Small Regulatory RNAs in Mammals: Genomics, Function and Evolution

by

Jinkuk Kim

B.S., Korea Advanced Institute of Science and Technology (2004)

Submitted to the Harvard-MIT Division of Health Science and Technology
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Signature of Author .....
Harvard-MIT Division of Health Science and Technology
March 31, 2011

Certified by .....
David P. Bartel, Ph.D.
Professor of Biology
Supervisor

Accepted by .....
Ram Sasisekharan, Ph.D.
Director, Harvard-MIT Division of Health Science and Technology / Edward Hood Taplin
Professor of Health Science and Technology and Biological Engineering

Small Regulatory RNAs in Mammals: Genomics, Function and Evolution

by

Jinkuk Kim

Submitted to the Harvard-MIT Division of Health Science and Technology
on March 31, 2011 in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Bioinformatics and Integrative Genomics

ABSTRACT

This thesis explores two aspects of small regulatory RNAs in mammals: (1) the genomic origin of mammalian piwi-interacting RNAs (piRNAs), (2) the evolutionary and functional implication of the seed-based target recognition mechanism of microRNAs (miRNAs).

First, we participated in the discovery of mammalian piRNAs from adult rat testes. Our initial characterization of mammalian piRNAs by high-throughput sequencing revealed the peculiar features of their genomic origin: they predominantly derive from long single-stranded RNA precursors that are encoded at ~100 loci with no preferential association to repeat elements.

Second, we measured the efficacy of polymorphic miRNA target sites in mammals. A large part of the miRNA-target recognition is determined by the 7—8-nt match between the seed region of miRNAs and the 3'UTR of mRNAs. Because of the small informational complexity of the specificity, spontaneous point mutations in 3'UTRs often create or disrupt miRNA target sites. The resulting polymorphisms in the target sites may contribute to gene expression diversity. By experimentally measuring the efficacy of such polymorphic target sites, we were able to conclude that between two unrelated mammalian individuals of the same species more than 100 genes are likely differentially regulated due to the target-site polymorphisms. Some of the expression diversity might translate into phenotypic diversity, providing substrates for the natural selection to act upon.

We also constructed a miRNA library covering nearly all ~16,000 theoretically possible seed sequences. Under the assumption that the functionality of a miRNA is approximately defined by the identity of the seed, the library is a resource that may enable the systematic exploration of the phenotypic consequences of nearly all possible functionally distinct miRNA species.

Thesis Supervisor: David P. Bartel, Professor of Biology, MIT

ACKNOWLEDGEMENTS

First, I thank my advisor Dave for always encouraging me to follow my heart. No matter what I decided to work on, he supported me by providing with all the freedom, resource and advice. I am deeply grateful for the precious opportunities.

I thank Tyler and Chris. It was an extraordinary honor to have, in my committee, two of the scientists whom I have admired greatly.

I thank my academic adviser Leonid Mirny for his warm counseling every semester, and my program director Zak Kohane for the opportunity to study in these great institutions.

I thank my collaborators: Nelson Lau, Anita Seto and Bob Kingston for the piRNA project; Tim Harkins for the SNP project; Su Wu, Jen Grenier, Wenjun Guo, Wendy Johnston, Andrew Grimson, Rudolf Jaenisch, Bob Weinberg and Whitehead Institute Genome Technology Core for the miRNA screen project; Alan Buckler, Mike Lam and Christine Mayr for the cancer 3'UTR mutation project; Vijay Shankaran and Harvey Lodish for the miRNA in trisomy 13 project; Junghwan Sung for the miRNA in prostate cancer project; Wei Chen and H. Hilger Ropers for the miRNA in mental retardation project.

I feel very fortunate about having been able to work every day with so many talented colleagues in Bartel lab.

I thank my bay-mate Sue-Jean for her support in many respects; Calvin and Mike Lam for their help especially during the period when I was learning experiments; Laura and Lori for the numerous favors.

I also thank Huili, Jinwoo, Garcia, Wendy, Anna, Vincent and many others in the lab, for being such good friends and helping colleagues.

I thank Robin Schanche for the literary assistance on a part of this thesis.

I thank Hyungseok, Seokhee, Sangwon, Jiwoon, Sungwhan, Euiheon, Kwonmoo, Ha, Byungsup, Joonsik and my many other Korean friends, for making my Boston life fun and meaningful.

Last, I thank my mom, my dad and my sister for everything.

TABLE OF CONTENTS

CHAPTER I

Introduction

Small Regulatory RNAs in Mammals

MicroRNAs (miRNAs) are ~22-nt, endogenous RNAs that direct post-transcriptional repression of protein-coding genes (Bartel, 2004). The first miRNA, *lin-4*, was discovered serendipitously in 1993 during forward genetic screens in worms (Lee et al., 1993). But it was only since 2001 that miRNAs were starting to be appreciated as a large class of molecules with a wide range of functionalities, encoded at hundreds of loci in the genomes of most metazoan species (Reinhart et al., 2000; Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001).

