nature genetics

Article

https://doi.org/10.1038/s41588-023-01445-4

Integrated single-cell chromatin and transcriptomic analyses of human scalp identify gene-regulatory programs and critical cell types for hair and skin diseases

Received: 1 September 2022

Accepted: 17 June 2023

Published online: 27 July 2023

Check for updates

Benjamin Ober-Reynolds[1], Chen Wang[2,3,4], Justin M. Ko[2], Eon J. Rios[2,3], Sumaira Z. Aasi[2], Mark M. Davis[4,5,6], Anthony E. Oro[2,7] & William J. Greenleaf[1,8,9]

Genome-wide association studies have identified many loci associated with hair and skin disease, but identification of causal variants requires deciphering of gene-regulatory networks in relevant cell types. We generated matched single-cell chromatin profiles and transcriptomes from scalp tissue from healthy controls and patients with alopecia areata, identifying diverse cell types of the hair follicle niche. By interrogating these datasets at multiple levels of cellular resolution, we infer 50–100% more enhancer–gene links than previous approaches and show that aggregate enhancer accessibility for highly regulated genes predicts expression. We use these gene-regulatory maps to prioritize cell types, genes and causal variants implicated in the pathobiology of androgenetic alopecia (AGA), eczema and other complex traits. AGA genome-wide association studies signals are enriched in dermal papilla regulatory regions, supporting the role of these cells as drivers of AGA pathogenesis. Finally, we train machine learning models to nominate single-nucleotide polymorphisms that affect gene expression through disruption of transcription factor binding, predicting candidate functional single-nucleotide polymorphism for AGA and eczema.

Skin consists of a community of cell types from diverse developmental origins that perform coordinated functions underlying tissue homeostasis. For example, skin contains hair follicles that progress through cycles of growth (anagen), regression (catagen) and resting (telogen), guided by paracrine signals from their surrounding stromal and immune niche[1–4]. Disruption of these cellular communities causes human skin and hair diseases such as alopecia areata, when normal hair follicle cycling is prevented by autoreactive T cells, or androgenetic alopecia (AGA), where hair follicles gradually miniaturize as a result of a poorly understood interplay of genetic and hormonal factors.

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [2]Department of Dermatology, School of Medicine, Stanford University, Stanford, CA, USA. [3]Division of Dermatology, Department of Medicine, Santa Clara Valley Medical Center, San Jose, CA, USA. [4]Institute of Immunity, Transplantation and Infection, School of Medicine, Stanford University, Stanford, CA, USA. [5]Department of Microbiology and Immunology, School of Medicine, Stanford University, Stanford, CA, USA. [6]Howard Hughes Medical Institute, School of Medicine, Stanford University, Stanford, CA, USA. [7]Program in Epithelial Biology, Stanford University School of Medicine, Stanford, CA, USA. [8]Department of Applied Physics, Stanford University, Stanford, CA, USA. [9]Chan Zuckerberg Biohub, San Francisco, CA, USA. e-mail: wjg@stanford.edu

Understanding the pathobiology of these and other skin and hair diseases therefore depends on approaches capable of identifying perturbations across multiple candidate cell types and states.

While genome-wide association studies (GWAS) have identified numerous distinct genomic loci associated with complex hair- and skin-related disorders[5–9], identification of specific causal variants and interpretation of their molecular function remains challenging. Most GWAS disease risk variants reside in noncoding genomic regions, and many are predicted to exert their effects through disruption of cell-type-specific *cis*-regulatory elements (CREs)[10] that may not exert their effects on the nearest gene. Identifying causal variants and interpreting their function thus requires analysis of gene-regulatory networks in disease-relevant cell types.

Although single-cell genomics—primarily single-cell RNA sequencing (scRNA-seq)—has enabled identification and characterization of the diverse cell types in human skin in healthy and disease contexts[11–17], many of these studies are limited by incomplete cell-sampling approaches. While scRNA-seq assays the transcriptional state of cell types within a tissue the underlying CREs are not observed, precluding deeper insights into how noncoding CRE variation influences disease phenotypes.

In this study we characterize gene-regulatory networks in healthy and diseased skin and hair follicles using paired, single-cell atlases of gene expression and chromatin accessibility in human scalp. We identify enhancer–gene linkages at multiple scales of cellular resolution, yielding 50–100% more enhancer–gene links than previous multiomic studies. We identify a subset of cell lineage genes with a disproportionately large number of CREs, and show that expression of these highly regulated genes (HRGs) is driven by distinct combinations of enhancer modules. We predict gene targets of transcription factors (TFs) driving keratinocyte differentiation trajectories. We integrate our data with skin and hair disease GWAS loci to identify critical cell types and putative target genes. AGA GWAS signals were strongly and specifically enriched in dermal papilla (DP) open-chromatin regions and linked to target genes enriched for roles in WNT signaling. Finally, we train machine learning models to nominate potential causal variants based on their predicted effects on cell-type-specific chromatin accessibility, identifying 47, 19 and 19 prioritized SNPs for AGA, eczema and hair color, respectively.

Results

A paired transcriptomic and epigenetic atlas of human scalp

We created paired, single-cell transcriptomic and chromatin accessibility atlases from primary human scalp tissue. We obtained tissue from three sources: punch biopsies from healthy control volunteers (C_PB, $n = 3$), patients with alopecia areata ($n = 5$) and discarded normal peripheral surgical tissue (C_SD, $n = 7$) (Fig. 1a,b and Supplementary Table 1). We dissociated tissue and prepared scRNA-seq and single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) libraries using the 10X Genomics Chromium platform. We obtained 54,288 single-cell transcriptomes and 45,896 single-cell chromatin accessibility profiles following quality control and filtering (Extended Data Fig. 1a–c,f,g), identifying 22 cell clusters in both scRNA-seq and scATAC-seq datasets (Fig. 1c,d).

To annotate clusters we examined the gene expression and gene activity scores of known marker genes (Fig. 1e,f and Extended Data Fig. 2a,b)[18,19]. To better resolve the heterogeneity of broad cell groupings we subclustered five major cell classes (keratinocytes, T lymphocytes, myeloid lineage cells, fibroblasts and endothelial cells) in both scRNA- and scATAC-seq datasets (Extended Data Fig. 3a–c). We identified 42 scRNA-seq and 38 scATAC-seq 'high-resolution clusters', revealing rare cellular subtypes including DP cells (*HHIP*, *WNT5A* and *PTCH1*)[20], eccrine gland cells (*AQP5* and *KRT19*)[21] and TREM2-positive macrophages (*TREM2* and OSM)[22]. Both low- and high-resolution cluster profiles were highly reproducible using subsampled datasets (Extended Data Fig. 4).

Using the high-resolution scATAC-seq clusters we identified 589,294 'peaks' of open chromatin corresponding to CREs[18]. We identified 182,498 differentially accessible peaks between the broad scATAC clusters (Wilcoxon false discovery rate (FDR) ≤ 0.1, $\log_2$ (fold change (FC)) ≥ 0.5; Extended Data Fig. 2c). These cluster-specific peaks were enriched for lineage-determining TF motifs, such as RUNX and ETS factors in T lymphocytes[23,24], SPI (PU.1) factors in myeloid lineage cells[25], TP63 in keratinocytes[26] and MITF in melanocytes[27] (Fig. 1g).

All scRNA and scATAC clusters were composed of cells spanning the majority of patient donors (Extended Data Fig. 1d,e). However, certain cell types were more abundant in particular sample groups: samples from patients with alopecia areata had increased T lymphocytes and depletion of follicular keratinocytes (Fig. 1h,i and Extended Data Fig. 2d,e). These observations align with alopecia areata pathophysiology, which involves peribulbar hair follicle T cell infiltration and disruption of normal hair follicle cycling[28].

HRGs use distinct enhancer modules to tune gene expression

We bioinformatically integrated our scATAC and scRNA datasets using canonical correlation analysis[29] and observed high correspondence between cell types (Extended Data Fig. 3d,e). We used these integrated datasets to identify CREs with accessibility correlated to local gene expression ('peak-to-gene links')[18,19]. To detect peak-to-gene linkages relevant for both broad cell type identity and regulation of more similar cell subtypes, we performed integration and peak-to-gene linkage identification on both the full scalp dataset and each of the subclustered datasets (Methods and Fig. 2a,b). In total, we identified 146,088 peak-to-gene links (Extended Data Fig. 5a,b). Only 66,702 links were detected using the full dataset (Extended Data Fig. 5a), but linkages from any source were more likely to be evolutionarily conserved than unlinked peaks (Extended Data Fig. 5c) and were more likely to be corroborated by enhancer–gene pair predictions in a large activity-by-contact (ABC) model dataset than distance-matched, permuted linkages (Extended Data Fig. 5d)[30,31]. Most peaks (491,106, 83.3%) were not linked to any gene, consistent with the expected small effect size of most CREs[32]. CREs were linked to the nearest gene in only 47% of cases, a proportion supported by experimental estimates of enhancer–gene linkages (Extended Data Fig. 5e)[30].

Consistent with previous studies, we identified a subset of genes associated with especially large numbers of linked CREs[33,34]. We identified 1,739 such HRGs by ranking genes according to the number of linked peaks and retaining those that exceeded the inflection point at 20 peak-to-gene linkages (Fig. 2c). These genes include TFs driving cell identity (*RUNX1*, *TWIST2* and *MITF*) and those with cell-type-specific functions (*COL1A1*, *KRT14* and *ICOS*). Scalp HRGs were enriched for previously identified 'superenhancer'-associated genes (Fig. 2d)[35,36], and overlapped significantly with previously described domains of regulatory chromatin-associated genes in mouse skin (odds ratio (OR) = 6.18, one-sided Fisher's exact test $P = 6.27 \times 10^{-119}$)[33]. To explore the regulatory heterogeneity of HRGs, we clustered *k*-nearest neighbor pseudobulked samples by the accessibility of linked CREs, using *k*-means clustering to identify co-occurring regulatory modules (Fig. 2e). HRGs from each cluster were enriched for cell-type-specific Gene Ontology (GO) terms, including 'adaptive immune response' (myeloid), 'melanocyte differentiation' (melanocytes) and 'hair follicle development' (follicular keratinocytes) (Fig. 2e).

Whereas many HRGs were expressed in one or a few closely related cell types, several HRGs such as *RUNX3* were expressed in multiple distinct cell types (Fig. 2b). To explore the regulatory heterogeneity of individual HRGs we clustered *k*-nearest neighbor pseudobulks using the accessibility of peaks linked to a single HRG (Methods). For many HRGs we observed multiple 'modules' of coaccessible CREs in distinct cell types (Fig. 2f–i). Some modules were shared by multiple cell types while others were highly cell type specific. Interestingly, the aggregate accessibility observed across linked peaks

https://doi.org/10.1038/s41588-023-01445-4

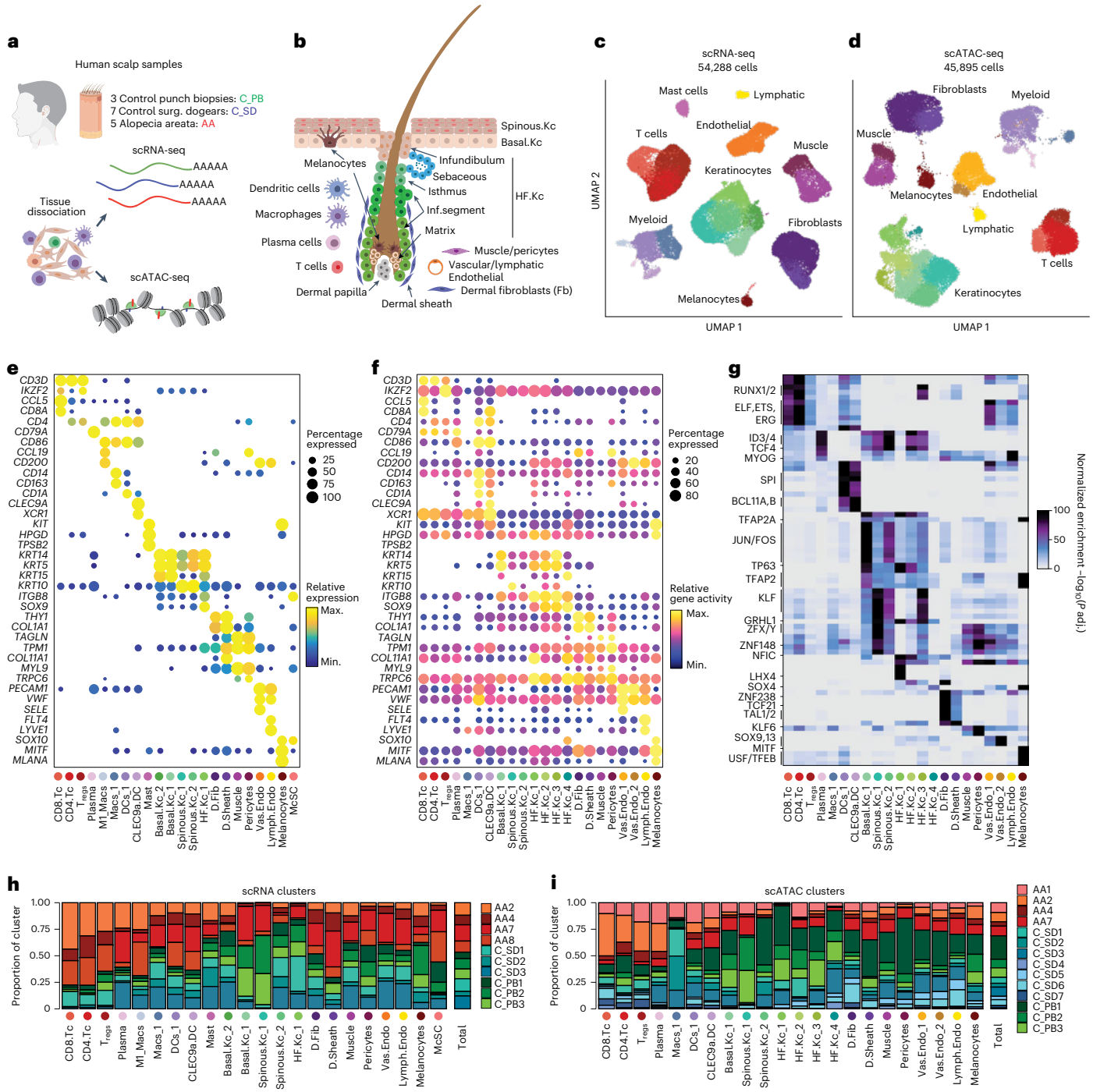


Fig. 1 | Multiomic single-cell atlas of primary human scalp. a, Samples and profiling methods used in this study. **b**, Schematic representation of cellular diversity within human scalp. **c**,**d**, UMAP representation of all scRNA-seq (**c**) and scATAC-seq (**d**) cells passing quality control, colored by annotated clusters. Broad cell types are labeled on UMAP and higher-resolution labels are shown in **e**,**f**. **e**, scRNA gene expression for selected marker genes for each scRNA-seq cluster. Color indicates relative expression across all clusters and dot size indicates the percentage of cells in that cluster expressing the gene. **f**, scATAC gene activity scores for the markers shown in **e**. **g**, Hypergeometric enrichment of TF motifs in marker peaks for each scATAC-seq cluster. **h**, Fraction of each sample comprising each scRNA-seq cluster. Samples from control punch biopsies are shown in shades of green, control surgical tissue in shades of blue and patients with alopecia areata in red. Total proportions for each sample are shown in the rightmost column. **i**, Same as **h** but for scATAC-seq clusters.

was correlated with expression of the linked gene (Extended Data Fig. 5f). These findings support an additive, modular model of enhancer activity—a model substantiated by genetic perturbation studies of individual enhancers for alpha-globin[37] and Myc[38], studies of enhancers involved in limb development[39] and genomic-scale measures of enhancer activity[40,41].

Gene-regulatory diversity of scalp keratinocytes

Whereas the transcriptional heterogeneity of interfollicular[12,17] and follicular[42] keratinocytes is known, our peak-to-gene linkage analysis enabled deeper interrogation of the gene-regulatory logic of these populations (Fig. 3a,b and Extended Data Fig. 6a,b). To focus on the gene-regulatory mechanisms delineating keratinocyte subsets, we

https://doi.org/10.1038/s41588-023-01445-4

used peak-to-gene linkages specific to keratinocytes (28,991 links) for subsequent analyses. *K*-means clustering of linkages revealed extensive gene-regulatory diversity within keratinocyte subtypes, with clusters of coaccessible peaks enriched for distinct TF motifs (Extended Data Fig. 6c,d). To identify TFs with a regulatory role in specifying keratinocyte subsets, we first identified motifs with variable accessibility between keratinocyte populations (Extended Data Fig. 6e, *y* axis); to differentiate between TFs with similar motifs, we correlated TF expression with motif activity across cell types (Extended Data Fig. 6e, *x* axis). This approach identified TFs with known roles in skin (TP63, FOSL1 and KLF4)[26,43–45] and hair differentiation (SOX9, LHX2 and HOXC13)[36,46,47]. Some TFs were active in multiple related cell types, such as TP63 in basal keratinocyte clusters and SOX9 in follicular keratinocyte clusters, while others were more cell type specific, like LHX2 in the inferior segment of the hair follicle and RUNX3 in sebaceous gland cells (Fig. 3c).

Gene targets of TFs driving keratinocyte differentiation

Interfollicular keratinocytes undergo continuous replacement by coordinated differentiation and outward migration of basal keratinocytes to spinous, granular and, finally, cornified keratinocytes. To identify TF drivers of this differentiation in a human in vivo context we constructed a semisupervised pseudotemporal trajectory between basal keratinocytes and differentiated spinous keratinocytes (Fig. 3d). Visualization of the most variable 10% of peaks along this trajectory revealed a continuous, gradual opening and closing of accessible chromatin (Fig. 3e). The most variable 10% of genes included known transcriptional changes during keratinocyte differentiation, with early trajectory cells expressing basal keratins (*KRT15*, *KRT5* and *KRT14*) and hemidesmosome components (*ITGA6*, *ITGB1* and *COL17A1*) and later cells expressing suprabasal keratins (*KRT1* and *KRT10*) (Fig. 3e)[48,49]. Genomic tracks of *ITGB1*, active in basal keratinocytes, and *KRT10*, active in spinous layer keratinocytes, demonstrate coordinated changes in linked enhancer accessibility and target gene expression across differentiation (Fig. 3f,g). By correlation of TF motif activity with expression using cells along the differentiation trajectory we identified TFs with known, sequential roles in interfollicular keratinocyte differentiation, such as TP63 followed by KLF3/4, RORA and then CEBPA/D (Fig. 3h)[43,50].

We next sought to identify potential regulatory gene targets of TFs driving keratinocyte cell identity. For TFs identified as potential differentiation drivers (Fig. 3h and Extended Data Fig. 6e) we correlated TF motif activity with the integrated gene expression of all expressed genes. Next, for each gene we selected all linked peaks containing the TF motif and computed a 'linkage score' aggregating peak-to-gene linkage strength and the confidence of embedded motif matches. Using this approach we identify potential TF regulatory targets as genes with expression correlated to global motif activity and a high TF linkage score (Pearson correlation >0.25 and linkage score >80th percentile). We identified 175 potential TP63 regulatory targets (Fig. 3i) and found enrichment of genes that were downregulated (OR = 1.95, one-sided Fisher's exact test $P = 0.0002$), but not upregulated (OR = 0.71), in keratinocytes with inactivating TP63 mutations[51]. These targets included basal keratins (*KRT5* and *KRT14*) and genes involved in anchoring keratinocytes to the basement membrane (*LAMC2*, *ITGA6* and *COL17A1*), consistent with the known role of TP63 in regulation of adhesion[52]. FOSL1, active in the intermediate stages of differentiation, was linked to targets enriched for cadherin binding functionality, a regulatory signal in early keratinocyte differentiation (Fig. 3j)[53]. For KLF4, a TF involved in terminal differentiation, targets included regulators of keratinocyte differentiation (*DMKN* and *KRTDAP*) and structural components of spinous and granular keratinocytes (*KRT1*, *2* and *IVL*; Fig. 3k). Predicted KLF4 targets were also enriched for genes downregulated (OR = 1.87, one-sided Fisher's exact test $P = 4.9 \times 10^{-7}$) but not upregulated (OR = 0.95) in keratinocytes following KLF4 knockdown[54]. We also identified gene targets of TFs involved in follicular keratinocyte function, identifying those associated with WNT–protein binding for LHX2, a TF expressed in inferior segment follicular keratinocytes (Extended Data Fig. 6f–h).

Selective preservation of HFSCs in alopecia areata

Alopecia areata results in disruption of hair follicle cycling by autoreactive cytotoxic T lymphocytes, but we did not observe a clear phenotypic distinction between T lymphocytes originating from areata versus control samples (Supplementary Note). However, we found that selected follicular keratinocyte populations appeared to be depleted in areata samples (Extended Data Fig. 7a). Using Milo[55] we confirmed that, compared with control samples, areata samples had fewer cells corresponding to the inferior segment of the hair follicle (Fig. 4a,b). Further subclustering of these cells revealed six populations of follicular keratinocytes in the bulbar and suprabulbar regions of the hair follicle (Fig. 4c,d and Extended Data Fig. 7b,d). We annotated these as quiescent hair follicle stem cells (HFSCs: *KRT15*, *CD200*, *LHX2* and *NFATC1*)[36,56–58], two populations of sheath cells (Sheath_1/2: *SOX9*, *KRT5* and *KRT75*)[42], matrix cells (Matrix: *LEF1*, *KRT81* and *HOXC13*)[59,60] and hair germ cells (HG: *CD34*, *LGR5*, *CDH3* and *WNT3*)[61]. Using Milo, we found that areata samples demonstrated preservation of HFSCs but a depletion of sheath populations (FDR < 0.1; Fig. 4e), consistent with the known nonscarring, relapsing–remitting nature of alopecia areata and supporting the theory that sheath cells in the hair bulb are especially affected by the disrupted immune environment[28,62,63].

