Minimal effect of gene clustering on expression in *Escherichia coli*

Lusha W. Liang*, Razika Hussein*, Dena H. S. Block* and Han N. Lim*

*Department of Integrative Biology, University of California, Berkeley, USA

§ contributed equally
Running title: Effect of clustering on expression

Key words: Gene colocalization; gene clustering; gene expression noise; genome organization; chromosome position effect.

Corresponding author: Han N. Lim

Department of Integrative Biology
1005 Valley Life Sciences Building
Mail Code 3140
University of California, Berkeley
Berkeley, CA 94720-3140, USA.
Email: hanlim@berkeley.edu
Phone: (510) 643 5915
Fax: (510) 643 6264
ABSTRACT

Genes that interact or function together are often clustered in bacterial genomes and it has been proposed that this clustering may affect gene expression. In this study, we directly compared gene expression in non-clustered arrangements and three common clustered arrangements (codirectional, divergent and operon) using synthetic circuits in *Escherichia coli*. We found that gene clustering had minimal effect on gene expression. Specifically, gene clustering did not alter constitutive expression levels or stochastic fluctuations in expression (“expression noise”). Remarkably, the expression of two genes that share the same chromosome position with the same promoter (operon) or with separate promoters (codirectional and divergent arrangements) was not significantly more correlated than genes at different chromosome positions (non-clustered arrangements). The only observed effect of clustering was increased transcription factor binding in codirectional and divergent gene arrangements due to DNA looping but this is not a specific feature of clustering. In summary, we demonstrate that gene clustering is not a general modulator of gene expression and therefore any effects of clustering are likely to occur only with specific genes or under certain conditions.
INTRODUCTION

Bacterial genes that have products which interact or function within a common pathway are often adjacent to one another on the chromosome (i.e. the genes are “clustered”) [reviewed in (ROCHA 2008)]. Genes that are commonly clustered include those that encode proteins in a metabolic pathway or proteins in multi-protein complexes. Regulatory genes that encode transcription factors and non-coding RNAs are also commonly located next to their target genes. Common arrangements of clustered genes are: 1. codirectional (→→, genes on the same DNA strand with separate promoters); 2. divergent (←→, genes on opposite DNA strands with separate promoters); and 3. operons (→, genes on the same DNA strand with a shared promoter) (KORBEL et al. 2004). Clustered genes that are located on opposite DNA strands with convergent transcription (→←) are less common.

A common functional explanation for gene clustering is that it results in more correlated gene expression. That is, genes that are next to each other on the chromosome are more likely to be affected in the same way by DNA compaction and supercoiling (ALLEN et al. 2006; JEONG et al. 2004; KEPES 2004; MATHELIER and CARBONE 2010; PETER et al. 2004), gene dosage (CHANDLER and PRITCHARD 1975; COOPER and HELMSTETTER 1968; SCHMID and ROTH 1987; SOUSA et al. 1997) and neighboring sequences (DE and BABU 2010). Therefore the expression of clustered genes is expected to be more likely to increase and decrease in synchrony (due to variation in these local chromosome position effects) resulting in more correlated expression than occurs with non-clustered genes. It has been proposed that more correlated gene expression helps maintain the optimal stoichiometry of the gene products thereby increasing efficiency and/or decreasing the concentration of toxic intermediates in metabolic pathways (IBER 2006).
While the hypothesis that bacterial gene clustering leads to more correlated expression is appealing there is no clear evidence to support it. This hypothesis cannot be easily evaluated by comparing the expression patterns of native genes that are clustered and non-clustered because genes in clusters often share the same regulators (Hershberg et al. 2005) and/or regulate each other (Korbel et al. 2004). As a consequence it is unclear whether a higher correlation in clustered genes compared to non-clustered genes (Korbel et al. 2004) is actually an effect of clustering itself or because their regulation is more likely to be coupled. In yeast, synthetic circuits have shown that neighboring genes are more correlated due to local chromosome position effects (Becskei et al. 2005). However, it is uncertain how applicable these findings are to bacteria given their different mechanisms of transcription and translation. Furthermore bacteria have different chromosome structures (Rocha 2008) and lack histones which are largely responsible for position effects on gene silencing (Brand et al. 1985) and gene expression noise in eukaryotes (Batenchuk et al. 2011).

Given that gene clustering is common, any effect it has on gene expression could have a major impact on analyses of gene regulation and the construction of synthetic gene circuits. Therefore we sought to directly test whether gene clustering increases correlations in expression and whether it affects other fundamental properties of gene expression. We performed the study using synthetic non-clustered and clustered arrangements consisting of two fluorescent reporter genes (cfp and yfp) that allow the measurement of gene expression in single cells. The clustered gene arrangements were assembled at two independent loci and both coregulated (i.e. genes regulated by identical promoters) and non-coregulated (i.e. genes regulated by different
promoters) arrangements were created. Gene expression in these synthetic clustered and non-clustered arrangements can be directly compared because they have identical genes and their output and regulation is decoupled from a cell’s physiology and feedback mechanisms. The first part of the study quantified maximum and minimum expression levels, cooperativity, and the inducer concentration needed for half-maximal expression in clustered and non-clustered gene arrangements. The second part of the study assessed the impact of gene clustering on stochastic fluctuations and correlations in gene expression.
MATERIALS AND METHODS

