Tomancak and Ohler *Genome Biology* 2010, **11**:115
http://genomebiology.com/2010/11/4/115

Genome **Biology**

RESEARCH HIGHLIGHT

Mapping the complexity of transcription control in higher eukaryotes

Pavel Tomancak[1] and Uwe Ohler[2,3,4]*

Abstract

Recent genomic analyses suggest the importance of combinatorial regulation by broadly expressed transcription factors rather than expression domains characterized by highly specific factors.

The complexity of life does not correlate with an increased size of the list of parts (the genes) from which organisms are built, but rather with an increased complexity in how these parts are regulated and combined into networks to specify the correct tissue-specific expression of genes. Analyses of yeast had shown a fairly simple hierarchical regulatory architecture, in which master regulators drive expression of many genes and any given gene is typically regulated by at most a handful of transcription factors (TFs) [1]. Some studies in animals, including studies of the early development of *Drosophila*, suggested a straightforward extension of the concept of a small number of highly specific TFs that define expression domains. Recent studies, including one by Adryan and Teichmann in this issue of *Genome Biology* [2], put the idea to the test by evaluating large genomic datasets, and their conclusions challenge this hypothesis.

Adryan and Teichmann's study is based on datasets obtained by two popular methods for analyzing gene expression [3,4]. Transcriptional profiling using microarrays requires substantial amounts of biological material and is thus typically used on intact multicellular specimens or cultured cell lines. RNA *in situ* hybridization is used to visualize spatial and temporal gene expression, but is limited for several reasons: some classes of eukaryotic genes, such as microRNAs, are difficult to study in this way; many tissues, such as brains, cannot be permeabilized enough to deliver the probe throughout the sample; temporal resolution is limited; and there is a lack of reliable quantification methods. Systematic RNA *in situ* surveys are therefore routinely combined with microarray analysis to counter the drawbacks of the two methods [4].

Drosophila embryonic development is particularly amenable to analysis by both *in situ* hybridization and microarray analysis. Large numbers of approximately staged embryos enable the isolation of sufficient amounts of RNA for microarray experiments or fixed specimens for *in situ* labeling. Several microarray time-courses profiling embryogenesis have been assembled so far, and these have been instrumental in understanding the major patterns of gene expression, defining gene batteries characteristic of maternal deposition, the maternal-to-zygotic transition, neurogenesis and organogenesis. Two major RNA *in situ* hybridization screens in embryos, focusing on tissue specificity of gene expression and RNA localization, documented expression patterns of about 60% of the genes in the genome with more than 100,000 images. Both surveys used controlled vocabulary annotations provided by experts to describe the patterns observed in the images. Using these annotations, similar patterns have been grouped by clustering approaches. Incorporating time-course microarray data into the clustering enabled the distinction to be made between broadly expressed genes and highly restricted tissue-specific expression [4]. Both studies were unbiased with respect to the types of genes analyzed and reported a spectacular diversity of gene expression regulation that defies easy attempts at classification.

Integrative analyses of genome-wide gene expression datasets

Adryan and Teichmann [2] have taken a fresh look at these available *Drosophila* datasets, focusing primarily on spatial patterns of gene expression, as summarized by controlled vocabulary annotations [4], and integrating them with recent microarray studies [3]. The study [2] concentrates on TFs, as they are arguably at the core of the gene regulatory networks governing embryonic development, and follows previous work by the authors [5] that defined a curated set of TFs in the *Drosophila* genome using protein sequence features (binding domains).

*Correspondence: uwe.ohler@duke.edu
[2]Institute for Genome Sciences & Policy, Duke University Medical Center, 101 Science Drive, Durham, NC 27708, USA
Full list of author information is available at the end of the article

The authors [2] made several noteworthy observations regarding TF activity on a genome-wide scale. Almost the entire complement of TFs is used during both embryogenesis and in adults, implying that the entire transcriptional regulatory machinery is used at multiple stages of the *Drosophila* life cycle. The authors [2] also see little relationship among the types of adult and embryonic tissues that a given TF is expressed in, which suggests that, on a genome-wide level, there is no support for the idea that TFs maintain their expression along developmental lineages. The embryo and adult fly are two largely distinct animals separated by an autonomous larval stage and transformed into one another during metamorphosis, and from this perspective, the findings [2] are sensible.

More surprising are patterns observed within embryogenesis, in which many TFs show tissue-specific gene expression during early stages (blastoderm stage and around gastrulation) and late stages (organogenesis) that do not follow developmental trajectories [2]. *Drosophila* embryologists might object that these patterns are not the rule and back up their argument with the examples of master regulators that specify and mark developmental lineages, such as Single minded, which specifies the midline cells of the nervous system. On the other hand, counter-examples are readily available, such as the extensively studied Hunchback TF, which has distinct and unrelated functions in early body-plan patterning (gap gene function) and nervous system development (sequential cell fate specification). The key to the argument is statistics; when looking at the class of TFs as a whole, there is no significant trend of respecting developmental lineages, and the examples that might be used to object to this model are important exceptions, but not the rule.