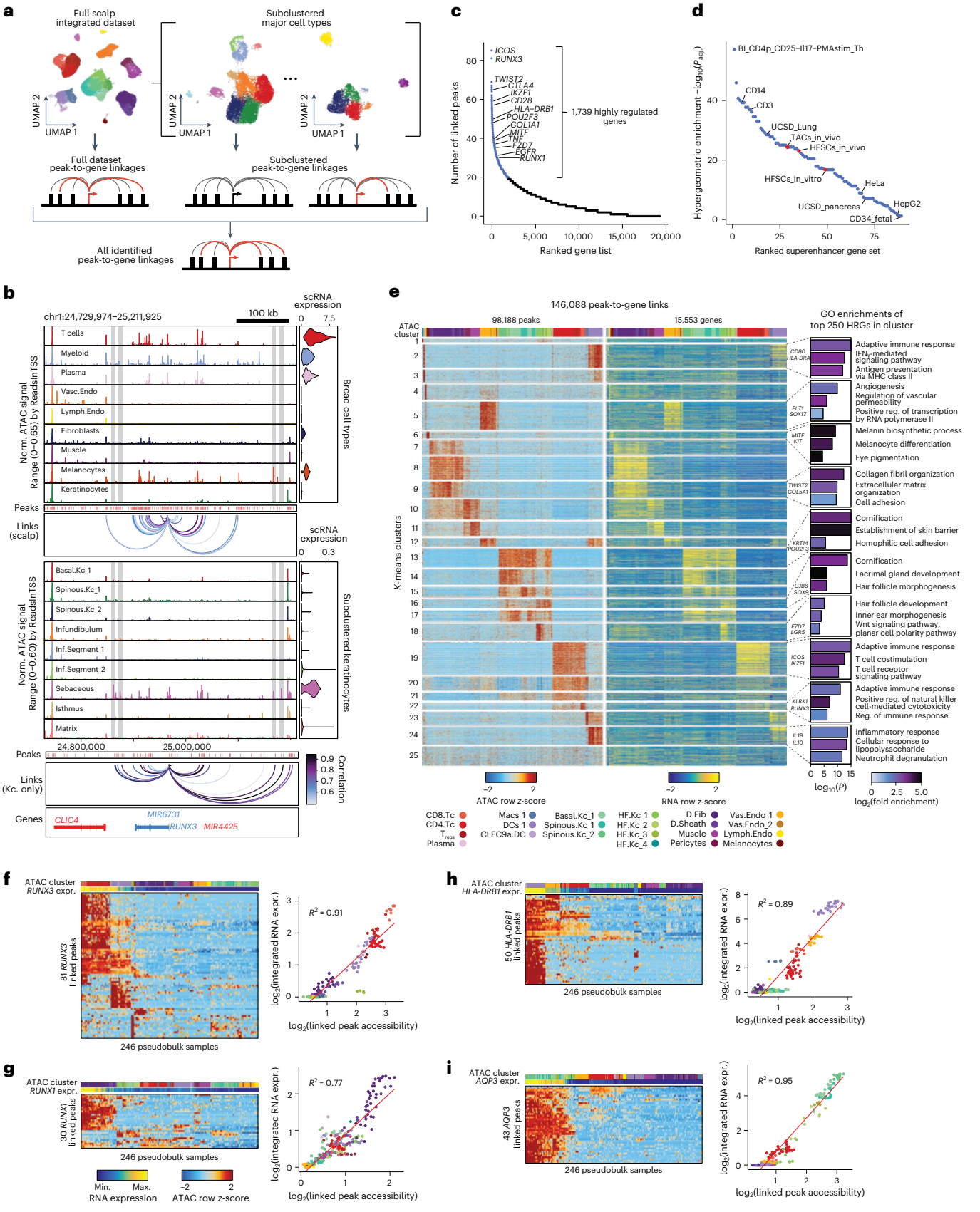
WNT pathway dynamics in hair keratinocyte differentiation

The WNT signaling pathway plays an essential role in hair follicle development, cycling and regeneration after wounding[33,46,64–68]. However, most studies of WNT pathway activity in hair follicle cycling used in vitro or mouse in vivo systems. To explore WNT signaling dynamics in human hair follicles we constructed a semisupervised pseudotemporal trajectory from quiescent HFSCs to matrix cells (Fig. 4f). We correlated TF motif activity with expression along this trajectory to identify putative drivers of differentiation (Extended Data Fig. 7e). Consistent with mouse studies, NFATC1 was active in quiescent HFSCs,

Fig. 2 | Gene-regulatory dynamics and modularity in human scalp. a, Peak-to-gene linkages were identified on the integrated scATAC and scRNA full datasets, and on the five major cell type subclustered datasets. Linkages identified in each dataset are merged to form the full set of peak-to-gene linkages. **b**, Genomic tracks for chromatin accessibility around the *RUNX3* locus. Right: integrated *RUNX3* expression levels are shown in the violin plot for each cell type. Loops shown below the top panel indicate peak-to-gene linkages identified on the full dataset. Bottom: genomic tracks for accessibility around *RUNX3* for subclustered keratinocytes. Loops shown below these tracks indicate peak-to-gene linkages identified on the subclustered dataset. Gray vertical bars spanning both panels highlight selected peaks linked to *RUNX3* expression that were identified in subclustered keratinocytes but not using the full integrated dataset. **c**, Genes ranked by the number of peak-to-gene links identified for each gene: 1,739 HRGs had >20 peak-to-gene linkages. **d**, Hypergeometric enrichment of superenhancer-linked genes in human scalp HRGs for multiple cell and tissue types. Red dots represent enrichment of hair follicle superenhancer-linked genes. **e**, Heatmap showing chromatin accessibility (left) and gene expression (right) for 146,088 peak-to-gene linkages, which were clustered using *k*-means clustering ($k = 25$). Sample top HRGs for selected clusters are shown to the right of the gene expression heatmap. Right: GO term enrichments for the top 200 genes ranked by number of peak-to-gene linkages for selected *k*-means clusters. **f**, Heatmap showing chromatin accessibility at *RUNX3*-linked peaks for 246 pseudobulked scATAC-seq samples. Cell type labels are shown in the bar above the heatmap, and *RUNX3* expression levels for each pseudobulk below. Right: scatter plot showing the relationship between linked peak accessibility and resulting gene expression for each of the pseudobulked samples shown in the heatmap on the left. Red line indicates line of best fit. **g**, Same as in **f** but for *RUNX1*. **h**, Same as in **f** but for *HLA-DRB1*. **i**, Same as in **f** but for *AQP3*.

https://doi.org/10.1038/s41588-023-01445-4

the WNT-regulating TFs TCF3 and TCF4 became active in intermediate sheath cells and LEF1 activity surged in matrix cells[69–71]. GO term analysis on the most variably expressed genes across this trajectory revealed enrichment of WNT signaling pathway genes (Extended Data Fig. 7f). To visualize WNT signaling dynamics during HFSC differentiation we plotted the expression of WNT signaling factors and receptors across pseudotime (Fig. 4g). HFSCs robustly expressed WNT receptors *FZD1* and *7*, but also soluble WNT inhibitors (*DKK3*, *SFRP1*) and the soluble FZD



https://doi.org/10.1038/s41588-023-01445-4

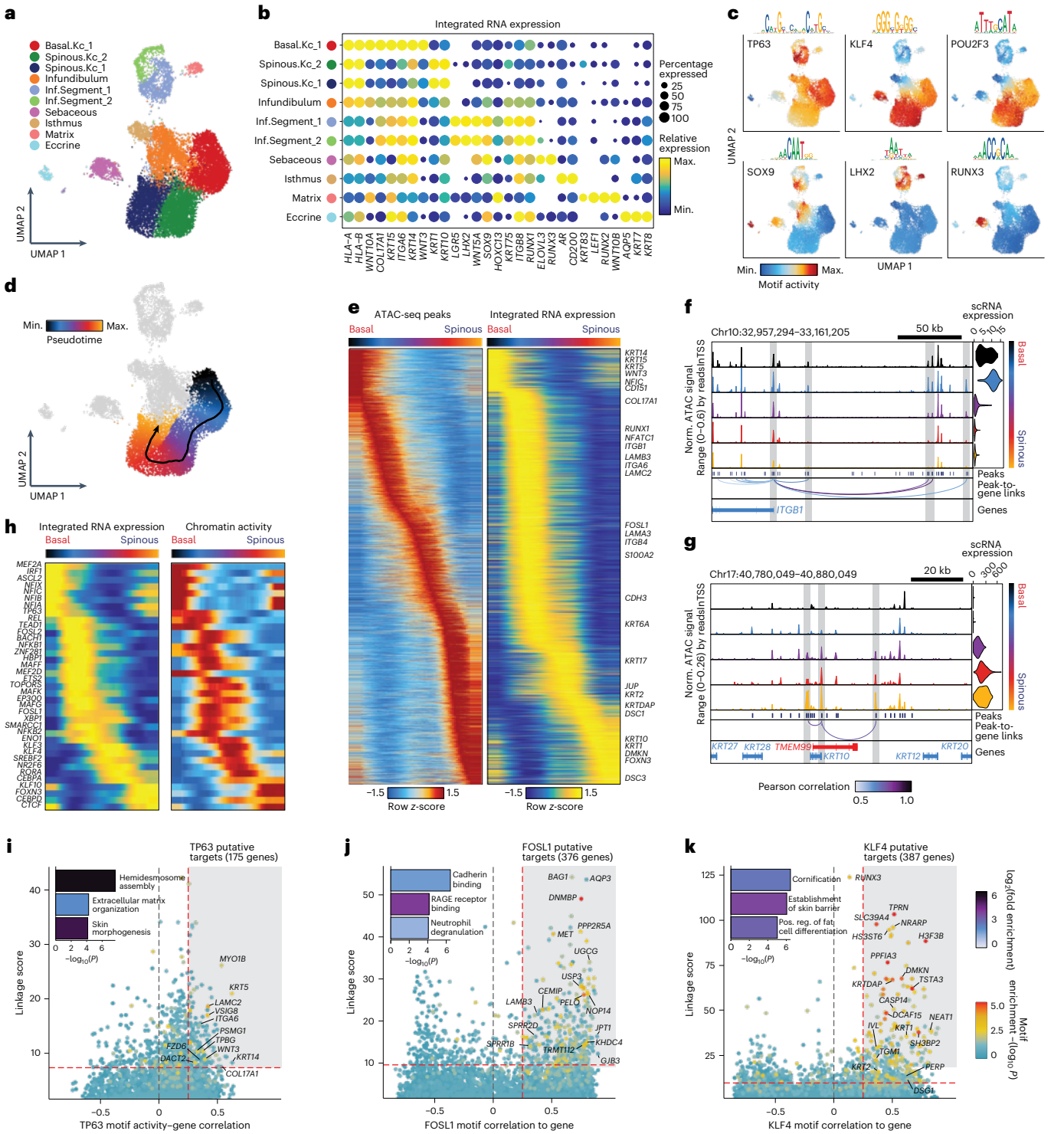


Fig. 3 | Scalp keratinocyte diversity and regulatory control of interfollicular keratinocyte differentiation. a, UMAP representation of subclustered keratinocytes in the scATAC-seq dataset. **b**, Integrated gene expression for select markers across keratinocyte subtypes. Color indicates relative expression across all clusters and dot size indicates the percentage of cells in that cluster expressing the gene. **c**, ChromVAR deviation *z*-scores showing TF motif activity for selected TFs. **d**, Slingshot differentiation trajectory starting with basal interfollicular keratinocytes and progressing to upper layer spinous keratinocytes. **e**, Heatmap of 10% most variable peaks ($n = 31,333$) and 10% most variable genes ($n = 2,127$) along the trajectory from basal to spinous keratinocytes. **f**, Genomic tracks of accessibility around the *ITGB1* promoter. Tracks are pseudobulked samples ordered along the interfollicular differentiation trajectory. Right: integrated *ITGB1* expression levels are shown in the violin plot for each pseudobulk. **g**, Same as in **f** but for the *KRT10* promoter. **h**, Paired heatmaps of positive TF regulators whose TF motif activity (left) and matched gene expression (right) are positively correlated across the interfollicular keratinocyte differentiation pseudotime trajectory. **i**, Prioritization of gene targets for TP63. The *x* axis shows Pearson correlation between TF motif activity and integrated gene expression for all expressed genes across all keratinocytes; the *y* axis shows TF linkage score (for all linked peaks, sum of motif score scaled by peak-to-gene link correlation). Color of points indicates hypergeometric enrichment of the TF motif in all linked peaks for each gene. Top gene targets are indicated in the shaded area (motif correlation to gene expression >0.25, linkage score >80th percentile). Inset, GO term enrichments for top gene targets. **j**, Same as in **i** but for FOSL1. **k**, Same as in **i** but for KLF4. Norm., normalized. ReadsInTSS, reads in transcription start sites.

https://doi.org/10.1038/s41588-023-01445-4

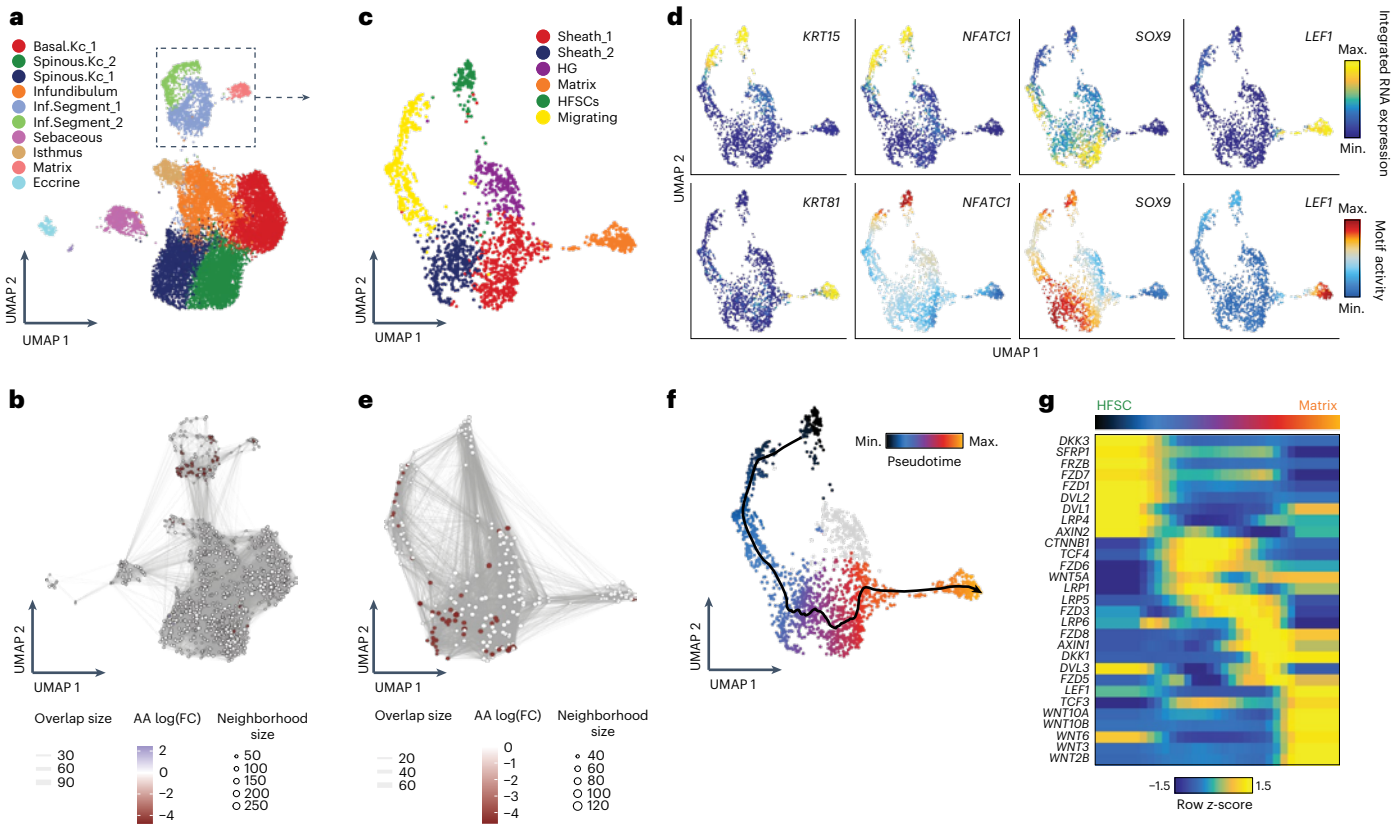


Fig. 4 | Regulatory dynamics of human hair follicle cycling. a, Subclustered keratinocytes in scATAC-seq space. The inferior segment of the hair follicle is highlighted. **b**, Differential abundance of cycling hair follicle keratinocytes between alopecia areata and control samples using Milo. Colored spots represent neighborhoods that are differentially abundant with spatial FDR < 0.1. **c**, Subclustered hair follicle keratinocytes from the inferior segment of the hair follicle. **d**, Selected marker gene expression and TF motif activity deviation *z*-scores for subclustered inferior segment hair follicle keratinocytes. **e**, Same as in **b**, except for subclustered cycling hair follicle keratinocytes. Hair sheath cells are differentially depleted relative to HFSCs. **f**, Differentiation trajectory from HFSCs to matrix cells. **g**, Heatmap of variable expression of members of the WNT signaling pathway during hair follicle cycling.

receptor *FRZB*, suggesting that these cells may be primed to respond to paracrine WNT signaling but maintain quiescence by blocking these signals. As differentiation progresses, expression of WNT pathway inhibitory signals decreases and expression of beta-catenin (*CTNNB1*) and WNT-regulating TFs (*TCF3* and *TCF4*) increases. Consistent with studies in mice, dividing matrix cells in the hair bulb express activating WNT effectors (*WNT3*, *WNT5A* and *WNT10A/B*) and the TF *LEF1* (Fig. 4d,g and Extended Data Fig. 7e)[33,72].

Identification of critical cell types for skin and hair traits

Many skin and hair diseases are highly polygenic, and most associated variants reside in noncoding genomic regions[5,7,9,73]. To identify cell types involved in the pathoetiology of skin and hair disease we used cell-type-specific open-chromatin regions to perform linkage disequilibrium score regression (LDSC) using GWAS for 13 traits spanning skin and hair disease, autoimmune disease and several nonskin phenotypes (Fig. 5a and Extended Data Fig. 8a,b)[74,75]. We observed enrichment of AGA according to SNP heritability across fibroblast open-chromatin regions, with the strongest enrichment in DP peaks—the component of the hair follicle reported to have the highest androgen receptor activity[76,77]. We also found modest but significant enrichment of AGA GWAS signal in several follicular keratinocyte clusters. Autoimmune skin diseases, including psoriasis and eczema, had significant enrichment in T lymphocyte open-chromatin regions while tanning and hair pigment color were most enriched in melanocyte open-chromatin regions. Traits not related to scalp, such as schizophrenia and body mass index, did not demonstrate any cell-type-specific enrichment (Fig. 5a). Additional LDSC analyses are discussed in Supplementary Note and Extended Data Fig. 8c–f.

Because LDSC requires full GWAS summary statistics, which were unavailable for several traits including alopecia areata, we also examined enrichment of fine-mapped SNPs in cell-type-specific open-chromatin regions[78–80]. Fine-mapped SNPs for skin, hair and autoimmune disorders were more likely to overlap scalp CREs than those for neurodegenerative and psychiatric disorders, and this gap widened with increasing fine-mapping posterior probability (Fig. 5b and Extended Data Fig. 8f). We observed cell-type-specific enrichment of fine-mapped SNPS for several diseases (Fig. 5c). Alopecia areata SNPs were most enriched in CD4 T cell (OR = 3.91, one-sided Fisher's exact test adjusted $P = 0.00018$) and T regulatory cell (T_{reg}; OR = 4.16, one-sided Fisher's exact test adjusted $P = 0.0012$) open-chromatin regions, but were also enriched in several myeloid lineage clusters (for example, M2.macs_2: OR = 3.25, one-sided Fisher's exact test adjusted $P = 0.00027$). Interestingly, although body height-associated SNPs were broadly enriched in fibroblast clusters there was little enrichment in the DP cluster, while AGA SNPs were most strongly enriched in DP open-chromatin regions (OR = 5.72, one-sided Fisher's exact test adjusted $P = 4.3 \times 10^{-34}$; Fig. 5c).