**Plasmids and strains:** Details of the plasmids, strains and oligonucleotides are provided in Tables S1-S3. The plasmids were constructed with *cfp* and *yfp* reporters under the control of the PLlacO-1 and PLtetO-1 promoters amplified from the pZ system of plasmids (LUTZ and BUJARD 1997). Monomeric forms of *cfp* and *yfp* were PCR amplified from plasmids provided by R. Tsien (UC San Diego, CA, USA) (SHANER et al. 2004; ZACHARIAS et al. 2002). We used the highly efficient RBS (T7 10 5’ UTR sequence) from the pET-11a plasmid (Stratagene) to enable single chromosomal copies of the genes to be easily visualized. Plasmids were constructed as templates for PCR amplification and the PCR product consisting of the promoter, fluorescent gene and/or terminator was inserted into the chromosome using the lambda Red system (DATSENKO and WANNER 2000).

Non-clustered gene arrangements were created by two sequential chromosome integrations. Codirectional gene arrangements were constructed by inserting one gene into the genome followed by integration of a second gene downstream of the first gene. Divergent genes were constructed by inserting the second gene on the opposite strand and upstream of the first integration. Operons were created by removing the promoter and terminator sequences separating the two genes in codirectional arrangements resulting in a “scar” sequence of approximately 82 nucleotides separating the first and second genes. Terminators were included in the gene arrangements to prevent transcription out of the fluorescent reporter genes. We found no evidence of transcription read-through into the gene arrangements or transcriptional collisions between genes in the arrangements and neighboring sequences (see **Results** section). All strains were verified by DNA sequencing.
**Measurement of gene expression:** Single colonies grown overnight in LB media were diluted 1/25,000 into fresh LB media with the appropriate IPTG concentration and grown for a further 3.75 hours at 37°C and 200 rpm. The samples were harvested and placed on ice. Immediately prior to microscopy, 1.5 mL of culture was concentrated by centrifugation at 16,100 rcf for 1 min, resuspended in LB and 3 μL was placed on a glass slide with a cover slip (Corning Company). Approximately 5 images were captured per slide using a Pixus 1024 CCD camera (Princeton Instruments) on a Nikon TE2000E inverted microscope with an X-cite 120PC lamp (Exfo) as previously described (LIM *et al.* 2011). Single cells were identified in each image and their average fluorescence was measured using software (Metamorph 7.0, Molecular Devices). There were ~360 identified cells per slide. We measured the background fluorescence levels in a control strain without fluorescent genes which was subtracted from the above measurements.

**mRNA measurements:** The protocol and probes used for the Northern blots were identical to that previously reported (LIM *et al.* 2011). Total RNA was extracted from two separate cultures for each strain at OD$_{600} = 0.1$ to 0.5 using the RNeasy mini kit and RNase free, DNase I (Qiagen). The membranes were probed with a 127 b.p. PCR product that is complementary to the 5’ end of *cfp* and *yfp* or a 115 b.p. PCR product that is complementary to 16S RNA (loading control). The probes were labeled using the DIG High Prime DNA Labeling kit (Roche) and detected with the Detection Starter Kit II (Roche) and radiographic film (Amersham Hyperfilm ECL, GE Healthcare). Digital images were captured on a Gel Doc XR imaging system (Bio-Rad) by transillumination of the film and band intensity was quantified on non-saturated exposures with Quantity One Analysis software (Bio-Rad).
**Models and simulations:** Simplified models of non-operon (non-clustered, codirectional and divergent) and operon gene arrangements were constructed as shown in Figure 7A. Our experiments showed no convincing evidence that non-clustered, codirectional and divergent genes had different constitutive expression or correlation coefficients therefore they are described by the same model; that is, two genes regulated by independent promoters (“non-operon”). The operon has a single promoter that can produce a full length mRNA containing the sequence of two genes or a partial length mRNA containing all the coding sequence of the first gene but not the second gene due to premature termination. Premature termination occurs randomly for 50% of transcription events.

Constitutive expression by the different gene arrangement models was simulated to support the proposed bases for the findings in Figure 6. An mRNA is produced at an average rate of 0.3 mRNA molecules/min and degraded at a rate proportional to the mRNA concentration and the rate constant $k_{m}$ (0.15 per min). The mRNA degradation rate constant was based on a typical half-life of 5 minutes [i.e. $\ln(2)/5$ min] (BERNSTEIN et al. 2002) and this was used to calculate the production rate based on a relatively high steady state mRNA concentration of 2 molecules/cell due to the strong promoter (TANIGUCHI et al. 2010). Protein is generated at a rate determined by the product of the mRNA concentration and the translation rate constant $k_{p}$ (4.0 proteins/mRNA molecule/min), and degraded at a rate proportional to the protein concentration and the rate constant $k_{d}$ (0.03 per min). The protein degradation rate constant was determined by the growth rate as previously reported (BLOCK et al. 2012) and this was used to calculate the translation rate constant assuming the protein concentrations (267 proteins/cell) are higher than
average (TANIGUCHI et al. 2010) due to high mRNA concentrations and the efficient RBS from the T7 10 gene (OLINS and RANGWALA 1989). It should be noted that the experimental values for the mRNA and protein concentrations may be higher than our simulated values therefore the simulated contribution of “intrinsic” noise to gene expression and the effect of operon arrangements on the correlation coefficient should be considered to be an upper bound.

