The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes

Jelle Jacobs1,2, Mardelle Atkins3,4, Kristofer Davie1,2, Hana Imrichova1,2, Lucia Romanelli3,4, Valerie Christiaens1,2, Gert Hulselmans1,2, Delphine Potier1,2, Jasper Wouters1,2, Ibrahim I. Taskiran5, Giulia Paciello6, Carmen B. González-Blas1,2, Duygu Koldere1,2, Sara Aibar1,2, Georg Halder3,4 and Stein Aerts1,2*

Transcriptional enhancers function as docking platforms for combinations of transcription factors (TFs) to control gene expression. How enhancer sequences determine nucleosome occupancy, TF recruitment and transcriptional activation in vivo remains unclear. Using ATAC-seq across a panel of Drosophila inbred strains, we found that SNPs affecting binding sites of the TF Grainy head (Grh) causally determine the accessibility of epithelial enhancers. We show that deletion and ectopic expression of Grh cause loss and gain of DNA accessibility, respectively. However, although Grh binding is necessary for enhancer accessibility, it is insufficient to activate enhancers. Finally, we show that human Grh homologs—GRHL1, GRHL2 and GRHL3—function similarly. We conclude that Grh binding is necessary and sufficient for the opening of epithelial enhancers but not for their activation. Our data support a model positing that complex spatiotemporal expression patterns are controlled by regulatory hierarchies in which pioneer factors, such as Grh, establish tissue-specific accessible chromatin landscapes upon which other factors can act.

Each cell type in the human body expresses a unique set of genes. Deciphering the regulatory programs that govern these transcriptional states requires predictive models that can link the genome sequence with the recruitment of TFs and other DNA-binding proteins, and then link this protein binding to chromatin state, the chromatin state with enhancer function, and the enhancer function with target gene expression. A better understanding of the DNA regulatory code will advance interpretation of noncoding sequence variation and may ultimately provide new ways to detect and manipulate cell states, for example in cancer cells or for regenerative medicine.

About a decade ago, the main bottleneck for deciphering gene regulation was to locate all of the enhancers involved in the control of a certain cell state. Today, that problem has largely been solved thanks to advances in epigenomics1–3. Indeed, high-throughput techniques such as ChIP–seq4 against TFs, cofactors and histone modifications, DNaseI-seq5 and ATAC–seq6 for open-chromatin profiles, and DNA methylation profiling7 have delivered genome-wide enhancer predictions for a multitude of healthy and diseased tissues and cell types2,3.

For TP53, it was recently shown that, after binding, it not only out-competes nucleosomes but also activates gene expression22,24. For Drosophila, the only pioneer factor known to date is Zelda, which establishes the chromatin landscape in early Drosophila embryos during the maternal-to-zygotic transition16,27. Here we set out to study how the three layers—namely, sequence information, chromatin accessibility and enhancer function—are linked in epithelial cells. To this end, we used a combination of chromatin accessibility (bulk, cell-sorted and single-cell ATAC–seq),
natural variation, machine learning, evolutionary variation and in vivo enhancer reporter assays in epithelial tissues in Drosophila. Our results provide strong evidence for a hierarchical model of enhancer control that is conserved across animals.

Results

Natural variation in chromatin accessibility predicts potential chromatin regulators. To identify DNA features that are essential for establishing chromatin accessibility in a specific tissue, we profiled open chromatin across a cohort of inbred Drosophila strains. In particular, we performed 30 ATAC-seq experiments on epithelial tissues (eye–antennal imaginal discs), covering 23 distinct strains from the Drosophila Genetic Reference Panel (DGRP)\(^\text{28,29}\) (Methods). The open-chromatin profiles were highly similar (\(\rho = 0.76–0.96\) within the eye–antennal discs in this cohort of flies and differed substantially (\(\rho = 0.25–0.50\)) from the open-chromatin profiles of non-epithelial tissues, such as the adult brain\(^\text{30}\) (Fig. 1a and Supplementary Fig. 1).

We identified 30,774 accessible regions across the 30 eye–antennal disc samples (Methods). To link variation in chromatin accessibility with underlying sequence variation (cis variation), we applied a generalized linear model (GLM) on all 297,000 SNPs that were present in accessible regions. This analysis identified 4,289 \((~1.5\%)\) SNPs that correlated significantly (false discovery rate \((FDR) < 0.05\)) with accessibility changes in their encompassing regions (Fig. 1b), termed chromatin accessibility quantitative trait loci (caQTLs)\(^\text{\text{31}}\) (Methods). The 4,289 caQTLs were located in 2,048 regions (Fig. 1b,c) were the most strongly enriched in Grh-specific ChIP-seqings thus predict that a Grh-binding site can causally determine the in vivo accessibility of an enhancer-sized region.

Grainy head has a key role in the chromatin landscape of epithelial cells. Notably, Grh is a highly conserved TF with essential roles in epithelial cell-fate specification and wound healing across animals\(^\text{32–36}\). Of the 30,774 accessible regions of the eye–antennal discs, 10.5% (3,246 regions) had at least one Grh-binding site. These 3,246 potential Grh target regions were located near a large set of 1,786 genes that are strongly enriched for Gene Ontology (GO)\(^\text{37}\) terms such as epithelium development \((P = 10^{-36})\) (Supplementary Table 2). The Grh target regions contained 22.2% of the mapped reads and were in general the most accessible regions of the epithelial chromatin landscape \((P = 2.251 \times 10^{-146})\) by Welch’s \(t\) test (Fig. 2a). We hypothesized that Grh target regions are the most accessible because Grh stably binds its target sites in a large fraction of cells in the eye–antennal discs. Supporting this hypothesis is the fact that Grh proteins are ubiquitously expressed in basically all cells of the eye–antennal and wing imaginal discs (Fig. 2b,c).

To test whether Grh is actually bound to the predicted Grh-binding sites inside the accessible regions of eye–antennal discs, we performed ChiPmentation against GFP-tagged Grh\(^\text{39}\) and also reanalyzed published anti-Grh ChIP–seq data\(^\text{40}\). The same Grh motifs that concordantly changed with chromatin accessibility (Fig. 1c,d) were the most strongly enriched in Grh-specific ChIP
peaks (normalized enrichment score (NES) = 12.30)\(^{41,42}\). Of note, the ChIP-seq signal across the 3,246 Grh target regions correlated quantitatively (\(\rho = 0.92\)) with the ATAC-seq signal (Fig. 2d) and covered previously characterized Grh-dependent enhancers\(^{43,44}\) (Fig. 2e).

We then evaluated the occupancy of the Grh-binding sites in an independent assay by performing in vivo DNA footprinting analysis\(^{5,15}\) (Methods) and found that Grh-binding sites had a protection profile that was as strong as that seen for nucleosomal or silent DNA (Fig. 2e). These findings indicate that, whenever a region with a Grh motif is accessible, Grh is stably bound there. Overall, these data suggest that Grh has a key role, both in breadth (many target regions) and in depth (the highest peaks), in the accessible chromatin landscape of epithelial cells.

**Grainy head–binding sites are essential for enhancer activity.** So far, we have shown that changes in a Grh-binding site could generate or destroy the accessibility of an enhancer-sized region in developing epithelia. To test whether these regions were indeed functional enhancers, we cloned four enhancer pairs, each of which represented different SNP sequences, comprising one sequence with and one sequence without the caQTL that affected a Grh-binding site. These regions were individually cloned into GFP reporter vectors\(^4\) and stably integrated into the same position in the fly genome (Methods) (Fig. 3a). Notably, the accessibility profile, as assayed by ATAC-seq of these integrated fragments was entirely determined by their sequence and was independent of the local 3D chromatin context (Fig. 3a–d).

