A pathway-centric view of spatial proximity in the 3D nucleome across cell lines

Hiren Karathia[1], Carl Kingsford[2], Michelle Girvan[3], Sridhar Hannenhalli[1*]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD

[2]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA

[3]Department of Physics, University of Maryland, College Park, MD

Authors' emails:

Hiren Karathia – hiren@umiacs.umd.edu

Carl Kingsford – carlk@cs.cmu.edu

Michelle Girvan - mgirvan@gmail.com

[*]Corresponding author

Sridhar Hannenhalli

3104G Biomolecular Sciences Building (#296)

University of Maryland, College Park, MD 20742, USA

301 405 8219 (v) 301 314 1341 (f)

sridhar@umiacs.umd.edu

Key words: Hi-C, Chromatin structure, Pathways, Transcriptional regulation

Abstract

Spatial organization of the genome is critical for condition-specific gene expression. Previous studies have shown that functionally related genes tend to be spatially proximal. However, these studies have not been extended to multiple human cell types, and the extent to which context-specific spatial proximity of a pathway is related to its context-specific activity is not known. We report the first pathway-centric analyses of spatial proximity in six human cell lines. We find that spatial proximity of genes in a pathway tends to be context-specific, in a manner consistent with the pathway's context-specific expression and function; housekeeping genes are ubiquitously proximal to each other, and cancer-related pathways such as p53 signaling are uniquely proximal in hESC. Intriguingly, we find a correlation between the spatial proximity of genes and interactions of their protein products, even after accounting for the propensity of co-pathway proteins to interact. Related pathways are also often spatially proximal to one another, and housekeeping genes tend to be proximal to several other pathways suggesting their coordinating role. Further, the spatially proximal genes in a pathway tend to be the drivers of the pathway activity and are enriched for transcription, splicing and transport functions. Overall, our analyses reveal a pathway-centric organization of the 3D nucleome whereby functionally related and interacting genes, particularly the initial drivers of pathway activity, but also genes across multiple related pathways, are in spatial proximity in a context-specific way. Our results provide further insights into the role of differential spatial organization in cell type-specific pathway activity.

Introduction

Recent advances in Chromosome Confirmation Capture (3C) and its high throughput derivative, Hi-C, have enabled genome-wide identification of spatially proximal genomic regions [1-3]. Comparative analysis of Hi-C data across cell lines and species reveals a conserved framework of 3D architecture, represented by topologically associating domains (TADs) and further context-specific variation in distal interactions [4].

Among other things, these 3D maps of chromosomes help explain, in part, spatio-temporal regulation of gene expression by distal enhancers, aided by long-range DNA looping [5-7]. Similarly, previous studies have shown that groups of spatially clustered enhancers exhibit co-activity across cell types and this co-activity is reflected in co-expression of proximal genes, which are often functionally related [8, 9]. More specifically, genes involved in the same pathway have been shown to be spatially proximal in *Saccharomyces cerevisiae* [10, 11], *Plasmodium falciparum* [12] and *Homo sapiens* lymphoblastoid cell lines [13] . However, these previous studies have not been extended to multiple human cell lines, and it is not clear to what extent spatio-temporal activity of pathways is related to the spatial proximity of the constituent genes. More generally, the broader characterization of physically proximity of genes in the context of functional pathways is missing and could reveal organizing principles underlying spatial proximity of pathway genes as they relate to pathway activity.

In this work, we perform a comparative pathway-centric analysis of Hi-C-derived spatial proximity data in 6 ENCODE [14] cell types - *HEK293 [15]*, *hESC [4]*, *IMR90*, *BT483[16]*, *GM06990[16]*, and *RWPE1[17]*, each with replicate data. Our analysis of two large sets of pathways – KEGG [18], and NetPath [19] reveals several properties of spatial proximity of pathway genes: We find that in general, genes in a pathway tend to be spatially proximal and this tendency is even greater for gene pairs that belong to multiple pathways. Our expression analysis shows that genes that are co-localized in nuclear space with other genes have higher expression, and this effect is especially prominent when they are proximal to a gene in the same pathway. We also found that spatial proximity of pathway genes is strongly correlated with cell type-specific pathway activity. As an expected corollary, housekeeping genes, by virtue of being ubiquitously active, exhibit ubiquitous spatial proximity. Surprisingly though, we found that the protein products of spatially proximal genes in a pathway have a significantly greater tendency to physically interact than various controls. Functional enrichment analysis suggests that spatially proximal pathway genes are enriched for specific functional classes such as transcription factor and transmembrane genes, and they occupy

higher levels in the regulatory hierarchy. Finally, we look at higher-level spatial organization of functional pathways by quantifying spatial proximity for all pairs of pathways. Using this, we identify a network of spatially proximal pathways that is consistent with their functional roles.

Overall, this first comprehensive pathway-centric analysis of spatial proximity in multiple human cell lines shows a strong link between spatial proximity and context-specific gene expression and pathway activity. Our analysis also reveals surprising links between spatial proximity and interaction between the corresponding protein products. Functional analysis of proximal genes within pathways strongly suggests a regulatory hierarchical bias in physical proximity of pathway genes. Taken together, these results are consistent with a mechanism in which early regulatory components of a pathway are brought into spatial proximity in a condition specific manner.

Results

1. Software pipeline for Hi-C processing - overview

Fig. 1 shows the overall pipeline that, starting from the raw reads obtained from a Hi-C experiment, produces significant pair-wise gene interactions. Details of the pipeline are provided in the Methods section, and we highlight a few pertinent features here. The pipeline allows the user to select a resolution at which significant interactions are identified. We performed our analysis at 100 kb resolution because we are interested in gene-centric interactions and 100 kb is expected to cover ~1 gene; smaller resolution yields fewer significant interactions due to loss in power and larger resolution results in ambiguous gene-segment mapping. We have further discussed the choice of the resolution later in the discussion section. The pipeline uses the normalization step of the Homer tool [20] to control for the genomic distance-dependent features of Hi-C counts (that proximal genomic regions are more likely to interact). Using this pipeline, we processed 6 sets of pooled replicates for 6 ENCODE cell lines - HEK293, hESC, IMR90, BT483, GM06990, and RWPE1. Table 1 shows the data obtained for the 6 cell types at the default interaction p-value threshold of 0.001 and FDR <= 0.1.

Assessment of spatial proximity of pathway genes

We quantify cell-line-specific spatial proximity of genes in a pathway using *edge fraction (EF)*, which is essentially the fraction of all possible gene pairs in the pathway that are spatially proximal. This measure was previously shown to be effective [10] . We then quantify

significance of *EF* based on a sampling procedure (see Materials & methods), obtaining a Z-score and the corresponding p-value and multiple testing corrected q-value; a higher Z-score is indicative of spatial proximity of the pathway genes above expectation. We studied 164 KEGG pathways [21] with at least 10 genes and estimated their spatial proximity Z-scores in 6 cell lines. Fig. 2 shows that overall biological pathways tend to be spatially proximal, consistent with previous reports [10]. Interestingly, as shown in Fig. 2, we found that spatial proximity for subsets of genes shared between two pathways is even greater, suggesting that such genes, coordinating multiple pathways, may be under a greater constraint to be spatial proximal.

Fig. 3 shows the cell-type-specific Z-scores for a representative set of pathways and Supplementary Fig. 1 shows the same for all pathways. Consistent with the fact that the KEGG database is dominated by essential and broad cellular processes, we found that spatially proximity of KEGG pathways are not only generally high (Fig. 2), but a large fraction of pathways exhibit a significant level of spatial proximity in many cell types (Fig. 4). In particular, given the ubiquitous expression and function of housekeeping genes, we tested whether these genes tend to be ubiquitously spatially proximal or whether their ubiquitous expression is decoupled from their spatial proximity to one other. Based on 3800 housekeeping genes [22], we found that housekeeping genes exhibit significant spatial proximity to other housekeeping genes in 5 out of 6 cell lines tested (Fig. 3). Other ubiquitously proximal pathways include *Metabolism of xenobiotics by cytochrome P450*, *1- and 2-Methylnephthalene degradation* and *Gamma-hexachlorocyclohexane degradation*, all involved in drug metabolisms in animals.

Several cases of cell-type-restricted spatial proximity are worth noting. For instance, 'Cell cycle' genes are expected to be active in pluripotent stem cells. This pathway is significantly proximal in only two cell types, one of which is the human embryonic stem cell (hESC). Cytokine-mediated signaling is critical in immune response and consistently, Cytokine-cytokine receptor interactions are uniquely proximal in immune B-cell (GM06990); B-cell-specific proximity of cytokines CCL23 and CCL4 is consistent with their known role in increased monocyte recruitment during inflammation [23]. Likewise we found the *Androgen and estrogen metabolism* pathways to be proximal in breast cancer cell lines, where the role of this pathway is well known [24]. Interestingly, we found the *Androgen-Estrogen receptor* pathway to be proximal in Kidney cell line as well (Z-score = 6.5, FDR = 0.12), consistent with the role of this pathway in glucuronidation activity that involves communication between thyroid and kidney [25]. Unexpectedly, we see proximity of '*Type-II diabetic mellitus*' in lung

fibroblast-derived IMR90. This is however consistent with recently observed connections between diabetes and lung functions [26]. Finally, one the only two cell lines where Melanogenesis genes is found to be proximal are the prostate epithelial RWEP1 and mammary epithelial BT483; melanogenesis in prostate epithelial cells has been previously reported [27].

As an additional set of pathways, we examined a set of annotated cancer-related pathways from NetPath [19]. As shown in Supplementary Fig. 2, genes in cancer-related pathways exhibit much more subdued spatial proximity patterns in cell lines not derived from primary tumors. Another noticeable trend is that many pathways exhibit high intra-pathway gene proximity in hESC, consistent with previously noted similarities between cancer and stem cell processes [28]. We found that *IL3-signaling* pathway is uniquely proximal in hESC consistent with its known role in cell cycle control [29].

As a useful resource, in Supplementary Table 1 we have provided, for all pathways considered, the specific gene pairs that were proximal in different cell lines.

