molecular
systems
biology

Accurate measurements of dynamics and reproducibility in small genetic networks

Julien Dubuis, Reba Samanta, Thomas Gregor

Corresponding author: Thomas Gregor, Princeton University

Review timeline:		
	Submission date:	07 September 2012
	Editorial Decision:	22 October 2012
	Revision received:	29 November 2012
	Acceptance letter:	10 December 2012

Editor: Thomas Lemberger

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision   22 October 2012

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the three referees who agreed to evaluate your manuscript. As you will see from the reports below, the referees find the topic of your study of potential interest. They raise, however, several concerns on your work, which should be carefully addressed in a revision of the present work.

Overall, the reviewers acknowledge the technical quality of the work. They feel however that considerably more efforts should be made to explain the implication of this work for the understanding of developmental patterning and early morphogenesis. In addition, important clarifications are requested with regard to the methodology.

Prior to our decision, we have circulated the reports among the three referees ('pre-decision cross-commenting'). As feedback to the additional experiment suggested by Reviewer #3, ie re-quantifying the gap genes+bicoid, reviewer #2 indicated that these experiments would be useful but beyond the scope of this paper, in particular if they involve staining 5 genes on the entire set. We agree with reviewer #2 that this should not represent a pre-condition for publication.

In view of the quantitative nature of your work, it will be important to make the full dataset available to the community. This would include the stainings of the 163 embryos reported here, the extracted raw expression profiles and the 'detrended' normalized profiles.

In addition, important quantitative data displayed in the figures (eg the time-dependent furrow depth in Figure 2B) should also be supplied as 'source data files'. These files are distinct from normal supplementary information, are directly to a specific figure panel and can be downloaded directly from the figure (example: <http://tinyurl.com/365zpej>). See more information: <http://www.nature.com/msb/authors/index.html#a3.4.3>

If you feel you can satisfactorily deal with these points and those listed by the referees, you may wish to submit a revised version of your manuscript. Please attach a covering letter giving details of the way in which you have handled each of the points raised by the referees. A revised manuscript will be once again subject to review and you probably understand that we can give you no guarantee at this stage that the eventual outcome will be favorable.

Referee reports:

Reviewer #1 (Remarks to the Author):

This paper by Dubuis et al describes new tools and methods to quantify levels and variance of gene expression patterns in the early embryo of Drosophila melanogaster. The authors then use these newly developed methods to characterize the temporal dynamics and reproducibility of the gap gene system. Despite its focus on methodology, the paper is ideally suited for publication in Molecular Systems Biology for the following two main reasons: (1) The rigorous, thorough, and precise characterization of experimental versus biological contributions to pattern variability constitutes a landmark for the quantitative analysis of spatio-temporal gene expression. It is extremely relevant beyond Drosophila as a model system. (2) The measurements presented allow a highly interesting and relevant inference concerning the underlying gene regulatory mechanism, namely, the fact that cross-regulation among gap genes indeed increase the reproduciblity of pattern formation. This establishes, rigorously, that morphogen gradients are not sufficient to ensure precise patterning in this system, which settles a controversy in the field which has been going on for more than ten years. In summary, this is a landmark paper. I cannot support publication in MSB strongly enough!

Adressing a few minor issues would further improve the quality of the paper:

The authors mention on p7 that the Hb/Kr, Kr/Kni, Kni/Gt, and Gt/Hb borders should be controlled by mutual repression. This is a valid interpretation of their results, which seemingly contradicts the mechanism of gap domain shifts postulated by Jaeger et al. 2004, which requires asymmetric repression between these genes. However, Crombach et al. 2012 have slightly revised the nature of this mechanism, showing that only the net overall effect of repression needs to be asymmetric (one interaction stronger than the other one) for this mechanism to work. It would be interesting to discuss this point in a sentence or two.

Two papers by Manu et al. (2009; PLoS Biol and PLoS Comp Biol) and a paper by Gursky et al. (2011; BMC Syst Biol) propose explicit regulatory mechanisms by which reproducibility of gap gene expression could be increased over time. It would be appropriate to mention this and to to cite these papers in the Discussion.

One methodological aspect that was not clear to me is how the authors established the orientation of their embryos. This is not explained in the current version of the manuscript. How is it possible to rigorously estimate azimuth angles, or to align an embryo for imaging of its dorsal surface, if no markers for DV expression (eg twist) were used? What does 'eye selection' (eg caption, suppfig 4) mean? Please clarify.

I have two questions about Fig 1 and the issue of cross-talk: (1) Why is there so much cross-talk in the 633 channel? The apsorption spectra in Fig 1A would not lead me to expect that. (2) It is generally assumed that anti-rat and -mouse secondaries show high levels of cross-binding to their respective primaries. I do not see that confirmed in the results. It would be worth commenting on this.

At times, the manuscript is somewhat difficult to read. There are many very long sentences, and some of the figures are overly complicated. I list a few suggestions for clarification/simplifiation below.

Finally, figure panels (eg Fig 2) could be made larger. They are sometimes hard to read as they are.

Some more detailed comments:

p3 and thereafter: positions are indicated in % EL with 0% at the anterior pole; convention usually has 0% at the posterior pole; this should be clarified when first using this measure.

p3, citations after 'to a complex network' are rather haphazard (some primary literature and a review); a more balanced selection of references should be provided; similarly, I am not sure whether Spirov03 and Manu09 are appropriate references where they are mentioned just a few sentences further.

p4, typo: 'an methods' should be 'and Methods'.

p4 (and elsewhere): please clarify if the cross-talk estimate is used to correct the measured expression patterns; I guess so but it is not clear from the current wording.

p4, the statement 'ending with gastrulation' is confusing since gap genes are expressed a little bit beyond gastrulation; I guess the authors take gastrulation as the end point of their analysis? Please clarify.

p6: I am not sure I understand why 163 embryos were stained in total, but only 80 use in the analysis; please clarify.

p12, the sentence citing a 'Fig. 1C' must be a left-over from an earlier version of the manuscript; there seems to be no such figure in the present version. Please correct.

p12 and fig 1: please clarify which channel corresponds to which row in the matrix. This is a bit confusing.

caption, fig2: 'profile in A shown in blue': I don't know what this refers to.

Panels 3b and suppfig 3b are the same. Is it required to reproduce them twice?

fig6: please provide a key to lines in panels D and E; I have no clue what the dashed and dotted lines in E are.

fig7 is completely overloaded; please declutter, eg by leaving out or dimming the expression patterns; currently they hide the actual results shown in this figure and make it very hard to read.

caption, suppfig1: descriptions of panels A and B are mixed up.

caption, suppfig4: typo, 'allows allows'.

Reviewer #2 (Remarks to the Author):

"Accurate measurements of dynamics and reproducibility in small genetic networks"
Submitted to MSB
Dubuis, Samanta, and Gregor

SUMMARY
The overall goal of this paper is to understand the variance in the gap gene network that patterns the Drosophila blastoderm embryo. This is part of a general line of research that addresses precision in developmental processes, attempting to understand its origins and its relationship to environmental and genetic variation. Variance and co-variance can also be used to infer underlying network architecture. Because the observed variance is a combination of biological fluctuations and experimental error, the authors aim to precisely quantify the degree of experimental error in their measurements. To do so, they make three technical contributions to the field: 1) increased precision in staging of embryos (from ~10 min to 1-2 minutes) 2) quantification of 8 sources of experimental

measurement errors. 3) definition of a position-dependent variance metric that they use to analyze the precision of the network over time. They conclude experimental errors sum to less than 20% of total variance for any gene, allowing them to analyze the remaining 80% of measured variance to understand the precision of the gap gene network. By analyzing the co-variance among gap gene expression profiles, they demonstrate that positional errors decrease two fold over the course of cellularization (from one internuclear distance to 0.5), and that they reach a minimum 15 minutes prior to gastrulation, synchronously across all genes (except the Hb anterior domain). These results demonstrate the utility of quantitative measurements, and provide some biological insights.