Next, by examining the expression pattern of the GFP reporter gene in eye and wing imaginal discs, we tested the direct effect of the caQTLs on the activity of these potential enhancers. Notably, the accessible sequences with an intact Grh-binding site drove GFP expression in specific and reproducible patterns for all four fragments (Fig. 3e,g,i,k). Their counterparts, on the other hand, which lacked Grh-binding sites, were predominantly inactive (Fig. 3f,h,j,l). Thus, the Grh-binding sites are necessary for both enhancer accessibility and activity.

**Grainy head directly regulates enhancer accessibility but not activity.** Our transgenic reporters showed a clear correlation between the presence of a Grh-binding site and GFP expression. The four expression patterns were, however, distinct from each other and not ubiquitous (Fig. 3e,g,i,k). Taking into account the idea that Grh binding nevertheless determines chromatin accessibility and that Grh is ubiquitously expressed in the discs, we hypothesized that Grh binding opens or ‘primed’ its target enhancers, without necessarily activating them per se. Regions that do not yield any
ρ15 showed activity in the eye–antennal disc (Fig. 4a). Of note, there were additional Grh target enhancers (Supplementary Table 3), of which primes, its target enhancers without directly activating them.

To test whether Grh was able to bind to enhancers without directly activating them, we measured the reporter activity of 21 Grh-bound enhancers and the number of cells in which the GFP reporter was active (Fig. 4b). Furthermore, using single-cell ATAC–seq, we found that the actual number of single cells in which the Grh-bound enhancers were accessible did correlate with bulk ATAC–seq signals (Fig. 4c and Supplementary Fig. 2). These results suggest that Grh binding opens its target enhancers without necessarily activating them.

Next, to demonstrate that these enhancers could indeed be accessible in cells in which they were not active, we used flow cytometry to sort cells and performed ATAC–seq on a specific subpopulation of cells from the eye–antennal disc (example in Fig. 5b). These two criteria were nearly 100% predictive of enhancers that were actually bound by Grh (Fig. 2d,e). Next, we scanned these enhancers for conserved motifs from our library of 18,832 TF position-weight matrices (PWMs), using a branch length score (BLS) (Fig. 5a), and identified several co-conserved TF-binding motifs (Fig. 5c and Supplementary Table 5). Notably, one of the top co-conserved motifs was CANNTG, an E-box motif that is active in a subset of cells in imaginal discs. These E-boxes were conserved in 92 Grh enhancers (area defined by the asterisk in Fig. 5d; example in Fig. 5b), which were located near Ato-induced genes (Supplementary Fig. 3), suggesting that the ‘Grh + Ato’ enhancers are functional. Furthermore, of 20 experimentally validated Ato target enhancers, we found 6 to be co-bound by Grh and Ato (Supplementary Fig. 3) and identified evolutionary comparison of Grainy head–bound enhancers identifies candidate activators. The tested Grh-bound enhancers showed a large variety of expression patterns, suggesting that multiple other TFs can cooperate with Grh. To identify candidate co-TFs, we performed cross-species motif analysis, which is a powerful approach to identify functional TF-binding sites. We thus performed ATAC–seq analysis of the eye– antennal discs of ten additional Drosophila species and selected conserved enhancers that had a Grh motif and were accessible in the other species (example in Fig. 5b). These two criteria were nearly 100% predictive of enhancers that were actually bound by Grh (Fig. 2d,e). Next, we scanned these enhancers for conserved motifs from our library of 18,832 TF position-weight matrices (PWMs), using a branch length score (BLS) (Fig. 5a), and identified several co-conserved TF-binding motifs (Fig. 5c and Supplementary Table 5). Notably, one of the top co-conserved motifs was CANNTG, an E-box motif for which the best candidate activator is Atonal (Ato), a basic helix-loop-helix (bHLH) factor that is active in a subset of cells in imaginal discs. These E-boxes were conserved in 92 Grh enhancers (area defined by the asterisk in Fig. 5d; example in Fig. 5b), which were located near Ato-induced genes (Supplementary Fig. 3), suggesting that the ‘Grh + Ato’ enhancers are functional. Furthermore, of 20 experimentally validated Ato target enhancers, we found 6 to be co-bound by Grh and Ato (Supplementary Fig. 3) and identified noticeable transcriptional activity, despite being accessible or bound by TFs, have also been reported in other studies.

Next, to demonstrate that these enhancers could indeed be accessible in cells in which they were not active, we used flow cytometry to sort cells and performed ATAC–seq on a specific subpopulation of cells from the eye–antennal disc (Fig. 4d and Supplementary Note 1). The ATAC–seq experiments were done once for each line. A magnified view of the affected domain (Fig. 4f). These findings indicate that Grh binding opens, or primes, its target enhancers without directly activating them.
4 new Grh target enhancers that were active in Ato-expressing cells (BL48037, BL46823, BL38727 and BL50129; Fig. 4a). We conclude from these results that the activity of enhancers primed through Grh binding requires the recruitment of additional factors like Ato.

**Loss-of-function and gain-of-function experiments indicate that Grainy head is a pioneer factor.** The data presented so far suggest that Grh is a pioneer TF for epithelial cells. The main function attributed to pioneer TFs is binding of their recognition sites within closed chromatin, making their target regions accessible and available for other proteins to bind. This means that removal of a pioneer factor should directly reduce the accessibility of its target regions whereas these regions should become accessible after ectopic expression of the factor.

We tested both predictions and first investigated how lack of Grh proteins impacted the open-chromatin landscape of epithelial cells. We thus performed bulk ATAC-seq analysis on eye-antennal discs that were largely mutant for Grh proteins (Fig. 3). Comparison of the accessible chromatin landscape of Grh mutant discs with those from matching controls identified 1,076 regions with reduced accessibility (log(fold change) < −0.5; P < 0.1) in the Grh mutant discs. Notably, only Grh motifs (from our 18,000 motifs) were strongly enriched in these regions (i-cisTarget NES = 21). Known Grh target enhancers near genes normally expressed in epithelia, such as Cad99C or jar, lost accessibility in the grh-mutant discs (Fig. 6b). Overall, the 3,246 Grh target regions were significantly less accessible in the Grh-mutant tissue (P = 3.28 × 10⁻⁸ by Welch’s t test; Fig. 6c). In summary, these results suggest that loss of Grh directly reduces the accessibility of its target regions.

