Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks

Siqi Wu[a,b], Antony Joseph[a,b,c], Ann S. Hammonds[b], Susan E. Celniker[b], Bin Yu[a,d,1], and Erwin Frise[b,1]

[a]Department of Statistics, University of California, Berkeley, CA 94720; [b]Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [c]Walmart Labs, San Bruno, CA 94066; and [d]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720

Contributed by Bin Yu, March 6, 2016 (sent for review October 26, 2015; reviewed by Richard Bonneau and Michael S. Waterman)

Spatial gene expression patterns enable the detection of local covariability and are extremely useful for identifying local gene interactions during normal development. The abundance of spatial expression data in recent years has led to the modeling and analysis of regulatory networks. The inherent complexity of such data makes it a challenge to extract biological information. We developed staNMF, a method that combines a scalable implementation of nonnegative matrix factorization (NMF) with a new stability-driven model selection criterion. When applied to a set of *Drosophila* early embryonic spatial gene expression images, one of the largest datasets of its kind, staNMF identified 21 principal patterns (PP). Providing a compact yet biologically interpretable representation of *Drosophila* expression patterns, PP are comparable to a fate map generated experimentally by laser ablation and show exceptional promise as a data-driven alternative to manual annotations. Our analysis mapped genes to cell-fate programs and assigned putative biological roles to uncharacterized genes. Finally, we used the PP to generate local transcription factor regulatory networks. Spatially local correlation networks were constructed for six PP that span along the embryonic anterior–posterior axis. Using a two-tail 5% cutoff on correlation, we reproduced 10 of the 11 links in the well-studied gap gene network. The performance of PP with the *Drosophila* data suggests that staNMF provides informative decompositions and constitutes a useful computational lens through which to extract biological insight from complex and often noisy gene expression data.

principal patterns | stability selection | sparse decomposition | spatial gene expression | spatially local networks

Biological processes in multicellular organisms depend on spatial and temporal control of gene expression. Gene products function in the context of other spatially localized gene products and these interactions have been well characterized for development and tissue differentiation. Recent studies of prenatal (1) and adult human brain (2) revealed widespread anatomical variability in gene networks, which is reflective of developmental processes and of the distribution of major cell types. Spatially resolved studies of tumors uncovered widespread intratumor heterogeneity (3–8). Given the importance of spatiotemporal gene expression, many efforts are underway to characterize it genome-wide. Systematic datasets include *Drosophila* gene expression during embryogenesis [Berkeley *Drosophila* Genome Project (BDGP) (9)], and oogenesis (10), subcellular mRNA localization (11), and in brain (12), imaginal discs (13), central nervous system (14), and other developmental model systems [e.g., *Xenopus* (15), *Ciona* (16), and mouse (17–19)].

Spatial datasets are complex and quickly surpass the human ability to interpret them. To represent, search, and analyze such large spatial expression datasets, they are commonly curated with defined controlled vocabulary (9, 17–21). Curation using ontologies is time-consuming and requires expert knowledge. Despite significant progress toward automatic computer annotation through supervised learning based on human labels (22–26), the subtleties inherent in spatial expression patterns are difficult to capture and finding related patterns is challenging. An alternative, complementary to ontologies, is the spatial expression information extracted directly from images (12, 17–19, 22, 27–30). We discovered putative gene interactions by correlating gene expression and performing cluster analysis (27), and others have used sparse Gaussian graphical models (30) to do the same. Due to data complexity and the large size of image collections, image-based approaches are not routinely used for modeling.

Organ systems develop through the combinatorial action of gene regulatory networks (21, 31), and gene function and regulatory interactions can markedly differ depending on the spatial location (32). Studies of genomic enhancer elements have shown that wild-type spatial expression patterns are actually the product of multiple genomic elements. These previous studies dissected biological enhancers and discovered that complex expression patterns could be subdivided into smaller regions (33, 34). In *Drosophila*, clustering early embryonic gene expression patterns recovered groups of cells that likely interact with one another, contributing to the formation of organs and tissues (27, 33). These regions are similar to those identified in studies using laser ablation to determine cell lineage and function (35, 36). Yakoby et al. proposed an innovative method to model spatial gene expression in *Drosophila* follicle cells as a Boolean combination of smaller building blocks (10). Due to the small number of gene expression patterns in their work, they were able to produce

Significance

Despite the abundance of spatial gene expression data, extracting meaningful information to reveal how genes interact remains a challenge. We developed staNMF, a method that combines a powerful unsupervised learning algorithm, nonnegative matrix factorization (NMF), with a new stability criterion that selects the size of the dictionary or the set of principal patterns (PP). We demonstrate that PP give rise to a novel and concise representation of the *Drosophila* embryonic spatial expression patterns and they correspond to biologically meaningful regions of the *Drosophila* embryo. Furthermore, this new representation was used to automatically predict manual annotations, categorize gene expression patterns, and reconstruct the local gap gene network with high accuracy.

Author contributions: S.W., B.Y., and E.F. designed research; S.W., A.J., and B.Y. performed research; S.W., S.E.C., and B.Y. contributed new reagents/analytic tools; S.W., A.J., A.S.H., B.Y., and E.F. analyzed data; and S.W., A.S.H., S.E.C., B.Y., and E.F. wrote the paper.