In the biogenesis of miRNAs (Kim et al., 2009), one of the most characteristic features is that miRNAs are derived from precursors with a small hairpin structure, which folds back to form a stem loop structure. The precursors in animals, which are usually generated as introns of protein coding genes or as independent non-coding pol-II transcripts, are subject to stepwise processing by multiple protein machineries. They are first subject to Drosha/DGCR8-mediated cleavage, which releases the hairpin from the rest of the precursor. The hairpin is then exported out of the nucleus with the help of exportin-5 for Dicer-mediated cleavage in cytoplasm. The cleavage eliminates the loop part of the hairpin, so that the remaining small RNA duplex can be put into action by being loaded into RISC

(RNA-induced silencing complex). Some exceptions to this processing pathway include mirtrons and endogenous shRNA, which bypass the Drosha processing (Okamura et al., 2007; Ruby et al., 2007a; Babiarz et al., 2008; Chiang et al., 2010), and miR-451, which bypasses the Dicer processing (Cheloufi et al., 2010).

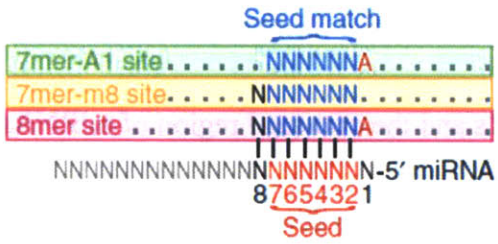


Figure 1. Three types of miRNA target sites.

Once loaded into RISC, miRNAs direct gene repression by providing RISC with the specificity on the genes to be targeted for repression. The most common mode of target recognition is through partial base-pairing of the miRNAs with messenger RNAs (Bartel, 2009). In this mode, the interaction between miRNAs and target messages is nucleated by so-called "seed-pairing", the perfect consecutive 6-nt base-pairing of the miRNA "seed" (position 2—7) with 3' untranslated regions (3'UTRs) of the messages (Lewis et al., 2003; Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005). But the 6-nt seed-pairing is

only marginally effective in mediating repression, unless it is augmented (1) by an additional base-pairing at position 8 (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005), forming "7mer-m8" sites in 3'UTRs, (2) by a invariant nucleotide "A" across miRNA position 1(Lewis et al., 2005), forming "7mer-A1" sites, or (3) by both, forming "8mer" sites (Figure 1). Efficacy of the three types of sites are known to be also affected by other properties, such as local AU-content, distance of the site from the center of the 3'UTR, and supplemental pairing between the 3'UTRs and the non-seed region of miRNAs (Grimson et al., 2007). The seed-based interaction leads to repression of gene expression, mostly through destabilization of the engaged messages, and to also a lesser extent through translational inhibition (Baek et al., 2008; Filipowicz et al., 2008; Hendrickson et al., 2009; Guo et al., 2010).

Another mode of target recognition is through the extensive base-pairing between miRNAs and messenger RNAs (Yekta et al., 2004; Davis et al., 2005). In this mode, miRNAs direct RISC to catalyze an endonucleolytic cleavage of the recognized messages at the phosphodiester bond between the nucleotides that are paired to miRNA nucleotides 10 and 11, resulting in degradation of the messages. The base-pairing near the cleavage site seems especially important, because as few as 11 consecutive base-pairings between

messages and the miRNA center region can sometimes lead to repression of the messages with or without detectable endonucleolytic cleavage (Shin et al., 2010). These non-seed-based regulations, however, are far less common than the seed-based regulations, as protein-coding genes rarely have extensive complementarity to any miRNAs.

Endo-siRNAs (endogenous small interfering RNAs) are a class of endogenous metazoan RNAs of similar size (~22-nt) as miRNAs (Babiarz et al., 2008; Tam et al., 2008; Watanabe et al., 2008). These small RNAs differ from miRNAs in that they are usually generated from long double-stranded RNA precursors usually formed from anti-sense pseudo-gene and sense coding-gene transcripts, in only very limited types of cells like ES cells or oocytes. Examination of the consequences of inhibiting the production of these RNAs suggested that they mediate important functions in gene regulation and chromosomal segregation (Murchison et al., 2007; Kaneda et al., 2009; Ma et al., 2010; Suh et al., 2010), but the precise mechanisms of action and function are yet to be elucidated.

RNA interference (RNAi) is a phenomenon in which double-stranded RNAs introduced into cells, after being processed into ~22-nt small interfering RNAs (siRNAs) by Dicer, result in silencing of the gene that has extensive homology to the introduced RNA (Fire et al., 1998; Elbashir et al., 2001; Zamore and Haley, 2005). Although this

phenomenon can now be simply explained as exogenous double-stranded RNAs being fed into the endogenous small RNA pathway for the cleavage-dependent repression, the discovery of RNAi contributed to the understanding of the endogenous small RNA pathway: some of the key molecular components of small RNA biogenesis and function, such as the Dicer and Argonautes, were first characterized by experiments trying to elucidate the mechanism of RNAi (Hammond et al., 2000; Hammond et al., 2001; Hannon, 2002). Moreover, RNAi has enabled loss-of-function studies on individual genes in mammalian systems without recourse to time-taking, laborious genetic methods, and thus, profoundly contributed to our understanding of gene function. RNAi as a technique, however, is not without a disadvantage. siRNAs, which are designed to be extensively complementary to the gene of interest, not only act through the intended cleavage-dependent mechanism, but also through the seed-based mechanism, accidentally repressing numerous "off-target" genes that happen to harbor 7-nt matches to the seed regions of the siRNAs (Birmingham et al., 2006).

