fortune Journals

Research Article

JOURNAL OF BIOINFORMATICS AND SYSTEMS BIOLOGY

ISSN: 2688-5107

Analysis and Modeling of Early Estradiol-induced GREB1 Single Allele Gene Transcription at the Population Level

Seyyed Mahmood Ghasemi [1,2], Pankaj K. Singh[2,3], Hannah L. Johnson[2,4], Ayse Koksoy[2,3], Michael A Mancini[2,3,4,5,6], Fabio Stossi [2,4,5] and Robert Azencott[1,*]

Abstract

Single molecule fluorescence in situ hybridization (smFISH) can be used to visualize transcriptional activation at the single allele level. We and others have applied this approach to better understand the mechanisms of activation by steroid nuclear receptors. However, there is limited understanding of the interconnection between the activation of target gene alleles inside the same nucleus and within large cell populations.

Using the transcriptional coactivator GREB1 gene as an early estrogen receptor (ER) response target, we applied smFISH to track E2-activated GREB1 allelic transcription over early time points to evaluate potential dependencies between alleles within the same nucleus. We compared two types of experiments where we altered the initial status of GREB1 basal transcription by treating cells with and without the elongation inhibitor flavopiridol (FV).

E2 stimulation changed the frequencies of active GREB1 alleles in the cell population, and this was independent of FV pre-treatment. In FV treated cells, the response time to hormone was delayed, albeit still reaching at 90 minutes the same levels as in cells not treated by FV. We show that the joint frequencies of GREB1 activated alleles observed at the cell population level imply significant dependency between pairs of alleles within the same nucleus. We identify probabilistic models of joint alleles activations by applying a principle of maximum entropy. For pairs of alleles, we have then quantified statistical dependency between their GREB1 activations by computing their mutual information. To further analyze the time course of GREB1 activation observable at the population level, we have introduced a stochastic model compatible with allelic statistical dependencies, and we have fitted this model to our data by intensive simulations. This provided estimates of the average lifetime for degradation of GREB1 introns and of the mean time between two successive transcription rounds. Our approach informs on how to extract information on single allele regulation by the estrogen receptor from within a large population of cells, and should be applicable to many other genes.

Keywords: Gene Transcription; GREB1; smFISH; Single Allele;

Introduction

Time-dependent modulation of gene transcription is necessary for a cell to respond to stimuli in a dynamic and reversible manner. The difficulty in dissecting the complex mechanisms of biological responses is enhanced by the fact that gene transcription in individual cells within a population appears to be vastly heterogeneous [1–8].

Affiliation:

[1]Department of Mathematics, University of Houston, Houston, TX, USA

[2]GCC Center for Advanced Microscopy and Image Informatics, Houston, TX, USA

[3]Center for Translational Cancer Research, Institute of Biosciences and Technology, Texas A&M University, Houston, TX

[4]Integrated Microscopy Core, Baylor College of Medicine, Houston, TX, USA

[5]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA

[6]Department of Pharmacology and Chemical Biology, Baylor College of Medicine, Houston, TX, USA

***Corresponding author:**
Robert Azencott, Dept of Mathematics, University of Houston, Houston, TX, USA,

Citation: S.Mahmood Ghasemi, Pankaj Singh K, Hannah Johnson L, Ayse Koksoy, Michael Mancini A, Fabio Stossi and Robert Azencott. Analysis and Modeling of Early Estradiol-induced GREB1 Single Allele Gene Transcription at the Population Level. Journal of Bioiformatics and Systems Biology. 7 (2024): 108-128.

Received: September 04, 2023
Accepted: September 11, 2023
Published: June 10, 2024

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

We and others have used single molecule RNA FISH (smFISH) to study the effect of stimuli on gene transcription by population analysis of fixed samples [3,7], which facilitates the capture of a large number of events, their spatial location, and the nascent RNA from individual alleles; however, this is at the expense of time dynamics. Specifically, we have focused on the estrogen receptor (ER), a well-established model for transcriptional response to hormones (*i.e.,* 17β-estradiol, E2), using GREB1 as a prototypical early ER target gene [2,8,9]. In our previous study [8], we identified that GREB1 responded to E2 in a cell- and allele-dependent manner, and that the frequency of allele activation was tunable by specific epigenetic inhibitors, indicating that the cell has mechanisms in place to control the frequency of allelic responses to stimuli.

To complement the population analysis, several time dynamic studies have been performed. The method of choice has been engineering one or more copies of a gene to contain repeat sequences (*e.g.,* MS2, PP7) that are recognized by specific fluorescent proteins [10-12]. These live studies proposed that gene transcription occurs in stochastic bursts, where a gene is ON for a short period of time, followed by longer periods of inactivity. Live cell experiments of engineered GREB1 alleles [2], and another prototypical E2-target, TFF1 [1], have shown a highly-dynamic response to hormone in individual cells. From these pivotal studies, the following observations have been made: 1) E2 regulates the frequency of bursting by reducing the promoter OFF times; 2) extrinsic noise governs the cell-by-cell heterogeneity in response; 3) there is some correlation between alleles in the same nucleus; and, 4) at the level of individual cells, live imaging experiments can be accurately fitted by a stochastic model driven by a two-states promoter.

To model the stochastic gene transcription bursts observed in individual cells, several papers [1,2,13-20] have introduced and simulated engineered gene promoter models randomly switching between one active state and several inactive states. The most popular have been two states models where the gene promoter is either ON or OFF. For instance, in Fritzch et al [2]., "two-states" models were fitted to GREB1 transcription bursts observed live in single cells, with model parameters exhibiting quite strong fluctuations from cell-to-cell, which encouraged our present population-level study of endogenous GREB1 gene expression.

Here, we sought to study and model the initial phases of hormone stimulation by using smFISH on fixed MCF-7 breast cancer cells in culture. In our experiments, smFISH images are acquired every 15 minutes, at times $T_0 = 0$, $T_1 = 15$ min,...,$T_6 = 90$ min. At each time $T = T_j$ a large cell population pop(T) of N(T) cells is imaged, with N(T) ranging from 400 to 1000. We compared two types of initial conditions:

- in (FV+E2)-experiments, a flavopiridol (FV) pretreatment of cells started two hours before T_0 to block ongoing transcriptional elongation [21] until FV release at T_0, which synchronizes the initiation step of the transcriptional cycle.

- in E2-experiments, cells were maintained in "native" state at $T = T_0$, so that the transcription cycle was random before T_0 At T_0 each experiment started from seven distinct initial cell populations {init(0), ..., init(7)}. Each population init(j) evolved separately from time T_0 until T_j, and the state of $pop(T_j)$ was only imaged at time T_j. This approach does not enable tracking the same cells and active alleles across time. At each time $T = T_j$, image analysis of smFISH data provided GREB1 transcription statistics across all cells of $pop(T_j)$. For k=0,1,2,3,4 we computed the frequencies Q_k(T) of nuclei exhibiting "k" detectable nascent mRNAs ("active alleles"). The frequencies Q_k(T) aggregate transcription activities over the N(T) cells of pop(T). As seen in [1,2,13-20], transcription bursts of these N(T) individual cells are *random* and clearly *non synchronous*, so that our transcription data which aggregate GREB1 transcriptions of individual cells across a large population pop(T) provide population-level activation frequencies which evolve smoothly in time with no significant bursts. To model the smooth time dynamics of the frequencies Q_k(T), we introduced a *"population-level"* stochastic transcription model involving four key parameters:

1) mean waiting time "A" between successive productive transcription rounds;
2) mean lifetime "L" of nascent mRNA;
3) mean elongation time "MTD" to complete one mRNA; and,
4) the minimal number "VTH" of RNA molecules enabling fluorescence detection.

Parameters estimation for our population-level model was implemented by massive stochastic simulations to reach a good fit to smFISH imaging data across biological replicate experiments.

A key step was to show that at each time $T = T_k$, the population-level transcription activation frequencies indicated *statistical dependencies* between pairs (AL_1, AL_2) of GREB1 alleles within individual nuclei. At each time $T = T_j$, we applied a maximum entropy principle to fit a probabilistic model to the frequencies of joint allele activations observed at the population level. This enabled the quantification of dependencies between pairs of alleles by computing their Mutual Information. While the present study focused on GREB1 gene transcription, we expect that the algorithmic modeling and data analysis techniques that we developed will be applicable for other hormone-induced genes transcription.

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

Results

Time course analysis of early E2-induced GREB1 gene transcription by smFISH before and after treatment by flavopiridol

As we have shown previously [8], E2 activates GREB1 gene transcription in a cell- and allele-dependent manner, as measured by smFISH using spectrally separated exon and intron probe sets. Here, we sought to focus on the initial phase of hormonal stimulation (first 90 minutes) by measuring cell population pop(T) responses at 15 minutes intervals (T_0 = 0, T_1 = 15 min, …, T_6 = 90 min) under two types of cell state initial conditions and three independent biological replicates per condition type. The first initial condition was to consider the initial state of gene transcription as random, i.e. all individual alleles already have their own past "history" with RNA polymerases II located at random phases during either initiation or elongation, which is represented by cells grown for 48 hours in hormone-depleted media. Cell population transcriptional data after this initial cell state are displayed by RED curves and labeled "E2-curves" in every Figure. The second initial condition was designed to arrest transcription elongation by using the reversible CDK9 inhibitor, flavopiridol (FV), for 2 hours before E2 treatment at T_0, causing elongation to stop and RNA polymerase II to stall at gene promoters [21]. We then released the FV block by three washes, and stimulated GREB1 gene transcription by hormone treatment (E2, 10nM); in the Figures the corresponding population data are displayed by BLUE curves and labeled "FV+E2".

For each independent biological replicate, the analyzed pop(T) ranged from approximately 400 to 1000 cells, captured by high resolution (60x/1.42NA) epifluorescence deconvolution microscopy (representative images are in **Figure 1A**). Our temporal resolution has limitations as GREB1 probe visualization requires ~20 individually labeled fluorescent oligo probes (out of 48) to bind to nascent RNA, which, based on their location on the gene, occurs when the GREB1 is ~60-70% transcribed; for sequences of oligos refer to [8,22]. On average, the speed of RNA Polymerase II in mammalian cells (i.e., ~2-2.5Kb/min, [21,23,24]), suggests that detection of newly made, partial RNAs could occur only after ~30 minutes of E2 induction.

As smFISH experiments do not directly follow the activation of individual alleles live, the best proxy is to analyze a time series that evaluates transcriptional events across the population pop(T) of size N(T) using several statistical approaches describing the set of observable active alleles. In each nucleus, we used custom automated image analysis

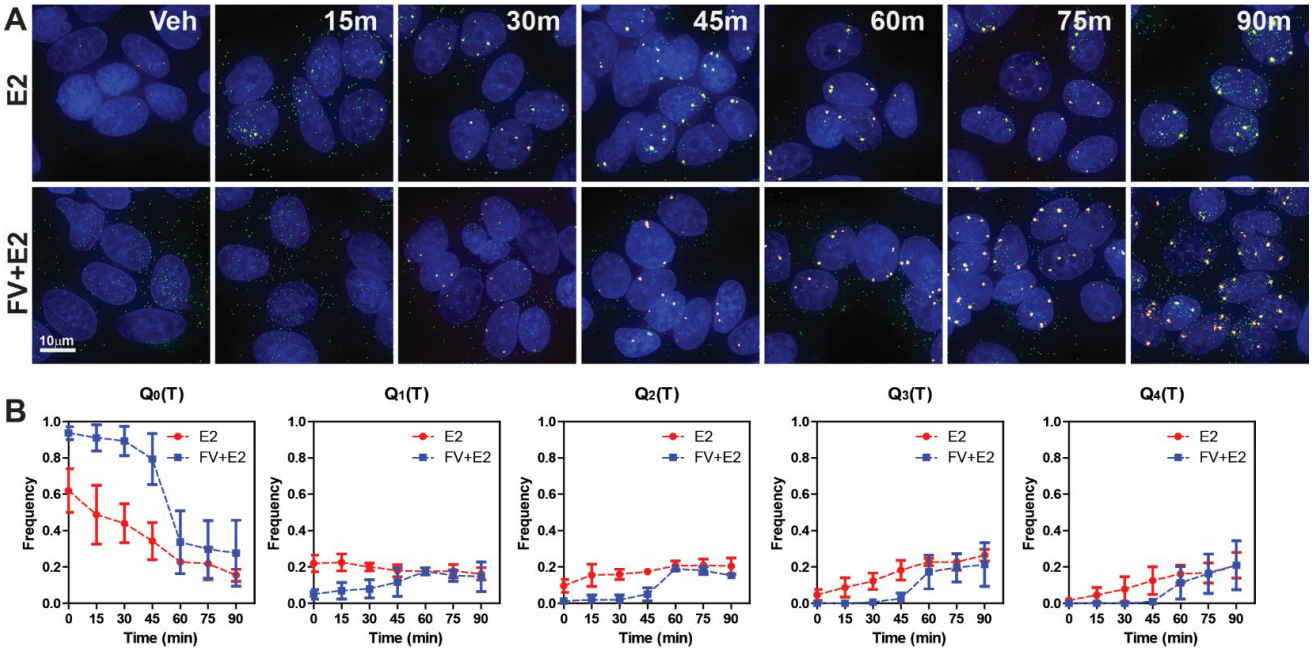

Figure 1: GREB1 smFISH time course analysis in MCF-7 breast cancer cells with and without flavopiridol block/release. A). MCF-7 cells were treated for the indicated times with 10nM E2 and GREB1 smFISH was performed at each time point. Images are at 60x/1.42, deconvolved and max projected. Red spots represent intronic and green spots exonic probe sets. Samples labeled as FV+E2, were pretreated with 1μM flavopiridol (FV) for 2 hours, followed by three washes and E2 treatment. Scale bar: 10μm. B). The time courses of five frequencies $\{Q_0(T),Q_1(T),Q_2(T),Q_3(T),Q_4(T)\}$ of active GREB1 alleles/cell are shown as follows. The red curves display the frequencies $Q_k(T)$ after averaging over three E2 experiments. The blue curves display the $Q_k(T)$ after averaging over three FV+E2 experiments. The vertical bars display the dispersion of $Q_k(T)$ values over three similar independent experiments. Note that the dispersion of the $Q_k(T)$ values across 3 experiments is much larger than the standard error of estimation affecting $Q_k(T)$ in each experiment. At the end of all experiments (T= 90min), all $Q_k(T)$ stabilize to a value ≈ 20%.

