Summer Research: Form-Function Divergence in Induced Trisomy-7

Roy Siegelmann

Abstract

Aneuploidy is an abnormality in the number of a specific chromosome. Cells with three copies of chromosome 7, an effect called trisomy-7, display a high correlation with the presence of colorectal cancer. To date, the accepted explanation for disruption in phenotype caused by trisomy has been that of increased gene expression leading to increased protein production, which in turn alters phenotypic expression. In an effort to investigate mechanisms of trisomy-7, we conducted in-depth analyses of Hi-C (architectural) and RNASeq (functional) data, via tools of spectral analysis, network theory, and statistical measures of correlation. We demonstrate the presence of differences in genome-wide network features and a divergence of cellular form and function measures in the aneuploidic cell. Finally, we conclude that these effects are the result of global folding and supercoiling, which when combined with local upregulation that leads to a relatively insignificant impact on expression, precipitates the observed divergence. Further expansion into this field may yield the ability to identify cancerous development prior to functional and macro-scale changes, and be used in the treatment of problematic levels of expression.

1 Introduction

Common wisdom has long attributed phenotypic variation resulting from aneuploidy to variation in gene expression levels, which amplifies effects in the genetic pathway due to an abnormal amount of proteins resulting from the atypical number of genes. The addition of an entire chromosome, as in trisomy, impairs cellular homeostasis and often cripples the cell, which could be explained by the expected increased expression of up to 50%, yet this is often not the case. Numerous studies have investigated this effect, including some which specifically focus on trisomy-7, an aneuploidy which displays this lack of expected levels of expression and is particularly interesting for a number of reasons. Foremost among these is the strong correlation between cells with trisomy-7 and colorectal cancer, despite the fact that post-zygotic trisomy-7 is lethal in developing embryos. Current research in the field has focused on functional investigations of the qualification and quantification of genetic differential expression. Our study focuses on the analysis of the physical architecture of a cell affected by trisomy-7 based on comparative Hi-C data, and its comparison with global functional RNASeq data.

We began our study with this research question: What, if any, local and global changes does trisomy have on the architecture (form), and expression (function) of the cell, and how does this connect to cancer susceptibility? We hypothesized that: Trisomy leads to genome-wide supercoiling followed by upregulation in non-aneuplodic chromosomes, disrupting genomewide form-function correlation and destabilizing the cell.

Our plan for conducted the research was as follows. Using a biological chromosome conformation capture method known as **Hi-C**, we would obtain data about the physical spatial connection between the entire genome. Separate Hi-C matrices would be constructed for the healthy and trisomy-7 genomes, so that analysis could be performed properly. Each Hi-C matrix would be considered as an adjacency matrix for a network with nodes as stretches of chromosome, and edges as inter-segmental connectivity. Additionally, a genome-length vector was to be obtained for each through a method known as **RNA-Seq**, which as the name suggests, sequences the RNA transcripts throughout the genome to provide data regarding the level of activity (transcription) occurring at each point. Each of the above methods would be performed separately on a 1-megabase scale and on a 100-kilobase scale, a resolution that is ten times more in-depth, though potentially less accurate. We would then synthesize the two and analyze them using bioinformatical methods to characterize changes and correlate structural and functional differences.

Our discoveries throughout research were as follows:

1. The trisomy-7 variant of the genome displays higher connectivity, fewer and larger highly-connected components genomewide, derived from both Hi-C visual analysis and TAD-scale analysis (explained in section 4).

2. The DNA stretches have more neighbors, and thus decreased 'network importance', derived from centrality features.

3. Changes in form and function between the healthy cell and trisomy-7 variant are minor and diverge genomewide, leading to a clear decrease in form-function correlation, derived from RNASeq-Fiedler vector analyses and comparisons.

The above results suggest a mechanism wherein the genome has folded and supercoil in on itself beyond the normal levels in an attempt to reduce the expression the additional chromosome 7. Additionally, to correct the other, now-underexpressing chromosomes, the cell performs targeted upregulation on them to bring their expression to normal levels.

The remainder of this paper will be organized as follows: Sections 2 and 3 both focus on mathematical concepts used in data analysis. Section 4 includes an in-depth exploration of our results. Section 5 contains our conclusions and discussion.

2 Graphs, Laplacians, and the Fiedler Vector

A **graph** G is defined by the ordered pair $G = (V, E)$, where V is the set of **vertices** (also called nodes or points) and E is the set of **edges** connecting them. An edge is defined by the pair (v, u) of vertices it connects. In an **undirected graph** the pair of nodes is undirected, whereas in a **directed graph**, the direction of the edge is from vertex v ("the tail") to vertex u ("the head"). The **degree** of a vertex is the count of edges touching it. $|V| = n$ is the total number of vertices.

