Promoter-mediated transcriptional burst dynamics underlying the expansion of an actin gene family

Edward Tunnacliffe1, Adam Corrigan1, Jonathan R Chubb1

1University College London

Submitted to Proceedings of the National Academy of Sciences of the United States of America

During the evolution of gene families, functional diversification of proteins often follows gene duplication. However, many gene families expand while preserving protein sequence. Why do cells maintain multiple copies of the same gene? Here we have addressed this question for an actin family with 17 genes encoding an identical protein. The genes have divergent flanking regions and are scattered throughout the genome. Surprisingly, almost the entire family showed similar developmental expression profiles, with their expression also strongly coupled in single cells. Using live cell imaging, we show that differences in gene expression were apparent over shorter timescales, with family members displaying different transcriptional bursting dynamics. Strong bursty behaviours contrasted steady, more continuous activity, indicating different regulatory inputs to individual actin genes. To determine the sources of these different dynamic behaviours, we reciprocally exchanged the upstream regulatory regions of gene family members. This revealed that dynamic transcriptional behaviour is directly instructed by upstream sequence, rather than features specific to genomic context. A residual minor contribution of genomic context modulates the gene OFF rate. Our data suggest promoter diversification following gene duplication expands the range of stimuli that can regulate the expression of essential genes. These observations contextualize the significance of transcriptional bursting.

transcriptional bursting | stochastic gene expression | single cell transcriptomics | Dictyostelium | gene family

Introduction

Gene duplication is recognised as an important process for generating complexity in evolution (1). Following duplication, gene sequences are present in at least two copies in the genome. Assuming these sequences are identical and subject to the same regulatory constraints, they will perform the same function—they are redundant. Over time, duplicate genes typically diverge in sequence and function, however in some cases strong selection acts to maintain identical amino acid or nucleotide sequences over long periods of evolution. Examples include histones, where humans have 14 genes for histone H4, each encoding the same protein (2). Similarly, ribosomal RNA genes are present in hundreds to thousands of copies in eukaryotes with extremely high sequence conservation between family members (3).

Why does an organism require so many genes encoding an apparently identical end product? One explanation is that a large amount of gene product is required and multiple genes allow more transcription. However, while histone genes can be under coordinate control during the cell cycle (4), they have different promoter elements and show varying contributions to total histone content in normal and cancer cells, suggesting regulatory differences between family members.

To understand how differences in gene regulation have influenced the evolution of multigene families, we investigated the actin gene family of the amoeba Dictyostelium discoideum. This organism has more than 30 actin genes, of which 17 (the act8 group) encode an identical amino acid sequence (5). Act8 family genes produce more than 95% of total actin (6) and are dispersed throughout the genome (5). This actin family organisation is a broadly applicable evolutionary strategy, shared by species that diverged more than 400 million years ago (Table S1)(7).

Dictyostelium cells are highly motile, so may require lots of actin, perhaps beyond the production capacity of a single gene. However, estimates of their actin content are of the same order as skeletal muscle, which derives its actin from only one gene (8, 9). Divergent flanking sequences (10, 11) and different genomic contexts of the act8 genes suggest different regulatory dynamics and responses—for example during development. The expansion of the family may also buffer against gene expression noise—it may be undesirable for the expression of an essential protein to be unpredictable, and additional genes may average out noise.

Here, we evaluate the potential for different regulatory dynamics within the gene family. The family shows comparatively similar expression profiles over development, and strong coupling between genes in single cells. However, the genes differ in the dynamics of their transcriptional bursts and show different bursting responses upon induction of development. Switching promoters of actin genes demonstrates that transcriptional dynamics are instructed predominantly by upstream sequence, rather than genomic context.

