On Transcription And Approximations

Shengtong Sun Raiyan Raya Todd Pocuca Louis Ward

BIOPHYS 3G03
McMaster University

Abstract

The decreasing cost of RNA sequencing has made transcriptome analysis a cornerstone of biological research. While statistical methods based on the negative binomial distribution are widely used to analyze transcript counts, mismatches between model and reality have emerged. Particularly, the presence of bimodal transcript count distributions cannot be sufficiently explained by a single negative binomial. To investigate whether the fundamental process of transcription could inherently generate such patterns, we developed a stochastic, agent-based model of transcription derived from mechanistic first principles. Simulations across a broad range of parameter values revealed that our model consistently produced transcript count distributions exhibiting underdispersion relative to the best-fitting negative binomial, and no instances of bimodality were observed. These findings suggest that the core transcriptional machinery in isolation may not fully account for the overdispersion seen in experimental data, potentially highlighting the role of other cellular factors or measurement noise. Furthermore, the absence of bimodality in our simulations lends support to the hypothesis that bimodal expression patterns in seemingly homogenous cell populations may arise from underlying cellular heterogeneity or more complex regulatory mechanisms beyond the minimal transcription system modeled here. Our work provides a baseline for understanding the statistical properties of transcript counts generated from mechanistic principles and underscores the need for continued investigation into the factors contributing to transcriptomic variability.

Introduction

The decreasing cost of high-throughput sequencing has made whole-transcriptome experiments an increasingly popular tool for life science researchers investigating biological phenomena. These experiments provide a snapshot of the RNA molecules present within a biological sample, generating transcript count profiles from bulk tissue or single cells.

Despite the prevalence of transcriptome data, the statistical methods used to analyze the resulting counts were developed without detailed knowledge of the data generating process underlying transcription or the effect of the inherent noise and artifacts introduced by the RNA sequencing process. The two most popular libraries for downstream transcriptome analysis—`DESeq`[1] and `edgeR`[2] —were both developed primarily based on established statistical frameworks for count data, adapting and extending the negative binomial model for transcriptomics.

Scenarios where model assumptions clashed with reality have, however, accumulated over time, particularly with observations of bimodal count distributions in seemingly homogenous cell populations[3]. These populations serve as a baseline for models of transcriptome data, as only a single set of parameters should be sufficient to explain all observed behaviour; discrepancies with the ground-truth could indicate flaws in the models used for transcriptome analysis. An alternative explanation for these mismatches between model and reality, however, is that the cell populations studied were not truly homogenous; instead, multiple cell states were generated due to processes such as differentiation and cell cycling. Indeed, this is proposed as the main model in Z. S. Singer *et al.* [3]. Whether the mechanism(s) of transcription could generate a bimodal distribution without requiring multiple cell states, however, is an open question and could imply that an alternative to the negative binomial can yield more efficient methods.

To address this gap, we sought to construct a stochastic model of transcription that follows from mechanistic first principles. The process of gene transcription, by which RNA polymerase(RNAP) synthesizes an RNA molecule from a DNA template, can be broadly divided into distinct stages: initiation, elongation, and termination.

Initiation involves the assembly of the preinitiation complex (PIC) at the gene promoter. Under the classic model of PIC assembly, general transcription factors (GTFs) and RNAP sequentially bind to the promoter

in a specific order(reviewed in [4]). Promoter escape then marks the transition to elongation, characterized by the phosphorylation of the RNAP II C-terminal domain(CTD) and the exchange of initiation factors for elongation factors[4].

During elongation, RNAP traverses the gene through a sequence of distinct conformational changes that result in the addition of each nucleotide. This translocation cycle can be conceptualized as a series of distinct kinetic states transitioning between each other at their own intrinsic rate(Fig. 1)[5]. These rates however can be modified by factors such as the availability of NTP and the torsional stress generated by the unwinding of the DNA helix ahead of the polymerase[6], which can reduce the forward translocation rate if the polymerase cannot generate sufficient force.