Citation: S.Mahmood Ghasemi, Pankaj Singh K, Hannah Johnson L, Ayse Koksoy, Michael Mancini A, Fabio Stossi and Robert Azencott. Analysis and Modeling of Early Estradiol-induced GREB1 Single Allele Gene Transcription at the Population Level. Journal of Bioinformatics and Systems Biology. 7 (2024): 312-332.

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

(described in detail in the Methods section), to identify the number k = 0,1,2,3,4 of active GREB1 alleles. From this we derived the total number $N_k(T)$ of nuclei exhibiting k active alleles, thus yielding the five frequencies $Q_k(T) = N_k(T)/N(T)$ The behavior of the cell population at each time point is then represented by the five GREB1 activation frequencies $\{Q_0(T),Q_1(T),Q_2(T),Q_3(T),Q_4(T)\}$,which naturally add up to 1 (**Figure 1B**). These frequencies were calculated across more than 400 cells per replicate, with standard error margins of the order of 2.5% (see Supplemental Materials). The effects of flavopiridol block/release on E2-induced gene transcription (**Figure 1B**, blue curves) appear to result in: 1) a significant increase in $Q_k(T)$ at all the time points T< 60 min; and, 2) a corresponding decrease in all the other $Q_k(T)$, collectively indicating that the FV pre-treatment was effective. More noteworthy are the two following observations: 1) the cell-to-cell and allele-by-allele variation in responses is maintained if transcription elongation is manipulated indicating that this part of the transcription cycle is not controlling synchronicity of responses to hormone; and, 2) the final frequencies of activation at time points T larger than 60min are virtually identical whether we used FV pre-treatment or not, indicating that the E2 response "catches up" independently of the starting conditions, so that a random cell state starting condition (i.e. no FV pre-treatment) does not offer an advantage in term of response to hormone over time.

Allelic activations by E2-induced GREB1 transcription exhibit signi icant statistical dependency.

We explored whether GREB1 activation occurred independently at the four alleles present in each of the aneuploid MCF-7 nuclei by estimating the probabilities of joint activation for pairs of alleles AL_1, AL_2 At time T, in any nucleus NUC_n, each allele can either be ON or OFF, therefore yielding 16 distinct possible joint activation states $\{S_0,S_1,S_2,..., S_{15}\}$for the 4 alleles AL_1,AL_2,AL_3,AL_4. Denote $prob_n(S_k)$ the probability that the four alleles in NUC_n are in the joint state S_k. The 16 probabilities $prob_n(S_0)$, $prob_n(S_1),...,prob_n(S_{15})$ add up to 1, and depend on time T. Due to population heterogeneity, $prob_n(S_k)$ will depend on many cell extrinsic and intrinsic factors specific to each NUC_n ; hence, our image data could only record averages $F_T(S_k)$ of the probabilities $prob_n(S_k)$ over all nuclei NUC_n of pop(T). Concretely, since image resolution did not enable allele matching between distinct cells, the probabilities $F_T(S_k)$ were not directly computable from image analysis, which could only provide the five observed frequencies $Q_0(T),Q_1(T),Q_2(T),Q_3(T),Q_4(T)$ shown in Figure 1B. For each time point T, the algorithmic challenge was hence to compute 16 unknown probabilities $F_T(S_0),....,F_T(S_{15})$ starting only from the 5 observed frequencies $Q_0(T),Q_1(T),Q_2(T),Q_3(T),Q_4(T)$. We identified the five explicit linear relations (see Methods Equation 1) expressing the $Q_k(T)$ in terms of the $F_T(S_k)$, but this only provided five explicit linear constraints on our 15 unknowns $F_T(S_k)$. To handle this estimation problem, a natural first approach was to assume that under the probability distribution F_K, one had statistical independence of activations among the four GREB1 alleles. In practical terms, independence means that there is no quantifiable mechanism through which any allele interferes or influences activation potential in other alleles in the same nucleus. We have proved (see Methods Equation 2) that, if the probability F_T of joint activations involved statistical independence between alleles activations, the observed frequencies $Q_0(T),Q_1(T),Q_2(T),Q_3(T),Q_4(T)$ would have to verify extremely restrictive polynomial constraints (see Methods MM5 and Equation 2).

All our experiments revealed that these polynomial constraints were never satisfied by the observed $Q_0(T),Q_1(T),Q_2(T),Q_3(T),Q_4(T)$. Thus, to evaluate the $F_T(S_k)$ from frequencies of joint alleles activations observed at population level, we had to reject the hypothesis of statistical independence between activations of the four distinct alleles within a nucleus.

We have confirmed this theoretical result, by comparing, at each time point, the experimental activation frequencies $Q_0(T),Q_1(T),Q_2(T),Q_3(T),Q_4(T)$ with the analogous activation frequencies generated by joint probability models F_T based on independence between allele activations. Indeed, independence would imply that each probability model F_T is fully determined once one specifies activation frequencies at time T for each single allele AL_1,AL_2,AL_3,AL_4. For each time T, we have generated 10^6 such models F_T based on the hypothesis of independence between alleles activations and computed the associated virtual activations frequencies $[virQ_0(T),virQ_1(T),virQ_2(T),virQ_3(T), virQ_4(T)]$ to compare them to the observed $[Q_0(T), Q_1(T), Q_2(T), Q_3(T), Q_4(T)]$, Our results clearly show that none of these 106 sets of virtual frequencies $[virQ_0(T),virQ_1(T),virQ_2(T),virQ_3(T),virQ_4(T)]$ could match the experimentally observed $[Q_0(T), Q_1(T), Q_2(T), Q_3(T), Q_4(T)]$. This was visualized by scatter plots, each one displaying a pair $[Q_i(T), Q_j(T)]$ observed over time for each experiment. For example, **Figure 2** separately displays scatter plots for the two pairs $[Q_0(T), Q_1(T)]$ and $[Q_0(T), Q_4(T)]$. The red dots represent these pairs of frequencies observed via smFISH or E2 experiments. The blue dots similarly display the same pairs for FV+ E2 experiments. Arrows indicate the observations of frequencies $[Q_0(T), Q_1(T)]$ and $[Q_0(T), Q_4(T)]$. at successive times T for an individual experiment. The 10^6 green dots represent the virtual pairs $[virQ_0(T), virQ_1(T)]$ or $[virQ_0(T), virQ_4(T)]$. associated to 10^6 virtual joint probability models F_T based upon the independence hypothesis. As shown in **Figure 2**, both blue and red dots are positioned well away from green dots, thus confirming that, for population level modeling, one must reject the hypothesis of independence between alleles.

Citation: S.Mahmood Ghasemi, Pankaj Singh K, Hannah Johnson L, Ayse Koksoy, Michael Mancini A, Fabio Stossi and Robert Azencott. Analysis and Modeling of Early Estradiol-induced GREB1 Single Allele Gene Transcription at the Population Level. Journal of Bioinformatics and Systems Biology. 7 (2024): 312-33.

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

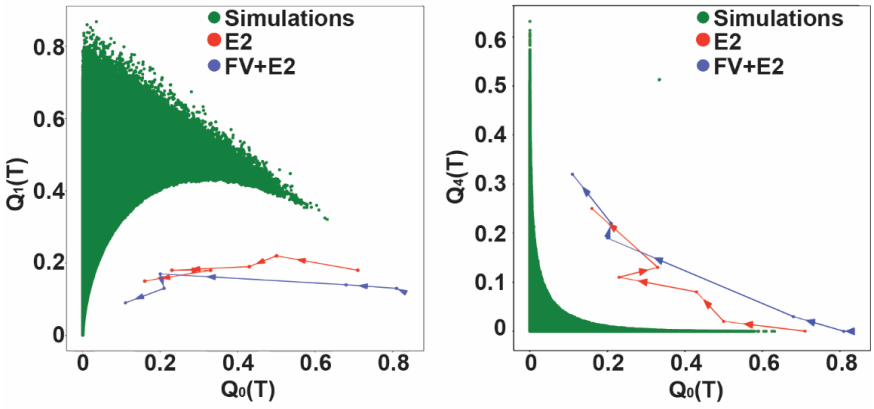

Figure 2: Statistical dependency between GREB1 alleles activations. Scatter plot representation of two frequency pairs $[Q_0(T), Q_1(T)]$ on the left and $[Q_0(T), Q_4(T)]$ on the right. The blue (FV+E2) and red (E2) curves display the real data time evolutions for the pair of frequencies observed in two experiments. Arrows indicate the time direction. On the left panel, each of the 10^6 green dots represent one pair of frequencies $[q_0(T), q_1(T)]$ generated by at least one model with independent alleles. Note that the green dots always remain distinct from the experimental red and blue dots. Similar graphs were obtained for any pair $Q_i(T)$, $Q_j(T)$ with $i \neq j$. This indicates that probabilistic modeling of transcription activities aggregated at cell population level requires assuming some dependency between alleles activations.

As just showed above, to be compatible with experimental data, the unknown average frequencies $F_T(S_0),\ldots, F_T(S_{15})$ of jointly activated alleles across cell population pop(T) must exhibit dependencies between alleles activations. Each one of the 5 observed frequencies $Q_k(T)$ is an explicit linear combination of the 16 unknowns $F_T(S_0),\ldots, F_T(S_{15})$ (see Methods, Equation1). Since there was no probabilistic model F_T achieving zero dependencies between alleles activations and also verifying these 5 linear constraints, we decided to seek a model F_T compatible with these 5 constraints and *minimizing dependencies* between alleles activations. For fitting a joint probability distribution F_T to data under linear constraints, a generic principle is that minimizing dependencies is approximately equivalent to *maximizing the entropy Ent(F_T)* of the probability model F_T under the same linear constraints (see Methods MM6). This maximum entropy principle is well established in the physics of gases or of spinglass magnets arrays, and has also successfully been used to model images by Gibbs distributions [25,26]. Here we have applied this principle to theoretically compute the unique joint probability F_T of activated alleles which has maximum entropy among all probabilities compatible with the five observed frequencies $Q_0(T)$, $Q_1(T)$, $Q_2(T)$, $Q_3(T)$, $Q_4(T)$. We then proved that this max-entropy probability model F_T has full symmetry, meaning that arbitrary permutations of the four alleles do not change the frequencies of their joint state of activations. For instance, due to average probability modeling of the whole population pop(T), each of the four alleles has the same activation probability $P_{AL1}(T)$ at time T. We have thus obtained explicit formulas expressing each one of the 16 unknowns $F_T(S_0),\ldots, F_T(S_{15})$ in terms of the five observed frequencies $Q_k(T)$ (see Methods, equation 1). These formulas also gave us explicit expressions for two key probabilities (see Methods, Equation 3), namely: 1) for each single allele AL_1, the probability $P_{AL_1}(T)$ that AL_1 will be active at time T; and 2) for each pair of alleles AL_1, AL_2 in the same nucleus, the probability $P_{AL1,AL2}(T)$ that AL_1 and AL_2 will be simultaneously active at time T. The probabilities $P_{AL_1}(T)$ and $P_{AL_1 AL_2}(T)$ do not change if we replace AL_1, AL_2 by any other two alleles AL_i, AL_j within the same nucleus. This is due to the full symmetry of the joint probabilities $F_T(S_k)$, a property which was derived from the maximum entropy principle.