Results

Developmental dynamics of actin gene expression

Having multiple genes encoding the same protein dispersed throughout the genome may have enabled diversification and refinement of actin expression. Consistent with this view, there is considerable diversity of upstream regulatory sequence between act8 family members, with TATA and 3’ UAS motifs conserved across most of the family, while other motifs such as a G-box are found in only some promoters (Fig. S1A). While some promoters

Significance

Gene transcription occurs in discontinuous bursts. Although bursts are conserved in all forms of life, the causes and implications of bursting are not clear. Here we delineate a specific cause of bursts, and contextualize the significance of bursting, using analysis of a gene family encoding 17 identical actin proteins. Although the genes show similar developmental expression, which is coupled in single cells, they show strong differences in bursting dynamics. These distinct bursting patterns indicate that different signals regulate the individual genes, and imply expansion of the gene family allowed diversification of actin gene regulation. By exchanging the promoters of genes, we show that the dominant driver of bursting dynamics is the gene promoter, not the genome context.

Reserved for Publication Footnotes
contain several elements, others such as act8, show little complexity, with large runs of A and T. Conserved elements were also found at the 3’ end of act genes (Fig. S1A), which could enable further regulatory diversification, although earlier studies showed no strong differences in RNA turnover within the family (12).

To test whether actin genes are differentially regulated, we used RNA sequencing data to determine developmental profiles of gene expression (13). Fig. 1A shows the developmental expression patterns of all 17 act8 family genes. All genes are induced upon differentiation onset with a peak between 1-3h. Most genes show decreased expression during mid development (6-10 h) and by 16h, show little expression. Promoter differences may explain the subtle variations in developmental expression—genes with similar promoters, such as act9, act13 and act14, show more similar expression during development. An exception to the general pattern is act1, which has additional peaks in expression later in development, although the gene shows comparatively few read counts. Despite minor differences, the remaining 16 genes display broadly similar expression, suggesting the generation of different expression profiles during development is unlikely to have been a major influence in the expansion of the family.

**Single cell coupling of actin gene expression**

To obtain single cell resolution and more accurate quantita-
tion of actin gene expression, we measured the relative abundance of act8 family transcripts using single molecule RNA FISH (smFISH). To sample genes with contrasting promoter architecture, we chose act1, act5, act6 and act8. 24 MS2 stem loops (14) were targeted at the 5’ of coding sequences, causing the MS2 loops to be included in the transcribed RNA. Cells with MS2-tagged genes were probed with a fluorescent oligonucleotide complementary to the MS2 array. Cytoplasmic particles corresponding to single RNAs were counted (Fig. 1B). The act5, act6 and act8 genes were all strongly expressed, with act8 the strongest and act6 the weakest, but each gene showed 10s to 100s of RNAs per cell. In contrast, act1 expression was close to background (cells without MS2), indicating that in undifferentiated cells, actin expression at full capacity is not required. In line with this, we found that disruption of up to 4 act8 family genes had no effect on cell doubling times and a 6 gene mutant showed only a weak growth defect (Fig. S1B and C).

Each act gene showed considerable expression variability (Fig. 1B). Although variability is useful in cellular decision-making (15), it can also be disruptive, effectively making some cells overexpress the gene, while others are deficient. Noisy gene expression may create variability for one gene, but uncorrelated fluctuations in the 16 other genes could dilute this noise, allowing an optimal actin level for each cell. To test this reasoning, we used single-cell RNA sequencing (scRNAseq) data (16), to compare the expression of act genes in individual cells. Single cell read counts for all act8 genes were clustered as heatmaps showing correlation values between pairs of genes (Fig. 1C). Some genes showed only weak correlations, consistent with the possibility of multiple family members diluting out stochastic variation. How-

**Distinct bursting patterns of actin genes.** (A) Live imaging of transcription dynamics of three different actin genes. Intensity of transcription spots fluctuates over time to varying extents for each gene (time in min:sec). (B) Transcription spot intensity traces for populations of cells. Each panel represents an individual experiment for each actin gene. Each row represents the intensity trace for a single cell. Black indicates the cell was outside the field of view. Each block in a panel is a field of view. Imaging was carried out over 3 experimental days for each gene, with multiple fields of view per experiment: act1 (152 cells), act5 (476), act6 (384), act8 (275). (C) Probability density functions of spot intensity distributions for all experiments. Horizontal lines represent quartiles and median.
ever, most genes were strongly correlated in their expression in undifferentiated single cells, and across multiple developmental time points (Fig. S1D). These data imply most act8 family genes are co-ordinately regulated in single cells, with cell-cell variability (high versus low expression) a decision taken over most of the family. This suggests the expansion of the family was not selected primarily for buffering molecular noise in gene expression.