2. Spatial proximity and gene expression

Next, we assessed whether spatial proximity of genes is correlated with their expression levels. Furthermore, we also assessed among the spatially proximal genes whether belonging to same pathway has any association with expression level. This analysis was done in all 6 cell lines using RNA-seq data available in GEO (see Materials & Methods). We compared cell type specific expression levels for three disjoint groups of genes (Fig. 5). The first group consisted of genes that are proximal to another gene in the same pathway (*proximal-intra-pathway*). The second group consisted of genes proximal to another gene but excluding *proximal-intra-pathway* genes; this group was designed to assess whether shared pathway membership impinges on gene expression. The last group consisted of all other genes not proximal to any gene (*non-proximal-generic*). Fig. 6 shows that in pooled result from all 6 tissues, while genes that are spatially proximal to other genes have a greater expression than non-proximal genes, the expression is greater for genes that are proximal to a gene in the same pathway. The results for individual cell lines are qualitatively similar and are shown in Supplementary Fig. 3.

We directly assessed the correlation between cell type specific pathway proximity and pathway activity. We used two measures to approximate pathway activity (see Materials & Methods).

The first measure captured the mean expression of all genes in the pathway and the second measure captured the ratio of mean expressions of proximal and non-proximal genes in the pathway. Each measure was converted into a based on random sampling (see Materials & Methods). As shown in Fig. 7, spatial proximity is highly correlated with mean pathway expression (Spearman rho = 0.77, p = 5.9e-62). Note that this analysis does not rely on any Z-score cutoff.

Our observed highly positive correlation between spatial proximity of a pathway genes and the pathway activity would imply that the pathway genes are localized in active compartments and not in repressive or inactive compartments (in which case spatial proximity would result in pathway suppression – an opposite effect). A previous paper [48] has shown that in GM06990, relative to the inactive (B) compartment, the active (A) compartment (1) is highly enriched for genes, (2) has higher expression, (3) has more accessible chromatin, and (4) is loosely packed, i.e., has fewer interactions. To assess our hypothesis that interacting pathway genes are preferentially in A-compartment, we estimated the compartmentalization in GM06990 cell line using HOMER tool and found 18256 (75%) genes in A-compartment and 6371 (25%) genes in B-compartment, consistent with previous report. Next, we compared for 4 groups of genes (defined in Fig. 5) the tendency to belong to A-compartment. As shown in Supplementary Fig. 4 there is a robust monotonic trend whereby the genes proximal to another gene in the same pathway have the greatest tendency to belong to A-compartment and both spatial proximity and co-pathway membership contribute to this tendency. Specifically there a large difference between the first class (proximal-intra-pathway) and the non-proximal genes (p-value = 1.6E-53, Odds-ratio = 6.3). Thus, it seems that our observed positive correlation between pathway proximity and activity is due to the fact that most within-pathway proximal genes are in A-compartment.

3. Spatial proximity and protein-protein interaction

We have found that genes in a pathway tend to be spatially proximal. Previous studies have shown that the proteins in a pathway have a greater tendency to physically interact with each other [30]. We therefore directly assessed the correlation between spatial proximity of a gene pair and the physical interaction of their products. We obtained the protein-protein interactions (PPI) from HPRD database [31] and STRING [32]. Fig. 8 shows, for pooled data from all the 6 cell lines, for each of the 5 groupings of gene pairs (defined in Fig. 5), the fraction of all gene pairs that have evidence for physical interaction (fractional PPI). Results

for other individual cell lines are provided in Supplementary Fig. 5. Taken together, these data suggest that, both pathway membership and spatial proximity of a gene pair is equally associated with PPI between their products, and the effect is partly independent, that is, PPI tendency is much greater for gene pairs that are both in the same pathway and physically proximal.

4. Functional enrichment for spatially proximal genes in pathways

Only a small fraction of genes in a pathway are spatially proximal to other genes in the same pathway. This could partly be due to low coverage and false negatives in the Hi-C derived interactions. However, given the differences between proximal and non-proximal genes above, we investigated whether specific functional terms are enriched among *proximal-intra-pathway* genes relative to non-proximal-intra-pathway genes (see Fig. 5 for definition). We pooled the data for all pathways into two groups and performed the functional enrichment analysis using GOrilla software [33]. Fig. 9 suggests that regulatory (both RNA processing, and splicing) and protein binding functions are highly enriched among the spatially proximal pathway genes. We emphasize that our background set of genes in this analysis included those that are in the pathways and are spatially proximal to some other gene not belonging to the same pathway. Thus the observed functional enrichment is not due to spatial proximity alone and is a specific property of spatially proximal genes within pathways.

5. Spatial proximity and regulatory hierarchy

The comparative analyses of biological properties of spatially proximal pathway genes relative to other pathway genes thus far suggests that the upstream genes in a pathway may be more likely to be spatially proximal, ensuring their robust expression and consequently robust pathway activity. We derived the hierarchical level of all pathway genes based on directed pathway edges (see Materials & Methods), and compared the hierarchical levels of proximal and non-proximal genes, pooled overall pathways. Our approach for assigning hierarchical level does not partition the pathway into a strict hierarchy, and can accommodate cycles. We found that, in 3 of the 6 cell lines, the hierarchical levels of proximal genes were higher than the rest; Wilcoxon test p-values: IMR90 ($p = 4.4E-04$), GM06990 ($p = 9.7E-04$) and hESC ($p = 0.05$). Lack of significance in other cell lines (p-values = 0.07, 0.16, and 0.2) may be attributed to due to insufficient data; when we pool the data from the other 3 cell lines, the result is significant ($p = 1.41E-05$). If we apply Chi-square test based on hierarchy level of 2 as the partitioning criterion, an additional cell line yields significance and importantly the odds ratio in all six cell lines range favorably from 1.7 to 3.2.

6. Spatial proximity between pathways

Next, we investigated, whether certain pairs of pathways might occupy neighboring spaces in the nucleus, suggesting their functional relatedness. Analogous to intra-pathway spatial proximity estimation, for each pair of pathways, after excluding the common genes, we obtained the number of interactions between genes across pathways and estimated its Z-score based on 1000 controlled random samplings (for computational tractability) of 2 sets of genes representing the 2 pathways. Using Z-score > 2 as the threshold, we identified a total of 3109 pathway pairs across all cell lines, 73 of which were proximal in at least 4 tissues and 20 were proximal in at least 5 tissues (Supplementary Table 2).

Next, analogous to intra-pathway proximity analysis, we estimated inter-pathway spatial proximity between housekeeping genes and all other KEGG pathways. We found that housekeeping genes are significantly proximal to 95 out of 164 KEGG pathways, at Z-score > 2 threshold in at least 1 cell line; the full distribution of number of housekeeping-proximal pathways shared in 1 or more cell lines is provided in Supplementary Fig. 6, and pathways proximal to housekeeping genes in 4 or more cell lines are provided in Supplementary Table 2. These results suggest a central role housekeeping genes may play in the genome organization as well as in coordinating the activities of other pathways.

Several of the pathway pairs deemed to be spatially proximal in our analysis (Supplementary Table 2) are very likely to be functionally related, for instance, pathways for metabolism of various amino acids and derivatives thereof. Importantly, however, these data also reveal non-trivial relationships, which have some support in literature. We discuss a few next.

- Consistent with the proximity of steroid hormone metabolism and proteolysis, direct functional links between these two processes have been noted [34].
- Links between amino acid metabolism and cell cycle and cancer (Chronic myloid leukemia in our case) is not surprising given the metabolic requirements during cell division and growth.
- JAK-STAT pathway and SNARE complexes are found to be proximal. While we did not find a direct link between the two, they are known to be co-targeted by TGF-beta signaling [35].
- Abnormal *steroid hormone metabolism* has been reported in *Myeloid leukemia* patients [36], which is consistent with the detected proximity of these two pathways.

- *Butanoate metabolism* and *Glioma* are deemed to be spatially proximal in our analysis. Treatment with butanoate (sodium butyrate) is known to induce differentiation of c6 Glioma cells [37].
- While a direct link between *Cholera* and *DRPLA*, deemed to be proximal, is not clear, we note that *DRPLA* is a spinocerebral degeneration disease, and *cholera toxin subunit b* is known to be dispersed to brain and spinal cord neurons [38].
- *Synthesis and degradation of ketone bodies* was found to be proximal to genes involved in several different cancers, including pancreatic cancer, and cell cycle. Previous papers have shown growth inhibitory effects of *ketone* bodies on pancreatic cancer could be mediated by reduced c-Myc expression [39]. Moreover c-Myc is overexpressed in pancreatic cancer [40] and its inhibition has been shown to result in regression of *lung cancer* [41] and pancreatic cancer [42].
- *Glycosylphosphatidylinositol* (GPI) and *Huntington* (a neurological disease) were deemed proximal. While a direct link between these two pathways is not clear, GPI-anchor cleavage is known to modulate notch signaling and promoter neurogenesis [43].

Overall, these results suggest that spatial proximity between pathway genes is somewhat associated with functional interactions between the pathways.

Discussion

In this work, we have presented the first comprehensive analysis of intra- and inter-pathway spatial proximity in multiple *Homo sapiens* cell lines. Previous studies have shown that in *Saccharomyces cerevisiae* [10], *Plasmodium falciparum* [12] and *H.sapiens* lymphoblastoid cell line [13], broadly, functionally related genes tend to be spatially proximal. Our goal here was to not only extend these previous observations to multiple human cell lines and assess the relationships between spatial proximity and pathway activity based on gene expression, but equally importantly, to further functionally characterize proximal genes within pathways and examine higher-order physical and functional interactions between pathways.

Previous similar analysis in *S. cerevisiae* [10] are based on only inter-chromosomal segment interactions, and *H. sapiens* lymphoblastoid cell line results [13] are based on low-resolution (1 Mb) segments, which can result in spurious interactions at the level of individual gene loci. Importantly, however, a greater tendency for genomically proximal regions to be spatially proximal, i.e., autocorrelation, unless appropriately controlled for, can result in false positives in inferring significant spatial proximity from Hi-C data. The absence of effective tools to

control for autocorrelation has forced previous studies to exclude intra-chromosomal interactions from consideration, significantly impacting their statistical power [10]. In contrast the Homer tool [20] satisfactorily controls of autocorrelation in estimating significance, consequently enabling us to include all significant interactions, both inter- and intra-chromosomal, in our analyses, while obviating an explicit control for inter-gene distances in random sampling procedures. We note that despite an explicit control for autocorrelation, genomically proximal regions (within 500 kb) have a slightly higher tendency to be spatially proximal (data not shown). However, we explicitly tested if this biases our pathway proximity assessment as follows. We compared intra-pathway gene distances with those for randomly selected genes from the same chromosome and found that for none of the pathways there was a significant difference between the two sets of distances.