A **vector v** is a one-dimensional l-long ordered list, and both v_i and $v(i)$ denote the i-th entry of the vector. A **matrix** A is a two-dimensional ordered list with r rows and c columns. Here $A_{i,j}$ denotes the entry in the r-th row and the j-th column. Alternatively, the j-th column can be denoted as the vector $\mathbf{A_j}$.

Vectors are said to be **linearly independent** of each other if none can be written as a linear combination of the other vectors. The **rank** of a matrix is defined as the largest number of its column vectors which are linearly independent. This is equal the the largest number of row vectors which are linearly independent, so in fact $\text{rank}(A) \leq min(r, c)$.

Eigenvectors of matrix A are the vectors which satisfy the equation $A\mathbf{v} = \lambda\mathbf{v}$, i.e. when multiplying A by a vector $\mathbf{v}$ you obtain the same vector $\mathbf{v}$ only multiplied by a constant λ where the latter is the **eigenvalue** corresponding to eigenvector $\mathbf{v}$. The number of nonzero eigenvalues is at most the rank of the matrix.

A graph G can be described by an **adjacency matrix** A with the number of rows and columns equal to the number of the vertices $|V|$. A is a binary matrix; it has the value 1 in entry (i, j) is and only if the graph contains the edge (i, j); and 0 in all other entries. An adjacency matrix of an undirected graph is always symmetric since the same edge is written both as (i, j) and (j, i). For example, the adjacency matrix of graph $G = (V, E)$ with $V = a, b$ and $E = (a, b)$ would look like (a) if undirected and like (b) if directed.

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

(a) (b)

Given an adjacency matrix A, its **degree vector d** is $|V|$ long, and its entries are defined by

$$d_{\text{i}} = \sum_{j=1}^{|V|} A_{i,j}$$

. This can be quickly turned into a **degree matrix** D by multiplying the identity matrix by that vector.

$$D = I\mathbf{d}$$

The **Laplacian matrix** L, which is a very useful low-level representation of a graph G as a matrix, is defined as as

$$L = D - A$$

.

The Laplacian matrix is further normalized to a standard scale and turned into the **normalized Laplacian matrix** , which is defined as

$$\hat{L} = D^{-1/2}LD^{-1/2}$$

Note: Since D is a diagonal matrix, $D^{-1/2}$ is identical to taking each element to the -1/2 power.

Since each entry of D is equal to the sum of that row of A,

$$\mathbf{l}_i = d_i - \sum_{n=1} \mathbf{a}_n = \mathbf{0}$$

Therefore, given an n-dimensional vector $\mathbf{x}$ (where each of the above matrices are n by n) such that each element is a '1',

$$L\mathbf{x} = \mathbf{0}$$

Finally, since D is a diagonal matrix, $D^{-1/2}$ is also a diagonal matrix, which means that $D^{-1/2}\mathbf{x} = \mathbf{x}D^{-1/2}$. Therefore,

$$\hat{L}\mathbf{x} = D^{-1/2}LD^{-1/2}\mathbf{x} = D^{-1/2}L\mathbf{x}D^{-1/2} = D^{-1/2}\mathbf{0}D^{-1/2} = \mathbf{0}$$

We have just derived the fact that the normalized Laplacian matrix has $\mathbf{x}$ as an eigenvector, with an eigenvalue of 0.

Additionally, we can prove that the Laplacian (and thus also the normalized Laplacian) matrix has no negative eigenvalues, meaning it is **positive semidefinite**.

We can rewrite the Laplacian as

$$L = \sum_{(i,j)\in E} (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T$$

where $\mathbf{e}_i$ denotes the i-th standard basis vector (i.e. a vector with 0's everywhere except the i-th entry, which has a '1') and E denotes the set of all edges within G, the graph from which the matrix Laplacian is derived.

The undirected Laplacian is clearly a symmetric matrix (since if $(a, b) \in E, (b, a) \in E$), and a theorem states that for symmetric matrices, the following three statements are equivalent:

(i) A is positive semidefinite.

(ii) A = VV^T for some matrix V.

(iii) A has all non-negative eigenvalues.

We observe that if the matrices A and B are both positive semidefinite, so is (A + B), since $\mathbf{x}^T(A+B)\mathbf{x} = \mathbf{x}^T Ax + \mathbf{x}^T Bx \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.

By (ii), $(\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T$ is positive semidefinite. By summing up all these terms we receive L, so based on the observation above we can say the Laplacian matrix is positive semidefinite, and thus has no negative eigenvalues.

Therefore, since the Laplacian has no negative eigenvalues, we can find the smallest nonzero eigenvalue. This is called the **Fiedler number** (also known as the **algebraic connectivity** of the graph G), and its corresponding eigenvector is called the **Fiedler vector**.

