Optimal transport analysis of single-cell transcriptomics directs hypotheses prioritization and validation

Rohit Singh[1#]*, Joshua Shing Shun Li[2]*, Sudhir Gopal Tattikota[2]*, Yifang Liu[2], Jun Xu[2], Yanhui Hu[2], Norbert Perrimon[2,3#], Bonnie Berger[1,4#]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [2]Department of Genetics, Blavatnik Institute, Harvard Medical School, Harvard University, Boston, MA USA; [3]Howard Hughes Medical Institute, Boston, MA 02115, USA; [4]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

*contributed equally

#correspondence: rsingh@csail.mit.edu; perrimon@genetics.med.harvard.edu; bab@mit.edu

Abstract

The explosive growth of regulatory hypotheses from single-cell datasets demands accurate prioritization of hypotheses for *in vivo* validation, but current computational methods fail to shortlist a high-confidence subset that can be feasibly tested. We present Haystack, an algorithm that combines active learning and optimal transport theory to identify and prioritize causally-active transcription factors in cell lineages. We apply Haystack to single-cell observations, guiding efficient and cost-effective *in vivo* validations that reveal causal mechanisms of cell differentiation in *Drosophila* gut and blood lineages.

Main Text

Temporal control of lineage-determining transcription factors (TFs) is crucial to tissue development and homeostasis. Single-cell transcriptomics (scRNA-seq) enable an unprecedented reconstruction of cell lineages in a pseudotemporal manner which can then be correlated with each TF's expression to hypothesize the regulators of cell differentiation. However, causality can only be hypothesized, but not confirmed from observational scRNA-seq studies[1]. The perturbational experiments required to test these hypotheses remain slow and expensive (**Supplementary Note 1**), even as high-throughput scRNA-seq predictions of causally-active TFs become available.

Hypothesis prioritization approaches are thus crucial in allocating downstream validation efforts to a small set of high-confidence regulatory hypotheses, but existing scRNA-seq analysis approaches are not selective enough to generate such hypotheses. Typically, these infer gene regulatory networks by correlating one TF's expression with another, with some methods also incorporating pseudotime[2–4]. Prioritizing high-confidence causal TFs from these methods is challenging because they quantify a TF's activity solely from its transcript counts and are vulnerable to the noise and sparsity of scRNA-seq data[5].

We present a hybrid computational-experimental method, Haystack, to identify TFs involved in differentiation from scRNA-seq datasets. TFs are computationally prioritized and then validated in an active learning framework[6] that iteratively prunes validation targets (**Fig. 1a**). Towards accurate and robust TF prioritization, the key conceptual advance of our work is leveraging optimal transport (OT) to synthesize two previously disparate views of regulatory activity: one that relates a TF to its transcriptional targets and the other that considers expression changes over differentiation to estimate gene regulatory networks. We begin by computing the TF modules (SCENIC[7] regulons) of a scRNA-seq dataset and map them to the

pseudotime landscape. Each regulon relates to the activity of one TF, quantified in terms of its own expression and its putative targets, thus producing a more robust estimate of TF activity than using transcript-counts alone. We then pursue the intuition that lineage-determining TF modules should be active primarily in undifferentiated cells with targets active in differentiated cells.

We frame the challenge of accurately and robustly characterizing TF activation profiles along a pseudotime trajectory as an optimal transport problem. Transport theory calculates the optimal coupling of two probability distributions under a cost function, with OT-based metrics having emerged as powerful approaches to compare irregular and noisy datasets. For example, Schiebinger et al.[8] temporally-coupled separate scRNA-seq datasets using OT with a cost metric defined by whole transcriptome changes. In our context, we model individual TF activity in single lineages as a probability distribution over pseudotime. The OT distance (over the one-dimensional pseudotime axis) between each TF distribution and the baseline, or between two TFs, is indicative of the location and concentration of a TF's activity during differentiation (**Methods**). For example, TFs localized to a specific differentiation state will have a large OT distance from the baseline probability distribution. Our OT-based metric offers the advantage of enabling non-parametric testing, thereby not assuming an underlying distribution (unlike the t-test). It therefore captures not only the differences of means but also the higher moments. Finally, measuring TF activity along pseudotime and not in discrete cell clusters means that our approach does not rely on cell-type annotations.

To improve robustness and precision, we incorporate additional analyses and public epigenetic data. The OT-based TF scores are ensembled with a Schema-based[9] feature-selection analysis that prioritizes TFs most predictive of a cell's pseudotime score. The latter addresses potential false positives where a non-localized bimodal TF distribution may have high OT scores. The combined TF rankings are further refined by considering potential source-target

TF pairs, i.e. where the binding site (from the cisTarget database[7]) of TF_{source} is upstream of TF_{target}'s genomic locus and the former is active earlier in the pseudotime landscape. The OT distance between each putative pair is computed, and the source TFs corresponding to the highest OT distances are selected as the initial set of high-confidence hypotheses for experimental validation— these are the TFs expected to be active in progenitor cells, regulating the differentiation into terminal lineages.

We applied our OT prioritization scheme on scRNA-seq studies of mouse intestine[10] (**Fig. 1b,c**) and human leukemia cells[11] (**Fig. 1e,f**), finding that it identified biologically relevant TFs (**Fig. 1d,g**). For each dataset, we aggregated the top-scoring source TF modules across lineages and found that 75-85% of the prioritized TFs have substantial support in published literature for their involvement in progenitor differentiation (**Table S1,S2; Supplementary Note 2**). Previously, several methods[2–4] for gene network inference that incorporate pseudotime have considered individual TFs' transcript counts (which can be noisy) while methods that infer TF activity more robustly (e.g. SCENIC[7])have not considered the differentiation landscape. Haystack benefits by combining these hitherto disconnected views. As quantified by a text-mining based gene enrichment metric (**Table S4; Supplementary Note 2**), Haystack achieved substantially higher precision than existing methods (**Fig. S1**).

In *Drosophila*, where predictions can be readily tested with genetic perturbations, we experimentally validated Haystack (**Supplementary Note 3**) on scRNA-seq datasets of fly gut (**Fig. 2a,b,S3**) and blood (**Fig. 2g,h,S4**)[12–14]. We take an active learning approach, starting with a preliminary assay whose results guide further validations. Specifically, the initial OT-based shortlist of TFs is investigated with qRT-PCR of cell-type markers, with changes in cell-type composition estimated by a novel aggregate-fold-change metric (**Methods**). TFs with supporting evidence are then assayed with microscopy (**Fig. 1a**). Haystack identified both known and novel TFs of biological relevance in *Drosophila* (**Fig. 2c,i**). In particular, validation in the fly gut

resolved previous conflicting observations of *peb* roles, where two independent groups had identified *peb* to play opposite functions in the fly midgut[15,16]. Consistent with *Baechler et al.*, our iterative validation found *peb* as a driver of enterocyte differentiation (**Fig. 2d-f**). Towards regulatory cascade discovery, our method made accurate source-target TF predictions and suggested that *peb*-dependent enterocyte differentiation may be mediated via *Myc* (**Fig. S3**). In the fly blood, Haystack predictions were successfully validated in multiple lineages (**Fig. S4**), where we focused in particular on the poorly studied lamellocyte (LM) lineage, identifying *Xbp1* and *CG3328* as novel regulators of LM differentiation (**Fig. 2h-l**). Altogether, Haystack can reliably be applied to various scRNA-seq datasets across species to shed light on novel and high-confidence determinants of cell lineages.

Single-cell genomics has led to an exponential growth of data with computational analyses that generate a multitude of hypotheses[17]. However, more data does not imply more insight— our knowledge of the causal mechanisms of cell development has not kept pace with such data explosion. Currently, a key bottleneck in research progress is validating causal hypotheses generated from observational scRNA-seq studies. The active learning framework of Haystack represents a general, principled approach to this challenge: a combination of robust inference, hypothesis prioritization, and iterative experimentation can efficiently discover regulatory mechanisms in diverse biological systems.

Methods

Trajectory analysis and transcription factor (TF) module identification

Trajectory analysis and pseudotime computation of the scRNA-seq data was used to estimate the differentiation time-course; in the case of multiple branching trajectories, we repeat our analysis for each branch separately and aggregate the results. Haystack can be applied with a preferred pseudotime (or RNA velocity) estimation program; here we present results with SlingShot[18] , diffusion pseudotime[19], and Monocle 3[20], choosing the method used in the original scRNA-seq study whenever available. The optimal transport analysis in Haystack does not require any knowledge of the cell types in the dataset and the differentiation stage of a cell is inferred solely from pseudotime analysis. When cell types are known, they may be used to limit Haystack analysis to a lineage of interest. TF regulons are mapped over the differentiation time-course, where regulons are inferred from the cisTarget database of TF binding sites.

Characterizing TF localization through optimal transport

To identify TFs that are active in just one differentiation stage and not broadly active, we arrange cells along a one-dimensional pseudotime axis (**Fig. 1a**). Each TF activity is characterized as a probability distribution along the pseudotime axis, computed from the histogram of per-cell regulon activity indicators (for each regulon, SCENIC[7] reports the cells with statistically significant activity of the regulon). We also compute the baseline probability distribution (i.e. histogram) of all cells along this axis. Using optimal transport (OT), we compute for each TF the distance between its distribution and the baseline. OT is a mathematical formulation for measuring the distance between two probability distributions under some cost function. In our context, the cost between two cells corresponds to the difference of their score along pseudotime axis, capturing the distance between their differentiation stages; other

measures like RNA velocity could be similarly incorporated. Intuitively, TFs that are active only in a limited region of the time-course will be at a substantial OT distance from the baseline. We note that the OT metric captures not only the differences of mean between distributions (first moment) but also differences in higher moments: e.g. even if the probability distribution of a TF has the same mean as the baseline, if the former is concentrated around this location (i.e. has a lower variance than the baseline), the OT distance can be large. The use of OT offers key advantages over alternative approaches. In OT, the empirical probability distribution of TFs over cells is directly computed upon, so we do not need to assume that the underlying distribution obeys specific properties (e.g. those required by statistical tests like the t-test). Unlike some metrics over probability distributions (e.g. mutual information or Jensen-Shannon distance)[21], the OT formulation incorporates the concept of cost which allows us to account for the pseudotemporal distance between two TFs (or between a TF and the background).

If a TF has high activity in two separate pseudotime regions distinct from the baseline (e.g. at both the start and end of the time-course), the OT metric may also score such non-localized bimodal distributions highly. To address such potential false positives, we incorporate an additional measure of TF localization: we represent each cell as feature-vector of TF activations and identify the features (i.e. TFs) that are most informative of the cell's pseudotime score. We built upon Schema[9], a metric learning approach for feature selection in multimodal single-cell data, to solve a quadratic program to identify TFs whose activation is most informative of the cell's pseudotime score. We ensembled the two approaches by converting their outputs to rank-scores of TFs and computing a weighted combination of the two. Small-scale explorations indicated that results were robust to the choice of weight around 0.5, which we have chosen for all results presented here. We then limited our analysis to the top-third of the TFs.