Reviewers: R.B., New York University; and M.S.W., University of Southern California.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: binyu@stat.berkeley.edu or erwin@fruitfly.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1521171113/-/DCSupplemental.

building blocks manually. Such an approach is intuitive and conceptually supported by the aforementioned works on genomic enhancers.

In this paper, we describe stability-driven nonnegative matrix factorization (staNMF), a method that interprets, represents, and analyzes comprehensive spatial gene expression datasets. staNMF partitions biological spatial data or images into spatial building blocks, called principal patterns (PP). Specifically, we adapted a powerful unsupervised learning algorithm, NMF (37), to learn data-driven representations from large and complex datasets. We invented a new stability criterion to address the challenge of NMF model selection to arrive at staNMF, which is scalable, tolerates experimental noise, and moderates image registration variance. We applied staNMF to a dataset of spatial gene expression images during early *Drosophila* embryogenesis. The output of staNMF, i.e., PP, is humanly interpretable and biologically meaningful. Using the PP, we grouped genes into overlapping categories corresponding to regions involved in coherent developmental programs. Finally, we built spatially local networks based on the learned PP, correctly reproducing 10 of 11 links in the well-studied gap gene network. This PP-based approach can be applied to extract biological insight from other complex and noisy spatial gene expression datasets, for example, from the extant zebrafish and mouse brain studies.

Results

Learned staNMF PP Correspond to Biologically Important Regions. We used a set of BDGP *Drosophila* lateral embryonic gene expression images at developmental stages 4–6 (1 h 20 min–3 h after egg laying at 25°C). The 1,640 images were derived from 701 genes with spatially restricted expression patterns. We extracted embryos using segmentation of images taken with differential interference contrast microscopy. The objects containing the embryos were resized in both axes, and registered to a predefined elliptical template. The intensities of the gene expression patterns were determined with a least-squares approach to distinguish the dye intensity from the background introduced by the imaging modality (*SI Appendix*). Each resulting image is an ellipse of 405 pixels with gene expression values between 0 and 1 (Fig. 1*A*).

In *Drosophila* development, cell fates are determined before any visible morphological features become apparent (35, 36) and are preceded by the coordinated coexpression of cohorts of genes in defined spatial regimes that divide the embryo into areas with unique regulatory profiles (27, 33). We model the embryo as a topological map where genes are either absent or expressed at a positive value. Thus, we think of each spatial gene expression as an additive and nonnegative linear combination of a set of regions of the embryo. To identify these additive and positively valued regions, we used NMF (37). For a given positive integer K, NMF finds a data-driven dictionary such that each expression image can be represented by a nonnegative linear combination of the K dictionary columns (Fig. 1*B*). We converted the pixel intensities of the preprocessed expression pattern into a linear vector and decomposed the vector with NMF, aiming to solve the following nonconvex optimization problem:

$$\min_{\mathbf{D}\geq 0,\mathbf{A}\geq 0} \|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2,$$

where $\mathbf{D}$ is the sought-after dictionary, $\mathbf{X}$ the linearized image data, $\mathbf{A}$ a linear nonnegative coefficient matrix, and $\|.\|_F$ the matrix Frobenius norm. To account for replicate images of the same gene, we introduced a weight factor (*SI Appendix*). The columns of the dictionary are the PP. The nonnegativity constraints on both the dictionary and coefficients enforce the PP to have nonnegative contributions to the gene image, resulting in a "parts-based" representation. The constraints also implicitly impose sparsity on both the PP and the linear coefficients.

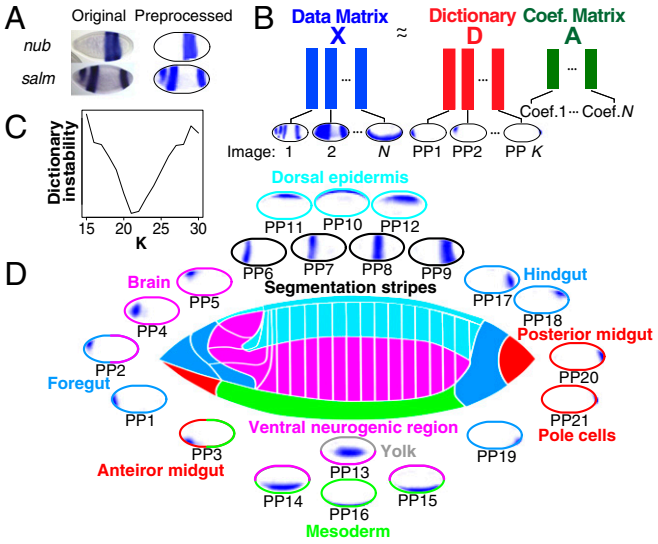

Fig. 1. Learning PP by staNMF from spatial gene expression patterns. (*A*) Expression patterns of two genes, *nub* and *salm*, in *Drosophila* embryos. (*B*) For a given number K, NMF factorizes the nonnegative data matrix $\mathbf{X}$, the columns of which are gene expression images, into the product of two nonnegative matrices: dictionary $\mathbf{D}$, which contains the K PP, and coefficient matrix $\mathbf{A}$, which contains the nonnegative coefficients of the images. (*C*) staNMF identified $K=21$ to be the optimal number of PP for $15\leq K\leq 30$. (*D*) The *Drosophila* fate map (center) (35, 36), surrounded by the 21 PP learned by staNMF. The PP are arrayed according to the corresponding regions of the fate map. See *SI Appendix* for how we mapped the PP to the fate map.