Linking fine-mapped SNPs to potential target genes

After nominating disease-relevant cell types in the scalp, we sought to identify specific genes associated with fine-mapped SNPs. For a given phenotype we aggregated the posterior probability of fine-mapped SNPs overlapping linked peaks for each gene then plotted the

https://doi.org/10.1038/s41588-023-01445-4

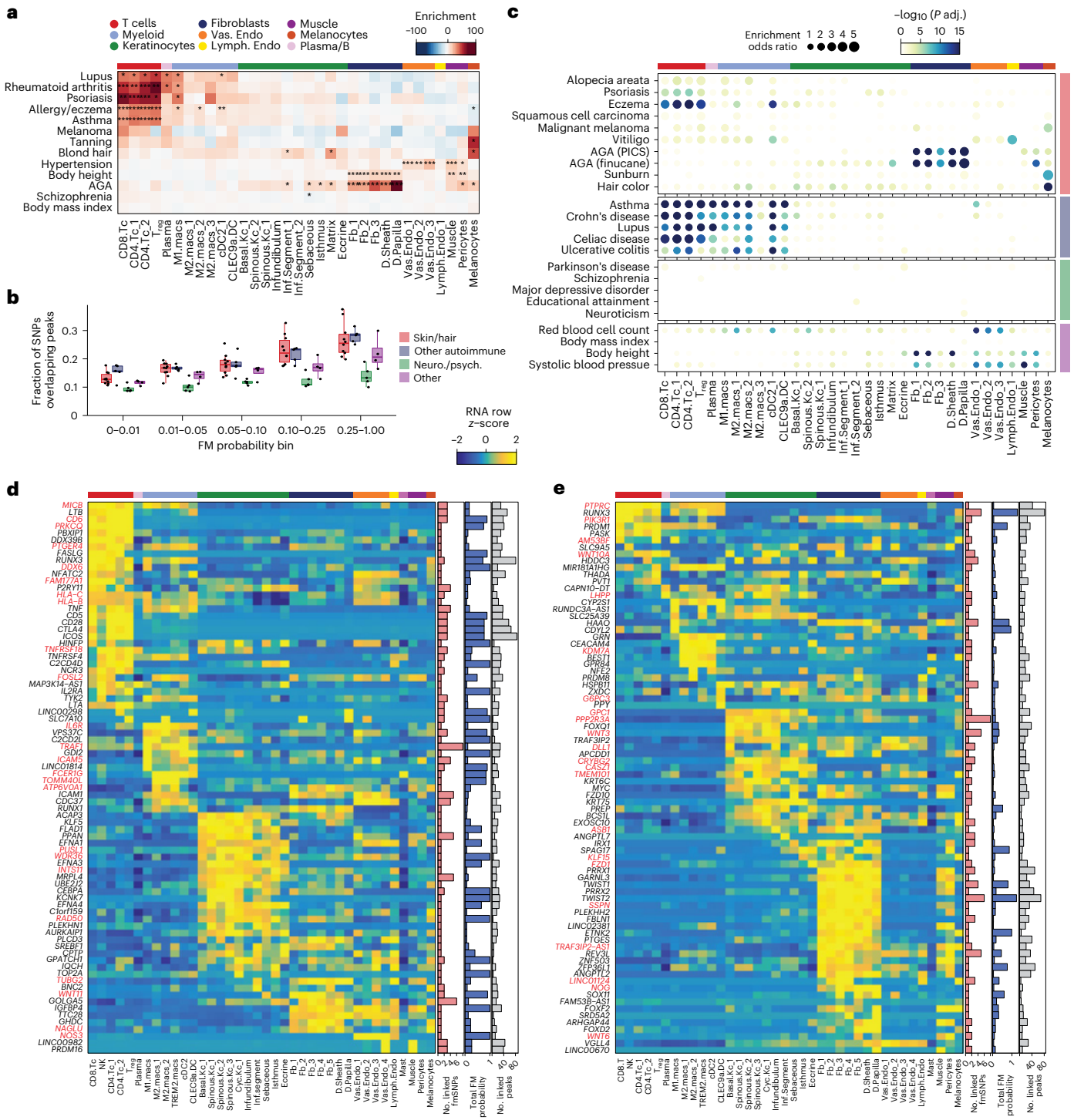


Fig. 5 | Identification of cell types and genes associated with hair, skin and autoimmune diseases. a, LDSC identifies enrichment of GWAS SNPs for various skin- and nonskin-related conditions in peak regions specific to subclustered cell types in human scalp. FDR-corrected *P* values from LDSC enrichment tests are overlaid on the heatmap (*FDR < 0.05, **FDR < 0.005, ***FDR < 0.0005). **b**, Fraction of fine-mapped (FM) SNPs overlapping scalp open-chromatin regions binned by increasing fine-mapping posterior probability. Each dot represents one trait and boxplot color indicates the group of traits being plotted; the number of traits per group is shown in **c**. Boxplots represent the median, 25th and 75th percentiles of the data and whiskers represent the highest and lowest values within 1.5 times the interquartile range of the boxplot. **c**, One-sided Fisher's exact test enrichment for fine-mapped, trait-related SNPs in peak regions specific to subclustered cell types in human scalp. Dot color indicates FDR-corrected $-\log_{10} P$ value and dot size indicates enrichment OR. Traits are grouped as in **b**. **d**, The top genes linked to peaks containing fine-mapped SNPs for eczema. The heatmap shows relative gene expression for each high-resolution scRNA cluster. The number of linked fine-mapped SNPs per gene is indicated in the red bar plot to the right, and the sum of fine-mapped posterior probability for linked SNPs is indicated in the blue bar plot. The gray bar plot shows the total number of identified peak-to-gene linkages for that gene. Gene names colored red indicate fine-mapped SNP-to-gene linkages supported by GTEx expression quantitative trait loci. **e**, Same as in **d** but for AGA.

expression of linked genes across high-resolution scRNA clusters to identify cell-type-specific expression of genes linked to fine-mapped SNPs (fmGWAS-linked genes). We identified 137 eczema fmGWAS-linked genes, the majority of which were expressed in T cell or keratinocyte clusters (Fig. 5d). These genes included modulators of immune signaling (*TNF*, *CTLA4* and *FASLG*) and previously nominated GWAS gene targets (*IL6R*, *PUS10* and *IL2RA*)[9,81]. We identified 130 AGA fmGWAS-linked genes, most of which were expressed in keratinocyte or fibroblast clusters (Fig. 5e). These genes were enriched for TFs (OR = 3.18, one-sided Fisher's exact test $P = 4.1 \times 10^{-7}$), including *TWIST2*, *RUNX3* and *SOX11*, and were also enriched for members of the WNT signaling pathway (*WNT3*, *WNT10A*, *FZD1* and *FZD10*; Extended Data Fig. 9a). We identified only 31 alopecia areata fmGWAS-linked genes, but these included examples involved in T cell functions such as *IL21*, *ICOS* and *IRF4* (Extended Data Fig. 9b). *IL21* had multiple linked SNPs, is known to support the persistence of cytotoxic CD8 T cells in chronic viral infections[82,83] and has been implicated in the etiology of several autoimmune diseases[84]. We also identified 158 hair color fmGWAS-linked genes, principally expressed in keratinocyte subpopulations and melanocytes (Extended Data Fig. 9c).

Nominating functional SNPs for skin and hair phenotypes

After nominating cell types and gene targets associated with skin and hair disease, we sought yet-higher-resolution information by identifying SNPs that might directly alter TF binding and enhancer function. To prioritize functional SNP candidates we implemented a gapped *k*-mer support vector machine (gkm-SVM) learning framework to score the allelic effect of a SNP on cell-type-specific chromatin accessibility, a proxy for differential TF binding (Methods and Fig. 6a)[85–88]. These models demonstrated accurate and stable performance on held-out data in a tenfold cross-validation scheme (Extended Data Fig. 10a–d). We used GkmExplain to predict the per-base impact of variants in a target sequence by providing models with sequences containing both the reference and alternative allele for a candidate SNP[89]. To create a set of prioritized SNPs for AGA, eczema and hair color we selected SNPs that (1) had fine-mapping posterior probability ≥0.01, (2) overlapped scalp CREs and (3) were predicted to disrupt chromatin accessibility in our model. Prioritized SNPs for eczema were enriched in keratinocyte and T cell clusters relative to random trait CRE-resident fine-mapped SNPs, while prioritized SNPs for hair color were enriched more specifically in follicular keratinocytes and melanocytes (Fig. 6b and Extended Data Fig. 10e). We did not observe cluster-specific enrichment of AGA-prioritized SNPs, perhaps due to the specificity of this trait for DP cells and the lack of a DP-specific model given the rarity of these cells in our dataset (Methods and Extended Data Fig. 10f). We filtered prioritized SNPs to include only those linked to a target gene using our peak-to-gene linkage analysis, increasing the interpretability of potential causative variants. Using these criteria we identified 47, 19 and 19 prioritized SNPs for AGA, eczema and hair color, respectively (Supplementary Table 12 and Extended Data Fig. 10g–j).

One high-effect eczema SNP is rs2058622, which resides in an *IL18R1* intron (Fig. 6c). This candidate SNP overlapped a CRE preferentially accessible in the CD4 helper T cell cluster and, although this CRE was within an *IL18R1* intron, this peak was linked to *IL18RAP* expression. Our CD4 T cell model suggested that the alternative allele of this SNP increases cell-type-specific chromatin accessibility at this peak, possibly by creating a RUNX motif (Fig. 6d). Interestingly, T-bet (encoded by *TBX21*), which also contains a central 'GTG' in its binding motif, has been shown to bind to this SNP region in a genotype-specific manner, suggesting multiple candidates for transfactors with differential binding to the major and minor allele of this regulatory element[90]. Furthermore, this SNP had been identified as a significant eQTL for *IL18RAP* expression in blood, with the G allele increasing expression ($P = 4.8 \times 10^{-54}$, normalized effect size 0.28). While this locus is one of those most strongly associated with eczema[9,81], it is a region with substantial LD and multiple potential gene targets, making identification of causal SNPs for this locus challenging and highlighting the utility of our multitiered approach (Fig. 6e). *IL18RAP* encodes an accessory protein required for potentiation of IL-18 signaling[91] and IL-18 overexpression in mouse skin induces a phenotype similar to eczema[92], suggesting a mechanistic pathway for this causal variant.

One high-effect AGA SNP is rs72966077, located immediately downstream of the *WNT10A* gene body (Fig. 6f). This SNP has also been implicated in acne vulgaris, another hair follicle- and androgen-associated disease[93]. The overlapping CRE is accessible in multiple keratinocyte clusters, although *WNT10A* expression was highest in basal keratinocytes and infundibular follicular keratinocytes. Our model demonstrates that the alternative allele of this SNP disrupts an ERG family TF motif (Fig. 6g). *ERG2* is expressed in infundibular, isthmus and inferior segment hair follicle keratinocytes, and these cell populations also have higher ERG2 motif activity than other keratinocyte populations (Fig. 6h). Patients with *WNT10A* mutations exhibit multiple skin appendage-related phenotypes, including hair thinning that resembles AGA[94]. Furthermore, depletion of ERG2 (also known as Krox20)-positive follicular keratinocytes in mice resulted in arrest of hair growth[95]. These converging evidences highlight the importance of the WNT signaling pathway in the pathobiology of AGA and show that, while the strongest AGA GWAS signal enrichment is in DP cells, there may also be keratinocyte-intrinsic genetic factors that contribute to this complex trait.

Discussion

We generated epigenomic and transcriptomic atlases of human scalp, a complex tissue harboring dynamic and precisely regulated hair follicles. We identified principles of variable gene expression across diverse cell types, defined gene-regulatory networks across keratinocyte subpopulations and prioritized cell types, genes and causal variants implicated in the pathobiology of skin and hair phenotypes.

The aggregate accessibility of linked enhancer modules predicts HRG expression (Fig. 2f–i), supporting an additive model of enhancer activity wherein expression is proportional to the integrated effect of multiple, generally interchangeable, CREs. We posit that this regulatory strategy makes expression of core function genes resistant to perturbation but also allows for tunable expression across cellular contexts. Thus, a mutation disrupting one of relatively few CREs controlling expression may have a greater biological impact than disruption of an HRG CRE (Fig. 5d,e and Extended Data Fig. 9b,c). This mode of enhancer activity is consistent with a recent study of enhancer–promoter interactions showing that variability in intrinsic enhancer activity is low compared with intrinsic promoter activity, and that enhancer elements are generally functionally interchangeable[41].

To identify TF regulatory targets we combined the correlation between TF activity and target gene expression, with a linkage score accounting for linked CREs containing TF binding motifs. Typically, high linkage scores were associated with positive correlation between TF activity and target gene expression (Fig. 3i,k and Extended Data Fig. 6f,g) but, for some TFs such as FOSL1 and POU2F3 (Fig. 3j and Extended Data Fig. 6h), several gene targets had high linkage scores but negative correlation to TF motif activity. This may imply a gene-silencing role for these TFs for selected targets. Indeed, selective transcriptional repression has been described for both FOSL1 (ref. 96) and POU2F3 (refs. 97,98). Several genes also had high linkage scores but little correlation to TF activity (Fig. 3i–k and Extended Data Fig. 6f–h). This may be due to biologically relevant TF binding despite lower global TF activity, or CRE accessibility may be driven by a different TF with co-occurring binding motifs. Emerging single-cell methodologies, such as single-cell CUT&Tag[99] or NEAT-seq[100], may help differentiate between these possibilities.

Our analyses using LDSC and fine-mapped SNP enrichment in CREs revealed driver cell types for hair and skin diseases. While we

https://doi.org/10.1038/s41588-023-01445-4

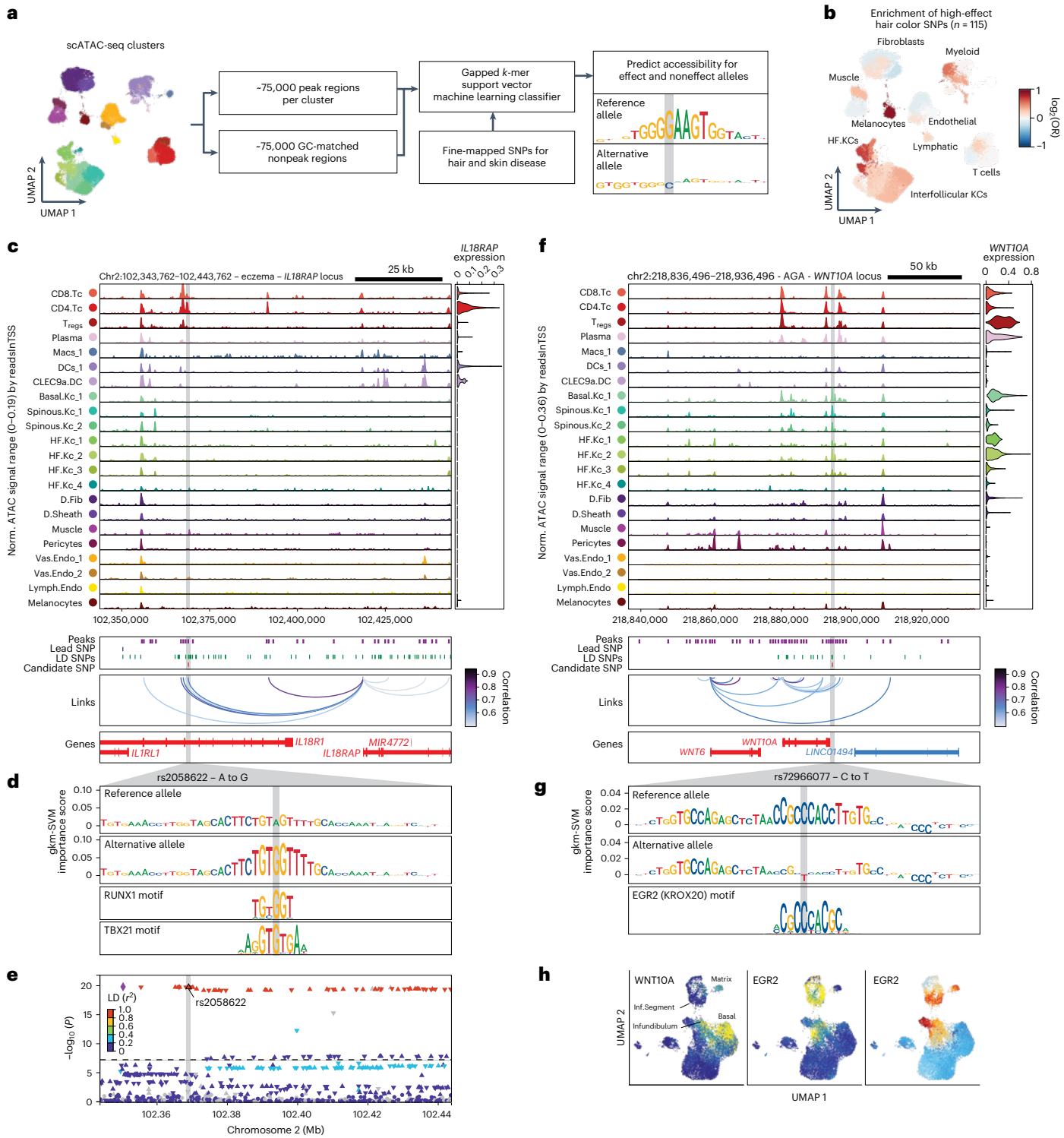


Fig. 6 | Candidate causal variants in skin and hair disease. a, Schematic of strategy used for identification of potential causative variants. **b**, Enrichment of high-effect fine-mapped SNPs from select skin and hair traits relative to random fine-mapped SNPs in *cis*-regulatory regions. **c**, Normalized chromatin accessibility landscape for cell-type-specific pseudobulk tracks around the *IL18RAP* locus. Integrated *IL18RAP* expression levels are shown in the violin plot for each cell type to the right. The position of ATAC-seq peaks, the GWAS lead SNP, the fine-mapped SNP candidates in LD with the lead SNP and the candidate functional SNP are shown below the ATAC-seq tracks. Significant peak-to-gene linkages are indicated by loops connecting the *IL18RAP* promoter to indicated peaks. **d**, GkmExplain importance scores for the 50-base-pair region surrounding rs2058622, an eczema-associated SNP that disrupts a RUNX motif in a CRE linked to *IL18RAP* expression. The effect and noneffect alleles for the gkm-SVM model correspond to the model trained on the CD4 helper T cell cluster. **e**, LocusZoom plot of the region shown in **c**, highlighting the strong LD and high overall GWAS signal of this locus. **f**, Same as in **c** but for the *WNT10A* locus. **g**, GkmExplain importance scores for the 50-base-pair region surrounding rs72966077, an AGA-associated SNP that disrupts an ERG motif in a CRE linked to *WNT10A* expression. The effect and noneffect alleles for the gkm-SVM model correspond to the model trained on the infundibular keratinocytes cluster. **h**, UMAP projection of high-resolution keratinocyte subclustering, showing expression of *WNT10A*, *EGR2* and EGR2 ChromVAR motif activity.

see enrichment of GWAS signals for AGA in some follicular keratinocyte subpopulations, the most significant enrichment was in DP CREs (Fig. 5). This is consistent with functional studies showing that DP cells have robust AR expression[76] and exhibit distinct expression profiles when isolated from both balding and nonbalding individuals[101,102]. Interestingly, autoimmune diseases such as eczema and psoriasis, with clear keratinocyte phenotypes clinically and histopathologically, showed little GWAS signal enrichment in keratinocytes relative to T lymphocytes, suggesting that the genetic susceptibility to these diseases is primarily immunological and due less to genetic variation intrinsic to keratinocytes.

Finally we used machine learning models of chromatin accessibility to nominate functional SNPs for hair and skin diseases, tracing the regulatory effect of single-base changes to disruption of target gene expression in the relevant cell type. However, we were unable to identify a potential causal SNP for many GWAS loci, perhaps because the affected CREs are observed only in the disease state or because the relevant cell type was not recovered. Some traits may be the result of a developmental process, with relevant regulatory regions dormant in adult tissues. While these analyses provide a valuable framework for linking genetic variation to disease phenotypes, individual SNP-to-gene linkages will require experimental validation in appropriate cellular contexts to be claimed as bona fide regulatory interactions. We anticipate that future studies will be able to fill these gaps as the costs of single-cell sequencing decrease, experimentally tractable model systems improve and models of gene regulation are refined.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01445-4.