Identifying lineage-determining TFs

Lineage-determining TFs typically have more than one downstream target[22], with their ability to influence a broad transcriptional program being key to fate determination. Accordingly, we queried the cisTarget database[23] to identify TF pairs (TF_{source}–TF_{target}) where the binding site of TF_{source} was found upstream of TF_{target}, suggesting that TF_{source} might regulate TF_{target}. From the shortlist of well-localized TFs, we considered every pairwise combination of TFs (say, TF_1 and TF_2) such that TF_1 is active earlier in the pseudotime landscape than TF_2 and with the binding site of TF_1 upstream of TF_2. We then applied OT to compute the pseudotime distance between the two TFs. From these pairs, we extracted the subset corresponding to high OT distance. The source TFs in these pairs are the candidate TFs we generate as the initial set of high-confidence hypotheses prioritized for experimental validation. Within this set, we rank TFs by the number of well-separated pairs in which the TF occurs as a source.

Estimating cell-type decomposition from qRT-PCR assays of marker genes

Quantitative Real Time Polymerase Chain Reaction (qRT-PCR)-based validation provides a medium-throughput validation that is more efficient than imaging and phenotypic studies. For each of the shortlisted TFs, we perform perturbation experiments (e.g., for *Drosophila* gut and blood studies, we overexpressed or knocked down the TFs of interest). From the original scRNA-seq study, we identified the cell types/clusters of interest and applied differential expression analysis to identify a limited set of markers (typically, 1-3) per cell type. In our *Drosophila* assays, this resulted in less than 15 markers in total, and thus amenable to a single qRT-PCR study. We assayed these markers to assess changes in cell type composition as a result of the perturbation, by comparing qRT-PCR values in the wild type tissue against the overexpressed/knocked-down tissue. Furthermore, since the markers are the same for each

perturbation, the same qRT-PCR primers and setup can be used across all perturbations in parallel, speeding up the process.

A confounding issue in estimating cell-type composition changes using qRT-PCR is that a perturbation may vary the overall proliferation rate of the tissue vis-a-vis the organismal background. Since qRT-PCR cycle threshold (CT) values are typically normalized against the background CT value of a housekeeping or ribosomal gene, this can confound cell type composition analysis. Accordingly, we introduce a fold-change metric to adjust for this confounding factor and robustly recover estimates of cell type compositions: we first compute robust estimates of cell type expression by averaging the markers for each cell type. We then choose one cell type's abundance as the baseline (typically, the progenitor cell type), and express all other cell type abundances as a ratio against this baseline. Wild-type and perturbation lines can now be compared using a fold-change metric on this ratio to identify which, if any, differentiated cell types have increased or decreased.

Drosophila **stocks and culture**

Flies were reared in humidified incubators at 25°C on standard lab food composed of 15 g yeast, 8.6 g soy flour, 63 g corn flour, 5 g agar, 5 g malt, 74 ml corn syrup per liter with 12/12 hr dark/light cycles. For all Gal80^{ts} experiments, crosses were reared at 18°C. After eclosion, flies were kept at 18°C for 3 days before shifting to 29°C (permissive temperature) for 10 days. For all blood experiments, fly larvae of respective genotypes were grown on the standard lab food until late third larval instar (LL3) at 25°C.

The following stocks were obtained from the Bloomington *Drosophila* Stock Center (BL), DGRC (NIG) and FlyORF: *UAS-Luc-i* (BL36303) *UAS-peb-i* (BL28735), *UAS-peb* (BL5358), *UAS-Psi-i*

(BL31301), *UAS-Psi* (BL16371), *UAS-drm* (BL7072), *UAS-Tet-i* (BL62280), *UAS-Mondo-i* (BL27059), *UAS-Mondo* (BL20102), *UAS-cnc* (BL17502), *UAS-Tet-i* (BL62280), *UAS-tbp-i* (NIG 9874R-1), *UAS-FoxK* (F000615), *UAS-Mondo* (F001398), *UAS-Xbp1* (BL60730), *UAS-E(spl)mbeta-HLH* (BL26675), *UAS-CG3328-i* (BL55211) and *UAS-Xbp1-i* (BL36755). The following Gal4 lines used to perturb genes in guts and hemocytes, respectively, were: *esg-Gal4* and *w[1118];Hml-Gal4.Delta,UAS-2xEGFP* (BL30140), hereafter referred to as *HmlGFP*. *Oregon R* (*OreR*) control flies were obtained from the Perrimon Lab stock. The *BcF6-mCherry* (a crystal cell reporter)[24] stock obtained from Dr. Tsuyoshi Tokusumi (Schulz Lab), was crossed with the *HmlGFP* line to obtain *HmlGFP;BcF6-mCh* stock. To drive the expression of *CG3328* in blood cells in a Cas9-based transcriptional activation (*CRISPR*[a]) manner[25], the *Hml-Gal4,UAS-EGFP;dCas9-VPR* (*HmlGFP;dCas9-VPR*) was crossed to *CG3328-sgRNA* fly line (BL80297).

RNA extraction and qRT-PCR

<u>*Drosophila* midguts</u>: 7-10 midguts were dissected in 1xPBS and homogenized in 300uL of TRIzol (ThermoFischer, cat# 15596-026) using RNase-free pestles. RNA was extracted using Zymo Direct-zol RNA MicroPrep kit (cat# R2060) and subsequently DNase-treated using Turbo DNA free (cat# AM1907). 400-450ng of the resulting RNA was reverse transcribed using Bio-Rad iScript Select cDNA synthesis kit (cat# 708896) and SyBr green (cat# 1708880) based qRT-PCR was performed to determine the levels of gene expression. qRT-PCR primers were designed using FlyPrimerBank[26]. The efficiency of primers was determined by running qRT-PCR on serial dilutions of pooled cDNA. Only primers in the range of 85% to 110% efficiency were selected for further use. See Table S3.

<u>*Drosophila* hemocytes</u>: RNA isolation of larval blood was performed as described previously with minor modifications[14], where hemolymph from ~15-20 larvae (or ~50 for a better yield) are

sufficient for RNA isolation per biological replicate. For a detailed protocol regarding hemocyte isolation, RNA/cDNA preparation and qRT-PCR set up, see https://en.bio-protocol.org/prep1155.

Immunostaining and imaging

Drosophila guts: Whole midguts were dissected in PBS and fixed in 4% PFA in PBS at room temperature for 30 minutes. Fixed guts were washed once in 0.1% Triton X-100 in PBS (PBST), then blocked with a blocking buffer (0.1% Triton, 5%NDS in PBS) for 30 minutes at RT. Primary antibodies were incubated overnight at 4°C in the blocking buffer. Guts were washed 3x in the blocking buffer and incubated with secondary antibodies overnight at 4°C along with DAPI (1:1000 of 1mg/ml stock). After antibody staining, guts were washed 3 times in PBST and mounted in Vectashield antifade mounting medium (Vector Laboratories cat# H-1200). Tape was used as a spacer to prevent coverslips from crushing the guts. Antibody dilutions used were as follows: chicken anti-GFP (1:2000, Abcam cat# ab13970), donkey anti-rabbit 565 (1:2000, Molecular Probes cat# A31572), goat anti-mouse 633 (1:2000, Thermo Scientific cat# A-21240) and goat anti-chicken 488 (1:2000, Thermo Fisher Scientific cat# A-11039). Guts were imaged on a spinning-disk confocal system, consisting of a Nikon Ti2 inverted microscope equipped with a W1 spinning disk head (50um pinhole, Yokogawa Corporation of America) and a Zyla 4.2 Plus sCMOS monochrome camera (Andor).

Drosophila hemocytes: 20 late third instar larvae (LL3) from each genotype were vortexed, and bled in 300 ul of Schneider's media in a 9-well spot glass plate. Next, the media with hemocytes was transferred to 8-well chambered cover glass slide (VWR, cat# 62407-296) and the cells were allowed to settle at room temperature for ~30 min. Alternatively, 96-well glass bottom plates (Cellvis, cat# P96-1.5H-N) were also used to plate hemocytes from 10 LL3 larvae per

biological replicate per well. Next, 4% (final concentration) paraformaldehyde (Electron Microscopy Services, cat# 15710) was added to each well with Schneider's media and hemocytes and incubated on a rocker for 20 min. Later, the fixed hemocytes were washed three times with 1x PBS (Gibco, cat# 10010-023) and blocked with a blocking buffer (5% BSA in 1x PBS containing 0.1% Triton-X) for 10 min. The cells were incubated with 1:100 dilution of anti-Atilla L1abc antibody[27] overnight at 4°C. The next day, the cells were washed three times with 1x PBS and incubated with corresponding secondary antibody (1:500 dilution, anti-mouse alexa fluor 633) for 1 h at room temperature. Finally, the cells were washed three times with 1x PBS and Vectashield containing DAPI (Vector Laboratories Inc., cat# H-1200) was added before imaging the cells using Nikon Ti2 Spinning Disk Confocal Microscope. Cell counts were performed on 3-4 independent regions of interest (ROIs) per well (biological replicate) captured by the Nikon Ti2 Spinning Disk Confocal Microscope or GE IN Cell Analyzer 6000 Cell Imaging System. All images were analyzed by Fiji ImageJ software[28].

Code and Data Availability

Python code and instructions for use of the Haystack framework are available at:

https://cb.csail.mit.edu/cb/haystack/

Source data from the imaging and qRT-PCR assays is available upon request.

Acknowledgments

We thank the assistance provided by the Microscopy Resources on the North Quad (MicRoN) core and the GE IN Cell Analyzer 6000 Cell Imaging facility of the *Drosophila* RNAi Screening Center (DRSC) at Harvard Medical School. We thank the members of the Berger and Perrimon labs for helpful discussion and feedback. RS and BB were partially supported by NIH NIGMS R35GM141861. JSSL was supported by the Croucher fellowship for Postdoctoral Research from the Croucher Foundation. This work was supported by NIH NIGMS P41 GM132087 and

vBBSRC-NSF (NP). NP is an investigator of Howard Hughes Medical Institute. This article is subject to HHMI's Open Access to Publication policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

References

1. Gianicolo, E. A. L., Eichler, M., Muensterer, O., Strauch, K. & Blettner, M. Methods for Evaluating Causality in Observational Studies. *Dtsch. Arztebl. Int.* **116**, 101–107 (2020).
2. Matsumoto, H. *et al.* SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**, 2314–2321 (2017).
3. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst* **5**, 251–267.e3 (2017).
4. Deshpande, A., Chu, L.-F., Stewart, R. & Gitter, A. Network inference with Granger causality ensembles on single-cell transcriptomics. *Cell Rep.* **38**, 110333 (2022).
5. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**, 147–154 (2020).
6. Hie, B., Bryson, B. D. & Berger, B. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. *Cell Syst* **11**, 461–477.e9 (2020).
7. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
8. Schiebinger, G., Shu, J., Tabaka, M. & Cleary, B. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* (2019).
9. Singh, R., Hie, B. L., Narayan, A. & Berger, B. Schema: metric learning enables

interpretable synthesis of heterogeneous single-cell modalities. *Genome Biol.* **22**, 131 (2021).

