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

I. Supplementary Notes

Demonstration of how peak-to-gene linkages are missed on full dataset

To demonstrate how enhancer-gene links can be missed by using only the full dataset, we examined peaks linked to RUNX3, a gene that is expressed in multiple distinct cell types. Using the full, non-sub clustered dataset, we identified 43 linked peaks for RUNX3, the majority of which were accessible in T-lymphocytes, myeloid lineage cells, and melanocytes. Relatively little chromatin accessibility was observed in keratinocytes (Figure 2B). However, performing peak-to-gene identification in the sub-clustered keratinocytes revealed numerous RUNX3-linked peaks that were strongly accessible in sebaceous gland cells. Repeating this process on each of the sub-clustered data sets resulted in a non-redundant set of 81 RUNX3- linked peaks, nearly doubling the number of RUNX3-linked enhancers identified using only the full dataset.

Discussion on absence of T-lymphocyte differences between alopecia areata and control samples

While we observed increased overall abundance of T-lymphocytes in samples from patients with alopecia areata (Fig. 1H,I, Extended Data Fig. 2D,E), we did not observe any discrete populations of T-lymphocytes unique to patients with alopecia areata, nor did we identify an appreciable number of differentially accessible open chromatin regions between alopecia areata and control T-lymphocytes. This perhaps could be because the cellular phenotype of auto-

reactive T-lymphocytes in alopecia areata is relatively subtle, or is driven by rare subpopulations of cells that we could not identify.

Additional LD score regression analyses

To assess how much additional information the cell type–specific chromatin profiles provided, we repeated LD score regression using cell type–specific marker genes instead of peaks (Methods, Extended Data Fig. 8C-E). We found that while the overall pattern of enrichments was similar, the degree of enrichment and the statistical significance tended to be lower when using cell type–specific genes compared to open chromatin regions. We additionally tested further restricting the cell type-specific accessible chromatin regions used in linkage disequilibrium score regression (LDSC) to only cis-regulatory elements linked to expression of a gene through a peak-to-gene linkage (reducing total number of usable chromatin regions from 589,294 to 98,188 peaks). We found that with this adjustment, the overall pattern of cell type enrichments was largely the same when using all cell type-specific accessible chromatin regions (Extended Data Fig. 8F). We found that using only CREs involved in peak-to-gene linkages generally increased the magnitude of enrichment for certain interactions (e.g. T-cell subclusters and most autoimmune disease GWAS datasets), however the lower number of peak regions used in the analysis predictably resulted in less significant enrichment p-values.

Additional discussion of hierarchical peak-to-gene linkage analyses

By identifying peak-to-gene linkages at multiple levels of cellular resolution, we captured both broad, cell-class regulatory differences, as well as regulatory programs delineating more subtle cell type differences. We anticipate that using both ‘low-resolution’ clustering and ‘high-resolution’ sub-clustering of complex datasets can be adapted to increase the resolution of predicted gene-regulatory networks in future studies, whether they use computationally integrated or experimentally linked multi-omic datasets. However, the fact that so many

additional enhancer-gene predictions can be made using this tiered approach highlights the dependence of this type of analysis on the particular datasets under examination. The cellular diversity of a dataset should be an important consideration for all such analyses, and experimental validation of predicted enhancer-gene linkages will ultimately be required.