A few facts about the Fielder value and vector - the Fielder value is greater than 0 if and only if G is a connected graph, as a corollary of the fact that the number of times 0 appears as an eigenvalue in the Laplacian is the number of connected components in the graph. The magnitude of the Fielder value reflects how well connected the overall graph is, meaning that a high Fielder value denotes a high level of connectivity, while a low one denotes the opposite.

3 Measures of Centrality and Principal Component Analysis

When discussing networks and the graphs they generate, it becomes incumbent upon us to identify a level of how centralized the graph is (called, aptly, **centralization**) and the most "central" or "important" node, i.e. that with the highest degree of **centrality**. Centrality is often standardized, as this gives us a measure of the relative centrality of different nodes. Given c_i is the centrality of the node i and c_{max} is the maximum centrality that can exist with this sort of centrality, that node's **standard centrality** is denoted by $c_i = c_i/c_{max}$.

There are four primary types of centrality, being degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality.

Degree centrality is defined as the degree (number of connections) of each node. In an undirected graph, a vertex's degree centrality is simply the number of edges it is part of, so that $c_i^d = \sum_{j:j\neq i} e_{i,j}$, where $e_{i,j}$ denotes an edge between vertex i and vertex j, being '1' if such an edge exists and '0' if one does not. The standardized degree centrality is quite simply $c_i^d/(n-1)$. The **degree centralization** of a graph, $C^d(G)$ is equal to the sum of the difference of each vertex's centrality and that of the maximum centrality found within this network (denoted c^*), divided by the maximum value that this sum could potentially be for a graph (which is $(n-1)(n-2)$, where n is the number of vertices, for a star graph). As such,

$$C^d(G) = \frac{\sum_i (c^{d*} - c_i^d)}{(n-1)(n-2)}$$

Closeness centrality is defined as the inverse of the average geodisic distance from the particular node to all other nodes. $d_{i,j}$ is the geodisic distance between nodes i and j, i.e. the minimum path length from i to j. Therefore,

$$c_i^c = \frac{1}{\sum_{j:j\neq i} d_{i,j}} = \frac{1}{(n-1)\bar{d}_i}$$

where $\bar{d}_i$ is the arithmetic mean of the geodisic distance between node i and all other nodes. As such, the standard closeness centrality can be defined as $\frac{1}{\bar{d}_i}$. Similarly to degree centralization, **closeness centralization** is defined as

$$C^c(G) = \frac{\sum_i (c^{c*} - c_i^c)}{(n-1)(n-2)}$$

Betweenness centrality involves considering a node as a bridge between other nodes, so that the vertex with highest betweenness centrality will be part of the shortest path between the highest number of other nodes (involved in the most **geodisics**, as it was). As such,

$$c_i^b = \sum_{j<k} \frac{g_{j,k}(i)}{g_{j,k}}$$

where $g_{j,k}$ is the number of geodisics (paths with the shortest distance) between nodes j and k, and $g_{j,k}(i)$ is the number of such geodisics that pass through node i. The standardized betweenness centrality is defined in a complicated manner that ultimately can be reduced to

$$\overline{c_i^b} = \frac{2c_i^b}{(n-1)(n-2)}$$

In a similar manner, the betweenness centralization of a graph can be defined as

$$C^b(G) = \frac{2\sum_i (c^{b*} - c_i^b)}{(n-1)^2(n-2)}$$

Finally, the basic idea of **eigenvector centrality** is that a central node is connected to other central nodes, so that the centrality of each node is proportional to the sum of the centralities of its neighbors. Given the adjacency matrix A and vector c^e, where c_i^e is the eigenvector centrality of node i, we can define

$Ac^e = \lambda c^e$, taking c^e to be the eigenvector corresponding to the highest possible eigenvalue λ. Therefore, we can finally define the eigenvector centrality of a node as

$$c_i^e = \frac{1}{\lambda} \sum_{j:j \neq i} e_{i,j} c_j^e$$

Beyond centrality, another important way of finding important information about graphs is through **principal component analysis (PCA)**. The basic idea of PCA is a desire to break matrices down into vectors for a variety of single-dimensional uses. In actuality, PCA breaks down matrices into a series of vectors equal in number to the rank of the matrix, ranked by importance in composition of the matrix, with subsequent vectors having exponentially less importance to the composition than those which precede them. As a result, only the first vector (or maybe two) are actually used, and this enables operations to be performed on "matrices" which can normally be performed only on vectors.