Our aim was a generalizable method contingent only on data and with little need of prior knowledge. NMF depends on a single parameter, the number K of PP. We reasoned that a useful definition of an optimal NMF-generated dictionary would be reproducibly independent of the initialization values. To identify the number of PP, we optimized on a metric that measures the instability of the learned PP relative to the initial starting points. In detail, for each K, the NMF algorithm was repeated multiple times with an initial PP dictionary randomly sampled from the columns of $\mathbf{X}$. We measured the instability of the PP by computing the average dissimilarity of all learned dictionary pairs ($\mathbf{D}$ and $\mathbf{D}'$) using their cross-correlation matrix ($\mathbf{C}$) and a new Amari-type quantity (38):

$$diss(\mathbf{D},\mathbf{D}')=\frac{1}{2K}\left(2K-\sum_{j=1}^{K}\max_{1\leq k\leq K}\mathbf{C}_{kj}-\sum_{k=1}^{K}\max_{1\leq j\leq K}\mathbf{C}_{kj}\right).$$

We repeated the process for each K and selected the K where the learned dictionaries achieve the lowest instability (*SI Appendix*). Among all dictionaries with the optimal size K, the dictionary with the minimum NMF objective function value was chosen. We called this stability-based NMF model selection method "staNMF" and validated our method with multiple synthetic datasets (*SI Appendix*, Figs. S4, S6, and S7). When applied to our spatial gene expression pattern dataset with range $15\leq K\leq 30$, staNMF identified the number of PP with the lowest instability at $K=21$ (Fig. 1*C*; see also *SI Appendix*).

The 21 learned PP divided the *Drosophila* embryo into contiguous pretissue and organ regions (Fig. 1*D*). Compared with principal component analysis, independent component analysis (39), and a recently proposed sparse Bayesian factor model (24), only PP recapitulate the underlying biology of cell and tissue fate maps (*SI Appendix*, Figs. S8 and S9). Each PP is spatially coherent: The intensity is locally continuous and the regions defined by the PP are interconnected. We grouped the 21 PP into four

categories: PP1–5: anterior patterns; PP6–9: vertical (gap) segmentation stripes; PP10–16: horizontal ventral–dorsal patterns; and PP17–21: posterior patterns. We compared the PP and the categories to the *Drosophila* fate map (35, 36), an experimentally determined functional mapping of spatial regions before availability of gene expression data. We associated each PP or a group of PP with a region in the fate map of similar size and shape (Fig. 1*D*; see also *SI Appendix*). We found that the PP refined the fate map in the dorsal epidermal region, the ventral neurogenic region, the mesoderm, and the hindgut. Some of the refinements are already biologically supported. For example, the vertical stripes are known to be the result of gap, pair-rule, polarity, and segmentation genes that eventually establish 14 refined stripes that become morphologically distinguishable in a later-stage embryo (40).

l_1 Regularization to Provide Sparse PP Data Representations. We evaluated the ability of PP to provide a compact representation for spatial gene expression patterns. A sparse decomposition of complex expression patterns into additive smaller components offers a simple and intuitive computational representation of spatial gene expression (Fig. 2). Whereas the nonzero coefficients of the matrix **A** (Fig. 1*B*) provided such a decomposition, they tended to select more PP than necessary in our simulations (*SI Appendix*, Fig. S11). Instead, we used the least absolute shrinkage and selection operator (LASSO) (41) for PP selection. Nonnegative constraints were put on the linear coefficients. For each expression pattern, we chose the LASSO regularization parameter using a 10-fold cross-validation and refitted the coefficients with nonnegative least squares on the selected PP (*SI Appendix*). We call these coefficients "sparse PP" (sPP) coefficients or representation. The average number of PP chosen by this procedure is 10.4, and the average correlation between the original expression pattern and the reconstructed pattern is 0.854 (*SI Appendix*, Fig. S12 *A* and *B*). Considering the small number of the selected PP, the correlation measure indicates that our model selection and fitting procedure achieved a reasonably good reconstruction quality. As expected, the correlation increases as the number of PP increases (*SI Appendix*, Fig. S12*C*). We investigated cases with poor performance and found such gene expression patterns are either faint or have poorly defined boundaries. In addition, nonsparse representations almost always correspond to ubiquitously expressed genes (*SI Appendix*, Fig. S12*D*). As illustrated by the residual images, errors are most likely to occur at expression pattern boundaries (Fig. 2).

PP Provide a Data-Driven Alternative to Human Expert Annotations. Expert curators annotated BDGP spatial gene expression patterns with a controlled vocabulary whose terms represent anatomical regions of the developing embryo, similar to the fate map

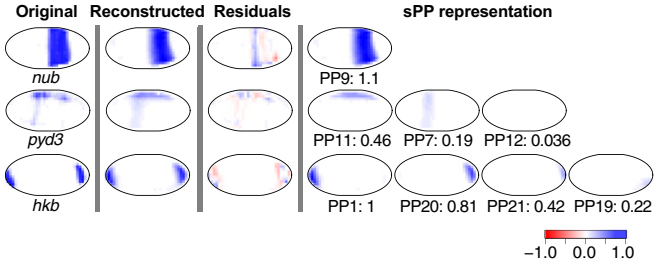

Fig. 2. Sparse decomposition of spatial gene expression patterns using the 21 PP. Shown are a sample of three gene expression patterns (original), their reconstructed patterns using the sPP representation (reconstructed), the difference between the original and the reconstructed patterns (residuals) and the contributions from the 21 PP (sPP representation).

