Unifying Gene Expression Complexity and Cellular Operational Efficiency: Application of the Locality and Caching Principles via a Cell-to-Computer Analogy

Wen Jiang[a], Denis Feliers[d], W. Jim Zheng[e], Fangyuan Zhang[c] and Degeng Wang[a,b,f]

[a]Department of Environmental Toxicology, [b]The Institute of Environmental and Human Health (TIEHH), [c]Department of Mathematics & Statistics, Texas Tech University, Lubbock, TX 79409.
[d]Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, CA 92121
[e]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin Street, Suite 600, Houston, TX 77030.

[f]: Corresponding author. Email: degeng.wang@ttu.edu.

Abstract

Gene expression is time-consuming, and the delay from transcription activation to produced proteins is sequentially longer from bacteria to yeast and to humans. How human cells bypass the delay and attain operational efficiency, *i.e.*, quick proteomic response to signals, is not well understood. The computer has endured the same system latency issue due to much slower information retrieval (hard drive (HD) to memory and to CPU) than CPU execution, and mitigated it via efficient memory management, namely, the spatiotemporal locality principles that control specialized user functions and the permanent caching of core system functions – the operating system (OS) kernel. Thus, in this study, we unified gene expression and HD-memory-CPU information flow as instances of the Shannon information theory, both supporting the respective system operations and consisting of three components: information storage, the execution/decoding step, and the channel for the dynamic storage-to-execution information flow; the gene expression machinery and their regulators, and the OS kernel, were deemed as the respective channels. This abstraction prompted a multi-omic comparative analysis, generating experimental evidence that transcriptome regulation shares the computer memory management principles. First, the temporal locality principle explains the mRNA stabilization-by-translation regulatory mechanism and controls specialized cellular functions. Second, the caching principle explains cytoplasmic mRNA sequestration and the defiance of the locality principle by highly sequestered mRNAs. Third, strikingly, in both systems, the caching principle controls the information channels; similar to permanent caching of OS kernel, basic translation/transcription machinery and their regulators are the top most sequestered mRNAs. Summarily, the locality and the caching principles differentially regulate specialized functions and core system functions, respectively, integrating the complexity of transcriptome regulation with cellular operational latency mitigation.

Introduction

Gene expression is one of, if not the, most fundamental cellular processes. It goes hand-in-hand with the control of essentially all biochemical pathways, activating and inhibiting a pathway by increasing and decreasing the abundance of corresponding proteins, respectively. On the other hand, the multi-stepped process is time consuming, and sequentially more so from bacteria to yeast and to mammalians; for instance, transcription elongation rate is 42-54 nucleotides per second in bacteria, but only 22-25 nucleotides per second in human cells. This prolonged delay from transcription activation to protein production in human cells would, if not mitigated, render it impossible for the cells to make rapid responses to environmental signals – a severe system latency issue. On the other hand, the gene expression process consumes ATP and GTP as energy sources, and nucleotides and amino acids as building blocks, levying significant metabolic overhead on the cells. Thus, gene expression regulation is a critical system optimization problem; human cells must have evolved elaborate regulatory strategies to strike the optimal balance, *i.e.*, bypassing the latency and achieving operational efficiency without incurring excessive wasteful metabolic expenditure.

At the mRNA level, extensive studies have characterized many regulatory mechanisms, and thus revealed tremendous complexity; for instance, the binding site mapping for hundreds of RNA binding proteins and the identification of a large number of non-coding RNAs (1,2). MicroRNA (miRNA), perhaps the most studied non-coding RNAs, controls >60% of mRNAs (3). While transcription regulation is the primary controller of mRNA production, these posttranscriptional regulators control mRNA translation and/or degradation via cognate binding sites, which are mostly in mRNA un-translated regions (UTR), especially 3'-UTR. The prevalence of these regulatory mechanisms is underscored by the observation that UTR occupies, on average, 50% of the length of a mRNA (4); in case of the CREB1 gene, it is 90%. This complexity is manifested as various levels of discrepancies among key gene expression parameters, namely, the transcription rate, mRNA abundance, mRNA translation activity and protein abundance (5-14). The discrepancy was noticed prior to the genomic era (15,16), observed by the first simultaneous genome-wide analysis of multiple gene expression parameters (17), and then quickly confirmed in numerous studies (18-26).

Currently, our understanding of transcriptome regulation is fragmentary, and we have not gotten to fully comprehend the functional advantages these regulatory mechanisms confer to the cells. While biochemical details of many regulatory mechanisms are known, it is unlikely that individual mechanisms always act alone. Instead, multiple mechanisms might act in concert with one another to support efficient execution of specific cellular processes. It is critical to uncover the underlying principles for the cellular orchestration of multiple regulatory mechanisms among all the genes.

Thus, in this study, we strived to understand how the cells mitigate the potential system latency at the mRNA level, *i.e.*, the time delay from transcription to mRNA translation during the gene expression process. To achieve the mitigation, the cells need to orchestrate transcription, mRNA degradation and mRNA translation activities among all the genes. Efficient mitigation of the latency has two seemingly contradictory objectives – expression of mRNAs in advance of cellular needs for their proteins and minimization of wasteful expression of mRNAs that are

degraded without ever being translated. Additionally, selective degradation of existing mRNAs occur constantly, conceivably to free up resources to accommodate incoming nascent mRNAs; ideally, only those mRNAs that are neither currently translating nor expected to be so in the near future are subject to degradation. Unfortunately, we do not have a cohesive model for how the cells achieve these objectives.

This latency issue, on the other hand, is not unique to the cells. The computer information retrieval process from the hard drive (HD) to memory, and then to central processing unit (CPU) caches, is strikingly similar to the gene expression process (14,27,28). They both function to support system operation through dynamic retrieval of permanent stored digital information–the binary information in the HD in computers and the quadruple (A, T, C and G) genetic information in the genomes in cells; and both are via an intermediate step–the memory in computers and the mRNA level in the cells. Similar to the tardiness of gene expression in human cells, the HD-to-CPU information retrieval is much slower than CPU execution cycle; the CPU might stay idle for extended periods of time while waiting for information for the next cycle–a latency issue once severely plaguing the computing community (29). Memory management/optimization principles, namely, spatiotemporal locality and caching, were developed; the goal is to enable speculative loading and purging of information, maximizing the possibility that the memory only contains information currently used, or to be used in the near future, by the CPU (29,30). Implementation of these principles successfully mitigated the latency issue in the computers. Consequently, that these memory management and optimization principles also apply to transcriptome regulation becomes an interesting, and powerful, hypothesis.

To test the hypothesis, we performed a multi-omics comparative study of gene expression to the HD-to-CPU information flow. Our results support the hypothesis. First, the gene expression process becomes sequentially more selective, *i.e.*, focuses on sequentially less genes, from transcription to translation; the trend is similar to the sequential HD-to-CPU capacity decreases, and thus selectivity increases, in computers. Second, the temporal locality principle that governs selective memory information purging in computers is equivalent to the mRNA stable-when-translating regulatory mechanism, and controls many cellular functions, as exemplified by major metabolic pathways. Third, similar to permanent memory caching of the operating system (OS) kernel in computers, many cellular functions, as exemplified by basal transcription and translation machineries, defy this regulatory mechanism. Instead, high levels of mRNA cytoplasmic sequestration, *i.e.*, stable but not translating, were observed. Strikingly, these functions and the computer OS kernel are equivalent to each other, acting as the information flow channel in the respective system; that is, memory caching is equivalent to mRNA sequestration. Summarily, the locality and the caching principles, and their centrality to computer system optimization, enabled an integrative view of the stabilization-by-translation and the cytoplasmic sequestration of mRNAs in the context of cellular latency mitigation, integrating the complexity of transcriptome regulation with cellular operational efficiency.

Materials and Methods

Tissue Culture and mRNA Isolation for RNA-seq Analysis

As previously described (31-33), the human HCT116 cells were cultured in a serum-free medium (McCoy's 5A (Sigma) with pyruvate, vitamins, amino acids and antibiotics) supplemented with 10 ng/ml epidermal growth factor, 20 μg/ml insulin and 4 μg/ml transferrin. Cells were maintained at 37 °C in a humidified incubator with 5% CO_2.

To extract mRNA for RNA-seq analysis, RNeasy kit (Qiagen) was used to extract total RNA from the HCT116 cells according to manufacturer's specification. GeneRead Pure mRNA Kit (Qiagen) was then used to isolate mRNA from the total RNA for Illumina NGS sequencing according to manufacturer's specification.

GRO-seq Analysis

Global run-on was done as previously described (4,34-37). Briefly, two 100 cm plates of HCT116 cells were washed 3 times with cold PBS buffer. Cells were then swelled in swelling buffer (10 mM Tris-pH 7.5, 2 mM MgCl2, 3 mM CaCl2) for 5 min on ice. Harvested cells were re-suspended in 1 ml of the lysis buffer (swelling buffer with 0.5% IGEPAL and 10% glycerol) with gentle vortex and brought to 10 ml with the same buffer for nuclei extraction. Nuclei were washed with 10 ml of lysis buffer and re-suspended in 1 ml of freezing buffer (50 mM Tris-pH 8.3, 40% glycerol, 5 mM MgCl2, 0.1 mM EDTA), pelleted down again, and finally re-suspended in 100 μl of freezing buffer.

For the nuclear run-on step, re-suspended nuclei were mixed with an equal volume of reaction buffer (10 mM Tris-pH 8.0, 5 mM MgCl2, 1 mM DTT, 300 mM KCl, 20 units of SUPERase-In, 1% Sarkosyl, 500 μM ATP, GTP, and Br-UTP, 2 μM CTP) and incubated for 5 min at 30 °C. Nuclei RNA was extracted with TRIzol LS reagent (Invitrogen) following manufacturer's instructions, and was resuspended in 20 μl of DEPC-water. RNA was then purified through a p-30 RNAse-free spin column (BioRad), according to the manufacturer's instructions and treated with 6.7 μl of DNase buffer and 10 μl of RQ1 RNase-free DNase (Promega), purified again through a p-30 column. A volume of 8.5 μl 10 × antarctic phosphatase buffer, 1 μl of SUPERase-In, and 5 μl of antarctic phosphatase was added to the run-on RNA and treated for 1 hr at 37 °C. Before proceeding to immuno-purification, RNA was heated to 65 °C for 5 min and kept on ice.

Anti-BrdU argarose beads (Santa Cruz Biotech) were blocked in blocking buffer (0.5 × SSPE, 1 mM EDTA, 0.05% Tween-20, 0.1% PVP, and 1 mg/ml BSA) for 1 hr at 4 °C. Heated run-on RNA (~85 μl) was added to 60 μl beads in 500 μl binding buffer (0.5 × SSPE, 1 mM EDTA, 0.05% Tween-20) and allowed to bind for 1 hr at 4 °C with rotation. After binding, beads were washed once in low salt buffer (0.2 × SSPE, 1 mM EDTA, 0.05% Tween-20), twice in high salt buffer (0.5% SSPE, 1 mM EDTA, 0.05% Tween-20, 150 mM NaCl), and twice in TET buffer (TE pH 7.4, 0.05% Tween-20). BrdU-incorporated RNA was eluted with 4 × 125 μl elution buffer (20 mM DTT, 300 mM NaCl, 5 mM Tris-pH 7.5, 1 mM EDTA, and 0.1% SDS). RNA was then extracted with acidic phenol/chloroform once, chloroform once and precipitated with ethanol overnight. The precipitated RNA was re-suspended in 50 μl reaction (45 μl of DEPC water, 5.2 μl of T4 PNK buffer, 1 μl of SUPERase_In and 1 μl of T4 PNK (NEB)) and incubated at 37 °C for 1 hr. The RNA was extracted and precipitated again as above before being processed for Illumina NGS sequencing.

Fragmentation of cytoplasmic extract and mRNA extraction

Cytoplasmic extract fragmentation was performed as previously described (38,39). Briefly, the HCT116 cells were incubated with 100μg/ml cycloheximide for 15 minutes, washed three times with PBS, scraped off into PBS, and then pelleted by micro-centrifugation. Cell pellet was homogenized in a hypertonic re-suspension buffer (10 mM Tris (pH 7.5), 250 mM KCl, 2 mM MgCl2 and 0.5% Triton X100) with RNAsin RNAse inhibitor and a protease cocktail. Homogenates were centrifuged for 10 min at 12,000 g to pellet the nuclei. The post-nuclear supernatants were laid on top of a 10-50% (w/v) sucrose gradient, followed by centrifugation for 90 min at 200,000 g. The polysomal and non-polysomal fractions were identified by OD_{254} and collected. RNeasy kit (Qiagen) was used to extract RNA from both fractions according to manufacture's specification. GeneRead Pure mRNA Kit (Qiagen) was then used to isolate mRNA for Illumina NGS sequencing from the RNA according to manufacture's specification.

Illumina NGS Sequencing

Sequencing libraries were generated with the Illumina TruSeq RNA Sample Preparation Kit. Briefly, RNA molecules were fragmented into small pieces using divalent cations under elevated temperature. The cleaved RNA fragments are copied into first strand cDNA synthesis using reverse transcriptase and random primers. This was followed by second strand cDNA synthesis using DNA Polymerase I and RNase H. These cDNA fragments were end-repaired using T4 DNA polymerase, Klenow polymerase and T4 polynucleotide kinase. The resulting blunt-ended fragments were A-tailed using a 3′–5′ exonuclease-deficient Klenow fragment and ligated to Illumina adaptor oligonucleotides in a 'TA' ligation. The ligation mixture was further size-selected by AMPure beads and enriched by PCR amplification following Illumina TruSeq DNA Sample Preparation protocol. The resulting library is attached and amplified on a flow-cell by cBot Cluster Generation System.

The sequencing was done with an Illumina HiSeq 2000 sequencer. Multiplexing was used to pool 4 samples into one sequencing lane. After each sequencing run, the raw reads were pro-processed to filter out low quality reads and to remove the multiplexing barcode sequences. The GRO-Seq, RNA-seq and polysome-seq datasets are available through the NCBI GEO database (accession number GSE111222). The non-polysome-Seq data is being submitted to the GEO database.

