

IDENTIFYING DETAILS THAT MATTER:  
FRUIT FLY DEVELOPMENT, GENETIC  
REGULATION, AND MICROBIAL ECOLOGY

MIKHAIL TIKHONOV

A DISSERTATION  
PRESENTED TO THE FACULTY  
OF PRINCETON UNIVERSITY  
IN CANDIDACY FOR THE DEGREE  
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE  
BY THE DEPARTMENT OF PHYSICS

ADVISERS: PROFESSOR WILLIAM BIALEK  
PROFESSOR THOMAS GREGOR

SEPTEMBER 2014

© Copyright by Mikhail Tikhonov, 2014.

All rights reserved.

# Abstract

The wealth and complexity of the known microscopic detail of biological processes and pathways make the search for universality particularly challenging and appealing for a physicist. This dissertation investigates several examples drawn from three different biological contexts.

First, I discuss the gene regulatory network responsible for segment patterning in the fruit fly. The fruit fly embryo is one of the best-studied examples of precision in biological processes. However, a novel technique I developed with my collaborators demonstrates that even in this system transcription is intrinsically noisy, as previously observed in bacteria. Using single-molecule-precision measurements of the transcriptional activity of four critical patterning genes, we exhibit universality of expression noise parameters and show how precision is recovered through spatiotemporal averaging. On a theoretical level, I demonstrate how these experimental findings help understand the multi-tier architecture of the patterning network: the diffusion-mediated non-locality of transcriptional response makes a cascade of readouts the optimal gradient response strategy, even if each readout is intrinsically noisy.

Second, I investigate the importance of microscopic parameters of networks at the scale of their global function. The fields of neural and genetic networks have exactly opposite assumptions on the matter, the former concentrating on synapse strength and the latter solely on network topology. I present a class of simple perceptron-based Boolean models within which the relative importance of topology vs. interaction strengths becomes a well-posed problem. I show that optimizing interaction strengths is a better strategy of achieving high complexity, defined as the number of fixed points the network can accommodate, and comment on the implications for real networks and their evolution.

Third, I discuss the so-called 16S tag sequencing method of studying microbial communities. The standard approach to 16S data, which relies on clustering reads

by sequence similarity into Operational Taxonomic Units (OTUs), underexploits the accuracy of modern sequencing technology. I present a novel, clustering-free approach that exploits cross-sample comparisons to achieve sub-OTU resolution, and demonstrate that this new level of detail can provide new insight into factors shaping community assembly.

Finally, I discuss some common themes in the conclusions from these projects.



# Acknowledgements

I am indebted to William Bialek for introducing me to the very concept of theoretical biophysics; the class he taught in the spring of 2010 inspired me to leave the world of high-energy theory and venture into a wholly unfamiliar territory, a switch that I never regretted. It was eye-opening to discover that the world of living organisms harbors so many fundamental questions that physicists could help address—questions, in fact, that have direct connections with experiments and could, quite plausibly, have implications for our everyday lives. For someone with a heavily theoretical background like myself, this was a new and enticing thought. It seemed natural, therefore, to begin my journey in an experimental group. I thank Thomas Gregor for teaching me what it means to do careful experiments on a biological system. I have been very lucky to be co-advised by Thomas and Bill: working with both of them taught me just how much the detailed understanding of an experiment shapes the theoretical interpretation of its results; this lesson changed the way I think about experimental data and will stay with me wherever I go from here.

I am very grateful to Ned Wingreen, who found the time to be available; the metagenomics project we tentatively started together evolved into a most rewarding collaboration. I am also thankful to Thomas and Bill for giving me the freedom to pursue this direction with Ned. I sincerely thank Curtis Callan and Eric Wieschaus for their valuable advice.

My friends and colleagues at the biophysics theory group made Icahn lab a welcoming and intellectually stimulating place, and I would like to thank them all for the great moments both in and out of the lab: Ji Hyun Bak, Gordon Berman, Farzan Beroz, Anne-Florence Bitbol, Chase Broedersz, Michele Castellana, Mark Ioffe, Dima Krotov, Ben Machta, Leenoy Meshulam, Armita Nourmohammad, Thibaud Taillefumier, David Schwab, and DJ Strouse. I also thank all the members of the Gregor lab, for everything they taught me and for all the fun times we spent together: Bryan

Chun, Julien Dubuis, Hernan Garcia, Sri Iyer Biswas, Albert Lin, Shawn Little, Feng Liu, Robert Malcolm, Mariela Petkova, Allyson Sgro, and Eric Smith. I thank Lee Morgan and Jessica Heslin for their administrative help.

On a separate note, I would like to thank all my friends at L'Avant-Scène and particularly Florent Masse. Princeton would not have been the same without you.  
*C'était une belle traversée !*

During my time at Princeton, I was fortunate to meet many wonderful people. Far too numerous to be named here, you all left a mark on who I am; for that, I am grateful. And finally, I am thankful that friendship can defy distance. Bruno and Charlotte, I think of you.

## Publications and preprints associated with this dissertation

1. Little SC\*, Tikhonov M\*, Gregor T. (2013) “Precise developmental gene expression arises from globally stochastic transcriptional activity.” *Cell* **154**, 789-800. (\*equal contribution)
2. Garcia HG, Tikhonov M, Lin A and Gregor T. (2013) “Quantitative imaging of transcription in living *Drosophila* embryos links polymerase activity to patterning.” *Curr Biol* **23**, 1-6.
3. Tikhonov M, Leach RW, Wingreen NS. (2014) “Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution.” *ISME J* (in press).
4. Tikhonov M & Bialek W, “Complexity in generic biochemical circuits: topology versus strength of interactions” (in review; arXiv:1308.0317 [q-bio.MN]).

I am grateful to my collaborators Hernan Garcia, Robert Leach, Albert Lin and Shawn Little, and to the many people who provided feedback on these projects, most notably Robert Edgar, Daniel Fisher, Rob Knight, Simon Levin, Stephen Pacala, Sarah Preheim, Michael Rosen, Gertrud Schupbach, and Gasper Tkačik.

Materials from this dissertation have been publicly presented at the following conferences:

1. International Conference on Biological Physics 2011, San Diego, CA.
2. BPS Annual Meeting 2013, Philadelphia, PA.
3. APS March Meeting 2013, Baltimore, MD.

Finally, I would like to acknowledge funding support through NSF Physics of Living Systems program, NSF Center for the Science of Information, and DARPA Biochronicity program.

To my parents.

# Contents

Abstract . . . . .	iii
Acknowledgements . . . . .	v
List of Tables . . . . .	xiv
List of Figures . . . . .	xv
<b>1 Introduction</b>	<b>1</b>
<b>2 Segment patterning in fruit fly: Experiments</b>	<b>3</b>
2.1 Introduction . . . . .	6
2.2 Measuring absolute numbers of mRNA transcripts . . . . .	8
2.3 Precision of cytoplasmic <i>hb</i> mRNA and protein distributions . . . . .	12
2.4 Determining instantaneous transcriptional activity . . . . .	14
2.5 Variation in nascent transcription site activity <i>vs.</i> cytoplasmic output	17
2.6 Gap genes share expression characteristics . . . . .	20
2.7 Discussion . . . . .	23
<b>Technical details</b>	<b>29</b>
2.A Experimental procedures . . . . .	29
2.B Detection of individual <i>hb</i> transcripts: the workflow . . . . .	29
2.B.1 Pre-processing of image stacks . . . . .	31
2.B.2 Identifying fluorescent particles . . . . .	32
2.B.3 Defining summation volumes . . . . .	33

2.B.4	Calibrating total fluorescence . . . . .	35
2.B.5	Extracting features of cytoplasmic profiles . . . . .	37
2.C	Detection of individual <i>hb</i> transcripts: FishToolbox v2.0.12 . . . . .	38
2.D	Supplementary figures . . . . .	39
	Figure 2.S1: Detection of individual <i>hb</i> transcripts . . . . .	39
	Figure 2.S2: Particle counts and total fluorescence . . . . .	42
	Figure 2.S3: Comparison with Hb protein, and <i>hb</i> mRNA lifetime . . . . .	44
	Figure 2.S4: Detection and interpretation of transcription dynamics . . . . .	45
	Figure 2.S5: Maximum gene expression rates . . . . .	47
	Figure 2.S6: Spatial and temporal averaging . . . . .	48
2.E	Control experiment demonstrating single-molecule resolution . . . . .	49
2.F	Noise level predicted by the two-state model of transcription . . . . .	51
2.G	Transcriptional activity of loci on sister chromatids . . . . .	52
2.H	Efficiency of temporal and spatial averaging . . . . .	53
<b>3</b>	<b>Segment patterning in fruit fly: Theory</b>	<b>55</b>
3.1	Introduction: Local channel picture is insufficient . . . . .	57
3.2	Transcriptional readout in presence of averaging . . . . .	60
3.2.1	Model for readout . . . . .	60
3.2.2	The linear approximation . . . . .	62
3.3	Optimal amplification . . . . .	65
3.3.1	The benefits of amplification . . . . .	65
3.3.2	Patterning capacity . . . . .	69
3.3.3	Optimal amplification . . . . .	70
3.3.4	Reexamining the definition of patterning capacity . . . . .	72
3.4	Conclusion . . . . .	72

<b>Technical details</b>	<b>74</b>
3.A Commutation relations . . . . .	74
<b>4 Genetic networks: going beyond topology</b>	<b>76</b>
4.1 Introduction . . . . .	77
4.2 The model . . . . .	80
4.2.1 Definitions . . . . .	80
4.2.2 Parameter space geometry . . . . .	82
4.2.3 Weighting sectors: a formal definition . . . . .	84
4.3 Computing complexity . . . . .	87
4.4 Discussion . . . . .	90
<b>Technical details</b>	<b>93</b>
4.A Weighting sectors for topological in-degree $K = 3$ . . . . .	93
4.B Computational details . . . . .	95
4.B.1 Targeted search for high-complexity weightings . . . . .	97
4.B.2 Computing the mean complexity of a topology . . . . .	97
4.B.3 Targeted search for high-complexity topologies . . . . .	100
4.C TSP complexity . . . . .	102
<b>5 Microbial communities</b>	<b>106</b>
5.1 Introduction . . . . .	108
5.2 Cluster-free filtering . . . . .	111
5.2.1 Data selection and quality filtering . . . . .	111
5.2.2 Cluster-free filtering . . . . .	113
5.2.3 Cluster-free filtering — the denoiser . . . . .	115
5.3 Sequence similarity <i>vs.</i> ecological similarity . . . . .	117
5.4 Distinct subpopulations with high dynamical similarity . . . . .	121
5.5 Clustering vastly underestimates ecological richness . . . . .	126

5.6	Sequence identity <i>vs.</i> sequence similarity . . . . .	128
5.7	Discussion . . . . .	131
<b>Technical details</b>		<b>134</b>
5.A	Supplementary methods. . . . .	134
5.A.1	Motivation: sequencing noise is low . . . . .	134
5.A.2	Estimating rates of one-nucleotide substitutions . . . . .	136
5.A.3	The algorithm for filtering substitution errors . . . . .	141
5.A.4	Other error types, including chimeras and PCR indels . . . . .	143
5.A.5	Cluster-free filtering software package . . . . .	144
5.A.6	Mock community validation and comparison with DADA . . . . .	146
5.A.7	Runtime comparison with DADA . . . . .	149
5.A.8	Other applications: environmental cross-sectional 454 data . . . . .	150
5.A.9	How many samples is enough? . . . . .	152
5.B	Supplementary information for Figure 5.2 . . . . .	154
5.B.1	Anticorrelated subpopulations example . . . . .	154
5.B.2	Best expected correlation of two time traces . . . . .	154
5.B.3	Distance metric for sequence pairs . . . . .	157
5.C	Supplementary information for Figure 5.3 . . . . .	157
5.C.1	Estimating correlation time from autocorrelation function . . . . .	157
5.C.2	Dynamics on long time scales . . . . .	158
5.C.3	Persistence of difference: the null model . . . . .	159
5.C.4	Persistence of difference for non-longitudinal data . . . . .	160
5.D	Supplementary information for Figure 5.4 . . . . .	161
5.D.1	Over-estimation of OTU quality scores . . . . .	161
5.E	Supplementary information for Figure 5.5 . . . . .	162
5.E.1	Cross-individual analysis of fecal samples . . . . .	162
5.E.2	Cross-individual analysis at 97% OTU level . . . . .	164



<b>6 Conclusion</b>	<b>165</b>
<b>Bibliography</b>	<b>169</b>

# List of Tables

4.S1	Weighting sectors for in-degree $K = 3$ . . . . .	93
4.S2	Link states and compatible weighting sectors . . . . .	96
5.S1	Substitution error rates as measured from the data . . . . .	139
5.S2	Accuracy comparison with DADA on mock community data . . . . .	147
5.S3	Runtime comparison with DADA . . . . .	150

# List of Figures

2.1	A schematic of pattern refinement during development. . . . .	4
2.2	Counting of absolute transcript number in <i>Drosophila</i> embryos . . .	9
2.3	Precision and reproducibility of cytoplasmic <i>hb</i> profiles . . . . .	11
2.4	Variability of transcriptional activity at nascent transcription sites .	15
2.5	Fluctuations in <i>hb</i> transcription are dominated by intrinsic noise . .	19
2.6	Universal properties of transcripts of all gap genes . . . . .	21
2.S1	Detection of individual <i>hb</i> transcripts . . . . .	41
2.S2	Particle counts and total fluorescence . . . . .	41
2.S3	Comparison with Hb protein, and <i>hb</i> mRNA lifetime . . . . .	43
2.S4	Detection and interpretation of transcription dynamics . . . . .	45
2.S5	Maximum gene expression rates . . . . .	48
2.S6	Spatial and temporal averaging . . . . .	49
3.1	The readout operator. . . . .	61
3.2	The model of the readout . . . . .	61
3.3	The linear approximation of input/output function . . . . .	62
3.4	Non-local readout as a local measurement of a non-local quantity . .	63
3.5	Hb boundary formed over multiple nuclear cycles . . . . .	64
3.6	Canonical form of an arbitrary circuit . . . . .	64
3.7	Using morphogen <i>c</i> to discriminate between neighboring nuclei . . .	66
3.8	A linear amplifying readout with dynamic range restriction . . . . .	69

3.S1	Commutation relations . . . . .	75
4.1	Explanation of notations . . . . .	81
4.2	Weighting sectors: the simplest case . . . . .	83
4.3	The average complexity does not grow with $N$ . . . . .	88
4.4	Complexity distributions for different topologies overlap considerably	89
4.S1	Distribution of average complexity of all $N = 6$ topologies . . . . .	94
4.S2	Shortcomings of capacity as a measure of complexity . . . . .	101
4.S3	TSP complexity <i>vs.</i> capacity . . . . .	103
5.1	Clustering underexploits the quality of modern sequence data . . . .	109
5.2	Sequence similarity need not imply dynamical similarity, and vice versa	119
5.3	Dynamical similarity versus 16S similarity . . . . .	123
5.4	Clustering reads into OTUs underestimates dynamical diversity . . .	127
5.5	Cohabiting individuals share bacterial subpopulations . . . . .	130
5.S1	Rank distribution of the top 5 sequences . . . . .	136
5.S2	The complete error cloud of Seq. #1 . . . . .	137
5.S3	Reproducibility of inferred substitution error rates . . . . .	138
5.S4	Dependence of the inferred error rates on quality filtering parameters	140
5.S5	A more detailed version of the error cloud cartoon . . . . .	142
5.S6	The workflow of cluster-free filtering software package . . . . .	145
5.S7	Cross-sectional environmental 454 data . . . . .	151
5.S8	A pair of sequences representing anticorrelated subpopulations . . .	154
5.S9	Best expected correlation $c_{\max}$ . . . . .	156
5.S10	Dynamics on long time scales . . . . .	158
5.S11	Cross-individual analysis of fecal samples . . . . .	163
5.S12	Dynamical similarity between pairs of common 97% OTUs . . . . .	164

# Chapter 1

## Introduction

Why can we send a robot to Mars but not predict when antibiotic-resistant bacteria will evolve? How is it that the nucleus of an atom and the center of the Sun, though infinitely harder to access experimentally, are better understood than an amoeba? Physics as a search for fundamental laws of nature has been remarkably successful in understanding inanimate matter at scales ranging from subatomic to cosmological. There is, however, an entire world that we have barely begun to understand — the world of living organisms. Why is it proving so challenging?

A thread running through all of physics is the idea that physical phenomena occurring at a particular scale of time, space or energy must be described in terms of degrees of freedom that belong to that same scale. But for a living system, whether it is a living cell, an organism or a community, organization and structure span multiple scales. It is unclear how to integrate all microscopic detail into effective degrees of freedom when changing a few nucleotides in a single DNA molecule in an embryo can change the adult body morphology, while changing countless others may have no apparent effect.

For physicists and biologists studying similar questions, the degree of attention to microscopic detail is perhaps one of the most important points of friction. Many

physicists believe that detailed modeling of biological complexity is an enterprise with too many parameters to have predictive power; in contrast, biologists often perceive the physicists' approach as too reductionist and simplistic: physicists tend to dismiss molecular details, and in biology, details matter.

However, *which* details matter depends, of course, on the questions being asked. It appears that some features, e.g. the fundamentally multi-scale nature described above, are shared across living systems. It is an intriguing possibility that these shared features could be quantitatively understood within some common framework that we are currently missing. My motivation and long-term goal is to use the interface of physics and biology to begin constructing such a framework.

In the meantime, searching for commonalities requires studying a diverse set of cases: we should not be drawing our examples from what already looks like the same box. The unique environment of the Biophysics Theory group at Princeton made it possible for me to work on several projects from different fields, each presented in a separate chapter of this Dissertation. The final chapter identifies some common themes in the conclusions from individual projects.

## Chapter 2

# Segment patterning in fruit fly: Experiments

One of the best-studied examples of precision and reproducibility in biological processes is the early embryonic development of the fruit fly, specifically the patterning along the major body axis. The mother fly breaks the symmetry of the egg by supplying a few proteins that form “maternal gradients” within the embryo. Proteins serve as transcription factors, regulating how strongly various genes are “expressed”, i.e. read out and made into other proteins. Thus maternal gradients serve as input to a “genetic network” that patterns the embryo by laying out the “blueprint” of the future adult body barely 3.5 hours after fertilization, and with remarkable precision. The patterning network is multi-tier, each next level reading out the previous one to form a finer pattern (Fig. 2.1). The input is a gradient spanning the entire embryo; at the next tier we have proteins whose concentration profiles delimit broad domains. Genes located further downstream have expression patterns forming 7 stripes, precursors of the future body segments. At the final tier of the network, this pattern is subdivided into 14 stripes, at which point the combination of genes expressed in each nucleus reliably distinguishes it from its more anterior or more posterior neighbor.

Theoretical arguments have suggested that the precision and reproducibility of protein profiles in this system are so high they may be close to the physical limits imposed by the stochasticity of single-molecule chemistry (Tkacik et al., 2008). (Here, precision and reproducibility are defined as the variation in protein concentration between nuclei at similar locations within one embryo and across embryos, respectively.) This striking precision is in contrast with the general expectation that biological processes are “noisy”, and motivated a number of experiments whereby the profiles of patterning proteins were measured as carefully as possible (Gregor et al., 2007; Dubuis et al., 2013), probing the origin of such precision. However, proteins do not set the concentration of other proteins directly: at a microscopic level, the concentration of protein A will regulate how strongly gene B is “transcribed”, i.e. made into RNA copies; these are then “translated” into protein. In a collaboration with Shawn Little, we have developed a new technique allowing us to gain access to all the intermediate steps in this cascade, and measure, with single-molecule precision and in absolute units, the transcriptional activity of individual nuclei and the accumulation of RNA transcripts of four critical patterning genes. This enabled us to track noise through the entire process, and we found that transcribing loci exhibit an intrinsic noise of  $\approx 45\%$ , independent of specific promoter-enhancer architecture. However, this noise is quickly filtered out through simple physical mechanisms, namely temporal and

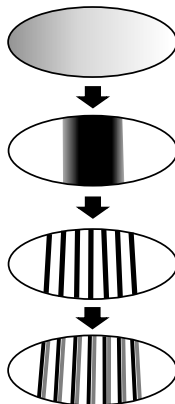


Figure 2.1: A schematic of pattern refinement during development.



spatial averaging, without involving any regulatory feedback, and the accumulation of mRNAs fluctuates by only  $\approx 8\%$  between neighboring nuclei, generating precise protein distributions.

Transcriptional noise is well-documented in bacteria and mammalian cell culture, but simply observing large noise does not mean it cannot be eliminated when necessary: a bacterium engineered to produce a fluorescent protein for our convenience has no reason to “care” about precision, and one can always argue that perhaps transcriptional readout could be made much more precise in the right circumstances. In contrast, our experiments, for the first time, probed an endogenous (naturally occurring) gene in a higher animal, in a situation where precision of expression has direct consequences for the cell-fate decision. The universality of noise parameters observed across the genes we studied further suggests that fluctuations in mRNA production are context-independent and are a fundamental property of transcription.

The work presented in this chapter is the subject of a paper on which Shawn and I are co-first authors: Little SC, Tikhonov M, Gregor T. “Precise developmental gene expression arises from globally stochastic transcriptional activity.” *Cell* 154 (2013), 789-800. In addition, the analysis code I designed was used in: Garcia, Tikhonov, Lin and Gregor (2013).

## 2.1 Introduction

A fundamental question in biology concerns the degree of precision that cellular systems exhibit in their responses to a given set of environmental conditions, extracellular signals, or other input stimuli (Lagha et al., 2012; Lander, 2013; Little and Wieschaus, 2011). Achieving physiologically or environmentally appropriate cellular behavior constrains the magnitude of molecular fluctuations (Rao et al., 2002). But the production and interaction of molecules are intrinsically stochastic, limiting the ability of cells to control gene expression and biochemical activities (Raser and O’Shea, 2005). In most contexts, it is unknown how closely cellular activity and phenotypic behavior rely on precise control of gene expression.

Many features of *Drosophila* embryogenesis suggest that strict control of gene expression determines reproducible and precise cell fate establishment. In *Drosophila* embryos, patterned gene expression in the early syncytium of  $\approx 6000$  nuclei is triggered by asymmetrically distributed, maternally supplied cues (Sauer et al., 1996). Among these is the transcription factor Bicoid (Bcd), the anterior-posterior (AP) concentration gradient of which shows remarkably reproducible distributions between embryos (Gregor et al., 2007). Moreover, within an embryo, the nuclei at similar AP coordinates differ in Bcd concentration by less than 10% (standard deviation over mean), a degree of precision sufficiently high for each row of cells along the AP axis to discern its position from its immediate neighbors (Gregor et al., 2007). These observations suggest stringent control mechanisms over Bcd expression. Bcd precision correlates with highly precise protein distribution of zygotically expressed target genes (Dubuis et al., 2013; Gregor et al., 2007). Early gene expression events confer cells with distinct gene expression programs necessary to determine individual fates within under three hours following fertilization (Gergen et al., 1986; Kornberg and Tabata, 1993).

These observations suggest a model in which tightly regulated transcriptional inputs give rise to rapidly established, highly precise outputs. However, it is unclear

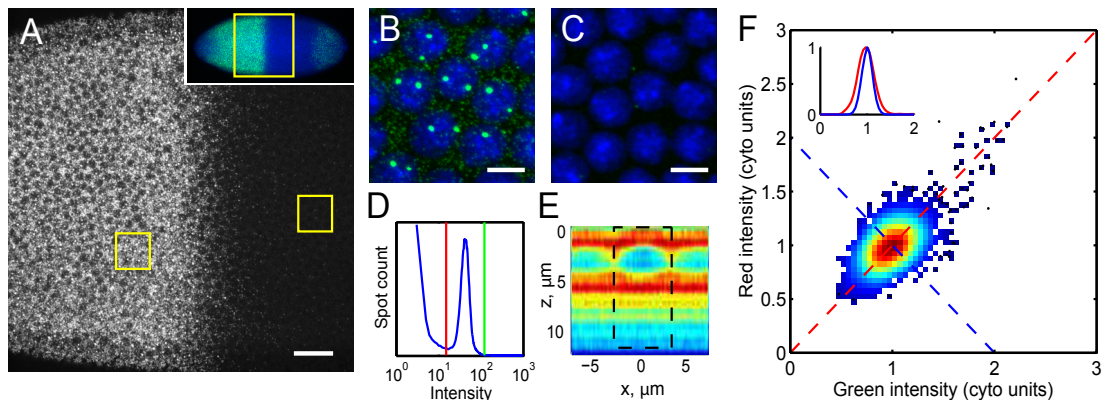
whether or how precise zygotic gene expression is achieved at the transcriptional level. In all contexts assayed from prokaryotes to mammalian cells, transcriptional activity between otherwise identical cells varies by at least  $\approx 50\%$ , and is typically much higher (Cohen et al., 2009; Gandhi et al., 2011; Golding et al., 2005; Pare et al., 2009; Taniguchi et al., 2010; Raj et al., 2006, 2010; Reiter et al., 2011; Sigal et al., 2006; Zenklusen et al., 2008). Quantitative observations support the idea that the process of transcription is intrinsically stochastic (Kaern et al., 2005; Li and Xie, 2011). In the context of development, it is unknown if the relatively small fluctuations in controlling inputs impact transcriptional output, and whether embryogenesis requires the activity of specialized filtering and/or feedback mechanisms to ensure fidelity in the rapid establishment of gene expression programs.

In a collaboration with Shawn Little, we have developed an enhanced method of fluorescence *in situ* hybridization (FISH) to label and detect individual zygotically expressed mRNA molecules in fixed tissue. A combination of a new experimental protocol built by Shawn and new data analysis tools that I created allowed us, for the first time, to measure in absolute molecular counts the magnitude and fluctuations in the earliest gene expression events of the *Drosophila* embryo. In the work presented here, we show that the earliest expressed genes share common expression characteristics: despite their expression in spatially distinct territories, their rates of production are identical, and all display intrinsically stochastic transcriptional activity. These similarities suggest that expression rate and variability result from fundamental, global features of transcriptional regulation that limit the attainable degree of precision. Nevertheless, the stochastic expression results in precise and nearly uniform transcript accumulation, achieved by straightforward spatiotemporal averaging, rather than complex regulatory feedback.

## 2.2 Measuring absolute numbers of mRNA transcripts in early *Drosophila* embryos

Previous work in *Drosophila* embryos has documented that nuclei positioned at the same AP coordinate express nearly the same protein amount of the gap gene Hunchback (Hb) with fluctuations of  $<10\%$  (Gregor et al., 2007). Given that the primary activating input to *hb* transcription is the transcription factor Bcd, whose variability between nuclei at a given position is of the same order as Hb (Gregor et al., 2007), the most straightforward explanation of minimal Hb variation downstream of precise Bcd activity is via a precise transcriptional response. To quantitatively evaluate transcription of *hb*, we built on a FISH method developed previously (Little et al., 2011) to label *hb* mRNAs using multiple fluorescently labeled antisense DNA oligonucleotides (Fig. 2.2A). By scanning confocal microscopy, we detect two broad classes of objects: first, sparse and brightly labeled spots corresponding to sites of nascent transcript production (e.g., Wilkie et al., 1999), and second, numerous relatively dim diffraction limited spots,  $\approx 90\%$  of which are located in the internuclear space (Fig. 2.2A-C). For clarity, I refer to this class of objects as cytoplasmic particles. These particles exhibit sufficiently high contrast to be readily distinguished from background imaging noise using automated image processing (Fig. 2.2D). Each particle is detected on at least three adjacent 250 nm separated confocal imaging sections (Fig. 2.S1A-B) with 3-dimensional structure identical to the measured point spread function (PSF; Fig. 2.S1C-D). To test detection efficiency, we applied probes with alternating fluorophore colors. A minimum of 85% of cytoplasmic particles detected in one channel is also found in the other, indicating that  $>94\%$  mRNA are detected in at least one of the channels (Fig. 2.S1E-G).

Intensity distributions of cytoplasmic particles are unimodal and tightly clustered around the mean (Fig. 2.2D), suggesting that the cytoplasmic particles are simi-



**Figure 2.2: Counting of absolute transcript number in *Drosophila* embryos.**

**A:** Confocal section through the nuclear layer of a WT embryo during interphase 13 labeled with 114 fluorescent oligonucleotide probes against *hunchback*, oriented anterior to the left. Scale bar: 25  $\mu\text{m}$ . Inset: Low magnification image identifying the region shown in A. **B,C:** Magnified views of anterior (B) and posterior (C) boxed regions in (A). Scale bars: 5  $\mu\text{m}$ . **D:** Intensity histogram of detected particles, showing thresholds for separation of transcripts from imaging noise (red line) and from the long tail of rare, bright spots corresponding to transcription sites (green line). **E:** *hb* transcript distribution in axial cross-section through a nucleus centered at  $x = 0$ .  $z = 0$  represents apical surface; spherical nuclei are compressed by approximately 40% in  $z$  direction. Color indicates mean particle density in relative units (red=high, blue=low). Dashed box: cylindrical volume of summation with  $z$ -depth of 13  $\mu\text{m}$ . **F:** Scatter plot of intensities as detected in two channels using 114 probes against *hb* of alternating colors. Color indicates relative density of data points. Inset: Cross-sections of the scatter plot in (F) along the correlated (red) and anti-correlated direction (blue) reveal Gaussian distributions with  $\sigma = 20\%$  (red) and  $\sigma = 12\%$  (blue). Values are normalized to the mean cytoplasmic particle intensity (1 “cyto unit”) in each channel. (See Sec. 2.B.4)

lar in mRNA content and consistent with previous work indicating that gap gene mRNAs neither package into multimers nor exhibit directional transport (Davis and Ish-Horowicz, 1991). Deviation from mean intensity results from at least two phenomena: particles can be bound by different probe numbers, and multiple particles can overlap and be detected as single spots. To determine the relative contributions of each, we examine the correlation of intensities for each particle in a 2-color detection experiment. Correlation between the two channels is weak (Fig. 2.2F), implying that the fractional standard deviation of mRNA content in detected particles is at most 16% (see Sec. 2.E). To determine the number of mRNAs per particle, we compared counts of maternally deposited *hb* mRNA particles from imaging entire

embryos at nuclear cycle (nc) 3-4 to those from quantitative RT-PCR in similarly aged embryos (Fig. 2.S1H). From these independent measures, we found an average of  $1.2 \pm 0.5$  mRNAs per imaged particle. Given a probability  $p$  of finding  $n$  mRNAs per particle distributed over a discrete set of integers, virtually all detected spots correspond to either 1 or 2 mRNA molecules ( $p(n) \approx 0$  for  $n \geq 3$ ); the constraint that  $\sigma[p(n)] \leq 0.16$  thus entails  $p(1) \geq 0.97$  (see Experimental Procedures). In addition, comparing counts between wild-type (WT) and *hb* hemizygous embryos yields a 2-fold concentration difference (Fig. 2.S1I). We conclude that 97% of detected particles are individual mRNA molecules.

As the density of zygotically produced *hb* mRNA increases above about 1 molecule per  $\mu\text{m}^3$ , the PSFs of individual molecules begin to overlap. To extend particle counting to arbitrarily high density, we take advantage of the naturally large dynamic range of expression. In high-density regions, we determine absolute counts by measuring total fluorescence collected from all mRNA per volume, and calibrating to low expressing regions where individual mRNA can be counted directly (Fig. 2.S2A). We can thereby measure absolute concentration and local fluctuation with accuracies of 12% and 5% respectively (Fig. 2.S2B-C). These two methods of measuring transcript concentration have overlapping domains of applicability: the first, direct counting, accurate for average transcript concentrations  $\leq 0.5$  molecules/ $\mu\text{m}^3$ ; the second, using total fluorescence, for concentrations  $\geq 0.35$  molecules/ $\mu\text{m}^3$  (Fig. 2.S2C). Thus our FISH method is suitable for high-precision measurements of absolute mRNA counts at any concentration.

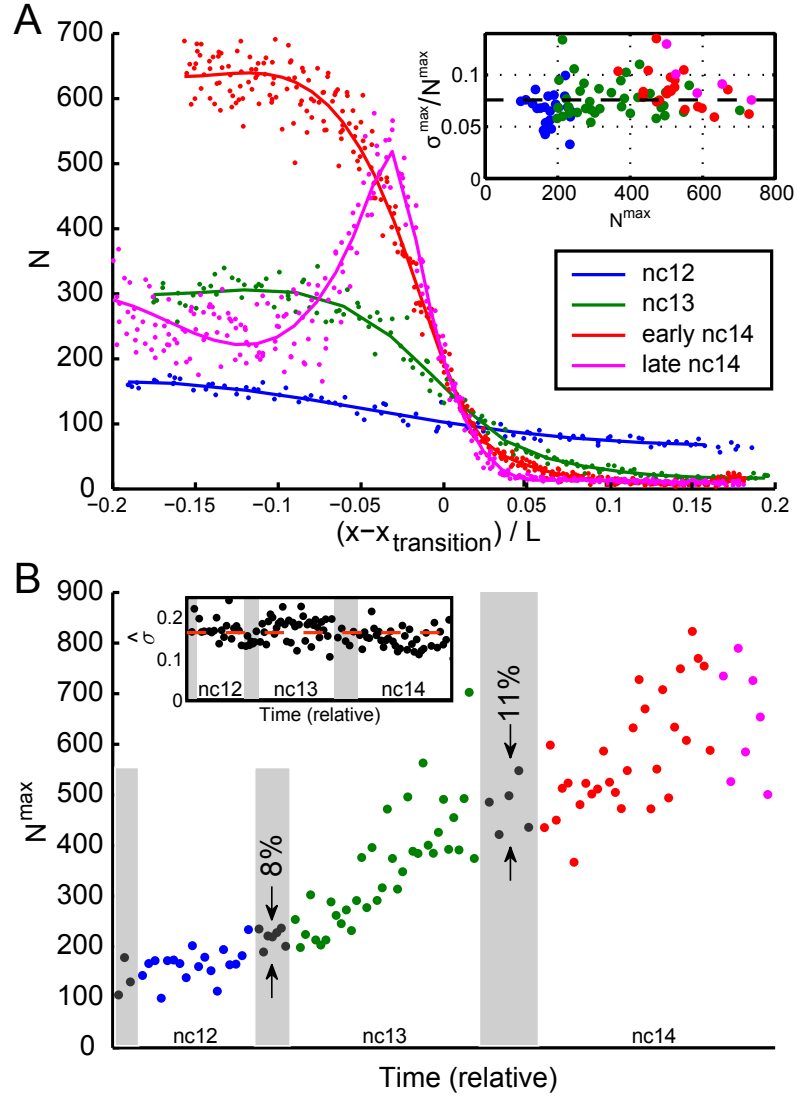


Figure 2.3: **Precision and reproducibility of cytoplasmic *hb* profiles.**

**A:** Absolute cytoplasmic *hb* mRNA counts per standardized volume as a function of AP position. Data for four embryos are depicted at nuclear cycle 12 (blue), 13 (green), early 14 (red), and late 14 (magenta). Position is shown as distance from the individual profiles' inflection points  $x_{\text{transition}}$  (see also Fig. 2.S3C). Inset: fractional standard deviation  $\sigma^{\text{max}} / N^{\text{max}}$  within the spatial domain of highest mRNA accumulation as a function of the mean count within that domain ( $N^{\text{max}}$ ) for 101 embryos between nc12 and nc14. Dashed line at 8%.

**B:** Cytoplasmic *hb* mRNA counts ( $N^{\text{max}}$ ) for 101 embryos as a function of time. Ages estimated by visual inspection of DAPI staining; relative width of mitoses (gray shading) and interphases according to Foe and Alberts (1983). Reproducibility of counts of embryos in 12th and 13th mitoses is 8% and 11%, respectively. Inset: estimated reproducibility  $\hat{\sigma}$  as a function of time. Data points: running averages of root-mean-square displacement from the smoothed timeline over 15 consecutive data points, normalized to mean. Dashed line: average  $\hat{\sigma}$  (17%).

## 2.3 Cytoplasmic *hb* mRNA and protein distributions display similar levels of precision

To assess fluctuations in transcript number between nuclei, we measure mRNA concentration in cylinders separated by one internuclear distance and with a depth of 12  $\mu\text{m}$  beneath the plasma membrane, where the majority of zygotically expressed transcripts are found (Fig. 2.2E, Fig. 2.S3A). As expected from prior observations (e.g., Tautz et al., 1987), *hb* transcripts accumulate dramatically in the embryo anterior during early blastoderm (Fig. 2.3A). Transcription is terminated early in nc14 except near the embryo midpoint, and maternally supplied transcripts are continuously lost from the posterior (Fig. 2.3A, Fig. 2.S3B-C). *hb* mRNA expression profiles correlate well with observed protein levels (Fig. 2.S3D).

As an initial quantification of precision, we ascertain the degree of variability that is independent of putative regulatory inputs. We focus specifically on the spatial domain of maximal transcript accumulation, i.e. on nuclei found in regions of the highest observed gene product levels. Here, expression noise (the fractional standard deviation of *hb* concentration) is  $8 \pm 2\%$  as early as nc12; thus *hb* mRNA levels exhibit equal or better precision than Hb protein (Gregor et al., 2007). In age-ordered embryos we observe a monotonically increasing trend in counts through mid-nc14 (Fig. 2.3B), with an approximately constant fractional standard deviation across embryos ( $17 \pm 3\%$ , Fig. 2.3B, inset). Ambiguous age determination in fixed samples results in large systematic fluctuations across embryos of approximately the same age. This effect is minimized in embryos undergoing mitosis, during which transcriptional activity is silenced (Shermoen and Ofarrell, 1991), and which can be unambiguously temporally ordered. In these embryos, the counts were reproducible at  $<11\%$ , similar to Hb protein profiles (Gregor et al., 2007). The actual precision and reproducibility are likely to be higher, since our measurements contain systematic errors arising



from the FISH procedure such as physical distortion (5% measurement error) and error in counts (2-3% measurement error; see Fig. 2.S2C). Importantly, the variation of cytoplasmic profiles is nearly at the level of Poisson counting noise, i.e. at the lowest bound that can be attained by a stochastic process. For  $N = 500$  molecules per volume, as observed in late nc13 or early nc14, counting noise amounts to 5%, matching the lower bound of our measurements (Fig. 2.3A, inset). Large mRNA counts provide a natural buffer against potential fluctuations in translation that have been observed in other systems (Bar-Even et al., 2006; Newman et al., 2006; Taniguchi et al., 2010), yielding precise Hb expression. By comparison, in genome-wide studies, the most highly (and therefore most precisely) expressed genes in yeast and *E. coli* exhibit cell-to-cell fluctuations exceeding 50% in mRNA count (Gandhi et al., 2011; Taniguchi et al., 2010). Thus, early embryos exhibit an extraordinary degree of precision, rarely observed in other contexts.

Our timeline suggests that *hb* transcript lifetime is large, as there is no apparent decrease in counts during the 12th and 13th mitoses (Fig. 2.3B). To verify this assessment, we measure *hb* mRNA lifetime directly by disrupting transcription with  $\alpha$ -amanitin injection and subsequently monitoring the loss of zygotic *hb* (Fig. 2.S3E-F). We find a lifetime of  $\approx 60$  minutes, consistent with an estimate from imaging (transcript loss of  $< 11\%$  (Fig. 2.3B) in 5 minutes of mitosis (Foe and Alberts, 1983) corresponds to a lifetime of  $> 45$ min). These results show that the accumulation of transcripts is only mildly impacted by degradation.

## 2.4 Determining instantaneous transcriptional activity by measuring total nuclear nascent mRNA content

The low noise of *hb* cytoplasmic mRNA counts suggests that all nuclei in the fully active region should produce transcripts at nearly equivalent rates. However, all systems studied to date, including *E. coli*, yeast, cultured cells, and late *Drosophila* embryos (Golding et al., 2005; Larson et al., 2011; Pare et al., 2009; Raj et al., 2006; Zenklusen et al., 2008), produce transcripts through brief intervals of dense output interspersed with long quiescent periods of stochastic duration (Li and Xie, 2011). Production of this type seems incompatible with the near uniformity of cytoplasmic *hb* mRNA content. To determine the extent of variability in *hb* transcriptional activity, we developed a novel measure of transcription using the fluorescence intensities of nascent transcription sites.

Consistent with previous results (Wilkie et al., 1999), we observe that the maximum number of detectable nascent sites per nucleus increases as a result of DNA replication during the course of interphase, from 2 sites in early interphase to 4 in mid to late (Fig. 2.S4A). Because sister chromatid loci remain in close physical proximity until mitosis, and because transcription sites may occasionally occupy overlapping focal volumes, the number of active loci is challenging to discern. Therefore, we used the total fluorescence of all transcription sites in a nucleus as an alternate measure of instantaneous transcriptional activity. Under the assumption that nascent and mature mRNAs are equally accessible to probes, nascent site intensities can be represented as an equivalent number of finished, mature mRNAs in the cytoplasm by normalizing to “unit” intensity, i.e. the mean intensity of a completed transcript. We thus measure transcriptional activity in absolute units of total mRNA content. To determine the extent of measurement error arising from differences in probe binding

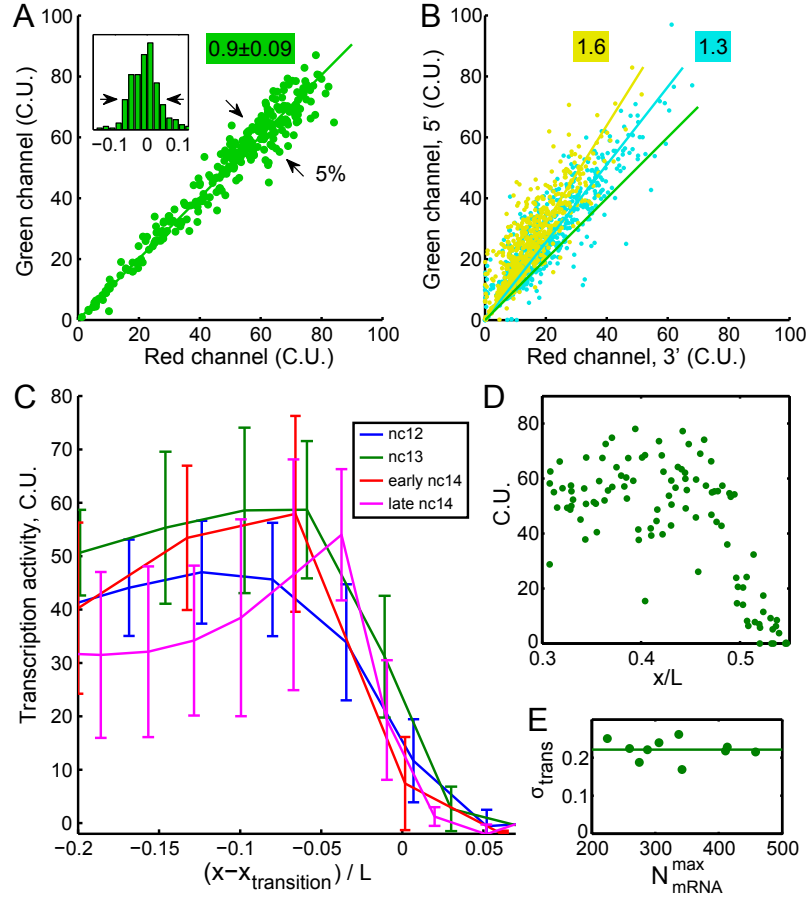


Figure 2.4: **Variability of transcriptional activity at nascent transcription sites.**

**A:** Scatter plot of total nascent *hb* mRNA per nucleus as measured using 114 probes with alternating colors for an embryo in nuclear cycle 13. Nascent mRNA content is reported after normalization to the mean cytoplasmic particle intensity (C.U.) in each channel. Transcription intensities tightly follow a direct proportionality relation with slope  $0.90 \pm 0.09$  ( $n = 5$  embryos). Inset: root mean square normalized deviation from the linear fit displays scatter of 5% (arrows) for the embryo shown in A. **B:** Two-color scatter plot of nascent mRNA content in which probes bearing the same fluorophore are clustered on the 5' (green channel) and 3' (red channel) portions of the transcript. Cyan: measurements using 57 green and 57 red-labeled probes; observed slope: 1.3. Yellow: results with 78 green and 36 red-labeled probes; observed slope: 1.6. Green line in A is plotted for comparison. **C:** Transcriptional activity per nucleus in cytoplasmic units as a function of position along the AP axis for four embryos: nuclear cycle 12 (blue), 13 (green), 14 early (red), and 14 late (magenta) in binned averages of 10, 20, 40 and 40 nuclei, respectively. Error bars are standard deviations in the respective bins. Position is shown as distance from the individual profiles' inflection points  $x_{\text{transition}}$  (see Experimental Procedures). **D:** Transcriptional activity per nucleus as a function of absolute AP position for the embryo in interphase 13 in C. **E:** Transcription noise for 10 embryos in nuclear cycle 13, plotted as fractional standard deviation across nuclei as a function of cytoplasmic *hb* counts within the spatial domain of highest accumulation. Transcription activity noise remains constant throughout interphase at  $22 \pm 3\%$ .

affinity and/or the subsequent normalization procedure, we use probes of alternating fluorophore colors. Ideally, for a given nascent site, the number of cytoplasmic units (C.U.) will be identical in both colors. Plotting nascent mRNA content of one color as a function of the other yields points on a line with a slope close to unity (between 5 embryos the mean slope ( $\pm$  SD) is  $0.90 \pm 0.09$ ), with a scatter of 5% (Fig. 2.4A). Therefore, we measure transcriptional activity with an error of 5%, and we can relate it to absolute mRNA content with an uncertainty under 20% (i.e. the largest deviation of  $0.90 \pm 0.09$  from 1).<sup>1</sup>

Three lines of evidence support the idea that nascent mRNA content reflects instantaneous transcriptional activity. First, the appearance of loci is coupled to the nuclear mitotic cycle: they are observed during interphase and absent during mitosis. Second, transcription in anterior nuclei initiates slightly earlier than in those closer to the center of the embryo, both because anterior nuclei inhabit a region of higher concentration of Bcd and because of metasynchronous nuclear divisions propagating as a wave towards the embryo center (Foe and Alberts, 1983). Consistent with this expectation, during the first minute of the 13th interphase we observe a gradient of nascent mRNA content along the AP axis (Fig. 2.S4B). Third, we designed probe sets to label the 5' and 3' portions of the completed transcript with fluorophores of green and red colors, respectively. If nascent sites are composed of incomplete transcripts, then 5' sequences must be more numerous than 3' sequences (Fig. 2.S4C), resulting in an increase of green signal at the expense of red. In agreement, our measurements reveal the enrichment of green signal after normalization to cytoplasmic unit intensity (Fig. 2.4B). Importantly, greater 5' enrichment is observed as the fraction of transcript labeled with green fluorophore increases (increasing slopes of fit lines in Fig. 2.4B). Thus, nascent *hb* loci are largely composed of unfinished transcripts and serve as a measure of transcriptional activity.