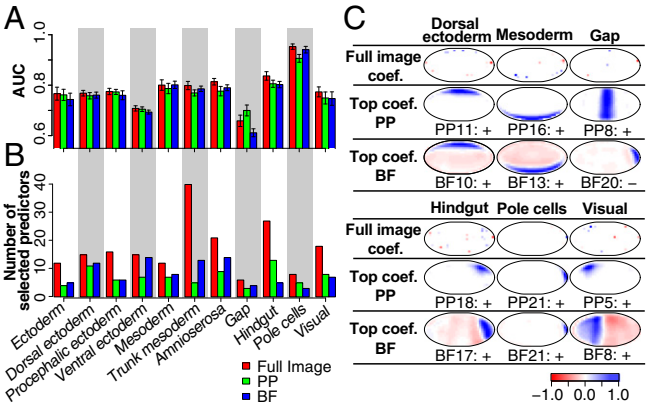

Fig. 3. Predicting annotation terms based on 405 image pixels, the sPP, and the BF (sparse BF) representations. (*A*) Prediction accuracy as evaluated by the AUC value. Data are expressed as mean ± SEM. (*B*) Number of selected predictors in the optimal model. (*C*) Interpretability of the L1LR under all three representations. The pixel-based full image representation: all coefficients are shown as pixel values within the embryo; the sPP-based and the BF-based representations: only the PP or BF that corresponds to the largest L1LR coefficient is shown. "+": the largest L1LR coefficient is positive; "−": the largest L1LR coefficient is negative.

discussed above. To compare the 21 learned PP with the anatomical vocabulary, we used the sPP coefficients as predictors in a supervised learning approach and labeled each image with annotation terms. We selected 11 stage 4–6 annotation terms with more than 100 images: ectoderm anlage in statu nascendi (AISN), dorsal ectoderm AISN, procephalic ectoderm AISN, ventral ectoderm AISN, mesoderm AISN, trunk mesoderm AISN, amnioserosa AISN, gap, hindgut AISN, pole cells, and visual AISN. For each of the 11 terms, we labeled images annotated this term as "1," the rest as "0," and fitted an l_1-penalized logistic regression (L1LR) with the sPP coefficients as predictors. A 10-fold cross-validation was performed for each term to choose the regularization parameter in the L1LR. To compare with sPP, we also trained L1LR using the full expression pattern with 405 pixels, and the sparse Bayesian factors (BF) of ref. 24 (*SI Appendix*). We generated 21 BF to compare directly to the 21 PP.

The prediction performance of the three methods is very similar, as measured by the cross-validation AUC (area under the receiver operating characteristic curve, valued between 0 and 1) (Fig. 3*A*). On average, the AUC value for the sPP representation is 0.772, compared with 0.787 for the pixel-based representation and 0.767 for the BF representation. Taking into account the SE of the AUC for each annotation term, none of the three methods significantly outperforms the others. In terms of model complexity, on average 17 predictors are selected for the pixel-based L1LR, 7 for our sPP-based approach, and 8 for the BF-based model (Fig. 3*B*).

Among the three L1LR models, the sPP-based model is the most interpretable and most biologically meaningful (Fig. 3*C*). For the pixel-based model, we created a visualization of the 405 predictors for each annotation term by plotting the L1LR coefficient values as pixels in our elliptic embryo shape. To compare with this visualization, we selected the (top L1LR) PP and BF corresponding to the largest L1LR coefficients for their respective L1R1 models. The pixel-based predictors consist of scattered points and the top L1LR BF contains negative values, both of which are difficult to interpret. In contrast, the top L1LR PP consistently showed the annotation term exactly as a curator would annotate the gene expression (Fig. 3*C*; see *SI Appendix*, Figs. S13–S17 for all L1R1 results as evidence for the benefits of

the sPP models). In addition, the pixel-based and the BF-based representations resulted in unstable predictor sets between cross-validation runs (*SI Appendix*, Fig. S18). This instability further reduces the interpretability of the two models. In the automatic label prediction task, the better interpretability of the top L1LR PP and the comparable prediction accuracy of sPP-based models provided additional support for the PP as a biologically attractive decomposition.

PP Associated Gene Functions and Relationships Between Spatial Regions. We use the term "function" as defined by the experimentally generated fate map that describes the locations of larval/adult progenitor cells in the blastoderm. These cells give rise to particular tissues and organs during development. By systematically associating genes to PP, we can group known and uncharacterized genes and assign putative biological roles. We assigned a gene to PP category k if the kth sPP coefficients of the gene exceeded 0.1. The number of genes in each of the 21 PP categories is, on average, 300 genes ranging from 184 to 395. PP categories 6–9 contain fewer, on average, 223 genes (Fig. 4*A*, *Right* and *SI Appendix*, Fig. S19 and Tables S2 and S3). In addition, we also found a significant presence of previously uncharacterized computed genes (CG) in all PP categories: The average percentage of CG per PP category is 23.4%.