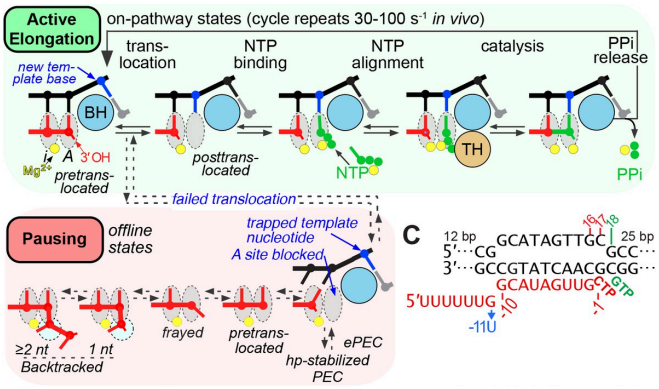


Figure 1: From J. Saba *et al.* [5]

Finally, termination signals the end of transcription, leading to the release of the nascent RNA transcript and the subsequent dissociation of the polymerase from the DNA. Following its synthesis, each transcript within the cellular environment undergoes a process of degradation. The balance between the stochastic birth of transcripts and their subsequent degradation ultimately gives rise to a "steady-state" or *stationary* distribution of transcript counts within the cell.

To investigate the characteristics of this stationary distribution, and specifically to evaluate the goodness-of-fit of the negative binomial model, we developed a stochastic model based on our mechanistic understanding of transcription. This framework allowed us to generate synthetic transcript count data under a wide range of plausible parameter values to assess how well the negative binomial can approximate the resulting count distributions

Methods

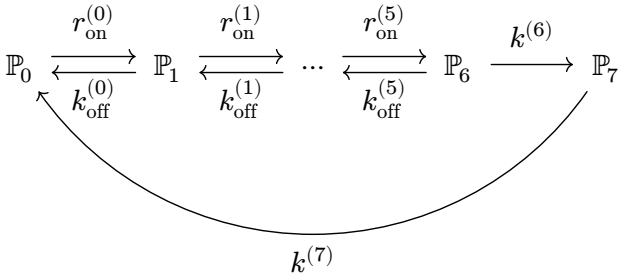
Model Construction

We implemented a stochastic, agent-based model in `NetLogo` that follows from the mechanistic underpinnings of transcription. The model tracks the state of the promoter, individual RNAP agents, and the population of RNA transcripts.

Transcription Initiation

Adopting the classical model of PIC assembly described in I. Baek [4], the promoter status $\mathbb{P}$ of a hypothetical gene is modelled through a series of states $\{\mathbb{P}_0, \mathbb{P}_1, ..., \mathbb{P}_6, \mathbb{P}_7\}$ representing the sequential assembly of the preinitiation complex (PIC) and phosphorylation of RNAP, ranging from an unbound promoter $\mathbb{P}_0$ to a fully assembled PIC and phosphorylated RNAP $\mathbb{P}_7$ (Fig. 2). Transitions between the first six states are reversible with their intrinsic rates modified by the concentrations of the general transcription factors (TFII-A, B, D, E, F, H) and RNAP. However, since only the TFIIF-RNAP dimer is relevant for TFIIF and RNAP dynamics, we only encode the dimer's concentration and kinetic rates in the model.

Promoter escape is then modelled as two irreversible steps ($\mathbb{P}_6 \to \mathbb{P}_7 \to \mathbb{P}_0$) dependent on the phosphorylation rate of the polymerase CTD and the exchange of initiation factors for elongation factors, which completely disassembles the PIC and spawns a new polymerase agent at the transcription start site.



$$r_{\text{on}}^{(0)} = [\text{TFIIF}]k_{\text{on}}^{(0)}$$
$$r_{\text{on}}^{(1)} = [\text{TFIIA}]k_{\text{on}}^{(1)}$$
$$\vdots$$
$$r_{\text{on}}^{(5)} = [\text{TFIIH}]k_{\text{on}}^{(5)}$$

Figure 2: Depiction of transcription initiation.