We have performed our analyses based on 100 kb resolution to detect interactions. However, we have also assessed the impact of using a higher resolution of 10 kb. We found that the number of interactions detected was much greater when using 100 kb resolution, especially the inter-chromosomal interactions, due to a greater statistical power for interaction detection at this resolution. For example, for cell type HEK293, only 7889 intra- and only 14 inter-chromosomal gene-gene interactions are detected using 10kb resolution and 43439 intra- and 1538 inter-chromosomal interactions at 100 kb. Moreover, at 10 kb resolution, large fractions of detected interactions are within a gene, which does not contribute to our analyses. We note that using a 100 kb resolution does not substantively influence the gene-gene interaction inference, as only 4% of 100 kb segments have multiple genes. Thus, to maximize statistical power with relative small fraction of ambiguous (but not necessarily biologically wrong) gene-gene interaction calls, we chose to perform all downstream analyses based on 100 kb resolution.

Hi-C data, like most genome-wide datasets, comes with a level of false positives, as does the pathway data. This is an important issue, and one that cannot be addressed by computational means alone. We have relied on the published interaction data that have undergone quality control measures, and used robust tool with recommended controls to perform the analyses. For pathway proximity, we have relied on a well-controlled randomized gene set to assess the significance of proximity. Despite the controls, a certain fraction of data is likely to be false positive. However, the noise in the data, as long as the tests are properly controlled, is not likely to generate strong consistent signals across multiple cell types simply by chance.

We performed a number of checks to ensure the robustness of our conclusions against several potential biases. First, note that in quantifying the significance of the pathway spatial proximity, we control for the lengths of the pathway genes. Second, we ensured that our ability to detect the spatial proximity of a pathway is independent of the number of genes in the pathway (Spearman correlation between pathway size and spatial proximity Z-score was statistically insignificant). A recent paper [9] has suggested a link between detection of Hi-C interaction of a gene and a gene's codon usage (which is related to its expression) and the GC content of the gene's genomic locus. Third, we ascertained that the codon-usage does not bias the detection of Hi-C interactions (Spearman correlation = -0.09).

Finally, with regards to the potential GC bias, indeed highly interacting loci tend to have lower GC composition, as noted previously [11]. Although the GC composition near the restriction sites can present a technical bias [44], several papers also suggest that GC composition may be an inherent property of the physical proximity [45]. Therefore, it seems that an explicit control for GC content may not be ideal. However to explore the extent to which GC-controlled analyses would affect our results, for HEK293 cell line we reprocessed the data with specific GC control option provided in HOMER tool and compared the downstream results of our analyses with and without the GC control. We found that 80% of interactions detected without GC control are also detected with GC control and overall correlation between gene-wise interaction degrees between the two is 0.92. Next, we re-estimated the pathway proximity Z-scores based on GC-controlled interactions and found those to be highly correlated with the Z-score based on uncontrolled interaction detection (Spearman rho = 0.85). Lastly, we re-calculated the correlation between pathway proximity and pathway activity and found that too is as significant as the estimated without GC control (Spearman rho = 0.80, relative to 0.83 without GC control). Thus, our conclusions are not substantively biased by these various potentially confounding factors.

Our primary resource of biological pathways – KEGG, is dominated by essential and broadly utilized cellular pathways, and therefore it is encouraging to see that by and large KEGG pathways are not only highly significantly proximal (Fig. 2), but a large fraction of these pathways are proximal in multiple cell lines (Fig. 4). However, as we show, this is not true for a different set of pathways relevant to cancer, where the overall z-scores are much more subdued, and spatially proximity is less ubiquitous (Supplementary Fig. 2). Despite, general ubiquity of spatial proximity of KEGG pathways, we still see a strong correlation between cell type-specific spatial proximity and pathway activity as approximated by gene expression (Fig. 7).

Previous studies have noted greater transcription in spatially clustered regions [8]. Independently, earlier studies have shown the existence of so called transcription factories [46] - nuclear locales with enriched core transcriptional machinery components where transcripts are synthesized and processed. Moreover, links between transcription factories and chromatin organization have been noted [47]. Taken together, these previous results are consistent with our observation that genes in spatially proximity with other genes have much higher expression. However, interestingly, in addition to spatial proximity alone, functional relationship between the spatially proximal genes, i.e., membership in the same pathway, makes a small but significant additional contribution to the gene expression level (Fig. 6).

Our analysis reveals an unexpected association between spatial proximity of a gene pair and the interaction between their protein products. Among the physically proximal gene pairs, the well-annotated genes, i.e., annotated in some KEGG pathway, have much greater PPI propensity than genes that do not belong to an annotated pathway ((III) vs. (IV) in Fig. 8); this may be explained by a greater representation of well-studied genes in PPI databases. Functionally related genes have been shown to have a greater propensity to physically interact [49], consistent with our findings (compare (I) and (III) in Fig. 8). However, we found that spatial proximity is independently associated with protein interaction in both pathway ((I) versus (II) in Fig. 8) and non-pathway ((IV) versus (V) in Fig. 8) contexts. The gene-pairs that are both spatially proximal and belong to same pathway have the highest PPI propensity ((I) in Fig. 8). These trends are identical in all 6 cell lines, suggesting that both spatial proximity and pathway membership contribute independently to PPI propensity.

Scrutinizing each of the pathways, we found the spatially proximal genes in a pathway to have distinguishing functional characteristics relative to other pathway genes that are not spatially proximal to any other gene in the same pathway. In terms of biological processes (Fig. 9), such genes are overwhelmingly involved in transcription, splicing and intracellular transport and localization. Interestingly, the genes in a pathway that are spatially proximal to other genes in the same pathway tend to occupy a higher level in the regulatory hierarchy, related to other genes in the pathway, that also are spatially proximal to other genes but none in the same pathway. Overall, these results ascribe, for the first time, a special functional status to spatially proximal genes in pathways – such genes tend to perform higher-level regulatory functions.

We found that housekeeping genes, consistent with their ubiquitous expression and activity, tend to be broadly and highly spatially proximal. Previous studies have observed clustering of housekeeping genes into so call transcription factories [50], and have suggested that

interactions between housekeeping genes may play a role in the spatial organization of the chromatin [51]. Our results confirm these previous observations through the first genome-wide assessment of spatial proximity of housekeeping genes in multiple cell lines. In addition to spatial proximity of housekeeping genes, we also found that as a group housekeeping genes are 'centrally' located in the nucleus and act as a link between numerous other pathways. A mechanistic interpretation of this intriguing observation, as well as of the causal link between the expression of housekeeping genes and their spatial proximity, will require further analysis.

Conclusion

Overall, based on the first comprehensive pathway-centric analysis of spatial proximity in multiple cell lines, our results suggest that (i) context-specific regulation of pathways is associated with their context-specific spatial proximity; in doing so, our analysis provides mechanistic insights into cell type-specific activity of certain pathways, (i) spatial-proximity of pathway genes is associated with physical interaction among their gene products, and (iii) specific classes of genes within pathways, likely occupying higher regulatory levels, have a greater tendency to be spatially proximal. Our results also provide insights into correlated activity of multiple pathways by showing that the genes in these pathways are spatially proximal.

Methods

Hi-C processing pipeline. We downloaded paired-end Hi-C raw reads FASTQ files of sample replicates for the following tissues from GEO database (www.ncbi.nlm.nih.gov/geo): (i) HEK293 (GSM1081530, GSM1081531) [15], (ii) IMR90 (GSM1055800, GSM1055801) [52], (iii) hESC (GSM862723, GSM892306) [4], (iv) GM06990 (GSM1340639) [16], (v) RWPE1 (GSM927076) [17] and (vi) BT483 (GSM1340638, GSM1340637) [16]. We mapped the reads onto hg19 human genome using *BWA* tools [53] with default parameters. The resulting SAM files were converted to BAM files using "*samtools view*" program [54] and processed to removing PCR duplicates using "*samtools sort*" and "*Picard"* tools.

We then processed the non-redundant reads for Hi-C analysis using various HOMER tools [20]: we ran "*makeTagDirectory*" program using options "*tbp -1*" (to ensure that any genomic location is mapped by a unique read), "*–restrictionSite*" (only keep reads if both ends of the paired-end read have a restriction site within the fragment length estimate 3' to the read), "*-removePEbg*" (removing read pairs separated by less than 1.5x the sequencing insert

fragment length) , "*-removeSelfLigation*" (remove re-ligation events), and "*-removeSpikes*" (remove high tag density regions).

Normalization of Hi-C interactions. Having output from the previous steps, next we normalized the data to create background of the Hi-C interactions at 100kb resolution using HOMER *"analyzeHiC"* program. The program (i) divides the genome into 100kb regions, (ii) calculates total read coverage in each region, (iii) calculates the fraction of interactions spanning any given distance with respect to read depth, (iv) optimizes a read count model to assign expected interaction counts in regions with uneven sequencing depth and (v) calculates variation in interaction frequencies as a function of distance. For fragments *i* and *j*, the procedure the estimated expected number of reads supporting the interaction as:

$$e_{i,j} = \frac{f(i-j) * n_i * n_j}{N}$$

where N = total number of reads, n = number of reads in a region, and f represents the expected frequency of Hi-C reads as a function of distance.

The background model were created for the entire data for a given sample then applied for selection of Hi-C interacting reads at default p-value cutoff of 0.001 using "*-interaction*" parameters using "*analyzeHiC*" program. We further filtered the Hi-C interactions using FDR cutoff <= 0.1 and considered the resulting set as significant Hi-C interactions. The significant Hi-C interacting reads were then passed through "*annotateInteractions*" program for mapping them onto the annotated genomic features (i.e., 5'UTR, CDS, introns, exons, intergenic etc.), from which, we selected only those interactions mapping to gene regions (i.e., Promoter, 5'UTR, exon, intron and 3'UTR), resulting in a comprehensive set of spatially proximal gene pairs (SGP) for each tissue.

Codon usage. To assess the effect of codon bias on interaction detection, we used Biopython's "*SeqeUtils::ModuleCodonUsage*" package and calculated Codon Adaptation Index (CAI) for each gene of the *H. sapiens* genome.