To begin performing PCA, we take an m x n matrix (which we will call A). Each row of A is then summed up to create an n-dimensional vector called the **mean vector**, which is denoted by the Greek letter μ. Following this, we can subtract the mean vector from each column of A to create the **centered matrix** of A, denoted A_{cent}, which in the case of A being a square matrix returns a symmetric and positive semi-definite matrix. Then, the **covariance matrix** (which we will call C) is computer for A_{cent} (which is equivalent to $\frac{1}{n-1} A^T A$). Finally, we find the eigenvectors of C, and sort them in order of decreasing corresponding eigenvalues. The first eigenvector is then the **first principal component**, which is the vector which has the most bearing on the identity of the matrix, with the next eigenvector being the second principal component, and so on and so forth.

4 Research Summary and Discoveries

The following investigation was performed on cells with **induced trisomy-7**, which is the introduction of a third copy of chromosome 7 into a cell[1]:

Figure 1 contains the Hi-C matrix of the entire genome in 1-megabase resolution, in the healthy cell on the left, the trisomy-7 cell in the center, and the difference between them on the right. The matrix is symmetrical and the i-th row refers to the same chromosomal segment as the i-th column. As such, the main diagonal contains the connections of a 1-megabase segment with itself, which is why the main diagonal is so bright (brighter color meaning higher connectivity). Along this diagonal are small, bright squares, representing the chromosomes (which clearly have far higher intra-connectivity than inter-connectivity). Continuing to Figure 2a-b, we focus in on some individual chromosomes (7 and 4), and observe their individual Hi-C matrices. We still see these squares around the main diagonal. These squares represent **topologically associating domains (TADs)**[2], which are segments of the genome that form structures that primarily interact between themselves. Part c-d of the same figure contain data about the number of TADs and mean size of TADs for all chromosomes, which we see the former decrease and latter increase between the healthy and aberrant cells.

Figure 3a-b contains plots of each type of centrality for the selected chromosomes, along with comparisons (in magenta on the right) of the healthy cell (in blue on the left) and trisomy variant (in red in the center). Parts c-d of this figure have the norms and means of changes in each kind of centrality by chromosome respectively.

Figure 4a-b graphs structural and functional information collected about the same selected chromosomes. Graphs on the left (in blue) are for the healthy chromosome, in the center (in red) are for the trisomy-7 chromosome, and on the right (in magenta) compare the two. The first row contains graphs that display log-scaled and standardized ($log_2(R + 0.5) + 1$) where R is the RNASeq vector) data about RNASeq, a measure of the activity level throughout the chromosome. The second row contains the Fiedler vector, graphed along the length of the chromosome (which has biological meaning, as will be explained shortly). The final entry in this row contains a comparison in the Fiedler vectors, with a value of '1' denoting a change from negative in the healthy chromosome to positive in the trisomy-7 variant, a value of '-1' denoting the reverse, and '0' being no change in the sign of the Fiedler vector at that point.

In the literature[3], it has been established that there is a strong correlation between the RNASeq vector and the the first principal component, which is expanded to the Fiedler vector due to the near-perfect

correlation between PC1 and the Fiedler vector [2]. The areas where the Fiedler vector is positive generally correspond to **euchromatin**, the more loosely-packed and highly transcribed area of the DNA, which has high RNASeq values (also known as the **A compartment**). Similarly, the negative areas correspond to the **B compartment**, which contains **heterochromatin**, the highly packed and infrequently-transcribed area of the DNA with low RNAseq values. The correlation data between each mathematical Hi-C vector component and the RNASeq can be seen in Figure 4c in both the 100kb and 1mb resolutions.

As seen, there is some sort of chromosome-wide effect from adding an extra chromosome 7, one which cannot be clearly explained by the status quo in biological thought. First and foremost, Figures 1 and 2 display clearly that there is a difference in the TAD block-structure throughout the entire genome in a positive direction in the Hi-C data, i.e. additional interactions between genomic loci throughout the entire genome. This is further enhanced by the TAD quantification and qualification seen in Figure 2c-d, which show a similar structure in the TAD connectivity patterning (the TADs are no less or more connected by topological region in one than the other), but there are much fewer, and thus also larger, TADs. However, as seen throughout Figure 3, there is relatively little difference in the expression of the genes (RNASeq values), even in chromosome 7, which has an additional copy. Many of the changes seen in RNASeq are not seen in the Fiedler vector, which has different, yet equally minor differences.

This phenomenon is well-captured by Figure 4c, which shows that the correlation between the expression of the healthy and trisomy-7 variants of the genome with their corresponding Fiedler vectors and first principal components are different. There is a much lower correlation between the latter (which are visualizations of the form, or structure, of the genome) and the RNASeq (the function) in cells with trisomy-7. This form-function divergence is furthermore supported by the data in Figure 4a-b, the centralities. Between the healthy cells and the trisomy-7 variant, there is little difference in the closeness and eigenvector centralities (although the former is ever-so-slightly higher for reasons that will soon be apparent). However, in the variant, there is significantly higher degree centrality and lower betweenness centrality.

