ets transcription factor EHF as a regulator of cornea epithelial cell identity. *J. Biol. Chem.* **288**, 34304–34324 (2013).

78. Tiwari, A., Loughner, C. L., Swamynathan, S. & Swamynathan, S. K. KLF4 Plays an Essential Role in Corneal Epithelial Homeostasis by Promoting Epithelial Cell Fate and Suppressing Epithelial–Mesenchymal Transition. *Invest. Ophthalmol. Vis. Sci.* **58**, 2785–2795 (2017).

79. Cieślar-Pobuda, A. *et al.* Human induced pluripotent stem cell differentiation and direct transdifferentiation into corneal epithelial-like cells. *Oncotarget* **7**, 42314–42329 (2016).

80. Li, G. *et al.* Transcription Factor PAX6 (Paired Box 6) Controls Limbal Stem Cell Lineage in Development and Disease. *J. Biol. Chem.* **290**, 20448–20454 (2015).

81. Swamynathan, S. K. Ocular Surface Development and Gene Expression. *J. Ophthalmol.* **2013**, 103947 (2013).

82. Webster, J. D. & Vucic, D. The Balance of TNF Mediated Pathways Regulates Inflammatory Cell Death Signaling in Healthy and Diseased Tissues. *Front. Cell Dev. Biol.* **8**, (2020).

83. Kimura, K., Teranishi, S., Fukuda, K., Kawamoto, K. & Nishida, T. Delayed disruption of barrier function in cultured human corneal epithelial cells induced by tumor necrosis factor-alpha in a manner dependent on NF-kappaB. *Invest. Ophthalmol. Vis. Sci.* **49**, 565–571 (2008).

84. Kimura, K. *et al.* Protection of Human Corneal Epithelial Cells From TNF-α–Induced Disruption of Barrier Function by Rebamipide. *Invest. Ophthalmol. Vis. Sci.* **54**, 2752–2760 (2013).

85. Wang, X., Shan, X. & Gregory-Evans, C. Y. A mouse model of aniridia reveals the in vivo downstream targets of Pax6 driving iris and ciliary body development in the eye. *Biochim. Biophys. Acta Mol. Basis Dis.* **1863**, 60–67 (2017).

86. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinforma. Oxf. Engl.* **32**, 2847–2849 (2016).

87. Shetty, A. *et al.* A systematic comparison of FOSL1, FOSL2 and BATF-mediated transcriptional regulation during early human Th17 differentiation. *Nucleic Acids Res.* **50**, 4938–4958 (2022).

88. Shinde, V. *et al.* RNA sequencing of corneas from two keratoconus patient groups identifies potential biomarkers and decreased NRF2-antioxidant responses. *Sci. Rep.* **10**, 9907 (2020).

89. Latta, L. *et al.* Human aniridia limbal epithelial cells lack expression of keratins K3 and K12. *Exp. Eye Res.* **167**, 100–109 (2018).

90. Schlötzer-Schrehardt, U. *et al.* Dysfunction of the limbal epithelial stem cell niche in aniridia-associated keratopathy. *Ocul. Surf.* **21**, 160–173 (2021).

91. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).

92. Sande, M. van der *et al.* seq2science. (2022) doi:10.5281/zenodo.6034404.

93. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

94. Heeringen, S. J. van. genomepy: download genomes the easy way. *J. Open Source Softw.* **2**, 320 (2017).

95. Choudhary, S. pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive. Preprint at https://doi.org/10.12688/f1000research.18676.1 (2019).

96. Crusoe, M. R. *et al.* The khmer software package: enabling efficient nucleotide sequence analysis. Preprint at https://doi.org/10.12688/f1000research.6924.1 (2015).

97. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).

98. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

99. Picard Tools - By Broad Institute. http://broadinstitute.github.io/picard/.

100. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

101. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).

102. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

103. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).

104. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

105. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).

106. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

107. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

108. Stephens, M. False discovery rates: a new deal. *Biostatistics* **18**, 275–294 (2017).

109. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize Implements and enhances circular visualization in R. *Bioinforma. Oxf. Engl.* **30**, 2811–2812 (2014).

110. Pagès, H., Carlson, M., Falcon, S. & Li, N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. (2022) doi:10.18129/B9.bioc.AnnotationDbi.

111. Pathview: an R/Bioconductor package for pathway-based data integration and visualization | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/29/14/1830/232698.

112. BEDTools: a flexible suite of utilities for comparing genomic features | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/26/6/841/244688.

113. Chen, X., Miragaia, R. J., Natarajan, K. N. & Teichmann, S. A. A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.* **9**, 5345 (2018).

114. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

115. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).

116. Preface. in *Ophthalmic Genetic Diseases* (ed. Couser, N. L.) xi (Elsevier, 2019). doi:10.1016/B978-0-323-65414-2.05001-7.

117. Khawaja, A. P. *et al.* Genetic Variants Associated With Corneal Biomechanical Properties and Potentially Conferring Susceptibility to Keratoconus in a Genome-Wide Association Study. *JAMA Ophthalmol.* **137**, 1005–1012 (2019).

118. Bykhovskaya, Y. & Rabinowitz, Y. S. Update on the genetics of keratoconus. *Exp. Eye Res.* **202**, 108398 (2021).

119. Hardcastle, A. J. *et al.* A multi-ethnic genome-wide association study implicates collagen matrix integrity and cell differentiation pathways in keratoconus. *Commun. Biol.* **4**, 266 (2021).