---

<sup>1</sup>The FISH labeling protocol was subsequently improved, and this two-color calibration control experiment now stably generates values within a few percent of 1 (data not shown).

## 2.5 Variation in nascent transcription site activity is 6-fold higher than variation in cytoplasmic output

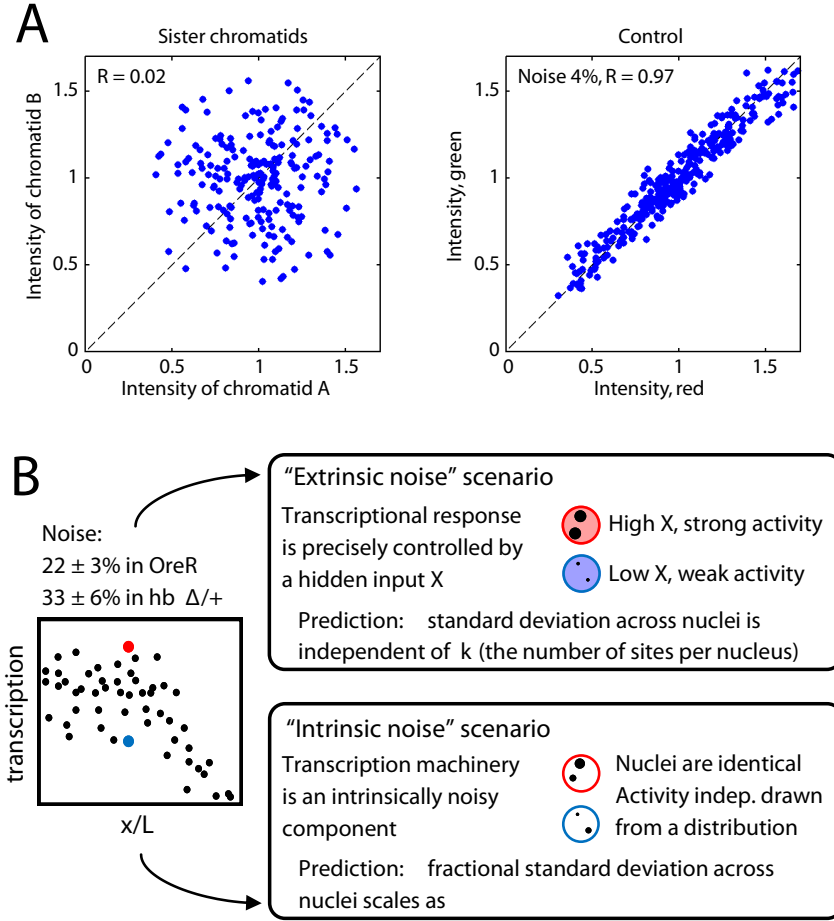
Given the low noise in cytoplasmic counts, we expected that, within the domain of greatest transcript accumulation, the nascent mRNA content at all available genomic loci would rise simultaneously until the locus is saturated with RNA polymerase II (RNAP), in principle reaching and sustaining some maximum nascent mRNA content. However, our measurements of nascent mRNA content show otherwise (Fig. 2.4C-E). The nuclear nascent mRNA content varies by  $22 \pm 3\%$ , 3-fold higher than that observed in cytoplasmic counts. Importantly, to be certain that this variability does not result merely from the delay in attaining steady-state maximum activity after mitosis (Fig. 2.S4D), we specifically confined our analysis to mid and late interphase 13 embryos. We observe this degree of variation even when loci are allowed the full temporal extent of the 13th interphase to reach any putative maximum value (Fig. 2.4E). These results indicate that *hb* loci fail to sustain any amount of uniform maximum content.

We note that the 22% variation we observe is a measure of fluctuations across a possible maximum of 4 active genomic loci in WT embryos. In embryos heterozygous for a *hb* deficiency, the total nascent mRNA content per nucleus, summed over a maximum of only two loci, varies by  $33 \pm 6\%$ , as expected if each locus performs an independent readout with a variability of 45% ( $22\% \times 2 = 44\%$ ,  $33\% \times \sqrt{2} = 47\%$ ), or a 6-fold increase over fluctuations in cytoplasmic counts. Analyzing closely apposed alleles on sister chromatids with sufficient separation to reliably gauge intensities reveals no correlation in their activities (Fig. 2.5A), indicating independent activity even for recently duplicated loci. Importantly, if the observed fluctuations result from variability in any input factor controlling *hb* expression (i.e., “extrinsic” noise), then

the variation in total nuclear activity would show no dependence on the number of loci in the nucleus. However, as the noise scales with the number of loci (Fig. 2.5B), the fluctuations we observe in the maximally expressed domain are intrinsic to the process(es) of transcription and not determined by variability in the controlling inputs.

Transcriptional activity will necessarily exhibit some degree of noise arising from stochastic single-molecule events, but the fluctuations we observe exceed the Poisson expectation considerably. From our observations, we can estimate the number of RNAP engaged in transcription per nucleus and thereby determine the expected degree of fluctuations; we find the predicted noise magnitude of at most 11% (see Sec. 2.F). The observed fluctuations of  $22 \pm 3\%$  are at least 2-fold greater than this prediction, ruling out a model in which transcriptional fluctuations in the region of maximum expression are determined by a single rate-limiting step of RNAP loading (Fig. 2.S4C).

From these observations, we conclude that, first, even in the domain of maximal expression, *hb* is not saturated with the maximum possible density of RNAP; and, second, despite the near uniformity of cytoplasmic transcript concentration, instantaneous activity of individual *hb* loci is intrinsically stochastic. The estimated variation in transcriptional activity at an individual locus is very similar to the minimum value of  $\approx 50\%$  observed for differences in mRNA numbers for the most highly expressed genes in yeast and *E. coli* (Gandhi et al., 2011; Taniguchi et al., 2010). In all contexts, variation between cells is significantly higher than that predicted for a process with a single rate-limiting step (Chubb et al., 2006; Golding et al., 2005; Le et al., 2005; Raj et al., 2006). These similarities across such diverse contexts suggest that the observed fluctuations are globally inherent features of the activity of otherwise “fully activated” genes. In the context of a rapidly developing embryo, the highest attainable expression rate would serve to minimize cell-to-cell fluctuations to



**Figure 2.5: Fluctuations in *hb* transcription are dominated by intrinsic noise.** **A:** Transcriptional activity (measured as difference-of-Gaussian intensities, see Sec. 2.G) of *hb* loci on optically resolved sister chromatids is uncorrelated (Pearson correlation coefficient  $R = 0.02$ ), compared to the tight correlation ( $R = 0.97$ ) in a control experiment using probes of alternating colors (with 4% imaging noise). **B:** Transcriptional variability arises from fluctuations in inputs (extrinsic noise) and from the process of transcription itself (intrinsic noise). Two extreme scenarios are presented in cartoon form. Upper panel: a fluctuating extrinsic input leads to correlated activities of transcription sites within a given nucleus; its contribution to the fractional standard deviation is independent of the number of transcribing loci  $k$ . Lower panel: intrinsic mechanistic noise affects all transcription sites independently; the fractional standard deviation scales as inverse square root of available transcription sites. Left: the measured transcription noise in WT and *hb* $\Delta/+$  embryos ( $22 \pm 3\%$  and  $33 \pm 6\%$ , respectively) shows scaling behavior characteristic of intrinsic noise with magnitude  $\approx 45\%$  ( $\sqrt{4} \times 22\% = 44\%$ ;  $\sqrt{2} \times 33\% = 47\%$ ).

the fullest possible extent and thereby promote precision. The observed tolerance of fluctuations, linked with the apparent inability to sustain saturating RNAP density, suggests that this degree of imprecision cannot be circumvented even in this highly precise developmental context.

## 2.6 Gap genes share expression characteristics and are produced at equal rates

The transcription of *hb* fluctuates around a mean polymerase density that is about half the level that is physically obtainable. What sets the magnitude of the mean activity level? Specific features of the *hb* promoter may limit activity. Alternatively, the maximum rate may not be specific to *hb* but shared among early expressed genes. If so, this would suggest that the maximum obtainable output, and its related noise level, are not set by any specific promoter-enhancer arrangement or any patterning cue, and instead are determined by general physical considerations.

To discern between these possibilities, we measured the accumulation of transcripts of four genes primarily responsible for the earliest patterning events: *hb*, *Krüppel* (*Kr*), *knirps* (*kni*), and *giant* (*gt*) (Fig. 2.6A). We found that all four “gap” genes display nearly uniform accumulation of cytoplasmic mRNAs accompanied by >3-fold higher fluctuations in instantaneous transcriptional activity, essentially identical to the characteristics of *hb* (Fig. 2.6B-C). These results strongly suggest that all early transcriptional events are subject to common constraints.

To closely compare transcript accumulation between genes, we took advantage of the observation that for *Kr* and *kni* cytoplasmic mRNA density increases monotonically between early nc12 and well into interphase 14 (Fig. 2.S5C-D), in contrast to *hb*, which ceases accumulating broadly in early nc14 (Fig. 2.3A). Hence we use the *Kr* or *kni* transcript counts in their respective regions of maximal expression as a proxy for



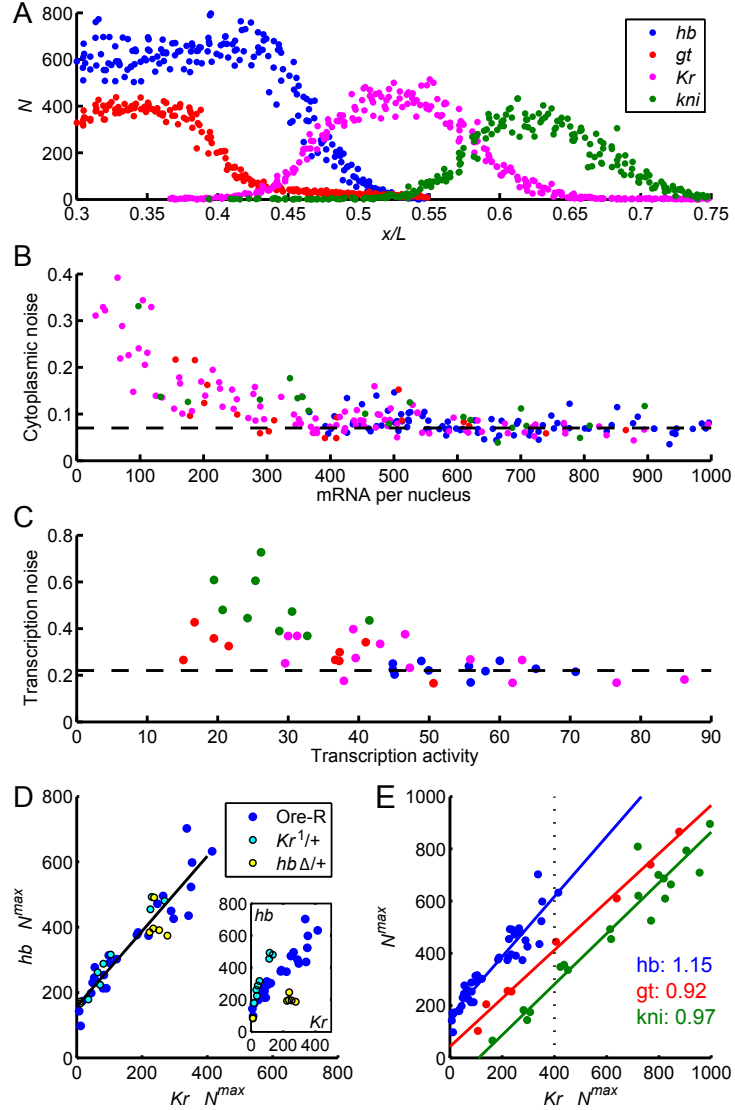


Figure 2.6: **Universal properties of transcripts of all gap genes.** **A:** Cytoplasmic profiles of four gap genes (absolute concentration per standard volume versus position along AP axis), measured in two embryos of the same age (second half of nuclear cycle 13; indicated by dotted line in panel E) processed with *hb* (blue) & *gt* (red) and with *Kr* (magenta) & *kni* (green) probes. **B-E:** Gap gene expression characteristics within each gene's region of maximum expression. **B:** Noise in cytoplasmic counts as a function of counts per nucleus (dashed line is at 8%). **C:** Noise in transcriptional activity as a function of activity level (dashed line is at 23%). **D:** mRNA expression, measured as mean absolute mRNA count per standard volume, in embryos from nuclear cycles 12 to early 14 co-stained with FISH probes against *hb* and *Kr* mRNA. Data from WT embryos (blue) coincides with those from embryos deficient for one copy of *hb* (*hb*Δ/+; yellow) or *Kr* (*Kr*<sup>1/+</sup>; cyan) when the concentration of the respective mRNA is rescaled by a factor of 2. Inset: Raw data (not rescaled). **E:** Expression levels of *hb* (blue), *kni* (green) and *gt* (red) versus expression of *Kr*. Data from WT, *hb*Δ/+ and *Kr*<sup>1/+</sup> embryos is combined by rescaling as in D (see also Fig. 2.S5C-D). *hb* data as in A; *kni* and *gt* were assessed in nuclear cycles 13 and 14. Slopes of fit lines indicate ratio of absolute production rates. These ratios are within 15% of unity.

time, reducing staging uncertainty when comparing different genes. We performed dual color labeling with probes against pairs of gap genes and report cytoplasmic counts of *hb*, *gt* and *kni* mRNA as a function of *Kr* (Fig. 2.6).

Fig. 2.6C displays the expression of *hb* and *Kr* in WT (blue), *hb* hemizygous (green), and *Kr*<sup>1</sup> heterozygous (red) embryos during nc12 and nc13. Counts in deficiency or mutant heterozygous embryos expectedly deviate substantially from those in WT (Fig. 2.6C inset), but after multiplying the counts of the deficient gene by 2, all points collapse onto the same line (Fig. 2.6C), demonstrating the absence of compensatory mechanisms. We observed the same behavior for *gt-Kr* and *kni-Kr* expression pairs. Unexpectedly, for the three sets of gene pairs, linear fitting yields lines with slopes between 0.9 and 1.15 (Fig. 2.6E); that is, in their regions of maximal expression, the four genes are produced at nearly identical rates. The differences between absolute levels within these regions (Fig. 2.6E) reflect differences in the timing of when *kni*, *Kr*, and *gt* transcripts begin to accumulate, and for *hb* the perdurance of maternal mRNA. These maximal production rates are independent of Bcd activator concentration: although genetically altering the dosage of *bcd* between 50% and 280% of WT shifts the expression domains along the AP axis (Liu et al., 2013), this manipulation does not alter either accumulation rate or precision (Fig. 2.S5A-B).

These results are consistent with the idea that these transcripts are produced at the same rate. This strong similarity occurs despite the fact that these genes are expressed maximally in non-overlapping spatial domains. The transcriptional activity in the maximally expressed domain and the magnitude of transcriptional noise are therefore set independently of the inputs that determine spatially patterned expression, which are specific to each gene. By focusing on the regions of maximal expression, we could isolate the features of transcription that appear to be universal across gap genes and, furthermore, match the noise characteristics previously observed in bacteria and cell cultures. This suggests that the failure to sustain a maximal

loading of RNAPs on the gene and the intrinsic noise of 45%, which adds on top of other possible noise sources, are a common feature of transcriptional activation across diverse biological contexts.

## 2.7 Discussion

The fundamental question of how embryos achieve precise control over the earliest transcriptional events is largely unanswered. General models of embryogenesis posit that early developmental events are dominated by molecular noise and imprecision in the control of gene expression (Arias and Hayward, 2006; Manu et al., 2009; Rao et al., 2002), a view consistent with observations of wide fluctuations in transcriptional activity in the majority of systems yet assayed quantitatively (Li and Xie, 2011). Indeed, the finding in fly embryos that instantaneous transcriptional activity varies between loci by nearly 50% suggests that the earliest transcriptional events of the fly embryo obey rules of stochastic activity as observed in other systems where output can vary by a similar degree, and is often much higher (Munsky et al., 2012). Stochastic variation appears to be a universal feature of transcription from single cell organisms grown in culture (Raj et al., 2006; So et al., 2011; Stewart-Ornstein et al., 2012) and for cells in certain developmental settings (Pare et al., 2009; Raj et al., 2010; Saffer et al., 2011). *Drosophila* embryos display an extraordinary degree of precision in the rapid establishment of distinct gene expression programs; nevertheless, even this system cannot circumvent stochastic transcriptional activity. This finding supports the idea that control systems that might overcome stochastic molecular activity are difficult to design, costly to implement, and rarely if ever found (Lestas et al., 2010).

It is possible that cultured yeast and bacteria exist at sufficiently high densities such that precision is not required to ensure survival of a large fraction of the population; indeed, in several cases stochastic expression serves to maximize survival

options (Balaban et al., 2004; Maamar et al., 2007; Mirouze et al., 2011; Nachman et al., 2007). In addition, for prokaryotes and haploid yeast and unlike early embryos, the presence of a single genomic locus precludes the possibility of noise filtering by averaging over independent loci. Alternatively, precision might be required to ensure the survival of single cell organisms when grown in their endogenous conditions, which may be difficult to study in a laboratory setting. However, in stark contrast to single cell organisms, many developing embryos possess large fields of cells that must coordinately undertake rapidly determined fate decisions, thus mandating high precision and low expression noise such that the appropriate gene expression programs are induced at the correct time and place. If in *Drosophila* patterning mRNAs accumulate in a precise manner minimizing expression noise, as we have shown here, how then can the embryo achieve this near uniformity?

### **Spatiotemporal averaging reconciles highly variant transcription with precise accumulation and recovers the input-output relationship**

The large differences we observe in nascent transcript content, if sustained over sufficiently long periods, would inevitably result in unequal transcript production, inconsistent with our observations of nearly homogeneous cytoplasmic transcript concentration. As noted above, the lifetime of *hb* transcripts is sufficiently long to allow substantial accumulation during the course of the syncytial blastoderm stage. If instantaneous nascent mRNA content is not maintained continuously throughout interphase but instead fluctuates about the mean as a result of fluctuating numbers of RNAP, then cytoplasmic accumulation serves as a natural time-averaging filter. The impact of time averaging can be estimated in two independent ways. First, transcript accumulation reflects temporal integration over the duration of the interphase of a signal fluctuating with a characteristic time  $t_0$ , the time it takes a polymerase to traverse the 3.2 kbp of *hb* gene. This time can be estimated using the RNAP proces-

sivity of 1.1-1.4 kbp/min reported in the literature (Irvine et al., 1991; O'Brien and Lis, 1993; Shermoen and Ofarrell, 1991; Thummel et al., 1990). This method provides a rough estimate consistent with the observed noise filtering (see Sec. 2.H). Alternatively, a more careful estimate (Fig. 2.S6B) allows obtaining a theoretical bound on the maximum efficiency of temporal averaging based on the quantities we measure directly; most crucially, the absolute number of engaged polymerases per nucleus. In the case of *Kr* mRNA profile, it shows that by the time the mean expression level reaches 800 molecules per nucleus, pure temporal averaging can at most reduce the expression noise to 8%. For these late embryos, however, our measurements show a consistently lower noise level of  $6 \pm 2\%$ , suggesting that an additional noise filtering mechanism must be at play.

This additional filtering can be readily provided by a small degree of spatial averaging, i.e. by the exchange of mRNA between neighboring cytoplasmic volumes during nc13 and early nc14 before the partitioning of the syncytial blastoderm. Our data shows that mRNA possesses some mobility: both *hb* and *Kr* transcript numbers increase at  $>10\mu\text{m}$  from their sites of production in nuclei (Fig. 2.S3A, Fig. 2.S6A). We note that because the cylindrical summation volumes we assay possess a diameter of one internuclear distance (and thus contact each other), the mRNA traveling distance required for us to observe some effective spatial averaging is very small. A straightforward estimate (see Sec. 2.H) shows that attributing the excess noise filtering to spatial averaging requires only 4% of produced transcripts to be exchanged between neighboring volumes. Thus, even a limited degree of spatial averaging is completely sufficient to account for the appearance of low variation in cytoplasmic accumulation from stochastic transcription.

These results have several implications. First, we note that the observed variation in cytoplasmic concentration is likely to contain error introduced by our measurement, and the variation we observe is nearly at the level of counting noise (Fig. 2.3A, inset).

This might indicate that spatial averaging predominates the filtering of transcription noise; however, the degree to which RNAP numbers fluctuate, and therefore the extent of purely temporal averaging, can only be determined with measurements of transcription in living embryos. Second, both spatial and temporal averaging mechanisms effectively relax a requirement for rapid, tightly controlled transcriptional responses to modulating inputs, thereby minimizing the need for additional layers of feedback or other control systems. In turn, the fluctuations of putative inputs must approach the same degree of noise as the intrinsic variability of the transcriptional process itself before any effect on gene expression is realized.

It is well established that the position of the Hb expression boundary depends on Bcd genetic dosage (Driever and Nüsslein-Volhard, 1988), and that the concentration of Hb protein along the AP axis depends upon and is at least as precise as Bcd concentration (Gregor et al., 2007). Superficially, a highly stochastic transcriptional response would appear to render irrelevant any link between Bcd precision and Hb output: the 10% fluctuations observed for Bcd (Gregor et al., 2007) cannot directly impact a transcriptional process whose noise is  $>40\%$ . However, it is vital to recall that each nucleus employs averaging mechanisms to reduce the effect of intrinsic noise. Because of the central role played by time averaging, the relative importance of various noise sources depends on the time scale of observation. The immediate readout (on a scale of minutes) is dominated by intrinsic transcriptional noise which renders the precision of the input irrelevant. However, after averaging over all active loci within a nucleus, over time and to a certain degree over neighboring nuclei, the contribution of the intrinsic noise decreases to become comparable with the input (or extrinsic) fluctuations, and thanks to these mechanisms, on a long time scale (such as the 3 hours of development) the precision of patterning decisions becomes limited by the latter. A precise response to Bcd input will be recovered as long as the mean activity of *hb* transcription is correlated with Bcd concentration, as proposed

previously (Erdmann et al., 2009). This reconciles the apparently stochastic behavior of *hb* transcriptional activity with the precisely positioned boundary of expression (Porcher et al., 2010). In this manner, Hb activity and fluctuations in the boundary domain retain the previously observed dependence upon levels of and fluctuations in Bcd concentration.

### **Limitations to precise control of gene expression**

We have shown that in the context of the gap genes, transcript output in the maximally expressing region does not equate with the actual maximum attainable density of RNAP loading. This maximum is attained by only a small fraction of nuclei at any given moment. Thus, it is currently unclear what determines the mean density of RNAP loading common to these four genes and what prohibits all nuclei from continuously activating the achievable maximum density. It is possible that the output rate is determined by a common, maternally supplied and spatially ubiquitous factor, for example Zelda or BSF (De Renzis et al., 2007; Liang et al., 2008), which calibrates the RNAP density of these four genes to give rise to the observed transcript output rate. Conversely, from the perspective of noise minimization, it would seem advantageous to design these genes' promoter-enhancer architecture such as to obtain the actual maximum possible density, since higher output achieves greater noise reduction. However, a biological system likely cannot be readily engineered to produce transcripts at an arbitrarily rapid rate. Hence, it seems likely that mean RNAP loading, and hence transcript output, is strongly influenced by physical considerations, such as transcription factor binding, promoter melting, enhancer looping, and/or chromatin accessibility, that might be difficult to overcome by any simple means. Future work will determine the extent to which the mean polymerase density we observe for these four genes is a shared feature of early expression and the extent to which this rate can be manipulated according to cellular context. Moreover, further studies will be

required to determine the timescale of fluctuations of an active locus during interphase: that is, whether variations arise largely from “bursts” of dense RNAP loading followed by quiescent periods, or conversely if the variations result from RNAP loading rates that are maintained continuously during interphase but differ dramatically between loci.

The formation of cellular membranes during the 14th interphase prohibits spatial exchange. It is thus improbable that spatial averaging mechanisms can play a role in ensuring precise responses at this time. Moreover, the transcripts of the pair-rule genes are directed to the apical surface where they accumulate (Davis and Ish-Horowicz, 1991). Differential cellular behavior, presaging the formation of morphological structures, emerges in the latter part of the 14th interphase. Thus it is likely that shortly after the onset of the 14th interphase, individual cells begin accumulating gene products required for their specific behaviors, thus rendering spatial averaging a hindrance to differentiation. It is therefore likely that temporal averaging and/or other mechanisms such as regulatory feedback ensure the precise distribution of patterning factors at this time. The degree of precision of transcriptional events over the course of the 14th interphase will be the subject of future investigations.

In summary, this work demonstrates the power of the *Drosophila* embryo as a system for quantitative evaluation of transcriptional regulation. Early fly embryos possess a number of advantages enabling such studies, including modulatory patterning inputs spanning large dynamic ranges, a complete list of essential gene network components, and an abundance of modern analysis tools. This presents the unique opportunity to uncover the biological and physical design features of a system evolutionarily constructed to achieve rapid and precise establishment of cell fates in an intact, physiologically meaningful context.



# Technical details

## 2.A Experimental procedures

For details of fly strain and embryo manipulation and FISH protocol, see the published version of this work, Little et al. (2013). My work on this project required me to learn and apply these experimental techniques, and the hands-on experience with all stages of the protocol proved invaluable for pushing mRNA detection to the limits of precision reported here. However, I was not directly involved in their development or fine-tuning, except through identifying performance-limiting factors and designing metrics to quantitatively evaluate the effect of protocol modifications explored by my experimental collaborator, Shawn Little. Here, I concentrate on my primary contribution, namely designing the data analysis approach.

## 2.B Detection of individual *hb* transcripts: the workflow

The analysis protocol described below was developed in close contact with the data, informed by experimental limits and needs. In this section, I provide an overview of the data analysis methodology employed in this work.

A single entry in our dataset library contains the following set of objects:

- Two low-magnification DAPI images of the whole embryo in register with each other: one in the mid-sagittal plane and one at the surface;
- A high-magnification DAPI image of the surface of the embryo in register with the FISH channel(s);
- One or more channels of FISH data, each containing a flat-field image and a confocal stack of images;
- A “tag file” containing all the information describing the embryo and the channels.

To achieve dynamic range required for an accurate measurement of spots as dim as single cytoplasmic transcripts and as bright as sites of nascent transcription, we acquire two FISH stacks for each fluorophore color: a low-power stack for measuring transcription intensity, followed by a high-power stack for measuring cytoplasmic counts. For experiments that did not include transcription intensity measurements, only high-power stacks are collected for each gene of interest. I should note that since this work was published, the acquisition of new single-photon counting (HyD) detectors for the microscope allowed us to switch to a single-power data acquisition protocol.

Image analysis can be split into five major steps:

1. Pre-processing of image stacks and collection of information on the embryo;
2. Identifying fluorescent particles;
3. Defining summation volumes for cytoplasmic counts and transcription sites;
4. Calibrating total fluorescence to measure transcription intensity and cytoplasmic counts;
5. Extracting features of cytoplasmic profiles.

### 2.B.1 Pre-processing of image stacks

Raw image stack is flat-field corrected and realigned (up to 2 pixel  $xy$  shift between individual frames) to compensate for stage drift. Alignment shifts are determined by maximizing cross-correlation between central 300-by-300 pixel regions of successive frames. Low-magnification mid-sagittal DAPI image is thresholded to obtain the embryo mask. The points of the embryo that extend the furthest along the major axis of this ellipsoid-like mask are designated as ends of the antero-posterior axis (the embryo is always imaged in the same orientation, allowing the software to correctly select the anterior and posterior tips). High-magnification DAPI image is resized to the same scale as low-magnification DAPI image of the embryo surface and correlation analysis (search for  $xy$  location providing the best match between two images) is used to identify the exact location of the high-magnification imaged region within the embryo.

Next, the high-magnification DAPI image of the embryo surface is used to identify the locations of nuclei centers. Detection is performed by an automatic routine and is verified and corrected by human input, particularly for embryos in mitosis where, by convention, one nuclear center is selected for each dividing nucleus. This resolves any ambiguity in cases where the embryo was fixed in the middle of a mitotic wave.

Finally, we manually select a region along the midline of the embryo (at least 5 nuclei wide in nc12, and at least 20 nuclei wide in nc14) where tissue deformation, which can be visually estimated via the change in density of nuclei, is minimally inhomogeneous. Cytoplasmic counts are measured in fixed physical volumes and so are affected by the degree of tissue compression; we therefore only measure them in nuclei belonging to this minimally deformed “center band”. In contrast, our measure of total transcription activity of a nucleus is insensitive to tissue deformation, so we use all nuclei to improve statistics. Using only nuclei in the center band does not reduce the measured noise of transcription activity.

### 2.B.2 Identifying fluorescent particles

Raw images are filtered using a Difference-of-Gaussians (DoG) filter with inner and outer Gaussian parameters 1.2 and 2.2 pixels, respectively. This is a balanced filter giving a measure of local contrast (zero response on a constant image and high response on a spot, i.e. a locally bright region surrounded by a dark one). A “master threshold” is chosen (in a manner that will be described shortly) and local maxima exceeding this threshold in the DoG-filtered images are detected in each frame individually. This yields a list of several million “bright spots”, on the order of 30 thousand per frame. Given that the step in  $z$  direction in our confocal stacks is several times smaller than the width of the PSF of the microscope (Fig. 2.S1D), each true point-like source of fluorescence is clearly visible at the same location on at least 3 consecutive slices. We call these multiple images of the same real particle “shadows” of each other. In contrast, random fluctuations in the background are extremely unlikely to happen at the same position in two independent frames. This provides a powerful filtering criterion: we arrange all detected bright spots into columns of “shadows” (allowing a  $\pm 2$  pixel relative shift in  $xy$  position to account for a possible shift due to imaging noise or physical movement) and reject all columns of height less than 3. All columns with two clear peaks in  $z$  are separated; finally, the brightest shadow in each column is identified as a candidate true particle. This selection process eliminates  $>90\%$  of bright spots and allows us to obtain remarkable signal-noise separation (Fig. 2.2D). Reducing the required number of shadows to 2 increases the total number of detected particles by  $<3\%$  percent.

The master threshold is adjusted iteratively so that in the end, the particle intensity histogram (Fig. 2.2D) exhibits two peaks of equal heights when plotted on linear intensity scale: the peak of spots due to random imaging noise and the peak of cytoplasmic mRNA. This ensures that the master threshold is low enough that even the dimmest cytoplasmic particles had all 3 shadows detected, but not too low

so as to lead to an explosion of noise peak (the shadow filtering scheme breaks down if local maxima are detected at a density where spontaneous creation of columns is ubiquitous). A bleaching factor (typically  $\approx 5\%$  intensity loss per 10 frames) is determined from a linear fit of logarithm of particle intensity versus index of frame, and raw intensities of frames are bleach-corrected accordingly. After correction, intensity distributions of detected particles coincide for all depths.

This entire analysis step, including threshold choice, is fully automated.

### 2.B.3 Defining summation volumes

We begin by labeling all candidate particles as transcription sites, cytoplasmic transcripts or noise. The labeling is performed using simple global thresholding of bleaching-corrected particle intensity. The threshold separating cytoplasmic transcripts from noise is defined as the bottom of the valley between the two peaks on the particle intensity distribution histogram as in Fig. 2.2D (red line). The threshold between cytoplasmic transcripts and transcription sites (Fig. 2.2D, green line) is determined from the observation that unlike cytoplasmic particles, transcription sites are tightly clustered in  $z$ . We list candidate particles in order of decreasing intensity and construct a running standard deviation of their  $z$  coordinate in a window of 50 particles. We observe a sharp transition between “tightly clustered” and “equally likely to appear at all depths”; and use it to set the threshold. Both threshold-setting procedures are automatic; the latter is approved by human input and corrected if necessary, e.g. in embryos undergoing mitosis where no transcription sites are visible by eye nor expected to be present. Typical transcription site threshold intensity set in this manner is 3-4 times the mean intensity of a cytoplasmic particle. A flexible threshold outperforms a fixed one (a constant number of cyto units in each embryo), because the optimal threshold is lower in younger embryos (with fewer cyto spots)

than in older embryos, where a large number of cytoplasmic transcripts makes them more likely to randomly exceed the same threshold.

Particles identified as transcription sites are each enclosed in a parallelepiped of  $19 \times 19 \times 9$  pixels ( $1.4 \times 1.4 \times 2.3 \mu\text{m}^3$ ); total fluorescence collected from these volumes is attributed to transcription activity, and assigned to nuclei using a Voronoi tessellation based on nuclear centers. The choice of parallelepipeds rather than cylinders was motivated by computational efficiency; all total fluorescence-based measurements are background corrected (see below) and were checked not to depend on the choice of shape of summation volume for transcription sites.

For cytoplasmic mRNA, we defined a measure of local expression based on the number of transcripts contained in a band of  $12 \mu\text{m}$  from the plasma membrane. This choice was motivated as follows. On the one hand, we would like our measure to be local, reflecting the transcriptional output of an individual nucleus, inasmuch as it is possible in a syncytium. On the other hand, transcripts are largely depleted from nuclei during the majority of interphase (Fig. 2.2E) and as nuclei change size, their spatial distribution varies accordingly; in nuclear cycle 14, as nuclei elongate, this spatial reorganization can extend up to a depth of  $11\text{-}12 \mu\text{m}$ . Counting all transcripts within a band of  $12 \mu\text{m}$  is therefore a natural compromise between keeping the summation volume small and reducing sensitivity to spatial reorganization of RNA density, focusing on transcript accumulation with age.

Specifically, we enclose each nucleus in a cylinder extending from the embryo surface down to  $12 \mu\text{m}$  from the plasma membrane, centered at the nucleus center and of a diameter equal to the average inter-nuclear distance in the “center band” of the embryo as defined above. We then exclude from this volume the parallelepipeds containing transcription sites, and define “mRNA count per nucleus” as the total number of transcripts in this volume, called summation cylinder. Note that, for consistency, we exclude transcription site parallelepipeds from cytoplasmic summation

volume even when measuring direct counts of cytoplasmic particles (not based on total fluorescence). Since mRNA transcripts are largely excluded from nuclei, this modification of summation volumes changes the direct counts by less than 1%.

We stress that summing these numbers for all nuclei does *not* correspond to the total number of mRNA in the embryo; our goal is not to account for all mRNA molecules, but to quantify noise in transcription activity and output. For the same reason, we measure mRNA counts in cylinders of fixed volumes rather than a Voronoi-type tessellation of the embryo surface.

Note that the volume of the summation cylinder changes with age and decreases twofold with every nuclear division. To compare transcript density between embryos of different ages, we define the “standard volume” as the volume of the summation cylinder at nuclear cycle 14, and rescale measured “counts per nucleus” into counts per standard volume, which we also call “counts per cell”: for an embryo in nuclear cycle  $k$ , we define  $N_{\text{cell}} = N_{\text{nucleus}}/2^{14-k}$ . This quantity measures the number of transcripts in equivalent volumes and is directly comparable for embryos of different ages. The name “counts per cell” reflects the fact that at 14th nuclear cycle the embryo exits syncytial stage and cellularization occurs.

## 2.B.4 Calibrating total fluorescence

Measurements of cytoplasmic concentration in dense regions and of transcription activity are based on total fluorescence. To convert these to “cyto units”, we use the following calibration procedure. A large number of test summation volumes are placed at regularly spaced positions within the center band in the embryo and transcription sites are excluded as described above (this allows us to obtain good statistics, since relation between total fluorescence and direct counts holds for any volume, and we need not be limited by the number of nuclei we sample). For each summation volume, we calculate the total fluorescence collected from it as well as its total volume in voxels.

The former is tightly correlated with direct counts measured in the same volumes (see Fig. 2.S2A), and in a broad range of particle densities is well described by a linear dependence. For data in this linear regime (whose bounds are selected manually), we fit a two-parameter model:  $F = \alpha D + \beta V$ , where (F, D, V) is the data (total collected fluorescence, direct particle count and volume in voxels, respectively, for each summation volume), and the two parameters are  $\alpha$ , total fluorescence of a single transcript, and  $\beta$ , background fluorescence per voxel. We call  $\alpha$  “the cytoplasmic unit of total fluorescence”, or “cyto unit” for short. Note that the explanatory fit on Fig. 2.S2A was plotted as if summation volumes were all identical; in practice this is true only approximately. This assumption was made only for clarity of presentation, and the analysis software performs the fit taking into account the (almost negligible) differences in volume. Knowing  $\alpha$  and  $\beta$ , we can calculate “total fluorescence-based counts”  $C_F$  of mRNA per nucleus:  $C_F = (F - \beta V)/\alpha$ . The same formula allows converting total transcriptional intensity per nucleus into activity expressed in cyto units.

Finally, cytoplasmic counts are corrected for tissue deformation, estimated using nuclear density fluctuations. For each nucleus the correction factor is determined as the square of mean inter-nuclear distance to local inter-nuclear distance ratio, where the latter is defined as the average distance to 6 nearest neighbors.

“Intensity” of a cytoplasmic particle always means DoG intensity (Fig. 2.2DF; Fig. 2.S1E-G). Total fluorescence is always measured in raw pixel values (Fig. 2.S2A; all transcription intensity measurements). Note that Fig. 2.2F uses a “cyto unit of intensity”, i.e. the mean DoG intensity of cytoplasmic particles. In all other cases “cyto unit” refers to the parameter  $\alpha$  as defined above (“cyto unit of total fluorescence”).



### 2.B.5 Extracting features of cytoplasmic profiles

For each dataset, the diagnostic plot as on Fig. 2.S2A is inspected to determine whether total-fluorescence-based counts or direct counts provide a better estimate of the expression in the “fully on” region. In young embryos (e.g. most *nc12* embryos), cytoplasmic densities are low enough that the coordinated deviation from linearity is imperceptible, i.e. negligible compared to the spread of datapoints in the direction perpendicular to the roughly linear global trend. In these embryos, using total fluorescence would not change the mean level of cytoplasmic counts in any given bin along antero-posterior embryo axis, but would add noise. Consequently, for these embryos direct counts are used as more reliable; in cases such as depicted on Fig. 2.S2A, total fluorescence counts are used instead. For every embryo the more reliable measure of cytoplasmic counts per nucleus is denoted  $N$ , and the plot of  $N$  versus normalized nuclear locations along antero-posterior axis ( $x/L$ ) constitutes the “cytoplasmic profile”. For every profile, its region of maximal expression is selected manually as a band of 5-10% embryo length for *hb* and *gt*, and 4-7% embryo length for *Kr* and *kni*. The mean cytoplasmic expression level within the maximal expression domain is denoted  $N_{\max}$ . For every profile, a smoothed profile is obtained using a global cubic spline fit, constrained to zero derivative at the edges of the imaged region, with up to 4 internal breakpoints whose positions are adjusted to minimize  $\chi^2$ ; the number of breakpoints is a function of embryo age and reflects the fact that profile shape becomes more complex with time (Fig. 2.S3C). Expression noise is defined as the root-mean-square deviation of datapoints from the smoothed profile in the region of maximal expression.

## 2.C Detection of individual *hb* transcripts:

### **FishToolbox v2.0.12**

The first stages of image processing described below, namely the detection of particles in sub-PSF-sliced confocal stacks, is at the core of this analysis, and after several years of improvement and usage it is the most streamlined and versatile part of the code I developed. I implemented it as a package of MatLab routines (**FishToolbox**), thoroughly commented and designed to be modular and easily adaptable for other applications. Since its creation, **FishToolbox** was used internally by many members of Prof. Gregor’s group, externally on zebrafish data, and now my code also serves as a base for the live transcription analysis software **Live\_mRNA** developed by H. Garcia.

The core of **FishToolbox** is a routine that takes a stack of images and detects locations and properties of point fluorescent sources within this stack; it also includes tools to maintain a library of datasets labeled using tags (e.g., specifying nuclear cycle, protocol version, date of imaging etc.) and to run the analysis on all or a subset of available datasets (for example, all embryos of cycle 13 imaged after a certain date). Some additional routines are available to visually inspect and check the quality of the performed fits, as well as some other specialized routines for pre- and post-processing the data.

A software package is an evolving entity. Any usage instructions I could provide here would quickly become outdated: concurrently with this writing, I am working on implementing some modifications in the code, to ensure seamless integration with the software developed by H. Garcia. For the most up-to-date installation and usage information, complete with example parameter files and a “getting started” tutorial, I refer the reader to the instruction manual provided with the package distribution available at <https://github.com/PrincetonUniversity/FishToolbox/>.