Variation in global factors such as RNA polymerase, RNase E, ribosomes and cell growth were included in the simulations by assuming the rate constants randomly varied from cell to cell by up to ± 50% of the mean. These random differences in the global factors that alter the rate constants for the different reactions steps (transcription, mRNA degradation, translation, and protein clearance) were fixed over the time course of the simulations based on reported observations (ROSENFELD et al. 2005) and uniformly distributed over the range. The simulations were performed with Gillespie’s algorithm (GILLESPIE 1977) using Matlab R2008a (The Mathworks).
RESULTS

Experimental system: Clustered and non-clustered gene arrangements were constructed using the cfp and yfp genes encoding monomeric CFP and YFP respectively (SHANER et al. 2004; ZACHARIAS et al. 2002) (Figures 1A, 3A and Figures S1-S2). The non-clustered genes were placed at separate chromosome locations (intS and galK) approximately equidistant from the origin of replication to minimize differences in gene dosage (COOPER and HELMSTETTER 1968; SOUSA et al. 1997). For the clustered group, three common gene arrangements were examined: 1. codirectional; 2. divergent; and 3. operons. The clustered gene arrangements were constructed at both the intS and galK sites (ELOWITZ et al. 2002) and the order of cfp and yfp was swapped at the two sites so we could distinguish general effects from effects that are limited to a particular locus or gene order. For simplicity, we refer to gene arrangements as being at “intS” if cfp or yfp is located at intS or the adjacent gtrAB genes and at “galK” if cfp or yfp is located at galK or the adjacent galM gene. CFP and YFP expression were measured by fluorescence microscopy (Materials and Methods).

We were unable to insert the different gene arrangements into the chromosome in a single step so they were assembled piece by piece. As a consequence, in some cases the kanamycin antibiotic resistance gene (KanR) had to be left in situ to prevent loss of the inserted genes due to flanking FRT sites and in other cases KanR was removed to permit the insertion of a second fluorescent gene. Therefore some genes had KanR upstream of the promoter and others had a terminator upstream (Figure 1A). We made two different non-clustered gene arrangements (HL5086 and HL5087) with the upstream sequences of cfp and yfp swapped so that we could account for the effect of the different upstream sequences on gene expression. Terminators were located
downstream of each gene except for one of the two genes in each divergent arrangement (for reasons that are unclear these arrangements could not be constructed when both genes had a terminator). There was no decrease in gene expression or the correlation coefficient of non-clustered genes that would indicate that transcriptional collisions and read-through were occurring and affecting gene expression (see sections below).

We confirmed the expected transcript lengths with Northern blots using a 127 b.p. *cfp* probe complementary to the 5’ end of *cfp* and *yfp* (100% and >99% nucleotide identity respectively) (*Figure S3*). The blots showed the transcription terminators function as expected and there is no transcription read-through from native flanking sequences into the gene arrangements. The blots also indicated that operons generated partial length mRNAs as well as full length mRNAs. Partial length operon mRNAs are not uncommon (KHEMICI and CARPOUSIS 2004; NILSSON and UHLIN 1991; PATEL and DUNN 1995; YARCHUK et al. 1992). In these arrangements, the partial length mRNAs contain the 5’ sequence and are about the length of the first gene therefore they are probably due to premature transcription termination at the “scar” sequence, which can form a hairpin loop between the first and second genes.

**Gene clustering does not alter constitutive expression levels:** In the first set of experiments we measured the “induction curves”, which specify the steady state gene expression at various IPTG concentrations (*Figure 1B-G*), in arrangements where *cfp* and *yfp* transcription were both controlled by the PLlacO-1 promoter (termed “coregulated”). Expression from PLlacO-1 was varied by adding isopropyl-β-D-thiogalactopyranoside (IPTG) to the media which prevents the inhibition of transcription by LacI repressor expressed from the native *lacI* gene. For each
induction curve we determined the maximum and minimum expression (α and δ respectively; both with units: fluorescence a.u.), the IPTG concentration required for half-maximal induction (K, units: μM) and the Hill coefficient (n, unitless) by fitting the induction curves to a Hill-type function (Figure 2A). K is a measure of transcription factor binding affinity and n specifies the amount of cooperativity and steepness of the induction curves.

Maximum expression, which is a measure of constitutive expression, was similar among clustered and non-clustered gene arrangements except for genes in the second position of operons and genes lacking terminators (Figure 2B and Table S4). The reduced expression of the second gene in operons is probably due to premature transcription termination (see above) and the reduced expression of genes without terminators may be due to differences in the 3’ mRNA sequence altering its folding and translation or its degradation rate. The important point from these measurements is that highly expressed genes can be clustered without affecting each other’s constitutive expression.

The minimum expression, dynamic range (ratio of maximum to minimum expression), IPTG concentration required for half-maximal induction (K) and Hill coefficient (n), which are all measures of transcription factor binding, varied among the arrangements (Figure 2C-F and Table S4). Non-clustered and operon gene arrangements had very similar values for all these properties. In contrast, codirectional and divergent gene arrangements had reduced minimum expression and increased dynamic range, Hill coefficient, and IPTG concentration required for half-maximal induction; together these changes indicate increased transcription factor binding.
The increased transcription factor binding is probably due to DNA looping (Oehler et al. 1994; Oehler et al. 1990) which allows LacI bound at neighboring promoters to interact.