5 Conclusions and Discussion

Our results suggest a mechanism wherein the genome has folded and supercoiled beyond the normal levels in an attempt to reduce the expression of the aberrant chromosome 7. This mechanism is likely imprecise, which explains why there are less TADs (and thus larger highly-connected areas) and the fact that all of aforementioned effects take place in all chromosomes, as opposed to just in chromosome 7, which would be the case were this mechanism targeted. This likely works well regarding the expression of chromosome 7 (which is why the RNASeq values in the trisomy-7 variant are quite close to those in the healthy cells), but the other chromosomes should not have decreased expression, which we would expect to see from this supercoiling. Therefore, in an effort to maintain proper levels of transcription, there must be local upregulation via an undetermined mechanism on a lower scale *within* each chromosome, leading to our observation of relatively normal RNASeq genome-wide. Thus, with global differences in form and folding, yet similar function and expression, the divergence of form and function seen in trisomy-7 is explained.

These conclusions also supports the observations we have made regarding centrality measures. Closeness centrality is a measure of how close the individual node is to the remainder of the system; it changes little in our discoveries, due to the main distance being from far-away genes that have not come closer. Our closeness centrality becomes just slightly higher due to the local folding making the already-near nodes even more so. The eigenvector centrality displays little difference due to the fact that 'relative connectedness' of neighbors (i.e. the global structure of the network) will not change by supercoiling. The large change in betweenness and degree centralities can also be explained through graph theory. With more highly connected coiled segments, each node is connected to many more local nodes, greatly increased degree centrality (simply a measure of how many nodes the target is connected to). Additionally, the highly connected graph has many paths to pass through a specific connected segment (geodisics), decreasing the proportion of paths that will pass through an individual node, and thus the betweenness centrality.

The divergences we have observed in form and function, caused by a number of disparate methods of up- and downregulation with the goal of returning homeostasis to the cell, are accompanied by an unintended consequence. Both expression and architecture, while averaging to relatively normal levels, are as a whole erratic and unstable, certain areas being slightly too high and other areas being too low for one metric or the

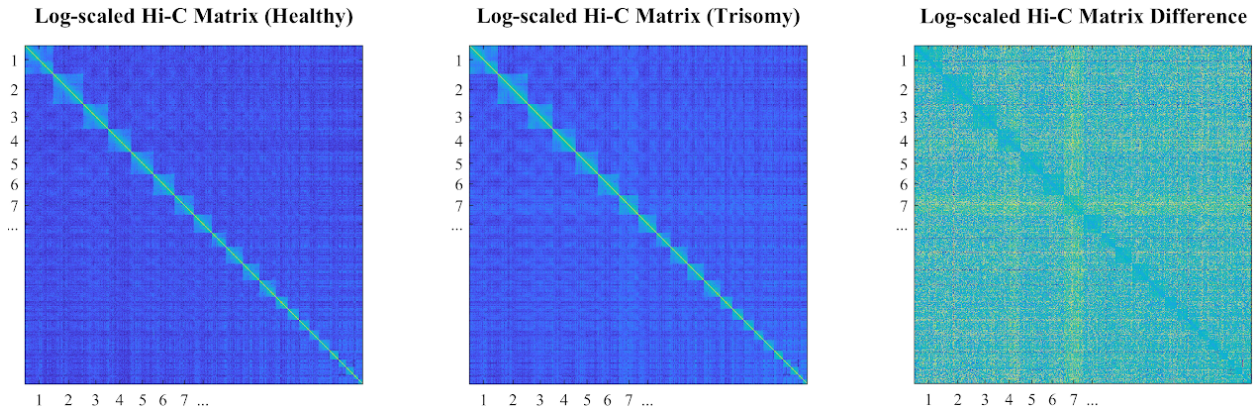


Figure 1: Genome-Wide Hi-C matrix, scaled logarithmically base 2

other. This leaves the cell busy and spent, striving to return to homeostasis and live at any cost, lowering its defenses and opening the door to cancer-like behaviors. Additionally, gene-level analysis has revealed an underexpression of genes related to aging, apoptosis, and regulation of development, and overexpression of genes related to cell growth, prolificity, and oncogenes, all markers of pre-carcinomatic cells.