II. Supplementary Methods

Bulk ATAC-seq of subset of control samples

Bulk ATAC seq was performed on dissociated cells from four of the surgical dogear control samples (C_SD4, C_SD5, C_SD6 and C_SD7). These samples were thawed quickly in a 37℃ water bath and then 1 mL of prewarmed media (RPMI 1640 w/ 10% FBS) was added to each sample. Samples were centrifuged at 300 x g for 5 minutes at 4℃ and the supernatant was aspirated. Each sample was resuspended in ice cold 1x PBS with 0.5% BSA and then split into two aliquots of 200,000 cells. If fewer cells were present for a sample, the number of cells was split evenly. One aliquot from each sample was pooled and this pool was used for two lanes of 10x single cell ATAC-seq v2. The remaining aliquot of each sample was used for bulk ATAC-seq, which was performed similarly to as described previously[1]. Briefly, cell aliquots were centrifuged at 300 x g for 5 minutes at 4℃ and the supernatant was aspirated. Each pellet was resuspended in 100 µL of ice cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA, 0.01% Digitonin, 0.1% Tween-20, and 0.1% NP40) and incubated on ice for 3 minutes. The lysis reaction was then diluted by the addition of 1 mL of ice cold RSB-washout buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA, 0.1% Tween-20). Samples were centrifuged at 500 x g for 10 minutes at 4℃. The supernatant was aspirated and each nuclei pellet was resuspended in 50 µL of Transposition solution (10 mM Tris-HCl, pH 7.4, 5 mM $MgCl_2$, 10% dimethyl formamide, 0.33x PBS, 0.01% digitonin, 0.1% Tween-20, 100 nM Illumina Tn5 Transposase (Illumina, 20034197)). Samples were incubated at 37℃ in a thermomixer rotating at 1000 RPM for 30 minutes. The remainder of the bulk ATAC-seq library generation was performed as described previously. Resulting bulk-ATAC libraries were pooled and sequenced on an Illumina NextSeq 550 using paired-end 36-bp reads.

Genotyping and sample deconvolution using demuxlet

Bulk ATAC-seq FastQ files were processed and aligned to the hg38 reference genome using Bowtie2 (v2.2.6), and then peaks were identified using MACS2 (v2.1.1)[2]. Peak regions identified from these bulk ATAC-seq samples were genotyped using SAMtools mpileup (v1.5) and VarScan mpileup2snp (v2.4.3), and then used as input for Demuxlet to identify the patient sample for each cell for the pooled scATAC-seq samples[3,4].

Comparison of fresh vs cryopreserved samples

To determine if there were any systematic differences between the fresh vs cryopreserved samples, we tested for differential genes and differential peaks between fresh and cryopreserved samples for each of the major cell groupings described above (Keratinocytes, Fibroblasts, Endothelial, T-lymphocytes, and Myeloid lineage cells). For each of these major cell groupings, we used ArchR to identify differential scATAC-seq peaks using the 'getMarkerFeatures' function with the default 'wilcoxon' test method and correcting for TSS enrichment and log10(nFrags) bias (Extended Data Fig. 1F). For each of the major cell groupings, we pseudo-bulked each individual sample and then removed mitochondrial, ribosomal, and chrY genes. We then used the DESeq2 R package (v1.30.1) to test for differentially expressed genes between fresh vs frozen samples while controlling for sex and disease status (alopecia areata vs control) as covariates (Extended Data Fig. 1G).

Sub-clustering of major cell types

To improve identification of rare cell types, we sub-clustered 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[5], hair-follicle keratinocyte clusters exhibited high gene activity and expression of the transcription factor SOX9[6], 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[7]. We observed a relatively large scRNA-seq cluster expressing high levels of mast-cell markers including beta tryptases (TPSB1/2) and HPGD [8–10], but we did not observe a corresponding scATAC-seq cluster, perhaps due to the tendency for granulocyte chromatin to spontaneously decondense during nuclear isolation [11,12]. After labeling clusters in each modality, we sub-clustered major cell types in each dataset (Keratinocytes, Fibroblasts, Endothelial cells, T-lymphocytes, and Myeloid lineage cells) (Extended Data Fig. 3A-C). For scRNA-seq subclustering, we used the same iterative LSI dimensionality reduction procedure described above, except that we used two rounds of LSI instead of three, and we used the 'Harmony' R package (v0.1.0) to reduce sample batch effects since this effect was more pronounced on sub-clustered datasets relative to the full dataset [13]. For both scRNA-seq and scATAC-seq subclustering, the number of variable features and the resolution of clustering was tuned to attempt to balance the number of identified clusters between modalities and to match the cellular heterogeneity of each sub-clustered group. For the sub-clustered scRNA-seq datasets, we used 2,000 variable genes and 15 SVD dimensions, and a clustering resolution of either 0.1 or 0.2 in the first round, followed by a clustering resolution of 0.3 in the final round. To generate UMAPs for each sub-clustered scRNA-seq dataset, we used n.neighbors=35 or 40, min.dist=0.4, and metric=cosine. For scATAC-seq subclustering, we again used ArchR's implementation of iterative LSI dimensionality reduction, and we used Harmony to reduce sample batch effects. For each sub-clustered group, we used either 25,000 or 50,000 variable features, 25 dimensions, and between 0.2-0.4 resolution for clustering. To generate UMAPs for each sub-clustered scATAC-seq dataset, we used n.neighbors=35, min.dist=0.4, and metric=cosine.

