Final Progress Report for Research Projects Funded by Health Research Grants

Instructions: Please complete all of the items as instructed. Do not delete instructions. Do not leave any items blank; responses must be provided for all items. If your response to an item is "None", please specify "None" as your response. "Not applicable" is not an acceptable response for any of the items. There is no limit to the length of your response to any question. Responses should be single-spaced, no smaller than 12-point type. The report **must be completed using MS Word**. Submitted reports must be Word documents; they should not be converted to pdf format. Questions? Contact Health Research Program staff at 717-783-2548.

1. **Grantee Institution:** The Trustees of the University of Pennsylvania

2. **Reporting Period (start and end date of grant award period):** 1/1/2012-12/31/2013

3. **Grant Contact Person (First Name, M.I., Last Name, Degrees):** Gearline R. Robinson-Hall, BSF

4. **Grant Contact Person's Telephone Number:** 215-746-6821

5. **Grant SAP Number:** 4100050912

6. **Project Number and Title of Research Project:** 6 - High-Throughput High-Resolution Tools for RNA-Based Functional Genomics

7. **Start and End Date of Research Project:** 1/1/2010-12/31/2013

8. **Name of Principal Investigator for the Research Project:** Junhyong Kim, PhD

9. **Research Project Expenses**.

   9(A) Please provide the total amount of health research grant funds spent on this project for the entire duration of the grant, including indirect costs and any interest earned that was spent:

   $ 1,780,680.91

   9(B) Provide the last names (include first initial if multiple individuals with the same last name are listed) of **all** persons who worked on this research project and were supported with health research funds. Include position titles (Principal Investigator, Graduate Assistant, Post-doctoral Fellow, etc.), percent of effort on project and total health research funds expended for the position. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name, First Name	Position Title	% of Effort on Project	Cost
Chekholko, Alexander	System Programmer Sr.	Yr. 1- 13% Yr. 2 – 19%	Yr. 1- $13, 286.59 Yr. 2 –$6,860.59
Chen, Stephanie L	Student Worker	Yr.2 – 8% Yr. 3 – 30%	Yr.2 -$362.50 Yr. 3 - $2,342.50
Grosser, Tilo	Research Assistant Professor	Yr. 2 – 83% Yr. 3 – 42%	Yr. 2 - $8,483.90 Yr. 3 - $8,483.90
Lawson, John A	Associate Director	Yr. 2 – 82% Yr. 3 – 36%	Yr. 2 - $22,157.32 Yr. 3 - $22,157.32
Liang, Xue	Research Fellow	Yr. 2 – 75% Yr. 3 – 43%	Yr. 2 - $6,125.01 Yr. 3 - $6,125.01
Pawelski, Sven-Christian	Postdoctoral Researcher	Yr. 2 – 74% Yr. 3 – 59%	Yr.2 -$10,769.00 Yr. 3 - $10,769.00
Peritz, Adam	Resource Technologist	Yr. 1 - 9% Yr. 2 – 7% Yr. 3 – 2%	Yr. 1 - $10,417.52 Yr. 2 - $8,188.90 Yr. 3 - $1,940.10
Pizarro, Angel	IT Technical Director	Yr. 1 -100% Yr.2 – 25%	Yr. 1 - $23,755.54 Yr. 2- $23,755.54
Price, Thomas Stephen	Research Assistant Professor	Yr. 2 – 86% Yr. 3 – 86%	Yr. 2 - $17,000.01 Yr. 3 - $17,000.01
Saigal, Ashmita	Master's Student	Yr. 1 - 100% Yr. 2 – 89% Yr. 3 – 51%	Yr. 1 - $2,330.00 Yr. 2 - $19,729.86 Yr. 3 - $17,399.86
Song, Wenliang	Research Assistant Professor	Yr. 2 – 78% Yr. 3 – 78%	Yr. 2 - $16,333.36 Yr. 3 - $16,333.36
Suzuki, Tatsunori	Postdoctoral Researcher	Yr. 2 – 74% Yr. 3 – 36%	Yr. 2 - $12.346.90 Yr. 3 - $12,346.90
Yang, Guangrui	Research Associate	Yr. 2 – 83% Yr. 3 – 41%	Yr. 2 - $19,661.68 Yr. 3 - $19,661.68
Zolekar, Ashwini	Student Worker	Yr. 2 – 90% Yr. 3 – 57%	Yr. 2 - $5,310.00 Yr. 3 - $5,310.00

9(C) Provide the names of **all** persons who worked on this research project, but who *were not* supported with health research funds. Include position titles (Research Assistant, Administrative Assistant, etc.) and percent of effort on project. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name, First Name	Position Title	% of Effort on Project
Kim, Junhyong	Professor	20%
Eberwine, James	Professor	10%
Hogenesch, John	Professor	10%
Dueck, Hannah	Graduate Student	50%
Fisher, Steve	Research Associate	10%
McGroty, Sean	Research Technician	20%
Middleton, Sarah	Graduate Student	25%

Lahens, Nicholas	Graduate Student	20%
Khaladkar, Mugdha	Postdoctoral Fellow	20%

9(D) Provide a list of **all** scientific equipment purchased as part of this research grant, a short description of the value (benefit) derived by the institution from this equipment, and the cost of the equipment.

Type of Scientific Equipment	Value Derived	Cost
Illumia HiSeq 2000		$317,000

10. Co-funding of Research Project during Health Research Grant Award Period. Did this research project receive funding from any other source during the project period when it was supported by the health research grant?

Yes___X_____ No_____

If yes, please indicate the source and amount of other funds:

NIMH R01MH088849--$ 800,000

11. Leveraging of Additional Funds

11(A) As a result of the health research funds provided for this research project, were you able to apply for and/or obtain funding from other sources to continue or expand the research?

Yes____X____ No_____

If yes, please list the applications submitted (column A), the funding agency (National Institutes of Health—NIH, or other source in column B), the month and year when the application was submitted (column C), and the amount of funds requested (column D). If you have received a notice that the grant will be funded, please indicate the amount of funds to be awarded (column E). If the grant was not funded, insert "not funded" in column E.

Do not include funding from your own institution or from CURE (tobacco settlement funds). Do not include grants submitted prior to the start date of the grant as shown in Question 2. If you list grants submitted within 1-6 months of the start date of this grant, add a statement below the table indicating how the data/results from this project were used to secure that grant.

A. Title of research project on grant application	B. Funding agency (check those that apply)	C. Month and Year Submitted	D. Amount of funds requested:	E. Amount of funds to be awarded:
Evolution of novel cytoplasmic introns in mammalian neurons	X NIH ☐ Other federal (specify:_____) ☐ Nonfederal source (specify:_)	October, 2013	$ 2,224,840	Pending
High performance IBM iDataPlex/SONAS computing cluster for genomics	XNIH ☐ Other federal (specify:_____) ☐ Nonfederal source (specify:_)	July, 2012	$1,954,859	$1,953,909.77

11(B) Are you planning to apply for additional funding in the future to continue or expand the research?

Yes__X_____ No_____

If yes, please describe your plans:

We will continue to expand the combinatorial library research program for effecting transdifferentiation. Because this is a technology development project, it is difficult to acquire NIH/NSF funding. We will seek to explore other avenues including the DARPA research program.

12. Future of Research Project. What are the future plans for this research project?

This project helped establish single cell analysis as a major strength of research at Penn. The NIH has initiated multiple strategic programs targeting single cell level analysis. We will be creating a new internal center program to expand on the current research and to respond to external opportunities.

13. New Investigator Training and Development. Did students participate in project supported internships or graduate or post-graduate training for at least one semester or one summer?

Yes___X_____ No_____

If yes, how many students? Please specify in the tables below:

	Undergraduate	Masters	Pre-doc	Post-doc
Male				1
Female	1			
Unknown	1	1		1
Total	**2**	**1**		**2**

	Undergraduate	Masters	Pre-doc	Post-doc
Hispanic				
Non-Hispanic	1			
Unknown	1	1		2
Total	**2**	**1**		**2**

	Undergraduate	Masters	Pre-doc	Post-doc
White				
Black				
Asian	1			
Other				
Unknown	1	1		2
Total	**2**	**1**		**2**

14. **Recruitment of Out-of–State Researchers**. Did you bring researchers into Pennsylvania to carry out this research project?

    Yes_____ No____X_____

    If yes, please list the name and degree of each researcher and his/her previous affiliation:

15. **Impact on Research Capacity and Quality**. Did the health research project enhance the quality and/or capacity of research at your institution?

    Yes___X____ No_____

    If yes, describe how improvements in infrastructure, the addition of new investigators, and other resources have led to more and better research.

    This project played a key role in establishing a state-of-art next generation sequencing facility for transcriptome characterization. As described in the main report section, we completed more than 70 full RNA sequencing runs per year, involving more than 1,500 different samples. The facility has served members of the Penn Genome Frontiers Institute as well as the broad Penn biomedical research community.

16. Collaboration, business and community involvement.

16(A) Did the health research funds lead to collaboration with research partners outside of your institution (e.g., entire university, entire hospital system)?

Yes____X____ No_____

If yes, please describe the collaborations:

Our sequencing facilities were used by others in the local Philadelphia research community including researchers at the Wistar Institute, the Monell Institute, and the National Disease Research Interchange

16(B) Did the research project results in commercial development of any research products?

Yes_____ No__X_____

If yes, please describe commercial development activities that resulted from the research project:

16(C) Did the research lead to new involvement with the community?