Pathway datasets. A 'pathway' for our purpose is a set of genes. We downloaded two sets of pathways from KEGG [21] and NetPath [19] databases, only retaining those with at least 10 genes, resulting in 164 and 32 pathways respectively. We also included the set of 3800 housekeeping genes [22], as an additional 'pathway'.

Defining classes of edges and nodes relative to pathways. For various analyses we have defined 5 disjoint sets of genes and gene pairs in the context of a pathway and spatial proximity (see Figure 5 for illustration). All spatially proximal gene pairs were partitioned into three groups: *proximal-intra-pathway:* intra-pathway spatially proximal gene pairs, *proximal-inter-pathway:* Spatially proximal pair of genes where each gene is in a different pathway, and *proximal-generic:* Spatially proximal pair of genes such that at least one of the genes is not in any pathway. Similarly all non-proximal gene pairs were partitioned into two categories: *non-proximal-intra-pathway:* a non-proximal gene pair within a pathway, and *non-proximal-generic:* any pair of genes that are not spatially proximal to any other gene and not within any pathway.

Edge fraction and its significance in estimating pathway spatial proximity. Intra-pathway gene spatial proximity was estimated as *Edge Fraction (EF)* – number of pairwise gene-gene interaction in the pathway normalized by the number of total possible interactions. To quantify significance of the *EF* for pathway with *N* genes, we randomly sampled *N* genes, such that number of genes in each chromosome is identical to real pathway, and each sampled gene's length was within 20% of the matched pathway gene's length (this controls for length-based bias in interaction detection). We generated 1000 such samples and calculated 1000 corresponding *EF*s. We then obtained the Z-score corresponding to the EF for actual pathway relative to 1000 controls. We estimated the Z-score for each pathway in each cell line resulting in 165 x 6 Z-score matrix.

Inter-pathway proximity. Analogous to the intra-pathway proximity analysis above, for a pair of pathways, after excluding the shared genes, we estimate the edges between genes in two pathways, and estimate its significance based on randomly sampling 2 gene sets (instead of 1 as above), with identical controls as above. We thus estimated a Z-score for inter-pathway proximity for all pairs of pathways.

Processing RNA-Seq data. We downloaded raw RNA-Seq FASTQ files for three of the tissues in which matching RNA-Seq was available: (i) HEK293 (GSM1081534, GSM1081535) [15], (ii) IMR90 (GSM1154029) [52] and (iii) RWPE1 (GSM927074) [17], (iv) hESC (GSM758566) [55], (v) GM06990 (GSM958747) [56] and (vi) BT483 (GSM1172854) [57]. The raw FASTQ reads were mapped and processed up to de-duplication steps using the same pipeline that we used to apply for Hi-C data analysis, and then used to quantify expression levels using cufflinks [58] tool with default parameters, yielding gene-wise RPKM values.

Significance of pathway activity. For genes with multiple transcripts, we take the maximum expression over all transcripts for the genes. For a pathway, we estimated its activity as the average expression of the genes in the pathway; we selected only the *proximal-intra-pathway* genes to estimate activity. To quantify significance of the pathway activity, we followed a sampling approach similar to that for estimating EF significance above, yielding a z-score for each pathway and cell line.

Estimating regulatory hierarchy. The regulatory analysis of a pathway requires a directed graph in which all defined interactions suggest direction of the signal flow. KEGG does not provide the directed graph by default. Therefore, we downloaded *KGML* (KEGG Markup Language) and parsed the files using *KEGGgraph* package in Bioconductor and created directed graph for each pathway using *NetworkX* package in Python. In order to investigate hierarchy of genes we first assigned a synthetic root connecting to all pathway nodes with zero in-degree, and then calculated the shortest path length (SPL) from root to every other node. Low SPL indicates higher level of hierarchy. For Chi-square and Fisher tests (see Results), we created two gene sets: (i) genes with SPL = 1 (top level) and (ii) genes with SPL >= 3 (lower hierarchy). We pooled these two sets across all pathways.

Author contributions

S.H. conceived the project with input from C.K. and M.G. H.K. designed and implemented the software pipeline and performed all analyses. S.H. wrote the manuscript with help from all authors.

Additional Data Files

“Supplementary File” contains supplementary figures and tables. File ‘Supplementary Fig1’ contains the Supplementary Fig. 1.

Acknowledgement

This work was funded by NIH HG007104 to C.K. and NIH GM100335 to S.H.

Competing interests

None of the authors has any competing interests.

Figure Legends

Figure 1. Overview of the Hi-C processing pipeline and flow of downstream analysis (see Methods for details).

Figure 2. Z-score distribution of intra-pathway spatial proximity. The figure shows the distributions of spatial-proximity z-scores for three sets of gene sets, pooled from 6 cell lines. Blue: KEGG pathway. Red: Random gene-sets matching each KEGG pathway controlled for gene lengths and chromosomal distributions. Green: intersection of each pair of KEGG pathways.

Figure 3. Spatial proximity of *intra-pathway* genes in selected KEGG pathways across six cell lines. Note the ubiquitous proximity of Housekeeping genes, hESC-specific proximity of p53 signaling pathway.

Figure 4. *Intra-pathway* pathway proximity is shared across tissues. The figure shows number of pathways (Y-axis) with high ***intra-pathway*** proximity (Z-score >= 2) in different number of cell lines (X-axis). The table values show number of pathways whose ***intra-pathway*** genes proximity (Z-score >= 2) is unique to a cell line (diagonal) or shared between a pair of cell lines (off-diagonal).

Figure 5. Schematic defining sets of genes and gene-pairs. Outermost rectangle represents all genes. Inner rectangle represents all genes that are in spatial proximity to at least one other gene. Circles represent annotated pathways. Nodes represent genes and edges represent spatially proximal gene pairs. All spatially proximal gene pairs are partitioned into three groups: ***proximal-intra-pathway:*** intra-pathway spatially proximal gene pairs (orange nodes and edges), ***proximal-inter-pathway:*** Spatially proximal pair of genes where each gene is in a different pathway (light blue nodes and edges), and ***proximal-generic:*** Spatially proximal pair of genes such that at least one of the genes is not in any pathway (red nodes and edges). Similarly all non-proximal gene pairs are partitioned into two categories: ***non-proximal-intra-pathway:*** a non-proximal gene pair within a pathway, and ***non-proximal-generic:*** any pair of genes that are not spatially proximal to any other gene and not within any pathway (pairs of dark blue nodes).

Figure 6. Spatial proximity and gene expression. The figure shows box-plots of gene expression (FPKM) values of the genes in three different groups (see Figure-5) pooled from all 6 cell lines. A~B Wilcoxon test p-values = 3.4E-35. A~C Wilcoxon test p-value = 8.43E-72. See Supplementary Fig. 4 for results of 6 individual cells.

Figure 7. Spatial proximity versus mean pathway expression. This figure shows scatter plot between proximity z-scores of pathways versus z-scores of expression values of the proximal-intra-pathway genes pooled from all 6 cells. Spearman rho = 0.77, p-value: 5.90e-62.

Figure 8. Pathway membership, spatial proximity and PPI. The figure shows 6 cells pooled fraction of gene pairs (Y-axis) in different gene groups (X-axis; see Figure 5) whose protein products physically interact. The inset shows Fishers tests p-values for various pair-wise comparisons. See supplementary Figure 6 for other tissues.

Figure 9. GO enrichment analysis. This figure enriched GO terms (–log (q-value) >= 2) ***proximal-intra-pathway*** genes relative to other spatially proximal genes (see Fig. 5).

566 **Table 1. Hi-C analysis summary.**

Cell line	Rep-1 Hi-C mapped fragments in million	Rep-2 Hi-C mapped fragments in million	# Unique genes in HiC interactions	# Unique genes in HiC interactions
IMR90	66.69	299.86	14550	101502
HEK293	92.03	210	11978	44977
hESC	42.49	234.33	10713	34531
GM06990	58.62	96.33	9019	11447
BT483	41.8	64.7	5948	5792
RWPE1	73.65	82.33	9416	14124

567

568

References

1. Shaw PJ: **Mapping chromatin conformation.** *F1000 biology reports* 2010, **2**.
2. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J: **Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions.** *Nature biotechnology* 2013, **31:**1119-1125.
3. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES: **Hi-C: a method to study the three-dimensional architecture of genomes.** *Journal of visualized experiments : JoVE* 2010.
4. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions.** *Nature* 2012, **485:**376-380.
5. Aguilar-Arnal L, Sassone-Corsi P: **Chromatin landscape and circadian dynamics: Spatial and temporal organization of clock transcription.** *Proceedings of the National Academy of Sciences of the United States of America* 2014.
6. Duggal G, Wang H, Kingsford C: **Higher-order chromatin domains link eQTLs with the expression of far-away genes.** *Nucleic acids research* 2014, **42:**87-96.
7. Sanyal A, Lajoie BR, Jain G, Dekker J: **The long-range interaction landscape of gene promoters.** *Nature* 2012, **489:**109-113.
8. Malin J, Aniba MR, Hannenhalli S: **Enhancer networks revealed by correlated DNAse hypersensitivity states of enhancers.** *Nucleic acids research* 2013, **41:**6828-6838.
9. Diament A, Pinter RY, Tuller T: **Three-dimensional eukaryotic genomic organization is strongly correlated with codon usage expression and function.** *Nature communications* 2014, **5:**5876.
10. Wang H, Duggal G, Patro R, Girvan M, Hannenhalli S, Kingsford C: **Topological properties of chromosome conformation graphs reflect spatial proximities within chromatin.** In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. pp. 306-315. Wshington DC, USA: ACM; 2013:306-315.
11. Dekker J: **GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p.** *Genome biology* 2007, **8:**R116.
12. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert JP, Noble WS, Le Roch KG: **Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression.** *Genome research* 2014, **24:**974-988.
13. Thevenin A, Ein-Dor L, Ozery-Flato M, Shamir R: **Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome.** *Nucleic acids research* 2014, **42:**9854-9861.
14. **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
15. Zuin J, Dixon JR, van der Reijden MI, Ye Z, Kolovos P, Brouwer RW, van de Corput MP, van de Werken HJ, Knoch TA, van IWF, et al: **Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells.** *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111:**996-1001.
16. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, Nagano T, Andrews S, Wingett S, Kozarewa I, et al: **Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C.** *Genome research* 2014, **24:**1854-1868.
17. Rickman DS, Soong TD, Moss B, Mosquera JM, Dlabal J, Terry S, MacDonald TY, Tripodi J, Bunting K, Najfeld V, et al: **Oncogene-mediated alterations in chromatin conformation.** *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109:**9083-9088.
18. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG.** *Nucleic acids research* 2014, **42:**D199-205.

19. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, Venugopal AK, Telikicherla D, Navarro JD, Mathivanan S, Pecquet C, et al: **NetPath: a public resource of curated signal transduction pathways.** *Genome biology* 2010, **11:**R3.
20. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.** *Molecular cell* 2010, **38:**576-589.
21. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic acids research* 2000, **28:**27-30.
22. Eisenberg E, Levanon EY: **Human housekeeping genes, revisited.** *Trends in genetics : TIG* 2013, **29:**569-574.
23. Starr AE, Dufour A, Maier J, Overall CM: **Biochemical analysis of matrix metalloproteinase activation of chemokines CCL15 and CCL23 and increased glycosaminoglycan binding of CCL16.** *The Journal of biological chemistry* 2012, **287:**5848-5860.
24. Yager JD, Davidson NE: **Estrogen carcinogenesis in breast cancer.** *The New England journal of medicine* 2006, **354:**270-282.
25. Basu G, Mohapatra A: **Interactions between thyroid disorders and kidney disease.** *Indian journal of endocrinology and metabolism* 2012, **16:**204-213.
26. I AAE-A, Hamdy G, Amin M, Rashad A: **Pulmonary function changes in diabetic lung.** *Egyptian Journal of Chest Diseases and Tuberculosis* 2013, **62:**513-517.
27. Kudva R, Hegde, P.: **Blue nevus of the prostate.** *Indian J Urol* 2010, **26:**301-302.
28. Dreesen O, Brivanlou AH: **Signaling pathways in cancer and embryonic stem cells.** *Stem cell reviews* 2007, **3:**7-17.
29. Seita J, Weissman IL: **Hematopoietic stem cell: self-renewal versus differentiation.** *Wiley interdisciplinary reviews Systems biology and medicine* 2010, **2:**640-653.
30. Kelley R, Ideker T: **Systematic interpretation of genetic interactions using protein networks.** *Nature biotechnology* 2005, **23:**561-566.
31. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, et al: **Human protein reference database--2006 update.** *Nucleic acids research* 2006, **34:**D411-414.
32. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ: **STRING v9.1: protein-protein interaction networks, with increased coverage and integration.** *Nucleic acids research* 2013, **41:**D808-815.
33. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z: **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.** *BMC bioinformatics* 2009, **10:**48.
34. Umpleby AM, Russell-Jones DL: **The hormonal control of protein metabolism.** *Bailliere's clinical endocrinology and metabolism* 1996, **10:**551-570.
35. Asirvatham AJ, Gregorie CJ, Hu Z, Magner WJ, Tomasi TB: **MicroRNA targets in immune genes and the Dicer/Argonaute and ARE machinery components.** *Molecular immunology* 2008, **45:**1995-2006.
36. Gallagher TF, Hellman L, Zumoff B, Miller DG: **Steroid Hormone Metabolism in Chronic Myelogenous Leukemia.** *Blood* 1965, **25:**743-748.
37. Sun SH, Ou HC, Jang TH, Lin LB, Huang HM: **Altered phospholipid metabolism in sodium butyrate-induced differentiation of C6 glioma cells.** *Lipids* 1997, **32:**273-282.
38. Alisky JM, van de Wetering CI, Davidson BL: **Widespread dispersal of cholera toxin subunit b to brain and spinal cord neurons following systemic delivery.** *Experimental neurology* 2002, **178:**139-146.
39. Shukla SK, Gebregiworgis T, Purohit V, Chaika NV, Gunda V, Radhakrishnan P, Mehla K, Pipinos, II, Powers R, Yu F, Singh PK: **Metabolic reprogramming induced by ketone bodies diminishes pancreatic cancer cachexia.** *Cancer & metabolism* 2014, **2:**18.
40. Buchholz M, Schatz A, Wagner M, Michl P, Linhart T, Adler G, Gress TM, Ellenrieder V: **Overexpression of c-myc in pancreatic cancer caused by ectopic activation of NFATc1 and the Ca2+/calcineurin signaling pathway.** *The EMBO journal* 2006, **25:**3714-3724.

41. Soucek L, Whitfield J, Martins CP, Finch AJ, Murphy DJ, Sodir NM, Karnezis AN, Swigart LB, Nasi S, Evan GI: **Modelling Myc inhibition as a cancer therapy.** *Nature* 2008, **455:**679-683.
42. Kunnumakkara AB, Guha S, Krishnan S, Diagaradjane P, Gelovani J, Aggarwal BB: **Curcumin potentiates antitumor activity of gemcitabine in an orthotopic model of pancreatic cancer through suppression of proliferation, angiogenesis, and inhibition of nuclear factor-kappaB-regulated gene products.** *Cancer research* 2007, **67:**3853-3861.
43. Park S, Lee C, Sabharwal P, Zhang M, Meyers CL, Sockanathan S: **GDE2 promotes neurogenesis by glycosylphosphatidylinositol-anchor cleavage of RECK.** *Science* 2013, **339:**324-328.
44. Yaffe E, Tanay A: **Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture.** *Nature genetics* 2011, **43:**1059-1065.
45. Vinogradov AE: **DNA helix: the importance of being GC-rich.** *Nucleic acids research* 2003, **31:**1838-1844.
46. Jackson DA, Hassan AB, Errington RJ, Cook PR: **Visualization of focal sites of transcription within human nuclei.** *The EMBO journal* 1993, **12:**1059-1065.
47. Rieder D, Trajanoski Z, McNally JG: **Transcription factories.** *Frontiers in genetics* 2012, **3:**221.
48. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326:**289-293.
49. Maier T, Guell M, Serrano L: **Correlation of mRNA and protein in complex biological samples.** *FEBS letters* 2009, **583:**3966-3973.
50. Razin SV, Gavrilov AA, Yarovaya OV: **Transcription factories and spatial organization of eukaryotic genomes.** *Biochemistry Biokhimiia* 2010, **75:**1307-1315.
51. Gushchanskaya ES, Artemov, A.V., Ulyanov, S.V., Penin, A.A., Logacheva, M.D., Razin, S.V., Gavrilov, A.A.: **Spatial organization of housekeeping genes in interphase nuclei.** *Molecular Biology* 2014, **48:**886-895.
52. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B: **A high-resolution map of the three-dimensional chromatin interactome in human cells.** *Nature* 2013, **503:**290-294.
53. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
54. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
55. Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, Tanzer A, Lagarde J, Zaleski C, See LH, et al: **Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression.** *Nature communications* 2015, **6:**5903.
56. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al: **Landscape of transcription in human cells.** *Nature* 2012, **489:**101-108.
57. Daemen A, Griffith OL, Heiser LM, Wang NJ, Enache OM, Sanborn Z, Pepin F, Durinck S, Korkola JE, Griffith M, et al: **Modeling precision treatment of breast cancer.** *Genome biology* 2013, **14:**R110.
58. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nature biotechnology* 2010, **28:**511-515.

Figure 1.