We next examined gene regulation in a set of non-coregulated gene arrangements where one PLlacO-1 promoter in each arrangement was replaced by the constitutively transcribed PLtetO-1 promoter (Figure 3A). Gene expression was again measured at varying IPTG concentrations (Figure 3B-E) and the induction curves for the PLlacO-1 promoter were fitted to the Hill function. Most of the non-coregulated codirectional and divergent gene arrangements had similar values as non-clustered genes (excluding genes that lack a terminator) for minimum expression, dynamic range, IPTG concentration for half-maximal induction, and Hill coefficient (Figure 3F-J and Table S4). That is, the codirectional and divergent gene arrangements no longer had increased transcription factor activity and cooperativity without the second copy of the PLlacO-1 promoter to enable LacI binding and DNA looping between neighboring promoters.

**Gene clustering has minimal impact on gene expression noise:** We characterized the impact of gene clustering and gene arrangements on stochastic fluctuations in expression (i.e. gene expression noise). Gene expression noise was quantified by the coefficient of variation (C.V.), which is simply the standard deviation of the expression divided by the mean. The C.V. was determined for all strains as well as an additional set of non-clustered genes without downstream terminators (configuration shown in Figure S2). The latter serve as controls for the divergent gene arrangements to determine whether any difference in expression noise is due to the arrangement itself or due to one of the genes lacking a terminator.
Our analysis showed that gene expression noise generally decreased as the mean expression increased in agreement with previous models (Paulsson 2004) and experiments (Elowitz et al. 2002; So et al. 2011) (Figure 4A-H). We found no consistent difference in this relationship between expression noise and mean expression in clustered and non-clustered arrangements. That is, the “curve” did not move up or down in the different arrangements (with the exception of the coregulated, codirectional arrangements), which would have indicated increased or decreased expression noise for a given mean expression. The coregulated, codirectional gene arrangements displayed increased gene expression noise for both CFP (Figure 4C) and YFP at intermediate IPTG concentrations (Figure S4C). The increased gene expression noise in codirectional arrangements is likely to be due to very high levels of cooperativity (Figure 2F) and the steep induction curves (Paulsson 2004; Pedraza and van Oudenaarden 2005).

Non-clustered genes without terminators (Figures 4B, S4B and S5B) had much lower mean expression than non-clustered genes with terminators but the maximum and minimum levels of expression noise are almost the same (resulting in the curve shifting left as opposed to up or down). That is, the free LacI concentration at 0 and 1000 μM IPTG produces the same amount of gene expression noise in non-clustered genes with and without terminators even though their mean expression is different. Therefore the amount of gene expression noise is not primarily determined by the mean expression itself but rather by the free LacI concentration and thus activity of the transcription factor. This finding is consistent with our previous study (Block et al. 2012) and other reports in bacteria and yeast (Elowitz et al. 2002; Pedraza and van Oudenaarden 2005; Rosenfeld et al. 2005; Skupsky et al. 2010). In summary, our
measurements show that gene clustering and arrangements do not directly influence gene expression noise but can indirectly modulate it by altering transcription factor binding.

The correlation coefficient is modulated by transcription factor binding: It has been proposed that stochastic fluctuations in gene expression will be more correlated for clustered genes (particularly for operons) compared to non-clustered genes (ROCHA 2008; SWAIN 2004; VAN HOEK and HOGEWEG 2007). To investigate this, we calculated the correlation coefficient (R), which is a scale independent metric for assessing the linear dependence of CFP and YFP expression (TAYLOR 1997) (Figure S6), for all gene arrangements (Figure 5). R varies from 1 (completely correlated) to 0 (completely uncorrelated) to -1 (completely anti-correlated).

Because the correlation coefficient can be sensitive to outliers we repeated the analyses with only those cells within 2 S.D. of the mean expression and we obtained similar results (Figure S7). In our system, CFP and YFP do not have equivalent intensity distributions therefore the “intrinsic” (uncorrelated variation in expression) and “extrinsic” (correlated variation in expression) components of the gene expression noise could not be calculated (ELOWITZ et al. 2002; PAULSSON 2005). However, a positive correlation coefficient typically indicates the presence of extrinsic noise, and as the extrinsic noise increases the correlation coefficient approaches a value of 1.

The correlation coefficient for coregulated, non-clustered genes showed little change with increasing IPTG (Figure 5A, B). Therefore the LacI concentration (and consequently its activity) appears to have little effect on the correlation coefficients of non-clustered genes in contrast to
its effects on gene expression noise. The high correlation at maximum induction when LacI is unable to bind, suggests that fluctuations in global factors [including RNA polymerases, ribosomes and the growth rates of individual cells (ELOWITZ et al. 2002; STROVAS et al. 2007; TSURU et al. 2009)] causes the constitutive expression of the non-clustered genes to increase and decrease together.

