

Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics

Carmen Bravo González-Blas^{1,2}, Xiao-Jiang Quan^{1,2}, Ramon Duran-Romaña¹, Ibrahim Ihsan Taskiran^{1,2}, Duygu Koldere^{1,2}, Kristofer Davie¹, Valerie Christiaens^{1,2}, Samira Makhzami^{1,2}, Gert Hulselmans^{1,2}, Maxime de Waegeneer^{1,2}, David Mauduit^{1,2}, Suresh Poovathingal¹, Sara Aibar^{1,2} & Stein Aerts^{1,2*}

¹ VIB Center for Brain & Disease Research, Leuven, Belgium

² Department of Human Genetics, KU Leuven, Leuven, Belgium

* Corresponding author: stein.aerts@kuleuven.vib.be

Abstract

Single-cell technologies allow measuring chromatin accessibility and gene expression in each cell, but jointly utilizing both layers to map *bona fide* gene regulatory networks and enhancers remains challenging. Here, we generate independent single-cell RNA-seq and single-cell ATAC-seq atlases of the *Drosophila* eye-antennal disc and spatially integrate the data using a virtual latent space that mimics the organization of the 2D tissue. To validate spatially predicted enhancers, we use a large collection of enhancer-reporter lines and identify ~85% of enhancers in which chromatin accessibility and enhancer activity are coupled. Next, we infer enhancer-to-gene relationships in the virtual space, finding that genes are regulated by multiple redundant enhancers. Exploiting cell-type specific enhancers, we deconvolute cell-type specific effects of bulk-derived chromatin accessibility QTLs. Finally, we discover that Prospero drives neuronal differentiation through the binding of a GGG motif. In summary, we provide a comprehensive spatial characterization of gene regulation in a 2D tissue.

Introduction

Cellular identity is defined by Gene Regulatory Networks (GRNs), in which transcription factors bind to enhancers and promoters to regulate target gene expression, ultimately resulting in a cell type-specific transcriptome. Single cell technologies provide new opportunities to study the mechanisms underlying cell identity. Particularly, single-cell transcriptomics allow measuring gene expression in each cell; while single-cell epigenomics, such as single-cell ATAC-seq, serves as a read-out of chromatin accessibility¹. Although these technologies and computational approaches are recently evolving to include spatial information²⁻⁶, most approaches currently target single-cell transcriptomes. It remains a challenge how to exploit single-cell epigenomics data for resolving spatiotemporal enhancer activity and GRN dynamics, both experimentally and computationally.

In addition, while ATAC-seq is a powerful tool for predicting candidate enhancers, not all accessible regions correspond to functionally active enhancers⁷. For example, accessible sites can correspond to ubiquitously accessible promoters or binding sites for insulator proteins⁸; to repressed or inactive regions due to binding of repressive transcription factors^{7,9-11}; or to primed regions that are accessible across a tissue, but become only specifically activated in a subset of cell types¹². Importantly, single-cell ATAC-seq has not been fully exploited to explore these aspects yet. While most scATAC-seq studies have been carried out in mammalian systems - in which enhancer testing is not trivial -, Cusanovich *et al.* evaluated 31 cell-type specific enhancers predicted from scATAC-seq in the *Drosophila* embryo, finding that ~74% showed the expected activity patterns¹³.

Another current challenge in the field of single-cell regulatory genomics is how to integrate epigenomic and transcriptomic information. Although some experimental approaches have been developed for profiling both the epigenome and the transcriptome of the same cell¹⁴⁻¹⁶, currently either the quality of the measurements, or the throughput, is still significantly lower compared to each independent single-cell assay. For example, sci-CAR or SNARE-seq on human cells achieved a median of 1,000-4,000 UMIs (scRNA-seq) and 1,500-3,000 fragments (scATAC-seq) per cell; while the coverage with non-integrative methods, such as 10X, is around 20,000 UMI per cell and 10,000 fragments per cell for scRNA-seq and scATAC-seq, respectively^{14,15,17}. Methods that achieve high sensitivity, such as scCAT-seq¹⁶, are based on microwell plates rather than droplet microfluidics, making their throughput limited.

Given the current limitations of combined omics methods, the computational integration of independent high-sensitivity assays provides a valuable alternative. For example, Seurat¹⁸ and Liger¹⁹ have been used to integrate independently sequenced single-cell transcriptomes and single-cell epigenomes. Nevertheless, these methods require the “conversion” of the genomic region accessibility matrix to a gene-based matrix, and how to perform such a conversion is an unresolved issue. Some studies have

used the accessibility around the TSS as proxy for gene expression²⁰; others aggregate the accessibility regions that are co-accessible (i.e. correlated) with the TSS of the gene in a certain space²¹. However, promoter accessibility is not always correlated with gene expression. Furthermore, enhancers can be located very far from their target genes -upstream or downstream, up to 1 Mbp in mammalian genomes, or up to 100-200kb in *Drosophila*, often with intervening non-target genes in between - and relationships between enhancers and target genes are often not one-to-one (i.e. an enhancer can have multiple targets, and a gene can be regulated by more than one enhancer)⁷. Enhancer-promoter interactions can also be predicted using Hi-C approaches at the bulk level²², however these methods have limited sensitivity at single-cell resolution²³.

The *Drosophila* third instar larval eye-antennal disc provides an ideal biological system for the spatial modelling of gene regulation at single cell resolution. The eye-antennal disc comprises complex, dynamic, and spatially restricted cell populations in two dimensions. The antennal disc consists of 4 concentric rings (A1, A2, A3, and arista), each with a different transcriptome and different combinations of master regulators. For example, both Hth and Cut regulate the outer antennal rings (A1 and A2), with additional expression of Dll in A2; while Dll, Ss, and Dan/Danr are key for the development of the inner rings (A3 and arista), among others^{24,25}. On the other hand, a continuous cellular differentiation process from anterior to posterior occurs in the eye disc, in which progenitor cells differentiate into neuronal (i.e. photoreceptors) and non-neuronal (i.e. cone cells, bristle and pigment cells) cell types. This differentiation wave is driven by the morphogenetic furrow (MF). Posterior to the MF, R8 photoreceptors are specified first, and then they sequentially recruit R2/R5, R3/R4 and R7 photoreceptors and cone cells to form hexagonally packed units called ommatidia²⁶ (Fig 1a). In summary, the heterogeneity of cell types and differentiation trajectories results in diverse - static and dynamic - GRNs, which can be modelled with a combination of experimental and computational approaches.

In this work, we first generate a scRNA-seq and a scATAC-seq atlas of the eye-antennal disc. Second, taking advantage of the fact that the disc proper is a 2D tissue, we spatially map these single-cell profiles on a latent space that mimics the eye-antennal disc, called the virtual eye-antennal disc. Next, by exploiting publicly available enhancer-reporter data²⁷, we assess the relationship between enhancer accessibility and activity. Third, we use these virtual cells, for which both epigenomic and transcriptomic data is available, to derive links between enhancers and target genes using a new regression approach. Fourth, we use a panel of 50 bulk ATAC-seq profiles across inbred lines to predict cell-type specific caQTLs (chromatin accessibility QTLs). Finally, we use our findings to characterize the role of Prospero in the differentiation of photoreceptors. In summary, we provide a comprehensive characterization of gene regulation in the eye-antennal disc, using a strategy that is applicable to other tissues and organisms. Our results can be explored as a resource on Scope²⁸

(http://scope.aertslab.org/#/Bravo_et_al_EyeAntennalDisc) and the UCSC Genome Browser (http://genome.ucsc.edu/s/cbravo/Bravo_et_al_EyeAntennalDisc).

Results

A single-cell transcriptome atlas of the eye-antennal disc

First, we set out to identify the different cell populations in the eye-antennal disc, and obtain their transcriptomes. We profiled 3,531 high-quality cells using scRNA-seq on the 10X Genomics platform, with a median of 20,761 UMIs and 3,094 expressed genes per cell, respectively (Fig S1a, b). Analysis with Seurat revealed 17 clusters, most of which map to spatially located cell types (Fig 1b). Importantly, the structure in the tSNE - and UMAP (Fig S1c) -, reveals two main branches, one corresponding to the antennal disc, in which clusters represent the antennal rings from outer to inner; and one corresponding to the eye disc, in which progenitors differentiate into ommatidial (i.e., photoreceptors and cone cells) and interommatidial cell types. We verified that cell clustering was driven by cell identity and not affected by batch effects, using three independent biological replicates (Fig S1d). We also found a subset of ommatidial cells with a high number of UMIs and genes expressed, which was annotated as doublets by DoubletFinder²⁹ (Fig S1e-g). The higher proportion of doublets in this group is not unexpected, since ommatidia are tightly packed and are more difficult to dissociate.

To annotate these 17 cell clusters we combined two approaches. First, we used known marker genes from literature (Fig 1c, S1h). For example, we find *Dfd* expressed in the peripodial membrane clusters; with *dpp* expressed in the lateral peripodial membrane³⁰; and *oc* as key marker of the head vertex³¹. In the eye disc we find a gene expression gradient starting from *Optix* expression in progenitors and precursors; to *ato* expression in the MF, and then *gl* expression in the ommatidial and interommatidial cells. Importantly, we find *Gasp* as key marker of the interommatidial cells (Fig 1c), which plays a role in extracellular matrix integrity and assembly³². Indeed, Gene Ontology (GO) enrichment of the genes differentially expressed in this group reveals terms related to cell-cell junction assembly and organization (p-val: 10^{-16}). Meanwhile, in the ommatidial groups we observed a gene expression gradient of markers from early photoreceptors (R8, *sens*), to intermediate (R3-4, *svp*) and late-born PRs and cone cells (R7 and cone cells, *sv*)³³⁻³⁵. In fact, semi-supervised analysis of these populations (see *Methods*) subdivides the ommatidial classes into the different photoreceptor types and cone cells (Fig S2), largely finding R8, R3/R4 and R1/R6 in the early-born PRs cluster, and R7 and cone cells in the late-born photoreceptors and cone cells cluster (only 26 R2/R5 cells are detected). On the other hand, markers of the antennal rings form a gradient along the antennal cell types, from *ct* (A1 and A2), to *Dll* (A2, A3 and arista) and *ss* (A3 & arista)²⁵. Interestingly, within A2, we find a subpopulation of cells expressing *ato* and *sens*, corresponding to the Johnston Organ Precursors (JOPs)^{36,37}. We also identify a population of glial cells, based on the enrichment of the transcription factor *repo*³⁸; a cluster of

hemocytes, enriched for *pnr*³⁹; and a small group of cells with high expression the transcription factor *twi*, corresponding to ad epithelial cells (mesodermal myoblasts), which are known to reside in most imaginal discs^{40,41}. Accordingly, GO term enrichment using the differentially expressed genes in this group reveals terms related to mesoderm development (p-val: 10⁻⁴). Finally, we found a population of 299 cells coming from the brain expressing *Oli*⁴²; which represent contaminating cells from the brain due to the dissections.

To validate and further extend our cell type annotations, we used a publicly available Drop-seq dataset from the eye disc containing 11,500 single cell profiles⁴³, with a median of 517 genes detected per cell (Fig S3a,b). Using Seurat's label transferring, we mapped the previously annotated cell types on our dataset (and vice versa) and found that both annotations agreed (Fig 1d, Fig S3c,d). These labels permitted to subdivide our glial cell cluster into wrapping glia, subperineural glia and perineural glia; and to annotate a small population of cells just posterior to the MF as the second mitotic wave (SMW), which is a round of synchronous cell division that occurs right after cells exit the MF (Fig S3e). On the other hand, no *twi*⁺ cells are found in the Drop-seq dataset. This is likely due to the fact that these cells are located in the antennal disc, which is missing in Drop-seq dataset. Indeed, the activity of a *twi* enhancer (Fig S3f) is observed in the antennal disc rather than in the eye disc²⁷.

Next, we used SCENIC to identify master regulators and gene regulatory networks in the eye-antennal disc⁴⁴, resulting in 175 regulons (159 motif-based regulons and 16 regulons based on ChIP-seq tracks; see *Methods*). While some regulons are enriched across the entire tissue (such as Grh¹²), many are cell-type specific. For the antennal rings, we find Lim1 (A1), TfAP-2 (A1, A3 and arista) and ss (arista and A3), in agreement with literature⁴⁵⁻⁴⁷. In the eye disc, regulons recapitulate the GRN dynamics during the differentiation process, with Optix and Ey in the progenitor and precursor cells, Ato in the morphogenetic furrow, So in photoreceptors and B-H2 in interommatidial cells, among others^{48,49}. We further validated the Ato regulon using previously published RNA-seq data from Ato gain-of-function and loss-of-function mutants⁵⁰ (Fig S4). Indeed, genes included in the Ato regulon are significantly upregulated in the Ato gain-of-function (NES: 2.44) and downregulated in the *ato* mutant (NES: -2.57), respectively.

In conclusion, using scRNA-seq we have identified the different populations in the eye-antennal disc and the interplay of GRNs that underlie the developmental program of this system. We provide this data as loom files that can be explored in SCoPe at: http://scope.aertslab.org/#/Bravo_et_al_EyeAntennalDisc.

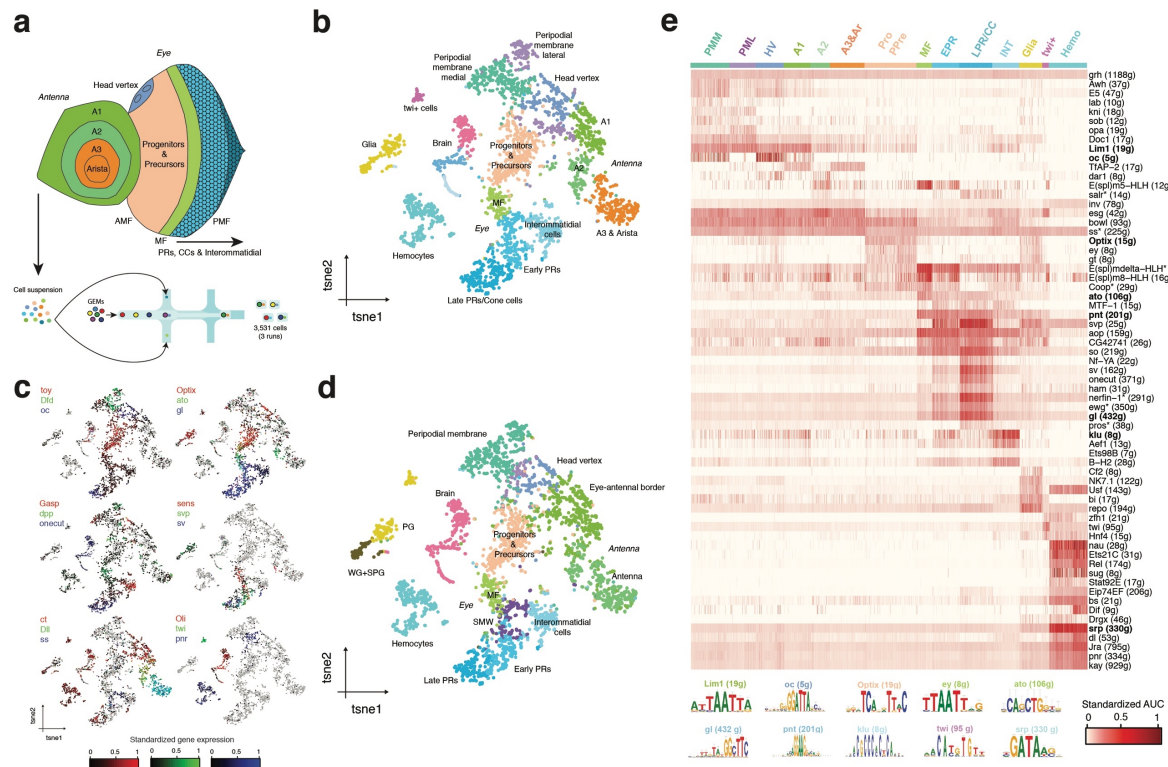


Figure 1: scRNA-seq recapitulates cellular diversity and GRNs in the eye-antennal disc. **a.** Experimental approach. scRNA-seq was performed in eye-antennal discs using 10X Genomics, resulting in a data set with 3,531 high quality cells. Main spatial compartments in the eye-antennal disc are annotated. **b.** tSNE representation of the scRNA-seq data (with 3,531 cells). **c.** tSNE colored by the standardized gene expression of known cell type markers in the eye antennal disc. In each plot three marker genes are shown, using RGB encoding. **d.** tSNE annotated by label transfer with Seurat v3¹⁶ using the scRNA-seq eye disc data set from Ariss et al.⁴³ **e.** Cell-to-regulon heatmap showing the standardized enrichment or Area Under the Curve (AUC) from SCENIC⁴⁴ for each selected regulon based on RSS in each cell. Top enriched motifs for representative regulons are shown below. Regulons marked with * are based on ChIP-seq track enrichment. AMF: Anterior to the morphogenetic furrow. PMF: Posterior to the morphogenetic furrow. PR: Photoreceptor. CC: Cone cell. PMM: Peripodial membrane medial. PML: Peripodial membrane lateral. HV: Head Vertex. Pro: Progenitors. Pre: Precursors. MF: Morphogenetic furrow. EPR: Early photoreceptors. LPR/CC: Late photoreceptors and cone cells. INT: Interommatidial cells. Hemo: Hemocytes. SMW: Second mitotic wave. PG: Perineurial glia. WG: Wrapping Glia. SPG: Subperineurial glia.

A single-cell ATAC-seq atlas of the eye-antennal disc

Next, we performed scATAC-seq to explore the chromatin accessibility landscape of the eye-antennal disc. Using 10x Chromium, we obtained 15,766 scATAC profiles (Fig 2a). We assessed the quality of the data set based on the relative enrichment of fragments around the TSS and the correlation with bulk ATAC-seq on the same tissue ($R^2=0.9615$; Fig 2b), among other quality control measures; and filtered

out a total of 379 cells based on the number of fragments within bulk peaks and the total number of fragments (Fig S5). This resulted in a data set with 15,387 single-cell epigenomes.

We ran cisTopic²⁰ using three different sets of regulatory regions: (1) narrow peaks as called by MACS2 from the bulk ATAC-seq profile of the wild type *Drosophila* eye-antennal disc; (2) bulk peaks defined by extending +/- 250 bp from the summits called by MACS2; and (3) cisTarget regions, defined by partitioning the entire non-coding *Drosophila* genome based on cross-species conservation, resulting in more than 136,000 bins with an average size of 790 bp⁵¹. We found that the cisTopic analysis performed with cisTarget regions resulted in the highest resolution compared to using bulk peaks or summits (Fig S6). For example, small subpopulations such as brain and *twi*⁺ cells could be revealed, which were otherwise mixed with ommatidial and glial cells, respectively. Hence, we used this model, with 49 topics, for further analysis.

Clustering on the topic-cell distributions (i.e., the contribution of each regulatory topic to each cell) resulted in 22 clusters, most of which map to spatially located cell types (Fig 2c). Despite the fact that cell clustering is not driven by read coverage, we find two groups that likely correspond to doublets based on read depth and percentage of reads in peaks (Fig S7a-c). Annotation of cell types from scATAC-seq is not as straightforward as for scRNA-seq, because the cluster markers now represent regulatory regions instead of genes. To address this, we exploit four different approaches: (1) motif (and ChIP-seq track) enrichment on the regulatory topics; (2) enrichment of epigenomic signatures of FAC-sorted cell types; (3) a novel method for deriving gene activity scores from cisTopic distributions; and (4) label transferring from our previously annotated scRNA-seq data set.