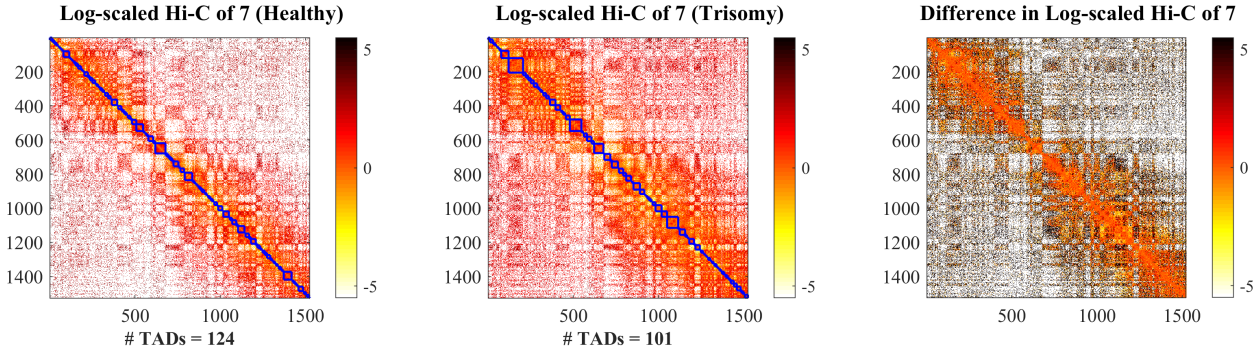
Our discoveries may have far-reaching implications on the world of medicine. First and foremost, we have found a method that can be used to quickly determine whether a specific area in a person's body is at high-risk to cancer, via analysis of the form-function correlation within cells. This can be used in regular screenings and to enhance anti-cancer methods of preventative medicine. Additionally, we are now aware of the reason for the relation between aneuploidies and susceptibility to cancer, which opens the door to future research on this topic. Other potential research includes the reason for the viability of trisomy-13, 18, and 21 embryos, the mechanism used within the cell for targeted upregulation, and the expansion of this study to other aneuploidies for comparison.

Bibliography

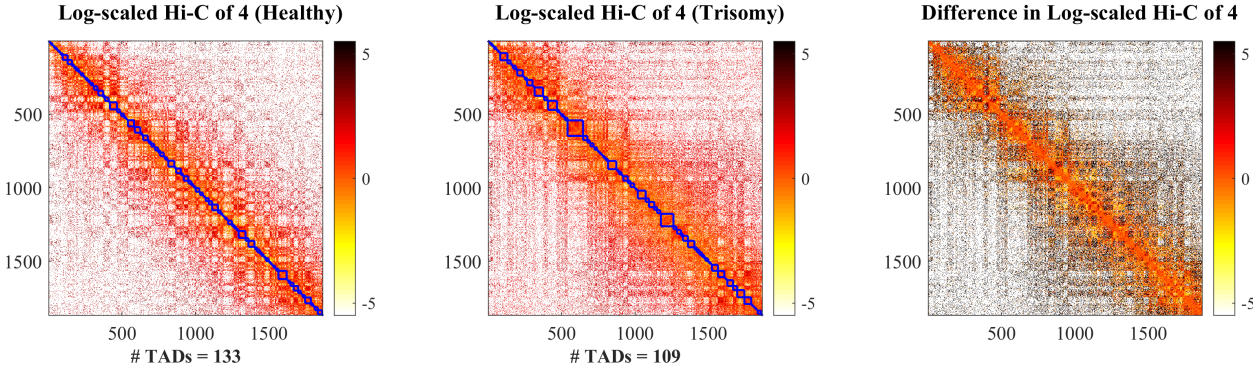
[1] Rüdiger Braun et al. "Single chromosome aneuploidy induces genome-wide perturbation of nuclear organization and gene expression". In: *Neoplasia* 21.4 (2019), pp. 401–412.

[2] Jie Chen, Alfred O Hero III, and Indika Rajapakse. "Spectral identification of topological domains". In: *Bioinformatics* 32.14 (2016), pp. 2151–2158.

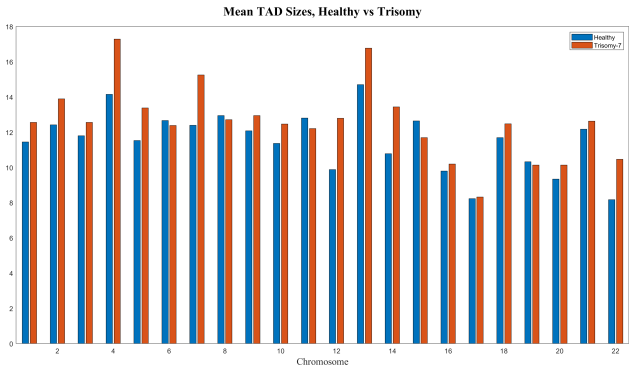
[3] Erez Lieberman-Aiden et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". In: *science* 326.5950 (2009), pp. 289–293.