We infer no strong differences in act gene post-transcriptional regulation. Protein expression from act genes (monitored using mNeonGreen knock-ins) reflected the differences in transcript level seen by smFISH, with no differences observed in the variability of protein expression or localisation to actin structures (Fig. S1E).

**Actin genes show different transcriptional bursting patterns**

With only small differences in act gene developmental expression, despite divergent control sequences, we tested whether differential regulation is apparent over smaller timescales, using MS2-tagged act genes. Upon transcription, nascent MS2-tagged RNA can be detected with a MCP-GFP fusion protein as a fluorescent spot at the transcription site. In live cells, the spot intensity varies over time (Fig. 2A) reflecting the fluctuating transcriptional activity of the gene (12).

Different actin genes exhibit very different transcriptional behaviours in undifferentiated cells (Fig. 2A). No act1 spots were detected, indicating the gene is inactive or active below the detection threshold (5 RNAs (17)). Both act5 and act6 showed bursts of activity followed by periods of inactivity, with spot intensity greater for act5 than act6. In contrast, act8 activity fluctuations were small compared to the other genes. These example images were representative of hundreds of cells (Fig. 2B). Act5 was strongly ON in most cells, most of the time, whereas act6 was mostly OFF, with the occasional moderate intensity burst. Act5 was between these extremes, retaining frequent switching between ON and OFF. These dynamics are summarised in spot intensity distributions (Fig. 2C). The bursting of act5 and act6 is apparent in the thin upper tails of the distributions, and a large proportion of cells in the OFF state. In contrast, act8 was mostly ON, with an approximate normal distribution of spot intensities in the active state. Act5 and act6 transcription showed significantly higher variance than act8 (Fig. S2A).

To assess the type of control mechanisms regulating act genes, we determined how different measures of variability change with increasing gene activity. Increasing mean expression while decreasing noise (\( \langle C \rangle^2 = \langle C^2 \rangle - \langle C \rangle^2 \)) indicates burst frequency control of gene expression (18, 19). In contrast, increasing noise strength (Fano factor, \( \nu \)) with mean expression indicates burst size control. All three genes showed a strong negative correlation between spot intensity and noise (Spearman’s rank correlation, act5: \( \tau = -0.92 \), \( p = 0.006 \), act6: \( \tau = -0.79 \), \( p = 5 \times 10^{-6} \), act8: \( \tau = -0.83 \), \( p = 1 \times 10^{-4} \) ) (Fig. 3A). In contrast, only act6 intensity correlated with noise strength (act5: \( \tau = 0.13 \), \( p = 0.62 \), act6: \( \tau = 0.83 \), \( p = 1 \times 10^{-10} \), act8: \( \tau = -0.36 \), \( p = 0.17 \)) suggesting act6 modulates both burst frequency and size to change gene expression, while act5 and act8 modulate frequency alone.

Noise decreases with increasing gene expression (20). The difference in noise between act5 and act8 is unlikely to be caused by differences in expression level, as for a given spot intensity, act5 variability was higher than act8 (Fig. 3A). The noise–mean relationship of act5 and act6 can be explained by the same exponential function, as the data will lie on the same linear regression line (Fig. 3A), indicating the difference in variance between act5 and act6 arises because of reduced noise at higher expression.

We directly measured bursting dynamics from live cell data. By imposing a threshold separating ON and OFF phases, using act1 traces to estimate the measurement noise (Fig. S2B,C), ON and OFF durations are represented as cumulative frequency plots (Fig. 3B), with act8 showing short OFF and long ON phases, act6 the opposite, with act5 intermediate.