Of the 49 predicted topics, two represent a batch effect of the run, and one represents a female sex-specific topic (Fig S7d-g). The remaining topics represent general, cell-type specific and low contribution topics (Fig 2d, Fig S8, Fig S9a). Among the cell-type specific topics we find a topic for each antennal ring (topics 19, 26, 40 and 22; respectively), with a subdivision of A2 in two groups (A2a and A2b, respectively). Regions in these topics, from the outer to the inner ring, are enriched for motifs (and/or ChIP-seq tracks) linked to known master regulators, such as Hth in A1, Dll in A2, and Ss in A3 and arista (Fig S8). Additionally, we identify a subpopulation of cells in A2b with accessible regions controlled by Ato, which correspond to the Johnston Organ Precursors (JOPs). Similarly, retinal developmental topics recapitulate the dynamic changes in chromatin during differentiation, with the Optix motif enriched in regions specific to the domain anterior to the MF; the Ato motif in MF specific regions; the Glass, Sine oculis (So), and Onecut motifs in the clusters representing ommatidial cells; and the Glass, So, and Lozenge motifs in interommatidial cell types; among others (Fig 2e). Furthermore, we also find a new, highly enriched GGG motif in the genomic regions specific to ommatidial development, which can be linked to a relatively large set of candidate TFs based on motif-

to-TF mappings, as will be discussed further below. We also discovered generally accessible topics, highly enriched for promoters (Fig S9b), some of which decrease in accessibility during ommatidial development. These epithelial topics are represented by genomic regions bound by the pioneer transcription factors Trl and Grh, based on motif and ChIP-seq enrichment (Fig 2e). Indeed, Grh has been shown to be expressed and promote chromatin opening in all epithelial cells, decreasing upon neuronal differentiation¹²; which is also supported by our scRNA-seq data set (Fig S9c). We also identify other cell-type specific topics for other subpopulations; such as a topic enriched for the Twist motif that identifies the *twi*⁺ ad epithelial cells; a topic enriched for the Serpent (Srp) motif, corresponding to hemocytes; and a topic enriched for the Repo motif, corresponding to glial cells. Finally, we also identify two small subpopulations with topics enriched for Stripe (Sr), which correspond to brain cells likely attached during the dissection.

To further validate these cell type annotations, we used our previously published ATAC-seq data from FAC-sorted cells located specifically anterior to the morphogenetic furrow, based on the activity of the Optix2/3 enhancer driving GFP (Optix-GFP⁺, Fig 2f)^{12,52}. We find that regions specifically accessible in these cells compared to the rest of the eye are accessible in the cells identified as precursors in the scATAC-seq data; and also show enrichment for the motif of the transcription factor Optix, in agreement with the topic specific to this population. We also re-used our previous single-cell ATAC-seq data, obtained on the Fluidigm C1, of Optix-GFP⁺ FAC-sorted cells¹² and we performed an additional Fluidigm C1 run with cells FAC-sorted based on the activity of the sens-F2 enhancer⁵³ (sens-GFP⁺), which correspond to the intermediate groups in the MF and R8 photoreceptors. When mapping these cells into the topic space, we find that they cluster within the correct cell types of the 10X sc-ATAC-seq data (Fig S9d). Accordingly, we also find that the activity of the Optix 2/3 enhancer and the sens-F2 enhancer agrees with the accessibility of these regions in the matching cell types (Fig S9e-f).

Next, we developed a new approach for deriving a “gene activity matrix” from the topic-cell and region-topic distributions (Fig 2g). Briefly, we first multiply the region-topic and topic-cell distributions to obtain a region-cell distribution, which indicates the probability of accessibility of each region in each cell. Then, for each gene, we aggregate the probabilities of the surrounding regions (in this case, 5kb around the TSS plus introns), resulting in a gene activity score. This new matrix, which contains scATAC-seq cells as columns and gene activities as rows, can be analyzed as a gene expression matrix. For example, we used it to score SCENIC regulons on the scATAC-seq cells to validate the master regulators found in the topics (Fig 2g). We find the Optix regulon enriched anterior to the morphogenetic furrow; the Ato regulon enriched in the MF; Onecut enriched in late ommatidial cells; and Grh enriched across all cell types except late ommatidial cells. Furthermore, we also used DoubletFinder²⁹, developed for scRNA-seq data, and labelled a group of cells enriched in both ommatidial and interommatidial topics as doublets (Fig S9g). In addition, we used this matrix for label

transferring with our scRNA-seq data set using Seurat v3¹⁸, finding a strong agreement between our independent RNA and ATAC-based annotations (Fig 2h, Fig S9h). Interestingly, we find that when mapping groups from scRNA-seq to scATAC-seq some cells in the eye are assigned to the next developmental cluster (i.e. some cells annotated as MF are labelled as early PRs). This lag effect may be explained by the fact that chromatin accessibility changes could occur slightly before changes in the steady-state transcriptome during differentiation.

Importantly, we find regions enriched for a specific motif that are located in the surroundings of genes (learned from the scATAC-seq data) that are co-expressed with the corresponding transcription factor (learned from the scRNA-seq data), likely representing *bona fide* functional enhancers. For example, we find 2,769 regions enriched for the Optix/so motif; out of which 505 and 894 are in the surroundings of genes co-expressed with Optix and so, respectively. Similarly, out of the 1,859 and 1,128 regions enriched for the Atonal and the Glass motifs, 285 and 452 are close to co-expressed genes (Figure S10; Table S1).

In summary, we provide a thorough characterization of the chromatin accessibility landscape of the eye-antennal disc, corroborated by our scRNA-seq data set. This data can also be explored at SCoPe (http://scope.aertslab.org/#/Bravo_et_al_EyeAntennalDisc) and UCSC (http://genome.ucsc.edu/s/cbravo/Bravo_et_al_EyeAntennalDisc).

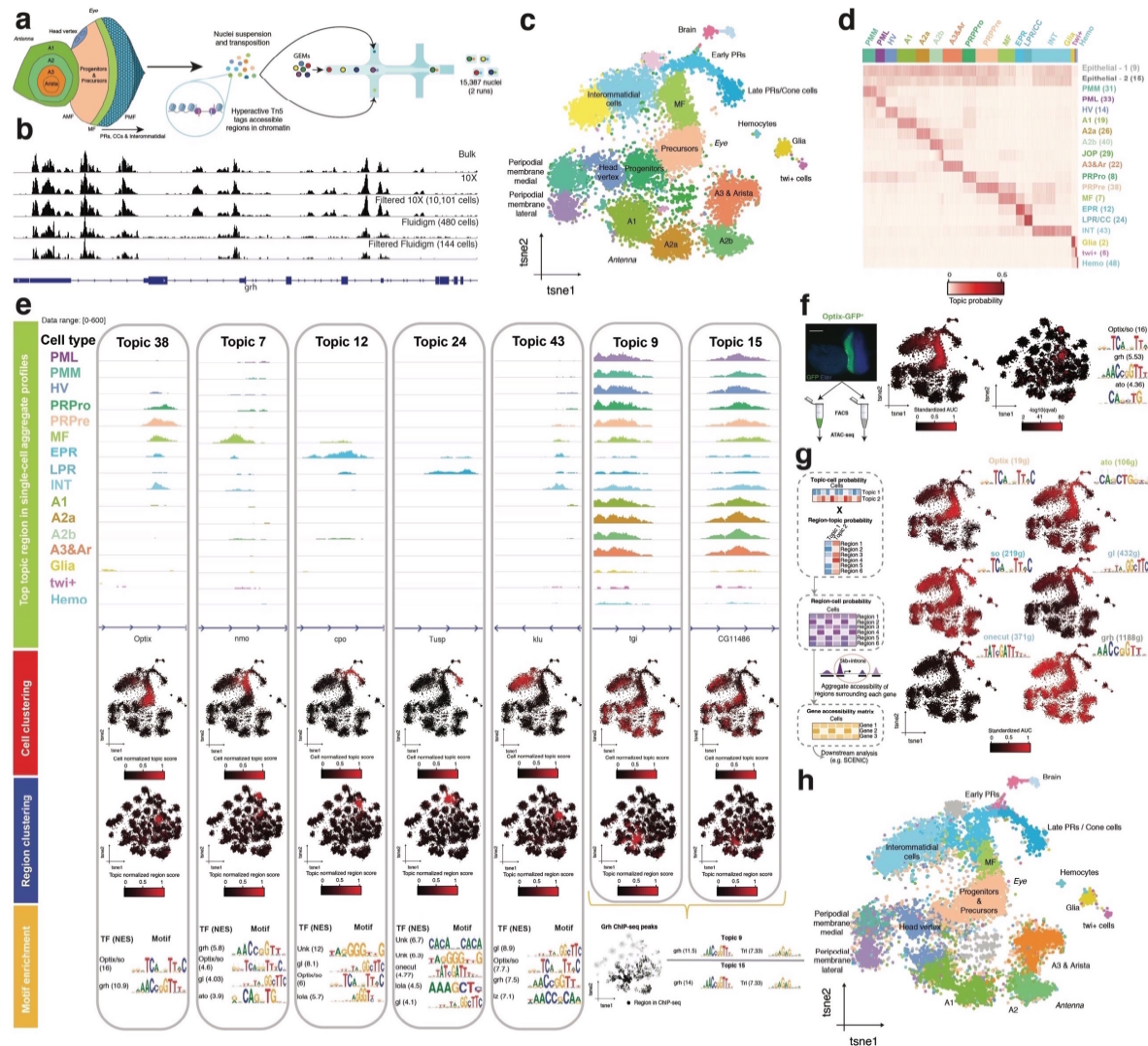


Figure 2: scATAC-seq recapitulates cell diversity in the eye-antennal disc. **a.** Experimental set up. 15,387 nuclei were profiled using 10X scATAC-seq. **b.** Comparison of bulk ATAC profiles, scATAC-seq aggregates with 10X (from the 1st run) and scATAC-seq aggregates from FACS sorted populations (Optix-GFP⁺ and sens-GFP⁺) profiled with Fluidigm C1. Number of cells in each aggregate is indicated between brackets. **c.** cisTopic cell tSNE (15,387 nuclei) colored by annotated cell type. **d.** Topic-cell enrichment heatmap with selected topics. **e.** Topic modelling recapitulates the dynamic chromatin changes during differentiation in the eye disc. Top: Aggregate profiles per cell type in the top region of the indicated topic. Middle-top: cisTopic cell tSNE colored by topic enrichment. Middle-bottom: cisTopic region tSNE colored by topic enrichment. Bottom: Selected enriched motifs in each topic. For topics 9 and 15, the cisTopic region tSNE colored by Grh ChIP-seq peaks. **f.** Bulk ATAC was performed on Optix-GFP⁺ and Optix-GFP⁻ FACS sorted cells (based on the activity of the Optix2/3 enhancer). cisTopic cell tSNE and region tSNE are colored based on the enrichment of regions that are differentially accessible between Optix-GFP⁺ and Optix-GFP⁻. Motifs enriched in the regions differentially accessible in Optix-GFP⁺ cells are shown. Scale bar, 100 μ m. **h.** cisTopic topic-cell and region-topic distributions can be exploited to predict the probability of each region in each cell. By aggregating the probabilities of regions around the TSS of each gene (in this case, 5kb upstream and introns), a gene accessibility matrix can be derived. cisTopic cell tSNE is colored based on the enrichment of regulons derived from scRNA-seq, evaluated in the gene

accessibility matrix. **g.** cisTopic cell tSNE colored by the scRNA-seq annotation after label transferring with Seurat v3 using cisTopic gene accessibility matrix. PMM: Peripodial Membrane Medial. PML: Peripodial Membrane Lateral. HV: Head Vertex. Pro: Progenitors. Pre: Precursors. MF: Morphogenetic Furrow. EPR: Early photoreceptors. LPR/CC: Late photoreceptors and cone cells. INT: Interommatidial cells. Hemo: Hemocytes. JOP: Johnston Organ Precursor.

Spatiotemporal mapping of single-cell omics couples enhancer accessibility with functionality

Since most of the cell types in the eye-antennal disc map to locally restricted populations, we developed a strategy to map the scRNA-seq and scATAC-seq profiles to their putative position of origin in the tissue using a template of the eye-antennal disc with 5,058 virtual cells, corresponding to the 5,058 pixels in our eye-antennal disc representation (Fig 1a, Fig 2a, Fig 3a). Briefly, we first order antennal and eye cells by pseudotime (see *Methods*), which correspond to the proximal-distal and anterior-posterior axes in the antenna and the eye, respectively (Fig S11). For each cluster, we divide real and virtual cells into bins based on pseudotime and position in the corresponding axis, respectively. Finally, we map real cells onto the virtual cells in the matching bin in the virtual eye-antennal disc, with a 1-to-1 matching. When there are fewer real cells than virtual cells in the bin, real cells are sampled randomly more than once; and when more real cells are available than virtual cells, N real cells are sampled, where N is the number of virtual cells in that bin.

Using the mapped scRNA-seq data we can visualize previously known gene expression patterns (Fig 3b). For example, our spatial map recapitulates expression of *hth*, *salm*, *danr*, *ct*, *Dll* and *ss* in the antennal rings; as shown by Emerald *et al.*²⁵. In the eye part, patterns from anterior to posterior, with the expression of *oc* in the head vertex, *Optix* and *toy* anterior to the MF, *ato* and *dpp* in the MF, and *gl* posterior to the MF, among others, agree with literature⁴⁹.

To validate the scATAC-seq mapping we used image data of more than 700 enhancer-reporter lines from the Janelia Flylight project²⁷. In short, in each line a specific enhancer controls the expression of GAL4, and when crossed with a UAS-GFP reporter line the activity of the enhancer is recapitulated by the GFP signal. These enhancer activity patterns were registered onto the virtual eye-antennal disc using a custom landmark-based pipeline (see *Methods*). By comparing the predicted accessibility pattern for each region (based on ATAC-seq) with its activity (based on the GFP reporter), we find that accessibility and activity are correlated in 77% of the enhancers (Fig 3c); however, there are cases in which accessibility and activity are uncoupled (Fig 3d). We find that specific enhancers (with a high gini score) tend to agree in their accessibility and activity; while ubiquitously accessible enhancers (with a low gini score) do not show corresponding accessibility and activity patterns (Fig 3e). In addition,

motifs linked to transcription factors with a restricted expression, such as Glass (posterior to the MF) and Ocelliless (head vertex), are found in the specifically accessible enhancers; while motifs linked to Grainyhead, an epithelial transcription factor, are found in the generally accessible regions. Indeed, Jacobs *et al.* showed that Grh is a pioneer transcription factor which directly promotes opening of all its target regions throughout the epithelial tissue of the eye-antennal disc, while their activity is restricted to certain cell-types¹². Our data confirms that Grh binding results in a general ATAC-seq signal, but not necessarily in activity. For example, among 20 Atonal target enhancers found earlier⁵⁰, 6 are bound by Grh and are ubiquitously accessible, yet activated only in Ato positive cells; while the other 14 enhancers are not bound by Grh and these show cell-type specific accessibility (Fig S12).

Thus, the scATAC-seq data corroborates a model consisting of two classes of enhancers: (1) primed enhancers, with general accessibility (e.g. by Grh binding) but specific activity based on the presence of other transcription factor/s (e.g. Ato) and (2) unprimed regions, in which accessibility (e.g. by binding of a TF/s, as Ato) and activity are coupled (Fig 3f). Most of the enhancers of the first class belong to the general topics (with a total of 4,500 binarized regions on the representative general topics); while regions from the second class are spread across the cell-type topics (with a total of 26,500 regions classified in cell type specific topics). In summary, accessibility can be used as a proxy for enhancer activity for the majority of enhancers, but there are ~15% of enhancers that form an exception.

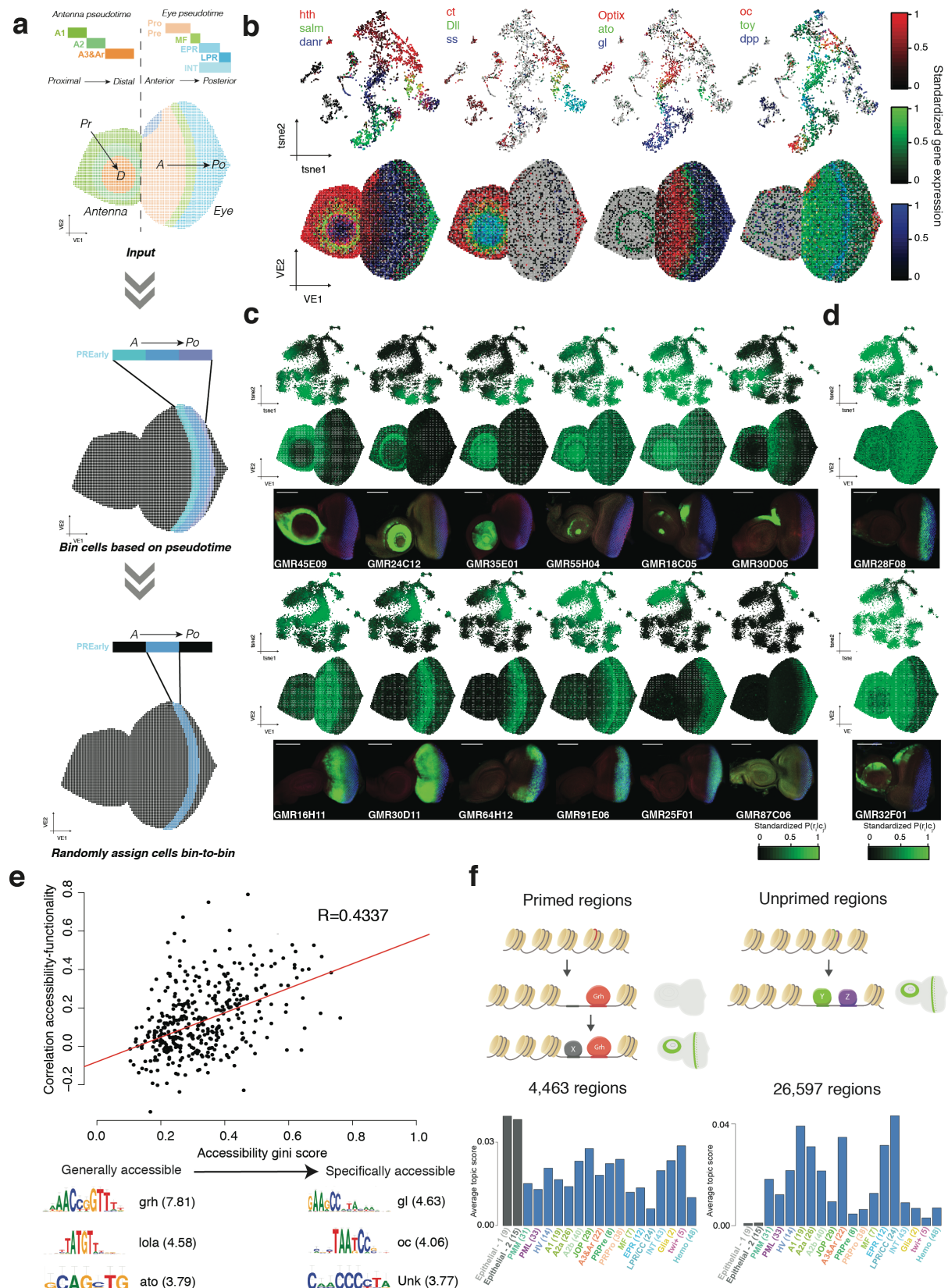


Figure 3: Spatiotemporal mapping of single cell RNA-seq and single cell ATAC-seq data. **a.** Computational approach for mapping single cell RNA or single cell ATAC-seq data into the virtual eye-antennal disc. Briefly, cells are ordered by pseudotime, corresponding to the proximal-distal axis in the antennal disc and the anterior-posterior axis in the eye disc. For each cluster, real and virtual cells are divided into the same number bins based

on pseudotime and axis position, respectively. Finally, cells are mapped into the virtual cells in the matching bin.

b. Gene expression correspondence between the Seurat tSNE and the virtual eye. The expression of three genes are shown per plot, using RGB encoding. **c.** Correspondence between region accessibility and activity for 12 *Janelia-Gal4* enhancers. Top row: cisTopic cell tSNE colored by the accessibility probability of each region in each cell. Middle row: Virtual eye colored by the accessibility probability of each region in each cell. Bottom row: Confocal images showing the activity (GFP, green) of each region in eye-antennal discs. Scale bar, 100 μ m.

d. Discordance between region accessibility and activity for 2 *Janelia-Gal4* enhancers. Top row: cisTopic cell tSNE colored by the probability of each region in each cell. Middle row: Virtual eye colored by the probability of each region in each cell. Bottom row: Confocal images showing the activity (GFP, green) of each region in eye-antennal discs. Scale bar, 100 μ m. **e.** Relationship between the correlation between the accessibility and the activity of the regions and their distribution (as gini score). Below, representative motifs enriched in generally and specifically regions, with low (< 0.2) and high (> 0.4) gini score, respectively, are shown. **f.** Model describing the two classes of enhancers found. On one hand, some enhancers (such as *grh* targets) are generally accessible, but only become functional with a specific co-factor(s) binds; on the other hand, for other enhancers, accessibility is more specific and is couples with activity (based on the binding of one or more TFs). Histograms shown the average topic score for enhancers of both classes are shown. PMM: Peripodial Membrane Medial. PML: Peripodial Membrane Lateral. HV: Head Vertex. Pro: Progenitors. Pre: Precursors. MF: Morphogenetic Furrow. EPR: Early photoreceptors. LPR/CC: Late photoreceptors and cone cells. INT: Interommatidial cells. Hemo: Hemocytes. JOP: Johnston Organ Precursor.