10. Böttcher, A. *et al.* Non-canonical Wnt/PCP signalling regulates intestinal stem cell lineage priming towards enteroendocrine and Paneth cell fates. *Nat. Cell Biol.* **23**, 23–31 (2021).
11. Petti, A. A. *et al.* A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat. Commun.* **10**, 3660 (2019).
12. Hung, R.-J. *et al.* A cell atlas of the adult midgut. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 1514–1523 (2020).
13. Hung, R.-J., Li, J. S. S., Liu, Y. & Perrimon, N. Defining cell types and lineage in the Drosophila midgut using single cell transcriptomics. *Curr Opin Insect Sci* **47**, 12–17 (2021).
14. Tattikota, S. G. *et al.* A single-cell survey of blood. *Elife* **9**, (2020).
15. Zeng, X. *et al.* Genome-wide RNAi screen identifies networks involved in intestinal stem cell regulation in Drosophila. *Cell Rep.* **10**, 1226–1238 (2015).
16. Baechler, B. L., McKnight, C., Pruchnicki, P. C., Biro, N. A. & Reed, B. H. Hindsight/RREB-1 functions in both the specification and differentiation of stem cells in the adult midgut of Drosophila. *Biol. Open* **5**, 1–10 (2015).
17. Hie, B. *et al.* Computational Methods for Single-Cell RNA Sequencing. (2020).
18. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
19. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848

(2016).

20. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
21. MacKay, D. J. C., Mac Kay, D. J. & MacKay, V. J. C. *Information Theory, Inference and Learning Algorithms*. (Cambridge University Press, 2003).
22. Peter, I. S. & Davidson, E. H. A gene regulatory network controlling the embryonic specification of endoderm. *Nature* **474**, 635–639 (2011).
23. 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).
24. Tokusumi, T. *et al.* Screening and Analysis of Janelia FlyLight Project Enhancer-Gal4 Strains Identifies Multiple Gene Enhancers Active During Hematopoiesis in Normal and Wasp-Challenged Larvae. *G3* **7**, 437–448 (2017).
25. Ewen-Campen, B. *et al.* Optimized strategy for in vivo Cas9-activation in. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9409–9414 (2017).
26. Hu, Y. *et al.* FlyPrimerBank: an online database for Drosophila melanogaster gene expression analysis and knockdown evaluation of RNAi reagents. *G3* **3**, 1607–1616 (2013).
27. Kurucz, E. *et al.* Definition of Drosophila hemocyte subsets by cell-type specific antigens. *Acta Biol. Hung.* **58 Suppl**, 95–111 (2007).
28. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
29. Hie, B., Cho, H., DeMeo, B., Bryson, B. & Berger, B. Geometric Sketching

Compactly Summarizes the Single-Cell Transcriptomic Landscape. *Cell Syst* **8**, 483–493.e7 (2019).

30. Ohlstein, B. & Spradling, A. The adult Drosophila posterior midgut is maintained by pluripotent stem cells. *Nature* **439**, 470–474 (2006).
31. Micchelli, C. A. & Perrimon, N. Evidence that stem cells reside in the adult Drosophila midgut epithelium. *Nature* **439**, 475–479 (2006).
32. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
33. Banerjee, U., Girard, J. R., Goins, L. M. & Spratford, C. M. Drosophila as a Genetic Model for Hematopoiesis. *Genetics* **211**, 367–417 (2019).
34. Márkus, R., Kurucz, E., Rus, F. & Andó, I. Sterile wounding is a minimal and sufficient trigger for a cellular immune response in Drosophila melanogaster. *Immunol. Lett.* **101**, 108–111 (2005).
35. Hao, Y. & Jin, L. H. Dual role for Jumu in the control of hematopoietic progenitors in the Drosophila lymph gland. *Elife* **6**, (2017).
36. Lebestky, T., Chang, T., Hartenstein, V. & Banerjee, U. Specification of Drosophila hematopoietic lineage by conserved transcription factors. *Science* **288**, 146–149 (2000).
37. Anderl, I. *et al.* Transdifferentiation and Proliferation in Two Distinct Hemocyte Lineages in Drosophila melanogaster Larvae after Wasp Infection. *PLoS Pathog.* **12**, e1005746 (2016).
38. Amcheslavsky, A., Ito, N., Jiang, J. & Ip, Y. T. Tuberous sclerosis complex and Myc

coordinate the growth and division of Drosophila intestinal stem cells. *J. Cell Biol.* **193**, 695–710 (2011).

39. Ren, F. *et al.* Drosophila Myc integrates multiple signaling pathways to regulate intestinal stem cell proliferation during midgut regeneration. *Cell Res.* **23**, 1133–1146 (2013).

FIGURE 1:

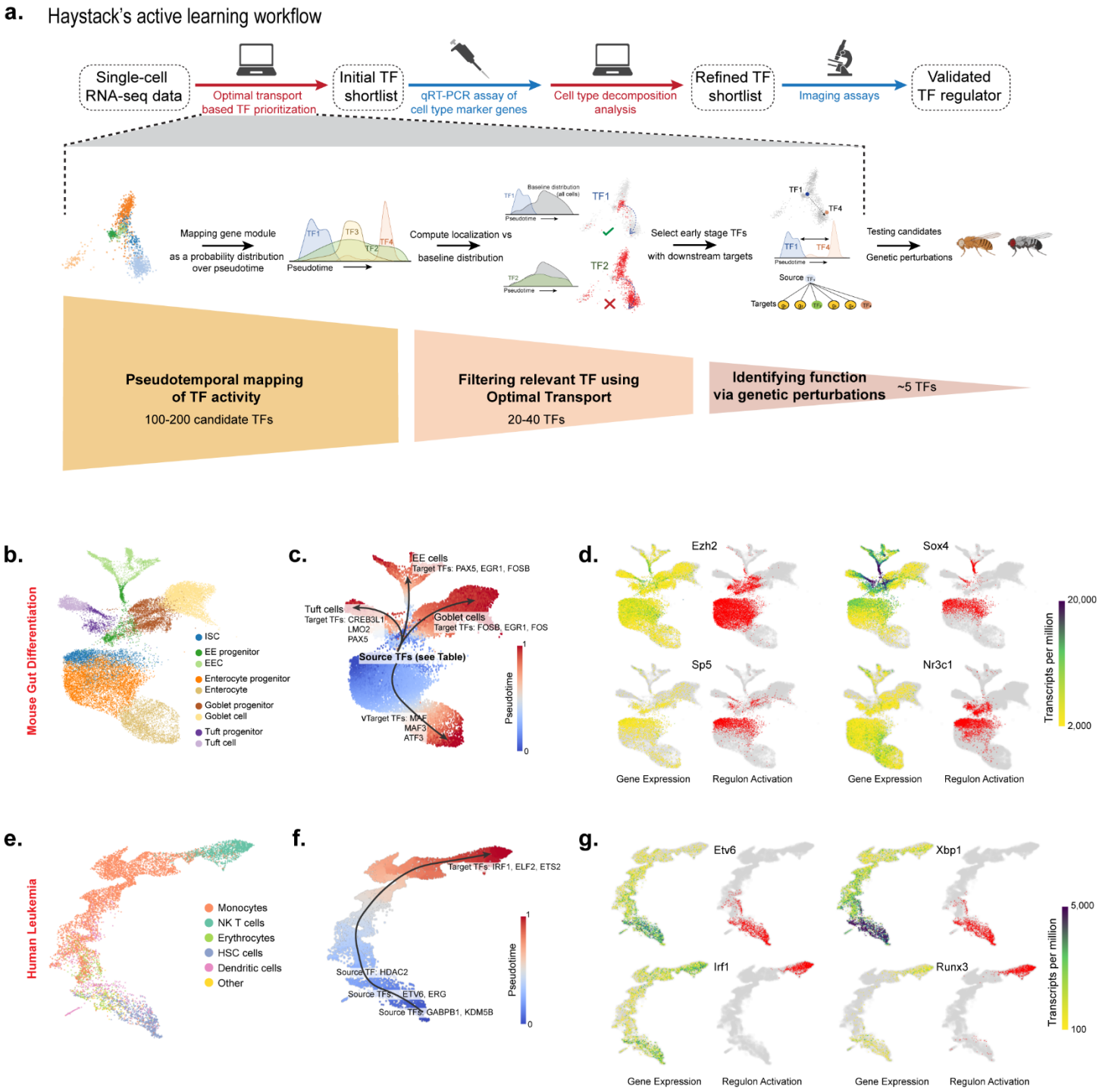

a.
Haystack's active learning workflow
Single-cell RNA-seq data
Optimal transport based TF prioritization
Initial TF shortlist
qRT-PCR assay of cell type marker genes
Cell type decomposition analysis
Refined TF shortlist
Imaging assays
Validated TF regulator
Mapping gene module as a probability distribution over pseudotime
Compute localization vs baseline distribution
Select early stage TFs with downstream targets
Testing candidates Genetic perturbations
Pseudotemporal mapping of TF activity
100-200 candidate TFs
Filtering relevant TF using Optimal Transport
20-40 TFs
Identifying function via genetic perturbations
~5 TFs
b.
Mouse Gut Differentiation
ISC
EE progenitor
EEC
Enterocyte progenitor
Enterocyte
Goblet progenitor
Goblet cell
Tuft progenitor
Tuft cell
c.
EE cells
Target TFs: PAX5, EGR1, FOSB
Tuft cells
Target TFs: CREB3L1 LMO2 PAX5
Goblet cells
Target TFs: FOSB, EGR1, FOS
Source TFs (see Table)
Target TFs: MAF MAF3 ATF3
Pseudotime
d.
Ezh2
Sox4
Sp5
Nr3c1
Gene Expression
Regulon Activation
Transcripts per million
20,000
2,000
e.
Human Leukemia
Monocytes
NK T cells
Erythrocytes
HSC cells
Dendritic cells
Other
f.
Target TFs: IRF1, ELF2, ETS2
Source TF: HDAC2
Source TFs: ETV6, ERG
Source TFs: GABPB1, KDM5B
Pseudotime
g.
Etv6
Xbp1
Irf1
Runx3
Gene Expression
Regulon Activation
Transcripts per million
5,000
100

Fig.1. Haystack identifies biologically relevant transcription factors (TFs).

a. Haystack's hybrid computational-experimental workflow relies on concepts of active learning and optimal transport to determine regulatory TFs along a trajectory inferred from an observational scRNA-seq study. Applying transport theory, we combine robust estimates of TF activity (also considering a TF's targets) with pseudotime information to prioritize TFs. Cell-type decomposition inferred from qRT-PCR experiments further selects TFs whose perturbations result in changes in differentiation. Focused imaging-based validations are subsequently conducted on the highest-confidence candidates.

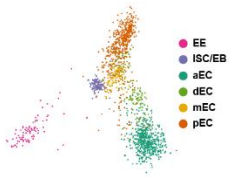
b-d. Uniform Manifold Approximation and Projection (UMAP) plots depicting the re-clustering analysis of (**b**) scRNA-seq data of mouse gut (Bottcher et al., 2021); (**c**) Monocle 3-based pseudotime analysis reveals that ISCs can give rise to four lineages: Goblet cells, EECs, Tuft cells, and Enterocytes; and (**d**) gene expression (yellow-green color scale) of the TFs *Ezh2*, *Sox4*, *Sp5* and *Nr3c1*, contrasted with their TF-module activations (in red, inferred by SCENIC) in the respective intestinal clusters.

e-g. Re-clustering of leukemia scRNA-seq data[11] (**e**) identifies known blood cell populations; (**f**) Monocle 3-based pseudotime analysis reveals a single lineage trajectory with HSCs as the source; and (**g**) UMAP plots showing the expression (yellow-green color scale) of the TFs *Etv6*, *Xbp1*, *Irf1*, and *Runx1* compared to their TF-module activity (red).