## 2.D Supplementary figures

### Figure 2.S1: Detection of individual *hb* transcripts

**A-D:** 3d detection and point spread function. Fluorescent particles appear on 6 consecutive confocal slices spaced by 250 nm. A cytoplasmic mRNA transcript (**A**), a nascent transcription site (**B**), and a fluorescent bead (**C**) are shown in  $42 \times 42$  pixel windows; pixel size is 76 nm. **D:** Axial intensity profiles (point-spread-functions in  $z$ ) for particles in (A, blue), (B, red) and (C, green), normalized to unit integral. For comparison, lateral intensity profile versus  $x$  is shown on the same axis (black dashed line; normalization scaled to fit axis). **E-G:** Particle detection efficiency. After labeling mRNA with alternating colors, thresholds were found for each channel as in Fig. 2.2D. Because a particle may be detected at two different positions in each channel (typically  $\pm 1$  pixel, either from imaging noise or physical movement of the sample), we choose one channel for detection and measure intensity in the other (sample). This procedure therefore provides a lower bound on detection efficiency. Plots show 2d intensity histogram for all detected spots. **E:** 2d histogram showing green intensity at the exact locations of red particles versus their red intensities. Distribution is clearly bimodal in 2d, demonstrating good co-localization. Red and green lines indicate thresholds for each channel. Of red spots classified as mRNA particles, at most 22% escape detection in green channel (are classified as “noise”). **F:** Same as E, but with detection and sample channel inverted. Of red spots classified as mRNA particles, at most 27% escape detection in green channel. The probability for an mRNA molecule to go undetected in both channels is therefore at most  $0.22 * 0.27 = 6\%$ , i.e. the detection efficiency is better than 94%. **G:** As a control, we repeat the procedure in (E), but plot green intensities not at the exact locations of red particles, but shifted by 5 pixels ( $0.38 \mu\text{m}$ ) in both  $x$  and  $y$ . The high degree of correlation (i.e. color co-localization) exhibited on panel E is lost. **H:** Absolute RT-qPCR control measurement. Threshold cycle  $C_T$  as a function of number of plasmid DNA (blue) or in vitro synthesized *hb* mRNA (red) molecules per PCR with (for mRNA) or without (for plasmid) reverse transcription. Nucleic acid concentration was determined by spectrophotometry with a measurement error of 1.1%.  $C_T$  for each amount of

nucleic acid was assayed in 4 independent reactions. Lower inset:  $C_T$  for individual embryos 30-60 min in age processed and diluted as described in Experimental Procedures ( $n = 6$  whole embryos,  $n = 6$  embryos serially diluted). Linear fits to data and error propagation were performed as described in Extended Experimental Procedures. Upper inset: the mean and standard deviation in  $C_T$  when  $2.4 \cdot 10^7$  molecules of standard mRNA were analyzed by RT-qPCR either by adding directly to the reaction (input, 4 independent samples) or after processing with same RNA extraction protocol used for embryos (recovered,  $n = 10$  independently processed samples). Using mRNA standard curve to convert  $C_T$  to mRNA number, the extraction protocol yields recovery of  $0.77 \cdot 10^7 / 2.4 \cdot 10^7 = 32\%$  of input, and increases the measurement variability (fractional standard deviation) to 31%, compared to 12% without the extraction protocol. By RT-qPCR, an Ore-R embryo contains  $8.7 \pm 3.1 \cdot 10^5$  mRNAs. **I:** Maternal hb mRNA density in  $hb\Delta/+$  and WT embryos differs 2-fold. Density (molecules per  $\mu\text{m}^3$ ) of maternal hb mRNA as a function of position along the AP axis. Data shown for two WT embryos (blue traces, both in nc8), and two  $hb\Delta/+$  embryos (nc7 and 8). Solid blue line is at half the mean of WT traces. Density was measured by counting all transcripts in a band running along the center of the embryo of width equal to 30% egg length (EL), in a stack of 20 frames ( $5 \mu\text{m}$  in  $z$ ), binned in regions 5% EL wide. These volumes contain about 4500 maternal mRNA molecules in a WT embryo.

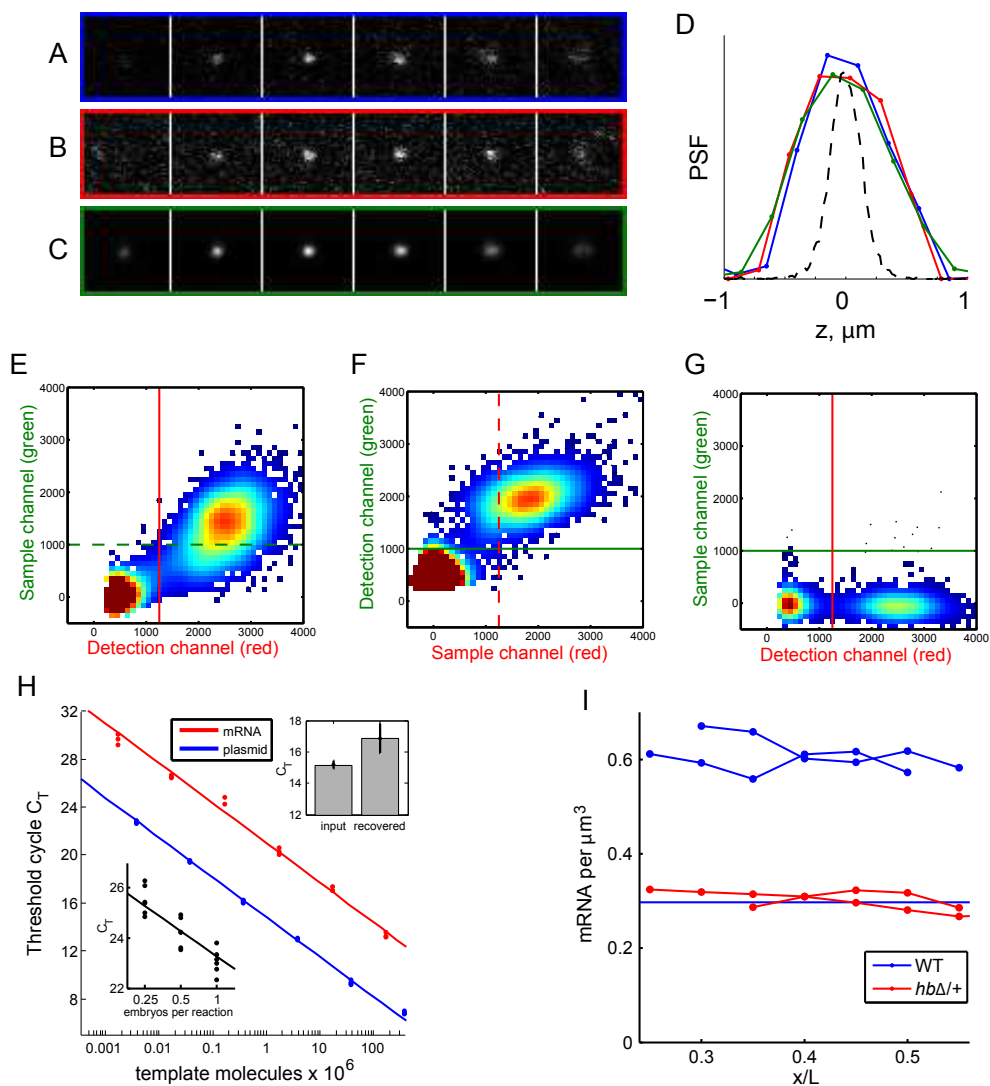


Figure 2.S1: Detection of individual *hb* transcripts.

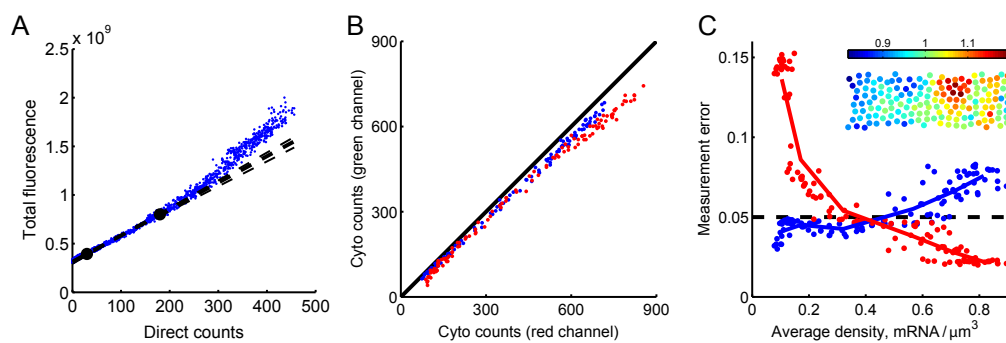


Figure 2.S2: Particle counts and total fluorescence.

## Figure 2.S2: Particle counts and total fluorescence

**A:** Correction factor for densely populated regions. The total fluorescence of labeled cytoplasmic *hb* mRNA is shown as a function of number of resolved particles (“direct counts”) in the same averaging volume. The deviation from a linear dependence at high counts occurs when spots become too dense to be resolved reliably, at approximately 1 molecule per  $\mu\text{m}^3$ . Each data point corresponds to a summation volume of  $2.8 \cdot 10^5 \text{ pixel}^3$ , i.e.  $400 \mu\text{m}^3$  (cylinders with radius  $\approx 3 \mu\text{m}$  and height  $13 \mu\text{m}$  drawn at regular intervals from a band parallel to AP axis). Bold dashed line is a fit to the data between manually selected black dots; its slope is the total fluorescence collected per mRNA molecule ( $2.7 \cdot 10^6$  pixel intensity units) and its offset is background fluorescence ( $\approx 1.1 \cdot 10^3$  pixel intensity units per voxel). These parameters provide absolute calibration to convert total fluorescence into absolute mRNA number in high density regions. Error bars of the fit (light dashed lines) are 12%; this includes the uncertainty arising from the choice of the range used for the linear fit (location of the black dots). This 12% error defines the uncertainty in absolute measurement of cytoplasmic concentration in dense regions. **B, C:** Precision of total fluorescence measurements and of direct counts as a function of mRNA density. Cytoplasmic counts of *hb* mRNA were measured using probes of alternating colors, using direct particle counts or calibrated total fluorescence in each of the two color channels. **B:** Scatter plot of cytoplasmic counts per nucleus as measured in red and green channels (direct particle counts, blue; calibrated total fluorescence, red). The error of absolute concentration measurement is 10% for direct counts and 12% for total fluorescence counts, determined as deviation from 1:1 slope. This error is dominated by uncertainty of intensity threshold selection when assigning direct counts, and by the normalization procedures when calibrating total fluorescence. **C:** Precision of relative measurements made using direct particle counts (blue) or calibrated total fluorescence (red), as a function of average (not peak) mRNA density in summation volume. Error is defined as the root-mean-square deviation of data points from the linear slope in (B), scaled by the mean (distance to (0,0)). As expected, direct counts (total fluorescence) are more precise at low (high) densities, respectively. Overall the precision of relative measurements of cytoplasmic counts in a given volume is 5% (dashed line); in the region of maximal ex-

pression it goes down to 2-3%. Inset: The uncertainty of expression noise measurement is dominated not by uncertainty of counting, but by the uncertainty in summation volume definition arising from tissue deformation after hybridization. Shown are nuclei in a stripe along the midline of an embryo, color-coded to highlight inhomogeneity of nuclear density in fixed samples. Color indicates density (number of nuclei per unit area) relative to the average density; this is used as a correction factor for cytoplasmic counts to compensate for tissue deformation. Absolute magnitude of correction can reach 15% in deformed embryos (as shown); local fluctuations of correction factor (imperfect compensation) are 5% and dominate the uncertainty in measuring expression noise (compare with 2-3% of counting uncertainty in a given volume).

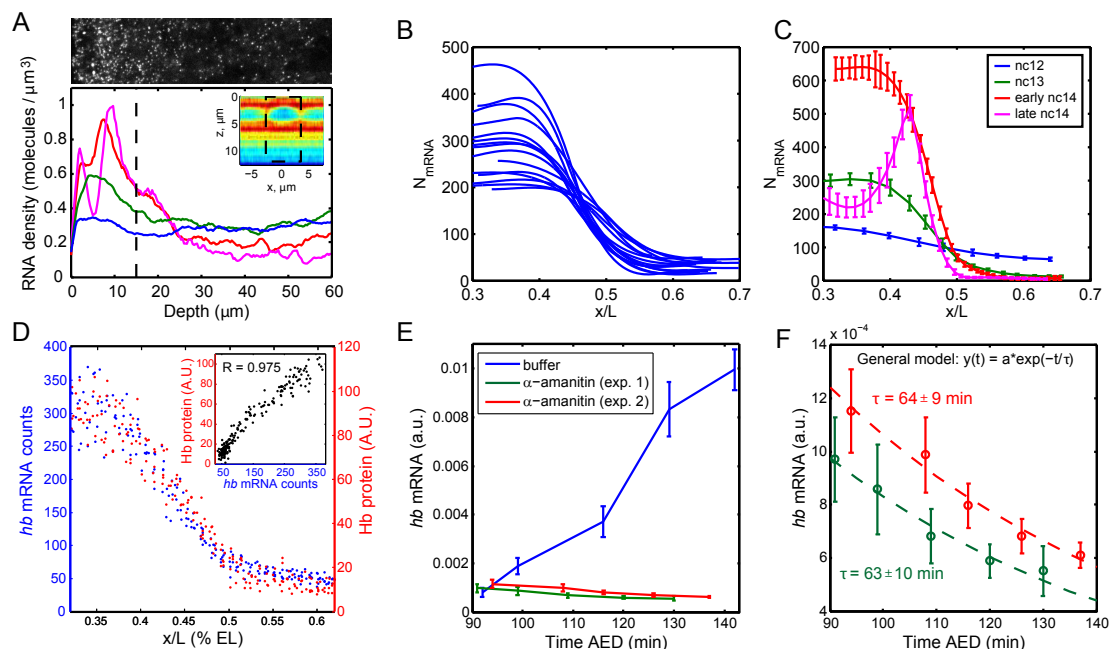


Figure 2.S3: **Spatial distribution of *hb* mRNA and protein, and *hb* mRNA life-time.**

## Figure 2.S3: Comparison with Hb protein, and *hb* mRNA lifetime

**A:** Upper: midsagittal view of a  $1\ \mu\text{m}$  projection of a  $15\mu\text{m} \times 60\mu\text{m}$  stripe of an nc14 embryo labeled with *hb* probes. Lower: Average particle density was computed over a stack of 20 slices in a window of 15% EL in AP width for 4 embryos in nuclear cycle 12 (blue), 13 (green), early 14 (red) and late 14 (magenta). Dotted line shows an approximate depth of summation cylinders we use for our measure of local mRNA output: we use  $12\ \mu\text{m}$  from the surface in flattened (compressed) embryos, which roughly corresponds to  $15\ \mu\text{m}$  in the uncompressed mid-sagittal plane used for this figure; the compression factor is determined from the aspect ratio of the imaged nuclei that, up to early nc14, are approximately spherical prior to mounting. Fig. 2.2E is shown for comparison (inset). Apical spot density depends on the exact structure of nuclear layer and therefore on age within interphase (compare red and magenta traces). Setting the threshold at  $12\ \mu\text{m}$ , we eliminate this source of variation and obtain a measure of total expression that is largely insensitive to spatial reorganization of mRNA. A resulting source of error due to uncertainty of  $z$ -positioning (detection of embryo surface and the varying degree of embryo deformation) can be estimated from this depth distribution to 5% per  $\mu\text{m}$  of  $z$ -error (i.e. for these embryos, if the threshold is set at  $13\ \mu\text{m}$  instead of at  $12\ \mu\text{m}$ , 5% more spots would be counted). This is the single largest source of error in data points on Fig. 2.6. **B:** Cytoplasmic *hb* mRNA profiles for 15 embryos during interphase 13 (counts per standard volume versus position along AP axis). Smooth profiles were obtained using spline fits; individual data points (not shown) follow these profiles with  $8 \pm 2\%$  root-mean-square deviation. **C:** Cytoplasmic *hb* mRNA profiles of Fig. 2.3A plotted versus absolute AP position rather than relative, showing smoothed profiles during interphase 12 (blue), 13 (green), 14 early (red) and 14 late (magenta). Lines are best spline fits with 1, 2, 3 and 4 internal breakpoints, respectively. Error bars are root-mean-square deviations from smoothed profile, calculated in windows of 10 nuclei. **D:** Profiles of *hb* mRNA and Hb protein in an embryo simultaneously processed for FISH and immunofluorescence show a strong correlation (Spearman's correlation coeffi-



cient  $R = 0.975$ ). Note that dual labeling significantly impairs quantification performance. **E:** Amount of zygotically expressed *hb* mRNA as a function of time after egg deposition (AED) as determined by RT-qPCR relative to tubulin56D at various times after injection of embryos with  $\alpha$ -amanitin (two separate experiments, green and red lines) or buffer only (blue line). First time point at  $\approx 92$  min corresponds to injection time. Error bars are standard deviations across technical replicates. **F:** Results of two independent injection experiments were fit to a model of exponential decay to yield a *hb* mRNA lifetime  $\tau$  of about 60 minutes (mean and standard deviation indicated in figure are from non-linear least-squares fits using a restricted step method).

## Figure 2.S4: Detection and interpretation of transcription dynamics

**A:** An overlay of *hb* FISH data showing active sites of nascent transcription (green) and DAPI staining of DNA (blue) for four nuclei in a single embryo at interphase 13. Red

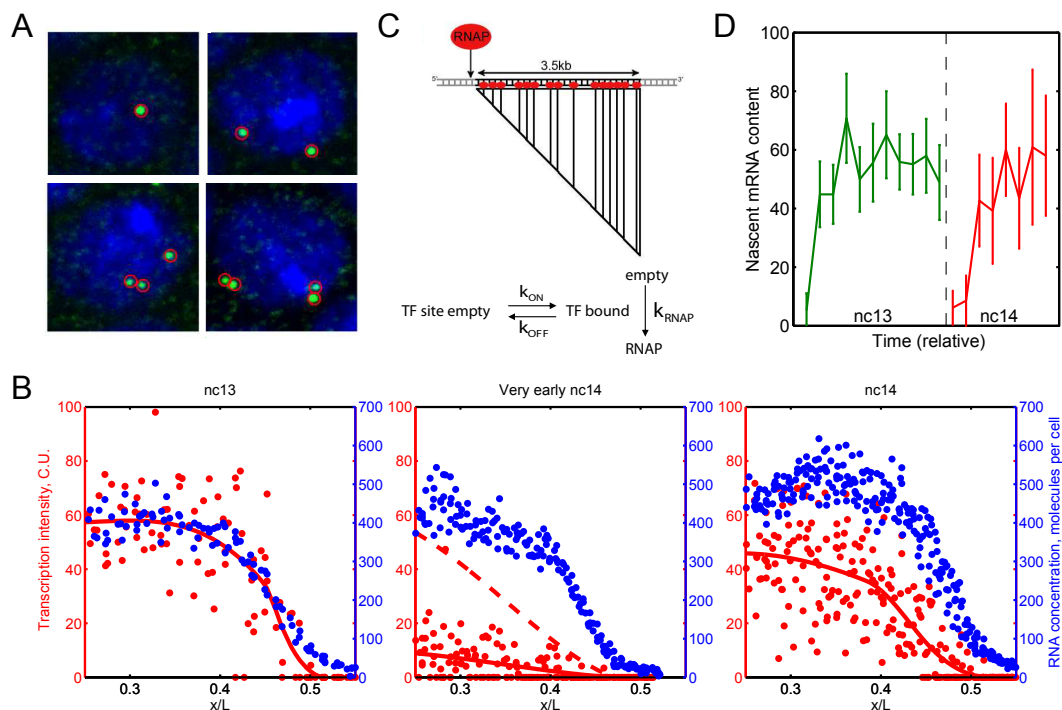


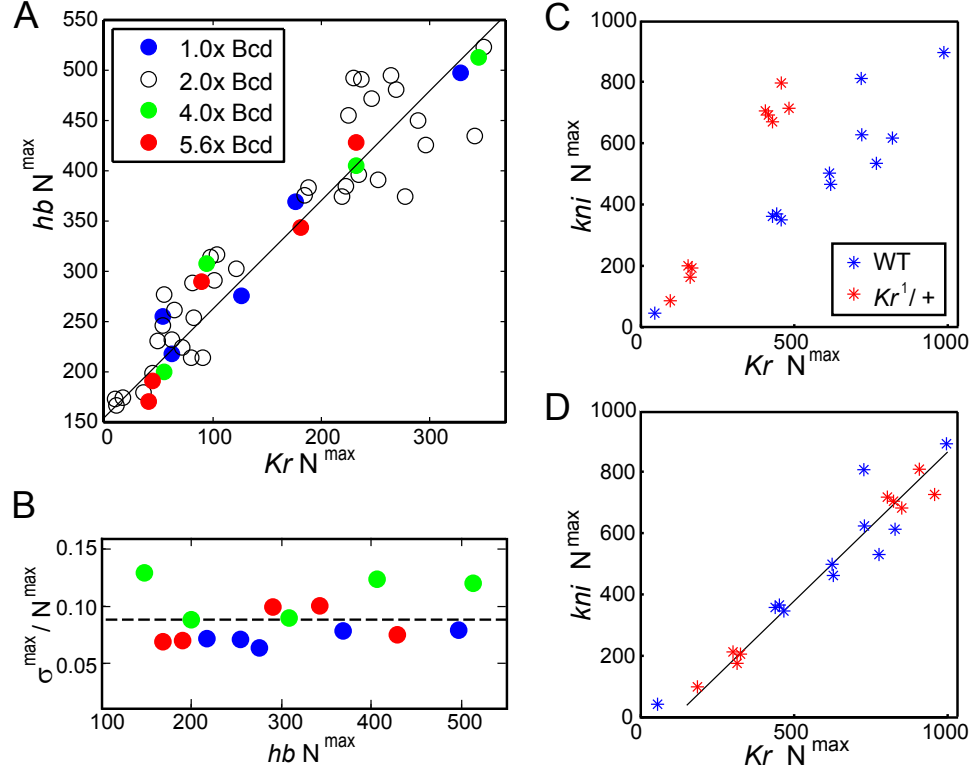
Figure 2.S4: Detection and interpretation of transcription dynamics.

circles indicate active transcription sites. Up to four transcription sites can be distinguished; however, resolving sister chromatids is only possible if they are well-separated and if this separation occurs in the lateral plane (due to impaired confocal resolution in the axial direction; see Fig. 2.S1D). **B:** Transcriptional activity per nucleus in cyto units (red) and cytoplasmic mRNA counts per standard volume (blue) for embryos of three similar ages: late interphase 13 (left), first minute of interphase 14 (middle), and early interphase 14 (right). Smoothed profiles are depicted as solid red lines. During interphase, the shape of transcription activity profile is very similar to that of cytoplasmic mRNA concentration (left and right panels). However, in approximately the first minute after completion of mitosis, anterior nuclei activate *hb* earlier than those located more posterior (middle panel). If the nascent mRNA content of active loci is dominated by the loading and progression of RNA polymerases along the length of the genomic template, then in the first minute of interphase we expect the nuclei located in the domain of maximal expression to exhibit a gradient of mean transcription activity, as indeed observed (middle panel). Red dashed line is the smoothed transcription profile multiplied by an arbitrary factor of 5 so as to visually highlight its distinction from the step-like shape observed on the other panels. Note the presence of transcriptionally silent nuclei along the entire length of AP axis on the middle panel. Their presence is a signature of a very early post-mitotic stage. Rare in the very anterior and more and more abundant towards the posterior, these nuclei shape the gradient of the averaged profile. **C:** In the simplest description of initiation-rate-limited transcription, a polymerase, once initiated, will travel the length of a gene (e.g.  $\approx 3.5$  kbp for *hb*) at some effective speed and produce a transcript. The statistics of initiation thus translate directly into the statistics of transcriptional output. Upper panel: at a given moment, an active transcription site will have a number  $N$  of actively transcribing polymerases distributed along its length (red dots), each carrying a partially finished transcript (vertical black lines) that is already capable of binding some of the fluorescent probes. For a particular transcription site, its fluorescent intensity in “cytoplasmic units” measures the total length of all unfinished transcripts in units of length of a complete transcript; this depends on the exact location of all polymerases and bound probes. On average, however, one finds

a very simple linear relation:  $F = \alpha N$ , where  $\alpha$  depends only on the location of probe binding sites on the transcript. If probes are homogeneously distributed along the length of the gene, unfinished transcripts form a triangle as shown and  $\alpha = 0.5$ . For a general arrangement, simple geometric considerations give  $0.5 < \alpha < 1$  if the probe distribution is biased toward 5' end, and  $\alpha < 0.5$  when biased toward 3'. Lower panel: two-state model of transcription. Transcription initiation can occur only if the promoter is in the “Transcription factor (TF) bound” state. The assembly of RNAP complex is treated as a process with a single rate-limiting step  $k_{\text{RNAP}}$ . In this model, at saturation of input ( $k_{\text{ON}} \times [\text{TF}] \gg k_{\text{OFF}}$ ), RNAP loading becomes a single-step process with Poisson statistics. Our results rule out this model, demonstrating that transcription noise is significantly super-Poissonian even at saturated input. **D:** Transcription activity as a function of time in embryos ordered by approximate age based on DAPI staining. Nascent mRNA content, which is low immediately following mitosis, attains a nearly constant level that is maintained for the majority of interphase.

## Figure 2.S5: Maximum gene expression rates are similar and independent of input genetic dosage

**A:** *hb* and *Kr* mRNA counts per standard volume were determined in the maximal expression region in embryos derived from females carrying 1, 4, or 6 genomic copies of *bcd*. The resulting Bcd protein content (shown in legend; Liu et al., 2013) ranges from 50% of WT (1.0x Bcd, blue) to 280% of WT (5.6x Bcd, red). Data from Fig. 2.6E (circles) shown for comparison. Time points span nc13 and early nc14. Although expression boundaries of *hb* and *Kr* shift along the AP axis as a function of Bcd dosage (not shown), production rates are unaltered in the domain of highest expression. **B:** Fractional standard deviation  $\sigma^{\text{max}}/N^{\text{max}}$  within the spatial domain of highest *hb* mRNA accumulation as a function of the mean count for the embryos shown in A. Expression noise across embryos in this experiment is  $9 \pm 2\%$ . **C, D:** *kni* mRNA expression versus *Kr* mRNA expression in embryos from nuclear cycles 13 and 14. mRNA expression is measured as mean absolute mRNA



**Figure 2.S5: Maximum gene expression rates are similar and independent of input genetic dosage.**

count per standard volume in the region of maximum expression of the respective gene. As in Fig. 2.6D-E, data for WT embryos (blue) coincides with data for embryos deficient in one chromosomal copy of *Kr* ( $Kr^1/+$ ; red) if *Kr* concentration in heterozygous embryos is multiplied by 2, demonstrating linearity of the final mRNA output in number of available loci. Raw data shown in (C), rescaled in (D).

## Figure 2.S6: Spatial and temporal averaging

**A:** *Kr* transcript density as a function of distance from the apical surface (depth). Nuclei inhabit the region surrounding a depth of  $\approx 5\mu\text{m}$ . Transcripts accumulate several  $\mu\text{m}$  basally from nuclei as development proceeds, indicating transcript mobility and the presence of spatial averaging. **B:** For a source with super-Poissonian noise statistics, with a standard deviation  $\sigma_{\text{nuc}}$  per  $N_0$  mRNA molecules produced, the maximum efficiency of noise reduction

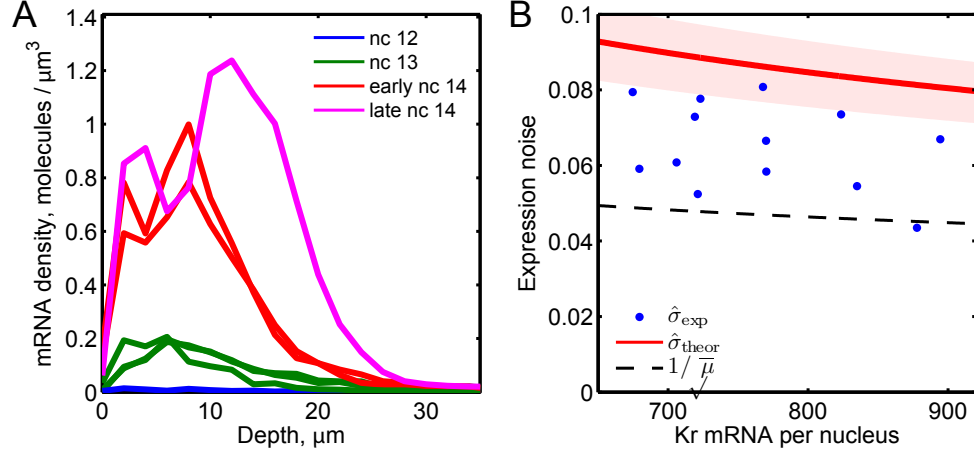


Figure 2.S6: **Spatial and temporal averaging.**

via temporal integration is achieved by a maximally uncorrelated process: independently running the process  $m$  times, we find that at mean  $\mu = mN_0$  the standard deviation of the accumulated output is given by  $\sigma_{\text{nuc}}\sqrt{m}$ . The predicted relation (red line) between the mean and the observed fractional noise of the mRNA profile is therefore bounded from below by  $\hat{\sigma}_{\text{cyto}} = \sqrt{\eta^2 + \hat{\sigma}_{\text{nuc}}^2 N_0 / \mu}$ , where  $N_0 = 100 \pm 20$  and  $\hat{\sigma}_{\text{nuc}} = \sigma_{\text{nuc}} / N_0 = 22 \pm 3\%$  are measured directly (nascent mRNA content and its standard deviation, respectively) and a conservative estimate for the measurement noise  $\eta$  is 3% (see Fig. 2.S2C). Shaded red area corresponds to uncertainty due to error in parameter estimation. The experimentally observed cytoplasmic expression noise as a function of Kr accumulation (blue points) shows that mRNA profiles exhibit better precision, suggesting the presence of spatial averaging. Poisson (counting noise;  $1/\sqrt{\mu}$ ) fluctuations are shown for comparison (dashed line).

## 2.E Control experiment demonstrating single-molecule resolution

Embryos in nuclear cycles 11-13 were processed with a set of FISH probes against *hb* mRNA, conjugated to fluorophores whose colors alternated along the transcript length. For cytoplasmic particles that were independently detected in both channels,

we constructed a two-dimensional histogram of their intensities as detected using both colors (Fig. 2.2F). If all cytoplasmic particles corresponded to a single mRNA transcript, the distribution of red ( $r$ ) and green ( $g$ ) particle intensities, expressed in cyto units, would be a two-dimensional Gaussian  $G_1(r, g)$  centered at  $(1, 1)$ , since probes conjugated to red and green fluorophores target different binding sites and binding events are, to a good approximation, independent. In practice, red and green measurements exhibit a weak correlation (Fig. 2.2F), which can be due to several reasons, one of which is correlation through mRNA content. Let  $p(n)$  be the probability for a given particle to contain  $n$  mRNA transcripts, then if we attribute all of the observed correlation on Fig. 2.2F to this source of correlation, we find that the intensity distribution function  $G(r, g)$  is given by a convolution of  $G_1(r, g)$  with  $F(r, g) = \sum_n p(n)\delta(r - n)\delta(g - n)$ . Since variances add under convolution, the variance of  $p(n)$  can be found as the difference between the variance of the cross-section of  $G(r, g)$  along the correlated direction (line  $r = g$ ) and the variance of  $G_1(r, g)$  which, in turn, is equal to the variance of  $G(r, g)$  along the anti-correlated direction (line  $r = 2 - g$ ).

We thus find an upper bound of 16% on the fractional standard deviation of  $p(n)$ . In an independent measurement using RT-qPCR (below), we show that an embryo contains, on average,  $1.2 \pm 0.5$  mRNA molecules per observed cytoplasmic particle. This constrains the mean of  $p(n)$ , and notwithstanding the large error bar due to imprecision of RT-qPCR, we can conclude that  $p(n)$  is concentrated on  $n = 1$  and  $n = 2$ , with higher contributions being negligible. As a consequence, the result  $\sigma[p(n)] < 16\%$  becomes highly constraining: denote  $p = p(1)$ , then

$$\frac{\sigma}{\mu} = \frac{\sqrt{p(1-p)}}{p + 2(1-p)} = 0.16,$$

which entails  $p = 0.97$ , or 97% of observed spots are single molecules.

## 2.F Noise level predicted by the two-state model of transcription

At saturation of transcription factor concentration, the two-state model of transcription (Fig. 2.S4C) reduces to a single-step description of transcription initiation. To a first approximation, this predicts a Poisson-distributed number of initiation events in a given time window. Using probes distributed evenly along the length of the mRNA, a mean transcriptional activity of about 50 C.U. in total nascent mRNA content (Fig. 2.4D) corresponds to 100 uniformly distributed RNAP across all loci. This is in agreement with a lower bound of a cytoplasmic accumulation rate of at least 500 transcripts during the 15 minutes of interphase nc13 (Fig. 2.3B), indicating a minimum of 33 transcripts made by each nucleus per minute. Consistent with previous estimates of RNAP processivity (Shermoen and Ofarrell, 1991) and with our measurements of mRNA lifetime, this requires at least  $\frac{33}{min} \times \frac{3.2kbp}{1.4kbpmin^{-1}} = 80$  engaged RNAP at a given instant. If RNAP numbers fluctuate according to a Poisson distribution, the predicted fractional error should be at most  $\sqrt{80}/80 \approx 11\%$ . As noted, the actual mean number of RNAP is higher than 80 per nucleus, but larger numbers lower the expected noise even further.

However, if the initiation rate is large enough that the density of bound RNAPs becomes comparable to the maximum attainable density of polymerases, the assumption of independent binding is no longer valid: there is a minimum time delay that must separate consecutive binding events, since the next RNAP can only bind after the previous one has cleared the landing site (RNAP crowding). To numerically study the consequence of this effect on the statistics of binding events, we performed Gillespie simulation of RNAP binding in various parameter regimes. (Note that the simulation has only two parameters, the mean output rate and the minimum time delay between binding events, and the first is fixed by our measurements.) We find

that, in general, RNAP crowding effect leads to a mild reduction of output noise. This can be intuitively understood as follows: at density close to maximal, the polymerases are forced to “walk” in tight synchrony, since no polymerase can overtake another. In the limit of low density, crowding becomes irrelevant and we obtain the simple Poisson regime. As we report in this work, for gap gene transcription, the average RNAP density is low enough to allow large ( $\approx 50\%$ ) fluctuations of polymerase load of individual sites. One would not, therefore, expect RNAP crowding to play a significant role, and indeed, simulations show that for RNAP exclusion footprint of 50-80bp, the reduction in noise level compared to Poisson statistics is from 11% (Poisson prediction) to 10%.

## 2.G Transcriptional activity of loci on sister chromatids

Pairs of loci representing sister chromatids were selected automatically in  $n = 4$  embryos, heterozygous for deficiency in *hb* (and consequently possessing at most two transcription sites per nucleus). The selection criterion was for the pair of spots to exceed a manually selected brightness threshold and be detected at least 4 pixels apart while still belonging to the same nucleus. The brightness threshold was chosen conservatively to exclude any possibility that a single cytoplasmic transcript could be mistaken for a transcription site. Automatically selected pairs were manually inspected for misdetections. Transcriptional activities of the two loci were measured using the difference-of-Gaussian intensity estimator (DoG; see Sec. 2.B.2) that is linearly related to the nascent mRNA content. The slope and offset of this linear relation are hard to estimate precisely (and so, when measuring transcriptional activity per nucleus, we use total fluorescence instead of the DoG estimator); however, correlation between variables are preserved by linear transformations, and consequently, the



analysis of correlation between the activities of sister loci can be performed directly on DoG values (Fig. 2.5B). As a control, we used  $n = 4$  embryos labeled using *hb* probes of alternating colors and compared the DoG intensities of all transcription sites that were detected in both channels. The tight correlation (Fig. 2.5D) demonstrates that the lack of correlation on the panel 2.5C cannot be attributed to the intrinsic noisiness of the DoG estimator itself, and indicates that the transcriptional activities of sister chromatids are indeed uncorrelated.

## 2.H Efficiency of temporal and spatial averaging

The RNAP processivity of 1.1-1.4 kbp/minute (Irvine et al., 1991; Shermoen and Ofarrell, 1991; Thummel et al., 1990) corresponds to a minimum time of 2.3 min for an RNAP to traverse the 3.2 kbp of the *hb* gene (the RNAP “correlation time”). A reasonable estimate of the time available for transcription during the thirteenth interphase is 15 min, i.e. this is the available “integration time” over which accumulation of transcripts may occur. With these values, noise reduction by temporal averaging alone may be estimated as  $\sqrt{15/2.3}$ , or a factor of  $\approx 2.6$ . The number of accumulated transcripts produced from one nucleus during nc13 can be found using starting and ending *hb* counts of approximately  $400 \pm 30$  and  $1000 \pm 70$ ; the noise in production is therefore  $\sqrt{70^2 - 30^2} = 63$ . Thus, the precision in production of  $63/600 \approx 11\%$  is similar to the reduction in nascent transcript noise by 2.6-fold, from 22% to about 9%. Therefore, according to this estimate, temporal averaging can easily provide the required noise filtering by allowing stable mRNA to accumulate while RNAP numbers fluctuate during the course of interphase.

A more careful estimate (see Fig. 2.S6), however, allows to obtain a theoretical bound on the maximum efficiency of temporal averaging based on the quantities we measure directly. In the case of *Kr* mRNA profile, it shows that by the time the

mean expression level reaches 800 molecules per nucleus, pure temporal averaging can at most reduce the expression noise to 8%. For these late embryos, however, our measurements show a consistently lower noise level of  $6 \pm 2\%$ . This subtle discrepancy can be accounted for by spatial averaging.

Let  $p$  be the probability that an mRNA produced by one nucleus is found within the volume assigned to a particular neighboring nucleus. In a hexagonal lattice, the total fraction of exchanged transcripts is  $6p$ . Considering that all exchange events are independent (and so the variances add), we find that spatial averaging reduces noise level  $\sigma$  to  $\sqrt{(1 - 6p)^2\sigma^2 + 6(p\sigma)^2} \equiv \sigma\chi$ , where  $\chi$  denotes the fold reduction in noise. The observed excess filtering compared to what can be achieved by temporal averaging corresponds to  $\chi = 8\%/6\% = 1.3$ . Solving for  $p$ , we find  $p = 0.04$ , which corresponds to just 32 transcripts exchanged between neighboring volumes during the entire development time during which 800 transcripts per nucleus were produced. Thus, even a limited degree of spatial averaging is completely sufficient to account for the appearance of low variation in cytoplasmic accumulation from stochastic transcription.

# Chapter 3

## Segment patterning in fruit fly: Theory

The experimental work presented in the previous chapter demonstrated that transcriptional process is characterized by a large contribution of intrinsic noise, and underlined the importance of spatial averaging in *Drosophila* embryo patterning. I see this as an example of a theory-driven experiment: reducing measurement errors to unprecedented 2-3% and achieving single-molecule resolution was a major investment that only made sense because of the theoretical motivation, and led to new insight. Now, completing the loop back to theory, these experimental results raise a major question. If, even in this system, transcriptional readout is noisy, then why have a patterning network made of multiple tiers? Each readout introduces new noise, so, intuitively, basing cell-fate decisions on genes that are regulated by a noisy version of a noisy version of the maternal input seems like a bad idea.

In this chapter, I show that the solution to this apparent paradox, again, lies in the diffusion-mediated non-locality of transcriptional-response. Non-locality and intrinsic noise can be naturally incorporated in a simple formalism, showing that optimizing decision-making accuracy is not the same as optimizing precision of ex-

pression, and the diffusion-mediated cell-to-cell “communication” can make a cascade of noisy readouts the optimal gradient response strategy. This work is unpublished.

### 3.1 Introduction:

#### Local channel picture is insufficient

The patterning problem naturally lends itself to an information-theoretic approach (Gregor et al., 2007; Tkacik et al., 2008), whereby nuclei are seen as acquiring “positional information” by measuring the local concentration of patterning cues. A precise input is in principle capable of generating a precise output, and so the patterning system is seen as a channel transmitting information at its input (concentration of a maternal patterning cue, or cues) into an acquired cell identity. The “positional information” can be precisely defined and carefully measured (Dubuis et al., 2013) as the mutual information between position along the antero-posterior (AP) axis of the embryo and the local concentration of morphogens  $I(x, \{c_i\})$ . Applied to the fruit fly, this approach has exhibited a number of intriguing features of the patterning system, such as homogeneity of information distribution along the AP axis or the dynamical nature of gap gene information content. It also provides an interesting new angle from which to consider mutants. However, the experimental results presented in the previous chapter exhibit two shortcomings of this picture.

First, the same amount of information can be packaged in ways that would make it more or less easy for the system to access. Consider a hypothetical morphogen (i.e. a protein serving as a patterning cue) whose concentration  $X$  in the embryo forms a perfect step:  $X = a$  in the anterior and  $X = b$  in the posterior. Imagine it is held constant and perfectly noiseless, and thus carries exactly 1 bit of information for any  $a$  and  $b$ . It is mathematically true that no (local) patterning network can extract more than 1 bit from such an input. It is clear, however, that for  $a/b \gg 1$  reading out this information is trivial, whereas for  $a/b = 1.01$  the task becomes very difficult, since the intrinsic noisiness of transcriptional readout makes it hard to detect such a small change in concentration. The difference between these two cases is, obviously, of

central importance for the embryo, and at each next tier of the patterning network, the expression domains are delimited by increasingly sharp boundaries, so that in the end distinguishing neighboring nuclei can be performed reliably. However, this distinction is missing from the current framework: we can measure the amount of information that is contained, mathematically, in a given morphogen profile, and ask whether it would be enough, *in principle*, to pattern the whole embryo, but we cannot ask how accessible it is or how the system would go about extracting it.

There is also a second problem: experiments suggest that the total local information contained in morphogen profiles may be increasing with time. Specifically, the joint information content of the first tier of patterning genes is significantly larger than that of the primary maternal input Bicoid (Dubuis et al., 2013). This is inconsistent with the picture of a local information channel, since channels obey a strict mathematical inequality: information can only be lost, but not gained, in transmission. One possible solution to this paradox is that the “missing” information could be provided by other inputs. But before we make that conclusion, we need to realize that there is a deeper reason why local information can indeed appear to be “created”. Nuclei do not function in isolation: in *Drosophila*, all of patterning occurs at the syncytial stage of development, i.e. nuclei are not separated by cellular membranes. The entire embryo remains a single cell with a number of nuclei doubling every 6-10 minutes; only at the end of the 13th division, when the blueprint is already in place, the membranes separating nuclei into individual cells are formed. Previous work (Gregor et al., 2007) has already pointed out that diffusion leads to a non-negligible exchange of protein products between nuclei, and we have additionally shown that this exchange also affects the mRNA, to the extent that it noticeably reduces expression noise, as described in the previous chapter. In other words, in the embryo, nuclei do not function as independent computational units; they are coupled into a tissue-wide structure, and this non-locality allows them to collectively make their decisions more

precise. This departure from the “local channel” view means that the performance of the system need not be limited by the local information content at any given time.

In this chapter, I show how non-locality and intrinsic noise can be naturally incorporated into a simple formalism leading to new experimentally verifiable predictions. I first introduce a simple model of transcriptional readout that captures the two essential properties discussed above: some amount of irreducible intrinsic noise and an element of non-locality. This allows me to formally define “accessible information” as the information contained in a signal after adding a noise component of fixed magnitude: the idea is that the embryo does not have access to true concentration of a transcription factor, but only to its noisy version seen through the prism of its noisy transcriptional machinery. I then address the problem of an apparent “creation” of local information through non-locality by showing that any circuit based on transcriptional readouts is mathematically equivalent to a purely local readout of a non-local quantity, namely the input averaged over a volume. As time progresses, this effective averaging volume increases; in other words, the patterning system “collects” information from a volume and channels it into a form that is locally accessible to individual nuclei. The amplification of a signal through transcriptional readout leads to a tradeoff: on the one hand, adding extra noise reduces the total amount of information in the signal, but on the other hand, the amplification of dynamic range increases the fraction of this amount that the system can access. I identify the regime in which this results in a net increase of accessible information, and show that for the conditions realized in the fly embryo, extra tiers of readouts can be beneficial.

For simplicity, in what follows I will use the terminology of the Bcd/Hb system, although the framework can be applied more generally.

## 3.2 Transcriptional readout in presence of averaging

In the *Drosophila* embryo, the control of one transcription factor (TF) concentration by another is a fundamentally non-local operation. Local transcriptional readout leads to the production of mRNAs that are exported into the cytoplasm, which is common for all nuclei. Translation happens in the cytoplasm and the newly-made protein is pumped back into the surrounding nuclei; in this manner, local concentration of Hb is necessarily a function of Bcd concentration in several neighboring nuclei. As shown in the previous chapter, the effect is further increased, since even mRNA mobility is non-negligible in this context, and not just that of protein. Finally, additional spatial averaging occurs when protein products are released into the cytoplasm during mitosis and reabsorbed by the newly-formed nuclei, with exchange between neighbors, again, inevitable.

### 3.2.1 Model for readout

Our model for transcriptional readout needs to include three key components. First, readout introduces some noise. Second, it is characterized by an input/output relation. Third, as I just described, it must involve some mixing between neighbors.

I will be describing various processes occurring within an embryo as operators acting on concentration vectors  $\vec{g} = \{g_i | i \in \text{Nuclei}\}$ . Here  $g_i$  is the local concentration of substance  $g$  (protein or mRNA) in or at nucleus  $i$ . I will use a graphical notation for operators, representing them as circuit elements (boxes of some shape) connected by lines representing the concentration vectors they act on. For instance, the transcriptional readout is an operator  $\vec{g}^{\text{out}} = R \vec{g}^{\text{in}}$  (Fig. 3.1).

I will model the readout as the following sequence of steps:

1. take input: transcription factor concentration in each nucleus  $g_i^{\text{in}}$ ;





Figure 3.1: The readout operator.

2. add noise of some fixed relative magnitude  $\eta_0$ : this is the transcription factor concentration measured by transcription sites;
3. apply a deterministic input/output function  $F$  to obtain the amount of target protein synthesized in the cytoplasm from mRNAs made in nucleus  $i$ ;
4. finally, to obtain the new local concentration of the target protein  $g_i^{\text{out}}$ , apply spatial averaging, modeled as a convolution with a Gaussian kernel of width  $\sigma_r$  ( $r$  for “readout”). This implements the unavoidable cross-talk between neighboring nuclei as described above.