Exploiting the latent space to link enhancers to target genes

The virtual eye-antennal disc acts as a latent space in which both transcriptomic and epigenomic profiles are available in the same virtual cell. Hence, we developed a computational strategy to infer enhancer-to-gene relationships. Particularly, we investigated to what extent enhancers in a large space around the TSS of a gene (i.e. ± 50 kb from the TSS plus introns) can predict the expression of the gene (Fig 4a). For each gene, we calculated: (1) the correlation between gene expression and the accessibility probability of each candidate region across all the virtual cells and (2) the importance of each candidate region for predicting the expression of the gene using random forest regression models, which assess non-linear relationships. We used the sign of the correlation score to classify links as positive (> 0) or negative (< 0) and the random forest importance as measurement of the confidence of the links. After pruning low confidence links (see *Methods*), we obtained a total of 183,336 enhancer-to-gene relationships.

To verify these predicted enhancer-to-gene links, we used validated associations from literature. For example, we predict *sens* expression to be exclusively regulated by one enhancer, *sens-F2*, as proven by Pepple *et al.*⁵³ (Fig 4a). In other cases, we find that gene expression is a result of combinations of enhancers. For instance, *dac* expression is mainly controlled by two redundant enhancers (3EE and

5EE), as shown by Pappu *et al.*⁵⁴. Both enhancers are accessible in the precursor cells, where *dac* is expressed (Fig S13a). As a more complex example, *gl* expression is regulated by a combination of 14 enhancers, out of which 3 enhancers have been validated by Fritsch *et al.*⁵⁵ (Fig 4b, Fig S13b). While *gl* is expressed in all cell types posterior to the MF (ommatidial and interommatidial); some of these regions are exclusively accessible and active in interommatidial cells (i.e. subregion in enhancer 3), while others are only accessible and active in photoreceptors (i.e. enhancer 2, subregion in enhancer 3 and enhancer 4); suggesting that different enhancer combinations are involved in ommatidial cells versus interommatidial cells. Overall, there is a median of 22 enhancers linked to each gene and only 2.4 % of all genes are regulated by one enhancer (Fig 4d). These results indicate that gene expression is regulated by an intricate network of enhancer interactions. Further corroborating these links, we find that ~62% of the *Janelia* enhancers for which accessibility and activity are coupled are positively linked (with a correlation > 0.2) to a target gene.

Interestingly, TF genes are regulated by significantly more enhancers compared to non-TF genes (H_0 : average number of positive links for TF genes (13) \leq Average number of positive links for non-TF genes (11); p-value: 2×10^{-4} ; Fig S13c). This is further supported by Gene Ontology analysis with GOrilla (Genes ordered by decreasing number of links, p-value: 5×10^{-4} ; Fig S13d-e). Indeed, it has been hypothesized that TF genes require a tighter regulation because abnormalities in their expression can cause more dramatic effects compared to defects in the expression of terminal effector genes⁵⁶. In addition, we find enhancer-enhancer pairs linked to the same gene with a high correlation in accessibility (with a median of 4 enhancer-enhancer pairs with a correlation > 0.8; equivalent to 3-4 redundant enhancers), being significantly higher between enhancers linked to a TF gene compared to those linked to a non-TF gene (H_0 : average number of enhancer-enhancer pairs with a correlation above 0.8 for TF genes (13) \leq Average number of enhancer-enhancer pairs with a correlation above 0.8 for non TF genes (7); p-value: 0.006; Fig S13f). In agreement, ~73% of the enhancer-enhancer pairs involving *Janelia* enhancers also show correlation between their activity patterns (Fig S14). The multiplicity of enhancers with the same function, known as shadow enhancers, has an evolutionary basis and provides robustness during development⁵⁷. In fact, redundant enhancers can compensate when an enhancer is affected by a loss-of-function mutation or deletion^{58,59}. Altogether, we conclude that genes are regulated by many enhancers, likely with a redundant function to ensure an accurate regulation of gene expression.

Based on the 183k enhancer-to-gene associations, we investigated the distance between regions and their predicted genes, the genomic annotation and the motif composition of the enhancers involved in these networks. Firstly, we found that enhancers do not necessarily act on their closest gene, although the nearest gene is overall the most likely target (Fig 4e). Secondly, most regions linked to a target gene fall in non-promoter regions (75%) (Fig S13g). Indeed, for 84% of the genes that show cell-type specific

expression (with adjusted p-value < 0.05 and average log FC > 1), the accessibility of the promoter is not correlated with the expression of the gene (correlation < 0.5), as promoters tend to ubiquitously accessible (H_0 : Proportion of promoters in generally accessible topics (0.54) <= Proportion of promoters across all topics (0.36), p-value < 2.2×10^{-16}). Interestingly, enhancer accessibility can be positively (95,484; of which 13,125 are uniquely positive with a correlation > 0.1) or negatively (87,229; of which 2,927 are uniquely negative with a correlation < -0.1) correlated with target gene expression. Negative correlation is suggestive of gene repression, whereby a repressor binds to the enhancer, creating an accessible region, only in the cells where the gene is not expressed. On the 'activating' enhancers we find Lola-T/K, AP-1 and Onecut related motifs enriched, among others. Interestingly, the GGG motif previously found in the ommatidial cell types enhancers is also enriched in these links. On the other hand, on the 'repressive' enhancers we identified motifs linked to so/Optix, Lz and Blimp-1, among others. While Blimp-1 and Lz can act as repressors in *Drosophila*^{60,61}, so and Optix have been suggested to act as either activators or co-repressors (anterior to the morphogenetic furrow) during eye development⁶². For instance, by looking at enhancer-to-gene links related to *hth*, a gene potentially repressed by So⁶³, we find a repressive enhancer (chr3R:10563160-10564462) with a So binding site (based on ChIP-seq) that is also enriched in the Optix-GFP⁺ FAC-sorted cells (Fig S15). In fact, this enhancer is specifically accessible in the cells in which both Optix and so are expressed, while *hth* is repressed in these cells. This suggests that Optix and so may cooperate to repress *hth* in the eye precursor cells via this regulatory region.

Next, we used the inferred enhancer-to-target genes in an attempt to improve the inference of a “gene activity matrix” from the scATAC-seq data (i.e., predicting gene expression from ATAC-seq peaks). Briefly, instead of aggregating the probability of all the regions around a certain space around the TSS (i.e. 5kb upstream the TSS and introns as used above) of the gene of interest, we calculate the gene activity score by the weighted sum (weighted by importance) of the accessibility probabilities of the enhancers linked to each gene. We were able to recapitulate previously observed gene expression patterns (Fig 1c, Fig 4g), supporting the robustness of the inferred links. For instance, we found the expression gradient of *ct*, *Dll* and *ss* from outer to inner antennal rings, and a gradient from *Optix*, to *ato*, and to *gl* driving differentiation in the eye^{25,49}.

We then exploited these enhancer-to-gene associations to create new regulons, now being able to extend the search space for motif discovery around each gene. Particularly, comparison to the SCENIC workflow⁴⁴, in which after deriving co-expression modules per TF the target genes are selected based on the enrichment of the motif/s linked to the TF in the entire sequence space around the TSS (i.e. 5kb upstream the TSS and introns); we evaluated motif enrichment restricted to the regions that are linked to each potential target gene. Out of the 161 regulons predicted in this manner, 91 have a canonical SCENIC counterpart, and have average size 2.6 times smaller than the SCENIC regulons (Fig S16a).

In addition, this approach identifies new regulons that were not found with SCENIC, such as Toy and Zld, which are involved in differentiation of ommatidial cell types; Salm in ommatidial cell types; Ct in R7 and cone cells and A1; Dll in A2, A3 and arista and Dfd in the peripodial membrane^{25,30,49,64,65} (Fig S16b, c). We further validated the link-based regulons using differential expression rankings from Ato gain and loss-of-function mutant (versus WT), GMR-GFP⁺ cells (versus GMR-GFP⁻), and a loss-of-function mutant of *onecut* (versus WT). We found that the predicted genes in the Ato regulon are upregulated in the GOF mutant and downregulated in the LOF mutant; Glass predicted target genes are enriched in GMR-GFP⁺ cells; and the predicted *onecut* regulon is downregulated in the *onecut* LOF (Fig 4h). In addition, we find an overlap of 24% when comparing the predicted Glass binding sites with Glass ChIP-seq binding sites in the embryo⁶⁶.

In summary, we provide a new method to infer GRNs involving distal enhancers, and a resource of enhancer-to-gene relationships that can be exploited to validate basic principles of gene regulation and infer detailed gene regulatory networks.

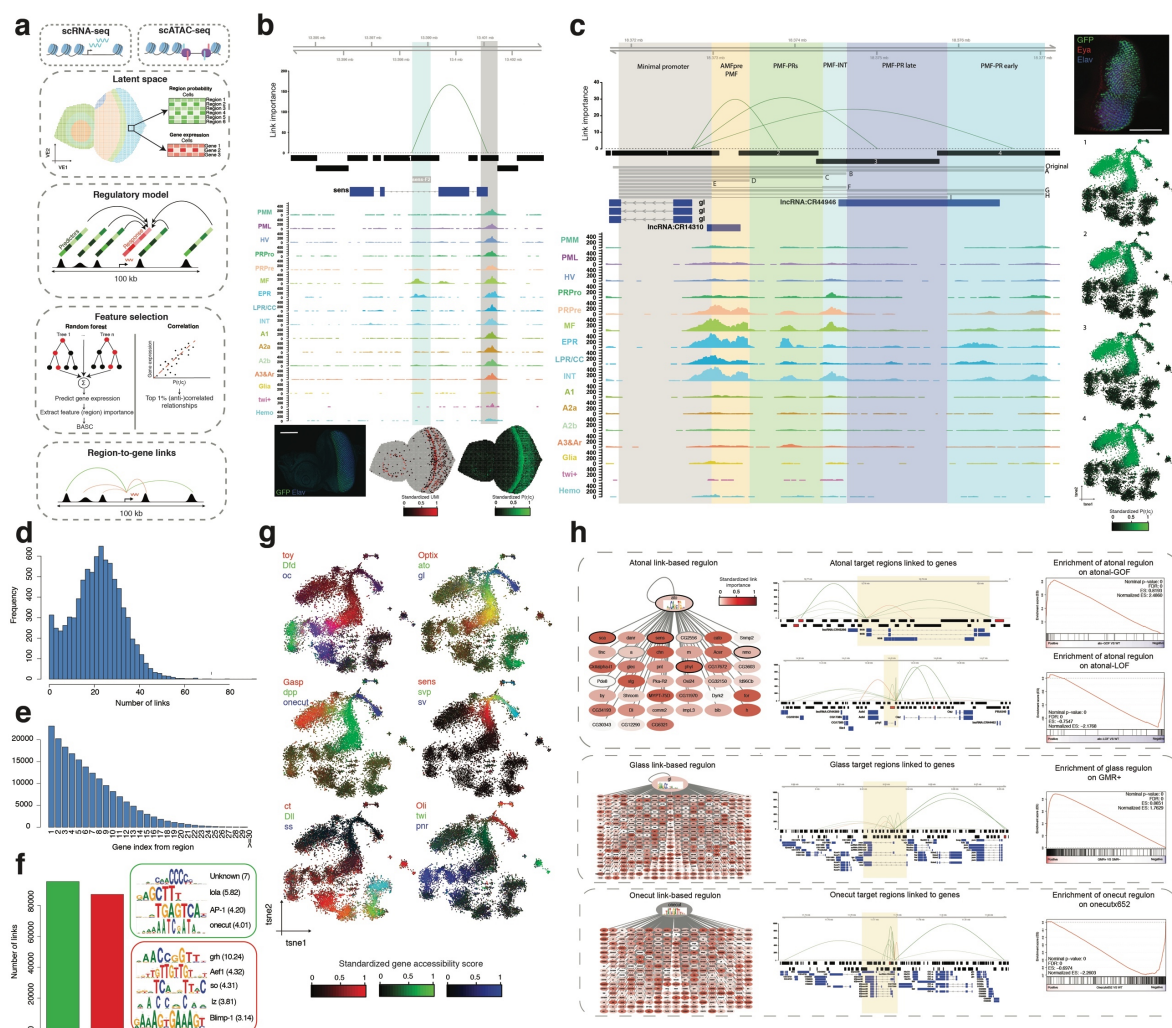


Figure 4: Enhancer-to-target links unveil a complex multi-level regulation of gene expression. a. Computational approach for linking enhancer to target genes. **b.** Example of a gene (*sens*) controlled by uniquely

one enhancer (sens-F2). Linkage importance is shown on top, followed by the cisTarget regions, the sens-F2 enhancer, gene annotation and the aggregated ATAC-seq profiles of the different cell types. The activity of sens-F2 is shown below, together with the virtual eye-antennal disc colored by sens gene expression (red) and sens-F2 accessibility probability (green). Scale bar: 100 μ m. **c.** Example of a gene (gl) controlled by multiple enhancers, validated by Fritsch *et al.*⁵⁵. Linkage importance is shown on top, followed by the cisTarget regions, the constructs tested by the authors, gene annotation and the aggregated ATAC-seq profiles of the different cell types. Highlighted areas indicate the cell types in which those segments of the sequence result in activity. The activity of the Glass Multimer Reporter (GMR) is shown, together with the cisTopic cell tSNE colored by the accessibility of the marked regions. **d.** Number of enhancer-to-gene links per gene **e.** Number of links with genes in the ranked position based on distance from the enhancer. **f.** Number of positive and negative links, with representative enriched motifs in each category with Normalized Enrichment Score (NES). **g.** Predicted gene expression (or gene accessibility) based on the signed aggregation of the probabilities of the enhancers linked to each gene weighted by importance. Three genes are shown per plot, using RGB encoding. **h.** Link-based regulons for ato, gl and onecut, built using GRNBoost co-expression modules and motif enrichment on the regions linked to each potential target gene. Left: Cytoscape view of the link-based regulons. Color scale indicates the average importance of the regions enriched in the transcription factor motif for each gene. Middle: Examples of target genes, showing the enhancer-to-region links (top), cisTarget regions (middle) and gene annotation. cisTarget regions in which the motif for the transcription factor is enriched are shown in red. The area highlighted in yellow corresponds to the motif enrichment search space used in SCENIC⁴⁴. Right: GSEA plots comparing the link-based regulons with differentially expressed genes in a compendium of conditions compared to wild type. We score the atonal regulon against both gain and loss of function mutants described in Aerts *et al.*⁵⁰; the gl regulon, against GMR⁺ FAC sorted cells from Potier *et al.*; and the onecut regulon, against a loss-of-function mutant of onecut, also presented by Potier *et al.*⁶⁷. PMM: Peripodial Membrane Medial. PML: Peripodial Membrane Lateral. HV: Head Vertex. Pro: Progenitors. Pre: Precursors. MF: Morphogenetic Furrow. EPR: Early photoreceptors. LPR/CC: Late photoreceptors and cone cells. INT: Interommatidial cells. Hemo: Hemocytes. JOP: Johnston Organ Precursor.

Cell-type specific caQTL analysis reveals key transcription factor binding sites that impact chromatin accessibility

Having established a gene regulatory landscape at single-cell resolution, we next asked whether it can be exploited to interpret the effects of cis-regulatory variation on enhancer function. To this end, we identified chromatin accessibility quantitative trait locus (caQTLs) using a cohort of bulk eye-antennal disc ATAC-seq profiles across inbred lines¹² (Fig 5a). While 21 of these samples were profiled by Jacobs *et al.*¹², we performed 29 additional bulk ATAC-seq experiments, resulting in a panel with 50 samples with highly robust ATAC-seq profiles (correlation between samples: 0.5-1, Fig 5b).

To identify caQTLs (i.e. SNPs or indels that correlate with ATAC-seq signal), we used a generalized linear model (GLM) on all the 456,893 SNPs present in the 38,179 accessible regions, finding 10,969 SNPs (2.4%) that correlated significantly with accessibility changes in the regions where they are

located (adjusted p-val < 0.05). These ~10k caQTLs are found across 4,853 genomic regions (Fig 5c). Compared to the reference allele, 6,781 of these caQTLs promote chromatin closure, while the remaining 4,188 result in chromatin opening (Fig 5c). Next, we evaluated whether these caQTLs either create or break a TF motif, using a collection of more than 24,000 transcription factor motifs^{51,68}. Particularly, for each motif we compared the motif score between the reference enhancer sequence and the enhancer carrying the SNP, obtaining a Delta score for each caQTL and each motif. In agreement with Jacobs *et al.*¹², we found that the motif linked to Grh is significantly more associated to caQTLs than to control SNPs (adjusted p-val: 10^{-29} by Fisher's exact test) and directly explains the accessibility of 158 regions (with $\text{abs}(\text{delta}) > 2$, Fig S17a). However, in Jacobs *et al.* we failed to detect any additional enriched motifs similarly affected by caQTLs.

Here, we exploited our cell-type specific topics to perform the motif enrichment analysis for each topic separately. This effectively changes the null model, and aims to detect motifs that are significantly more altered in caQTLs in cell-type specific regions, compared to SNPs in cell-type specific regions. This strategy of using cell-type specific null models indeed revealed 33 additional motifs (log(Fisher test adjusted p-value) > 8, in at least one topic), which explain 2,061 extra caQTLs genome-wide (with $\text{abs}(\text{delta}) > 2$), increasing the number of explained caQTLs compared to performing bulk analysis where only the Grainyhead motif was found with comparable significance¹².