Yes_____ No___X_____

If yes, please describe involvement with community groups that resulted from the research project:

17. Progress in Achieving Research Goals, Objectives and Aims.

List the project goals, objectives and specific aims (as contained in the grant agreement). Summarize the progress made in achieving these goals, objectives and aims for the period that the project was funded (i.e., from project start date through end date). Indicate whether or not each goal/objective/aim was achieved; if something was not achieved, note the reasons why. Describe the methods used. If changes were made to the research goals/objectives/aims, methods, design or timeline since the original grant application was submitted, please describe the changes. Provide detailed results of the project. Include evidence of the data that was generated and analyzed, and provide tables, graphs, and figures of the data. List published abstracts, poster presentations and scientific meeting presentations at the end of the summary of progress; peer-reviewed publications should be listed under item 20.

This response should be a DETAILED report of the methods and findings. It is not sufficient to state that the work was completed. Insufficient information may result in an unfavorable performance review, which may jeopardize future funding. If research findings are pending

publication you must still include enough detail for the expert peer reviewers to evaluate the progress during the course of the project.

Health research grants funded under the Tobacco Settlement Act will be evaluated via a performance review by an expert panel of researchers and clinicians who will assess project work using this Final Progress Report, all project Annual Reports and the project's strategic plan. After the final performance review of each project is complete, approximately 12-16 months after the end of the grant, this Final Progress Report, as well as the Final Performance Review Report containing the comments of the expert review panel, and the grantee's written response to the Final Performance Review Report, will be posted on the CURE Web site.

There is no limit to the length of your response. Responses must be single-spaced below, no smaller than 12-point type. If you cut and paste text from a publication, be sure symbols print properly, e.g., the Greek symbol for alpha (α) and beta (ß) should not print as boxes (□) and include the appropriate citation(s). DO NOT DELETE THESE INSTRUCTIONS.

Project Overview

Genomics has impacted every area of biological and biomedical science and its concepts and technologies have become fundamental to understanding human diseases. A key aspect of genome science is that it has been powered by the development of new technologies that allow efficient measurement and manipulation of biomolecules in ever larger numbers, at ever greater throughput, and at ever finer resolutions. In particular, the measurement and manipulation of RNA populations have led to the development of functional genomics and to the proposals for RNA-based therapeutics. While the study of genome-wide RNA populations is already more than 15 years old, continued discovery of novel functional RNA elements is still opening up entirely new areas of biology. Quantitative assessment and manipulation of RNA at high-resolution is just starting. <u>The goal of this proposal is to develop new high-throughput tools to manipulate and measure the broad functional RNA populations of cells at high resolution, including at the level of single cells, and to examine the potential for quantitative RNA manipulations for therapeutics. The specific aims of this project are:</u>

1. Establish a facility for next generation RNA sequencing at high resolution including up to single-cell RNA sequencing.
2. Prototype a new tool for generating a quantitative library of RNA that can be used for transfection and other quantitative cell manipulation studies.
3. Use the facilities and the tools established in Aims 1 and 2 to understand RNA population variability at multiple levels in mammalian systems and to investigate the roles of noncoding expressed RNA.

We report the following progress and findings for each of the three aims.

Aim 1: Establish a facility for next generation RNA sequencing at high resolution with single cell sequencing.

We developed a new Next Generation high-throughput sequencing (HTS) facility within the Penn Genome Frontiers Institute (PGFI). Using partial support from this grant and additional institutional support, we purchased one ABI Solid 4 system and two Illumina HiSeq 2000 systems. Subsequently the machines were upgraded to the SOLiD 4 system to SOLiD 5500 system that gained 2-fold additional capacity. We additionally upgraded the Illumina HiSeq to model 2500 using internal funds and also acquired an Illumina MiSeq platform. While other facilities are available on campus, this facility specialized in RNA sequencing and single cell RNA sequencing. This facility carried out an average of ~80 HiSeq flow cells per year (approximately 1,300 samples) starting in October of 2010. On balance 75% of the sequencing runs were for high-resolution RNA/transcriptome analysis. We have amassed more than 500 single cell RNA sequencing data using this facility and created a new data repository for single cell data. In 2013, the facility was moved under the Institute for Translational Medicine and Advanced Therapeutics to better assist in translational studies. The upgraded machines continue to serve the community for RNA characterization.

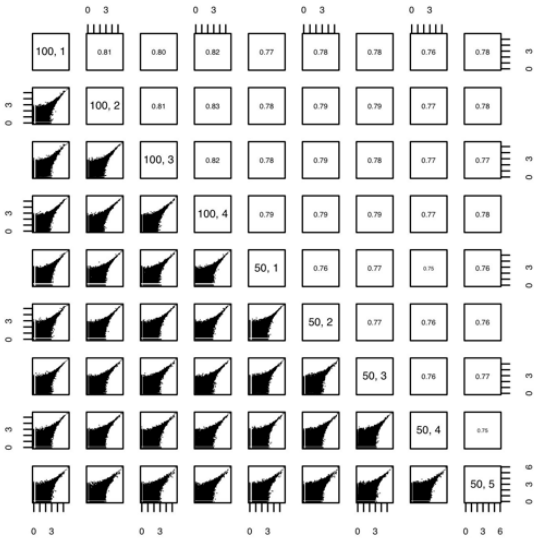
Next-generation sequencing requires considerable computational processing to generate human-usable information from the raw data produced by the sequencing machines. For RNA sequencing application using 100 million sequence reads, we have benchmarked computational loads using the data produced by our HTS facility. For complete mapping using a Penn-developed program, RUM (RNA Unified Mapper), 1800 CPU hours (average of 2 Ghz AMD CPUs) using 6 GB of main memory and 250-500 GB of output space is required. The amount of output is data dependent and is a function of sequence redundancy of the genome. To meet these computing demands, we have concentrated on upgrading both our hardware infrastructure and establishing analysis pipelines. During this project, we added 390 TB of new Isilon network storage device to Penn Genome Frontiers Institute's High-Performance Computing cluster to service the RNA sequencing aims of this grant. To help efficient data processing, we originally developed a computational analysis pipeline for RNA sequencing data that utilizes both our compute resources and external cloud vendor environments. The core of this pipeline was the RNA Unified Mapper developed at Penn (Grant et al., Bioinformatics 27:2518). This step of the pipeline was modified to include the STAR aligner (Dobin et al. Bioinformatics 29:15), which led to more than an 100-fold speed up with little loss of sensitivity. We leveraged the funding of this project and the large amount of RNA sequencing data that was being generated by the facility established in Aim 1 to apply for a NIH large instrumentation grant. We were successful in obtaining a ~$2 million grant and used this to create a new computing facility for next generation sequencing and genomic medicine. The data generated by this project was invaluable in establishing the rationale for this grant.

In addition to the standard alignment pipeline, we have developed several computational infrastructures and algorithms to help support RNA/DNA analysis from next generation sequencing. One key pipeline was to help process single cell transcriptome data from In Vitro Transcription (IVT) amplification (called aRNA procedure). Single cell data involves several special characteristics. First, the sequencing reads can contain unique artifacts that arise out of the amplification chemistry including long stretches of poly-A sequences, IVT primers, etc. We

have developed special filters to eliminate such sequences from the reads before alignment. Second, because the starting material is so minute, the process can be subject to considerable contamination problems. We have developed a procedure to randomly sample a selection of reads and use a full BLAST search against a known nucleotide database. The output then is mapped to taxonomic groups and reported in a table. This Quality Control (QC) procedure, for example, helped us track down a problem where the contamination originated in a reagent supplied by a commercial vendor. This pipeline is available as open source in GitHub (https://github.com/safisher/ngs/wiki). We also refined several RNA quantification and normalization methods for single cell sequencing and characterized control datasets. For example, Fig. 1 show an example of technical control experiment where we diluted bulk mouse cardiomyocyte RNA into 50pg and 100pg total RNA (estimated 1pg and 2pg of mRNA) and independently amplified and sequenced the samples to ~30 million read depth. These control samples range from 0.75 to 0.83 in pairwise Pearson correlation and pass high QC filters. A log-reads vs. log-rank plot of the dilution controls (most other biological single cell datasets show similar plots), showing a power-law relationship for the read distribution with a sharp inflection point around rank 5,000, most likely reflecting technical dropouts of low expressed genes. Using a Poisson model for dilution replicates as well as an analysis of the relationship between fano factor and mean expression, we calculated that we could recover with high degree of quantitative accuracy approximately two mRNA molecules per sample. Assuming recovery efficiency from cellular extracts of 25-50%, we estimate that we will be able to quantitatively assay single cell transcriptomes from genes with 4-8 mRNA molecules per cell. Additional theoretical computation based on the coupon collector's model for approximately 300,000 estimated mRNA molecules in a human B cell also suggests that 30 million reads will recover most of the transcriptome with high probability.

In addition to these computational procedures for single cell transcriptomes we also developed a Bayesian integration scheme to quantify RNA expression from RNA sequencing where reads that map to redundant positions are probabilistically integrated. Briefly, one of the problems in quantifying RNA molecules from sequence data is that typically a portion (10-50%) of sequence reads map to multiple locations in the genome. This leads to a conundrum in how to convert the presence of multiply-mapped reads to the number of RNA molecules transcribed from the mapped location of the genome. Recently, we developed a Bayesian scheme for estimating Single Nucleotide Polymorphism (SNP) from multiply mapped sequences (Simola and Kim, published in Genome Biology doi:10.1186/gb-2011-12-6-r55). The main idea here was to generate a prior model of possible genomic locations for sequence origination and then a conditional probability model for sequence reads dependent on the reference genome. We are working to adopt this model to RNA transcriptome quantification. The main approach is to utilize a transfer function from probabilistic distribution of mapped read densities to output RNA quantities.