NGS Data Analysis

The sequencing reads were mapped to the UCSC hg19 human genome sequences with the STAR software (40), using the default input parameter values. For each sample, at least 80% of the reads were successfully mapped. For the sake of consistency across the four transcriptome regulation parameters, we counted the reads for each gene for the exon regions only. The counting was performed with the HTSeq-count software (41), and the counts were then transformed into Reads Per Kilo-base Per Million Mapped Reads (RPKM) values. 12921 genes have a minimal RPKM value of 1 for at least one of the three parameters, and were considered

expressed in the HCT116 cells. Linear regression of log-transformed data was used to examine consistence between biological replicate samples.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) gene sets

The KEGG and GO functional gene sets were downloaded from the Molecular Signatures Database (MSigDB) version 7.4 at the Gene Set Enrichment Analysis (GSEA) website (https://www.gsea-msigdb.org/gsea/index.jsp) (42,43). The KEGG gene set contains 186 gene sets. The GO molecular function (MF) set contains 1708 sets, and the biological process (BP) and the cellular component (CC) sets 7481 and 996 sets, respectively.

Statistical Analysis

The R open source statistical software (version 3.6.2) installed on a Mac Pro desktop computer was used for statistical analysis. Outlier identification, student t-test, descriptive statistical parameter calculation, correlation coefficient calculation, linear regression and other statistical procedure are all done with this R software.

Statistics of the sequestration index in gene pairs

We calculated the differences between the sequestration indexes ($sequ^{diff}$) of the two genes in each gene pair. For a set of gene pairs, the seq^{diff} parameter should follow a normal distribution with a mean of 0 and a variance (var) given by the following equation:

$$var_{(index1-index2)} = var_{index1} + var_{index2} - 2*cor(index1, index2)$$

For the whole set of gene pairs, the following steps are used to calculate the var and standard deviation (SD):

$$var_{index1} = var_{index2} = var_{sequestration\ index} = 0.359$$

$$cor(index1, index2) = 0$$

$$var_{(index1-index2)} = 2 * var_{sequestration\ index} = 2 * 0.359 = 0.718$$

$$SD_{(index1-index2)} = \sqrt{0.718} = 0.847$$

GO Similarity Analysis

Pairwise GO similarity scores between human genes were computed as previously described (44-46). Briefly, for each gene, we first generated GO fingerprint – a set of ontology terms enriched in the PubMed abstracts linked to the gene, along with the adjusted p-value reflecting the degree of enrichment of each term. The GO similarity score quantifies similarity between the GO fingerprints of corresponding gene pair. For detail about GO fingerprint generation and similarity calculation, please see description in previous publications (45).

Results

The analogy between gene expression and information retrieval in computers and the system latency issue

It is long noticed that the cells and the computers share many common architectural and operational features, though the former rely on controlled biochemical processes, *i.e.*, the biochemical reactions/flow in the biochemical/regulatory network, and the latter on controlled electrical processes (47-50). As briefly mentioned above, we have previously reported the similarity between gene expression and the HD-to-CPU information retrieval and their centrality to the operation of the respective systems (14,27,28). In this study, we unified the two as respective implementation of the components of the Shannon information theory: information encoding/storage, the execution level where the information is decoded into actions, as well as the channel through which the information is dynamically retrieved from the storage level to the execution/decoding level (51) (Figs. 1 A and B). In the cells, the quadruple genetic (A, T, C and G) information is stored in the chromosomes; in computers, it is the binary information stored in the hard drive (HD). At the execution/decoding level, the binary information retrieved into the CPU cache is decoded into patterns of electrical flow in the CPU to execute specific Boolean algebraic transformation in computers; in the cells, mRNAs form polysome complex with ribosomes, translating the quadruple information into proteins with specific 3-dimensional structures and functional domains to carry out corresponding biochemical functions – often to catalyze a biochemical reaction. Intriguingly, the information channel encompasses a middle step between the storage and the execution/decoding levels in both systems – the mRNA step in the cells and the main memory in computers. The molecular biology central dogma outlines the genome-mRNA-protein information flow process, and we have argued that the gene expression machineries and their regulators constitute the information channel in the cells (28); in computers, the HD-memory-CPU information flow is tightly controlled, to be discussed in details later, by the OS kernel.

As mentioned earlier, in both systems, the retrieval process inevitably imposes potential latency due to the time delay from the storage level to the execution level. Though aware that the delay becomes sequentially longer from bacteria to yeast and to humans, we do not yet systematically understand how human cells overcome the latency.

On the other hand, we fully understand this issue and its mitigation in computers. The HD-to-memory retrieval is much slower than that from memory to CPU cache, which is, in turn, slower than a CPU decoding/execution cycle (29). The resulting latency was mitigated by key memory management and optimization principles. These principles stipulate speculative HD-to-memory loading, maximizing the likelihood that the information needed by the CPU shortly is already in the memory. Additionally, the capacity becomes smaller and smaller from HD to memory and to CPU cache (Fig. 1B). There is a constant need to purge information already in the memory to free up space for the incoming information. Hence, these principles also stipulate speculative purging, ensuring that only the information not to be needed by the CPU shortly get purged. In a word, the memory management principles maximally gear the memory towards current and near-future CPU needs (29).

However, the speculation is impossible to be entirely flawless. Counterproductive speculation inaccuracy is certain to occur. For loading, a portion of the speculatively loaded information might get purged without ever being requested by the execution step, wasting memory space and the system resource dedicated to the loading operation. Similarly, the side effect of speculative purging is the unfortunate removal of soon-to-be-needed information and having to reload it

quickly, a situation termed "thrashing" that is even more wasteful and detrimental than a speculative loading error. In general, higher levels of speculation, *e.g.*, a larger volume of loaded information per loading operation, lead to higher levels of speculation inaccuracy. For maximal latency mitigation, computer memory management always has to strike an optimal tradeoff between the incurrence of counterproductive speculation inaccuracy and other parameters of the speculation process. Whether gene expression regulation shares similar principles and techniques, as denoted by the question marks (?) in figure 1B, became extremely interesting to us (29).

A multi-omics analysis revealed sequentially higher selectivity from the storage to the execution levels – another similarity to information retrieval in computers

Thus, we further explored the computer-cell analogy experimentally by performing a multi-omics experiment, *i.e.*, simultaneous measurement of multiple transcriptome regulation parameters (Figs. 1 B and C). As usual, the RNA-seq was used to measure mRNA abundance (RA) (Fig. 1C, purple text boxes and arrows). Additionally, we used GRO-seq to measure transcription rate (TR), monitoring transcription as cellular equivalent to computer HD-to-memory information loading (4,34-37) (Figs. 1 B and C, brown text boxes, lines, and arrows). Polysome-profiling was used to measure translation activity (TA), monitoring translation as the equivalent to information fetching from memory to CPU cache (38,39) (Figs. 1 B and C, light blue text boxes, lines, and arrows). And we used TR and RA together, to be described in detail later, to estimate mRNA degradation rate (Figs. 1 B and C, green lines and arrows). The TR, RA, and TA data were previously described (4,37). In this study, we analyzed them in the context of the analogy between gene expression and computer information retrieval. Additionally, the cytoplasmic non-polysome-associated mRNA abundance (NP) data was incorporated into this analysis to enable, to be discussed in detail later, a more comprehensive analysis of the analogy (Figs. 1 B and C, blue text boxes, lines, and arrows).

As reported, our analysis revealed complexity of transcriptome regulation in the form of discrepancy among the gene expression parameters (Figs. 2 and 3). A trend of sequentially enhanced gene expression selectivity from transcription to translation was observed. Two aspects of the boxplot in figure 2 illustrate the trend schematically. First, the median parameter RPKM values decrease from TR to RA, to NP, and to TA, with all Mann-Whitney-U-test p-values smaller than 1E-16. Second is the sequential TR-RA-NP-TA increases of the variances of the distributions, with F-test p-values all smaller than 1E-16; sequentially more genes are expressed at extreme (high or low) levels. The trend is also shown quantitatively (Table 1). In a word, the cells focus the gene expression resources, *e.g.*, the building blocks and the RNA binding proteins, on sequentially less genes as the genetic information flow from the genome to the proteome.

This trend is, as previously reported (4), also reflected in the slopes in linear regression analyses (Fig. 3). The $\log_2$(RA) *vs.* $\log_2$(TR) regression gave a slope of 1.17 – significantly greater than 1 with a smaller than 1E-16 p-value, reflecting that more genes exhibit extreme high or low values in $\log_2$(RA) than $\log_2$(TR) (Fig. 3A). Not surprisingly, the $\log_2$(TA) *vs.* $\log_2$(RA) regression gave a slope of 1.1 – also significantly greater than 1 with a smaller than 1E-16 p-value, reflecting that more genes exhibit extreme high or low values in TA than RA (Fig. 3C). The trend is

significantly beyond experimental noise, as illustrated by the two RA biological replicates in figures 3 A and C.

Put in other words, this is another architectural similarity of gene expression to the computer information retrieval process. As shown in figure 1B, the capacity becomes sequentially smaller from hard drive to memory and to CPU cache in computers (29). Hence, the information retrieval process has to become sequentially more selective, *i.e.*, focusing on smaller sets of information, necessitating efficient memory management and usage.

Given the overall architectural analogies from gene expression to computer information retrieval as well as the importance of memory management in system operational efficiency in computers, we explored our dataset to investigate whether and how much the computer memory management principles would shed insights into gene expression regulation and cellular operations.

The locality principles: spatial locality and transcriptional coregulation

Next, we investigated the applicability of the spatiotemporal locality principles – the spatial locality for speculative HD-to-memory information loading and the temporal locality for information purging (30) (Fig. 1B) – in explaining common observations in gene expression analysis.

The spatial locality principle governs HD-to-memory information loading (29,30) (Fig. 1B). It takes advantage of the observation that the CPU tends to repeatedly request information from the same HD neighborhood. Upon a CPU memory request miss, *i.e.*, the requested information is absent from the memory, the computer will speculatively load the whole neighborhood of the requested information instead of just the requested piece. Such speculative loading reduced the latency, as the information needed by the CPU in the near future is already pre-loaded into the memory. As the CPU will likely request the other information in the neighborhood soon, wasteful loading is also minimized; that is, the preloaded information will likely be used prior to being swapped out of the memory. In a word, the spatial locality principle strikes an optimal balance between the need for speculative loading prior to CPU request to minimize latency and the incurrence of wasteful loading. As the answer to the first question (?[1]) in figure 1B, this is obviously equivalent to co-transcriptional regulation of functionally related genes in the cells. In bacteria, the genes are often organized into poly-cistron operons, which is analogous to the organization of related information into a continuous string in a flat file in computers. In eukaryotic, though each gene is an independent transcription unit, functionally related genes often share common index, *e.g.*, transcription factor binding sites; this scheme is analogous to the organization of related information into a computer database file in that each piece of information is directly addressable, but related pieces share common addressing indices.

The locality principles: temporal locality and the mRNA stabilization-by-translation regulatory mechanism

The temporal locality principle governs the other aspect of memory management–identification of information already in the memory for purging to free up space for incoming information

(29,30) (Fig. 1B). This principle takes advantage of the observation that the CPU requests and usage tend to cluster at specific memory locations within a short period of time–hence the temporal locality. Thus, only memory locations not recently used by the CPU are identified for purging, to avoid having to quickly re-load them.

Degradation of mRNAs is the cellular equivalent to the memory information purging (Fig. 1B), and elaborate control of this process should be crucial for efficient cellular operation (52). Briefly, stabilization of an mRNA reduces the need for its transcriptional production and the associated energetic and metabolic overhead, as transcription consumes significant amount of ATP and nucleotide building blocks. It also enables the cell to bypass the time-consuming steps of transcription, processing and nucleus export for quick protein production. On the other hand, over stabilization of mRNAs generates added regulatory demands, for instance, to guard against aberrant protein production from the mRNAs, levying extra pressure on the regulatory machinery such as the RNA-binding proteins. Thus, mRNA degradation control should be a major factor in the "economics" of transcriptome regulation.

Stabilization-by-translation is a major mRNA degradation control mechanism (53). It is functionally equivalent to the computer temporal locality principle in that they both protect information that is frequently requested/used at the execution level from being purged out of the intermediate step of the retrieval process. For instance, mRNAs with optimized amino acid codon composition generally have higher translation efficiency and, thus, are more stable (54). One major underpinning of this mRNA stability control is the EIF4F complex, which integrates the 7-methylguanosine-cap structure at the 5' end of mRNAs and the PolyA tails at the 3' end, recruits ribosomes and prevents exonuclease-mediated mRNA decay (55,56). Stabilization of actively translating mRNAs is intuitively understandable, as an actively translating mRNA is likely producing mission critical proteins. Moreover, the genome-to-proteome trend of enhanced selectivity underscores the importance of studying mRNA degradation, the causative regulatory force for the TR-RA discrepancy described in figures 3 and 4A.

Thus, to address the 2nd question (?^{2}) in figure 1B, we hypothesized this stable-when-translating regulatory mechanism as the cellular equivalent to the temporal locality principle. As previously described, our dataset enabled a test of this hypothesis (4). First, the steeper-than-1 slope of the RA *vs.* TR regression (Fig. 3A) reflects the increased selectivity from TR to RA due to mRNA degradation. It is shown that RNA processing is extensively coupled to transcription and that transcription is on average more than three folds slower than RNA processing (10,11,57); for most genes, transcription is the rate-limiting step in mRNA production, and TR closely correlates with mRNA production rate (11). Thus, TR was used as mRNA production rate in this analysis. Under our experimental condition, most mRNAs should be in steady state condition and follow the equations below:

$$\frac{dRA}{dt} = TR - a*RA = 0 \quad \text{a: the degradation rate}$$

$$TR = a*RA$$

$$a = RA/TR$$

That is, for most mRNAs, the log-ratio $\log_2(RA/TR)$ should estimate the degradation activity (Figs. 1 B and C, purple lines and arrows); low degradation activity leads to higher RA than TR and thus positive log-ratio values, and high degradation activity the opposite. Second, the steeper than 1 slope of the TA *vs.* RA regression (Fig. 3C) reflects the increased selectivity from RA to

TA due to translation regulation; high translation activity leads to higher TA than RA, and low translation activity the opposite. The $\log_2$(TA/RA) log-ratio was used, as commonly practiced, as mRNA translation index.