Structural analysis of the key protein component of RISC, called Argonautes (AGOs), revealed their structural level mechanisms of action. Among four domains of AGOs, Mid and PAZ domains have single-stranded nucleic acid binding motifs,

respectively accommodating the 5' and 3' ends of small RNAs (Song et al., 2004). The third domain, PIWI, contains RNase H motif, which is responsible for the endonucleolytic cleavage of the highly complementary target mRNAs. In mammalian genomes, AGO proteins are encoded as four paralogs, which can be separated into two groups, AGO1/3/4 and AGO2, with respect to both genomic distribution and enzymatic activity. AGO1/3/4, which are linked within an ~190 kilobase locus, lack the endonucleolytic cleavage activity (Cheloufi et al., 2010); only AGO2, encoded at a distant locus away from the other three, shows detectable endonuclease activity (Liu et al., 2004; Meister et al., 2004; Song et al., 2004). The lack of the endonuclease activity in three of the four paralogs might reflect the low demand for cleavage-dependent repression in mammals, as a great majority of miRNAs seem to act via a seed-based mechanism.

Protein family analysis of the mammalian proteome showed that the four AGO-like proteins, AGO1/2/3/4, constitute only one subclade of the greater Argonaute family of proteins, which also includes the piwi-like subclade of proteins, such as hiwi, hili, hiwi2, hiwil3 in humans (Tolia and Joshua-Tor, 2007). Because the piwi-like proteins contain all four domains of AGO-like proteins, it had been speculated that piwi-like proteins also mediated RNA-guided function. In fact, fly homologs of piwi-like proteins, aubergine and

piwi, were found to be responsible for the silencing of repeat elements in the male germline (Schmidt et al., 1999; Pal-Bhadra et al., 2002). Moreover, the silencing of the repeat elements was found to give rise to small RNAs that are a little longer than the previously known small regulatory RNAs (Aravin et al., 2001 Pal-Bhadra, 2002 #68); the small RNAs were accordingly dubbed repeat-associated small interfering RNAs (rasiRNAs).

With the growing interest in the mammalian counterparts of the rasiRNAs, four research groups converged on the discovery of the small RNAs associated with mammalian piwi-like proteins (Aravin et al., 2006; Girard et al., 2006; Grivna et al., 2006; Lau et al., 2006). As is described in detail in Chapter II, our collaborators used a biochemical approach to fractionate adult rat testes extract, tracking down the fractions containing RNAs of size ~30-nt. The RNAs were found to be co-fractionated with riwi (rat piwi-like protein), as well as RecQ1 helicase, which was previously implicated in quelling, which is a form of gene silencing in *Neurospora* (Lau et al., 2006).

The new mammalian small RNAs, although similar to rasiRNAs in terms of their association with piwi-like proteins in the male germline and a strong 5' terminal nucleotide bias toward U, were different in that they were a few nucleotides longer (29–30-nt versus 24–27-nt), and more importantly in that they predominantly mapped to only one strand of

genomic loci without preferential association to repeat elements. Therefore, reflecting on the most basal commonality between the small RNAs in flies and mammals, which is the association with piwi-like proteins, the small RNAs in both flies and mammals were collectively referred to as piwi-interacting RNAs (piRNAs), thereby overriding the old name for fly piRNAs: rasiRNAs. Mammalian piRNAs were mapped to ~100 clusters in the genome. On close examination of each cluster, piRNAs seemed to be frequently derived from two single-stranded RNA precursors divergently transcribed from the center of the cluster, suggesting the presence of a bidirectional promoter at the center. Although the primary sequences of the piRNAs were found to be poorly conserved, the production of the piRNAs seemed strongly conserved between mammalian species, as the mouse genomic loci known to be orthologous to most rat piRNA clusters were also found to produce piRNAs.

Follow-up studies revealed that a subclass of the mammalian piRNAs is expressed at a very narrow window of time during sperm development, called the pre-pachytene stage. The piRNAs expressed in this stage, called pre-pachytene piRNAs, are different from regular mammalian piRNAs, and rather closer to fly piRNAs in terms of characteristics, in that (1) they preferentially map to repeat elements, and (2) they frequently map to both

strands of the repeat elements. These piRNAs were found to be responsible for the silencing of the repeat elements, a process accomplished via DNA methylation (Aravin et al., 2007). Additional details were subsequently elucidated: the pre-pachytene piRNAs are selectively sorted into two different mouse piwi-like proteins, mili and miwi2, which are two different members of the piwi-like proteins expressed specifically in the pre-pachytene stage (Aravin et al., 2008). At first, piRNAs—specifically those with strong U-bias at the 5' terminal nucleotide derived from the sense strand of the repeat elements—are loaded into mili. Because anti-sense is also transcribed, the mili-loaded piRNAs recognize the anti-sense transcripts, and direct endonucleolytic cleavage, contributing to the generation of the secondary piRNAs with signature A-bias at nucleotide position 10. The secondary piRNAs loaded into miwi2, in turn recognize the sense transcript, contributing to the generation of primary piRNAs with U-bias at the 5' terminal nucleotide. This feed-forward amplification cycle, often called the "ping-pong" cycle, was initially discovered in *Drosophila* (Brennecke et al., 2007; Gunawardane et al., 2007) and was found to be conserved to sponges (Grimson et al., 2008) and mammals (Aravin et al., 2008). Despite these recent breakthroughs on the nature of pre-pachytene piRNAs, many aspects of mammalian piRNA biology remain to be elucidated, especially those of non-pre-pachytene piRNAs.