Fig 1: Independent dilution control experiments of 100pg and 50pg (labeled on the diagonal boxes).



By establishing a standard protocol for single cell RNA measurements we were able to obtain a grant supplement of ~$1 million over a four years (2013 to 2016) for a related single cell analysis project from NIH to serve as a data coordination group for three different institutions.

SA 2: Prototype a new tool for generating a quantitative library of RNA that can be used for transfection and other quantitative cell manipulation studies.

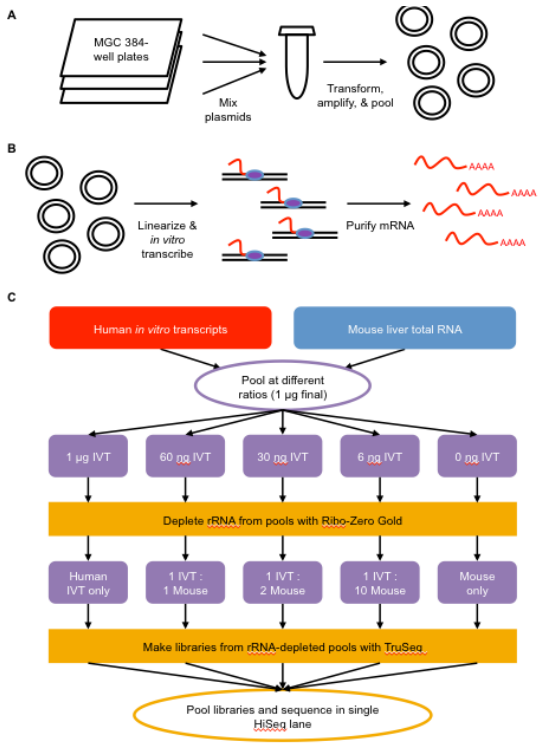


Fig 2. Experimental design to assess IVT and RNA sequencing quantification bias.

The original aim of SA 2 was to develop a combinatorial library of In Vitro Transcribed (IVT) RNA that can be used in a quantitative formula to transdifferentiate cell types and also to perturb the cells in a systematic manner for functional genomic studies. While we have executed many of the fundamental foundations for this goal, because of the technical difficulties, we are behind on the final steps of this aim. Here, we describe the current progress and then describe the steps towards completion of this aim.

We utilized an expression library in the laboratory of co-investigator John Hogenesch (Mammalian Gene Collection; Temple et al., Genome Research 19:2324). The original plans required quantification of the target cell's transcriptome to generate appropriate combinatorial formula and also quantification

of the IVT libraries. Early on, we realized that the RNA sequencing methods for very different IVT product may have biases in quantification. Therefore, we generated a project to assess possible bias. The IVT libraries from 384-well plates were expressed, quantified at the total level, and then split into complex pool mixture design (Fig 2). This allowed us to assess different degrees of bias arising out of RNA sequencing and also from the IVT procedure. An example is shown in Fig 3. This shows a pile up plot of RNA sequencing reads from five different treatments for the locus BC003353. The first line, labeled "Simulated" shows expected read density recovered from an in silico simulation of RNA expression and sequencing reads. As expected, the coverage across the locus is uniform. The second line labeled "Plasmid" shows mapping of a direct sequencing from the DNA (the expression vector), which also shows relatively low levels of variation across the locus. The next line shows "No selection" where the total RNA was used for sequencing. The results show some variability but less than the next two lines. The fourth line shows read depths for using a subtraction protocol to deplete rRNA and the fifth line shows the read depths for using poly-A selection. As can be seen in these two lines both of these protocols result in highly variable read depths across the locus. Other data suggests that this bias also depends on the locus. The bias for the poly-A samples is more predictable with expected 5' to 3' gradient. The bias for the rRNA depletion procedure is much more unpredictable from locus to locus. These results suggested to us that we need to treat the quantitative data much more carefully as described next. We have submitted these control results as a manuscript to Genome Biology.

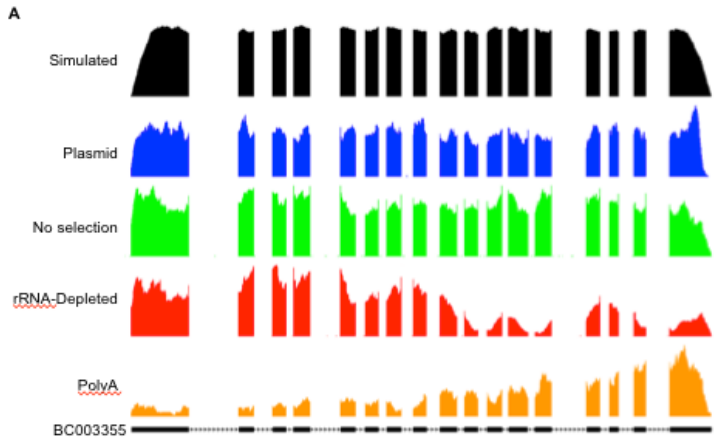


Fig 3: Pile up of next generation sequencing reads for one example locus from In Vitro Transcription samples. Line 1 shows simulated data; line 2 shows DNA sequencing from the expression vector; line 3 shows RNA sequencing using total RNA; line 4 shows RNA sequencing using a subtractive procedure for removing rRNA; line 5 shows using poly-A selection.

Because of our control data described above, we modified our combinatorial transfection strategy to use qualitative on/off data from only the transcription factor subset of the

transcriptome. This is the same strategy used in the generation of the induced pluripotent stem (ips) cells in Yamanaka et al. (2006; Cell 126:663). In a related but separate study, we characterized the single cell transcriptome of 91 individual cells from five different cell types in mice (13 brown adipocytes, 19 cardiomyocytes, 19 cortical pyramidal neurons and 18 hippocampal pyramidal neurons from embryonic mouse and 22 serotonergic neurons from adult mouse). We examined the transcription factor compartment of these single cell data sets and used a noise factor to discretize the values into ON and OFF. To focus on genes likely to be involved in developing and maintaining cell identity, we first selected a subset of genes with significantly low within-cell type expression variation relative to between-cell type variation (Bonferonni-corrected p-value < 0.01). Using this set of 247 transcription factors, we found the minimum spanning network, or union of minimum spanning trees, that connects all cells with the fewest possible on/off expression state changes of the transcription factors.

Though this network (Fig 4) finds the most parsimonious description of transcriptional differences across individual cells, we observed a large amount of expression variability over the graph. On average, two connected cells demonstrate 47 differences in transcription factor expression, or differences in nearly 1/5 of the input genes. Expression differences are large even between cells of the same type, with on average 46 genes switched on or off between connected cells. While variation across cell types is larger, averaging 52 genes over these edges, it is not significantly different from that within each cell type (Welch's one-sided t-test, p=0.06) consistent with the above results of large cell-to-cell variability in gene expression. Despite this, most connections in the transcription factor network are between cells of the same type (91 out of 104 edges), suggesting cell type-specific covariation patterns in the underlying transcription factor expression profiles despite large variation at the individual gene level. Central nervous system (CNS) cells are connected to non-CNS cells by a single edge, indicating greater consistency among expression profiles within each of these broad phenotypic categories. Pyramidal neurons of different brain regions demonstrate many paths with minimal transcriptional changes, unlike that observed at boundaries between other cell types, again suggestive of underlying structure in expression profiles that correlates with phenotypic similarity.

We hypothesized that transcription factors demonstrating limited numbers of on/off state switches across the network may be important for producing and maintaining cell type identity. Fig. 4 depicts the expression changes of 23 genes demonstrating 10 or fewer switches across the network. All 23 demonstrate expression changes over edges across cell types, or edges emanating from the bridging cells. The path to cardiomyocytes involves expression switches in transcription factors related to heart development (*Ankrd1, Gata6, Irx4* and *Smyd1*) and sarcomere function (*Abra* and *Actn2*), and includes many genes demonstrating heart-related mutant phenotypes (*Tgif2*, in addition to *Abra, Ankrd1, Gata6, Irx4* and *Smyd1*). Transitional edges to brown adipocytes involve genes associated with fat cell differentiation (*Pparg* and *Trib3*) and one with a mutant phenotype of impaired thermogenesis (*Stat5a*). Interestingly, *Irf7* switches at these edges and is expressed in all brown adipocytes, but has no known role in brown adipocyte identity. An expression change in *Basp1* marks the edge connecting CNS and non-CNS cells, whose knockout produces abnormal neuron and axon morphology. Edges transitioning to Dorsal Raphe neurons demonstrate expression changes in *Fev*, a gene associated with serotonergic neuron identity. Over network edges transitioning to pyramidal neurons, transcription factors switch on that are involved in neurogenesis and neuron cell fate determination (*Bhlhe22, Foxg1, Lhx2, Lhx6, Neurod1*, and *Neurod2*) and axon guidance and

migration (*Arx*, *Bhlhe22*, *Foxg1*, and *Lhx2*). Several, on knockout, increase risk of seizures (*Arx*, *Lhx6*, *Neurod1*, and *Neurod2*) and produce abnormal hippocampus and cerebral cortex morphology (*Arx*, *Bhle22*, *Foxg1*, *Lhx2*, *Lhx6*, and *Neurod1*). Notably, three genes found at this transition have been used in lineage conversion experiments to generate neuronal cell types (*Neurod1*, *Neurod2*, and *Foxg1*). These results suggest that minimal switching genes are candidates for development and maintenance of cell type identity and we hypothesize that sets of low variability genes with observed switches at cell type boundaries might be promising candidates for direct lineage conversion experiments. We also note that no gene on this list is ubiquitously expressed within one cell type and also ubiquitously unexpressed in all other cells, suggesting that ensembles of genes, rather than individual regulators, likely underlie cellular phenotype with the result that multiple pathways in transcriptome states exist between cell phenotypes.