The correlation coefficients for codirectional and divergent gene arrangements were reduced at low and high IPTG concentrations and elevated at intermediate IPTG concentrations (Figure 5C, D). Therefore in contrast to non-clustered genes, codirectional and divergent gene arrangements are modulated by LacI activity. At low IPTG concentrations, the correlation coefficient may be reduced for several reasons (see further explanation in the Discussion): (i) there is strong repression resulting in decreased mean expression and increased intrinsic noise; (ii) the induction curve is “flat” at low IPTG concentrations in these arrangements (Figure 1) making it less sensitive to fluctuations in the LacI concentration; and (iii) bursts of expression may occur from only one gene due to DNA looping and partial dissociation of the bound LacI complexes (CHOI et al. 2008). At intermediate IPTG concentrations, DNA looping may cause coordinated switching of the two genes between the repressed and unpressed states. These correlated switching events plus the variation in global factors explain the slightly higher correlation in the expression of genes in the codirectional arrangement compared to the non-clustered arrangement (compare average of the peaks in Figure 5C and Figure S7C with Figure 5A, B and Figure S7A, B respectively). At high IPTG concentrations, LacI binding cannot occur so there is no coordinated switching of gene expression and the correlation coefficient is less than at intermediate IPTG concentrations but still relatively high due to fluctuations in global factors.
In operons, the correlation coefficients at most IPTG concentrations were relatively constant and similar to the non-clustered arrangements (compare Figure 5E with Figure 5A). While this result was initially surprising, it is completely consistent with all our other findings. The non-clustered genes showed that most of the correlations in gene expression arise from fluctuations in global factors (*i.e.* extrinsic noise) and that stochastic gene-to-gene variation (*i.e.* intrinsic noise), which contributes to anti-correlated expression, is relatively low. Therefore further reducing the latter by transcribing genes in the same mRNA to eliminate anti-correlated transcription events has minimal effect on the correlation coefficient (see supporting models below).

The correlation coefficient for most of the non-coregulated genes in clustered and non-clustered arrangements generally increased with the IPTG concentration (Figure 5F-H). The exception appeared to be the non-coregulated, codirectional arrangement (HL3953, Figure 5G) but this was primarily due to a small number of outliers (Figure S7G). The general increase in the correlation coefficients can be explained by fluctuations in the global factors having opposing effects on PLtetO-1 and PLLacO-1. Global factors that cause gene expression from PLtetO-1 and PLLacO-1 to increase will also increase LacI production and thus act to repress some of the expression from PLLacO-1. The net result is a decrease or a smaller increase in PLLacO-1 expression compared to PLtetO-1 and thus a reduced correlation for the non-coregulated genes. At high IPTG concentrations, LacI cannot bind and therefore PLtetO-1 and PLLacO-1 expression increase and decrease together with fluctuations in global factors resulting in a higher correlation coefficient.
In summary, gene clustering by itself does not appear to directly modulate the amount of correlation in gene expression but certain arrangements can do so indirectly via DNA looping and/or highly cooperative transcription factor binding.

**Constitutive gene expression is highly correlated:** The above experiments at saturating IPTG concentrations strongly suggested that variation in global factors causes highly correlated constitutive expression. We deleted the lacI gene in strains with non-clustered and clustered gene arrangements to eliminate the possibility that LacI might have influenced the correlation coefficients even at saturating concentrations of IPTG. We found that with lacI deleted the correlation coefficient for constitutive expression was still very high and it was similar in both non-clustered and clustered gene arrangements (Figure 6). In particular, we again observed that the correlation coefficient was not significantly higher in operons than in other gene arrangements. The correlation coefficient for the divergent arrangements was slightly lower than other arrangements and the significance and basis of this is unclear.

It has been shown theoretically that the constitutive expression of genes in operons should be more correlated than the expression of genes in non-operon arrangements (i.e. non-clustered, codirectional and divergent arrangements) (SNEPPEN et al. 2010; SWAIN 2004). A likely reason that this was not experimentally observed is because these models did not include variation in global factors, which minimizes the contribution of gene-to-gene variation to the total gene expression noise and the correlation coefficient. To demonstrate this point we created models of non-operon and operon gene expression (i.e. any arrangement where two genes are transcribed in separate mRNAs and an arrangement where two genes share a single mRNA respectively). There
were two models for the operon; in one a single promoter produced only full length mRNAs and in the other a single promoter produced equal amounts of full length and partial length mRNAs as was observed experimentally (the partial length mRNA only has the complete sequence for the first gene due to premature transcription termination).

We simulated the two models with each reaction step (transcription, mRNA degradation, translation, and protein degradation) occurring stochastically (Figure 7A and Materials and Methods). These simulations without variation in global factors showed the expression of genes in operons was highly correlated compared to non-operon arrangements because each stochastic transcription and degradation event caused the mRNA for both genes to be created and destroyed at the same time. In contrast, the creation and destruction of the separate mRNAs for each gene in non-operon arrangements was independent and therefore constitutive expression was less correlated. In addition, the simulations under these conditions showed that premature transcription termination only has a small effect on the correlation coefficient of operon genes (Figure 7B).