Next, we tested whether Grh was able to bind to nucleosomal DNA and make its target regions accessible, as this is a crucial feature of pioneer factors. To evaluate this, we analyzed larval brains, which mainly comprise neurons and glial cells that do not express Grh and in which we found that the majority (74.3%) of Grh target regions were inaccessible. If Grh were a true pioneer factor, it would be able to bind to its target regions within this closed chromatin, making them accessible. To test this, we ectopically expressed the epithelial isoform of grh in Drosophila larval brains (Fig. 6d), by using a pan-neuronal elav driver. We then performed ATAC-seq analysis and compared the accessible chromatin landscape of GrhN-expressing brains with that of wild-type brains. Of note, practically all 3,246 Grh target regions identified in the eye-antennal discs specifically gained accessibility in response to ectopic Grh expression (P = 1.9 × 10⁻¹⁰ by Welch’s t test; Fig. 6e), whereas the neuronal cis-tronome did not change (Supplementary Note 2). Furthermore, only Grh motifs were significantly enriched (NES = 26.6) in the 1,774 regions with the strongest accessibility gain (log(fold change) > 0.5; P < 0.05) (Fig. 6f shows two examples). Potential target genes near these regions were enriched for GO terms that are normally not present in brain tissues, such as embryonic dorsal epidermis.
Additionally, dinucleotide- and trinucleotide-repeat sequences bound or accessible Grh motifs (Fig. 7a). By combining five different Grh motif matches, we identified a set of 1,300 regions with recurrent epidermal tissues (Fig. 7a). From these 10,000 highest-scoring Grh motif matches in the genome and ordered them according to Grh occupancy (ChIP–seq meta-analysis; Methods). The Grh-specific motif matches in the genome and ordered them according to Grh occupancy (ChIP–seq meta-analysis; Methods). The Grh-specific ChIPmentation and ATAC signal in wing imaginal discs clearly followed this ranking, indicating that, despite the different fates of these cells, Grh is able to bind to the same target regions in both epithelial discs of 11 Drosophila species. The predicted nucleosome occupancy profile for precision-recall curves, see Supplementary Fig. 5). Hence, a suitable Grh motif in a local ‘favorable’ sequence context is predictive for its in vivo binding.

Because nucleosome positioning is also strongly affected by local sequence context21, we investigated whether there was a difference in nucleosome affinity between the functional and nonfunctional motifs. We found that the predicted nucleosome occupancy profiles differed substantially, with a pronounced dip in nucleosome occupancy at nonfunctional motifs and a wider increase around functional Grh motifs (Fig. 7f). This indicates that Grh preferentially binds to DNA sites in regions that have a high affinity for nucleosomes (which are likely to be bound by nucleosomes in the absence of Grh), similar to other pioneer factors such as PU.1, FOXA, SOX2 and TP5318,20–22,24,56. Thus, in vivo Grh binding is highly predictive and associated with nucleosome displacement.

Grainy head homologs have similar roles in human cells. We have identified Grh as the principal pioneer factor of the epithelial accessible chromatin landscape in Drosophila. Mammals have three Grh-like TFs (GRHL1, GRHL2 and GRHL3) with known and conserved target genes involved in cell adhesion58–60. Because the DNA-binding domain of Grh proteins is highly conserved across animals32,33, we investigated whether Grh homologs could have similar pioneering functions in mammalian cells.

First, using publicly available ChIP–seq datasets for GRHL238,39, we confirmed that the DNA-binding motif for Grh was conserved between Drosophila and mammals (Fig. 8a). We then identified a core functional Grainy head motifs are embedded in a specific DNA context. Having established that Grh acts as a pioneer factor, capable of binding and opening its target regions in different tissues, we sought to investigate which sequence elements determine in vivo Grh binding. We first assessed whether the genomic Grh binding pattern was simply determined by affinity for its binding sites. We collected the 10,000 highest-scoring Grh motif matches in the genome and ordered them according to Grh occupancy (ChIP–seq meta-analysis; Methods). The Grh-specific ChIPmentation in D. melanogaster and D. yakuba DNA context.
set of regions that are recurrently bound by GRHL2 (Fig. 8b, ChIP meta-analysis; Methods) and compared the predicted nucleosome-occupancy profiles of functional versus nonfunctional GRHL2 motifs. We found that GRHL2 binding occurred more frequently in regions with a high preference for nucleosomes (Fig. 8c), which are strongly enriched for GO terms such as epithelium development (FDR = 1.43 × 10^{-45}) and cell junction (FDR = 4.21 × 10^{-49}). Combined with previous studies, these results indicate that the GRHL TFs have a key role in the accessible chromatin landscape in mammalian epithelial cells.

Finally, we investigated whether the GRHL TFs also have a direct impact on chromatin accessibility. The GRHL1 and GRHL2 genes are highly expressed in MCF-7 cells, and we had identified a set of putative GRHL target regions. To test whether the accessibility of these regions was due to GRHL binding, we performed omni-ATAC–seq analysis on control MCF-7 cells and on MCF-7 cells in which expression of GRHL1, GRHL2 and GRHL3 was knocked down for 48h using siRNAs (Fig. 8d and Methods). We found that the regions with reduced accessibility after knockdown of these genes were strongly enriched for GRHL target regions (P = 2.11 × 10^{-47}) by Fisher’s exact test) and GRHL-binding sites (NES = 12.8). These regions included GRHL-regulated enhancers near target genes that encode proteins involved in cell adhesion, such as PCDH1 and SPINT1 (Fig. 8e). These findings suggest that the function of Grh as a pioneer factor of the epithelial chromatin landscape is conserved between Drosophila and mammals.
Discussion

We performed an in vivo caQTL study that links sequence variation with changes in chromatin accessibility and enhancer activation. Previous caQTL studies have been performed in human lymphoblastoid or induced pluripotent stem cells33,34. Performing a caQTL study in a complex, developing tissue with a diversity of spatiotemporal expression patterns enabled us to link regulatory variation with developmental enhancers. We discovered that SNPs that change a Grh-binding site can causally determine the in vivo accessibility of an enhancer-sized region. Even though these enhancers also contain binding sites for other TFs, their accessibility can be attributed to Grh binding. Indeed, accessibility to epithelial cell enhancers was instigated when Grh was ectopically introduced in non-epithelial cells and abolished when Grh was removed from epithelial tissue. We found that functional Grh-binding sites are generally located in regions with a high affinity for nucleosomes, suggesting that in the absence of Grh its target enhancers are repressed owing to strongly bound nucleosomes. Such an elegant scenario, in which pioneer target enhancers are kept ‘off’ in other cell types by nucleosome binding, was recently proposed as the ‘default off’ model20.

Grh is ubiquitously expressed in the imaginal disk, and its stable binding across multiple cell types is in agreement with a recent study in the *Drosophila* embryo35, where Grh was found stably bound to its target genes throughout embryonic development. Even though Grh was bound to its targets in all cells of a tissue, the tested Grh-bound enhancers were active in different patterns or at different stages. This observation is reminiscent of Zelda in the embryo19,26,27, where the spatiotemporal enhancer activity of Zelda targets is induced by binding of combinations of other factors, such as the maternal activator Bicoid and the factors encoded by the gap genes hunchback, giant and Kruppel35. Also, for Grh we discovered many potential co-regulatory TFs, such as Ato, which is expressed and active in well-defined subparts of the eye–antennal disc34. On the other hand, there are some Grh target genes, such as *coracle*34, *Fasciclin3*34 and *grh* itself, that are ubiquitously expressed in epithelia. Identifying the cofactors or additional mechanisms behind this ubiquitous expression could be an interesting future challenge.

Of note, only a subset of the enhancers that are active in imaginal discs are dependent on Grh. Many eye-specific enhancers, such as Optix or p53 target enhancers (Supplementary Notes 1 and 3), likely become accessible owing to binding of the respective factor. Thus, multiple regulatory programs run in parallel within the same tissue, each with their array of target enhancers that are nucleosome free either as a result of binding of pioneer (or lineage) factors or through the cooperative binding of multiple TFs35. The possibility that multiple regulatory layers are simultaneously active in the same cell also became apparent in neurons after ectopic expression of Grh, which triggered ectopic accessibility of the entire epithelial cistrome without affecting the neuronal cistrome (Supplementary Note 2). This further suggests that pioneer factors work in parallel, where each pioneer factor finds and opens its specific target regions, jointly establishing the accessible chromatin landscape upon which other factors can act.