(a) Chromosome 7 Hi-C matrix, scaled logarithmically base 2



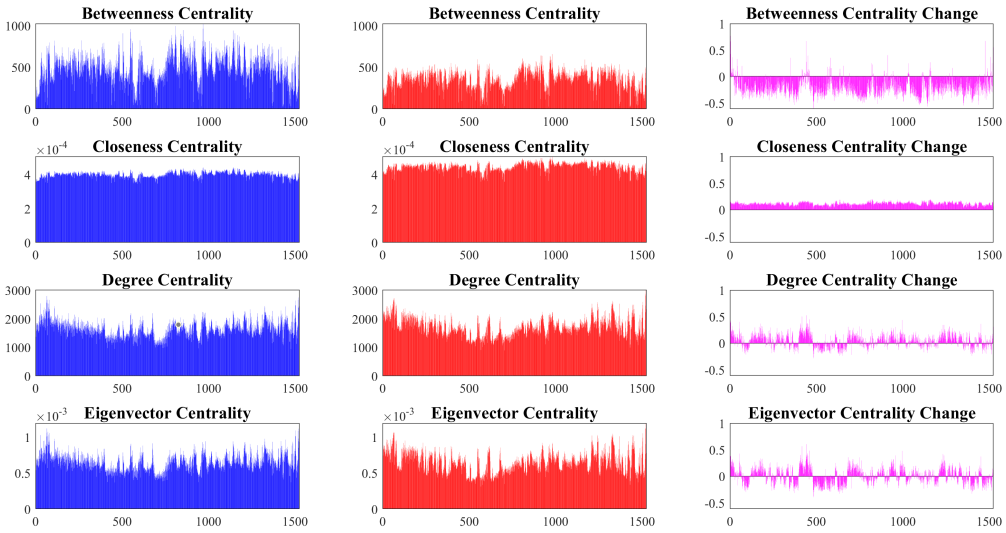
(b) Chromosome 4 Hi-C matrix, scaled logarithmically base 2