To test the hypothesis, we computed predicted RA (RA^{Pred}) and TA (TA^{Pred}) values by the respective $\log_2$(RA) *vs.* $\log_2$(TR) (Fig. 3A) and $\log_2$(TA) *vs.* $\log_2$(RA) (Fig. 3C) regressions, *i.e.*, the fitted values (or the corresponding points) on the regression lines; the regression lines masked out the scatteredness of the data points due to the function-specific patterns of gene expression regulation and experimental noise, thus better reflecting the TR-to-RA and RA-to-TA increases in gene expression selectivity. We then computed the $\log_2(RA^{Pred}/TR)$ and $\log_2(TA^{Pred}/RA)$ log-ratios. If our hypothesis is correct, a positive correlation between the two was expected. This is, indeed, the case. As shown in a scatter plot in figure 4A, the two log-ratios have a correlation coefficient of 0.7, with a p-value lower than 1E-16, supporting mRNA stabilization-by-translation as the answer to the second question ($?^2$) in figure 1B.

Nevertheless, masking out the expected function-specific pattern of gene expression regulation is a significant drawback. Thus, a scatter plot of $\log_2$(TA/RA) *vs.* $\log_2$(RA/TR) is also shown (Fig. 4B). While positive correlation between the two log-ratios still exists as previously shown, the data points became more scattered. The correlation coefficient decreases to 0.39, reflecting the effect of function-specific mode of regulation.

The function-specific pattern was illustrated with the mRNAs for proteins in the TCA cycle, the lysosome and the ribosome (Figs. 3B, 3D and 4C). The three mRNA groups all exhibited overall higher-than-expected $\log_2$(RA) levels (Fig. 3B), but the ribosome mRNAs exhibited lower-than-expected TA levels (Fig. 3D). Consistently, ribosome mRNAs showed lower-than-expected $\log_2$(TA/RA) values, and lower $\log_2$(TA/RA)-$\log_2$(RA/TR) correlation than the lysosome and the TCA-cycle mRNAs (Fig. 4C). In short, the ribosome mRNAs defy the stabilization-by-translation regulatory mechanism, *i.e.*, the temporal locality principle, while the TCA cycle and the lysosome mRNAs follow it.

Next, we strived for a transcriptome-wide delineation of the function-specific pattern. We also aimed to uncover the alternative mechanism that underpins the defiance of the mRNA stabilization-by-translation regulatory mechanism and thus the modesty of the $\log_2$(TA/RA)-$\log_2$(RA/TR) correlation. As described in detail below, there is another aspect of computer memory management – the differential management of user programs such as an Internet browser and core system programs (the OS kennel) (Fig. 1B). We thus investigated whether and how the cells also differentially regulate mRNAs for specific functional domains and those for core cellular functions that are critical for all cellular processes, that is, to answer the 3rd question ($?^3$) in figure 1B.

<u>Caching of OS kernel in computers and cytoplasmic sequestration of mRNAs for the gene expression machinery and their regulators in the cells</u>

In computers, the temporal locality principle does not control all software programs; some are permanently cached in the memory (29). The software programs are divided into two categories: user programs and core system programs commonly known as the OS kernel, which use different

memory areas (Fig. 1B). The two are controlled by different policies. User programs, as exemplified by Internet browsers and Microsoft Word, are developed for special purposes, not allowed to access kernel memory area and controlled by the spatiotemporal locality principles. To the contrary, the kernel manages all the resources of the computer, *e.g.*, CPU time, the memory and the hard drive. Instead of focusing on individual user programs, the kernel ensures a robust running environment for all of them; individual user program spontaneously interacts with the kernel indirectly, *e.g.*, via OS system calls, to either request services for its own running needs or respond to kernel demands. When a kernel service is requested, a memory miss is operationally much costlier, potentially crashing the system. Thus, the kernel is permanently cached in the designated kernel memory area, as opposed to being controlled by the locality principles, to ensure its constant availability.

Similarly, the moderate $\log_2$(TA/RA)-$\log_2$(RA/TR) correlation (Fig. 4B) suggests that many mRNAs defy the stabilization-by-translation regulatory mechanism, as exemplified by the ribosome protein mRNAs (Fig. 4C). Consistently, many other post-transcription regulatory mechanisms control mRNA stability in such manners, in that they stabilize mRNAs independent of, sometimes even accompanied by inhibiting, mRNA translation activity. That is, they maintain caches of mRNAs that are sequestered away from translation. For instance, there are a large number (> 400) of RNA-binding proteins (RBP), many of which inhibit translation while the inhibited mRNAs remain stable; for instance, the regulation of insulin-like growth factor 2 (IGF2), beta-actin (ACTB) and beta-transducin repeat-containing protein (BTRC) mRNAs by insulin like growth factor 2 mRNA binding protein 1 (IGF2BP1) (58). Additionally, the microRNA (miRNA) regulatory system suppresses the translation of their target mRNAs and destines the mRNAs for sequestration in the p-body and/or the GW-body (59). While some target mRNAs are de-capped and degraded, a significant portion remains stable. The sequestration regulatory actions might occur at specific dedicated sub-cellular foci. For instance, the GW-body stores translationally inactive mRNAs (60). The P-body is also a repository for such mRNAs, though a portion of stored mRNAs is degraded. Another example is the cellular formation of stress granules upon exposure to environmental stressors. Generally, these regulatory mechanisms act through cognate binding sites in the mRNA 3'- or 5'-untranslated regions (UTR). As described earlier, their prevalence and importance in human cells are underscored by the observation that, on average, ~50% of a human mRNA are UTRs (4).

The caching principle in computer is strikingly similar to the mRNA cytoplasmic sequestration, in that they both retain selective information in the intermediate step regardless of their request/usage by the execution/decoding step. Thus, we investigated the applicability of the caching principle, *i.e.*, explaining the mRNA sequestration and the modesty of the $\log_2$(TA/RA)-$\log_2$(RA/TR) correlation. For this purpose, it is necessary to determine which, and to what extent, individual mRNAs are subject to sequestration. Thus, we incorporated the cytoplasmic non-polysome-associated mRNA data into the analysis (Figs. 1 B and C, light blue textboxes, lines and arrows). A $\log_2$(NP) *vs.* $\log_2$(RA) scatter plot is shown, in comparison with the $\log_2$(RA1) *vs.* $\log_2$(RA2) plot, in figure 5A. Once again, a function-specific pattern was observed (Fig. 5B). The ribosomal protein mRNAs exhibit higher NP relative to RA values, which is consistent with their low TA values and defiance of the stabilization-by-translation regulatory mechanism. To the contrary, as shown in figures 3D and 4C, the TCA cycle and lysosome mRNAs have relatively high TA values and are regulated by the stabilization-by-translation mechanism.

Consistently, they exhibit relatively low NP values. This pattern supports the notion that cytoplasmic sequestration level reflects whether the mRNA is controlled by the stabilization-by-translation regulatory mechanism.

To systematically examine this trend, we calculated a mRNA sequestration index. This is done by first performing a $\log_2(NP) \sim \log_2(RA)*\log_2(TA)$ linear regression, thus taking both RA and TA into consideration. The residue of this regression – the difference between observed $\log_2(NP)$ value and expected value – quantifies raw mRNA cytoplasmic sequestration. It is well known that mRNA ORF length affects ribosome density on the mRNA and thus its distribution among the fractions in a cytoplasmic sucrose gradient fractionation experiment (61,62). As shown in figure 5C, our analysis also observed the effect of ORF length. The raw mRNA sequestration level negatively correlates with $\log_2$(ORF length). The negative correlation was illustrated by a loess regression. As expected, the ribosome protein mRNAs exhibited higher sequestration levels than the TCA cycle and lysosome mRNAs (Fig. 5D). Thus, we adjusted the raw sequestration level with the loess regression, *i.e.*, using the loess regression residue as our mRNA sequestration index, eliminating the ORF-length imposed complication in downstream functional analysis.

To test whether the function-specific mRNA sequestration pattern illustrated in figure 5D is genome wide, we calculated the differences between the two sequestration indices (index1 – index2), termed $sequ^{diff}$, in individual gene pairs. We then performed two systematic unbiased analyses of the $sequ^{diff}$ parameter. In the first analysis, we also calculated the similarities of Gene Ontology (GO) fingerprints of the two genes (see Materials and Methods for detail). Then, the distribution of $sequ^{diff}$ of gene pairs with different levels of GO similarities was compared. The $sequ^{diff}$ should follow normal distributions with a mean of 0 and variances given by the following equation:

$$var_{(index1-index2)} = var_{index1} + var_{index2} - 2*cor(index1, index2)$$

If the function specific pattern exists, the correlation between index1 and index2 should increase as GO similarity increases. Consequently, the $sequ^{diff}$ distribution should exhibit decreasing variance, *i.e.*, the level of dispersion (see Materials and Methods for detail). As shown in figure 6A, this is indeed the case. The box-plots all indicated the same mean value of 0, but exhibited steadily decreasing levels of dispersion, *i.e.*, lower and lower proportions of gene pairs have significantly non-zero $sequ^{diff}$ values, along with increasing GO similarities. To display the trend quantitatively, the gene pairs were ordered by the GO similarity score and binned, with a bin size of 10000 and a 50% overlap between adjacent bins. As shown by a scatter plot in figure 6B, as the median GO similarity score in the bins increases, the SD of $sequ^{diff}$ in the bins decreases; the blue horizontal line indicates the SD of the whole distribution (0.847) (see Materials and Methods for detail), and the black line the SD for gene pairs with no significant GO similarity. When the GO similarity scores are lower than 50, the data-points displayed scatteredness, though the overall trend is obvious. Above the similarity score of 50, the scattering disappears. Consequently, the median distances between the two sequestration indices of the gene pairs, *i.e.*, the absolute values of $sequ^{diff}$, decreases with increasing GO similarity scores (Figure 6C). In the second analysis, we used the protein-protein interaction data. The same trend as that in figure 6A was observed. The distributions of $sequ^{diff}$ exhibited lower dispersion for gene pairs with interacting proteins, and the dispersion decreased along with increasing interaction quality scores (Figure 6D). Thus, a systematic function-specific pattern of mRNA sequestration was observed.

Next, to test whether high mRNA cytoplasmic sequestration levels are associated with the defiance of the stable-when-translating regulatory mechanism, the genes were ordered by their sequestration index. A sliding window analysis of the $\log_2$(TA/RA)-$\log_2$(RA/TR) correlation coefficient versus the sequestration index is shown in figure 7A, displaying a clear negative correlation. When the sequestration index is low, the genes have high $\log_2$(TA/RA)-$\log_2$(RA/TR) correlation, obeying the stable-when-translating regulation and thus the temporal locality principle. As the sequestration index increases, the $\log_2$(TA/RA)-$\log_2$(RA/TR) correlation steadily decreases. And mRNA cytoplasmic sequestration, thus the caching principle, progressively becomes the dominant regulatory mechanism. Additionally, we adjusted the $\log_2$(RA/TR) log-ratio with the sequestration index based on a linear regression ($\log_2$(TA/RA) vs $\log_2$(RA/TR)*(sequestration index)). Not surprisingly, the adjustment resulted in a much improved correlation with the $\log_2$(TA/RA) index (Fig. 4D). These results support cytoplasmic mRNA sequestration as cellular equivalent to the caching of OS kernel in computer and the mechanistic underpinning for the defiance of the stable-when-translating regulatory mechanism.

Moreover, our dataset provided an opportunity to identify cellular functions that are controlled by the cytoplasmic mRNA sequestration mechanism. Whether this mechanism regulates cellular functions that are equivalent to those of OS kernel in computers could then be determined (Fig. 1B). Thus, we downloaded the curated KEGG functional groups and calculated mean sequestration index value for each group. We also calculated the $\log_2$(TA/RA) *vs.* $\log_2$(RA/TR) correlation for each group. As shown in figure 7B, a negative correlation between the sequestration index value (in the form of t-scores) and the $\log_2$(TA/RA)-$\log_2$(RA/TR) correlation was observed. Significantly over sequestered groups exhibited lower correlation than under sequestered groups (Fig. 7C). Once again, a function-specific pattern was observed. As shown by the boxplot in figure 7D, the distribution of the t-scores of the KEGG groups exhibited significantly higher level of dispersion than those of same-sized groups of randomly selected genes.

Table 2 listed the respective top-20-ranked functional groups with highest- or lowest-sequestration levels. Metabolic pathways dominate under-sequestrated groups; 9 of the top 20 are metabolic functions, with TCA cycle and glycolysis ranked at the 6th and 5th, respectively. Some sub-cellular organelles are also under-sequestered; lysosome was ranked the 1st , and proteasome the 20th (Table 2).

On the other hand, the gene expression machineries and cellular signaling dominate over-sequestrated functions (Table 2). Ribosome and basal transcription factors were ranked at the 1st and 11th, respectively; ubiquitin mediated proteolysis was rated at the 17th. 10 of the top 20 functions are cellular signaling related; the WNT, MAPK, JAK-STAT and mTOR signaling were ranked at the 2nd, 6th, 12th and 18th.

<u>Equivalence between computer OS kernel and cellular gene expression machinery and their regulators as the information flow channel</u>

Given the equivalence between the computer memory caching and cellular mRNA sequestration, we asked whether the functions controlled by the two – computer OS kernel and cellular gene expression machinery and their regulators – are equivalent as well. As mentioned earlier, the

gene expression machineries and their regulators act as the dynamic genome-to-proteome information flow channel. Briefly, the basic gene expression machineries, such as RNA polymerase II and ribosome, perform the task; cellular signaling, on the other hand, contributes to the specificity, *i.e.*, which genes are transcribed and which mRNAs are translated at any particular moment, by controlling cognate transcription factors and translation regulators.

This prompted us to re-examine the functions of OS kernel in computer operation. The kernel is commonly described in the literature as managing the hardware resource, *i.e.*, dynamically and optimally allocating the resource among competing user programs. Intriguingly, the hardware management functions, as we argue here, can be re-phrased as collectively carrying out the dynamic HD-to-CPU information retrieval process. Memory management is one major kernel function; the other functions are the file system and process control. The file system implements efficient hard drive information organization, keeps track of the locations of individual pieces of information and, upon requests, quickly retrieves the information into the memory; it is thus equivalent to the transcription machinery and its regulators, in that they both perform dynamic information retrieval from the storage to the intermediate step. As described above, memory management controls both information loading into, and purging out of, the memory; the spatiotemporal locality principles control part of the memory management actions. The kernel's process control function, together with the hardware-implemented cache memory controller, control the information flow from memory to CPU cache and then to CPU execution, which is equivalent to the translation step and catalysis of biochemical reactions by the produced proteins. In a word, the OS kernel, along with the cache memory controller, acts as the channel for the dynamic HD-to-CPU information flow.