Transcription Elongation

Once initiated, individual polymerase agents move along a discrete one-dimensional lattice representing the gene. Following from J. Saba *et al.* [5], elongation is modelled as a branched kinetic process with distinct

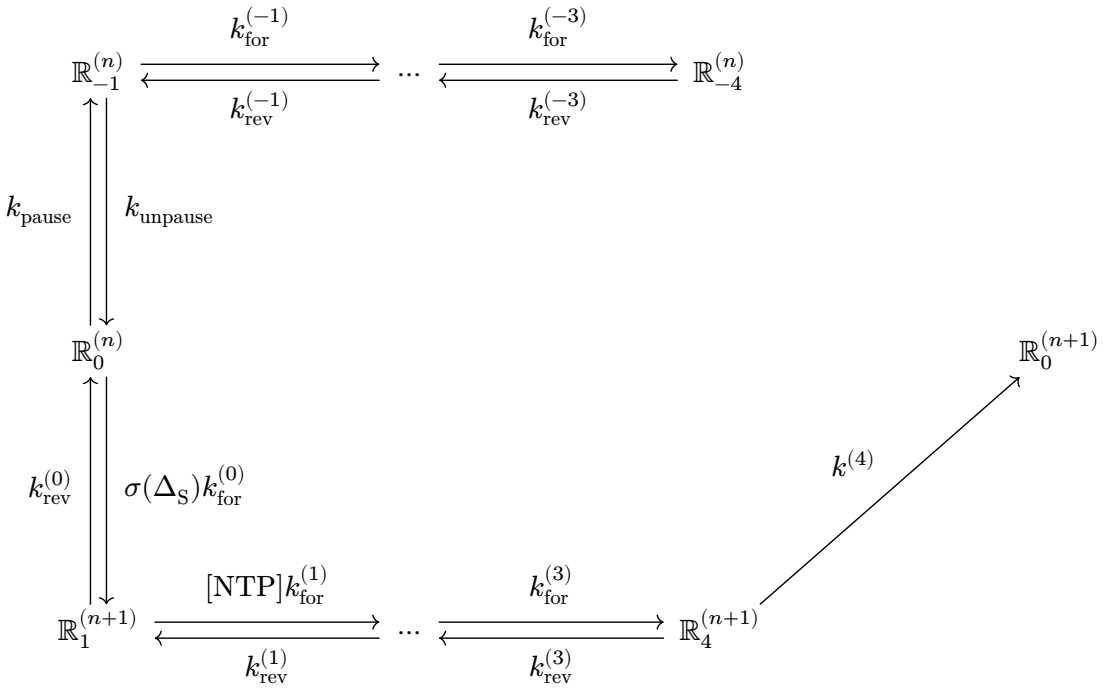


Figure 3: Depiction of transcription elongation

"on-pathway" states and "offline" paused states, with transitions between these states occurring at intrinsic rates modified by NTP concentration if relevant(Fig. 3).

The status of a given RNAP $\mathbb{R}_x^{(n)}$ is encoded by its position n and its state i. Transitions between the "on-pathway" states $\mathbb{R}_0^{(n)}, \mathbb{R}_1^{(n+1)}, ..., \mathbb{R}_4^{(n+1)}$ correspond to the sequential steps of elongation: translocation, NTP binding, NTP alignment, catalysis, and PPi release[5]. With the exception of PPi release($\mathbb{R}_4^{(n+1)} \to \mathbb{R}_0^{(n+1)}$), which we model as irreversible, all transitions within the "on-pathway" are reversible.

The "offline" states $\mathbb{R}_{-1}^{(n)}, \mathbb{R}_{-2}^{(n)}, ..., \mathbb{R}_{-4}^{(n)}$ represent various conformations of the paused elongation complex [5], and transitions between these states are all reversible. While the paused complex can exhibit behaviors such as backtracking, these steps are known to be significantly slower, particularly beyond $\mathbb{R}_{-4}^{(n)}$. Given that transcriptional pausing primarily impacts elongation by diverting RNAP from the "on-pathway" conformations to these paused states, we have focused our modeling efforts on this primary mechanism of pausing and have not explicitly included the extremely slow backtracking transitions within the "offline" states.

The "on-pathway" and "offline" states are interconnected only at a branching point between $\mathbb{R}_0^{(n)}$ and $\mathbb{R}_{-1}^{(n)}$. Transitions at this point represent the entry into and exit from transcriptional pausing.

