Chromatin interaction network as mediator of error mitigation in genome

A dissertation submitted for partial fulfillment of BS-MS dual degree in Science

Meenakshi Bagadia

MS09083

Indian Institute of Science Education and Research Mohali

Knowledge City, Sector 81, SAS Nagar, Manauli, PO 140306

April 2014

Certificate of Examination

This is to certify that the dissertation titled “Chromatin interaction network as mediator of error mitigation in genome” submitted by Meenakshi Bagadia (Reg. No. MS09083) for the partial fulfillment of BS-MS dual degree program of IISER Mohali has been examined by the thesis committee duly appointed by the institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr. Somdatta Sinha	Dr. Chanchal Kumar (Co-supervisor)	Dr. Kuljeet Singh Sandhu (Supervisor)

Dated:

Declaration

The work presented in this dissertation has been carried out by me under the guidance of Dr. Kuljeet Singh Sandhu at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Meenakshi Bagadia

(Candidate)

Dated: April 25th , 2012

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Kuljeet Singh Sandhu

(Supervisor)

Acknowledgement

I would like to express my gratitude to my supervisor Dr. Kuljeet Singh Sandhu for the useful comments, remarks and engagement through the learning process of this master thesis. He is an excellent teacher and guide who inspired me to excel academically as well as personally. His support and guidance throughout the project has been invaluable to me and his enthusiasm for his work has always inspired me.

I would like to thank Dr. Chanchal Kumar for supporting and co-supervising this work. It was always encouraging to have him.

I would like to give respect to the thesis committee consisting of Dr. Somdatta Sinha, Dr Chanchal Kumar (Co-Supervisor) and Dr. Kuljeet Singh Sandhu (Supervisor) for giving valuable suggestions during different phases of the project.

I would also like to thank the Department of Mathematics and Department of Biological Sciences, IISER Mohali for providing me the opportunity to work in an interdisciplinary area. I would also like to thank Library, IISER Mohali for providing the facilities which allow us to do healthy research.

I would especially like to thank Nitesh Tayal, an ex- MS student under Dr. Kuljeet Singh Sandhu. He has always been helping me whenever I faced problem during my work.

I am also thankful to all the members of Computational Biology Lab Adhikar, Arashdeep, Ashutosh, Kanwal Puneet, Preeti, Priya, Rivi and Srishti. I would like to thank them for building up a positive environment and making my stay in lab a wonderful experience.

At last, I would like to thank my friends, who have supported me throughout entire process, both by keeping me harmonious and helping me putting pieces together. I will be grateful for their love and support.

Meenakshi

Contents

List of Figures

Abstract

Our cells continuously experience several kinds of non-genetic errors, which need to be mitigated in order to keep the robustness of the cellular system. Here we studied two types of error.

1) First, even if all cells are precisely in same conditions, there is stochastic variation in gene expression among the cells, which is termed as intrinsic noise. If the expression of an essential gene is altered to a significant level, then it may be lethal for the cell. Thus for the stable functioning of the cell it is necessary to keep the expression noise minimum for essential genes. How exactly cells achieve this is not clear. We hypothesize that physical attachment of genes to the sub-nuclear compartments like transcription factories might reduce its mobility and consequently the noise in its transcription.

2) Second, altered expression at certain loci could disproportionate the required concentrations of functionally related gene-products, which are generally positioned in the genomic neighbourhood. This error can be mitigated by simultaneous alteration in the expression of neighbouring genes, termed as *transcriptional ripple*. The underlying mechanism is not understood. We propose that physical interactions among neighbouring genes influence their transcriptional states. To simulate the phenomenon, we made perturbation transmission model inspired by communicating vessels principle, which essentially captures the ripple-effect and can be used study the phenomenon and its functional consequences at genome scale.

Keywords: gene expression, transcriptional noise, chromatin interaction, transcriptional ripple, communicating vessel model

Chapter 1 Stochastic variation in the gene expression

Introduction

An individual in a population is unique. Most of the population variation is due to genetic difference. However, it has been studied that genetically identical individuals can be very different and most important source of this variability are random fluctuation in the expression of individual genes. Fundamentally, this is because the gene expression involves a set of discrete and random biochemical reactions that control the abundance of gene products. Because DNA, RNA, proteins are present in very low numbers, even small fluctuations can generate huge population variation thus gene expression must be thought as a stochastic process, with the randomness in transcription and translation leading to cell-cell variation in mRNA and protein levels[1,2].

We here refer *noise* in gene expression as the stochastic variation in mRNA concentration among the isogenic cells. The noise can be intrinsic as well as extrinsic. Even when cells are at same identical state, the reactions leading to transcription and translation of a gene of particular interest would occur at different times, in different orders in different cells[3]. Such stochastic variability is termed as intrinsic expression noise. Whereas extrinsic expression is variation in the level of gene expression due to different environmental conditions. Here we are mainly interested in intrinsic transcriptional noise.

Why it is important to study the expression noise? Consider an essential gene, which needs to be transcribed consistently in the cell. If the expression level of such a gene is highly varied, then it might be harmful for the cell. So it is necessary to minimize the expression noise for essential genes.

The project aimed to decipher the determinants of the transcriptional heterogeneity in the cells, with the *hypothesis* that transcriptional noise is modulated by relative mobility of gene loci in the three dimensional nuclear space.

Chromatin interaction network (ChINs)

Complexity of chromosomes architecture has been known since the end of nineteen century, when chromatin loops were first observed[4]. Although genetic information is stored in the linear sequence of base pairs that make up the DNA, but further it has been found that DNA is intricately folded in higher-order three- dimensional structure which involves the formation of chromatin loops, where distal elements of the chromatin fibre come into close physical proximity with each other. Also it has been documented that this 3-D organization of genome inside the nucleus has consequences for the regulation gene expression and/or propagation of genome[4].

The chromatin loops describe short range and long range interactions in *cis* whereas chromatin bridges depict long range interactions in *trans*. Both the interacting partners must reach beyond the confines of its chromosome territory for interactions[4].

To enable loop formation, the chromatin fibres must physically encounter each other. Formation of chromatin loops may be formulated by DNA condensation, super coiling, higher affinity protein-protein interactions or additional force applied by strong DNA binding proteins[5].

