© 2014. Published by The Company of Biologists Ltd | Development (2014) 141, 1161-1166 doi:10.1242/dev.105288

RESEARCH REPORT

TECHNIQUES AND RESOURCES

BLIND ordering of large-scale transcriptomic developmental timecourses

Leon Anavy[1], Michal Levin[1], Sally Khair[1], Nagayasu Nakanishi[2], Selene L. Fernandez-Valverde[2], Bernard M. Degnan[2] and Itai Yanai[1,*]

ABSTRACT

RNA-Seq enables the efficient transcriptome sequencing of many samples from small amounts of material, but the analysis of these data remains challenging. In particular, in developmental studies, RNA-Seq is challenged by the morphological staging of samples, such as embryos, since these often lack clear markers at any particular stage. In such cases, the automatic identification of the stage of a sample would enable previously infeasible experimental designs. Here we present the 'basic linear index determination of transcriptomes' (BLIND) method for ordering samples comprising different developmental stages. The method is an implementation of a traveling salesman algorithm to order the transcriptomes according to their inter-relationships as defined by principal components analysis. To establish the direction of the ordered samples, we show that an appropriate indicator is the entropy of transcriptomic gene expression levels, which increases over developmental time. Using BLIND, we correctly recover the annotated order of previously published embryonic transcriptomic timecourses for frog, mosquito, fly and zebrafish. We further demonstrate the efficacy of BLIND by collecting 59 embryos of the sponge *Amphimedon queenslandica* and ordering their transcriptomes according to developmental stage. BLIND is thus useful in establishing the temporal order of samples within large datasets and is of particular relevance to the study of organisms with asynchronous development and when morphological staging is difficult.

KEY WORDS: *Amphimedon* transcriptomic timecourse, Single-embryo RNA-Seq, Developmental timecourse, Large-scale datasets, Principal components analysis, Traveling salesman problem

INTRODUCTION

High-throughput sequencing methods have produced two important innovations for the analysis of transcriptomes: the amount of RNA starting material required has dropped to as little as a single cell or lower (Hashimshony et al., 2012; Islam et al., 2011; Ramsköld et al., 2012) and the number of samples that may be affordably processed is dramatically higher owing to the inherent multiplexed nature of the available methods (Hashimshony et al., 2012; Islam et al., 2011). These innovations allow for high-resolution analysis of gene expression, but also markedly impact on how these high-throughput experiments are designed.

[1]Department of Biology, Technion – Israel Institute of Technology, Haifa 32000, Israel. [2]Centre for Marine Science, School of Biological Science, The University of Queensland, Brisbane, QLD 4072, Australia.

*Author for correspondence (yanai@technion.ac.il)

Received 23 October 2013; Accepted 17 December 2013

The construction of a coherent transcriptomic timecourse typically requires a staging process in which the developmental stages of the samples must first be determined and then, typically, collected as pools to increase the starting amounts (Levin et al., 2012; Yanai et al., 2011). For synchronous processes, staging relies on the sampling time, whereas for asynchronous processes it is necessary to stage by morphology, which can be difficult and time consuming. These constraints limit the use of transcriptomic timecourse analyses in biological processes missing either visual morphological markers or synchronous development. In these cases, a method is required allowing for the random collection of many transcriptomes (i.e. embryos) followed by the determination of their developmental order at the analysis stage. Here, we present BLIND, a method for the analysis of large and complex transcriptomes and demonstrate its ability to accurately infer developmental ordering of transcriptomic timecourses.

RESULTS AND DISCUSSION

Developmental transcriptomes form a path in the principal components plane

From an analysis of previously published transcriptomic developmental timecourses of frog, mosquito, fly and zebrafish (Akbari et al., 2013; Lott et al., 2011; Yanai et al., 2011; Yang et al., 2013), we observed that the samples can be ordered from the transcriptomes alone (Fig. 1; supplementary material Fig. S1). For each timecourse we applied principal components analysis (PCA), a linear method that enables the reduction of the dataset dimensionality to a few 'principal components' that capture as much of the variation as possible. Fig. 1A shows the first two principal components of 14 transcriptomes from embryonic stages of the frog *Xenopus laevis*. The position of the samples in the PCA plane can be viewed as a path representing the progress of embryonic development. The same phenomenon is observed for developmental transcriptomes in the timecourses of the other species (supplementary material Fig. S1).