Following sub-clustering in each dataset, we assigned sub-cluster labels based on known marker genes (Extended Data Fig. 3A-C - 'FineClust').

Subsampling of full dataset to assess cluster reproducibility

To assess the reproducibility of the aforementioned low and high cellular resolution clustering, we subsampled the full filtered dataset in two ways: first, by subsampling the patient samples used in the analysis (removing one sample from each category: AA4, C_SD3, and C_PB3– these samples were selected as they were present in both the scRNA and scATAC-seq datasets), and again by randomly removing 25% of the total number of cells from each of the scRNA and scATAC-seq datasets.

Using these two subsampled datasets ('sample subsampled' and 'cell subsampled'), we repeated clustering of both scRNA and scATAC datasets at both low and high cellular resolution as described previously. We found that with both subsampling strategies, we recovered highly highly reproducible cluster profiles at both high and low resolutions (Extended Data Fig. 4A,B,D,E). In a few instances, the cluster profile of sub-clustered cell groups from subsampled datasets did not separate rare cell populations (e.g. Eccrine gland cells in the sample subsampled keratinocyte scRNA dataset or NK cells in the cell subsampled T-cells).

We repeated the CCA integration for each of these subsampled datasets and found that as in the full dataset, corresponding clusters from the scRNA and scATAC datasets were accurately integrated (Extended Data Fig. 4C,F).

Peak calling in scATAC-seq datasets

After sub-clustering of major cell types, we transferred the sub-cluster labels ('FineClust') back to the full scATAC-seq dataset for peak calling to maximize our ability to detect open chromatin regions specific to rare cell subtypes. Peak calling was carried out using the standard ArchR workflow. Pseudo-bulk group coverages were calculated for each cluster using the ArchR

function ‘addGroupCoverages’, which were then used to call peaks using ‘addReproduciblePeakSet’. This function uses MACS2 (v2.1.1) to call fixed-width 500bp peaks on each cell type, merges the peak set from each cell type, and then iteratively removes overlapping peaks by dropping the lower scoring peak of overlapping pairs until no overlapping peaks remain [2,4]. This procedure resulted in identification of 589,294 unique peaks for the entire dataset.

For each of the sub-clustered scATAC-seq datasets, only a relatively small subset of the full union peak set will be accessible in any of the cell types present. This results in excessive numbers of completely inaccessible regulatory regions being used for analyses on the sub-clustered datasets, which decreases statistical power. However, simply repeating peak-calling on each sub-clustered dataset would result in a peak set that does not represent a true subset of the full peak set due to the iterative overlapping peak removal procedure used to obtain the full union peak set. To obtain a peak set that is specific to a given sub-clustered dataset but is also a true subset of the full union peak set, we loaded the original peak calls for each cell type in the sub-clustered dataset and then kept only the subset of peaks from the union peak set that overlapped with peaks called on the sub-clustered cell types. This sub-clustered peak set was used for the keratinocyte– specific analyses shown in figures 3 and 4.

Identification of potential regulatory target genes of TF regulators

We used the following strategy to identify potential regulatory gene targets of a given TF. Using the sub-clustered keratinocytes, we identified ~500 low overlapping pseudo bulk samples of KNN with k=100 and obtained the mean normalized integrated gene expression and the mean chromVAR motif deviation score for each pseudo bulk sample. For our subset of candidate TF regulators, we used these pseudo bulk samples to calculate 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 identifying 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:

$$LS_g = \sum_{k=1}^{n} R_k^2 MS_k$$

Where 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, more strongly correlated linked peaks containing the TF motif, and/or linked peaks that contain highly confident instances of the motif. Finally, we also calculated the hypergeometric enrichment p-value for the TF motif in all linked peaks for a given gene. We defined potential gene regulatory targets of a TF regulator as those that have an absolute TF motif to gene expression correlation of >0.25 and a linkage score greater than the 80th percentile across all genes. We performed Gene Ontology (GO) enrichment analyses on the putative direct regulatory gene targets using the topGO (v2.42.0) R package [14], using all expressed genes as background.