GENERAL COMMENTS
Technically, this work is very strong. The methods are rigorous, and its refreshing to see quantitative measurements taken so seriously. However, the biological insights are less novel. They build on concepts already solidified in the field (such as the canalization of the network and cross-regulation in the gap gene network). As written, the paper is focused on the methods, which is fine given that's where it makes its most substantial contribution. But the writing is quite technical and could benefit from more explanation of the high-level logic of their approach. By providing more context in the introduction, and by explaining their logic throughout the results, I think the authors could improve the accessibility of their manuscript a great deal. The paper would then not only be technically impressive, but it could also make the case more broadly for how careful quantitative measurements in wild-type embryos can yield biological insights.

MAJOR POINTS
Here I focus on places that could use additional conceptual or technical description.

"The final macroscopic outcome of developmental processes in multicellular organisms results in structures that are remarkably similar between individuals of a given species." (p. 3) This sentence would benefit from some references and further description. What are examples of this type of morphological precision? Which features are particularly notable for the Drosophila embryo?

This sentence in the intro presents a hypothesis about morphological precision as proven: "This similarity has its origins in the reproducible spatial patterns of morphogen concentrations in the early embryo." (p. 3) But my understanding is that this is an active area of research, and is in fact the central question of the paper. Can reproducibility/precision in morphology be traced to precision in the underlying molecular patterning mechanisms? The alternative is that patterning is itself messy, but cleaned up later. This may of course vary across different patterning mechanisms in different systems. Examples of active models beyond Drosophila would be useful to mention here. My understanding of the overall justification of this line of research is to ask whether the precision observed in the patterning system is sufficient to account for the precision observed in morphology. Or alternatively, focusing on just the quantitative features of the network itself, to understand which features are variable, how this relates to network architecture, and how these features respond to environmental and genetic variation. This paper is then situated as an attempt to precisely measure variation itself, which is technically challenging. I suggest that this is described more explicitly in the introduction.

"Conceptually, we are seeking..." (p. 3) This sentence lists many network properties that could be measured, but does not state how the measurements will be useful in the long-term. Clarifying this would help the reader grasp the logic of this type of study. For understanding this question of precision in patterning, what are the most relevant measurements? Should we be most interested in invariant features or those that vary, or both? This study is focused on one genotype, but could extending these types of studies to additional genotypes be fruitful? Focusing the readers' attention on a specific example/scenario might be helpful.

Some discussion of the value of studying inherent fluctuations in the system in the introduction would be useful for the more novice reader. Some of this is presented in the opening paragraph in the discussion, but moving it to the introduction would give the reader more context. This would also improve the justification of their careful measurements.

I find it strange that after all of their quantification of error, they choose to work with the uncorrected total variance, rather than the corrected variance. Granted, after the effort of determining the experimental error is less than 20%, one can be more confident in the

measurements. However, this choice should be further justified either in the results or in the discussion. This is even more confusing since they proceed to bin their measurements into 8 equally populated bins, which makes their measurements comparable to other previous studies (such as Reinitz/FlyEx and the BDTNP).

Minor points
Here I focus on wording that could be clarified. There also are numerous grammatical errors throughout, but these should be addressed after revision since the text may change a great deal.

p. 3 "the level of reproducibility that is actually relevant to biological processes" is an ambiguous phrase. Does it mean measuring reproducibility at the appropriate scale (i.e. molecules, cells, tissues, etc.) or does it mean the level of variation that is tolerated by the system while still maintaining function?

p. 3 "inherent experimental error due to variable conditions" isn't very precise. What types of experimental errors? My understanding is that the major issue is with these techniques is averaging many stains together - therefore potentially convoluting the variation in the quantitative relationship between genes. Please clarify the logic here and how staining for multiple proteins in the same embryo circumvents these limitations.

p.3 The authors present the use of fixed tissue as a limitation of previous methods, but employ a fixed tissue approach themselves. This is confusing, especially since they are employing a similar method for obtaining dynamic information from fixed tissue (extent of invaginating cell membranes), but at a higher level of resolution. (Previous studies used something like 10 min bins, while this study uses 1-2 min bins). This should be stated more clearly.

p.3 "In particular, what is the final reproducibility at gastrulation and how does it compare to the reproducibility of the Bcd gradient?" The first part of this sentence is ambiguous - reproducibility of what feature of gastrulation?

p.5 "After profile alignment and correction for tissue shrinkage due to the fixation process ..." The logic of this procedure is unclear from the main text. It appears they will first relate furrow depth to absolute time using live bright-field imaging. They can then align these time traces to determine the error in assigning the appropriate time from a single time-point. To extend this procedure to fixed embryos they need to normalize to their live measurements, since fixation is known to cause embryos to shrink. Even in the Methods, this normalization calculation isn't particularly clear. The AP and DV dimensions of the embryo shrink by ~5%, but how this is used to normalize the live furrow depth isn't immediately obvious. I suggest explaining the logic of their procedure more completely in the main text, and elaborating on the normalization calculation in the Methods.

p. 5 The paragraph beginning "The precision of 1-2 min with which..." would be easier to understand if the conclusion was presented as the topic sentence. Something like "Previous quantitative studies of expression in this system have used roughly 10 minutes time intervals. By staging embryos in 1-2 min intervals, we determined that 10-20% of the variance in these 10 minute time intervals is due to underlying expression dynamics."

p.5 "Hence, an overwhelming majority of the raw profile-to-profile variability seen in Figure 2C can be attributed the dynamics of the expression levels." This conclusion is obvious from previous studies as well. The more relevant result from this particular experiment is the ability to decompose the variance, as stated in the next sentence.

p.5 "These errors induce extra variance in the profile levels across a population of embryos, masking the actual biological fluctuations resulting from the natural embryo-to-embryo variance." This sentence could be reworded to improve clarity. The observed variance is a combination of 1) the natural variation in the system and 2) experimental error. They are interested in quantifying 1, and therefore must determine 2 to subtract it from the observed variance.

p.6 "To understand how the average levels of gene expression change with time, we monitor the immunofluorescence intensity of the boundary inflection points as well as the absolute value of the slope at these points as a function of the embryo age". What do these measurements and their timed

coordination tell you about the underlying network? Seems to me that some context about the known cross-repressive interactions between the gap genes is needed here, to help readers understand that the dynamics are likely due to interactions between the gaps as their expression patterns mature. The conclusion that their interactions (as tracked by the boundary position and the sharpening of the boundaries) is coordinated, might be expected from the known network topology, but the precision of the coordination is remarkable. And what do they mean by "special properties" at the end of this paragraph?

p.7 "In particular the constancy of sigma x/L across the Hb/Kr, Kr/Kni, Kni/Gt and Gt/Hb borders hints at a mutual regulation of these genes, rather than a uni-directional repression, in which case we would see a smaller variance in the expression profile that is regulated." This is a nice conclusion, again supported by previous work on the network topology (see for example work from Reinitz and Papetsenko). This should be cited and discussed here.

METHODS
Measurements of the invagination depth of the membrane furrow channels
p.9 "The depth of (i)fc of the FC was monitored as a function of absolute time until gastrulation as shown on Figure 2A." This sentence doesn't indicate how this measurement was actually made. Were embryos manually annotated, or was automated image processing used? In either case, what were the defining features used to delineate the boundaries?