Thus, the gene expression machinery and their regulators have functional equivalence to the computer OS kernel, in that they both control the dynamic storage-to-execution information flow in support of system operations. That is, they constitute the information flow channel in the respective systems. It is therefore not surprising that the machinery and their regulators are regulated by cytoplasmic mRNA sequestration – the cellular equivalent to memory caching of the OS kernel in computers. And it is likely not a coincident that the cache memory controller is hardware-implemented and the ribosome protein mRNAs are extremely sequestered (Fig. 7B), as the two are functionally equivalent to each other, performing the extremely crucial step of information retrieval from the intermediate step to the execution/decoding level.

<u>Analysis of the Gene Ontology (GO) gene sets further support control of the genome-to-proteome information channel by the mRNA sequestration regulatory mechanism</u>

Encouraged by these observations, we investigated the function-specific pattern more comprehensively by performing the same analysis with GO gene sets (Table 3). As the KEGG functional group typically comprises of genes for different categories of biochemical activities, molecule-function-specific patterns might get lost in the above analysis. Indeed, the analysis of GO molecular function (MF) gene sets revealed such cases, and provided further evidence for cytoplasmic sequestration of mRNAs for gene expression machineries and their regulators. Transcription regulation and cellular signaling functions dominate the top 30 GO MF terms. Most significantly, 11 of the 30 terms are related to transcription regulation; the top 4 ranked terms are all transcription factor (TF) related (Table 3, red text). Instead of forming their own

biological process/pathway module, these transcriptional regulators often participate in different biochemical or signaling pathways. Thus, they are missed in our analysis of the KEGG biological process gene sets. Additionally, 4 of the 25 terms (ranked at 16-18th, and 21st) are related to cellular signaling (Table 3, green text).

The GO cellular component (CC) and biological process (BP) gene set analysis further confirmed over-sequestration of mRNAs for component of the information flow channel, as categorized below:

- The transcription and the translation machineries, RNA binding proteins(RBP), translation regulators and protein transportation (Table 3, blue text). In addition to the transcription factors/regulators, other components of the gene expression process are also enriched. Notably, 3 of the top 5 GO CC terms (1st, 4th and 5th) are all ribosome related; ranked at the 2nd is the mRNA-RBP complex. The term "fibrillar_center" is related to ribosome biogenesis and ranked at the 23rd. The "nuclear _speckle", the sub-nuclear structure enriched of proteins involved in transcription and mRNA splicing, is ranked at the 28th. As for GO MF terms, two translation related terms are ranked at the 10th and 12th. And three of the GO BP terms (ranked at 14th, 19th and 28nd) are related to translation regulation and protein transportation (Table 3, blue text). As expected, many transcription regulation terms are also among the GO CC and BP trms (Table 3, red text).
- The mRNA and the protein degradation machineries. Proteolysis and mRNA degradation are equivalent to information purging out of the CPU cache and the memory, respectively. Consistently, "ubiquitin-mediated-proteolysis" is the 17th-ranked most sequestered KEGG term (Table 2). Not surprisingly, two related terms are ranked at the 9th and 11th among GO MF terms (Table 3, orange text). Additionally, the GO CC gene set analysis revealed over-sequestration (ranked at the 29th) of mRNAs for the CCR4-NOT complex (Table 3, orange text).
- Intracellular signaling terms (table 3, green text). 4 of the GO CC terms are related to intracellular signaling, with "PI3-kinase complex I" and "phosphatase_complex" ranked at the 6th and 8th, respectively. And four such terms are ranked at the 3rd, 6th, 22nd and 23rd among the GO BP terms.

Moreover, the analysis of GO gene sets also revealed significant enrichment of epigenomic terms. Chromatinization and further condensation of the genome imposed further gene expression time delay and, if not mitigated, severe system latency in eukaryotic cells. Proteins for the epigenomic functions are thus part of the genome-to-proteome information flow channel. It is therefore not surprising that their mRNAs are controlled by cytoplasmic sequestration as well (Table 3, purple and italic text). The term "chromatin" and "PML_body" are the 3rd and 12th ranked most sequestered GO CC terms, respectively; the term "histone_h3_k36_trimethylation" is the 21st ranked GO BP term; and "histone_acetyltransferase_activity_h3_k23_specific" is the 24th ranked GO MF terms. Additionally, the terms "chromatin_organizationm", "covalent_chromatin_modification" and "chromatin_remodeling" are all ranked within the top 400 (236th, 290th and 352nd) out of the 7481 GO BP terms (Table S1).

In summary, we revealed equivalents between the computer OS kernel and the cellular gene expression machinery and their regulators, and the equivalents are multi-faceted. First, they are both part of the information body that dynamically flows from the storage level to the

execution/decoding level. Second, they are master controller of the dynamic information flow process. That is, as discussed earlier, they act as the information channel in the respective system. Third, they are similarly self-regulated in terms of being granted preferential status. In the computer, the kernel permanently resides in the memory, not subject to the temporal locality memory purging policy. And the cache memory controller is usually hardware implemented due to its operational importance. In the cells, the mRNAs are regulated by cytoplasmic sequestration, defying the stable-when-translating regulatory mechanism. Not surprisingly, the mRNAs for the ribosome, the cellular equivalent to the computer cache memory controller, are extremely sequestrated. In short, mRNA sequestration acts as the equivalent to permanent caching of OS kernel in the computer, answering the 3rd question mark (?3) in figure 1B.

Differential regulation of cell surface receptors and intracellular signaling pathways

The differential regulation of specialized and core cellular function is exemplified by an interesting observation about the regulation of cellular signaling. Given the observation of over-sequestration of many signaling pathways in our analysis of KEGG and GO MF functional groups, we were startled to see that protein kinase mRNAs as a whole group were not identified to be over-sequestrated, even though the kinases form the backbone of cellular signaling networks. A closer examination revealed that a subset of the kinases – the trans-membrane receptor tyrosine kinases – were significantly under-sequestrated, ranked as the 18th least sequestered GO MF terms (Figure 8 and Table S1). When the receptor kinases are removed, the remaining intracellular protein kinases became significantly over sequestrated. Thus, cell surface receptor kinase and intracellular protein kinase mRNAs are differentially regulated. Arguably, this exemplifies differential regulation of special purpose functions and core system functions. The cell surface receptors transmit specific signals carried by their ligands, serve specific functional purposes and are thus controlled by the stable-when-translating mechanism. Similarly, the G-protein coupled receptors are also under-sequestered, ranked as the 111th least sequestered GO MF gene set (out of the 1708 sets) (Table S1); not surprisingly, the 2nd and 4th ranked least sequestered KEGG terms are related to cell surface functions (Table 2), and "cell_surface" is the 3rd least sequestered GO CC term (Table S1). On the other hand, the intracellular signaling pathways are often parts of numerous distinct cellular processes, and more likely part of the core system functions; for instance, the MAP kinase, the mTOR and the PI-3 kinase signaling modules are integral to essentially all major cellular processes. Not surprisingly, they are over-sequestrated (Tables 2 and 3). Similarly, the mRNAs for protein phosphatases, another major category of proteins that control intracellular signaling – are also significantly sequestrated (Table 3).

Discussion

The genome-to-proteome information flow process and its centrality to cellular operation are encapsulated in the molecule biology central dogma. However, how the cells mitigate the latency imposed by the tardiness of this process remains unknown; and our understanding of the long-observed complexity of the process is fragmented. We studied this process as an instance of the Shannon information theory, and hence the gene expression machinery and their regulators as the information flow channel. Briefly, the abstraction prompted a multi-omic comparative analysis with the dynamic HD-memory-CPU information flow–the instance of the same theory in support

of system operation in computers. The study enabled a theoretical integration of the complexity of transcriptome regulation, via an analogy to proven principles of computer memory management, with the latency mitigation.

As discussed, transcriptome regulation in human cells exhibits inherent, prohibitively enormous, complexity. Conceivably, the complexity evolved hand-in-hand with the complexity of cellular architecture and processes from bacteria to eukaryotes and to metazoans such as humans. The subcellular components and their associated functions become increasingly compartmentalized or modularized. For instance, the formation of the nucleus and resultant spatial separation of translation from transcription. While the gene counts sequentially increase, chromatinization and increasing chromosome condensation render transcription activation and elongation progressively slower and costlier. Not surprisingly, the evolutionary emergence of a myriad of post-transcriptional regulatory mechanisms accompanies that of metazoans. The examples include the microRNA regulatory system and the sets of RNA binding proteins, and more are sure to be un-covered. Unfortunately, the high levels of complexity also render direct theoretical inquiry for the underpinning principles difficult, if not impossible, in many situations.

Our cell-to-computer analogy based on the respective implementation of the information theory in support of system operations seems helpful. The analogy led to a multi-omic measurement of vital gene expression steps, intending to test the explanatory power of computer latency-mitigation principles. The analysis revealed additional architectural similarity; the sequential genome-to-proteome enhancement of selectivity is parallel to the sequentially HD-to-CPU lower capacity, and thus higher selectivity, in the computer; and mRNA degradation is equivalent to computer memory information purging. We also revealed regulatory similarities. The genomic organization of the genes was re-interpreted in the spirit of the spatial locality principle, and we generated experimental evidence that mRNA stabilization-by-translation is analogous to the temporal locality principle that governs selective information purging from the memory. That is, the spatiotemporal locality principle and its critical role in computer latency mitigation facilitate an interpretation of genomic organization and co-transcription regulation, as well as mRNA stability control by translation activity, in the context of cellular operational efficiency.

Moreover, our dataset provided an opportunity to tackle another transcriptome regulation complexity – the defiance of the locality principle. As discussed above, many other mechanisms regulate mRNAs, often resulting in mRNA sequestration/caching, i.e., stabilization of translationally inactive mRNAs. mRNA sequestration/caching analysis mandates simultaneous monitoring of mRNA stability and translation activity. The stability data cannot tell, by itself, whether the stabilization is due to translation activity or the caching mechanisms. Similarly, the translation activity data alone is powerless in assessing mRNA stability. Our transcriptome-wide depiction of the relationship between mRNA stability and translation activity met the prerequisite. This relationship, along with the public functional annotation of the genes, provided a framework for assessing mRNA sequestration and determinating the mode of regulation of individual cellular functions, i.e., the extent to which their mRNAs are controlled by either caching or translation activity. Indeed, our analysis of cytoplasmic non-polysome-associated mRNAs confirmed that highly sequestered mRNA defies the stable-when-translating regulatory mechanism. We observed that mRNAs for the gene expression machinery are highly cached, and so are mRNAs for their regulators and intracellular cellular signaling proteins. On the other hand,

mRNAs for metabolic pathways, such as the TCA cycle, are much under-cached, and so are the mRNAs for structural proteins of some subcellular organelles, such as the lysosome and proteasome.

This mRNA sequestration pattern reinforced the cell-to-computer analogy (27,50) and the application of the Shannon information theory. The gene expression machinery, their regulators, and the signaling molecules constitute the information flow channels in the cells; the basic machinery carries out the flow, and their regulators and cellular signaling determine specificity, *i.e.*, dynamic allocation of the channel capacity among biochemical pathways. In computers, the HD-to-CPU information flow channel is arguably the OS kernel; the well-defined kernel functions (file system, memory management, and process control/scheduling) collectively accomplish the task. It is striking that, in both systems, channel components are themselves part of the flow process and given preferential status – mRNA sequestration and memory caching, respectively. While the similarity between computers and cells has been widely noticed and computers always have an OS, whether the cells have equivalent to the operating system was questioned (50). Our results suggest that the gene expression machinery, regulators of the machinery, and intracellular signaling are arguably the answer. Thus, this study benefited from, and also further extended, the cell-to-computer analogy and the application of the information theory.

Our usage of computer optimization principles as a stepping stone to tackle, indirectly, the complexity of the cellular operation is consistent with the tradition of using simpler model systems in biological research. For instance, the yeast *Saccharomyces cerevisiae* has long served as a model organism for understanding higher eukaryotes. Similarly, neural science and development biology have long used the worm *Caenorhabditis elegans* conveniently as the model. It is also consistent with an underlying notion in systems biology, that is, to explore similarities between biological and engineered complex systems (63).

To the best of our knowledge, this is the first report of the quantification of mRNA caching activity. And our results point to future investigation to further advance our understanding of this pervasive phenomenon. This analysis represents a snapshot of actively growing log-phase cells, which are not homogeneous cell populations. Instead, they are a mixture, for instance, of cells in different cell cycle stages. On the other hand, it is reasonable to assume that the mRNA caching activity is dynamic, *i.e.*, cell cycle stage, cell type, and growth condition specific. Thus, there exists a need to apply our approach to multiple cell types and time-series experiments, during which the cells transition through physiological processes or adjust to stresses. This way, we will know when and why mRNA caching occurs. We will also know when the cell releases cached mRNAs into the polysomes for quick protein production. We will be able to delineate the functional and operational advantages endowed to the cells by mRNA caching.

We would also point out that this analysis quantified the overall mRNA caching activity, not focusing on individual caching mechanisms or subcellular foci. On the other hand, we have learned a lot about specific caching mechanisms. For instance, the P-body has long been known as a repertoire for non-translating mRNAs. It was shown to act as sub-cellular foci for miRNA-mediated mRNA caching/sequestration (59). Some stored mRNAs get degraded, but some recycle back for active protein production. Recently, the GW-body was identified as another

subcellular foci for translationally repressed mRNAs. And a large number of RBPs have been discovered in the human genome. Some of the RBPs are undoubtedly involved in mRNA caching (1,2). However, we believe there is a lot to be learned yet. Our results should be a good framework for the functional study of these subcellular foci and RBPs.