References

1. Paus, R. & Cotsarelis, G. The biology of hair follicles. *N. Engl. J. Med.* **341**, 491–497 (1999).
2. Woo, W.-M. & Oro, A. E. SnapShot: hair follicle stem cells. *Cell* **146**, 334–334 (2011).
3. Schneider, M. R., Schmidt-Ullrich, R. & Paus, R. The hair follicle as a dynamic miniorgan. *Curr. Biol.* **19**, R132–R142 (2009).
4. Hsu, Y.-C. & Fuchs, E. Building and maintaining the skin. *Cold Spring Harb. Perspect. Biol.* **14**, a040840 (2022).
5. Petukhova, L. et al. Genome-wide association study in alopecia areata implicates both innate and adaptive immunity. *Nature* **466**, 113–117 (2010).
6. Betz, R. C. et al. Genome-wide meta-analysis in alopecia areata resolves HLA associations and reveals two new susceptibility loci. *Nat. Commun.* **6**, 5966 (2015).
7. Pirastu, N. et al. GWAS for male-pattern baldness identifies 71 susceptibility loci explaining 38% of the risk. *Nat. Commun.* **8**, 1584 (2017).
8. Hagenaars, S. P. et al. Genetic prediction of male pattern baldness. *PLoS Genet.* **13**, e1006594 (2017).
9. Paternoster, L. et al. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* **47**, 1449–1456 (2015).
10. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS era: from association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
11. Gur, C. et al. LGR5 expressing skin fibroblasts define a major cellular hub perturbed in scleroderma. *Cell* **185**, 1373–1388 (2022).
12. Reynolds, G. et al. Developmental cell programs are co-opted in inflammatory skin disease. *Science* **371**, eaba6500 (2021).
13. Borcherding, N. et al. A transcriptomic map of murine and human alopecia areata. *JCI Insight* **5**, e137424 (2020).
14. He, H. et al. Single-cell transcriptome analysis of human skin identifies novel fibroblast subpopulation and enrichment of immune subsets in atopic dermatitis. *J. Allergy Clin. Immunol.* **145**, 1615–1628 (2020).
15. Hughes, T. K. et al. Second-strand synthesis-based massively parallel scRNA-seq reveals cellular states and molecular features of human inflammatory skin pathologies. *Immunity* **53**, 878–894 (2020).
16. Gellatly, K. J. et al. scRNA-seq of human vitiligo reveals complex networks of subclinical immune activation and a role for CCR5 in T function. *Sci. Transl. Med.* **13**, eabd8995 (2021).
17. Wang, S. et al. Single cell transcriptomics of human epidermis identifies basal stem cell transition states. *Nat. Commun.* **11**, 4239 (2020).
18. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
19. Granja, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
20. Rendl, M., Lewis, L. & Fuchs, E. Molecular dissection of mesenchymal-epithelial interactions in the hair follicle. *PLoS Biol.* **3**, e331 (2005).
21. Iizuka, T., Suzuki, T., Nakano, K. & Sueki, H. Immunolocalization of aquaporin-5 in normal human skin and hypohidrotic skin diseases. *J. Dermatol.* **39**, 344–349 (2012).
22. Wang, E. C. E., Dai, Z., Ferrante, A. W., Drake, C. G. & Christiano, A. M. A subset of TREM2 dermal macrophages secretes oncostatin M to maintain hair follicle stem cell quiescence and inhibit hair growth. *Cell Stem Cell* **24**, 654–669 (2019).
23. Collins, A., Littman, D. R. & Taniuchi, I. RUNX proteins in transcription factor networks that regulate T-cell lineage choice. *Nat. Rev. Immunol.* **9**, 106–115 (2009).
24. Muthusamy, N., Barton, K. & Leiden, J. M. Defective activation and survival of T cells lacking the Ets-1 transcription factor. *Nature* **377**, 639–642 (1995).
25. Nerlov, C. & Graf, T. PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. *Genes Dev.* **12**, 2403–2412 (1998).
26. Truong, A. B., Kretz, M., Ridky, T. W., Kimmel, R. & Khavari, P. A. p63 regulates proliferation and differentiation of developmentally mature keratinocytes. *Genes Dev.* **20**, 3185–3197 (2006).
27. Levy, C., Khaled, M. & Fisher, D. E. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends Mol. Med.* **12**, 406–414 (2006).
28. Pratt, C. H., King, L. E. Jr, Messenger, A. G., Christiano, A. M. & Sundberg, J. P. Alopecia areata. *Nat. Rev. Dis. Prim.* **3**, 17011 (2017).
29. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
30. Fulco, C. P. et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
31. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
32. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 1516 (2019).
33. Ma, S. et al. Chromatin potential identified by shared single-cell profiling RNA chromatin. *Cell* **183**, 1103–1116 (2020).
34. Trevino, A. E. et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053–5069 (2021).
35. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
36. Adam, R. C. et al. Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. *Nature* **521**, 366–370 (2015).

37. Hay, D. et al. Genetic dissection of the α-globin super-enhancer in vivo. *Nat. Genet.* **48**, 895–903 (2016).
38. Bahr, C. et al. A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **553**, 515–520 (2018).
39. Osterwalder, M. et al. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
40. Choi, J. et al. Evidence for additive and synergistic action of mammalian enhancers during cell fate determination. *eLife* **10**, e65381 (2021).
41. Bergman, D. T. et al. Compatibility rules of human enhancer and promoter sequences. *Nature* **607**, 176–184 (2022).
42. Takahashi, R. et al. Defining transcriptional signatures of human hair follicle cell states. *J. Invest. Dermatol.* **140**, 764–773 (2020).
43. Kim, D. S. et al. The dynamic, combinatorial cis-regulatory lexicon of epidermal differentiation. *Nat. Genet.* **53**, 1564–1576 (2021).
44. Rubin, A. J. et al. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nat. Genet.* **49**, 1522–1528 (2017).
45. Segre, J. A., Bauer, C. & Fuchs, E. Klf4 is a transcription factor required for establishing the barrier function of the skin. *Nat. Genet.* **22**, 356–360 (1999).
46. Lien, W.-H. et al. In vivo transcriptional governance of hair follicle stem cells by canonical Wnt regulators. *Nat. Cell Biol.* **16**, 179–190 (2014).
47. Jave-Suarez, L. F., Winter, H., Langbein, L., Rogers, M. A. & Schweizer, J. HOXC13 is involved in the regulation of human hair keratin gene expression. *J. Biol. Chem.* **277**, 3718–3726 (2002).
48. Moll, R., Divo, M. & Langbein, L. The human keratins: biology and pathology. *Histochem. Cell Biol.* **129**, 705–733 (2008).
49. Fuchs, E. & Green, H. Changes in keratin gene expression during terminal differentiation of the keratinocyte. *Cell* **19**, 1033–1042 (1980).
50. Lopez, R. G. et al. C/EBPalpha and beta couple interfollicular keratinocyte proliferation arrest to commitment and terminal differentiation. *Nat. Cell Biol.* **11**, 1181–1190 (2009).
51. Qu, J. et al. Mutant p63 affects epidermal cell identity through rewiring the enhancer landscape. *Cell Rep.* **25**, 3490–3503 (2018).
52. Carroll, D. K. et al. p63 Regulates an adhesion programme and cell survival in epithelial cells. *Nat. Cell Biol.* **8**, 551–561 (2006).
53. Charest, J. L., Jennings, J. M., King, W. P., Kowalczyk, A. P. & García, A. J. Cadherin-mediated cell-cell contact regulates keratinocyte differentiation. *J. Invest. Dermatol.* **129**, 564–572 (2009).
54. Fortunel, N. O. et al. KLF4 inhibition promotes the expansion of keratinocyte precursors from adult human skin and of embryonic-stem-cell-derived keratinocytes. *Nat. Biomed. Eng.* **3**, 985–997 (2019).
55. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).
56. Garza, L. A. et al. Bald scalp in men with androgenetic alopecia retains hair follicle stem cells but lacks CD200-rich and CD34-positive hair follicle progenitor cells. *J. Clin. Invest.* **121**, 613–622 (2011).
57. Morris, R. J. et al. Capturing and profiling adult hair follicle stem cells. *Nat. Biotechnol.* **22**, 411–417 (2004).
58. Horsley, V., Aliprantis, A. O., Polak, L., Glimcher, L. H. & Fuchs, E. NFATc1 balances quiescence and proliferation of skin stem cells. *Cell* **132**, 299–310 (2008).
59. Lien, W.-H. et al. Genome-wide maps of histone modifications unwind in vivo chromatin states of the hair follicle lineage. *Cell Stem Cell* **9**, 219–232 (2011).
60. Erjavec, S. O. et al. Whole exome sequencing in alopecia areata identifies rare variants in KRT82. *Nat. Commun.* **13**, 800 (2022).
61. Jaks, V. et al. Lgr5 marks cycling, yet long-lived, hair follicle stem cells. *Nat. Genet.* **40**, 1291–1299 (2008).
62. Islam, N., Leung, P. S. C., Huntley, A. C. & Gershwin, M. E. The autoimmune basis of alopecia areata: a comprehensive review. *Autoimmun. Rev.* **14**, 81–89 (2015).
63. Gilhar, A., Etzioni, A. & Paus, R. Alopecia areata. *N. Engl. J. Med.* **366**, 1515–1525 (2012).
64. Huelsken, J., Vogel, R., Erdmann, B., Cotsarelis, G. & Birchmeier, W. beta-Catenin controls hair follicle morphogenesis and stem cell differentiation in the skin. *Cell* **105**, 533–545 (2001).
65. Andl, T., Reddy, S. T., Gaddapara, T. & Millar, S. E. WNT signals are required for the initiation of hair follicle development. *Dev. Cell* **2**, 643–653 (2002).
66. Van Mater, D., Kolligs, F. T., Dlugosz, A. A. & Fearon, E. R. Transient activation of beta -catenin signaling in cutaneous keratinocytes is sufficient to trigger the active growth phase of the hair cycle in mice. *Genes Dev.* **17**, 1219–1224 (2003).
67. Ito, M. et al. Wnt-dependent de novo hair follicle regeneration in adult mouse skin after wounding. *Nature* **447**, 316–320 (2007).
68. Gat, U., DasGupta, R., Degenstein, L. & Fuchs, E. De novo hair follicle morphogenesis and hair tumors in mice expressing a truncated beta-catenin in skin. *Cell* **95**, 605–614 (1998).
69. DasGupta, R. & Fuchs, E. Multiple roles for activated LEF/TCF transcription complexes during hair follicle development and differentiation. *Development* **126**, 4557–4568 (1999).
70. Clevers, H. & Nusse, R. Wnt/β-catenin signaling and disease. *Cell* **149**, 1192–1205 (2012).
71. Keyes, B. E. et al. Nfatc1 orchestrates aging in hair follicle stem cells. *Proc. Natl Acad. Sci. USA* **110**, E4950–E4959 (2013).
72. Adam, R. C. et al. Temporal layering of signaling effectors drives chromatin remodeling during hair follicle stem cell lineage progression. *Cell Stem Cell* **22**, 398–413 (2018).
73. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
74. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
75. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
76. Hodgins, M. B. et al. Androgen receptors in dermal papilla cells of scalp hair follicles in male pattern baldness. *Ann. N. Y. Acad. Sci.* **642**, 448–451 (1991).
77. Midorikawa, T., Chikazawa, T., Yoshino, T., Takada, K. & Arase, S. Different gene expression profile observed in dermal papilla cells related to androgenic alopecia by DNA macroarray analysis. *J. Dermatol. Sci.* **36**, 25–32 (2004).
78. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
79. Weeks, E. M. et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. Preprint at *medRxiv* https://doi.org/10.1101/2020.09.08.20190561 (2020).
80. Taylor, K. E., Ansel, K. M., Marson, A., Criswell, L. A. & Farh, K. K.-H. PICS2: next-generation fine mapping via probabilistic identification of causal SNPs. *Bioinformatics* **37**, 3004–3007 (2021).
81. Sliz, E. et al. Uniting biobank resources reveals novel genetic pathways modulating susceptibility for atopic dermatitis. *J. Allergy Clin. Immunol.* **149**, 1105–1112 (2022).
82. Elsaesser, H., Sauer, K. & Brooks, D. G. IL-21 is required to control chronic viral infection. *Science* **324**, 1569–1572 (2009).

83. Fröhlich, A. et al. IL-21R on T cells is critical for sustained functionality and control of chronic viral infection. *Science* **324**, 1576–1580 (2009).
84. Ren, H. M., Lukacher, A. E., Rahman, Z. S. M. & Olsen, N. J. New developments implicating IL-21 in autoimmune disease. *J. Autoimmun.* **122**, 102689 (2021).
85. Corces, M. R. et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).
86. Sheng, X. et al. Mapping the genetic architecture of human traits to cell types in the kidney identifies mechanisms of disease and potential treatments. *Nat. Genet.* **53**, 1322–1333 (2021).
87. Turner, A. W. et al. Single-nucleus chromatin accessibility profiling highlights regulatory mechanisms of coronary artery disease risk. *Nat. Genet.* **54**, 804–816 (2022).
88. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**, e1003711 (2014).
89. Shrikumar, A., Prakash, E. & Kundaje, A. GkmExplain: fast and accurate interpretation of nonlinear gapped k-mer SVMs. *Bioinformatics* **35**, i173–i182 (2019).
90. Soderquest, K. et al. Genetic variants alter T-bet binding and gene expression in mucosal inflammatory disease. *PLoS Genet.* **13**, e1006587 (2017).
91. Cheung, H. et al. Accessory protein-like is essential for IL-18-mediated signaling. *J. Immunol.* **174**, 5351–5357 (2005).
92. Konishi, H. et al. IL-18 contributes to the spontaneous development of atopic dermatitis-like inflammatory skin lesion independently of IgE/stat6 under specific pathogen-free conditions. *Proc. Natl Acad. Sci. USA* **99**, 11340–11345 (2002).
93. Petridis, C. et al. Genome-wide meta-analysis implicates mediators of hair follicle development and morphogenesis in risk for severe acne. *Nat. Commun.* **9**, 5075 (2018).
94. Xu, M. et al. WNT10A mutation causes ectodermal dysplasia by impairing progenitor cell proliferation and KLF4-mediated differentiation. *Nat. Commun.* **8**, 15397 (2017).
95. Liao, C.-P., Booker, R. C., Morrison, S. J. & Le, L. Q. Identification of hair shaft progenitors that create a niche for hair pigmentation. *Genes Dev.* **31**, 744–756 (2017).
96. Evellin, S. et al. FOSL1 controls the assembly of endothelial cells into capillary tubes by direct repression of αv and β3 integrin transcription. *Mol. Cell. Biol.* **33**, 1198–1209 (2013).
97. Jang, S. I., Karaman-Jurukovska, N., Morasso, M. I., Steinert, P. M. & Markova, N. G. Complex interactions between epidermal POU domain and activator protein 1 transcription factors regulate the expression of the profilaggrin gene in normal human epidermal keratinocytes. *J. Biol. Chem.* **275**, 15295–15304 (2000).
98. Sugihara, T. M., Kudryavtseva, E. I., Kumar, V., Horridge, J. J. & Andersen, B. The POU domain factor Skn-1a represses the keratin 14 promoter independent of DNA binding. A possible role for interactions between Skn-1a and CREB-binding protein/p300. *J. Biol. Chem.* **276**, 33036–33044 (2001).
99. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* **39**, 825–835 (2021).
100. Chen, A. F. et al. NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat. Methods* **19**, 547–553 (2022).
101. Chew, E. G. Y. et al. Differential expression between human dermal papilla cells from balding and non-balding scalps reveals new candidate genes for androgenetic alopecia. *J. Invest. Dermatol.* **136**, 1559–1567 (2016).
102. Kitagawa, T. et al. Keratinocyte growth inhibition through the modification of Wnt signaling by androgen in balding dermal papilla cells. *J. Clin. Endocrinol. Metab.* **94**, 1288–1294 (2009).

Methods

Sample acquisition and patient consent

Primary human scalp samples were obtained either in the form of 4 mm punch biopsies or from excess discarded scalp tissue from patients undergoing dermatological surgeries (surgical 'dogears'). Samples were collected from either Stanford University or Santa Clara Valley Medical Center with Stanford University Institutional Review Board approval, and all patients provided written informed consent. Following collection, samples were stored in 1× PBS at 4 °C until dissociation and downstream processing. Samples were stored no longer than 5 h before dissociation. A 4 mm punch was performed on surgical dogear samples before proceeding with sample dissociation.

Single-cell dissociation and fluorescent activated cell sorting

Scalp punch biopsies were rinsed with ice-cold 1× PBS and then lightly diced into 1–2-mm pieces with a sterile razor blade. Diced samples were then dissociated using the Miltenyi Biotec Human Whole Skin Dissociation Kit (catalog no. 130-101-540) according to the manufacturer's directions. Briefly, samples were incubated in 0.5 ml of dissociation solution containing the indicated volumes of enzymes P, A and D for 3 h at 37 °C. Following incubation, 0.5 ml of ice-cold RPMI 1640 with 10% FBS was added to each sample and samples were then mechanically dissociated using the gentleMACs dissociator with the 'h_skin_01' program. Following dissociation, samples were briefly centrifuged and then filtered through a 70 µm cell strainer. The dissociation tube was washed with additional ice-cold medium and samples were then centrifuged for 10 min at 300*g* in a swinging-bucket centrifuge. After aspiration of supernatant, samples were either resuspended in 0.1 ml of BamBanker freezing medium (Wako Chemicals, catalog no. 302-14681) and cryopreserved at −80 °C or subjected immediately to staining for fluorescent activated cell sorting (FACS). We did not observe any systematic clustering differences between samples that had been sorted immediately after dissociation and those that had been cryopreserved, even without Harmony or other batch correction methods (Extended Data Fig. 1d–g and Comparison of fresh versus cryopreserved samples in Supplementary Methods). Cryopreserved samples included C_SD4, C_SD5, C_SD6, C_SD7, AA7 and AA8. All remaining samples were sorted immediately after dissociation without cryopreservation.

Cells were stained with anti-CD90 PE Cy7 (BD Pharmingen, no. 561558) for 30 min at 4 °C in FACS staining buffer (PBS with 0.5% bovine serum albumin) then washed with FACS buffer. Live cells were distinguished using the LIVE/DEAD Fixable Aqua Dead Cell Stain Kit (ThermoFisher, catalog no. L34957) according to the manufacturer's directions. For cryopreserved samples, cells were thawed at 37 °C for 3 min, resuspended in RPMI + 10% FBS and washed with FACS buffer before staining. Aqua-negative live cells were sorted as fibroblast ($CD90^+$) and nonfibroblast ($CD90^-$) populations. Sorted cells were counted and the $CD90^+$ population reduced by half before recombination with the $CD90^-$ population for further processing by scATAC-seq and/or scRNA-seq.

scRNA library generation, sequencing and alignment

Following sorting, cell suspensions were centrifuged at 300*g* for 5 min at 4 °C and resuspended in 1× PBS with 0.5% bovine serum albumin. Samples were counted using a hemocytometer, and the required volume of cells was aliquoted for generation of scRNA-seq libraries. scRNA-seq libraries were prepared using the 10X Genomics Chromium Next GEM Single Cell 3′ RNA v.3.1 protocol, targeting 8,000 cells per sample. Completed libraries were sequenced on an Illumina NextSeq 550 platform with 28/8/0/91 base-pair cycles. Raw sequencing data were converted to fastq format using the command 'cellranger mkfastq' (10X Genomics, v.3.1.0). Resulting fastq files were then aligned to the hg38 reference genome (cellranger-GRCh38-3.0.0) and quantified using the command 'cellranger count'.

scATAC library generation, sequencing and alignment

After the required number of sorted cells were aliquoted for generation of scRNA-seq libraries, the remaining sample volume was used for generation of scATAC-seq libraries. The remaining cell volume was used to prepare nuclei according to the 10X ATAC nuclei isolation protocol for 'low cell input nuclei isolation' (CG000169, Rev B). scATAC-seq libraries were prepared using the 10X Genomics Chromium Next GEM Single Cell ATAC v.1.1 protocol, targeting 6,000 cells per sample. Completed libraries were sequenced on an Illumina NextSeq 550 platform with 33/8/16/33 base-pair cycles. Raw sequencing data were converted to fastq format using the command 'cellranger-atac mkfastq' (10X Genomics, v.1.2.0). Resulting fastq files were aligned to the hg38 reference genome (cellranger-atac-GRCh38-1.2.0) and quantified using the command 'cellranger-atac count'.

scRNA-seq quality control, dimensionality reduction and clustering

Unless otherwise indicated, all subsequent analyses were performed using R v.4.0.2. Following alignment and quantification, scRNA-seq count matrices were further processed using the Seurat R package (v.4.0.4)[29]. Initial quality control was performed on each sample independently. First, cells were removed if they had fewer than 200 genes expressed, fewer than 1,000 unique sequenced reads (unique molecular identifiers) or greater than 20% of counts corresponding to mitochondrial genes. Doublets were identified and removed using the 'DoubletFinder' R package (v.2.0.3)[103]. Because we observed evidence of ambient RNA contamination in several samples, we used the 'DecontX' method in the 'celda' R package (v.1.6.1) to estimate and remove contaminating ambient RNA from each cell[104]. After carrying out each of these quality-control steps, samples were merged into a single Seurat object for clustering. Decontaminated count data were scaled to 10,000 and then $\log_2$ normalized.

We adapted an iterative latent semantic indexing (LSI) approach to dimensionality reduction and clustering[19]. First we removed mitochondrial genes, sex chromosome genes and genes associated with cell cycle (Seurat's 'cc.genes') to minimize sample batch effects in variable feature selection. Next we identified the top 4,000 variable genes across all cells and calculated term frequency–inverse document frequency (TF–IDF) for these variable genes. We performed singular value decomposition (SVD) on the TF–IDF matrix and used the first 25 dimensions as input into Seurat's sharing-nearest-neighbor clustering with an initial resolution of 0.2. Counts from single cells in each of these resulting clusters were summed, transformed with the logCPM transformation 'edgeR::cpm(mat, log=TRUE, prior.count=3)' and then used to identify the top 4,000 variable genes for the next round of LSI. TF–IDF transformation followed by SVD was again performed using the new set of 4,000 variable genes, and clustering was repeated with an increased resolution of 0.4. The previously described variable gene selection, TF–IDF transformation and SVD were performed once more and clustering was repeated with a final resolution of 0.8. The 25 LSI dimensions from the final round were used to generate two-dimensional representations using the uniform manifold approximation and projection (UMAP) implementation from the Seurat and 'uwot' R packages (v.1.0.10; n.neighbors=50, min.dist=0.5, metric=cosine).