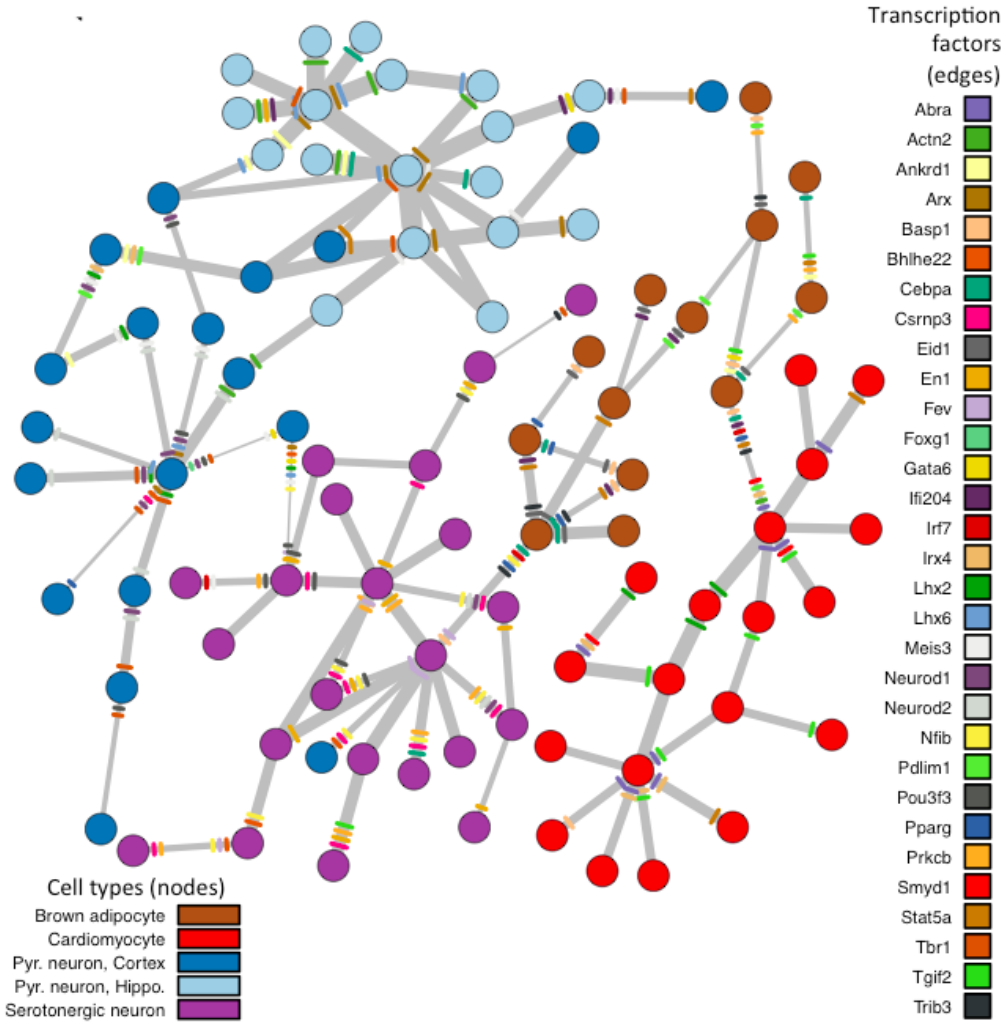


Fig. 4. Similarity network of single cells based on binary transcription factor expression. Cells (nodes) are connected by a minimum spanning network, a graph that connects all cells with minimal changes in transcription factor expression. Edge width indicates the number of differences in transcription factor expression with wide edges indicating greater similarity. Colored hash marks

indicate differences in transcription factor expression, with hashes immediately next to a cell indicating factors expressed by that cell but not by the neighboring cell. 23 transcription factors with ten or fewer expression changes across the network are shown.

We used the information above to design combinatorial transfection experiments using the factors that define the brown adipose tissue cells. This target cell type is especially appropriate as it has well-defined phenotype (lipid inclusions). We carried out our first phototransfection experiments in 2014 and we are waiting on the transfection results. We hope to continue to optimize the protocols, especially in terms of transfection efficiency and the combination of transcription factors.

SA 3: Use the facilities and the tools established in Aims 1 and 2 to understand RNA population variability at multiple levels in mammalian systems and to investigate the roles of noncoding expressed RNA.

As mentioned above, we have been using the RNA sequencing facility and the computational pipelines to characterize a large number of single cell transcriptomes. In particular, we used our high-resolution transcriptome techniques to examine the variability and the role of non-coding RNA. Here, we report on two papers, one published and another under revision. In the first study, we used subcellular RNA dissection to characterize non-coding intronic sequences that are found in dendrites of mouse and rat neurons. In the second, we developed a new algorithm to identify novel functional RNA structures.

We performed a comprehensive transcriptome wide survey of the prevalence of cytoplasmic intron sequences retained transcripts (CIRTs) in rat and mouse dendritic population. Two batches of dendritic mRNA from primary rat hippocampal neurons and three batches of dendritic mRNA from primary mouse hippocampal neurons, each consisting of 150–300 individually dissected dendrites were independently aRNA amplified and subjected to Illumina NextGen sequencing. Of the reads mapping to intronic regions, only the uniquely mappable reads were used by paired-end Bowtie alignment. Mechanically dissected dendrites are expected to contain only non-nuclear RNA. However, to overcome the potential issue of nuclear hnRNA contamination as well as the presence of cryptic transcriptional units (e.g., cryptic exons, anti-sense RNA, etc), we carried out computational filtering to obtain the most probable set of cytoplasmic intron sequences. First, each intron with putative cytoplasmically retained sequence was also filtered for overlap with predicted genes using the tools GENSCAN (Burge and Karlin, 1997; J. Mol. Biol. 268:78) and TWINSCAN (Korf et al., 2001; Bioinformatics 17:S140) to minimize any other possible cryptic translated sequences. Second, we also screened the introns with retained sequence for the presence of intronic poly(A)-sites by using the PolyA_DB2 database which reports poly(A) sites obtained using alignments between cDNA/ESTs and genome sequences, and removed those that contained a poly(A) site. Second, we examined the remaining reads for pattern of sequence-read coverage and found no significant enrichment in intergenic reads, nor did we find any systematic pattern of directional bias in intron presence or absence across all genes that might occur with partial products of nuclear splicing. Third, we analyzed the possibility that

independently transcribed ncRNAs may be present within the introns that give rise to the CIRTs. Fourth, another possible origin of the CIRTs is from non-specific anti-sense transcription. Since the RNA-seq protocol that we followed is not strand specific, it is difficult to completely rule out the possibility of antisense transcription resulting in these intronic reads. However, to get a lower limit on the prevalence of antisense transcription in the regions overlapping with the retained introns that we have reported here, we made use of the multi-exonic antisense transcripts from mouse embryonic stem cells (ESC), neural progenitor cells (NPC) and mouse lung fibroblasts (MLF) reported by a recent study that carried out *ab inito* reconstruction of these transcriptomes (Guttman et al., 2010; Nat. Biotech 28:503). We found <5% of the retained introns from mouse samples to be having at least a partial overlap with these antisense transcripts. Thus, although we cannot completely rule out all instances of antisense transcription, this comparison suggests that a majority of the retained introns are not likely to be a result of antisense transcription. Finally, a small fraction of the reads mapped to the exon-intron junctions and were the only reads that mapped to the particular intron. It is possible that these reads represent an extension of the exonic sequence and correspond to un-annotated alternate exon start or end sites. Hence, we also excluded these introns from all further analysis.

The main findings that we report in Khaladkar et al. (2013; PLoS One 8(10):e76194) is that there is a surprising number of genomic loci with evidence of cytoplasmic intron sequences. We found ~44% rat genes and ~59% of the mouse genes in dendritic samples showing evidence for the presence of CIRTs (Table 1). The median read density in the intronic regions of these genes ranged between 3-7 reads per 50 bp while the maximum read density was in the range of 950-2350 reads per 50 bp. We also identified a set of transcripts in dendritic samples of rat and mouse wherein the sequence evidence comes only from within the intronic regions of a gene model (7-14%) (Table 1). To estimate the length frame of the retained intronic sequence, we clustered the reads that were separated by no more than 500 bp to form contiguous regions, which we call contigs. We chose to have this larger distance threshold in order to account for the reads arising from repetitive sequences that tend to be enriched in intronic regions and may not be uniquely mappable. This distance is also roughly equal to two back-to-back paired-end reads. The average fraction of retained introns per gene amongst all introns of that gene was 1. The fraction of retained introns per gene was not correlated with read depths ($r=0.03$, p-value = 0.009) suggesting that the detection of a restricted subset of possible introns is not due to lack of read coverage. Thus, the putative cytoplasmically retained intronic reads arise from a specific subset of introns rather than through random retention of intronic sequences. A large fraction of these introns gave rise to a single contig; on average 86% of all retained intronic sequences are within a single contig. Since the mRNA amplification protocol recognizes the poly-A tail for amplification, there is a possibility for a bias towards 3'-most introns in the detection of intronic retention. We analyzed whether the set of retained introns occurs primarily from the 3' half of the transcript. The results showed a roughly equal proportion of retained introns from the 5' half and 3' half of all the genes with >= 5 introns combined and there was no pattern of 5' or 3' bias of introns in the CIRTs and the frequency of CIRTs was uniform as a function of 5' to 3' introns of a gene model. We also found that ~47% of all retained introns contained an internal pA tract of length >= 10 bp. It is likely that these pA tracts are used during mRNA

amplification and thus they allow us to sample intronic sequences distal from the poly-A end of a transcript. We scanned these contigs for the maximal open reading frame (ORF) length in all six frames (three in the sense and three in the antisense orientation) and found a large majority of the contigs (>75%) to have an ORF length of less than 60 amino acids. Also, >80% of the retained intronic contigs have at least one stop codon in all six frames in the sense and antisense orientation. Therefore, the possibility that these contigs code for proteins is unlikely.