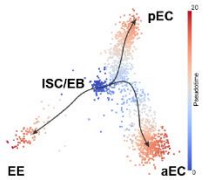
FIGURE 2:

Drosophila gut

a.



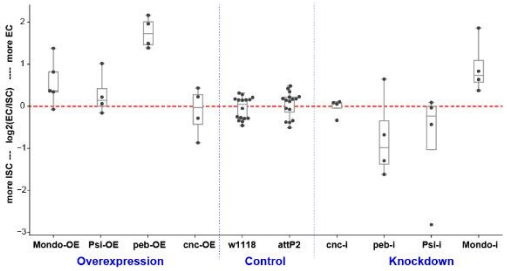
b.



c.

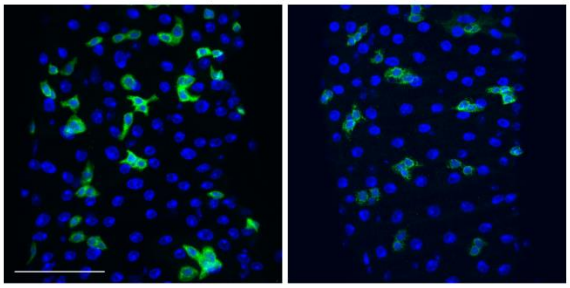
Source TF	Lineage
FoxK	ISC/EB → aEC, pEC
cnc	ISC/EB → pEC
peb	ISC/EB → pEC, EE
Psi	ISC/EB → pEC
drm	ISC/EB → pEC
Tbp	ISC/EB → pEC
Tet	ISC/EB → aEC, pEC
Mondo	ISC/EB → EE

d.

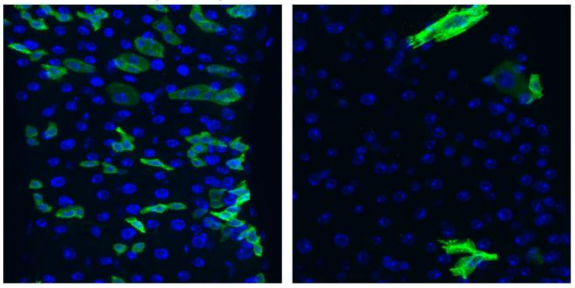


esg-Gal4, UAS-GFP, Tub-Gal80^{TS} (10D)

e. *Control (attP2)* *peb-i*



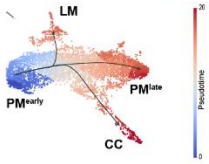
f. *Control (w1118)* *peb-OE*



Drosophila blood

g.

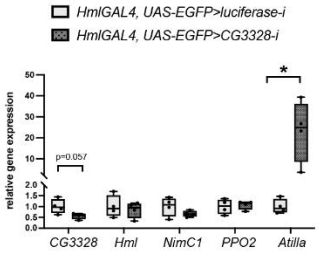
h.



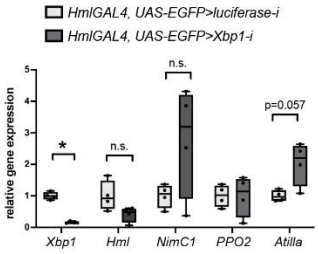
i.

Source TF	Lineage
Dp	PM^{early}→CC, LM
Jumu	PM^{early}→CC, LM
Lz	PM^{early}→CC
Myc	PM^{early}→CC
E(spl)mβ-HLH	PM^{early}→PM^{late}
Xbp1	PM^{early}→LM
CG3328	PM^{early}→LM

j.



k.



Hml-Gal4, UAS-2xEGFP

l. *Control (luci-i)* *CG3328-i* *Xbp1-i*

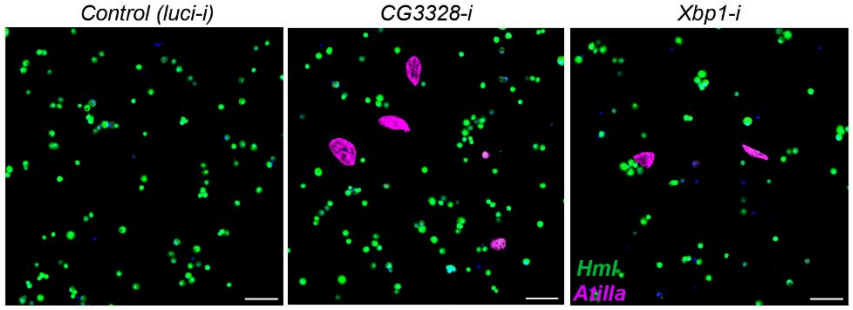


Fig. 2. Analysis of *Drosophila* midgut and larval blood scRNA-seq data by Haystack.

a. *Drosophila* midgut scRNA-seq dataset[12] shows six known intestinal cell populations: enteroendocrine (EE), intestinal stem cells/enteroblasts (ISC/EB), enterocytes (ECs) of the different regions of the gut [anterior (aEC), differentiating (dEC), middle (mEC), and posterior (pEC)].

b. Pseudotime analysis using Slingshot reveals three distinct lineages from the ISC/EB cluster: EE, aEC, and pEC.

c. Haystack identifies source TFs specific to three lineages.

d. qRT-PCR-based lineage analysis of ISC and EC cell-types in the midguts of *Mondo*, *Psi*, *Peb*, and *cnc* perturbations compared to their respective controls (w^{1118} and *attp2*). Only *Peb* displays mutually consistent overexpression and knockdown phenotypes.

e-f. Confocal images of *Drosophila* midguts of *peb-i* and *peb-OE* show a decrease in progenitors. Scale bar = 100um.

g. *Drosophila* larval blood scRNA-seq[14] shows plasmatocyte (PM), crystal cell (CC), and lamellocyte (LM) clusters. CC1 and LM1 represent putative immature while CC2 and LM2 represent mature CCs and LMs, respectively.

h. Monocle 3-based pseudotime analysis shows that three lineages (PM^{late}, CC, and LM) emerge from the oligopotent PM^{early} cluster.

i. Table representing source TFs identified by Haystack.

j-k. qRT-PCR analysis of larval blood upon knockdown of *CG3328* (**j**) and *Xbp1* (**k**) shows an increase in the LM marker gene *Atilla*. Non-parametric multiple t tests were used and n.s. and * represent not significant and $P<0.05$, respectively. N=4 biological replicates.

l. Confocal imaging of blood cells derived from *HmlGFP>luci-i*, *CG3328-i*, and *Xbp1-i* larvae. Scale bar = 50um.

Supplementary Information

Supplementary Note 1

Cost of perturbations

Consider a relatively simple scenario, where one needs to analyze 10 homozygous mice that can only be generated from breeding heterozygous (het) animals. A total number of 40 offspring would be required to obtain these 10 (25%) homozygotes. Assuming a breeding female averages 6 pups/litter, a total of 7 het breeding pairs (7 females and 7 males; 7*6=42) would be required to generate one experimental cohort in one round of breeding (3 months). If het animals are not available, rearing wild type animals with a het animal generates ~50% het offspring. Thus, to generate 14 het mice would require 28 offspring. As such, ~5 WT x het breeding pairs are needed. Altogether, starting with 5 WT X het breeder pairs, the 10 homozygous mice would be ready in about 6 months. For just one experiment, approximately 80 mice will be generated. It would cost ~$2800 for housing the mice with the assumption of $1.25 per diem per animal (Boston University, https://tinyurl.com/yc5akwu8). This is an underestimation since in reality there are costs associated with reagents for genotyping and extra crosses that buffer for unsuccessful breedings. Additionally using qRT-PCR as a proxy for cell-type composition bypasses the need to buy costly reagents like antibodies or *in situ* probes to label cell-types. This highlights why selectively prioritizing high-confidence TFs for perturbation experiments is crucial for reducing cost.

Supplementary Note 2

Applying Haystack on mouse gut and acute myeloid leukemia cells

We applied Haystack on the scRNA-seq studies of mouse gut differentiation[10] (GSE152325) and human acute myeloid leukemia (AML) cells[11] (10.5281/zenodo.3345981). We used these datasets to first reconstruct cell lineages (diffusion pseudotime algorithm for the mouse gut as in the original study; Monocle 3 for the AML study) and identified TF modules along the trajectories. In the mouse gut, intestinal stem cells (ISC) differentiate along four lineages: enteroendocrine (EE) cells, tuft cells, goblet cells and enterocytes (ECs). Applying Haystack on each of the lineages, we identified the source TFs localized to or near the ISC cluster. Interestingly, when we visualized individual TFs using transcript levels alone (eg. *Ezh2*, *Sox4*, *Sp5* and *Nr3c1*) (**Fig. 1d**), a diffuse expression pattern was observed. Instead, visualizing TF modules revealed that the same diffusely expressed TFs had very localized activity. We aggregated these results across the lineages by counting the total number of downstream targets per source TF and ranked them within each lineage. Of the 24 putative source TFs predicted by Haystack, 75% had substantial support in the literature for their involvement in the development of the gut, most of which were related to differentiation (see **Table S1**).

For the human AML dataset, the pseudotime analysis revealed that hematopoietic stem cells (HSC) differentiate into dendritic cells, erythrocytes, NK T cells and monocytes in a linear trajectory. We found that TF modules exhibited clearly localized activity while the raw gene expression was diffusely expressed (e.g., *Etv6*, *Xbp1*, *Irf1* and *Runx3*) (**Fig. 1g**). We identified a total of 25 high-confidence TF predictions, ~85% of which had substantial evidence in the literature to support involvement in blood function (see **Table S2**).

Benchmarking Haystack

Generating a reference set of genes from text mining: A standard approach to evaluating gene prioritization methods is to assess the subset of genes they prioritize against a reference list of genes documented to be involved in the process of interest. To our knowledge, for mouse gut or human AML differentiation no such reference gene sets are available. We therefore approximated the reference gene sets for the two tissues by text mining of research studies. Though these reference sets have not been derived by manual curation, we believe that our systematic approach results in unbiased errors, making these sets a useful validation benchmark. For each tissue, a collection of relevant studies was first acquired by querying the Pubmed database for publications whose title or abstract indicate that the study likely corresponds to the tissue of our interest. The queries used were:

- Mouse gut: `("mouse"[Title/Abstract]) AND (gut[Title/Abstract] OR intestine[Title/Abstract]) AND (development[Title/Abstract] OR differentiation[Title/Abstract])`

- Human AML cells: `(human[Title/abstract]) AND (hematopoiesis[Title/Abstract] or "blood differentiation"[Title/Abstract] or leukemia[Title/Abstract]) AND regulation[Title/Abstract]`

An alternative querying approach would have been to limit ourselves to MeSH (Medical Subject Headings) gene and function annotations. However, we found the MEDLINE annotation of Pubmed articles by MeSH terms to miss important genes for our biological processes of interest. For example, the following query which uses only MeSH terms to search for publications of *Ezh2* activity in mouse intestine development returned zero hits at the time of submission. In contrast, our approach identified four relevant publications (e.g., see Ezh2-related reference in **Table S1**)

- `(("Ezh2 protein, mouse" [Supplementary Concept]) OR ("Polycomb Repressive Complex 2"[Mesh])) AND ("Intestines/growth and development"[Mesh])`

From the titles and abstract of the queried list of publications, we identified gene names. The set of protein-coding gene names and their synonyms was sourced from the NCBI Gene database. We excluded gene synonyms that are also common English words. For instance, for the *WASP actin nucleation promoting factor* we included the synonyms *IMD2, SCNX, THC, THC1, WASP*, and *WASPA*; but we excluded the synonym *WAS*. The final reference set for each tissue was chosen as the genes listed in at least two publications (**Table S4**).