With technological and methodological advancements in biology and with the availability of whole genome sequencing methods, it has become possible to capture the genome wide chromatin interaction profile of a cell using a technique called chromosome conformation capture (3C) and the related techniques like circular chromosome conformation capture (4C), ChIA-PET

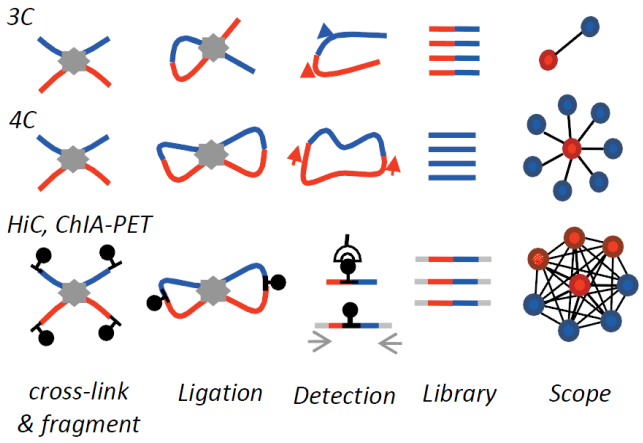


Fig 1.1: Exploring ChINs : Shown are the proximity ligation based methods to identify chromatin interaction[5]

Materials and methods:

Data source

- The genome wide single cell gene expression dataset[7] of embryonic stem cell of mouse consists of expression of 24435 genes across 13 cells from ICM.
- RNA-pol2 associated chromatin interaction data[8] consists of pair wise interactions of genes which are mediated by RNA-pol2.

Data normalization

Normalisation is required before any analysis which involves comparison of two or more data sets. Here, the genome wide single cell gene expression data was quantile normalised[6] in R, using the DNAMR library (refer to Fig S1).

The main principle of quantile normalization is to make the distribution of probe intensities for each array in a set of arrays the same. The algorithm for normalising a set of data vectors by given them the same distribution is:

1. Given n arrays of length p, construct matrix X of dimension p x n, where each array is a column.
2. Sort each column of X and represent this new matrix by X_{sort}
3. Take average across rows of X_{sort} and assign this average to each element in the row to get X'_{sort}
4. Rearrange each column of X'_{sort} to have the same ordering as original X, to get $X_{normalized.}$

   Here is an example of the above algorithm.

Original			Sorted			Averaged			Re-ordered		
5 (i)	4 (i)	3 (i)	2 (ii)	1 (ii)	3 (i)	2 (ii)	2 (ii)	2 (i)	5.67 (i)	4.67 (i)	2 (i)
2 (ii)	1 (ii)	4 (ii)	3 (iii)	2 (iv)	4 (ii)	3 (iii)	3 (iv)	3 (ii)	2 (ii)	2 (ii)	3 (ii)
3 (iii)	4 (iii)	6 (iii)	4 (iv)	4 (i)	6 (iii)	4.67 (iv)	4.67 (i)	4.67 (iii)	3 (iii)	5.67 (iii)	4.67 (iii)
4 (iv)	2 (iv)	8 (iv)	5 (i)	4 (iii)	8 (iv)	5.67 (i)	5.67 (iii)	5.67 (iv)	4.67 (iv)	3 (iv)	5.67 (iv)
X			X_{sort}			X'_{sort}					
$X_{normalize}$											

Calculation of Expression noise

The expression noise of genes across the cell was calculated in terms of coefficient of variance. Mathematically it is defined as,

$$Epression\ noise\ CV_X = \frac{\sigma_X}{\mu_X}$$

where σ_X is the standard deviation and μ_X is the mean value of the expression across the cells (X) (for further analysis refer Fig S2).

Another way of defining noise is through fano factor, i.e.

$$F_X = \frac{\sigma_X^2}{\mu_X}$$

Network construction and analysis

The undirected network was constructed from the RNA-pol2 associated chromatin interaction data, using igraph library on R-package. Number of interactions (degree) of each gene was calculated using 'degree' function on igraph (using a Perl script) (for analysis refer to Fig S3)

Calculating shortest path

We were then interested in studying the spatial organization of genes especially the low noise genes. For this, first from the noise data, the genes were distributed on the basis on low, medium and high noise respectively. Then the interaction network for each noise type was extracted from the entire genome interaction network (using Perl script). For each gene, shortest path to every other gene from this chosen gene was calculated using 'shortest.path' function on igraph and were averaged to get the average shortest path for the chosen gene for each noise type. Furthermore, average path length of the entire noise type was calculated using 'average.path.length' function on igraph library on R package. The analysis was done through boxplot in R.

The 'shortest.path' function uses Breadth First Search algorithm.

Breadth First Search:

Given a graph G and a vertex s in G, breadth-first search algorithm traverses all the vertices reachable from s. Algorithm for BFS is as follows:

1. Let $L_0 = \{s\}$
2. Let L_1 to be all the neighbours of L_0
3. Let L_2 to be all nodes that do not belong to L_0 or L_1, and that have an edge to a node in L_1
4. Similarly L_{i+1} to be all nodes that do not belong to an earlier layer and that have an edge to a node in L_i

For each i, Li consists of all nodes at distance exactly i from s. There is a path from s to t if and only if t appears in some layer.

To further scrutinize the observation, we used mRNA degradation data and enrichment of H3K4me3/H3K27me3 which are described below.

mRNA degradation dataset

We have database for mRNA half-life of 19977 genes of pluripotent and differentiating mouse embryonic stem cell[9]. The data consist of two mouse ESC lines: MC1 and MC2 in different environment conditions (LIF+, LIF- and RA+ respectively). The noise and degree data was mapped to this data to study the dependence of mRNA decay rate on noise and degree.