A consequence of the hierarchical enhancer model is that co-regulatory factors can also be repressors, which could explain why key TFs, like Grh, have been given both repressive and activating roles35,36. For many such factors, it was unclear how one TF could activate some target genes while repressing others at the same time. We show here that Grh merely primes its target regions, making them accessible for other factors to bind. The open enhancers can then be subject to binding by either transcriptional activators or repressors that exert their effects on gene expression, which can explain why Grh removal simultaneously leads to the reduced expression of some target genes while increasing the expression of others.

Mammalian GRHL factors have recently been implicated in a number of human diseases, all involving issues with an epithelial cell fate. In ovarian cancer, GRHL2 expression counteracts epithelial-to-mesenchymal transition (EMT) and increases the overall survival of patients35,36. Mutations in the genes encoding GRHL factors are also linked to pulmonary fibrosis37 and several craniofacial disorders38. We found that mammalian GRHL homologs are required for setting
Fig. 8 | Human GRHLs have properties similar to their Drosophila homolog. a, Most enriched motifs found in the GRHL2 ChIP peaks. b, Seqminer plots visualizing the GRHL2 ChIP signal of six ChIP experiments on ovarian cancer (OVCA3, OVCA429 and PEO1) and airway epithelium (d29, d36 and d44) cell lines; regions were ordered using a ChIP meta-analysis (Methods). c, Predicted nucleosome preference of regions with functional GRHL motifs (red) and regions with nonfunctional GRHL motifs (black), centered on the GRHL motif. d, Seqminer plots visualizing the accessible chromatin of MCF-7 cells in GRHL-bound regions, public DHS data on untreated cells, omni-ATAC on MCF-7 cells 48h after transfection with non-targeting (control) siRNA, and omni-ATAC on MCF-7 cells 48h after transfection with a mix of GRHL1-, GRHL2- and GRHL3-targeting siRNAs (collectively referred to as siGRHL) (omni-ATAC-seq experiments were done once for each sample). e, Omni-ATAC-seq tracks of the MCF-7 cell line, treated with non-targeting (control) siRNA (black) and 48h after treatment with GRHL1-, GRHL2- and GRHL3-targeting shRNAs (gray) (omni-ATAC-seq experiments were done once for each sample). Predicted GRHL target regions (red) near two genes with epithelial expression, PCDH1 and SPINT1, are shown.

up the accessible epithelial chromatin landscape and that the biochemistry of GRHL binding and nucleosome displacement is highly analogous to that in Drosophila. Our findings could help provide better understanding of the role of Grh (and GRHLs) in human disease. In conclusion, we set out to determine how sequence information is linked to chromatin accessibility and enhancer activity in vivo. We identified Grh as a pioneer factor of the epithelial chromatin landscape in Drosophila, with likely conserved roles across animals. Furthermore, our results support a model in which pioneer factors, such as Grh, sit at the top of regulatory hierarchies, establishing tissue-specific accessible chromatin landscapes upon which other factors can act.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0140-x.

Received: 26 July 2017; Accepted: 6 April 2018;
Published online: 4 June 2018

References

1. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* 28, 1045–1048 (2010).
2. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636–640 (2004).
3. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515, 355–364 (2014).
4. Robertson, G. et al. Genome-wide profiles of STAT1–DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4, 651–657 (2007).
5. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* 2010, pdb.prot5384 (2010).
6. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. F. 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).
7. Li, Y. & Tollefsbol, T. O. DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol. Biol.* 791, 11–21 (2011).
8. Davidson, E. H. Emerging properties of animal gene regulatory networks. *Nature* 468, 911–920 (2010).
9. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13, 613–626 (2012).
10. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074–1077 (2013).
11. Melnikov, A. et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30, 271–277 (2012).
12. Patwardhan, R. P. et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* 30, 265–270 (2012).
13. Kvon, E. Z. et al. Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* 512, 91–95 (2014).
14. Pfeiffer, B. D. et al. Tools for neuroanatomy and neurogenetics in Drosophila. *Proc. Natl. Acad. Sci. USA* 105, 9715–9720 (2008).
15. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: intelligent nucleosomes to initiate reprogramming. *J. Cell. Biochem.* 94, 890–898 (2005).
16. Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.* 43, 73–81 (2017).
17. Shlyueva, D., Stampfle, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272–286 (2014).
18. Soufi, A. et al. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* 161, 555–568 (2015).
19. Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming. *Genes Dev.* 28, 2679–2692 (2014).
20. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* 25, 2227–2241 (2011).
21. Barozzi, I. et al. Co-regulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Mol. Cell* 54, 844–857 (2014).
22. Younger, S. T. & Rinn, J. L. p53 regulates enhancer accessibility and activity in response to DNA damage. *Nucleic Acids Res.* 45, 9889–9900 (2017).
23. Zhang, S. & Cui, W. SOX2, a key factor in the regulation of pluripotency and neural differentiation. *World J. Stem Cells* 6, 303–311 (2014).

24. Verfaillie, A. et al. Multiplex enhancer-reporter assays uncover unsophisticated Z. A. enhancer logic. *Genome Res.* 26, 882–895 (2016).

25. Boyer, L. A. et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* 122, 947–956 (2005).

26. Liang, H.-L. et al. The zinc-finger protein Zeda is a key activator of the early zygotic genome in *Drosophila*. *Nature* 456, 400–403 (2008).

27. Foer, S. M. et al. Zeda potentiates morula gene expression by increasing chromatin accessibility. *Curr. Biol.* 24, 1341–1346 (2014).

28. Mackay, T. F. C. et al. The *Drosophila melanogaster* Genetic Reference Panel. *Nature* 482, 173–178 (2012).

29. Huang, W. et al. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel genomes. *Genome Res.* 24, 1193–1208 (2014).

30. Chen, X., Rahman, R., Guo, F. & Rosbash, M. Genome-wide identification of neuronal-activity-regulated genes in *Drosophila*. *eLife* 5, e19942 (2016).

31. Degner, J. F. et al. DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature* 482, 390–394 (2012).

32. Venkatesan, K., Maniatis, H. R., Mello, C. C., Smith, T. F. & Hansen, U. Functional conservation between members of an ancient duplicated transcription factor family, LSFI (grainy head). *Nucleic Acids Res.* 31, 4304–4316 (2003).

33. Paré, A., Kim, M., Juarez, M. T., Brody, S. & McGinnis, W. The functions of grainy-head-like proteins in animals and fungi and the evolution of apical extracellular barriers. *PLoS One* 7, e36254 (2012).

34. Narasimha, M., Uv, A., Krejci, A., Brown, N. H. & Bray, S. J. Grainy head promotes expression of septate junction proteins and influences epithelial morphogenesis. *J. Cell Sci.* 121, 747–752 (2008).

35. Nevil, M., Bondra, E. R., Schulz, K. N., Kaplan, T. & Harrison, M. M. Stalactite binding of a conserved transcription factor grainy head to the target genes throughout *Drosophila melanogaster* development. *Genetics* 205, 605–620 (2017).