Given RNAP movement is constrained by torsional strain[6], we model the effect of this strain Δ_s by increasing downstream supercoiling(s_+) and decreasing upstream supercoiling(s_-) upon successful translocation. The translocation rate of a polymerase is then scaled by a logistic function dependent on the difference between downstream and upstream supercoiling $\Delta_s = s_+ - s_-$, and reduces the rate drastically at a critical point set to 100(see Appendix B). Topoisomerase activity acts to relieve this torsional strain as it does within cells and is modelled as a random binding event along the gene occuring at a given rate. Once successfully bound, the closest polymerases within a neighbourhood of 50nt have their corresponding number of supercoils set to 0.

Transcription Termination

Upon reaching the end of the gene, a polymerase agent is removed and a global counter for the number of RNA transcripts N_{RNA} is incremented. Transcripts are then subject to a first-order stochastic degradation process with a defined rate.

Stochastic Simulation of the Stationary Distribution

To efficiently explore the parameter space and generate sufficient samples from the stationary distribution for each parameter set, we implemented the model in the `Rust` programming language. While the initial model

development was conducted in `NetLogo`, the computational intensity of long-duration simulations across many parameter combinations made it unsuitable for this study, requiring an implementation in `Rust`.

Model parameters describing transcription initiation (binding/unbinding rates and concentrations of TFs and RNAP), elongation (translocation, pausing, NTP binding, catalysis rates), torsional strain effects (topoisomerase activity), and RNA degradation were randomized for each independent simulation run. Each simulation was then run for sufficient time to reach the stationary distribution and sampled infrequently to minimize the effect of autocorrelation. The collection of observations for each parameter set then formed the basis for our downstream analysis.

Evaluation of the Negative Binomial Fit

Let $\mathcal{X}_i = \left\{x_1^{(i)}, ..., x_n^{(i)}\right\}$ denote the sample of the stationary distribution for the i-th simulation. The negative binomial approximation is fit to the data in `R` using the `optim` function to maximize the log-likelihood:

$$\ell(\mathcal{X}_i; \mu_i, \phi_i) = \sum_{j=1}^{n} \ln\Big(P\Big(x_j^{(i)} \mid \mu_i, \phi_i\Big)\Big)$$

See Appendix C for parameterization of P used. Using the estimated parameters, the theoretical Fano factor is:

$$F_i = \frac{\mu_i + \frac{\mu_i^2}{\phi_i}}{\mu_i}$$

Which is compared to the empirical Fano factor of the collected data:

$$\hat{F}_i = \frac{\mathrm{var}(\mathcal{X}_i)}{\mathrm{mean}(\mathcal{X}_i)}$$

Additionally, the bimodality of $\mathcal{X}_i$ is assessed by computing the bimodality coefficient of the data:

$$b_i \approx \frac{\gamma_i^2 + 1}{\kappa_i}$$

Where γ_i is the skewness, κ_i is the kurtosis, and a bimodality coefficient above $\frac{5}{9} \approx 0.55$ is generally seen as evidence for bimodality, as the uniform has a bimodality coefficient of $\frac{5}{9}$.

Results

Goodness-of-Fit of the Negative Binomial

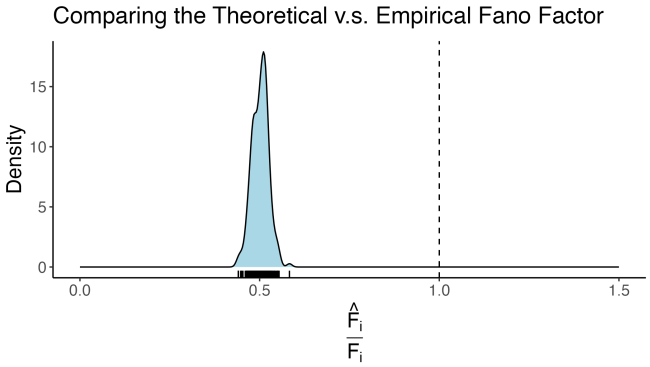


Figure 4: Comparison of empirical and theoretical Fano factors. Ticks to the left of the vertical dashed line indicate underdispersion of the simulated data.