Indeed, cell-type specific motif enrichment permits to infer in which cell types these caQTLs are relevant. For example, caQTLs found within accessible regions anterior to the morphogenetic furrow and interommatidial cells significantly affect Optix and So binding sites (adjusted p-val: 10^{-2} by Fisher's exact test); while caQTLs in photoreceptor and cone cell regions mainly impact the GGG motif (adjusted p-val: 10^{-4} by Fisher's exact test) (Fig 5d, S17b), among others; suggesting that transcription factors linked to these motifs play an important role in chromatin regulation in these specific cell types. In addition, when evaluating caQTLs genome-wide (instead of binarized topic regions) affecting the binding sites of Grh (158), Optix/so (53), Gl (49) and the GGG motif (29) we observe accessibility in epithelial cell types, anterior to the MF and interommatidial cells, interommatidial cells and photoreceptors and cone cells, and photoreceptors, cone cells and brain neurons; respectively (with $\text{abs}(\text{delta}) > 2$, Fig 5e,f, S17c).

In summary, cell-type specific signatures derived from single-cell ATAC-seq can be exploited to assess cell-type specific effects of caQTLs derived from a panel of bulk ATAC-seq profiles, providing a higher resolution and sensitivity compared to a bulk data analysis.

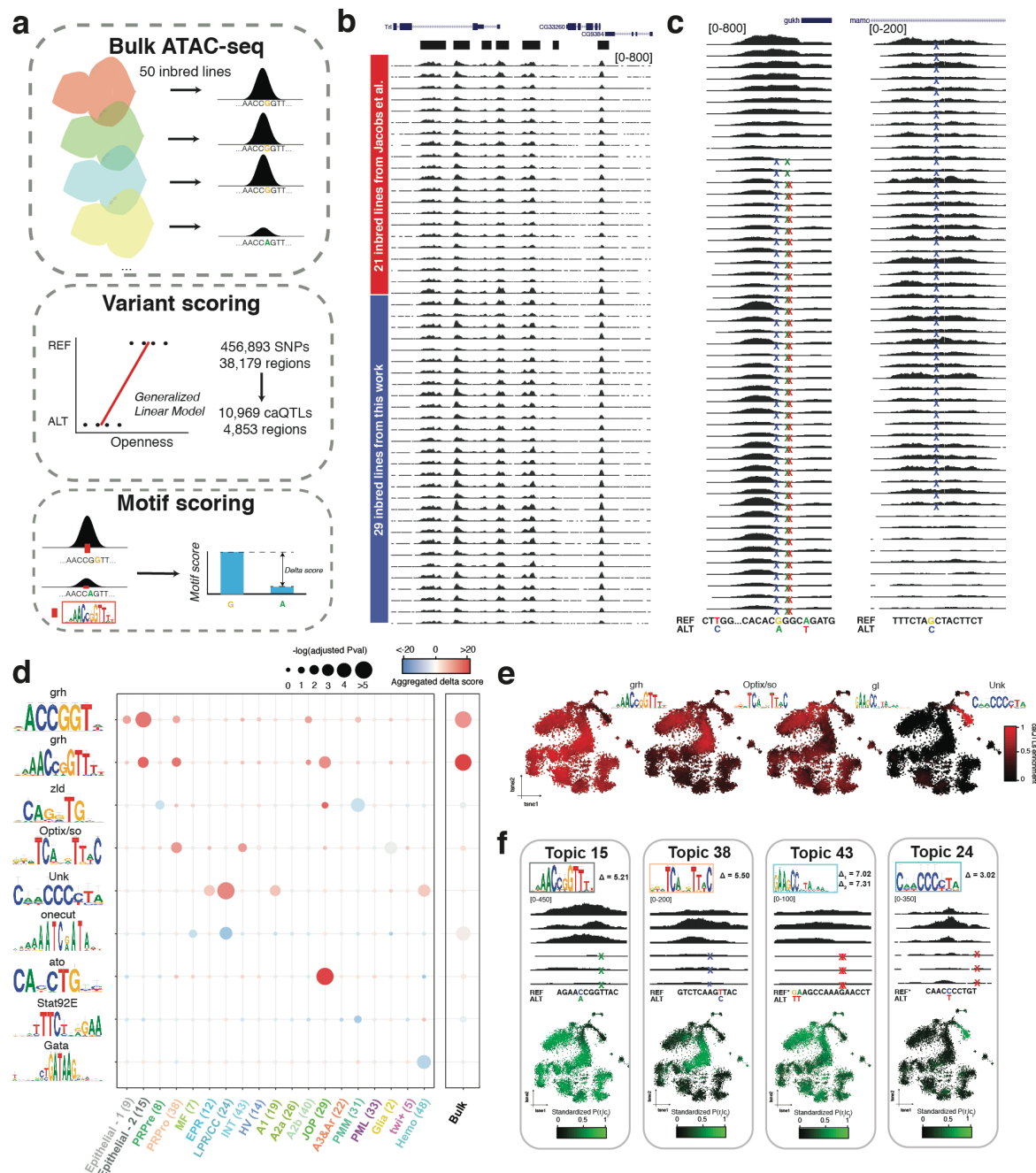


Figure 5: Cell-type specific caQTL analysis reveals key transcription factor binding sites that model the chromatin landscape during the development of the eye disc. **a.** Approach for the identification of genome-wide caQTLs using bulk ATAC-seq profiles of 50 inbred *Drosophila melanogaster* lines. Briefly, after identifying the SNPs among the lines, a generalized linear model (GLM) is used to assess whether the presence of the SNP has an effect in chromatin accessibility. Once these caQTLs are identified, we estimate the effect they have on transcription factor binding sites by comparing the motif score with the reference and alternative SNP (i.e. delta score). A positive delta score indicates that the presence of the motif is related to chromatin opening, while a negative delta score reflects that the motif cause chromatin closeness. **b.** Bulk chromatin profiles of the 50 inbred lines. While 21 of these ATAC-seq experiments were performed by Jacobs et al.¹², we generated 29 additional profiles. Peak calling defined regions are shown in black on the top. **c.** Examples of caQTLs linked to openness

(left) and closeness (right). **d.** Adjusted p-value by Fisher exact test comparing the proportion of caQTLs versus random SNPs affecting each motif and aggregated delta score per topic and bulk regions. **e.** cisTopic cell tSNE colored by the enrichment of regions whose accessibility is affected by caQTLs that alter the highlighted binding sites. **f.** Examples of caQTLs in regions that belong to different topic and affect a certain binding site. Top: Motif with delta score. Middle: Representative bulk ATAC-seq profile on lines with the reference and the alternative allele. Bottom: cisTopic cell tSNE colored by the accessibility of the region affected by the caQTL. The caQTLs coordinates are, from left to right: chr3L:17392596, chr3R:14076593, chr2R:18674001 and chr2R:18674002, and chr3R:29376820. PMM: Peripodial Membrane Medial. PML: Peripodial Membrane Lateral. HV: Head Vertex. Pro: Progenitors. Pre: Precursors. MF: Morphogenetic Furrow. EPR: Early photoreceptors. LPR/CC: Late photoreceptors and cone cells. INT: Interommatidial cells. Hemo: Hemocytes. JOP: Johnston Organ Precursor.

Prospero mediates terminal photoreceptor differentiation by binding the GGG motif

While the GGG motif plays an important role in regions specifically accessible in photoreceptor neurons, the transcription factor/s that bind to it are currently unknown. In fact, this motif is enriched in regions specifically accessible in photoreceptors (Fig 2); the accessibility of regions with this motif is tightly correlated with their activity (Fig 3); these regions are related to gene activation rather than repression (Fig 4); and caQTLs affecting this motif are enriched in photoreceptor-specific enhancers (Fig 5).

To find potential transcription factors that bind to this motif, we first collected candidate TFs that are expressed in photoreceptors, and that have a GGG-like motif, based on the *Drosophila* motif, or the motif of the orthologous factor in other species. We also analyzed the entire modERN collection of ChIP-seq data by motif enrichment, and identified three TFs (Pros, Nerfin-1, and l(3)neo38), that have a very strong GGG motif enriched in their ChIP-seq peaks (*cisTarget* Normalized Enrichment Score (NES) of 10.40, 5.93 and 5.61). In total we selected 14 candidate TFs, namely Pros, Lola (isoforms L and T), Nerfin-1 (FlyORF constructs CC and HA), l(3)neo38, Sp1, Ttk (isoforms Ttk88 and Ttk69), Lz, Lov, Psq and Fru (alleles EY09280 and E0Y2366). Next, we overexpressed each of these 14 TFs in the posterior part of the eye disc using GMR-GAL4 as driver, and for each TF we analyzed phenotypic changes as well as bulk ATAC-seq.

Firstly, we assessed whether over-expression of these TFs in the posterior part of the eye disc resulted in an adult eye phenotype, which was the case for 9 of the 14 TFs (Fig S18). Of these, overexpression of Pros and Lola-T had the most severe phenotype, resulting in lethality in the early pupa stage. Overexpression of l(3)neo38, Nerfin-1 and Sp1 caused a rough eye phenotype, provoked by defects in the development of photoreceptors⁶⁹; the overexpression of Ttk69 gave rise to a small eye and loss of photoreceptors; and overexpression of Lz and Lola-L led to loss of pigment and rough eye phenotype.

To assess the changes caused in the chromatin landscape by the overexpression of these TFs, and to investigate whether GGG-regions are affected, we performed bulk ATAC-seq on the eye-antennal disc for each TF gain-of-function. We clustered all ATAC-seq data across all TFs using cisTopic (on bootstrapped data, see *Methods*), revealing two topics whose regions are highly enriched in the GGG motif, namely topics 4 and 18 (Fig 6a,b). Both topics represent regions that become highly accessible upon overexpression of Pros, with regions in topic 18 also weakly increasing in accessibility upon overexpression of other TFs, including Nerfin-1 and l(3)neo38 (Fig S19a,b).

Importantly, only Pros overexpression results in a strong opening of both early and late-born photoreceptor GGG enriched regions, while overexpression of other TFs has a weak effect (Fig 6c, S19c,d). On the other hand, topic 4, which contains regions uniquely accessible upon Pros overexpression, is more strongly enriched in the late-born photoreceptor regions found in the scATAC-seq data compared to regions in topic 18; which contains regions that slightly increase in accessibility upon overexpression of other TFs, such as Nerfin-1 and l(3)neo38 (Fig S19e,f). These results agree with the phenotype observed in the third instar larvae eye disc: Pros overexpression has a strong impact throughout photoreceptor development, while the effects of Nerfin-1 and l(3)neo38 are milder and largely affect the structure of early ommatidia (Fig S19g).

In the wild-type eye antennal disc, Nerfin-1 and l(3)neo38 are expressed in early and late-born photoreceptors, while Pros expression is limited to late-born photoreceptors (Fig 6d). This suggests that Nerfin-1 and l(3)neo38 could be the early openers of the GGG enriched regions, while Pros would act in late-recruited photoreceptors. In fact, the embryonic ChIP-seq profiles of these transcription factors support their binding to the photoreceptor GGG enriched regions, especially for Pros and Nerfin-1 (Fig S19h). When comparing the GGG regions bound by these factors in the embryo, we find that 50-65% of the sites are shared by the three transcription factors (Fig 6e). Differential motif enrichment analysis between shared versus transcription factor specific binding sites reveals that the shared sites are highly enriched for GGG motifs (adjusted p-value: 10^{-14}), meaning that the three TFs can bind to regions with strong GGG motifs. On the other hand, regions specifically bound by l(3)neo38 are enriched for the canonical l(3)neo38 binding site (adjusted p-value: 10^{-17}); regions uniquely bound by Pros are enriched for a GATC motif, previously reported as being associated with Prospero binding sites and linked to Lola-N⁷⁰ (adjusted p-value: 10^{-10}); and regions uniquely bound by Nerfin-1 are enriched for the Ara/Caup/Mirr motif (adjusted p-value: 10^{-5}). Indeed, both Nerfin-1 and Mirr have been reported to be involved in axon guidance^{71,72}.

In summary, given the high enrichment of the GGG motif within Pros ChIP-seq peaks in the embryo, the strong opening of GGG enriched regions upon Pros overexpression and its expression in late photoreceptors, we propose Prospero as the key regulator of late-born photoreceptors (R7) and cone

cells through the binding of the GGG motif. In addition, our data suggests that in early photoreceptors, in which Pros is not expressed, Nerfin-1 and l(3)neo38 can be weaker binders of strongly GGG enriched regions.

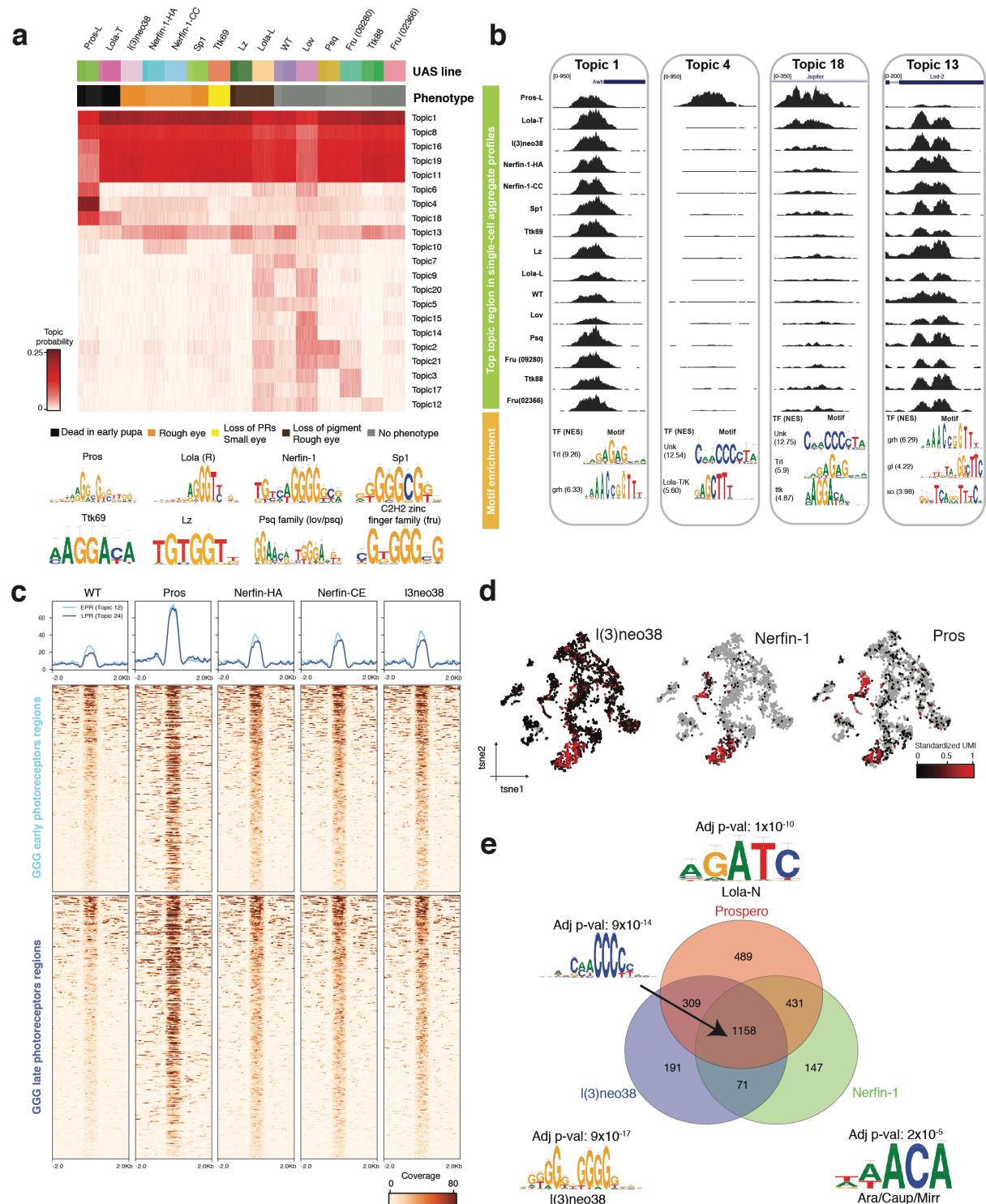


Figure 6: Prospero mediates terminal photoreceptor differentiation by binding the GGG motif. a. cisTopic topic-cell heatmap, based on a model with 21 topics. For running cisTopic, 50 single cell profiles were bootstrapped from the 15 bulk ATAC-seq profiles of the GMR-GAL4 UAS-TF and wild type lines included in the screen. **b.** Highlighted topics showing a representative topic region (top) and representative enriched motifs

with their Normalized Enrichment Score (NES). **c.** Heatmaps showing the normalized coverage of the early photoreceptor GGG enriched regions and late GGG enriched photoreceptor regions on the selected GMR-GAL4 UAS-TF lines. **d.** Seurat cell tSNE colored by the expression of *l(3)neo38*, *Nerfin-1* and *Prospero*. **e.** Venn diagram showing the overlap between the GGG enriched binding sites of *Prospero*, *Nerfin-1* and *l(3)neo38*. Differentially enriched motif in each class are shown with their adjusted p-value.

Discussion

Single cell technologies provide unprecedented insights into the dynamics of gene regulation across all cell types within a tissue. However, these techniques require the dissociation of the tissue, resulting in the loss of spatial information. While new experimental techniques are arising to preserve spatial information while profiling single cells, these mainly target single-cell transcriptomics and methods that profile genome-wide transcription are limited in resolution^{6,73,74}. Alternatively, new computational approaches have been developed, such as *novoSpaRc*⁵; however, *de novo* spatial relationships are only possible on one-dimensional tissues, and otherwise require of a gene expression reference map⁴. In this work, we present a semi-supervised approach to map omics data into a virtual template by extracting axial information via pseudotime ordering. The main limitations of this approach are that (1) it can be currently applied to 1D or 2D tissues, (2) it requires a priori information about at least one landmark between the real and the virtual cells and the direction of the axis and (3) it assumes symmetry around the axes, meaning that other gradients may be lost as cells are spread randomly in each bin. Nevertheless, the spatial gene expression atlas resulting from the mapping of scRNA-seq accurately recapitulates known gene expression patterns, and allows to generate virtual gene expression profiles for any gene, at a resolution comparable with *novoSpaRc*⁵.

Whereas spatial inference has been reported based on scRNA-seq data, in this work we generate the first spatial map of a tissue from scATAC-seq data. This accessibility atlas effectively predicts enhancer-reporter activity for more than 700 enhancers from the Janelia FlyLight Project, with ~85% enhancers showing matching accessibility and activity patterns. The remaining enhancers (~15%) are binding sites of the epithelial pioneer transcription factor *Grainyhead*¹², which primes these regions in all the epithelial cells without resulting in enhancer activity. Indeed, pioneer transcription factors are able to displace nucleosomes, resulting in an ATAC-seq signal; and despite that they are necessary, their binding is not sufficient for activity¹². Thus, enhancer accessibility can be achieved either by the binding of pioneer factors or through the cooperative binding of multiple TFs. These results highlight both the power of using scATAC-seq as a proxy of enhancer activity, as well as the need for caution when dealing with pioneer factors.