In summary, we described a mechanistic and functional study of the complexity of transcriptome regulation via a computer-to-cell analogy. Human gene expression is metabolically expensive, *e.g.*, consuming large amounts of building blocks and ATP as the energy source. The process is also time-consuming, imposing potential operational delays. Thus, there is a need to fine-tune the process for an optimal tradeoff among multiple cellular "economics" factors. Our analyses suggest the spatiotemporal locality and the caching principles are integral to such optimization schemes. The functional and regulatory equivalent between the storage-to-execution information flow channels of the cell and the computer is striking and begs for further investigation. They might be a pure coincidence of the biological evolution of the cells and the competition-driven evolution of computers in the market. Alternatively, they might be the manifestation of fundamental organizational and operational principles of complex systems. We will not be surprised, though, if the latter turns out to be the case.

1. Gebauer, F., Schwarzl, T., Valcarcel, J. and Hentze, M.W. (2021) RNA-binding proteins in human genetic disease. *Nat Rev Genet*, **22**, 185-198.
2. Uhl, M., Tran, V.D., Heyl, F. and Backofen, R. (2021) RNAProt: an efficient and feature-rich RNA binding protein binding site predictor. *Gigascience*, **10**.
3. Bartel, D.P. (2018) Metazoan MicroRNAs. *Cell*, **173**, 20-51.
4. Jiang, W., Guo, Z., Lages, N., Zheng, W.J., Feliers, D., Zhang, F. and Wang, D. (2018) A Multi-Parameter Analysis of Cellular Coordination of Major Transcriptome Regulation Mechanisms. *Sci Rep*, **8**, 5742.
5. Hayles, B., Yellaboina, S. and Wang, D. (2010) Comparing transcription rate and mRNA abundance as parameters for biochemical pathway and network analysis. *PLoS One*, **5**, e9908.
6. Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R. *et al.* (2015) Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science*, **347**, 1259038.
7. Li, J.J., Bickel, P.J. and Biggin, M.D. (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, **2**, e270.
8. Liu, Y. and Aebersold, R. (2016) The interdependence of transcript and protein abundance: new data--new complexities. *Mol Syst Biol*, **12**, 856.
9. McManus, J., Cheng, Z. and Vogel, C. (2015) Next-generation analysis of gene expression regulation--comparing the roles of synthesis and degradation. *Mol Biosyst*, **11**, 2680-2689.
10. Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N. *et al.* (2011) Metabolic labeling of RNA

uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat Biotechnol*, **29**, 436-442.

11. Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D.J., Pauli, A., Hacohen, N., Schier, A.F., Blackshear, P.J., Friedman, N. *et al.* (2014) High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell*, **159**, 1698-1710.
12. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337-342.
13. Vogel, C. and Marcotte, E.M. (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet*, **13**, 227-232.
14. Wang, D. (2008) Discrepancy between mRNA and protein abundance: insight from information retrieval process in computers. *Comput Biol Chem*, **32**, 462-468.
15. Anderson, L. and Seilhamer, J. (1997) A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, **18**, 533-537.
16. Gygi, S.P., Rochon, Y., Franza, B.R. and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*, **19**, 1720-1730.
17. Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R. and Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929-934.
18. Flory, M.R., Lee, H., Bonneau, R., Mallick, P., Serikawa, K., Morris, D.R. and Aebersold, R. (2006) Quantitative proteomic analysis of the budding yeast cell cycle using acid-cleavable isotope-coded affinity tag reagents. *Proteomics*, **6**, 6146-6157.
19. Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K. and Weissman, J.S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737-741.
20. Griffin, T.J., Gygi, S.P., Ideker, T., Rist, B., Eng, J., Hood, L. and Aebersold, R. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Mol Cell Proteomics*, **1**, 323-333.
21. Le Roch, K.G., Johnson, J.R., Florens, L., Zhou, Y., Santrosyan, A., Grainger, M., Yan, S.F., Williamson, K.C., Holder, A.A., Carucci, D.J. *et al.* (2004) Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Res*, **14**, 2308-2318.
22. Tian, Q., Stepaniants, S.B., Mao, M., Weng, L., Feetham, M.C., Doyle, M.J., Yi, E.C., Dai, H., Thorsson, V., Eng, J. *et al.* (2004) Integrated genomic and proteomic analyses of gene expression in Mammalian cells. *Mol Cell Proteomics*, **3**, 960-969.
23. Washburn, M.P., Koller, A., Oshiro, G., Ulaszek, R.R., Plouffe, D., Deciu, C., Winzeler, E. and Yates, J.R. (2003) Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, **100**, 3107-3112.
24. Cheng, Z., Teo, G., Krueger, S., Rock, T.M., Koh, H.W., Choi, H. and Vogel, C. (2016) Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Mol Syst Biol*, **12**, 855.

25. Griffin, T.J., Gygi, S.P., Ideker, T., Rist, B., Eng, J., Hood, L. and Aebersold, R. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Mol Cell Proteomics*, **1**, 323-333.
26. Pai, A.A., Cain, C.E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J.B., Degner, J.F., Gaffney, D.J., Pickrell, J.K., Stephens, M. *et al.* (2012) The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genet*, **8**, e1003000.
27. Wang, D. and Gribskov, M. (2005) Examining the architecture of cellular computing through a comparative study with a computer. *J R Soc Interface*, **2**, 187-195.
28. Wang, D.G. (2005) "Molecular gene": Interpretation in the right context. *Biol Philos*, **20**, 453-464.
29. Tanenbaum, A.S.A.T. (2017) *Structured computer organization*.
30. Denning, P.J. (2005) The locality principle. *Commun. ACM*, **48**, 19–24.
31. Jiang, W., Tillekeratne, M.P., Brattain, M.G. and Banerji, S.S. (1997) Decreased stability of transforming growth factor beta type II receptor mRNA in RER+ human colon carcinoma cells. *Biochemistry*, **36**, 14786-14793.
32. Wang, D., Li, W., Jiang, W., Humphrey, L.E., Howell, G.M. and Brattain, M.G. (1998) Autocrine TGFalpha expression in the regulation of initiation of human colon carcinoma growth. *J Cell Physiol*, **177**, 387-395.
33. Wang, D., Patil, S., Li, W., Humphrey, L.E., Brattain, M.G. and Howell, G.M. (2002) Activation of the TGFalpha autocrine loop is downstream of IGF-I receptor activation during mitogenesis in growth factor dependent human colon carcinoma cells. *Oncogene*, **21**, 2785-2796.
34. Core, L.J., Waterfall, J.J. and Lis, J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, **322**, 1845-1848.
35. Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A. and Ren, B. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290-294.
36. Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M.U., Ohgi, K.A. *et al.* (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, **474**, 390-394.
37. Wang, D. and Karamyshev, A.L. (2020) Next Generation Sequencing (NGS) Application in Multiparameter Gene Expression Analysis. *Methods Mol Biol*, **2102**, 17-34.
38. Day, R.T., Cavaglieri Rde, C., Tabatabaimir, H., Mantravadi, V., Lee, M.J., Barnes, J.L., Kasinath, B.S. and Feliers, D. (2010) Acute hyperglycemia rapidly stimulates VEGF mRNA translation in the kidney. Role of angiotensin type 2 receptor (AT2). *Cell Signal*, **22**, 1849-1857.
39. Feliers, D., Duraisamy, S., Barnes, J.L., Ghosh-Choudhury, G. and Kasinath, B.S. (2005) Translational regulation of vascular endothelial growth factor expression in renal epithelial cells by angiotensin II. *Am J Physiol Renal Physiol*, **288**, F521-529.
40. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15-21.

41. Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166-169.
42. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**, 15545-15550.
43. Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E. *et al.* (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, **34**, 267-273.
44. Qin, T., Matmati, N., Tsoi, L.C., Mohanty, B.K., Gao, N., Tang, J., Lawson, A.B., Hannun, Y.A. and Zheng, W.J. (2014) Finding pathway-modulating genes from a novel Ontology Fingerprint-derived gene network. *Nucleic Acids Res*, **42**, e138.
45. Qin, T., Tsoi, L.C., Sims, K.J., Lu, X. and Zheng, W.J. (2012) Signaling network prediction by the Ontology Fingerprint enhanced Bayesian network. *BMC Syst Biol*, **6 Suppl 3**, S3.
46. Tsoi, L.C., Boehnke, M., Klein, R.L. and Zheng, W.J. (2009) Evaluation of genome-wide association study results through development of ontology fingerprints. *Bioinformatics*, **25**, 1314-1320.
47. Bray, D. (1990) Intracellular signalling as a parallel distributed process. *J Theor Biol*, **143**, 215-231.
48. Bray, D. (1995) Protein molecules as computational elements in living cells. *Nature*, **376**, 307-312.
49. Brent, R. and Bruck, J. (2006) 2020 computing: can computers help to explain biology? *Nature*, **440**, 416-417.
50. Condon, A., Kirchner, H., Lariviere, D., Marshall, W., Noireaux, V., Tlusty, T. and Fourmentin, E. (2018) Will biologists become computer scientists? A truly interdisciplinary effort by computer scientists and biologists to understand how cells process information may yield new insights for both fields. *EMBO Rep*, **19**.
51. Shannon, C.E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**, 379-423.
52. Schoenberg, D.R. and Maquat, L.E. (2012) Regulation of cytoplasmic mRNA decay. *Nature Reviews Genetics*, **13**, 246-259.
53. Roy, B. and Jacobson, A. (2013) The intimate relationships of mRNA decay and translation. *Trends Genet*, **29**, 691-699.
54. Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R. *et al.* (2015) Codon optimality is a major determinant of mRNA stability. *Cell*, **160**, 1111-1124.
55. Shyu, A.-B., Wilkinson, M.F. and van Hoof, A. (2008) Messenger RNA regulation: to translate or to degrade. *Embo J*, **27**, 471-481.
56. Jacobson, A. (1996) Poly (A) metabolism and translation: the closed-loop model. *Translational control.*, 451-480.
57. Pandit, S., Wang, D. and Fu, X.D. (2008) Functional integration of transcriptional and RNA processing machineries. *Curr Opin Cell Biol*, **20**, 260-265.

58. Bell, J.L., Wächter, K., Mühleck, B., Pazaitis, N., Köhn, M., Lederer, M. and Hüttelmaier, S. (2013) Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): post-transcriptional drivers of cancer progression? *Cell Mol Life Sci*, **70**, 2657-2675.
59. Liu, J., Valencia-Sanchez, M.A., Hannon, G.J. and Parker, R. (2005) MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nat Cell Biol*, **7**, 719-723.
60. Patel, P.H., Barbee, S.A. and Blankenship, J.T. (2016) GW-Bodies and P-Bodies Constitute Two Separate Pools of Sequestered Non-Translating RNAs. *PLoS One*, **11**, e0150291-e0150291.
61. Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O. and Herschlag, D. (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A*, **100**, 3889-3894.
62. Heyer, E.E. and Moore, M.J. (2016) Redefining the Translational Status of 80S Monosomes. *Cell*, **164**, 757-769.
63. Csete, M.E. and Doyle, J.C. (2002) Reverse engineering of biological complexity. *Science*, **295**, 1664-1669.

Figure Legends

Figure 1. Analogizing the gene expression process to the computer information retrieval process via the Shannon information theory (A & B), and our strategy for experimental exploration of the analogy (C). Panel **A** outlines the three components of the information theory, with green text and lines denoting the information channel and the enclosed intermediate step. Panel **B** outlines the implementation of the theory in the cell and in the computer. The sequential narrowing of the textboxes from hard drive (HD) to central processing unit (CPU) cache denotes the sequential capacity decrease in the computer memory hierarchy. The parenthesized italic texts specify the optimization principles at the corresponding steps in the computer, and the parenthesized question marks (?) denote our investigation into whether equivalent principles act in the cells. Panel **C** outlines our experimental strategy to answer these questions. Log-phase HCT116 cells were split into multiple aliquots. One aliquot was used for RNA-seq analysis to measure steady-state mRNA abundance (RA) (purple textboxes and arrows). One aliquot was used for GRO-seq analysis to measure transcription rate (TR) (brown textboxes, lines and arrows). Another aliquot was used to isolate and quantify both polysome-associated mRNA/translation activity (TA) (light blue textboxes and arrows) and non-polysome-associated cytoplasmic mRNA (NP) (blue textboxes and arrows). The arrowed lines between panels B and C map experimentally measured parameters to the targeted gene expression steps.

Figure 2. Comparison of $\log_2$(TR), $\log_2$(RA), $\log_2$(NP) and $\log_2$(TA) distributions. A box plot of the RPKM values of the four parameters is shown, illustrating the sequentially lower median values, and higher dispersion levels, from TR to TA. The trends are consistent with the sequential narrowing of the textboxes in figure 1B.

Figure 3: Linear regression analyses to illustrate the discrepancy among TR, RA and TA. A: Scatter plot of $\log_2$(RA) (y-axis) versus $\log_2$(TR) (x-axis) (red) and $\log_2$(RA biological replicate 2) (Y-axis) versus $\log_2$(RA biological replicate 1) (x-axis) (black). Both linear regression lines are also shown. **B:** The $\log_2$(RA) versus $\log_2$(TR) scatter plot for three functional groups: ribosome proteins, lysosome and TCA cycle. The linear regression line ($\log_2$(RA) versus $\log_2$(TR)) from A is also shown. **C:** Scatter plot of $\log_2$(TA) (y-axis) versus $\log_2$(RA) (x-axis) (red) and of the two RA biological replicates (black). Both linear regression lines are also shown. **D:** The $\log_2$(TA) versus $\log_2$(RA) scatter plot for three functional groups: ribosome proteins, lysosome and TCA cycle. The linear regression line ($\log_2$(TA) versus $\log_2$(RA)) from C is also shown. The same two RA biological replicates are used in A and C. The patterns in A and C are consistent with the corresponding narrowing of the textboxes in figure 1B. Outliers were removed from the three functional groups.

Figure 4. Scatter plots illustrating the correlation between mRNA stability index and translation index. In **A**, the indices were calculated with predicted RA and TA values based on the linear regressions in figures 4A and 4C, respectively, *i.e.*, $\log_2(RA^{Pred}/TR)$ and $\log_2(TA^{Pred}/RA)$. The strong positive correlation supports mRNA stabilization-by-translation as the answer to the second question mark ($?^2$) in figure 1B. In **B**, experimentally determined RA and TA values were used. The linear regression line is also shown. **C**: a scatter plot for the three KEGG functional groups. The linear regression line from B is also shown. In **D**, the stability index was adjusted with the mRNA sequestration index (see figure 6 and text for details).