To summarize the above, we found a broad prevalence of CIRTs in the neurons in both species (44-60% of the expressed transcripts). The sequence patterns, including stereotypical length, biased inclusion of specific introns, and intron-intron junctions, suggested CIRT-specific nuclear processing. Our analysis also suggested that these cytoplasmic intron-sequence retaining transcripts may serve as a primary transcript for ncRNAs. Our results show that retaining intronic sequences is not isolated to a few loci but may be a genome-wide phenomenon for embedding functional signals within certain mRNA. The results hypothesize a novel source of cis-sequences for post-transcriptional regulation. Our results hypothesize two potentially novel splicing pathways: one, within the nucleus for CIRT biogenesis; and another, within the cytoplasm for removing CIRT sequences before translation. We also speculate that release of CIRT sequences prior to translation may form RNA-based signals within the cell potentially comprising a novel class of signaling pathways.

Table 1. Prevalence of CIRTs in dendritic and soma mRNAs from rat and mouse hippocampus neurons

Sample	Total Genes	Genes with exonic reads	Genes with exonic and intronic reads	Genes with only intronic reads	Total Genes with intronic reads	Genes with a given intron retained in all samples
Mouse/Dendrite	14,478	12,344	6,473 (52.4%)	2,134 (14.7%)	8,607 (59.4%)	648 (7.5%)
Rat/Dendrite	10,104	9,308	3,689 (39.6%)	796 (7.8%)	4,485 (44.4%)	687 (15.3%)
Rat/Soma	8,750	8,308	2,079 (25%)	442 (5.1%)	2,521 (28.8%)	110 (4.4%)
Mouse/Soma	12,200	11,255	3,990 (35.4%)	945 (7.7%)	4,935 (40.4%)	115 (2.3%)

Finally, because many non-coding RNAs are hypothesized to form secondary structures with possible molecular physiological functions, we developed a new algorithm to predict novel functional RNA without folding the RNA. In work we submitted to RNA (Middleton and Kim, 2013), we designed an approach to RNA structure clustering that does not require structure prediction or alignment. Our approach is inspired by the idea of an “empirical kernel”, where the distance between any two objects is computed within an observation-spanned subspace by comparing each object to a set of empirical examples or models. To measure the distance between the structures of two RNAs, we construct an empirical feature space using covariance

models (CMs), which are a form of stochastic context-free grammar used by the Rfam database to model the consensus sequence and secondary structure of RNA structure families. The raw feature space consists of 1,973 dimensions, each corresponding to one of the 1,973 CMs in the Rfam v.10.1 database. To obtain the coordinates of an arbitrary RNA sequence within this space, we score the sequence against each CM using the *cmscore* module of Infernal (v.1.0.2; Nawrocki et al., 2009; Bioinformatics 25:1335) and use the resulting bitscores as the coordinates along each axis. These bitscores indicate how well the sequence matches each CM, taking into account compensatory base changes that maintain conserved pairing interactions. Thus, the feature space maps RNA sequences according to their similarity to known structures. To test our procedure, we created three artificial structures of the same length but with different numbers of hairpins and randomly generated sequences that had the appropriate base complementarity to form each of these structures. A total of 50 sequences were generated for each structure. The sequences for each structure appeared random on the primary sequence level (25% average pairwise sequence identity), making the three structure groups indistinguishable by sequence alone. We scored each sequence against the Rfam CMs and projected them into the RESS. As an initial assessment of the relative positioning of each artificial structure's sequences within the RESS, we performed PCA on the projection. The different structural sequences formed three well-separated clusters along the first and second PC axes, indicating that the RESS mapped the sequences with similar structure closer together than sequences of different structure.

We next sought to use NoFold to identify potentially novel structure motifs in experimental datasets. We chose two datasets that we hypothesized might contain biologically relevant RNA structure motifs: dendritically localized RNAs identified in rat neurons described above, in which we search for structures that might serve as localization signals; and non-canonical translation initiation sites in humans identified through ribosome profiling, in which we search for upstream motifs that may promote initiation at these sites. The results of these analyses are summarized in the following sections, and full data on all identified motifs will be made available on our website (kim.bio.upenn.edu/software/nofold.shtml).

As a first step towards identifying structural Dendritic Targeting Elements (DTEs), we compiled a list of 211 transcripts with experimental evidence for dendritic localization (see Methods). From each transcript, we obtained the 3'UTR sequence as well as the sequence of any cytoplasmically retained introns (see above), which have previously been shown to harbor DTEs. We created 50nt and 150nt sliding window sets for the retained intron and 3'UTR sequences of the dendritically localized transcripts and searched these regions for motifs using NoFold. The results of these analyses are summarized in Table 2. NoFold identified a total of 290 motifs that occurred three or more times across the sequences. Previously, Buckley and colleagues found that a ~74nt hairpin structure within the retained introns of several dendritic transcripts was sufficient to confer dendritic localization in rat hippocampal neurons (Buckley et al., 2011; Neuron 69:877) These structures were instances of the ID element, a type of rodent SINE retrotransposon that likely arose from the dendritically-localized BC1 gene. We asked whether the ID element structure was among the motifs found by NoFold in our intron sequences. We found two motifs in the 50nt set and one motif in the 150nt set that had high sequence homology to the ID element, all of which were significantly enriched in the dendritic introns (Fisher's exact test, FDR-adjusted $p < 0.05$). The 150nt motif was additionally predicted to form a highly similar structure to the ID hairpin (Fig. 5A). This motif contained sequences overlapping 10 of the 12

BLAST hits for the ID element within the intron sequences, and additionally contained one extra instance of the ID element not found by BLAST. Although this extra sequence had low sequence identity with the ID hairpin sequence (59%), it was structurally conserved (SCI = 0.83) and was predicted to form a similar hairpin structure.

In addition to the ID element, we also identified several motifs with similarity to known localization elements from *Drosophila*. Most strikingly, we found that 37 motifs were annotated as having the K10 transport/localization element CM (K10_TLS; RF00207) among their top ten best CMs, with five motifs having an average Z > 5 and 28 having a Z > 3 for this CM. The K10_TLS is a 44nt stem-loop structure that mediates localization of the K10 mRNA during *Drosophila* oocyte development (Serano and Cohen, 1995; Development 121:3809). The majority of our K10_TLS-like motifs are predicted to have a stem-loop consensus structure enriched with AU base pairs, similar to K10_TLS (Fig. 5B), though none show obvious homology on the primary sequence level. Overall, these 37 clusters encompassed a total of 60 unique genes, which is 28% of the total genes in the datasets. We also found nine motifs with another *Drosophila* localization structure, the Wingless localization element 3 (WLE3; RF01046), within their top ten CMs (but only one with an average Z > 3). To our knowledge, a role for these motifs has not yet been described in mammals. Additionally, we identified several potentially novel motifs with stable and conserved structure, such as hairpin motif M172, which is found in six dendritic transcripts, and double-hairpin motif M158, which is found in four transcripts (Fig 5C).

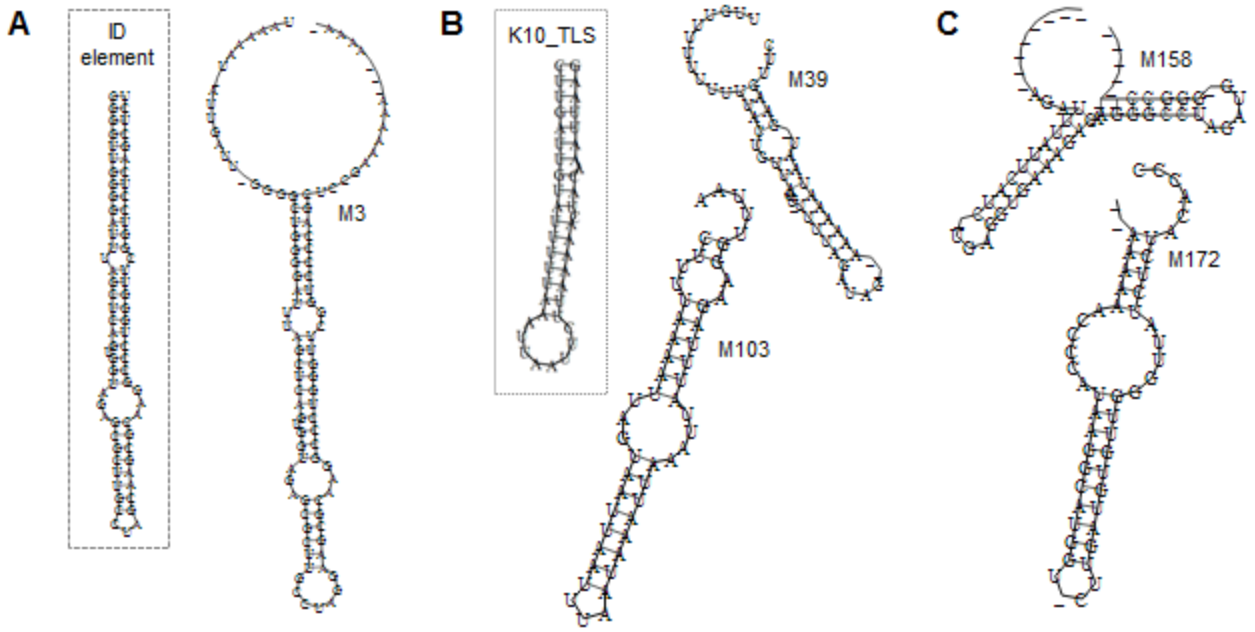


Figure 5: Example novel RNA structures predicted by NoFold algorithm

TABLE 2. Clustering sensitivity of NoFold and GraphClust for three test conditions on the Rfam benchmark dataset.