The virtual map also acts as a latent space in which scATAC-seq and scRNA-seq data are available for each virtual cell. While experimental approaches for the simultaneous profiling of epigenome and

transcriptome are emerging¹⁴⁻¹⁶, these do not yet achieve the same throughput and sensitivity compared to the independent assays. Computationally, Granja *et al.*⁷⁵ have taken a similar approach, in which cells are mapped into the same latent space and for each single cell transcriptome the aggregate scATAC-seq profile of the closest neighbors is assigned. The resulting integrated profiles allow inferring relationships between enhancers and target genes. While Pliner *et al.*²¹ have tackled this problem uniquely using scATAC-seq data, Granja *et al.*⁷⁵ used Pearson correlation between the chromatin accessibility and gene expression. In this work, we extend this approach by also using Random Forest models to assess non-linear relationships. Of note, these approaches are not robust to pioneer sites, whose accessibility and activity are unpaired. For example, in our approach a validated intronic enhancer of Atonal and Grainyhead in *sca*⁵⁰ is missed, as the enhancer is ubiquitously accessible while only functional in the morphogenetic furrow, where the gene is expressed. Nevertheless, for the remaining 85% of the enhancers in which accessibility and activity are coupled, in this system, we have been able to reconstruct novel and validated enhancer-to-target gene links.

The predicted links between enhancers and target genes support that (1) the probability of an enhancer regulating a gene decreases exponentially with the distance and the number of non-intervening genes in between, as also reported by other authors^{7,76,77}; and (2) genes are regulated by several - and in some cases, redundant - enhancers, with a median of 22 enhancers linked to each gene. Indeed, Cannavó *et al.* reported in the *Drosophila* embryo that ~64% of the mesodermal loci has redundant (or shadow) enhancers, of which ~60% contain more than one pair of shadow enhancers⁷⁸. In agreement, we find that ~80% of the genes are regulated by shadow enhancers (6,937 out of 8,307 genes), out of which ~72% are regulated by at least 3 shadow enhancers (4,900 out of 6,937 genes). Importantly, transcription factors are more tightly regulated, being linked with a higher number of enhancers (with an average of 13 positive links per gene) and having almost twice the number of redundant enhancers compared to non-TFs genes (with an average of 13 enhancer-enhancer pairs, equivalent to ~5 shadow enhancers). As abnormalities in the expression of transcription factor genes can have more severe phenotypes compared to final effector genes, having more - and redundant - enhancers may provide evolutionary robustness.

Of note, almost ~50% of the inferred links are negatively correlated with their target genes. While Polycomb mediated repression has been shown to reduce region accessibility⁷⁹, other studies suggest that, although repressed enhancers are less accessible than active enhancers, they still show accessibility compared to the non-regulatory genome⁸⁰. Such an effect can be observed in the embryonic *eve stripe 2* enhancer, which is active (and more accessible) in the second embryonic stripe, while repressed (and less accessible) in the rest⁸¹. Meanwhile, in the eye antennal disc, where it is not active nor repressed, there is no accessibility (Fig S20). Thus, accessible regions can not only correspond to promoters or insulator regions, or primed or active enhancers, but also to repressed enhancers.

Several works have focused on the inference of GRNs from single-cell data, mostly exploiting scRNA-seq to infer co-expression patterns between TFs and potential target genes⁸². In an attempt to reduce the number of false positive targets due to activating cascade effects, we introduced SCENIC⁴⁴, which additionally evaluates the enrichment of binding sites for the TF around the TSS of the putative target genes. On the other hand, other studies have exploited single-cell ATAC-seq to find target enhancers with binding sites for specific TFs. For example, chromVAR⁸³ aggregates regulatory regions based on motif enrichment, and then evaluates these modules on single-cell ATAC-seq data; while cisTopic²⁰ performs motif enrichment on sets of co-accessible enhancers inferred from scATAC-seq profiles (i.e. topics) to find common master regulators. However, none of these approaches incorporates knowledge about the TF nor target genes expression. Here, we aim for the first time to integrate all these layers - transcription factor binding sites, chromatin and gene expression - to infer *bona fide* GRNs, by deriving co-expression modules between genes and transcription factors (from the scRNA-seq data) and pruning them based on the enrichment of the TF motif in the enhancers that regulate these genes (based on the enhancer-to-target gene links derived from the integration of scATAC and scRNA-seq data). In other words, single-cell Gene Regulatory Networks are not built with TF-Gene relationships, but we rather expand them to introduce the enhancers as nodes (TF-Enhancer-Gene).

As bulk profiles may mask true biological signal (due to the proportions of the different cell types), single-cell data has been used to deconvolute cell-type specific signals from RNA-seq bulk data⁸⁴, permitting to exploit bulk omics population panels only requiring one single-cell analysis. In fact, when evaluating chromatin accessibility QTLs, if a binding site is created, or destroyed, by a mutation in an enhancer that is active in a subset of cells, its effect at the bulk level will be less pronounced and may remain undetected. In this work we exploit for the first-time cell-type specific enhancers, learned from the scATAC-seq data, to account for cell-type specific genomic variation. For example, we revealed the relevance of Atonal binding sites for opening of Johnston's organ precursor specific regions and the GGG motif, previously unlinked to any transcription factor, in photoreceptors; among others. Interestingly, Atonal has been shown as key transcription factor for the specification of sensory neurons³⁷; and bHLH proteins have been proposed to act as pioneer transcription factors on certain contexts⁸⁵, such as the mammalian family member Ascl1⁸⁶.

The importance of the GGG motif in neuronal enhancers was evident in most of our analyses; however, its interpretation was a challenge because the binding TFs were unknown. While yeast one-hybrid (Y1H) experiments have been previously used to reverse-engineer which transcription factors can bind a motif of interest, lowly expressed TFs may be underrepresented in the cDNA library and interactions that occur *in vivo* may be missed (such as those dependent of post-transcriptional modifications)^{70,87}. Here, we have used a novel *in vivo* approach, in which we identify the changes that overexpression of potential TF candidates cause in chromatin accessibility through bulk ATAC-seq. Although this

strategy allows to characterize the effects of TF overexpression directly on the tissue of interest, it also has limitations, such as the limited throughput of *in vivo* genetic screens (one TF per experiment, compared to dozens of TFs that can be tested by Y1H or Perturb-ATAC⁸⁸ *in vitro*). This requires making a stringent selection of potential candidates, that can be further bounded by the existence of compatible tools, such as UAS-TF lines. In addition, the changes in chromatin may not be direct; but these effects can be ruled out using external data available, such as ChIP-seq.

We have also identified the neuronal precursor transcription factor Prospero as the strongest binder of the GGG motif, followed by Nerfin-1 and l(3)neo38. In fact, overexpression of each of them, but especially Prospero, results in the opening of GGG regions; and all three transcription factors, especially Pros and Nerfin-1, can bind to the GGG motif. Based on the expression of these transcription factors, we hypothesize that Nerfin-1 - and l(3)neo38 - are the early binders of the GGG motif, while Pros can bind to these regions in the late-born photoreceptors, where it is expressed. In fact, Pros and Nerfin-1 have been reported to share direct targets during CNS differentiation⁸⁹ and have been found to be key regulators during the photoreceptor and retinal differentiation in other organisms, such as zebrafish, chicken and mammals⁹⁰⁻⁹³.

Finally, we provide a comprehensive and user-friendly single-cell resource of the *Drosophila*'s eye-antennal disc. We envision that our computational strategies and enhancer resource will be of value not only to the *Drosophila* community, but also to the field of single-cell regulatory genomics in general.

Methods

Fly husbandry and genotypes

A detailed description of the lines used in this work is provided in Table S2. A wild type line, hybrid of DGRP-551, DGRP-360, DGRP-907 and DGRP-913, was used on the single-cell RNA-seq and single-cell ATAC-seq experiments with 10X Genomics. For cell sorting (followed by bulk and single-cell ATAC-seq with Fluidigm C1), we used a sens-F2B-GFP transgenic line⁵³. For measuring enhancer activity in a subset of lines from the Janelia Flylight Project, we selected the stocks (with Bloomington number): 49564, 49076, 45619, 39134, 47330, 49534, 47473, 48098, 47166, 49127, 40482, 45172 and 49359; and crossed them with a UAS-eGFP line (Bloomington number: 4776). For the analysis of caQTLs, we performed bulk ATAC-seq on 29 lines from the *Drosophila* Genetics Reference Panel (with Bloomington number): 25189, 25191, 25194, 25198, 25201, 28129, 28136, 28138, 28140, 28141, 28176, 28177, 28185, 28189, 28194, 28198, 28212, 28229, 28233, 28235, 28238, 28239, 28250, 29652, 29658, 55015, 55018, 55028 and 55030. For the genetic screen, we used the following lines from the Bloomington *Drosophila* Stock Center: 32244, 28828, 7361, 33836, 28829, 16994, 17551, 7360 and 15564; and the following from FlyORF: F000093, F000461, F004559, F001783 and F004846. These

lines were crossed with a GMR-GAL4 line. All flies were raised and crossed at 25°C on a yeast based medium.

Dissociation of eye-antennal discs into single cells

Wandering third instar larvae were collected and a total of ~30 eye-antennal discs were dissected and transferred into a tube containing 200 µl of ice-cold PBS. The sample was centrifuged at 800 g for 5 minutes, and after removing the supernatant, 50 µl of dispase (3mg/ml; Sigma-Aldrich_D4818-2mg) and 70 µl of collagenase (100 mg/mL; Invitrogen_17100-017) were added. The tissue was dissociated during 45-60 minutes at 25°C at 550 rpm, pipetting up and down every 15 minutes to disrupt clumps of cells. Cells were washed with 1 mL of ice-cold PBS, and resuspended in 400 µl of PBS 0.04% BSA. The cells were passed through a 10 µm pluriStrainer (ImTec Diagnostics_435001050) and cell viability and concentration were assessed by the LUNA-FL Dual Fluorescence Cell Counter.

Single-cell RNA-seq (10X Genomics)

Single-cell libraries were generated using the GemCode Single-Cell instruments and the Single Cell 3' Library & Gel Bead Kit v2 and CHIP Kit from 10X Genomics, following the protocol provided by the manufacturer. Briefly, the eye-antennal disc cells were suspended in PBS 0.04% BSA. About 8,700 cells were added in each reaction with a targeted cell recovery of 5,000 cells. Following the generation of nanoliter-scale Gel bead-in-EMulsions (GEMs), GEMs were reverse transcribed in a C1000 Touch Thermal Cycler (Bio Rad) programed at 53°C for 45 min, 85°C for 5 min, and hold at 4°C. After reverse transcription, single-cell droplets were broken and the single-strand cDNA was isolated and cleaned with Cleanup Mix containing DynaBeads (Thermo Fisher Scientific). cDNA was then amplified with a C1000 Touch Thermal Cycler programed at 98°C for 3 min, 12 cycles of (98°C for 15 s, 67°C for 20 s, 72°C for 1 min), 72°C for 1 min, and held at 4°C twice. Subsequently, the amplified cDNA was fragmented, end-repaired, A-tailed and index adaptor ligated, with SPRIselect Reagent Kit (Beckman Coulter) with cleanup in between steps. Post-ligation product was amplified with a C1000 Touch Thermal Cycler programed at 98°C for 45 s, 14 cycles of (98°C for 20 s, 54°C for 30 s, 72°C for 20 s), 72°C for 1 min, and hold at 4°C. The sequencing-ready library was cleaned up with SPRIselect beads.

Dissociation of eye-antennal discs into single nuclei

Wandering third instar larvae were collected and a total of ~30 eye-antennal discs were dissected and transferred into a tube containing 200 µL of ice-cold PBS. The sample was centrifuged at 800 g for 5 minutes, and after removing the supernatant, resuspended in 500 µl of nuclei lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3mM MgCl₂, 0.1% Tween-20, 0.1% Nonidet P40, 0.01% Digitonin, 1%

BSA, and water) and transferred to a dounce homogenizer (Sigma-Aldrich-D8938_2mL). After incubating the sample for 5 minutes on ice, 25 strokes were applied with the loose pestle. The sample was incubated for 10 minutes on ice and after applying 25 strokes with the tight pestle, transferred to a 2 mL tube. The homogenizer and the pestle were rinsed with wash buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3mM MgCl₂, 0.1% Tween-20, 1% BSA and water), and the solution was also transferred to the 2mL tube. The sample was washed once with wash buffer and resuspended on 50 µl of 1X diluted nuclei buffer (10X Genomics). The nuclei were passed through a 10 µm pluriStrainer (ImTec Diagnostics_435001050) and cell viability and nuclei concentration were assessed by the LUNA-FL Dual Fluorescence Cell Counter.

Single-cell ATAC-seq (10X Genomics)

Single-cell libraries were generated using the GemCode Single-Cell instruments and the Single Cell ATAC Library & Gel Bead Kit and CHIP Kit from 10X Genomics, following the protocol provided by the manufacturer. Briefly, the eye-antennal disc nuclei were suspended in 1X diluted nuclei buffer (10X Genomics). About 8,700 nuclei were added in each reaction with a targeted nuclei recovery of 5,000 nuclei. The samples were incubated at 37°C for 1 hour with 10 µl of transposition mix (per reaction, 7 µl ATAC Buffer and 3 µl ATAC Enzyme (10X Genomics)). Following the generation of nanoliter-scale Gel bead-in-EMulsions (GEMs), GEMs were reverse transcribed in a C1000 Touch Thermal Cycler (Bio Rad) programed at 72°C for 5 min, 98°C for 30 sec, 12 cycles of 98°C for 10 sec, 59°C for 30 sec and 72°C for 1 min, and held at 15°C. After reverse transcription, single-cell droplets were broken and the single-strand cDNA was isolated and cleaned with Cleanup Mix containing DynaBeads (Thermo Fisher Scientific). cDNA was then amplified with a C1000 Touch Thermal Cycler programed at 98°C for 3 min, 12 cycles of (98°C for 15 s, 67°C for 20 s, 72°C for 1 min), 72°C for 1 min, and held at 4°C twice. Subsequently, the amplified cDNA was fragmented, end-repaired, A-tailed and index adaptor ligated, with SPRIselect Reagent Kit (Beckman Coulter) with cleanup in between steps. Post-ligation product was amplified with a C1000 Touch Thermal Cycler programed at 98°C for 45 s, 14 cycles of (98°C for 20 s, 54°C for 30 s, 72°C for 20 s), 72°C for 1 min, and hold at 4°C. The sequencing-ready library was cleaned up with SPRIselect beads.

Cell sorting

Wandering third instar larvae were collected and a total of 200 eye-antennal discs were dissected in ice-cold PBS and placed in SF900 medium. For dissociation, the tissue was placed in 400 µl of trypsin in 0.05% EDTA. The eye-antennal discs were then incubated at 37 °C for 1 h with agitation; being mixed every 20 min with a pipette. After dissociation, cells were centrifuged at 800 g for 5 min at 4 °C and washed with PBS. Finally, the cells were resuspended in 400 µl of PBS, filtered using a 40 µm cell

strainer and stained with propidium iodide (PI; final concentration 1 µg/ml) to exclude dead cells. The cells were sorted on a BD Aria I, selecting against the presence of PI and for the presence of GFP.

As many cells as possible were sorted into a microcentrifuge tube, pelleted by centrifugation at 800 g for 5 min at 4 °C and resuspended at a concentration of 1,000 cells/µl. Single-cell ATAC-seq was performed as previously described^{20,94}, using 5- to 10-µm Open App integrated fluidic circuits (IFCs) on the Fluidigm C1 and with no cell washing step. Briefly, cells were loaded (using a 40:60 ratio of RGT:cells) on a primed Open App IFC (5-10 µm, the protocol for ATAC-seq from the C1 Script Hub was used). After cell loading, the plate was visually checked under a microscope and the number of cells in each of the capture chambers was noted. Next, the sample preparation was performed on the Fluidigm C1 during which the cells underwent lysis and ATAC-seq fragments were prepared. In a 96-well plate, the harvested libraries were amplified in a 25 µl PCR reaction. The PCR products were pooled and purified on a single MinElute PCR purification column for a final library volume of 15 µl. Quality checks were performed using the Bioanalyzer high sensitivity chips. Fragments under 150 bp were removed by bead-cleanup using AMPure XP beads (1.2x bead ratio) (Beckman Coulter).

ATAC-seq

For the DGRP panel lines we used the ATAC-seq protocol for eye-antennal discs as previously described^{94,95}. Briefly, ~10 eye-antennal discs were dissected and lysed in 50 µl ice-cold ATAC lysis buffer (10 mM Tris-HCl, pH 7.4, 10mM NaCl, 3mM MgCl₂, 0.1% IGEPAL CA-630). Lysed discs were then centrifuged at 800 g for 10 minutes at 4°C and the supernatant was discarded. The rest of the ATAC-seq protocol was performed as described previously^{94,95}, using the following primers: Fwd:- ‘AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG’ and Rev:- ‘CAAGCAGAAGACGGCATACGAGATXXXXXXGTCTCGTGGGCTCGGAGATGT’ (where X indicates barcode nucleotides). The final library was purified using a Qiagen MinElute kit (Qiagen) and Ampure XP beads (Ampure) (1:1.2 ratio) were used to remove remaining adapters. The final library was first checked on an Agilent Bioanalyzer 2000 for the average fragment size. Resulting successful libraries were sequenced with 75bp, single end reads on the Illumina NextSeq 500 platform. Single end sequencing was chosen for this part of the study because we were not interested in the fragment contents (i.e., how many nucleosomes are placed between two insertion sites), rather just the profile of insertion sites, and also made the comparison with the previously existing data (i.e. the bulk ATAC-seq DGRP panel and Optix-GFP from Jacobs *et al.*¹²) easier.

For the genetic screen samples we used the Omni-ATAC-seq protocol, as previously described²⁰. Briefly, ~10 eye-antennal discs were dissected and lysed using 50 µl of cold ATAC-Resuspension Buffer (RSB) (see Corces *et al.*⁹⁶ for composition) containing 0.1% NP40, 0.1% Tween-20 and 0.01%

digitonin, by pipetting up and down three times and incubating the cells for 3 min on ice. The lysis was washed out by adding 1 mL of cold ATAC-RSB containing 0.1% Tween-20 and inverting the tube three times. Nuclei were pelleted at 500 RCF for 10 min at 4°C, the supernatant was carefully removed and nuclei were resuspended in 50 µl of transposition mixture (25 µl 2x TD buffer (see Corces *et al.*⁹⁶ for composition), 2.5 µl transposase (100 nM), 16.5 µl DPBS, 0.5 µl 1% digitonin, 0.5 µl 10% Tween-20, 5 µl H₂O) by pipetting six times up and down, followed by 30 minutes incubation at 37°C at 1,000 RPM mixing rate. After MinElute clean-up and elution in 21 µl elution buffer, the transposed fragments were pre-amplified with Nextera primers by mixing 20 µl of transposed sample, 2.5 µl of both forward and reverse primers (25 µM) and 25 µl of 2x NEBNext Master Mix (program: 72°C for 5 min, 98°C for 30 sec and 5 cycles of [98°C for 10 sec, 63 °C for 30 sec, 72°C for 1 min] and hold at 4°C). To determine the required number of additional PCR cycles, a qPCR was performed (see Buenrostro *et al.*⁹⁷ for the determination of the number of cycles to be added). The final amplification was done with the additional number of cycles, samples were cleaned-up by MinElute and libraries were prepped using the KAPA Library Quantification Kit as previously described⁸⁶. Samples were sequenced on an Illumina NextSeq 500 High Output chip, with 50bp single-end reads.