We validated inferred TF regulatory targets using previously published datasets of RNA-seq performed on keratinocytes with TF mutations or TF knockdown. For TP63, we downloaded the differentially expressed genes (Table S1D from Qu et al. 2018) identified between control human keratinocytes and keratinocytes containing a mutant, binding incompetent form of TP63[15]. We calculated the enrichment of downregulated genes from TP63 mutant keratinocytes in our predicted TP63 regulatory targets using a one-sided Fisher's exact test. We also compared the enrichment of upregulated genes from these TP63 mutant keratinocytes in our predicted TP63 regulatory targets. For KLF4, we downloaded the raw, unnormalized counts matrix from a recent study that performed shRNA knockdown of KLF4 in human adult keratinocytes (GSE111786_counts_raw.csv.gz)[16]. We removed genes that had fewer than 10

counts across all samples, and then performed normalization and differential testing using the DESeq2 R package (v1.30.1)[17]. We again calculated the enrichment of downregulated and upregulated genes from KLF4 knockdown keratinocytes in our predicted KLF4 regulatory targets.

LD Score Regression using scATAC-seq data

We used linkage disequilibrium score regression (LDSC, v1.0.1) to estimate the heritability of multiple skin, hair and other traits in each high-resolution clustered cell type in our dataset [18]. Briefly, this method determines if a functional category is enriched for heritability of a given trait by determining if SNPs with high LD to that category tend to have higher χ^2 statistics than SNPs with low LD to that category, conditioned on a baseline set of annotations. Cluster-specific peak regions were used as input functional categories for LDSC. To obtain these cluster-specific peaks, we first removed clusters that had fewer than 40 cells total, as these clusters generally had too few cells to identify sufficient numbers of confident cell type–specific peaks. For remaining clusters, we identified which peaks from the union peak set had been 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 (9 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. We formatted these cluster specific peaks using the 'make_annot.py' script, and LD scores were computed for each annotation using the 'ldsc.py' script with default parameters. 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. In addition to using skin and hair-related GWAS traits, we selected several other traits related to neurologic or psychiatric conditions (e.g. major depressive

disorder, schizophrenia, neuroticism, parkinson's disease) that would not be expected to demonstrate much, if any, cell type–specificity in our scalp dataset. Other broad, but highly powered GWAS studies (e.g. BMI, body height, systolic blood pressure (SBP)) were selected to demonstrate that even though cells in the scalp may not be obviously involved in these traits, there may be some areas of biological overlap (e.g. fibroblasts being enriched for body height GWAS signal and endothelial cells being enriched for SBP GWAS signal). We followed the recommended guidelines for cell type–specific partitioned heritability analysis using the 1000G EUR phase 3 population reference and the hg38 baseline model (v2.2). We used the 'ldsc.py' script to calculate partitioned heritability for each trait in the cluster specific peak sets. We used Benjamini-Hochberg FDR correction to adjust heritability enrichment p-values. We also repeated this analysis using cluster-specific marker gene regions. To identify cell type–specific marker genes, we used Seurat's 'FindAllMarkers' function with default settings to identify a set of genes differentially highly expressed in each high-resolution scRNA cluster. For each cluster, we then retained only genes that had been identified in no more than 25% of all clusters (10 out of a possible 42 clusters). We added a 100-kb window around each gene region and used these regions as input for LD score regression analysis as described above.