Figure 5. Function-specific pattern of mRNA sequestration and determination of the sequestration index. A: Scatter plot of $\log_2$(NP) (y-axis) versus $\log_2$(RA) (x-axis) (red) and $\log_2$(RA biological replicate 2) (Y-axis) versus $\log_2$(RA biological replicate 1) (x-axis) (black). The two linear regression lines are also shown. The same two RA biological replicates as in figures 4A and 4C are used. **B:** The $\log_2$(NP) versus $\log_2$(RA) scatter plot for three functional groups: ribosome proteins, lysosome and TCA cycle. The linear regression line ($\log_2$(NP) versus $\log_2$(RA)) from A is also shown. **C**: Scatter plot of the raw mRNA sequestration level versus $\log_2$(ORF length). The red dots denote the prediction of a loess regression, whose residue is used as the mRNA sequestration index. **D:** Scatter plot (raw mRNA sequestration level versus $\log_2$(ORF length)) for the three KEGG functional groups. The loess regression prediction from C is also shown.

Figure 6. Function-specific mRNA sequestration pattern illustrated by genome-wide analyses. For each gene pair, their GO similarity and the differences between their sequestration index (termed $sequ^{diff}$ in text) were calculated. **A:** Comparative box plot of the sequestration index differences of gene pair sets with increasing levels of GO similarity. For **B and C**, the gene pairs were binned according to the GO similarity score. In **B**, a scatter plot of the standard deviation (SD) of the sequestration index differences versus the median GO similarity scores in the bins is shown; the black and blue horizontal lines denote the SDs for gene pairs with no GO similarity and all gene pairs, respectively. In **C**, a scatter plot of the median absolute values of the sequestration index differences versus the median GO similarity scores in the bins is shown. **D:** Comparative box plot of the sequestration index differences of gene pairs with varying protein-protein interaction scores.

Figure 7. The negative relationship between mRNA sequestration index and the mRNA translation-stability correlation as well as the function-specific pattern among the KEGG functional gene sets. A: Sliding window analysis of the relationship between mRNA sequestration index and the mRNA translation-stability correlation. **B-D:** Association of high mRNA sequestration level with low mRNA translation-stability correlation among the KEGG functional groups. For each functional group, the mRNA sequestration level (in the form of t-score of the index values) and the correlation coefficient between the translation and stability indices were calculated. **B:** Scatter plot of the correlation coefficients versus mRNA sequestration levels is shown. The red data points denote groups with statistically insignificant sequestration levels, *i.e.*, p-value larger than 0.05. The ribosome protein is labeled to illustrate the extreme sequestration level. **C:** comparative scatter plot of the correlation coefficients of under- and over-sequestrated KEGG functional groups. **D:** comparative scatter plot of the sequestration levels of the set of KEGG groups and the randomized set, illustrating the functional-specific mRNA sequestration pattern. The randomized set has the same number of groups and group sizes. Together with Tables 2 and 3, these results support sequestration of mRNAs for basal gene expression machinery, their regulators and intracellular cellular signaling. The mRNA sequestration is equivalent to permanent memory caching of operating system (OS) kernel in the computer, answering the 3rd question mark ($?^3$) in figure 1B. See text for details.

Figure 8. Comparative box plot of the mRNA sequestration index values of all protein kinases (All), intracellular kinases (Intracellular) and the cell surface receptor protein kinases (Receptor).

Table 1: Comparison of statistical features of the profiles of gene expression parameters.

	Standard Deviation	Mean	Value Range	
			From	To
$\log_2$(TR)	1.53	3.85	-2.68	10.41
$\log_2$(RA)	2.38	3.18	-6.02	12.33
$\log_2$(NP)	2.56	3.03	-6.41	13
$\log_2$(TA)	2.77	2.74	-7.18	13.04

Table 2: Top 20 ranked KEGG functions with least or most mRNA sequestration. Gene expression related terms are in red text, cellular signaling related terms in blue and metabolism terms in green text.

Rank	Least Sequestered	Most Sequestered
1	lysosome	ribosome
2	ecm_receptor_interaction	wnt_signaling_pathway
3	other_glycan_degradation	b_cell_receptor_signaling_pathway
4	cell_adhesion_molecules_cams	renal_cell_carcinoma
5	glycolysis_gluconeogenesis	long_term_potentiation
6	citrate_cycle_tca_cycle	mapk_signaling_pathway
7	dna_replication	chronic_myeloid_leukemia
8	asthma	colorectal_cancer
9	antigen_processing_and_presentation	neurotrophin_signaling_pathway
10	propanoate_metabolism	t_cell_receptor_signaling_pathway
11	valine_leucine_and_isoleucine_degradation	basal_transcription_factors
12	autoimmune_thyroid_disease	jak_stat_signaling_pathway
13	sphingolipid_metabolism	chemokine_signaling_pathway
14	galactose_metabolism	erbb_signaling_pathway
15	graft_versus_host_disease	acute_myeloid_leukemia
16	allograft_rejection	long_term_depression
17	steroid_biosynthesis	ubiquitin_mediated_proteolysis
18	o_glycan_biosynthesis	mtor_signaling_pathway
19	pyruvate_metabolism	pancreatic_cancer
20	proteasome	phosphatidylinositol_signaling_system

Table 3: Top 30 ranked GO CC, MF and BP Terms with Highest mRNA Sequestration Levels. Transcription factor/regulator related terms are in red text; protein and mRNA degradation related terms are in brown text; epigenetic terms are in purple text; other gene expression related terms are in blue text. Cellular signaling related terms are in green text.

Rank	CC	MF	BP
1	cytosolic_ribosome	DNA_binding_transcription_factor_activity	bone_minerlization_involved_in_bone_maturation
2	messenger_ribonucleoprotein_complex	transcription_regulator_activity	enucleate_erythrocyte_differentiation
3	chromatin	cis_regulatory_region_sequence_specific_dna_binding	immune_response_inhibiting_signal_transduction
4	ribosomal_subunit	sequence_specific_dna_binding	t_helper_1_cell_cytokine_production
5	cytosolic_large_ribosomal_subunit	interleukin_1_receptor_activity	negative_regulation_by_host_of_viral_genome_replication
6	phosphatidylinositol_3_kinase_complex_class_I	interleukin_1_binding	regulation_of_interleukin_1_mediated_signaling_pathway
7	transcription_regulator_complex	DNA_binding_transcription_repressor_activity	receptor_localization_to_non_motile_cilium
8	phosphatase_complex	DNA_binding_transcription_activator_activity	negative_regulation_of_transcription_by_RNA_polymerase_II
9	RNA_polymerase_II_transcription_regulator_complex	ubiquitin_ligase_inhibitor_activity	negative_regulation_of_nucleobase_containing_compound_metabolic_process
10	cytosolic_small_ribosomal_subunit	translation_termination_factor_activity	regulation_of_branching_involved_in_salivary_gland_morphogenesis
11	SUMO_ligase_complex	ubiquitin_protein_transferase_inhibitor_activity	tricuspid_valve_development
12	PML_body	structural_constituent_of_ribosome	tricuspid_valve_morphogenesis
13	PAM_complex_TIM23_associated_import_motor	intracellular_camp_activated_cation_channel_activity	negative_regulation_of_biosynthetic_process
14	sperm_connecting_piece	RNA_polymerase_II_c_terminal_domain_phosphoserine_binding	negative_regulation_of_cytoplasmic_translation
15	small_ribosomal_subunit	RNA_polymerase_II_transcription_coactivator_binding	positive_regulation_of_renal_sodium_excretion
16	large_ribosomal_subunit	protein_serine_threonine_phosphatase_inhibitor_activity	positive_regulation_of_transcription_by_RNA_polymerase_II
17	ribosome	phosphatidylinositol_bisphosphate_kinase_activity	oligodendrocyte_progenitor_proliferation

18	protein_acetyltransferase_complex	GDP_dissociation_inhibitor_activity	activation_of_phospholipase_d_activity
19	MRE11_complex	potassium_channel_inhibitor_activity	translational_initiation
20	transcription_factor_tftc_complex	transcription_factor_binding	negative_regulation_of_vascular_associated_smooth_muscle_cell_migration
21	BLOC_1_complex	phosphatase_regulator_activity	histone_h3_k36_trimethylation
22	nuclear_body	SUMO_specific_protease_activity	negative_regulation_of_interleukin_1_mediated_signaling_pathway
23	fibrillar_center	DNA_binding_transcription_factor_binding	thrombopoietin_mediated_signaling_pathway
24	WAS_complex	histone_acetyltransferase_activity_h3_k23_specific	positive_regulation_of_nucleobase_containing_compound_metabolic_process
25	nuclear_cyclin_dependent_protein_kinase_holoenzyme_complex	bile_acid_sodium_symporter_activity	negative_regulation_of_striated_muscle_contraction
26	eukaryotic_translation_initiation_factor_2_complex	nucleoside_triphosphatase_regulator_activity	negative_regulation_of_cardiac_muscle_contraction
27	protein_phosphatase_type_2a_complex	LBD_domain_binding	negative_regulation_of_insulin_secretion_involved_in_cellular_response_to_glucose_stimulus
28	nuclear_speck	transcription_coregulator_activity	cotranslational_protein_targeting_to_membrane
29	CCR4_NOT_complex	ornithine_decarboxylase_regulator_activity	protein_deglutamylation
30	SAGA_type_complex	protein_geranylgeranyltransferase_activity	positive_regulation_of_antigen_processing_and_presentation

Figure 1

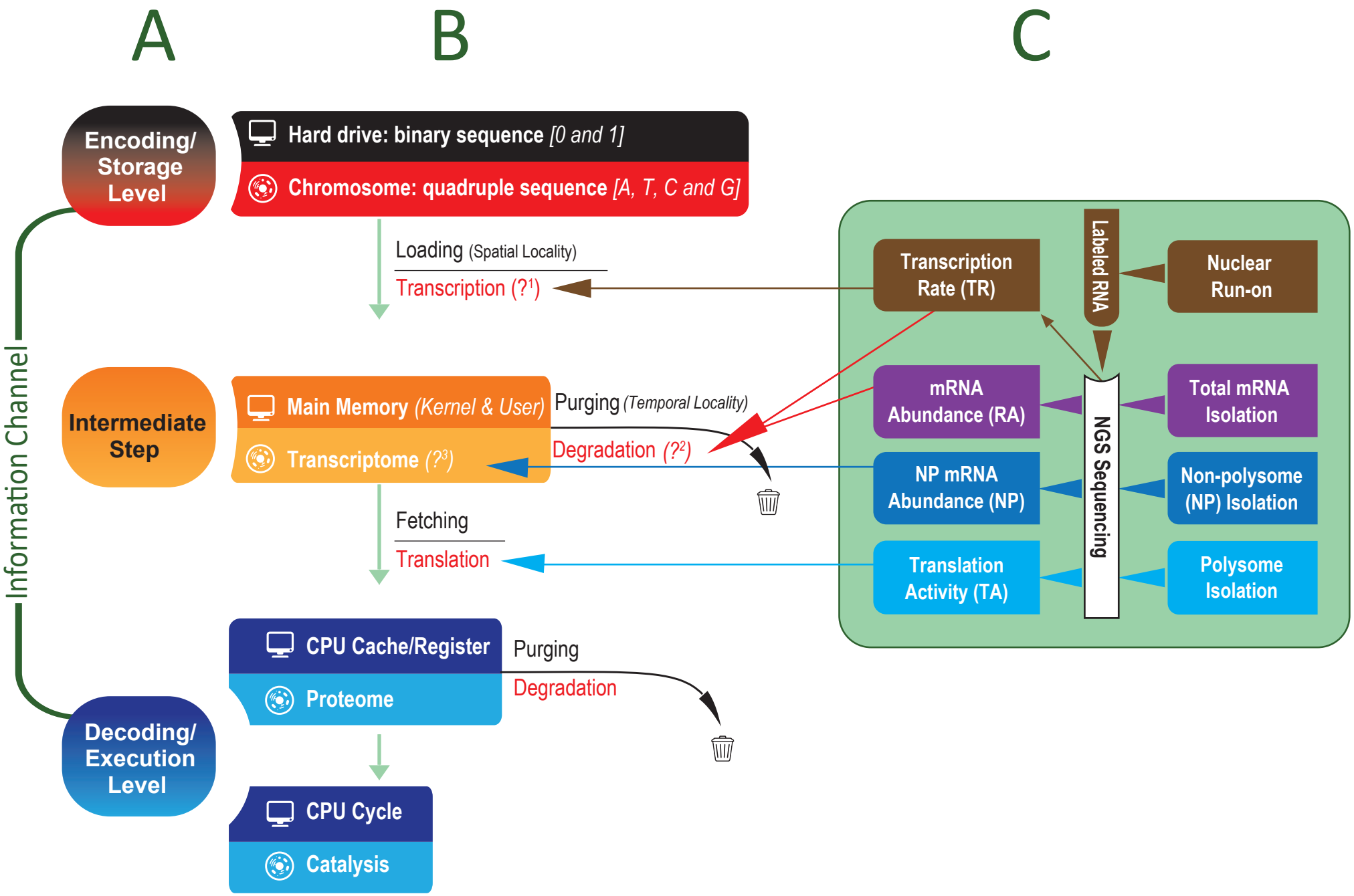

A
B
C
Encoding/ Storage Level
Intermediate Step
Decoding/ Execution Level
Information Channel
Hard drive: binary sequence [0 and 1]
Chromosome: quadruple sequence [A, T, C and G]
Loading (Spatial Locality)
Transcription (?1)
Main Memory (Kernel & User)
Transcriptome (?3)
Purging (Temporal Locality)
Degradation (?2)
Fetching
Translation
CPU Cache/Register
Proteome
Purging
Degradation
CPU Cycle
Catalysis
Transcription Rate (TR)
Labeled RNA
Nuclear Run-on
mRNA Abundance (RA)
Total mRNA Isolation
NP mRNA Abundance (NP)
Non-polysome (NP) Isolation
Translation Activity (TA)
Polysome Isolation
NGS Sequencing