Immunohistochemistry

Imaginal eye-antennal discs from third-instar larvae were dissected and fixed in 4% formaldehyde at room temperature for 30 min. Next, they were washed in 1X PBT (PBS + 0.3% Triton X-100) during 15 min for 3 times and blocked in 3% BSA for 1 hour at room temperature. To test enhancers, tissues were incubated with a primary antibody mixture (rabbit anti-GFP (Invitrogen) 1:1000; rat anti-Elav (DSHB, 7E8A10) 1:50; and mouse anti-pros (DSHB) 1:200) at 4 °C overnight. The samples were then washed 3 times with 1X PBT for 15 min at room temperature, followed by 2 hours incubation with secondary antibody mixture (Goat Anti-Rabbit - Alexa Fluor® 488 antibodies; donkey anti-rat Alexa Fluor® 647; and donkey anti-mouse Alexa Fluor® 555) (Invitrogen/Life Technologies) at room temperature in the dark. The samples were washed again 3 times as mentioned above before mounting the eye-antennal discs on slide with Vectashield (Vector Laboratories). For imaging, an Olympus FV1200 confocal microscope was used (20X dry). Fiji⁹⁸ (ImageJ v2.0.0-rc-69/1.52p) was used to merge and process the images.

Analysis of single-cell RNA-seq data

The 10X eye-antennal disc samples were processed (alignment, barcode assignment and UMI counting) with the Cell Ranger (version 2.0.2) count pipeline, using the *cellranger aggr* command with *--normalize=mapped*, and building the reference index upon the 3rd 2017 FlyBase release (*D. melanogaster* r6.16)⁹⁹. Lowly expressed genes detected in less than 11 cells (0.3% of the cells) and with less 32 UMI counts across the data set (3 counts in 0.3% of the cells) were filtered, resulting in a

data set with 8,744 genes and 3,531 cells that was analyzed using Seurat (v2.3.4). Briefly, data was log-normalized with a scale factor of 10^4 and latent variables, defined as the number of UMIs, were regressed out. For further downstream analysis, the most variable genes (1,495) were selected using *FindVariableGenes()* with default parameters. Next, we used PCA to reduce the dimensionality of the original matrix, selecting the first 102 PCs based on a cross-validation step. These 102 PCs were used as input for the Shared Nearest-Neighbor (SNN) graph method implemented in Seurat, with a resolution of 1.2, resulting in 17 cell clusters. Differentially expressed genes for each cluster were estimated with the function *FindAllMarkers()*, using a Wilcoxon Rank Sum test with a logFC threshold of 0.25. tSNE and UMAP were performed with default parameters, using the first 102 PCs. In addition, DoubletFinder²⁹ (v2.0.1) was run using the first 102 PCs, with an estimated pK value of 0.04. Assuming a doublet formation rate of 7.5%, 246 high confidence doublets were found. For the semi-supervised clustering of photoreceptor subclasses, singlet cells in the early photoreceptors and late photoreceptors and cone cells were selected and Seurat (v2.3.4) was run as previously explained using marker genes for each photoreceptor subclass and cone cells as listed in Flybase, comprising a total of 86 unique genes, using the first 7 PCs based on a cross-validation step. PySCENIC^{44,100} (v0.9.1) was run with default parameters, using motif and ENCODE ChIP-seq based databases (as in i-cisTarget⁶⁸), resulting in 175 regulons (159 motif-based regulons). Regulon Specificity Scores (RSS) were calculated as described by Suo *et al.*¹⁰¹, and the Atonal regulon was used as input for GSEA analysis using as rankings the genes ordered by log fold change values calculated by GEO2R for eye-antennal disc RNA-seq profiles of a Gain-Of-Function Atonal mutant and a Loss-Of-Function Atonal mutant (versus WT)⁵⁰. A representative gene regulatory network with regulons enriched in the morphogenetic furrow was built using Cytoscape.

Seurat (v3.0.1) was also used for transferring cluster labels between the eye disc data set from Ariss *et al.*⁴³ and this data (and vice versa). Brain cells from our data set were not included in the analysis, resulting in a data set with 3,232 cells and 8,744 genes, and the eye disc data set from Ariss *et al.* was filtered to keep cells with more than 1,000 UMI counts and 500 genes expressed, resulting in a data set with 5,630 cells and 7,801 genes. Label transferring was performed with default parameters and PCA as dimensionality reduction method, using *vst* as selection method and 2,000 features for finding the variable features and the first 30 PCs for finding anchors and transfer the data. Antennal cell types were not transferred between our data set and Ariss *et al.* eye disc data set. Loom files with the results of these analyses were created using SCoPeLoomR⁹⁵ (v0.4.0) and are available at http://scope.aertslab.org/#/Bravo_et_al_EyeAntennalDisc and processed data can be visualized at http://genome.ucsc.edu/s/cbravo/Bravo_et_al_EyeAntennalDisc.

Analysis of FAC-sorted ATAC-seq data

ATAC-seq reads were first cleaned for adapters using *fastq-mcf* (ea-utils v1.12) and a list of sequencing primers. Cleaned reads (FastQC v0.1) were then mapped to the 3rd 2017 FlyBase release (*D. melanogaster* r6.16) genome using Bowtie2 (v2.2.5) with default parameters, and sorted bam files were produced using SAMtools (v1.2). Single-cell profiles were aggregated using *samtools merge*. Normalized bigwigs were generated using the Kent software (UCSC).

The single-cell data was deduplicated using *picard MarkDuplicates*. Aggregation plots were produced using in-house scripts available at: <https://github.com/aertslab/ATAC-seq-analysis>, and cells were filtered manually based on the aggregation plot profiles, resulting in 74 and 72 Optix-GFP⁺ and sens-GFP⁺ single cell ATAC-seq profiles (out of 96 and 384 sequenced cells, respectively). Downstream analysis was done using cisTopic (v0.2.2)²⁰.

On the bulk samples, peaks were called on mapped reads using MACS2 (v2.1.2.1) with the following additional options: *-nomodel -call-summits --nolambda*. Peaks in the independent samples were merged, and fragments per peak (and ctx region) in each sample were counted using *featureCounts* (Subread v2.0.0). Deseq2 (v1.18.1) was used to obtain differentially accessible peaks between positive and negative cells (with logFC > |1| and p-value < 0.05).

Analysis of ChIP-seq data

ChIP-seq reads were first cleaned for adapters using *fastq-mcf* (ea-utils v1.12) and a list of sequencing primers. Cleaned reads (FastQC v0.1) were then mapped to the 3rd 2017 FlyBase release (*D. melanogaster* r6.16) genome using Bowtie2 (v2.2.5) with default parameters, and sorted bam files were produced using SAMtools (v1.2). Single-cell profiles were aggregated using *samtools merge*. Normalized bigwigs were generated using the Kent software (UCSC). Peaks were called on mapped reads using MACS2 (v2.1.2.1) with the following options: *-g dm -nomodel -bdg -t Samples -c Control*.

Analysis of single-cell ATAC-seq data

The 10X eye-antennal disc samples were processed (alignment and barcode assignment) with a customized version of the Cell Ranger ATAC (version 1.0.0) pipeline, in which the parameter PEAK_MERGE_DISTANCE was set to 50 (instead of 500) and the parameter PEAK_ODDS_RATIO was set to 4 (instead of 1/5). In addition, the reference index was built upon the 3rd 2017 FlyBase release (*D. melanogaster* r6.16)⁹⁹. Sex was assigned to each cell based on the percentage of reads mapped to the X chromosome, as shown by Cusanovich *et al.*¹³.

Downstream analysis was performed with cisTopic²⁰ (v0.2.2). Briefly, fragments within defined regulatory regions (such as ctx regions) were counted, resulting in a matrix with 129,553 regulatory regions and 15,387 cells, after filtering out a total of 379 cells based on the number of fragments within bulk peaks and the total number of fragments. Topic modelling was performed using 2, 10, 20, 30 to 50 (1 by 1), 60, 70, 80, 90 and 100 topics, and 500 iterations, out of which 250 were used as burn-in. Based on the highest log-likelihood, the model with 49 topics was selected. The cell-topic tSNE representation was obtained by using tSNE on the normalized topic-cell matrix (by Z-Score), without using the PCA reduction and with a perplexity of 100. Cell clustering was performed on the normalized cell-topic matrix (by Z-Score) using the Shared Nearest-Neighbor (SNN) graph method implemented in Seurat (v2.3.4), with a resolution of 1.2, resulting in 22 cell clusters. For identifying topics potentially related to batch effects (mainly experimental run and sex), we binarized the cell-topic distributions and used a proportion test comparing the proportion of cells corresponding to each experimental run or sex versus their proportion in the entire population. Two batch effect topics significantly related to the experimental run: topic 46 with run 1 (Bonferroni adjusted p-value for topic 46: 10^{-29}) and topic 18 with run 2 (Bonferroni adjusted p-value for topic 18: 10^{-217}); and a topic was found to be related to the female sex (Bonferroni adjusted p-value for topic 4: 10^{-21}).

On the other hand, region-topic distributions were binarized with a probability threshold of 0.985. The region-topic tSNE was performed with similar parameters as before, using a perplexity of 200. The annotation of regions was done with default parameters. RcisTarget⁴⁴ (v1.5.0) and i-cisTarget^{51,68} were run to assess motif enrichment on the binarized topics, using a ROC threshold of 0.01, a maximum rank of 5,000 and the version 8 motif database, containing more than 20,000 motifs. The probability of each region in each cell (region-cell) was calculated using the *predictiveDistribution()* function, in which the topic-cell and the region-topic matrices are multiplied. For the enrichment of epigenomic signatures, region sets were mapped to the regions in the data set with a minimum overlap of 40% and the enrichment of the signatures in the cells was estimated using a maximum AUC rank of 12,956 (10% of the total number of regions) and cell-region rankings based on the region-cell probability matrix, while the enrichment of signatures in topics was estimated using a maximum AUC rank of 3,887 (3% of the total number of regions) and the region-topic distributions as rankings. Additionally, we projected the FAC-sorted single cell profiles (Optix-GFP⁺ and sens-GFP⁺) with at least 70% of the fragments within regulatory regions into the existing topic space. Briefly, the topic-cell distributions of the new cells were estimated by multiplying the binary count matrix (cell-regions) by the region-topic distributions of the existing models. The estimated topic-cell contributions were merged with the topic-cell distributions of the original cells, normalized (by Z-Score) and batch effects were corrected with Harmony (v1.0)¹⁰².

Gene activity scores were estimated by aggregating the region probabilities of the regions surrounding the TSS of each gene (5kb upstream and introns), as used for cisTarget enhancer-to-gene associations⁶⁸, and probabilities were multiplied by 10^6 and rounded before creating the loom file. These gene-activity based matrix was used to assess the enrichment of the regulons derived from the analysis with pySCENIC (v0.9.1) in the single-cell RNA-seq data, using AUCell⁴⁴ (v1.5.2) with default parameters, with a maximum AUC rank of 439 (5% of the total number of genes). In addition, we also used DoubletFinder (v2.0.1) on this matrix, using the first 102 PCs, with an estimated pK of 0.27. Assuming a doublet formation rate of 20%, we find 13,848 high confidence singlets. Finally, we performed label transferring between the scRNA-seq and the scATAC-seq (gene activity based) data sets (and vice versa) with Seurat (v3.0.1). Label transferring was performed with default parameters and CCA as dimensionality reduction method, using *vst* as selection method and 3,000 features for finding the variable features and the first 20 dimensions for finding anchors and transfer the data. Loom files with the results of these analyses were created using SCoPeLoomR⁹⁵ (v0.4.0) and are available at http://scope.aertslab.org/#/Bravo_et_al_EyeAntennalDisc and processed data can be visualized at http://genome.ucsc.edu/s/cbravo/Bravo_et_al_EyeAntennalDisc.

Projection of single-cell omics data into a virtual latent space

The eye-antennal disc representation (Fig 1a, 2a) was used to generate the virtual eye template coordinates. Importantly, for representing non-spatially restricted groups (i.e. *twi*⁺ cells, hemocytes, glia, peripodial membrane groups) and clarify cell types posterior to the morphogenetic furrow (i.e. early photoreceptors, late photoreceptors and cone cells, and interommatidial cells; for both scRNA-seq and scATAC-seq) or in the antenna and anterior to the morphogenetic furrow (i.e. antennal rings A2a and A2b and precursors and progenitors, respectively, on the scATAC-seq analysis), circles were added to the representations. The template was reduced to a size of 100x100 pixels and was split into one image per cell type (in red color). Each image was read using the jpeg (v0.1-8) R package, and the background (in white color) was removed using k-means clustering on the RGB pixel values. Since interommatidial cells and photoreceptors are mixed posterior to the morphogenetic furrow, we intercalated photoreceptors and interommatidial cells in the early and late compartments posterior to the morphogenetic furrow. The resulting template coordinates were annotated per cell type, resulting in 5,058 cells on the eye-antennal disc representation, and a total of 5,379 and 5,526 cells for the scRNA-seq and scATAC-seq maps considering the non-spatially mapped cell types and detailed groups.

For mapping the scRNA-seq and the scATAC-seq data, antennal and eye disc cell types were ordered by pseudotime in each data set using the *DPT()* function from the destiny¹⁰³ (v3.0.0) R package, using Seurat PCs and topic contributions of the singlet cells, respectively, as input for estimating the diffusion components. The pseudotime order represents the distal-proximal axis for the antennal cells, and the

anterior-posterior axis in the eye cell types. Each cell type was divided in 10 bins based on their pseudotime order. Similarly in the virtual eye-antennal disc template, for each spatially located cell type in eye we calculated the distance to a reference vertical line located in the morphogenetic furrow (i.e. distance is calculated on the X axis between the landmark point on the same Y coordinate); and for each spatially located group in the antenna we calculated the distance of each virtual cell to a reference point in the center of the arista (i.e. the length of the X-Y vector from the cell and the reference). Each cell type was then divided in 10 bins based on their distance to the reference landmark. For each cell type, we assigned a real profile from the matching bin to each virtual cell randomly (e.g. the cells in the first bin of a pseudotime ordered cell type are assigned to the virtual cells in the first bin of that cell type based on the distance to the landmark in the virtual eye). Progenitors and precursors and antennal rings A2a and A2b in the scATAC-seq mapping were assigned together to the anterior to the morphogenetic furrow and antennal ring A2 compartments based on pseudotime. For non-spatially located cell types and detailed groups cells were sampled randomly without binning. If there are more real cells than virtual ones, random sampling is done without repetition, if there are more virtual cells than real ones, real profiles are assigned more than once. The scRNA-seq (i.e. gene expression) and scATAC-seq (i.e. region-cell probabilities) of the virtual cells are those of their matching real cell. Loom files with the results of these analyses were created using SCoPeLoomR⁹⁵ (v0.4.0) and are available at http://scope.aertslab.org/#/Bravo_et_al_EyeAntennalDisc.

Comparison of accessibility and activity profiles in the virtual latent space

The enhancer activity GFP signal was mapped into the virtual eye representation using a customized script which leverages Matlab's Image Toolbox for landmark-based image registration. Briefly, signals in the antenna and the eye were mapped independently, using projective and polynomial transformations, respectively; and manually selecting 6-8 landmarks per image. The GFP channel from the transformed images was read into R using the jpeg package (v0.1-8) and overlapped with the virtual eye template coordinates; if GFP signal was detected on a cell a value of 1 was given, if not, a value of 0 was assigned. After removing images with low or unclear signals, with signal out of the disc proper (e.g. remaining of the peripodial membrane or glial cells), with unsuccessful mapping, and duplicates, we obtained a matrix recapitulating the activity 390 enhancers, with each enhancer being active in a median of 106 virtual cells. Since *Janelia* enhancers are quite broad (i.e. 1-5kb) and may include more than one cisTarget region, the accessibility probability of each *Janelia* enhancer was calculated by aggregating the region-cell probabilities of the regions falling within it. For comparing accessibility and activity in these regions, we calculated the Spearman correlation between the accessibility probabilities and the activity patterns, and the accessibility gini scores using the *gini.index()* function from the R package lawstat (v3.2). Motif enrichment was performed in the generally accessible regions (with gini index < 0.2) and specific regions (with gini index > 0.4) using i-cisTarget⁶⁸. For scoring the Atonal

enhancers validated by Aerts et al., PWMs were scored in the enhancer sequences using Cluster-Buster¹⁰⁴, and visualized with TOUCAN¹⁰⁵.

Linkage of enhancers to target genes

For each gene, we identified as potential regulatory regions those included in a genomic space of +/- 50kb around the TSS of the gene, including introns, resulting in a median of 54 potential regulatory regions per gene. For genes with more than one TSS, we selected one TSS position randomly. We then combined two approaches to establish relationships between enhancer and target genes, (1) a linear strategy by calculating the Pearson correlation between the enhancer probabilities and gene expression in the virtual cells and (2) a non-linear strategy based on Random Forest models, using the enhancer probabilities as predictors, the gene expression as response and the GENIE3 R package⁴⁴ (v1.8.0) to build each - gene specific - model, with 1,000 trees and default parameters. Importantly, we used all virtual cells profiles except for those representing detailed subgroups, covering 5,253 virtual cells. Correlation-based relationships were filtered to keep those below -0.181 and above 0.194, corresponding to the 1st and 99th percentiles of the normal distribution fitted to all the correlations derived with the fitdistrplus R package (v1.0-11). Random forest derived relationships, based on the importance given to each region in each model, were filtered to keep the top relationships for each gene by binarizing the region importances per gene using BASC binarization as implemented in the Binarize R package (v1.3). We classified links as positive if they were positively correlated with their target genes (>0) and negative if they were negatively correlated with their target genes (<0). This resulted in a total of 183,336 enhancer-to-gene relationships, with a median of 22 links per gene. The Gviz R package (v1.22.3) was used to make figures representing links, and link tracks (with scores and sign of the relationship) are available at http://genome.ucsc.edu/s/cbravo/Bravo_et_al_EyeAntennalDisc.

For estimating GO terms related to the genes with the most and the least links we used GOrilla¹⁰⁶ and visualized the results with REVIGO¹⁰⁷. The list of transcription factors for comparing features in TF and non-TF genes was obtained from the RcisTarget Drosophila database⁴⁴. For estimating the number of redundant enhancers, we considered that two enhancers linked to the same gene were correlated if the Pearson correlation between their region-cell probabilities was above 0.8. For estimating the correlation in activity between enhancer-enhancer pairs, we evaluated the 63 combinations for which the activity of both enhancers was mapped into the virtual eye (with correlation > 0.1). Gene activity scores were calculated by aggregating the region-cell probabilities of the regions linked to each gene, weighted by their signed (positive or negative effect) random forest importance. For integrating these links in the pySCENIC pipeline, we use the modules derived from GRNBoost¹⁰⁸ (Arboreto v0.1.5) and performed the motif enrichment step (with RcisTarget⁴⁴ (v1.5.0), using a ROC threshold of 0.01 and a maximum AUC rank of 5,000) on the regions linked to each gene in the module (using the region-based

cisTarget databases) instead of around the TSS of the gene (using the gene-based cisTarget databases implemented in the original workflow). Genes linked to regions in which motifs linked to the transcription factor in each module were enriched ($NES > 3$), were kept as part of the regulon. The regulons were evaluated on the cells using AUCCell⁴⁴ (v1.5.2). For validating the link-based regulons, we used as input for GSEA the regulons and the genes ordered by decreasing logFC for (1) Atonal Gain-of-Function and Loss-of-Function mutants (versus WT, using GEO2R), (2) GMR⁺ versus GMR⁻ populations and (3) onecutx562 (Loss-of-Function mutant) versus WT, as provided by Potier *et al.*⁶⁷. Loom files with the results of these analyses were created using SCoPeLoomR⁹⁵ (v0.4.0) and are available at http://scope.aertslab.org/#/Bravo_et_al_EyeAntennalDisc.

caQTL analysis

Data preprocessing

Adapter sequences were trimmed from the raw reads using *fastq-mcf* (ea-utils v1.1.2, with default parameters and using a list containing the common Illumina adapters) and the quality of the cleaned reads was checked with FastQC (v0.1). All experiments were mapped using Bowtie2 (v2.2.5) to their personalized version on 3rd 2017 FlyBase release (*D. melanogaster* r6.16) genome.