					Plain sequences		Embedded sequences		Plain seqs with BG	
Family	**Rfam ID**	**#Seqs**	**Avg % ID**	**Avg Len**	**NoFold**	**GraphClust**	**NoFold**	**GraphClust**	**NoFold**	**GraphClust**
5S_rRNA	RF00001	100	49%	116	1.00	1.00	0.20	**1.00**	**1.00**	0.99
5_8S_rRNA	RF00002	22	54%	149	0.91	**0.95**	**0.86**	0	0.86	**0.95**
U1	RF00003	20	48%	162	0	0	0	0	0	0
U2	RF00004	70	47%	188	1.00	1.00	1.00	1.00	1.00	1.00
tRNA	RF00005	100	40%	73	**0.92**	0.91	**0.72**	0	**0.91**	0.90
Vault	RF00006	52	50%	101	**0.96**	0.94	0.50	**0.94**	0.94	**0.96**
U12	RF00007	27	46%	165	1.00	1.00	**1.00**	0.85	0.89	**1.00**
Hammerhead_3	RF00008	13	45%	55	**0.85**	0	0	0	0.85	**0.92**
RNaseP_nuc	RF00009	68	32%	303	**0.74**	0.62	0.49	**0.54**	0.50	**0.60**
RNaseP_bact_a	RF00010	100	49%	360	1.00	1.00	1.00	1.00	1.00	1.00
RNaseP_bact_b	RF00011	41	53%	357	0	**1.00**	1.00	1.00	1.00	1.00
U3	RF00012	38	41%	204	0.92	0.92	0.87	**0.95**	**0.82**	0
6S	RF00013	86	38%	181	**0.98**	0.90	**0.77**	0.60	0.79	**0.99**
U4	RF00015	61	45%	145	**0.97**	0.95	0.66	**0.95**	**0.97**	0.95
SNORD14	RF00016	7	44%	110	0	0	0	0	0	0
Metazoa_SRP	RF00017	17	45%	290	0.94	0.94	0.94	**1.00**	0.94	0.94
CsrB	RF00018	7	53%	340	**1.00**	0	**1.00**	0	**1.00**	0
Y_RNA	RF00019	24	47%	97	1.00	1.00	0.96	**1.00**	1.00	1.00
U5	RF00020	82	44%	117	**1.00**	0.99	1.00	1.00	**1.00**	0.99
Histone3	RF00032	43	45%	46	**0.86**	0.65	**0.26**	0	0.79	**0.91**
Background	-	3000	25%	215	-	-	-	-	0	0
				Avg sensitivity	**0.80**	0.74	**0.66**	0.59	**0.81**	0.76
				Avg precision	0.98	**0.99**	**0.99**	0.98	**0.99**	0.98

18. Extent of Clinical Activities Initiated and Completed. Items 18(A) and 18(B) should be completed for all research projects. If the project was restricted to secondary analysis of clinical data or data analysis of clinical research, then responses to 18(A) and 18(B) should be "No."

18(A) Did you initiate a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

____Yes
__X_No

18(B) Did you complete a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

_____Yes
__X__No

If "Yes" to either 18(A) or 18(B), items 18(C) – (F) must also be completed. (Do NOT complete 18(C-F) if 18(A) and 18(B) are both "No.")

18(C) How many hospital and health care professionals were involved in the research project?

_____Number of hospital and health care professionals involved in the research project

18(D) How many subjects were included in the study compared to targeted goals?

_____Number of subjects originally targeted to be included in the study
_____Number of subjects enrolled in the study

Note: Studies that fall dramatically short on recruitment are encouraged to provide the details of their recruitment efforts in Item 17, Progress in Achieving Research Goals, Objectives and Aims. For example, the number of eligible subjects approached, the number that refused to participate and the reasons for refusal. Without this information it is difficult to discern whether eligibility criteria were too restrictive or the study simply did not appeal to subjects.

18(E) How many subjects were enrolled in the study by gender, ethnicity and race?

Gender:
_____Males
_____Females
_____Unknown

Ethnicity:
_____Latinos or Hispanics
_____Not Latinos or Hispanics
_____Unknown

Race:
_____American Indian or Alaska Native
_____Asian
_____Blacks or African American
_____Native Hawaiian or Other Pacific Islander
_____White
_____Other, specify:_____
_____Unknown

18(F) Where was the research study conducted? (List the county where the research study was conducted. If the treatment, prevention and diagnostic tests were offered in more than one county, list all of the counties where the research study was conducted.)

19. Human Embryonic Stem Cell Research. Item 19(A) should be completed for all research projects. If the research project involved human embryonic stem cells, items 19(B) and 19(C) must also be completed.

19(A) Did this project involve, in any capacity, human embryonic stem cells?
_____Yes
_X___ No

19(B) Were these stem cell lines NIH-approved lines that were derived outside of Pennsylvania?
_____Yes
_____ No

19(C) Please describe how this project involved human embryonic stem cells:

20. Articles Submitted to Peer-Reviewed Publications.

20(A) Identify all publications that resulted from the research performed during the funding period and that have been submitted to peer-reviewed publications. Do not list journal abstracts or presentations at professional meetings; abstract and meeting presentations should be listed at the end of item 17. **Include only those publications that acknowledge the Pennsylvania Department of Health as a funding source** (as required in the grant agreement). List the title of the journal article, the authors, the name of the peer-reviewed publication, the month and year when it was submitted, and the status of publication (submitted for publication, accepted for publication or published.). Submit an electronic copy of each publication or paper submitted for publication, listed in the table, in a PDF version 5.0.5 (or greater) format, 1,200 dpi. Filenames for each publication should include the number of the research project, the last name of the PI, and an abbreviated title of the publication. For example, if you submit two publications for Smith (PI for Project 01), one publication for Zhang (PI for Project 03), and one publication for Bates (PI for Project 04), the filenames would be:

Project 01 – Smith – Three cases of isolated
Project 01 – Smith – Investigation of NEB1 deletions
Project 03 – Zhang – Molecular profiling of aromatase
Project 04 – Bates – Neonatal intensive care

If the publication is not available electronically, provide 5 paper copies of the publication.

Note: The grant agreement requires that recipients acknowledge the Pennsylvania Department of Health funding in all publications. Please ensure that all publications listed acknowledge the Department of Health funding. If a publication does not acknowledge the funding from the Commonwealth, do not list the publication.

Title of Journal Article:	Authors:	Name of Peer-reviewed Publication:	Month and Year Submitted:	Publication Status (check appropriate box below):
1. IVT-seq reveals extreme bias in RNA-sequencing	Lahens, N.F., Kavkli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R.S., Grant, G.R., Hogenesch, J.	Genome Biology	12/2014	XSubmitted ☐Accepted ☐Published
2. NoFold: RNA structure clustering without folding or alignment	Middleton, S. and J. Kim	RNA	10/2014	X Submitted ☐Accepted ☐Published
3. Cell identity by single-cell transcriptome variation	Dueck, H., Khaladkar, M., Kim, T.K., Spaethling, J.M., Francis, C., Suresh, S., Fisher, S.A., Seale, P., Beck, S.G., Bartfai, T., Kuhn, B., Eberwine, J., and J. Kim	Genome Research	10/2013	X Submitted ☐Accepted ☐Published
4. Subcellular RNA sequencing reveals broad presence of cytoplasmic intron-sequence retaining transcripts in mouse and rat neurons	Khaladkar M1, Buckley PT, Lee MT, Francis C, Eghbal MM, Chuong T, Suresh S, Kuhn B, Eberwine J, Kim J.	PLoS One 8(10):e7619 4	11/2013	☐ Submitted ☐Accepted X Published
5. Sniper: improved SNP discovery by multiply mapping deep sequenced reads	Simola, D. and J. Kim	Genome Biology	01/2011	☐ Submitted ☐Accepted X Published

20(B) Based on this project, are you planning to submit articles to peer-reviewed publications in the future?

Yes___X___ No_____

If yes, please describe your plans:

We are still carrying out the combinatorial library transfection experiments for cell phenotype remodeling. Our goal is to demonstrate the effectiveness of such a scheme and publish a methodology paper in 2014.

21. Changes in Outcome, Impact and Effectiveness Attributable to the Research Project. Describe the outcome, impact, and effectiveness of the research project by summarizing its impact on the incidence of disease, death from disease, stage of disease at time of diagnosis, or other relevant measures of outcome, impact or effectiveness of the research project. If there were no changes, insert "None"; do not use "Not applicable." Responses must be single-spaced below, and no smaller than 12-point type. DO NOT DELETE THESE INSTRUCTIONS. There is no limit to the length of your response.