Based upon this observation, we developed BLIND for the basic linear index determination of transcriptomes. BLIND considers the distance between every two samples on the principal components plane as the developmental distance between the samples. If the distances are faithfully representative then the shortest path through the samples corresponds to the developmental progress across the samples. Finding the shortest path on the principal components plane is an instance of the general traveling salesman problem (Held and Karp, 1970). This problem has been shown to be a non-deterministic polynomial-time hard problem (NP-hard) and therefore the optimal solution cannot be retrieved in polynomial time (Papadimitriou, 1977). For an approximation, BLIND invokes a genetic algorithm that 'evolves' a path by starting with a random set of possible paths and iteratively selecting the shorter ones to combine and generate a new set (Larrañaga et al., 1999). The resulting path is inferred as the developmental order of the samples

Development (2014) doi:10.1242/dev.105288

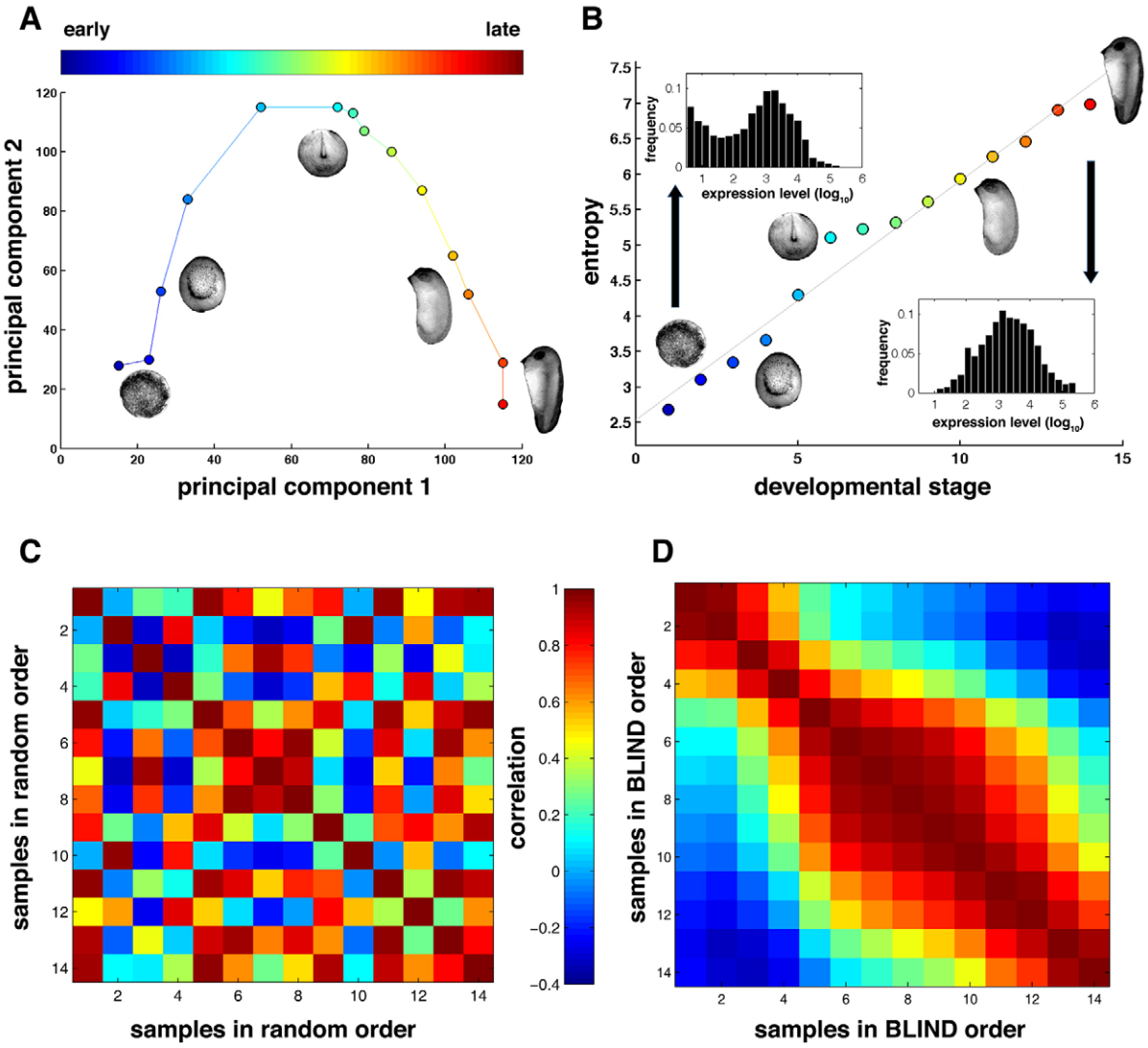

Fig. 1. The BLIND method for ordering transcriptomic timecourses. (A) Principal components analysis (PCA) on a developmental timecourse for *X. laevis* (Yanai et al., 2011). Each circle represents the transcriptome of a single embryo, where the color indicates the relative developmental stages of the samples. The connecting line is the developmental path calculated by BLIND. (B) Entropy of gene expression levels across developmental time in the *X. laevis* timecourse. The inset distributions show the gene expression levels at the indicated stages. The line is a linear fit. (C,D) Pairwise similarities between the *X. laevis* transcriptomes when samples are randomized (C) and BLIND ordered (D).