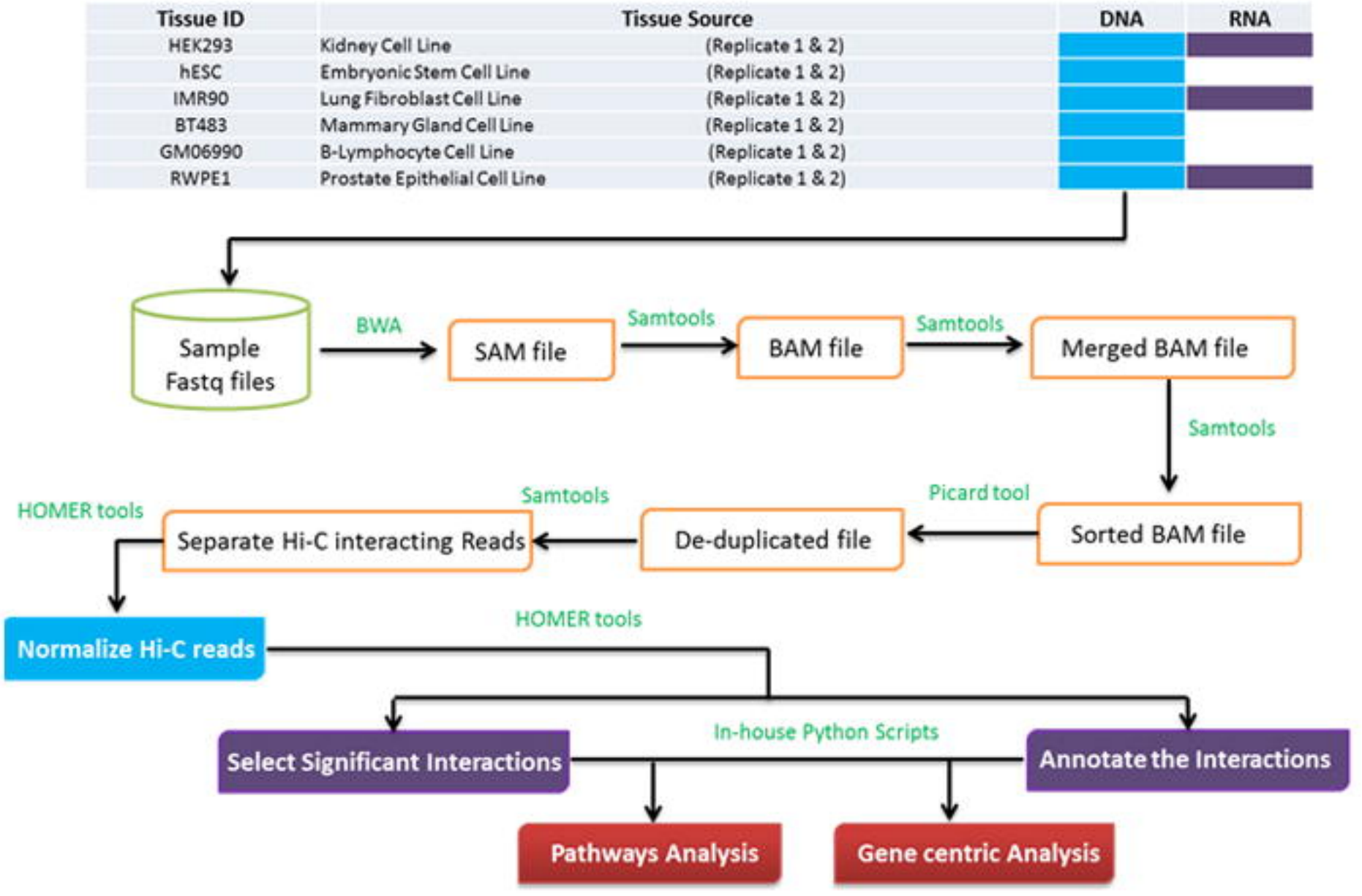
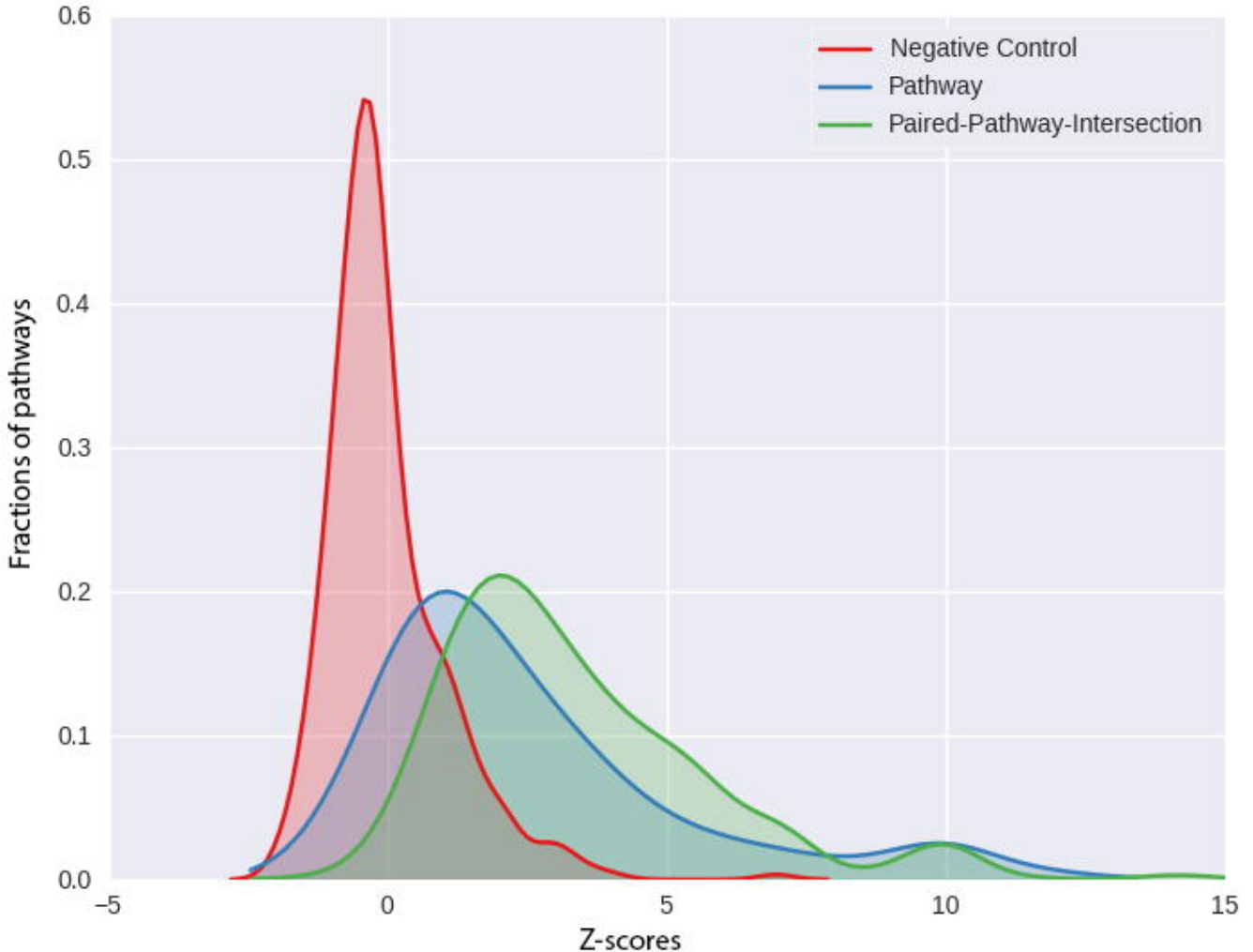

Tissue ID
Tissue Source
DNA
RNA
HEK293
Kidney Cell Line
(Replicate 1 & 2)
hESC
Embryonic Stem Cell Line
(Replicate 1 & 2)
IMR90
Lung Fibroblast Cell Line
(Replicate 1 & 2)
BT483
Mammary Gland Cell Line
(Replicate 1 & 2)
GM06990
B-Lymphocyte Cell Line
(Replicate 1 & 2)
RWPE1
Prostate Epithelial Cell Line
(Replicate 1 & 2)
Sample Fastq files
BWA
SAM file
Samtools
BAM file
Samtools
Merged BAM file
Samtools
Sorted BAM file
Picard tool
De-duplicated file
Samtools
Separate Hi-C interacting Reads
HOMER tools
Normalize Hi-C reads
HOMER tools
Select Significant Interactions
In-house Python Scripts
Annotate the Interactions
Pathways Analysis
Gene centric Analysis

Figure 2.


Negative Control
Pathway
Paired-Pathway-Intersection
Fractions of pathways
Z-scores
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−5
0
5
10
15

KEGG PATHWAYS Z-SCORE HEATMAP

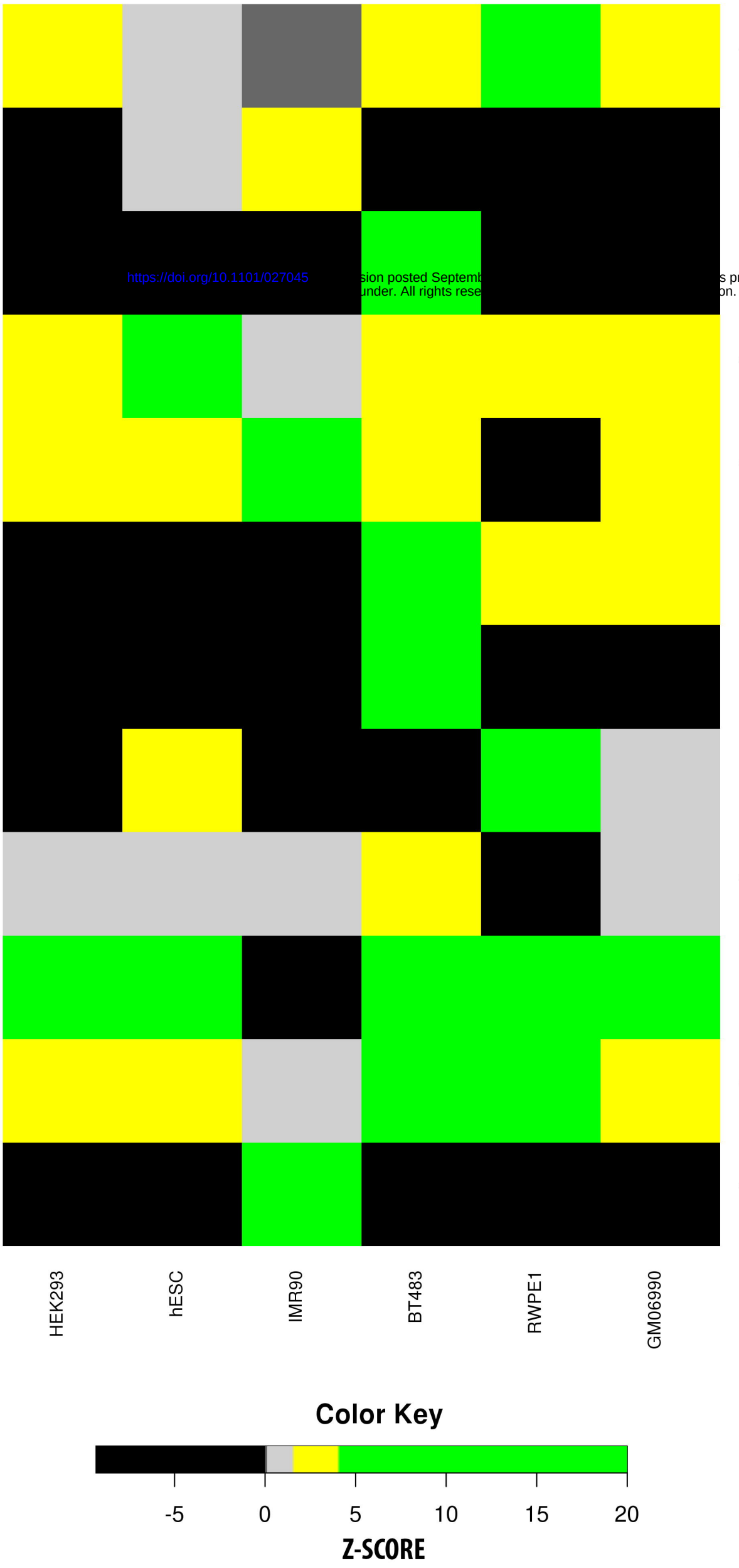
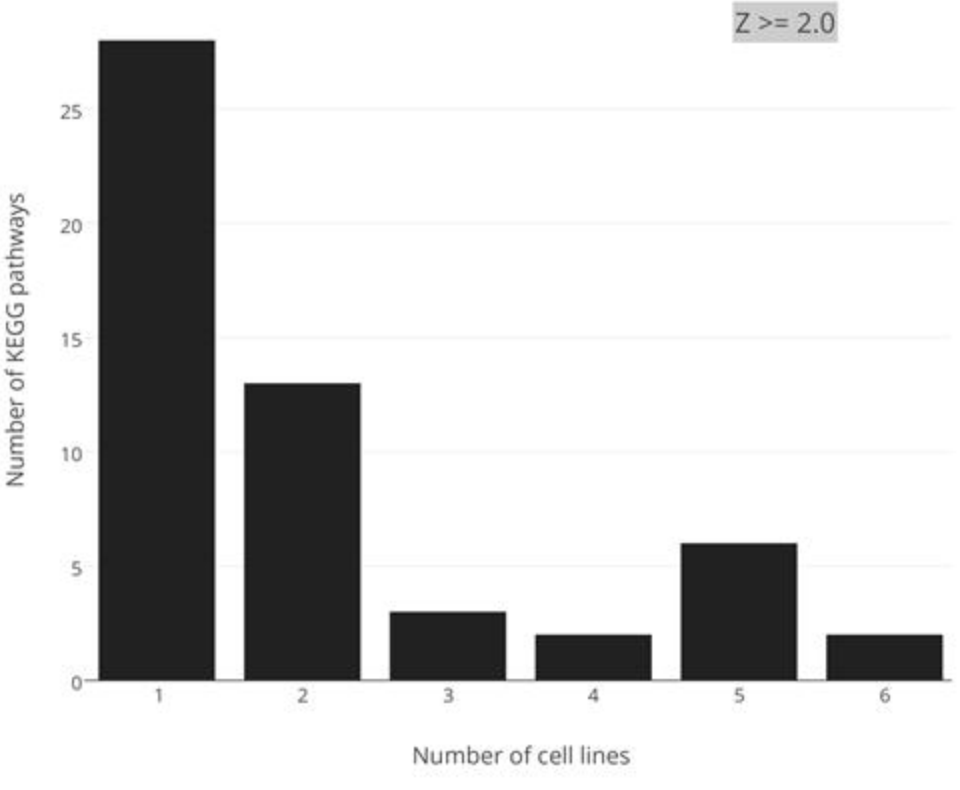


Figure 4.