We repeated the simulations with the inclusion of variation in global factors such as the concentrations of RNA polymerases, RNA degradosomes and ribosomes, as well as the growth rate which determines the clearance rate of stable proteins such as CFP and YFP. Variation in global factors caused the constitutive expression of genes in non-operon and operon arrangements to become more highly correlated (Figures 7C and S10). However, the relative increase in the correlation coefficient was greater for non-operon arrangements and consequently
the difference in the correlation coefficients between operon and non-operon arrangements was much smaller with variation in global factors just as we observed experimentally.

In summary, the experiments and simulations demonstrate that fluctuations in global factors cause highly correlated constitutive expression in our system. As a consequence, gene clustering has relatively little impact on the correlation of constitutive gene expression.
DISCUSSION

In this study we systematically assessed the general effects of clustering and common gene arrangements on expression using synthetic circuits in *E. coli* that allowed expression between non-clustered and clustered genes to directly be compared at identical loci and without feedback control. It would have been very difficult to characterize the effects of clustering and gene arrangements by analyzing the expression of native genes because there are too many confounding factors. Our system also enabled the assessment of multiple facets of gene expression including constitutive expression, transcription factor activity, gene expression noise, and correlations in gene expression.

Our results showed that gene clustering and arrangements generally have minimal effects on expression. Gene clustering did not directly alter constitutive expression, gene expression noise or correlations in gene expression. However, some clustered arrangements (*e.g.* codirectional and divergent) have the potential to alter transcription factor binding via DNA looping and this can indirectly alter the noise and correlations in gene expression. It must be stressed that DNA looping is not specific to clustering and can be achieved by simply having multiple transcription factor binding sites in the vicinity of the gene (GEANACOPOULOS et al. 2001; LEWIS et al. 2011; OEHLER et al. 1994; OEHLER et al. 1990; SCHLEIF 2010). The absence of direct effects of clustering on gene expression is very informative; as we further discuss below it indicates that spatial differences within bacteria and along the chromosome generally have minimal impact on gene expression.
Our demonstration that gene clustering does not alter \textit{constitutive} expression indicates that there is no significant competition or local recruitment of global factors even with highly expressed genes (otherwise the maximum expression of clustered genes would have been lower or higher than non-clustered genes). Therefore there must be an excess of the global factors required for gene expression and rapid diffusion of these factors in the cell under our experimental conditions (\textit{i.e.} exponential growth in rich media at 37\(^\circ\)C with vigorous mixing and aeration). However, this may not be the case during stress when cellular processes compete for ribosomes (Scott \textit{et al.} 2010).

We found the total gene expression noise in our system was not \textit{directly} altered by gene clustering. However, in codirectional arrangements an \textit{indirect} increase in gene expression noise does occur at intermediate IPTG concentrations due to their more switch-like response curves. The minimal impact of clustering and arrangements on gene expression noise is consistent with our recent finding that altering chromosome position (and thus altering local factors such as DNA supercoiling and neighboring DNA sequences) does not affect gene expression noise (Block \textit{et al.} 2012).

Our measurements showed that the two major contributors to gene expression noise in our system are the transcription factor (LacI) and global factors. Global factors may affect any aspect of gene expression including transcription, translation, mRNA degradation, and protein degradation (Maheshri and O'Shea 2007; Shahrezaei \textit{et al.} 2008). Other studies using fluorescent reporter genes have also shown that global factors are important contributors to the total gene expression noise (Elowitz \textit{et al.} 2002; Pedraza and van Oudenaarden 2005;
RASER and O'SHEA 2005; TANIGUCHI *et al.* 2010). Fluctuations in global factors also affect LacI production and/or degradation causing its concentration to fluctuate. Therefore stochastic variation in global factors can both directly and indirectly (via LacI) contribute to total gene expression noise. Global factors tend to fluctuate over long timescales resulting in cell-to-cell variation in expression (*i.e.* extrinsic noise) rather than gene-to-gene variation in expression within a cell (*i.e.* intrinsic noise) (MAHESHRI and O'SHEA 2007; SHAHREZAEI *et al.* 2008).

Intrinsic noise arises from stochasticity in the production and degradation of mRNAs and proteins and it generally increases as their concentrations decrease (*i.e.* intrinsic noise increases as the mean expression decreases) (ELOWITZ *et al.* 2002; PAULSSON 2004). We were unable to directly estimate the contribution of intrinsic noise to the total gene expression noise because the reporter genes did not have equivalent intensity distributions (ELOWITZ *et al.* 2002; PAULSSON 2005). However, we can get some sense of the impact of intrinsic noise from the correlation coefficient because it contributes to anti-correlated expression as well as correlated expression whereas extrinsic noise only contributes to correlated expression. That is, if there is a significant amount of intrinsic noise we would expect anti-correlated expression to increase (and therefore the correlation coefficient should decrease) as the mean expression decreases (ELOWITZ *et al.* 2002; PAULSSON 2004).

In non-clustered and operon arrangements, the correlation coefficient was high at all expression levels (*Figure 5A, 5E*), which suggests the intrinsic noise is relatively small in these arrangements. In contrast, the correlation coefficient in codirectional and divergent gene arrangements was reduced at low mean expression levels. The decreased correlation coefficient
could be due to potent repression causing very low mean expression and therefore a greater amount of intrinsic noise \([i.e.\] increased anti-correlated expression noise\). However, it may also be due to other effects of DNA looping and cooperativity including a “flat” induction curve at low IPTG concentrations \([i.e.\] less correlated expression noise because fluctuations in the LacI concentration have less effect on the target gene\) and partial dissociation of LacI complexes causing bursts of expression from only one gene \([i.e.\] increased anti-correlated expression noise\) (Choi et al. 2008). That is, increased intrinsic noise is only one possible cause for the decreased correlation coefficient in these arrangements.