(see Materials and Methods for a full description of the algorithm). Fig. 1A shows the concordance between the BLIND path through the samples and their published developmental order. BLIND also recovers the correct order in the other published timecourses (Table 1). In summary, BLIND starts with unordered transcriptomes (Fig. 1C) and sorts them according to a genetic algorithm pathfinder on the principal components of the transcriptomes (Fig. 1D).

Transcriptomic entropy increases over developmental time

The path extracted by the traveling salesman algorithm does not indicate the direction of development. We found, however, that for developmental timecourses the direction of time can be deduced by computing for each transcriptome the entropy, which is a measure of the variability in total gene expression levels. As shown in Fig. 1B, entropy in expression levels increases with developmental time in the *Xenopus laevis* timecourse. The rise in transcriptomic entropy with developmental time was also observed in two of the four other previously published timecourses (supplementary material Fig. S2). In the remaining timecourses entropy was not dynamic, perhaps owing to the restricted span of the timecourse.

From a developmental perspective, the rise in entropy suggests that the maternal deposit of RNA present in the single-cell embryo is relatively ordered, whereas the increasingly complex embryo

Table 1. Performance of BLIND in previously published timecourses

Species	*N*	R^2	*P*-value
Xenopus laevis	14	1	–
Xenopus tropicalis	14	1	–
Drosophila melanogaster (female)	12	0.99	$<10^{-9}$
Drosophila melanogaster (male)	12	0.91	$<10^{-6}$
Aedes aegypti	24	0.98	$<10^{-19}$
Danio rerio	9	1	–

BLIND-ordered samples were compared with the published order by computing Pearson's correlation coefficient between the two ordered vectors.