Figure 2

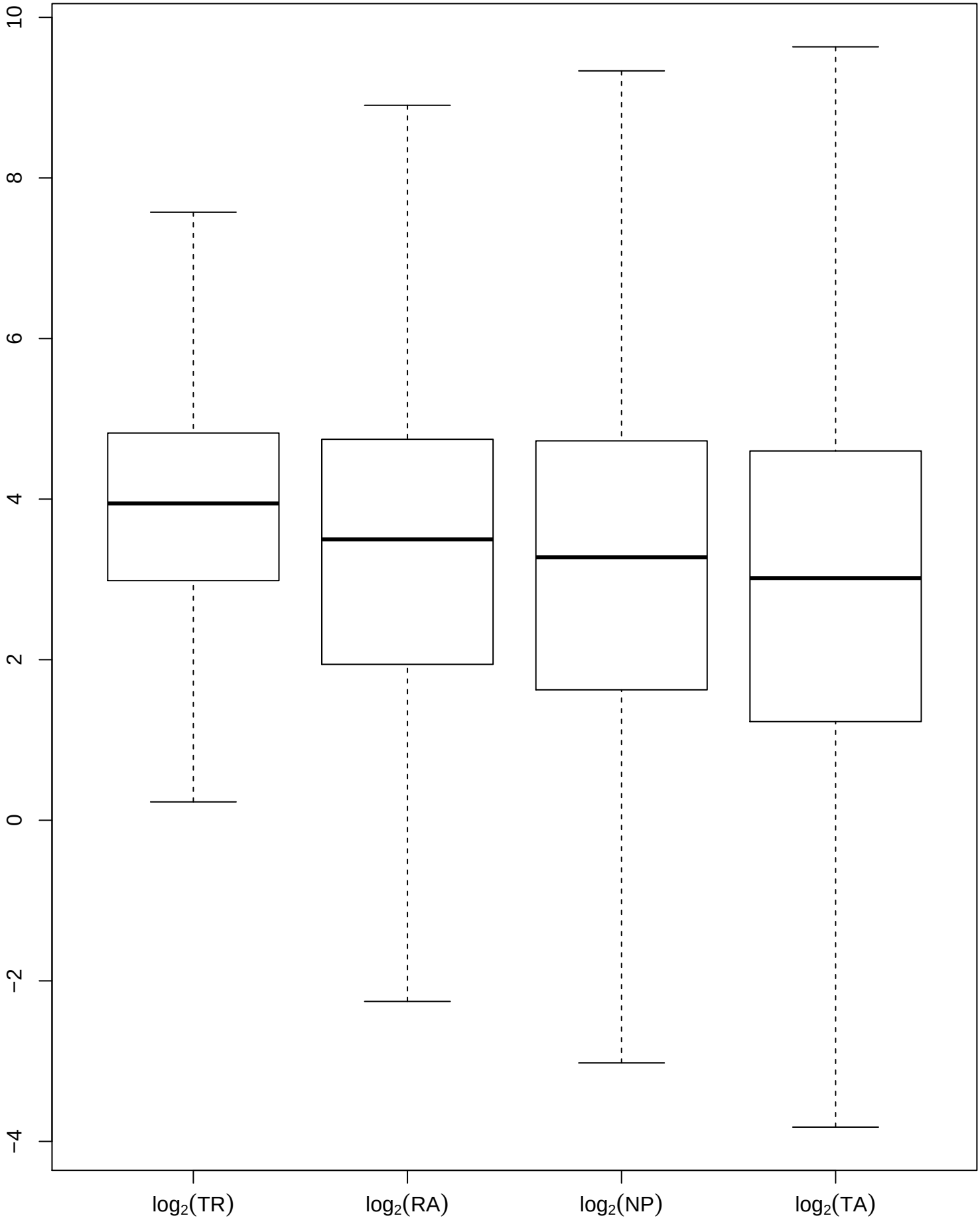


Figure 3

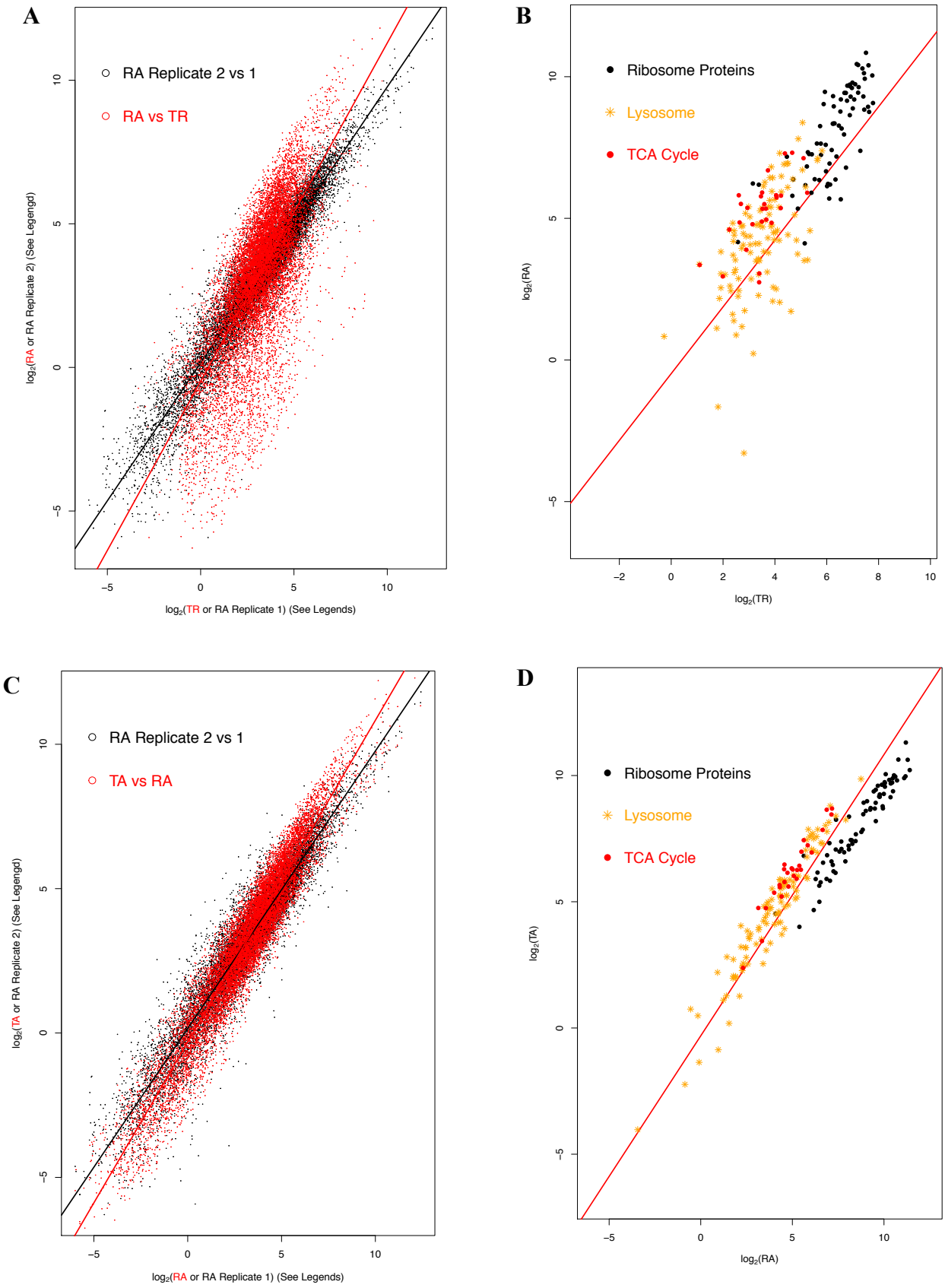

A
RA Replicate 2 vs 1
RA vs TR
log2(RA or RA Replicate 2) (See Legengd)
log2(TR or RA Replicate 1) (See Legends)
B
Ribosome Proteins
Lysosome
TCA Cycle
log2(RA)
log2(TR)
C
RA Replicate 2 vs 1
TA vs RA
log2(TA or RA Replicate 2) (See Legengd)
log2(RA or RA Replicate 1) (See Legends)
D
Ribosome Proteins
Lysosome
TCA Cycle
log2(TA)
log2(RA)

Figure 4

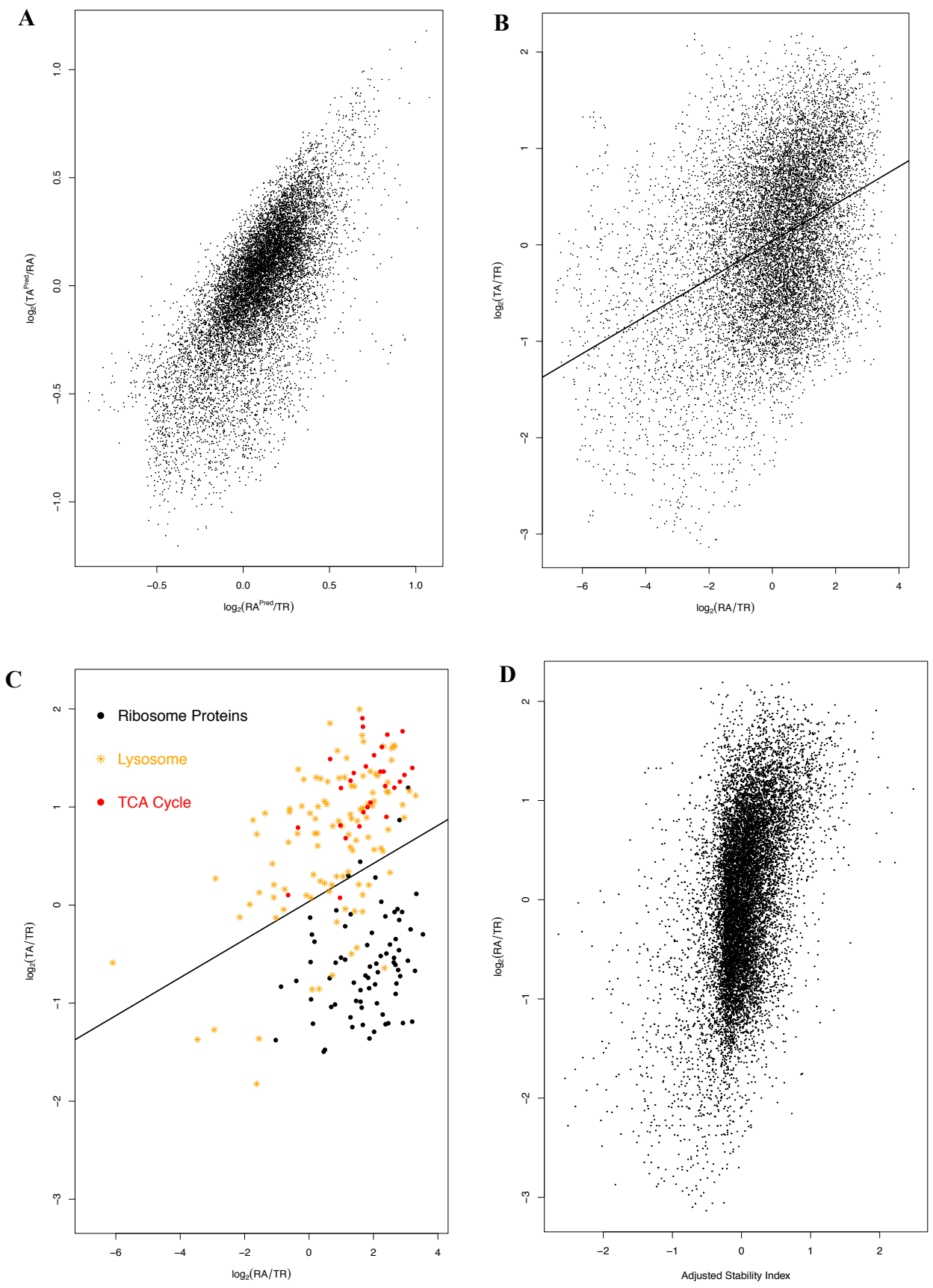

A
log2(TA^Pred/RA)
log2(RA^Pred/TR)
B
log2(TA/TR)
log2(RA/TR)
C
Ribosome Proteins
Lysosome
TCA Cycle
log2(TA/TR)
log2(RA/TR)
D
log2(RA/TR)
Adjusted Stability Index

Figure 5

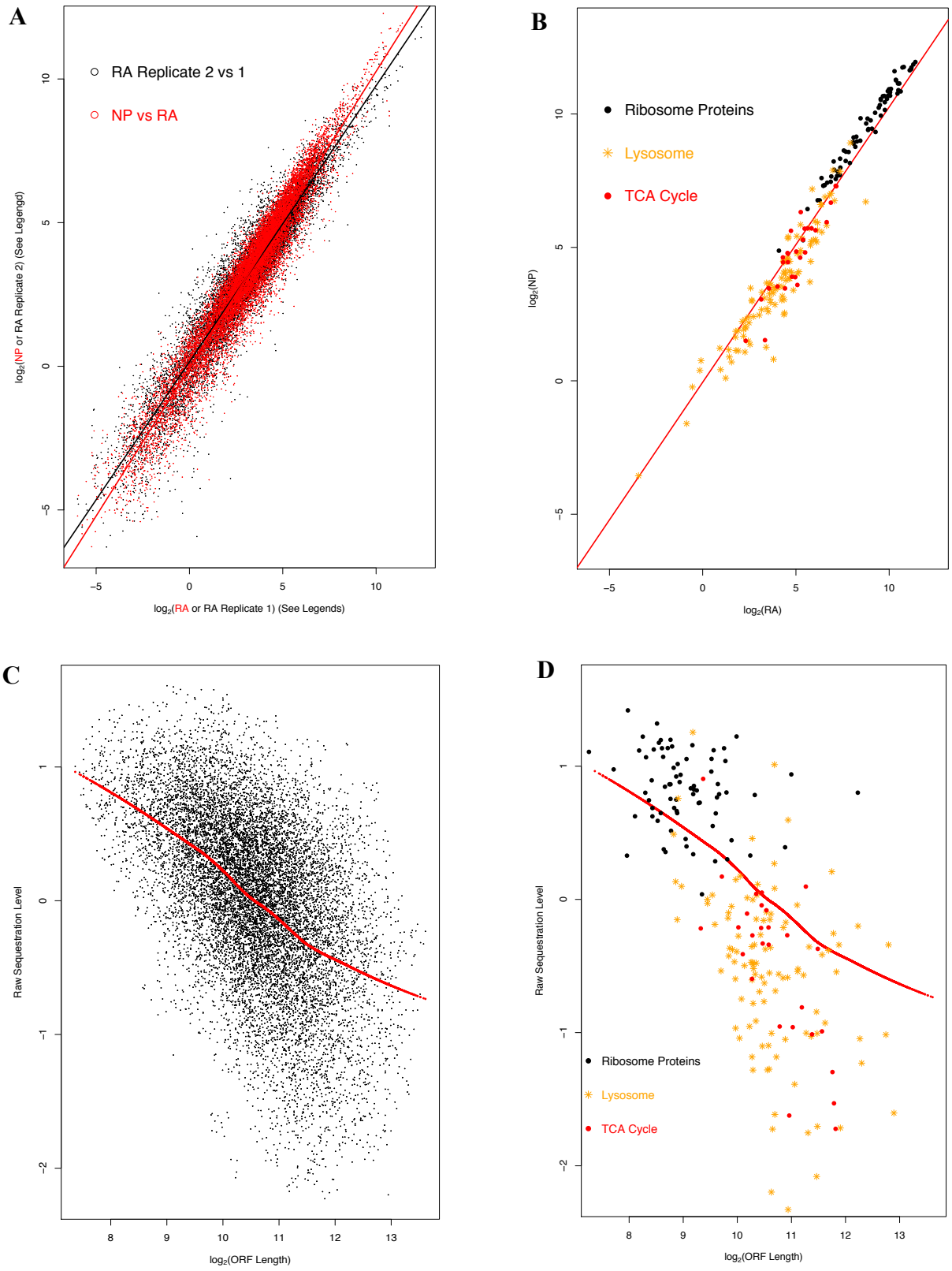

A
RA Replicate 2 vs 1
NP vs RA
log2(NP or RA Replicate 2) (See Legend)
log2(RA or RA Replicate 1) (See Legends)
B
Ribosome Proteins
Lysosome
TCA Cycle
log2(NP)
log2(RA)
C
Raw Sequestration Level
log2(ORF Length)
D
Ribosome Proteins
Lysosome
TCA Cycle
Raw Sequestration Level
log2(ORF Length)