Enrichment dataset (H3K4/27me3)

Next, we were interested in studying the effect of epigenetic. For this, the mm8 assembly of mouse was downloaded from UCSC Genome Bioinformatics website. This was made unique on the basis of gene names, by taking the gene which is having largest transcript length. Then from this data, the region of 5000 base pair near the TSS i.e. TSS±5000 were calculated. Using the java script borrowed from Dr. Guoliang Li (ex-colleague of Dr. Kuljeet Singh Sandhu) and the region profile mouse data[10], the enrichment of H3K4me3 and H3K27me3 were calculated with the bin size of 1000 and 100 base pair. For each gene total enrichment of H3K4me3 and H3K27me3 respectively was calculated in the region TSS±0.5 kb and TSS±1 kb. In this data, the noise and degree data were mapped using unique gene name. Thus, we were then able to study the enrichment of H3K4me3 and H3K27me3 with noise and degree in both the regions.

Using the coordinates of intervals enriched for H3K4Me3 and H3K27Me3 in mouse ES cells, the enrichment of H3K4me3 near the promoter region (TSS±0.5kb, TSS±1kb) was checked (1 or 0); similar procedure was done for H3K27me3 enrichment. Then genes having both the markers H3K4me3 and H3K27me3 were calculated and the fractions of genes having both the markers were studied with noise and degree.

Time course differentiation dataset

The next dataset was time course differentiation dataset[11] of three different mESC strains (J1, R1, and V6.5). Geometric mean was used for each ES (0 hour) and EB (14 hour) in each strain. Fraction of genes having EB/ES ratio >=2 was then calculated for all the three germ lines of mouse, namely R1, J1and V6.5.

In order to ensure the results, the entire data analysis was done again using new set of RNA-pol-2 interaction data[8]. Here the interaction was captured in higher resolution i.e. 1500 bp.

Results:

To test the hypothesis that the association with transcription factories reduces the transcriptional noise, we started with two datasets: the genome wide single cell gene expression dataset of embryonic stem cell of mouse and RNA-polymerase 2 associated chromatin interaction data. Using the above data, expression noise and number of interaction of genes were calculated respectively.

In Fig 1.2, there was strong negative correlation ($\rho = -0.51$, $p < 2.2e^{-16}$) between transcriptional noise and degree highlighting that low noise genes are associated with high degree and high noise genes are associated with low degree. Furthermore, in order to nullify the impact of relative abundance of mRNA-copies, the observation was also scrutinized by calculating the abundance corrected noise using LOWESS in R package and plotted against the degree (refer to Fig S5). Furthermore, low noise genes, due to their higher degree, were more proximal to each other in the chromatin interaction network when compared to high noise genes (refer to Fig S4).

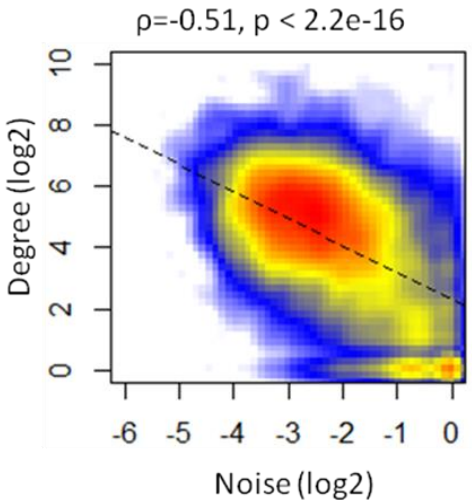


Fig1.2: Correlation between expression noise and degree.

We further scrutinized our observation using yet another dataset. Topologically associated domains (TADs) are large (~1Mb average size) chromatin domains which have very dense intra-domain chromatin interactions[12]. The boundaries of these domains are

enriched with high transcriptional activity and it has been proposed that the boundaries of distant domains collide with each other possibly at the site of transcription factory. Therefore, proximity of a gene from these boundaries can be a proxy to an association with transcription factory. For each gene, distance from nearest TAD boundary was calculated and plotted with noise. Fig1.3, shows significantly positive correlation (ρ = 0.12, $p < 2.2e^{-16}$) between noise and the distance from the TAD boundary, supporting our hypothesis.

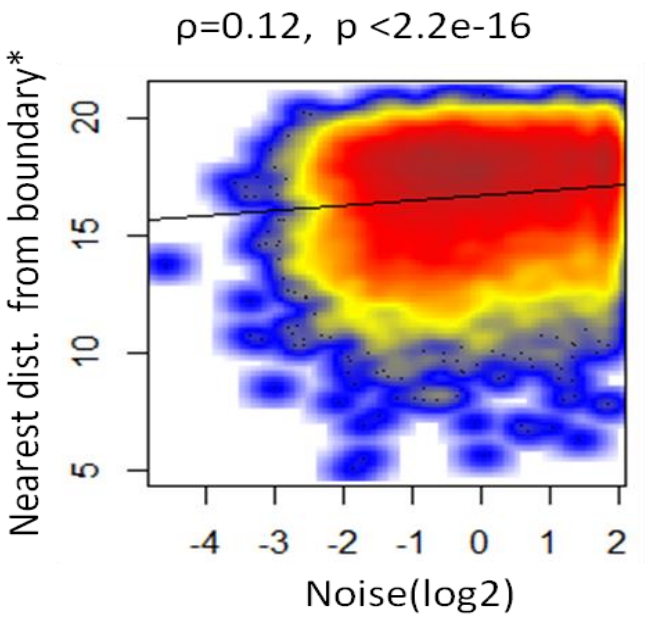


Fig 1.3: Correlation between noise and distance from nearest boundary

Are there any other factors responsible for transcriptional noise?

Degradation of mRNA might be one of the factors that control the steady state level of gene expression. For this, the mRNA decay rate of genes was obtained from microarray analysis of RNA samples obtained from mouse embryonic stem (ES) cells. Fig1.4 shows, a very weak correlation (ρ=-0.037, p =0.002) between mRNA decay rate and transcriptional noise.

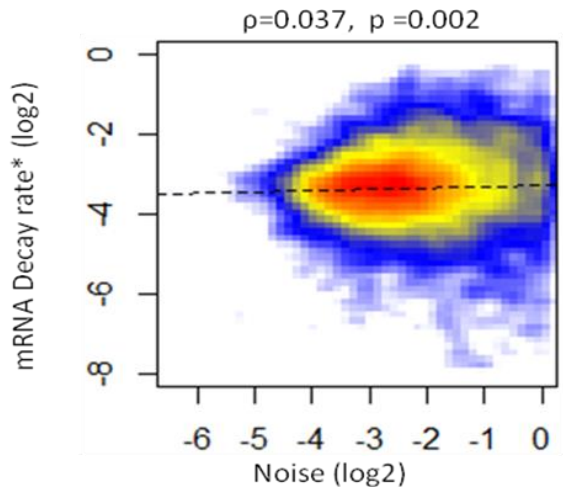


Fig1.4: Correlation between mRNA decay rate and transcriptional noise.