Key Evidence of Seed-based MicroRNA-Target Recognition

Sequence Conservation Study

The initial awareness of the seed as the centerpiece of miRNA target recognition was made possible by the analysis of sequence conservation between multiple mammalian species. The genome sequences of multiple mammalian species, which were beginning to be available in early 2000s (Lander et al., 2001; Waterston et al., 2002; Gibbs et al., 2004), enabled systematic assessment of the functionality of DNA elements, through the analysis of the selective sequence conservation of the elements throughout multiple species. The availability of this comparative-genomic methodology for mammals was well-timed with the discovery of the second miRNA, let-7, in 2000 and many other miRNAs in 2001, which collectively showed that miRNAs are not just worm-specific oddities, but are actually a large class of molecules, present in broad metazoan species (Reinhart et al., 2000; Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). A study published in 2003 doing a cross-genome comparison of humans, mice, and rats, reported distinctively preferential conservation of sites with 6–8-nt motifs in 3'UTRs that are complementary to the 5' region of miRNAs (Lewis et al., 2003). Especially notable was that if at least one

mismatch or wobble-pairing is included within the seed-pairing, then the quality of selective conservation almost completely disappears.

This study was followed by many independent studies confirming this observation (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Xie et al., 2005). Among these studies was one by Xie *et al.* that comprehensively surveyed the conservation of random motifs in 3'UTRs and promoters; the study reached the conclusion that 9 out of the top-10 most conserved 8-mers found in 3'UTRs contained matches to the seed region of known miRNAs (Xie et al., 2005). Later studies incorporating an ever-growing repertoire of sequenced genomes indicated that these highly conserved seed-matches are extremely prevalent; indeed, most human genes were found to have selectively maintained at least one seed-match (Friedman et al., 2009). Although the conservation analyses clearly demonstrated the prevalence and biological significance of the seed-base regulations, a question still remained about whether this seed-pairing is the dominant modality by which miRNA recognizes targets, or not. It could have been the case that, for example, the motif discovery strategies might have missed some unknown, complicated base-pairing pattern or esoteric physicochemical properties of target sites. Moreover, some of the first identified miRNA-target relationships during genetic screens, namely the ones between *let-7* and *lin-*

41 (Reinhart et al., 2000), violated the seed-based model, which further challenged the notion that the seed-pairing is the dominant modality, or that it is only one of a few major modalities of miRNA-mediated target recognition.

Microarray Study

The answer to the question at least in part came from a study based on microarray experiments. In the study published in 2005, microarrays were used to measure any change in gene expression at mRNA level in response to miRNA transfection into cultured mammalian cells (Lim et al., 2005). This ground-breaking study yielded at least three surprising findings. First, contrary to the popular belief that the miRNAs repress gene expression mainly through translational inhibition, it was detected that a large number of genes were downregulated at the mRNA level, thus indicating a high prevalence of mRNA destabilization as a mechanism of miRNA-mediated gene repression. Second, among the large number of downregulated messages, the majority contained at least one seed-match to the transfected miRNA, suggesting that seed-pairing is likely the dominant mechanism of change, at least at the mRNA level. Third, the genes that were found downregulated not only included genes with conserved seed-matches, but also included a lot of genes with

non-conserved matches. Although the non-conserved targets were 2-3 fold less frequently downregulated than were the conserved targets, the fact that the non-conserved sites worked as well as conserved sites, roughly one out of 2–3 times, suggested that the determinants provided by the simple seed-matched site are more frequently sufficient than expected. This potential sufficiency of seed-matched sites may have had an important effect on evolutionary processes: if the seed-matched sites are sufficient to a certain degree, it could be a likely story that the 3'UTRs of genes that would have been under selective pressure to maintain a high level of expression would have been depleted with seed-matched sites to coexpressed miRNAs.

Selective Avoidance Study

The aforementioned hypothetical story was tested by the study that was reported shortly after the microarray study; the subsequent study examined natural selection against creation of seed-matches in 3'UTRs—in other words, selective avoidance of undesirable seed matches in 3'UTRs (Farh et al., 2005). For example, consider a set of genes that are highly and specifically expressed in muscle. Since these genes are likely the ones that serve important functions in muscle, they are likely to have evolved to avoid having any negative

regulatory elements that could diminish their tissue specific expression, such as the target sites to muscle-specific miRNA, miR-1. If a 7mer miR-1 site is assumed to be always sufficient to mediate repression, one would expect to find only very few miR-1 sites in the 3'UTRs of these genes. Here, the study found that miR-1 sites are observed in 3'UTRs of these genes, but 50% less frequently than are expected by chance. This observation indicated that at least 50% of the randomly emerging 7mer seed-matched sites are functional, because the depleted 50% of miR-1 sites must have been functional to be actively swept away by natural selection.