None

22. Major Discoveries, New Drugs, and New Approaches for Prevention Diagnosis and Treatment. Describe major discoveries, new drugs, and new approaches for prevention, diagnosis and treatment that are attributable to the completed research project. If there were no major discoveries, drugs or approaches, insert "None"; do not use "Not applicable." Responses must be single-spaced below, and no smaller than 12-point type. DO NOT DELETE THESE INSTRUCTIONS. There is no limit to the length of your response.

None

23. Inventions, Patents and Commercial Development Opportunities.

23(A) Were any inventions, which may be patentable or otherwise protectable under Title 35 of the United States Code, conceived or first actually reduced to practice in the performance of work under this health research grant? Yes_____ No___X_____

If "Yes" to 23(A), complete items a – g below for each invention. (Do NOT complete items a - g if 23(A) is "No.")

a. Title of Invention:

b. Name of Inventor(s):

c. Technical Description of Invention (describe nature, purpose, operation and physical, chemical, biological or electrical characteristics of the invention):

d. Was a patent filed for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?
Yes_____ No___

If yes, indicate date patent was filed:

e. Was a patent issued for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?
Yes_____ No___
If yes, indicate number of patent, title and date issued:
Patent number:

Title of patent:
Date issued:

f. Were any licenses granted for the patent obtained as a result of work performed under this health research grant? Yes_____ No____

If yes, how many licenses were granted?_____

g. Were any commercial development activities taken to develop the invention into a commercial product or service for manufacture or sale? Yes___ No____

If yes, describe the commercial development activities:

23(B) Based on the results of this project, are you planning to file for any licenses or patents, or undertake any commercial development opportunities in the future?

Yes_____ No___X_____

If yes, please describe your plans:

24. Key Investigator Qualifications. Briefly describe the education, research interests and experience and professional commitments of the Principal Investigator and all other key investigators. In place of narrative you may insert the NIH biosketch form here; however, please limit each biosketch to 1-2 pages. *For Nonformula grants only – include information for only those key investigators whose biosketches were not included in the original grant application.*

BIOGRAPHICAL SKETCH Provide the following information for the key personnel and other significant contributors in the order listed on Form Page 2.Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**			
NAME Junhyong Kim		POSITION TITLE Professor	
eRA COMMONS USER NAME junhyong			
EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*			
INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Seoul National University, Seoul, Korea	B.S.	1984	Microbiology
SUNY at Stony Brook, Stony Brook , NY	MS/PhD	1986/1992	Ecology and Evolution
University of Arizona, Tucson, AZ	Postdoc	1992-1994	Drosophila Genetics

A. Personal Statement:

My laboratory works at the intersection of genomics, evolutionary biology, and quantitative biology. I have 25 years of experience in computational and theoretical biology as well as 15 years of experience in molecular biology and genomics. In my lab, we typically ask evolutionary questions applied to genomic models systems using experimental and quantitative approaches. Recently we have been working on various aspects of cell regulation mechanisms including cell differentiation, single-cell transcriptome variation, and RNA-dependent dendritic function in mammalian neurons. In particular, our key goal is to understand the regulation and evolution of single cell function in vertebrates. My lab carries out both computational and experimental work and we have also worked on developing new genomic technologies as well as algorithms. We have published in the fields of evolution, ecology, systematics, human genetics, anthropology, computational biology, genomics, neuroscience, and biophysics. In training, I have personally mentored 40 undergrads, graduate students, and postdocs. I have created new diversity training programs within Penn Genome Frontiers Institute and have developed a public high school genomics curriculum.

B. Positions and Honors

Edmund J. and Louise W. Kahn Professor, July 2002—present, Department of Biology, University of Pennsylvania, joint appointment in Department of Computer and Information Science.
Co-Director, Penn Genomics Institute, Oct 1, 2006--present
Jan 1994 – June 2002: Assistant and Associate Professor (with tenure), Department of Biology (currently Ecology and Evolutionary Biology), Yale University. Joint appointment in Department of Statistics, Department of Molecular, Cellular, and Developmental Biology, and Biomedical Engineering program.

Sept. 2001- June 2002: Director, BBS graduate track in Bioinformatics and Computational Biology, Yale Univ.

Ellison Medical Foundation Senior Scholar in Aging, 2010
Guggenheim Fellow, 2010
Yale Senior Faculty Award
Yale Seessel Anonymous Award for studies in Biochemistry
Alfred Sloan Foundation Young Investigator Award in Molecular Evolution

C. SELECTED RECENT PUBLICATIONS

1. Anafi, R.C., Lee, Y., Sato, T.K., Venkataraman, A., Ramanathan, C., Hughes, M.E., Baggs, J., Liu, A.C., Kim, J., and Hogenesch, J.B., 2014. Large-scale data integration identifies Gm129/Chrono as a novel repressor of the mammalian circadian clock. In press, *PLoS Biology*.
2. Middleton, S. and **J. Kim** 2014 NoFold: RNA structure characterization without folding or alignment. In press, *RNA*.
3. Khaladkar, M., Buckley, P.T., Lee, M., **Eberwine, J**. and **J. Kim** 2013. A genome-wide survey of cytoplasmically retained intron sequence transcripts through single cell sequencing. *PLoS One 8(10):e76194*.
4. Kim, T.K., Sul, J.-Y., Kim, J. and J. Eberwine. 2013. Translational dynamics of dendritic glutamate receptor mRNAs show micron-scale spatial dependence. *Cell Reports 5(1):114-125*.
5. Daugharty, E., Goodman, A., and **J. Kim** 2012. Pervasive antisense transcription is conserved in budding yeast. *Mol. Biol. Evol.*, doi:10.1093/molbev/mss240.
6. **Eberwine, J.,** Lovatt D., Buckley, P., Dueck, H., Francis, C., Kim, T.K., Lee, J., Lee, M., Miyashiro K., Morris, J., Peritz, T., Schochet, T., Spaethling, J., Sul, J.-Y., and **J. Kim** 2012. Quantitative biology of Single Cells. J. *Roy. Soc. Interface* doi:10.1098/rsif.2012.0417.
7. Simola, D., and **J. Kim**. 2011. Using multi-locus Bayesian model and redundantly mapped sequence reads to enhance SNP discovery with next-generation sequencing. Genome Biology 12:R55.
8. Kim, T., Sul, J.Y., Peternko, N., Lee, M., Patel, V., Kim, J. and J. Eberwine. 2011 Transcriptome transfer provides a model for understanding the phenotype of cardimyocytes. PNAS doi:10.1073/pnas.1101223108.
9. Buckley, P.T., Lee, M.T., Sul, J.-Y., Miyashiro, K.Y., Bell, T.J., Fisher, S.A., **Kim, J. and J. Eberwine**. 2010. Broad pattern of intron retention promotes multiple mechanisms for localization of dendritic RNA. Neuron 69:877.-884

BIOGRAPHICAL SKETCH

NAME **James H. Eberwine**	POSITION TITLE Professor of Pharmacology and Psychiatry
eRA COMMONS USER NAME eberwine	

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE	YEAR(s)	FIELD OF STUDY
Yale University, New Haven, CT	B.S.	1978	Biochem and Philosophy
Columbia University, New York, NY	Ph.D.	1984	Biochemistry
Stanford University, Stanford, CA	Postdoc	1984-1989	Molecular Neurobiology

A. Personal Statement.

Over the last 20 years I have pioneered the molecular analysis of pharmacological and disease-associated mRNA from small amounts of tissue including single cells by developing approaches to enable this level of analysis. For example, using RNAs isolated from single dendrites followed by aRNA amplification (developed in my lab), we found that dendrites have a greater RNA complexity than was suspected (~1000 mRNAs localized in dendrites). In another study, we demonstrated protein synthesis in dendrites by transfecting an epitope-tag encoding mRNA into isolated live dendrites and showing that the tag was translated. Such analyses of dendrites provide insight into the functioning of this important substrate of cognition. With regard to molecular analysis of CNS disorders the aRNA procedure has been used to generate probes for >70% of all published mRNA microarray studies thereby dramatically impacting biomarker discovery and therapeutic development. To make this scientific progress, I have had to develop many techniques including those involved in RNA binding protein and mRNA cargo identification (PAIR and APRA procedures), phototransfection for introducing mRNA into individual dendrites and cell somas and single cell PCR and the aRNA procedures for the analysis of mRNA in single cells. Among these discoveries is the surprising fact that many dendritically localized mRNAs have retained introns that can be spliced in the cytoplasm of the cell to give rise to the functionally relevant mRNAs. This is a developing story in my lab. Recently our TIPeR (transcriptome induced phenotype remodeling) paper pioneered an approach for converting one cellular phenotype to a distinct phenotype through the transfer of RNA populations.