Functional characterization of transcriptional noise and degree

We have the distribution of genes based on their functions. This was studied with respect to the noise and degree. We saw that genes having housekeeping functions like translation, cell-cycle regulation, metabolism, cytoskeleton etc are associated with low noise and high degree, whereas the genes having regulatory functions like development, signal transduction etc. are associated with high noise and low degree.

Interpretation: The cells have evolved a mechanism to minimize transcriptional noise of genes important for the cell survival and to ascribe sufficient noise to genes which require some plasticity or adaptation to the environment.

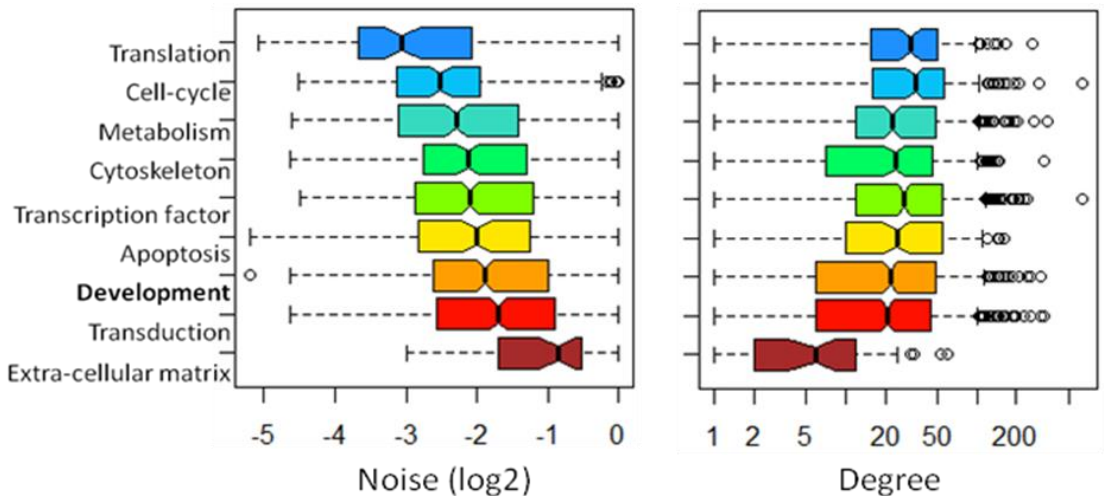


Fig1.5: Box plot showing the functional consequence of genes with noise and degree respectively.

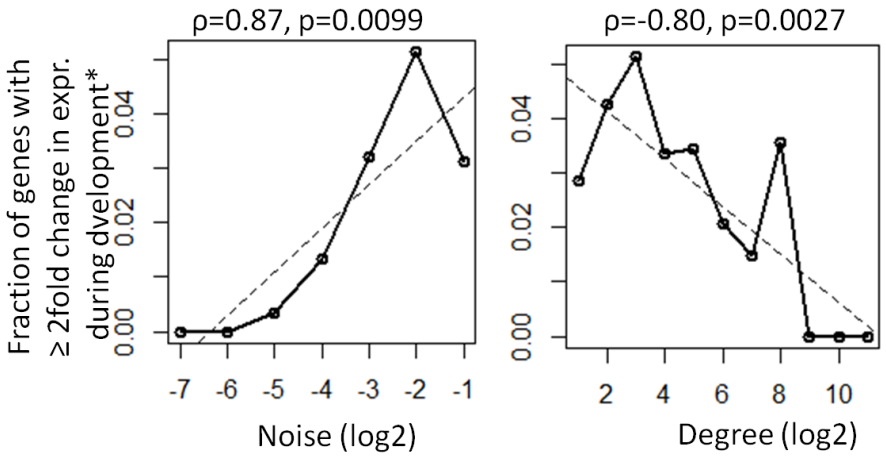
Since our experimental system is mouse embryonic stem cells, we further explored the association of transcriptional noise and degree with the development. In time course differentiation data, when the fractions of genes with > 2 fold change in expression during development was plotted against noise and degree, we observe that it has positive correlation with noise and negative correlation with degree(Fig 1.6 (A)), while the genes important for ES cell self renewal are associated with lower noise and higher degree.

Further, we observed strong positive correlation of fraction of genes having bivalent histone modification (H3K4/K27me3) with noise and negative correlation with degree (Fig 1.6 (B)).

Interpretation: The analyses suggests that genes which are required for ES cell differentiation exhibit bursty expression in ES cells and that the ES cells can be stochastically pre-poised to differentiate into particular lineage. On the other hand, the genes involved in ES cell self-renewal seems to be associated with transcription factories in order to reduce transcriptional noise of pluripotency related genes and keeping the potential of self renewal potential of ES cell. The association with bivalent histone modifications might suggest two things: 1) Bivalency, i.e., having activation and repression potential together, might ascribe noise to transcription. 2) H3K4me3 and H3K27me3 marks on a histone might not be present in the same cell, but rather represents

active and inactive state of the gene in different cells. This is a speculation and needs further evidence.