Evolution of Targeting through Point Mutations in 3'UTRs

The studies described above indicated that only 7-nt long target sites are in most cases necessary and in many cases sufficient for mediating repression; it implies that most genes are only a single nucleotide mutation away from gaining or losing repression by miRNAs expressed in the same tissue. Therefore, spontaneous point mutations that frequently accumulate in 3'UTRs can lead to numerous subtle variations in gene expression. The point mutations can then be subject to natural selections of various types, depending on the functional importance of the target sites created or disrupted by the mutations.

Mutations under Negative Selection

If a mutation has disrupted a preexisting target site that is biologically beneficial, then it will be subject to negative selection. Negative selection is alternatively called purifying or stabilizing selection, and if this occurs, then the mutant allele would decrease in frequency, and ultimately reach extinction. The evidence supporting this selection against losing target sites is overwhelming. As was discussed, when genomes of multiple species were compared, nucleotide sequences of target sites were much more highly conserved, compared to control sequences (Lewis et al., 2003; Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Xie et al., 2005; Friedman et al., 2009). Within the same species, when genotypes of individuals were compared, single nucleotide polymorphism (SNP) density was significantly lower in conserved seed-matched sites, than in control sites (Chen and Rajewsky, 2006; Saunders et al., 2007). More sophisticated population genetic analysis identified the classic signatures of negative selection, specifically (1) derived (non-ancestral) SNP alleles disrupting miRNA target sites showed lower allelic frequency than those disrupting control sites, and (2) the rate of divergence (inter-species variation) normalized by the rate of polymorphism (intra-species variation) was smaller for SNP

alleles disrupting miRNA target sites than for those disrupting control sites (Chen and Rajewsky, 2006).

Another type of mutation that would likely be negatively selected would be the ones that create new target sites that turn out to be biologically harmful. This selection against gaining target sites is often called "anti-targeting", and is widely observed by many independent studies. For instance, 3'UTRs of highly and tissue-specifically expressed genes were found to be depleted of target sites to the miRNAs coexpressed in the specific tissue (Farh et al., 2005; Stark et al., 2005). Population genetic analysis also showed that the SNP alleles creating target sites to coexpressed miRNAs tend to have low frequency, which is indicative of anti-targeting (Chen and Rajewsky, 2006).

Mutations under Positive Selection

The negative selection against loss or gain of target sites, described above, is the process where the mutations that have negative influence on the fitness of the carriers are penalized by natural selection. On the other hand, natural selection could promote mutations that have positive influence on the fitness; such mutations include the ones that either create beneficial target sites or disrupt harmful target sites.

This positive selection or alternatively called adaptive selections could lead to increase in frequency, and eventually fixation, of the favorable mutant allele. Positive selection is far less common mechanism of action for natural selection than negative selection, conceivably because it is likely to be more difficult for random mutations to improve biological systems that are already evolutionarily optimized than to disrupt them. Because positive selection is such a low-frequency event, detection of any evidence of the selection requires a large number of functional variants, upon which natural selection can act. High frequency of point mutations in 3'UTRs and low informational complexity of miRNA targeting specificity qualify the mutations that create/disrupt target sites into the category of exceptionally common functional variants, which are likely to have left detectable signs of positive selection. In fact, studies have found that some SNP alleles altering miRNA target sites in the human genome show signs of recent positive selection, such as long-range linkage disequilibrium or a high degree of allelic frequency discordance between subpopulations (Chen and Rajewsky, 2006; Saunders et al., 2007).

Neutral Mutations

A mutation might be under neither negative nor positive selection if the created/disrupted target site has no biological effect that strongly influences the fitness of the carriers. One obvious reason why the mutation has no biological effect could be that the created/disrupted target site is incapable of mediating repression, maybe because it is in a disadvantageous context. However, the reason could also be that the altered target site is functional, but its biological effect is not dramatic enough to change the fitness. Such mutations that alter functional but biologically non-critical sites may have important roles especially in supplying a population with phenotypic diversity, which could later become valuable substrates for natural selection, in the case of, for example, an environmental change.

Single-Nucleotide Polymorphisms for Studying 3'UTR Mutations

In Chapter III, I describe a study that examined SNPs for studying the impact of 3'UTR mutations on gene expression diversity (Kim and Bartel, 2009). Because SNPs represent point mutations that occurred in the recent past yet have not reached fixation or extinction, it could be that by examining SNPs, scientists could obtain estimates on the extent of the impact of the mutations on the diversity in a given population at the present time. In

addition, analyzing correlations between SNPs and specific phenotypic features or disease states may enable the identification of the genetic roots of these phenotypes or diseases.

To start estimating the scope of the diversity contributed by the 3'UTR mutations, it is important to acknowledge that not all types of mutations described above are equally represented as SNPs: Because the mutations that create/disrupt functional (competent for repression) target sites are more likely to be biologically consequential, they are more likely to be actively pushed into either fixation or extinction by natural selection. Therefore, the mutations that affect functional sites are underrepresented as SNPs. In other words, the set of polymorphic target sites are likely enriched with non-functional sites.