36. Varma, S. et al. The transcription factors Grainyhead-like 2 and NK2-homeobox 1 form a regulatory loop that coordinates lung epithelial cell morphology and differentiation. *J. Biol. Chem.* 287, 37282–37295 (2012).

37. Lyons, R. et al. FlyMine: an integrated database for *Drosophila* and Anopheles genomics. *Genome Biol.* 13, R29 (2012).

38. Schmidt, C., Rendeiro, A. F., Sheffield, N. C. & Bock, C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat. Methods* 12, 96–96 (2015).

39. modENCODE Consortium. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330, 1787–1797 (2010).

40. 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).

41. 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).

42. Imrichová, H., Hulselmans, G., Atak, Z. K., Potier, D. & Aerts, S. i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nature* 523, W57–W64 (2015).

43. Mace, K. A., Pearson, J. C. & McGinnis, W. An epidermal barrier wound repair pathway in *Drosophila* is mediated by grainy head. *Science* 308, 381–385 (2005).

44. Wang, S. et al. The tyrosine kinase Stitchever activates grainy head and Zelda to compete for binding to the promoters of the earliest-expressed *Drosophila* genes. *Dev. Biol.* 345, 248–255 (2010).

45. Ye, T. et al. seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.* 39, e35 (2011).

Acknowledgements
We would like to thank F. Casares for helpful discussions, L. Vandecroek (KU Leuven) for sharing the DGPRI lines with us, S. Bray (University of Cambridge) for sharing the USP5-grh line and for insightful discussions, and M. Harrison (UW School of Medicine and Public Health) for sharing antibody to the C terminus of Grh. Stocks obtained from the Bloomington *Drosophila* Stock Center (NIH grant no. P40OD018537) were used in this study. Computing was performed at the Flemish Supercomputer Center (VSC). This work was supported by funding from PWO project grants (grant no. G.0640.13 (S. Aerts), G.0791.14 (S. Aerts), and G.0954.16 N (G. Halder)), Special Research Fund (BOF) KU Leuven grants (grant no. PF/10/016 and OT/13/103; to S. Aerts), the Foundation Against Cancer (grant no. 2012-F2 and 2016-070; to S. Aerts), an ERC CoG grant (724226, cis-CONTROL; S. Aerts), PhD fellowships from the Flemish Agency for Innovation by Science and Technology (J.J., K.D. and H.J.) and a postdoctoral research fellowship from Komo op tegen Kanker (Stand up to Cancer), the Flemish Cancer Society (IJW).

Author contributions
J.J. and S. Aerts conceived and designed the experiments; J.J. and V.C. performed all of the bulk ATAC-seq analyses and generated flies; K.D. and V.C. performed single-cell ATAC-seq and sorted ATAC-seq; D.P. and V.C. performed ATAC-seq on the *Drosophila* species; V.C. performed Grh-specific ChIPmentation; L.R., M.A., V.C., D.K. and J.J. performed imaging disc dissections, staining and imaging; J.J. analyzed the data, with assistance from G. Hulselman on the event selection and BLS, assistance from I.T.T. and G.P. on the random forest analyses, assistance from C.B.G.-B. and K.D. on the single-cell analysis and assistance from S. Aibar on DNA footprinting; J.W. performed omni-ATAC-seq on MCF-7 cells and GRHL knockdown; H.J. analyzed the human GRHL data; and J.J. and S. Aerts wrote the manuscript, with insightful feedback from M.A. and G. Halder.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0140-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.A.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Methods

Fly husbandry and genotypes. We used the following lines from the Drosophila Genetics Reference Panel (Bloomington numbers): 25174, 25185, 25186, 25187, 25193, 25194, 25199, 25208, 28202, 28222, 28224, 28243, 28260, 28262, 28265, 28275, 29651, 29652, 55026 and 55031. For Grh-specific genetics reference panel (Bloomington numbers): 25174, 25185, 25186, 25187, 25193, 25194, 25199, 25208, 28123, 28153, 28202, 28206, 28222, 28243, 28260, 28262, 28265, 28275, 29651, 29652, 55026 and 55031. For Grh-specific ChIP, we used [w1118]; P{Bac[+ w+DnaT]} w+ mCherry-grh-GFP, Optix2/3 GMR, 28260, 28262, 28265, 28275, 29651, 29652, 55026 and 55031. For Optix2/3 GMR, we used the Optix2/3-GFP transgenic line. To generate the grh− mutant–eye–antennal discs, we recombined flies with the grh^mut1 mutation on the FRT42 chromosome to generate the following stock: ey−/ey−, FRT42 grh^ChR/CyO, mCherry-RFP and crossed them to w; FRT42, ey−; FRT42 cell-let/fishCyO, mCherry-RFP and crossed them to w; FRT42, ey−, FRT42 cell-let/fishCyO, mCherry-RFP. As a preclearing step, 20 µl of protein A−protein G magnetic beads (Merck, 5 µg OpenApp integrated fluidic circuits (IFCs) on the Fluidigm C1 and with no cell washing step.

ChIPmentation. Grh-GFP-expressing (Bloomington stock 42272) third-instar larvae were dissected in ice-cold PBS, and the carcasses with attached discs were fixed for 25 min at room temperature by mixing in 1 ml of cross-linking solution (1.8% formaldehyde, 50 mM Hepes pH 7.9, 1 mM EDTA, 100 mM NaCl). The carcasses were washed, and the eye–antennal imaginal discs were dissected and sonicated until chromatin fragments reached an average size of 500 bp. The sonicated chromatin was centrifuged at top speed for 10 min at 4 °C. As a precloning step, 20 µl of protein A−protein G magnetic beads (Merck, Millipore) was added, incubated for 1 h at 4 °C and removed by centrifugation at 3,000 × g for 2 min. Anti-GFP (ab290, Abcam) was added to ‘fixed’ chromatin aliquots, incubated overnight at 4 °C. Immunocomplexes were recovered by adding protein A−protein G magnetic beads to the sample followed by incubation for 3 h at 4 °C. Magnetic beads with precipitated chromatin were washed. Beads were resuspended in elution buffer, RNase A was added to the immunoprecipitated chromatin and the mixture was incubated for 30 min at 37 °C. The immunoprecipitated DNA was then purified. To incorporate sequencing adaptors, we combined the purified cDNA with 4 µl of Nextera TD buffer (Illumina) and 1 µl of Nextera TN5 enzyme (Illumina) on ice and incubated the sample at 55 °C for 5 min. The tagmented cDNA was purified again on a MinElute column and eluted in 20 µl of EB buffer. To PCR amplify the fragments, we added 25 µl of NEBnext PCR master mix (Biomek) and 5 µl of Nextera primer mix and incubated the sample at 72 °C for 5 min, then at 98 °C for 30 s, followed by 15 cycles of 98 °C for 10 s, 63 °C for 30 s and 72 °C for 3 min. We PCR purified the AMPlicons with 55 µl of AMPure beads (Agencourt).