Putting this together, mathematically, we have

$$\vec{g}^{\text{out}} = G_{\sigma_r}(\{F(g_i^{\text{in}}(1 + \xi_i))\}),$$

where  $\xi_i$  are i.i.d. drawn from a Gaussian distribution of width  $\eta_0$ , and  $G_{\sigma_r}$  denotes the spatial averaging operator. From now on I will only use pictures, and represent such expressions as on Fig. 3.2.

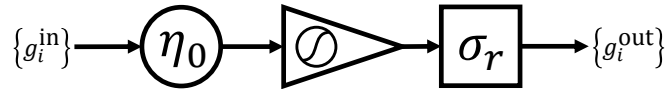


Figure 3.2: The model of the readout: add noise (circle), apply a deterministic input-output function (triangle), apply spatial averaging (square).

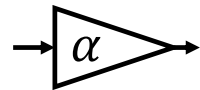
Note that of the three basic operations on this figure, the only non-local one is the the spatial averaging operator (the square). We can think of Fig. 3.2 as placing

one transcription site under the control of a TF and accumulating the corresponding protein output for some fixed amount of time, which I will denote  $\tau_0$ .

A fixed-magnitude input noise is a very simplistic model. The extensive literature on noise in biochemical circuits has investigated much more realistic scenarios, and even in the specific context of *Drosophila* patterning network, the effect of different sources of noise on information transmission through the network has been studied in great detail (Tkacik et al., 2009). My goal here, however, is to investigate the interplay between noise and non-locality. Given that transcriptional readout is intrinsically noisy, does this have any general implications for the patterning strategy? For this purpose, the simplest model of a fixed-magnitude input noise will suffice.

### 3.2.2 The linear approximation

Specifically, I consider the problem of forming an expression boundary positioned at  $x_0$  (an expression profile where the concentration is low for  $x < x_0 - \delta x$  and high for  $x > x_0 + \delta x$ ). I place myself at the center of this boundary  $x_0$  and limit my consideration to a small region around it. In this region, the input/output function is linear with some slope  $\alpha > 0$  (for example, for a Hill function with coefficient  $n_H$ , we have  $\alpha = n_H/2$ ). To stress that I will now be working in this linear approximation, I will from now on denote the i/o function operator as linear amplification with coefficient  $\alpha$ , as presented on Fig. 3.3. (I assume that morphogen profiles span the range  $[0, g_{\max}]$ .)



$$F_\alpha: g \mapsto \frac{g_{\max}}{2} + \alpha(g - g(x_0))$$

Figure 3.3: The linear approximation of input/output function

In this approximation, the transcriptional readout is a composition of three linear operators. Conveniently, linear operators commute (subject to appropriate re-scaling

of noise); in particular, we can bring the spatial averaging operator all the way to the left (Fig. 3.4). In other words, although the transcriptional readout is an intrinsically

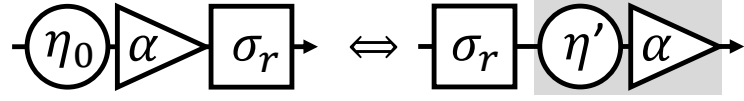


Figure 3.4: Recasting the non-local readout as a local measurement (shaded) of a non-local quantity

non-local operation, it is equivalent to *local* readout (with a different noise) of a non-local quantity, namely the Gaussian-averaged version of the transcription factor. Taking the example of the Bcd/Hb system, the local concentration of Hb in a nucleus is the Gaussian average of noisy readouts of Bcd in several neighboring nuclei. But if I prefer to think in local terms, then I can invert the order of operations and equivalently write local Hb as a (less noisy) function of the local value of the Gaussian-averaged Bcd profile. It follows that modeling transcriptional regulation by local equations is an approximation and needs to be used with caution: although it can provide an accurate description of the mean expression profile, the propagation of noise or expression fluctuation statistics are necessarily affected by the non-locality of the readout.

In reality, when Hb profile is established as a readout of Bcd, the process is much more complicated than the single readout diagram presented in Fig. 3.4. Time averaging during a single nuclear cycle can be seen as performing multiple single readouts and adding the result. Each mitosis applies additional spatial averaging operators. The full diagram that describes the Hb protein boundary at the start of nuclear cycle 14 would be described by a complicated circuit looking somewhat like Fig. 3.5.

This figure is still a simplification; most notably, it ignores the fact that the number of nuclei changes with each nuclear cycle. It is clear, however, that for an arbitrary circuit of this type, the explicit commutation rules for the basic operators

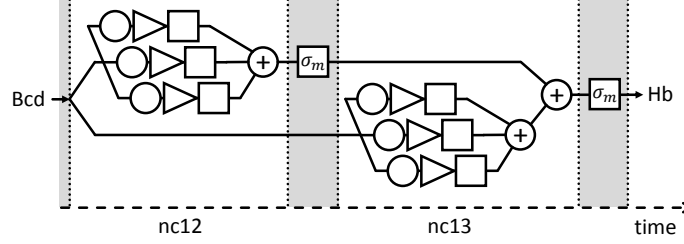


Figure 3.5: Hb boundary is formed over multiple nuclear cycles. The diagram shown describes the fate of Hb protein synthesized during nc12 and nc13. Empty shapes represent the standard operators as on Fig. 3.2; the “ $\oplus$ ” operator stands for summation and  $\sigma_m$  represents spatial averaging during mitosis (shaded).

(see Sec. 3.A) allow me to commute all the spatial averaging operators to the left of the diagram as a straightforward generalization of the calculation presented on Fig. 3.4. Therefore, an arbitrary circuit diagram can be reduced to the same basic circuit (Fig. 3.6). At any point in time, Hb concentration in the vicinity of its boundary

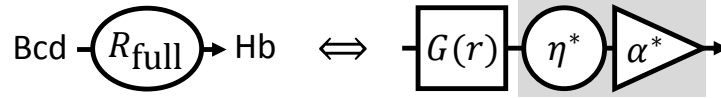


Figure 3.6: Arbitrary circuits can be reduced to the canonical form: a local readout (shaded) of a non-local quantity, obtained by a convolution of the input with a general kernel  $G(r)$  (not necessarily Gaussian).

can be expressed as a noisy, but purely local function of the Bcd input concentration convolved with a spatially extended kernel. From Fig. 3.5, it is clear that the noise of the function decreases and the width of the averaging kernel increases with each nuclear cycle (in the general case, this kernel is a sum of Gaussians, with one term for each contributing nuclear cycle). As time goes on, our simple patterning system (here, just one gene reading out the maternal input) converts information gathered from an increasingly large volume into a quantity that is accessible locally (the target protein concentration).

It is clear that the argument can be generalized to extra readout layers (e.g., adding a pair-rule gene reading Hb). But before we do this, we need to ask ourselves: why have multiple tiers of transcriptional readout at all?

### 3.3 Optimal amplification

#### 3.3.1 The benefits of amplification

Every readout introduces noise, so some information must be lost. Yet in the patterning system of the fly embryo, the segment polarity genes read pair-rule genes, which were established by reading out gap genes, whose concentration had in turn been determined by reading out Bcd. Naïvely, then, the information present in the maternal gradient appears to be used suboptimally. To understand whether this is indeed the case, let us first consider the problem of distinguishing just two neighboring nuclei by their local concentration of a morphogen  $c$ . Mathematically, a simplified version of this problem can be formulated as follows: we have a generator of random variables  $c_i$  that draws values from a Gaussian distribution of width  $\sigma$  and mean located either at  $+\Delta\mu/2$  or  $-\Delta\mu/2$ . How hard is it to distinguish which of the two distributions the generator is using? Biologically,  $\Delta\mu$  is the expected difference in protein concentration between neighbors, and  $\sigma$  is the expression noise. This formulation simplifies the problem because we assume that the expected mean is one of two known values; this means we ignore reproducibility issues and only consider the problem of finite precision.

There are two approaches to this problem. First, I can think of this in intuitive terms: how many samples (readouts) do I need to reliably determine from which of the two distributions I am drawing? Alternatively, I can directly calculate the mutual information between the variable  $c$  and the binary variable describing from which of the two distributions  $c$  is being drawn. Both of these approaches are useful.

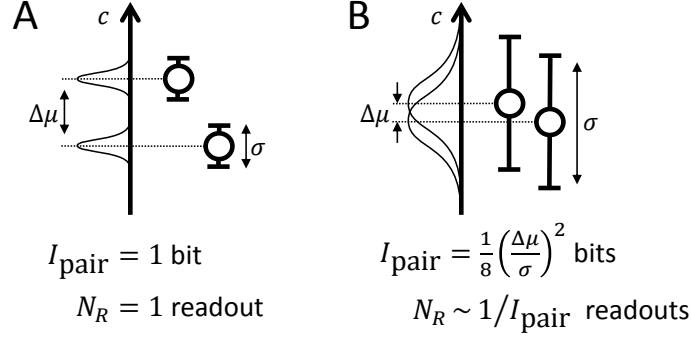


Figure 3.7: Using morphogen  $c$  to discriminate between neighboring nuclei: the information  $I_{\text{pair}}$  carried by the morphogen, or equivalently, the required number of readouts  $N_R$ , is determined by the ratio  $\sigma/\Delta\mu$ , where  $\Delta\mu$  is the expected concentration difference and  $\sigma$  is the standard deviation of fluctuations seen by the nuclei. **A:** low-noise regime, **B:** high-noise regime.

To determine the number of readouts required to reliably distinguish the neighbors, we note that after  $N$  independent measurements, the standard error of the mean is  $\sigma/\sqrt{N}$ . Therefore,  $N$  must be chosen to satisfy

$$C \frac{\sigma}{\sqrt{N}} < \frac{\Delta\mu}{2},$$

where  $C$  is set by the probability of mistake that we are willing to tolerate – for example,  $C = 2$  for approximately 98% reliability (a mistake is a “ $2\sigma$ -event”). Therefore, for given  $\Delta\mu$  and  $\sigma$ , the number of readouts  $N_R$  required to distinguish the neighbors goes as

$$N_R \propto \left( \frac{\sigma}{\Delta\mu} \right)^2.$$

Alternatively, let me calculate the mutual information  $I(c, a)$ , where  $a$  is a binary variable describing whether I am drawing  $c$  from the first or the second distribution ( $a \in \{1, 2\}$ ).

$$I(a, c) = S[p(a, c)] - \langle S[p(c|a)] \rangle_a$$

The joint distribution is a sum of two Gaussians displaced by  $\Delta\mu$  relative to each other, whereas the conditional distribution is a single Gaussian of width  $\sigma$ :

$$S\left(\frac{G(-\delta\mu/2, \sigma) + G(\delta\mu/2, \sigma)}{2}\right) - S(G(0, \sigma)).$$

In the small-noise limit, the two peaks in the double-Gaussian distribution are well separated and the mutual information tends to 1 bit as expected. We are more interested in the large-noise limit; for  $\Delta\mu/\sigma \ll 1$  the double Gaussian is approximately a single, slightly wider Gaussian, of variance  $\sigma^2 + (\Delta\mu/2)^2$ , and since  $S(G(\mu, \sigma)) = \frac{1}{2} \log(2\pi e\sigma^2)$ , we find

$$I_{\text{pair}} \equiv I(a, c) = \frac{1}{8} \frac{\Delta\mu^2}{\sigma^2} + o((\Delta\mu/\sigma)^2).$$

As expected, in the large-noise limit,  $I_{\text{pair}} \propto 1/N_R$ .

Importantly, in our model, the nuclei have to make their decisions based on a noisy version of the input; in other words, the variance in the denominator includes both the variance of the input  $\xi^2$  and the added variance of the readout noise  $(\mu\eta_0)^2$ :

$$I_{\text{pair}} = \frac{1}{8} \frac{\Delta\mu^2}{\xi^2 + \mu^2\eta_0^2}.$$

Here  $\mu$  is the mean magnitude of the input (remember that  $\eta_0$  is a fixed fractional noise). However precise the input signal may be, a single measurement can only convey a finite amount of information  $I_{\text{pair}} \leq \frac{1}{8} \frac{\Delta\mu^2}{\mu^2\eta_0^2}$ : the information contained in the signal may be infinite, mathematically, but the accessible information is bounded. The purpose of the patterning system is to encode positional information in a manner that can be accessed with a single local readout so each nucleus can activate appropriate developmental programs. If the “dynamic range”  $\Delta\mu$  of the patterning cue is too

small, this goal cannot be achieved no matter how efficiently noise is being filtered, and amplification steps are required.

How does amplification affect the amount of information accessible to the system? Applying the readout operation as defined on Fig. 3.2 to a morphogen with mean  $\mu$ , variance  $\xi^2$  and mean concentration difference between neighbors  $\Delta\mu$  transforms these parameters in the following way:

$$\begin{aligned}\mu &\mapsto \mu \\ \Delta\mu &\mapsto \Delta\mu' = \alpha\Delta\mu \\ \xi^2 &\mapsto \xi'^2 = \alpha^2 \frac{\xi^2 + \mu^2 \eta_0^2}{N_{\text{eff}}}.\end{aligned}\tag{3.1}$$

The mean remains unaffected (by the definition of our input/output function), dynamic range is amplified, and the variance is transformed in a way that can be derived with the commutation rules (see 3.A), but is in fact intuitive: the variance of the morphogen as measured by transcribing nuclei is  $\xi^2 + \mu^2 \eta_0^2$ , and this is faithfully amplified by the linear input/output function. However, spatial and temporal averaging reduce this variance by a factor  $1/N_{\text{eff}}$ , where  $N_{\text{eff}}$  is the effective number of measurements (proportional to integration time and to  $\sigma_r^2$ ).

How does our ability to discriminate between neighbors change under this transformation?

$$I_{\text{pair}} = \frac{1}{8} \frac{\Delta\mu^2}{\xi^2 + \eta_0^2 \mu^2} \mapsto I_{\text{pair}}^{(\alpha)} = \frac{1}{8} \frac{\Delta\mu^2}{\frac{\xi^2 + \mu^2 \eta_0^2}{N_{\text{eff}}} + \frac{\mu^2 \eta_0^2}{\alpha^2}}.$$

Note that amplification affects both  $\Delta\mu$  and  $\xi^2$  and so what increases information is not the naive dynamic range increase. Rather, the dynamic range amplification reduces the relative importance of measurement noise compared to the input noise ( $\alpha^2$  in the denominator of the second term).



Note that if  $N_{\text{eff}} = 1$ , we find

$$I_{\text{pair}}^{(\alpha)} \Big|_{N_{\text{eff}}=1} = \frac{\Delta\mu^2}{\xi^2 + \mu^2\eta_0^2 + \frac{\mu^2\eta_0^2}{\alpha^2}} < I_{\text{pair}}.$$

This says that for a purely local channel (no spatial averaging), readout always leads to loss of information. This makes perfect sense, because applying a noisy function to a signal degrades its information content. However, for  $N_{\text{eff}} > 1$  (which is necessarily the case for the embryo) an amplifying transcriptional readout can lead to an increase in accessible information.

### 3.3.2 Patterning capacity

Is there such a thing as optimal amplification? It seems that we have a problem, because  $I_{\text{pair}}^{(\alpha)}$  is an increasing function of  $\alpha$ : maximizing  $I_{\text{pair}}^{(\alpha)}$  calls for an infinite amplification. But this is no surprise: if our goal were to separate a given pair of neighbors, i.e. place a single boundary at a location defined by morphogen concentration  $c_0$ , the best solution is a step-like input/output function with a discontinuous transition at precisely  $c_0$ .

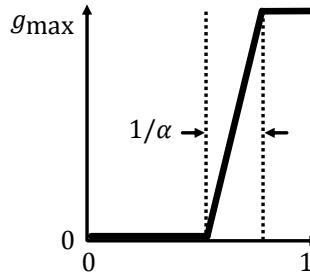


Figure 3.8: Profile of a patterning gene (“Hb”) after a linear amplifying readout of the initial gradient (“Bcd”) with dynamic range restriction.

Our goal, however, is *not* to place a single boundary, but to pattern the entire embryo. A sharp transition is necessarily narrow, and better resolution comes at a price of covering fewer nuclei. The width of the transition region goes as  $1/\alpha$

(Fig. 3.8), and the smaller this region, the more genes are required to establish a sufficient number of boundaries. The evolutionary cost of using an extra gene appears impossible to quantify, and exactly what compromise is “best” for the embryo at the gap gene stage is anything but obvious. This is the general caveat of framing a biophysical question as an optimization problem. To make progress, let me postulate a plausible definition of “patterning capacity”  $C$  of a morphogen by multiplying the two quantities that should both be large if few gap genes are to be used, namely  $I_{\text{pair}}^{(\alpha)}$  and the width of the patterned region  $1/\alpha$ :

$$C = \frac{1}{\alpha} I_{\text{pair}}^{(\alpha)}.$$

I will first investigate the properties of  $C$ , and then return to the question of its definition.

### 3.3.3 Optimal amplification

The patterning capacity as defined above is given by

$$C^{(\alpha)} = \frac{1}{\alpha} I_{\text{pair}}^{(\alpha)} \propto \frac{\Delta \mu^2}{\alpha \frac{\xi^2 + \mu^2 \eta_0^2}{N_{\text{eff}}} + \frac{1}{\alpha} \mu^2 \eta_0^2}.$$

This function has a maximum, and the optimal amplification value is given by:

$$\alpha_{\text{opt}} = \max \left( 1, \sqrt{\frac{\mu^2 \eta_0^2 N_{\text{eff}}}{\xi^2 + \mu^2 \eta_0^2}} \right)$$

(recall that all the previous formulas assumed  $\alpha > 1$ ).

Let us examine this result. We find that amplification only serves a purpose ( $\alpha_{\text{opt}} > 1$ ) if

$$N_{\text{eff}} - 1 > \frac{\xi^2}{\mu^2 \eta_0^2}. \quad (3.2)$$

There are two ways how this condition may fail:

- When readout is a purely local operation ( $N_{\text{eff}} = 1$ ), it only makes matters worse (degrades information content at each readout step).
- For  $N_{\text{eff}} > 1$ , if the readout noise is low enough, we can use the original morphogen to directly pattern the whole embryo, whereas an amplifying readout ( $\alpha > 1$ ) will restrict patterning to only a fraction of the length of the embryo (of size  $1/\alpha$ ; Fig. 3.8).

Assuming (3.2) is satisfied, we find that the optimal amplification value for gap gene readout

$$\alpha_{\text{opt}} = \sqrt{\frac{N_{\text{eff}}}{1 + \frac{\xi^2}{\mu^2 \eta_0^2}}} \quad (3.3)$$

can be expressed in terms of noise properties and  $N_{\text{eff}}$  that encodes spatial averaging magnitude, all measurable in principle. Our experimental work on transcriptional noise shows that  $\frac{\xi}{\mu \eta_0}$  is of order 1, and so

$$\alpha_{\text{opt}} \approx \sqrt{N_{\text{eff}}/2}. \quad (3.4)$$

This relation constitutes a prediction relating the Hill coefficient of Hb readout and the number of effective measurements of Bcd that the Hb readout system has time to perform ( $N_{\text{eff}}$  quantifies the amount of temporal and spatial averaging). For Hb readout, the Hill coefficient is approximately  $n_H = 5$  (Gregor et al., 2007), i.e.  $\alpha \approx 5/2$ , which would correspond to  $N_{\text{eff}} \approx 12$ . This should be compared to our expectation that the spatial averaging kernel must extend to the six direct neighbors but not much further, and that temporal averaging over a few nuclear cycles will contribute an additional factor. This order-of-magnitude agreement is encouraging.

### 3.3.4 Reexamining the definition of patterning capacity

Whether the prediction of eq. (3.4) should be verified more carefully on experimental data depends on our degree of faith in the optimization framework proposed here, and this hinges on our *ad hoc* definition of patterning capacity. This definition was not based on first principles and postulates a particular functional form of the tradeoff faced by the system. Although the conclusions are plausible, without further justification the relation (3.3) should not be over-interpreted. Nevertheless, the argument above provides useful new insight into the functioning of the patterning system: Incorporating noise and non-locality into our description suggests that, rather than continuously refining and improving the pattern, the network is in fact sacrificing precision of later tiers in exchange for an amplified dynamic range, which is necessary for unambiguous readout by downstream processes.

## 3.4 Conclusion

The framework I described explores the implications of two general features of transcriptional readout, noise and non-locality. How specific is all this to the fruit fly? Syncytial development is a unique property of *Drosophila* among all model organisms. However, we have reasons to believe that intrinsic noise could be an inherent feature of eukaryotic transcriptional machinery in general, so we can expect that responding to patterning cues reliably is a problem faced by all developing organisms. Further, gradient response circuits frequently involve a weakly diffusible molecule, or an intermediate that cells can exchange (e.g. Dpp and EGF signaling). This suggests that the same framework can apply to morphogenesis more generally. It would also be intriguing to look for the same strategy employed in a different context of collective decision-making, e.g. collective sensing of chemical gradients by bacteria.

It is important to remember that I have been discussing precision only, leaving the question of reproducibility aside. The natural next step is to investigate how cell-to-cell communication enables the patterning network to perform system-wide error correction. Although later tiers of patterning may sacrifice some precision, the *reproducibility* of domain boundaries specified by patterning proteins is known to increase during development time. One can experimentally modify the maternal input in ways that should cause global shifts in the final pattern and make various body parts of the adult larger or smaller at the expense of others. However, many such changes get corrected during development and the adult flies have normal body proportions. The mechanism of this error-correction is not understood. From a theoretical standpoint, this means that the tissue-wide network of cells is capable of “sensing” its boundaries, and this is somehow converted into positional information propagating into the tissue. We do not have a framework for understanding such propagation. The fruit fly provides an excellent motivation to begin developing such understanding in toy model scenarios, and can eventually be used for experimental verification of predictions.

# Technical details

## 3.A Commutation relations

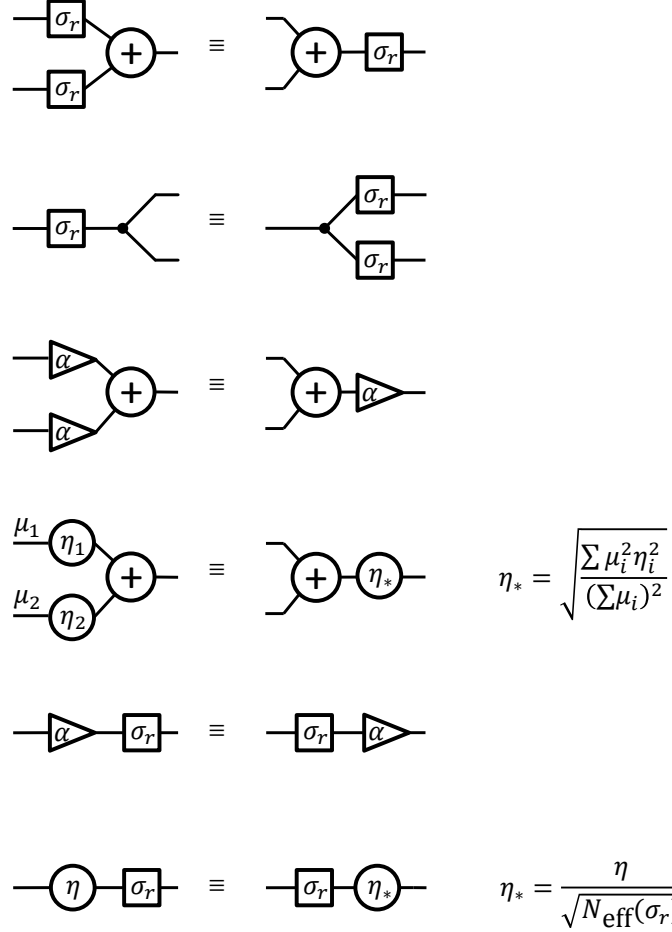


Figure 3.S1: The commutation relations used to derive the transformation rules (3.1). For the fourth relation,  $\mu_1$  and  $\mu_2$  denote the mean magnitude of the signal that can vary across inputs and affects the transformation of the noise term.  $N_{\text{eff}}(\sigma_r)$  is the effective number of readouts corresponding to the spatial averaging radius  $\sigma_r$ . If  $f_i$  denote the convolution weights ( $\sum_i f_i = 1$ ) with which different nuclei contribute to the average, then I define  $N_{\text{eff}} = (\sum_i f_i)^2 / (\sum_i f_i^2)$ . Note that  $N_{\text{eff}}(\sigma_r)$  scales as  $\sigma_r^2$ .

# Chapter 4

## Genetic networks: going beyond topology

Genetic regulatory networks are defined by their topology and by a multitude of continuously adjustable parameters, such as the strength of interactions between nodes. Here I present a class of simple perceptron-based Boolean models within which the relative importance of topology *vs.* interaction strengths becomes a well-posed problem. I find that optimizing interaction strengths is a better strategy of achieving high complexity, defined as the number of fixed points the network can accommodate, and discuss the possible implications for real networks and their evolution.

The work presented in this chapter is the subject of a publication currently in review: Tikhonov M & Bialek W, “Complexity in generic biochemical circuits: topology versus strength of interactions”.



## 4.1 Introduction

The state of a living cell is determined largely by the concentrations of various proteins. But the instructions for making the proteins are encoded in DNA, and the rates at which this information is read out are determined, in part, by the concentration of other proteins. Thus, there is a network of interactions in which genes encode proteins, and proteins control the reading-out of the genes. Such “genetic regulatory networks” are not the whole story of how cells control their states, but this is a good starting point, and the image of cellular states as the states of an interacting network has shaped quite a bit of thinking about cellular function (see Levine and Davidson, 2005, , which is an introduction to a series of articles on this topic).

During embryonic development of multicellular organisms, the states of the relevant genetic networks are thought to encode the body plan of the adult (Gerhart and Kirschner, 1997; Lawrence, 1992), and so the ability of the network to adopt a richer set of states corresponds to building a more complex organism. What is it about the network that controls this complexity? How do the changes in DNA sequence allow the emergence of greater complexity over the course of evolutionary history (Carroll, 2005)?

Much of what we know about the structure of genetic networks comes from classical genetics—experiments when a mutation knocks out one element of the network. This leads to information about network topology: the protein encoded by gene A represses the read-out of gene C, activates the read-out of gene D, and does nothing to genes B and E. It is much more difficult to measure the strength of these interactions. Perhaps because of this experimental situation, there has been a considerable focus on network topology as a determinant of biological function.

Ideas about network topology include several themes. One approach aims at a statistical characterization of topologies, focusing on the number of connections to a single node (Barabasi and Oltvai, 2004) and their relation to node centrality or

essentiality (Navlakha et al., 2014), or on the presence of local motifs in which patterns of connections among small numbers of genes are over-represented (Lee et al., 2002; Milo et al., 2002). Another idea is that relatively small changes in DNA sequence in the regions where proteins bind and regulate the expression of genes can change the effective topology of the network, and thus there is a path for topology to evolve quickly.<sup>1</sup> Finally there is the idea that the difficulty of defining interactions strengths is a problem not only for us but for the cell itself, and hence that important cellular functions must be “robust” against variations in these parameters<sup>2</sup>; taken literally, this means that function should be encoded in topology alone.

It has been repeatedly observed that the topology of real biological circuits is indeed non-generic (e.g., Goentoro et al., 2009; Lee et al., 2002; Milo et al., 2002; Tyson et al., 2003, and references therein). Nevertheless, reducing the description of a network to its topology appears unacceptably coarse. Depending on the quantitative parameters, a system consisting of just two interacting chemical components can exhibit complex behaviors such as bistable switches or pattern-forming instabilities, and three can already be chaotic (Bintu et al., 2005). There are known cases when certain features of network behavior may have a purely topological origin (Shinar and Feinberg, 2010), but these results attracted significant attention precisely because they constitute a surprising exception rather than the general rule. Examples where interaction strengths in a network can play the determining role on its global properties include, for example, the work on synchronization properties of coupled oscillator networks (Arenas et al., 2008), bistability in chemical reaction networks (Feinberg, 1987), the spin-glass literature (Mezard et al., 1987), or the whole field of machine learning, based entirely on the fact that by adjusting the interactions in a

---

<sup>1</sup>Early versions of these ideas can be found in (King and Wilson, 1975; Zuckerkandl and Pauling, 1965); see also Carroll (2005).

<sup>2</sup>For an example of these ideas in the context of development, see von Dassow et al. (2000). For a review of this and related ideas about fine tuning vs. robustness in biological networks, see Bialek (Chapter 5 in 2012).

network with a fixed architecture, one can “train” it to perform staggeringly different and complex tasks. In the biological context, the role played by quantitative parameters has been highlighted, for example, in ensuring circuit multi-stability (Bagowski et al., 2003) or in shaping the behavior of excitable systems (Rue and Garcia-Ojalvo, 2011). For these reasons, although in the genetic network literature the focus on network topology is widespread, it seems clear that understanding these systems will require at least some level of quantitative detail. Rather than being irrelevant, parameters could be optimized to transmit the maximum amount of information through a network (Tkacik et al., 2008), to achieve the maximum signal-to-noise ratio for weak signals (Andrews et al., 2006), or to ensure that events occur in a precisely timed sequence (Ronen et al., 2002).

In this chapter, my goal is to define a class of models within which the relative contributions of topology and parameters to the complexity of a network can be dissected. To make their comparison a well-posed problem, I have to address two challenges. First, I have to say what I mean by complexity. Second, I have to define a measure on the space of interaction strengths. The claim, for example, that “typical” parameter values lead to certain behaviors depends on the shape of the distribution over parameter space. Since interactions are determined by the rates and equilibrium constants of biochemical reactions (e.g., binding of a protein to a site along the DNA (Bintu et al., 2005)), these parameters are continuous and exponentially sensitive to perhaps more natural parameters such as binding energies; the question of what constitutes a natural measure on such a space is not trivial.

Here I introduce a highly simplified model in which these issues have a natural formulation, and in particular where the continuous parameter space breaks into discrete sectors with equal weight, so that varying parameters and varying topology both become matters of enumeration. Within this model we will see that complexity

is dominated by the choice of parameters, and that it is very difficult to evolve greater complexity by changing topology without optimizing parameters.

## 4.2 The model

### 4.2.1 Definitions

To make a simplified model, I imagine that every gene  $i$  has a binary state,  $s_i = \pm 1$ , where  $s_i = +1$  indicates that protein encoded by gene  $i$  is being synthesized, and thus is present at a high concentration, while  $s_i = -1$  indicates that this protein is at near-zero concentration. The state is determined by inputs from other proteins, and I assume that these inputs add; the gene is “on” if the total input exceeds a threshold:

$$s_i = \text{sgn} \left[ \sum_j \hat{J}_{ij} s_j - H_i \right]. \quad (4.1)$$

The matrix  $\hat{J}$  encodes both the topology of the network and strength of the interactions. To separate these I write  $\hat{J}_{ij} = J_{ij} T_{ij}$ , where the elements of the topology matrix  $T_{ij}$  are assigned values  $+1$ ,  $-1$ , or  $0$  depending on whether the protein encoded by gene  $j$  activates, represses, or does nothing to gene  $i$  (Fig. 4.1). The interaction matrix  $J_{ij}$  then can be assigned all positive elements. In a similar spirit, I write  $H_i = c_i h_i$ , where  $c_i = \pm 1$  and  $h_i \geq 0$ ;  $c_i = +1$  means that gene  $i$  would be “on” in the absence of inputs (“constitutively active,” in biological terms), and conversely for  $c_i = -1$ . Thus, Eq. (4.1) becomes

$$s_i = \text{sgn} \left[ \sum_j T_{ij} J_{ij} s_j - c_i h_i \right]. \quad (4.2)$$

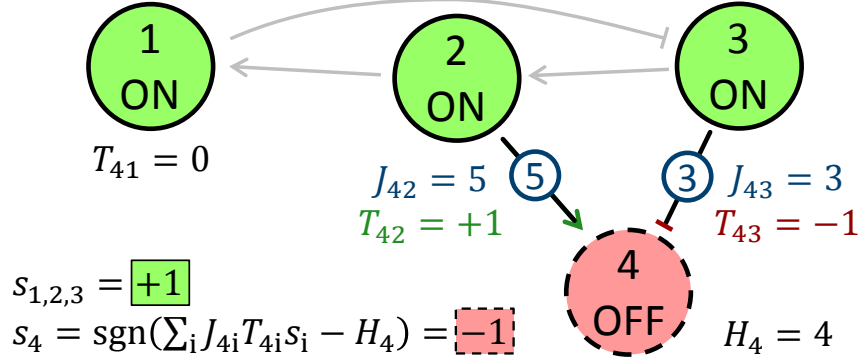


Figure 4.1: Node 4 is activated by node 2 with strength 5 and repressed by node 3 (strength 3). It is constitutively repressed with strength  $H_4=4$ . The regulatory rule dictates node 4 to remain “off.”

For technical reasons I will set  $T_{ii} = 0$ , forbidding explicit auto-regulation.<sup>3</sup> The network is defined by its topology  $\Gamma \equiv \{T_{ij}, c_i\}$  and its continuous parameters or weightings  $W \equiv \{J_{ij}, h_i\}$ .

A configuration of on/off states  $\{s_i = \pm 1\}$  that satisfies Eq. (4.2) at every node is a fixed point of the network; fixed points describe configurations of static gene activity patterns within a single cell. Throughout this chapter, I explicitly avoid model-dependent considerations of dynamics and only count fixed points of this effective interaction network. Their number may correspond, for example, to the number of cell types this network can encode during development, and so is a natural measure of network complexity. In the language of neural networks (Amit, 1989; Hopfield, 1982), one can call it the *capacity* of a network  $c(\Gamma, W)$ . It can also be thought of as a measure of information-processing capability: for example, a network possessing just two fixed points  $\{s_i^{(1)}\}$  and  $\{s_i^{(2)}\}$  can be seen as taking one input (state of any one node  $s_k$  such that  $s_k^{(1)} \neq s_k^{(2)}$ ) and setting the state of other nodes to well-defined values that depend on this input ( $s_i^{(1)}$  or  $s_i^{(2)}$ , respectively). A network with more fixed points is in principle capable of distinguishing more combinations of inputs and

<sup>3</sup>In the presence of explicit auto-regulation, the allowed state of a node is no longer a single-valued function of its other inputs: consider, for example, bistability associated with strong auto-activation. The effect of auto-regulation can be implemented in this model as a feedback loop involving an extra node.

adjusting the outputs accordingly, and so performs a more complex computation. My task, then, is to compute  $c(\Gamma, W)$  (Kauffman, 1969).

### 4.2.2 Parameter space geometry

Although the parameters  $\{J_{ij}, h_i\}$  are continuous and, in principle, unbounded, if we are only interested in the *fixed points* of the network, there is a natural compact geometry to the parameter space, and this geometry also breaks into discrete subspaces. To see this, let's denote by  $\vec{w}_i$  the vector of all parameters that “feed” into gene  $i$ ,  $\vec{w}_i \equiv \{J_{i1}, J_{i2}, \dots, h_i\}$ . With  $N$  genes, there are  $N$  separate vectors  $\vec{w}_i$ , and together they define the parameter space of the model. But Eq. (4.2) has a symmetry, where the fixed points of the system are invariant under independent scaling of the parameters feeding into each node,  $\vec{w}_i \rightarrow \alpha_i \vec{w}_i$ .<sup>4</sup> Thus I can choose  $|\vec{w}_i| = 1$  for each gene  $i$ , so that the relevant parameter space is a direct product of positive segments of unit spheres.

The relevant parameter space being a portion of the unit sphere, there is a natural measure, namely the uniform distribution. But we can do more, because whole sectors of the parameter space produce the same fixed points. Indeed, Eq. (4.2) says that each node computes a Boolean function of its inputs, and there is only a finite number of Boolean functions with  $N$  inputs. Further, my model generates only a small subset of these, the perceptrons Minsky and Papert (1969).

In the simplest case (Fig. 4.2A), a gene  $i$  receives input from one other gene (with strength  $J_{i1}$ ) and compares this to a local threshold  $h_i$ ; the “unit sphere” is just a quarter circle. But if  $J_{i1} < h_i$ , then gene  $i$  will have the same state,  $s_i = \text{sgn } c_i$ , no matter what the state of the input  $s_1$  might be. On the other hand, if  $J_{i1} > h_i$ , then

---

<sup>4</sup>Part of my simplification is to ignore noise, so that Eq. (4.2) is deterministic. With noise the symmetry is only approximate, since noise magnitude provides a scale. Most of the literature on network complexity also considers the noiseless limit, so this seems like an appropriate starting point. There is a separate literature about design principles related to combating noise and optimizing information transmission in genetic networks (Andrews et al., 2006; Bialek, 2012; Tkacik et al., 2008), but connecting these different views is beyond the scope of this chapter.

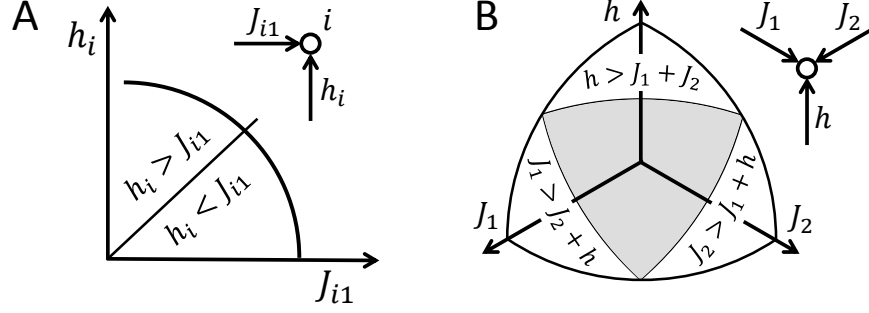


Figure 4.2: Parameter space splits into discrete weighting sectors. **A:** The relative strength of two inputs to node  $i$ ,  $J_{i1}$  and  $h_i$ , is parameterized by a point on a circle. There are two sectors:  $h_i > J_{i1}$  and  $J_{i1} > h_i$ . **B:** For three inputs, the parameter space is a 2-sphere and splits into 4 sectors: three in which one input dominates, and the unique non-dominating sector (shaded).

the state of gene  $i$  is determined uniquely by the state of the input,  $s_i = \text{sgn}(T_{i1}s_1)$ .

These sectors have the same weight under the uniform distribution.

This simplest case also alerts us to a problem, namely that some combinations of parameters aren't very interesting. In one of the two sectors, gene  $i$  is essentially uncoupled from the network. In the other, it is redundant with gene 1. Somehow neither of these cases sounds much like a “network.”

The next simplest case is where gene  $i$  takes two inputs and compares their sum to a threshold (“in-degree” 2). Now the vector  $\vec{w}_i$  is three-dimensional, so that the relevant space is a segment of the familiar two-sphere embedded in three dimensions (Fig. 4.2B). Again the continuous space breaks into discrete regions, and within each region the fixed points of the network are the same. There are four regions, and three of them are uninteresting in the same way that we saw for the two-dimensional case. In one sector, the state of gene  $i$  is dominated by the threshold  $h_i$ , and so is independent of all other nodes. In two other sectors, one of the interactions  $J_{ij}$  is so large that it dominates all other inputs, and hence gene  $i$  is redundant with one other gene  $j$ . Only in the remaining fourth sector the state of gene  $i$  depends in a nontrivial way on the combination of its inputs.

The picture of dominating sectors in Fig. 4.2B—regions of parameter space where one node becomes either decoupled from the network or redundant with another node—is the quantitative expression of the intuition that making interactions continuously weaker is not obviously separable from changing topology by erasing these interactions altogether. In the dominating sectors, interactions can be erased without changing the set of fixed points of the network. Therefore, parameter settings in such dominating sectors are functionally equivalent to networks with simpler topology, so to compare the role of continuous parameters to that of network topology, I will exclude these regions of parameter space.

### 4.2.3 Weighting sectors: a formal definition

To define parameter sectors in the general case, it is convenient to formalize the symmetry between parameters  $J$  and  $h$  that became apparent in Fig. 4.2. The  $-H_i = -c_i h_i$  term in Eq. (4.2) can be seen as an additional input from a constitutively expressed virtual node  $s_{\cancel{i}} \equiv 1$ . This node acts as a repressor if  $c_i = +1$  and an activator if  $c_i = -1$ , and the internal threshold  $h_i > 0$  is the weight of this interaction. In the expanded graph containing this virtual “always on” node, the regulatory links and the internal thresholds can be treated in a unified way:  $c_i$  becomes part of the extended topology  $\tilde{T}_{ij}$ , and  $h_i$  a part of the extended interaction matrix  $\tilde{J}_{ij}$ :

$$\begin{aligned}\tilde{J}_{ij} &= J_{ij} & \tilde{J}_{\cancel{i}i} &= 0 & \tilde{J}_{i\cancel{i}} &= h_i \\ \tilde{T}_{ij} &= T_{ij} & \tilde{T}_{\cancel{i}i} &= 0 & \tilde{T}_{i\cancel{i}} &= -c_i.\end{aligned}$$

For a fixed topology and a given state of nodes  $\{s_i\}$ , denote  $f_{i \rightarrow j} \equiv s_i \tilde{T}_{ji}$  for each link  $i \rightarrow j$ . Borrowing the language of spin glass literature (Mezard et al., 1987), I can say that  $f_{i \rightarrow j}$  describes whether the link is “satisfied” or “frustrated” in this node configuration, i.e. whether node  $j$  is in the state where regulation by  $i$  is pushing it



(in particular, a virtual link is satisfied for a constitutively activated node that is *on* or a constitutively repressed node that is *off*).

The full parameter space is a direct product of the parameter spaces describing individual nodes, so to simplify notation, let me focus on one node  $i_0$ . Denote  $U(i_0)$  the complete set of its inputs (genes that regulate  $i_0$ , as well as the virtual node  $\mathbb{K}$ ), and  $K + 1$  their number:

$$U(i_0) = \{j \mid \tilde{T}_{ij} \neq 0\}.$$

As before, let  $\vec{w}$  be the  $(K + 1)$ -element vector of the strengths of the interactions regulating gene  $i_0$ :  $\vec{w} = \{\tilde{J}_{i_0j} \mid j \in U(i_0)\}$ . These parameters define a Boolean function of  $K + 1$  arguments  $\phi_{\vec{w}}: \{1, -1\}^{K+1} \mapsto \{1, -1\}$ :

$$\phi_{\vec{w}}(b_1, b_2, \dots, b_{K+1}) = \text{sgn} \left( \sum_j b_j w_j \right) = \text{sgn}(\vec{b} \cdot \vec{w}). \quad (4.3)$$

This function has the following interpretation: when applied to the set  $\{f_{j \rightarrow i_0} \mid j \in U(i_0)\}$ , it tells us if this combination of satisfied/frustrated links is allowed by the regulatory rule at node  $i_0$ . Note that this is *not* the function that maps the states of input nodes into the state of the target node (the input/output function); using  $\phi$  allows me to exhibit the symmetry between all  $K + 1$  inputs, whereas the input-output function must treat the internal threshold separately.

I now arrive at my definition: a “weighting sector” at node  $i_0$  is the equivalence class of vectors  $\vec{w}$  that define the same Boolean function  $\phi_{\vec{w}}$ . In the  $K = 2$  example considered above (Fig. 4.2), the three “dominating sectors” are described by Boolean functions  $\phi_a(b_1, b_2, b_3) = b_a$ , where  $a \in \{1, 2, 3\}$ . In the unique non-dominating sector, Eq. (4.2) is satisfied whenever any two of the links are satisfied, and so the Boolean function is given by

$$\phi(b_1, b_2, b_3) = (b_1 \wedge b_2) \vee (b_2 \wedge b_3) \vee (b_3 \wedge b_1).$$

For  $K = 3$ , the parameter vector  $\vec{w}_i$  is four-dimensional, and there are twelve discrete sectors of the 3-dimensional parameter sphere  $|\vec{w}|^2 = 1$  that generate different fixed points (see Sec. 4.A). An important feature of the model I describe is that the relative probabilities of different sectors are set by the microscopic structure of the problem rather than postulated. Remarkably, for  $K = 3$  direct integration shows that all sectors have equal probability under the uniform measure on the parameter sphere. Using standard spherical coordinates, the volume of a sector  $\Omega$  is given by

$$V(\Omega) = \int_{\Omega} \sin^2(\theta) \sin(\phi) d\theta d\phi d\psi.$$

For instance, the sector dominated by weight  $w_1$  is defined by the conditions  $w_2 > 0$ ,  $w_3 > 0$ ,  $w_4 > 0$ , and  $w_1 > w_2 + w_3 + w_4$ . Rewriting the last condition as

$$\theta < \operatorname{arccot}(\cos \phi + \sin \phi \cos \psi + \sin \phi \sin \psi)$$

and integrating, one finds that this volume is equal to  $\pi^2/96$ , or  $1/12$  of the total volume of the positive sector of the unit 3-sphere.