FIGURES
General comments
The parameter names are often used as axis labels. To make the manuscript more accessible, they may also wish to use a descriptive text label. For example in Figure 2D, the x-axis could also be labeled "Furrow channel depth".

The figure legends also have descriptive titles. It might be easier for readers to grasp the purpose of the figure if the titles presented conclusions instead. For example, Figure 2: Precise temporal staging allows quantitation of time dependent variation in gene expression at every position.

Figure 2:
A: Inset is difficult to see - red/orange label is obscured. FC should be defined in the legend for readers that go through figures independently of the text. Mitosis 13 is used to label the figure, but the legend uses n.c. 14. Best to be consistent with this nomenclature.
B: Individual traces and points are extremely hard to see. Granted, the point is that they all overlay very precisely, but it may be worth increasing the size of this figure to make the features named in the legend easier to identify.
C: The paper discusses intervals in terms of time rather than FC depth (which is what is used in the legend for this panel). I think this should be made consistent. I vote for time since it is more obviously linked to the technical advance in this paper. Also, there are multiple dotted lines in the figure so referring to "the grey dotted line" should be changed to "the vertical grey dotted line". Vertical axis should be "Gt Intensity".
D: I presume this is also a trace for Gt at a single position, but this is not made clear in the legend.

Figure 3:
A: I find this schematic very hard to read. The yellow is nearly invisible, and there are a number of different molecular species depicted. The diagram is more complicated than the concept, which is to compare the signal from 2 different Hb antibodies detected with 2 different secondary antibodies in the same sample. I suggest showing a cartoon of a single Hb molecule, bound to two different primary and secondary antibodies, with the excitation wavelengths labeled (since those are what's used in panel B).
B: The axis labels could be more intuitive. GP-anti Hb / 546 - anti GP and Rat-anti Hb / 488 anti-Rat for example.

Figure 4:
D: The grey dotted line is difficult to track. Is it necessary for it to be dashed?

Figure 5:
The gap genes are not labeled, either in the figure or in the legend. Looks like kni, Kr, Gt, Hb, left to

right.

A: To contrast with the colored traces, I suggest that the dark blue trace be made black. It will make the description of lighter/darker colors in the legend easier to decipher. It would also allow them to use a light blue/darker blue pair to replace the yellow/green pair, where the yellow line is nearly impossible to see. I also find it confusing to have both the mean expression level and the variances plotted together this way. If the goal is to support their claim that "For any of the gap genes, it represents on average less than 20% of the gene expression variance measured across embryos", it might be fine to eliminate the total gene expression profile. If there is a further point to be made, which is that the degree of variance correlates with features of the expression pattern, then that should be stated explicitly.

Figure 6:

I find all of the symbols here to be overwhelming. The point is actually graphically pretty simple-the position of the borders changes over time (e.g. from low on the y-axis to high on the y-axis.) I think this would be easier to see with single line traces, or even dots (if there's concern about overlap, these could be open or transparent circles). The triangles are unnecessary, and the dashed lines are confusing. Its also worth noting that not all boundaries change position; this isn't mentioned in the text.

Figure 7:

I find the triangles distracting on this plot, and I don't think they are necessary. The overall point is simply to grasp that the black line has high values in some areas early on but progressively gets smaller.

PRE-DECISION CROSS-COMMENTING -- ADDITIONAL REMARKS

1. More context about the biological question being addressed, both in terms of framing the line of inquiry and citing relevant literature.

Each reviewer has numerous suggestions about how to do this. It will likely be impossible to incorporate them all while retaining a coherent narrative. I do not require every suggestion of mine to be incorporated to support publication, but I do think that the paper needs substantial revision to place the work in proper context. In particular, I would like to see a discussion of how this work furthers the hypothesis that network architecture improves patterning precision. As Reviewer 3 correctly points out, this paper is consistent with this hypothesis, but does not prove it. I'm not sure that this could be proven without similar characterization of alternative network topologies, which is clearly beyond the scope of this paper. But I think it is appropriate to more thoroughly discuss the contributions and limitations of this paper in this context.

2. Some methodological clarifications including
a. how they define the orientation of the embryos
b. the precise method used to stage the embryos (manual annotation of images or image processing)
c. use of the uncorrected variance in their network analysis
d. the calculation used to normalize time between fixed and live embryos
e. justification of time binning
f. justification of using a subset of the total data

Reviewer 3 also calls for an additional experiment - including Bcd stains in the current dataset. To incorporate this into their current analytical framework, the authors would have to find parameters and reagents to allow simultaneous imaging of 5 genes (not just the 4 presented here). Moreover, they would then have to reacquire all images using the 5-stain protocol. While I agree that it would be useful to have Bcd data included, I also believe that this is beyond the scope of this paper.

Reviewer #3 (Remarks to the Author):

The authors perform quantitative studies of gap gene expression in early Drosophila embryos. They focus on four gap genes in their current work, hb, Kr, gt and kni, through the use of both newly-generated and existing antibodies against the products of these genes. The authors provide careful and systematic evaluations of the experimental errors in quantifying the immunostaining data in whole mount embryos. They provide evidence suggesting that immunostaining is a viable approach

for quantitatively measuring gene expression profiles in embryos (an important idea that is not entirely new). They also provide data demonstrating that the reproducibility of the gap gene expression profiles increases two-fold during the course of nuclear cycle 14 (a notion that has also been available in the literature).

This is a carefully-designed and well-executed technical study. The systematic evaluations of the experimental errors should be useful to the field. Unfortunately, the current study remains at a technical level without significantly advancing our knowledge of the gene regulatory network that instructs embryonic patterning. In particular, a lack of quantitative data for the morphogen gradient of Bicoid under their current (and likely improved) technical framework makes it difficult to assess how the developmental system operates mechanistically. The current work conspicuously lacks a specific hypothesis presented and tested. The conclusion presented in the Discussion "Reproducibility inherent in the Bcd gradient is transferred to the gap genes early in n.c. 14, reduced two-fold within the gap gene network..." appears to imply, but does not clearly state, that the network itself is responsible for this two-fold reduction, a notion that has been championed by Reinitz and his associates, among others. This conclusion also appears to ignore completely the potential effects of time- or space-averaging, a hypothesis that has been available in the field (including Dr. Gregor's own previous studies). One of the claims made by the authors "We use this analysis to extract gap gene profile dynamics with ~1 min accuracy..." (in Abstract) is misleading because, although they have done calibrations for embryo time, their analysis was based on data from embryos that are grouped into 8 time classes (each lasting much more than 1 min). After reading through this nicely-executed technical work, one is left with an impression that after all we really have not gained much new mechanistic insights.

The existing quantitative data (in addition to the Bcd data that should be generated under their current framework) could potentially yield some new mechanistic insights. For example, the authors have the co-staining data for all the four gap gene products in individual embryos, an advantage highlighted repeatedly by the authors but not fully utilized by them. Is it feasible to extract information that can add to our understanding of how these gap gene products co-evolve at the level of individual embryos? Would an analysis that is truly at a ~1 min temporal accuracy allow the extraction of new information about how the system operates? How does embryo length variation impact the reproducibility of the patterning outcome?

A technical comment: the authors tend to make statements throughout the manuscript as if they are not aware of the existing knowledge in the literature. The discussion about the two-fold increase in the reproducibility of gene expression profiles (above) is one example. As another example, in the first paragraph in Discussion, the authors state that "Data of the latter approach has been missing thus far..." when referring to quantitative relationships in wt embryos. Do they really mean that? Accurately relating their experiments and findings to the literature and the existing knowledge might improve the manuscript (What is the current state-of-the-art in the field? What specific aspects has the current work advanced and how?).