	HEK293	IMR90	hESC	RWPE1	GM06990	BT483
HEK293	2	8	4	11	10	9
IMR90		2	4	13	10	8
hES			6	5	3	3
RWPE1				15	12	13
GM06990					2	11
BT483						3

Z-score >= 2.0

Figure 5.

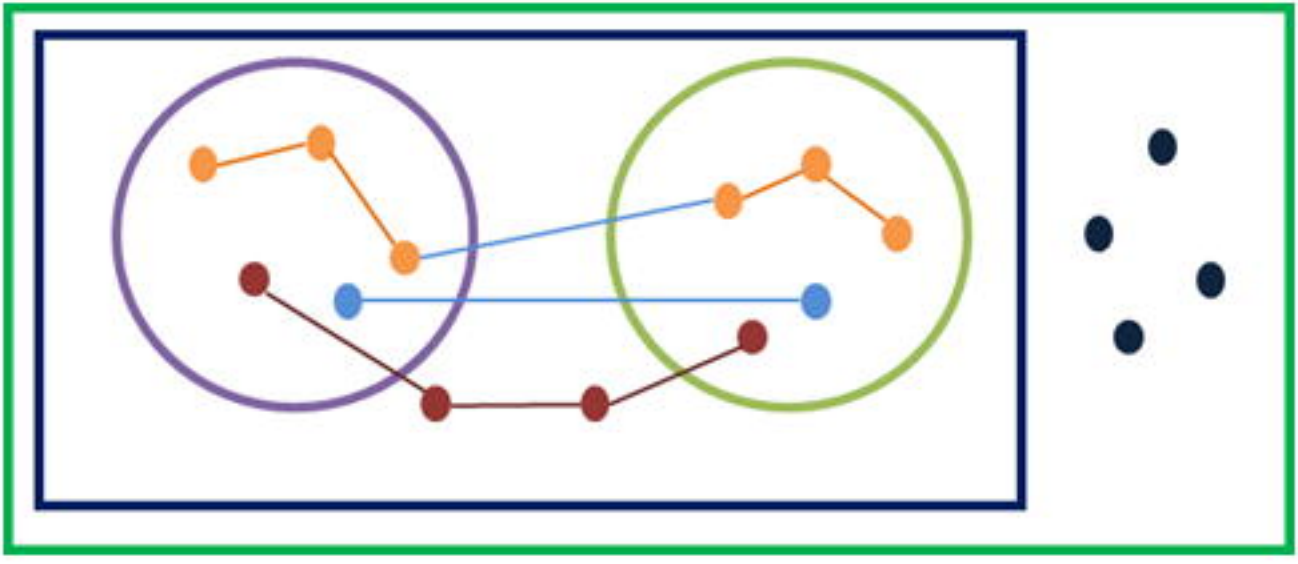


Figure 6.

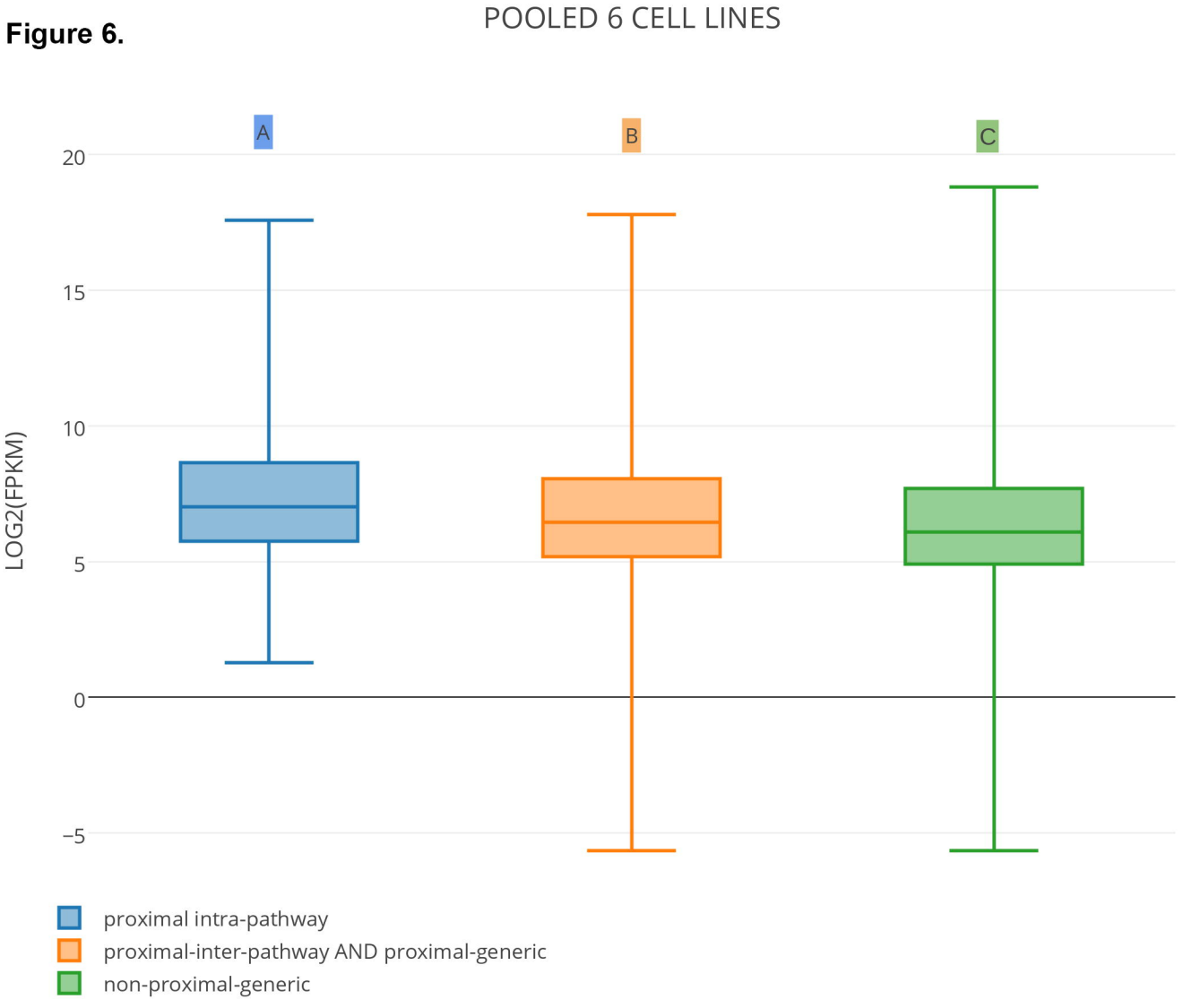

POOLED 6 CELL LINES
A
B
C
20
15
10
5
0
−5
LOG2(FPKM)
proximal intra-pathway
proximal-inter-pathway AND proximal-generic
non-proximal-generic

Figure 7.