<u>Benchmarking:</u> We computed the top TFs prioritized by Haystack as well as other methods (**Fig. S1**). We evaluated three gene network inference methods: SCODE[2], PIDC[3], and SINGE[4]. The gene network inference methods seek to build a regulatory graph of TFs based just on TF gene expression; SCODE and SINGE also make use of pseudotime input. While SINGE reports both a regulatory graph and a ranked list of TFs (which we used), SCODE and PIDC only report the regulatory graph. Therefore, we computed a TF ranking as the sum of unsigned edge-weights (i.e., both activation and repression) originating from the TF. This approach is similar to the TF ranking procedure described in Matsumoto et al.[2] Due to SINGE's long run-time (10 days for a dataset with 3,000 cells), we applied it to a random subset of 3,000 cells in human AML tissue and 2,000 cells in mouse gut tissue. For each method we selected the top 30 hits (fewer if the method reported less than 30 TF hits). We also shortlisted TFs by a differential expression analysis, performing the Wilcoxon rank-sum test to identify TFs differentially expressed between progenitor and other cell types as annotated in the original studies. The requirement of a pre-curated annotation might be infeasible for a poorly-studied tissue; therefore, neither Haystack nor the gene network inference methods require such pre-curated annotations. We also evaluated the full set of SCENIC regulons (93 in the case of mouse gut; 219 for human AML cells); this is an unordered set since SCENIC does not consider the differentiation landscape. For each gene set, its enrichment in the reference set was evaluated, computed as the fold-change over the expected enrichment of an equal-sized random set of protein-coding genes.

As another control, we computed the fold-enrichment of the set of all TFs in the species, finding this set to be enriched in both tissues. We interpret this as supporting our text-mining approach

to extracting relevant studies: it would have been surprising if published studies of differentiation in these tissues did not highlight TFs more than other genes. We note that the set of SCENIC regulons is substantially more enriched than the set of all TFs, indicating that its module discovery process enhances the TF signal in the data. However, SCENIC by itself is not sufficient— it outperformed single-gene differential expression in mouse gut but underperformed the latter in human AML cells. In comparison, the set of TFs prioritized by Haystack not only displayed higher enrichment than either SCENIC or differential expression, but also higher than the gene network inference methods (**Fig. S1**).

Runtime and memory usage: We assume here that pseudotime trajectories have been already computed by the researcher's method of choice. Once TF modules and pseudotime have been computed, the optimal transport computation in Haystack runs under 5 minutes and requires less than 8 GB of RAM. The preprocessing step of computing TF modules via SCENIC (https://github.com/aertslab/pySCENIC) is more time intensive: on a 10,000 cell dataset, computing TF modules using the pySCENIC Docker instance (using the GRN2 sub-module and with parallelization enabled) required 22 minutes on a 24-core Intel Xenon 3.5 GHz server with peak memory consumption under 30 GB. The run-time of SCENIC increases with the number of cells and we therefore recommend sketching approaches[29] to downsample datasets with hundreds of thousands of cells.

Supplementary Note 3

Applying Haystack on *Drosophila* midgut

The fly midgut consists of a monolayer of absorptive enterocytes (ECs) and secretory enteroendocrine cells (EEs) that are replenished by self-renewable intestinal stem cells (ISCs)[30,31]. Previously, we had performed scRNA-seq on fly whole guts and identified a total of 22 clusters mainly consisting of finer sub-classifications of EEs and ECs[12,13]. Some of these subtypes are distinguished by their spatial location whereas others were intermediate states between ISC/EB and a specific terminal state[13]. We applied Haystack on this scRNA-seq dataset. To avoid the confounding effects of unknown cell clusters, we limited our analysis to the known midgut cell-types; including EEs, ISC/EBs, anterior ECs, differentiating ECs, middle ECs and posterior ECs (**Fig. S2a**). Using SlingShot[18], we mapped individual cells onto a pseudotime trajectory consisting of three lineages with the starting point set as ISC/EB and end points as EE, aEC or pEC (**Fig. 2a,b**). By mapping TF module activity along these trajectories, Haystack identified eight TFs that were localized to the ISC/EB starting point (Source TFs). These included *Forkhead box K (FoxK)*, *cap-n-collar (cnc)*, *pebbled (peb)*, *P-element somatic inhibitor (Psi)*, *drumstick (drm)*, *TATA binding protein (Tbp)*, *Ten-Eleven Translocation family protein (Tet)* and *Mondo* (**Fig. S2b**).

As a broadly applicable intermediate validation, we measured mRNA levels (by qRT-PCR) of gene markers of specific cell types to approximate the cell-type composition. Using differential gene analysis, we selected a total of 11 markers that were distinctly expressed in ISC/EB (*Dtg, N, LanB1*), aEC (*CG6295, Npc2f*), pEC (*LManVI, Gs2, Mur29B*) or EE (*esg*, *AstC, IA-2*) (**Fig. S2b**). In the original Hung et al. study, these markers had been identified for individual cell clusters using a differential expression analysis in Seurat[32] and for each cell type, we chose a combination of markers such that all sub-clusters of a cell type were covered. We knocked-down or overexpressed the eight TFs specifically in adult ISC/EB with available reagents. Among these, we found five RNAi lines (*peb-i, Psi-i, Tet-i, Mondo-i, cnc-i*) and seven overexpression (OE) lines (*peb-OE, Psi-OE, FoxK-OE, drm-OE, two Mondo-OE, cnc-OE*) that successfully reduced or increased, respectively, the gene of perturbation when assaying mRNA extracted from the whole midgut (**Fig. S2d**). We measured the levels of the 11 markers for each perturbation, which amounted to a total of 132 observations. We calculated the fold change (FC_{RpL32}) in comparison to a control, using *RpL32* as a reference. In all cases, perturbation led to at least one, and up to 11, significant fold change(s) in marker gene expression. The FC_{RpL32} in markers gave us a proxy for the cell-type composition of the gut. For example, knockdown of *Psi* in intestinal progenitors caused a decrease in EE and pEC markers whilst ISC/EB and aEC markers remained unchanged (except for *Npc2f*). This suggests that *Psi* is involved in promoting the differentiation of progenitors to EEs and ECs in the posterior midgut (**Fig. 2d**). *Drm* overexpression caused an increase in ISC/EB markers with no changes in most terminal cell-type markers other than an increase in *Mur29B*. This suggests that *drm* promotes ISC proliferation and may not be involved in differentiation (**Fig. S2d**).

To overcome the caveat that terminal markers are confounded by the proliferation of progenitors, we used the EC–ISC mean marker ratio to compare our perturbations with the control. For *peb* perturbations, we knocked-down or overexpressed *peb* in adult intestinal progenitors and used confocal microscopy to observe cellular defects (see **Fig. 2e,f**). Intestinal progenitors were labeled by GFP expression and ECs were recognized by their polyploid nuclei (**Fig. S2e-f**). Compared to control, the raw cell counts (including GFP-positive cells, EEs, and ECs) were lower in both the *peb-i* and *peb-OE*, with a decrease more prominently observed in the posterior midgut than the anterior region (**Fig. S2e-f**). This similarity in the cell-count profile, because of opposite perturbations of the same gene, is surprising and could be a reason why previous studies[15,16] arrived at different conclusions. We reasoned that the differences between the opposite perturbations might be clarified if we refined our parsing of cell types. Although mature ECs are polyploid and can be readily distinguished from other cell types, premature ECs undergoing endocycling can be hard to discern from ISCs. Additionally, ECs that have rapidly differentiated from ISCs can be GFP-positive as a result of the perdurance of the protein expressed in the progenitors. Thus, we measured nuclei area as a non-biased way to profile ISC/EB differentiation. In the anterior region, *peb-i* or *peb-OE* in intestinal progenitors resulted in no change in the mean nuclei area. The same perturbations, in the posterior midgut, caused a statistically significant but moderate increase in the mean nuclei area. Interestingly, differences could be observed by looking at the frequency distribution of the nuclei area. When compared to control, *peb* knockdown displayed a higher percentage of small nuclei cells (progenitors), a reduction in endocycling ECs, and an increase in large ECs in the anterior midgut. In the same region, *peb-OE* showed a reduction in the percentage of progenitors. In the posterior midgut, *peb-i* caused a decrease in early endocycling ECs but an increase in the late and mature ECs. *Peb-OE* causes a notable decrease in the proportion of progenitors and an increase in large ECs.

Applying Haystack on *Drosophila* blood

Drosophila hemolymph consists of three populations of blood cells or hemocytes: macrophage-like plasmatocytes (PM), platelet-like crystal cells (CC), and giant-cell like lamellocytes (LM), which express the known marker genes *NimC1*, *PPO1/2*, and *Atilla*, respectively[33]. We applied our method on blood cell scRNA-seq pertaining to larvae upon wounding, which is sufficient to activate blood cells and induce LMs[14,34]. For simplicity, PM clusters from the original study have been combined into two groups, PM^{early} and PM^{late}, corresponding to the early (oligopotent) and late-stage (mature) PMs, respectively (**Fig. 2g**). The CC and LM cell types are also subdivided into two clusters, with CC1 and LM1 representing the putative immature states of mature CCs (CC2) and LMs (LM2) (**Fig. 2g**). We re-applied Monocle 3 on this dataset[20] and identified three main lineage trajectories with the source set at oligopotent PMs (PM^{early}, in blue) (see **Fig. 2h**). For the $PM^{early} \rightarrow$ CC lineage, Haystack identified four source TFs: *Dp*, *Jumu*, *Lz*, and *Myc* (**Fig. 2i**). The latter three TFs (*Jumu*, *Lz*, and *Myc*) have been implicated in the CC lineage[35,36], suggesting the power of Haystack in accurately predicting lineage-determining TFs. For the $PM^{early} \rightarrow$ LM lineage, Haystack identified two novel source TFs *CG3328* and *Xbp1* (**Fig. 2i**), both of which displayed a diffused pattern of expression around the PM^{early} cluster (**Fig. S4b,f**). We note that the human ortholog of *Xbp1* was also identified in the Haystack analysis of human

leukemia data (see **Fig. 1g**), which suggests an evolutionarily conserved role for this TF in blood cell differentiation.