This initial clustering procedure identified 29 clusters. After identification of marker genes for each cluster using Seurat's 'FindAllMarkers' function and inspection of sample representation of each cluster, we identified a small number of clusters that appeared to be doublet clusters (clusters 18, 26, 28 and 29). Each of these clusters was composed entirely, or nearly entirely, of a single sample, did not have unique marker genes when compared with other clusters or expressed biologically incompatible combinations of marker genes. We removed all cells belonging to these clusters and repeated the previously described iterative LSI clustering procedure on the remaining cells, this time using clustering resolutions of 0.1, 0.3 and 0.6 for the three rounds.

We regenerated UMAP with the same parameters used previously; this final filtered and clustered dataset contained 21 clusters. Visualization of gene expression on UMAP representations was smoothed using the MAGIC diffusion algorithm[105]. To minimize the risk of 'oversmoothing' expression patterns, the application of MAGIC was restricted to data visualization[106].

scATAC-seq quality control, dimensionality reduction and clustering

Following alignment, ATAC-seq fragment data were further processed using the 'ArchR' R package (v.1.0.1)[18]. For each cell we computed the number of unique sequenced fragments and transcription start site (TSS) enrichment, which serves as a signal-to-noise metric for ATAC-seq data[19]. We plotted all barcoded droplets on a scatter plot using these two metrics and observed that, while some samples had a clear separation between true cells (high TSS and number of unique fragments) others had a more continuous distribution between true cells and droplets containing contamination-free DNA (lower number of unique fragments and lower TSS enrichment). To label droplets as probable true cells we used an expectation maximization-based approach. For each sample we used the 'mclust' R package (v.5.4.7) to fit up to four two-dimensional gaussians to $\log_{10}$ nFragments (number of ATAC-seq fragments) by TSS enrichment joint distribution ('Mclust(df, G=2:4, modelNames=VVV'). Cells classified as originating from the Gaussian with the greatest mean TSS enrichment were labeled as true cells while the remaining droplets were filtered from the project. Cells with a TSS of below five or nFragments below 1,000 were all filtered from the project, regardless of their expectation maximization classification label. This approach was functionally similar to setting a hard filter for TSS and nFragments for samples that had clearly defined true cell populations, but enabled exclusion of more contaminating droplets for samples that had a less clearly defined population of true cells (Extended Data Fig. 1a).

Following initial quality control, doublets were identified and filtered using the ArchR 'addDoubletScores' and 'filterDoublets' functions, with a filter ratio of 1. We then used ArchR's implementation of iterative LSI dimensionality reduction using the 'addIterativeLSI' function with 50,000 variable features and 25 dimensions. We identified clusters using the ArchR function 'addClusters' with a resolution of 0.6 and then generated a two-dimensional representation of the data using the 'addUMAP' ArchR function, with nNeighbors=50, minDist=0.4 and metric=cosine. This initial clustering procedure identified 22 clusters. We identified marker genes for each cluster using the 'getMarkerFeatures' function with the accessibility around each gene (the 'Gene Activity Score') as a proxy for gene expression[18]. We identified a small number of poor-quality clusters (clusters 7, 13, 15 and 18). These clusters were composed entirely, or nearly entirely, from a single sample, did not have unique marker genes, had systematically lower TSS enrichment or were enriched for high doublet scores. Cells belonging to these clusters were removed from the project, and dimensionality reduction and clustering was repeated on the filtered project using 50,000 variable features and 50 dimensions for 'addIterativeLSI', and then a resolution of 0.7 for 'addClusters'. We regenerated the UMAP using nNeighbors=60, minDist=0.6 and metric=cosine. This final filtered and clustered dataset contained 22 clusters. Visualization of gene activity scores on UMAP was similarly smoothed using the MAGIC algorithm[105]. Smoothed data were used only for visualization purposes.

Subclustering of major cell types

To improve identification of rare cell types we subclustered several major cell groups from the full scRNA- and scATAC-seq datasets. For scRNA-seq data, cluster labels were assigned based on known cell type markers (Fig. 1e, 'NamedClust'). Cluster labels for scATAC-seq data were assigned in a similar manner, using gene activity scores as a proxy for gene expression (Fig. 1f, 'NamedClust'). For example, basal keratinocyte clusters exhibited high gene activity and expression of the basal keratin *KRT15* (ref. 107), hair follicle keratinocyte clusters exhibited high gene activity and expression of the TF SOX9 (ref. 108), T lymphocyte clusters exhibited high gene activity and expression of the cell surface marker *CD3D* and fibroblast clusters exhibited high gene activity and expression of the cell surface marker *THY1* (ref. 109). We observed a relatively large scRNA-seq cluster expressing high levels of mast cell markers, including beta tryptases (*TPSB1/2*) and *HPGD*[110–112], but did not observe a corresponding scATAC-seq cluster, perhaps due to the tendency for granulocyte chromatin to spontaneously decondense during nuclear isolation[113,114]. After labeling clusters in each modality we subclustered major cell types in each dataset (keratinocytes, fibroblasts, endothelial cells, T lymphocytes and myeloid lineage cells; Extended Data Fig. 3a–c). See Supplementary Methods for clustering details and information about peak calling across subclustered datasets.

Integration of scRNA- and scATAC-seq datasets

Starting with the full dataset, we matched each scATAC-seq cell with its closest corresponding scRNA-seq cell using a previously described multimodal dataset integration technique based on canonical correlation analysis. Specifically we used the ArchR function 'addGeneIntegrationMatrix', which employs Seurat's 'FindTransferAnchors' function to integrate datasets[18,29]. We then used nGenes=3,000 for integration of the full dataset. We computed the Jaccard index between scRNA- and scATAC-seq cluster labels of integrated metacells and observed high correspondence (Extended Data Fig. 3e). Furthermore we identified the same major cell types in each dataset, with the exception of mast cells, which were observed only in the scRNA-seq dataset (Fig. 1c,d). We repeated this integration procedure for each of the previously described subclustered datasets (keratinocytes, fibroblasts, endothelial cells, T lymphocytes and myeloid lineage cells) using nGenes=2,000. For each subclustered dataset we similarly observed high correspondence between scRNA- and scATAC-seq-derived cluster labels (Extended Data Fig. 3d).

Linkage of gene-regulatory elements to gene expression using integrated datasets

CREs were linked to their potential gene targets ('peak-to-gene links') using a correlation-based approach[115]. This procedure involves the creation of up to 500 partially overlapping pseudobulks of 100 *k*-nearest-neighbors integrated single cells ('low-overlapping cell aggregates'). The peak counts of each pseudobulk are summed, as are the gene expression counts of the corresponding integrated scRNA-seq transcript profiles. Candidate peak–gene pairs are then identified by first associating peaks within a genomic distance of 250 kb to the TSS of each gene and then computing the Pearson correlation coefficient of $\log_2$-normalized accessibility and gene expression counts. This procedure was carried out using the 'addPeak2GeneLinks' function in ArchR[18]. High-confidence peak-to-gene links were obtained by retaining those with a Pearson correlation coefficient of >0.5.

Because this correlation procedure is dependent on dimensionality reduction of the particular dataset used, and because dimensionality reduction in turn is dependent on variable gene selection across the full dataset, we found that using the entire scalp dataset for this analysis robustly identified peak-to-gene links corresponding to regulatory interactions defining major cell types (for example, keratinocytes versus T cells), but was less efficient at recovering regulatory interactions between more closely related cell subtypes (for example, specific hair follicle keratinocyte subsets; Fig. 2b). To increase our sensitivity in detection of peak-to-gene linkages distinguishing more fine-grained cell subtypes, we repeated the previously described peak-to-gene linking procedure on each subclustered major cell type using only the subset of peaks relevant to a specific subclustered dataset as described above. To create a consensus peak-to-gene link set we combined all identified peak-to-gene links from the full dataset and

each subclustered dataset, sorted peak-to-gene links by their Pearson correlation coefficients and removed duplicate peak-to-gene links, resulting in a consensus peak-to-gene link set of 146,088.

Validation of inferred peak-to-gene linkages using conservation and ABC model predictions

Following identification of peak-to-gene linkages on the full scalp dataset and on each of the subclustered datasets (keratinocytes, fibroblasts, endothelial, T lymphocytes and myeloid), peak-to-gene links were validated using two strategies. First we used the 'gscores' function from the 'GenomicScores' R package (v.2.2.0) to compute mean phastCons 100-way vertebrate evolutionary conservation scores for peaks linked in the full dataset and in each of the subclustered datasets, as well as for peaks that were not linked in any analysis. For each group of peak-to-gene linkages (that is, the full dataset linkages and each of the subclustered datasets) we used a Wilcoxon rank-sum test to compare linked and unlinked peaks (Extended Data Fig. 5c).

Second, we compared our peak-to-gene linkages with predicted enhancer–gene interactions from a recently published ABC dataset generated from 131 human tissues and cell types[31]. We downloaded the full dataset of all 131 tissues (https://www.engreitzlab.org/resources/) and cell types and converted enhancer coordinates from hg19 to hg38 using the 'liftover' function from the 'rtracklayer' R package (v.1.50.0). For validation of our peak-to-gene link inferences we required both that the linked peak had to overlap an enhancer region in the ABC model dataset and that the corresponding linked gene had to match. We used all possible peak-to-gene linkages (that is, all peak–gene pairs separated by <250 kb) as background to test for enrichment of ABC model-predicted enhancer-gene links in our inferred peak-to-gene links (Extended Data Fig. 5d, top bar). To account for the skewed length distribution for inferred peak-to-gene links compared with all possible peak-to-gene links, we also compared the enrichment of ABC model-predicted enhancer–gene links in inferred peak-to-gene links with a distance-matched background set of peak-to-gene links (Extended Data Fig. 5d, second bar). To do this we first computed the distance between gene promoter and linked peak for all inferred peak-to-gene links. We divided these distances into 20 contiguous equal-sized bins and assigned background peak-to-gene links to each of these. We sampled 146,088 peaks from the background peak-to-gene link set while matching the distance distribution of the inferred peak-to-gene links, and then calculated the number of background peak-to-gene links that overlapped ABC enhancer–gene pair predictions. We repeated this sampling procedure 100 times and used the mean number of overlapping background peak-to-gene links to calculate the enrichment of ABC enhancer–gene pair predictions in our inferred peak-to-gene linkages using a hypergeometric enrichment test. We calculated the enrichment of ABC model-predicted enhancer–gene pairs in inferred peak-to-gene linkages for linkages identified on the full, nonsubclustered dataset ('full scalp'), and for each of the subclustered datasets (Extended Data Fig. 5d, bottom six bars).

Identification and analysis of HRGs

Following creation of our consensus peak-to-gene link set we ranked all expressed genes by their number of peak-to-gene links, finding that a subset of genes had notably more peak-to-gene linkages than others. We set a cutoff near the inflection point ('elbow') of 20 linked peaks per gene to identify a subset of HRGs, 1,739 genes (Fig. 2c). We compared these HRGs with a dataset of previously identified superenhancer-associated genes from a variety of tissues and cell lines[35]. We also compared these HRGs with the human homologs of previously identified mouse hair follicle-associated superenhancer genes[36]. We calculated the enrichment of superenhancer-associated genes from various tissues in our set of 1,739 scalp HRGs using a hypergeometric enrichment test (Fig. 2d). We additionally compared HRGs with previously identified domains of regulatory chromatin-associated genes following conversion of mouse genes to their human orthologs[33]. We calculated the significance of this overlap using a one-sided Fisher's exact text. In Fig. 2e we list two of the top HRGs for each *k*-means cluster to the right of the peak-to-gene heatmap. We performed GO enrichment analyses on the top 200 genes ranked by number of peak-to-gene linkages for each of the *k*-means clusters using the topGO (v.2.42.0) R package[116]. For this and all subsequent GO term enrichment analyses we use the topGO 'weight01' method for calculation of enrichment *P* values. Because *P* values calculated using this method are conditioned on neighboring terms in the GO topology, term tests are not independent and multiple testing theory does not directly apply. As the authors of the package suggest, we therefore do not apply further multiple hypothesis testing correction. See section 6.2 of the topGO manual for further details: http://www.bioconductor.org/packages/release/bioc/vignettes/topGO/inst/doc/topGO.pdf.

Analysis of modular enhancer usage in HRGs

To visualize the heterogeneity of enhancer usage between cell types expressing the same gene, we generated 246 pseudobulks of *k*-nearest neighbor cells with $k = 250$. To plot peak using pseudobulk heatmaps we normalized pseudobulk accessibility by summing the peak counts for each pseudobulk, depth normalization- and $\log_2$-transformed counts data and then quantile normalization using the 'normalize.quantiles' function from the 'preprocessCore' R package (v.1.52.0). For each individual HRG we then calculated the *z*-score for the normalized accessibility of each linked peak across all pseudobulk samples. Peaks were ordered using hierarchical clustering, with euclidean distance as the dissimilarity measure and complete linkage as the agglomeration method. For scatter plots comparing pseudobulk-linked peak accessibility with linked gene expression we calculated the mean normalized integrated gene expression for each pseudobulk sample and applied $\log_2$ transformation. To calculate total linked chromatin accessibility we summed the depth-normalized counts of linked peaks for a given gene and then applied $\log_2$ transformation. Pseudobulk labels in both heatmaps and scatter plots were determined by selection of the most frequent cluster label from the 250 cells comprising each pseudobulk.

ChromVAR motif analysis

We used chromVAR (v.1.12.0) to measure enrichment of TF motifs in accessible chromatin across single cells[117]. Specifically, we first used the ArchR function 'addMotifAnnotations' to identify all cisbp motif matches in the peak set, used 'addBgdPeaks' to identify a set of genomic copy- and accessibility-matched background peaks and then used the 'addDeviationsMatrix' function to calculate motif deviation *z*-scores for each cisbp motif.

Trajectory analysis for interfollicular and hair follicle keratinocytes

For analysis of epigenetic and gene-regulatory dynamics over the course of differentiation of interfollicular keratinocytes we used the R package 'slingshot' (v.1.8.0)[118]. To apply slingshot to our integrated scATAC-seq data for interfollicular keratinocytes we used the ArchR function 'addSlingShotTrajectories' with 'embedding=UMAP', restricting available clusters to interfollicular keratinocyte clusters (Basal.Kc_1, Spinous.Kc_1 and Spinous.Kc_2) and designating the basal keratinocyte cluster as the origin of differentiation. To identify TF regulator candidates for this differentiation trajectory we used two complementary approaches. First, using all keratinocyte clusters, we calculated the correlation between a given TF's chromVAR motif deviation *z*-scores and that same TF's integrated gene expression across low-overlapping cell aggregates. Correlating these measures can help distinguish which specific TF in a larger TF family is responsible for the motif activity observed in a given cell type. These TF correlations were plotted against the maximum difference in chromVAR motif *z*-scores between clusters, highlighting TFs exhibiting more dynamic regulatory activity across

cell types (Extended Data Fig. 6e). To identify TFs more specific to the interfollicular keratinocyte differentiation trajectory, we selected integrated gene expression values and chromVAR deviation scores along the previously determined slingshot differentiation trajectory using the ArchR function 'getTrajectory' with groupEvery=1.5. We then correlated these trajectories using the ArchR function 'correlateTrajectories' with default parameters.

For analysis of the differentiation trajectory of the inferior segment of the hair follicle we further subclustered these cells as described above. We used the keratinocyte scATAC-seq clusters Inf.Segment_1, Inf.Segment_2 and Matrix and scRNA-seq cluster Inf.Segment. For the subclustered scRNA-seq datasets we used 1,500 variable genes, 20 SVD dimensions and a clustering resolution of 0.2 in the first round, followed by a clustering resolution of 0.4 in the final round. To generate UMAPs for the subclustered scRNA-seq dataset we used n.Neighbors=20, min.Dist=0.1 and metric=cosine. For scATAC-seq subclustering we again used ArchR's implementation of iterative LSI dimensionality reduction. We used 25,000 variable features, 30 dimensions and 0.4 resolution for clustering. To generate UMAPs for the subclustered scATAC-seq data we used n.Neighbors=20, min.Dist=0.1 and metric cosine. We reintegrated these subclustered datasets and reidentified peak-to-gene linkages as described above. This hair follicle inferior segment subclustering was used only for analysis of hair follicle differentiation trajectory (Fig. 4), and the peak-to-gene links identified on this dataset were not used for any other analyses. Identification of TF regulators for the hair follicle differentiation trajectory was performed using slingshot as described above, providing the HFSC, Migratory, Shaft_1, Shaft_2 and Matrix clusters as being involved in the trajectory and designating the HFSC cluster as the origin.

Identification of potential regulatory target genes of TF regulators

To identify potential gene targets of a TF we calculated the Pearson correlation coefficient between the candidate TF regulator's chromVAR motif activity and the integrated gene expression of all expressed genes. Next we calculated a linkage score for each gene and TF pair. This score is calculated by identification of all peak-to-gene links for that gene for which the linked peak contains an instance of the candidate TF motif, and then summing the product of the squared peak-to-gene linkage correlation with the the motif score:

$$\mathrm{LS_g} = \sum_{k=1}^{n} R_k^2 \mathrm{MS}_k$$

where $\mathrm{LS_g}$ is the linkage score of gene g, n is the number of linked peaks for gene g, R is the peak-to-gene Pearson correlation coefficient for peak k and MS_k is the motif score for the motif occurring in peak k. The linkage score is thus higher for genes that have multiple linked peaks containing the TF motif, have more strongly correlated linked peaks containing the TF motif and/or have linked peaks that contain highly confident instances of the motif. See Supplementary Methods for additional details.

Differential cell type abundance testing using Milo

We used the 'miloR' R package (v.1.1.0) to perform k-nearest neighbor graph-based differential cell type abundance testing between alopecia areata and unaffected control samples (C_PB and C_SD)[55]. Although miloR was originally designed to be applied to scRNA-seq data, the algorithm depends only on having a cell–cell similarity structure to the dataset and thus can be similarly applied to scATAC-seq data. We applied miloR to our integrated scATAC-seq data by creating a 'SingleCellExperiment' R object from the counts matrix of our keratinocyte ArchR project, and then used ArchR LSI dimensionality reduction as the reduced.dim input for miloR in the 'buildGraph' function. For comparison of differential abundance across all keratinocytes we used only samples that had at least 50 cells in the subclustered dataset, k = 30 for the 'buildGraph' function and prop=0.1 for the 'makeNhoods' function. For comparison of differential abundance across only the lower, cycling portion of hair follicle keratinocytes (Fig. 4e) we used only samples that had at least ten cells in the subclustered dataset, k=30 for the 'buildGraph' function and prop=0.3 for the 'makeNhoods' function. We plotted differentially abundant cell neighborhoods with SpatialFDR = <0.1 using the 'plotNhoodGraphDA' function.

LDSC using scATAC-seq data

We used LDSC (v.1.0.1) to estimate the heritability of multiple skin, hair and other traits in each high-resolution clustered cell type in our dataset[75]. Cluster-specific peak regions were used as input functional categories for LDSC. To obtain these cluster-specific peaks we first removed clusters with fewer than 40 cells in total, because these clusters generally had too few cells for identification of sufficient numbers of confident cell-type-specific peaks. For the remaining clusters we identified which peaks from the union peak set were originally identified in a given cluster by overlapping the union peak set with the MACS2 peak calls from that specific cluster. For each cluster we then retained only peaks that had been identified in no more than 25% of all clusters (nine out of a possible 36 clusters). This strategy enabled us to both filter out common 'housekeeping peaks' that are accessible in the majority of cell types while retaining peaks that are unique to, at most, a few clusters. Formatted summary statistics for partitioning can be downloaded from https://console.cloud.google.com/storage/browser/broad-alkesgroup-public-requester-pays/sumstats_formatted. We followed the recommended guidelines for cell-type-specific partitioned heritability analysis using the 1000 G EUR phase 3 population reference and the hg38 baseline model (v.2.2). We used the 'ldsc.py' script to calculate partitioned heritability for each trait in cluster-specific peak sets. We used Benjamini–Hochberg FDR correction to adjust heritability enrichment P values. See Supplementary Methods for additional details.

Analysis of fmGWAS variants

We obtained fine-mapped SNPs from multiple sources. First we downloaded a compendium of fine-mapped SNPs for 94 UK Biobank traits (www.finucanelab.org/data) and used the male pattern balding ('Balding_Type4'), body mass index and systolic blood pressure ('SBP') traits for downstream analyses[79]. Second, we downloaded precomputed PICS fine-mapped SNPs for a variety of traits in the GWAS catalog (https://pics2.ucsf.edu/Downloads/PICS2-GWAScat-2021-06-11.txt.gz)[78,80]. Details of trait definitions are available from either the UK Biobank (https://www.ukbiobank.ac.uk/) or the GWAS catalog (https://www.ebi.ac.uk/gwas/). We calculated enrichment of fine-mapped SNPs with a fine-mapping posterior probability of ≥0.01 from selected traits in the previously described cluster-specific peak sets, using one-sided Fisher's exact test with a background SNP set containing all fine-mapped SNPs (also with a fine-mapping posterior probability of ≥0.01) across all traits. Enrichment P values were adjusted using Benjamini–Hochberg FDR correction. See Supplementary Methods for additional details.

In regard to identification of genes associated with fine-mapped SNPs for selected traits, we identified those with a fine-mapping posterior probability of ≥0.01 and that overlapped a scATAC-seq peak region. Next, for each gene we identified all fine-mapped SNPs that fell within a peak linked to the expression of that gene then summed the fine-mapping posterior probability for these linked SNPs. Genes linked to a peak containing a fine-mapped SNP with a high posterior probability, or those linked to multiple linked peaks containing fine-mapped SNPs with appreciable fine-mapping posterior probability, were assumed more likely to represent genes whose expression is associated with the trait of interest. We plotted row-scaled gene expression for the top 80 genes (by total associated fine-mapping probability) in each of our high-resolution scRNA-seq clusters in a heatmap then plotted the number of linked peaks and cumulative fine-mapping posterior probability to the right of each gene.

gkm-SVM machine learning classifier training and testing

We adapted a previously published strategy for trained gkm-SVM models using scATAC-seq data[85]. See Supplementary Methods for details on model training, testing and SNP prioritization.