1st Revision - authors' response   29 November 2012

Reply: We thank the reviewer for his enthusiasm for our approach and for his helpful comments that will hopefully strengthen our manuscript and make it more accessible to a broader audience. We have modified the introduction, added numerous references and clarified the raised technical issues. Please find our specific to each point answers below.

Reply: We changed the relevant sentence and cited the suggested references, as well as Papatsenko 2011, as recommended by Referee #2. We also added a sentence at the end of this paragraph highlighting the point of extracting regulatory mechanisms from profile fluctuations.

"*In particular, the constancy of σx/L across the Hb/Kr, Kr/Kni, Kni/Gt, and Gt/Hb borders hints at a mutual regulation of these genes, as hypothesized before (Kraut 1991, Papatsenko 2011, Crombach 2012), rather than a uni-directional repression (Jaeger 2004a,b), in which case we would see a smaller variance in the expression profile that is regulated than in the expression profile that regulates. A full characterization of the underlying regulatory mechanisms (e.g.*

direction and strength) extracted from the profile variabilities is beyond the scope of this paper, but we believe that rigorous experimental quantification is a necessary first step in this direction."

Two papers by Manu et al. (2009; PLoS Biol and PLoS Comp Biol) and a paper by Gursky et al. (2011; BMC Syst Biol) propose explicit regulatory mechanisms by which reproducibility of gap gene expression could be increased over time. It would be appropriate to mention this and to to cite these papers in the Discussion.

Reply: We have included these 3 references in the last section of the results and in the discussion. The following two sections have been altered:

"*The dynamics of gap gene reproducibility are remarkably correlated with the overall gap gene dynamics uncovered in Figure 6, giving us a hint of a functional significance for the collective culmination network dynamics. In particular, this synchronous increase in reproducibility from one internuclear distance to half an internuclear distance could be the result of specific regulatory interactions among the zygotic gap genes, as previously suggested (Lacalli 1988, Edgar 1989, Manu et al. 2009a,b, Gursky 2011).*"

"*This reproducibility increases further to a half-internuclear distance at the peak of gap gene expression, 15 min before gastrulation. Explicit regulatory mechanisms by which reproducibility of gap gene expression could be increased over time have been proposed (Lacalli 1988, Manu 2009a&b, Gurski 2011), and should be subjected to a quantitative test using our new set of data.*"

One methodological aspect that was not clear to me is how the authors established the orientation of their embryos. This is not explained in the current version of the manuscript. How is it possible to rigorously estimate azimuth angles, or to align an embryo for imaging of its dorsal surface, if no markers for DV expression (eg twist) were used? What does 'eye selection' (eg caption, suppfig 4) mean? Please clarify.

Reply: We agree with the reviewer that this point was not at all covered in our manuscript. We essentially split our original data set of 163 embryos in half, and only retain 87 embryos that we consider to be oriented along DV (instead of LR) such that the dorsal side is easily identifiable. This choice is made by eye inspection of each individual embryo, i.e. non-computerized. We acknowledge in the manuscript that this choice constitutes the single largest source of systematic error in our profile estimations, which is why we performed two independent estimations of this systematic error. We added an additional sub-section entitled "*Post-imaging embryo selection*" to the Materials and Methods section of the revised version of our manuscript to make this selection process explicit.

I have two questions about Fig 1 and the issue of crosstalk: (1) Why is there so much cross-talk in the 633 channel? The absorption spectra in Fig 1A would not lead me to expect that. (2) It is generally assumed that anti-rat and -mouse secondaries show high levels of cross-

Reply: (1) The reason why there is crosstalk in the 670/40 optical channel at all is because the absorption spectrum of the 594 nm fluorophore is mildly excited by the 633 nm laser line, and because the emission spectrum of the 594 nm fluorophore has a non-negligible overlap with the 670/40 channel. Please also note that the y-axis is in log units, which means that the crosstalk in the 670/40 channel is 100 times lower than the lowest part of the actual fluorescence signal, and it is on average more than 10 times lower than the strongest crosstalk in the 520/40 channel. To make this point clearer to the reader, we modified the sub-section "*Fluorophore cross-talk*" in the Materials and Methods section of our revised manuscript.

(2) To prevent cross-binding between the rat and mouse primary antibodies we first incubated the embryos in Ginea Pig, Rat and Rabbit primaries, followed by their respective secondaries; subsequently we re-blocked the embryos in blocking buffer, after which we applied the mouse primary and secondary antibodies. Since we think that this protocol might have wider applicability we upgraded the section "*Antibody staining and confocal microscopy*" of the Materials and Methods to clarify this point to the reader.

Reply: We increased the size of Figures 1 and 2 to make the individual panels easier to read. We also increased the font size of the labels in Figures 3 and 4.

Reply: We clarified our choice of origin on p3, and we adjusted the caption in Figure 1C to make this point explicit. We also added tick labels on the x and y axes of that figure and we increased the size of the tick marks.

Reply: We changed the citations in both places with more appropriate references.

p4, typo: 'an methods' should be 'and Methods'.

Reply: We corrected this typo.

p4 (and elsewhere): please clarify if the cross-talk estimate is used to correct the measured expression patterns; I guess so but it is not clear from the current wording.

Reply: The main goal of estimating the different sources of systematic errors in our measurements was to make sure that the sum total of such errors only constitutes a small fraction of the total variance in our measurements. In that case, the majority of the overall measured variance can be regarded as a statistical fluctuation in our measurements and used for further statistical analysis. Note that the cross-talk contributes to less than 1% of the total variance and constitutes the smallest source of error.

Because the various sources of systematic errors all have the potential to be correlated, we did not feel comfortable to subtract any one of them, but rather work with the total variance in our measurements, estimate that the sum total of the systematic errors is below 20%, and then any inference we extract from the total variance will be an upper bound. That is the case for example for our positional error estimates: half a nuclear distance is an upper bound, and it could in principle be even lower than that, weren't it for our systematic errors. We added two sentences clarifying this point at the end of the sub-section "*Experimental error quantification*" of the main text.

"*In the remaining sections we will continue to work with the measured total variance; this is a cautious choice, as subtracting measurement errors from raw data is only feasible if there are no correlations among the various error sources, which is often unknown. The only systematic error that we take out is the time correction (detrending) because our age classification of the embryos (staging) gives us some additional information about the structure of the error and we know the underlying model.*"

p4, the statement 'ending with gastrulation' is confusing since gap genes are expressed a little bit beyond gastrulation; I guess the authors take gastrulation as the endpoint of their analysis? Please clarify.

Reply: We changed this sentence to: "The Drosophila gap genes are endogenously transcribed for a time span of 1-2h, peaking roughly 2.5h after the onset of embryonic development."

p6: I am not sure I understand why 163 embryos were stained in total, but only 80 use in the analysis; please clarify.

Reply: Of our 163-embryo data set we only retain those 87 embryos that have a DV orientation for our analysis. Out of those 87, we discarded 7, who had already started gastrulation, which made the profile extraction complicated given the misalignment of the nuclei. We have clarified

this in the new sub-section entitled "*Post-imaging embryo selection*" in the Materials and Methods section of our revised manuscript.

p12, the sentence citing a 'Fig. 1C' must be a leftover from an earlier version of the manuscript; there seems to be no such figure in the present version. Please correct.

Reply: We changed the sentence accordingly.

p12 and fig 1: please clarify which channel corresponds to which row in the matrix. This is a bit confusing.

Reply: We have adjusted the section about "*Fluorophore crosstalk*" in the Materials and Methods section to make the matrix more explicit, and we have added the gene abbreviations and fluorophore excitation wavelengths to Figure 1 to identify each row.

caption, fig2: 'profile in A shown in blue': I don't know what this refers to.