In order to estimate which proportion of the polymorphic target sites are capable of mediating repression, we developed a high-throughput-sequencing-based method called allelic imbalance sequencing. The key idea of the method is to use cells that are heterozygous for the SNPs that alter miRNA target sites so that both target and non-target alleles are present in the same cell. In such cases, if the cognate miRNA for the target site is expressed in the cell, then the miRNA will destabilize the mRNAs from the target allele, but not the mRNAs from the non-target allele, thereby contributing to the allelic imbalance in the level of mRNAs. To measure the allelic imbalance, we first performed high-

throughput sequencing analysis on the RT-PCR products targeted toward 67 selected heterozygous SNPs and then compared the number of reads that originated from each allele. The measured allelic imbalance enabled us to infer the degree of repression mediated by the polymorphic target sites.

This study indicated that the 67 target sites mediated the repression of mRNA by 12%, on average. It also indicated that, by a conservative estimate, at least 18% of the target sites created/disrupted by SNPs were capable of mediating repression. Extrapolating these numbers to all the 3'UTR SNPs that have been identified in human and mouse revealed that between two unrelated mammalian individuals of the same species, more than 100 genes are differentially regulated due to the 3'UTR SNPs affecting the miRNA target sites. This result allowed us to conclude that the 3'UTR mutations, and the resulting gain/loss of miRNA-mediated repression, have substantially contributed to the diversity of gene expression in mammals.

Some of the diversity in gene expression is likely to translate into phenotypic diversity. Efforts to associate variations in phenotypic features or disease susceptibilities have frequently implicated the SNPs that alter miRNA target sites (Sethupathy and Collins, 2008; Ryan et al., 2010). In a study pursuing genetic determinants of muscularity of Taxel

strain of sheep, a SNP that creates a miR-1 site in the 3'UTR of myostatin in Taxel strain was found to explain the superior muscularity compared to other strains of sheep (Clop et al., 2006). Moreover, when FGF20 (a gene whose overexpression is causally linked to Parkinson's disease) was subjected to a family-based association study, the strongest association was noted to be from a 3'UTR SNP; the risk allele of the SNP was subsequently shown to cause overexpression of FGF20 through the disruption of a target site for miR-433, a miRNA highly expressed in the brain (Wang et al., 2008). In addition, many studies have revealed that the susceptibilities of many types of cancers are associated with the polymorphisms in miRNA target sites (Ryan et al., 2010). For example, independent studies found that increased risks of cancer — of the lung, mouth, and ovaries — are found associated with an allele of the SNP in a let-7 target site in KRAS; the allele was reported to disrupt the site and thereby cause overexpression of the oncogene (Chin et al., 2008; Christensen et al., 2009; Ratner et al., 2010).

In summary, research has elucidated the impact of the 3'UTR mutations on the evolution of miRNA targeting. The low informational complexity of the miRNA target specificity allows 3'UTR point mutations to frequently alter miRNA-mediated repression for individual genes. The resulting diversity in gene expression and phenotypes provides

valuable substrate for natural selection to act on, potentially facilitating phenotypic inventions and adaptation to changing environments.

Evolution of Targeting through Emergence of Novel MicroRNAs

As could genetic changes in 3'UTRs, spontaneous genetic changes to miRNA genes could also contribute to evolution. Among various types of genetic changes that could happen for miRNA genes, the changes that give rise to a miRNA with a novel seed could in principle have widespread impact on the transcriptome through repression of numerous seed-matched targets at once.

Mechanism of MicroRNA Emergence

Comparative analysis of miRNAs in multiple closely related species, especially in flies, showed that many miRNAs are lineage specific; this indicates that miRNAs have been emerging relatively frequently even in the recent past (Lu et al., 2008; Berezikov et al., 2010; Nozawa et al., 2010). Close inspection of the miRNAs suggested that miRNAs have emerged from at least three different origins: (1) from duplication of pre-existing miRNAs, (2) from transposable elements, or (3) from random hairpin structures.

First, miRNAs could emerge from duplication of pre-existing miRNA genes. Paralogous comparison of miRNA genes has indicated that many are apparently homologous to each other, suggesting their shared origins (Ruby et al., 2007b; Nozawa et al., 2010). Especially so are the miRNA genes in the same cluster (Tanzer and Stadler, 2004). For example, analysis of the miR-17 clusters in various metazoan organisms revealed that the clusters have been expanding through a series of local and non-local duplications at various time points of the metazoan evolution. At the early vertebrate lineage, the cluster was composed of three non-homologous miRNAs, miR-17, miR-19, and miR-92, linked within a single locus. Nowadays, in mammalian genomes, the three ancient miRNAs manifest as 14 miRNAs encoded at three different loci; all 14 miRNAs bear significant homology to one of the three founding miRNAs (Tanzer and Stadler, 2004). Duplication, especially tandem duplication, of a miRNA produces an additional copy of miRNAs with the same tissue-specific expression pattern as the original. The redundant copy may stay within the same miRNA seed family, consequently solidifying tissue-specific control of gene expression through the augmented overall expression of the miRNA family in the specific set of tissues. Alternatively, the redundant copy may diverge to form another seed family, consequently providing additional complexity of tissue-