ATAC-seq and ChIP−seq analysis. Adaptor sequences were trimmed from the raw reads using ‘fastx−mcf’ (default parameters using a list containing common Illumina adaptors). Cleaned reads went to a quality-control step using FastQC from Babraham Bioinformatics. We avoided mapping artifacts, as we mapped the ATAC−seq reads for these 30 DGRP lines (25174_IR, 25185_IR, 25186_IR, 25187_IR, 25193_IR, 25194_IR, 25199_IR, 25208_IR, 28123_IR, 28153_IR, 28202_IR, 28222_IR, 28243_IR, 28260_IR, 28262_IR, 28265_IR, 28275_IR, 29651_IR, 29652_IR, 55026_IR and 55031_IR) and linked each region to the annotated and additionally called SNPs.
for these lines. 297,000 SNPs were assigned to their encompassing region using bedtools.2.26.0 ‘intersectBed’ on the extended vcf file.

For each region, we then had the normalized reads for each of the 30 lines as one vector and all SNPs called inside this region as a binary matrix for the 30 lines (present = 1, absent = 0, unknown = NA). We searched for correlating region–SNP vectors using the GLM function in R (Version 3.2.2; http://stat.ethz.ch/R-manual/R-devel/library/stats/html/glm.html) (accessed 20 July 2017). P values were adjusted using the Benjamini–Hochberg procedure in R. Using an FDR of 0.05, we identified 4,289 highly correlating SNP–region pairs, which we refer to as caQTLs.

Delta motif scores. To single out motifs that correlated significantly with open-chromatin changes, a delta motif score was calculated for each of the 18,832 unique motifs in our collection. The sequence for each of the 2,048 variable regions, which contained at least one caQTL, was extracted by using ‘bedtools getfasta’. Next, we mutated these sequences with their encompassing caQTLs according to their effect on open chromatin by using ‘seqekt mutfa’ . For each of the 2,048 regions, we then had two sequences, one for the accessible chromatin and one for the less accessible (closed) chromatin. Each time, we scored both sequences with the 18,832 motifs by using ‘cluster buster --m 0 --c 0’ and retained (for every motif) the highest delta motif score for that region. For every motif, we calculated the sum of all of these delta scores, and we used these delta motif scores for ranking the region–motif pairs. For every motif, we calculated the sum of all of these delta scores, and we used these delta motif scores for ranking the region–motif pairs.

In this method, used by the DREAM5 consortium87, we obtained a final ranking based on the ChIP signal from all three experiments by generating precision-recall curves and calculating the area under the curve (AUPRC) on 1,300 positives and 4,000 negatives (unbalanced set): 1. No training, random ordering (gray PR curve, AUPRC = 0.5, AUPRC = 0.245, PA = 1, gray); 2. One motif (brown PR curve, AUPRC = 0.613, AUPRC = 0.345, increase versus (1) P = 0.001, brown); 3. Top CRM scores for five Grh motifs (red PR curve, AUPRC = 0.766, AUPRC = 0.55, increase versus (2) P = 0.001, orange); 4. Integrated Cister score for repeats (GA, CAr, CAA, AAA, A, CAr, GA, Car, GA, AAA) over the entire region (612 bp) and closer to the Grh motif (one nucleosome around the motif (306 bp)) (green PR curve, AUPRC = 0.82, AUPRC = 0.634, increase versus (3) P = 0.001, green); 5. GC fraction for each region (one nucleosome (±147 bp) around the Grh motif) (purple PR curve, AUPRC = 0.842, AUPRC = 0.674, increase versus (3) P = 0.001, purple); 6. All previous features (Grh motifs, repeats and GC fraction) combined (blue PR curve, AUPRC = 0.87, AUPRC = 0.736, increase versus (4) P = 0.001, increase versus (3) P = 0.03, blue).

The improvement in AUPRC or AUPRC compared to the random forest model with one feature less was calculated using bootstrapping.

DNA shape and nucleosome predictions. Each time we compared the 1,300 bound Grh motifs with the 4,000 unbound motifs (extended by 1 kb on each side).

For the nucleosome prediction data, we calculated an average score for each base pair for the bound and non-functional Grh motifs. The nucleosome prediction data were obtained from https://genie.weizmann.ac.il/software/nucleo_genomes.html. For human nucleosome prediction, we used the top 1,000 GRHL2-bound regions versus the bottom 1,000 regions obtained after order statistics.

Drosophila species and branch length score. We performed ATAC–seq analysis on the eye–antennal discs of third-instar larvae from ten related Drosophila species: Drosophila simulans, Drosophila sechellia, Drosophila yakuba, Drosophila erecta, Drosophila anasi, Drosophila pseudoobscura, Drosophila persimilis, Drosophila willistoni, Drosophila mojavensis and Drosophila virilis. ATAC–seq reads were mapped to the respective genomes, and peaks were called using MACS292 with standard parameters. We used Kent tools ’liftover –minMatch’ to select the top 10,000 highest-scoring motifs (none overlapping) and used ‘Cister’91, a tool for detecting cis-element clusters with optimized parameters for the repeats (± 0–w 1000), to score the entire sequence (612 bp). We then integrated the Cister output results over the sequence to obtain one score (feature) per repeat element. This way the high variability of repeat lengths was taken into account, such that longer repeat sequences resulted in a higher feature score.

The score for the five Grh motifs (Supplementary Fig. 5) was obtained using ‘Cluster-Buster’87; we used each of the five PWMs to score the Grh sequences and retained the top CRM score.

The GC fraction was calculated by selecting the DNA sequence one nucleosome upstream and downstream of the Grh motif (±147 bp) and simply counting the occurrence of G and C base pairs divided by the total number of base pairs.

The final features that were used to train and evaluate the random forest were as follows: AUROC on 1,300 Grh-bound regions (positives) and 1,300 non-functional (balanced) motifs (negative) and all caQTLs accordingly set as area under the precision-recall curve (AUPRC) on 1,300 positives and 4,000 negatives (unbalanced set):

1. No training, random ordering (gray PR curve, AUPRC = 0.5, AUPRC = 0.245, PA = 1, gray);
2. One motif (brown PR curve, AUPRC = 0.613, AUPRC = 0.345, increase versus (1) P = 0.001, brown);
3. Top CRM scores for five Grh motifs (red PR curve, AUPRC = 0.766, AUPRC = 0.55, increase versus (2) P = 0.001, orange);
4. Integrated Cister score for repeats (GA, CAr, CAA, AAA, A, CAr, GA, Car, GA, AAA) over the entire region (612 bp) and closer to the Grh motif (one nucleosome around the motif (306 bp)) (green PR curve, AUPRC = 0.82, AUPRC = 0.634, increase versus (3) P = 0.001, green);
5. GC fraction for each region (one nucleosome (±147 bp) around the Grh motif) (purple PR curve, AUPRC = 0.842, AUPRC = 0.674, increase versus (3) P = 0.001, purple);
6. All previous features (Grh motifs, repeats and GC fraction) combined (blue PR curve, AUPRC = 0.87, AUPRC = 0.736, increase versus (4) P = 0.001, increase versus (3) P = 0.03, blue).

The improvement in AUROC or AUPRC compared to the random forest model with one feature less was calculated using bootstrapping.
Regions were assigned to neighboring genes (5 kb upstream or intronic), and these gene lists were used in GSEA in generation of transgenic lines. Genomic DNA was extracted from an adult fly from lines 25208, 28123 and 28222. The four enhancer pairs of interest were obtained by genomic PCR from the specific lines, using the primers listed in Supplementary Table 6. PCR products were purified on a 1.2% agarose gel. The correct band was cut out and further purified using the Qagen Gel Extraction kit. Constructs were cloned into an entry vector, using the PENTR/D-TOPO cloning kit following the standard protocol. Plasmids were transformed into chemically competent DH5α cells. Constructs were confirmed by sequencing. In a second cloning step, we set up LR Gateway reactions (Invitrogen) between the entry clone and a modified pSHStinger vector. The final constructs were stably and site-specifically (VEK3721232233) integrated into the D. melanogaster genome by injecting it into embryos (done by GenetiVision), using the PhC31 system.