In this study, we focused on the LM lineage as the regulators of PM->LM differentiation are not well established. Hence, for *in vivo* perturbations, we knocked-down or overexpressed *CG3328* and *Xbp1* in blood cells using the *Hml-Gal4, UAS-2xEGFP* driver (hereafter as *HmlGFP*), where EGFP marks most PMs[37]. To identify changes in blood cell lineages upon perturbation of *CG3328* and *Xbp1*, we first performed qRT-PCR on total RNA derived from the larval hemolymph containing circulating and sessile blood cells of the various genotypes. RNAi-mediated knockdown of *CG3328* resulted in an induction of the LM marker gene *Atilla* (**Fig. 2j**), indicating activation of the LM lineage. To address the role of gain-of-function of *CG3328*, we utilized CRISPR-mediated activation (*CRISPR*[a]) approach using the *Hml-Gal4, UAS-EGFP; UAS-dCas9* (*HmlGFP;dCas9*) driver. However, increasing the levels of *CG3328* in PMs did not affect the expression of any of the lineage marker genes (**Fig. S4c**), suggesting that forced expression of *CG3328* does not impact the blood cell type composition. With regards to the perturbation of *Xbp1*, we observed an increase in *Atilla* expression (**Fig. 2k**), akin to the knockdown of *CG3328*. On the other hand, overexpression of *Xbp1* resulted in decreased expression patterns of *NimC1* and *Atilla*, while *Hml* and *PPO2* remain unchanged compared to *OreR* controls (**Fig. S4g**), suggesting that forced expression of *Xbp1* disallows both PM^{late} and LM lineages. To further validate our findings from the qRT-PCR data pertaining to the role of these two TFs in regulating the LM lineage, we performed confocal imaging of blood cells in respective genotypes. We identified that knockdown of *CG3328* and *Xbp1* led to an increase in the fraction of LMs compared to their respective controls (see **Fig. 2l; S4d,e,h,i**). Lastly, besides the LM lineage, we also tested the role of *E(spl)mbeta* in the $PM^{early} \rightarrow PM^{late}$ lineage, as predicted by Haystack (**Fig. 2i, S4j**). qRT-PCR analysis shows that overexpression of this TF decreased the expression levels of *NimC1* (**Fig. S4k**), while the cell type compositions (of PMs and CCs) remain unchanged with no detection of Atilla+ LMs (**Fig. S4l**). Altogether, these results indicate the predictive power of Haystack in shortlisting biologically relevant TFs for downstream lineage analyses.

Using Haystack to predict source-target TF pairings

Towards identifying signaling cascades of TFs, we next used Haystack to identify TFs localized to the end-points of trajectories (i.e., target TFs) in *Drosophila* gut differentiation. Using the cisTarget database, we limited our analysis of TFs that were putative transcriptional targets of the 8 source TFs described previously. We identified a total of 54 that satisfied this criterion and further narrowed them down to 13 TFs that localized to cells at the three endpoints. Although not all target TFs were downstream of each of the 8 source TFs, we measured the levels of all 13 target TFs under the 12 “source” perturbations (7 overexpression; 5 knockdown) to assess the validity of our putative predictions; we also measured all 8 source genes to confirm the perturbations. This amounted to a total of 252 observations. Among the 24 perturbations tested for putative source-target pairings, 14 (58%) were correct (**Fig. S3**). That is, down regulating a

source TF led to a decrease in a putative target TF or vice versa for overexpression. One of the novel source-target pairs was peb>*Myc*. *Peb-i* resulted in a decrease in *Myc* levels, whilst *peb-OE* increased the levels of *Myc*. This is consistent with previous studies of Myc function in the midgut where the overexpression of *Myc* increases nuclei size and knockdown of *Myc* reduces ISC numbers[38,39].

Supplementary Figures and Tables

Fig. S1

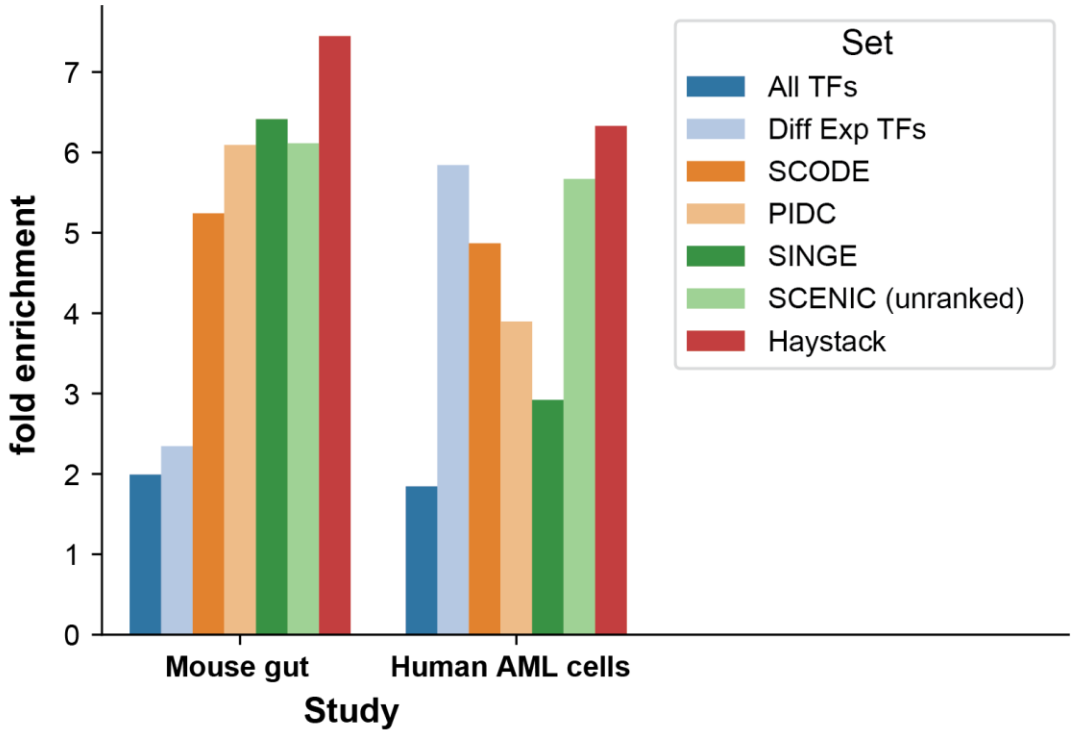


Fig. S1. Benchmarking Haystack with other gene network inference methods.

The bar graph represents enrichment scores using a text-mining based gene enrichment metric. Briefly, a collection of relevant studies was first acquired by querying the Pubmed database for publications whose title or abstract indicate that the study likely corresponds to the tissue of interest. The top 30 TFs (or fewer, as available) predictions derived from each gene network inference method are used to determine if they are enriched in the collection of Pubmed studies. This is displayed as a fold enrichment compared to the control of 30 random genes. Since SCENIC does not rank TF modules, its enrichment scores correspond to the entire set of TF modules discovered by SCENIC (93 in Mouse gut, 219 in Human AML cells).

Fig. S2

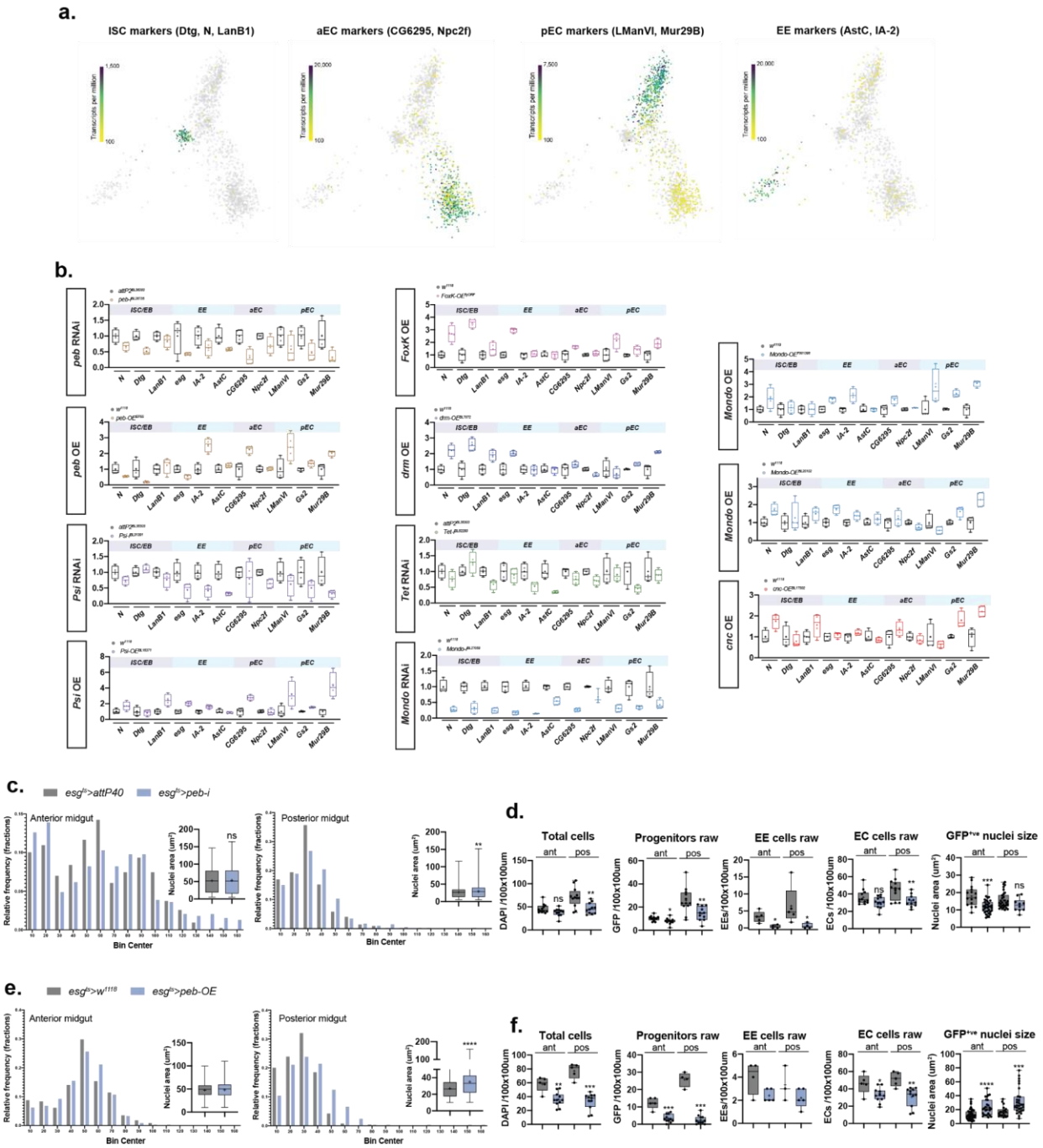


Fig. S2. Analysis of the *Drosophila* midgut using qRT-PCR and confocal microscopy.

a. UMAP plots representing the expression of intestinal cell marker genes for progenitors (*Dtg, N, LanB1*), aEC (*CG6295, Npc2f*), pEC (*LManVI, Mur29B*) or EE (*AstC, IA-2*)

b. Bar graphs represent marker gene expression pertaining to the gut lineage validated by qRT-PCR of *Drosophila* gut mRNAs upon knockdown (KD) or overexpression (OE) of various source TFs such as *peb*, *Psi*, *FoxK*, *drm*, *tet*, *Mondo*, and *cnc*. The y-axis represents fold change compared to control normalized to *RpL32*.

c-f. Quantification of confocal micrographs after *peb* perturbation. **(c,d)** *peb* knockdown and **(e,f)** *peb* overexpression in intestinal progenitors. Parametric t-tests were used to calculate statistics for nuclei area. Non-parametric t-tests were used to calculate the statistics where *, **, ***, **** represent p values <0.05, 0.01, 0.001, 0.0001, respectively.