specific control of gene expression. The most intuitive mechanism of such divergence of a miRNA into another family is seed mutation. For example, either one of the miR-17 (with seed sequence, AAAGUGC) or miR-18 (with seed sequence, AAGGUGC) in the miR-17 cluster appears to have emerged at the early vertebrate lineage through the tandem duplication of the other miRNA, followed by a single nucleotide transition within the seed region (Tanzer and Stadler, 2004). Other mechanisms of the miRNA seed family divergence may include mutations in important structural determinants of hairpins that change the cutting site recognized by Dicer or Drosha enzymes, often leading to the shift in seed register in the hairpins. In addition, mutations in hairpins influencing preferences on which strand gets loaded into RISC may also result in the divergence. A study on recently emerged *Drosophila* miRNAs estimated that at least 10% of the miRNA genes have emerged through duplication of pre-existing miRNAs (Nozawa et al., 2010). The actual number might be bigger than 10% considering possible false negatives that might have dropped out just because a search of homology against one another failed to identify the miRNAs of the common origin.

Second, miRNAs could also emerge from transposable elements (Smalheiser and Torvik, 2005; Borchert et al., 2006; Piriyapongsa and Jordan, 2007). For example, the

hairpin of miR-28, a miRNA found in multiple mammals, appears to be formed by two LINE-2 elements juxtaposed in opposite orientation, with each element contributing to each arm of the hairpin (Smalheiser and Torvik, 2005). At least, three more human miRNAs (miR-95, miR-151, miR-325) are in such configuration with LINE-2 elements. All four of these LINE-2-associated miRNAs are found in introns, suggesting that random integration of transposons into introns is a common mechanism for gaining transcriptional competency. A recent study estimated that 10% of all miRBase-registered human miRNAs have significant homology to transposable elements (Piriyapongsa et al., 2007; Kozomara and Griffiths-Jones, 2011); this observation, however, needs to be interpreted with a caution, because many of the miRBase-registered miRNAs might not be bona-fide miRNAs (Chiang et al., 2010). Most of the 10% are poorly conserved miRNAs, many are species-specific, and none is conserved beyond mammals. The limited conservation could be due to the possibility that the transposon-mediated miRNA emergence happened only after mammalian speciation, as no transposon-associated miRNAs have been found in flies or worms (Smalheiser and Torvik, 2005). Alternatively, perhaps all transposon-derived miRNAs that emerged before mammalian speciation shed their characteristic features of transposons, such that the transposons are no longer visible by the RepeatMasker algorithm.

Third, a substantial proportion of the miRNAs whose origins cannot be tracked down to other miRNAs or transposons are likely to have emerged, *de novo*, from random transcripts capable of forming hairpin structure. Because the informational complexity of the hairpin structure recognizable by the miRNA processing machinery is relatively small, a series of spontaneous genetic changes in genomic regions that are already competent for transcription, such as introns or non-coding genes, might have given birth to hairpins suitable for recognition by the machinery. Another possibility is the opposite scenario, where already eligible hairpins in a non-transcribed genomic region happen to acquire transcription competency through spontaneous genetic changes.

Natural Selection on Newly Emerged MicroRNAs

Numerous miRNAs found at the present time are the ones that have survived the iterative and intensive test of negative selection. Considering that the emergence of a new miRNA with novel seeds is supposed to perturb the transcriptome away from the state that has been evolutionarily optimized for such a long time, it is intriguing to think about how any such change could have possibly survived negative selection.

One possible explanation could be the following: Although most of the novel miRNAs would have proven themselves to be deleterious, it is a reasonable possibility that in some rare cases, novel miRNAs might have had advantageous effects significant enough to compensate for disadvantageous effects, such that the net effect would be positive. One speculation on how random transcriptomic perturbation could possibly have substantially advantageous effects is based on the following assumption: Because the gene regulatory network is a convoluted, dynamic system with large orders of complexity, it may be the case that not all of the theoretically possible transcriptomic states are stable; many might be unstable. The stable states, which may correspond to distinct phenotypic states, might have evolved in a way that each stable state confers robustness against spurious state transitions by minor fluctuations in gene expression. It is reasonable to conjecture that when a transcriptome is pushed into an unstable/transient state, it will tend to converge towards a stable state that is the most topologically proximal for a given vector field. Under this assumption, the perturbation of transcriptome by a novel miRNA might allow the transcriptome to make a quantum transition from one stable state ultimately into a different stable state, which might happen to correspond to a more favorable phenotypic state in a given environmental context. The assumption is consistent with the previous observations

that the transitions between defined cellular states—for example, from fibroblasts to iPSCs—can be achieved through multiple different recipes, and therefore through multiple different trajectories in the state space (Frazer et al., 2007; Takahashi et al., 2007). Once the miRNA has survived negative selection and is under weak "net" positive selection, the ensuing 3'UTR evolution that prunes harmful targeting of the miRNA and adds beneficial targeting could allow full integration of the miRNA into the transcriptome.