Analysis of fine-mapped GWAS variants

We obtained fine-mapped SNPs from multiple sources. First, we downloaded a compendium of fine-mapped SNPs for 94 UKBB traits (www.finucanelab.org/data), and used the male pattern balding ('Balding_Type4'), body mass index ('BMI') and systolic blood pressure ('SBP') traits for downstream analyses [19]. Second, we downloaded pre-computed 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)[20,21]. Details of trait definitions are available from the UK Biobank (https://www.ukbiobank.ac.uk/), or from the GWAS catalog (https://www.ebi.ac.uk/gwas/). For example, 'educational attainment' (GWAS catalog) refers to years of education, 'hair color'

(GWAS catalog) is a summary category for studies related to hair pigmentation, and 'tanning' (UK Biobank) is a self-reported measure of ease of skin tanning.

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. To compare the differences between using open chromatin data verses simply assigning fine-mapped SNPs to their nearest gene in 1-dimensional genomic space, we repeated this fine-mapped SNP enrichment analysis by linking fine-mapped SNPs to their nearest gene and calculating the enrichment of cell type–specific genes linked (i.e. nearest) to fine-mapped SNPs (Extended Data Fig. 8C,D,G). We again found that the broad pattern of cell type–specificity of this analysis was similar to our more restrictive analysis of using direct overlap with cell type specific open chromatin regions. By forcing each fine-mapped SNP to be associated with its nearest gene, however, the number of SNP to marker gene 'overlaps' dramatically increases relative to the direct overlap of SNPs with the specific genomic coordinates of a set of marker peak regions (each SNP is associated with at least one of ~20,000 possible genes, ~1000 of which will be marker genes for a given cell type). The ultimate effect of this difference is that the statistical significance of this 'nearest-gene' analysis is notably higher than the statistical significance of the original analysis (both using one-sided Fisher enrichment tests). We also note, however, that some traits appeared to be less cell type–specific using this version of the analysis. For example, the odds ratio (OR) of enrichment for fine-mapped AGA SNPs was weaker using nearest gene SNP associations, and body height SNPs showed significant enrichment in dermal papilla-associated marker genes–an enrichment not observed in the open chromatin version of this analysis.

To identify genes associated with fine-mapped SNPs, for selected traits we identified fine-mapped SNPs that had a fine mapping posterior probability of ≥ 0.01 and overlapped a

scATAC-seq peak region. Next, for each gene, we identified all fine-mapped SNPs that fell within a peak that was linked to the expression of that gene, and we 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 genes linked to multiple linked peaks containing fine-mapped SNPs with appreciable fine-mapping posterior probability, are assumed to more likely represent genes whose expression is associated with the trait of interest. We plotted the 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, and plotted the number of linked peaks and the 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 gapped k-mer support vector machine (gkm-SVM) models using scATAC-seq data[22]. For each scATAC-seq cluster, we trained a gkm-SVM classifier to predict whether a given genomic sequence is likely to be accessible or inaccessible in that cell type[23,24]. We trained models for all clusters that had at least 200 cells to reduce training biases from spurious binding events from noisier, sparser data. We used training sequences of 1001 bp, expanding peaks on each side to reach this length and removing peaks that contained any N bases. As positive training data, we first identified the top 7,500 (by FDR) marker peaks for each cluster using the ArchR function 'getMarkers' with an FDR cutoff of 0.1 and a Log2FC cutoff of 0.5. These peaks represent the set of peaks that are most specific to a given cluster. We next identified which peaks from the union peak set had been originally identified in a given cluster by overlapping the union peak set with the MACS2 peak calls from that specific cluster. We sorted these by decreasing MACS2 score, and selected the top N peaks to combine with the previously identified marker peaks such that each cluster had ≤ 75,000 total unique peaks for training. For clusters that had fewer than 75,000 peaks, we used all peaks originally identified from that cluster as training peaks. Clusters that had fewer than

56,250 total peaks were not used for model training. To obtain negative training data, we generated 4,000,000 random genomic regions of 1001 bp (after masking assembly gaps (AGAPS) and intra-contig ambiguities (AMB)) and calculated the GC content of each of these regions. For each cluster, we then generated 20 equal-size bins of GC-content percentile from the N (≤ 75,000) positive training sequences. We labeled the random sequences according to these cluster-specific GC-content bins and sampled a total of N random regions while matching the binned distribution of GC-content from the original data.