(A)



(B)

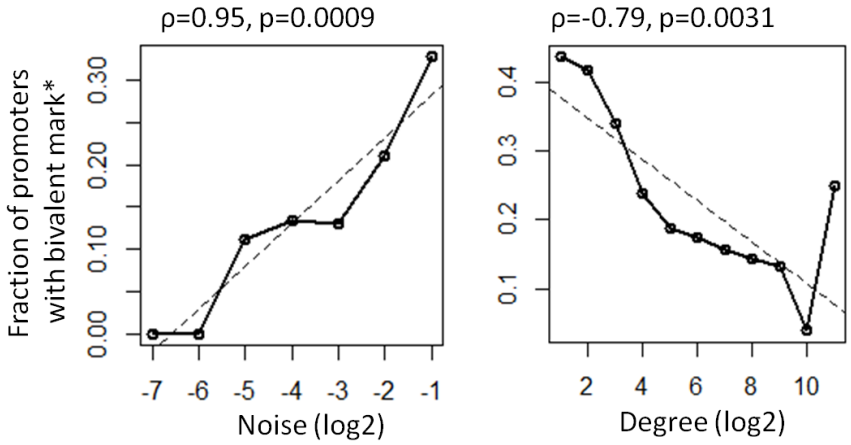


Fig 1.6: Association of noise and degree with ESC differentiation and bivalent histone modification

Chapter 2 Disproportionate concentration of gene products

Introduction

Network is a collection of nodes which are connected by edges. This simple notion of network is now use to study the complex dynamics of biological system. Here, we use networks as a tool to understand the propagation of error or perturbation in the system.

In cells, when the gene expression of a certain loci is altered, then it might lead to the disproportionate concentration of functionally related gene-products. This error needs to be mitigated for the robustness of the cell. One of the solutions can be the transcriptional ripple i.e. simultaneous alteration in expression of neighbouring genes. For this we propose a perturbation transmission model, so that error propagates in the neighbourhood to alter the expression and thus balances the required proportion of gene-products.

The model is inspired by communicating vessel model[14]. In general communicating vessels is the name giving to the set of vessels which are connected by a pipe, containing homogeneous fluid. Liquid will continue to flow in order to balance the level in all the vessels. The same principle can be applied to biological networks for instance chromatin-chromatin interaction network. Here the vessels will be the nodes representing interacting loci or genes and the connecting pipe can be the interaction frequency between the nodes.

The algorithm[13] of the model is:

In each time step, every node transfers a proportion of its available energy through every available edge, proportional to

1. The duration of the time step
2. Weight of the edge
3. Difference of the energy states on the two ends of the edge

In most dynamical system there is always dissipation constant associated with each node. It is the amount of energy dissipated by the node.

Thus the differential equation for the model is:

$$\frac{dS}{dt} = -\sum_{i=0}^{l}\left[\frac{S - S_i}{2} w_i\right] - D_0$$

where,

S : energy of current node

l : no of edges of the current node

$w_{i:}$ weight of the i^{th} node

S_i : current energy of the node on the other end of i^{th} edge

D_0 : amount of energy dissipated by a node in a given time step

Discrete form of the above equation is:

$$S[t+1] = -\sum_{i=1}^{l}\left[\frac{S[t] - S_i[t]}{2} w_i\right] - D_0$$

*The notations are same as above.

The weights w_i should be chosen such that $\sum_{i=1}^{l} w_i \leq 1$, otherwise more energy will propagate outwards then the amount contained in the node, thus resulting in negative energy.

Material and Methods:

Data source

The program was studied in great details for dummy networks. In order to check the program for real network perturbation, the paper Ripples from neighbouring transcription[14] was referred.

- The time course expression by ERK MAP kinase during cell cycle progression from G0/G1 to S phases was taken from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4739. The data consist of expression values of affymetrix ids for the time points 0h, 1h, 4h, 7h, 12h in different experiment conditions.

Removal of redundancy

The affymetrix ids were first converted to gene names using http://refdic.rcai.riken.jp/tools/xrefconv.cgi. 38519 ids out of 45695 ids were mapped to corresponding gene names. Since there can be many to one map between affymetrix ids and gene name. To make the map unique, the ids and symbols were mapped to the data which contains their chromosome locations and made unique on the basis on gene names, by taking the gene which is having largest transcript length. Now this data with gene name with chromosome location was mapped to the original time expression data. For simulations we have used only time course data for experiment1.

Network construction

From the reference paper, we have early immediate genes which were the source node for the program. We start with intra chromosomal interactions. For this the IEG gene with all the genes present on this chromosome were extracted and the sub-network from the initial genome wide interaction network was constructed. Thus, we have the network for the IEG gene of our interest.

Adjacency list construction

Next, we converted the network into adjacency list using igraph library in R package. Here, the edge weight for all the edges was kept constant. The node energy values were the expression values of genes at time 0h. The perturbed value of the IEG gene were the expression value at time 1h. The program was then simulated once the inputs were ready in required format.

Example of input data: There are two input parameters for the program; Node energy file and Adjacency list file. Following is the example of network for the chromosome 10,

where source node is nab2, whose initial value at 0h was 1244.5, which was perturbed to 4472.6 at 1 hour.

Node energy data:

Node name	Node energy
nab2	1244.5
lrp1	1580.8
stat6	2613
...	...

Adjacency list data:

nab2 lrp1:0.00033 ppp1r12a:0.00033 stat6:0.00033 zc3h10:0.00033...

where lrp1, ppp1r12a, stat6, zc3h10 etc are the nearest neighbour of nab2, the IEG gene. The edge weight is taken to be 0.00033 for all the edges, such that sum of all the weight is less than 1.

Results

With the addition of fibroblast growth factor, there is a steep increase in the expression of immediate early gene, which is accompanied by increase in expression of neighbouring genes. For example there is steep increase in the expression of IEG gene Ier3, and with time the expression of the nearest neighbour gene Nrm also increases. When the expression value of Ier3 was changed from 0.22 to 0.82 and was simulated via the model, we observe the increase in the expression value of Nrm which qualitatively agrees with the experimental results. Similar simulation pattern was observed, with steep increase in expression pattern of IEG gene Junb and Nab2. These results indicate that if we presume that our hypothesis of cross promoter interactions is true, the transcriptional ripples can be simulated using communicating-vessels model.