Briefly, called variants in this genome assembly were retrieved from ftp://ftp.hgsc.bcm.edu/DGRP/freeze2_Feb_2013/liftover_data_for_D.mel6.0_from_William_Gilks_Oct_2015/ and for each of the 50 DGRP lines we adapted the consensus genome (r6.16) using *seqtk mutfa* (*seqtk* (v1.0)), each time including their SNPs (previously called from whole genome sequencing). After the first mapping round, additional SNPs were called on the ATAC reads using SAMtools (v1.2), with the command *samtools mpileup -B -f r6.16.fasta DGRP_lineX.bam | varscan.sh mpileup2snp --output-vcf 1*. Newly called homozygous SNPs (several thousands per line) were added to the existing vcf files using VCFtools (v0.1.14). The genomes were again updated to obtain a final personalized genome for every DGRP line, strongly reducing mapping errors and increasing the sensitivity of subsequent analyses. Cleaned reads were mapped onto the final genomes using Bowtie2 (v2.2.5) again and SAMtools (v1.2) was used for sorting and indexing.

Peaks were called on the mapped reads using MACS2 (v2.1.2.1), with the command *macs2 callpeak -g dm --nomodel--keep-dup all --call-summits*. The narrow peak files (bed format) for all the DGRP lines were merged into a single file that contained a total of 39,879 regions accessible in at least one DGRP line. After filtering out chrU, chrUextra, chrHet and chrM regions and removing regions enriched in repeats (>25% of the sequence) using bedtools (v2.28.0) with the command *intersectBed -v -f 0.25*, we obtained 38,179 accessible regions across this DGRP panel. For every ATAC-seq sample we counted

the number of reads falling into each accessible region using *featureCounts* (Subread v2.0.0).
Normalized bigwig files were generated using the Kent software from UCSC.

Determination of caQTLs

Next, 209 regions with a low coverage for every DGRP line were removed (coverage of the region below 0.2pb for every DGRP lines), ending up with 37,990 accessible regions. For each region, we extracted the normalized ATAC-seq reads for these 50 DGRP lines and linked each region to the annotated and additionally called SNPs for these lines. 676,916 SNPs were assigned to their encompassing region using bedtools.2.26.0 *intersectBed* on the extended vcf file.

In this way, we obtained for each region the normalized reads for each of the 50 lines as one vector and all SNPs called inside this region as a binary matrix for the 50 lines (present =1, absent=0, unknown=NA). We searched for correlating region-SNP vectors using the generalized linear model function in R (version 3.5.0). The p-values were adjusted using the Benjamini-Hochberg procedure in R. We identified 10,969 highly correlating SNP-region pairs referred to as caQTLs (Chromatin Accessibility Quantitative Trait Loci ; adjusted p-value < 0.05).

Delta motif scores

To single out motifs that correlate significantly with the open chromatin changes, a Delta motif score was calculated for every of the 24,454 unique motifs in our collection. The sequence for each of the 4,853 variable regions, that contained at least one caQTL, was extracted using *bedtools getfasta* (Bedtools v2.28.0). Next, we mutated these sequences with their encompassing caQTLs according to their effect on the open chromatin using *seqtk mutfa* (seqtk (v1.0)). For each of the 4,853 regions we obtained 2 sequences, one for the accessible chromatin and one for the less accessible/closed chromatin. We scored every time both sequences with the 24,454 motifs using Cluster-Buster¹⁰⁴, with the options – *m 0 -c 0*, and retained for every motif the highest CRM score for each sequence. By subtracting the CRM score of the less accessible/closed region from the encompassing accessible region we obtained a delta motif score for that region.

Motif significance

For the general significance, we summed all delta scores from the 4,853 regions to obtain a cumulative delta score for each motif. We calculated a Delta motif score, following the same procedure, on 20K random SNPs that were present in an accessible region but had no effect on chromatin accessibility (GLM FDR > 0.95). We then calculated for each motif whether it was significantly more affected ($|\Delta \text{score}| > 3$) by caQTLs compared to the non-correlating SNPs, using the Fishers exact test.

Out of the 10K caQTLs, 6,682 caQTLs fall within a topic (60.9%), having in average 362 caQTLs per binarized topic. For the cell-type specific analysis, we summed all delta scores from all the regions containing at least one caQTL per topic to obtain a cumulative delta score for each motif (in average, there are 362 caQTLs per topic). We calculated a Delta motif score, following the same procedure, on the 40K random SNPs that were present in an accessible region in a topic but had no effect on chromatin accessibility (GLM FDR > 0.95). We then calculated for each motif whether it was significantly more affected ($|\text{Delta score}| > 3$) by caQTLs compared to the non-correlating SNPs, using the Fishers exact test.

Genetic screen data analysis

ATAC-seq reads were first cleaned for adapters using *fastq-mcf* (ea-utils v1.12) and a list of sequencing primers. Cleaned reads (FastQC v0.1) were then mapped to the 3rd 2017 FlyBase release (*D. melanogaster* r6.16) genome using Bowtie2 (v2.2.5) with default parameters, with the single end option (to compare with the WT sample, which was single-end sequenced). Sorted bam files were produced using SAMtools (v1.2). Normalized bigwigs were generated using the Kent software (UCSC). Peaks were called on mapped reads using MACS2 (v2.1.2.1) with the following options: *-g dm -nomodel -bdg -t Sample/Control -c Sample/Control* (depending on whether we want to determine upregulated or downregulated peaks). ChIP-seq bam files were downloaded from ENCODE and normalized bigwigs were also generated using the Kent software (UCSC). Peaks were called on mapped reads using MACS2 (v2.1.2.1) with the following options: *-g dm -nomodel -bdg -t Samples -c Control*. Normalized bigwigs are available at: http://genome.ucsc.edu/s/cbravo/Bravo_et_al_EyeAntennalDisc.

For each sample, 50 single cells were simulated by bootstrapping 20,000 mapped reads (per cell) from the bulk bam files, resulting in a data set with 750 simulated single cells. Downstream analyses were performed with cisTopic²⁰. Briefly, we determined the number of ctx regions in which at least one read is mapped, and topic modelling was run using default parameters, with models including 2, 10 to 60 (one by one), 70, 80, 90 and 100 topics, using a total of 500 iterations, out of which 250 were used as burn-in. Based on the highest log-likelihood, we selected a model with 21 topics. Motif enrichment analysis was performed using RcisTarget and i-cisTarget, using a ROC threshold of 0.01 and maximum AUC rank of 5,000^{44,68}. The enrichment of epigenomic signatures in cells was performed using default parameters, using a maximum AUC rank of 12,320 (10% of the total number of ctx regions), while the enrichment of epigenomic regions within topics was done with default parameters. Coverage heatmaps were done using deepTools (v3.3.1). For identifying differentially enriched motifs between groups of regions, we first scored the ctx regions in the groups of interest with the 24,454 PWMs available in the cisTarget motif collection using Cluster-Buster¹⁰⁴, with the options *-m 0 -c*. Using this matrix, with ctx regions as columns, motifs as rows and the value of the best Cis-Regulatory Module (CRM) as value,

we performed a Likelihood Ratio test between the region groups of interest, as implemented in MAST (v1.4.1). P-values were adjusted using the FDR method.

Publicly available data used in this work

Eye disc Drop-seq data was obtained from GEO, with GEO accession number GSE115476, while dimensionality reduction coordinates and cell labels were retrieved from the supplementary data from Ariss *et al.*⁴³. Raw data from Optix-GFP⁺ single-cell and bulk ATAC-seq, Grh ChIP-seq and 21 bulk ATAC-seq profiles from were retrieved from GEO, with GEO accession number GSE102441. Raw Sine Oculis ChIP-seq data was retrieved from GEO, with GEO accession number GSE52943. Atonal Gain-Of-Function and Loss-Of-Function data was retrieved from GEO, with GEO accession number GSE16713. Differential expressed genes between GMR⁺ FAC sorted cells and GMR⁻ FAC sorted cells and onecutx562 versus WT was retrieved from the supplementary materials from Potier *et al.*⁶⁷. Glass, Prospero, Nerfin-1 and l(3)neo38 ChIP-seq profiles were retrieved from ENCODE, with the following experiment IDs, respectively: ENCSR472URU, ENCSR682YQM, ENCSR335NNR and ENCSR643EOU. ATAC-seq profiles on different embryonic domains were obtained from GEO, with GEO accession number GSE118240.

Resource description

SCope

- Ariss - WT 11416 cells
 - EAD_Ariss_WT_Seurat_SCENIC: Loom file containing dimensionality reductions (as shown by Ariss *et al.*, based on analysis with Seurat, and based on pySCENIC regulons), gene expression and regulon enrichment from pySCENIC (with regulons derived from pySCENIC in this data set) from the Drop-seq eye disc data set from Ariss *et al.*⁴³. This data set contains 11,416 cells, 7,801 genes and 140 regulons. The labels given by Ariss *et al.* ('Ariss labels') and the labels transferred with Seurat from the 10X scRNA-seq data ('10X labels') are given as metadata.
 - EAD_Ariss_WT_Seurat_SCENIC_regulonsfrom10X: Loom file containing dimensionality reductions (as shown by Ariss *et al.*, based on analysis with Seurat, and based on pySCENIC regulons), gene expression and regulon enrichment from pySCENIC (with motif-based regulons derived from pySCENIC in our 10X data set) from the Drop-seq eye disc data set from Ariss *et al.*⁴³. This data set contains 11,416 cells, 7,801 genes and 159 regulons. The labels given by Ariss *et al.* ('Ariss labels') and the labels transferred with Seurat from the 10X scRNA-seq data ('10X labels') are given as metadata.

- scATAC-seq - 15387 cells
 - Gene
 - EAD_scATAC_AggSumPredictiveDistribution_Gene_Regulons: Loom file containing the cisTopic cell-Topic tSNE coordinates, gene activity scores based on the aggregation of region probabilities around the TSS (5kb plus introns, multiplied by 10^6) and regulon enrichment (on the gene activity score matrix, using the regulons derived from the analysis with pySCENIC in the 10X scRNA-seq data set). This data set contains 15,387 cells, 16,892 genes and 175 regulons. The labels transferred from the 10X scRNA-seq data ('RNA labels') are given as metadata.
 - EAD_scATAC_AggSumPredictiveDistribution_Gene_Topics: Loom file containing the cisTopic cell-Topic tSNE coordinates, gene activity scores based on the aggregation of region probabilities around the TSS (5kb plus introns, multiplied by 10^6) and topic enrichment. This data set contains 15,387 cells, 16,892 genes and 49 topics. The labels transferred from the 10X scRNA-seq data ('RNA labels') are given as metadata.
 - EAD_scATAC_AggSignedImportancePredictiveDistribution_Gene_Regulon ss: Loom file containing the cisTopic cell-Topic tSNE coordinates, gene activity scores based on the aggregation of region probabilities based on the enhancer-to-gene links (multiplied by 10^8) and regulon enrichment (on the gene activity score matrix, using the regulons derived from the analysis with pySCENIC in the 10X scRNA-seq data set). This data set contains 15,387 cells, 8,347 genes and 175 regulons. The clusters derived from SNN clustering with Seurat on the topic-cell matrix ('Seurat_res_1.2') are given as metadata.
 - EAD_scATAC_AggSignedImportancePredictiveDistribution_Gene_Topics: Loom file containing the cisTopic cell-Topic tSNE coordinates, gene activity scores based on the aggregation of region probabilities based on the enhancer-to-gene links (multiplied by 10^8) and topic enrichment. This data set contains 15,387 cells, 8,347 genes and 49 topics. The clusters derived from SNN clustering with Seurat on the topic-cell matrix ('Seurat_res_1.2') are given as metadata.
 - Janelia
 - EAD_scATAC_AggSumPredictiveDistribution_JaneliaRegions_Topics: Loom file containing the cisTopic cell-Topic tSNE coordinates, Janelia region probabilities based on the aggregation of the probabilities of the ctx regions that overlap with the Janelia enhancer (multiplied by 10^6) and topic

enrichment. This data set contains 15,387 cells, 740 Janelia regions and 49 topics. The labels transferred from the 10X scRNA-seq data ('RNA labels') and the clusters derived from SNN clustering with Seurat on the topic-cell matrix ('Seurat_res_1.2') are given as metadata.

- Ctx Regions

- EAD_scATAC_PredictiveDistribution_CtxRegions_Topics: Loom file containing the cisTopic cell-Topic tSNE coordinates, ctx region probabilities (multiplied by 10^6) and topic enrichment. This data set contains 15,387 cells, 129,553 ctx regions and 49 topics. The labels transferred from the 10X scRNA-seq data ('RNA labels') and the clusters derived from SNN clustering with Seurat on the topic-cell matrix ('Seurat_res_1.2') are given as metadata.

- scRNA-seq - 3531 cells

- EAD_scRNAseq_LinkBasedandSeurat: Loom file containing dimensionality reductions (based on analysis with Seurat and based on pySCENIC regulons), gene expression and link-based regulon enrichment (regulons formed by performing the motif enrichment step of the SCENIC⁴⁴ workflow on the regions linked to each gene). This data set contains 3,531 cells, 8,744 genes and 161 regulons. The labels given by cell clustering with Seurat ('Seurat_res_1.2') and the experimental run ('Experiment run') are given as metadata.
- EAD_scRNAseq_SCENICandSeurat: Loom file containing dimensionality reductions (based on analysis with Seurat and based on pySCENIC regulons), gene expression and pySCENIC regulon enrichment (motif and ChIP-seq based). This data set contains 3,531 cells, 8,744 genes and 175 regulons. The labels given by cell clustering with Seurat ('Seurat_res_1.2'), the experimental run ('Experiment run'), the labels transferred from Ariss *et al.* ('Ariss labels'), and the labels transferred from the scATAC-seq data ('ATAC labels') are given as metadata.
- EAD_scRNAseq_SCENICandSeurat_regulonsfromAriss: Loom file containing dimensionality reductions (based on analysis with Seurat and based on pySCENIC regulons), gene expression and pySCENIC regulon enrichment (using the regulons derived from Ariss *et al.*). This data set contains 3,531 cells, 8,744 genes and 140 regulons. The labels given by cell clustering with Seurat ('Seurat_res_1.2'), the experimental run ('Experiment run'), the labels transferred from Ariss *et al.* ('Ariss labels'), and the labels transferred from the scATAC-seq data ('ATAC labels') are given as metadata.

- Virtual EAD - 5370 cells
 - Janelia
 - Janelia_Accessibility_AggSumProb: Loom file containing the virtual eye-antennal disc coordinates and the Janelia region probabilities based on the aggregation of the probabilities of the ctx regions that overlap with the Janelia enhancer (multiplied by 10^6). This data set contains 5,526 cells and 740 Janelia regions. The labelling of the cells in the virtual eye-antennal disc ('Zone') are given as metadata.
 - Janelia_Functionality_ImageRegistration: Loom file containing the virtual eye-antennal disc coordinates and the Janelia enhancer activity patterns mapped from the images into the virtual eye-antennal disc. This data set contains 5,058 cells and 454 Janelia mapped images (corresponding to 390 Janelia enhancers). The labelling of the cells in the virtual eye-antennal disc ('Zone') are given as metadata.
 - ATAC
 - Pseudotime-based_ATAC_VE_CtxRegions+Topics: Loom file containing the virtual eye-antennal disc coordinates, the ctx region probabilities (multiplied by 10^6) and the topic enrichment. This data set contains 5,526 cells, 129,553 ctx regions and 49 topics. The labelling of the cells in the virtual eye-antennal disc ('Zone') and the cell type labels based on the scATAC-seq data ('Cell type') are given as metadata.
 - RNA
 - Pseudotime-based_RNA_VE: Loom file containing the virtual eye-antennal disc coordinates, gene expression and the pySCENIC regulon enrichment (derived from the 10X scRNA-seq data). This data set contains 5,370 cells, 8,744 genes and 175 regulons. The labelling of the cells in the virtual eye-antennal disc ('Zone'), the cell type labels based on the scRNA-seq data ('Cell type'), the labels transferred from Ariss *et al.* ('Ariss labels') and from the scATAC-seq data ('ATAC labels') are given as metadata.

UCSC

- Custom tracks
 - Bulk ATAC DGRP regions: Bed file containing the 38,179 regions found accessible across the 50 bulk ATAC-seq profiles from *Drosophila* inbred lines.

- Color - Standardized R2G Coor+Genie3 50Kb: BigInteract track containing links between enhancers and target genes. The track is colored by the sign of the link, which can be positive (green) or negative (red). Default threshold: 0.
- Ctx_regions: Bed file containing the 129,553 ctx regions accessible in the eye-antennal disc in the *Drosophila* genome.
- Janelia lines: Bed file containing the coordinates of the enhancers tested by the Janelia FlyLight Project (with Janelia line ID).
- Optix-GFPVSRest_logFC1: Bed file containing the regions differentially accessible in the Optix-GFP⁺ cells compared to the Optix-GFP⁻ cells (with p-value < 0.05 and logFC > 1)
- RedFly (BED): Bed file containing the coordinates of the enhancers contained in the Redfly database¹⁰⁹.
- Score - Standardized R2G Coor+Genie3 50Kb: BigInteract track containing links between enhancers and target genes. The transparency of the links represents the Random Forest importance of the enhancer-to-gene link. Default threshold: 0.
- So ChIPseq peaks: Sine oculis (so) ChIP-seq peaks determined by MACS2 peak calling after remapping the data from Jusiak *et al.*¹¹⁰ to the 3rd 2017 FlyBase release (*D. melanogaster* r6.16) genome.
- Eye-Antennal Disc Hub (Bravo et al, 2019) @ aertslab.org
 - 10X topics: Topic bigwig files representing the region-topic scores obtained from the analysis of the 10X scATAC-seq with cisTopic²⁰ (v0.2.2).
 - Aggregate scATAC - Cell sorting: Aggregate profiles from the FAC-sorted Optix-GFP⁺ and sens-GFP⁺ cells as normalized bigwig files.
 - ATAC DGRP - Eye disc: Bulk ATAC-seq profiles from the 50 DGRP lines used in this study as normalized bigwig files.
 - ATAC Aggr EAD Clusters: Cell-type specific (based on clustering on the topic-cell matrix) aggregate profiles from the 10X scATAC-seq analysis as normalized bigwig files.
 - Bulk ATAC - Cell sorting: Bulk ATAC-seq profiles from the FAC-sorted Optix-GFP⁺ and sens-GFP⁺ cells as normalized bigwig files.
 - ENCODE ChIP-seq: Normalized bigwigs from the ChIP-seq experiments of Prospero, Nerfin-1 and l(3)neo38 (and controls) retrieved from ENCODE.
 - ENCODE Normalized ChIPseq: Control normalized bigwig files from the ChIP-seq experiments of Prospero, Nerfin-1 and l(3)neo38 retrieved from ENCODE.
 - Grh ChIP-seq: Normalized bigwigs from the Grainyhead ChIP-seq experiments performed by Jacobs *et al.*¹² after remapping to the 3rd 2017 FlyBase release (*D. melanogaster* r6.16) genome.