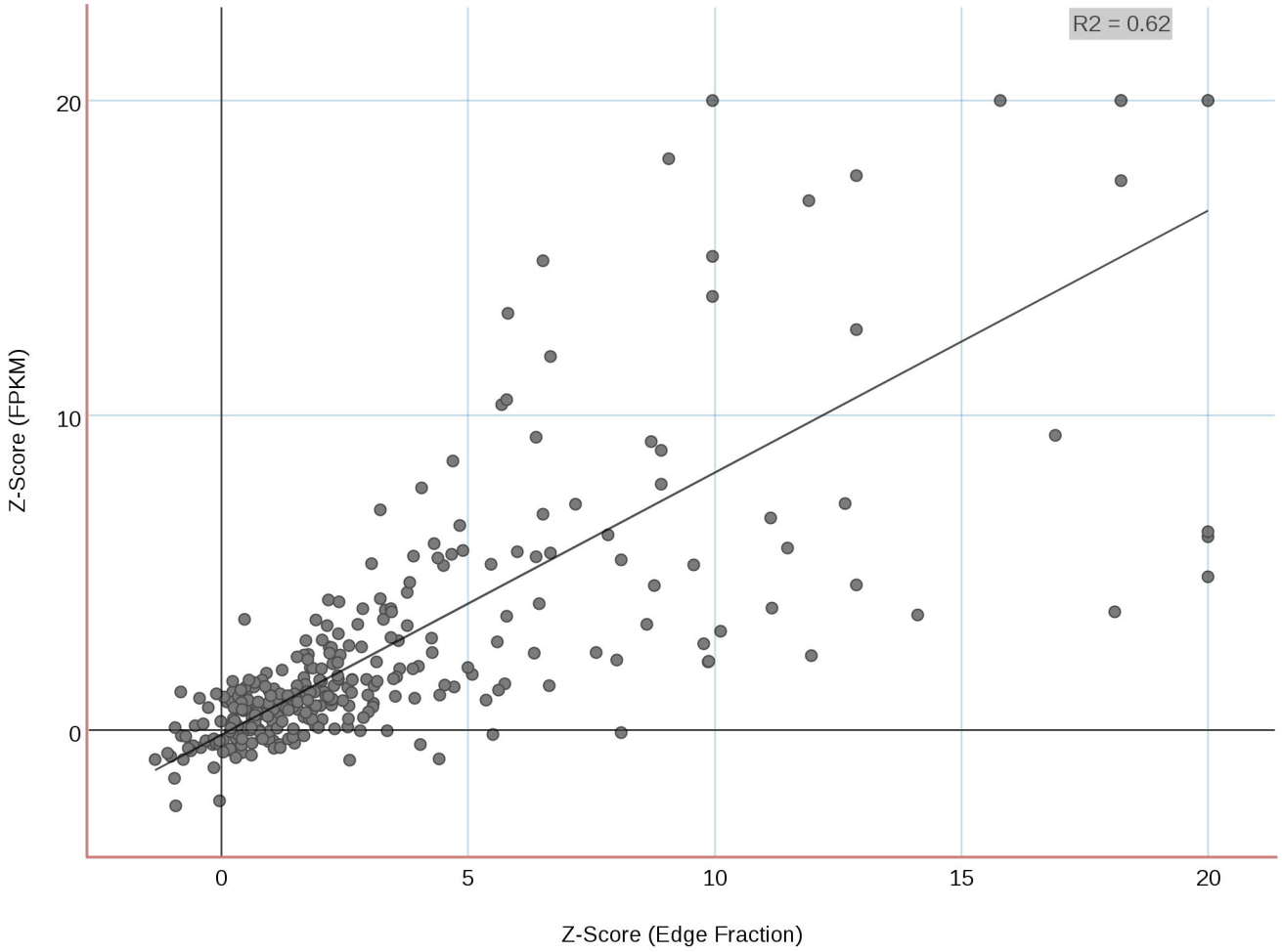
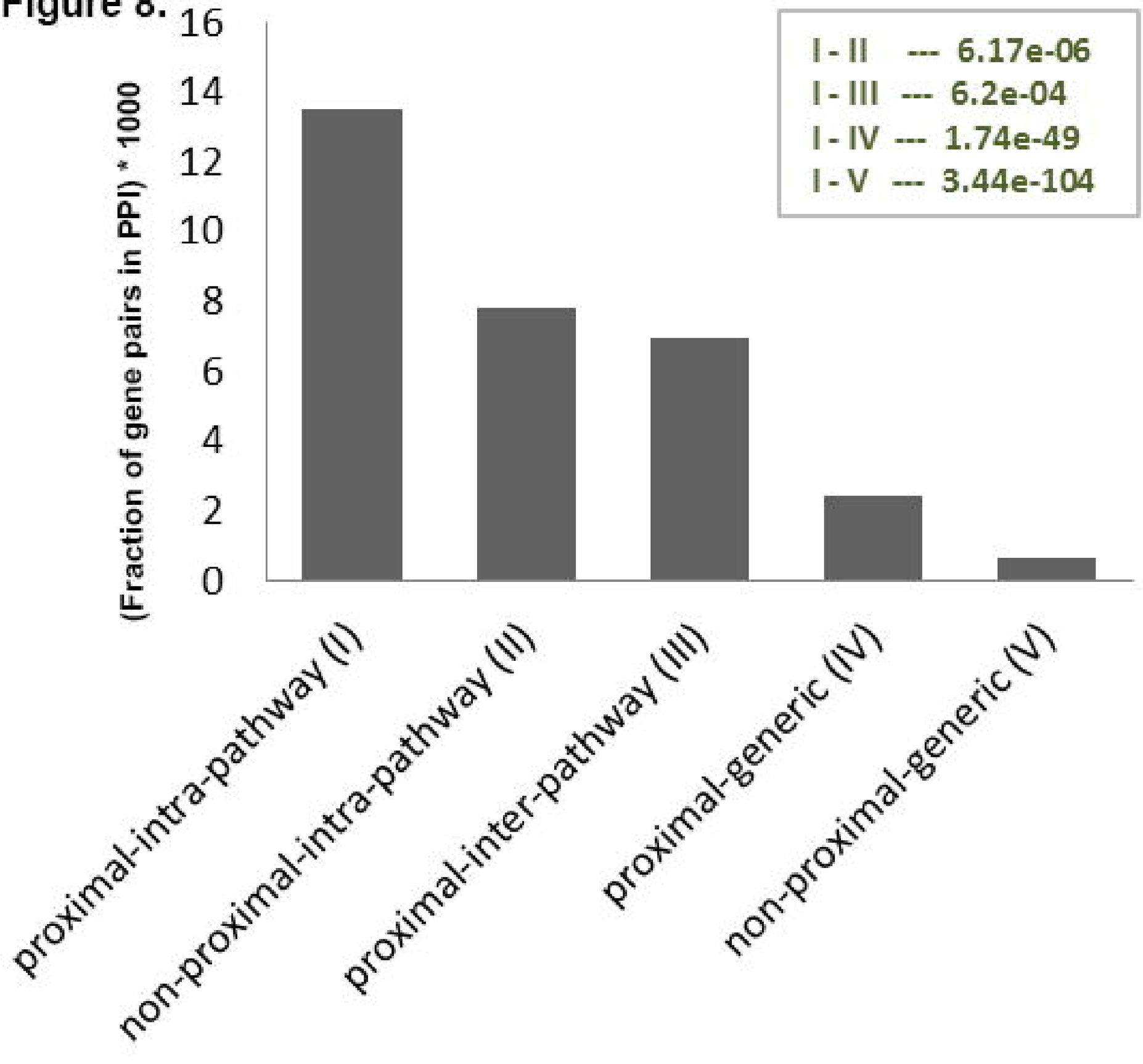
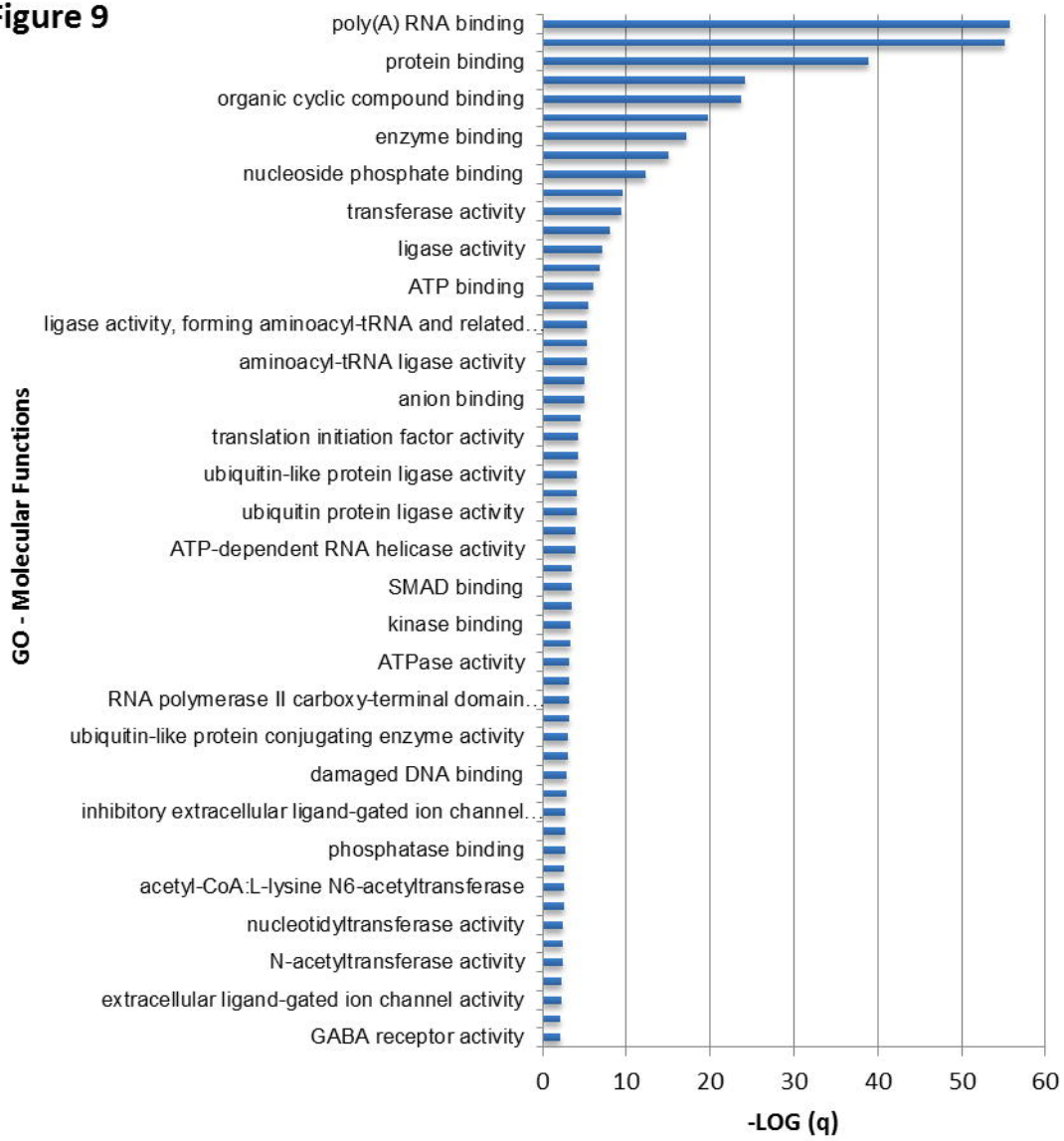


Figure 8.



POOLED 6 CELL LINES

Figure 9


poly(A) RNA binding
protein binding
organic cyclic compound binding
enzyme binding
nucleoside phosphate binding
transferase activity
ligase activity
ATP binding
ligase activity, forming aminoacyl-tRNA and related..
aminoacyl-tRNA ligase activity
anion binding
translation initiation factor activity
ubiquitin-like protein ligase activity
ubiquitin protein ligase activity
ATP-dependent RNA helicase activity
SMAD binding
kinase binding
ATPase activity
RNA polymerase II carboxy-terminal domain..
ubiquitin-like protein conjugating enzyme activity
damaged DNA binding
inhibitory extracellular ligand-gated ion channel..
phosphatase binding
acetyl-CoA:L-lysine N6-acetyltransferase
nucleotidyltransferase activity
N-acetyltransferase activity
extracellular ligand-gated ion channel activity
GABA receptor activity
GO - Molecular Functions
0
10
20
30
40
50
60
-LOG (q)