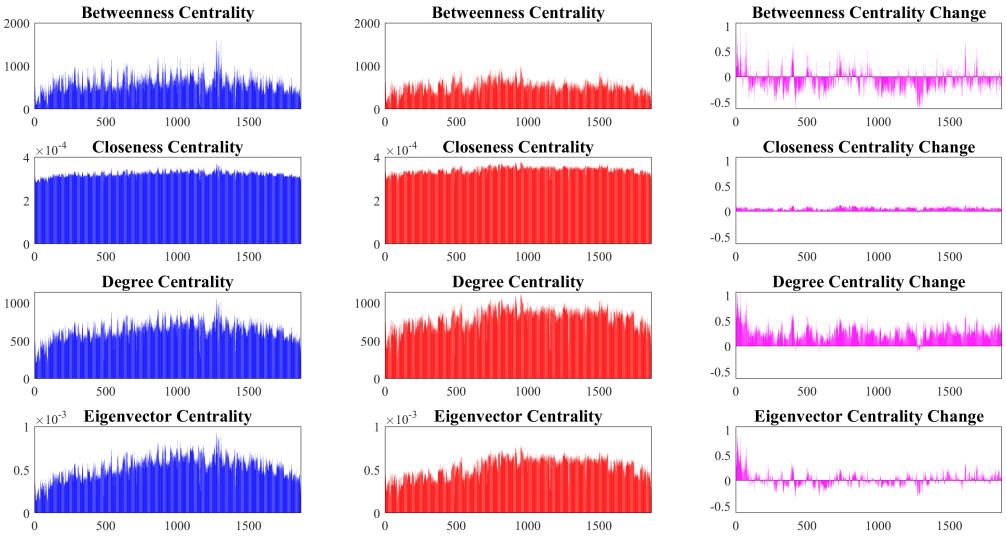
(c) Comparison of mean TAD sizes

(d) Comparison of number of TADs

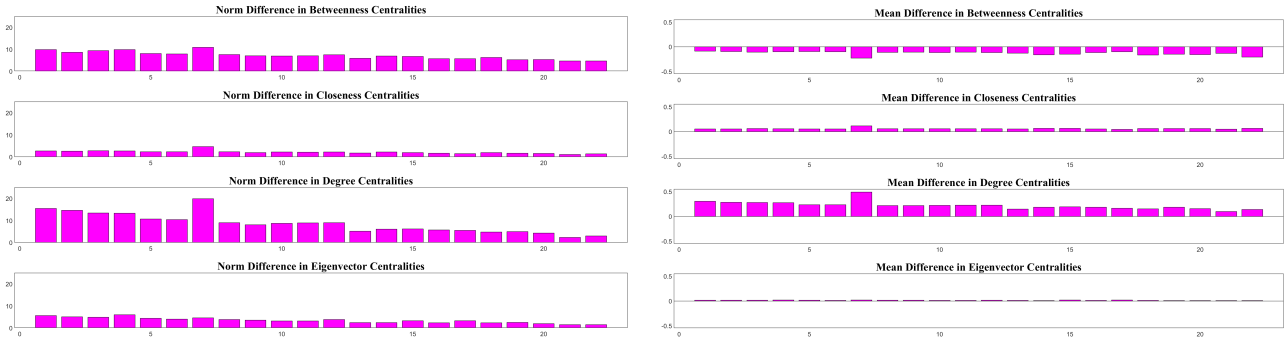
Figure 2



(a) Chromosome 7 Centralities



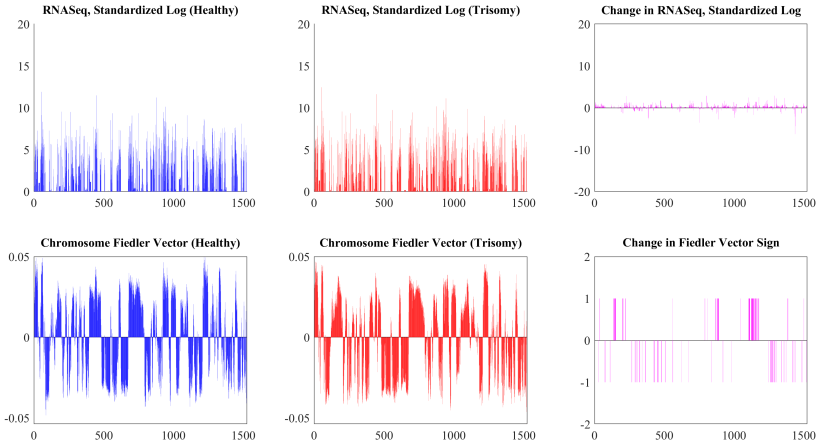
(b) Chromosome 4 Centralities



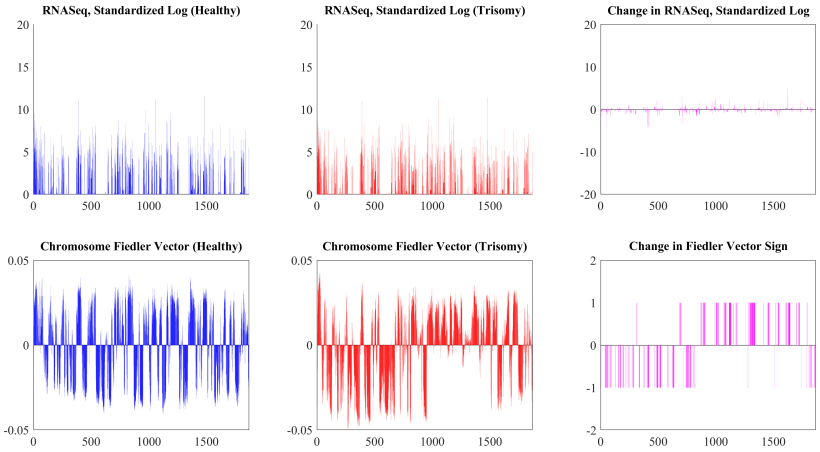
(c) Norm centrality differences

(d) Mean centrality differences

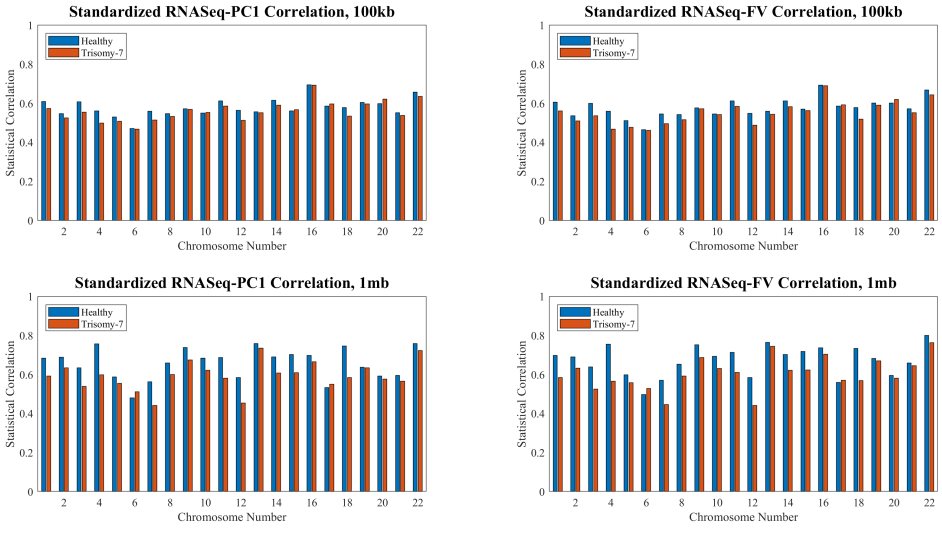
Figure 3



(a) Chromosome 7 Fiedler Vector and RNASeq



(b) Chromosome 4 Fiedler Vector and RNASeq



(c) Form-Function Correlation

Figure 4