There is a large literature describing genetic networks in Boolean terms, going back at least to Kauffman (1969). Following Kauffman, much of this work considers each gene as being controlled by an arbitrary Boolean function of its  $n$  inputs, drawn uniformly from the set of all  $2^{2^n}$  possibilities (the Random Boolean Network model, RBN). Biologically, however, some functions are harder to implement than others: for example, it is far from obvious how a cell might physically realize a regulatory node performing summation modulo 2. Individual RBN nodes are capable of performing arbitrarily complex calculations with equal ease. Here, I seek to understand how network-level complexity is constructed from simple local rules, and thus a new framework is required. More in the spirit of Hopfield (1982), I explicitly construct our model as a coarse-graining of an underlying microscopic model with a simple

sum-threshold rule. For networks in which each gene is influenced by three inputs, I find that a given topology has  $(12 - 4)^N = 8^N$  equiprobable realizations as distinct networks with weighted interactions (four sectors with a single dominating weight have been excluded; see Sec. 4.A). This means that I can do exhaustive enumerations up to values of  $N$  that are typical of real genetic networks. With four inputs there are  $(76)^N$  networks for each topology, so I will confine the discussion to the case of three inputs.

### 4.3 Computing complexity

Let me start with  $N$  genes connected in a particular topology, and choose parameters at random from the  $8^N$  equiprobable sectors described above. For each network I can measure the capacity, or number of solutions to Eq. (4.2), and then average over parameters at fixed topology. Motivated by the spin-glass intuition (Mezard et al., 1987) that the diversity of solution patterns arises from the phenomenon of link frustration, I chose to focus on networks consisting exclusively of repressing interactions<sup>5</sup> ( $c_i \equiv 1$  and  $T_{ij} \in \{0, -1\}$ ). Note that in my model, an inactive repressor acts as an activator, so this assumption is not overly restrictive.

In Fig. 4.3, I show the distribution of this mean complexity across topologies. We observe that in this case, the average complexity is independent of  $N$ . This may appear counterintuitive: we expect larger graphs to be capable of storing more patterns, but they also have more weightings with few or no fixed points. This result can be demonstrated analytically with a mean-field argument that holds independently of my simplifying assumptions such as constant in-degree 4.B. It shows that, quite generally, a larger network is not automatically more complex; rather, it has the potential for high complexity, but only realizes this potential with a careful choice of

---

<sup>5</sup>A network consisting exclusively of activators always trivially possesses the fixed point  $s_i \equiv 1$ , whereas for repressors the existence of even one fixed point is not guaranteed.

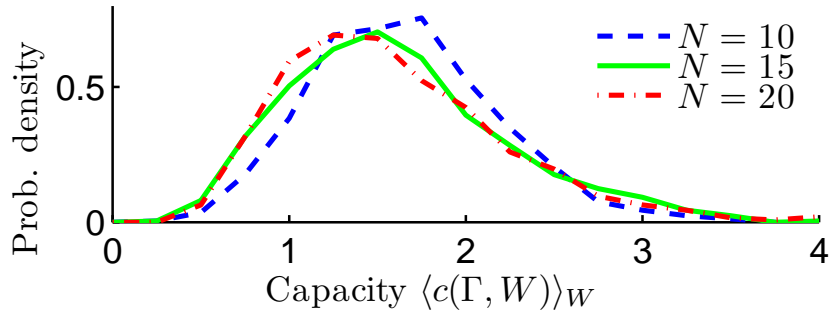


Figure 4.3: Distribution of average capacity  $\langle c(\Gamma, W) \rangle_W$  for 1000 random topologies  $\Gamma$  of in-degree  $K = 3$  and varying  $N$ . For each topology  $\Gamma$ , the capacity was averaged over all  $8^N$  realizations of  $\Gamma$  as a weighted graph (all possible weightings  $W$ ), which we have shown to be equiprobable for  $K = 3$ . This average complexity does not grow with  $N$ .

weights. It is worth mentioning that if I were to sample topologies non-uniformly, preferring those from a particular class (e.g. topologies generated by a particular network growth model), the expected complexity of the average realization would be modified and would generically begin to scale with  $N$ . The exponent of such scaling can be used to characterize the complexity of a *class* of topologies. For example, enrichment of  $\Gamma$  in certain motifs causes  $\langle c(\Gamma, W) \rangle_{\Gamma, W}$  to grow with  $N$ , providing a connection between the framework described here and the results of Lee et al. (2002) and Milo et al. (2002).

To highlight the difference between average and attainable complexity, I calculate, for a given topology, the distribution of  $c(\Gamma, W)$  over all weightings  $W$ . For  $N = 6$  I can enumerate all topologies; Fig. 4.4 shows three examples with the lowest, typical and highest average complexity. We observe that the complexity distributions overlap considerably, and the typical realizations of even the best topology are routinely outperformed by “lesser” topologies when their weights are optimized. This persists for larger  $N$ : the best  $N = 9$  topology found by a targeted search has average complexity  $\max_{\Gamma} \langle c(\Gamma, W) \rangle_W = 5.26$ . This is far in the tail of the distribution: uniform sampling of 1000 topologies gives an average complexity of only  $\langle c(\Gamma, W) \rangle_{\Gamma, W} = 1.7 \pm 0.5$ . However, optimizing weights of random topologies gives higher complexity in

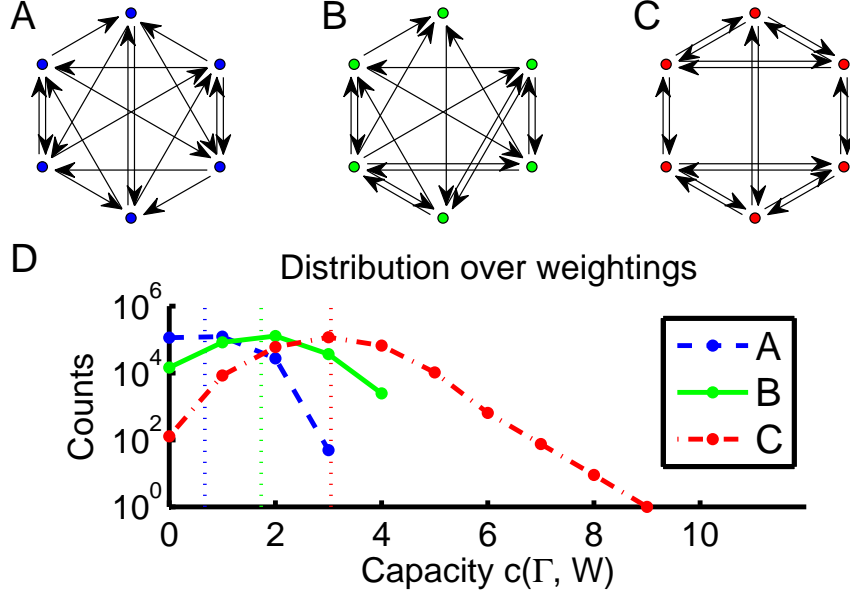


Figure 4.4: **A-C**:  $N = 6$  graphs with lowest, median and highest average capacity. All interactions are repressive (but denoted by pointed, rather than blunt, arrows for visual clarity). **D**: Distributions of capacity of these graphs over the choice of weighting overlap significantly. The curves shows the total number of weightings with a given capacity, out of  $8^6 \approx 2.6 \cdot 10^5$  total. Vertical lines show the average capacity (0.7, 1.7 and 3.0). Even the “worst” graph (A) can be optimized to attain the typical capacity of the best topology (C).

85% of the samples. In other words, if one had to pick only one feature to optimize, either weights or topology but not both, optimizing weights is the better strategy.

Before interpreting these results, I should be careful about my definition of complexity. First, counting fixed points treats them as equally important or likely. Second, highly disconnected graphs can achieve high capacity without being complex in any intuitive sense. For example, the maximal capacity of a network with  $M$  mutually repressing pairs (Fig. 4.S2A) grows exponentially,  $\max_W c(\Gamma, W) = 2^M$ ; other simple entropy-based definitions of complexity suffer from this problem as well. I note, however, that the maximal capacities that I find for  $N$ -gene networks ( $\geq 27$  for  $N = 9$ ,  $\geq 54$  for  $N = 10$ ) are larger than  $2^{\lfloor N/2 \rfloor}$ , i.e. storing patterns in a distributed way is more efficient than splitting the network into many disconnected components. Therefore, the highest capacity networks are not trivially so. Further, in the set of all

graphs, disconnected topologies are extremely rare and can be expected to provide but a small correction to the averages. One can construct a better definition, quantifying complexity of a network as the diversity of causal relations across its fixed points (see Sec. 4.C). This definition naturally handles all the problems mentioned above, yet for connected graphs it is in excellent agreement with  $c(\Gamma, W)$  (Fig. 4.S3), which is much easier to compute.

## 4.4 Discussion

In the current paradigm, topology is seen as the primary determining property of regulatory network complexity. For example, eukaryotes are said to be more complex than prokaryotes because of their more intricate regulation of a roughly similar number of genes. Here, “complex” is used synonymously with “more densely connected,” and evolution towards more complex systems is assumed to act by introducing new regulatory interactions. This topology-centric view is inconsistent with ample evidence indicating that sensitivity to some quantitative parameters (in the language of Gutenkunst et al. (2007), the “stiff” parameter combinations) is a generic phenomenon in both physical and biological contexts (Arenas et al., 2008; Bagowski et al., 2003; Bintu et al., 2005; Feinberg, 1987; Goentoro et al., 2009; Rue and Garcia-Ojalvo, 2011; Tyson et al., 2003). To reconcile this conflict, it is usually argued that biological networks are non-generic, and could have evolved to achieve reduced parameter sensitivity. While this may be true for some networks (e.g., Bagowski et al., 2003) and makes them appealing targets for study in view of the experimental challenges associated with quantitative measurements, this evidence is insufficient to conclude that quantitative parameters generally play a subdominant role in shaping the function of biological circuits.

Here, I constructed a class of models where the contributions of topology and interaction strengths to network complexity could be quantitatively dissected and compared. Within this class, I found that the interactions strengths are dominant: starting from a generic random network, adjusting the strengths of regulatory interactions is a better strategy for increasing its complexity than changing the topology or even adding new nodes.<sup>6</sup> As in the brain, it seems that topology provides the potential for complexity, but parameters must be adjusted carefully to realize this potential.

Function and complexity are not synonymous. For instance, a cascade transducing a signal from a receptor to a kinase need not be complex in the sense defined here, and its function will not qualitatively depend on the rates of individual steps. However, the computations performed by regulatory networks may be much more complex. For example, a stem cell integrates external cues into a decision to activate one of many specialization programs; somewhere in its internal circuitry is a module responsible for this information-processing task. Our analysis shows that for such circuits, focusing solely on topology would mean ignoring an entire dimension of complexity that biological networks possess and are known to exploit in other information-processing contexts, e.g. neural networks (Hopfield, 1982). Developmental decision-making circuits are largely believed to function as hierarchically structured cascades of binary decisions; however, encoding states in a distributed way is more efficient, albeit more sensitive to parameters (see the results on maximal capacity above; also Hopfield (1982)). The standing of real circuits with respect to this tradeoff is, I believe, not a settled question.

---

<sup>6</sup>In the context of evolution, what matters is not whether there exists a weighting of the same topology that has a larger number of fixed points, but whether weights can be adjusted continuously to add another fixed point to the set *while keeping all the old ones*. Our framework allows us to study this mutation landscape. In particular, one can show that if weightings  $W_1$ ,  $W_2$  have fixed point sets  $S_1$ ,  $S_2$ , and  $S_1 \subset S_2$ , then  $W_1$  can always be evolved into  $W_2$  via a path that never loses any of the fixed points in  $S_1$ .

In my model, maximally complex network function is accessible only in a small fraction of parameter space, but optimization does not lead to unique parameter settings. Instead, there is a whole sector of the underlying continuous parameter space that produces the same results. Thus, in these models, maximizing complexity leads to an interesting combination of tuning and tolerance.

Adjusting parameters is a more efficient method for increasing complexity, as illustrated in Fig. 4.4, but it still is possible that the evolution of complexity is associated with changes in network topology (Carroll, 2005). If continuous parameters can evolve more rapidly than topologies can change, which seems plausible, then today's organisms may be dominated by networks that are near optimal given their topology. In this scenario, today's more complex organisms must have networks with different topology than their less complex counterparts, but not because parameters are irrelevant—rather, topology becomes determining only once parameters have been optimized.



# Technical details

## 4.A Weighting sectors for topological in-degree

$$K = 3$$

In this work, I consider graphs of topological in-degree  $K = 3$ , which means we have 4 controlling links per node in the extended graph. In this case there are exactly 12 weighting sectors, summarized in Table 4.S1: 4 cases with a single “dominating” link (one weight is stronger than all others put together, so the state of this one link defines whether the whole configuration is allowed or not), 4 “sub-dominating” (the strongest link and any other must be satisfied, or all three weakest) and 4 “combinatorial” (any pair from a given set of three should be satisfied).

As mentioned in 4.2.2, a node regulated by a dominating link is either redundant with another node or disconnected from the rest of the network (if the dominating weight is the internal threshold). Such nodes cannot increase the complexity of a network, so one expects that forbidding dominating sectors should increase the complexity of a graph. This is indeed correct: Fig. 4.S1 shows the distribution of average

Sector name		Minimal configuration of satisfied links
Dominating	$D_a$	Link $a$ ( $a \in \{1, 2, 3\}$ )
Sub-dominating	$SD_a$	Link $a$ and any other, or all but $a$
Combinatorial	$C_a$	Any two excluding $a$

Table 4.S1: Weighting sectors for in-degree  $K = 3$ .

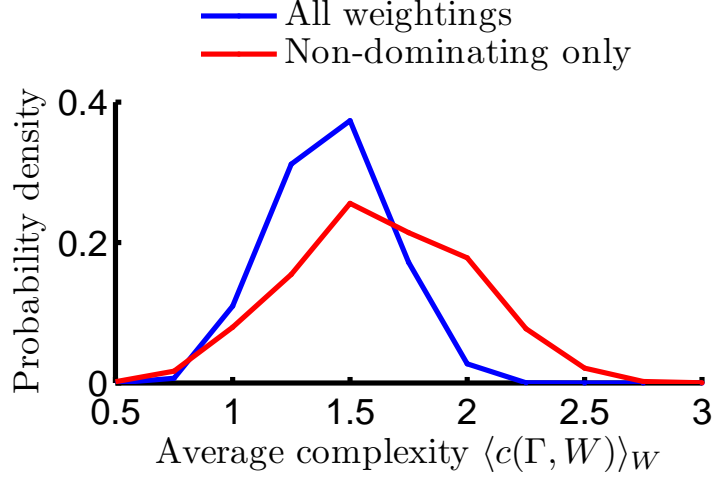


Figure 4.S1: Distribution of average complexity of all  $N = 6$  topologies. Restricting weighting sectors to only non-dominating ones gives a higher average complexity, consistent with expectation that dominating sectors correspond to an effectively simpler topology.

complexity  $\langle c(\Gamma, W) \rangle_W$  over all  $N = 6$  topologies. Restricting weighting sectors to only non-dominating ones increases the average complexity. This, however, should be seen as the effect of topology rather than parameter choice, because when we picked a topology of in-degree  $K = 3$ , we already recognized the need to have more than 1 regulatory input per node. Therefore, in this work, to distinguish between the effect of weights and topology, I considered non-dominating weighting sectors only. With  $K = 2$  we have only a single non-dominating sector (Fig. 4.2B), while with  $K = 4$  there are 76 (!). Correspondingly, I studied networks of in-degree  $K = 3$ , as it is the simplest nontrivial case.

It must be noted that in a combinatorial sector  $C_a$ ,  $a \in \{1, 2, 3\}$ , the weight of the link  $a$  is so weak it has no effect on the state of the target node. For example, a representative set of weights from the sector  $C_1$  is  $\{1, 5, 5, 5\}$ , and removing the first link will not affect the set of fixed points of this particular weighting. Despite this fact, the high-complexity graphs are highly enriched in combinatorial sectors, because they in fact perform the most nontrivial local calculation in the information-theoretic sense. To see this, I consider the Boolean input-output functions implemented by a

given  $K = 3$  node and compute the mutual information between any one input  $s_{\text{in}}^{(a)}$  and the output node  $s_{\text{out}}$ , defined as the reduction in entropy of the output brought by the knowledge of the state of a particular input:

$$\begin{aligned} I(s_{\text{out}}, s_{\text{in}}^{(a)}) &= S[p(s_{\text{out}})] - S[p(s_{\text{out}} | s_{\text{in}}^{(a)})] \\ &= 1 - S[p(s_{\text{out}} | s_{\text{in}}^{(a)})]. \end{aligned}$$

Here  $S[p(\cdot)]$  denotes the entropy of a probability distribution, and the unconditional entropy of the output is exactly 1 bit. The complexity of the local computation performed by a node can be quantified as the maximum information any one input brings about the output: the lower the information, the more complex the computation. The one performed by a dominating node is trivial: the output is entirely determined by the strongest input (which carries 1 bit). A sub-dominating node fares better: the strongest input carries only

$$1 + \frac{1}{8} \log_2 \left( \frac{1}{8} \right) + \frac{7}{8} \log_2 \left( \frac{7}{8} \right) \approx 0.46 \text{ bits.}$$

For the combinatorial node, no input carries more than  $\approx 0.19$  bits, and thus it performs the most interesting computation.

## 4.B Computational details

Computations were performed with a C++ code, on network topologies with in-degree 3 consisting of repressing interactions only. The choice to only use repressing interactions was motivated by the spin-glass intuition that the diversity of solutions to Eq. (4.2) fixed points arises from the phenomenon of link frustration mentioned in Sec. 4.2.3 (see also Mezard et al. (1987)). This assumption is not overly restrictive, since in my model, an inactive repressor acts as an activator. Note, however, that a

Link states	Compatible weighting sectors
$\{1, 1, 1, 1\}$	All
$\{1, 1, 1, 0\}$	All but $D_4$
$\{1, 1, 0, 0\}$	$SD_1, SD_2, C_3, C_4$
$\{1, 0, 0, 0\}$	$D_1$ only
$\{0, 0, 0, 0\}$	None

Table 4.S2: Weighting sectors compatible with a given pattern of satisfied/frustrated links. Dominating sectors included for illustration purposes.

network consisting exclusively of activators always trivially possesses the fixed point  $s_i \equiv 1$ , whereas for repressors, the existence of even one fixed point is not guaranteed.

Efficient computation was made possible by the following observation. Assume the topology of the graph is fixed. Determining which configurations are fixed points for a given weighting (i.e. solutions of Eq. (4.2)) is computationally hard. However, our constraints (4.2) possess a special structure: there is one local constraint per node, and it involves only the weighting sector associated with this node itself. In other words, the constraints are factorized over the local choice of a weighting sector. This makes the inverse question, “given a node state configuration  $\{s_i\}$ , for which weights is this a fixed point?” extremely simple. Each pattern of satisfied and frustrated incoming links (set by the node states) defines a list of compatible sectors (Table 4.S2). I can therefore construct a “weighting sector compatibility matrix”, a  $N \times 12$  Boolean matrix identifying, for each node, the list of allowed sectors. For illustration purposes, the dominating sectors are also included in this table; they are not included in the calculations.

The first step of my calculations is always to construct a list of all states that are ever solutions of Eq. (4.2), as well as their compatibility matrices, and takes very little time. These are then used for subsequent steps, to which I now turn.

### 4.B.1 Targeted search for high-complexity weightings

To determine the distribution of complexity  $c(\Gamma, W)$  over weightings  $W$ , I perform a recursive tree search by sequentially fixing weight sector choices at every node and calculating, at every step, which subset of the list of potential fixed points is compatible with the partially fixed weight sector sequence. The factorization property mentioned above ensures that when choosing a sector at each new node, the list of compatible fixed points can only shrink. I can thus perform a targeted search aimed at identifying high-complexity weightings: if I discover early on that the total number of potentially compatible fixed points falls below a certain threshold, I can drop the entire subtree of weightings described by the partial specification, and move on. Since most weightings have in fact very few fixed points, this allows to exactly enumerate *all* weightings of complexities exceeding a fixed threshold, while still keeping computation time low. If the “drop-out” threshold is set to zero, the entire distribution is calculated exactly, and the computation time is still considerably lower than the brute-force enumeration of weightings, taking on the order of a minute on a modern desktop computer for  $N = 7$ . The high-complexity tails can be readily studied up to  $N = 10$ .

To obtain an approximate distribution of the number of fixed points over all weightings for  $N = 8-10$ , when full weighting enumeration is not feasible, I first run a targeted search for high-complexity weightings as described above. I then sample a large number of weightings (e.g.,  $10^5$ ) at random and stitch the resulting distribution of low-complexity weightings with the already calculated exact tail of the distribution. This procedure gives excellent agreement for graphs with  $N \leq 7$  where a full distribution can be calculated exactly.

### 4.B.2 Computing the mean complexity of a topology

The average complexity over all weightings  $\langle c(\Gamma, W) \rangle_W$  for a given topology  $\Gamma$  can be determined without calculating the entire distribution of the number of fixed

points by using the following trick. Notice that the average number of fixed points is equal to  $P(\Gamma)/N_W$ , where  $N_W$  is the number of all possible weightings  $W$ , and  $P = \sum_W c(\Gamma, W)$  is the total number of pairs  $(\vec{s}, W)$ , where the state vector  $\vec{s}$  is a fixed point of the network  $(\Gamma, W)$ . To calculate  $P(\Gamma)$ , rather than summing the number of fixed points for each weighting  $W$ , I will sum, for all states  $\vec{s}$ , the number  $N_W(\vec{s})$  of weightings compatible with that state:

$$P(\Gamma) = \sum_W c(\Gamma, W) = \sum_{\vec{s}} N_W(\vec{s}).$$

To calculate  $N_W(\vec{s})$  I use the factorization property mentioned above and the Table 4.S2:

$$N_W(\vec{s}) = \prod_{i=1}^N N_{\text{sec}}[\{f_{j \rightarrow i}(\vec{s}) \mid j \in U(i)\}],$$

where  $f_{j \rightarrow i}$  denotes, as before, the frustrated/satisfied state of link  $j \rightarrow i$ , and  $N_{\text{sec}}$  is the number of weighting sectors compatible with a given pattern of incoming link states (see Table 4.S2). This trick allows me to rapidly calculate the exact value of typical complexity of graphs up to about  $N = 25$ , and eliminates the need to loop over the weightings themselves, a prohibitively large space for large values of  $N$ .

For a large  $N$ , I can also estimate the average complexity  $\bar{c} = \langle c(\Gamma, W) \rangle_{W, \Gamma}$  by making a mean-field approximation in Eq. (4.B.2). For a given node  $i$  in the network, each of the incoming links has equal probability of being satisfied and frustrated. Therefore, Table 4.S2 (after excluding the dominating sectors) tells us that the number of allowed weighting sectors at a randomly selected node is a random variable  $n$ , drawn from a distribution  $P$ :

$$n = \begin{cases} 8 & \text{with probability } 5/16 \\ 4 & \text{with probability } 6/16 \\ 0 & \text{with probability } 5/16 \end{cases}.$$

In the mean-field approximation, I can take this to hold independently for each node in the graph, so

$$\bar{c} = \frac{1}{N_W} \sum_{\vec{s}} N(\{\vec{s}\}) \approx \frac{1}{8^N} 2^N \prod_{i=1}^N n_i,$$

where  $n_i$  are random variables drawn from  $P$ . Note that every term in the sum has probability  $(11/16)^N$  to be nonzero, and nonzero terms can be rewritten in terms of a new random variable  $q$  drawn from  $Q = P|_{P>0}$ , i.e. from distribution  $P$  conditioned on positivity constraint. I conclude that

$$\bar{c} = \left(\frac{2 \cdot 11}{8 \cdot 16}\right)^N \prod_{i=1}^N q_i = \left(\frac{11}{64}\right)^N \exp(N \langle \ln q \rangle) = \alpha^N,$$

where  $\alpha = \frac{11}{64} \exp(\langle \ln q \rangle) \approx 1.01$  is suspiciously close to 1.

This result can be understood in more general terms. A network in my model is a system of  $N$  binary variables, constrained with  $N$  binary equations, each forbidding exactly half of the configuration space. One therefore expects, on average, a number of fixed points of order  $2^N \left(\frac{1}{2}\right)^N = 1$ , irrespectively of  $N$ . Note that this argument is very general and requires neither the assumption of a constant in-degree nor the fact that genes are modeled as binary variables; it relies only on the fact that the input-output function at each node maps each sets of inputs into exactly one output. The slight difference between the  $\alpha$  of the mean-field calculation and 1 comes from the weak convexity of the logarithm and is most likely within the error of the mean-field approximation. This simple argument shows that for random (uniformly sampled) topologies the average complexity is not expected to increase with  $N$ . While a large network is in principle capable of storing many patterns (and this can indeed be aided by a biased choice of topology), achieving high complexity requires a careful adjustment of weights.

I would like to contrast this result with the fact that spin glasses can have exponentially many locally stable states (Mezard et al., 1987). The apparent contradiction

comes from the fact that the ground state of a spin glass is not required to satisfy every single node; a ground state only minimizes the frustration in the entire system. In particular, a ground state always exists, in contrast to a fixed point of a network in the sense we consider here.

### 4.B.3 Targeted search for high-complexity topologies

Previous sections discussed techniques to find, for a fixed topology  $\Gamma$ , the distribution of available complexities  $c(\Gamma, W)$  and the optimal weightings that maximize it. Another relevant question to ask is which topologies  $\Gamma^*$  achieve the highest complexity, either on average  $\langle c(\Gamma^*, W) \rangle_W$  or after optimization of weights  $\max_W(c(\Gamma^*, W))$ . Finding these topologies becomes particularly important when we realize that disconnected graphs may have a high number of fixed points without being truly complex in information-processing sense described in Sec. 4.2.1 (when the network is seen as performing a mapping between a subset of nodes designated as “input” and the rest of the nodes, the “output”). For example, a disconnected network consisting of  $M$  mutually repressing pairs of nodes has  $2M$  nodes and can have  $2^M$  fixed points (Fig. 4.S2A). In the space of all  $N$ -node graphs, disconnected topologies are exponentially rare. Therefore, one does not expect disconnected topologies to significantly affect statistical properties, such as average complexity values of all topologies with a fixed  $N$ . However, to be able to interpret the results I report, I need to ensure that the highest-complexity networks identified by my measure are not pathological in this manner.

Luckily, the highest-capacity graphs are in fact connected and their capacity exceeds  $2^{N/2}$ : storing patterns in a network in a distributed way is more efficient than splitting it into many disconnected components. To see this, one needs a method to perform a targeted search for high-complexity topologies. Indeed, for  $N > 6$ , there



are too many topologies to be sampled exhaustively, and probing tails of heavy-tailed distributions by random sampling is extremely inefficient.

To solve this problem, I used the empirical observation that high-complexity topologies are enriched in mutual-repression motifs (compare Fig. 4.4, panels A, B and C). The reason for this becomes clear when one recalls the mean-field argument I used to calculate the expected number of fixed points of a graph. The mean-field approximation assumes, for all links, equal probability of being satisfied or frustrated. In a mutual repression motif, however, the two links are always either both satisfied, or both frustrated. In a topology enriched in mutually repressing pairs, every time a node is satisfied (i.e. has a sufficient number of satisfied incoming links), this increases the probability of its neighbors to be satisfied as well; consequently, such topologies tend to have weightings with larger number of fixed points. By sampling only topologies whose  $T_{ij}$  matrix (no tilde!) is close to symmetric, I frequently find graphs with complexities far above average. The best networks I found in this way have complexity 27 for  $N = 9$  and 54 for  $N = 10$ , all confidently above  $2^{N/2}$ . The highest complexity achievable for a given  $N$  may be higher still, demonstrating that the highest capacity networks are not trivially so.

I stress that the fact that symmetric  $T_{ij}$  matrices lead to higher complexity values is a consequence of my simplifying choice to only consider repressing interactions. Allowing interactions to be both repressing and activating will remove this special structure. Note that the discussion in this section shows how biasing the sampling of

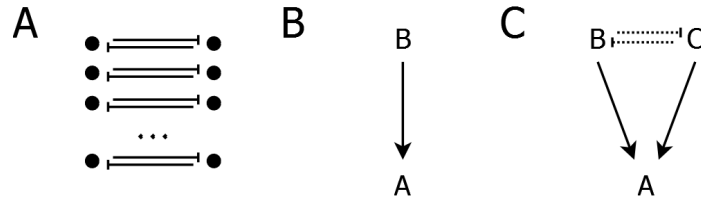


Figure 4.S2: **A:** A simple graph with high capacity. **B, C:** Two graphs of equal capacity but different regulatory complexity.

topologies towards an enrichment in certain local structures can enhance the average complexity of network function. Topologies with the greatest potential for complexity have local structures in common with those identified in real networks Lee et al. (2002); Milo et al. (2002). Within the framework presented here, the importance of quantitative parameters and the benefit of local topological features can thus be treated on the same footing, the latter entering through the choice of measure on the space of topologies. I leave a detailed exploration of this topic for a future investigation.

## 4.C TSP complexity

Another approach to deal with pathological cases such as Fig. 4.S2A is to redefine the complexity measure in a way that does not see such graphs as complex. I will now construct an improved definition of network complexity that quantifies complexity as the diversity of causal relations across the set of fixed points. I will now explain what I mean by this, and show that this definition, first, correctly handles pathological cases of disconnected topology, and second, for connected graphs is in excellent agreement with the more simple complexity measure used throughout this chapter.

I begin with an example. Compare two situations (Fig. 4.S2B, C): in the first, gene  $A$  is directly regulated by  $B$ , so they are both “on” or both “off”. In the second, gene  $A$  can be activated by either of its two inputs  $B$  or  $C$ , so we can have  $A$  “on” because  $B$  is “on”, or because  $C$  is “on”. Imagine that  $B$  and  $C$  are embedded in the network in such a way that they cannot both be active (shown as mutual repression on Fig. 4.S2C). In this case both networks have two states, but the regulation of  $A$  can be described as more “complex” in the second case, because the causal relations are more diverse.

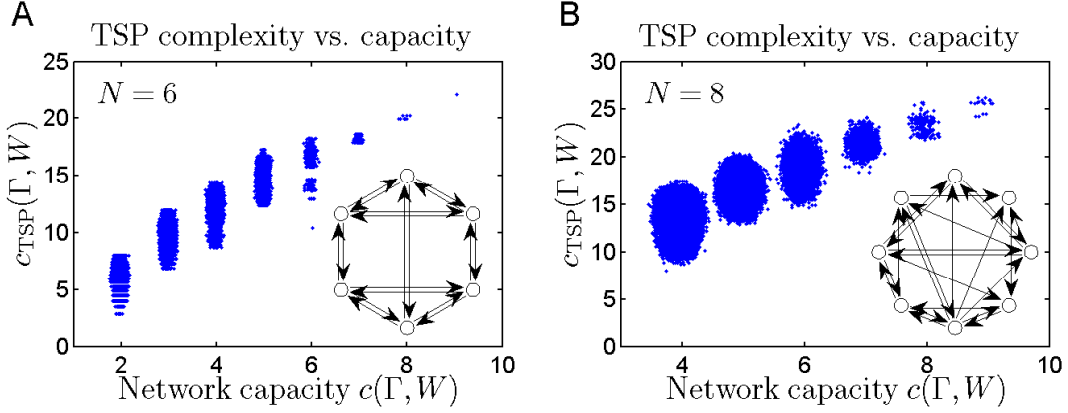


Figure 4.S3: For connected graphs, network capacity  $c(\Gamma, W)$  is in excellent agreement with the TSP complexity  $c_{\text{TSP}}(\Gamma, W)$ . Datapoints correspond to different weightings of the same topology shown in the inset. A small random component was added to the  $X$  axis for display purposes. **A:** the highest-complexity topology of  $N = 6$  (see Fig. 4.4C); scatter plot shows all weightings with at least 2 fixed points. Weightings with 0 and 1 fixed points have zero TSP complexity by definition. **B:** A random  $N = 8$  topology with the same maximal capacity (9 fixed points). All  $2.4 \times 10^6$  weightings with at least 4 fixed points are shown. The remaining  $1.4 \times 10^7$  have 3 fixed points or fewer.

To formalize the intuition gained from this example, I first introduce the notion of an “active link”. Define a satisfied link to be *active* if its satisfied state is essential for the regulatory rule (4.2) to be satisfied. In other words, for a given fixed point  $\{s_i\}$ , a link  $j_0 \rightarrow i_0$  is “active” if and only if substitution  $s_{j_0} \rightarrow -s_{j_0}$  upsets Eq. (4.2) at node  $i_0$ . (Clearly, frustrated links can never be active). Loosely speaking, an active link is “responsible” for setting the state of the node  $i_0$ . Each fixed point  $\vec{s}$  of a network  $(\Gamma, W)$  defines a pattern of active links  $\{a_k^{(\vec{s})}\}$ : a binary sequence specifying, for every link  $k$ , whether it is active ( $a = 1$ ) or not ( $a = 0$ ). For example, a “dominating” link is always the unique active input at the node it regulates, for any fixed point. For other combinations of link weights, we can have zero, one or more input links active simultaneously, and this pattern will vary from one fixed point to another.

I now take a moment to define the second ingredient of my definition, the diversity of a set of binary sequences. Given any two sequences, it is easy to define some measure of their difference; I will use the Euclidean distance. How can we quantify

the diversity of a *set* of sequences? Given a set of elements  $G = \{x_1, x_2, \dots, x_k\}$  and a metric of pairwise distances between elements  $d$ , we would like a reasonable measure of *diversity*  $D(G)$  to satisfy three properties:

1. Invariance under permutation:  $D(x_1, x_2, \dots, x_k) = D(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(k)})$  for any permutation  $\sigma$ .
2. Insensitivity to duplication of an element:  $D(x, x, y, z, \dots) = D(x, y, z, \dots)$ .
3. Additivity: if all  $x_i$  are pairwise equidistant,  $d(x_i, x_j) \equiv d_0$ , then

$$D(x_1, x_2, \dots, x_k) = kd_0.$$

A measure of diversity satisfying all these intuitive properties can be obtained by solving the travelling salesman problem (TSP):  $D_{\text{TSP}}(x_1, x_2, \dots, x_k)$  is defined as the length of the shortest closed path passing once through each of the “cities”  $x_i$ , with distances between cities being defined by the metric  $d$ .

I now have all the ingredients ready, and define the *TSP complexity* of a graph  $\Gamma$  as the TSP diversity of active link patterns across the set of fixed points of the network:  $c_{\text{TSP}}(\Gamma, W) \equiv D_{\text{TSP}}(\{a_k^{(s)}\})$ . This improved definition naturally accounts for varying pairwise similarity between fixed points. In particular, by focusing on the state of links rather than the nodes themselves, it correctly deals with “simple” graphs that may have many fixed points: note, for instance, that all the fixed points of the graph on Fig. 4.S2A have the exact same pattern of active links, and thus the graph receives zero TSP complexity score despite its large capacity.

In Fig. 4.S3 I show the comparison between  $c_{\text{TSP}}(\Gamma, W)$  and  $c(\Gamma, W)$ . We see that the agreement is quite good. Pathological cases when capacity overestimates true complexity are rare, and are all located at intermediate capacity values, in the bulk of the distribution. Therefore, once again, they neither alter the structure of the

high-complexity tail nor affect statistical properties of the distribution significantly. This improved measure of complexity, however, is computationally hard to evaluate, and the computational “tricks” discussed above no longer apply. Therefore, for the purposes of this work, I chose to use the simpler definition.

# Chapter 5

## Microbial communities

Host-associated microbial communities are known to be of tremendous importance for host fitness, improving nutrient uptake, training the immune system, and resisting invasion by pathogens (see, for example, Brestoff and Artis, 2013; Kamada et al., 2013; Fredricks, 2013). Our understanding of these communities, however, remains remarkably poor. The origin, maintenance, and importance of species diversity (Fierer and Lennon, 2011), the factors determining community stability and resilience (Shade et al., 2012), and the mechanisms of community assembly (Costello et al., 2012) are only some of the questions driving this rapidly expanding field.

Until recently, our ability to study these communities has been limited by the fact that most of their members cannot be cultured in a laboratory setting. However, advances in genome-sequencing technology now allow organisms to be probed in their natural environments. In particular, the 16S rRNA tag-sequencing approach identifies community members using fragments of DNA from the so-called hypervariable regions of the ribosomal 16S gene. The development of this technique (briefly introduced below) and the decreasing cost of high-throughput sequencing have prompted a large number of tag-sequencing experiments, including such large-scale efforts as the Human Microbiome Project or the Earth Microbiome Project. The amount of

collected data is growing exponentially. However, the standard approach to analyzing this data, which relies on clustering reads by sequence similarity into Operational Taxonomic Units (OTUs), underexploits the accuracy of modern sequencing technology. In this chapter, I present a clustering-free approach to multi-sample datasets that can identify independent bacterial subpopulations regardless of the similarity of their 16S tag sequences and therefore achieve sub-OTU resolution. Using published data from a longitudinal time-series study of human tongue microbiota, this approach is capable of resolving within standard 97% similarity OTUs up to 20 distinct subpopulations, all ecologically distinct but with 16S tags differing by as little as 1 nucleotide (99.2% similarity). I apply this approach to a comparative analysis of oral communities of two cohabiting individuals and demonstrate how the sub-OTU resolution can provide new insight into factors shaping community assembly.

For this project, I was supervised by Prof. Ned Wingreen, and the analysis code I designed was implemented as a distributable package by Robert Leach. This work is the subject of the following publication: Tikhonov M, Leach RW, Wingreen NS. (2014) “Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution”. *ISME J* (in press).

## 5.1 Introduction

Tag sequencing is a technique based on the remarkable properties of ribosomal RNA. The ribosome is a complex made of RNA and protein that performs a function essential for all life as we know it, translating RNA templates and synthesizing the proteins they encode. For this reason, its functional parts are highly conserved across all bacteria (the ribosomes of eukaryotes and archaea are also highly conserved, but differ from their bacterial analogs). The linkers that connect the functional parts are, however, not under selection and the mutations they acquire over the course of evolution are faithfully inherited by the progeny; they are called “hyper-variable regions” and differ significantly across bacterial “species”. Consequently, the DNA of each bacterium effectively carries a “tag” that identifies it, and these tags are flanked by highly conserved sequences that can be used to selectively amplify only these fragments via a DNA amplification technique called PCR. Given a sample of a bacterial community, we can extract the DNA from all the cells, amplify the tags identifying the community members (typically one of the hyper-variable regions of a ribosomal RNA gene called 16S) and sequence them to determine the composition of the community. This technique is culture-independent and can identify the presence of a bacterial strain even if it is a novel species that has never been seen before. In addition, for well-studied habitats, matching sequenced tags to databases of known species allows a fairly accurate taxonomic identification of the bacteria to which they belonged.

The de facto standard approach to 16S data analysis begins by clustering reads by sequence similarity into “Operational Taxonomic Units” (OTUs); see Fig. 5.1A (Quince et al., 2009; Kunin et al., 2010; Huse et al., 2010). A variety of clustering techniques have been developed and are widely used in popular software tools or packages (Hunt et al., 2008; Schloss et al., 2009; Edgar, 2010; Huang et al., 2010; Edgar et al., 2011; Quince et al., 2011; Schloss et al., 2011; Sul et al., 2011; Caporaso et al., 2012;



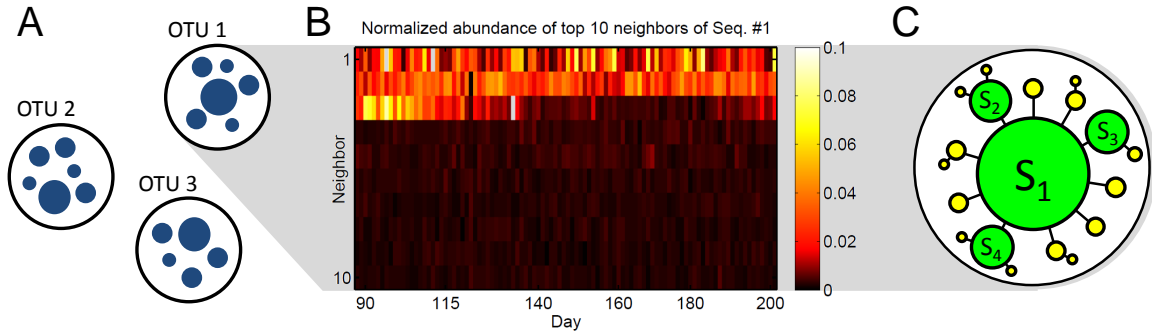


Figure 5.1: **Clustering reads by OTUs underexploits the quality of modern sequence data.** **A:** Cartoon illustrating OTU-based noise filtering. Due to sequencing errors, PCR errors or natural intra-strain variability, each bacterial “species” generates a cloud of similar 16S sequences (blue circles; the radius of a circle represents the abundance of a given 16S sequence in a sample, and spacing represents distance in sequence space). Clustering reads into OTUs by sequence similarity is a standard approach to filter this noise. **B:** Heat map of the abundance, for 100 consecutive samples, of the 10 highest-abundance direct neighbors (Hamming distance = 1) of Seq. #1, normalized for each sample to the abundance of Seq. #1 (4600 counts/day on average). Three specific direct neighbors are strongly and consistently overrepresented and exhibit distinct dynamics. **C:** Cartoon based on (B) of the expected structure of an “error cloud”. Each circle is a unique sequence, with size representing abundance in a sample. True biological sequences ( $S_1$ - $S_4$ ; green circles) generate “daughter” variants due to substitution errors (yellow circles). Black lines denote Hamming distance = 1 in sequence space.

Zheng et al., 2012; Morgan et al., 2013; Youngblut et al., 2013). Despite significant progress in the development of such software, all clustering-based approaches suffer from a major shortcoming (Prosser et al., 2007; Hamady and Knight, 2009; Schloss and Westcott, 2011). Although an OTU is a useful concept for coarse-graining sequencing data, its definition is not biologically motivated, but as its name acknowledges is purely operational. Sequences assigned to a particular OTU are generally presumed to be close phylogenetic relatives and therefore likely to derive from ecologically similar bacterial subpopulations. However, the assumption that 16S sequence similarity is a good proxy for ecological similarity is notoriously problematic (Prosser et al., 2007; Preheim et al., 2013). Moreover, OTU assignments are not definitive but depend on both the clustering algorithm and the random seed chosen (Schloss et al., 2011).

Several approaches have been proposed to improve the resolution of 16S data analysis beyond the standard 97%-similarity OTUs. Denoising algorithms exploit the predictable structure of certain error types to attempt to reassign or eliminate noisy reads (Huse et al., 2010; Quince et al., 2011; Rosen et al., 2012). These algorithms are widely used for identifying low-abundance (“rare”) species against a noisy background, often with the aim of improving estimates of ecological diversity. These objectives, however, remain very challenging due to issues that no denoiser can fully address. Any error model is necessarily approximate, and no denoising algorithm can deal with errors that are not adequately described by its error model; when calling low-abundance species this issue becomes particularly problematic. An alternative approach termed Distribution-Based Clustering (DBC; Preheim et al., 2013) aims to circumvent the limitations of conventional denoisers by using cross-sample comparisons, i.e. supplementing sequence information by ecological information (distribution of abundance across multiple biological samples). However, DBC as an OTU clustering algorithm also has important limitations: for low-count sequences, cross-sample comparisons necessarily become unreliable, and the execution time is prohibitively long even for moderately-sized datasets.

Here, I build on these methods to address a distinct question. Rather than trying to further improve the existing approaches to OTU clustering and rare species identification, I combine error-model based denoising and systematic cross-sample comparisons to resolve the fine (sub-OTU) structure of moderate-to-high abundance community members in 16S Illumina data. Importantly, this method does not rely on clustering similar sequences together. In this regard, it is similar to oligotyping (Eren et al., 2013), but the approach presented here does not require manual supervision and applies to an entire community rather than an isolated OTU. Using published data from a longitudinal study where the tongue community of two human individuals was sampled almost daily for several months (Caporaso et al., 2011), I

demonstrate that sequence similarity is a very poor predictor of ecological similarity, which I quantify for two bacteria as the correlation of their abundance time traces (“dynamical similarity”). Thus, most clustering-based approaches would erroneously group together bacterial subpopulations of high ecological diversity for this data set. However, a comparative analysis of the tongue communities of the two individuals also shows that when a pair of 16S tags is observed in both individuals, the dynamical similarity of the pair as measured independently in the two individuals is highly correlated. This correlation falls off substantially when sequences differing by 1 nt out of 130 are compared. In other words, the exact sequence of the 16S tag carried by a bacterial subpopulation is predictive of its ecology, while even 99.2% similarity between tags of different subpopulations is generally not predictive of dynamical similarity, as defined above. These results lend support to the recent idea that even a purely 16S-based study can provide insight into functional relatedness of community members (*cf.* PiCRUST, Langille et al., 2013), while also exhibiting and beginning to quantify the limitations of such methods. I demonstrate the applicability of my approach to a broad range of dataset types (host-associated longitudinal; environmental cross-sectional; mock community), providing examples when highly similar sequences were found to exhibit ecologically significant distinctions. Finally, I discuss how the single-nucleotide sub-OTU resolution of the presented method can provide new insights into factors shaping community assembly.