To directly relate genes to each other, we created a heatmap visualization of the sPP coefficients for 667 genes that belong to at least one PP category. We ordered the genes by first associating each of them to the PP with the maximum sPP coefficient, and then performing a hierarchical clustering of the genes assigned to the same PP (Fig. 4*A*, *Left*; see also *SI Appendix*). A surprisingly large fraction of genes (17.8%) exhibit their strongest expression in PP21 (pole cells) and have limited expression in other PP. We found that only 5.8% of the 156 transcription factors are among these PP21 specific genes, confirming previous results (21). Of the 667 genes, 4.5% have their strongest expression in segmentation patterns PP6–9, suggesting that only a small number of genes are dedicated to segmentation. Furthermore, 93.3% of these genes have been characterized, implying that we know most segmentation genes. We found genes with known roles in foregut development (*croc*, *hkb*, and *kni*) associated with PP1, segmentation specific genes (*Dfd*, *kn*, *Kr*, and *tsh*) associated with PP6–9, genes essential for mesoderm/ectoderm development (*Mes2*, *sna*, and *sog*) associated with PP15, genes essential for pole-cell formation associated with posterior PP21 (*lok*, *pgc*, and *rdx*) as well as previously uncharacterized genes such as *CG1663*, *CG8289*, *CG9514*, and *CG10479* in these PP categories (Fig. 4*B*). With additional later-stage organ system annotation data (21), we found genes expressed in PP16 (mesoderm) in stages 4–6 become expressed in the central nervous system (CNS) starting at stage 9 (*trx*, *sna*, *Traf4*, and *Caf1*). Early mesoderm genes with function during CNS development have been shown before (42), but here we demonstrate a systematic secondary function of mesoderm specific genes, including previously uncharacterized genes (e.g., *CG11247*).

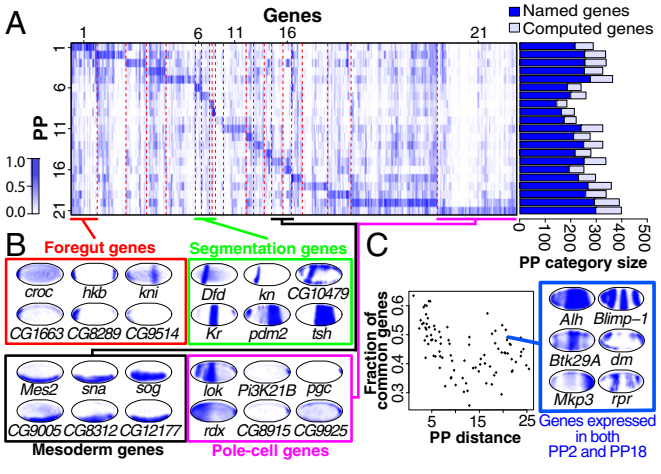

Fig. 4. PP-based gene categorization. (*A*) Left heatmap: PP expression profile of genes. Each column corresponds to the sPP coefficients of one gene. Between the red dashed lines are the genes with the strongest expression in the same PP (indicated by numbers on horizontal axis). Right barplot: numbers of named and computed genes (CG) in each PP category. (*B*) Genes with known functions and CG were found in the associated PP categories. (*C*) The relationship between the fraction of common genes in a pair of PP categories and the centroid distance of the two PP, for PP1–9, PP17–21. Each dot in the plot corresponds to a PP pair. Shown also are six genes expressed in both PP2 (brain/foregut) and PP18 (hindgut), a pair of distant PP.

Next, we investigated the relationship between the PP that span the anterior–posterior axis, i.e., PP1–9, PP17–21. We plotted the fraction of common genes in a pair of PP categories, defined as the Jaccard distance between the two categories, in relation to the pairwise PP centroid distance (Fig. 4*C* and *SI Appendix*, Fig. S20). Our results show that when the PP distance is small, the fraction of common genes is high. However, after the initial decrease, the fraction of common genes increases as the PP distance increases. An example is the set of genes (49% or 227) shared between the distant PP2 and PP18 that map to anterior foregut/brain and posterior hindgut (Fig. 4*C*). These genes include known foregut and hindgut development genes such as *Alh*, *Blimp-1*, *Btk29A*, *dm*, *Mkp3*, and *rpr*. This finding substantiates the previously identified common origins and gene expression signatures of foregut and hindgut that were based on manual annotations (21, 36). Similarly, 229 genes (52%) are shared between PP3 (anterior midgut/mesoderm) and PP19 (hindgut), including known midgut and hindgut genes, *ry*, *Ect4*, *Sdc*, *Pcl*, *larp*, and *emc*, suggesting a more general link between the anterior and posterior patterns.

PP-Based Correlation Network Inference Leads to Accurate de Novo Reconstruction of the *Drosophila* Gap Gene Network. Associations between two genes are routinely described by their correlation to each other (43). In terms of spatial relationships, positive gene interactions exhibit spatial overlap whereas repressive gene interactions exhibit spatial exclusivity. Below, we used the learned PP to construct spatially local correlation networks (SLCN) for 156 transcription factors (TF) from our expression data.