It has been predicted that the expression of genes in operons should be much more correlated than in non-clustered genes (Sneppen et al. 2010; Swain 2004). However, our experiments and simulations show the actual difference in the correlation in expression of non-clustered and operon genes will be relatively small when global factors are important contributors to the total gene expression noise. The expression of genes in operons will only be substantially more correlated than in other arrangements if the variation in global factors is much smaller than we observed, the transcription rate is very low so the intrinsic noise is greater (Elowitz et al. 2002; Paulsson 2005), fluctuations in the concentration and activity of the transcription factor are not causing correlated fluctuations in coregulated gene expression, and to a lesser extent if there is minimal premature transcription termination. Fluctuations in global factors depend on the environment (reviewed in Maheshri and O'Shea 2007), so it is possible that genes in clusters and operons could have measurably higher correlations under some conditions. It is unclear how common the regulatory and environmental requirements needed to produce more correlated expression for genes in native operons occur physiologically.
While we have found no evidence that gene clustering has general functional effects on gene expression, clustering may have specific functional effects for particular regulatory mechanisms and pathways. It has been proposed that clustering may aid the coordinated folding of proteins within multi-protein complexes because translation can occur at the site of transcription (Pal and Hurst 2004). The close proximity of clustered genes also enables transcription factors at their promoters to interact and modulate their activity as we observed in the codirectional and divergent gene arrangements. Native divergent genes often have overlapping promoters (Bendtsen et al. 2012; Korbel et al. 2004; Warren and Ten Wolde 2004b) which can result in the binding of a transcription factor (or RNA polymerase) at one promoter excluding transcription factor (or RNA polymerase) binding at the other (Choy et al. 1995; Hershberger et al. 1993; Wang et al. 1998). These overlapping promoters may generate anti-correlated expression and enhance the stability of genetic switches (Bendtsen et al. 2012; Warren and Ten Wolde 2004a). The arrangement of genes in operons may facilitate “just-in-time” synthesis of components in pathways thereby ensuring proteins are made as they are needed to maximize efficiency (Zaslaver et al. 2006; Zaslaver et al. 2004). Operons may also aid the evolvability of gene regulation because it is easier to evolve complex programs at a single promoter regulating multiple genes than at multiple promoters each regulating a single gene (Price et al. 2005).

Gene clustering may have arisen due to the many specific functional advantages it provides for gene regulation as we describe above. However, an alternative explanation is that clustering occurs because it increases the probability that multiple genes needed to acquire and regulate a
new trait (*e.g.* genes in a metabolic pathway) are inherited in a single horizontal gene transfer event due to their close genetic linkage as proposed in the “selfish operon hypothesis” (Lawrence and Roth 1996). Genetic linkage and functional explanations for the origin of gene clustering and gene arrangements are not mutually exclusive and both may be important.

Our study has practical applications for the engineering of synthetic circuits, which often require several functionally related genes to generate a compound or achieve a specific function (De Lorenzo 2008; Lee et al. 2008; Medema et al.). It was previously unknown whether the output of synthetic gene circuits could be improved by replicating the common arrangements that occur in the genome. We have shown that clustering by itself should not be a goal or a constraint on the design of gene circuits, at least under conditions of exponential growth and high expression which are common conditions for the use of many synthetic systems. However, in cases where genes are regulated by the same transcription factor and greater transcription factor activity and cooperativity is desired, it could be beneficial to have the genes in codirectional or divergent arrangements (at the cost of a reduced correlation in gene expression at low and high levels of induction). Conversely, if a graded response is required then genes should be placed within an operon or non-clustered arrangement.

Gene clustering is a common feature of genome organization but its impact on bacterial gene expression had not been comprehensively and systematically studied in an experimental system; this study represents an important step forward in addressing this knowledge gap. We found no difference in the constitutive expression levels, expression noise and correlations in expression of three common clustered arrangements compared to non-clustered arrangements. While our
data appear to exclude general effects of clustering on gene expression, it is possible there may be specific genes and/or environmental conditions where clustering does modulate expression.
ACKNOWLEDGEMENTS

We thank David Adamson for helpful discussions and comments on the manuscript, Emily Chang for assistance with fluorescence microscopy measurements, and Bernadett Gaal for constructing some of the plasmids that were used as templates for the PCR products that were integrated into the chromosome. This work was supported by start-up funds and a COR Faculty Research Grant (FRG) from University of California, Berkeley. The authors state there are no conflicts of interest. LWL, RH and DHSB made the strains and performed the measurements. HNL designed the experiments, performed the analyses and simulations, and wrote the paper.
REFERENCES


Liang, Hussein, Block & Lim


FIGURE LEGENDS

Figure 1. Induction curves for coregulated genes. Error bars indicate the s.e.m. of duplicate (HL5087), triplicate (HL3546, HL3573, HL3578, HL4343 and HL4348), and quadruplicate (HL3515 and HL5086) measurements. (A) Diagram showing the arrangements of coregulated genes where LacI regulates both genes. (B-G) CFP and YFP expression at different IPTG concentrations. Data symbols indicate gene arrangements shown in panel A and lines are fits obtained with the Hill function. # Expression values for HL4348 at low IPTG concentrations after background subtraction were ≤ 0 therefore a small offset (5 a.u.) was added to the expression simply to enable visualization of the data on logarithmic plots. 0* indicates an actual value of zero not 10^0.