After identifying the ≤ 75,000 positive and ≤ 75,000 negative training sequences for each cluster, we used a 10-fold cross-validation strategy to test model performance. We split training and testing sets by chromosome. The test sets for the 10 folds are as follows: Fold 1 consisted of chr 1; fold 2 consisted of chr 2 and 19; fold 3 consisted of chr 3 and 20; fold 4 consisted of chr 6, 13, and 22; fold 5 consisted of chr 5 and 16; fold 6 consisted of chr 4, 15 and 21; fold 7 consisted of chr 7, 14 and 18; fold 8 consisted of chr 11, 17 and X, fold 9 consisted of chr 9 and 12, fold 10 consisted of chr 8 and 10. We used similarly sized chromosome splits in our fold definition to prevent any data overlap between training and testing splits while also preserving the overall structure and relationships within the dataset. For each fold, we used sequences from all non-testing chromosomes for training. For each of the 10 folds for each of the 29 clusters, we used the sequences from the positive and GC-matched negative training data as input training gkm-SVM models. Specifically, we used the LS-GKM package with the following options for the 'gkmtrain' function[23,25]. We used the wgkmrbf kernel (t = 5), a word length of 11 (l = 11), 7 informative columns (k = 7), up to 3 mismatches to consider (d = 3), an initial value of 50 for the exponential decay function (M = 50), a half-life parameter of 50 (H = 50), and a precision parameter of 0.001 (e = 0.001). We assessed the performance of trained models from cross-validation folds by calculating AUROC and AUPRC using the 'PRROC' R package (v1.3.1) with negative testing data downsampled to match the number of positive testing data sequences. To examine model specificity, we used the fold 0 from each cluster to predict the

fold 0 testing data of every other cluster and again calculated the AUROC and AUPRC. Following assessment of model performance, we trained a full model for each cluster using all available training data which was used for estimating candidate SNP effects as described below.

Estimation of candidate SNP effect sizes using gkm-SVM models

We used our full gkmSVM models from each cell type to predict the change in accessibility for fine-mapped SNPs from GWAS for androgenetic alopecia ('Balding_Type4' from www.finucanelab.org/data and 'Male-pattern baldness' from PICS fine mapped), eczema (PICS fine mapped), hair color (PICS fine mapped), and randomly selected fine-mapped SNPs from any GWAS (PICS or Finucane data). We first selected only fine-mapped SNPs that overlapped a peak region, and then further filtered SNPs from all traits to those that had a fine-mapping posterior probability of ≥0.01. This resulted in 1,631 androgenetic alopecia SNPs, 612 hair color SNPs, 365 eczema SNPs, and a random selection of 2500 random SNPs to use as background. For each SNP that met the above criteria, we obtained the 250bp surrounding the SNP and created synthetic alternative allele sequences by replacing the reference allele at the center of the sequence with the SNP alternative allele. We then used the previously trained full models for each cluster to calculate cell type–specific GkmExplain importance scores for each base of both the reference and alternative allele sequences for each of the fine-mapped SNPs[26]. GkmExplain estimates the per-base contribution for an input sequence to the corresponding output prediction of a gkmSVM model. We used GkmExplain for mutation impact scoring instead of DeltaSVM, ISM, or SHAP because it tended to yield more directly interpretable importance scores at the motif level, and in aggregate has been previously shown to correlate extremely well with these other metrics [22]. For each SNP and each cluster model, we summed the GkmExplain importance scores for the central 50bp of both the reference and

alternative allele sequences, then subtracted the alternative allele score from the reference allele score to get the 'delta score' for the sequence immediately surrounding the SNP.