In **Figure 3 A**, we display the evolution of $P_{AL_1}(T)$ over time. For FV+ E2 experiments (blue curve), the initial $P_{AL_1}(0)$ is nearly 0 and $P_{AL_1}(T)$ remains practically equal to zero until T=30min since GREB1 transcription duration is of the order of 40 min and GREB1 transcriptions are nearly blocked by FV before T = 0. For E2 experiments (red curves), $P_{AL_1}(0)$ is naturally higher than for FV+E2 experiments (blue curves) due to some GREB1 transcription activity at low level before infusion of E2 at T=0. For all experiments, $P_{AL_1}(T)$ increases steadily with T and reaches a maximum ranging from 40% to 60% at T= 90min.

For each pair of alleles (AL_1, AL_2) in the same nucleus, their joint random activation states at time T can have only one of four possible configurations: active/active, active/inactive, inactive/active, or inactive/inactive. We have explicitly computed the probabilities of these four configurations in terms of the observed frequencies $Q_k(T)$ The probability $P_{AL_1 AL_2}(T)$ that the two given alleles AL_1 and AL_2 are simultaneously active at time T was then plotted for all experiments (see **Figure 3B**). For FV+ E2 experiments (blue curves), $P_{AL_1 AL_2}(T)$ remains quite low until T=30min since most polymerase elongations started after time T=0 is still too incomplete at T=30min to be reliably detectable. For all experiments, $P_{AL_1 AL_2}(T)$ increases with T, and reaches a maximum ranging from 25% to 45% at T = 90 min.

Citation: S.Mahmood Ghasemi, Pankaj Singh K, Hannah Johnson L, Ayse Koksoy, Michael Mancini A, Fabio Stossi and Robert Azencott. Analysis and Modeling of Early Estradiol-induced GREB1 Single Allele Gene Transcription at the Population Level. Journal of Bioinformatics and Systems Biology. 7 (2024): 312-33.

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

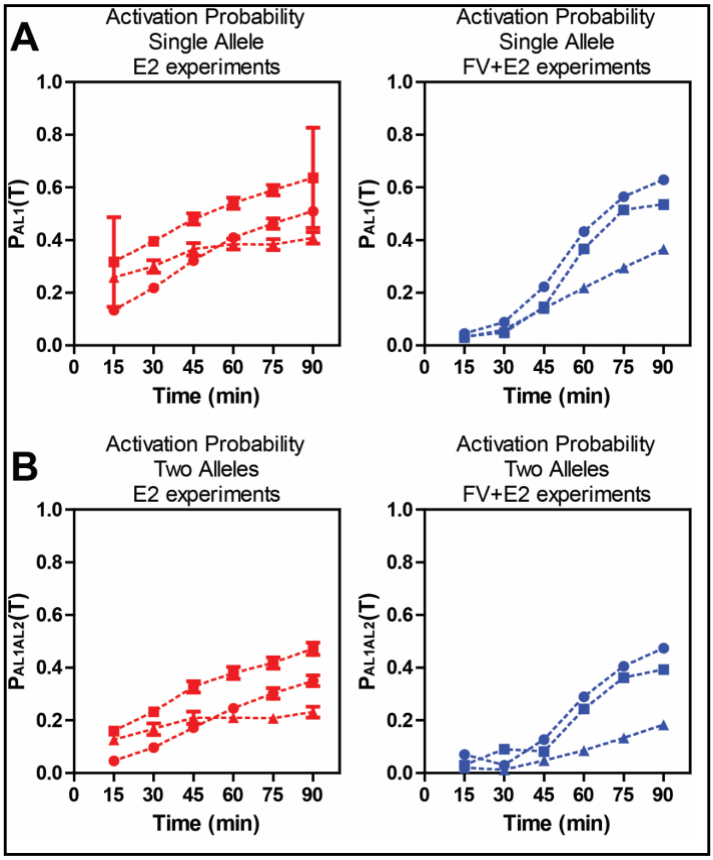

Figure 3: Time course for GREB1 activation probabilities $P_{AL_1}(T)$ and $P_{AL_1\,AL_2}(T)$**.** In (A), $P_{AL_1}(T)$ is the computed probability that a given single allele AL_1 will be GREB1 activated at time T, and in (B), $P_{AL_1\,AL_2}(T)$ is the joint probability that a given pair of alleles AL_1, AL_2 will both be activated at time T. The time courses of $P_{AL_1}(T)$ and $P_{AL_1\,AL_2}(T)$ between T=15min and T=90min are displayed by red curves for three E2 experiments and by blue curves for three (FV+E2) experiments. The vertical bars display the standard errors on these probabilities.

Table 1a: Activation probability $P_{AL1}(T)$ displayed in % for single allele AL_1

$P_{AL1}(T)$	E2 exp 1	E2 exp 2	E2 exp 3	FV+E2 exp 4	FV+E2 exp 5	FV+E2 exp 6
T = 0	8.8 +/- 0.9	25.2 +/- 1.6	NA	3.0 +/- 0.6	3.0 +/- 0.7	NA
T = 15 min	13.4 +/- 1.1	31.7 +/- 1.7	25.8 +/- 1.9	4.5 +/- 0.6	3.5 +/- 0.4	3.0 +/- 0.4
T = 30 min	21.8 +/- 1.4	39.5 +/- 1.7	30.0 +/- 2.4	8.8 +/- 0.9	4.8 +/- 0.5	6.0 +/- 0.5
T = 45 min	32.2 +/- 1.8	48.0 +/- 2.1	36.5 +/- 2.4	22.3 +/- 1.6	14.6 +/- 1.2	14.0 +/- 0.8
T = 60 min	41.0 +/- 1.8	54.0 +/- 2.1	38.5 +/- 2.0	43.2 +/- 1.5	36.6 +/- 1.4	21.8 +/- 0.9
T = 75 min	46.3 +/- 2.0	59.0 +/- 2.0	38.3 +/- 2.1	56.5 +/- 1.4	51.5 +/- 1.6	29.5 +/- 0.9
T = 90 min	51.0 +/- 1.9	63.6 +/- 1.9	40.8 +/- 2.2	62.9 +/- 1.3	53.5 +/- 1.4	36.5 +/- 1.1

Table 1b: Joint activation probability $P_{AL1\,AL2}(T)$ displayed in % for two alleles AL_1, AL_2

$P_{AL1\,AL2}(T)$	E2 exp 1	E2 exp 2	E2 exp 3	FV+E2 exp 4	FV+E2 exp 5	FV+E2 exp 6
T= 0	2.8 +/- 0.6	11.4 +/- 1.3	NA	0.5 +/- 0.2	0.3 +/- 0.2	NA
T= 15 min	4.6 +/- 0.7	16.0 +/- 1.5	12.6 +/- 1.6	0.7 +/- 0.2	0.3 +/- 0.1	0.2 +/- 0.1
T= 30 min	9.6 +/- 1.1	23.3 +/- 1.7	16.7 +/- 2.2	3.0 +/- 0.6	0.9 +/- 0.2	1.2 +/- 0.2
T= 45 min	17.2 +/- 1.6	32.7 +/- 2.2	21.0 +/- 2.3	12.7 +/- 1.4	8.1 +/- 1.0	4.7 +/- 0.5
T= 60 min	24.5 +/- 1.7	38.0 +/- 2.3	21.0 +/- 1.9	29.0 +/- 1.4	24.3 +/- 1.4	8.5 +/- 0.7
T= 75 min	30.2 +/- 2.0	41.8 +/- 2.2	20.8 +/- 1.9	40.5 +/- 1.5	36.2 +/- 1.7	13.3 +/- 0.7
T= 90 min	35.0 +/- 2.0	47.2 +/- 2.3	23.1 +/- 2.1	47.4 +/- 1.5	39.3 +/- 1.4	18.3 +/- 1.0

Citation: S.Mahmood Ghasemi, Pankaj Singh K, Hannah Johnson L, Ayse Koksoy, Michael Mancini A, Fabio Stossi and Robert Azencott. Analysis and Modeling of Early Estradiol-induced GREB1 Single Allele Gene Transcription at the Population Level. Journal of Bioinformatics and Systems Biology. 7 (2024): 108-128.

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

Statistical Validation of Dependency between GREB1 alleles activations after E2 treatment

For our probabilistic model F_T fitted by maximum entropy to actual GREB1 activation frequencies aggregated at the cell population level, our explicit computation of the probabilities $P_{AL_1}(T)$ and $P_{AL_1\,AL_2}(T)$ enabled us to test whether the activation of alleles AL_1, AL_2 is statistically dependent of each other, and to quantify their statistical dependency. Indeed at time T, statistical independence for the activation of alleles AL_1, AL_2 would classically imply the equality $P_{AL_1\,AL_2}(T) > P_{AL_1}(T) \times P_{AL_2}(T)$, In our experiments this equality is *significantly not satisfied*, as validated by our detailed analysis of estimation errors on $P_{AL_1\,AL_2}(T)$, $P_{AL_1}(T)$, $P_{AL_2}(T)$ (see Methods MM5,MM6).

At time T, the conditional probability that {AL_2 is active} given that {AL_1 is active} is classically computed by $prob_T(AL_2 \text{ active} \mid AL_1 \text{ active}) = P_{AL_1\,AL_2}(T)/P_{AL_1}(T)$, For all our 6 experiments (see **Figure 4**), we have $P_{AL_1\,AL_2}(T) > P_{AL_1}(T) \times P_{AL_2}(T)$ at all time points $T \geq 15$min. This forces the conditional probability $prob_T(AL_2 \text{ active} \mid AL_1 \text{ active})$ to be always larger than the unconditioned probability $P_{AL_2}(T) = prob_T(AL_2 \text{ active}$. This is a clear indicator of statistical dependency between the activations of AL_1 and AL_2. Indeed one can quantify the level of dependency between activations of AL_1 and AL_2 by comparing the **dependency ratio** $dep(T) = P_{AL_1\,AL_2}(T)/P_{AL_1}(T) \times P_{AL_2}(T)$ to the baseline value 1. Our detailed statistical study of estimation errors on dep(T) (see Methods section MM6) shows that *the inequality* $\boldsymbol{dep(T) > 1}$ *is significantly valid* at the 95% confidence level for all our

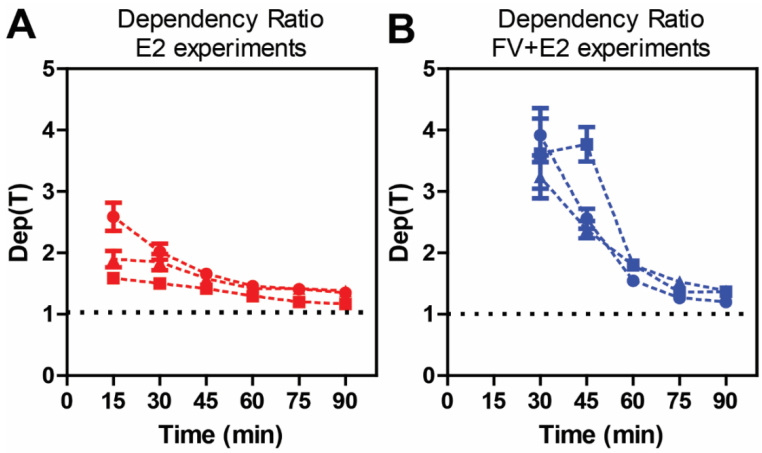

Figure 4: For pairs of alleles AL_1, AL_2, the dependency ratio $\boldsymbol{dep(T)}$ is significantly larger than 1 at confidence level 95%, which indicates significant statistical dependency between GREB1 activations of AL_1 and AL_2. We display the time course of $\boldsymbol{dep(T)}$ at all $T \geq 15$min for three E2 experiments (see A), and at all T ≥30min for three FV+E2 experiments (see B). The vertical bars display the standard errors of estimation on dep(T). Note that the ratio dep(T) remains larger than 1.15 in these time ranges.

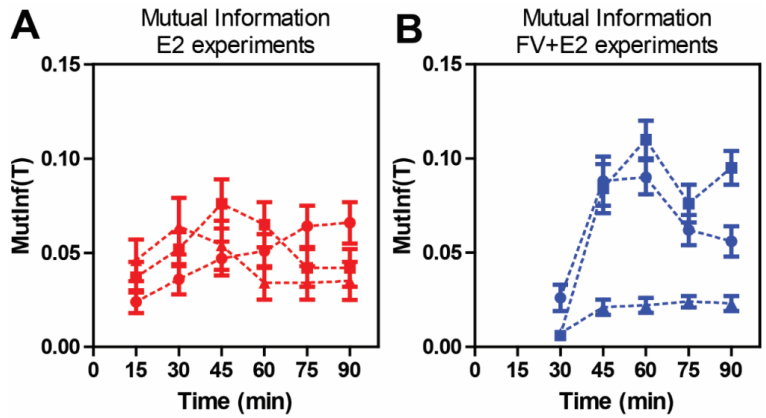

Figure 5: Mutual Information between pairs of alleles. For each one of our experiments, and at all time points T, we have computed the Mutual Information $MutInf_{AL_1,AL_2}(T)$ between the stochastic GREB1 activations of any given pair of alleles AL_1, AL_2 within the same nucleus. We display the time course of $MutInf_{AL_1,AL_2}(T)$ for three E2 experiments and $T \geq 15$ min (see A), as well as for three FV+E2 experiments and $T \geq 30$min (see B). The vertical bars display the standard errors on $MutInf_{AL_1,AL_2}(T)$ and show that in these time ranges, the mutual information is always significantly positive with confidence level 95%, which indicates a statistically significant level of dependency between GREB1 activations for pairs of alleles AL_1, AL_2. At times $T = 0$, $T = 15$min for FV+E2 experiments, and at $T = 0$ for E2 experiments, the joint probabilities AL_1, AL_2 are too small to accurately compute $MutInf_{AL_1,AL_2}(T)$.