Following similar reasoning, the authors [2] examined how the expression patterns of TFs differ from those of the non-TF remainder of the genome. A relatively small proportion of maternally expressed genes are TFs, but because the mRNA for most genes is provided by the mother, there are still surprisingly many TFs among them, far exceeding the well known examples that kick-start body patterning, such as Bicoid and Caudal. Adryan and Teichmann [2] reveal the full scale of the maternal transcription factor expression: regardless of the particular dataset, about 60% of TFs are maternally deposited, meaning that the cytoplasm of the early embryo is flooded with sequence-specific DNA binding activity that is largely unaccounted for in models of embryonic gene expression. Relatively little is known about the expression of proteins from these maternal TF transcripts, but the study of polysome association has suggested that the majority of them are in fact translated. What the impact of this indiscriminate loading of pleiotropic regulatory proteins into the early embryo is, and how it relates to the pervasiveness of TF binding sites in the genome, remains an interesting yet unanswered question.

Overall, the proportion of expressed genes that encode TFs is the highest during the crunch time of body-plan layout, around gastrulation (stages 4 to 8 in *Drosophila*) [2]. Later on, the authors [2] detected an intriguing dichotomy among germ layer derivatives. The enrichment of TFs in mesoderm and endoderm drops, whereas it remains high in ectoderm primordia and gets further restricted to the nervous system, where most of the TF 'action' seems to reside in late stages of embryogenesis. It is as if the regulatory traffic gets redirected to the nervous system, which still undergoes significant patterning decisions after other tissues have been specified; this lends further support to the notion that the activity of nodes in regulatory networks is not restricted to specific lineages but is flexibly reused when and where cell fate decisions are needed.

More specific analyses [2] address how broad TF subclasses defined by a common DNA binding domain are used in development. The authors [2] detect a trend for the largest domain families; members of the zinc-finger family tend to be expressed early in development, whereas basic helix-loop-helix (bHLH) and homeodomain TFs typically appear late. Why would that be the case? TFs from the same family derived from a common ancestor domain in the evolutionary past. The homeodomain-based regulatory system that patterns the anterior-posterior axis is ancient, as it is shared by all existing animal phyla. Could it be that expression constraints were carried over through countless duplication and diversification events and are still present? It would be interesting to see whether zinc-finger TFs, which are expressed predominantly early (because their mRNA is maternally contributed), show a similar bias to early expression in other animal phyla. Alternatively, the specific layout of gene regulatory networks early and late in development may require different classes of DNA binding trans-activators with different binding properties. The observation [2] that many of the early TFs are reused later argues against this interpretation. Once again, the observations reveal statistically significant genomic trends, and many exceptions to these broad rules can be found (for instance, some bHLH TFs are in fact maternally deposited).

Finally, Adryan and Teichmann [2] tackle the complex issue of combinatorial gene expression control. With the naïve hypothesis 'one tissue - one master regulator TF' rejected, they attempt to identify combinations of two or three TFs that would define developmental domains. Indeed, almost all possible combinations of TFs for which expression data are available from both sources (69,500 = $373^2/2$) are co-expressed in at least one tissue during development. Although these associations are highly dynamic, a significant fraction persists through time and through developmental intermediates, particularly during organogenesis. There is no evidence yet that these potential modules indeed interact at the same genomic regulatory target region, and the authors [2] note that the same level of association exists for non-TFs, but this may point to target genes of the combinatorial TF partners.

Broader implications for *cis*-regulatory regions

A new study from the FANTOM consortium [6] recently reported on combinatorial transcription regulation in mammals, integrating expression with protein-protein interaction (PPI) data. Again, individual TFs were found to be widely expressed, and the specification of tissue type relied on combinatorial control involving TFs. Therefore, two independent reports in different systems [2,6] arrive at the same conclusion that most TFs do not, by themselves, specify tissue restricted expression.

Sets of TFs could potentially co-regulate targets by exerting their influence on a common genomic regulatory region. The work of Ravasi *et al.* [6] implies a stricter model of combinatorial control, by including PPIs between TFs in addition to co-expression. PPIs can additionally 'disambiguate' between proteins with similar or identical binding sites, and this ability may be strictly necessary, given that TFs from the same family share sequence binding preferences [7,8] and that most TFs in flies belong to just a few classes that also happen to be co-expressed. It might therefore only be possible to identify functional targets in a specific manner by evaluating the binding of sets of interacting TFs. A known example of this is the mammalian E2F family, whose members can be activators or repressors despite the same binding preferences, which is achieved, at least partially, through specific interactions with other TFs.