Reply: We changed this sentence to: "The profile in A is shown in blue here."

Panels 3b and suppfig 3b are the same. Is it required to reproduce them twice?

Reply: This is an important panel to keep in the main body of the manuscript and we replot it in the supplement for completion. We leave it up to the editors to decide whether they want to omit it in the supplement.

fig6: please provide a key to lines in panels D and E; I have no clue what the dashed and dotted lines in E are.

Reply: We adjusted the caption to panels 6D and 6E for clarification. One of the red lines of figure 6E was mistakenly plotted as a dotted instead of a dashed line, which was indeed confusing for the reader. We corrected the figure accordingly.

fig7 is completely overloaded; please declutter, eg by leaving out or dimming the expression patterns; currently they hide the actual results shown in this figure and make it very hard to read.

Reply: We put the expression profiles in there to provide orientation w.r.t the AP-axis. We agree that for this goal we do not need the variances. Hence we only keep the mean for each profile and made it dimmer.

caption, suppfig1: descriptions of panels A and B are mixed up.

Reply: We corrected this mix-up.

caption, suppfig4: typo, 'allows allows'.

Reply: Also corrected.

SUMMARY
The overall goal of this paper is to understand the variance in the gap gene network that patterns the Drosophila blastoderm embryo. This is part of a general line of research that addresses precision in developmental processes, attempting to understand its origins and its relationship to environmental and genetic variation. Variance and co-variance can also be used to infer underlying network architecture. Because the observed variance is a combination of biological fluctuations and experimental error, the authors aim to precisely quantify the degree of experimental error in their measurements. To do so, they make three technical contributions to the field: 1) increased precision in staging of embryos (from ~10 min to 1-2 minutes) 2) quantification of 8 sources of experimental measurement errors. 3) definition of a position-dependent variance metric that they use to analyze the precision of the network over time. They conclude experimental errors sum to less than 20% of total
variance for any gene, allowing them to analyze the remaining 80% of measured variance to understand the precision of the gap gene network. By analyzing the co-variance among gap gene expression profiles, they demonstrate that positional errors decrease two fold over the course of cellularization (from one internuclear distance to 0.5), and that they reach a minimum 15 minutes prior to gastrulation, synchronously across all genes (except the Hb anterior domain). These results demonstrate the utility of quantitative measurements, and provide some biological insights.

GENERAL COMMENTS
Technically, this work is very strong. The methods are rigorous, and its refreshing to see quantitative measurements taken so seriously. However, the biological insights are less novel. They build on concepts already solidified in the field (such as the canalization of the network and cross-regulation in the gap gene network). As written, the paper is focused on the methods, which is fine given that's where it makes its most substantial contribution. But the writing is quite technical and could benefit from more explanation of the high-level logic of their approach. By providing more context in the introduction, and by explaining their logic throughout the results, I think the authors could improve the accessibility of their manuscript a great deal. The paper would then not only be technically impressive, but it could also make the case more broadly for how careful quantitative measurements in wild-type embryos can yield biological insights.

Reply: We thank the reviewer for a very careful reading and for the insightful and constructive comments in regard to our manuscript. We agree with most of the identified review points and we think that addressing them has helped greatly to advance our work. We have made major modifications to the introduction, and we included numerous additional references that should enhance the context of our work. We also attempted to clarify all open technical questions that the reviewer raised. Please see below for our specific answers.

MAJOR POINTS
Here I focus on places that could use additional conceptual or technical description.

"The final macroscopic outcome of developmental processes in multicellular organisms results in structures that are remarkably similar between individuals of a given species." (p. 3) This sentence would benefit from some references and further description. What are examples of this type of morphological precision? Which features are particularly notable for the Drosophila embryo?

Reply: We added two references after the opening sentence of our introduction and expanded its description into a full paragraph with further references.

This sentence in the intro presents a hypothesis about morphological precision as proven: "This similarity has its origins in the reproducible spatial patterns of morphogen concentrations in the early embryo." (p. 3) But my understanding is that this is an active area of research, and is in fact the central question of the paper. Can reproducibility/precision in morphology be traced to precision in the underlying molecular patterning mechanisms? The alternative is that patterning is itself messy, but cleaned up later. This may of course vary across different patterning mechanisms in different systems. Examples of active models beyond Drosophila would be useful to mention here. My understanding of the overall justification of this line of research is to ask whether the precision observed in the patterning system is sufficient to account for the precision observed in morphology. Or alternatively, focusing on just the quantitative features of the network itself, to understand which features are variable, how this relates to network architecture, and how these features respond to environmental and genetic variation. This paper is then situated as an attempt to precisely measure variation itself, which is technically challenging. I suggest that this is described more explicitly in the introduction.

Reply: We have completely rewritten the first two paragraphs of the introduction to incorporate the concerns and most of the suggestions outlined above.

"Conceptually, we are seeking..." (p. 3) This sentence lists many network properties that could be measured, but does not state how the measurements will be useful in the long-term. Clarifying this would help the reader grasp the logic of this type of study. For understanding this question of precision in patterning, what are the most relevant measurements? Should we be most interested in invariant features or those that vary, or both? This study is focused on one genotype, but could extending these types of studies to additional genotypes be fruitful?

Reply: We have greatly modified our introduction, and we believe this has helped clarify the overall logic of our study.

Reply: We incorporated the first paragraph of the discussion entirely into the introduction.

Reply: The main goal of estimating the different sources of systematic errors in our measurements was to make sure that the sum total of such errors only constitutes a small fraction of the total variance in the measurement, such that that majority of this variance could be further analyzed as a statistical fluctuation in our measurements. Because the various sources of systematic errors all have the potential to be correlated, we did not feel comfortable to subtract any one of them, but rather work with the total variance in our measurements, estimate that the sum total of the systematic errors is below 20%, and then any inference we extract from the total variance will be an upper bound. That is the case for example for our positional error estimates: half a nuclear distance is an upper bound, and it could in principle be even lower than that, weren't it for our systematic errors. We added two sentences clarifying this point at the end of the sub-section "Experimental error quantification" of the main text.

What our 8 equally populated time bins is concerned we had to make this choice for statistical reasons. For each time bin we are interested in a variance; but to estimate this variance we need at least 10 embryos per time window, which results in 8 bins for our 80 embryos. A larger collection of embryos would certainly lead to a larger number of windows as long as their widths do not fall below the 1-2 min resolution. A sample of N=10 allows us to undoubtedly distinguish a 1%EL reproducibility from a 2%EL reproducibility. With N=10, the variance (+/- standard error due to sample size) corresponding to $\sigma_x/L = 2\%EL$ early in n.c. 14 is $\sigma^2 = 0.0004 +/- 0.0002$, while the variance corresponding to $\sigma_x/L = 1\%EL$ late in n.c. 14 is $\sigma^2 = 0.0001 +/- 0.00004$. Therefore the increase in reproducibility that we observe in statistically relevant. To bring this to the attention to the reader we added the following sentence to section "Reproducibility of gap gene expression profiles":

"The binning step is necessary to increase statistical significance (i.e. the standard error on the variance sigma^2 is given by sigma^2*sqrt(2/N) for sample size N) in order to undoubtedly distinguish reproducibilities at the 1%EL level."

What makes our time binning different from previous studies is that 1) we have a 1-2 min accuracy on the borders of our time bins; and 2) we can afford to work with a time-detrended data set and do not have to resort to any other normalization procedures that are usuallly applied to compare different embryos; this means that essentially we can work directly with the raw data.