Citation: S.Mahmood Ghasemi, Pankaj Singh K, Hannah Johnson L, Ayse Koksoy, Michael Mancini A, Fabio Stossi and Robert Azencott. Analysis and Modeling of Early Estradiol-induced GREB1 Single Allele Gene Transcription at the Population Level. Journal of Bioinformatics and Systems Biology. 7 (2024): 312-33.

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

Table 2: E2 experiments: Parameters estimates for three Population Level Models:

Parameter	E2 exp. 1	E2 exp. 2	E2 exp. 3
MTD = Mean Transcription Duration	44 ± 2 min	44 ± 2 min	43 ± 2 min
L = mean lifetime of nascent mRNA	21 ± 1 min	21 ± 1 min	20.5 ± 1 min
A = mean time between transcriptions (after E2 treatment)	18 ± 1 min	15 ± 1 min	22 ± 1 min
A^+ = mean time between transcriptions (before E2 treatment)	36 ± 2 min	22 ± 1 min	29 ± 1 min
VTH = smallest # of RNA molecules to detect mRNA spots	2 molecules	2 molecules	2 molecules

Table 3: FV+E2 experiments: Parameters estimates for three Population Level Models

Parameter	FV+E2 exp. 4	FV+E2 exp. 5	FV+E2 exp. 6
MTD = Mean Transcription Duration	44 ±1 min	44 ± 1 min	44 ± 2 min
L = mean lifetime of nascent mRNA	21 ± 1 min	20.5 ± 1 min	20.5 ± 1 min
A = mean time between transcriptions (after E2 treatment)	14.5 ± 1 min	16 ± 1 min	23.5 ± 1 min
VTH = smallest # mRNA molecules to detect mRNA spots	2 molecules	2 molecules	2 molecules

E2 experiments as soon as $T \geq 15$min, and for all our FV+E2 experiments when $T \geq 30$ min. In this time range this proves *signi icant statistical dependency* between the activations of any alleles pair AL_1, AL_2. In fact, the dependency ratio $dep(T)$ remains larger than 1.15 at all analyzed time points. Hence, when AL_1 is active at time T, the conditional probability that AL_2 is also active is at least 15% higher than the unconditioned activation probability for AL_2. Note that for initial times $T = 0$ or $T = 15$min, the probabilities $P_{AL_1\,AL_2}(T)$ and $P_{AL_1}(T) \times P_{AL_2}(T)$ are typically too small for reliable estimation of the ratio $dep(T)$.

The probabilistic dependency between the activation states of two alleles AL_1 , AL_2can also be quantified by their *Mutual Information* $MutInf_{AL_1,AL_2}(T)$ (see formulas in Methods section MM6). Recall that $MutInf_{AL_1,AL_2}(T) \geq 0$ evaluates how knowing that AL_1 is active at time T improves the accuracy of predicting whether AL_2 is active at time T. Complete independence of AL_1 and AL_2 would imply $MutInf_{AL_1,AL_2}(T) = 0$, so strictly positive values of $MutInf_{AL_1,AL_2}(T)$ indicate dependency between the activation of AL_1 and AL_2. Since the population average probability F_T of jointly activated alleles has full symmetry, all allele pairs AL_i, AL_j must have mutual information identical to $MutInf_{AL_1,AL_2}(T)$. We have computed $MutInf_{AL_1,AL_2}(T)$ for all experiments, and all T. As detailed in Methods section MM6, the generic formula giving $MutInf_{AL_1,AL_2}(T)$ involves terms such as $P_{AL_1}(T)\, log(P_{AL_1}(T))$ and $P_{AL_1\,AL_2}(T) log(P_{AL_1\,AL_2}(T))$ for which the estimation errors become high when the activation probabilities $P_{AL_1}(T)$ and $P_{AL_1\,AL_2}(T)$ are very small. For (FV+E2) experiments, and for $T \leq 30$ min, both $P_{AL_1}(T)$ and $P_{AL_1\,AL_2}(T)$ are very close to 0, so that the natural estimates of $MutInf_{AL_1,AL_2}(T)$ become statistically reliable only for $T \geq 45$min. For E2 experiments, since GREB1 transcription activity starts before T=0, one can reliably estimate $MutInf_{AL_1,AL_2}(T)$ as soon as $T \geq 15$ min. In **Figure 5**, we display the time course of mutual information $MutInf_{AL_1,AL_2}(T)$ for each one of our experiments.

For FV+E2 experiments as well as for E2 experiments, and for $45min \leq T \leq 90min$, the values of $MutInf_{AL_1,AL_2}(T)$ roughly range from 0.04 to 0.09. For our ranges of mutual information values $m = MutInf_{AL_1,AL_2}(T)$, the dependency ratio $dep(T)$ can be roughly approximated by $(1 + sqrt[m/(1 - p)])$ where $p = P_{AL_1}(T)$. This mathematical approximation valid for small "m" explains why small mutual information values reflect much more sizeable positive values for the difference $[dep(T) - 1]$. The estimation errors indicate that for $45min \leq T \leq 90min$ and for all our experiments, the mutual information $MutInf_{AL_1,AL_2}(T)$ is *significantly positive* at the 95% confidence level. This confirms that, at the population level, we detect significant statistical dependency between jointly activated allele pairs within the same nucleus.

Population level stochastic model to emulate time course of GREB1 allele activation frequencies

We next sought to fit a *population level* stochastic model dedicated to emulating the dynamics of GREB1 transcriptional frequencies computed across large cell populations. Several papers (see 2,18,20) have modeled the dynamics of random gene transcription bursts observed in live *single cells* by stochastic "two-states" promoter models, in which gene promoters are viewed as stochastic automatons randomly cycling through an ON-state and an OFF-state. In these studies, the parameters of two-states promoter models are separately fitted to each single cell continuously observed at very short time intervals. As explicitly pointed out by (2), the estimated parameters of these single cell models vary quite strongly (up to 20%) from cell to cell, due to heterogeneities in cells biology and/or in their local chemical environment. In our smFISH experiments, the frequency $Q_k(T)$ of nuclei exhibiting "k" GREB1 activated alleles at time T is estimated by averaging across several hundred cells of $pop(T)$. Since random gene transcriptions bursts are *highly decorrelated* from cell to cell and have short duration, averaging joint activations frequencies across $pop(T)$ essentially smooths out

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

the random GREB1 transcription bursts occurring in single cells. We have verified this intuitive point by simultaneous simulation of N=400 "two-states" promoter models for GREB1 transcription, followed by averaging at each time T the GREB1 transcription bursts occurring at time T among these N simulations. In our experiments, which involve large cell populations, the observed frequencies $Q_k(T)$ indeed have rather smooth time evolutions, as well as the probabilities $P_{AL_1}(T)$ and $P_{AL_1\,AL_2}(T)$ derived from the frequencies $Q_0(T)$, $Q_1(T)$, $Q_2(T)$, $Q_3(T)$, $Q_4(T)$

To emulate the time course of GREB1 transcription frequencies observed across each cell population pop(T), we introduce a *population level* stochastic model, where successive ER DNA binding occurs randomly after exponentially distributed waiting times that can be followed by coregulator recruitment and transcription initiation. At these random transcription initiation times, GREB1 mRNA elongation proceeds with fixed *Mean Transcription Duration* (*MTD*). Several studies (21,23,24) indicate that gene transcription occurs at a roughly constant speed of 2 to 2.5 kb/min, which results in an MTD ≈ 44min for GREB1. Random time durations between successive rounds of GREB1 mRNA elongation are assumed to be independent of each other, and to have the same exponential density with *mean value A*, which is a model parameter. Such random time gaps are characteristic of Poisson stochastic processes.

We denote *"nas"* any complete nascent GREB1 mRNA, and "exonas" / "intnas" the *exonic* and *intronic* parts of *nas*. The (random) lifetimes of *exonas* and *intnas* are assumed to have exponential decay. The mean half-life of *exonas* has been empirically calculated via actinomycin D pulse-chase experiments and is approximately 3 hours. As our experiments last 90 minutes, the *exonas* decay does not significantly affect *nas* visibility during this time. However, the intronic component *intnas* was calculated to have a mean lifetime <35 min, which does directly affect the lifetimes of completed nascent mRNAs. The random lifetime of any complete nascent GREB1 mRNA (from completion to nearly full decay) is assumed to have an exponential density with unknown *mean value L.*

Our population level model is thus determined by 3 unknown parameters $\{A^+, A, L, MTD\}$. Since analysis of our smFISH images suggest that the smallest nascent mRNA spots may not be reliably detected, we introduce another unknown parameter, the *Visibility Threshold VTH* such that nascent mRNA spots are detectable on our images only if they contain at least *VTH* molecules. For any plausible values of $\{A^+, A, L, MTD, VTH\}$, this model enables rapid simulations generating frequency $F_{AL_1}(T)$ of activation at time T for a single allele. Quality of fit is evaluated by the differences $\left|F_{AL_1}(T) - P_{AL_1}(T)\right|$ over a range of time points T, where the probability $P_{AL1}(T)$ is derived as above from the frequencies $Q_0(T)$, $Q_1(T)$, $Q_2(T)$, $Q_3(T)$, $Q_4(T)$ computed via analysis of smFISH images.

We point out that our population level stochastic model does not attempt to model GREB1 transcriptions in *single cells*. Indeed, our model aims only to emulate the allele activation frequencies resulting from the *aggregation (at cell population level)* of the random allele activations generated by several hundreds of independent two-states stochastic models of GREB1 transcription activities, with two-states model parameters varying slightly from cell to cell.

Estimation of parameters for our population level model by intensive simulations

To fit the population level model to GREB1 transcription data provided by each FV+E2 experiment, we had to estimate the four parameters $\{A^+, A, L, MTD, VTH\}$ which were expected to belong to naturally pre-defined ranges (see Methods, section MM8). But for E2 experiments with no flavopiridol pretreatment, spontaneous GREB1 transcription events at low rates can start long before E2 treatment at T = 0. Some of the alleles activated within the last hour before $T = 0$ will generate incomplete nascent mRNAs that will not be detectable at $T = 0$ and will only be detected by image analysis after T = 15 min or T = 30 min. Taking into account these nascent mRNAs whose transcription started before T=0 complicates the data analysis for E2 experiments with no FV pretreatment, and requires introducing a new parameter, namely the mean value A^+ of waiting times between successive GREB1 transcription rounds before E2 treatment. As shown in [1,2], E2 treatment increases the frequency of genes transcriptions, so we should assume that $A^+ > A$. Thus, fitting population level data for E2 experiments with no FV pretreatment requires the estimation of five parameters $\{A^+, A, L, MTD, VTH\}$.

For the unknown values of the parameters $\{A^+, A, L, MTD\}$ natural ranges were identified from existing literature (see Methods, section MM8, MM9). For the small integer *VTH*, the potential range of values was evaluated by analyzing the rough number of molecules within mRNA smFISH spots detected in the images.

To actually fit the parameters of our population level models to each GREB1 experiment, we performed intensive simulations of this stochastic model to systematically explore the full discretized ranges of the 5 parameters $\{A^+, A, L, MTD, VTH\}$. For each combination of parameters values, and each time point T, these simulations yielded estimates for the activation frequency $F_{AL_1}(T)$ of single alleles across a large population of virtual cells. We only retained the model parameters with good fit to data, *i.e.*, ensuring that $\left|F_{AL_1}(T) - P_{AL_1}(T)\right| \leq 3\%$ where the probability $P_{AL_1}(T)$ of single allele activation was derived as above from image analysis. We selected final parameters values by enforcing parameter stability across all 6 experiments for *L,MTD,VTH*.

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

Tables 2 and 3 display the best fit parameter values for three experiments with no FV pre-treatment, and for three FV pre-treated experiments. Our best model parameters achieved a quality of fit ≈ 3% for each experiment, in good compatibility with the margins of error on the $P_{AL_1}(T)$ derived from image data.