Statistics and reproducibility

The statistical methods and tests used in various analyses are listed in their respective figure legends or section of Methods. No statistical method was used to predetermine sample size. The authors were not blinded to patient diagnosis during sample collection or analysis. All datasets generated that did not fail experimentally (for example, overloaded sample) were included in the study. Data (in the form of individual cells) were excluded from downstream analyses if they did not pass technical quality control thresholds in the initial data-processing stage, as described in Methods.

Ethics statement

All research described complies with the ethical guidelines for human subjects research under the approved Institutional Review Board protocol at Stanford University (no. 40524) for the collection and use of human tissue samples.

Reporting summary

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

Data availability

Sequencing data generated in this study have been deposited in the Gene Expression Omnibus (GEO) with accession code GSE212450. The full scalp dataset can be explored interactively at http://shiny.scscalpchromatin.su.domains/shiny_scalp/ (ref. 119). Reference genome files for alignment of single-cell data can be downloaded from https://support.10xgenomics.com/single-cell-gene-expression/software/release-notes/build. Predicted superenhancer-associated genes from 86 human cell types and tissues were downloaded from Supplementary Table 2 of https://doi.org/10.1016/j.cell.2013.09.053 (ref. 35). Predicted superenhancer-associated genes from mouse hair follicle cell populations were downloaded from Supplementary Table 1 of https://doi.org/10.1038/nature14289 (ref. 36). The ABC dataset generated from 131 human tissues and cell types was downloaded from https://www.engreitzlab.org/resources/ (ref. 31). Differentially expressed genes identified between control human keratinocytes and keratinocytes containing a mutant, binding-incompetent form of TP63 were obtained from Supplementary Table 1d of https://doi.org/10.1016/j.celrep.2018.11.039 (ref. 51). The counts matrix from short hairpin RNA knockdown of KLF4 in human adult keratinocytes is available on GEO with accession no. GSE111786 (ref. 54). Formatted summary statistics for partitioning heritability using LDSC can be downloaded from https://console.cloud.google.com/storage/browser/broad-alkesgroup-public-requester-pays/sumstats_formatted. Fine-mapped SNPs for 94 UK Biobank traits can be downloaded from www.finucanelab.org/data (ref. 79). Precomputed PICS fine-mapped SNPs for a variety of traits from the GWAS catalog are available at https://pics2.ucsf.edu/Downloads/ (refs. 78,80). Source data are provided with this paper.

Code availability

Custom code for data processing, peak-to-gene analyses and GWAS analyses is available on Github (https://github.com/GreenleafLab/scScalpChromatin and https://doi.org/10.5281/zenodo.7915926). Our analyses also make use of published software tools, with description of their use and parameter settings available in Methods and in the custom code above where applicable.

References

103. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337 (2019).
104. Yang, S. et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).
105. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
106. Tjärnberg, A. et al. Optimal tuning of weighted kNN- and diffusion-based methods for denoising single cell genomics data. *PLoS Comput. Biol.* **17**, e1008569 (2021).
107. Lloyd, C. et al. The basal keratin network of stratified squamous epithelia: defining K15 function in the absence of K14. *J. Cell Biol.* **129**, 1329–1344 (1995).
108. Vidal, V. P. I. et al. Sox9 is essential for outer root sheath differentiation and the formation of the hair stem cell compartment. *Curr. Biol.* **15**, 1340–1351 (2005).
109. Philippeos, C. et al. Spatial and single-cell transcriptional profiling identifies functionally distinct human dermal fibroblast subpopulations. *J. Invest. Dermatol.* **138**, 811–825 (2018).
110. Schwartz, L. B., Metcalfe, D. D., Miller, J. S., Earl, H. & Sullivan, T. Tryptase levels as an indicator of mast-cell activation in systemic anaphylaxis and mastocytosis. *N. Engl. J. Med.* **316**, 1622–1626 (1987).
111. Ren, S., Sakai, K. & Schwartz, L. B. Regulation of human mast cell beta-tryptase: conversion of inactive monomer to active tetramer at acid pH. *J. Immunol.* **160**, 4561–4569 (1998).
112. Stevens, W. W. et al. Activation of the 15-lipoxygenase pathway in aspirin-exacerbated respiratory disease. *J. Allergy Clin. Immunol.* **147**, 600–612 (2021).
113. Neubert, E. et al. Chromatin swelling drives neutrophil extracellular trap release. *Nat. Commun.* **9**, 3767 (2018).
114. Sollberger, G., Tilley, D. O. & Zychlinsky, A. Neutrophil extracellular traps: the biology of chromatin externalization. *Dev. Cell* **44**, 542–553 (2018).
115. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
116. Alexa, A., Rahnenführer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).
117. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
118. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
119. Sharma, A., Akshay, A., Rogne, M. & Eskeland, R. ShinyArchR.UiO: user-friendly,integrative and open-source tool for visualization of single-cell ATAC-seq data using ArchR. *Bioinformatics* **38**, 834–836 (2021).

Acknowledgements

We thank the Stanford FACS facility and Stanford Functional Genomics Facility for technical support. Figure schematics in Figure 1a,b were created with BioRender.com. This work was supported in part by grants from the NIH (nos. 2R37-ARO54780 to A.E.O. and RM1-HG007735, UM1-HG009442, UM1-HG009436, R01- HG009909, U19-AI057266, U54HG012723, UM1HG011972, R01NS128028 and U01-HG011762 to W.J.G). W.J.G. acknowledges support as a Chan-Zuckerberg Investigator (grant nos. 2017-174468 and 2018-182817). B.O.-R. was supported in part by Stanford MSTP training grant nos. T32-GM007365 and T32-GM145402. C.W. was supported in part by a Dermatology Foundation Physician Scientist Career Development Award.

Author contributions

Conceptualization was the responsibility of B.O.-R., C.W., A.E.O. and W.J.G. Investigation of single-cell experiments was carried out by B.O.-R. and C.W. B.O.-R. performed formal analysis. Resources and sample collection were done by C.W., A.E.O., J.M.K., E.J.R. and S.Z.A. B.O.-R. and

W.J.G. wrote the original draft. All authors participated in paper review and editing. A.E.O., M.M.D. and W.J.G. supervised the project. Funding acquisition was carried out by A.E.O., M.M.D. and W.J.G.

Competing interests

W.J.G. is a consultant for 10X Genomics, Guardant Health, Quantapore, Erudio Bio. and Lamar Health and cofounder of Protillion Biosciences and is named on patents describing ATAC-seq. The other authors declare no competing interests.

Additional information

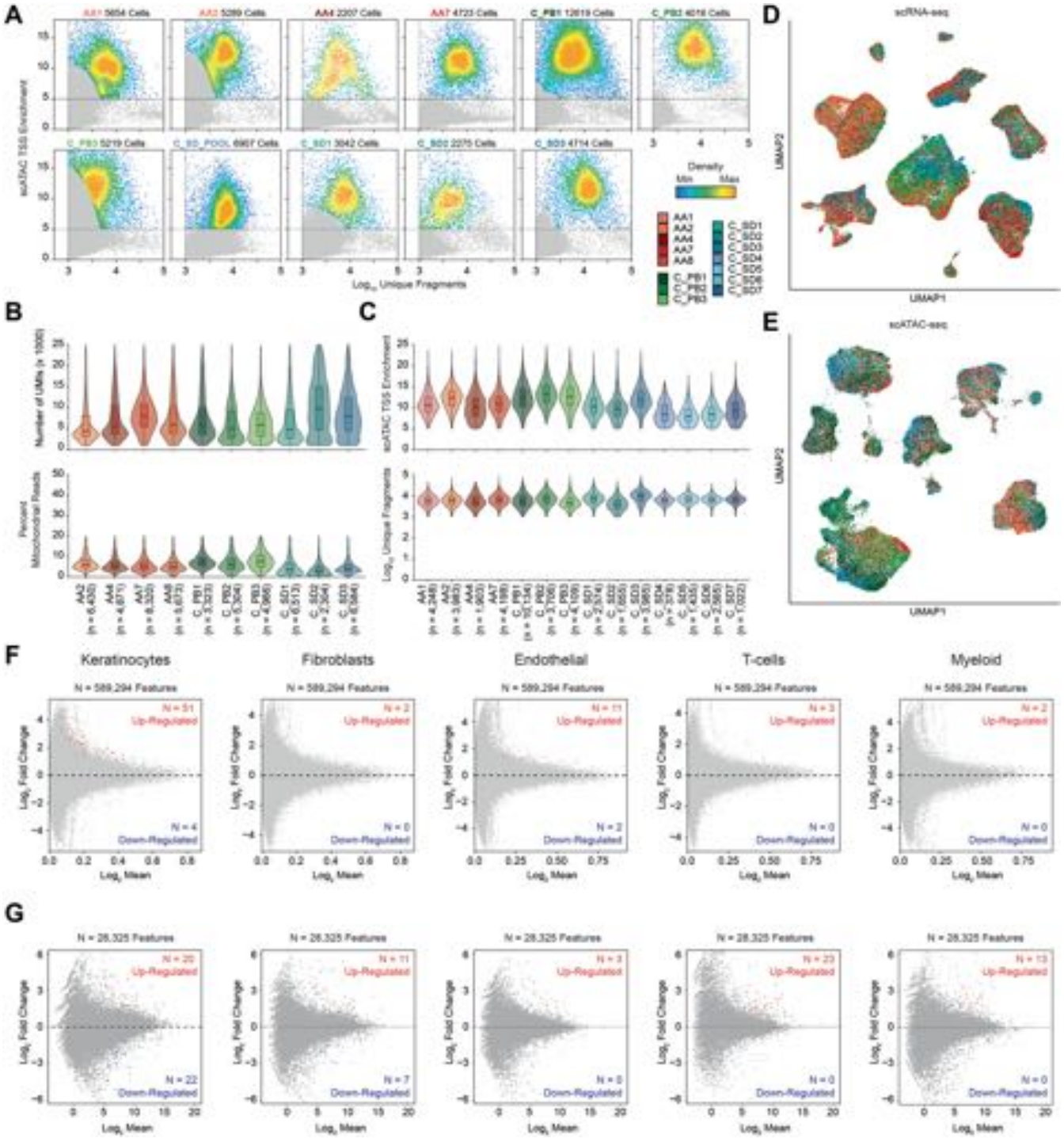
Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01445-4.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01445-4.

Correspondence and requests for materials should be addressed to William J. Greenleaf.

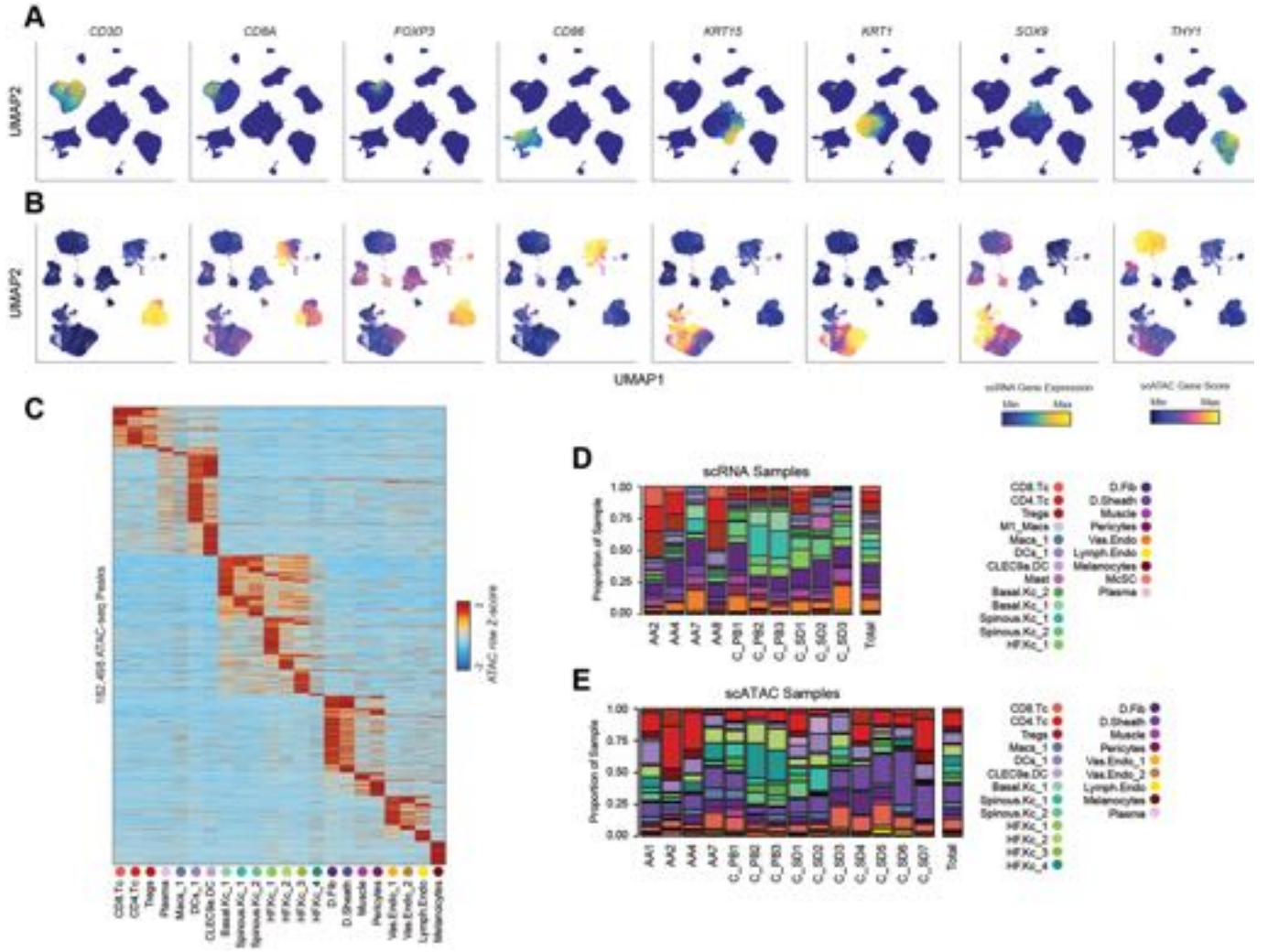
Peer review information *Nature Genetics* thank Matthias Huebenthal, Kerstin Meyer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-023-01445-4



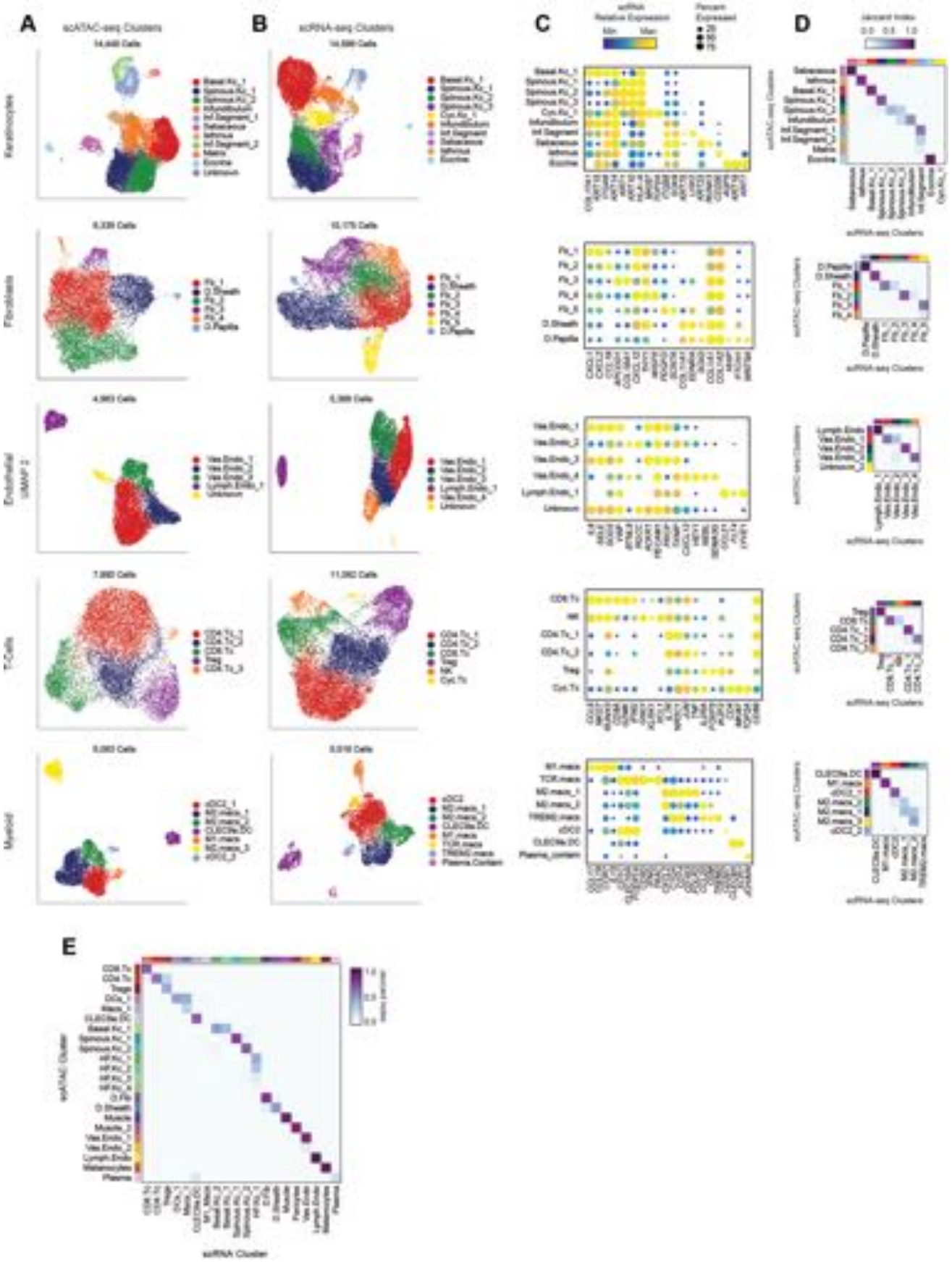
Extended Data Fig. 1 | Quality control of single cell RNA and ATAC datasets. (**a**) Scatter plots of the number of unique fragments by the transcription start site (TSS) enrichment for each of the scATAC-seq samples. Gray dots indicate cells that did not pass quality control filters (Methods). Colorbar indicates the density of points. (**b**) Violin plots of the number of unique reads (UMIs, top) and the percent of reads from mitochondrial genes (bottom) for each of the scRNA-seq samples. The inset box plot represent the median, 25th percentile and 75th percentile of the data, and whiskers represent the highest and lowest values within 1.5 times the interquartile range of the boxplot. (**c**) Violin plots of the TSS enrichment (top) and number of unique fragments (bottom) for each of the scATAC-seq samples. Box plot as in (B). (**d**) UMAP projection of full scRNA-seq dataset, colored by patient sample. (**e**) UMAP projection of full scATAC-seq dataset, colored by patient sample. (**f**) Differential scATAC-seq peaks between samples processed immediately after collection or after cryopreservation for each of the major cell groupings. Differential peaks (FDR < 0.1) are indicated by colored dots. (**g**) Differential scRNA-seq genes between samples processed immediately after collection or after cryopreservation for each of the major cell groupings. Differential genes (FDR < 0.1) are indicated by colored dots.

https://doi.org/10.1038/s41588-023-01445-4



Extended Data Fig. 2 | Annotation of single cell RNA and ATAC datasets. (**a**) UMAP projections of full scRNA-seq dataset colored by relative expression levels of representative cell compartment marker genes. (**b**) UMAP projections of full scATAC-seq dataset colored by relative gene activity scores of the same marker genes shown in (A). (**c**) Marker peaks (Wilcoxon FDR ≤ 0.1 and Log2 fold change ≥ 0.5) for each scATAC cluster. (**d**) The fraction of each scRNA-seq cluster comprising each sample. The total proportions for each cluster are shown in the rightmost column. (**e**) The fraction of each scATAC-seq cluster comprising each sample. The total proportions for each cluster are shown in the rightmost column.