Minor points
Here I focus on wording that could be clarified. There also are numerous grammatical errors throughout, but these should be addressed after revision since the text may change a great deal.

p. 3 "the level of reproducibility that is actually relevant to biological processes" is an ambiguous phrase. Does it mean measuring reproducibility at the appropriate scale (i.e. molecules, cells, tissues, etc.) or does it mean the level of variation that is tolerated by the system while still maintaining function?

Reply: We mean the latter and clarified this in the text.

p. 3 "inherent experimental error due to variable conditions" isn't very precise. What types of experimental errors? My understanding is that the major issue is with these techniques is averaging many stains together - therefore potentially convoluting the variation in the quantitative relationship between genes. Please clarify the logic here and how staining for multiple proteins in the same embryo circumvents these limitations.

Reply: This is a very important point that we wanted to bring across in our manuscript. We edited the sentence in question to make the statement less vague, and more importantly we added a paragraph to the results subsection "*Experimental error quantification*" that clarifies the various sources of experimental errors and which ones we deal with in this paper (errors within a given embryo, and errors across embryos within a given experiment) and which ones we don't (errors across experimental batches).

p.3 The authors present the use of fixed tissue as a limitation of previous methods, but employ a fixed tissue approach themselves. This is confusing, especially since they are employing a similar method for obtaining dynamic information from fixed tissue (extent of invaginating cell membranes), but at a higher level of resolution. (Previous studies used something like 10 min bins, while this study uses 1-2 min bins). This should be stated more clearly.

Reply: This paragraph was purely meant to point out the well-known limitations of the fixed-tissue/immunofluorescence approach. We adjusted the text to clarify that our approach is a fixed tissue approach as well; and we cited previous attempts to reconstruct dynamics from

fixed tissue.

"In particular, what is the final reproducibility at gastrulation and how does it compare to the reproducibility of the Bcd gradient?" The first part of this sentence is ambiguous - reproducibility of what feature of gastrulation?

Reply: We mean reproducibility of gene expression levels. We adjusted the text for clarification.

"After profile alignment and correction for tissue shrinkage due to the fixation process ..." The logic of this procedure is unclear from the main text. It appears they will first relate furrow depth to absolute time using live bright-field imaging. They can then align these time traces to determine the error in assigning the appropriate time from a single time-point. To extend this procedure to fixed embryos they need to normalize to their live measurements, since fixation is known to cause embryos to shrink. Even in the Methods, this normalization calculation isn't particularly clear. The AP and DV dimensions of the embryo shrink by ~5%, but how this is used to normalize the live furrow depth isn't immediately obvious. I suggest explaining the logic of their procedure more completely in the main text, and elaborating on the normalization calculation in the Methods.

Reply: We adjusted the main text (2nd paragraph of "*Developmental time measurements in fixed embryos*") and the methods section (end of "*Measurements of the invagination depth of the membrane furrow canals*") to clarify.

The paragraph beginning "The precision of 1-2 min with which..." would be easier to understand if the conclusion was presented as the topic sentence. Something like "Previous quantitative studies of expression in this system have used roughly 10 minutes time intervals. By staging embryos in 1-2 min intervals, we determined that 10-20% of the variance in these 10 minute time intervals is due to underlying expression dynamics."

Reply: We adjusted the mentioned paragraph accordingly.

"Hence, an overwhelming majority of the raw profile-to-profile variability seen in Figure 2C can be attributed the dynamics of the expression levels." This conclusion is obvious from previous studies as well. The more relevant result from this particular experiment is the ability to decompose the variance, as stated in the next sentence.

Reply: We added "*... as could be inferred from previous studies and now made clear by classifying embryos into smaller time classes...*" to this sentence, but we did not find any reference that actually makes the point about the variability stemming from the residual dynamics in the time interval. In fact, we did not find any reference that attempts to decompose the variability in its different sources of contribution (except Myasnikova et al. 2009, who analyzed the different sources of imaging noise). Therefore we did not put any citation here. If the reviewer feels strongly about this, however, we would be glad to hear which work he has in mind that has previously addressed this issue.

Reply: We reworded the sentence; but we want to point out again that we never subtract any variance and always keep working with the raw variance as we're seeking lower or upper bounds.

Reply: We agree with the reviewer that our description of the dynamics was not appropriately put into context with what is known in the literature. Therefore we added a paragraph to the results section "Dynamics of gap gene expression profiles":

"*Observation of these dynamic features of the network are a signature for the well-known mutually repressive interactions between the gap genes (cite Hulskamp, Kraut, Struhl, Kosman, Clyde). Although the coordination of the interactions might be expected from the known network topology (cite Jaeger, Manu, Papatsenko), the precision of the observed synchrony suggests that some intrinsic collective organization of the gap gene network is at play (cite Edgar, Lacalli, Clyde, Manu, Lander). In particular, the peak in the slopes at the intersection of the gene expression profiles of Kr and Kni at 42 min might be a point where the network is optimized for certain characteristics such as the accuracy with which it can determine cell fates.*"

and another paragraph to the Discussion:

"*Explicit regulatory mechanisms, by which reproducibility of gap gene expression could be increased over time, have been proposed (Lacalli et al, 1988; Manu et al, 2009a,b; Gursky et al, 2011), and can now be subjected to a quantitative test using our new set of data. Already at the level of Bcd it has been suggested that spatial averaging among neighboring nuclei is necessary to achieve the observed precision (Gregor et al, 2007a; Erdmann et al, 2009). Similarly, a combination of temporal and spatial averaging might be at play in achieving*

the further increase in reproducibility by the gap gene network. However, such averaging mechanisms have a blurring effect on sharp gap boundaries, and it will be interesting both theoretically and experimentally to understand whether a balance can be found."

Reply: We changed the relevant sentence and cited the suggested references, as well as Jaeger 2004 and Crombach 2012, as recommended by Referee #1. We also added a sentence at the end of this paragraph highlighting the point of extracting regulatory mechanisms from profile fluctuations. "*In particular, the constancy of σx/L across the Hb/Kr, Kr/Kni, Kni/Gt, and Gt/Hb borders hints at a mutual regulation of these genes, as hypothesized before (Kraut 1991, Papatsenko 2011, Crombach 2012), rather than a uni-directional repression (Jaeger 2004a and Jaeger 2004b), in which case we would see a smaller variance in the expression profile that is regulated than in the expression profile that regulates. A full characterization of the underlying regulatory mechanisms (e.g. direction and strength) extracted from the profile variabilities is beyond the scope of this paper, but we believe that rigorous experimental quantification is a necessary first step in this direction."*

Reply: We we expanded the sub-section entitled "Measurements of the invagination depth of the membrane furrow canals" in the Materials and Methods section of the revised manuscript to precisely explain our procedures. The sentence cited by the reviewer now reads: "*For each time frame t, the depth of the FC (defined as the distance between the furrow canal and the edge of the embryo – see inset of Figure 2A) was manually measured at three different places along the dorsal side of embryo i and then averaged to obtain a single value deltaFC(i,t), which was subsequently monitored as a function of absolute FC time until gastrulation, as shown in Figure 2A.*"

Reply: We opted to partially follow this suggestion. This is a quantitative work, symbols/ variables are used in equations, and the reader should not be confused in what is used where. We made sure, however, that we state clearly in the main text and in the caption what the symbols/variables we use mean. For intensity values we changed each label to the descriptive text label "Intensity" in all figures.

The figure legends also have descriptive titles. It might be easier for readers to grasp the purpose of the figure if the titles presented conclusions instead. For example, Figure 2: Precise temporal staging allows quantitation of time dependent variation in gene expression at every position.

Reply: We partially followed this suggestion and adjusted the titles for figures 2, 3, 5 and 8.
Figure 2: Staging of gene expression profiles in fixed embryos with minute precision.
Figure 3: Quantification of imaging noise and antibody non-specificities.
Figure 5: Summary of systematic errors of the gap gene expression profiles.
Figure 8: Temporal evolution of gap gene expression profile reproducibility.