These two tables exhibit good stability across all 6 experiments for the mean transcription duration $MTD \approx 44$ min, the mean lifetime $L \approx 21$ min of nascent mRNAs, and the minimum number $VTH = 2$ molecules of mRNA necessary to detect a nascent mRNA spot. The mean waiting time A between successive GREB1 transcription cycles after E2 treatment had a wider range between 15 min and 23 min among all 6 experiments. We have not identified the main factors influencing the waiting times A, but cell population heterogeneity is likely to strongly impact the variations in A observed from one experiment to the next. Indeed in (2), the authors mentioned that for their two-states model focused on GREB1 transcription observed on separate single cells, the estimated model parameters (such as our parameter A) did strongly vary from cell to cell, with relative variations up to 20%. It is also possible that A may not remain strictly constant in time during E2 induction.

Discussion

Stimulus-controlled gene transcription is one of the essential ways a cell senses and responds to environmental changes. While this process has been heavily studied in multiple models, a full understanding of how the regulation of events leading to gene transcription unfolds is constantly evolving. From numerous studies across species and models (1-8, 11, 14, 16, 18,19), it appears that cells respond to stimuli in a very heterogeneous manner and, even within the same nucleus, different copies of the same target gene respond asynchronously. Is this because of fully stochastic biological reactions or given the evolutionary development of regulated mechanistic steps in gene expression, is regulated allelic activation a way to finely-tune individual cell responses to external causes. In our earlier study (8), we suggested cells can use an epigenetic mechanism to control the frequency of active alleles in the nucleus. Here, we focused on the same biological system, the hormone (E2) stimulated GREB1 gene (2,8,9) in MCF-7 breast cancer cells to ask a few additional basic questions: 1) can we synchronize the response of individual alleles by altering transcription elongation? 2) can we determine if alleles in the same nucleus are acting independently or not? 3) can we develop a simplified model to emulate, at the cell population level, the first phases of hormonal response over time, with stability of the model parameters across independent biological replicates?

To address these questions, we compared GREB1 transcription activity in large cell populations under two types of initial conditions: 1) FV+E2 experiments where prior to E2 treatment of our cell populations at $T = 0$, transcription elongation was synchronized and then restarted by addition and wash-out of the reversible CDK9 inhibitor, flavopiridol (FV). 2) E2 experiments where at $T = 0$, cell populations are still in their natural random state after several hours without hormone treatment and with their transcription cycles left untouched. Our experimental data strongly indicate that "synchronizing" RNA Polymerase II at the elongation step is not sufficient to synchronize hormonal responses at the cell-by-cell or allele-by-allele levels. Indeed, for the two types of initial cell population conditions, at the end time point (90 minutes post E2), identical values are reached by key characteristics such as the activation probability of each single allele and the joint activation frequencies for pairs or triplets of alleles.

At each time T, our detailed analysis of observed frequencies for alleles jointly activated by GREB1 nascent mRNA spots demonstrated a *significant statistical dependency* between pairs of activated alleles within the same nuclei. This led us to apply, at each time T, a principle of *maximum entropy under constraints* to compute a probability distribution F_T for the joint activation states of the four alleles within typical nuclei. We then used the joint probability F_T to compute the mutual information $MutInf_{AL_1, AL_2}(T)$ between GREB1 activations of alleles AL_1 and AL_2 in order to quantify the dependency between pairs of alleles. A detailed error analysis for the estimated $MutInf_{AL_1, AL_2}(T)$ showed that this mutual information had statistically significant positivity for all $T \geq 30$ min, a clear indicator of moderate but significant dependency between activations for pairs of alleles. An interesting still open question is to identify biochemical factors enabling these dependencies, such as extrinsic chemical factors that can jointly affect all 4 alleles in each nucleus. Other cell-linked factors affecting GREB1 transcription of all 4 alleles were also invoked in (2) to explain the high variation of transcription model parameters fitted separately to single cells.

The probability distributions F_T were computed at each fixed time T from population level frequencies of joint alleles activations. To emulate the time dynamics of these probabilities F_T across time, we introduced a "population level" stochastic model, where random initializations of GREB1 transcriptions are driven by a Poisson process, and are always followed by actual elongation. We were led to introduce this *population level* model instead of the popular two-states models used for *single cell* transcription data (1,2) because averaging GREB1 transcription activity across large cell populations strongly smooths out the random transcription bursts occurring independently among individual cells. Since our smFISH image acquisition modalities do not enable the monitoring across time of single cells GREB1 transcription activity, we designed our population level model to roughly emulate the superposition of several hundreds of independent

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

two-states models of single cells gene transcription dynamics. For each one of our experiments (three FV+E2 experiments and three E2 only experiments) the parameters of our population level model were fitted to experimental data by intensive simulations exploring a very large set of combined parameter values. After this fitting of model parameters to data, the quality of fit was quite precise (less than 3% error on emulated frequencies of GREB1 activations), and across all 6 experiments we achieved good stability for the estimates of the three main parameters, namely the mean elongation duration $MTD \approx 44$ min, the mean lifetime L ≈ 21min of nascent mRNAs, and the number VTH = 2 of mRNA molecules necessary for reliable detection of a nascent mRNA spot. The estimated mean waiting time A between successive GREB1 transcription after E2 treatment had a wider range (15min to 23min) among our 6 experiments. This variation is quite compatible with the 20% variations range reported in (2) for the parameters of two-states models fitted separately to single cells.

We expect our innovative modeling approach for hormone-regulated target gene activity observed at population level to be applicable for many other genes and stimuli, a point we intend to validate through further experiments. An interesting and open challenge is to concretely identify the main cell-dependent factors simultaneously impacting transcriptional responses at individual alleles within the cell nucleus.

Materials and Methods

1. Cell culture, materials and treatments: MCF-7 cells were obtained from BCM Cell Culture Core, which routinely validates their identity by genotyping; cultures are constantly tested for mycoplasma contamination as determined by DAPI staining. MCF-7 were maintained in MEM plus 10%FBS media, as recommended by ATCC, except phenol red free and kept in culture for less than 60 days before thawing a fresh vial. Three days prior to experiments, cells were plated on round poly-L-lysine coated coverslips in media containing 5% charcoal-dextran stripped and dialyzed FBS-containing media. Treatments with 17β-estradiol (E2, Sigma) were performed as in [8]. For flavopiridol (FV+E2) experiments, cells were treated with FV 1μM for 2 hours, then removed and cells were washed 3x with media prior to E2 1nM treatment for the indicated times.

2. Single molecule RNA FISH (smFISH): GREB1 smFISH was performed as described in detailed protocols [8, 22]. Briefly, cells were fixed with 4% paraformaldehyde in PBS, on ice for 20 min. After a PBS wash, cells were left in 70%ethanol for a minimum of 4 hours prior to hybridization (o/n, 37C) with the previously validated GREB1 probe sets (LGC Biosearch Technologies) covering introns (Atto647N) and exons (Quasar570) of the GREB1 gene.

3. Imaging:
High resolution imaging for smFISH was performed on a Cytivia DVLive epifluorescence image restoration microscope with an Olympus PlanApo 60×/1.42NA objective and a 1.9k × 1.9k sCMOS camera. Z stacks (0.25 μm) covering the whole nucleus (~10 μm) were acquired before applying a conservative restorative algorithm for quantitative image deconvolution. Ten or more random fields of view (FOVs) were acquired for each time point.

4. Image analysis:

Each FOV has three fluorescence channels (DAPI, Q570 (exons) and A647N (introns)) in a 3D-image of size ≃ 1780 x 1780 x 25. Each 3D image channel was projected on its maximum intensity horizontal layer and then analyzed as a 2D image. In the DAPI channel, we first detect and identify cell nuclei. The main steps are: contrast thresholding, connected components detection, elimination of holes, and size filtering. After dilating the detected nuclear mask, we estimate cytoplasm boundaries by the "watershed" segmentation algorithm.

Classical image segmentation techniques are applied in the two other channels to separately detect exonic and intronic spots. Contrast analysis is implemented by OTSU thresholding [27] for the exonic channel (Q570), and by max-entropy thresholding [28] for the intronic channel (A647N). We tested moderate variations of these thresholds to define accuracy margins for these two key spots detections. Nascent mRNAs spots are detected within each nucleus by identifying all pairs (exonic spot + intronic spot) having non-empty intersection. At each time point, the number of active alleles detected per nucleus ranges from 0 to 4 since each cell has 4 GREB1 alleles [29]. Segmentation errors due to local cell packing or nuclei overlaps may occasionally generate spurious detections of more than 4 activation spots in a very small percentage of nuclei, which are then automatically discarded. Mature RNAs are identified as exonic spots (Q570) located within the cytoplasmic mask.

5. Probability distribution of joint alleles activations at ixed time T:

A key modeling step was, at each fixed time T, to estimate the probabilities of joint alleles activation for pairs, triplets, quadruplets of alleles within a cell. We could only aim to estimate averages over $pop(T)$ for each one of these joint probabilities, since in our experimental setup, distinct alleles are not identifiable. At time T, in any given cell, each allele AL_j (j = 1, 2, 3, 4) can either be active (ON = state "1"), or not (OFF = state "0"). The joint state of the four alleles AL_1, AL_2, AL_3, AL_4 is then described by a four-digit binary code, with $2^4 = 16$ possible joint states labeled S_0 = 0000, S_1 = 0001, S_2 = 0010, S_3 = 0011, S_4 = 0100, …, S_{14} = 1110, S_{15} = 1111

At time T, the cell population $pop(T)$ contains $N = N(T)$ nuclei, denoted NUC_1,…, NUC_N. For each nucleus NUC_N, the current joint state for the four alleles will be equal to S_k, with some unknown probability $prob_n(S_k)$. The 16 probabilities $prob_n(S_0)$, $prob_n(S_1)$,…, $prob_n(S_{15})$ add up to 1, and depend

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

on the time point T. But due to biochemical heterogeneity of the cells in the population at time T, the $prob_n(S_k)$ may also depend on unknown biochemical factors specific to nucleus *NUCn*. Since the observed frequencies $Q_k(T)$ of nuclei exhibiting k activated alleles at time T are computed across the whole of *pop*(*T*), they only provide reliable information on the average $F_T(S_k)$ of the probabilities $prob_n(S_k)$ over all nuclei indices $n = 1 \ldots N$. More precisely, for each possible joint state S_k of the four alleles, we want to estimate the average probability

$$F_T(S_k) = \left(\frac{1}{N}\right)[prob_1(S_k) + ... + prob_N(S_k)]$$

In a perfectly homogeneous cell population *pop*(*T*), the probabilities $prob_n(S_k)$ would not depend on the nucleus *NUCn* at all and would hence also be equal to the average probability $F_T(S_k)$. This ideal situation is blurred by the significant biological diversity of GREB1 transcriptional response from cell to cell, a point well documented in (2).

The frequencies $F_T(S_k)$ are not directly observable, since smFISH images do not enable specific matching of alleles AL_1, AL_2, AL_3, AL_4 from cell to cell. But as mathematically detailed in Methods and in Suppl. Materials 1 the observed activation frequencies $Q_k(T)$ are linked to the unknown frequencies:

$F_T(S_0) = F_T(0000)$, $F_T(S_1) = F_T(0001)$, …, $F_T(S_{15}) = F_T(1111)$ by the following five linear relations:

Equation 1

$$Q_0(T) = F_T(0000)$$

$$Q_1(T) = F_T(1000) + F_T(0100) + F_T(0010) + F_T(0001)$$

$$Q_2(T) = F_T(1100) + F_T(1010) + F_T(1001) + F_T(0110) + F_T(0101) + F_T(0011)$$

$$Q_3(T) = F_T(1110) + F_T(0111) + F_T(1011) + F_T(1101)$$

$$Q_4(T) = F_T(1111)$$

Since these five linear relations cannot determine the 16 unknown probabilities $F_T(S_k)$, we first checked if we could assume independence of alleles activation. Under the probability F_T of joint alleles activations, denote $f_j(T)$ the probability that the specific allele AL_j is activated. Assume temporarily that under F_T the activations of each allele AL_1, AL_2, AL_3, AL_4 are statistically independent of each other (i.e. there are no mechanisms through which alleles interfere / influence each other's activations). Independence implies that each unknown joint probability $F_T(S_k)$ can be expressed directly in terms of $f_1(T), f_2(T), f_3(T), f_4(T)$ by simple product formulas such as

$F_T(0100) = (1 - f_1(T)) f_2(T) (1 - f_3(T)) f_4(T)$, $F_T(1110) = f_1(T) f_2(T) f_3(T) (1 - f_4(T))$, …, etc.

Combining these product formulas with equation 1 we have formally proved (see Suppl. Materials 2) that statistical independence of single allele activations under the joint probability F_T forces the following polynomial equation of degree 4

$$Q_0(T)\, z^4 - Q_1(T)\, z^3 + Q_2(T)\, z^2 - Q_3(T)\, z + Q_4(T) = 0$$

Equation 2

to have *four positive and real valued solutions* z_1, z_2, z_3, z_4, We also showed that $f_1(T), f_2(T), f_3(T), f_4(T)$ are then given by

$f_j(T) = z_j / (1 + z_j)$ for $j = 1,2,3,4$.