Activity of Grh enhancers. Drosophila lines that had a ChIP and ATAC peak and a strong Grh motif were selected from the Janelia-Ga4 FlyLight enhancer project (Supplementary Table 3). These lines were crossed to a line with UAS-eGFP, and eye-antennal and wing imaginal discs from third instar larvae were dissected, fixed and stained.

Immunohistochemistry and image analysis. Imaginal eye and wing discs from third instar larvae were fixed in 5% formaldehyde at room temperature for 30 min. Next, they were washed in PBT (PBS + 0.3% Triton X-100) and blocked in 5% normal goat serum (NGS) at room temperature. To test Grh enhancer activity, discs were incubated with a mix of the primary antibodies rabbit anti-Grh (Invitrogen) and rat anti-Elav (DSHB, 7E8A10, 1:600) at 4°C overnight. Cy3-conjugated 488 donkey anti-rabbit 647 (Invitrogen/Life Technologies, A31573, lot 1826679) secondary antibody (Molecular Probes), 555 donkey anti-rat Cy3 (Jackson, 712-166-153, 1:600) and DAPI (1:1,000) were added for 2 h at room temperature, and the samples were then washed with PBNT, PBT and PBS (3 × 10 min). Samples were fixed in 4% formaldehyde for 10 min and washed (3 ×) with PBT before mounting the discs in Vectashield (Vector Laboratories). For the Grh-mutant tissue, we used rabbit anti-Grh-Term (1:500) to stain for Grh proteins, which was visualized using 488 donkey anti-rabbit 647 (Invitrogen/Life Technologies, A31573, lot 1826679) secondary antibody. Cleaved Drosophila Dcp-1 (Asp216) antibody (Cell Signalling, 9578, lot 2, 1:150) was used to stain apoptotic cells, which was also visualized using the 488 donkey anti-rabbit 647 secondary antibody.

For imaging, an Olympus FV1200 confocal microscope was used (20× oil, z stack). Image (Bio-formats Importer plug-in) was used to merge and process the images. Quantification of the GFP fraction of cells in the discs was also done in ImageJ, manually selecting the threshold for each disc before measurement.

Knockdown of expression of Grh-like transcription factors in human epithelial cell culture. Human epithelial breast carcinoma MCF-7 cells were cultured at 37°C and 5% CO2 in DMEM (Thermo Fisher Scientific) supplemented with 10% FBS (Lonza) and penicillin (penicillin/streptomycin/gentamicin, Thermo Fisher Scientific). The knockdown of GRHL1, GRHL2 and GRHL3 expression was performed by using a mix of the ON-TARGETplus GRHL1–3 siRNA SMARTpools (Dharmacon) according to a previously published protocol. In Fig. 6e, the parameters used were (n = 6,390.7, true difference in means not equal to 0, P = 1.76 × 10−10, t = 5.31, df = 4,692, true difference in means not equal to 0, P = 7.22 × 10−10) by a two-sided Fisher’s exact test (MCF-7: 15,555 accessible regions without a GRHL motif, of which 265 lost accessibility after GRHL knockdown; 4,017 with a GRHL motif, of which 283 lost accessibility after GRHL knockdown).

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

Data availability. The UCSC Genome browser hub with ATAC-seq tracks for all Drosophila lines-antennal and wing disc ATAC-seq of the animals with stably integrated enhancer reporter constructs (Fig. 3) and ChIPmentation against Grh-GFP in eye-antennal and wing discs can be found at ucsctracks.aertslab.org/users/jacobcs/DGRP_injections/big Wig.norm/hub norm.txt, as well as a hub with all cross-species ATAC-seq (evolution hub) at http://ucsctracks.aertslab.org/papers/evolution/hub.txt. To load these in the UCSC Genome Browser, go to “My Data—Track Hubs.” The raw and processed data are available from GEO under accession GSE102441.

References
70. 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).
71. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490 (2015).
72. Gramates, L. S. et al. FlyBase at 25: looking to the future. Nucleic Acids Res. 45, D663–D671 (2017).
73. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
74. Li, H. Seqtk: toolkit for processing sequences in FASTA/Q formats. https://github.com/lh3/seqtk (2017).
75. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
76. Daneczek, P. et al. The variant cell call and VCFtools. Bioinformatics 27, 2156–2158 (2011).
77. Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). Genome Biol. 9, R137 (2008).
78. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 924–930 (2014).
79. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).
80. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589 (2010).
81. Thomas-Chollier, M. et al. RSAT peak-motif motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res. 40, e31 (2012).
82. 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, 973–978 (2017).
83. Wei, T. et al. corplot: visualization of a correlation matrix. https://github.com/taiyun/corplot (2017).
84. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics Chapter 4, Unit 4.10 (2009).
85. Quinlan, A. R. BEDTools: the Swiss-army knife for genome feature analysis. Curr. Protoc. Bioinformatics 47, 11.2.1–11.2.34 (2014).
86. Fisher, R. A. The logic of inductive inference. J. R. Stat. Soc. 98, 39–82 (1935).
87. Weihrich, M. T. et al. Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. 31, 126–134 (2013).
88. Robin, X. et al. pROC: an open-source package for R and S-4 to analyze and compare ROC curves. BMC Bioinformatics 12, 77 (2011).
89. 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).
90. Pedregosa, F. et al. Scikit-learn: machine-learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
91. Frith, M. C., Hansen, U. & Weng, Z. Detection of cis-element clusters in higher-eukaryotic DNA. Bioinformatics 17, 878–889 (2001).
92. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277 (2000).
93. Mahony, S. & Benos, P. V. STAMP: a web tool for exploring DNA-binding motifs from ChIP-seq data. Bioinformatics 25, 45–52 (2009).
94. van Bergeijk, P., Heimiller, J., Uytarte, L. & Su, T. T. Genome-wide expression analysis identifies a modulator of ionizing-radiation-induced p53-independent apoptosis in Drosophila melanogaster. PLoS One 7, e36539 (2012).
95. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550 (2005).
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).
Experimental design

1. Sample size
   Describe how sample size was determined.
   We used a Generalized Linear Model to link SNP's inside an accessible region with the variability in accessibility of this region. We were mainly interested in the direct effect of SNPs inside an accessible region (we wanted to identify key transcription factor motifs important for our epithelial tissues). In a first round we performed ATAC-seq on 15 DGRP lines and could already identify the Grainyhead motif as being highly and significantly concordant with accessible chromatin. We decided to increase our sample size to 23 DGRP lines and added 2 additional replicates for 7 lines. As expected the GLM now detected more significantly correlating SNP-region pairs, but the overall conclusions didn’t change. This is why we decided this sample size was sufficient for our research goals. For all other experiments where we had to compare one group versus another group we always had at least two biological replicates.