Figure 2:
A: Inset is difficult to see - red/orange label is obscured. FC should be defined in the legend for readers that go through figures independently of the text. Mitosis 13 is used to label the figure, but the legend uses n.c. 14. Best to be consistent with this nomenclature.
B: Individual traces and points are extremely hard to see. Granted, the point is that they all overlay very precisely, but it may be worth increasing the size of this figure to make the features named in the legend easier to identify.
C: The paper discusses intervals in terms of time rather than FC depth (which is what is used in the legend for this panel). I think this should be made consistent. I vote for time since it is more obviously linked to the technical advance in this paper. Also, there are multiple dotted lines in the figure so referring to "the grey dotted line" should be changed to "the vertical grey dotted line". Vertical axis should be "Gt Intensity".
D: I presume this is also a trace for Gt at a single position, but this is not made clear in the legend.

Reply: We adjusted figure and caption to accommodate all suggestions:
A: We changed the color to red, define FC in caption and methods (as also suggested by reviewer #1), and we changed "the onset of n.c. 14" to "mitosis 13" in the caption.
B: We increased the overall size of the figure to emphasize these features.
C: We changed the caption to express the interval in terms of time; we altered the dotted line confusion; we modified the vertical axis label.
D: We added "... position x/L..." to the first sentence.

Figure 3:
A: I find this schematic very hard to read. The yellow is nearly invisible, and there are a number of different molecular species depicted. The diagram is more complicated than the

Reply: A: Schematic was adjusted following the suggested changes. For increased clarity, we also matched the color of the fluorophores with the color of the axes of panel B. We changed the axis labels as suggested. For consistency, we also changed them in Sup. Figure 3.

Reply: We made this line solid and moved it behind the black line (also on Sup Figures 4 & 5). We changed the captions accordingly.

Reply: We labeled the gap genes above each column. A: The reason for keeping the mean profiles are: 1) Show that the error is a small fraction of the variance which is itself a small fraction of the mean profile; and 2) Allow the reader to identify where the signal/noise ratio is less than one (i.e. where the curves for mean and variance cross), showing that this corresponds to the background noise and is purely an experimental artefact. We agree with the reviewer that these mean profiles are a bit dim and hard to identify, which is why we changed their colors to gray.

Reply: Our use of the symbols is conform with previously published data. Colors represent genes (consistently throughout the entire paper), triangles represent borders (left/right) and lines represent peaks. We prefer to keep this nomenclature, but we decreased the size of the triangles to declutter the figure. We added to the text a sentence that states that the dynamics of the boundaries are more important near the edges as previously described by Reinitz and colleagues.

Reply: The two curves (blue and black) represent different yet complementary visions of profile reproducibility. The fact that they fall on top of each other is in fact proof of their consistency. The last paragraph of the results section is making this point explicit. We also simplified the figure according to reviewer #1's comments by removing the variances of the gap gene profiles in the background and by making the mean profiles dimmer. Furthermore we changed the different triangles all to open round circles to make the blue curve easier to follow. We believe this removes most of the distracting elements in the figure.

Reply: We agree with reviewer 2 that a 5-stain protocol would go beyond the scope of the current manuscript and represents a significant experimental challenge. We would also like to point out that the influence of Bicoid beyond the 20 min mark in n.c. 14 is unclear. Bcd mRNA is fully degraded by this time (Little 2011), the protein gradient starts to decay, and in the view of the current manuscript Bicoid's job of conferring positional information and precision to the downstream network is done.

Reviewer #3 (Remarks to the Author):

The authors perform quantitative studies of gap gene expression in early Drosophila embryos. They focus on four gap genes in their current work, hb, Kr, gt and kni, through the use of both newly-generated and existing antibodies against the products of these genes. The authors provide careful and systematic evaluations of the experimental errors in quantifying the immunostaining data in whole mount embryos. They provide evidence suggesting that immunostaining is a viable approach for quantitatively measuring gene expression profiles in embryos (an important idea that is not entirely new). They also provide data demonstrating that the reproducibility of the gap gene expression profiles increases two-fold during the course of nuclear cycle 14 (a notion that has also been available in the literature).

This is a carefully-designed and well-executed technical study. The systematic evaluations of the experimental errors should be useful to the field. Unfortunately, the current study remains at a technical level without significantly advancing our knowledge of the gene regulatory network that instructs embryonic patterning. In particular, a lack of quantitative data for the morphogen gradient of Bicoid under their current (and likely improved) technical framework makes it difficult to assess how the developmental system operates mechanistically.

Reply: We agree with the reviewer that a version of the suggested experiment would be very interesting if n.c. 12 and 13 were to be added to the observed time frame. It would most likely give possible mechanistic insights into the system. However, as reviewer 2 correctly points out, technically such an experiment is extremely challenging as one would have to find 5 antibody sets that work together in the same embryo, which is currently not feasible. We have, however, started to look at Bicoid together with sets of three other gap genes, and we are preparing an additional manuscript that will report on our results regarding these experiments. We believe that adding them here would go beyond the scope of the current manuscript.

The current work conspicuously lacks a specific hypothesis presented and tested.

Reply: We have largely modified the beginning of our introduction, hopefully providing more context and a clear hypothesis. Here are two excerpts that we hope addresses directly the reviewer's concerns:

"Uncovering the origin of precise and reproducible structures in biological processes is a fundamentally quantitative question, the answer to which can be one of two very distinct concepts. In one view, each step in the process is noisy and variable, and noise reduction only occurs through integration of many elements or collectively within the whole network of elements. In the other view, each step in the process has been tuned to enhance its reliability, at times maybe even down to the limits set by basic physical principles. Can the precise and reproducible features observed in the patterning system be sufficient to account for the precision observed in morphology? Focusing on these quantitative features of the network will ultimately lead to our understanding of which network properties are truly reliable which ones are variable, how they relate to network architecture, and how they respond to environmental and genetic variation."

"As a complementary approach, precise measurements in an intact, wild-type system can reveal the quantitative relationships within the network in an unperturbed, natural state. Data of the latter approach has been missing thus far, and it will become essential for the mathematical modeling of gene interactions and for increasingly complex developmental systems."

The conclusion presented in the Discussion "Reproducibility inherent in the Bcd gradient is transferred to the gap genes early in n.c. 14, reduced two-fold within the gap gene network..." appears to imply, but does not clearly state, that the network itself is responsible for this two-fold reduction, a notion that has been championed by Reinitz and his associates, among others.

Reply: We have added references to articles by Manu and Gursky in the discussion.

"Reproducibility inherent in the Bcd gradient is transferred to the gap genes early in n.c. 14, reduced two-fold within the gap gene network [Manu et al. 2009 a & b, Gursky 2011], and subsequently relayed to the pair-rule genes later in n.c. 14."

This conclusion also appears to ignore completely the potential effects of time- or space-averaging, a hypothesis that has been available in the field (including Dr. Gregor's own previous studies).

Reply: This is a very important point, and we thank the reviewer for pointing this out to us. We have added a paragraph in the discussion to address averaging:

"Explicit regulatory mechanisms by which reproducibility of gap gene expression could be increased over time have been proposed (Lacalli:1988, Manu:2009b, Manu:2009a,

Gursky:2011), and could potentially be subjected to a quantitative test using our new set of data. Already at the level of Bcd it has been suggested that spatial averaging among neighboring nuclei is necessary to achieve the observed precision (Gregor:2007b, Erdmann:2009). Similarly, a combination of temporal and spatial averaging might be at play in achieving the further increase in reproducibility by the gap gene network. Such averaging mechanisms though have a blurring effect on sharp gap boundaries, and it will be interesting both theoretically and experimentally to understand whether a balance can be found."