## 5.2 Cluster-free filtering

### 5.2.1 Data selection and quality filtering

I used the raw data from a published long-term longitudinal sampling from four body sites (right and left palm, gut (feces), and tongue) of one male and one female individual (Caporaso et al., 2011). In this study, the hypervariable region V4 of the bacterial

16S rRNA gene was amplified and sequenced with Illumina GA-IIx. For details on collection and sequencing see the original reference (Caporaso et al., 2011). Quality-filtered data published with that study is available at MG-RAST:4457768.3-4459735.3 and is sufficient to reproduce the results presented below using the analysis scripts provided with the software. However, to investigate the performance of our filtering approach at different quality filtering settings, for this work I used the demultiplexed, but not quality-filtered FastQ data, kindly provided by the study authors. I split this data into per-sample FastQ files using a custom MatLab script and subjected it to minimal quality filtering using USEARCH v.7.0.1090 (Edgar, 2010), truncating reads at Phred quality score 2 (other thresholds were also evaluated; see Fig. 5.S4), trimming to a fixed length of 130 nt and eliminating reads with ambiguous characters (N). In addition, I removed reads with expected number of base call errors exceeding 1 (maxEE parameter in USEARCH). This criterion only eliminated 1% of trimmed reads. Notably, my approach does not rely on assumptions about a maximum number of errors in a read. Finally, to facilitate cross-sample comparisons, I compiled a library of all 1.4M unique reads ever observed and a global table listing the abundances of each sequence across samples. This was done using dereplication capabilities of USEARCH and a custom Perl script (**mergeSeqs.pl**). This script and others referenced in **bold** below are freely available at <https://github.com/hepcat72/CFR> and were implemented in collaboration with Robert Leach. Finally, I normalized the abundance table to  $2.4 \cdot 10^4$  total reads per sample, to correct for varying sample size.

Read quality varied across lanes, so the number of reads after quality filtering was highest in a subset of tongue and fecal samples. In this work, I focused primarily on the tongue samples, as these come closest to probing the internal dynamics of a community living in a well-defined location on the body; however, the analysis of fecal samples supports the same conclusions and is presented in Fig. 5.S11.

Tongue samples were distributed over two lanes. The lane 6 samples from the male subject from day 65 onwards (314 consecutive samples covering a period of 355 days,  $2.4 \pm 0.4 \cdot 10^4$  reads in quality-filtered samples before normalization) had approximately 4-fold more reads than those from the female subject and from days 1-64 of the male subject (all on lane 5). Consequently, the analysis below uses the data from the male subject from day 65 onwards, and, for the comparative analysis of the two individuals, also the 135 samples collected from the female subject. The early samples from the male subject (days 1-64) are only used for illustrative purposes (Fig. 5.3D).

To demonstrate the broad applicability of the method presented below I also employed other published data (Figs. 5.S7 and 5.S11); the data is described in the corresponding legends.

### 5.2.2 Cluster-free filtering

Clustering can be a useful strategy to coarse-grain 16S data while also reducing noise, but if sequencing noise is low enough, such coarse-graining may not be necessary. At low noise, each community member is predominantly represented by the same 16S sequence, surrounded by a cloud of low-abundance error sequences with the structure of the cloud determined by reproducible error rates. Prior work has described such error clouds in the data (Quince et al., 2009; Edgar, 2013), and the assumption that high-abundance sequences are likely to be error-free is used in several rank-based denoising and chimera-checking algorithms (SLP, Perseus, Uchime de novo, Uparse, AbundantOTU).

The treatment of reads that are very similar to high-abundance sequences is different across existing algorithms. For example, SLP (Huse et al., 2010) would consider any read differing by a single nucleotide from a higher-abundance sequence (its “direct neighbor” in sequence space) as an error. However, some of these reads may actually represent true community members (Preheim et al., 2013). A more nuanced

treatment can accept a sequence as likely to be real if its observed abundance is highly unlikely to have arisen in error, given some assumptions about error rates. This idea is at the foundation of error-model based denoising. It was used in AmpliconNoise (Quince et al., 2011), and its recent implementation in DADA (Rosen et al., 2012) makes DADA, to my knowledge, the best denoiser currently available.

However, no error model is perfect, and for all denoisers, errors not explicitly described by their model are labeled as true sequences. Thus a denoising algorithm alone is insufficient for achieving sub-OTU resolution: if two close sequences that would fall within a single OTU are both identified as “probably real”, one of these could still be an error. In the context of a single sample, confidently resolving close sequences as independently real requires a different experimental technique (Faith et al., 2013) or a complete, high-quality reference database of all bacteria in the sample, which in practice is available only for mock communities.

It is possible to resolve this problem in the framework of standard 16S experiments through a comparison of multiple samples, either longitudinal or cross-sectional (Preheim et al., 2013). As an example, Fig. 5.1B shows the abundances of the 10 highest-abundance direct neighbors of the overall top sequence of the tongue community, Seq. #1, for a representative set of 100 consecutive samples. We see that three specific direct neighbors are strongly and consistently overrepresented compared to the other neighboring sequences and, more importantly, exhibit a dynamical behavior of their own (consider, for example, the 3rd most abundant neighbor). This has a clear interpretation (Fig. 5.1C): these three sequences must belong to other, fairly abundant bacterial subpopulations, possibly related to Seq. #1, but distinct and with their own dynamics.

To achieve sub-OTU resolution, I adopt precisely this strategy, namely a cross-sample correlation analysis of individually denoised samples. Which denoiser should we use? DADA would be an excellent option; however, its estimated execution time

on the tongue dataset used here is  $2.3 \cdot 10^5$  sec (see Sec. 5.A.7). This is largely due to its exact treatment of probabilities, critically important for the processing of sequences with an abundance of just a few counts. However, for such sequences the imperfections of the error model become non-negligible and cannot be controlled, since cross-sample comparisons are interpretable only for sequences with sufficient abundance. I therefore designed a new, simplified denoiser. This algorithm, described below, takes two orders of magnitude less time to execute, yet for sequences of moderate abundance considered here achieves performance equal to DADA, as demonstrated using mock community data (Table 5.S2).

### 5.2.3 Cluster-free filtering — the denoiser

For 16S data obtained using the Illumina platform, the main sources of errors are PCR substitutions, PCR chimeras, and substitution errors due to Illumina base call errors. Of these, the substitution errors are responsible for generating the largest number of unique sequences (Fig. 5.S2; see also Edgar, 2013) and have the most predictable structure: their rates can be estimated directly from the data. To do so, I considered the error clouds around the top 10 sequences by overall abundance (in all tongue samples combined). Assuming that most of these sequences are in fact errors, I determined the rates of specific one-nucleotide substitutions (**errorRates.pl** with z-score threshold of 2; see Sec. 5.A.2). These inferred rates were consistent across error clouds observed in the data (Fig. 5.S3), with the average error rate of only 0.10% per nucleotide (Sup. Table 5.S1; compare with Quince et al., 2011, Table 2). I then used these error rates to predict the expected abundance of any given sequence if its presence were entirely due to independently generated sequencing errors of its more abundant neighbors (the “null model”; Fig. 5.S5; **nZeros.pl**). Sequences whose abundance exceeded a threshold of 10 counts and the null-model prediction by at least 10-fold (very conservative filtering parameters), were marked as “candidates”;

their presence cannot be explained as an error within a substitution-only error model (`getCandidates.pl`). Candidate sequences include true biological 16S sequences, but also sequences that arose through a different type of error, most notably PCR chimeras. Chimeric sequences were identified using UCHIME denovo (Edgar et al., 2011) on the pooled data from all samples. Most such sequences were already eliminated by the abundance threshold requirement: if I relax the abundance threshold to 2 (excluding singletons only), I find that the chimeras detected by UCHIME, when present in a sample, have abundance under 10 counts in 95% of cases. However, chimeras of highly abundant parents reproducibly occur at higher abundances (Haas et al., 2011) and are filtered at this step.

Candidate sequences that remained after filtering chimeras were labeled “real”. My highly conservative filtering criteria allow me to assume that this list contains only true biological sequences, i.e. that there are no false positives (*cf.* Sup. Table 5.S2), except possibly those due to some exceptionally frequent errors not described by my error model (see Sec. 5.A.4). This stringency comes at the expense of low-abundance false negatives (true biological sequences labeled as “possible noise”). My strategy is to retain all sequences marked “real” in 2 or more samples (out of 507). This makes my denoiser specifically adapted to multi-sample analysis: in each sample, only high-confidence detections are identified, which is very fast, and then a liberal criterion applied across samples retains all sequences that ever generated a high-confidence detection, except sample singletons. In particular, I stress that my detection threshold of 10 counts is not equivalent to removing all sequences with abundance below 10; the only sequences excluded from consideration are those that never rise to 10 counts in the entire set of 504 tongue samples, or do so only once. For such sequences, the measured counts are dominated by detection and counting noise.

In the interest of speed, and to ensure the robustness of reported sequence abundance values with respect to the details of the error model, I did not attempt to



remap noisy reads to their most probable source. My approach relies on the accuracy of measurement of relative abundances of true sequences. The error remapping process modifies sequence counts in a way that depends on the assumptions of the error model, distorting the relative abundance values whenever neighboring sequences are incorrectly classified as “reals” or “errors”. In contrast, discarding noisy reads leaves the relative abundances intact, as long as the probability of making zero errors is approximately constant across all sequences. This assumption is much weaker than adopting a particular error model. I estimate the zero-error probability at  $\approx 85\%$  (see Sec. 5.A.2); in other words, discarding noisy reads leads only to a  $\approx 15\%$  loss of sequencing depth. If read remapping is desired, the analysis described below can be applied to DADA denoiser output.

Since non-identical reads are never clustered together, what I describe is a single-nucleotide resolution approach. The complete workflow of cluster-free filtering is outlined in Fig. 5.S6 and detailed in Sec. 5.A.5. The code is freely available at <https://github.com/hepcat72/CFF>.

### **5.3 Sequence similarity need not imply ecological similarity, and vice versa**

The starting point for my analysis is a global sequence abundance table listing the abundances of each unique 16S sequence across samples. I retained the 307 sequences that passed the multi-sample filtering algorithm described in Sec. 5.2, and thus putatively belong to bacteria present in the population at least part of the time. I denote these sequences by their overall abundance rank: Seq. #1, #2, etc. In this list, 184 pairs of sequences were direct neighbors in sequence space (Hamming distance 1). These pairs had 99.2% sequence similarity but were resolved by my criteria as independently present in the community. The population of bacteria sharing the exact

same sequenced fragment of the 16S gene (at 100% identity) is the smallest taxonomic unit resolvable by 16S analysis. For notational convenience, throughout this work I call it the “subpopulation” identified by a sequence.

In the standard approach to tag-sequencing data, it is assumed that sequence similarity of 16S hypervariable regions can be used as a proxy for phylogenetic, and therefore ecological relatedness. My filtering method, applied to time-series data, allows me to bypass this assumption and assess ecological relatedness independently, based on the similarity of time traces, since each distinct subpopulation will respond in its own way to variation in environmental conditions (Youngblut et al., 2013), causing the abundance time traces to be more or less correlated (or possibly anticorrelated; see Fig. 5.S8). Fig. 5.2A-C illustrates this by showing time traces (normalized counts versus observation day) for three examples of sequence pairs. We find that sequences differing by as little as 1 nucleotide (99.2% similarity) can be ecologically distinct as evidenced by their very different time series (Fig. 5.2A); see also VandeWalle et al. (2012). For comparison, Fig. 5.2B shows another pair of sequences, also with 99.2% sequence similarity but whose abundance time traces appear indistinguishable. The remarkable correlation between these two traces provides an internal control and demonstrates that the much lower correlation of traces in Fig. 5.2A cannot be explained by measurement error but reflects a true ecological difference. Note that the abundances of the two sequences shown in Fig. 5.2B are not equal, but occur with a highly stable ratio. This could reflect a stable difference in abundance of the bacteria they represent, but is more likely caused by differential amplification efficiency of these sequences by the PCA primers (Turnbaugh et al., 2010; Klindworth et al., 2013) and/or a different number of genomic 16S copies per cell (Tourova, 2003). Panels A-B show that sequence similarity need not imply ecological similarity. Finally, Fig. 5.2C illustrates that the converse is also true: sequences exhibiting identical time-dependence may have as little as 81% sequence identity.

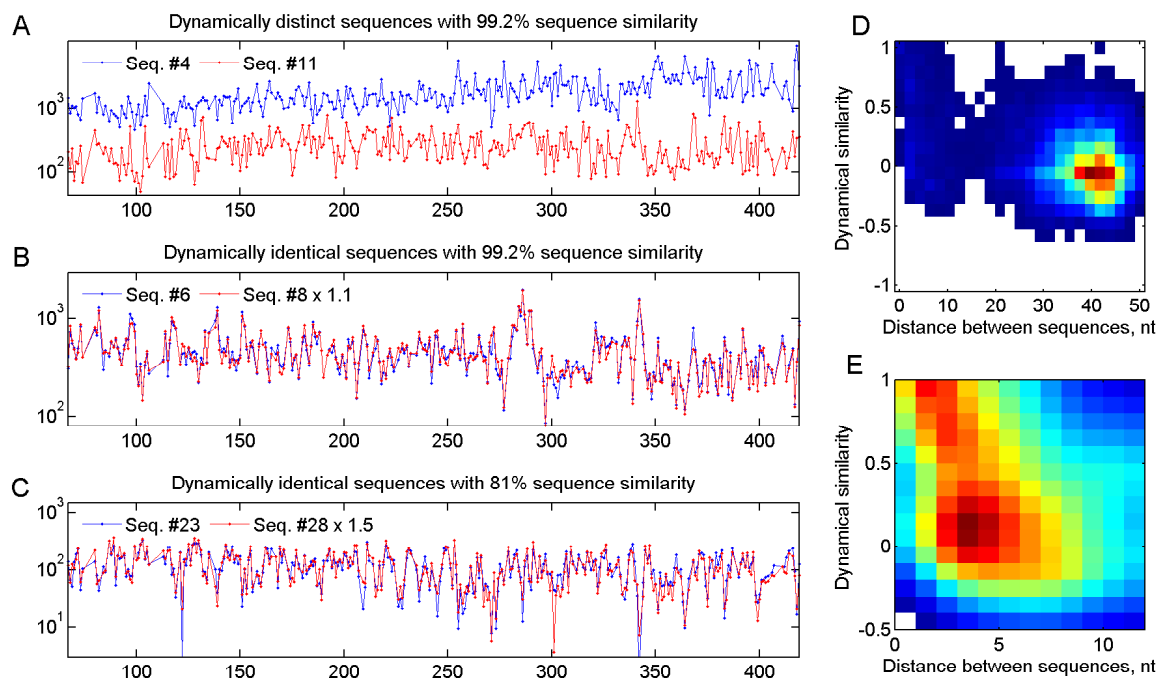


Figure 5.2: **Sequence similarity need not imply dynamical similarity, and vice versa.** Panels show sequence counts versus observation day, for days 65-420. **A:** Seq. #4 and #11, despite 99.2% sequence similarity, display significant differences in time dependence, indicating that these 16S tags belong to ecologically distinct bacterial subpopulations. **B:** For Seq. #6 and #8, 99.2% sequence similarity (1nt difference) is mirrored by near perfect correlation of time series. Red trace renormalized for best overlap. **C:** Seq. #23 and #28, with only 81% sequence similarity, nevertheless display near perfect correlation. Red trace renormalized for best overlap. **D:** 2D histogram of dynamical similarity (Pearson correlation of abundance traces, normalized by maximum expected correlation  $c_{\max}$ , see text) versus distance in sequence space (nt), for all pairs of the top 200 sequences (19900 data points). **E:** Zoom-in of D (1321 sequence pairs), showing the most similar sequences. Histogram smoothed for clarity.

To quantify the generality of these examples, it is useful to define a measure of the ecological similarity of the bacterial subpopulations represented by two sequences. A natural candidate metric is the Pearson correlation of the measured abundance traces. Note, however, that the maximum correlation one can expect between the time traces of two sequences depends on their abundance: for low-abundance sequences Poisson sampling noise becomes non-negligible and sets an upper bound on the correlation coefficient. I therefore define the “dynamical similarity” of two traces as the Pearson correlation of their abundance, normalized by their maximum possible correlation  $c_{\max}$ , computed as the correlation of the higher-abundance time trace with a Poisson-downsampled version of itself (see Sec. 5.B.2). For sequence distance, I use the Hamming distance between sequences after pairwise alignment (see Sec. 5.B.3). With these definitions, I can construct a 2D histogram of dynamical similarity *vs.* distance in sequence space for all sequence pairs constructed from the top 200 real sequences (Fig. 5.2D). As expected, most sequence pairs exhibit no significant dynamical similarity and are also far apart in sequence space, but a subset of closely similar sequences appears to display some degree of anticorrelation between the two measures. Zooming in on this region (Fig. 5.2E) makes this anticorrelation more apparent; however, even when restricted to the subset shown in Fig. 5.2E, the correlation coefficient remains weak ( $R = -0.3$ ). Sequences separated by up to 6-7 nt (95% sequence similarity) tend to be dynamically similar, the effect increasing for smaller distances, but this general trend is very loose and is not a reliable predictor of similarity for any particular pair. This result was not unexpected, and is frequently used in arguments against over-reliance on the 16S gene sequence (see, for example, Prosser et al., 2007), in favor of methods providing functional information, such as shotgun metagenomics. The novelty of Fig. 5.2E lies in the fact that it was obtained entirely within the framework of 16S tag sequencing methodology.

## 5.4 Cluster-free filtering can resolve distinct subpopulations with high dynamical similarity.

As explained in the previous section, 16S tags with low dynamical similarity clearly derive from distinct bacterial subpopulations, even if the sequences are themselves highly similar. I now consider pairs of sequences with highly correlated time traces such as observed in Fig. 5.2BC. Such correlated pairs could derive from the same bacterial cells (as multiple genomic copies of the 16S gene, or as exceptionally common PCR errors not included in our model). Alternatively, they could derive from distinct bacterial subpopulations that either occupy the same ecological niche or engage in a strong obligate symbiosis. Such pairs are thus of significant ecological interest, provided it can be shown that the sequences actually derive from different bacterial cells. In this section, I demonstrate that cross-sample correlation analysis can, in some cases, successfully make this subtle distinction between same-cell or different-cell sources.

To draw this distinction, I make use of the following observation. The abundance ratio of two sequences that derive from the same bacterium is set by some sample-independent parameter (e.g. involving differential amplification efficiency, 16S copy number, and/or PCR error rate); therefore, any fluctuation in their abundance ratio is due to measurement noise, and must be uncorrelated between samples. Any statistically significant time (or location; see Sec. 5.C.4) correlation of abundance ratio fluctuations, e.g. in consecutive (or proximate) samples, is therefore strong evidence that the two sequences are at least partially contributed by physically distinct subpopulations.

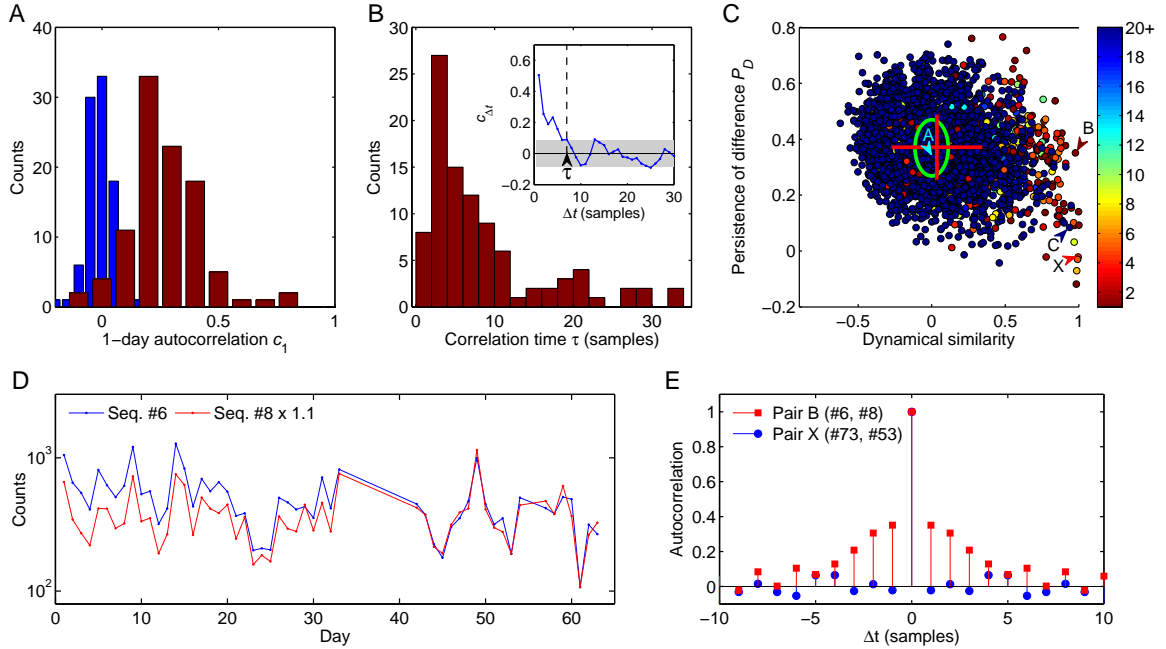
For this approach to succeed, the dynamics of individual subpopulations must be slow enough to allow correlations between consecutive samples to be observed. I therefore began by computing, for each of the top 100 sequences, the autocorrelation

function  $c_{\Delta t}$ , defined as the correlation between abundance fluctuations in samples separated by  $\Delta t$  time points, and normalized so that  $c_0 = 1$  (for simplicity, I treat samples as though they were equally spaced in time, which is approximately correct; the mean separation between samples was 1.1 days). The environment experienced by tongue microorganisms changes frequently, and one might have expected that daily sampling would probe the space of possible community states, but provide little information about community dynamics as these would occur on a faster time scale. Surprisingly, I found the time dependence of most sequences in the top 100 to have a significant autocorrelation despite the relatively low sampling rate (Fig. 5.3A). Although conditions on the tongue make fast abundance changes possible, as evidenced by the large, rapid fluctuations in Fig. 5.2A-C, I found the correlation time for the top 100 sequences to be surprisingly long, typically 2-4 days but often longer (Fig. 5.3B), sometimes exceeding a month (Fig. 5.S10).

These multi-day autocorrelations make it plausible that for physically distinct subpopulations, the fluctuations of their abundances relative to each other could be slow enough to be detectable even if their ecology is similar. Consider two sequences  $A$  and  $B$  whose abundance time traces are highly correlated. Denote by  $n_A(t)$ ,  $n_B(t)$  the two traces renormalized to the same mean for best overlap, as in Fig. 5.2BC, and let  $\Delta(t)$  be their fractional difference in a given sample (a quantity more robust to noise than the naïve abundance ratio):

$$\Delta(t) = \frac{n_A - n_B}{(n_A + n_B)/2}.$$

If  $n_{A,B}(t)$  reflects abundances of two distinct subpopulations, then  $\Delta(t)$  can be expected to exhibit an autocorrelation on par with that observed for the individual sequences. Intuitively, if on day 1, subpopulation  $A$  is, say, 10% more abundant than  $B$ , and the dynamics of both are slow, then  $A$  is likely to maintain its lead on



**Figure 5.3: Dynamical similarity versus 16S similarity.** **A:** 100 most abundant sequences of the population exhibit significant autocorrelation. Histogram of autocorrelation coefficients of sequence abundance for consecutive samples (red), and after randomly permuting sample labels (blue). **B:** Histogram of autocorrelation times of 100 most abundant sequences. I define the autocorrelation time  $\tau$  as the time shift  $\Delta t$  at which the autocorrelation function  $c_{\Delta t}$  falls below the threshold of statistical significance as illustrated in the inset (see Sec. 5.C.1). For 19 sequences the autocorrelation time exceeds 35 days (not shown). **C:** Persistence of difference  $P_D$  for all pairs of sequences from the top 100, plotted against the correlation of their abundances (normalized by maximum expected correlation  $c_{\max}$ ). Green ellipse indicates mean and standard deviation for the null model obtained by reversing in all pairs the time order for one of the sequences. Most pairs are consistent with the null model, except for a broadening of the correlation coefficient distribution (mean and standard deviations indicated by the red cross). Pairs to the right of the plot are dynamically similar (strong abundance correlation), often accompanied by high sequence similarity (color code indicates Hamming distance between aligned sequences in the pair; see Sec. 5.B.3). Of these, a subset (bottom right) also exhibit weak or negligible persistence of difference. These pairs, such as pair “X”, most likely correspond to genomic 16S variants found within a single bacterium. Letters A-C identify pairs shown in Fig. 5.2A-C. The large persistence of difference identifies pair B as coming from distinct bacterial cells. **D:** Sequence counts versus observation day for early samples of Seq. #6 and #8 (99.2% similarity), normalized as in Fig. 5.2B but excluded there due to relatively poor sequencing depth. The clear separation observed prior to day 40 confirms that these two sequences are contributed at least in part by distinct bacterial subpopulations. **E:** Autocorrelation functions of the relative difference  $\Delta(t)$  for two pairs identified in (C): pair “B” (red squares; high  $P_D$  indicative of distinct bacterial cells) and pair “X” (blue circles; low  $P_D$  indicative of 16S variants found within a single bacterium).

day 2. In contrast, if the two sequences are genomic variants contained within the same bacterium, then any difference between  $n_A(t)$  and  $n_B(t)$  must be due to measurement noise, and  $\Delta(t)$  will be uncorrelated between samples. I therefore introduce the persistence of difference  $P_D$  as the 1-day autocorrelation coefficient of  $\Delta(t)$ :

$$P_D = \frac{\langle \Delta(t)\Delta(t+1) \rangle}{\langle \Delta(t)^2 \rangle},$$

where angular brackets denote averaging over time.  $P_D$  characterizes the persistence of abundance fluctuations of two sequences relative to each other. For sequences arising from the same cells,  $P_D$  must vanish. Any pair of sequences exhibiting a statistically significant  $P_D$  must be contributed, at least in part, by two physically distinct bacterial subpopulations. Note that the absolute abundance of a sequence may change dramatically between days (e.g. more favorable conditions can cause both subpopulations to proliferate quickly), but the normalization of  $\Delta(t)$  makes  $P_D$  insensitive to such overall correlated behavior.

Summarizing the above, we have the following expectation for  $P_D$ : For a randomly chosen pair of sequences, with insignificant dynamical similarity,  $P_D$  should be significantly non-zero (due to the slow dynamics of the individual subpopulations; see Sec. 5.C.3), and form a unimodal distribution consistent with the null model of unrelated subpopulations. In contrast, pairs displaying high dynamical similarity come in two types, and the persistence of difference  $P_D$  should display a bimodal distribution: pairs of sequences found within the same bacterial cell will have vanishing or insignificant  $P_D$ , while pairs belonging to distinct subpopulations will likely exhibit a persistence of difference comparable with the null model prediction.

This is precisely what is observed. Fig. 5.3C shows, for all sequence pairs constructed from the top 100 sequences, a scatter plot of their persistence of difference  $P_D$  versus dynamical similarity as defined previously (the normalized Pearson correlation of their abundances). The mean and standard deviations of the distribution predicted by the null model (unrelated subpopulations) are indicated by the green ellipse, and



were computed directly from the data by reversing in all pairs the time order for one of the sequences. The mean and standard deviations of the actual data are indicated by the red cross. We find, as expected, that the  $P_D$  score of dynamically dissimilar sequence pairs is unimodal and consistent with the null-model prediction. In contrast, the  $P_D$  score of dynamically similar pairs exhibits the predicted bimodality (right side of the plot), with a subset exhibiting weak or negligible persistence of difference (bottom right). As explained above, I interpret these low- $P_D$  pairs as corresponding to genomic 16S variants found within a single bacterium. Letters A-C identify pairs shown on Fig. 5.2A-C. Note that the strong persistence of difference identifies the pair “B” as being contributed, at least in part, by distinct bacterial cells, despite 99.2% sequence similarity and an almost perfect correlation of abundances (Fig. 5.2B). Conversely, the low- $P_D$  pair “C” (with only 81% sequence similarity) likely corresponds to an example of two dissimilar 16S genes contained within a single bacterium. Note the enrichment of pairs with high sequence similarity among the dynamically similar pairs, as indicated by the color code (compare with Fig. 5.2D).

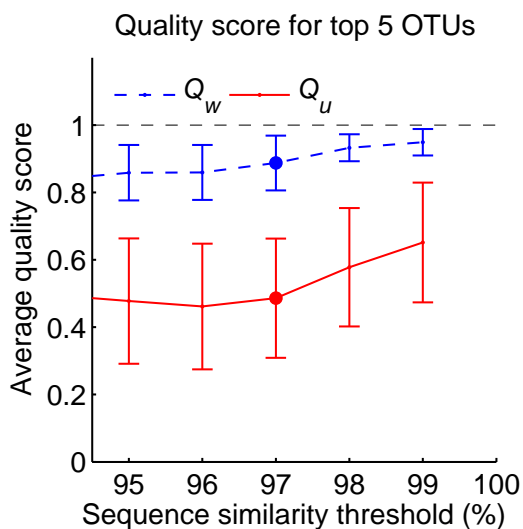
Remarkably, in the case of pair “B”, the conclusion of distinct bacterial subpopulations drawn from Fig. 5.3C can be confirmed directly. Panel D shows the time traces of this pair for days 1-64 (normalization as in Fig. 5.2B). Due to the relatively poor sequencing depth in these early samples, they were not included in Fig. 5.2B. The clear separation observed prior to day 40 provides an independent confirmation of our conclusion. I stress that these data were not used in the analysis presented in Fig. 5.3C, but the sensitivity of the autocorrelation method was sufficient to identify these sequences as deriving from physically distinct cells based solely on the data shown in Fig. 5.2B. The autocorrelation function of the fractional difference  $\Delta(t)$  for this pair is shown in Fig. 5.3E. I have verified that the persistence of difference for this pair does not change significantly if any window of 100 consecutive samples is used instead of the full time series (data not shown).

## 5.5 Clustering reads into OTUs vastly underestimates ecological richness

Figs. 5.2A, 5.S7, 5.S10, and 5.S11 provide examples of some fine features that standard OTU-based methods would fail to detect, but which become accessible with cluster-free filtering. I now ask whether such cases are the exception or the rule. For a given sequence similarity threshold, one can define, for each of the most abundant sequences, its would-be OTU, namely the ensemble  $\{S_i\}$  of all “real” sequences within the chosen similarity threshold. I construct the time trace of the abundance of this OTU as the sum of the abundances of all its members. I can now ask: how representative is this time trace of the true behavior of the member sequences? Let  $\{c_i\}$  be the correlation coefficients between time traces of individual members and the OTU itself, normalized to the maximum expected correlation as before. I define unweighted and weighted OTU quality scores  $Q_u$  and  $Q_w$  as, respectively, the simple average of  $\{c_i\}$ , and an average weighted by the abundance of the member:

$$Q_u = \frac{1}{K} \sum_i c_i \quad \text{and} \quad Q_w = \frac{\sum_i N_i c_i}{\sum_i N_i}$$

Here  $K$  is the number of subpopulations in the OTU and  $N_i$  is the average abundance of member  $i$ . The weighted quality score  $Q_w$  is always larger, because the most abundant sequence dominates the sum and so is better correlated with the OTU trace. Thus  $Q_w$  tells us how representative the OTU is of its most abundant member. The unweighted quality score  $Q_u$  tells us how diverse is the group of subpopulations lumped together into an OTU. If the sequences grouped into an OTU are all dynamically identical (are Poisson-resampled versions of each other at different abundances), both quality scores will be close to 1. If the OTU is dominated by one subpopulation, with other members dynamically different but very low in abundance, we will have



**Figure 5.4: Clustering reads into OTUs underestimates dynamical diversity.**

Average quality score for OTUs assembled around the top 5 sequences (defined as the ensemble of “real” sequences within a given sequence similarity threshold), as a function of similarity threshold. Error bars are standard deviations across 5 considered OTUs. Weighted quality score  $Q_w$  (dashed line; see text) is high, indicating that the OTU time traces are representative of the time traces of their most abundant members. However, the unweighted score  $Q_u$  (solid line) is dramatically lower, indicating that OTUs group together sequences with very different time traces. Thus OTUs combine sequences with high dynamical diversity. The commonly used “species-level” similarity threshold of 97% is highlighted.

$Q_w \approx 1$ , but  $Q_u \ll 1$ . Finally, if the OTU contains several dynamically distinct subpopulations at comparable abundances, both quality scores will be low.

The average quality scores for OTUs assembled around the top 5 sequences are presented in Fig. 5.4 as a function of sequence similarity threshold. The high weighted quality score  $Q_w$  means that an OTU time trace is, on average, fairly representative of its most abundant member. The unweighted score  $Q_u$  is, however, dramatically lower, indicating that the OTUs group together sequences from subpopulations with high dynamical diversity.

These quality scores rely on abundance time-trace correlations, which become contaminated with noise for low-abundance sequences. For the purposes of Fig. 5.4, to apply these definitions conservatively, I therefore only included high-abundance members of the OTU, considering only sequences from the top 200 by overall abundance. Further, my cluster-free filtering method also has finite resolution, as the sequences we analyze are only 130nt long and may derive from distinct 16S genes, implying some unresolved diversity. This limited resolution leads to an artificial inflation of OTU

quality scores as the similarity threshold approaches 100%. For both these reasons the true quality scores of OTUs are likely even lower (see Sec. 5.D.1).

## **5.6 Exact tag sequence identity is substantially more predictive of subpopulation dynamics than 99.2% sequence similarity**

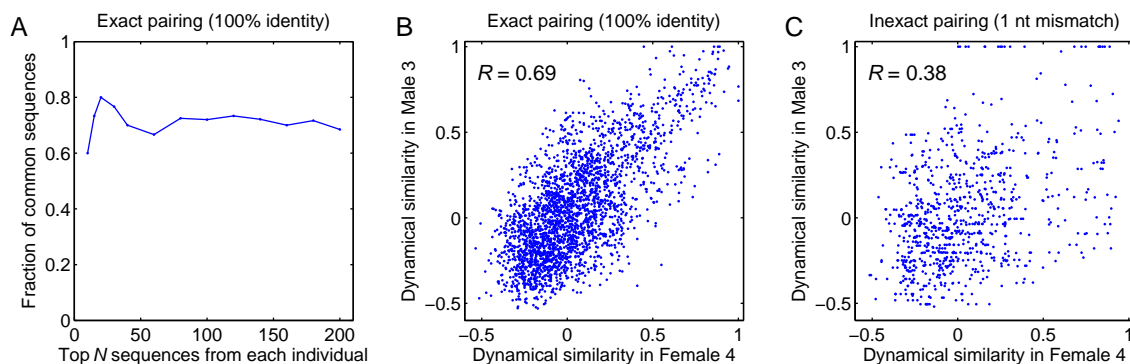
The fact that tag sequence similarity within the 16S gene is only loosely correlated with dynamical similarity (Fig. 5.2E) was not unexpected (see, for example, Prosser et al., 2007, and references therein). At a neutral mutation rate of order  $10^{-9}$  per base pair per generation (Ochman, 2003), an average difference of a single nucleotide out of 100 would already require divergence for millions of generations. A more precise estimate of divergence time should take into account the possibility of horizontal gene transfer, whose rate in an ecologically relevant setting is hard to assess. However, it is clear that, generically, two bacteria that differ by even 1 nt in a particular hypervariable region of the 16S gene likely diverged a long time ago. These bacteria are likely to also differ elsewhere in their 16S gene, and to carry even more significant differences in functional parts of their genome.

In contrast, what if we consider two bacteria whose sequenced portions of their 16S genes are identical? Since the length of the sequenced fragment is small (typically  $\sim 100$  nt) and the mutation rate is low, these bacteria could still have diverged a very long time ago (Lukjancenko et al., 2010). However, depending on circumstances, the actual time since the last common ancestor may be much shorter. For example, consider two communities that frequently exchange members. If two bacteria drawn from two such communities are 100% identical in their 16S tags, a likely explanation for this identity is a recent exchange event, in which case the entire genomes of these bacteria

may be close to identical. I conclude that in the presence of strain exchange between communities, exact sequence identity and near-identity may have fundamentally different implications. The study of Caporaso et al. sampled the tongue microbiota of two cohabiting individuals (Rob Knight, personal communication), and so strain exchange is likely to be a highly significant factor (Song et al., 2013). I hypothesized, therefore, that these communities would share some non-negligible number of subpopulations at 100% sequence identity, and that these common subpopulations might have similar ecology in both communities.

I began by identifying the fraction of common 16S sequences in the list of the top  $N$  for each individual (at 100% sequence identity). Based on the strain exchange hypothesis, I expected to see some matches, but still was surprised to find this fraction to be as high as 73% (Fig. 5.5A). Such a high proportion of perfect matches provides strong evidence that the identical sequences found in these two communities most likely diverged from a common ancestor more recently than any pair of close, but non-identical sequences within the same community. The same conclusion is supported by the analysis of fecal samples from the two individuals (Fig. 5.S11).

I then considered the 73 sequences that were found among the top 100 of both individuals and asked whether the behavior of these subpopulations was predominantly shaped by their presumed common origin (causing them to be similar) or by local adaptation (causing them to diverge while leaving the 16S region intact; see Lukjancenko et al. (2010)). To this end, for each pair of sequences  $(i, j)$  drawn from this list, I measured their dynamical similarity independently in the two datasets;  $S_{ij}^M$  for the male and  $S_{ij}^F$  for the female. If the effect of local adaptation were dominant, then the exactness of a match of 16S sequences would not carry much information: the ecologies and genomes would be no more similar between 100%-identical partners in the two communities than between any other sequences within the same bacterial “species” (OTU); this scenario is implicitly assumed by taxonomy-based methods.



**Figure 5.5: Comparative analysis at 100% sequence identity of oral community composition in two cohabiting individuals reveals shared subpopulations.** **A:** Fraction of shared 16S sequences, defined as the fraction of common tags (at 100% sequence identity) among the most abundant  $N$  sequences in each of the two individuals, plotted as a function of  $N$ . **B:** Scatter plot of the dynamical similarity of pairs of common sequences, as measured independently in the two individuals, for all possible pairs among the 73 common sequences shared within the top  $N = 100$ . **C:** Same as (B), but with intentionally inexact pairing of sequences across individuals (each sequence is mapped to a partner differing by exactly 1 nt). Despite 99.2% sequence similarity of such pairs, allowing the 1nt mismatch significantly decreases the degree to which dynamical similarity as observed in the two individuals is correlated.

Alternatively, if the ecology were determined primarily by the shared recent ancestor, then identical 16S tag sequences in the two communities would correspond to bacterial subpopulations with almost identical genomes. In this scenario, provided local adaptation did not modify the ecology of a subpopulation significantly,  $S_{ij}^M$  and  $S_{ij}^F$  should be strongly correlated, and unlike the first scenario, this correlation would be noticeably degraded for any less than 100% sequence identity. The latter is indeed what is observed (Fig. 5.5B-C). Fig. 5.5B demonstrates that subpopulations identified by the exact same 16S tags in the two individuals are dynamically similar; see also Figs. 5.S11D and 5.S12. To obtain Fig. 5.5C, I constructed an “inexact pairing” of sequences between individuals, whereupon each sequence from the top 100 in the female individual was matched to the highest-abundance sequence from the top 100 in the male individual that differed from it by exactly 1 nucleotide, when such a match existed. This matching corresponds to 99.2% sequence identity, yet already

substantially degrades the correlation between  $S_{ij}^M$  and  $S_{ij}^F$  (Fig. 5.5C). I conclude that 100% identity of tag sequences has qualitatively different implications from even 99.2% near-identity.

## 5.7 Discussion

In this work, I have demonstrated that cross-sample correlation analysis of denoised 16S data can be exploited to achieve sub-OTU resolution. The cluster-free filtering approach I presented reliably identified up to 20 distinct subpopulations within standard 97% similarity OTUs, and a comparative analysis of oral communities of two cohabiting individuals demonstrates that most such subpopulations are shared between the two communities. Furthermore, subpopulations identified by the exact same 16S tags in the two individuals are dynamically similar, whereas even a single nucleotide mismatch is enough to degrade this similarity. Overall, this analysis shows that coarse-graining sequence data into OTUs is not essential for ecological applications of 16S tag sequencing methodology.

The approach presented here combines two novelties. First and foremost, I do not cluster similar sequences together. Regrettably, in the literature the term “clustering” has multiple meanings. Most denoising algorithms aim to assign erroneous reads to their most likely source, to make the abundance estimates of true sequences more accurate. The same term “clustering” is used both for this read remapping and for merging multiple true sequences into a single OTU. However, these two practices are fundamentally different. Read remapping constitutes data denoising; as such, it is always advantageous, can be done in a principled way, and can be evaluated against an objective standard of performance. Adding it to my approach would likely somewhat improve the results. In contrast, OTU clustering is a form of data coarse-graining, and the optimal degree of coarse-graining is necessarily application-dependent. Im-

portantly, for some applications it may not be necessary or desirable. When studying coarse features of community composition and dynamics, e.g. comparing communities across habitats (Costello et al., 2009; Huttenhower et al., 2012) coarse-graining is appropriate. For example, metrics of community comparison such as UniFrac (Lozupone and Knight, 2005) are widely used precisely because, by construction, they are not sensitive to OTU sub-structure. However, when studying subtle differences between broadly similar communities, e.g. samples from similar habitats or repeated sampling of the same habitat, the sub-OTU structure becomes a valuable source of insight. This is the intended application for cluster-free filtering approach. Although I focused on longitudinal Illumina data, the denoising algorithm I developed does not assume short read length or low error rate and is directly applicable to a wide range of dataset types (see examples in Fig. 5.S7 and 5.S11), provided the error structure is consistent across samples (Preheim et al., 2013). I expect this approach to be useful for investigating the structure and dynamics of discrete community subtypes such as those observed in the vaginal community (Huttenhower et al., 2012).

The second novelty was to exploit the quantitative advantage offered by multi-sample (time-course or cross-sectional) data. Since the copy number of the 16S gene carried by a bacterium is typically unknown (Tourova, 2003), and the PCR amplification bias among different 16S fragments can sometimes reach orders of magnitude (Turnbaugh et al., 2010; Klindworth et al., 2013), the 16S data from a single sample carries very little quantitative information about community composition. In contrast, the ratios of sequence abundance are highly informative and can be measured very precisely, as demonstrated in Fig. 5.2B,C. Recently, time-course data collection has been gaining in popularity, as it was recognized that such experiments can offer valuable insight into community dynamics (Shade et al., 2013, and references therein). However, another major advantage of such datasets, namely that changes in sequence abundance ratios can be measured much more accurately than absolute abundances,



is only beginning to be explored. For this work, time-series data provides a context where sub-OTU resolution acquires its full power. Specifically, I have shown that cross-sample comparisons enable us to decouple sequence similarity from dynamical similarity while remaining fully within the framework of 16S tag sequencing. High-quality reference databases can complement this approach to facilitate paralog identification. The basic methodology described here should also be extendable to other marker genes.

The new approach described in this work is not a replacement for OTU clustering; it discards low-abundance sequences and so is unsuitable for studies of population-level alpha- or beta-diversity. However, the novel statistical and computational techniques I presented allow full utilization of the quantitative information carried by sequences with a moderate-to-high abundance. This has promising applications for the study of factors affecting community assembly. As discussed above, sub-OTU resolution can provide insight into the prevalence of strain exchange between communities, invasion / extinction dynamics of OTU subpopulations, and the time scale of ecological divergence relative to sequence divergence. In addition, the dynamics of individual-specific subpopulations could help characterize the role of host genetics or the host immune system on shaping the community, particularly in the context of highly controlled experiments with germ-free animals.