- 1495 ○ Predictive accessibility: Barchart track representing the region-cell probabilities
- 1496 (multiplied by 10^6) per cell type for each region.
- 1497 ○ RNA Aggr EAD Clusters: Cell type specific (based on Seurat clustering) normalized
- 1498 bigwigs containing 10X scRNA-seq reads.
- 1499 ○ scRNA Gene expression gene: Barchart track representing the normalized UMI counts
- 1500 per cell type for each gene.
- 1501 ○ scRNA Gene expression transcript: Barchart track representing the normalized UMI
- 1502 counts (multiplied by 10^2) per cell type for each transcript.
- 1503 ○ so ChIP-seq: Normalized bigwigs from the Grainyhead ChIP-seq experiments
- 1504 performed by Jusiak *et al.*¹¹⁰ after remapping to the 3rd 2017 FlyBase release
- 1505 (*D. melanogaster* r6.16) genome.
- 1506 ○ TF perturbations: Bulk ATAC-seq profiles from the GMR-GAL4 UAS-TF (and
- 1507 control) lines included in the genetic screen.

1508 **Data availability**

1509
1510 The data generated for this study have been deposited in NCBI's Gene Expression Omnibus and are
1511 accessible through GEO Series accession number GSE141590. We also provide a SScope session at
1512 http://scope.aertslab.org/#/Bravo_et_al_EyeAntennalDisc with the processed single-cell data and a
1513 UCSC hub (http://ucsctracks.aertslab.org/papers/Bravo_et_al_EyeAntennalDisc/hub.txt) and session at
1514 http://genome.ucsc.edu/s/cbravo/Bravo_et_al_EyeAntennalDisc with the processed aggregate and bulk
1515 ATAC-seq profiles, enhancer-to-gene links and ChIP-seq tracks, among others.

1516

1517 **Acknowledgements**

1518
1519 This work is funded by an ERC Consolidator Grant to S. Aerts (724226_cis-CONTROL); by the Special
1520 Research Fund (BOF) KU Leuven (grants PF/10/016, to S. Aerts), the Harry J. Lloyd Charitable Trust,
1521 the Foundation Against Cancer (2016-070, to S. Aerts) and a PhD fellowship from the F.W.O. to
1522 C.B.G.-B. (11F1519N). Stocks obtained from the Bloomington *Drosophila* Stock Center were used in
1523 this study. Single-cell infrastructure was funded by the Hercules Foundation (grant no. AKUL/13/41).
1524 Computing was performed at the Vlaams Supercomputer Center (VSC). The VIB BioImaging Core
1525 (Leuven platform) provided valuable insight on image processing. The authors thank to Maximilian
1526 Haeussler and Kate Rosenbloom (UCSC Genome Browser) for their help in data visualization on the
1527 UCSC Genome Browser; to the various groups that make curated position weight matrices publicly
1528 available, including T. Hughes (cis-bp), M. Bulyk (Uniprobe), A. Mathelier (Jaspar), V. Makeev
1529 (Hocomoco) and many others; to the Janelia FlyLight Project for publicly providing images and reporter
1530 lines to assess enhancer activity on imaginal discs and CNS in *Drosophila*; and to the ENCODE
1531 Consortium for publicly providing raw and processed data of a wide range of genomic assays.

Author contributions

S.Aerts, C.B.G.-B., K.D. and D.K. conceived the study; X.Q., I.I.T, K.D., D.K., V.C., S.M., D.M. and S.P. performed the experimental work; C.B.G.-B. conceptualized the computational approaches and performed the computational analyses with help of R.D.-R., I.T.T., K.D., D.K., G.H., M.W. and S.Aibar; C.B.G.-B. made the figures and C.B.G.-B. and S.Aerts wrote the manuscript.

Competing interest

The authors declare that no competing interests exist.

References

1. Fiers, M. W. E. J. *et al.* Mapping gene regulatory networks from single-cell omics data. *Brief. Funct. Genomics* doi:10.1093/bfpg/elx046.
2. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+ | Nature. <https://www.nature.com/articles/s41586-019-1049-y>.
3. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution | Science. <https://science.sciencemag.org/content/363/6434/1463>.
4. Karaikos, N. *et al.* The Drosophila embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
5. Nitzan, M., Karaikos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* **576**, 132–137 (2019).
6. Thornton, C. A. *et al.* Spatially-mapped single-cell chromatin accessibility. <http://biorxiv.org/lookup/doi/10.1101/815720> (2019) doi:10.1101/815720.
7. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
8. Xi, H. *et al.* Identification and Characterization of Cell Type-Specific and Ubiquitous Chromatin Regulatory Structures in the Human Genome. *PLOS Genet.* **3**, e136 (2007).
9. Gary, S. & Levin, M. Transcriptional repression in development. *Curr. Opin. Cell Biol.* **8**, 358–364 (1996).
10. Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
11. Li, L. M. & Arnosti, D. N. Long- and Short-Range Transcriptional Repressors Induce Distinct Chromatin States on Repressed Genes. *Curr. Biol.* **21**, 406–412 (2011).
12. Jacobs, J. *et al.* The transcription factor Grainyhead primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. *Nat. Genet.* **50**, 1011–1020 (2018).
13. Cusanovich, D. A. *et al.* The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).
14. Chen, S., Lake, B. B. & Zhang, K. Linking transcriptome and chromatin accessibility in nanoliter droplets for single-cell sequencing. <http://biorxiv.org/lookup/doi/10.1101/692608> (2019) doi:10.1101/692608.
15. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
16. Liu, L. *et al.* Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat. Commun.* **10**, 1–10 (2019).
17. Pervolarakis, N. *et al.* Integrated single-cell transcriptomics and chromatin accessibility analysis reveals novel regulators of mammary epithelial cell identity. *bioRxiv* 740746 (2019) doi:10.1101/740746.
18. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
19. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873–1887.e17 (2019).

20. Bravo González-Blas, C. *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
21. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858–871.e8 (2018).
22. Ghavi-Helm, Y. *et al.* Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* **51**, 1272–1282 (2019).
23. Nagano, T. *et al.* Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat. Protoc.* **10**, 1986–2003 (2015).
24. Dong, P. D. S., Dicks, J. S. & Panganiban, G. Distal-less and homothorax regulate multiple targets to pattern the Drosophila antenna. *Development* **129**, 1967–1974 (2002).
25. Emerald, B. S., Curtiss, J., Mlodzik, M. & Cohen, S. M. distal antenna and distal antenna related encode nuclear proteins containing pipsqueak motifs involved in antenna development in Drosophila. *Development* **130**, 1171–1180 (2003).
26. Roignant, J.-Y. & Treisman, J. E. Pattern formation in the Drosophila eye disc. *Int. J. Dev. Biol.* **53**, 795–804 (2009).
27. Jory, A. *et al.* A Survey of 6,300 Genomic Fragments for cis-Regulatory Activity in the Imaginal Discs of Drosophila melanogaster. *Cell Rep.* **2**, 1014–1024 (2012).
28. Davie, K. *et al.* A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. *Cell* **174**, 982–998.e20 (2018).
29. 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.e4 (2019).
30. Stultz, B. G., Park, S. Y., Mortin, M. A., Kennison, J. A. & Hursh, D. A. Hox proteins coordinate peripodial decapentaplegic expression to direct adult head morphogenesis in Drosophila. *Dev. Biol.* **369**, 362–376 (2012).
31. Blanco, J., Pauli, T., Seimiya, M., Udolph, G. & Gehring, W. J. Genetic interactions of eyes absent, twin of eyeless and orthodenticle regulate sine oculis expression during ocellar development in Drosophila. *Dev. Biol.* **344**, 1088–1099 (2010).
32. Tiklová, K., Tsarouhas, V. & Samakovlis, C. Control of Airway Tube Diameter and Integrity by Secreted Chitin-Binding Proteins in Drosophila. *PLOS ONE* **8**, e67415 (2013).
33. Frankfort, B. J., Nolo, R., Zhang, Z., Bellen, H. & Mardon, G. senseless Repression of rough Is Required for R8 Photoreceptor Differentiation in the Developing Drosophila Eye. *Neuron* **32**, 403–414 (2001).
34. Mlodzik, M., Hiromi, Y., Weber, U., Goodman, C. S. & Rubin, G. M. The drosophila seven-up gene, a member of the steroid receptor gene superfamily, controls photoreceptor cell fates. *Cell* **60**, 211–224 (1990).
35. Charlton-Perkins, M. *et al.* Prospero and Pax2 combinatorially control neural cell fate decisions by modulating Ras- and Notch-dependent signaling. *Neural Develop.* **6**, 20 (2011).
36. Nolo, R., Abbott, L. A. & Bellen, H. J. Senseless, a Zn Finger Transcription Factor, Is Necessary and Sufficient for Sensory Organ Development in Drosophila. *Cell* **102**, 349–362 (2000).
37. EBERL, D. F. & BOEKHOFF-FALK, G. Development of Johnston’s organ in Drosophila. *Int. J. Dev. Biol.* **51**, 679–687 (2007).
38. Yuasa, Y. *et al.* Drosophila homeodomain protein REPO controls glial differentiation by cooperating with ETS and BTB transcription factors. *Development* **130**, 2419–2428 (2003).
39. Minakhina, S., Tan, W. & Steward, R. JAK/STAT and the GATA factor Pannier control hemocyte maturation and differentiation in Drosophila. *Dev. Biol.* **352**, 308–316 (2011).
40. Beira, J. V. & Paro, R. The legacy of Drosophila imaginal discs. *Chromosoma* **125**, 573–592 (2016).
41. Furlong, E. E. M., Andersen, E. C., Null, B., White, K. P. & Scott, M. P. Patterns of Gene Expression During Drosophila Mesoderm Development. *Science* **293**, 1629–1633 (2001).
42. Oyallon, J. *et al.* Regulation of locomotion and motoneuron trajectory selection and targeting by the Drosophila homolog of Olig family transcription factors. *Dev. Biol.* **369**, 261–276 (2012).
43. Ariss, M. M., Islam, A. B. M. M. K., Critcher, M., Zappia, M. P. & Frolov, M. V. Single cell RNA-sequencing identifies a metabolic aspect of apoptosis in Rbf mutant. *Nat. Commun.* **9**, 1–13 (2018).
44. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083 (2017).

45. Ahn, Y., Zou, J. & Mitchell, P. J. Segment-specific regulation of the *Drosophila* AP-2 gene during leg and antennal development. *Dev. Biol.* **355**, 336–348 (2011).
46. Tsuji, T. Lim1 function in leg and antennal development. 9.
47. Emmons, R. B., Duncan, D. & Duncan, I. Regulation of the *Drosophila* distal antennal determinant spineless. *Dev. Biol.* **302**, 412–426 (2007).
48. Higashijima, S. *et al.* Dual Bar homeo box genes of *Drosophila* required in two photoreceptor cells, R1 and R6, and primary pigment cells for normal eye development. *Genes Dev.* **6**, 50–60 (1992).
49. Bessa, J., Gebelein, B., Pichaud, F., Casares, F. & Mann, R. S. Combinatorial control of *Drosophila* eye development by Eyeless, Homothorax, and Teashirt. *Genes Dev.* **16**, 2415–2427 (2002).
50. Aerts, S. *et al.* Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in *Drosophila* uncovers a regulatory basis for sensory specification. *PLoS Biol.* **8**, e1000435 (2010).
51. Herrmann, C., Van de Sande, B., Potier, D. & Aerts, S. i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res.* **40**, e114 (2012).
52. Ostrin, E. J. *et al.* Genome-wide identification of direct targets of the *Drosophila* retinal determination protein Eyeless. *Genome Res.* **16**, 466–476 (2006).
53. Pepple, K. L. *et al.* Two-step selection of a single R8 photoreceptor: a bistable loop between senseless and rough locks in R8 fate. *Dev. Camb. Engl.* **135**, 4071–4079 (2008).
54. Pappu, K. S. *et al.* Dual regulation and redundant function of two eye-specific enhancers of the *Drosophila* retinal determination gene *dachshund*. *Development* **132**, 2895–2905 (2005).
55. Fritsch, C. *et al.* Multilevel regulation of the glass locus during *Drosophila* eye development. *PLOS Genet.* **15**, e1008269 (2019).
56. Barolo, S. Shadow enhancers: Frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *BioEssays* **34**, 135–141 (2012).
57. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
58. Perry, M. W., Boettiger, A. N., Bothma, J. P. & Levine, M. Shadow Enhancers Foster Robustness of *Drosophila* Gastrulation. *Curr. Biol.* **20**, 1562–1567 (2010).
59. Frankel, N. *et al.* Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**, 490–493 (2010).
60. Agawa, Y. *et al.* *Drosophila* Blimp-1 Is a Transient Transcriptional Repressor That Controls Timing of the Ecdysone-Induced Developmental Pathway. *Mol. Cell. Biol.* **27**, 8739–8747 (2007).
61. Daga, A., Karlovich, C. A., Dumstrei, K. & Banerjee, U. Patterning of cells in the *Drosophila* eye by Lozenge, which shares homologous domains with AML1. *Genes Dev.* **10**, 1194–1205 (1996).
62. Anderson, A. M., Weasner, B. M., Weasner, B. P. & Kumar, J. P. Dual transcriptional activities of SIX proteins define their roles in normal and ectopic eye development. *Dev. Camb. Engl.* **139**, 991–1000 (2012).
63. Lopes, C. S. & Casares, F. Eye Selector Logic for a Coordinated Cell Cycle Exit. *PLOS Genet.* **11**, e1004981 (2015).
64. Hamm, D. C. *et al.* A conserved maternal-specific repressive domain in Zelda revealed by Cas9-mediated mutagenesis in *Drosophila melanogaster*. *PLoS Genet.* **13**, (2017).
65. Domingos, P. M. *et al.* Regulation of R7 and R8 differentiation by the spalt genes. *Dev. Biol.* **273**, 121–133 (2004).
66. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
67. Potier, D. *et al.* Mapping Gene Regulatory Networks in *Drosophila* Eye Development by Large-Scale Transcriptome Perturbations and Motif Inference. *Cell Rep.* **9**, 2290–2303 (2014).
68. Imrichová, H., Hulselmans, G., Kalender Atak, Z., Potier, D. & Aerts, S. i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Res.* **43**, W57–W64 (2015).
69. Iyer, J. *et al.* Quantitative Assessment of Eye Phenotypes for Functional Genetic Studies Using *Drosophila melanogaster*. *G3 GenesGenomesGenetics* **6**, 1427–1437 (2016).
70. Southall, T. D., Davidson, C. M., Miller, C., Carr, A. & Brand, A. H. Dedifferentiation of Neurons Precedes Tumor Formation in *lola* Mutants. *Dev. Cell* **28**, 685–696 (2014).

71. Kuzin, A., Brody, T., Moore, A. W. & Odenwald, W. F. Nerfin-1 is required for early axon guidance decisions in the developing *Drosophila* CNS. *Dev. Biol.* **277**, 347–365 (2005).
72. Karim, M. R. & Moore, A. W. Convergent Local Identity and Topographic Projection of Sensory Neurons. *J. Neurosci.* **31**, 17017–17027 (2011).
73. Burgess, D. J. Spatial transcriptomics coming of age. *Nat. Rev. Genet.* **20**, 317–317 (2019).
74. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
75. Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
76. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013).
77. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
78. Cannavò, E. *et al.* Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Curr. Biol.* **26**, 38–51 (2016).
79. Fitzgerald, D. P. & Bender, W. Polycomb group repression reduces DNA accessibility. *Mol. Cell. Biol.* **21**, 6585–6597 (2001).
80. Bozek, M. *et al.* ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the *Drosophila* blastoderm. *Genome Res.* **29**, 771–783 (2019).
81. Small, S., Blair, A. & Levine, M. Regulation of even-skipped stripe 2 in the *Drosophila* embryo. *EMBO J.* **11**, 4047–4057 (1992).
82. Chen, S. & Mar, J. C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* **19**, 232 (2018).
83. 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).
84. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3**, 346–360.e4 (2016).
85. Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555–568 (2015).
86. Pataskar, A. *et al.* NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program. *EMBO J.* **35**, 24–45 (2016).
87. Fuxman Bass, J. I., Reece-Hoyes, J. S. & Walhout, A. J. M. Gene-Centered Yeast One-Hybrid Assays. *Cold Spring Harb. Protoc.* **2016**, pdb.top077669 (2016).
88. Rubin, A. J. *et al.* Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* **176**, 361–376.e17 (2019).
89. Frolidi, F. *et al.* The transcription factor Nerfin-1 prevents reversion of neurons into neural stem cells. *Genes Dev.* **29**, 129–143 (2015).
90. Forbes-Osborne, M. A., Wilson, S. G. & Morris, A. C. Insulinoma-associated 1a (Insm1a) is required for photoreceptor differentiation in the zebrafish retina. *Dev. Biol.* **380**, 157–171 (2013).
91. Nelson, B. R., Hartman, B. H., Georgi, S. A., Lan, M. S. & Reh, T. A. Transient inactivation of Notch signaling synchronizes differentiation of neural progenitor cells. *Dev. Biol.* **304**, 479–498 (2007).
92. Dyer, M. A., Livesey, F. J., Cepko, C. L. & Oliver, G. Prox1 function controls progenitor cell proliferation and horizontal cell genesis in the mammalian retina. *Nat. Genet.* **34**, 53–58 (2003).
93. Edqvist, P. H. D., Myers, S. M. & Hallböök, F. Early identification of retinal subtypes in the developing, pre-laminated chick retina using the transcription factors Prox1, Lim1, Ap2alpha, Pax6, Isl1, Isl2, Lim3 and Chx10. *Eur. J. Histochem. EJH* **50**, 147–154 (2006).
94. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
95. Davie, K. *et al.* Discovery of Transcription Factors and Regulatory Regions Driving In Vivo Tumor Development by ATAC-seq and FAIRE-seq Open Chromatin Profiling. *PLOS Genet.* **11**, e1004994 (2015).
96. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).

97. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
98. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
99. Gramates, L. S. *et al.* FlyBase at 25: looking to the future. *Nucleic Acids Res.* **45**, D663–D671 (2017).
100. *pyscenic: Python implementation of the SCENIC pipeline for transcription factor inference from single-cell transcriptomics experiments.*
101. Suo, S. *et al.* Revealing the Critical Regulators of Cell Identity in the Mouse Cell Atlas. *Cell Rep.* **25**, 1436–1445.e3 (2018).
102. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* (2019) doi:10.1038/s41592-019-0619-0.
103. Angerer, P. *et al.* destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241–1243 (2016).
104. Frith, M. C., Li, M. C. & Weng, Z. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* **31**, 3666–3668 (2003).
105. Aerts, S. *et al.* TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res.* **33**, W393–W396 (2005).
106. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
107. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE* **6**, e21800 (2011).
108. Moerman, T. *et al.* GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159–2161 (2019).
109. Rivera, J., Keränen, S. V. E., Gallo, S. M. & Halfon, M. S. REDfly: the transcriptional regulatory element database for Drosophila. *Nucleic Acids Res.* **47**, D828–D834 (2019).
110. Jusiak, B. *et al.* Genome-wide DNA binding pattern of the homeodomain transcription factor Sine oculis (So) in the developing eye of Drosophila melanogaster. *Genomics Data* **2**, 153–155 (2014).