The *Drosophila* gap gene network has been studied for decades (40, 44, 45). It controls embryonic patterning by regulating the genes required to establish the anterior–posterior segmentation stripes and is primarily driven by well-studied activating and repressive interactions between eight TF. To reconstruct this network solely from our expression data of 156 spatially restricted TF, we selected six PP (PP6–9, PP17, and PP20) corresponding to the domains of the gap gene network. We called these six PP "gap-PP." For each gap-PP, we computationally constructed an SLCN. First, we identified all TF with sPP coefficient greater than 0.1 in the gap-PP, or its directly adjacent PP. Next we computed the weighted correlation between all pairs of selected TF using the pixel intensities in the gap-PP as weights. Finally, links of each resulting correlation network were filtered by thresholding the weighted correlations at a two-tail 5% cutoff, producing six SLCN with interactions among known and previously uncharacterized genes (*SI Appendix*, Figs. S21–S24). See *SI Appendix* for details of the SLCN construction.

We evaluated our SLCN construction by comparing interactions found in the six SLCN to known regulatory interactions of selected trunk and terminal gap genes, *giant*(*gt*), *hunchback*(*hb*), *knirps*(*kni*), *Krüppel*(*Kr*), *huckebein*(*hkb*), and *tailless*(*tll*). We compared the subnetworks of the SLCN containing only the six genes (Fig. 5*A*) to a schematic network diagram (Fig. 5*B*), as originally depicted in ref. 44. Although the diagram indicates that some gene interactions are contingent on spatial position, it

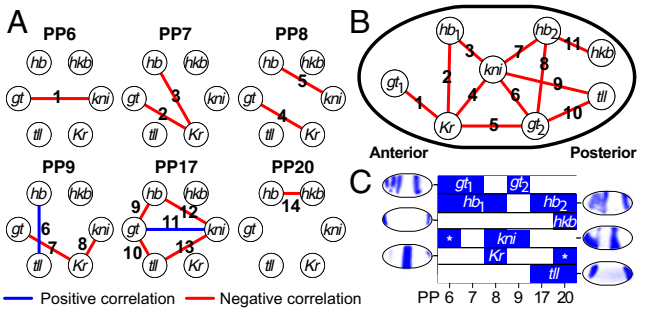

Fig. 5. Modeling and validation of the *Drosophila* gap gene network with SLCN. (*A*) The SLCN for six gap genes. For each of the six gap-PP, the subnetwork of the SLCN that contains the six gap genes is shown. Links are numbered from 1 to 14. (*B*) The gap gene network diagram depicting repressive interactions of six genes (44). Links are numbered from 1 to 11 and multiple occurrences of the same gene are subscripted by numbers (e.g., hb_1 and hb_2). The directions of the interactions are not indicated. (*C*) Expression patterns of the six gap genes and their linearly ordered PP representation. For each gene, the regions depicted in blue are the gap-PP with sPP coefficient greater than or equal to 0.1. The "*" symbol indicates a region of gene expression with no match in *B*.

does not provide precise locations of the interactions. To compare with our networks, we devised a method to match the links in the diagram to our SLCN. For each gap gene, we first created a linearly ordered PP representation by placing the six gap-PP anterior to posterior and associating a gap-PP to the gene if the sPP coefficient for the gap-PP exceeded a threshold of 0.1 (Fig. 5*C*). The gap-PP associated with each gap gene were then merged into one or more connected PP groups. Based on its relative location in the diagram, we then matched each gene node in the schematic diagram to a connected PP group for the same gene. We considered an interaction between two gene nodes in the schematic network diagram as successfully identified by our method if the same interaction exists in the any SLCN associated with the overlapping PP in the connected PP groups of the two gene nodes (*SI Appendix*, Fig. S25 and Table S4).

For example, the diagram depicts a repressive link between the anterior component of *gt* (i.e., gt_1) and *Kr*. Using our linearly ordered PP representation, we found the connected PP groups for gt_1 and *Kr* are PP6/7 and PP8, respectively (Fig. 5*C*). We searched for the *gt*-*Kr* interaction only in the SLCN of PP7 and PP8, because PP6 and PP8 do not overlap. In both networks, we found a repressive interaction (or negative correlation). Hence we considered the anterior *gt*-*Kr* link of the schematic gap gene network diagram as being identified with our model. See *SI Appendix*, Fig. S25 for the validation of the remaining links.

For the six gap genes, our SLCN reconstruction identified 14 interactions (Fig. 5*A*). Eight out of 11 links in the gap gene network diagram have a one-to-one mapping with 8 of the 14 SLCN interactions. In addition, the two *gt*-*Kr* links in the gap gene network (link 1 and 5 in Fig. 5*B*) are found in the SLCN of PP7–9 (links 2, 4, and 7 in Fig. 5*A*). The remaining *kni*-gt_2 link (link 6 in Fig. 5*B*) has no corresponding link in the SLCN. Three of the 14 SLCN links do not correspond to any interactions in the network diagram: the two *gt*-*kni* links in PP6 and PP17 (link 1 and link 11 in Fig. 5*A*), and the *hb*-*tll* link in PP9 (link 6 in Fig. 5*A*). Therefore, our SLCN recovered 10 out of 11 interactions in the gap gene network and discovered 3 interactions not described by the diagram. In contrast, when using correlation over the whole embryo to construct the TF network, we recovered only three out of nine unique links of the gap gene network (*SI Appendix*, Fig. S27).