Fig. S3

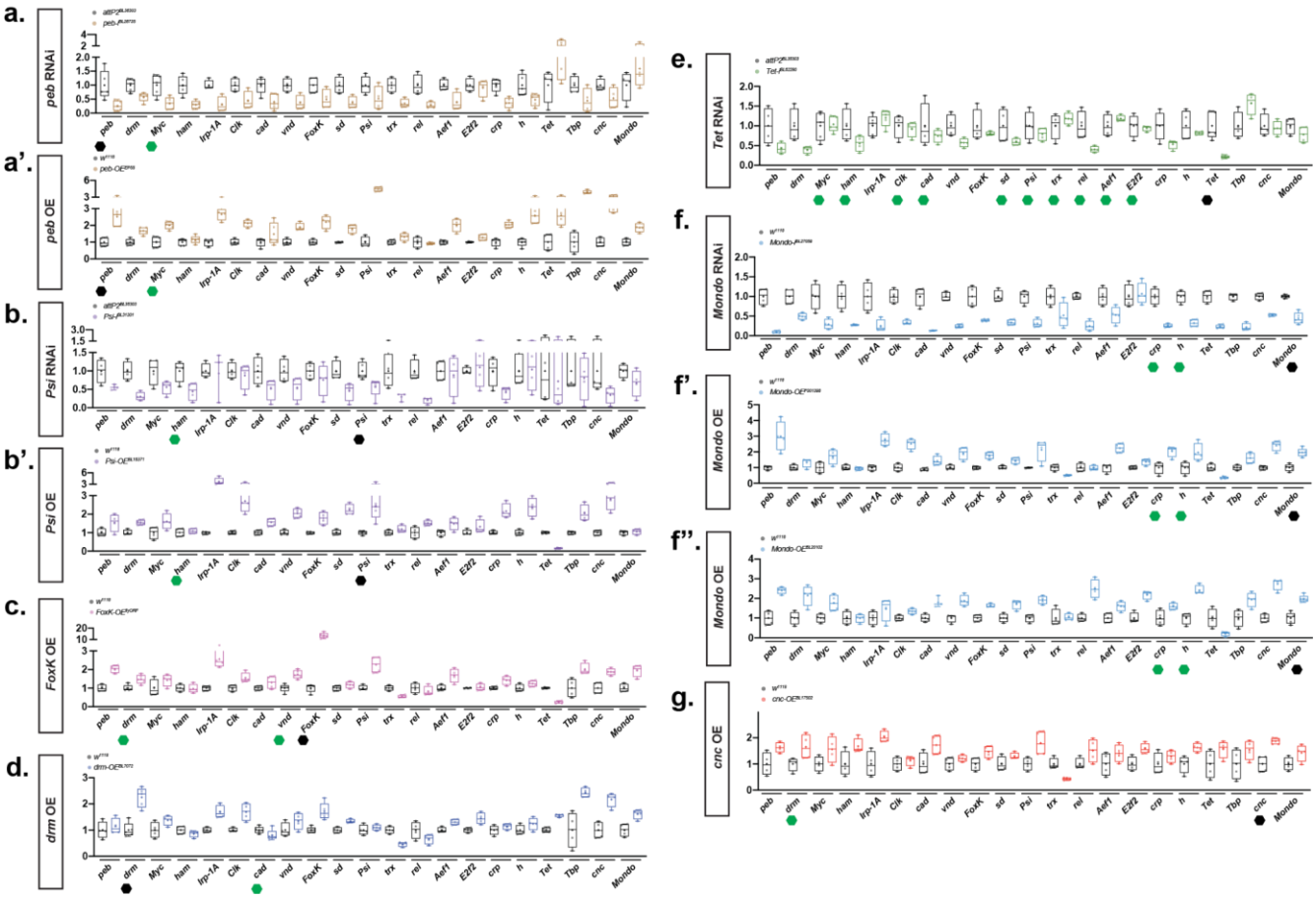

a.
peb RNAi
a'.
peb OE
b.
Psi RNAi
b'.
Psi OE
c.
FoxK OE
d.
drm OE
e.
Tet RNAi
f.
Mondo RNAi
f'.
Mondo OE
f''.
Mondo OE
g.
cnc OE
peb drm Myc ham Irp-1A Clk cad vnd FoxK sd Psi trx rel Aef1 E2f2 crp h Tet Tbp cnc Mondo

Fig. S3. Marker gene expression and Source - Target validation by qRT-PCR.

Bar graphs represent putative TF target gene expression pertaining to the gut lineage validated by qRT-PCR of *Drosophila* gut mRNAs upon knockdown (KD) or overexpression (OE) of various source TFs such as *peb* (**a, a'**), *Psi* (**b, b'**), *FoxK* (**c**), *drm* (**d**), *tet* (**e**), *Mondo* (**f-f''**), and *cnc* (**g**). Panels are bar graphs representing the source TF (black hexagon) - target TF (green hexagon) validations by qRT-PCR of the aforementioned TFs. The y-axis represents fold change compared to control normalized to RpL32.

Fig. S4:

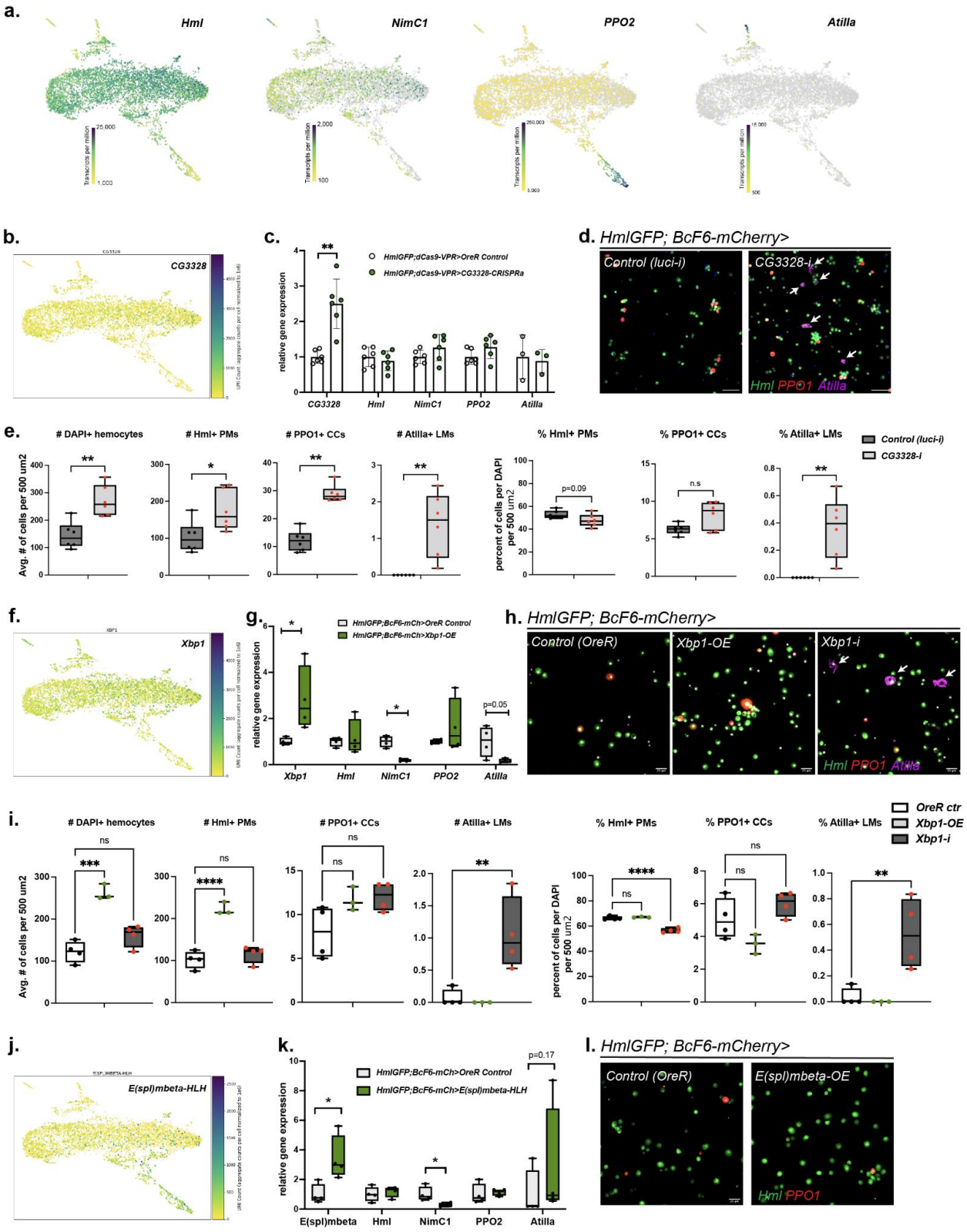

a.
Hml
NimC1
PPO2
Atilla
b.
CG3328
c.
HmlGFP;dCas9-VPR>OreR Control
HmlGFP;dCas9-VPR>CG3328-CRISPRa
relative gene expression
CG3328
Hml
NimC1
PPO2
Atilla
d. HmlGFP; BcF6-mCherry>
Control (luci-i)
CG3328-i
Hml PPO1 Atilla
e.
DAPI+ hemocytes
Hml+ PMs
PPO1+ CCs
Atilla+ LMs
% Hml+ PMs
% PPO1+ CCs
% Atilla+ LMs
Control (luci-i)
CG3328-i
Avg. # of cells per 500 um2
percent of cells per DAPI per 500 um2
p=0.09
n.s
f.
Xbp1
g.
HmlGFP;BcF6-mCh>OreR Control
HmlGFP;BcF6-mCh>Xbp1-OE
relative gene expression
p=0.05
Xbp1
Hml
NimC1
PPO2
Atilla
h. HmlGFP; BcF6-mCherry>
Control (OreR)
Xbp1-OE
Xbp1-i
Hml PPO1 Atilla
i.
DAPI+ hemocytes
Hml+ PMs
PPO1+ CCs
Atilla+ LMs
% Hml+ PMs
% PPO1+ CCs
% Atilla+ LMs
OreR ctr
Xbp1-OE
Xbp1-i
Avg. # of cells per 500 um2
percent of cells per DAPI per 500 um2
ns
j.
E(spl)mbeta-HLH
k.
HmlGFP;BcF6-mCh>OreR Control
HmlGFP;BcF6-mCh>E(spl)mbeta-HLH
relative gene expression
p=0.17
E(spl)mbeta
Hml
NimC1
PPO2
Atilla
l. HmlGFP; BcF6-mCherry>
Control (OreR)
E(spl)mbeta-OE
Hml PPO1