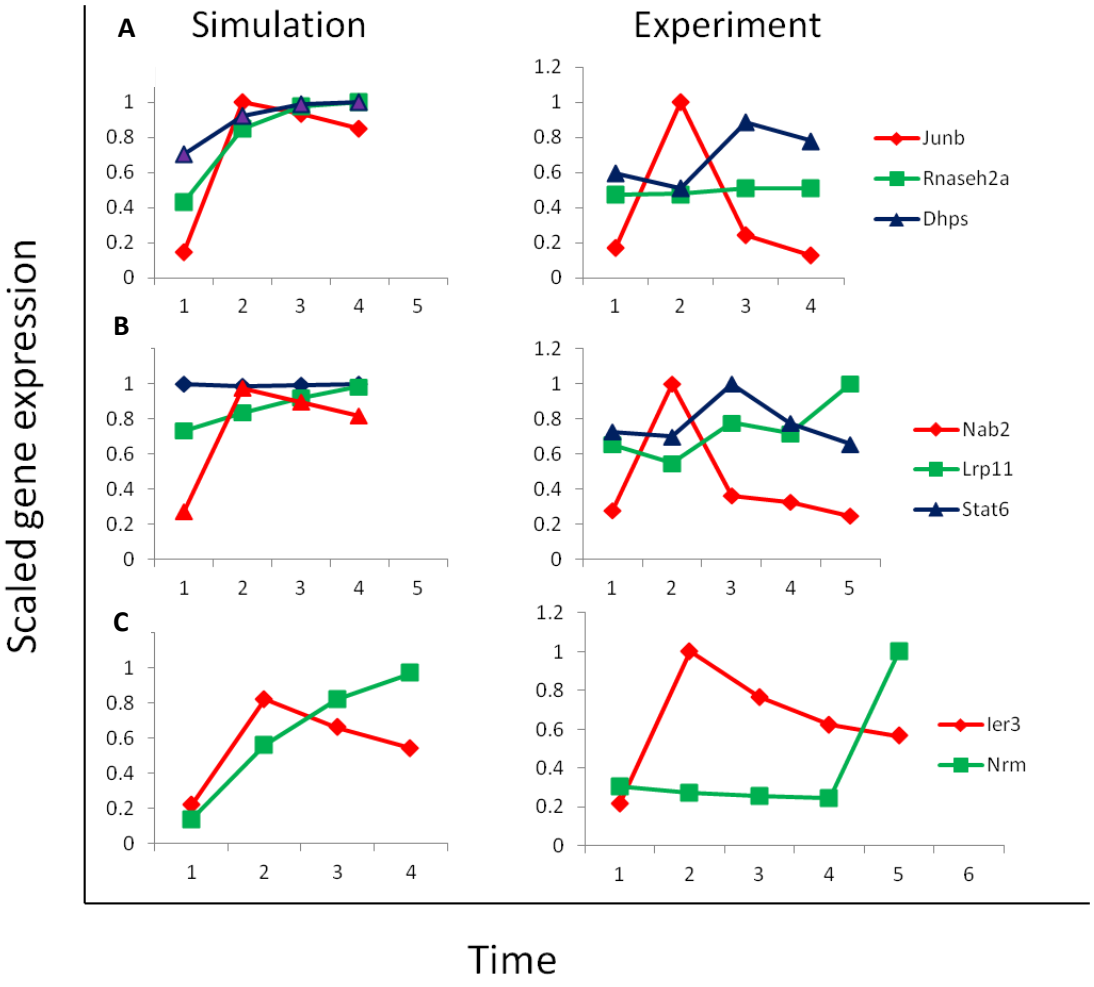


Fig2.1: Visualisation of temporal control of gene expression by (Left panel) simulations via perturbation transmission model, (Right panel) Experimental observations from literature. Here Ier3, Junb and Nab2 are the IEG genes on chromosome 17, 8 and 10 respectively.

Chapter 3 Discussion

Any living organism comprises of complex web of gene networks. An even small fluctuation in gene network is unavoidable. For example fluctuation in any of the number of factors affecting the cell growth can change the other cellular processes, which might prove lethal for the cell. For the robustness of the system, this error needs to be mitigated. Here we studied two type of fluctuations or error in living cell.

1. Stochastic variation in gene expression

This part, aimed to decipher the determinants of the transcriptional heterogeneity in the cells with the hypothesis that that transcriptional noise is modulated by relative mobility of gene loci in the three dimensional nuclear space. To test this, we started with single cell gene expression data and expression noise. The observation suggests that low noise genes are associated with high degree and high noise genes are associated with low degree. Furthermore, it was observed that low noise, genes due to their high degree are spatially proximal to each other, possibly to coordinate or synchronise their self renewal function. To further validate the observation, we studied the TAD dataset. The boundaries of topological domain are expected to be associated with transcriptional factories. Thus proximity of a gene from these boundaries can be a proxy to an association with transcriptional factories. The observation supports the hypothesis.

Are there any other factors responsible for transcriptional noise? What about degradation of mRNA? From the analysis, we have a very weak correlation between mRNA decay rate and Noise. Thus again, supporting the hypothesis.

Furthermore, we observed that genes with low noise are associated with self renewal of cells whereas the genes with high noise are associated with development.

2. Disproportionate concentrations of neighbouring gene-products

In chromatin interaction network, error at loci can create the disproportionate concentration of neighbouring genes, thus could create deregulation and comprise with the robustness of the system. The error can mitigated by transcriptional ripple. The underlying mechanism is not understood. We need to simulate and test theoretically. We thus proposed a perturbation transmission model inspired by communicating vessel

principle, so that error propagates in the neighbourhood and thus minimizing the error. To simulate the perturbation, we start with time course expression data by ERK MAP kinase during cell cycle progression from G0/G1 to S phases. We observed that the intense transcriptional activity at one locus spills over into its physical neighbouring loci, which is supports the experimental observation.

Communicating vessels model essentially captures the qualitative patterns of time-course expression curves reported in the literature.

Chapter 4 Conclusion

- ✓ Association with transcriptional factories can reduce the transcriptional noise of genes.
- ✓ Modulation of transcriptional noise might have served as an evolutionary constraint that shaped the 3D genome organisation.
- ✓ The perturbation transmission model can be used to study the Ripple Effect in transcription.
- ✓ Since networks are interdependent of each others. The model thus might suggests the propagation of genetic or epigenetic errors in one network to other interacting network

Chapter 5 Future prospective

To scrutinize our hypothesis against other structural and functional variables, we will perform Principal component regression and screen different variables which might determine transcriptional noise.