Discussion

We proposed staNMF that combines NMF with a new stability-based model selection criterion to decompose spatial gene expression patterns into local PP. When we applied staNMF to *Drosophila* embryonic expression data at early stages 4–6, the learned PP correspond to preorgan regions, and thus provide an informative representation of spatial gene expression data. We demonstrated that PP are a data-driven alternative to manual curation and facilitate the categorization of gene expression patterns. Our PP-based sparse representations (sPP) reduce large datasets to manageable scales. They allow suitable human interrogation and downstream computation on desktop computers while preserving quantitative relationships of full datasets. In addition, staNMF's utility was further substantiated by the agreement between our PP-based spatially local networks and the well-studied gap gene network.

Model selection or identification of a well-reasoned number of components for unsupervised learning has been a challenging problem. staNMF's underlying idea of stability was previously used to identify the number of clusters in cluster analysis using NMF (46). Our contribution is to use an Amari-type measure to evaluate dictionary stability, rather than clustering stability as in the previous work. We experimented with the method of ref. 46, and found that it failed to identify the correct number of PP in a number of synthetic datasets (*SI Appendix*). Recent Bayesian model selection approaches (47, 48) introduced additional hyperparameters, which in practice are generally not known in advance. We believe staNMF is an important advance because it does not depend on tuning parameters, and has been demonstrated to work well in both simulations and our *Drosophila* spatial gene expression data.

Our SLCN identified three network links previously not described in ref. 44. In PP6, we found a repression link between *gt* and *kni*. Gene expression images of *gt* and *kni* revealed a clear complementary pattern toward the anterior end with a negative local correlation of –0.720 in PP6 (*SI Appendix*, Fig. S25*D*). In the PP17 SLCN, an activation link between *kni* and *gt* was identified. Because our images covered an interval of around 1.5 h, the posterior part of *kni* expression pattern at the early developmental stages 4–6 might have been aligned to the *gt* gene posterior end at a later time point (*SI Appendix*, Fig. S25*D*). Experiments will be needed to confirm or refute these predicted links. Finally, although not described in ref. 44, the predicted *hb*-*tll* activation link in PP9 is supported by ref. 49.

Given the successes of our PP-based approach in the well-characterized early *Drosophila* embryo, we expect staNMF to be broadly applicable to derive meaningful data-driven representations of spatial gene expression for other systems such as zebrafish, *Caenorhabditis elegans*, and human histological samples. In conclusion, we have demonstrated with ample evidence the utility of sPP as an effective computational lens to reveal hidden structures in complex gene expression data.

Datasets and Software

Data are available as Datasets S1–S6, and code and datasets are provided under "Principal Patterns" on our website: insitu.fruitfly.org/downloads.

ACKNOWLEDGMENTS. We are indebted to BDGP members for their advice. We thank Ben Brown, Julien Mairal, and Sivaraman Balakrishnan for helpful discussions and comments. B.Y. acknowledges partial research support from National Science Foundation (NSF) Grants DMS-1107000, CDS&E-MSS 1228246, and DMS-1160319 (Focused Research Groups in the Mathematical Sciences), Army Research Office Grant W911NF-11-1-0114, Air Force Office of Scientific Research Grant FA9550-14-1-0016, and the Center for Science of Information, an NSF Science and Technology Center, under Grant Agreement CCF-0939370. National Institutes of Health Grants R01 GM076655 (to S.E.C. and A.S.H.) and R01 GM097231 (to E.F., S.E.C., S.W., and A.J.) supported this work. Work at Lawrence Berkeley National Laboratory was conducted under Department of Energy Contract DEAC02-05CH11231. S.W. was partially supported by National Human Genome Research Institute Grant 1U01HG007031-01, and Citadel Fellowship in the Department of Statistics at University of California, Berkeley.