B. Positions and Honors.

Positions and Employment

1990-1994 Assistant Professor, Depts. of Pharmacology and Psychiatry, Univ. of Penn School of Medicine

1994-1996 Associate Professor, Depts. of Pharmacology and Psychiatry, Univ. of Penn School of Medicine

1996-present Professor, Depts. of Pharmacology and Psychiatry, Univ. of Pennsylvania School of Medicine

2006-present Co-Director, PENN Genomics Institute

2007-present Elmer Holmes Bobst Chair of Pharmacology, University of Pennsylvania School of Medicine

Other Experience and Professional Memberships (selected)

1987-1993	Founding Course Director for "Cloning of Neural Genes" for first 7 years, Cold Spring Harbor
1992-1997	Established Investigator of the American Heart Association
1997	Presidential Lecturer for Society for Neuroscience Meeting
1998	Matthew Moore Distinguished Lecture, AANP
1998	Director's Lecture, NIH
1998	Dresel Lectureship - Dalhousie University
1998	Distinguished Visiting Scientist, Albany Medical College
1999	Plenary Lecturer at 1st Korea/USA Joint Workshop on Neuroscience
2000	Dean's Lecture at University of Kentucky
2001	Presidential Symposium Association for Research in Otolaryngology
2002	Presidential Symposium – Society for the Study of Reproduction
2003	Andre McInvale Memorial Lecturer, University of Colorado
2005-09	Course Co-Director, Advanced Technologies in Molecular Neuroscience, Cold Spring Harbor
2010	Keynote Lecturer at Royal Society International Symposium on "Making light deliver"
2012	Founding Course Director, Single Cell Analysis, Cold Spring Harbor
2012	Organizer – Society for Neuroscience Short Course entitled "Transcriptomics"
2012-15	Course Director, Single Cell Analysis, Cold Spring Harbor (1st Course of this type)
2013-18	External Advisory Board – Institute for Genomic Biology – Univ. of Illinois

Honors (Selected)

2001-2011	MERIT Award
2005	Stephen and Constance Lieber Distinguished Investigator for the (NARSAD)
2008	NIH Director's Pioneer Award Recipient
2009	EUREKA grant (co-recipient with Junhyong Kim)
2011	Ellison Medical Foundation Senior Scholar in Aging
2012	McKnight Technological Innovations in Neuroscience Award
2013	Clarenberg Lecture – Kansas State University

C. Selected Recent Peer-Reviewed Publications

1. Buckley, P., Lee, M., Sul, J-Y, Miyashiro, K., Bell, T., Fisher, S., Kim, J. and Eberwine, J.(2011): Cytoplasmic intron sequence-retaining transcripts (CIRTs) can be dendritically targeted via ID element retrotransposons. Neuron, 69:877-884. [PMID 21382548]
2. Kim, T-K., Sul, J-Y., Peternko, N., Lee, M., Patel, V., Kim, J. and Eberwine, J. (2011): tCardiomyocyte: a model for generating and understanding the phenotype of cardiomyocytes. Proc. Natl. Acad. Sci., 108(29):11918-23. [PMID:21730152].
3. Sul, J-Y., Kim, T. K., Lee, J. and Eberwine, J.: Perspectives on cell reprogramming with RNA. Trends in Biotechnology, *doi:10.1016/j.tibtech.2012.02.004, 2012.*
4. Kim, T-K, Kim, J. and Eberwine, J.: Translational Dynamics of Dendritic Glutamate Receptor mRNAs show Micron-Scale Spatial Dependence, *Cell Reports, in press*.

BIOGRAPHICAL SKETCH Provide the following information for the key personnel and other significant contributors in the order listed on Form Page 2.Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**			
NAME Stephen Fisher	POSITION TITLE Manager Research Project B		
eRA COMMONS USER NAME			
EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*			
INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
University of California, San Diego	B.A./B.A.	1991/1991	Biology/Psychology
Yale University	MS	1994	Psychobiology
Yale University	MPhl/Phd	1995/1999	Behavioral Neuroscience

B. <u>Personal Statement:</u>

For the past two decades I have focused my efforts on the application of computational techniques to the study of biological systems and varying levels of detail. My works began with the exploration of large cellular networks, progressing to interactions between individual cells and most recently to subcellular dynamics. At each stage I've worked to develop new computational tools and adapt existing tools to the biological systems being explored. I have 20 years of expertise in programming SQL, JAVA, Perl, R, and Python. I manage the primary computational analysis of all next generation sequencing data in the Kim lab and manage the data repository. I develop standard operating procedures for informatics and help manage the laboratory information systems. I am familiar with systems programming and high performance computing cluster administration. I have developed several bioinformatics tools including GLOP, GGO, and CRIMSON.

<u>B. Positions and Honors</u>

2002 - present, **Manager Research Project B**, Department of Biology, University of Pennsylvania.

2007, 2009, **Organizing Committee**, Workshop on Algorithms in Bioinformatics.

1999 - 2002, **Research Specialist**, Ecology and Evolutionary Biology, Yale University.

1996 - 1998, **NRSA Fellowship**, NIMH.

1992 - 1996, **Edwin S. Reid Fellowship**, Yale University.

<u>C. SELECTED RECENT PUBLICATIONS</u>

Buckley, P.T., Lee M.T., Sul, J.Y. Miyashiro, K.Y., Bell, T.J., Fisher, S.A., Kim, J. and Eberwine J. (2011). Cytoplasmic Intron Sequence-Retaining Transcripts Can Be Dendritically

Targeted via ID Element Retrotransposons. *Neuron*, **69**: 877-884. doi: 10.1016/j.neuron.2011.02.028.

Fisher, S.A., Fischer, T.M., and Carew, T.J. (1997). Multiple overlapping processes underlying short-term synaptic enhancement. *Trends. Neurosci.*, **20**: 170-177.

Jaffe, D.B., Fisher, S.A. and Brown, T.H. (1994). Confocal laser scanning microscopy reveals voltage-gated calcium signals within hippocampal dendritic spines. *J. Neurobio.*, **25**: 220-233. doi: 10.1002/neu.480250303.

Pouget, A., Fisher, S.A. and Sejnowski, T.J. (1993). Egocentric Representation in Early Vision. *J. Cog. Neurosci.*, **5**(2): 150-161.

BIOGRAPHICAL SKETCH Provide the following information for the key personnel and other significant contributors in the order listed on Form Page 2.Follow this format for each person. **DO NOT EXCEED FOUR PAGES.**	
NAME Mugdha Khaladkar	POSITION TITLE
eRA COMMONS USER NAME mugdha	Post-doctoral Researcher

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	YEAR(s)	FIELD OF STUDY
Pune University	B.E.	2004	Computer Engineering
New Jersey Inst. Of Technology	PhD	2004/2009	Computer Science (Bioinformatics)
University of Pennsylvania	Postdoc	2009-present	Computational Genomics in mammalian neurons

C. Personal Statement:

My primary strengths are in Computational Biology using which I investigate the fascinating interplay of different factors (chromatic structure, non-coding RNAs, RNA structure, cis and trans elements etc.) that govern the various cellular processes (polyadenylation, RNA localization, splicing etc.) ultimately leading to dynamic gene regulation. I have over eight years of experience working on several independent and collaborative projects in Functional genomics, Comparative genomics and Molecular evolution. At present my project seeks to understand the sub-cellular localization of mRNAs in the context of mammalian neurons. I am interested in discovering the elements involved in this process and in doing so gain insight into evolutionary mechanisms that have resulted in vast functional differences between the different species.

B. Positions and Honors

May. 2005 - May. 2009: Research Assistant in Department of Biochemistry and Molecular Biology, Medical school of New Jersey (UMDNJ).
June, 2007 – Aug 2007: Intern, Bell Labs, Alcatel-Lucent
June, 2006 – Aug 2006: Database Intern, Alcatel-Lucent
Treasurer, Graduate Students Association of New Jersey Inst. Of Technology
Certificate of Distinction at the 3rd and 4th Annual Provost's Student Research Showcase at New Jersey Institute of Technology– 2007, 2008
Student Recognition Award from the Alumni Association of New Jersey Institute of Technology – 2007
UPS Foundation Certificate for academic excellence – 2006

C. SELECTED RECENT PUBLICATIONS

1. **Mugdha Khaladkar**, Pete T. Buckley, Miler T. Lee, Chantal Francis, et al.: Subcellular RNA Sequencing Reveals Broad Presence of Cytoplasmic Intron-sequence Retaining Transcripts in Mouse and Rat Neurons. PLoS ONE 8(10): e76194, 2013.
2. Peter T. Buckley, **Mugdha Khaladkar**, Junhyong Kim and James Eberwine: Cytoplasmic intron retention, function, splicing and the sentinel RNA hypothesis. WIREs RNA. doi: 10.1002/wrna.1203, 2013
3. **Mugdha Khaladkar**, Sridhar Hannenhalli: Functional divergence of gene duplicates: a domain-centric view. *BMC Evol Biol*, Vol. 12, 2012
4. Lei Hua, Jason T.L. Wang, Xiang Ji, Ankur Malhotra, **Mugdha Khaladkar**, Bruce A Shapiro and Kaizhong Zhang: A Method for Discovering Common Patterns from two RNA Secondary Structures and it application to structural repeat detection. *Journal of Bioinformatics and Computational Biology*, Vol. 10, No. 4 2012
5. **Mugdha Khaladkar**, Mark Smyda and Sridhar Hannenhalli: Epigenomic and RNA Structural Correlates of Polyadenylation. *RNA Biology*,Vol. 8, Issue3, 529-537, 2011
6. **Mugdha Khaladkar**, Jianghui Liu, Dongrong Wen, Jason T.L. Wang and Bin Tian: Mining small RNA structure elements in untranslated regions of human and mouse mRNAs using structure-based alignment. *BMC Genomics*, 9, 189, 2008
7. **Mugdha Khaladkar**, Vandanaben Patel, Vivian Bellofatto, Jeffrey Wilusz and Jason T.L. Wang: Detecting conserved secondary structures in RNA molecules using constrained structural alignment. *Comput Biol Chem*, Vol. 32, Issue 4, 2008
8. **Mugdha Khaladkar**, Vivian Bellofatto, Jason Tsong-Li Wang, Bin Tian and Bruce A. Shapiro: RADAR: A Web Server for RNA Data Analysis and Research. *Nucleic Acids Research*, 35:W300-W304, 2007