# Technical details

## 5.A Supplementary methods.

### Cluster-free filtering: details and applications

In this section, I illustrate the idea of error-model-based denoising (see also the introduction in Rosen et al., 2012) and give a detailed description of the simple denoiser I designed for this work. I then describe the workflow of an open source software package I created with Robert Leach to implement this denoiser, and compare its performance on mock community data with DADA (Rosen et al., 2012). Finally, to illustrate that cluster-free filtering approach is not restricted to longitudinal Illumina data, I provide an example of its application to a very different dataset, specifically 454 sequencing data from a cross-sectional environmental sampling performed by Preheim et al..

#### 5.A.1 Motivation: sequencing noise is low

Clustering can be a useful strategy for filtering noise by coarse-graining data. However, such coarse-graining may not be a necessity: if the noise level is low, as suggested by known estimates of PCR and sequencing error rates (see, for example, Quince et al., 2011), then we can avoid clustering, since we expect each community member to be predominantly represented by the same 16S sequence.

I begin by illustrating this idea using the tongue microbiome data of Caporaso et al.. Since the tongue community is relatively stable (Costello et al., 2009), the low-noise scenario would predict that certain specific sequences should consistently dominate in each sample. Alternatively, if the noise were high, then the high-abundance community members would be represented by clouds of similar reads, none of which would clearly dominate.

To show that the data of Caporaso et al. supports the first (low-noise) scenario, I identified the top 5 sequences by overall abundance. These sequences were strongly different (Fig. 5.S1, inset), corresponding for the most part to bacteria from different phyla: in decreasing order of abundance, these were *Neisseria sp.* (phylum *Proteobacteria*, class *Betaproteobacteria*), *Haemophilus sp.* (phylum *Proteobacteria*, class *Gammaproteobacteria*), *Fusobacterium sp.* (phylum *Fusobacteria*), *Streptococcus sp.* (phylum *Firmicutes*), and *Prevotella sp.* (phylum *Bacteroidetes*). (Taxonomy assigned by a BLAST search (BLASTN 2.2.22, matrix (1, -1), gap/extension penalty (5, 2)) against GreenGenes database; DeSantis et al., 2006. All 5 sequences had a match with 100% identity over 100% of sequence length.) The sample-by-sample rank of these overall top 5 sequences was consistently in the top 10. I stress that the goal of Fig. 5.S1 is not to characterize the temporal stability of community composition (Costello et al., 2009, previously characterized, for example, in); rather, it serves to show that the community members that correspond to these highest-abundance tags are consistently represented by the same 130nt sequence (at 100% identity) across all samples. In other words, despite the presence of noise in the data, 100% sequence identity is not an unreasonable criterion: the error rate is low enough that the error-free sequence dominates over the “error cloud” of its variants (Edgar, 2013). This key observation is the foundation of the approach described in this work.

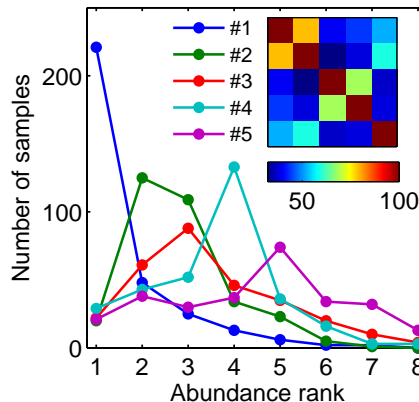


Figure 5.S1: The distribution of ranks for the top 5 sequences over all samples. Inset: pairwise sequence similarity (%). The top sequences are strongly distinct and their rank is consistent across samples.

### 5.A.2 Estimating rates of one-nucleotide substitutions

To estimate the rates of substitution errors observed in data after quality filtering, I used the “error clouds” around the high-abundance sequences in the dataset. Since all sequences were trimmed to a length of 130nt, each “mother” sequence has 390 direct neighbors in sequence space (Hamming distance = 1). For very high-abundance sequences such as Seq. #1, all 390 neighbors were observed in at least one sample of the time series. The time series of their abundances, normalized to the abundance of Seq. #1, is shown in Fig. 5.S2. For this figure, the neighbors were ordered by the type of substitution that differentiates them from the mother sequence, and, within these categories, by the position of the differing nucleotide along the sequence. We see that, with a few exceptions (most notably the three neighbors also shown in Fig. 5.1B), the abundance of a given neighbor is a constant fraction of the abundance of the mother sequence. This is precisely what we expect for neighbors that arise as PCR or sequencing errors of the mother sequence, and the abundance ratio is then the probability of that particular error.

We see that the error rate is set primarily by the type of substitution, and does not exhibit significant dependence on the position along the sequence. For long reads,

we would likely have seen an increase in error rates towards the end of the sequence, but these sequences are only 130nt long, well within the capabilities of accurate base-calling of the Illumina platform. I can therefore assign probabilities to substitution errors based solely on the substitution type (which nucleotide was replaced by which other), independent of the position along the read.

To determine these probabilities, I first identify the neighbors that are outliers in their substitution category; they likely correspond to true biological sequences physically present in the community, rather than sequencing errors. Outlier exclusion is done based on z-scores, i.e. for each sequence we compare its raw cumulative abun-

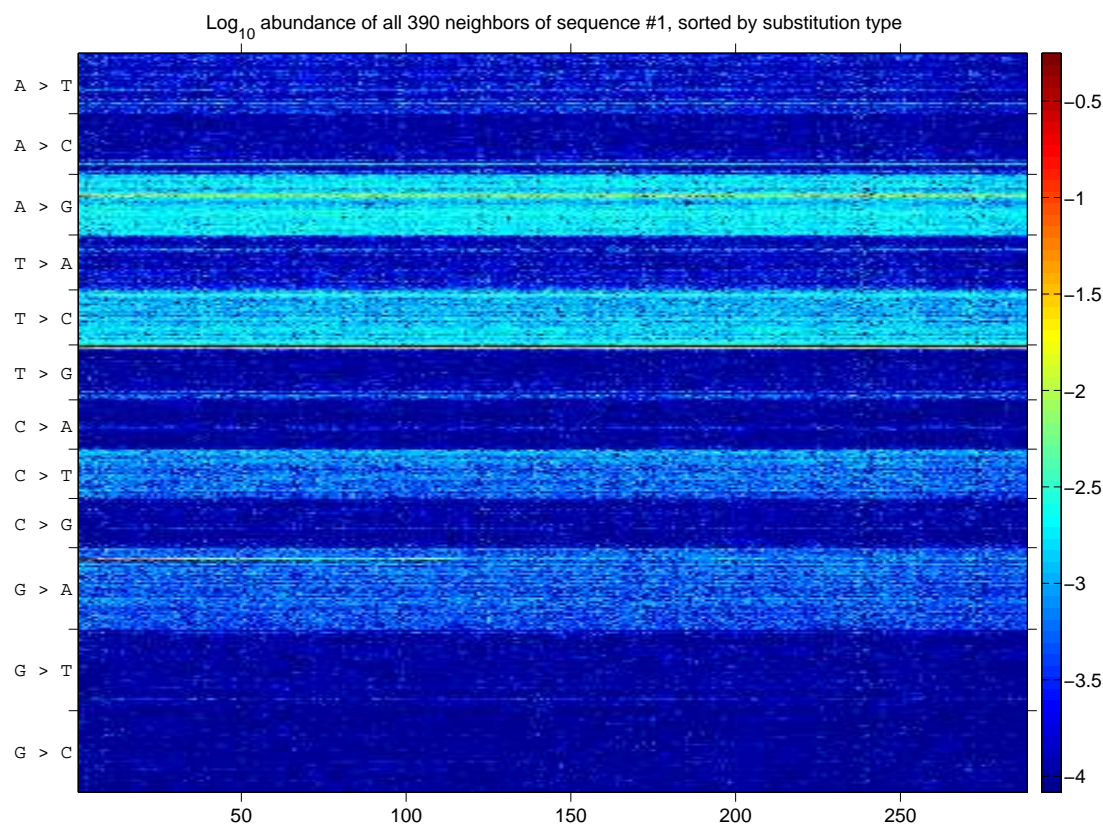


Figure 5.S2: The complete error cloud of Seq. #1. This extended version of Fig. 5.1B shows all 390 first neighbors of Seq. #1, ordered by the type of substitution (and within these classes, by the position of the substitution along the sequence). Color indicates abundance on a log scale, normalized to the abundance of Seq. #1. Except for a few overrepresented neighbors (*cf.* Fig. 5.1B), the substitution type accounts for most of the variance in neighbor abundance.

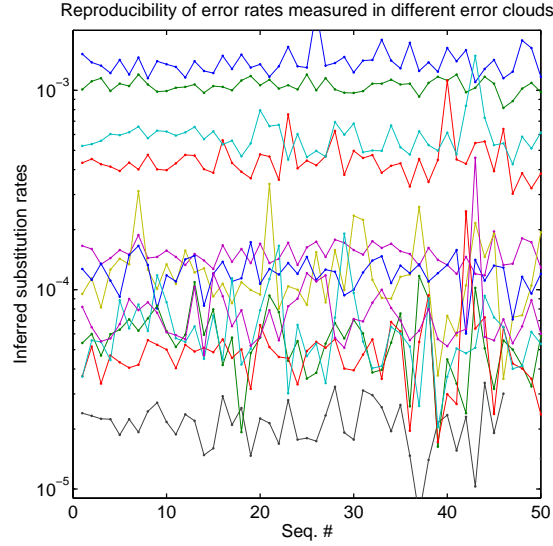


Figure 5.S3: The substitution error rates directly inferred from the error clouds of top 50 sequences by abundance are reproducible across error clouds. Each of 12 separate plots shows the inferred rate of a specific substitution (not labeled to reduce clutter; see also Fig. 5.S4 and Table 5.S1). Predictably, the variability increases when the error clouds of less abundant sequences are used.

dance over all samples to the mean in its substitution category, and normalize by the standard deviation in the category. A strong outlier differing from the mother sequence by a nucleotide substitution at location  $K$  will skew the error rate estimation at that location: some substitution type will appear to be unusually frequent. Therefore, I exclude nucleotide locations that correspond to the strongest outliers, those for which the z-score exceeds some threshold. The remaining locations are then used to estimate the error rates: for each of these locations, I count the number of times a particular substitution occurred, as well as the number of times the nucleotide was recorded correctly. After appropriate normalization, these counts give me the probability of each type of the error.

Fig. 5.S3 shows the inferred substitution rates for error clouds around the top 50 sequences by abundance, using minimally quality-filtered data processed as described in Sec. 5.2 (Phred score cutoff 2, z-score threshold 2). I find that the rates of different substitutions can differ dramatically (up to 50-fold), but the estimates are highly

	$\rightarrow A$	$\rightarrow C$	$\rightarrow T$	$\rightarrow G$	Total
$A \rightarrow$		$0.14 \pm 0.06$	$0.15 \pm 0.02$	$1.34 \pm 0.12$	$1.63 \pm 0.14$
$C \rightarrow$	$0.07 \pm 0.02$		$0.59 \pm 0.04$	$0.06 \pm 0.01$	$0.72 \pm 0.05$
$T \rightarrow$	$0.12 \pm 0.03$	$1.06 \pm 0.07$		$0.07 \pm 0.01$	$1.26 \pm 0.08$
$G \rightarrow$	$0.42 \pm 0.03$	$0.02 \pm 0.01$	$0.05 \pm 0.01$		$0.49 \pm 0.03$

Table 5.S1: Substitution error rates per nucleotide, multiplied by 1000, as measured from the “error clouds” of the top 10 sequences by abundance. Error bars are standard deviations across the 10 estimates.

reproducible (note the log scale on the Y axis), with variability predictably increasing if lower-abundance error clouds are used. Table 5.S1 lists the error rates estimated from the error clouds of the top 10 sequences (mean  $\pm$  standard deviation). Note that, in principle, this effective error probability includes both the base-call errors of the Illumina sequencer and the single-nucleotide substitution errors occurring during PCR. However, the approximate symmetry between rates of a substitution and its reverse-complement partner (e.g.  $p_{T \rightarrow A} \approx p_{C \rightarrow G}$ ), and a clear bias towards transitions as opposed to transversions, suggests that the observed substitutions are dominated by PCR errors (compare with Quince et al., 2011, Table 2).

Inferring error rates directly from the data offers multiple strong advantages. Specifying the error rate as an external parameter (e.g., Morgan et al., 2013) necessarily requires resorting to a conservative global upper bound. Different PCR conditions and different sequencing machines will have different error rates (for example, compare Fig. 5.S4 and Fig. 5.S7A). Further, substitutions vary strongly in probability: in the case at hand, using a single upper bound on error rates would have over-estimated the probability of certain error types by up to 50-fold, reducing my ability to resolve close sequences. In other words, measuring substitution rates directly from the data both reduces the number of algorithm parameters and improves performance.

To investigate how the measured substitution probabilities depend on the quality filtering parameters, I applied the same analysis to data filtered using different Phred quality score thresholds ( $Q_{\min} = 2, 10, 15, 20$ ) as well as different z-score thresholds

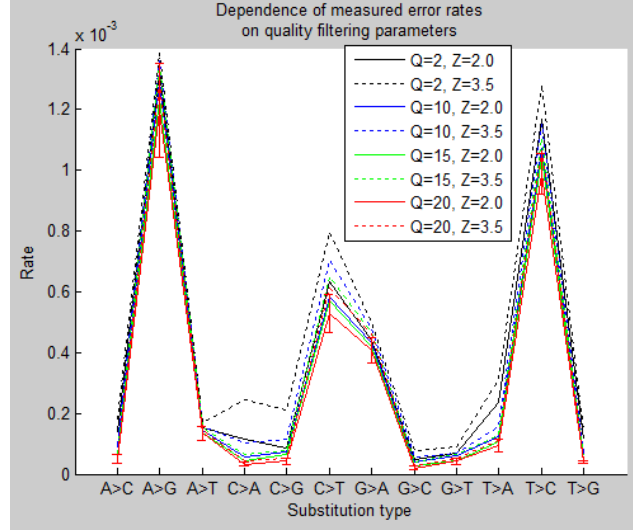


Figure 5.S4: The estimates of substitution error rates exhibit only a weak dependence on Phred quality score filtering parameters, as expected if the dominant source of substitution errors is PCR amplification rather than base call errors. Shown are average error rates as measured for top 10 sequences in the sample. Error bars on the ( $Q=20$ ,  $Z=2$ ) plot indicate standard deviation across the 10 estimates.

(2.0, 3.5). The results are presented in Fig. 5.S4. As expected, the average error rates increase as the Phred score threshold is lowered; however, the magnitude of this change is very small, comparable with the variability of error rate estimates across the top 10 sequences as indicated with the error bars on the plot corresponding to the most stringent filtering,  $Q_{\min} = 20$ ,  $Z = 2$ . This provides further evidence that the majority of substitution errors occur during PCR amplification rather than during sequencing, and thus are not captured by Phred quality scores. I conclude that strict Phred quality filtering unnecessarily reduces data quantity while only marginally improving its quality; for my analysis, I therefore subjected the reads to minimal quality filtering as described in Sec. 5.2.

The dependence on z-score threshold is also consistent with our expectations: a high z-score threshold increases the error rate estimate. Predictably, including stronger outliers ( $z = 3.5$ ) causes the measured error rate to vary significantly across filtering conditions; I used  $z = 2$  which provided excellent reproducibility.



The reproducibility of error rates as observed on Fig. 5.S3 justifies a posteriori the simplifying assumptions such as neglecting the probability of double substitutions in my calculation. Note that, according to the Table 5.S1, the average total error rate per nucleotide is only  $1.0 \cdot 10^{-3}/\text{nt}$ . Therefore, within my error model, assuming that errors occur independently, I estimate that a 130 nt-long sequence has 88% probability of being recorded with no errors. In practice, errors appear to correlate and the true zero-error probability is likely lower. As a different estimate, I calculate the total abundance of all sequences retained by my filtering (7 057 860 reads in 507 samples), and compare to the total number of reads before filtering (8 685 722 reads distributed across 1.4M unique sequences). I find that the filtering algorithm retained 81% of all reads. Since the algorithm intentionally disregards true sequences with low abundance, this estimate is conservative. Further, this estimate is largely insensitive to the error independence assumption: given my stringent filtering criteria, even an unexpectedly frequent double error will be discarded, provided it is less common than a single error. I conclude that  $> 81\%$  of reads in the dataset had no errors, which justifies my decision to discard noisy reads rather than attempting to remap them to their most likely source. For longer reads or noisier data, this approach remains applicable without changes; however, the fraction of error-free reads will be lower. In this case, to avoid significant loss of sequencing depth, I recommend replacing my simple denoiser by an algorithm such as DADA that performs read remapping.

### 5.A.3 The algorithm for filtering substitution errors

For the Illumina sequencing platform, substitution errors account for the bulk of the errors. As described above, these errors have a reproducible structure and their rates can be estimated directly from the data. Using these numbers, for any sequence  $S$  present in a given sample, I can estimate its null model abundance, denoted  $N^0$

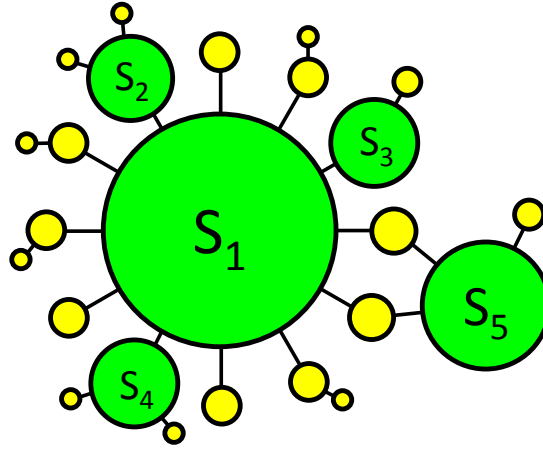


Figure 5.S5: A more detailed version of the error cloud cartoon in Fig. 5.1C. Each circle is a unique sequence, with size representing abundance in a sample. True biological sequences ( $S_1$ - $S_5$ ; green circles) generate “daughter” variants due to substitution errors (yellow circles). Black lines denote Hamming distance = 1 in sequence space. The error rates calculated from the error clouds (see Fig. 5.S3) can be used to calculate, for every sequence, its expected abundance under the assumption that it arose through substitution errors from its more abundant neighbors. Sequences whose abundance is significantly above this expectation are labeled as real (green circles). Note that sequences may arise as substitution errors of multiple “mother” sequences: common neighbors of  $S_1$  and  $S_5$  in this cartoon will have a larger abundance than other substitution errors of either  $S_1$  or  $S_5$ . However, if this increase in abundance is consistent with the null model, they will be correctly recognized as substitution errors.

(abundance derived from sequencing errors of its more abundant neighbors), as follows (Fig. 5.S5):

1. Order sequences by decreasing abundance:  $S_1, S_2$ , etc.
2. Set  $N_i^0 = 0$  for all  $i$
3. For each sequence  $S_i$  with abundance  $N_i$ :
  - (a) Find all  $j$  such that  $S_j$  is a first neighbor of  $S_i$  and  $N_j < N_i$ .
  - (b) For each  $j$ , use the substitution error table to determine the probability  $p_{ij}$  of  $S_i$  to be recorded as  $S_j$
  - (c) Set  $N_j^0 = N_j^0 + p_{ij}N_i$  (“spillover” from  $S_i$  into  $S_j$ )

This zero-parameter algorithm assigns, for each sequence, its null-model abundance expected in that particular sample, using error rates estimated directly from the data. I next use this information to identify “candidate sequences”: those whose presence cannot be explained by a substitution-only error model. Candidate sequences are selected through an abundance criterion, requiring their abundance to exceed the null model prediction ( $N^0$ ) by at least ten-fold, and be no less than 10 counts. I then retain all sequences that independently passed this stringent filtering in at least 2 samples. The reasoning behind this strategy is explained in Sec. 5.2.

#### **5.A.4 Other error types, including chimeras and PCR indels**

With Illumina sequencing, substitution errors account for most of the erroneous sequences in the data, and their occurrence appears to be adequately described by a simple quantitative model. This type of errors is therefore well-suited for error-model based denoising. The list of sequences retained after denoising includes true biological sequences, but also errors not described by the model. The latter category includes chimeras, PCR indels, and possibly other errors such as context-dependent PCR substitutions occurring much more frequently than expected within our model.

I am not aware of any quantitative model for PCR indel errors, which, in my experience, are strongly context-specific. Following Rosen et al., one could make the conservative decision that whenever two candidate sequences differ by pure indels, the lower-abundance should be treated as a possible error. A corresponding script is included in the cluster-free filtering pipeline. However, by definition, this makes it impossible to resolve true biological sequences differing by an indel. Retaining putative indel errors and comparing their abundance distribution across samples with their presumed “mother sequences” would allow identifying such cases. Since PCR indels are comparatively infrequent, for Illumina sequencing I consider indel filtering an optional step of the pipeline. In contrast, the 454 sequencing platform introduces

frequent indel errors at homopolymer regions of the sequence. For 454 data, therefore, proper indel treatment becomes a necessity. The indel-filtering script mentioned above provides one solution; however, since errors we seek to eliminate occur during PCR, while indels occur during 454 sequencing, the best indel treatment strategy for the 454 platform is to merge sequences into “indel families” (Rosen et al., 2012) prior to denoising. Implementing this functionality within our software package will improve its support of 454 data; at the moment, the better approach is to apply our cross-sample analysis to the output of the DADA denoiser (Rosen et al., 2012).

As for chimeras, in our analysis pipeline, we filter chimeric sequences with UCHIME de novo (Edgar, 2011). Following Robert Edgar (UCHIME documentation), I recommend applying chimera filtering to pooled data across samples.

### 5.A.5 Cluster-free filtering software package

The implementation of the denoising algorithm described here is freely available at <http://github.com/hepcat72/cff> as a suite of open-source Perl scripts. Fig. 5.S6 summarizes the workflow of the filtering process.

Applying the denoiser on a per-sample basis is a straightforward four-step process, optionally supplemented by indel filtering (Fig. 5.S6A). However, this denoiser is specifically designed to be run on large multi-sample datasets. The extended workflow appropriate for large datasets (Fig. 5.S6B) has three key differences:

1. The original samples are pooled to construct a library of all unique sequences ever observed, and sequences in the original samples are renamed so that the same sequence has the same identifier in all samples (**mergeSeqs.pl**).
2. The error rates are estimated using the pooled data from all samples (i.e. the library) for better accuracy.

3. The neighbor structure is constructed once, for all sequences in the library (**neighbors.pl**). In the per-sample workflow (Fig. 5.S6A), **neighbors.pl** is automatically invoked for every sample in a manner transparent to the user, which simplifies the workflow; however, many sequences are shared across samples, so for sufficiently large datasets, explicitly calculating the neighbor structure only once results in better performance.

The optional indel-filtering step uses MUSCLE aligner (Edgar, 2004) with modified gap penalty parameters, appropriate for detecting 454 homopolymer indels (`-gapopen -400 -gapextend -399`; see documentation).

The software package is provided with built-in documentation, a test dataset and two shell scripts that allow running the entire workflow presented above with a single command: `run_CFF_on_FastA.tcsh` and `run_CFF_on_FastQ.tcsh` (the latter script uses USEARCH to perform the minimal quality filtering as described in the

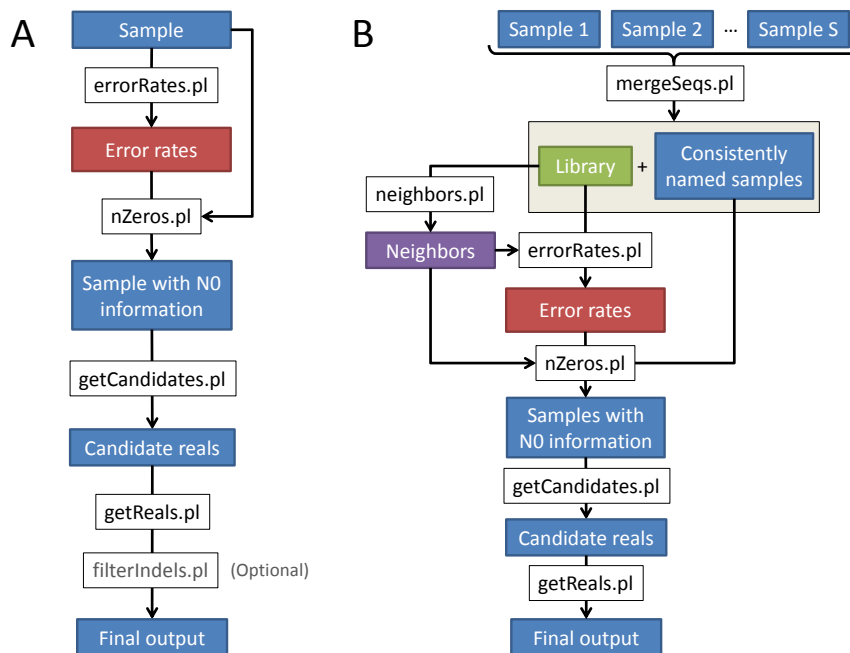


Figure 5.S6: The workflow of cluster-free filtering software package. **A:** The simplest way of running cluster-free filtering denoiser on a single sample. **B:** Extended workflow diagram appropriate for large multi-sample datasets. The optional indel filtering step is omitted for simplicity. In both cases, blue rectangles represent dereplicated FASTA files with sequences of identical length. **getReals.pl** includes chimera filtering (performed with UCHIME).

Methods). The flexible and thoroughly documented command-line interface makes it easy to incorporate cluster-free filtering into any existing pipeline. To reproduce our analysis of the data from Caporaso et al. (2011), download the quality-filtered data published with that study (available at MG-RAST:4457768.3-4459735.3), place it in a folder `CaporasoData` and run:

```
tcsh run_CFF_on_FastA.tcsh 130 analysisResults "CaporasoData/*.fna".
```

### 5.A.6 Mock community validation and comparison with DADA

To validate the performance of my simplified denoiser, I compared it with a state-of-the-art denoiser DADA (Rosen et al., 2012) using two mock community datasets (*Divergent* and *Artificial*; Quince et al., 2011) that Rosen et al. used to demonstrate DADA’s superior accuracy to AmpliconNoise. Quoting from the original publication, these datasets were constructed by amplifying the V5 region of the 16S rRNA gene from 23 and 90 clones, respectively, isolated from lake water. The *Divergent* clones were mixed in equal proportions and are separated from each other by a minimum nucleotide divergence of 7%, while the *Artificial* clones were mixed in abundances that span several orders of magnitude, with some of the clones differing by a single-nucleotide substitution. For purposes of comparison, I used the exact same sets of filtered reads (35 190 reads in *Divergent* set; 31 867 in *Artificial*), kindly provided by Michael Rosen.

The comparison of denoiser output and the reference set of Sanger clones was complicated by the imperfections of the reference set. A number of “reference” Sanger clones differed from their closest high-abundant matches in the 454 data at the same locations towards the beginning of the read, which is suggestive of errors in the reference sequences. Further, some reference sequences of the *Artificial* set had no

Category	DADA	CFF	Abundance
<i>Divergent</i> : 23 ref. seq.	23 true positives 0 false negatives	23 true positives 0 false negatives	231-1426 counts
Other detections	0	0	
<i>Artificial</i> : 49 ref. seq.	48 true positives 1 false negative	49 true positives 0 false negatives	18-3587 counts
Other detections	Seq. #35	Seq. #35 Seq. #95 Seq. #103 Seq. #119	163 counts 13 counts 12 counts 12 counts

Table 5.S2: Comparison of DADA and cluster-free filtering (CFF) denoiser on mock community data. Sequences numbered by decreasing abundance in the dataset.

close matches in the data; some Sanger clones differed at locations that were not part of the 454 sequenced fragments; and 454 sequences included 6 extra bases at the beginning of the sequence that were absent from the Sanger clones.

I therefore began by constructing “cleaned” reference sets as follows: for each reference Sanger clone, I found its closest match in the dataset that had at least 98% similarity and an abundance of at least 10 counts. This matching 454 read was used as the new reference sequence, and the differences, if any, were ascribed to Sanger clone errors. For the *Divergent* dataset, each reference sequence had exactly one clear match in the 454 data. For the *Artificial* set, of the 90 reference Sanger clones, we found that one was 29 nts away from the closest 454 read; for 3 other clones, no 454 read within  $\geq 98\%$  sequence similarity radius reached an abundance of 10 counts. Our algorithm intentionally disregards any sequences below this abundance threshold; therefore, for the purposes of this comparison these reference sequences were considered absent and we did not count them as false negative for any of the algorithms. Several groups of clones were not distinguishable by the 454 sequenced fragment. Altogether, the new reference set of sequences that were both present and distinct contained 49 reference sequences.

I then ran DADA and cluster-free filtering on both datasets. DADA was run with the same parameters as used for this data in the original publication, namely  $\Omega_a = 10^{-40}$  and  $\Omega_r = 10^{-3}$ . Cluster-free-filtering included indel filtering step, since this data was obtained using the 454 platform and indels appear frequently.

The results are presented in Table 5.S2. Both algorithms identified correctly all 23 reference sequences of the *Divergent* dataset. For the *Artificial* set, and due to the conservative parameters recommended by Rosen et al., one of the reference sequences was missed by DADA but was correctly identified by my algorithm. Sequence #35 (in order of decreasing abundance), absent from the reference set, was retained by both algorithms and is likely a true biological sequence. Cluster-free filtering generated 3 additional detections just above its threshold of 10 counts. It is instructive to trace the origin of these calls. For example, Seq. #95 was discarded by DADA as possibly an erroneous read generated by its closest reference sequence (Seq. #1) two substitutions away. Specifically, Seq. #95 differs from Seq. #1 by a T at location 23 and a G at location 118, a relation that we denote “Seq.#95 = Seq.#1 23T 118G”. If it were true that Seq. #95 is a substitution error of Seq. #1, we would generally expect single-error variants to be more abundant than double errors. In reality, Seq.#1 23T (=Seq. #587) and Seq.#1 118G (=Seq. #121) have abundances of just 4 and 12 counts, respectively, which is why my algorithm identified Seq. #95 as likely real. However, its unexplainably high abundance could also have arisen through amplification of a double substitution that occurred early in the PCR cycle, and the default parameters of DADA were chosen conservatively so as to eliminate such cases (Rosen et al., 2012). Whether or not these detections are false positives or true biological contaminants can be determined only by a cross-sample analysis as presented in the main text.



### 5.A.7 Runtime comparison with DADA

The methodology presented in this work was designed to perform cross-sample comparisons of sequence abundance in individually denoised samples. As explained in Sec. 5.2, the simplified denoiser I developed is meant to maximize performance on large datasets specifically for this application, taking advantage of my focus on moderate-to-high abundance sequences. Other denoisers can be used. To estimate the runtime of DADA on the tongue dataset considered here, I used a representative subset of 20 samples, 10 from lane 5 and 10 from lane 6 of Caporaso et al. Following the instructions in Rosen et al. (2012), I used ESPRIT to precluster sequences in each sample prior to processing them with DADA. The measured runtime is presented in Table 5.S3. Extrapolation to the full set of 507 samples yields the estimate of  $2.3 \cdot 10^5$  sec total runtime quoted in the text, compared to 626 sec actual runtime for cluster-free-filtering. As explained in the main text, one of the reasons for this speedup is that my multi-sample detection strategy allows me, in any given sample, to look for candidate sequences only among those with abundance  $\geq 10$  counts. This speedup can be applied to DADA as well; to this end, I removed all clusters that contained no sequences with abundance  $\geq 10$  counts, and measured DADA runtime after this filtering; this decreased the estimated runtime on the full dataset to  $5.5 \cdot 10^4$  sec. Using DADA in this way is the strategy I recommend for applying this cross-sample comparison methodology to 454 data with long reads where erroneous read remapping and indel family merging become advisable. Eliminating low-abundance sequences leads to a considerable improvement of DADA runtime; nevertheless, the total runtime remained two orders of magnitude slower than the cluster-free filtering approach, due primarily to the computational cost of preclustering.

	ESPRIT+DADA	same, $\geq 10$ counts only	CFF denoiser
Lane 5, 10 samples	969 + 1297 sec	969 + 49 sec	12 sec
Lane 6, 10 samples	924 + 5133 sec	924 + 186 sec	16 sec
Whole dataset	$2.3 \cdot 10^5$ sec (est.)	$5.5 \cdot 10^4$ sec (est.)	626 sec (actual)

Table 5.S3: Runtime comparison of DADA and the simplified cluster-free filtering (CFF) denoiser on two representative sets of 10 tongue samples (lane 5 and lane 6). Lanes were considered separately since samples in the two groups tended to differ significantly in the number of reads retained by quality filtering. The whole dataset consisted of 189 samples on lane 5 and 320 samples on lane 6. Comparisons were performed on an Intel Xeon CPU 2.83GHz.

### 5.A.8 Other applications: environmental cross-sectional 454 data

The approach described in this work does not explicitly rely on the longitudinal nature of the sampling. Most of this analysis can be readily applied to any multi-sample datasets, e.g. a cross-sectional sampling or a location series, provided samples were collected and processed in a similar way so that the error structure can be assumed to be similar. Further, and despite the caveats I described, the same method can be applied even to data collected using the 454 sequencing platform. To illustrate the broad applicability of my approach, I used data from a cross-sectional environmental sampling conducted by Preheim et al. (SRA accession number from SRP029470). Lake water microbiota were sampled at depths ranging from 0 m (surface) to 22 m with 1-meter depth intervals. The authors used this data to illustrate their sequence clustering algorithm (DBC) that also relies on cross-sample comparisons to distinguish between closely related OTUs; for details, see the original reference (Preheim et al., 2013). They report their algorithm worked best with stringent quality filtering whereupon sequences were trimmed to just 76 nt, and any reads containing bases with Phred quality scores at or below 16 were discarded. This filtering retained 7.78M total sequences (120K unique). Since my approach includes data denoising, I could use much more liberal quality score filtering and retain more data

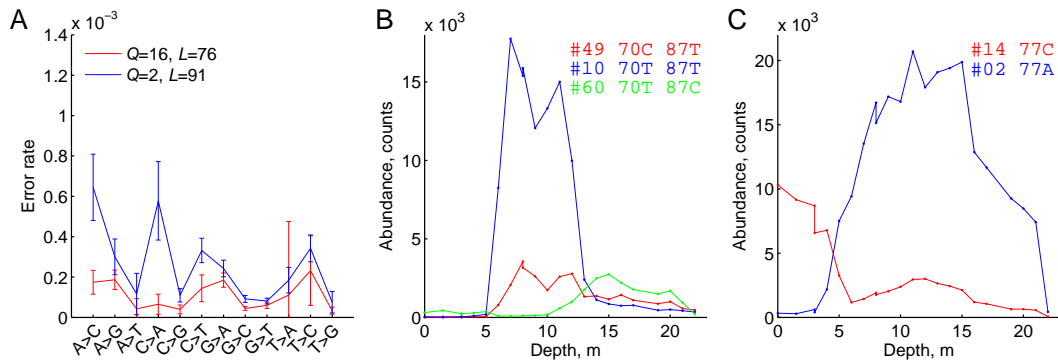


Figure 5.S7: Cross-sectional environmental 454 data: lake water microbiota (Lake Mystic) sampled at depths 0–22m. **A:** Substitution error rates inferred from the data for two sets of quality filtering parameters indicated in the legend;  $Q$  is the Phred quality score truncation threshold;  $L$  is read truncation length. **B:** Three sequences resolved by our analysis; Seq. #49 and Seq. #60 both differ from Seq. #10 by a single nucleotide at locations 70 and 87, respectively. Sequence abundance is shown as a function of depth; sequences are labeled by cumulative abundance rank. **C:** Same, for sequences Seq. #2 and Seq. #14 differing at nucleotide 77.

(USEARCH maxEE of 1 and truncating at Phred quality score 2). To compare runtime of our algorithm and DBC, I increased the read truncation length so as to keep the same total number of sequences. This set the quality-filtered sequence length to  $L = 91$  nt, 20% longer than used by the authors (7.98M sequences, 300K unique; `tcsh run_CFF_on_FastQ.tcsh 91 analysisResults "PreheimData/*.fastq"`).

Fig. 5.S7A shows the substitution error rates inferred from the data at both sets of quality filtering parameters. Note that these rates are significantly lower than those of Fig. 5.S4 (the scales of the two plots are identical), exhibit a very weak transition/transversion bias, and are more sensitive to Phred score quality filtering than what we have seen with Caporaso et al. data (Fig. 5.S4). This seems to indicate that the protocol used by Preheim et al. generates significantly fewer PCR substitution errors. This dependence of error rates on the experimental protocol highlights the advantage of being able to estimate error rates for a given dataset directly from the data, without the need for a separate calibration.

Figs. 5.S7BC provide examples of sequences differing by a single nucleotide exhibiting ecologically significant distinctions, as identified by my method in this envi-

ronmental dataset; compare with Fig. 5.2A, Fig. 5.S10 and Fig. 5.S11AB. Sequence abundance is shown as a function of depth (each sample was independently normalized to  $3.2 \cdot 10^5$  total quality-filtered reads per sample, to correct for varying sample size). The DBC method of Preheim et al. is also capable of identifying OTUs differing by a single nucleotide (compare Fig. 5.S7BC to Fig. 5b in the original reference); however, cluster-free filtering achieve higher resolution by retaining longer reads and took only 13 min single-core processor time (Intel Xeon CPU 2.83GHz), compared to 8 hours analysis time the authors report for parallelized DBC running on a cluster with 60-100 processes executing simultaneously. In fact, the true runtime difference is even greater, since the complexity of both algorithms scales with the number of *unique* sequences rather than total reads. For  $Q_{\min} = 17, L = 76$  as used by the authors of DBC, cluster-free filtering algorithm completes in only 210 seconds.

I stress, however, that DBC and cluster-free filtering seek to achieve different goals and are not directly comparable. DBC is an OTU clustering algorithm, whereas the goal of cluster-free filtering is to identify sub-OTU structure of moderate-to-high-abundance community members. However, to my knowledge DBC is the only existing tool that exploits cross-sample comparisons to inform the interpretation of sequencing data, and the performance comparison above serves to illustrate the drastically different computational cost of the two approaches.

### 5.A.9 How many samples is enough?

I have described a method that employs cross-sample comparisons to achieve sub-OTU resolution. The analysis presented in the main text uses data from a study with an uncommonly large number of samples; in contrast, the previous section demonstrates that the same method can be usefully applied to a dataset with only 22 datapoints. What is the minimum number of samples required by our method?

The answer is that the number of samples determines the resolution that can be achieved; more samples will always allow higher resolution, but coarser differences can be resolved with just a few. For example, just 2 samples (say, 0m and 10m) would have been enough to resolve the two subpopulations presented on Fig. 5.S7C. By contrast, the difference between depth traces of Seq. #10 and Seq. #49 (Fig. 5.S7B) is less pronounced and more samples are required. Finally, resolving the sequences in Fig. 5.2B would not have been possible with fewer than  $\sim 100$  samples. For high-abundance sequences where the complex structure of noise in the counts can be neglected, this tradeoff can be formally quantified using the Jensen-Shannon divergence as a measure of distance between abundance distributions of two sequences across samples; for details, see Preheim et al. (2013).

## 5.B Supplementary information for Figure 5.2

### 5.B.1 A pair of sequences representing strongly anticorrelated subpopulations

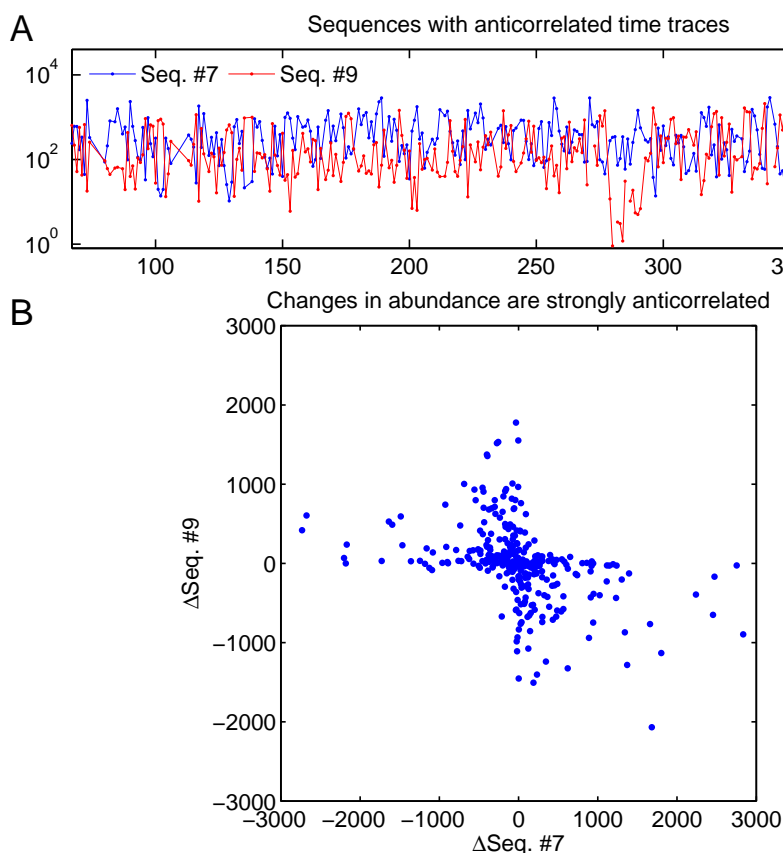


Figure 5.S8: **A.** Abundance time traces of Seq. #7 and Seq. #9. **B.** Scatter plot for the same sequences of their discrete derivatives of abundance (i.e. abundance changes from each day to the next). A BLAST search against the GreenGenes database identifies the likely taxonomy of Seq. #7 as *Streptococcus thermophilus*. Seq. #9 does not have a good match; the closest hit is an unclassified *Prevotella* sp. at only 88% identity.

### 5.B.2 Best expected correlation of two time traces

The maximum degree to which time traces of two sequences can be correlated is a function of their abundance: for low-abundance sequences the Poisson sampling noise

becomes non-negligible and sets an upper bound for the best achievable correlation coefficient. Consequently, to define a correlation as strong or weak, any measured correlation coefficient should be compared to this abundance-dependent quantity rather than to 1.

Let  $N(t)$  be the true abundance time trace of some bacterial strain (in units of cells, rather than sequence counts). Imagine that two sequences in the dataset were measuring the abundance of this exact same strain, but with different amplification efficiencies  $\lambda_1$  and  $\lambda_2$  (let  $\lambda_1 > \lambda_2$ ). Neglecting all sources of noise other than the Poisson counting noise, the abundance traces of these two sequences can be modeled by

$$n_{1,2}(t) = \text{Pois}[\lambda_{1,2}N(t)],$$

where  $\text{Pois}[\cdot]$  denotes adding Poisson noise. Since Poisson noise is unavoidable, the correlation coefficient between these two traces sets an upper bound for the correlation between  $n_1(t)$  and any other trace  $n^*(t)$  with the same mean abundance as  $n_2(t)$ . This maximum correlation depends on the shape of the trace  $N(t)$  and amplification efficiencies  $\lambda_1, \lambda_2$ , and can be expressed as follows:

$$c_{\max}[N(t), \lambda_1, \lambda_2] = \text{corr}(\text{Pois}[\lambda_1 N(t)], \text{Pois}[\lambda_2/\lambda_1 * \lambda_1 N(t)]).$$

And therefore, in terms of measurable quantities only:

$$c_{\max}[n_1(t), \langle n_2 \rangle] \approx \text{corr}(\text{Pois}[n_1(t)], \text{Pois}[\langle n_2 \rangle / \langle n_1 \rangle * n_1(t)]).$$

Here  $\langle \cdot \rangle$  denotes the average abundance, and I use the higher-abundance trace of the pair as the best estimate of the shape of the true abundance  $N(t)$ . The maximum correlation coefficient depends on the shape of the trace  $n_1(t)$  and on the

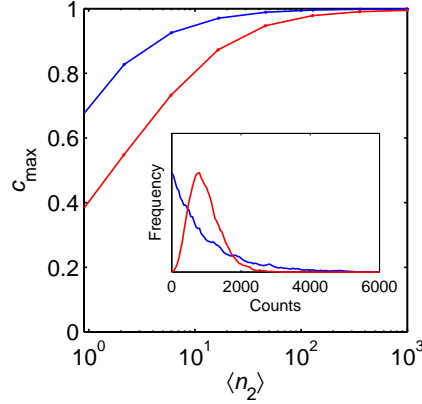


Figure 5.S9: Best expected correlation  $c_{\max}$  for a pair of abundance time traces  $n_{1,2}(t)$  is sensitive to the shape of the distribution of the daily counts, not just the average abundance  $\langle n_{1,2}(t) \rangle$ . The figure shows the best expected correlation  $c_{\max}[n_1(t), \langle n_2 \rangle]$  as defined in the text, for two different mock traces  $n_1(t)$  with the same mean ( $\langle n_1 \rangle = 1000$  counts/day) but with different distributions, modeled here by Gamma distributions with shape parameter 1 (blue) and 5 (red). The best expected correlation increases with the mean abundance  $\langle n_2(t) \rangle$ , but for the same mean it is higher for the blue trace whose distribution covers a wider dynamic range. Because of this nontrivial dependence on the distribution shape, in my definition of dynamical similarity I compute the best expected correlation individually for every pair of sequences.

mean abundance of the trace we compare it to; the lower the mean abundance is, the stronger the effect of Poisson noise and the lower the  $c_{\max}$ .