Development (2014) doi:10.1242/dev.105288

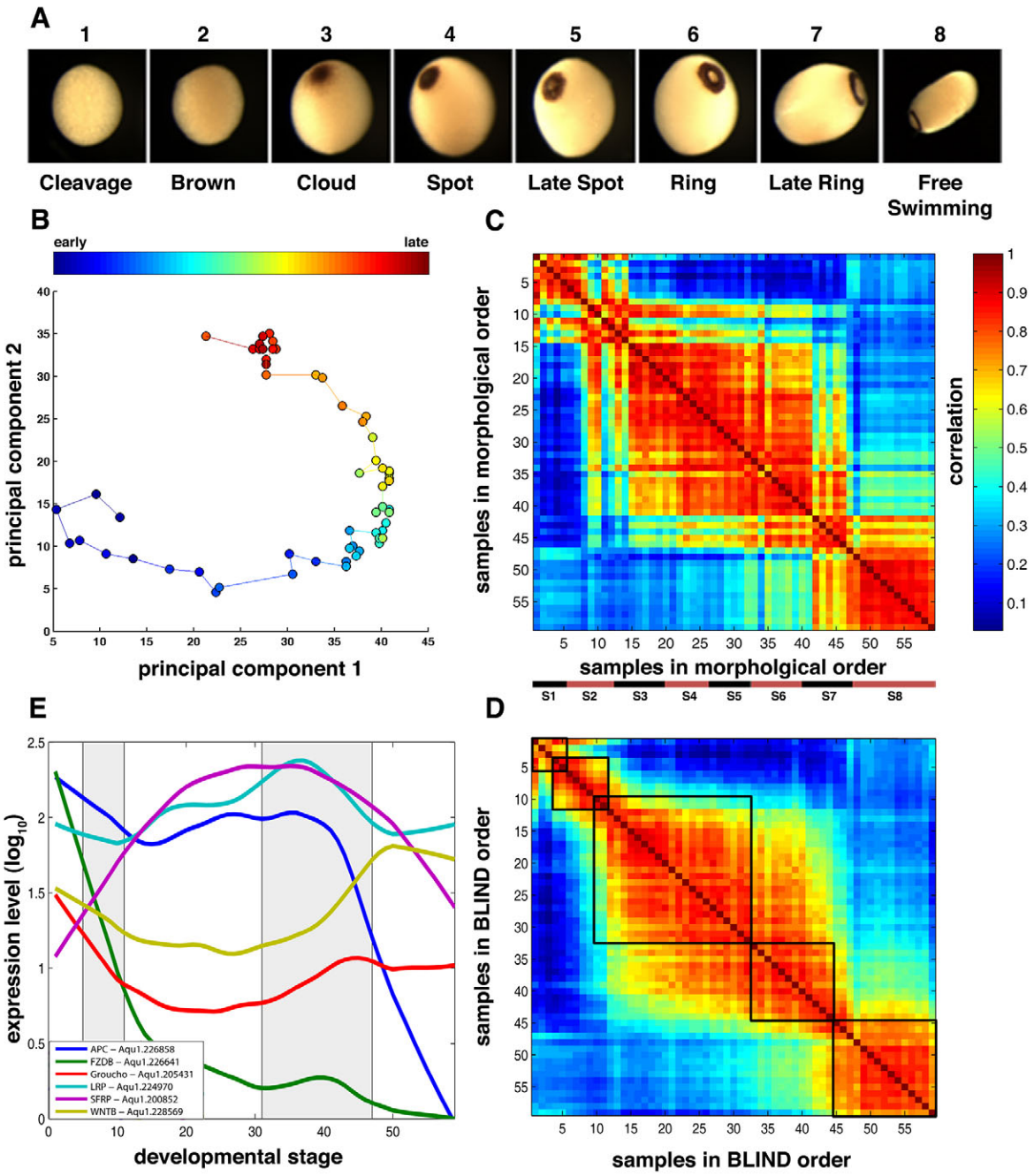

Fig. 2. BLIND-ordered samples in a single-embryo high-resolution developmental timecourse of the sponge *A. queenslandica*. (A) Micrographs of *A. queenslandica* embryos and larvae at the indicated developmental stages. (B) PCA on the transcriptomes of 59 samples from a developmental timecourse of *A. queenslandica*. Each circle is a sample colored according to the relative developmental stage as inferred upon collection. The samples are connected by lines representing the BLIND-deduced path. (C) Pairwise similarities between the transcriptomes comprising the *A. queenslandica* developmental timecourse. The embryos are ordered according to morphological staging. (D) Same as C, following the ordering of the samples by BLIND. Black boxes indicate observable transcriptome periods that are consistent with morphological transitions. (E) Gene expression profiles for the six indicated genes involved in Wnt signaling.

contains a more homogenous array of expression levels from many cells of many cell types. Indeed, the initial transcriptome is markedly different in distribution from the final transcriptome in the *Xenopus* timecourse according to its restriction of medium expression levels, thus leading to lower entropy (Fig. 1B, insets). This notion is also supported by the recent observation that individual cells have a bimodal distribution of expression levels, whereas for a complex collection of cells the expression levels are normally distributed because of the effect of averaging across many cells (Hebenstreit and Teichmann, 2011). The overall rise in expression entropy is exploited by BLIND to determine the polarity of the timecourse.