Requiring a polynomial of degree 4 to have four positive and real valued roots z_1, z_2, z_3, z_4 imposes very restrictive polynomial constraints on the polynomial coefficients $Q_0(T)$, $Q_1(T)$, $Q_2(T)$, $Q_3(T)$, $Q_4(T)$.

In Suppl. Materials 2, we give examples of these polynomial constraints. Our experiments showed very clearly that these constraints are never satisfied by the observed $Q_0(T)$, $Q_1(T)$, $Q_2(T)$, $Q_3(T)$, $Q_4(T)$, and hence that, at the level of cell populations averages, one had to *reject the hypothesis of statistical independence* between activations of distinct alleles.

6. Maximum entropy model and dependency between alleles activations:

Because full independence of the four alleles is not compatible with our experimental data, we computed, for each T, the joint probability F_T which minimizes dependency between alleles activation, and is still compatible with the observed $Q_0(T)$, $Q_1(T)$, $Q_2(T)$, $Q_3(T)$, $Q_4(T)$ via the linear relations imposed by Equation 1. For probability distributions verifying a set of linear constraints, a fairly generic principle is that minimizing dependencies is roughly equivalent to maximizing entropy. Recall that for any probability $F = [F_0, \ldots, F_{15}]$ on the set $S = [S_0, \ldots, S_{15}]$ of joint alleles activation, the entropy $Ent(F) \geq 0$ of F is given by Ent (F) = $- F_0 \log(F_0) - F_1 \log(F_1) - \ldots - F_{15} \log(F_{15})$.

In Suppl. Materials 3, we apply this principle to compute the unique probability of joint alleles activation frequencies, denoted $F_T = [F_T(S_0), \ldots, F_T(S_{15})]$ which has maximum entropy among all probabilities compatible with the observed frequencies $Q_0(T)$, $Q_1(T)$, $Q_2(T)$, $Q_3(T)$, $Q_4(T)$ via the five linear relations of Equation 1. Our formulas show that the maxiTmum entropy joint probability F_T must have full symmetry, meaning that permutations of the alleles AL_1, AL_2, AL_3, AL_4 do not change the frequencies of their joint activation states. Indeed, we have the explicit formulas

$$F_T(0000) = Q_0(T)$$

$$F_T(1000) = F_T(0100) = F_T(0010) = F_T(0001) = Q_1(T)/4$$

$$F_T(1100) = F_T(1010) = F_T(1001) = F_T(0110) = F_T(0101) = F_T(0011) = Q_2(T)/6$$

$$F_T(1110) = F_T(0111) = F_T(1011) = F_T(1101) = Q_3(T)/4$$

$$F_T(1111) = Q_4(T)$$

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

Since in our smFISH experimental images it is not possible to match specific alleles from cell to cell, full symmetry for the average probability F_T of joint alleles activations is a natural feature for compatibility with our image data.

Fix any single allele AL_1. Due to the full symmetry of F_T, the frequency

$$P_{AL_1}(T) = F_T\{allele\ AL_1\ is\ active\ at\ time\ T\}$$

takes the same value for all four alleles, namely (see Suppl. Materials 3),

$$P_{AL_1}(T) = Q_1(T)/4 + Q_2(T)/2 + 3Q_3(T)/4 + Q_4(T)$$

Equation 3

Consider now two alleles AL_1 and AL_2 in the same nucleus. At time T, the (random) state of AL_1 is either "active" or "inactive", which we denote by $AL_1 = 1$ or $AL_1 = 0$. Same remark for the random state of AL_2. The mutual information $MutInf_{AL_1,AL_2}(T)$ between the binary valued random states of AL_1 and AL_2 is given by the generic formula $MutInf_{AL_1,AL_2}(T) = MutInf_{AL_2,AL_1}(T) = \mathrm{Ent}(AL_1) + \mathrm{Ent}(AL_2) - \mathrm{Ent}(AL_1,AL_2) \geq 0$

where *"int"* denotes the entropy of a probability distribution. Higher values of $MutInf_{AL_1,AL_2}(T)$ indicate higher dependency between the random activation states of AL_1, AL_2 under the probability F_T. The maximum possible value of $MutInf_{AL_1,AL_2}(T)$ is 0.69, which can only be reached if there is a fully deterministic relation between the random activations of alleles AL_1 and AL_2. But $MutInf_{AL_1,AL_2}(T)$ values higher than 0.02 already indicate some significant level of dependency between AL_1 and AL_2. Conversely, values of $MutInf_{AL_1,AL_2}(T)$ extremely close to 0 reflect near independence between the activation states of AL_1 and AL_2. At time T, the pair (AL_1, AL_2) has 4 possible *joint states* (00), (01)(10), (11). Their frequencies $f_T00, f_T01, f_T10, f_T11$, are easily derived from the explicit expression (see equation 3) of the probability F_T, which yields

$$f_T11 = \frac{Q_2(T)}{6} + \frac{Q_3(T)}{2} + Q_4(T)$$

$$f_T00 = \frac{Q_1(T)}{2} + \frac{Q_2(T)}{2} + Q_0(T)$$

$$f_T01 = f_T10 = \frac{Q_1(T)}{4} + \frac{Q_2(T)}{3} + \frac{Q_3(T)}{4}$$

Equation 4

The probability $P_{AL_1,AL_2}(T) = f_T11$ that AL_1 and AL_2 are simultaneously active at time T is hence given by

$$P_{AL_1},P_{AL_2}(T) = \frac{Q_2(T)}{6} + \frac{Q_3(T)}{2} + Q_4(T)$$

Equation 5

Then the joint entropy $Ent(AL_1, AL_2)$ is given by the following

Equation 6

$$Ent(AL_1, AL_2) = -f_T00\log(f_T00) - f_T01\log(f_T01) - f_T10\log(f_T10) - f_T11\log(f_T11)$$

By definition of entropy, one has also

Equation 7

$$Ent(AL_1) = Ent(AL_2) = -P_{AL_1}(T)\log(P_{AL_1}(T)) - (1 - P_{AL_1}(T))\log(1 - P_{AL_1}(T))$$

The *mutual information* $MutInf_{AL_1,AL_2}(T)$ between the random activations of AL_1 and AL_2 can then be computed by

$$MutInf_{AL_1,AL_2}(T) = 2\,\mathrm{Ent}(AL_1) - \mathrm{Ent}(AL_1,AL_2)$$

Equation 8

The equations 4,5,6,7,8 clearly express $MutInf_{AL_1,AL_2}(T)$ in terms of the 5 observed activations frequencies $Q_0(0)$, $Q_1(0)$, $Q_2(0)$, $Q_3(0)$, $Q_4(0)$. Due to the full symmetry of joint probability F_T, the mutual information $MutInf_{AL_1,AL_2}(T)$ will be the same for all pairs of alleles (AL_i, AL_j) and this common value *quantifies the average amount of activation dependency* between pairs of alleles at time T.

7. Modeling the time course of GREB1 transcription frequencies observed at population level:

Since genes transcriptions are strongly decorrelated from cell to cell, the random transcription bursts occurring among (for instance) the 400 cells of population $pop(T)$ will be highly de-synchronized. Hence, averaging the random bursts that actual occur at fixed time T will clearly *smooth out* the impact of random bursts on the allele activation frequencies $Q_k(T)$, observed across $pop(T)$. We have validated this point by simulations of 400 two-states stochastic models of GREB1 transcription in single cells, and averaging at each time T the nascent mRNA outputs of these independent 400 models. As expected, in population averaged transcription activity, short transcription bursts were essentially no longer identifiable. So, to emulate the time course of our population averaged GREB1 transcription data, we have introduced a population level stochastic model.

In this simplified model successive GREB1 transcriptions are initialized at random times $t_1 < t_2 < \ldots < t_n$. Each initialized transcription launches the elongation of a GREB1 mRNA molecule at fixed linear speed and is completed after a fixed *Mean Transcription Duration* MTD which for GREB1 is ≈ 44 min. The time intervals ($t_{k+1} - t_k$ are random and assumed to be independent and to have the same *exponential density* with unknown *mean value A*. The successive occurrences t_k of transcription initializations define then a Poisson stochastic process. The complete nascent mRNA *"nas_k"* generated by GREB1 transcription started at time t_k will then become fully visible at time (t_k + *MTD*). Both exonic and intronic parts of nas_k, denoted $exonas_k$ and $intnas_k$, are naturally assumed to have exponential decay. The mean half-life of $exonas_k$ is about 3 hrs, and hence does not impact the visibility of nas_k during the time-course 90min of E2 treatment. However, the mean lifetime of intronic $intnas_k$ is known to be shorter than ≈ 35 min, and hence directly affects the first visibility time of nas_k. The random lifetime of $intnas_k$ from completion of nas_k to nearly full decay of $intnas_k$ is assumed to have an *exponential density* with unknown fixed *mean* {Our population level stochastic model thus has only 3 key parameters {*A, L,*

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

MTD}. But to take into account the limits imposed by image resolution, we introduce another integer valued parameter, the unknown *Visibility Threshold VTH* such that nascent mRNA spots are detectable only if they contain at least *VTH* molecules (after simulations outlined below, we obtained the estimate *VTH* = 2).

Our *population level model* is easy to simulate, and we have fitted its 4 parameters {A^+, *A, L, MTD, VTH*} to each FV+E2 experiment by intensive simulations as outlined below. For E2 experiments with no FV pre-treatments, we must take account of actual GREB1 transcription activity occurring at low frequency in our cell populations during the last hour before the time T= 0 of E2 treatment. This requires the introduction of another parameter, namely the mean waiting time A^+ between successive transcription initiations occurring *before* time $T = 0$.

8. Simulations and model fitting:

To fit our stochastic population level model to experimental data we performed intensive simulations to select the parameters {A^+, *A, L, MTD, VTH*} providing the best quality of fit to our smFISH image data. These parameters were constrained to have naturally pre-defined ranges:

- mean transcription duration MTD: $40min < MTD < 50\ min$ since RNA Polymerase II speed ≈ *2.5 kb /min* and GREB1 length ≈110 kb

- mean lifetime L of nascent mRNAs: $5min < L < 35\ min$, based on actinomycin D pulse-chase experiments

- mean waiting time A between rounds of transcription *after* $T = 0$: $5min < A < 35\ min$, based upon the on/off time ranges evaluated in (2)

- mean waiting time $A^+ > A$ between rounds of transcription *before* = 0: $A^+ < 60\ min$, based upon analysis of initial activations frequencies $Q_0(0)$, $Q_1(0)$, $Q_2(0)$, $Q_3(0)$, $Q_4(0)$. Note: A^+ is used only for experiments with no FV pretreatment.

The minimum number VTH of molecules needed for reliable detection of nascent mRNA spots had to be crudely pre-calibrated by image analysis. As detailed in section MM9 below, and similarly to (2), we calculated the integrated intensities of mature, cytoplasmic mRNA spots to roughly evaluate the number of molecules per detected nascent mRNA spot. This yielded a rough preliminary range $1 \leq VTH \leq 10$ molecules per detectable spot.

These parameters ranges were discretized into finite grids with accuracies of 0.5 min to 1 min for all time variables. This gave us a multigrid of roughly 10^6 possible parameter vectors PAR = {A^+, *A, L, MTD, VTH*}. For each potential vector PAR, we performed a first set of 1000 simulations of our population level model. Each such simulation outputs a random number NAS(T) of nascent mRNAs present at time T on a single virtual allele AL_1. Among the 1000 simulated $NAS\,(T)$, we compute the percentage $F_{AL_1}(T)$ of integers $NAS\,(T)$ which are larger than the visualization threshold *VTH*. Then $F_{AL_1}(T)$ is the model generated frequency of detectable single allele activations. For each experiment and each vector *PAR* we then evaluate the quality of fit between model and data by the distance $dist\,(model, data) = max_{\text{over all T}} \left| F_{AL_1}(T) - P_{AL_1}(T) \right|$

For the three FV+E2 experiments, the early values $P_{AL_1}(0)$, $P_{AL_1}(15min)$, $P_{AL_1}(30min)$ were practically 0 up to errors of estimations (0.03), and the simulated $F_{AL_1}(0)$, $F_{AL_1}(15min)$, $F_{AL_1}(30min)$ were identically zero since *MTD* was known to be of the order of 44 min. Therefore, for FV+E2 experiments, the distance between model and data was actually replaced by $dist\,(model, data) = max_{\text{over all T} \geq 45\text{min}} \left| F_{AL_1}(T) - P_{AL_1}(T) \right|$

For each experiment, the best choices for the model parameters vector PAR are obtained by optimizing the quality of fit, i.e. by minimizing the distance *dist* (*model*, *data*).