Fig. S4: Haystack identifies known and novel TFs in *Drosophila* blood cell differentiation

a. UMAP plots representing the expression of blood cell marker genes *Hml*, *NimC1* (PM), *PPO2* (CC), and *Atilla* (LM).

b-c. UMAP plot representing the expression of *CG3328* (**b**). qRT-PCR analysis of hemocytes derived from control or *CG3328-OE* shows that overexpression of *CG3328* has no substantial changes in the blood cell type composition based on the unchanged marker gene expression (**c**), N=3-6 biological replicates. Non-parametric multiple t tests were used to calculate the statistics where ** represents a p value <0.01.

d-e. Confocal images of *CG3328* knockdown (*CG3328-i*) shows production of Atilla+ LMs compared to luciferase RNAi (*luci-i*) controls (**d**, arrows). Bar graphs represent the cell counts which show an increase in the percentage of both PPO1+ CCs and Atilla+ LMs in *CG3328-i* (**e**). Note that total blood cell number (DAPI+ cells), Hml+ PMs, PPO1+, and Atilla+ LMs CCs are significantly increased upon *CG3328* knockdown. Non-parametric multiple t tests were used to calculate the statistics where * and ** represent p values <0.05 and 0.01, respectively. N=6 biological replicates.

f-g. UMAP plot representing the expression of *Xbp1* (**f**). qRT-PCR analysis shows that overexpression of *Xbp1* in blood cells decreases the expression of both *NimC1* and *Atilla* (**g**). N=4 biological replicates. Non-parametric multiple t tests were used to calculate the statistics where * represents a p value <0.05.

h-i. Confocal images (**h**) of *Xbp1*-overexpression (*Xbp1-OE*) and knockdown (*Xbp1-i*) shows increased blood cell number (in *Xbp1-OE*) and increased fraction of of Atilla+ LMs (arrows in **h**) compared to *OregonR* (*OreR*) controls (**i**). Note that *Xbp1-OE* causes an increase in the total blood cell numbers (DAPI+ cells) and Hml+ cells. One-way ANOVA was used to calculate the statistics where **, ***, and **** represent p values <0.01, 0.001, and 0.0001, respectively. N=3-4 biological replicates.

j-l. UMAP plot representing the expression of *E(spl)mbeta-HLH* (**j**). qRT-PCR analysis shows that overexpression of *E(spl)mbeta-HLH* in blood cells decreases the expression of *NimC1* (**k**, N=4 biological replicates), validating its role in the $PM^{early} \rightarrow PM^{late}$ lineage. Confocal images of *OreR* control and *E(spl)mbeta-OE* show no marked changes in cell type composition. Non-parametric multiple t tests were used to calculate the statistics where * represents a p value <0.05.

Table S1. Haystack source TF predictions in the mouse gut

Mouse gene	Mammalian function in gut related cell-types	PMID
ezh2	Enhances transcription of beta-catenin transcriptional complex at Wnt target promoters	24055345
myc	Required for the induction of crypt formation	20708588; 16107730
e2f1	Knockouts have increased p53 independent cell death of crypt intestinal cells	20016602
NR3C1	Deletion protects intestine against inflammation	33684964
egr1	Targeting Egr1 attenuates radiation induced apoptosis in the mouse small intestines	26206332
sox9	Required for Paneth cells differentiation	17681175; 26170137
nr1h4/fxr	FXR deficiency promotes cell proliferation, inflammation and tumorigenesis in the intestine	18981289
sox4	Sox4 promotes intestinal secretory differentiation toward tuft and enteroendocrine fates	30055169
fos	Expressed in villus epithelial cells, but not in crypt cells	11572941
nfic	-	-
pax6	Controls proglucagon gene expression in EE cells	10478839

lhx1	-	-
ascl2	A Wnt target. OE induces hyperplasia. Deletion leads to loss of Lgr5 stem cells	19269367
jun	Required for intestinal cancer development in Apc{min}/+ mice	16007074
foxa2	Control the differentiation of goblet and EE L- and D- cells	19737569
sp5	-	-
creb3l4	-	-
pou2f3	Pou2f3 null mice lack intestinal tuft cells	26762460
atf3	Loss of ATF3 decreases crypt numbers and shortens colon length during DSS-induced colitis	30455690
e2f8	Human E2F8 suppresses cell proliferation in colon cancer cells by modulating the NFkB pathway	31471336
mybl2	Knockdown induces the accumulation of cells in G2M with a concomitant decrease in G1 in Caco-2 cells	20857481
hoxb6	-	-
spi1	-	-

Table S2. HACKSTACK source TF predictions in human leukemia

Human gene	Mammalian function in blood related cell-types	PMID
etv6	TEL function is essential for the establishment of hematopoiesis of all lineages in the bone marrow	9694803
gabpb1	Necessary for stem/progenitor cell maintenance and myeloid differentiation	27100840
hdac2	Hdac1 and hdac2 are required for early hematopoiesis	24763403
kdm5b	Required for HSC self-renewal	25655602
taf7	-	-
erg	Promotes and required for HSC maintenance and restricts their differentiation	26385962; 21673349
polr2a	-	-
maz	Regulates erythroid differentiation program	34351390
kdm5a	Promotes NK cell activation by regulating interferon-gamma production	27050510
ybx1	Required for maintaining myeloid leukemia cell survival	33763698

spi1	Promotes B cell and macrophage differentiation at low and high concentrations respectively	10827957; 8896458; 8079170
smarca4	Required for leukemia cell expansion	24285714
xbp1	Required for plasma cell differentiation	12846805; 12969976
runx2	Ensures the expression of pDC-signature genes in leukemic cells	30971697
stat5a	Activation of STAT5A in HSCs results in their enhanced self-renewal and promotes differentiation toward erythroid lineage	15353555
gata2	Regulates dendritic cell differentiation; Mutations in Gata2 impair definitive hematopoiesis in CML	27259979; 34078881
tfdp1	-	-
foxp1	Represses human plasma cell differentiation; Negative regulator of Follicular Helper T cell differentiation.	26289642; 24859450
irf5	Regulates the plasma cell commitment factor Blimp-1 and B-cell terminal differentiation in mice	20176957
nfatc2	Control both T and B cell activation and differentiation	11163226
smad1	Depletion limits hematopoietic potential because of a block in mesoderm development	21515822

znf76	-	-
myc	c-Myc−/− mice develop severe thrombocytosis-anemia-leukopenia	19372257
tfec	Controls hematopoietic stem cell vascular niche during zebrafish embryogenesis	27402973
klf7	Increased expression inhibits myeloid cell proliferation in lymphoid leukemia	22936656

Table S3. Primers used for qRT-PCR

Gene	FlyPrimer bank ID / PMID	Forward Primer	Reverse Primer
aef1	PP29256	CACCTGACCACGCATAGTCC	GTGCTTAGCTGTCGAAAGCGA
AstC	PD44956	CTCACCCTGTTCTTTGCCCT	GGTCCTGTTTCGGCACCC
Cad	PP34104	AGCCGCCATACTTCGACTG	TTATCCTTGGTGCGGGTTTTG
CG6295	PP22809	CAACGTCCTGAACCCCGTC	GACCAGCCGTGGATAGTGA
clk	PP20992	GCCTCGGAAAGTATTACCTCCC	CCATCTCATAGGCCAGGTCATA
cnc	PA60393	CTGCATCGTCATGTCTTCCAGT	AGCAAGTAGACGGAGCCAT
crp	PP11311	GGTTGCCATCAAAACGGAGGA	TCGATGTGATAGTTCTCACCCC
CycT	PP22301	CCGGCCCGTCTGAAGTCTA	CCTTGCTGTTAGCTGTCCGAT
drm	PD44490	CACCAAGCCGTACAACCTGA	ACACCTCGCACGAATAGGTG
e2f2	PP6884	AGCGCAAAACCGCGAGTAT	GCCGAATCCACCTTCATCATC
esg	PP35234	ATACCCGAAATATCCCTGGAACA	CCCTGCTGATTGATGGTCCTG
foxk	PP30383	CGGATGCCGTGACAGTAATC	CGCGACACAAGGTTGTTCTC

h	PB60086	GCGTAACAGCAGCCAACAT	CATGATGGGCTTGTTCGAC
ham	PP25145	GGATGGCTAGAGCCCACAGA	TCGCCTATACAATCGTCCTGAA
IA-2	PP20886	GCACTCCGAGGTCTGCTAC	GTCTTCTCAATGTCCTCAACGTC
irp-1a	PP232	CCAGGAGTCATTCACCCAGGA	CACATGAAAGTTGTCACAGTTGC
LManVi	PP11567	ATGCCCCAAAACCAAGACGAA	CAGCTCAGCGATTACGGTATC
mondo	PP22673	TTTATACAGCCCAGTCTTGGTCC	CAAGCGTGTGGTTGGAATCAA
Myc	PP29594	TCGCAGACGACAGATAACACC	GACAGACCGTGTAGTCCAGAT
peb	PP19111	ATTTCGTCTGAATCGCTCGG	TGCTACTGTTACCCAGATAGCC
psi	PP10665	GTGCCAGTATTACTCAGGCAAT	ATCTGCTCCTCACAGCTTGTT
rel	PD70444	GGTGATAGTGCCCTGCATGT	CCATACCCAGCAAAGGTCGT
sd	PP22380	TACGGTCGCAACGAGCTAATC	AACTGACTTGCTTCCTGGTTC
TET	PP21254	ATCCCAACTACGGTAGGTCG	CATCGTCTTATTGAGGTCCGC
trx	PD70008	AATGCGGCGCGTTTCATTAA	GTCGTAGGTAAGCTCCTCGC
vnd	PP35690	TCCCCAGTTACCTCGGAAGTG	AGCTCTTGTAATCGCCGGAAA
vnd	PP35690	TCCCCAGTTACCTCGGAAGTG	AGCTCTTGTAATCGCCGGAAA

RpL32 (gut)	PD41810	AGCATACAGGCCCAAGATCG	TGTTGTCGATACCCTTGGGC
RpL32 (blood)	PMID:24240319	ATCGGTTACGGATCGAACAA	GACAATCTCCTTGCGCTTCT
Hml	PP16200	TGGTTATGGCGGGATAAAGACG	GTTGCCCTGACTTCCCTGG
NimC1	PP27067	TGCCCAACGGTATGTGGAAAA	GGAGAAGTTCGTTTGTAGCCAT
PPO1	PP22066	TTGGAACTGCCCGATTCCTTC	TTCAGATCCACGTCCTTAGAGAA
PPO2	PP20802	GCCTGGATCTGCCATCCTTC	CACCACAAAAGACTCCTCCCG
Atilla	PD42354	CAGTGCAAATCCCTCACGGA	CGCGGATGTTAGAGGCAGAA
CG3328	PP30800	CAGACGGATCTGGGCCAGTA	GTTGCTCGGGTTGATGATGAG
Xbp1	PD70455	CTCGAGTTCGGGATACGCAT	CCAGGTTAGATGGTCCAGGC
E(spl)mbeta-HLH	PP8427	CGCCGTGCCAGGATTAACA	GGAAACTCTCAGCGATGCTAAG