1. Miller JA, et al. (2014) Transcriptional landscape of the prenatal human brain. *Nature* 508(7495):199–206.
2. Hawrylycz MJ, et al. (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489(7416):391–399.
3. Almendro V, et al. (2014) Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell Reports* 6(3):514–527.
4. Almendro V, et al. (2014) Genetic and phenotypic diversity in breast tumor metastases. *Cancer Res* 74(5):1338–1348.
5. Bedard PL, Hansen AR, Ratain MJ, Siu LL (2013) Tumour heterogeneity in the clinic. *Nature* 501(7467):355–364.
6. Gerlinger M, et al. (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 366(10):883–892.
7. de Bruin EC, et al. (2014) Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* 346(6206):251–256.
8. Zhang J, et al. (2014) Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* 346(6206):256–259.
9. Tomancak P, et al. (2002) Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome Biol* 3(12):research0088.1-88.14.
10. Yakoby N, et al. (2008) A combinatorial code for pattern formation in Drosophila oogenesis. *Dev Cell* 15(5):725–737.
11. Lécuyer E, et al. (2007) Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* 131(1):174–187.
12. Jenett A, et al. (2012) A GAL4-driver line resource for Drosophila neurobiology. *Cell Reports* 2(4):991–1001.
13. Jory A, et al. (2012) A survey of 6,300 genomic fragments for cis-regulatory activity in the imaginal discs of Drosophila melanogaster. *Cell Reports* 2(4):1014–1024.
14. Manning L, et al. (2012) A resource for manipulating gene expression and analyzing cis-regulatory modules in the Drosophila CNS. *Cell Reports* 2(4):1002–1013.
15. Pollet N, et al. (2005) An atlas of differential gene expression during early Xenopus embryogenesis. *Mech Dev* 122(3):365–439.
16. Imai KS, Hino K, Yagi K, Satoh N, Satou Y (2004) Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: Towards a comprehensive understanding of gene networks. *Development* 131(16):4047–4058.
17. Smith CM, et al. (2007) The mouse gene expression database (gxd): 2007 update. *Nucleic Acids Res* 35(Database issue):D618–D623.
18. Richardson L, et al. (2009) Emage mouse embryo spatial gene expression database: 2010 update. *Nucleic Acids Res* 38(Database issue):D703–D709.
19. Lein ES, et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445(7124):168–176.
20. Tomancak P, et al. (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol* 8(7):R145.
21. Hammonds AS, et al. (2013) Spatial expression of transcription factors in Drosophila embryonic organ development. *Genome Biol* 14(12):R140.
22. Peng H, et al. (2007) Automatic image analysis for gene expression patterns of fly embryos. *BMC Cell Biol* 8(Suppl 1):S7.
23. Zhou J, Peng H (2007) Automatic recognition and annotation of gene expression patterns of fly embryos. *Bioinformatics* 23(5):589–596.
24. Pruteanu-Malinici I, Mace DL, Ohler U (2011) Automatic annotation of spatial expression patterns via sparse Bayesian factor models. *PLOS Comput Biol* 7(7):e1002098.
25. Pruteanu-Malinici I, Majoros WH, Ohler U (2013) Automated annotation of gene expression image sequences via non-parametric factor analysis and conditional random fields. *Bioinformatics* 29(13):i27–i35.
26. Yuan L, et al. (2014) Automated annotation of developmental stages of Drosophila embryos in images containing spatial patterns of expression. *Bioinformatics* 30(2): 266–273.
27. Frise E, Hammonds AS, Celniker SE (2010) Systematic image-driven analysis of the spatial Drosophila embryonic expression landscape. *Mol Syst Biol* 6:345.
28. Puniyani K, Faloutsos C, Xing EP (2010) SPEX2: Automated concise extraction of spatial gene expression patterns from Fly embryo ISH images. *Bioinformatics* 26(12): i47–i56.
29. Mace DL, Varnado N, Zhang W, Frise E, Ohler U (2010) Extraction and comparison of gene expression patterns from 2D RNA in situ hybridization images. *Bioinformatics* 26(6):761–769.
30. Puniyani K, Xing EP (2013) GINI: From ISH images to gene interaction networks. *PLOS Comput Biol* 9(10):e1003227.
31. Adryan B, Teichmann SA (2010) The developmental expression dynamics of Drosophila melanogaster transcription factors. *Genome Biol* 11(4):R40.
32. Schulz C, Tautz D (1994) Autonomous concentration-dependent activation and repression of Krüppel by hunchback in the Drosophila embryo. *Development* 120(10): 3043–3049.
33. Kvon EZ, et al. (2014) Genome-scale functional characterization of Drosophila developmental enhancers in vivo. *Nature* 512(7512):91–95.
34. Stanojevic D, Small S, Levine M (1991) Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science* 254(5036): 1385–1387.
35. Lohs-Schardin M, Cremer C, Nüsslein-Volhard C (1979) A fate map for the larval epidermis of Drosophila melanogaster: Localized cuticle defects following irradiation of the blastoderm with an ultraviolet laser microbeam. *Dev Biol* 73(2):239–255.
36. Hartenstein V (1993) *Atlas of Drosophila Development* (Cold Spring Harbor Laboratory Press, Plainview, NY).
37. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.
38. Amari S, Cichocki A, Yang H (1996) A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, MA), pp 757–763.
39. Hyvärinen A (1999) Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans Neural Networks* 10(3):626–634.
40. Nüsslein-Volhard C, Wieschaus E (1980) Mutations affecting segment number and polarity in Drosophila. *Nature* 287(5785):795–801.
41. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J Royal Statist Soc B* 58(1):267–288.
42. Ashraf SI, Hu X, Roote J, Ip YT (1999) The mesoderm determinant snail collaborates with related zinc-finger proteins to control Drosophila neurogenesis. *EMBO J* 18(22): 6426–6438.
43. Wang YX, Huang H (2014) Review on statistical methods for gene network reconstruction using expression data. *J Theor Biol* 362:53–61.
44. Jaeger J (2011) The gap gene network. *Cell Mol Life Sci* 68(2):243–274.
45. Nüsslein-Volhard C, Kluding H, Jürgens G (1985) Genes affecting the segmental subdivision of the *Drosophila* embryo. *Cold Spring Harb Symp Quant Biol* 50:145–154.
46. Brunet JP, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci USA* 101(12):4164–4169.
47. Tan VY, Févotte C (2013) Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Trans Pattern Anal Mach Intell* 35(7): 1592–1605.
48. Sun M, Zhang X (2015) A stable approach for model order selection in nonnegative matrix factorization. *Pattern Recognit Lett* 54:97–102.
49. Margolis JS, et al. (1995) Posterior stripe expression of hunchback is driven from two promoters by a common enhancer element. *Development* 121(9):3067–3077.