As a future perspective, we will explore for the more experimental evidence for transcriptional ripples in the genome through large scale analyses of available gene expression datasets in GEO database and test our algorithm. Once convinced, we will use this tool to study the transcriptional ripples and their genome wide functional consequences in the diseased conditions where sites of genetic or epigenetic perturbations are known. By integrating the other cellular networks like protein-protein interaction and DNA-protein interaction into the framework applying our algorithm on *interdependent networks*, we might be able to understand the, pleiotropic perturbations, if any, in the genome.

Bibliography

1. Raj Arjun, Oudenaarden Alexander van (2008) Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell, 135:216-226

2. Brian Munsky et.al (2012) Using Gene Expression Noise to Understand Gene Regulation. Science, 336:183 -187

3. Peter S. Swain et.al (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Natl Acad. Sci., 99(20): 12795–12800

4. Anita Göndör and Rolf Ohlsson (2009) Chromosome crosstalk in three dimensions. Nature, 461:212-217

5. Thesis: Dr. Klujeet Singh Sandhu, thesis, Karolinska Institutet 2010

6. Bolstad B.M. et.al (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19(2):185-93

7. Fuchou Tang et al (2010), Tracing the derivation of Embryonic Stem Cells from Inner Cell Mass by Single-Cell RNA-Seq Analysis. Cell Stem Cell, 6(5-2):468–478

8. Yubo et al (2013), Nature

9. Sharova L V et al (2009), Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells. DNA Research, 16(1):45-58

10. Mikkelsen TS (2007), Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature , 448(7153):553-60

11. Hailesellasse Sene K, Porter CJ, Palidwor G, Perez-Iratxeta C et al. (2007), Gene function in early mouse embryonic stem cell differentiation. BMC Genomics, 8:85.

12. Jesse R.Dixon et.al. (2012), Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485: 376-380

13. Kristof Z. Szalay and Peter Csermely(2013), Perturbation centrality and Turbine: a novel centrality measure obtained using a versatile network dynamics tool. PLoS ONE

14. Ebisuya Miki et al (2008), Ripples from neighbouring transcription. Nature Cell Biology, 10:1106 – 1113

Appendix

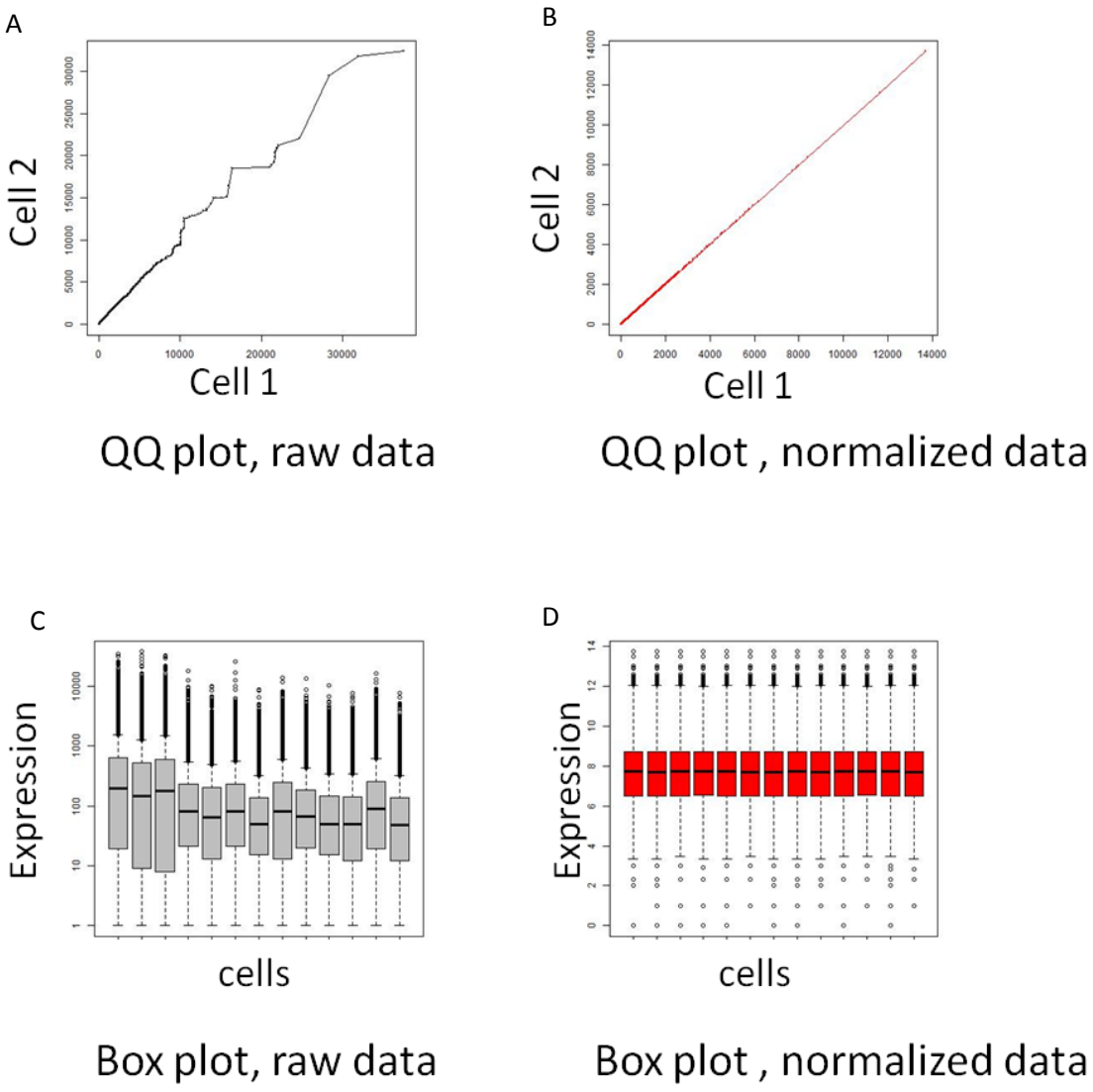


Fig S1: Data normalisation: (A) the QQ plot of raw data, (B) QQ plot of normalised data, (C) Box plot of raw data, (D) box plot of normalised data.