A high-resolution *Amphimedon queenslandica* timecourse

To demonstrate the capacity of BLIND to accurately infer developmental ordering we collected 59 single embryo and larvae samples of the sponge *Amphimedon queenslandica*. The brood chambers of *A. queenslandica* contain embryos at different

Development (2014) doi:10.1242/dev.105288

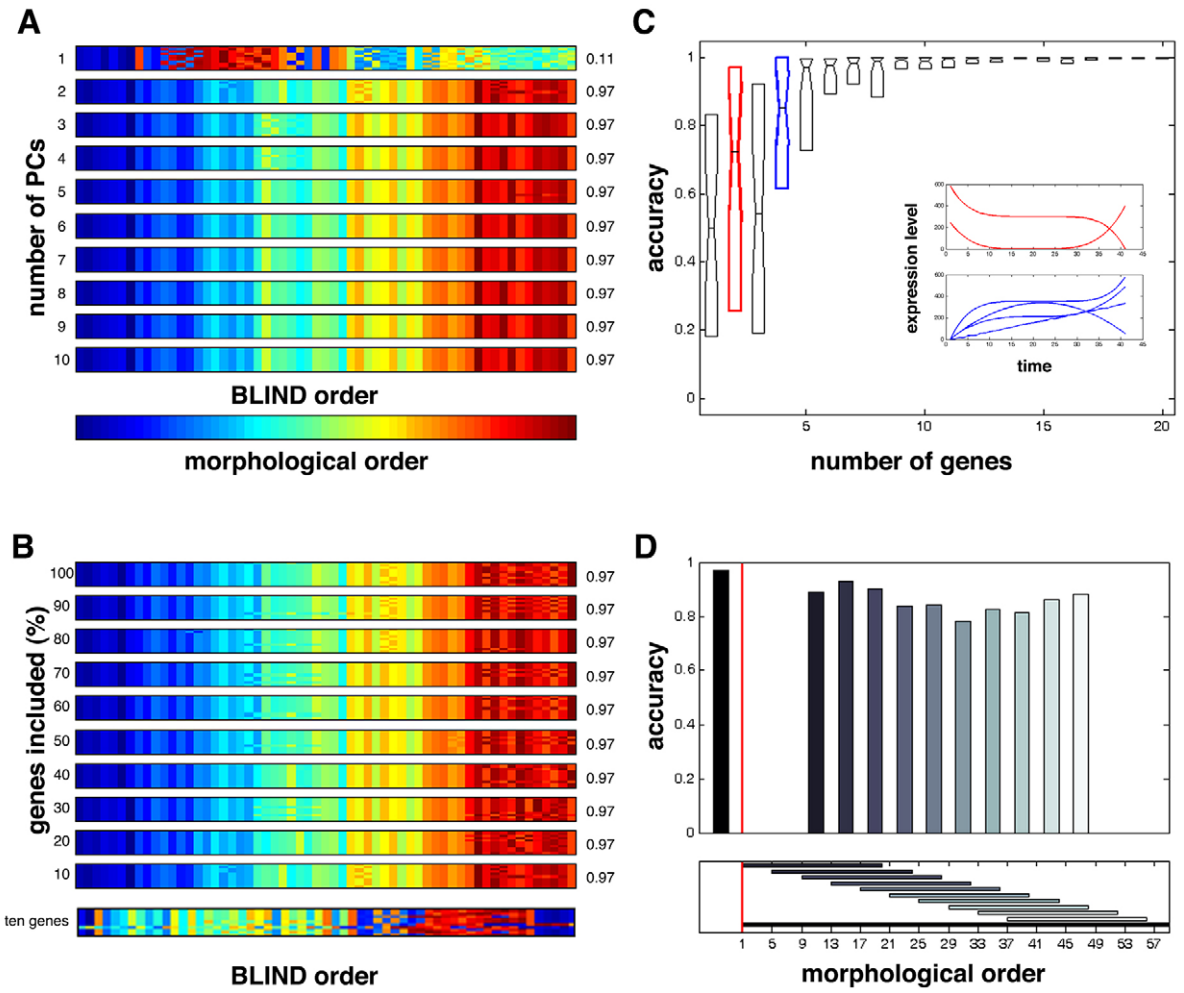

Fig. 3. Performance of the BLIND method. (A) Effect of the number of principal components used for BLIND ordering on performance. For each number of PCs, ten replicates (rows) are shown, where the annotated morphological order is indicated by color. BLIND accuracy, which is computed as the mean Pearson's correlation (n=59) between BLIND and morphological ordering, is indicated on the right. (B) Effect of the number of genes included on BLIND analysis performance (same format as A). (C) Performance of BLIND on simulated gene expression datasets. The inset shows two simulated datasets of two and four genes (see Materials and Methods). The boxplot shows the distribution of BLIND accuracies for independent simulations of a given number of genes. (D) Performance of BLIND on partial timecourses. The *Amphimedon* timecourse was examined using BLIND for the indicated overlapping partial windows. The accuracy of the BLIND ordering of the entire timecourse is shown on the left.

developmental stages (Fig. 2A) and morphologically staging them requires skilled assessment of individual embryos. Even then, at best, embryogenesis can be divided into a small number of broad stages. We thus asked whether BLIND could order the samples without such information. For each sample, the mRNA expression levels of all genes were measured using CEL-Seq (Hashimshony et al., 2012), which is an RNA-Seq method, resulting in a 29,883×59 gene expression matrix. Fig. 2B shows PCA of this dataset and the BLIND path among them. As the figure indicates, both the ordering and direction of the BLIND-ordered timecourse showed a strong correlation with those determined morphologically (Fig. 2D,E; supplementary material Figs S3, S4).

Examining the pairwise correlations among transcriptomes revealed five distinct transcriptomic periods that had a general agreement with morphological stages. The agreement decreases in the transition points between the stages (Fig. 2C,D), suggesting that subtle transcriptomic differences between samples might not be reflected at the morphological level. We also confirmed that the BLIND-ordered timecourse faithfully captured known gene expression programs. For example, Fig. 2E shows the gene expression profiles for six genes involved in the *wnt* pathway, consistent with their previously characterized developmental roles (Adamska et al., 2010).