We implemented the stochastic simulations of our population level model by the following algorithm. We first generate the times t_k by standard simulation of the sequence of independent random waiting times $(t_{k+1} - t_k)$ having the same exponential density with mean A. Elongation of the nascent mRNA nas_k begins at t_k and is completed at time $(t_k + MTD)$.

The random lifetime U_k of nas_k is provided by a separately simulated sequence of independent random lifetimes U_k having the same exponential density with mean L. Then at time T, the number of nas_k present at time T is determined by the number of nas_k such that $t_k + MTD < T < t_k + MTD + U_k$. This simulation algorithm is naturally faster than the Gillespie algorithm used for more complex stochastic models. Our Python simulation code is accessible in the publicly available software package on GitHub (https://github.com/smahmoodghasemi/BCM). This "brute force" approach to model fitting required intensive computing and was implemented on the "Sabine" multicore computing center at University of Houston. Once the simulations have been completed for 10^6 models, this large set of simulations outputs can be re-used as a fixed massive lookup table for all our past or future experiments. After these simulations were successfully completed, we also outlined a more efficient computing approach which could be used in similar explorations for other genes. Namely, one can implement a multi-scale exploration starting with a cruder grid of parameters vectors, and then focus on finer mesh grids tightly centered around promising first level estimates.

9. Estimation of number of molecules within detected nascent mRNA spots:

For each *nascent* mRNA spot *"nas"* detected within the nuclei present in an image J, we compute the integrated exonic intensity *EXO* (*nas*) as the sum of image intensities *EXO*(*x*) over all exonic pixels x of *"nas"*. For each *mature*

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

mRNA spot "*mat*" detected within the cytoplasm of all cells present in image J, we also compute the integrated exonic intensity *EXO*(*mat*) as the sum of image intensities *EXO*(*x*) over all pixels x of *"mat"*. We then compute the median *MED* of these integrated intensities *EXO*(*mat*) over all mature RNA spots detected in image J.

In the spirit of an approach explored in [2], the number *MOL*(*nas*) of RNA molecules present within any *detected* nascent mRNA spot *"nas"* is crudely estimated by the ratio *MOL*(*nas*) = *EXO* (*nas*)/*MED*. This can only be a very rough calibration of *MOL*(*nas*) since the observed *EXO*(*mat*) values have a high dispersion around their median *MED*, even at fixed time T. Nevertheless, to evaluate reasonable ranges for our model parameter *VTH* (Visualization Threshold) on any given image J, we have computed the low quantiles for the histogram of all *MOL*(*nas*) values extracted by computer analysis of image J. These low quantiles identified a potential range of 2 to 5 for the minimum number *VTH* of RNA molecules necessary to detect a nascent mRNA spot in our images. Since the error margins for these estimates were likely to be high, we simply assigned a much wider potential range of 1 to 10 for the unknown parameter *VTH*. After fitting of our population level to experimental data, our results reported above showed that the best estimate of *VTH* was always *VTH* = 2.

10. Estimation errors for frequencies $Q_k(T)$ and probabilities $P_{AL_1}(T)$, $P_{AL_1,AL_2}(T)$:

Tables given above and Table 6 in Supplemental Materials display the computed estimation errors for $P_{AL_1}(T)$, $dep(T)$, $MutInf_{AL_1,AL_2(T)}$, $P_{AL_1\ AL_2}(T)$. Here we outline how these errors are computed. Let $N(T)$ = # cells in population $pop(T)$. Use shorter notations Q_k for $Q_k(T)$, P_{AL_1}, for $P_{AL_1}(T)$, P_{AL_1,AL_2} for $P_{AL_1,AL_2}(T)$.

Let ΔQ_k, ΔP_{AL_1}, $\Delta P_{AL_1,\ AL_2}$ be the corresponding random errors of estimation. The 5x5 covariance matrix cov_Q of the 5 errors $[\Delta Q_0, \Delta Q_1, \ldots, \Delta Q_4]$ is classically given by (i, j)

$$cov_Q(i,j) = cov([\Delta Q_i, \Delta Q_j) = -Q_i Q_j / N(T) \text{ for all } i \neq j$$

$$cov_Q(i,i) = var(\Delta Q_i) = Q_i(1 - Q_i)/N(T)$$

Denote Q the column vector $[Q_0; Q_1; \ldots; Q_4]$ and for any matrix H denote H^{tr} the transpose of H. Due to Equations 3 and 5, the formulas for P_{AL_1} and $P_{AL_1\ AL_2}$ can be rewritten in matrix form as

$$P_{AL_1} = u * Q \text{ and } P_{AL_1\ AL_2} = v * Q \qquad \text{Equation 9}$$

Where u = [0,1/4,1/2,3/4,1] and v = [0,0,1/6,1/2,1]

Known statistical formulas then give the variances of ΔP_{AL_1} and $\Delta P_{AL_1,\ AL_2}$ as $var(\Delta P_{AL_1}) = u * cov_Q * u^{tr}$

and $var(\Delta P_{AL_1,\ AL_2}) = v * cov_Q * v^{tr}$

11. Estimation errors for the dependency ratio $dep(T) = P_{AL_1,AL_2} / P_{AL_1} \times P_{AL_2}$

Due to Equation 9 we have $dep(T) = v * Q/(u * Q)^2$ which is a nonlinear function $K(Q)$ of Q. The variance of the random error $\Delta dep(T)$ in the estimation of $dep(T)$ is then classically given by $var[\Delta dep(T)] = w * covQ * wtr$ where w is the gradient of $K(Q)$ with respect to Q. This gradient is given by $w = (1/u * Q)^2 * v - 2[v * Q/(u * Q)^3] * u$ which completes the computation of $var[\Delta dep(T)]$.

12. Estimation errors for the mutual information $MutInf_{AL_1, AL_2}(T)$

At any given time T, the pair of alleles (AL_1, AL_2) can be in one of their four joint activation states (00), (01), (10), (11) with corresponding joint probabilities given above by Equation 4. These formulas can be rewritten in matrix form as

$$f_T 00 = V00 * Q, f_T 01 = V01 * Q, f_T 10 = V10 * Q, f_T 11 = V11 * Q$$

where the line vectors *Vij* are given by

$V00$ = [1,1/2,1/6,0,0], $V01 = V10$ = [0,1/4,1/3, 1/4, 0], $V11$ = [0,0,1/6,1/2,1],

The entropies $E = Ent(AL_1) = Ent(AL_2)$ and $E_{12} = Ent(AL_1, AL_2)$ are given by

$$E = -P_{AL_1}(T) \log(P_{AL_1}(T)) - (1 - P_{AL_1}(T)) \log(1 - P_{AL_1}(T))$$

$$E_{12} = -f_T 00 \log(f_T 00) - f_T 01 \log(f_T 01) - f_T 10 \log(f_T 10) - f_T 11 \log(f_T 11)$$

which can be rewritten as

$$E = -(u * Q) \log(u * Q) - (1 - u * Q) \log((1 - u * Q)$$

$$E_{12} = -[(V11 * Q) \log(V11 * Q) + (V00 * Q) \log(V00 * Q) + 2(V10 * Q) \log(V10 * Q)]$$

The information $M = MutInf_{AL_1\ AL_2\ (T)}$ is then given by $M = 2E - E_{12}$ which is a non linear function $M = G(Q)$. The random estimation error ΔM on M has variance $var(\Delta M)$ which can be computed as above using the gradient $grad(G)$ of the function $G(Q)$ with respect to Q. We have the classical formula

$$var(\Delta M) = grad(G) * cov_Q * grad(G)^{tr}$$

Since $grad(G) = 2grad(E) - grad(E_{12})$, we compute the gradients of E and E_{12} from the preceding formulas to get

$$grad(E) = [-\log(u * Q) + \log(1 - u * Q)] * u$$

$$grad(E12) = -[1 + log(V11 * Q)] * V11 - [1 + log(V00 * Q)] * V00 - 2[1 + log(V10 * Q)] * V10$$

This clearly completes the computation of $var(\Delta M)$.

Acknowledgments

Imaging for this project was supported by the Integrated

Microscopy Core at Baylor College of Medicine and the Center for Advanced Microscopy and Image Informatics (CAMII) with funding from NIH (DK56338, CA125123, ES030285), and CPRIT (RP150578, RP170719), the Dan L. Duncan Comprehensive Cancer Center, and the John S. Dunn Gulf Coast Consortium for Chemical Genomics. Modeling and computing research for this project at University of Houston were supported by CAMII and Baylor College of Medicine with funding from CPRIT (RP170719). Intensive Computer Simulations for this project were implemented at the SABINE Cluster of RCDC (Research Computing Data Core, University of Houston) under a CPU-GPU computing time allocation to the University of Houston Department of Mathematics.

References

1. Rodriguez J, Ren G, Day CR, et al. Intrinsic Dynamics of a Human Gene Reveal the Basis of Expression Heterogeneity. Cell 176 (2019): 213-226.
2. Fritzsch C, Baumgärtner S, Kuban M, et al. Estrogen-dependent control and cell-to-cell variability of transcriptional bursting. Mol Syst Biol 14 (2018): 7678.
3. Raj A, Peskin CS, Tranchina D, et al. Stochastic mRNA synthesis in mammalian cells. PLoS Biol 4 (2006): 1707–1719.
4. Sepúlveda LA, Xu H, Zhang J, et al. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. Sci 351 (2016): 1218–1222.
5. Bohrer CH, Larson DR. The Stochastic Genome and Its Role in Gene Expression. Cold Spring Harb Perspect Biol 13 (2021): 040386.
6. Bahar Halpern K, Tanami S, Landen S, et al. Bursty gene expression in the intact mammalian liver. Mol Cell 58 (2015): 147–156.
7. Battich N, Stoeger T, Pelkmans L. Control of Transcript Variability in Single Mammalian Cells. Cell 163 (2015): 1596–610.
8. Stossi F, Dandekar RD, Mancini MG, et al. Estrogen-induced transcription at individual alleles is independent of receptor level and active conformation but can be modulated by coactivators activity. Nucleic Acids Res 48 (2020): 1800–1810.
9. Nwachukwu JC, Srinivasan S, Zheng Y, et al. Predictive features of ligand-specific signaling through the estrogen receptor. Mol Syst Biol 12 (2014): 864.
10. Vera M, Tutucci E, Singer RH. Imaging Single mRNA Molecules in Mammalian Cells Using an Optimized MS2-MCP System. Methods Mol Biol 2038 (2019): 3–20.
11. Hocine S, Raymond P, Zenklusen D, et al. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. Nat Methods 10 (2013): 119–121.
12. Schmidt A, Gao G, Little SR, et al. Following the messenger: Recent innovations in live cell single molecule fluorescence imaging. Wiley Interdiscip Rev RNA 11 (2020): 1587.
13. Paulsson J. Prime movers of noisy gene expression. Nat Genet 37 (2005): 925–936.
14. Suter DM, Molina N, Naef F, et al. Origins and consequences of transcriptional discontinuity. Curr Opin Cell Biol 23 (2011): 657–662.
15. Harper C v, Finkenstädt B, Woodcock DJ, et al. Dynamic analysis of stochastic transcription cycles. PLoS Biol (2011): 9(4).
16. Dar RD, Razooky BS, Singh A, et al. Transcriptional burst frequency and burst size are equally modulated across the human genome. Proc Natl Acad Sci U S A 109 (2012): 17454–17459.
17. Larson DR, Fritzsch C, Sun L, et al. Direct observation of frequency modulated transcription in single cells using light activation. Elife 2 (2013): 00750.
18. Zoller B, Nicolas D, Molina N, et al. Structure of silent transcription intervals and noise characteristics of mammalian genes. Mol Syst Biol 11(2015): 823.
19. Chubb JR, Trcek T, Shenoy SM, et al. Transcriptional pulsing of a developmental gene. Curr Biol 16 (2006): 1018–1025.
20. Ball AD, Adames NR, Reischmann N, et al. Measurement and modeling of transcriptional noise in the cell cycle regulatory network. Cell Cycle 12 (2019): 1–18.
21. Jonkers I, Kwak H, Lis JT. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife 3 (2014): 024027.
22. Mistry RM, Singh PK, Mancini MG, et al. Single Cell Analysis Of Transcriptionally Active Alleles By Single Molecule FISH. J Vis Exp 163 (2020): 1–15.
23. Ardehali MB, Lis JT. Tracking rates of transcription and splicing in vivo. Nat Struct Mol Biol 16 (2009): 1123–1134.
24. Danko CG, Hah N, Luo X, et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Mol Cell 50 (2013): 212–222.
25. Guyon X. Random fields on a network: modeling, statistics, and applications. Springer (1995).
26. Chalmond B. Modeling and Inverse Problems in Imaging Analysis. Springer New York 155 (2003).