Assuming that these general observations hold after further investigation, they have implications for the definition and identification of *cis*-regulatory modules. Early on, researchers in regulatory genomics have proposed the concept of *cis*-regulatory grammars: specific rules or constraints in terms of order, orientation, number and/or spacing between binding sites. Whether such grammars really exist has been under much debate; for instance, evolutionary patterns can wrongly suggest constraints when there are none [9]. If specific PPIs between TFs are necessary to define targets and specificity, these interactions will constrain the relative orientation of TFs and thus be reflected at the level of *cis*-regulatory organization. Although such rules may easily be lost in the noise of the vast landscape of a single regulatory genome, experimental profiling under more specific conditions, as well as conservation, will help us to narrow this down.

From high throughput to high resolution

New transcriptional profiling data are coming online daily, thanks to systematic efforts such as modENCODE [10], which aims to annotate all functional elements in model organism genomes such as *Drosophila melanogaster* and *Caenorhabditis elegans*. Quantitative expression measurements derived from complete samples could potentially be much better used if the spatial extent of expression is estimated from microscopy data. For such analyses, it is necessary to step back from the annotations and work directly with the primary image data. Fortunately, image analysis for spatial expression data has recently become a blooming research field of its own, and state-of-the-art computer vision techniques are now being used to classify and analyze patterns of gene expression automatically. Such approaches are unbiased and can lead to the definition of new expression domains, particularly when looking at combinations of patterns, and scale better to larger datasets for which tedious manual annotation efforts may simply prove infeasible. Several new projects using high-resolution microscopy techniques are under way to describe expression patterns at unprecedented cellular precision, but they have not yet reached the coverage required for making genome-wide statistical inferences. As the coverage increases in the near future, the global integrative analysis of such datasets will be possible. The work of Adryan and Teichmann [2] demonstrates the promise of the integration of quantitative measurements with spatial expression data and shows that this approach will be crucial to untangle the gene regulatory networks in development.

Acknowledgements
Together with Casey Bergman at the University of Manchester, the authors are currently recipients of a Human Frontier Science Program young investigator award.

Author details
[1]Max Planck Institute for Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany. [2]Institute for Genome Sciences & Policy, Duke University Medical Center, 101 Science Drive, Durham, NC 27708, USA. [3]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, 2301 Erwin Road, Durham, NC 27710, USA. [4]Department of Computer Science, Duke University, LSRC Building D101, 450 Research Drive, Durham, NC 27708, USA.

Published: 30 April 2010

References

1. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431**:99-104.
2. Adryan B, Teichmann SA: **The developmental expression dynamics of *Drosophila melanogaster* transcription factors.** *Genome Biol* 2010, **11**:40.
3. Hooper SD, Boue S, Krause R, Jensen LJ, Mason CE, Ghanim M, White KP, Furlong EE, Bork P: **Identification of tightly regulated groups of genes during *Drosophila melanogaster* embryogenesis.** *Mol Syst Biol* 2007, **3**:72.
4. Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM: **Global analysis of patterns of gene expression during *Drosophila* embryogenesis.** *Genome Biol* 2007, **8**:R145.

5. Pfreundt U, James DP, Tweedie S, Wilson D, Teichmann SA, Adryan B: **FlyTF: improved annotation and enhanced functionality of the *Drosophila* transcription factor database.** *Nucleic Acids Res* 2010, **38:**D443-D447.
6. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N, Carninci P, Daub CO, Forrest AR, Gough J, Grimmond S, Han JH, Hashimoto T, Hide W, Hofmann O, Kawaji H, Kubosaki A, Lassmann T, van Nimwegen E, Ogawa C, Teasdale RD, Tegnér J, Lenhard B, Teichmann SA, Arakawa T, Ninomiya N, *et al.*: **An atlas of combinatorial transcriptional regulation in mouse and man.** *Cell* 2010, **140:**744-752.
7. Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA: **Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites.** *Cell* 2008, **133:**1277-1289.
8. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X, Kuznetsov H, Wang CF, Coburn D, Newburger DE, Morris Q, Hughes TR, Bulyk ML: **Diversity and complexity in DNA recognition by transcription factors.** *Science* 2009, **324:**1720-1723.
9. Lusk RW, Eisen MB: **Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers.** *PLoS Genet* 2010, **6:**e1000829.
10. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH; modENCODE Consortium: **Unlocking the secrets of the genome.** *Nature* 2009, **459:**927-930.

doi:10.1186/gb-2010-11-4-115
Cite this article as: Tomancak P, Ohler U: **Mapping the complexity of transcription control in higher eukaryotes.** *Genome Biology* 2010, **11:**115.