Figure 2. Effects of gene clustering on coregulated gene expression. Error bars indicate the s.e.m. (number of replicates is stated in Figure 1 legend). (A) Hill function and its parameters (defined in main text). (B-F) Maximum expression, minimum expression, dynamic range, IPTG concentration required for half-maximal induction, and Hill coefficient for the different gene arrangements. Data symbols indicate the gene arrangements in panel A. Notes: * low expression of cfp and yfp in the second position of the operons; # cfp or yfp without a terminator; ND = not displayed because the values were ≤ 0 after background subtraction.

Figure 3. Induction curves for non-coregulated genes. Error bars indicate the s.e.m. of duplicate (HL5113 and HL5149) and triplicate (HL3953, HL3954, HL4302, and HL4342) measurements. (A) Diagram showing the non-coregulated gene arrangements where one gene has the PLLacO-1 promoter and the other has the PLtetO-1 promoter. (B-E) Induction curves for
CFP and YFP expression at different IPTG concentrations. Data symbols indicate gene arrangements shown in panel A. Lines are fits to the Hill function in panels B and D. Lines in panels C and E are a guide to the eye. 0* indicates an actual value of zero not 10^0. (F-J) Maximum expression, minimum expression, dynamic range, IPTG concentration required for half-maximal induction and Hill coefficient for the different gene arrangements shown in Figure 3A. # Measurements of cfp and yfp genes without terminators.

Figure 4. Gene expression noise for CFP in non-clustered and clustered gene arrangements. YFP expression noise is shown in Figure S4. (A-H) Each panel shows the coefficient of variance (C.V.) for a different gene arrangement. The non-clustered gene arrangement with terminators serves as a reference in all plots (blue shading indicates the range of the C.V. for this arrangement). Data symbols indicate gene arrangements shown in Figures 1A and 3A. In addition we include non-clustered genes without terminators (maroon symbols, panel B). An analysis that has only cells within 2 S.D. of the mean is shown in Figure S5. Error bars indicate the s.e.m. (number of replicates is stated in the legends of Figures 1 and 3).

Figure 5. Correlation coefficient as a function of IPTG concentration for clustered and non-clustered gene arrangements. (A-H) Each panel shows the correlation coefficient for a different gene arrangement. Data symbols indicate the gene arrangements shown in Figures 1A and 3A. In addition we include non-clustered genes without terminators (maroon symbols, panel B). Analyses including only cells within 2 S.D. of the mean expression are shown in Figure S7. The correlation coefficients are plotted as a function of mean expression in Figure S8. 0* indicates an actual value of zero not 10^0. Blue shaded regions show the approximate upper and
lower bounds of the non-clustered gene arrangements with terminators. Error bars indicate the s.e.m. (number of replicates is stated in the legends of Figures 1 and 3).

Figure 6. Correlation coefficient of gene expression in non-clustered gene arrangements with lacI deleted. Data symbols indicate the gene arrangements shown in Figure 1A. Error bars indicate s.e.m. of 5-7 replicate measurements. Correlation coefficients are calculated from all cells. An analysis with only the cells within 2 S.D. of the mean expression is shown in Figure S9.

Figure 7. Correlation coefficients from stochastic simulations of non-operon and operon arrangements. (A) Simplified schemes for non-operon and operon gene arrangements as described in the main text. (B) Correlation coefficients of non-operon and operon gene arrangements generated by stochastic simulations in the absence of variation in global factors of gene regulation. (C) Correlation coefficients of non-operon and operon gene arrangements generated by stochastic simulations with global fluctuations (± 50% of the mean) in transcription, mRNA degradation, translation and protein degradation.
A

NON-CLUSTERED

[Diagram showing genetic elements]

CLUSTERED

Codirectional

[Diagram showing genetic elements]

Operon

[Diagram showing genetic elements]

Divergent

[Diagram showing genetic elements]

= Asp term. = T1 term. = KanR

B

CFP expression (fluorescence, log10 a.u.)

C

YFP expression (fluorescence, log10 a.u.)

D

CFP expression (fluorescence, log10 a.u.)

E

YFP expression (fluorescence, log10 a.u.)

F

CFP expression (fluorescence, log10 a.u.)

G

YFP expression (fluorescence, log10 a.u.)

Figure 1. (Liang et al.)
$\text{Expression} = \alpha \left( \frac{[\text{IPTG}]^n}{[\text{IPTG}]+K^n} \right) + \delta$

Non-clustered (■ & □)
Codirectional (■ & □)
Operon (■ & □)
Divergent (■ & □)

**Figure 2.** (Liang et al.)
Figure 3. (Liang et al.)
Figure 4. (Liang et al.)
Figure 5. (Liang et al.)
Figure 6. (Liang et al.)
Figure 7. (Liang et al.)