In practice, for a pair of traces  $n_1(t)$ ,  $n_2(t)$ , I compute their best expected correlation as follows:

1. Take the more abundant trace  $n_1(t)$
2. Construct a renormalized trace  $n_2^{\text{mock}}(t) = \frac{\langle n_2 \rangle}{\langle n_1 \rangle} n_1(t)$
3. Poisson-resample both of these 10 times: denote these  $n_1^{(i)}$ ,  $n_2^{\text{mock}(i)}$ ,  $i = 1 \dots 10$ .
4. Compute the correlation coefficients between all pairs  $c_{ij} = \text{corr} \left( n_1^{(i)}, n_2^{\text{mock}(j)} \right)$ .
5. Set  $c_{\max}[n_1(t), \langle n_2 \rangle] = \langle c_{ij} \rangle$ .

The shape of the function  $c_{\max}[n_1(t), \langle n_2 \rangle]$  is illustrated on Fig. 5.S9.



### 5.B.3 Distance metric for sequence pairs

I use the BLAST definition, i.e. the ratio of the number of mismatches to the total number of columns after pairwise realignment, and multiply this ratio by the length of the sequence (130 nt). For close sequences that differ by a few substitution errors the alignment is trivial, and this normalization corresponds to the Hamming distance between sequences, in nt.

## 5.C Supplementary information for Figure 5.3

### 5.C.1 Estimating correlation time from autocorrelation function

I define the autocorrelation time  $\tau$  of a sequence as the time shift  $\Delta t$  at which the autocorrelation function  $c_{\Delta t}$  falls below the threshold of statistical significance. For reasons discussed above, the notions of strong (significant) or weak (insignificant) correlation of sequence time traces are abundance-dependent. Therefore, instead of using a fixed threshold value for all sequences, I proceed as follows. For a given sequence, I first compute its root-mean-square autocorrelation coefficient for time shifts between 70 and 100 samples:

$$c_{\text{null}} = \sqrt{\langle (c_{\Delta t})^2 \rangle_{\Delta t=70\dots 100}}.$$

If we assume that all autocorrelation observed at such large time shifts is entirely due to noise, then  $c_{\text{null}}$  provides a natural scale for statistical significance. I conservatively define the significance threshold at twice the magnitude of  $c_{\text{null}}$ .

Note that  $c_{\text{null}}$  provides an upper bound on a statistically significant correlation value. If some dynamical processes in the population are slow enough that they contribute to the autocorrelation function even at such large time shifts (*cf.* Fig. 5.S10),

this will increase  $c_{\text{null}}$  and cause me to underestimate the true autocorrelation time. This means that assuming  $c_{\text{null}}$  was entirely due to noise is a safe approximation to make: if it does not hold, it can only strengthen my conclusion that the sequence abundance time traces exhibit multi-day autocorrelations.

## 5.C.2 Examples of sequences exhibiting consistent dynamics on very long time scales

Fig. 5.S10AB shows examples of sequences exhibiting steady change in abundance for more than a month. In both cases, the slow-changing sequence is 99.2% similar to a very high-abundance community member and could not have been resolved by traditional OTU-based methods. Note the sharp jump in panel B at day 182 of the sequence representing the invading subpopulation (red) to an abundance value close to the equilibrium established after day 210. It is intriguing to speculate that this trace may document spatial invasion of a subpopulation already established elsewhere on the tongue, a region accidentally sampled on day 182.

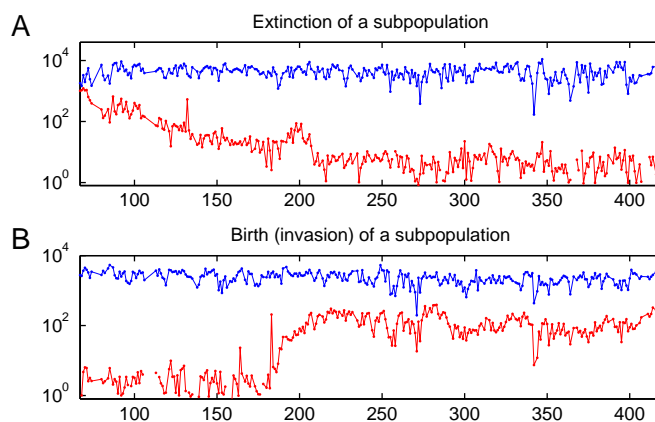


Figure 5.S10: **A.** Slow extinction of a subpopulation (red; cf. Fig. 5.1B, neighbor 3). From day 210 onwards the abundance of the red sequence is consistent with it being a substitution error of Seq. #1 (blue), which is a direct neighbor in sequence space. **B.** Slow birth/invasion of a subpopulation (red). The new sequence differs by 1nt from well-established Seq. #2 (blue), and prior to day 160 its abundance is consistent with being its substitution error. Note the high similarity of fluctuations from day 210 onwards.

### 5.C.3 Persistence of difference: the null model

To distinguish between 16S tags coming from distinct subpopulations or from physically the same bacterial cells, I introduced a quantity I called the *persistence of difference*  $P_D$ . For this, I first defined the fractional difference  $\Delta(t)$  between two time traces renormalized to the same mean  $n_{A,B}(t)$ :

$$\Delta(t) = \frac{n_A - n_B}{(n_A + n_B)/2}.$$

I then defined the persistence of difference  $P_D$  as the 1-day autocorrelation coefficient of  $\Delta(t)$ . If  $A$  and  $B$  are two genomic variants contained within the same bacterium, then any difference between  $n_A(t)$  and  $n_B(t)$  must be due to measurement noise, and  $P_D$  must vanish. If, however,  $n_{A,B}(t)$  reflect abundances of two distinct subpopulations, then  $\Delta(t)$  can be expected to exhibit some degree of autocorrelation due to the slow dynamics observed for most individual sequences. An intuitive argument for this was given in section 5.4. Here, to gain some extra intuition about the null model for  $P_D$ , I calculate it explicitly in the simplest case when the two traces  $n_{A,B}(t)$  are independent and can be approximated by a stationary, weakly fluctuating process:

$$n_A(t) = \mu(1 + \sigma_A \xi_A(t)) \tag{5.1}$$

$$n_B(t) = \mu(1 + \sigma_B \xi_B(t)) \tag{5.2}$$

Here  $\xi_{A,B}$  have zero mean, unit variance and are uncorrelated. Assuming  $\sigma_{A,B} \ll 1$ , we can write:

$$\Delta(t) \approx \sigma_A \xi_A(t) - \sigma_B \xi_B(t)$$

And therefore, making use of the independence assumption,

$$P_D = \frac{\langle \Delta(t)\Delta(t+1) \rangle}{\langle \Delta(t)^2 \rangle} \approx \frac{\langle \sigma_A^2 \xi_A(t)\xi_A(t+1) + \sigma_B^2 \xi_B(t)\xi_B(t+1) \rangle}{\sigma_A^2 + \sigma_B^2} = \frac{\sigma_A^2 c_{1A} + \sigma_B^2 c_{1B}}{\sigma_A^2 + \sigma_B^2}$$

Here  $c_{1A,B}$  are the one-day autocorrelation coefficients of the fluctuations of the two individual sequences.

The independence approximation made above is clearly not valid for the dynamics of most community members. For this reason, for the purposes of Fig. 5.3C, the null-model prediction was constructed directly from the data, by reversing in all pairs the time order for one of the sequences prior to the calculation of  $P_D$ . This removes any real correlations of the traces while preserving autocorrelation and other properties of the traces such as their fluctuation spectrum. Nevertheless, the calculation above is useful as it explains why the null-model expectation for  $P_D$  is non-zero when both sequences have slow internal dynamics.

Note that a sequence with an exceptionally long intrinsic time scale (as shown in Fig. 5.S10AB) will have a large  $P_D$  score when paired with any other sequence. These two sequences were therefore excluded from Fig. 5.3C.

#### 5.C.4 Persistence of difference for non-longitudinal data

None of the cross-sample comparison methodology described in this work is limited to time series data. The “persistence of difference” argument accompanying Fig. 5.3 is no exception; however, it does rely on two additional assumptions, namely that the composition of samples varies smoothly with some parameter labeling the samples, and that the sampling frequency is sufficiently high to allow correlations of fluctuations to be observed between consecutive samples. For the longitudinal data series of Caporaso et al. this parameter was time; for a location series one can expect commu-

nity composition to vary smoothly in space, and the same argument can be applied. In other words, the use of “persistence of difference”  $P_D$  need not be limited strictly to longitudinal datasets. However, autocorrelation-based analysis is particularly sensitive to the number of samples (see section 5.A.9). Determining whether  $P_D$  can be a useful concept for studying the spatial heterogeneity of populations requires further investigation.

## 5.D Supplementary information for Figure 5.4

### 5.D.1 Over-estimation of OTU quality scores

As described in the main text, for the purposes of Fig. 5.4, when calculating OTU quality scores, I used only sequences from the top 200 by overall abundance. Since most of the diversity is contributed by low-abundance species (Huttenhower et al., 2012), Fig. 5.4 underestimates the true diversity of an OTU. Including lower-abundance OTU members makes OTU quality scores drop continuously as new OTU members are added; however, it also becomes increasingly hard to separate dynamical diversity from the effects of noise. Consequently, in Fig. 5.4 I report my most conservative estimate of within-OTU diversity, where I use only the highest-abundance members out of all those resolved by cluster-free filtering (there was an average of  $18 \pm 4$  resolved sequences within a 97% OTU, and only  $9 \pm 2$  per OTU were used for Fig. 5.4).

In addition, OTU quality scores were calculated under the assumption that each sequence represents a separate subpopulation. Sequences that in fact derive from the same bacteria (16S paralogs or errors not in my model) appear in the defining equation as independent, dynamically identical subpopulations, increasing the apparent OTU quality score. This is another reason why the true quality scores of OTUs are likely even lower than reported in Fig. 5.4.

## 5.E Supplementary information for Figure 5.5

### 5.E.1 Cross-individual analysis of fecal samples

To confirm our conclusions from the analysis of tongue microbiome data presented in the main text, I repeated the same analysis using fecal samples of the two individuals, collected in the same study (Caporaso et al., 2011). There were 374 samples, 243 from the male subject and 131 from the female, with  $2.5 \pm 0.5 \cdot 10^4$  reads per sample. I normalized the observed abundances to  $2.5 \cdot 10^4$  total reads in each sample to correct for varying sample size. As before, sequences were labeled in order of decreasing overall abundance (pooling samples from both individuals): Seq #1F, Seq #2F, etc., where “F” reflects that we are now dealing with fecal samples rather than the tongue.

Again, I found that sequences differing by as little as a single nucleotide can exhibit ecologically significant differences in their dynamics. The most striking example is that the dominating sequence in individual “Male 3” differs from the dominating sequence in “Female 4” by a single nucleotide, and virtually no cross-contamination is observed (Fig 5.S12A). Both these sequences map to *Bacteroides* sp. in GreenGenes (DeSantis et al., 2006). Another example is presented in panel B. Finally, I repeated the cross-individual analysis presented, for tongue samples, in Fig. 5.5AB. We see that the two gut communities as probed by the fecal samples also share a large fraction of sequences at 100% identity. This once again supports the scenario whereby the communities exchange members with non-negligible frequency, although less so than the tongue samples. The observation of panel A is therefore unlikely to represent the effect of dispersal limitation, suggesting instead a functional difference between the representatives of *Bacteroides* sp. established in the two individuals or a resistance to invasion. Finally, I find that the dynamical similarity of shared sequences, when measured independently in the two individuals, is clearly correlated, just as it was for the tongue communities (Fig. 5.5B). With the number of shared sequences being lower

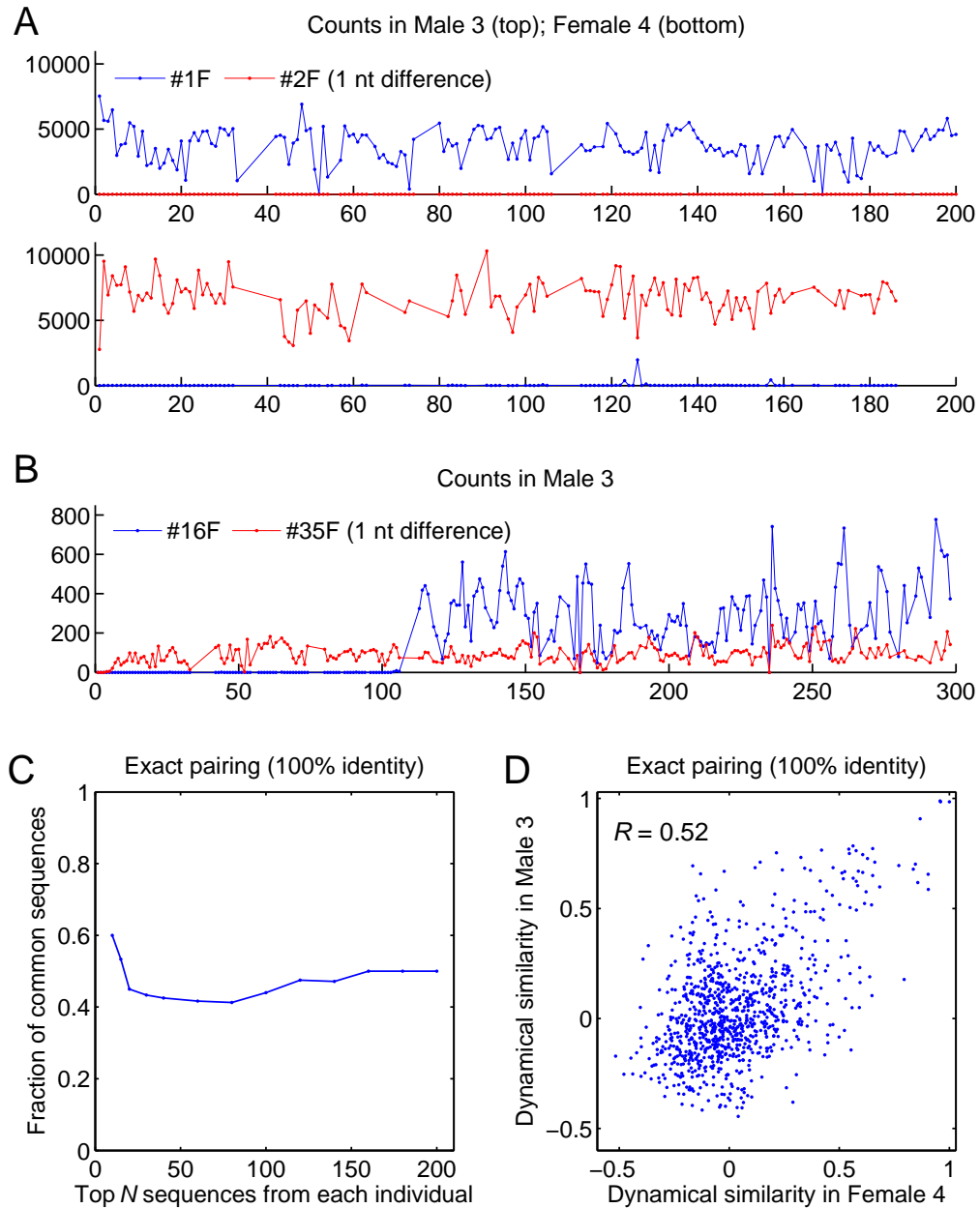


Figure 5.S11: **A.** Abundance time traces (sequence counts *vs.* observation day) for Seq. #1F and Seq. #2F, which differ by a single nucleotide and dominate in individuals Male 3 and Female 4, respectively. **B.** Another example of abundance time traces of two sequences that differ by a single nucleotide (99.2% similarity), yet exhibit strongly distinct dynamics and so derive from distinct bacteria. **C.** Fraction of shared 16S sequences, defined as the fraction of common tags (at 100% sequence identity) among the most abundant  $N$  sequences in the fecal samples of each of the two individuals, plotted as a function of  $N$  (compare with Fig. 5.5A.) **D.** Scatter plot of the dynamical similarity of pairs of common fecal sequences, as measured independently in the two individuals, for all possible pairs among the 44 common sequences shared within the top  $N = 100$  (compare with Fig. 5.5B).

for fecal samples than for tongue samples, the statistics were insufficient to compare dynamical similarity of “intentionally mismatched” sequences as in Fig. 5.5C.

## 5.E.2 Cross-individual analysis at 97% OTU level

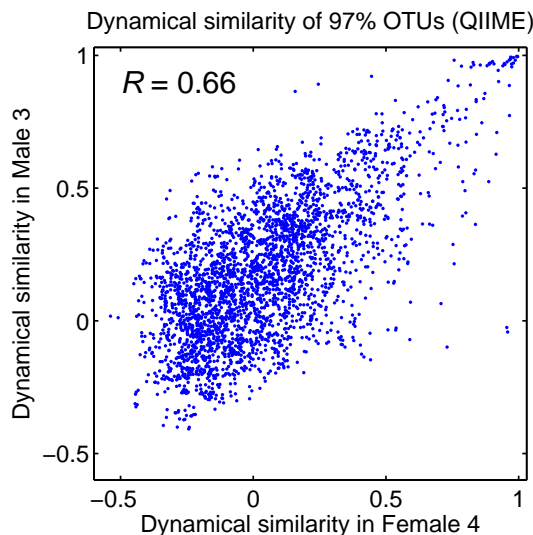


Figure 5.S12: Dynamical similarity between pairs of common 97% OTUs, as measured independently in the two individuals, for 78 common OTUs within the top 100, constructed using closed-reference OTU picking as implemented in QIIME.

The same analysis as in Fig. 5.5B can be performed for shared 97% OTUs rather than shared sequences (at 100% identity). I constructed OTUs using closed-reference OTU picking as implemented in QIIME (Caporaso et al., 2010), matching sequences at 97% sequence similarity against the GreenGenes database. Fig. 5.S12 shows the scatter plot of the dynamical similarity between pairs of common OTUs, as measured independently in the two individuals, for 78 common OTUs (those shared within the top 100). Note, however, that most OTUs are dominated by a single high-abundance sequence (as evidenced by the high weighted quality score on Fig. 5.4), and most of these dominating sequences are shared across the two communities (Fig. 5.5A). For these reasons, the plot shown here is very similar to Fig. 5.5B, but only because the within-OTU diversity is masked by dominating subpopulations.



# Chapter 6

## Conclusion

The questions discussed in the previous chapters are highly diverse. Their choice was shaped by my environment at Princeton and allowed me to gain exposure to a wide variety of both subjects and tools, from “wet lab” molecular biology techniques to sequencing data analysis and from confocal microscopy to chalk-and-board theory. Ultimately, the only connecting thread is that in all instances, I was a physicist studying a biological system, hence the title of this dissertation: in its broadest sense, the physicist’s approach is to look for details that matter, and be quantitative about it. So, can we identify any common features in the examples I have discussed, or draw comparisons between them?

Unsurprisingly, which details matter depends on the question being asked. To take the example of the fruit fly, details of promoter are of crucial importance for determining which gene will be expressed when and where during development – but not for setting the magnitude of noise, which appears to be universal, nor for the mechanism by which this noise is filtered (Chapter 2). This mechanism, specifically for the fly embryo, is diffusion in the syncytium. The syncytial embryonic stage, among model organisms, is unique to *Drosophila* and is, without doubt, a crucial element shaping the specifics of its developmental process. But from the point of

view of morphogenesis more generally, this fly-specific detail can be seen as a form of (diffusion-mediated) cell-to-cell communication that allows reducing noise through non-locality of gradient readout (Chapter 3). As such, it can shed light on common architectural features of gradient response pathways in general.

It is often asked if the structure of biological networks could be understood from the angle of robustness to noise, or error-correcting capability as a network-level property, something that individual nodes do not possess. In the case of the fruit fly, precision and reproducibility appear to be, at least in part, network-level phenomena in this sense. However, the network here is not just the regulatory architecture we could have naively read from the genome: we must also take into account that nuclei are spatially arranged in a tissue and can exchange protein products, so that local regulatory networks become spatially coupled into a “network of networks”. With sufficient understanding of promoters and enhancers (one we do not currently possess), one might be able to deduce the entire quantitative structure of the regulatory network from the genome sequence. Achieving such understanding is the Holy Grail of genomics, but here it would not have been sufficient, since we would not have seen the spatial coupling. What appears to be an idiosyncratic detail may in fact play a very important role in the phenomenon of interest, and in biology, I think, much more so than in physics, because idiosyncratic occurrences that can confer a fitness advantage will be amplified and exploited by evolution. In this manner, biological systems could be driven into highly non-generic regions of parameter space, much like in the abstract model considered in Chapter 4.

Of the examples considered here, the one I am most hopeful about pursuing in the future is the ecology of microbial communities (Chapter 5). For a physicist, the microbiota is an ideal system to study. Ecological interactions between members are far too numerous and intricate for any understanding of community dynamics to be achieved by mapping out all the microscopic details. At the same time, these highly

complex systems clearly exhibit universal features, such as a very large number of low-abundance species or strong functional stability of the community metagenome that contrasts with the high compositional diversity. This makes it possible to hope that these properties could be understood within some general theoretical framework. Of all examples of biological networks, microbial communities should be particularly amenable to an efficient statistical description. Both genetic and neural networks evolve as a whole; their structure is dictated by a genetically encoded “master plan” and is constrained by the evolutionary history of the organism. In contrast, each microbe in the gut is an independently acting agent. The community assembles *de novo* and differently in each newborn baby. In addition, horizontal gene transfer allows bacteria to swap pieces of their DNA and blurs the definition of a “species”. This is commonly seen as a very serious barrier for application of species-based models of theoretical ecology, but this limitation could become an advantage: the more fluid the concept of a species, the more hope there is that a “thermodynamic” approach might actually be the right language to describe ecological systems, rather than simply reflecting our inability to access all the microscopic details. Importantly, a theoretical understanding of this problem will likely have direct applications in both medicine and ecology, from disease diagnosis and effects of antibiotic treatments to the impact of climate change or biodiversity conservation efforts.

With the advent of high-throughput technology, in most medically relevant science the breadth of study is increasingly becoming synonymous with its perceived promise. However, large amounts of data do not, by themselves, lead to a theoretical framework; to the contrary, they require one to be interpreted. The Standard Model of particle physics could never have emerged only as a statistical analysis of the data from the Large Hadron Collider; instead, the Higgs particle was only discovered because theory told us exactly what to look for.

In life sciences, however, theory largely serves a subordinate role as a tool to fit the data. There is a good reason why in biology many more details matter than we physicists would have liked: if changing a detail can make an organism more likely to survive, evolution will make use of it. The reductionism of simple models, so effective in many areas of physics, is met with an often justified skepticism: in physics all pre-factors are of order 1, but in biology they rarely are. However, although two particles are more complicated than a single one, and a system of three is unsolvable, a powerful idea in physics is that a gas of  $10^{23}$  particles is again easy to describe, provided we ask the right questions. If living matter is a state where all details that could matter, do matter, could this staggering complexity harbor some emerging simplicity? I believe that, in some instances, we have good reasons to be hopeful, and the “unreasonable effectiveness of mathematics” (Wigner, 1960) may apply to biology as it did to physics.

# Bibliography

- Amit, D. J. (1989). *Modelling brain function : the world of attractor neural networks*. Cambridge University Press, Cambridge ; New York.
- Andrews, B. W., Yi, T. M. and Iglesias, P. A. (2006). Optimal noise filtering in the chemotactic response of *Escherichia coli*. *Plos Comput Biol* **2**(11), 1407–1418.
- Arenas, A., Diaz-Guilera, A., Kurths, J., Moreno, Y. and Zhou, C. S. (2008). Synchronization in complex networks. *Phys Rep* **469**(3), 93–153.
- Arias, A. M. and Hayward, P. (2006). Filtering transcriptional noise during development: concepts and mechanisms. *Nature Rev Genet* **7**(1), 34–44.
- Bagowski, C. P., Besser, J., Frey, C. R. and Ferrell, J. E. (2003). The JNK cascade as a biochemical switch in mammalian cells: Ultrasensitive and all-or-none responses. *Curr Biol* **13**(4), 315–320.
- Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L. and Leibler, S. (2004). Bacterial persistence as a phenotypic switch. *Science* **305**(5690), 1622–1625.
- Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O’Shea, E., Pilpel, Y. and Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *Nature Genet* **38**(6), 636–643.
- Barabasi, A. L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell’s functional organization. *Nature Rev Genet* **5**(2), 101–113.
- Bialek, W. (2012). *Biophysics : searching for principles*. Princeton University Press, Princeton.
- Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J. and Phillips, R. (2005). Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* **15**(2), 116–124.
- Brestoff, J. R. and Artis, D. (2013), Commensal bacteria at the interface of host metabolism and the immune system. *Nature Immunol* **14**(7), 676–684.
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**(5), 335–336.

- Caporaso, J. G., Lauber, C. L., Costello, E. K., et al. (2011). Moving pictures of the human microbiome. *Genome Biol* **12**(5), R50.
- Caporaso, J. G., Lauber, C. L., Walters, W. A., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* **6**(8), 1621–1624.
- Carroll, S. B. (2005). Evolution at two levels: On genes and form. *PLoS Biol* **3**(7), 1159–1166.
- Chubb, J. R., Trcek, T., Shenoy, S. M. and Singer, R. H. (2006). Transcriptional pulsing of a developmental gene. *Curr Biol* **16**(10), 1018–1025.
- Cohen, A. A., Kalisky, T., Mayo, A., et al. (2009). Protein dynamics in individual human cells: Experiment and theory. *PLoS One* **4**(4) e4901.
- Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I. and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science* **326**(5960), 1694–1697.
- Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M. and Relman, D. A. (2012). The application of ecological theory toward an understanding of the human microbiome. *Science* **336**(6086), 1255–1262.
- Davis, I. and Ish-Horowicz, D. (1991). Apical localization of pair-rule transcripts requires 3' sequences and limits protein diffusion in the *Drosophila* blastoderm embryo. *Cell* **67**(5), 927–940.
- De Renzis, S., Elemento, O., Tavazoie, S. and Wieschaus, E. F. (2007). Unmasking activation of the zygotic genome using chromosomal deletions in the *Drosophila* embryo. *PLoS Biol* **5**(5), 1036–1051.
- DeSantis, T. Z., Hugenholtz, P., Larsen, N., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**(7), 5069–5072.
- Driever, W. and Nüsslein-Volhard, C. (1988). The bicoid protein determines position in the *Drosophila* embryo in a concentration-dependent manner. *Cell* **54**(1), 95–104.
- Dubuis, J. O., Samanta, R. and Gregor, T. (2013). Accurate measurements of dynamics and reproducibility in small genetic networks. *Mol Syst Biol* **9**, 639.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5), 1792–1797.
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**(19), 2460–2461.
- Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**(10), 996–998.

- Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**(16), 2194–2200.
- Erdmann, T., Howard, M. and ten Wolde, P. R. (2009). Role of spatial averaging in the precision of gene expression patterns. *Phys Rev Lett* **103**(25), 258101.
- Eren, A. M., Maignien, L., Sul, W. J., et al. (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol* **4**(12), 1111–1119.
- Faith, J. J., Guruge, J. L., Charbonneau, M., et al. (2013). The long-term stability of the human gut microbiota. *Science* **341**(6141), 1237439–1237439.
- Feinberg, M. (1987). Chemical-reaction network structure and the stability of complex isothermal reactors.—I. The deficiency-zero and deficiency-one theorems. *Chem Eng Sci* **42**(10), 2229–2268.
- Fierer, N. and Lennon, J. T. (2011). The generation and maintenance of diversity in microbial communities. *Am J Bot* **98**(3), 439–448.
- Foe, V. E. and Alberts, B. M. (1983). Studies of nuclear and cytoplasmic behavior during the 5 mitotic-cycles that precede gastrulation in *Drosophila* embryogenesis. *J Cell Sci* **61**(May), 31–70.
- Fredricks, D. N. (2013). *The human microbiota : how microbial communities affect health and disease*. Wiley-Blackwell, Hoboken, N.J.
- Gandhi, S. J., Zenklusen, D., Lionnet, T. and Singer, R. H. (2011). Transcription of functionally related constitutive genes is not coordinated. *Nature Struct Biol* **18**(1), 27–34.
- Garcia, H. G., Tikhonov, M., Lin, A., and Gregor, T. (2013). Quantitative imaging of transcription in living *Drosophila* embryos links polymerase activity to patterning. *Curr Biol* **23**(21), 2140–2145.
- Gergen, J. P., Coulter, D. and Wieschaus, E. F. (1986). Segmental pattern and blastoderm cell identities, in J. G. Gall, ed., *Gametogenesis and the Early Embryo*. Liss, pp. 195–220.
- Gerhart, J. and Kirschner, M. (1997). *Cells, embryos, and evolution : toward a cellular and developmental understanding of phenotypic variation and evolutionary adaptability*. Blackwell Science, Malden, Mass.
- Goentoro, L., Shoval, O., Kirschner, M. W. and Alon, U. (2009). The incoherent feedforward loop can provide fold-change detection in gene regulation. *Mol Cell* **36**(5), 894–899.

- Golding, I., Paulsson, J., Zawilski, S. M. and Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell* **123**(6), 1025–1036.
- Gregor, T., Tank, D. W., Wieschaus, E. F. and Bialek, W. (2007). Probing the limits to positional information. *Cell* **130**(1), 153–164.
- Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R. and Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol* **3**(10), 1871–1878.
- Haas, B. J., Gevers, D., Earl, A. M., and the Human Microbiome Project Consortium (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21**(3), 494–504.
- Hamady, M. and Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res* **19**(7), 1141–1152.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA-Biol* **79**(8), 2554–2558.
- Huang, Y., Niu, B. F., Gao, Y., Fu, L. M. and Li, W. Z. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**(5), 680–682.
- Hunt, D. E., David, L. A., Gevers, D., Preheim, S. P., Alm, E. J. and Polz, M. F. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**(5879), 1081–1085.
- Huse, S. M., Welch, D. M., Morrison, H. G. and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* **12**(7), 1889–1898.
- Huttenhower, C., Gevers, D., Knight, R., and the Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486**(7402), 207–214.
- Irvine, K. D., Helfand, S. L. and Hogness, D. S. (1991). The large upstream control region of the *Drosophila* homeotic gene ultrabithorax. *Development* **111**(2), 407–424.
- Kaern, M., Elston, T. C., Blake, W. J. and Collins, J. J. (2005). Stochasticity in gene expression: From theories to phenotypes. *Nature Rev Genet* **6**(6), 451–464.
- Kamada, N., Chen, G. Y., Inohara, N. and Nunez, G. (2013). Control of pathogens and pathobionts by the gut microbiota. *Nature Immunol* **14**(7), 685–690.
- Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol* **22**(3), 437–467.



- King, M. and Wilson, A. (1975). Evolution at two levels in humans and chimpanzees. *Science* **188**(4184), 107–116.
- Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M. and Glockner, F. O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**(1), e1.
- Kornberg, T. B. and Tabata, T. (1993). Segmentation of the *Drosophila* embryo. *Curr Opin Genet Dev* **3**(4), 585–594.
- Kunin, V., Engelbrektson, A., Ochman, H. and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**(1), 118–123.
- Lagha, M., Bothma, J. P. and Levine, M. (2012). Mechanisms of transcriptional precision in animal development. *Trends Genet* **28**(8), 409–416.
- Lander, A. D. (2013). How cells know where they are. *Science* **339**(6122), 923–927.
- Langille, M. G. I., Zaneveld, J., Caporaso, J. G., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnol* **31**(9), 814–821.
- Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A. and Singer, R. H. (2011). Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science* **332**(6028), 475–478.
- Lawrence, P. A. (1992). *The making of a fly : the genetics of animal design*. Blackwell Scientific Publications, Oxford ; Boston.
- Le, T. T., Harlepp, S., Guet, C. C., Dittmar, K., Emonet, T., Pan, T. and Cluzel, P. (2005). Real-time RNA profiling within a single bacterium. *Proc Natl Acad Sci USA* **102**(26), 9160–9164.
- Lee, T. I., Rinaldi, N. J., Robert, F., et al. (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**(5594), 799–804.
- Lestas, I., Vinnicombe, G. and Paulsson, J. (2010). Fundamental limits on the suppression of molecular fluctuations. *Nature* **467**(7312), 174–178.
- Levine, M. and Davidson, E. H. (2005). Gene regulatory networks for development. *Proc Natl Acad Sci USA* **102**(14), 4936–4942.
- Li, G.-W. and Xie, X. S. (2011). Central dogma at the single-molecule level in living cells. *Nature* **475**(7356), 308–315.
- Liang, H.-L., Nien, C.-Y., Liu, H.-Y., Metzstein, M. M., Kirov, N. and Rushlow, C. (2008). The zinc-finger protein Zelda is a key activator of the early zygotic genome in *Drosophila*, *Nature* **456**(7220), 400–403.

- Little, S. C., Tikhonov, M. and Gregor, T. (2013). Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell* **154**(4), 789–800.
- Little, S. C., Tkacik, G., Kneeland, T. B., Wieschaus, E. F. and Gregor, T. (2011). The formation of the Bicoid morphogen gradient requires protein movement from anteriorly localized mRNA. *PLoS Biol* **9**(3), e1000596.
- Little, S. C. and Wieschaus, E. F. (2011). Shifting patterns: Merging molecules, morphogens, motility, and methodology. *Dev Cell* **21**(1), 2–4.
- Liu, F., Morrison, A. H. and Gregor, T. (2013). Dynamic interpretation of maternal inputs by the *Drosophila* segmentation gene network. *Proc Natl Acad Sci USA* **110**(17), 6724–6729.
- Lozupone, C. and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities, *Appl Environ Microbiol* **71**(12), 8228–8235.
- Lukjancenko, O., Wassenaar, T. M., Ussery, D. W. (2010). Comparison of 61 sequenced *Escherichia coli* genomes. *Microb Ecol* **60**, 708–720.
- Maamar, H., Raj, A. and Dubnau, D. (2007). Noise in gene expression determines cell fate in *Bacillus subtilis*. *Science* **317**(5837), 526–529.
- Manu, Surkova, S., Spirov, A. V., et al. (2009). Canalization of gene expression in the *Drosophila* blastoderm by gap gene cross regulation. *PLoS Biol* **7**(3), 591–603.
- Mezard, M., Parisi, G. and Virasoro, M. A. (1987). *Spin glass theory and beyond*. World Scientific lecture notes in physics, World Scientific, Singapore ; Teaneck New Jersey.
- Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298**(5594), 824–827.
- Minsky, M. L. and Papert, S. (1969). *Perceptrons : an introduction to computational geometry*. MIT Press, Cambridge, Mass.
- Mirouze, N., Prepiak, P. and Dubnau, D. (2011). Fluctuations in spo0A transcription control rare developmental transitions in *Bacillus subtilis*. *PLoS Genet* **7**(4), e1002048.
- Morgan, M. J., Chariton, A. A., Hartley, D. M., Court, L. N. and Hardy, C. M. (2013). Improved inference of taxonomic richness from environmental DNA. *PLoS ONE* **8**(8), e71974.
- Munsky, B., Neuert, G. and van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. *Science* **336**(6078), 183–187.
- Nachman, I., Regev, A. and Ramanathan, S. (2007). Dissecting timing variability in yeast meiosis, *Cell* **131**(3), 544–556.

- Navlakha, S., He, X., Faloutsos, C. and Bar-Joseph, Z. (2014). Topological properties of robust biological and computational networks. *J Roy Soc Interface* **11**(96), 20140283.
- Newman, J. R. S., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L. and Weissman, J. S. (2006). Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **441**(7095), 840–846.
- O'Brien, T. and Lis, J. T. (1993). Rapid changes in *Drosophila* transcription after an instantaneous heat-shock. *Mol Cell Biol* **13**(6), 3456–3463.
- Ochman, H. (2003). Neutral mutations and neutral substitutions in bacterial genomes. *Mol Biol Evol* **20**(12), 2091–2096.
- Pare, A., Lemons, D., Kosman, D., Beaver, W., Freund, Y. and McGinnis, W. (2009). Visualization of individual *Scr* mRNAs during *Drosophila* embryogenesis yields evidence for transcriptional bursting. *Curr Biol* **19**(23), 2037–2042.
- Porcher, A., Abu-Arish, A., Huart, S., Roelens, B., Fradin, C. and Dostatni, N. (2010). The time to measure positional information: maternal Hunchback is required for the synchrony of the Bicoid transcriptional response at the onset of zygotic transcription, *Development* **137**(16), 2795–2804.
- Preheim, S. P., Perrotta, A. R., Martin-Platero, A. M., Gupta, A. and Alm, E. J. (2013). Distribution-based clustering: Using ecology to refine the operational taxonomic unit. *Appl Environ Microbiol* **79**(21), 6593–6603.
- Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., et al. (2007). The role of ecological theory in microbial ecology. *Nature Rev Microbiol* **5**(5), 384–392.
- Quince, C., Lanzen, A., Curtis, T. P., et al. (2009), Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods* **6**(9), 639–641.
- Quince, C., Lanzen, A., Davenport, R. J. and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinform* **12**, 38.
- Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* **4**(10), 1707–1719.
- Raj, A., Rifkin, S. A., Andersen, E. and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. *Nature* **463**(7283), 913–918.
- Rao, C. V., Wolf, D. M. and Arkin, A. P. (2002). Control, exploitation and tolerance of intracellular noise. *Nature* **420**(6912), 231–237.
- Raser, J. M. and O'Shea, E. K. (2005), Noise in gene expression: Origins, consequences, and control. *Science* **309**(5743), 2010–2013.

- Reiter, M., Kirchner, B., Mueller, H., Holzauer, C., Mann, W. and Pfaffl, M. W. (2011). Quantification noise in single cell experiments. *Nucleic Acids Res* **39**(18), e124.
- Ronen, M., Rosenberg, R., Shraiman, B. I. and Alon, U. (2002). Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci USA* **99**(16), 10555–10560.
- Rosen, M. J., Callahan, B. J., Fisher, D. S. and Holmes, S. P. (2012). Denoising PCR-amplified metagenome data. *BMC Bioinform* **13**, 283.
- Rue, P. and Garcia-Ojalvo, J. (2011), Gene circuit designs for noisy excitable dynamics. *Math Biosci* **231**(1), 90–97.
- Saffer, A. M., Kim, D. H., van Oudenaarden, A. and Horvitz, H. R. (2011). The *Caenorhabditis elegans* synthetic multivulva genes prevent ras pathway activation by tightly repressing global ectopic expression of lin-3 EGF. *PLoS Genet* **7**(12), e1002418.
- Sauer, F., RiveraPomar, R., Hoch, M. and Jackle, H. (1996). Gene regulation in the *Drosophila* embryo. *Philos Trans R Soc Lond B Biol Sci* **351**(1339), 579–587.
- Schloss, P. D., Gevers, D. and Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**(12), e27310.
- Schloss, P. D. and Westcott, S. L. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl Environ Microbiol* **77**(10), 3219–3226.
- Schloss, P. D., Westcott, S. L., Ryabin, T., et al. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**(23), 7537–7541.
- Shade, A., Caporaso, J. G., Handelsman, J., Knight, R. and Fierer, N. (2013). A meta-analysis of changes in bacterial and archaeal communities with time. *ISME J* **7**(8), 1493–1506.
- Shade, A., Peter, H., Allison, S. D., et al. (2012). Fundamentals of microbial community resistance and resilience. *Front Microbiol* **3**, 417.
- Shermoen, A. W. and Ofarrell, P. H. (1991). Progression of the cell-cycle through mitosis leads to abortion of nascent transcripts. *Cell* **67**(2), 303–310.
- Shinar, G. and Feinberg, M. (2010). Structural sources of robustness in biochemical reaction networks. *Science* **327**(5971), 1389–1391.
- Sigal, A., Milo, R., Cohen, A., et al. (2006). Variability and memory of protein levels in human cells. *Nature* **444**(7119), 643–646.

- So, L.-H., Ghosh, A., Zong, C., Sepulveda, L. A., Segev, R. and Golding, I. (2011). General properties of transcriptional time series in *Escherichia coli*. *Nature Genet* **43**(6), 554–U84.
- Song, S. J., Lauber, C., Costello, E. K., et al. (2013). Cohabiting family members share microbiota with one another and with their dogs. *eLife* **2**, e00458.
- Stewart-Ornstein, J., Weissman, J. S. and El-Samad, H. (2012). Cellular noise regulations underlie fluctuations in *Saccharomyces cerevisiae*. *Mol Cell* **45**(4), 483–493.
- Sul, W. J., Cole, J. R., Jesus, E. D., Wang, Q., Farris, R. J., Fish, J. A. and Tiedje, J. M. (2011). Bacterial community comparisons by taxonomy-supervised analysis independent of sequence alignment and clustering. *Proc Natl Acad Sci USA* **108**(35), 14637–14642.
- Taniguchi, Y., Choi, P. J., Li, G.-W., et al. (2010). Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**(5991), 533–538.
- Tautz, D., Lehmann, R., Schnurch, H., Schuh, R., Seifert, E., Kienlin, A., Jones, K. and Jackle, H. (1987). Finger protein of novel structure encoded by *hunchback*, a second member of the gap class of *Drosophila* segmentation genes. *Nature* **327**(6121), 383–389.
- Thummel, C. S., Burtis, K. C. and Hogness, D. S. (1990). Spatial and temporal patterns of E74 transcription during *Drosophila* development. *Cell* **61**(1), 101–111.
- Tikhonov, M. and Bialek, W. (2014). Complexity in generic biochemical circuits: topology versus strength of interactions. (in review; arXiv:1308.0317 [q-bio.MN])
- Tikhonov, M., Leach, R. W. and Wingreen, N. S. (2014). Interpreting 16S metagenomic data without clustering to achieve sub-otu resolution. *ISME J*, in press.
- Tkacik, G., Callan, C. G. and Bialek, W. (2008). Information flow and optimization in transcriptional regulation. *Proc Natl Acad Sci USA* **105**(34), 12265–12270.
- Tkacik, G., Walczak, A. M. and Bialek, W. (2009). Optimizing information flow in small genetic networks. *Phys Rev E* **80**(3), 031920.
- Tourova, T. P. (2003). Copy number of ribosomal operons in prokaryotes and its effect on phylogenetic analyses, *Microbiol* **72**(4), 389–402.
- Turnbaugh, P. J., Quince, C., Faith, J. J., et al. (2010). Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci USA* **107**(16), 7503–7508.
- Tyson, J. J., Chen, K. C. and Novak, B. (2003). Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr Opin Cell Biol* **15**(2), 221–231.

- VandeWalle, J. L., Goetz, G. W., Huse, S. M., Morrison, H. G., Sogin, M. L., Hoffmann, R. G., Yan, K. and McLellan, S. L. (2012). Acinetobacter, aeromonas and trichococcus populations dominate the microbial community within urban sewer infrastructure. *Environ Microbiol* **14**(9), 2538–2552.
- von Dassow, G., Meir, E., Munro, E. M. and Odell, G. M. (2000). The segment polarity network is a robust development module. *Nature* **406**(6792), 188–192.
- Wigner, E. P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Commun Pure Appl Math* **13**(1), 1–14.
- Wilkie, G. S., Shermoen, A. W., O’Farrell, P. H. and Davis, I. (1999). Transcribed genes are localized according to chromosomal position within polarized *Drosophila* embryonic nuclei. *Curr Biol* **9**(21), 1263–1266.
- Youngblut, N. D., Shade, A., Read, J. S., McMahon, K. D. and Whitaker, R. J. (2013). Lineage-specific responses of microbial communities to environmental change. *Appl Environ Microbiol* **79**(1), 39–47.
- Zenklusen, D., Larson, D. R. and Singer, R. H. (2008). Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Struct Biol* **15**(12), 1263–1271.
- Zheng, Z. J., Kramer, S. and Schmidt, B. (2012). DySC: software for greedy clustering of 16S rRNA reads. *Bioinformatics* **28**(16), 2182–2183.
- Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins, in V. Bryson and H. J. Vogel, eds, *Evolving genes and proteins*. Academic Press, pp. 97–166.