Performance of the BLIND method

We next sought to test BLIND robustness to its two parameters: the number of principal components at the disposal of the traveling salesman algorithm and the number of dynamically expressed genes considered. Fig. 3A shows BLIND performance on the *Amphimedon* timecourse for different numbers of principal components, ranging from one to ten. For each number of principal components we invoked BLIND for ten replicates, each time recording its accuracy as the correlation between the BLIND order and the annotated morphological order. We found that running BLIND with a single principal component yields poor accuracy (R=0.11). However, for two principal components or more, the accuracy is $R \geq 0.97$, indicating that using at least two principal components is sufficient for robust performance. The coherence of the ten replicates in each set further reflects the reproducibility of BLIND despite its inherently heuristic nature.

Development (2014) doi:10.1242/dev.105288

We also found that BLIND is robust to the number of genes included in analysis. As demonstrated in Fig. 3B, examining the 10% most dynamic genes, or any higher fraction, produced good behavior. To gain insight into why few genes are apparently sufficient for BLIND performance, we tested BLIND on simulated temporal gene expression profiles (Fig. 3C, insets; see Materials and Methods). Invoking BLIND on these simulated datasets, we found that using even a few simulated profiles is sufficient to faithfully recover the ordering. The boxplots shown in Fig. 3C indicate that when using only a single simulated gene, BLIND generally gave poor results; however, with four genes it was already highly accurate (median R>0.85). From these simulations we conclude that the continuous nature of gene expression provides the crucial clue for the sorting of samples by BLIND. In contrast to the simulated profiles, which are perfectly continuous by their definition, invoking BLIND on only the ten most dynamic genes of the *Amphimedon* timecourse did not produce accurate results (Fig. 3B), indicating that the strength of the method is in its integration of information from many dynamically expressed genes, however noisy.