2. Data exclusions
   Describe any data exclusions.
   No data was excluded from the analyses.

3. Replication
   Describe whether the experimental findings were reliably reproduced.
   No attempts of replication failed - 7 lines were replicated in ATAC-seq; ChIP-seq was replicated using two techniques (ChIP-seq and ChIPmentation), and two epithelial tissues (eye and wing disc). ATAC-seq on reporter lines was replicated four times, once for each distinct reporter line, each time both the active and inactive enhancer. The ectopic expression of Grh was done 2 times.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   This is not relevant, as groups are determined by the presence or absence of a SNP in a regulatory region according to caQTL / eQTL analysis. We did however control our branch length score for each motif by calculating a BLS score on the real sequences and on shuffled sequences. We then subtracted the shuffled scores from the real scores. We also controlled for the “delta motif” scores calculated over all caQTLs for each motif, by performing the same procedure on 66k non-correlating SNP and calculating the enrichment for each motif using fisher’s exact test.

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   Blinding was not necessary as the computational analysis is fully reproducible. For the Random forest evaluation we did train on one half of the data and evaluated the other half.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.
6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

```
|   | Confirmed |
|---|-----------|
| n/a |           |
| The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) | ☑ |
| A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly. | ☑ |
| A statement indicating how many times each experiment was replicated | ☑ |
| The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) | ☑ |
| A description of any assumptions or corrections, such as an adjustment for multiple comparisons | ☑ |
| The test results (e.g. p values) given as exact values whenever possible and with confidence intervals noted | ☑ |
| A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range) | ☑ |
| Clearly defined error bars | ☑ |
```

See the web collection on statistics for biologists for further resources and guidance.

7. Software

Describe the software used to analyze the data in this study.

- fastq-mcf (default parameters using a list containing the common Illumina adapters) to trim adapter sequences from the raw reads.
- FastQC from Babraham Bioinformatics for quality control of trimmed reads.
- bowtie2 to map the reads to the Drosophila melanogaster Flybase r5.13 genome.
- Samtools to sort and index the mapped reads.
- seqtk mutfa to include the SNPs called for each DGRP line into the fasta sequence.
- SAMTools mpileup -B –f r5.13.fasta DGRP_lineX.bam | varscan.sh
- mpileup2snp --output-vcf 1. To call additional SNPs in the mapped ATAC reads.
- VCFtools to add the newly called homozygous SNPs to the original vcf file for each DGRP line.
- macs2 callpeak -g dm –nomodel (for ATAC only) –keep-dup all –call-summits for peak calling on the mapped ATAC-seq and ChiP-seq reads.
- featureCounts to count the number of reads falling into each accessible region.
- DESeq2 from Bioconductor used in R-studio version 3.2.2. to further process the raw counts matrices and to obtain the differential accessible regions sets, following the standard procedure for differential mRNA.
- picard tools markDuplicates to deduplicated the ScATAC data.
- ChromVAR to quantify the number of fragments falling within each of the accessible regions.
- intersectBed to identify overlapping regions between bed files.
- Generalized Linear Model function (GLM) in R (version 3.2.2) to identify correlating region-SNP pairs and variable regions.
- bedtools getfasta to extract fasta sequences
- cluster buster to score a fasta sequence with a position weight matrix from a motif.
- i-cisTarget, an online, sequence based motif enrichment tool that takes conservation into account, to identify motifs that were enriched in our sets of accessible regions
- bedtools multiBamCov to obtain the number of mapped reads in a certain region (BED).
- Seqminer to plot heatmaps that visualize the coverage over a set of regions.
- Random Forest classifier from the scikit-learn Python package (ensemble learning algorithm).
of 151 decision trees) to discriminate bound from unbound Grainyhead regions.

RSAT peak motif, an online tool to discover oligo’s, motifs and dyads that were significantly enriched in the bound versus the unbound Grainyhead regions.

RSAT-matrix-clustering, to cluster the co-conserved motifs into 23 distinct motif clusters.

Cister, a tool for detecting cis-elements clusters with optimized parameters for repeats.

Kent tools liftover –minMatch=0.1, to obtain the bed coordinates of the Grainyhead enhancers in other species.

shufflesseq from the EMBOSS package to shuffle fasta sequences.

Gene Set Enrichment Analysis, to identify enriched gene sets over a specific ranking.

ggplot2 in R (version 3.2.2) to make the graphs.

Adobe illustrator CS6 and Microsoft Power Point v15.36 to generate the figures.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The Nature Methods guidance for providing algorithms and software for publication may be useful for any submission.

### Materials and reagents

#### Policy information about availability of materials

**8. Materials availability**

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

The Grainyhead c-terminal antibody was from a limited batch, kindly send to us by Melissa Harrison (Department of Biomolecular Chemistry, UW School of Medicine and Public Health)

**9. Antibodies**

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

Primary antibodies:
For Grh-GFP-Chipmentation: anti-GFP Ab -ChIP grade (ab290, Abcam), 10 μg anti-GFP antibody per 200 eye-antennal discs, we analyzed the ChiP peaks and found the Grh motif as strongest enriched.

Stainings:
rabbit anti-GFP (Invitrogen)
Rat-Elav-7E8A10 anti-elav (DSHB, 1:600), Depositors Notes: This antibody is a marker for most differentiated neuronal cells in the central and peripheral nervous system. Validated in Drosophila: The activities of two Ets-related transcription factors required for Drosophila eye development are modulated by the Ras/MAPK pathway. O’Neill EM, Rebay I, Tjian R, Rubin GM Cell 1994 Jul 15.

Cleaved Drosophila Dcp-1 (Asp216) Antibody #9578 Lot.2 (cell Signalling, 1:150), validated in several studies in Drosophila. e.g. Drosophila Wnt and STAT Define Apoptosis-Resistant Epithelial Cells for Tissue Regeneration after Irradiation (PloS Biology).

Rabbit-anti-Grh-Cterm (1:500) (validated in. Grainyhead and Zelda compete for binding to the promoters of the earliest-expressed Drosophila genes. Dev. Biol. 345, 248–255 (2010).) was stained and visualized using Alexa647 conjugated Donkey-anti-Rabbit secondary antibody (Molecular probes).

The secondary antibodies:
488 donkey anti -Rabbit 647 Invitrogen/Life technologies, A31573, Lot 1826679.

555 donkey anti-Rat Cy3 Jackson 712-166-153  (1:600)
Nuclei were stained using DAPI (1:1000)
Data provided in the manuscript
10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used. Human epithelial breast carcinoma MCF7 cell line was used from ATCC.

b. Describe the method of cell line authentication used. Besides the correct epithelial morphology (pictures) of the MCF7 cultures, we also confirmed that the SNPs between our ATAC-seq data and published DHS data from encode on MCF7 match and used this information to authenticate our cell line.

c. Report whether the cell lines were tested for mycoplasma contamination. We used our ATAC-seq data as a proxy to test whether our cells were infected by mycoplasma. (we had done other ATAC experiments in the lab, where the cell lines did contain mycoplasma and as a result only ~10% of the sequenced reads mapped to the human genome, as the majority of reads mapped to mycoplasma). When we mapped the ATAC-seq reads from our MCF7 experiment, 95.6% and 85.5% of the reads mapped to the human genome, indicating that these samples were not contaminated by mycoplasma.

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use. no commonly misidentified cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study. Tissues from 11 Drosophila species were used; Melanogaster, Simulans, Sechellia, Yakuba, Erecta, Anasai, Pseudoobscura, Persimilis, Wilistoni, Mojavensis and Virilis.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants. The study did not involve human research participants.