Across all simulated parameter combinations, we observed a consistent pattern of underdispersion in the true transcript count distributions relative to the best-fitting negative binomial approximation. The empirical Fano factor calculated from the simulated data was lower than the theoretical Fano factor derived from the estimated negative binomial parameters in all simulations.

Assessing Bimodality

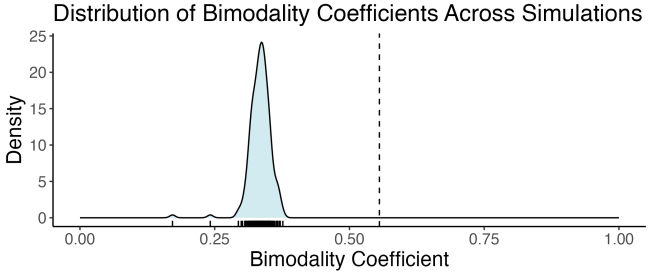


Figure 5: Distribution of bimodality coefficients. Ticks to the left of the vertical dashed line indicate unimodal distributions.

Analysis of the bimodality coefficient for the stationary transcript count distributions across all simulations yielded no values exceeding the threshold for bimodality (approximately 0.55). This indicates that within the parameter space explored, our mechanistic model of transcription did not produce any bimodal transcript count distributions.

Discussion

The underdispersion observed in our simulations contrasts the overdispersion commonly reported in experimental transcriptome data. While the negative

binomial model is often invoked to account for this overdispersion, our results suggest that a hierarchical model consisting of an inherently underdispersed hidden count distribution— obscured by additional layers of measurement error or stochastic fluctuations —may also sufficiently describe transcriptome data.

However, the limited complexity of our model likely contributes to the reduced variance in transcript counts we observed. In particular, the assumption of constant concentrations for all molecular species involved in transcription represents a significant simplification of cellular reality. *In vivo*, these concentrations are subject to their own dynamic processes of production and degradation, which could introduce substantial additional variability in transcript counts, potentially leading to overdispersion. For instance, our model maintains a static concentration of free nucleotide triphosphates (NTPs). Although we initially reasoned that the cellular pool of NTPs is generally large enough to remain relatively stable during the transcription of a single gene, neglecting the potential for these fluctuations could impact the overall dispersion of transcript counts.

Future directions can investigate if modelling the dynamics of NTPs concentration in the cell would result in greater dispersion. This is plausible as decreasing NTP concentrations also decrease the rate of elongation— and thus transcription —which would slow the production of RNA. More variations in the rate of transcription during the simulation would then result in greater stationary count dispersion.

Another limitation of our program is that we only modelled one gene. This gene doesn't contribute to the rate of its own transcription, but cells often have multiple interdependent genes which encode elements that influence transcription. If we were to stochastically model multiple genes, where each gene encodes one of the transcription factors, we may capture overdispersion and potentially bimodality.

The absence of bimodality in our simulations do, however, suggest that the transcriptional machinery modelled is insufficient to generate the observations of bimodality in experimental data. Although presence of multiple interacting genes could produce complex dynamics capable of yielding a bimodal count distribution, the minimal components of transcription are unlikely to produce this effect on their own. Thus, we provide some evidence supporting the presence of multiple cell states to generate bimodal distributions of transcript counts.

Conclusion

In this work, we developed a stochastic, agent-based model of transcription based on mechanistic first principles to investigate the statistical properties of the transcript count distribution found within cells. Our simulations, conducted across a broad range of parameter values, found that simulated transcript counts consistently exhibited underdispersion relative to the negative binomial distribution and did not yield any bimodality. This suggests that the core transcriptional machinery in isolation may not inherently produce the overdispersion observed experimentally, highlighting the potential importance of other cellular factors or measurement noise in the current statistical methods used for transcriptome analysis. This finding also lends support to the hypothesis that bimodal expression patterns in seemingly homogenous cell populations may arise from underlying cellular heterogeneity or more complex regulatory mechanisms not captured in our mechanistic understanding of transcription. While our work provides insight into the baseline statistical behavior of a minimal transcription system, future research incorporating more intricate biological processes, such as dynamic molecular concentrations and multi-gene interactions, could provide a more comprehensive understanding of transcript count variability.

Code Availability

The code used in this study is publicly available on GitHub at https://github.com/toddmccready/final-project.