Finally, we inquired whether BLIND is expected to produce accurate results on partial timecourses. The maternal and zygotic transcriptomes of animals are dramatically distinct (Levin et al., 2012; Yanai et al., 2011) suggesting that BLIND performance might be dependent upon overlap with the transition between these two. We tested different regions of the timecourse using overlapping windows, each of only 20 samples. If BLIND is dependent on early development it would be expected to fail for subsets of the data that include only the later time points. By contrast, we found that BLIND performed with fairly uniform accuracy across all subregions of the timecourse (Fig. 3D), suggesting its general applicability to development, in datasets with at least ten samples (supplementary material Figs S5, S6).

BLIND has some important limitations that may serve as points for its future development. As samples are randomly collected, BLIND may be modified to combine embryos that are extremely similar in age and appear essentially as replicates along the PCA path. Samples that fall beyond the natural PCA path might correspond to anomalous embryonic developments, perhaps accounting for dead embryos, and on this basis can be excluded from analysis. Finally, the BLIND method can be used to identify gaps along the PCA path that might correspond to developmental stages.

BLIND-ordered sampling of large-scale experiments has several important applications. Most readily, the method enables analysis of randomly collected embryos whose relative developmental stages are unknown. This is perhaps most advantageous for asynchronous embryos such as *Nematostella* (Fritzenwanker et al., 2007) and, in addition, to embryos that lack observable morphological markers (e.g. opaque embryos) or have to be acquired by environmental sampling (e.g. plankton tows). Large-scale transcriptomic approaches will perhaps be most valuable when studying processes at the single-cell level (Shapiro et al., 2013), such as tumor populations and B-cell maturation. Such scaling up to studying $>10^3$ samples will enable the high-resolution view necessary for understanding the gene regulation of complex processes.

MATERIALS AND METHODS

The BLIND method

The method begins with a gene expression matrix in which the rows correspond to genes and the columns to samples. Normalized expression values are transformed to $\log_{10}$ scale and then filtered to contain only the X most dynamically expressed genes, where X is a parameter set by the user. X is set to 10% throughout the analyses shown here, unless noted otherwise. Expression dynamics is computed as the range of expression values for each gene. In order to avoid outlier effects the range is taken as the difference between the second lowest and the second highest values. PCA is computed on the filtered expression matrix using the Matlab library function *princomp*. The first Y principal components were used to represent the samples, where Y is set by the user (Y=2 by default). The Y principal components are normalized and scaled by percentage of explained variance. The order of samples in the normalized PCA plane is determined using an implementation of a genetic algorithm for the traveling salesman problem (Kirk, 2008) in which, given a list of cities and the distances between each city-city pair, the task is to determine the shortest possible route visiting each city exactly once. The specific parameters used for this are: XY, a matrix containing the normalized first Y principal components of the samples; DMAT, an Euclidean distance matrix of the samples; POPSIZE, 100; NUMITER: 10^4; SHOWPROG, 0; SHOWRESULT, 0.

Transcriptomic entropy

Shannon's entropy was computed for each sample as $\sum_{i=1}^{G} p(k_i)\log(p(k_i))$, where $p(k_i)$ indicates the expression level of gene i divided by the sum of the expression levels of all genes. The samples' entropy across the traveling salesman problem-ordered dataset was then fitted using linear regression to identify the trend. In the case of a negative trend, the order was flipped to arrive at the final BLIND sample ordering.

BLIND web server

The BLIND method is available online at blind.technion.ac.il. Users can upload gene expression matrices, compute BLIND using selected parameters, view the ordering process and download the BLIND-ordered dataset.

A. queenslandica transcriptomics

Embryos, larvae and post-larvae were collected individually at Heron Island, the Great Barrier Reef, Australia. These were staged by morphology and imaged before being stored in 20 µl RNA later (Invitrogen). RNA was isolated using TRIzol as previously described (Levin et al., 2012). 5 ng total RNA was used as input for the CEL-Seq protocol (Hashimshony et al., 2012) using the published *A. queenslandica* genome and gene models (Srivastava et al., 2010). As previously described in the CEL-Seq protocol (Hashimshony et al., 2012), the resulting read counts were normalized to transcripts per million (TPM). The complete dataset has been deposited in the Gene Expression Omnibus with accession code GSE54364.