https://doi.org/10.1038/s41588-023-01445-4

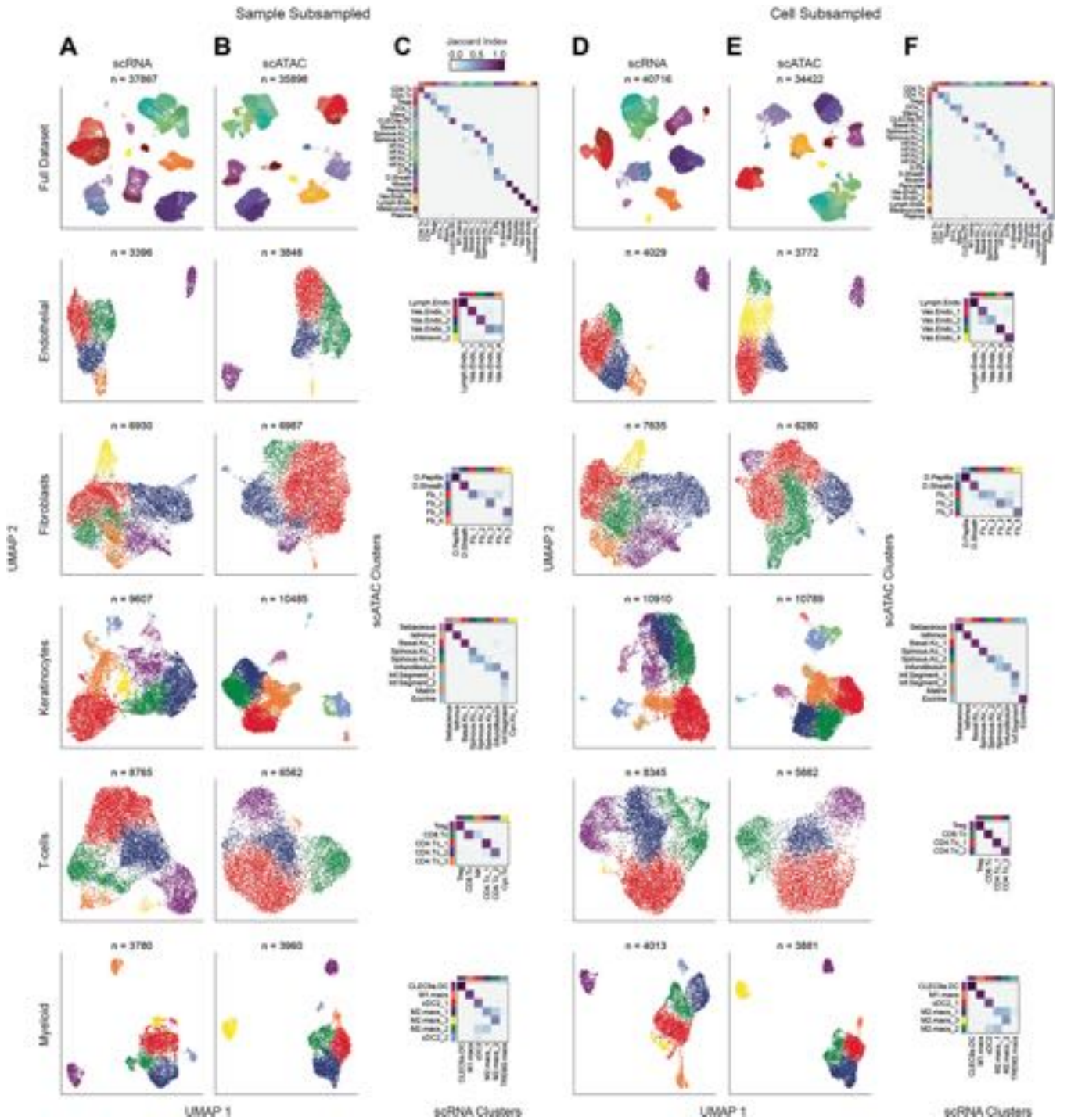


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41588-023-01445-4

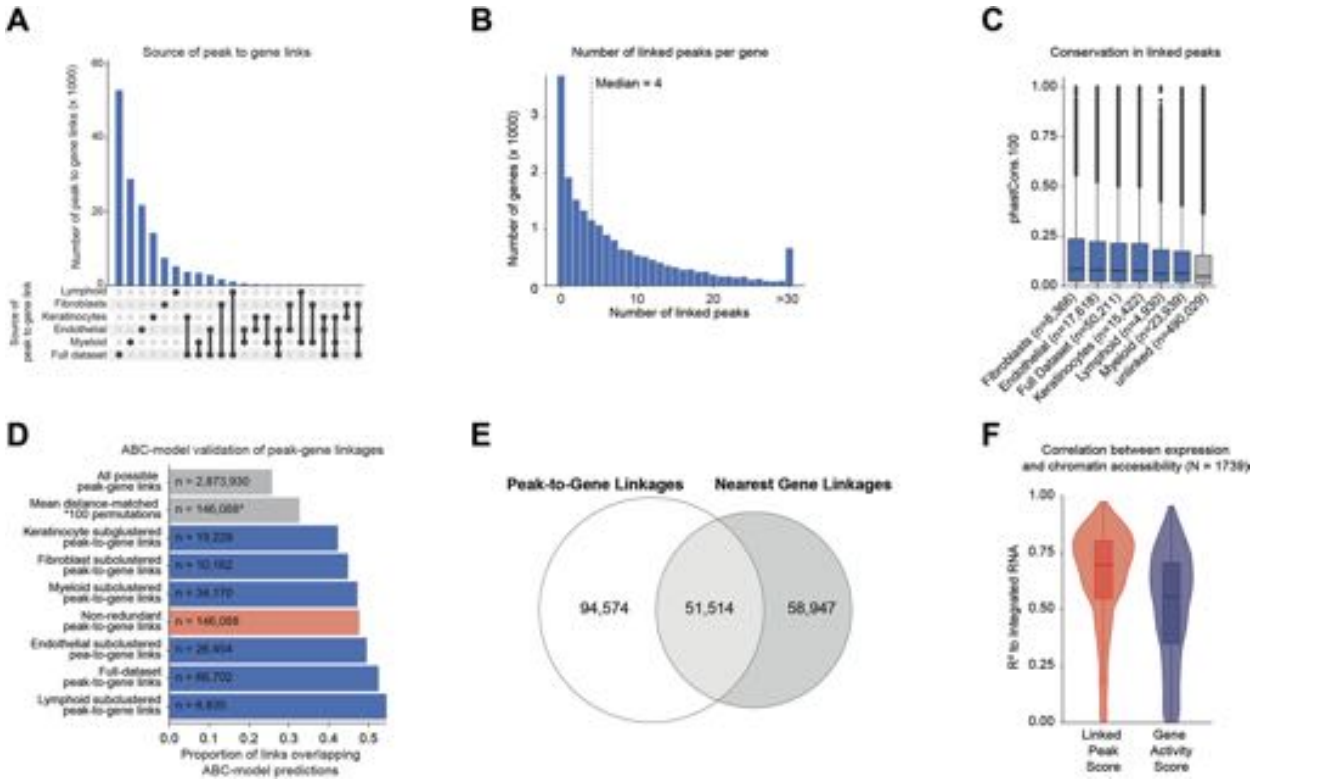
Extended Data Fig. 3 | Sub-clustering of major cell groups and integration of scRNA and scATAC datasets. (**a**) UMAP representations of sub-clustered major cell groups using scATAC data. Cell compartments are labeled on the left, and cells are colored according to their high-resolution cluster labels. (**b**) UMAP representations of sub-clustered major cell groups using scRNA data. Cell compartments are labeled on the right, and cells are colored according to their high-resolution cluster labels. (**c**) scRNA gene expression for selected marker genes for each high-resolution scRNA-seq cluster from each sub-clustered cell group. The color indicates the relative expression across all high-resolution clusters and the size of the dot indicates the percentage of cells in that cluster that express the gene. (**d**) Correspondence between scRNA and scATAC-seq cluster labels for high-resolution clusters in each of the sub-clustered datasets. Heatmaps are colored according to the Jaccard index of cluster label overlap between the scRNA and scATAC-seq datasets. (**e**) Correspondence between scRNA and scATAC-seq cluster labels in the full scalp dataset.

https://doi.org/10.1038/s41588-023-01445-4



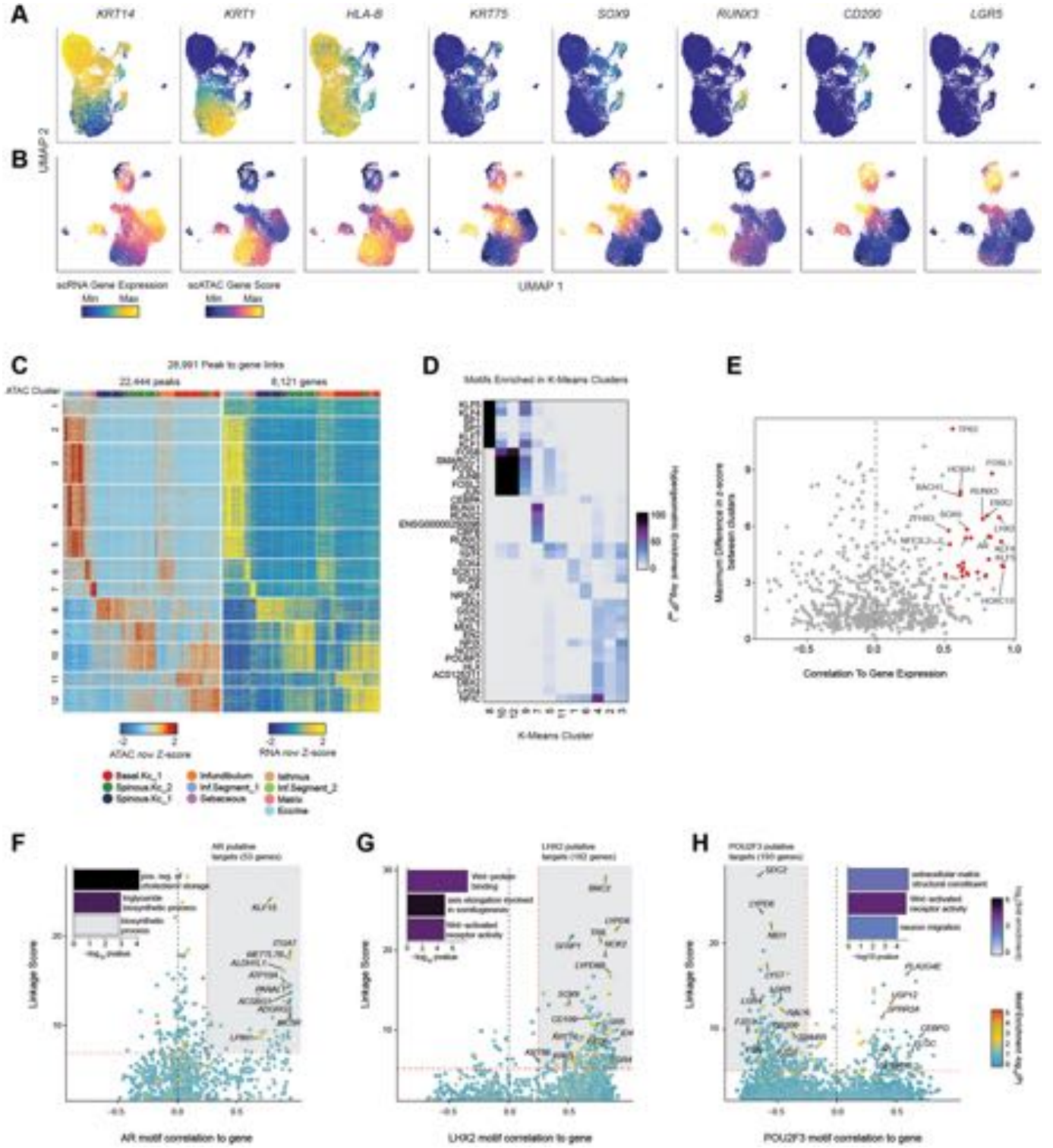
Extended Data Fig. 4 | Clustering and CCA-based integration robustness to subsampling. (A through C) Repeated dimensionality reduction and clustering of the scRNA and scATAC-seq datasets with three samples (AA4, C_SD3, and C_PB3) removed from the full dataset. (**a**) UMAP representations of the full subsampled dataset and sub-clustered major cell groups using scRNA data. Cell compartments are labeled on the left, and cells are colored according to their high-resolution cluster labels as shown in the x-axis in (C). (**b**) UMAP representations of the full dataset and sub-clustered major cell groups using scATAC data. Cell compartments are labeled on the left, and cells are colored according to their high-resolution cluster labels as shown in the y-axis in (C). (**c**) Correspondence between scRNA and scATAC-seq cluster labels for the low- and high-resolution clusters in each of the subsampled datasets. (**D** through F) Repeated dimensionality reduction and clustering of the scRNA and scATAC-seq datasets with 25% of the cells randomly removed from the full dataset. (**d**) Same as in (A), but for the cell-subsampled dataset. (**e**) Same as in (B), but for the cell-subsampled dataset. (**f**) Same as in (C), but for the cell-subsampled dataset.

https://doi.org/10.1038/s41588-023-01445-4



Extended Data Fig. 5 | Identification and characterization of peak-to-gene linkages. (**a**) Upset plot indicating the number of peak-to-gene linkages identified in the full dataset and in each of the sub-clustered datasets. (**b**) The distribution of the number of linked peaks per gene (median = 4). (**c**) The PhastCons 100-way vertebrate conservation scores for peaks with a linked gene in each dataset compared to unlinked peaks. Two-sided Wilcoxon rank-sum test comparing each dataset to unlinked peaks, $p < 2.2 \times 10{-}16$. Boxplots represent the median, 25th percentile and 75th percentile of the data, and whiskers represent the highest and lowest values within 1.5 times the interquartile range of the boxplot. (**d**) Bar plot showing the proportion of peak-to-gene linkages where both peak and gene were validated by a multi-tissue dataset of activity-by-contact (ABC) model enhancer-gene predictions. Categories compared included the space of all possible peak-to-gene links, the mean of 100 permutations drawn from all possible peak-to-gene links where for each permutation 146,088 peaks were selected to match the anchor distance distribution of true peak-to-gene links, and the set of true peak-to-gene links identified on each sub-clustered dataset. One-sided Fisher's exact test enrichment comparing each subgroup of true peak-to-gene links to a distance-matched background set, $p < 2.2 \times 10{-}16$. (**e**) Venn-diagram indicating the overlap of peak-to-gene linkages and peak-to-nearest-gene associations. (**f**) Comparison of the linked peak score (sum of accessibility at linked peaks) compared to the gene activity score for predicting gene expression for the 1739 HRGs. Plotted is the Pearson R2 from 246 pseudo-bulked samples per gene. Boxplots represent the median, 25th percentile and 75th percentile of the data, and whiskers represent the highest and lowest values within 1.5 times the interquartile range of the boxplot.

https://doi.org/10.1038/s41588-023-01445-4

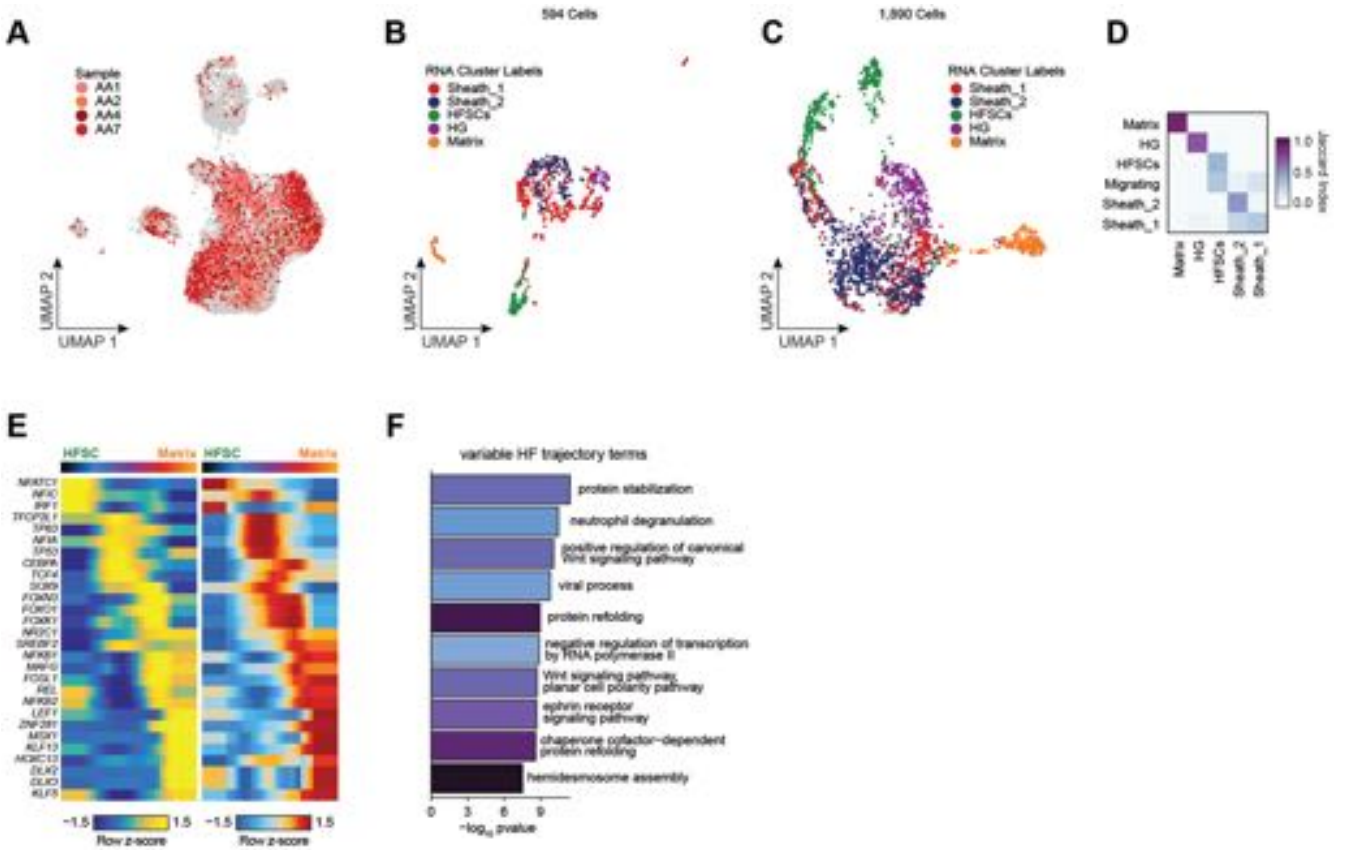


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41588-023-01445-4

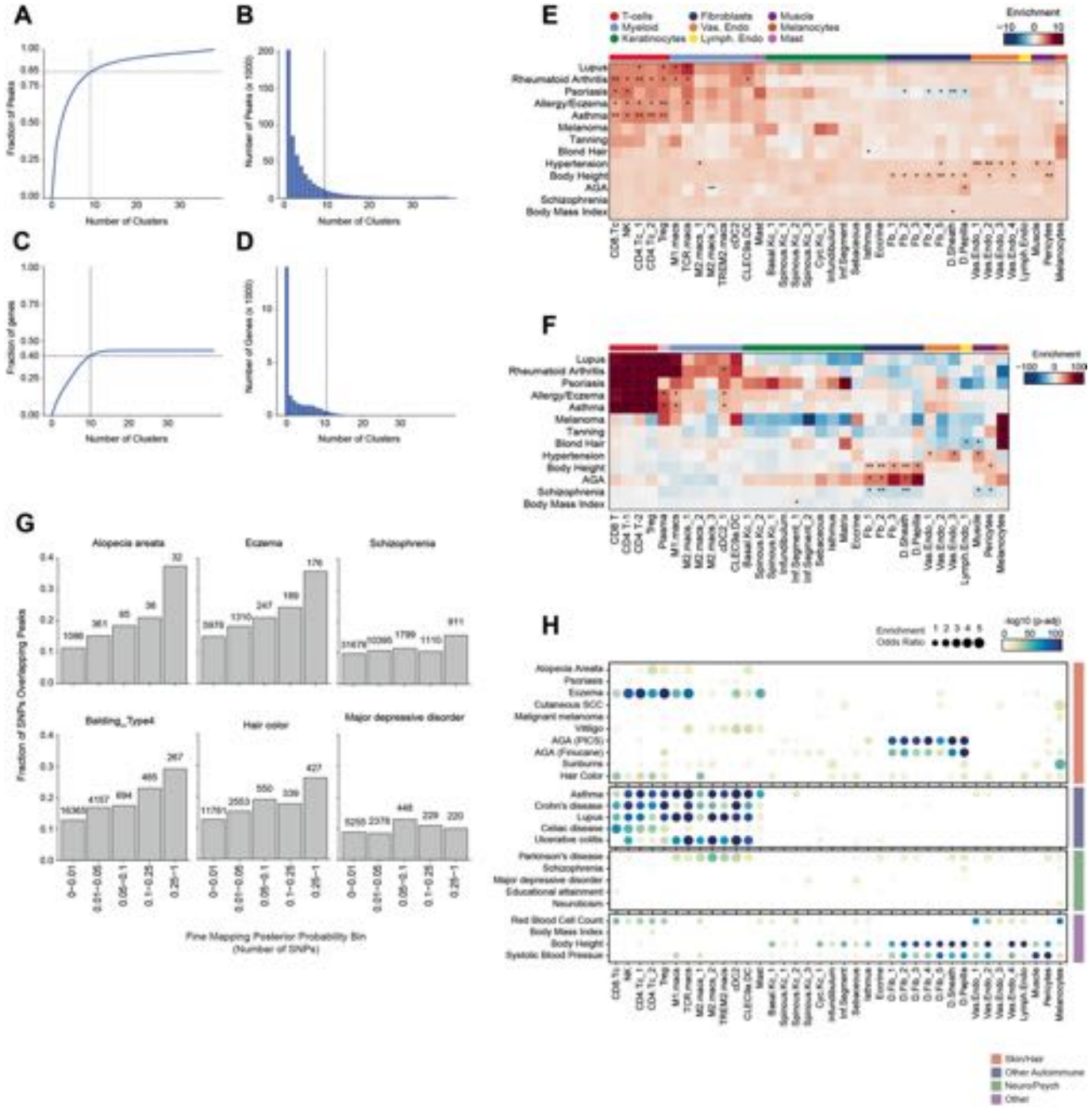
Extended Data Fig. 6 | Marker genes and cell type–specific TF regulator activity for sub-clustered interfollicular and hair-follicle associated keratinocytes. (**a**) UMAP projections of sub-clustered keratinocyte scRNA-seq dataset colored by expression levels of representative marker genes. (**b**) UMAP projections of sub-clustered keratinocyte scATAC-seq dataset colored by gene activity scores of the same marker genes shown in (A). (**c**) Heatmap showing the chromatin accessibility (left) and gene expression (right) for 28,991 keratinocyte-specific peak-to-gene linkages. Peak-to-gene linkages were clustered using k-means clustering (k = 12). Rows indicate peak accessibility and gene expression on the left and right heatmaps respectively. Each column is a pseudo-bulk sample, with the colorbar on top of each heatmap indicating the cluster identity of each pseudo-bulk sample. (**d**) Hypergeometric enrichment p-values of TF motifs in peaks from each of the k-means clusters from (C). (**e**) Plot of TF motif activity correlation to corresponding TF gene expression across sub-clustered dataset against the maximum difference in chromVAR deviation z-score between clusters. TF's with a maximum chromVAR difference in the top quartile and a pearson correlation greater than 0.5 are colored in red. (**f**) Prioritization of gene targets for LHX2. The x-axis shows the Pearson correlation between the TF motif activity and integrated gene expression for all expressed genes across all keratinocytes. The y-axis shows the TF Linkage Score (for all linked peaks, sum of motif score scaled by linkage correlation). Color of points indicates the hypergeometric enrichment of the TF motif in all linked peaks for each gene. Top gene targets are indicated in the shaded area (motif correlation to gene expression >0.25, linkage score >80th percentile). GO term enrichments for the top gene targets are shown in the inset bar plot. (**g**) Same as in (F), but for androgen receptor (AR). (**h**) Same as in (F), but for POU2F3.