**Signal regulation of bursting**

What generates the different act gene bursting patterns? One view is that the likelihood of a burst relates to the concentration or activity of a transcription factor (TF), with the TF responding to signalling. Different bursting patterns would result from different genes responding to different signals. Alternatively, all genes respond to the same signals, with different TF binding properties generating different burst dynamics. The complete lack of homology between the act8 and act5 promoters argues against this second view. However, if the first model is valid, there should be stimuli with different effects on different act genes.

Cell size and cell speed both correlate with act5, act6 and act8 transcription, suggesting larger, faster cells are more likely to express actin (Fig. S3A,B). However, these data do not distinguish...
between the genes. For undifferentiated cells, the major requirements for actin are in cytokinesis, phagocytosis and macropinocytosis. To identify signals that might differentially regulate actin bursts, we screened a panel of different culture environments to test how these and other cellular processes relate to actin protein expression (Table S2). Although subtle effects were observed, the relative ordering of expression level for act5, act6 and act8 did not change, so these different culture cues are unlikely to explain strong differences in bursting.

Differentiation in Dictyostelium is induced by starvation, followed by extracellular cAMP and other signals. The increase in relative mRNA for all act8 genes during early differentiation (Fig. 1A) is not suggestive of differences in the regulation of act transcription by these signals. However, global mRNA synthesis rates during differentiation are only ~15% of rates for undifferentiated cells (21), so the relative actin increase might reflect continuing act transcription while much of the genome falls silent, not a sudden transcription surge. To directly visualise transcription during the actin mRNA spike, we imaged act-MS2 cells during early starvation.

Starvation triggered different responses for different act genes. At the onset of starvation, relative levels of transcription were similar to those of undifferentiated cells (Fig. 3C). However, act5 showed a decline in output, with act6 transcription remaining low. The stronger
output of act5 occurred with more cells showing longer, more frequent bursts (Fig. S4C). These data indicate different act genes respond differently to specific stimuli.

**Evaluating the contributions of promoter and genomic context to bursting**

What are the nuclear determinants of bursting dynamics? Many factors acting over a wide range of length scales can control transcription (22-25). Cis-acting elements, chromatin structure and modification, genomic context and nuclear organisation are all potential inputs. However, the ordering and magnitude of these inputs to transcription are unclear, due to extensive crosstalk between the multiple levels of regulation.

Given the diversity in promoter architecture and genome position across the act8 family, we specifically tested the role of promoter sequences and genome context in regulating bursting dynamics. We exchanged the promoters of act6 and act8, the genes with the most distinct bursting behaviours, by replacing the promoter of one gene with that of the other at the endogenous locus. The sequences exchanged were the proximal 388bp (act8) and 543bp (act6) of promoters. Unlike earlier promoter replacement studies (26), we used live cells to directly monitor transcription dynamics. The act6 and act8 genes are exactly the same length, are on different chromosomes, have no introns and can be deleted without phenotype (Fig. S1C). This reciprocal switch enabled us to determine the effects of promoter sequence on gene activity, independently of native genomic context.

Analysis of the transcription dynamics of these switched cell lines identified the promoter as the dominant factor in the regulation of transcription. Fig. 4A shows that the ’A8P-A6G’ cell line, with the act6 promoter upstream of the act8 gene, had remarkably similar transcription dynamics to endogenous act8 – steady gene activity with low cell-to-cell variability. Similarly, the ’A8P-A6G’ gene, with the act6 promoter upstream of act8, was mostly OFF with infrequent short bursts of activity, resembling normal act6.

Spot intensity distributions were similar between endogenous and promoter-swapped genes with the same promoter (Fig. S5A). There may be some residual control from the genomic context, apparent when comparing act8 and A8P-A6G, where the spot intensity distribution of A8P-A6G is more skewed towards zero than the act8 distribution and there is a small, but significant, difference in variance for these genes (Fig. S5B). However, at a coarse level, the promoter is more instructive for nascent transcript levels.