Appendix

Appendix A: Multiple Stochastic Steps Can Have Less Variability

One of the questions asked during our presentation was how can multiple stochastic steps end up with less variability relative to a single step. Here is an example of how this can occur:

Let $X \sim \text{Exponential}(\lambda)$ denote an exponential random variable that represents the time until some event from some main process. Furthermore, suppose we additionally have $X_, ..., X_n$ independent exponential random variables with each governed by their own

rates $\lambda_1, ..., \lambda_n$ where the sum $X_1 + \cdots X_n$ represents the time until that same event occurs through a separate process.

Suppose that the mean time for both processes is the same, in other words:

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \cdots X_n]$$

This implies:

$$\frac{1}{\lambda} = \frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_n}$$

Since all rates are non-negative, then we have $\lambda_i > \lambda$ for all $i \in \{1, ..., n\}$(otherwise we force other terms to be negative, which is a contradiction).

Therefore, we have that:

$$\begin{aligned}
\mathrm{Var}(X) &= \left(\frac{1}{\lambda}\right)^2 \\
&= \left(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_n}\right)^2 \\
&= \left(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_n}\right)\left(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_n}\right) \\
&= \frac{1}{\lambda_1^2} + \cdots + \frac{1}{\lambda_n^2} + C
\end{aligned}$$

Where $C > 0$ absorbs all other terms.

Note however, that the variance of the separate process is thus smaller:

$$\begin{aligned}
\mathrm{Var}(X_1 + \cdots + X_n) &= \frac{1}{\lambda_1^2} + \cdots + \frac{1}{\lambda_n^2} \\
&< \frac{1}{\lambda_1^2} + \cdots + \frac{1}{\lambda_n^2} + C \\
&= \mathrm{Var}(X)
\end{aligned}$$

Therefore, we have an example where two processes have the same average time till some event, but one has significantly less variance. The same logic can be applied to the situation of observing underdispersion in the stationary distribution of our model.

However, when including factors such as reversible kinetics, transcriptional pausing, and torsional strain effects, we no longer get the guarentee that counts will be underdispersed(hence why we wanted to study this).

Appendix B: Form of Torsional Strain Effect

Let Δ_s denote the difference in supercoiling upstream and downstream a polymerase. Then the effect of torsional strain on the translocation rate is given by:

$$\sigma(\Delta_s) = \frac{1}{1 + \exp(\Delta_s - 100)}$$

The rate of the forward translocation is then $\sigma(\Delta_s) k_{\text{for}}^{(0)}$.

Appendix C: Negative Binomial Parameterization

We parameterize the negative binomial in terms of its mean μ and inverse overdispersion factor ϕ. Let $X \sim \mathrm{NegBinom}(\mu, \phi)$, then:

$$\begin{aligned}
P_X(x \mid \mu, \phi) &= \binom{x + \phi - 1}{x} \left(\frac{\mu}{\mu + \phi}\right)^x \left(\frac{\phi}{\mu + \phi}\right)^\phi \\
\mathbb{E}[X] &= \mu \\
\mathrm{Var}(X) &= \mu + \frac{\mu^2}{\phi}
\end{aligned}$$

Bibliography

[1] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, p. R106, Oct. 2010, doi: 10.1186/gb-2010-11-10-r106.

[2] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010, doi: 10.1093/bioinformatics/btp616.

[3] Z. S. Singer *et al.*, "Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells," *Molecular Cell*, vol. 55, no. 2, pp. 319–331, Jul. 2014, doi: 10.1016/j.molcel.2014.06.029.

[4] I. Baek, "Single Molecule Studies of RNA Polymerase II Transcription Initiation and Elongation," 2021. Accessed: Apr. 08, 2025. [Online]. Available: https://www.proquest.com/docview/2564137545

[5] J. Saba *et al.*, "The elemental mechanism of transcriptional pausing," *eLife*, vol. 8, p. e40981, Jan. 2019, doi: 10.7554/eLife.40981.

[6] J. Ma, L. Bai, and M. D. Wang, "Transcription Under Torsion," *Science*, vol. 340, no. 6140, pp. 1580–1583, Jun. 2013, doi: 10.1126/science.1235441.