https://doi.org/10.1038/s41588-023-01445-4



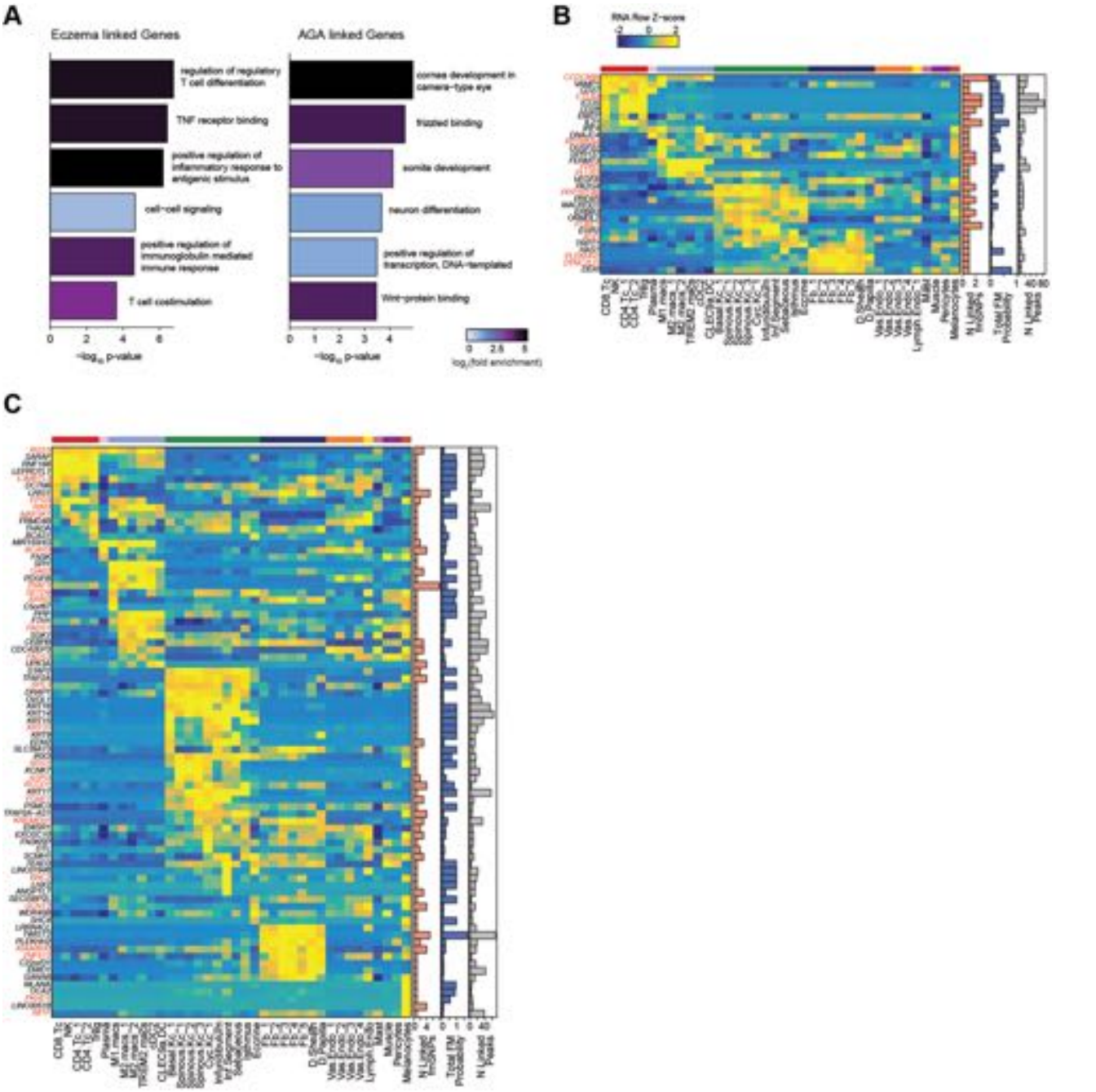
Extended Data Fig. 7 | Supplemental analyses of sub-clustered inferior segment hair follicle keratinocytes. (**a**) UMAP projection of sub-clustered keratinocytes showing cells originating from alopecia areata. Cells originating from control samples are colored gray and sorted to the back of the plot. (**b**) UMAP projection of sub-clustered scRNA inferior segment hair follicle keratinocytes. (**c**) UMAP projection of sub-clustered scATAC inferior segment hair follicle keratinocytes colored by matched nearest scRNA cluster. (**d**) Correspondence between scRNA and scATAC-seq cluster labels for integrated inferior segment hair follicle keratinocytes. (**e**) Paired heatmaps of positive TF regulators whose TF motif activity (left) and matched gene expression (right) are positively correlated across the hair follicle keratinocyte differentiation pseudotime trajectory. (**f**) GO term enrichments of the most variable 10% of genes across the hair follicle keratinocyte differentiation pseudotime trajectory.

https://doi.org/10.1038/s41588-023-01445-4



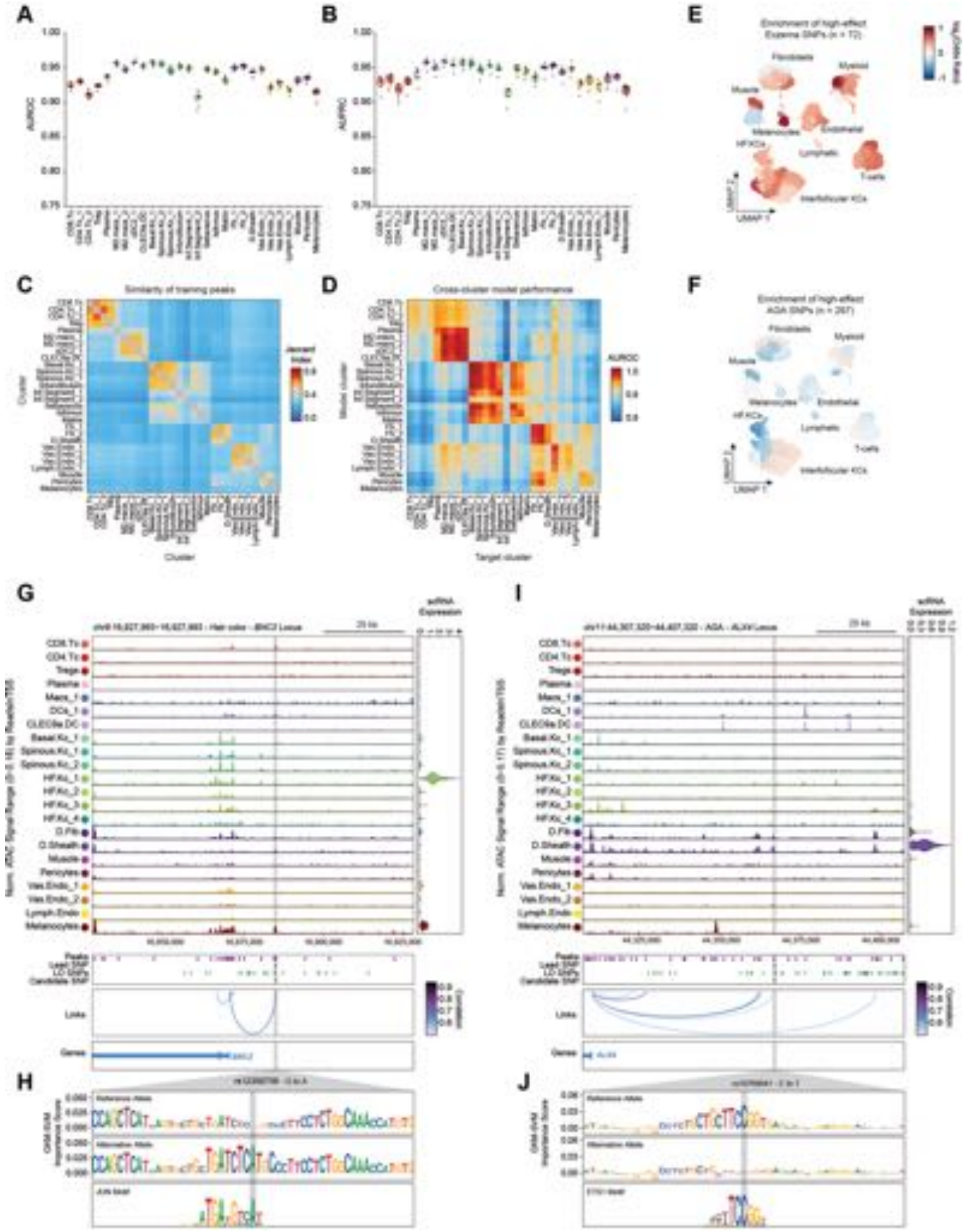
Extended Data Fig. 8 | Supplemental analyses of GWAS signal enrichment in cell type–specific open chromatin regions and cell type–specific genes. (**a**) Cluster-specificity of peaks used for LD score regression and for Fisher enrichment tests in Fig. 5. More than 50% of peaks are specific to <1/8 of high-resolution scATAC clusters, and 85% of peaks are specific to ≤ 1/4 of clusters. (**b**) Distribution of the number of clusters in which each peak is accessible. Peaks accessible in ≤ 1/4 of clusters (9 high-resolution clusters) were used for cluster-specific enrichment analyses. (**c**) Cluster-specificity of marker genes used for LD score regression and for Fisher enrichment tests in (E) and (G) respectively. (**d**) Distribution of the number of clusters identified as expressing a given marker gene. Marker genes expressed in ≤ 1/4 of clusters (10 high-resolution clusters) were used for cluster-specific enrichment analyses. (**e**) LD score regression identifies enrichment of GWAS SNPs for various skin and non-skin related conditions in gene regions specific to sub-clustered cell types (from the scRNA dataset) in human scalp. FDR-corrected P-values from LDSC enrichment tests are overlaid on the heatmap (*FDR < 0.05, **FDR < 0.005, ***FDR < 0.0005). (**f**) Same as in (E), but using only open-chromatin regions (from the scATAC dataset) that are implicated in peak-to-gene linkages (N = 98,188). (**g**) Fraction of fine-mapped SNPs for selected traits overlapping scalp CREs binned by fine-mapping posterior probability. (**h**) Fisher's exact test enrichment of the nearest gene for fine-mapped trait-related SNPs in cell type–specific genes for sub-clustered cell types in human scalp. The FDR-corrected -log10 p-value is indicated by the color of the dots, and the dot size indicates the enrichment odds ratio.

https://doi.org/10.1038/s41588-023-01445-4



Extended Data Fig. 9 | Supplemental analyses of fmGWAS-linked genes. (**a**) GO term enrichment for the top genes linked to fine-mapped SNPs by summed fine-mapping posterior probability in associated peak-to-gene linkages. (**b**) The top genes linked to peaks containing fine-mapped SNPs for alopecia areata. The heatmap shows relative gene expression for each high-resolution scRNA cluster. The number of linked fmSNPs per gene is indicated in the red bar plot to the right, and the total sum of fine-mapped posterior probability for linked SNPs is indicated in the blue bar plot. The grey bar plot shows the total number of identified peak-to-gene linkages for that gene in the entire scalp dataset. Gene names colored red indicate fine-mapped SNP to gene linkages supported by GTEx eQTLs. (**c**) Same as in (B), but for hair color.

https://doi.org/10.1038/s41588-023-01445-4



Extended Data Fig. 10 | See next page for caption.

https://doi.org/10.1038/s41588-023-01445-4

Extended Data Fig. 10 | Assessment of gkmSVM model performance and additional high-effect candidate fine-mapped SNPs. (**a**) The area under the receiver operator (AUROC), or (**b**) precision recall (AUPRC) curves for the gkm-SVM machine learning classifiers for each of the cluster models. Each dot indicates a cross-validation fold (n = 10). Boxplots represent the median, 25th percentile and 75th percentile of the data, and whiskers represent the highest and lowest values within 1.5 times the interquartile range of the boxplot. (**c**) The overlap of training data (peak sequences) between models. (**d**) The performance of each cluster model on predicting test sequences from a non-target cluster. (**e**) Enrichment of high-effect fine-mapped SNPs from eczema relative to random fine-mapped SNPs in cis-regulatory regions. (**f**) Same as in (**e**), but for AGA. (**g**) Normalized chromatin accessibility landscape for cell type–specific pseudo bulk tracks around the BNC2 locus. Integrated BNC2 expression levels are shown in the violin plot for each cell type to the right. The position of ATAC-seq peaks, the GWAS lead SNP, the fine-mapped SNP candidates in LD with the lead SNP, and the candidate functional SNP are shown below the ATAC-seq tracks. Significant peak-to-gene linkages are indicated by loops connecting the BNC2 promoter to indicated peaks. (**h**) GkmExplain importance scores for the 50 bp region surrounding rs12350739, a hair color associated SNP that creates a JUN motif in a CRE linked to BNC2 expression. (**i**) Same as in (G), but for the ALX4 locus. (**j**) GkmExplain importance scores for the 50 bp region surrounding rs10769041, an AGA associated SNP that disrupts an ETS motif in a CRE linked to ALX4 expression.

nature portfolio

Corresponding author(s): William Greenleaf

Last updated by author(s): May 30, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Code for generating fragments files for scATAC and counts matricies for single cell RNA was obtained from 10x genomics (go.10xgenomics.com/scATAC/cell-ranger-ATAC and https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger)
Data analysis	cellranger-atac-1.2.0 – alignment of ATAC data and generation of fragments files cellranger-3.1.0 – alignment of RNA data and generation of counts matrices macs2 2.1.1 – Software for peak calling R version 4.0.2 – R environment for all custom code ArchR - 1.0.1 - Software for analysis of scATAC-seq data Seurat_4.0.4 – Software for analysis of scRNA-seq data SAMtools mpileup v1.5 – Software for genotyping bulk ATAC data VarScan mpileup2snp v2.4.3 – Software for genotyping bulk ATAC data DoubletFinder_2.0.3 – Software for doublet removal for scRNA-seq celda_1.6.1 – Software used for ambient RNA decontamination (DecontX) BSgenome.Hsapiens.UCSC.hg38_1.4.3 – Package containing genomic DNA sequences uwot_1.0.10 – Used for UMAP harmony_1.0 – Software used for batch correction in subclustering analysis GenomicScores_2.2.0 – Used for computing evolutionary conservation of open chromatin regions topGO_2.42.0 – Software used for GO enrichments preprocessCore_1.52.0 – Used for data normalization chromVAR_1.12.0 – Used for measuring enrichment of transcription factor motifs in accessible chromatin

slingshot_1.8.0 – Used for estimating differentiation trajectories
miloR_1.1.0 – Used for differential abundance testing of alopecia areata vs control keratinocyte populations
edgeR_3.32.1 – Used for analysis of single-cell data
DESeq2_1.30.1 – Used for differential analysis of downloaded Klf4 knockdown data and for comparison of fresh vs cryopreserved samples.
LDSR_1.0.1 – Software for estimation of GWAS signal enrichment in cell-type specific chromatin regions
LSGKM-SVR (https://github.com/kundajelab/lsgkm-svr) – Software for predicting effects of genetic variation on chromatin accessibility
fitdistrplus_1.1.6 – Used for estimating distribution parameters

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Sequencing data generated in this study has been deposited in the Gene Expression Omnibus (GEO) with the accession code GSE212450. The full scalp dataset can be explored interactively at (http://shiny.scscalpchromatin.su.domains/shiny_scalp/) . Reference genome files for aligning single-cell data can be downloaded from https://support.10xgenomics.com/single-cell-gene-expression/software/release-notes/build. Predicted super enhancer associated genes from 86 human cell types and tissues were downloaded from table S2 from https://doi.org/10.1016/j.cell.2013.09.053. Predicted super enhancer associated genes from mouse hair follicle cell populations were downloaded from table S1 from https://doi.org/10.1038/nature14289. The activity-by-contact (ABC) dataset generated from 131 human tissues and cell types was downloaded from https://www.engreitzlab.org/resources/. Differentially expressed genes identified between control human keratinocytes and keratinocytes containing a mutant, binding incompetent form of TP63 were obtained from Table S1D from https://doi.org/10.1016/j.celrep.2018.11.039. The counts matrix from shRNA knockdown of KLF4 in human adult keratinocytes is available on GEO with the accession number GSE111786. Formatted summary statistics for partitioning heritability using LD score regression can be downloaded from https://console.cloud.google.com/storage/browser/broad-alkesgroup-public-requester-pays/sumstats_formatted. Fine-mapped SNPs for 94 UKBB traits can be downloaded from www.finucanelab.org/data. Pre-computed PICS fine-mapped SNPs for a variety of traits from the GWAS catalog are available at https://pics2.ucsf.edu/Downloads/.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	No attempt was made to exclude donors on the basis of age, gender, or sex. Individual donors self-reported their sex and consented to this information being shared. This information can be found at the GEO accession above. No sex- or gender-based analyses were performed in this study.
Population characteristics	Patients in this study were recruited from Stanford Healthcare and from Santa Clara Valley Medical Center. Ages of patients ranged from 20–80. The study included 5 males and 10 females. The study included 5 patients with active alopecia areata and 10 healthy controls.
Recruitment	For alopecia areata patients, no criteria other than active disease affecting >2% of the scalp and absence of current treatment was required for recruitment. For healthy control patients donating scalp samples either in the form of discarded surgical dogears or from healthy scalp punch biopsies, no criteria other than absence of hair disease (e.g. alopecia areata, androgenetic alopecia) in the affected tissue was required. Patients with alopecia areata were recruited from dermatology clinics at either Stanford University or Santa Clara Valley Medical Center. Patients donating healthy scalp samples were recruited from dermatology clinics at Stanford University. No attempt was made to exclude donors on the basis of age, gender, or sex. Patients undergoing dermatological surgeries typically had a non-melanoma skin cancer peripheral to the tissue used in this study. These patients thus tended to be older than the alopecia areata patients or the other healthy control patients. Patients volunteering to donate samples may also self-select for a number of reasons, such as personal interest in research, comfort with medical procedures, or socioeconomic status. These potential biases apply to all groups recruited in this study and are thus not expected to impact results.
Ethics oversight	All research described complies with the ethical guidelines for human subjects research under the approved Institutional Review Board (IRB) protocol at Stanford University (no. 40524) for the collection and use of human tissue samples.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was set based on the availability of patient donors during the recruitment period of the study. Sample size was sufficient to identify the expected cell types present in the human scalp in both healthy patients and those with active alopecia areata.
Data exclusions	All datasets generated that did not fail experimentally (e.g. overloaded sample) were included in the study.
Replication	Technical replicates for additional single-cell experiments were not performed as technical replicates are less informative than using multimodal data (scATAC and scRNA) for each sample. Biological replicates were obtained in the form of multiple patient donors from each disease state. Generation of single-cell data was performed only once for each patient sample. Selected findings (e.g. peak-to-gene linkages, enrichment of GWAS signals in cell type specific chromatin) were validated using external, orthogonal datasets and analyses.
Randomization	There was no randomization into experimental groups. All samples were processed to generate single-cell datasets individually as they became available.
Blinding	No blinding was performed in this study that focused on deep characterization alopecia areata and healthy control scalp tissue at a single point in time. No differential clinical intervention was performed or was being compared in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	PE-Cy7 Mouse Anti-Human CD90; Supplier:BD Pharmingen; Clone: 5E10; Catalog number: 561558
Validation	The CD90 antibody was pre-validated and conjugated by BD Pharmingen. It was purchased for FACS of dissociated scalp tissue.