Statistical significance and prioritization of candidate SNPs

We used multiple metrics derived from GkmExplain importance scores to obtain a statistical significance for each tested fine-mapped SNP. First, for each SNP, we generated 3 di-nucleotide shuffled sequences using the 'fasta-dinucleotide-shuffle.py' script from the MEME suite (v5.4.1) [27]. For each of these shuffled sequences, we generated a 'reference' and 'alternative' allele sequence corresponding to the original SNP by replacing the central position of the shuffled sequence with the SNP's reference or alternative allele base. We calculated GkmExplain importance scores for each of these shuffled sequences across all clusters as described above, and calculated the cluster-specific 'delta score' using the central 50 bp of each null reference and alternative allele pair. The delta scores from these shuffled sequences served as a null distribution for each cluster. In agreement with a similar previous analysis, we found that the t-distribution was a good fit for these GkmExplain delta score null distributions [22]. We used the 'fitdistrplus' R package (v1.1.6) to fit a t-distribution to each cluster's delta score null distribution and used these distributions to calculate p-values for each fine-mapped SNP in each cell cluster. To further prioritize SNPs that are more likely to be affecting predicted accessibility by disruption of a transcription factor binding site, we calculated another previously described metrics of SNP effect size, the 'prominence score'[22]. To calculate this score, we first identified the 'active' allele for each SNP by the sign of the delta score, with a positive delta score indicating that the reference allele is more likely to be accessible than the alternative allele. We then identified the subsequence surrounding the active allele SNP where each position's GkmExplain importance score exceeded the 97.5th percentile of the di-nucleotide shuffled background. Subsequence boundaries were determined by the position where two consecutive bases had importance scores falling below the threshold. If a SNP sequence did

not contain a subsequence of at least 7 bases, we used the central 7 bases surrounding the active allele as the subsequence. We used these subsequences to calculate the prominence score for each SNP. To calculate the prominence score, we took the sum of non-negative GkmExplain importance scores from the active allele subsequence and then divided by the sum of the non-negative importance scores for the entire 250bp sequence. This score can be thought of as a measure of the signal to noise ratio of the active allele for each SNP. We fit an exponential distribution to the prominence null distribution for each cluster and again used these distributions to calculate prominence p-values for each fine-mapped SNP in each cell cluster. To prioritize SNPs that have significant effects on predicted chromatin accessibility (large absolute delta score), likely through disruption of a transcription factor binding site (large prominence scores), we used the estimated p-values of these two scores determined from the shuffled sequence null distributions. We selected “high-effect” fine-mapped SNPs that had both a delta score p-value < 0.05, and had a prominence score p-value < 0.05. To increase interpretability and further filter for likely causal SNPs, we further filtered “high-effect” SNPs by requiring that they fall in a peak linked to expression of a gene. Using these criteria, we identified 47, 19, and 19 prioritized SNPs for AGA, eczema, and hair color respectively (Table S29).

References

1. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
2. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, (2018).
3. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
4. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
5. 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).
6. 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).
7. Philippeos, C. *et al.* Spatial and Single-Cell Transcriptional Profiling Identifies Functionally Distinct Human Dermal Fibroblast Subpopulations. *J. Invest. Dermatol.* **138**, 811–825 (2018).
8. 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).
9. 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).
10. Stevens, W. W. *et al.* Activation of the 15-lipoxygenase pathway in aspirin-exacerbated respiratory disease. *J. Allergy Clin. Immunol.* **147**, 600–612 (2021).
11. Neubert, E. *et al.* Chromatin swelling drives neutrophil extracellular trap release. *Nat.*

Commun. **9**, 3767 (2018).

12. Sollberger, G., Tilley, D. O. & Zychlinsky, A. Neutrophil Extracellular Traps: The Biology of Chromatin Externalization. *Dev. Cell* **44**, 542–553 (2018).
13. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
14. 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).
15. Qu, J. *et al.* Mutant p63 Affects Epidermal Cell Identity through Rewiring the Enhancer Landscape. *Cell Rep.* **25**, 3490–3503.e4 (2018).
16. 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).
17. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
18. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
19. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *medRxiv* 2020.09.08.20190561 (2020).
20. 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* (2021) doi:10.1093/bioinformatics/btab122.
21. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
22. 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).

23. 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).
24. Ghandi, M., Mohammad-Noori, M. & Beer, M. A. Robust k-mer frequency estimation using gapped k-mers. *J. Math. Biol.* **69**, 469–500 (2014).
25. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).
26. Shrikumar, A., Prakash, E. & Kundaje, A. GkmExplain: fast and accurate interpretation of nonlinear gapped k-mer SVMs. *Bioinformatics* **35**, i173–i182 (2019).
27. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–8 (2009).