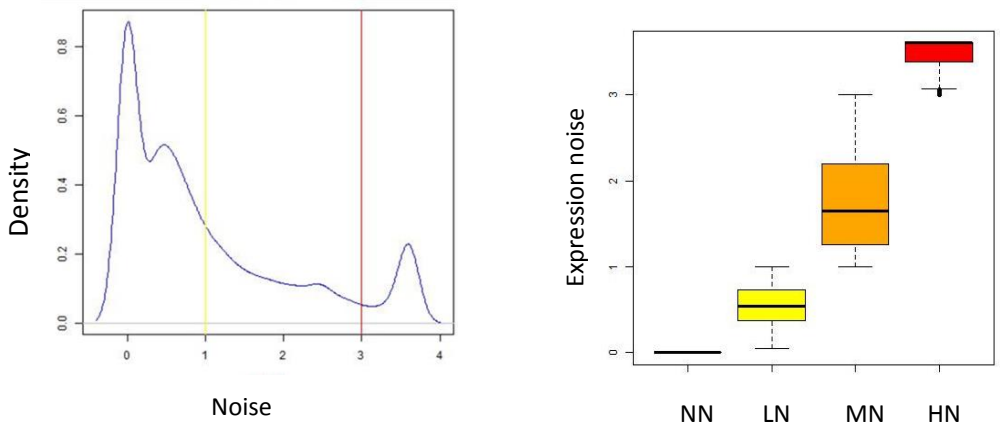


Fig S2: Noise distribution using (A) density plot showing the cut-off for noise NN = 0; 0 < LN < 1; HN > 3, (B) box plot

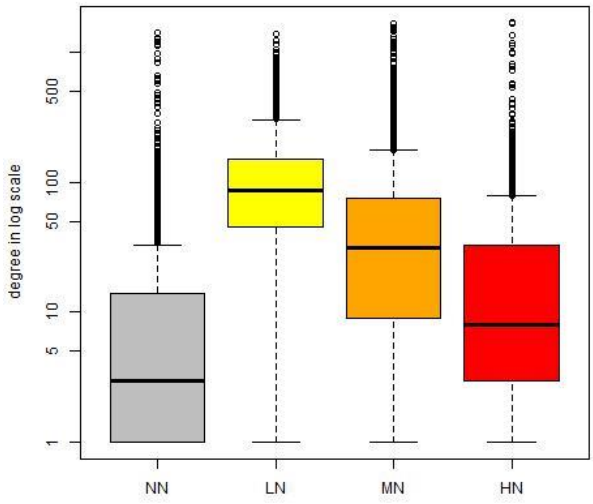


Fig S3: Degree distribution: low noise genes are associated with high degree and high noise genes are associated with low degree.

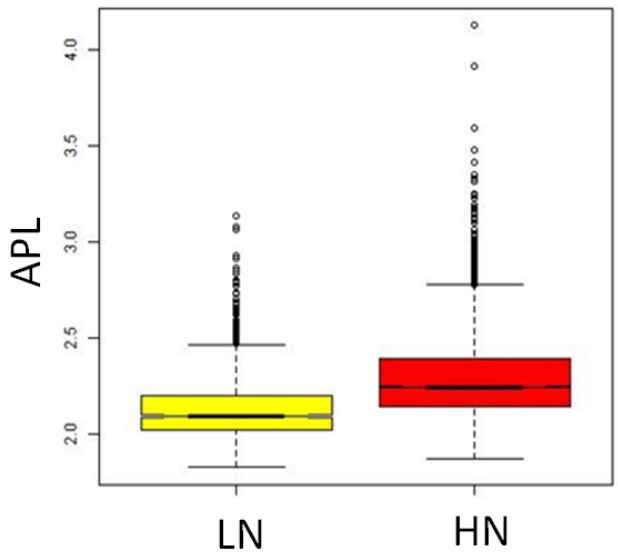


Fig S4: Average path length (APL) of genes with low and high noise.

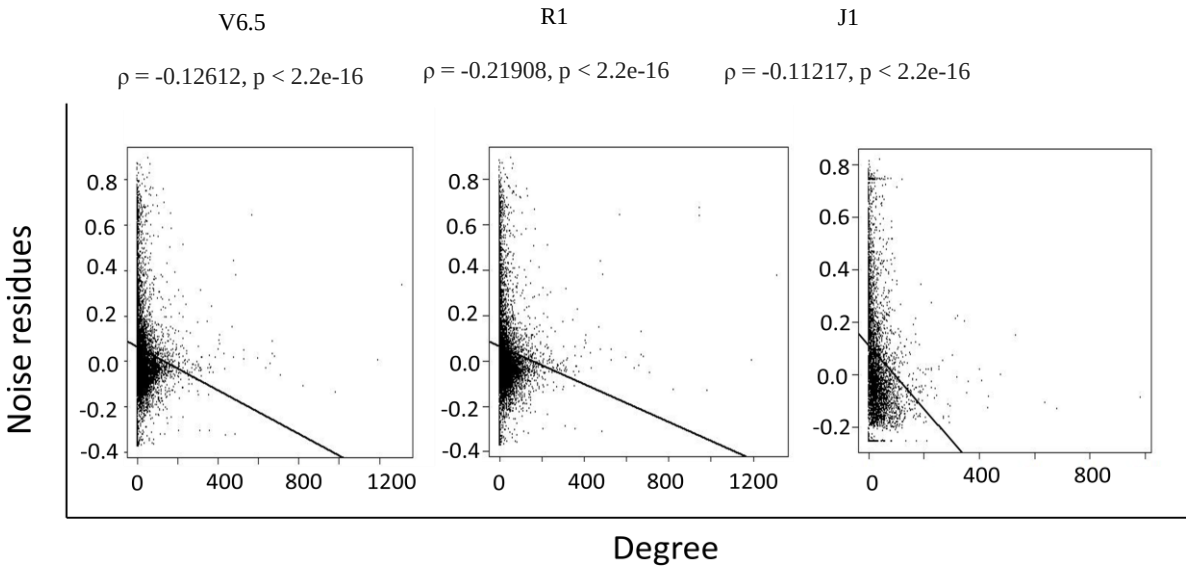


Fig S5: Correlation between abundant corrected noise and degree. The noise of genes was corrected by their abundance using LOWESS.