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

27. Otsu N. A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics 9(1979): 62–66.

28. Wong AKC, Sahoo PK. A gray-level threshold selection method based on maximum entropy principle. IEEE Transactions on Systems, Man, and Cybernetics 19 (1989): 866–871.

29. Kocanova S, Kerr EA, Rafique S, et al. Activation of estrogen- responsive genes does not require their nuclear co-localization. PLoS Genet 6 (2010): 1000922.

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

Supplementary Materials: Algorithmics and Tables:

Supplementary Material 0: ***Tables for observed frequencies*** $\boldsymbol{Q_0(T)\ Q_1(T)\ Q_2(T)\ Q_3(T)\ Q_4(T)}$

For each one of our 6 experiments, image analysis at time T yields five key frequencies $Q_j(T)$. In cell population $pop(T)$, $Q_j(T)$ is the frequency of nuclei exhibiting exactly "j" detected GREB1 nascent mRNA spots.

For two E2 experiments and two FV+E2 experiments, we display here the observed values of the $Q_j(T)$, listed as *percentages*. Similar results hold for our two other experiments.

Since $pop(T)$ has size $N(T) \geq 400$ cells, the estimation error on each $Q_j(T)$ is less than $\frac{1}{2\sqrt{400}} = 2.5\%$.

E2 experiment #1

T in minutes	$T = 0$	$T = 15$	$T = 30$	$T = 45$	$T = 60$	$T = 75$	$T = 90$
$Q_0(T)$	77	67.3	53.5	39.6	30	26	22
$Q_1(T)$	15	17.8	19.8	19.8	18	17	16
$Q_2(T)$	5	9.9	15.8	19.8	21	19	18
$Q_3(T)$	2	4	7.9	13.9	20	22	24
$Q_4(T)$	1	1	3	6.9	11	16	20

E2 experiment #2

T in minutes	$T = 0$	$T = 15$	$T = 30$	$T = 45$	$T = 60$	$T = 75$	$T = 90$
$Q_0(T)$	45.5	37.6	31	26	19.8	14	11.9
$Q_1(T)$	25.7	23.8	20	15	13.9	13	9.9
$Q_2(T)$	14.9	18.8	20	19	19.8	20	18.8
$Q_3(T)$	9.9	13.9	18	21	23.8	29	30.7
$Q_4(T)$	4	5.9	11	19	22.8	24	28.7

FV+E2 experiment #4

T in minutes	$T = 0$	$T = 15$	$T = 30$	$T = 45$	$T = 60$	$T = 75$	$T = 90$
$Q_0(T)$	91	85.9	78	59	32.3	18	13.9
$Q_1(T)$	6	10.1	13	15	15.2	13	9.9
$Q_2(T)$	3	4	6	10	16.2	18	16.8
$Q_3(T)$	0	0	2	10	20.2	27	29.7
$Q_4(T)$	0	0	1	6	16.2	24	29.7

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

FV+E2 experiment #5

T in minutes	$T = 0$	$T = 15$	$T = 30$	$T = 45$	$T = 60$	$T = 75$	$T = 90$
$Q_0(T)$	90.1	88	84.7	71.7	41.6	23	23.8
$Q_1(T)$	7.9	10	12.2	12.1	13.9	14	11.9
$Q_2(T)$	2	2	2	6.1	14.9	19	15.8
$Q_3(T)$	0	0	1	6.1	15.8	22	23.8
$Q_4(T)$	0	0	0	4	13.9	22	24.8

Supplementary Materials 1: Linear constraints on the average probability of joint alleles activations For any fixed nucleus NUC_n of the population $pop(T)$, each allele $AL_j, j = 1,2,3,4$ can either be active (state "1") or not (state "0"). There are 16 joint *alleles states* for AL_1, AL_2, AL_3, AL_4 naturally indexed by the first 16 binary numbers as follows

$$S_0 = 0000\,, S_1 = 0001\,, S_2 = 0010, \quad \ldots., \quad S_{15} = 1111$$

Denote $q_{k,n}$ the probability that NUC_n exhibits exactly k activation spots at time T, and let $prob_n$. be the joint probability of alleles activations in NUC_n; we then have the basic probabilistic relations

$q_{0,n} = prob_n(0000)$

$q_{1,n} = prob_n(1000) + prob_n(0100) + prob_n(0010) + prob_n(0001)$

$q_{2,n} = prob_n(1100) + prob_n(1010) + prob_n(1001) + prob_n(0110) + prob_n(0101) + prob_n(0011)$

$q_{3,n} = prob_n(1110) + prob_n(0111) + prob_n(1011) + prob_n(1101)$

$q_{4,n} = prob_n(1111)$

Note that $Q_k(T)$ is the average of the $q_{k,n}$ over all nuclei NUC_n in $pop(T)$, and that F_T is also the average of the $prob_n$ over n. After averaging over n, the preceding linear relations yield

Equation S1:

$$Q_0(T) = F_T(0000)$$

$$Q_1(T) = F_T(1000) + F_T(0100) + F_T(0010) + F_T(0001)$$

$$Q_2(T) = F_T(1100) + F_T(1010) + F_T(1001) + F_T(0110) + F_T(0101) + F_T(0011)$$

$$Q_3(T) = F_T(1110) + F_T(0111) + F_T(1011) + F_T(1101)$$

$$Q_4(T) = F_T(1111)$$

Supplementary Materials 2: Impact of independence on probabilities of joint alleles activations

Fix the time T. Under the average probability F_T of joint alleles activations just defined, let f_j be the probability that allele AL_j is activated at time T. Then $h_j = 1 - f_j$ is the frequency of non-activation for AL_j. Assume temporarily that under the joint probability F_T, the random activations of AL_1, AL_2, AL_3, AL_4 are *independent.* For any joint activation state $(b_1\ b_2\ b_3\ b_4)$, where each $b_k = 0\ or\ 1$, independence of alleles activations implies

$F_T(b_1\ b_2\ b_3\ b_4) = v_1\ v_2\ v_3\ v_4$, where $v_j = f_j$ if $b_j = 1$, and $v_j = h_j$ if $b_j = 0$

Combining these product formulas with the linear relations of equation S1 proves that the nuclei activation frequencies $Q_k = Q_k(T)$ must verify the five formulas

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

$$Q_4 = f_1 f_2 f_3 f_4$$
$$Q_3 = f_1 f_2 f_3 h_4 + f_1 f_2 h_3 f_4 + f_1 h_2 f_3 f_4 + h_1 f_2 f_3 f_4$$
$$Q_2 = f_1 f_2 h_3 h_4 + f_1 h_2 f_3 h_4 + f_1 h_2 h_3 f_4 + h_1 f_2 f_3 h_4 + h_1 f_2 h_3 f_4 + h_1 h_2 f_3 f_4$$
$$Q_1 = h_1 h_2 h_3 f_4 + h_1 h_2 f_3 h_4 + h_1 f_2 h_3 h_4 + f_1 h_2 h_3 h_4$$
$$Q_0 = h_1 h_2 h_3 h_4$$

Divide the first four of these equations by the last one and set $z_j = f_j / h_j$ for $j = 1, 2, 3, 4$, to obtain

$$\frac{Q_4}{Q_0} = z_1 z_2 z_3 z_4$$
$$\frac{Q_3}{Q_0} = z_1 z_2 z_3 + z_1 z_2 z_4 + z_1 z_3 z_4 + z_2 z_3 z_4$$
$$\frac{Q_2}{Q_0} = z_1 z_2 + z_1 z_3 + z_1 z_4 + z_2 z_3 + z_2 z_4 + z_3 z_4$$
$$\frac{Q_1}{Q_0} = z_1 + z_2 + z_3 + z_4$$

These formulas classically imply that z_1, z_2, z_3, z_4 must be the four roots of the polynomial equation

$$R(z) = Q_0 z^4 - Q_1 z^3 + Q_2 z^2 - Q_3 z + Q_4 = 0$$

Hence independence of allele activations under the joint probability F_T forces the polynomial $R(z)$ to have *four positive real valued solutions* $z_1 z_2 z_3 z_4$.

The relations $z_j = \frac{f_j}{h_j} = \frac{f_j}{1-f_j}$ then imply that the unknown probabilities f_1, f_2, f_3, f_4 are given by

$$f_j = z_j / (1 + z_j), for\ j = 1, 2, 3, 4.$$

The independence of alleles activations thus requires the polynomial $R(z)$ to have all its roots real and positive, a condition which imposes *very restrictive polynomial constraints* on the 5 observed frequencies $Q_j(T)$ for each time T. In particular, each one of the 10 pairs $Q_i(T), Q_j(T)$ with $i < j$ must verify very specific polynomial inequalities. For instance, the pairs Q_3, Q_4 and Q_1, Q_0 must verify

$$Q_3 \geq 4\,({Q_4}^{3/4} - Q_4)\ and\ Q_1 \geq 4\,({Q_0}^{3/4} - Q_0).$$

In all our experiments and at all positive times T, the polynomial $R(z)$ with coefficients $Q_0(T), \ldots, Q_4(T)$ derived for image analysis of smFISH data actually did NOT have four positive and real valued roots. This led us to *reject the alleles independence hypothesis* for the population average probability F_T of joint alleles activations

Supplementary Material 3: Maximum Entropy under Constraints

Fix time T and denote F for short the probability $F_T = [F_0\ F_1\ \ldots\ F_{15}]$ on the finite state space $S = [S_0, \ldots, S_{15}]$ of joint alleles activations. The entropy $Ent(F)$ is given by

$$Ent(F) = -F_0\, log(F_0) - F_1\, log\,(F_1) - \ldots. - F_{15}\, log\,(F_{15}).$$

The five frequencies $Q_j = Q_j(T)$ are known and fixed. We know that F must verify the 5 linear relations given by equation S1, which can be rewritten with more compact notations as

Equation S2

$$Q_0 = F_0$$

Ghasemi SM, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107084

$Q_1 = F_8 + F_4 + F_2 + F_1$

$Q_2 = F_{12} + F_{10} + F_9 + F_6 + F_5 + F_3$

$Q_3 = F_{14} + F_{13} + F_{11} + F_7$

$Q_4 = F_{15}$

To seek a probability F maximizing $Ent(F)$ under the 5 linear constraints of equation 2, and the linear relation $\{F_0 + \ldots + F_{15} = 1\}$, we introduce 6 Lagrange multipliers $L_1 \ldots L_6$. The partial derivative D_m of $Ent(F)$ with respect to F_m is equal to $[-1 - log(F_m)]$. The 16 classical Lagrange conditions for optimization under constraints are then

$D_0 = L_0 + L_6;$

$D_1 = D_2 = D_4 = D_8 = L_1 + L_6;$

$D_{12} = D_{10} = D_9 = D_6 = D_5 = D_3 = L_2 + L_6$

$D_7 = D_{11} = D_{13} = D_{14} = L_3 + L_6$

$D_{15} = L_4 + L_6$

Since $D_m = -1 - log(F_m)$ the 5 preceding equations show that

$F_1 = F_2 = F_4 = F_8$

$F_3 = F_5 = F_6 = F_9 = F_{10} = F_{12}$

$F_7 = F_{11} = F_{13} = F_{14}$

Reporting these equalities in Equation S2 yields directly

$F_0 = Q_0$

$F_1 = F_2 = F_4 = F_8 = Q_1 / 4$

$F_3 = F_5 = F_6 = F_9 = F_{10} = F_{12} = Q_2 /6$

$F_7 = F_{11} = F_{13} = F_{14} = Q_3 / 4$

$F_{15} = Q_4$

This fully determines $F = F_T$, and also proves that F_T has *maximum symmetry*, *i.e.,* is *unchanged by any permutation* of the alleles AL_1, AL_2, AL_3, AL_4. Indeed, the preceding expressions obtained for the probability $F = F_T$ can be rewritten

Equation S3

$F_T(0000) = Q_0(T)$

$F_T(0001) = F_T(1000) = F_T(0100) = F_T(0010) = Q_1(T) / 4$

$F_T(1100) = F_T(1010) = F_T(1001) = F_T(0110) = F_T(0101) = F_T(0011) = Q_2(T) /6$

$F_T(1110) = F_T(0111) = F_T(1011) = F_T(1101) = Q_3(T) / 4$

$F_T(1111) = Q_4(T)$

Due to the maximum symmetry of F_T, the probability $P_{AL_1}(T)$ that allele AL_1 is active at time T has the same value for all 4 alleles. By definition $P_{AL_1}(T)$ is given by the sum

$$P_{AL_1}(T) = F_T(0100) + F_T(1100) + F_T(0110) + F_T(0101) + F_T(1110) + F_T(0111) + F_T(1101) + F_T(1111)$$

The explicit formulas S3 just obtained for F_T yield then

$$P_{AL_1}(T) = \frac{Q_1(T)}{4} + \frac{Q_2(T)}{2} + \frac{3Q_3(T)}{4} + Q_4(T)$$

A similar computation provides the joint probability $P_{AL_1,AL_2}(T)$ as was outlined in "Methods".