To determine the drivers of transcription dynamics, we used co-occurrence matrices to assess changes in gene activity over short time lags, with each point representing the spot intensity at time t and t+1, where t is a lag (2 frames). Matrices for act6 and act8 both show thin bands parallel to both axes representative of significant short-term changes in activity (Fig. 4B). Data points closer to the diagonal represent more slowly fluctuating gene activity, and most act8 data is here. Plots for both genes show a clear separation between data where the gene slowly fluctuates, and switches rapidly. For the promoter-swapped genes, the overall shape of the distributions indicate the promoter underpins most of the structure of the data, however, for A8P-A6G, the separation between slow and fast switching states is less clear, indicating A8P-A6G and act8 dynamics are not identical.

To further evaluate the contributions of the promoter and genomic locus to bursting, we used an unbiased method of classification based on dynamic features of time series data. The method, "bag-of-patterns" (27), simplifies time-series into collections of discrete symbolic "words" representing the local structure of the data. The "words" within words represent different bins of signal intensity, and the sequences of letters within words reflect the spot behaviour as the intensity fluctuates between bins. Individual time-series are defined by the relative word usage, tallied as histograms (Fig. S6C). Comparison of time-series involves calculation of Euclidean distances between word histograms, with final representation as a dendrogram. Three parameters, word length (w), number of bins (g) and duration of time-series sub-sequence (n) were optimised. We used act5, act6 and act8 data as training sets to identify parameter combinations capable of clustering the data according to gene identity (Fig. 5A and S6D).

To test the contributions of promoter and genomic context to bursting, we used the optimised parameters to classify act6, act8, A8P-A6G and A6P-A8G according to their transcription dynamics. Genes sharing a promoter rather than a genomic locus display more similar dynamic behaviours. Act6 and A6P-A8G clustered more closely together compared to act8 and A8P-A6G (Fig. 5B and S6E). Two A8P-A6G datasets could not be clustered, however all 4 that were clustered fit within the act8 branch. Overall, the majority of the data clustered according to promoter identity, indicating the promoter is the dominant driver of bursting dynamics.

Comparisons between endogenous and promoter-switched genes showed that OFF periods of transcription were not different between pairs of genes sharing a promoter (KS test: act6 v A6P-A8G, p = 0.18; act8 v A8P-A6G, p = 0.16) (Fig. 5C, S7). However, A6P-A8G and A8P-A6G both spend significantly less time in the ON state than act6 and act8 respectively (KS test: act6 v A6P-A8G, p = 0.01; act8 v A8P-A6G, p = 3.9 x 10⁻⁶). This suggests that while promoter sequence controls the majority of the dynamic behaviour, features specific to the genomic locus may influence burst duration.

**Discussion**

To gain insight into the processes driving expansion of an actin gene family in the absence of protein sequence diversification, we evaluated the contribution of differential gene expression. Our data reveal the expansion of the Dictyostelium act8 family is not likely to have occurred to allow different temporal patterns of actin expression during development. Neither is expansion likely to have had a large contribution from a need to buffer noisy gene expression. Instead, we found the genes differ greatly in the dynamics of their transcriptional bursting. We interpret this effect as meaning that the expansion has occurred, at least in part, to enable actin expression to be regulated by an expanded set of signalling cues. In support of this idea, we show that act8 family members show different bursting responses to starvation, a trigger for the onset of differentiation. Genes such as act8 would act as the “workhorses” of the family, delivering much of the cellular actin, with genes such as act5 showing thermostat-like control, to top up levels as appropriate. Expansion may represent a solution to diversifying regulation of a gene with a high transcript load, in a genome too compact for extensive upstream control by multiple enhancers. We then tested whether bursting is driven by the specific promoter sequence of each gene, or via other features of genomic context, by switching promoters of actin genes on different chromosomes. The dominant contribution to bursting dynamics came from the promoter.