One of the claims made by the authors "We use this analysis to extract gap gene profile dynamics with ~1 min accuracy..." (in Abstract) is misleading because, although they have done calibrations for embryo time, their analysis was based on data from embryos that are grouped into 8 time classes (each lasting much more than 1 min).

Reply: The choice of regrouping our data in 8 equally populated time bins was made for statistical reasons. For each time bin we are interested in a variance; but to estimate this variance we need at least 10 embryos per time window, which results in 8 bins for our 80 embryos. A larger collection of embryos would certainly lead to a larger number of windows as long as their widths do not fall below the 1-2 min resolution. A sample of N=10 allows us to undoubtedly distinguish a 1%EL reproducibility from a 2%EL reproducibility. With N=10, the variance (+/- standard error due to sample size) corresponding to sigma_x/L=2%EL early in n.c. 14 is sigma^2=0.0004+/-0.0002, while the variance corresponding to sigma_x/L=1%EL late in n.c. 14 is sigma^2=0.0001+/-0.00004. Therefore the increase in reproducibility that we observe in statistically relevant. To bring this to the attention to the reader we added the following sentence to section "Reproducibility of gap gene expression profiles":

*"The binning step is necessary to increase statistical significance (i.e. the standard error on the variance sigma^2 is given by sigma^2*sqrt(2/N) for sample size N) in order to undoubtedly distinguish reproducibilities at the 1%EL level."*

What makes our time binning different from previous studies is that 1) we have a 1-2 min accuracy on the borders of our time bins; and 2) we can afford to work with a time-detrended data set and do not have to resort to any other normalization procedures that are usually applied to compare different embryos; this means that essentially we can work directly with the raw data.

After reading through this nicely-executed technical work, one is left with an impression that after all we really have not gained much new mechanistic insights.

Reply: The main goal of this paper was not to be mechanistically insightful, but to provide a framework for using the natural fluctuations of gene expression within a population of individuals. In particular, some features, such as the analysis of the temporal and spatial distribution of sigma_x/L, already give us clues about underlying regulatory interactions, namely: 1) the pattern borders are the result of mutual regulation of the neighboring genes, rather than unidirectional repression 2) cross-regulation among gap genes increases the reproducibility of

pattern formation.

To further emphasize these points in the paper, we performed several changes to the results section entitled "*Reproducibility of the gap gene expression profiles*": We added the references (Kraut 1991, Jaeger 2004, Papatsenko 2011, and Cromback 2012) recommended by reviewers #1 and #2. We added a sentence at the end of this paragraph highlighting the point of extracting regulatory mechanisms from profile fluctuations:

"*In particular, the constancy of σx/L across the Hb/Kr, Kr/Kni, Kni/Gt, and Gt/Hb borders hints at a mutual regulation of these genes, as hypothesized before (Kraut 1991, Papatsenko 2011, Crombach 2012), rather than a unidirectional repression (Jaeger 2004a,b), in which case we would see a smaller variance in the expression profile that is regulated than in the expression profile that regulates. A full characterization of the underlying regulatory mechanisms (e.g. direction and strength) extracted from the profile variabilities is beyond the scope of this paper, but we believe that rigorous experimental quantification is a necessary first step in this direction.*"

The existing quantitative data (in addition to the Bcd data that should be generated under their current framework) could potentially yield some new mechanistic insights. For example, the authors have the co-staining data for all the four gap gene products in individual embryos, an advantage highlighted repeatedly by the authors but not fully utilized by them.

Reply: We agree with the reviewer that we have not fully utilized the big advantage of co-stained gene expression profiles. Although, the collective network response would have been less evident if only pairs of genes would have been co-stained. The full power of this approach will be discussed in a follow-up paper, in which we apply it to measure the information content that resides in all four gap genes. Quantification of information relies on calculating the covariance matrix across all four major gap genes which is only possible when they are measured simultaneously and when the contribution of the experimental errors to the overall profile variances are small. We have improved the second paragraph of the discussion section that outlines potential applications (including a quantitative estimation of information in the system and a reference to a preprint version of our follow-up paper).

Is it feasible to extract information that can add to our understanding of how these gap gene products co-evolve at the level of individual embryos?

Reply: This is a great suggestion! We feel, however, that addressing this point in the current manuscript goes beyond the focus we were aiming for.

Would an analysis that is truly at a ~1 min temporal accuracy allow the extraction of new information about how the system operates?

Reply: As noted above, our current data set is statistically not suitable for a "truly 1 min temporal accuracy".

Reply: The impact of embryo size variation on reproducibility is very interesting and something the understanding of which the authors are actively pursuing. Natural size fluctuations in this system are only of the order of 3.5%. Thus it requires batches of large sample size to to be able to draw statistically meaningful conclusions. Furthermore, at the level of the gap genes these size fluctuations have largely been compensated for and approach the detection limit of our experimental approach (see Houchmandzadeh 2002). Therefore we estimate the effect of size variation on our conclusions to be small (albeit interesting), and we opted not to mention them in the current work.

A technical comment: the authors tend to make statements throughout the manuscript as if they are not aware of the existing knowledge in the literature. The discussion about the two-fold increase in the reproducibility of gene expression profiles (above) is one example. As another example, in the first paragraph in Discussion, the authors state that "Data of the latter approach has been missing thus far..." when referring to quantitative relationships in wt embryos. Do they really mean that? Accurately relating their experiments and findings to the literature and the existing knowledge might improve the manuscript (What is the current state-of-the-art in the field? What specific aspects has the current work advanced and how?).

Reply: We agree with the reviewer that our original version was lacking references about the increase of profile reproducibility. We have corrected it by citing Manu 2009a,b in the Results and Discussion sections. In addition to that, we have put references to Lacalli 1988, Edgar 1989 and Gursky 2011, which hypothesize how regulatory interactions can produce this increase in reproducibility.

We also agree with the reviewer that our sentence about missing data so far was misleading, and we now provide a revised version of the text: "*It is challenging to perform experiments that take advantage of the latter approach as it requires extreme care to minimize sources of experimental error that allow for the data to be accurately quantified. Nevertheless, such measurements are essential both for mathematical modeling of gene interactions and for understanding increasingly complex development systems.*"

Thank you again for sending us your revised manuscript. We are now satisfied with the modifications made and I am pleased to inform you that your paper has been accepted for publication.

With regard to the datasets, we would be grateful if you could supply the following:
- for each measured embryo, the individual raw and processed dorsal expression profiles, as well as the measured depth of the membrane furrow canal and the resulting inferred developmental time. These values can be provided either in a single combined table (tab delimited, csv or Excel) or in two tables--one for the raw data and one for the processed profiles.
- a table (tab delimited, csv or Excel) that provides the calibration curve for the time-dependent furrow depth.
- a zip files that contains the full set of images (with labels that allow to link the images to the profiles supplied in the file above), provided the entire dataset is *less than 1GB*. Otherwise, we would be grateful if you could deposit your data on your website or preferably in one of the general data repositories, such as Dryad (datadryad.org) or a suitable alternative.
- please include a README file that specifies the content of each file. Some additional guidelines for 'source data' in tabular format are available here <http://www.nature.com/msb/authors/source-data.pdf>.
- please a 'data availability' section to Materials and Methods where these datasets are listed and send us the amended text.
If the files are too large to be shared by email , please use a FTP site or a file sharing service such as drive.google.com , www.yousendit.com , dropbox.com.

Thank you very much for submitting your work to Molecular Systems Biology.