Gene expression simulations

Gene expression profiles were simulated using polynomial functions with degrees randomly selected from the range of zero (constant expression) to 5. A set of coefficients was then randomly generated from the range −10 to 10. Noise was added to each time point from a normal distribution (mean=0; standard deviation=0.5). For a given dataset of size N, N gene profiles were generated independently.

Acknowledgements

We thank David Silver for initial help with this project; members of the I.Y. lab for suggestions; and the Technion Genome Center for technical assistance and sequencing.

Competing interests

The authors declare no competing financial interests.

Author contributions

I.Y. and L.A. conceived the method. L.A. led the development of the method. N.N. isolated the embryos. M.L. performed the CEL-Seq method. S.L.F.-V. contributed analysis tools. L.A. and S.K. analyzed the data. B.M.D. coordinated the experimental design and analysis of the *Amphimedon* timecourse. I.Y. and L.A. drafted the manuscript, which was edited by all authors.

Funding

The research leading to these results has received funding from the European Research Council (ERC) under the European Union Seventh Framework Programme [FP7/2012-2017]/ERC grant agreement no. 310927 – EVODEVOPATHS to I.Y.

Supplementary material

Supplementary material available online at
http://dev.biologists.org/lookup/suppl/doi:10.1242/dev.105288/-/DC1

References

Adamska, M., Larroux, C., Adamski, M., Green, K., Lovas, E., Koop, D., Richards, G. S., Zwafink, C. and Degnan, B. M. (2010). Structure and expression of conserved Wnt pathway components in the demosponge Amphimedon queenslandica. *Evol. Dev.* **12**, 494-518.

Akbari, O. S., Antoshechkin, I., Amrhein, H., Williams, B., Diloreto, R., Sandler, J. and Hay, B. A. (2013). The developmental transcriptome of the mosquito aedes aegypti, an invasive species and major arbovirus vector. *G3 (Bethesda)* g3.113.006742.

Fritzenwanker, J. H., Genikhovich, G., Kraus, Y. and Technau, U. (2007). Early development and axis specification in the sea anemone Nematostella vectensis. *Dev. Biol.* **310**, 264-279.

Hashimshony, T., Wagner, F., Sher, N. and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666-673.

Hebenstreit, D. and Teichmann, S. A. (2011). Analysis and simulation of gene expression profiles in pure and mixed cell populations. *Phys. Biol.* **8**, 035013.

Held, M. and Karp, R. M. (1970). The traveling-salesman problem and minimum spanning trees. *Oper. Res.* **18**, 1138-1162.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P. and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160-1167.

Kirk, J. (2008). Open traveling salesman problem – genetic algorithm. *MATLAB* 7.12 (R2011a).

Larrañaga, P., Kuijpers, C. M. H., Murga, R. H., Inza, I. and Dizdarevic, S. (1999). Genetic algorithms for the travelling salesman problem: a review of representations and operators. *Artificial Intelligence Review* **13**, 129-170.

Levin, M., Hashimshony, T., Wagner, F. and Yanai, I. (2012). Developmental milestones punctuate gene expression in the Caenorhabditis embryo. *Dev. Cell* **22**, 1101-1108.

Lott, S. E., Villalta, J. E., Schroth, G. P., Luo, S., Tonkin, L. A. and Eisen, M. B. (2011). Noncanonical compensation of zygotic X transcription in early Drosophila melanogaster development revealed through single-embryo RNA-seq. *PLoS Biol.* **9**, e1000590.

Papadimitriou, C. H. (1977). The Euclidean travelling salesman problem is NP-complete. *Theor. Comput. Sci.* **4**, 237-244.

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C. et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777-782.

Shapiro, E., Biezuner, T. and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618-630.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E. A., Mitros, T., Richards, G. S., Conaco, C., Dacre, M., Hellsten, U. et al. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* **466**, 720-726.

Yanai, I., Peshkin, L., Jorgensen, P. and Kirschner, M. W. (2011). Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility. *Dev. Cell* **20**, 483-496.

Yang, H., Zhou, Y., Gu, J., Xie, S., Xu, Y., Zhu, G., Wang, L., Huang, J., Ma, H. and Yao, J. (2013). Deep mRNA sequencing analysis to capture the transcriptome landscape of zebrafish embryos and larvae. *PLoS ONE* **8**, e64058.