Figure 6

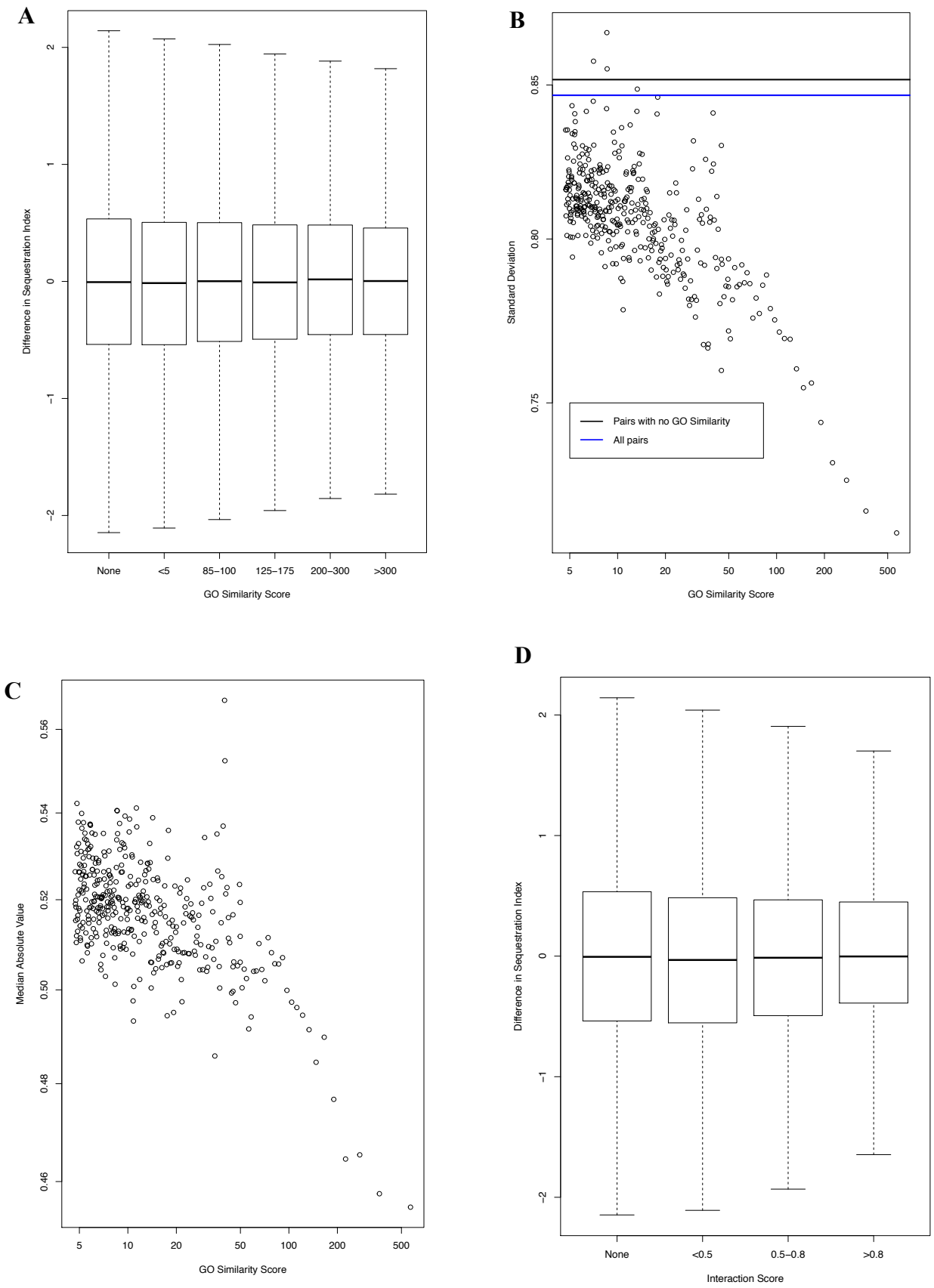
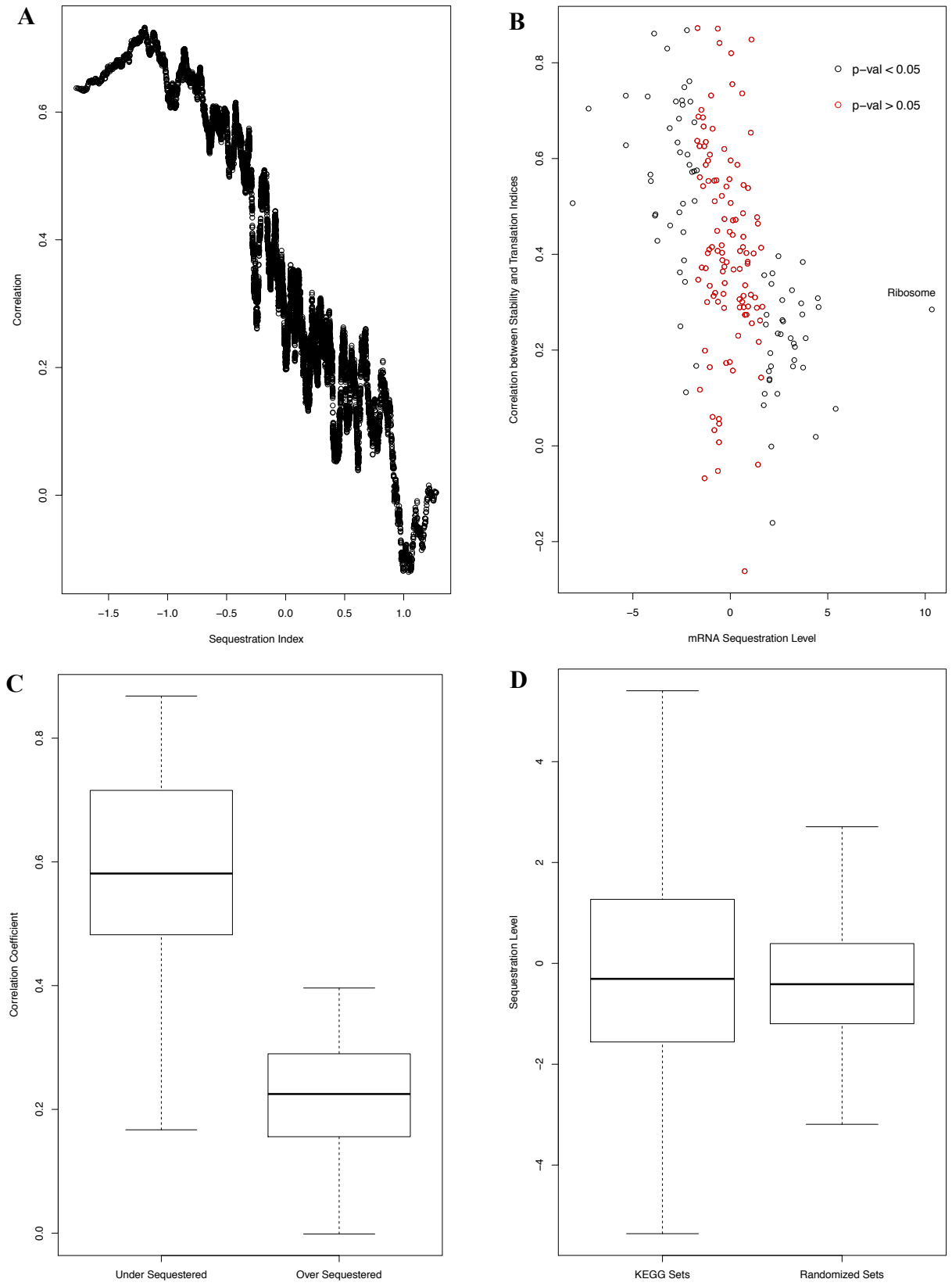

A
Difference in Sequestration Index
2
1
0
−1
−2
None
<5
85−100
125−175
200−300
>300
GO Similarity Score
B
Standard Deviation
0.85
0.80
0.75
Pairs with no GO Similarity
All pairs
5
10
20
50
100
200
500
GO Similarity Score
C
Median Absolute Value
0.56
0.54
0.52
0.50
0.48
0.46
5
10
20
50
100
200
500
GO Similarity Score
D
Difference in Sequestration Index
2
1
0
−1
−2
None
<0.5
0.5−0.8
>0.8
Interaction Score

Figure 7


A
Correlation
Sequestration Index
B
Correlation between Stability and Translation Indices
mRNA Sequestration Level
p−val < 0.05
p−val > 0.05
Ribosome
C
Correlation Coefficient
Under Sequestered
Over Sequestered
D
Sequestration Level
KEGG Sets
Randomized Sets

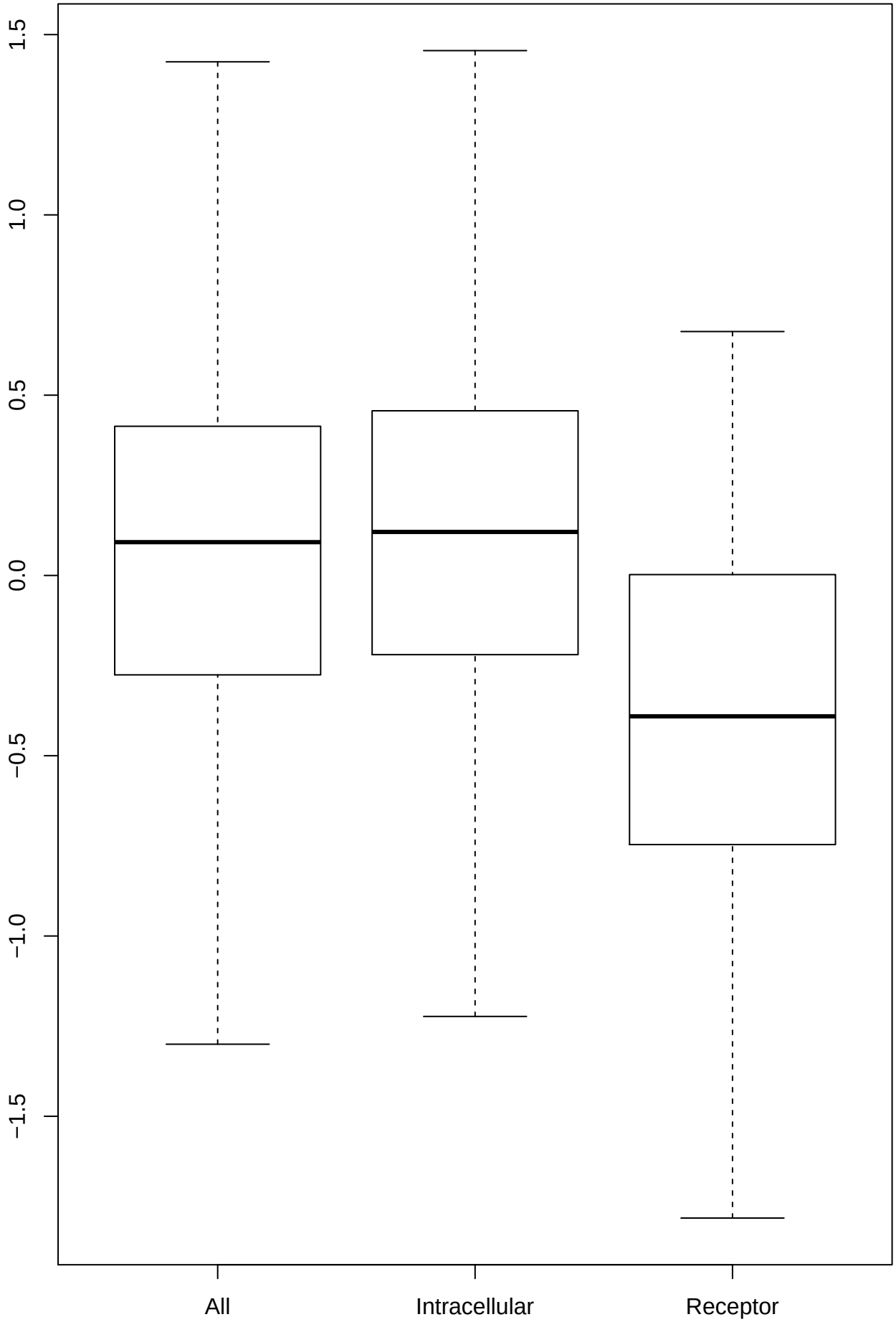



Figure 8