Residual effects on bursting dynamics that might be attributed to genomic context were detected in an increased OFF rate observed when a promoter operates at a non-native site. This could conceivably be due to destabilisation of local chromatin conformation by insertion of a non-local DNA sequence. For example, looping between the 5’ and 3’ ends of a gene can facilitate transcription re-initiation which may contribute to the repetitive transcription events of a burst (28). Incompatibility between 5’ and 3’ ends could inhibit looping and increase the OFF rate. This incompatibility scenario assumes specificity provided by the 5’ end of the gene- retaining the emphasis on control back to the promoter. We therefore propose that the increased OFF rate of the switched genes represents an upper limit on the contribution to bursting dynamics from the genomic context.
Our data do not imply chromatin structure and nuclear organisation are unimportant for bursting. We propose that the effects of chromatin and nuclear structure on a gene are first instructed by the DNA sequence and are conserved with many studies showing that chromatin modifications and nuclear organisation are imposed by transcription itself (29, 30). We emphasise that these dominant effects of the promoter have been observed for two genes in an apparently simple developmental eukaryote. This view may have to be modified when considering a metazoan genome with long-range transcriptional control. Dictyostelium has a small genome size with short intergenic regions—over 60% of the genome consists of non-coding DNA. Yet the organism has DNA and H3K9 methylation, mediator, a nuclear lamin and late-replicating peripheral heterochromatin—standard features of metazoan chromatin modification, topology and organisation. Metazoan genes for which genomic context is perhaps most strongly implicated are the globin and Hox clusters (25, 31). Here the genomic context is presumably maintained under strong selective pressure to keep the family together for coordinate control. The idea that the act family has undergone dispersed to sample the diversity of different genomic contexts is not strongly supported by our data.

Methods

Molecular biology and cell line generation

To confirm our in silico results, we targeted an MS2 cassette to the act, act6 and act8 genes respectively. Targeting vectors were designed so that the MS2 sequence was at a similar position in the coding sequence of all genes, 18-24 bp downstream of the ATG. Act5-MS2 cells have been described previously (17). To switch promoters, the same promoter fragments used as targeting arms in the act5-MS2 vector were cloned next to the MS2 sequence in a targeting vector for the other gene. To ensure targeting to the correct locus, we cloned regions upstream of the promoters as the 5′ homology arms of targeting vectors. In MS2 cell lines, the selectable marker was removed by transient CRE expression, to allow MS2 transcription to use endogenous terminators. For visualising nascent RNA, MS2-tagged cells were transfected with an extrachromosomal vector expressing the MCP-GFP fusion protein (17). To monitor protein levels we targeted codon-optimised mNeonGreen to the 3′ end of the endogenous genes in A.X3 cells, followed by removal of the selectable marker. For transcription imaging, we used Dicytostelium A.X3 cells with a stably expressed red fluorescent nuclear marker, H2B-RFP (17).

Imaging and data analysis

Cells were imaged in a low fluorescent medium in 8-well chambers (Nunc Lab-Tek II) and imaged on an UltraVIEW VoX spinning disc confocal microscope (Perkin Elmer) with an EM-CCD camera (C9100-13, Hamamatsu) using a 60x 1.4 NA objective. For imaging during development, cells were washed free of media and viewed under non-nutrient agar (12). Cell tracking, spot identification and motifility analysis was carried out using custom-built Matlab software (17). For smFISH, we used the protocol of (33). We used a single probe, end-labelled with Quasar 670, which binds the spacer between each MS2 stem-loop (17). Cells were imaged using the UltraVIEW with a 100x objective and 640 nm laser. Cytoplasmic mRNA counts were determined using FISH-quant (34).

To implement bag-of-patterns for our transcription dynamics data we concatenated individual cell tracks to generate time-series of around 1000-5000 frames for each cell line per imaging session. Concatenated tracks were log-transformed and normalised to give mean = 0 and SD = 1 (Turing, SE (A, B) for comparison of time series with different offsets and amplitudes. A MATLAB package was used to derive word histograms (https://ds.mnu.edu/~jessica/aux.html) before collating the bag-of-patterns for individual time-series (27). Parameter sets were defined empirically using act5, act6 and act8 data as a training set (Fig. S6C) before applying these to promoter-switched datasets.

Acknowledgments:

Work was supported by Wellcome Trust Senior Fellowship 202867/Z/16/Z to JRC, and MRC funding (MC.U122668) to the MRC LMB Cell Unit at UCL. Imaging was carried out at the MRC LMB Light Microscopy Facility. ET was supported by a London Interdisci-