Another possibility for explaining how some miRNAs could survive negative selection is the possibility that the survived miRNAs were lowly expressed at the critical period at and immediately after the time of birth. This possibility is supported by the observation that the miRNAs that emerged in the recent past are usually found lowly expressed, as the younger miRNAs that are found only in mammals are expressed 44-fold less strongly than the older miRNAs found beyond mammals (Chiang et al., 2010). These lowly expressed miRNAs are known to be less competent for target repression (Bartel, 2009; Chiang et al., 2010); they might repress only a small set of the seed-matched targets with the highest quality contexts. Since these miRNAs would have a less dramatic effect on the transcriptome, they are less likely to be subject to immediate negative selection and more likely to be neutral. And, a small minority of the lowly expressed miRNAs might

even be under slight positive selection. Whereas the neutral miRNAs will slowly disappear over time, the miRNAs with positive impact will be under selective pressure not only to preserve the miRNAs but also to gradually enhance their expression through the optimization of the processes of transcription, processing, and loading. But, the enhancement of miRNA expression, and the resulting expansion of the target repertoire, may elicit negative selection unless it is accompanied by coordinated natural selection on 3'UTRs, acting to remove/weaken disadvantageous target sites and add/strengthen advantageous sites. The coevolution of the miRNAs and 3'UTRs will augment the benefit of the miRNAs, further increasing the selective pressure to maintain and optimize the miRNAs and their targets.

Simulating MicroRNA Emergence in Cultured Cells

In Chapter IV, I describe an approach that can be viewed as simulating miRNA emergence in the mammalian cell culture system. The approach is to overexpress the miRNA library that contains miRNAs with all possible seed sequences, in order to find miRNAs that can be used to induce interesting and/or useful phenotypes. As might have been the case for the emergence of miRNAs in the evolutionary history, it may be that the negative selection will

prove to be a major obstacle against the success of the approach. However, the negative selection might be less of an obstacle in cultured cells than is in multicellular organisms. For example, a miRNA that can dedifferentiate terminally differentiated cells into the embryonic state would be disadvantageous if it spontaneously emerged in random cell types of a multicellular organism, but the same miRNA would not be disadvantageous to cells cultured *in vitro*; rather it could serve as a useful reagent in carefully designed and executed experiments, with perhaps economically and/or biologically fruitful results. Moreover, even though the random perturbation of the transcriptome by novel miRNAs would have certain disadvantageous effects even in cultured cells, the disadvantageous effects may not necessarily prevent the cells from manifesting a phenotype of interest in short-term *in vitro* assays designed to score a specific phenotype. Even though there is a risk that the miRNA found to induce a useful phenotype also induces certain negative ancillary effects, it is possible that transient, dose-controlled expression of the miRNAs might be useful in certain therapeutic or experimental settings for controlling phenotypes. In fact, in many cases of studies that utilize siRNAs with the purpose of silencing the specific target genes, the unintentional effects coming from the off-target effects do not impair the cells from manifesting phenotypes, as siRNAs have been successfully used in

numerous functional studies. Rather than off-target effects having worrisome or prohibitively risky associated impacts, it has been often the case that the off-target effects have yielded interesting phenotypes (Alvarez et al., 2006; Fedorov et al., 2006; Ali et al., 2009).

REFERENCES

Ali, N., Karlsson, C., Aspling, M., Hu, G., Hacohen, N., Scadden, D.T., and Larsson, J. (2009). Forward RNAi screens in primary human hematopoietic stem/progenitor cells. Blood *113*, 3690-3695.

Alvarez, V.A., Ridenour, D.A., and Sabatini, B.L. (2006). Retraction of synapses and dendritic spines induced by off-target effects of RNA interference. J Neurosci *26*, 7820-7825.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., *et al.* (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203-207.

Aravin, A.A., Naumova, N.M., Tulin, A.V., Vagin, V.V., Rozovsky, Y.M., and Gvozdev, V.A. (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. Curr Biol *11*, 1017-1027.

Aravin, A.A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K.F., Bestor, T., and Hannon, G.J. (2008). A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell *31*, 785-799.

Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G.J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science *316*, 744-747.

Babiarz, J.E., Ruby, J.G., Wang, Y., Bartel, D.P., and Blelloch, R. (2008). Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes Dev *22*, 2773-2785.

Baek, D., Villen, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64-71.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215-233.

Berezikov, E., Liu, N., Flynt, A.S., Hodges, E., Rooks, M., Hannon, G.J., and Lai, E.C. (2010). Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. Nat Genet *42*, 6-9; author reply 9-10.

Birmingham, A., Anderson, E.M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., Karpilow, J., *et al.* (2006). 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. Nat Methods *3*, 199-204.

Borchert, G.M., Lanier, W., and Davidson, B.L. (2006). RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol *13*, 1097-1101.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell *128*, 1089-1103.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of microRNA-target recognition. PLoS Biol *3*, e85.

Cheloufi, S., Dos Santos, C.O., Chong, M.M., and Hannon, G.J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. Nature *465*, 584-589.

Chen, K., and Rajewsky, N. (2006). Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet *38*, 1452-1456.

Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., Auyeung, V.C., Spies, N., Baek, D., Johnston, W.K., Russ, C., Luo, S., Babiarz, J.E., *et al.* (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes Dev *24*, 992-1009.

Chin, L.J., Ratner, E., Leng, S., Zhai, R., Nallur, S., Babar, I., Muller, R.U., Straka, E., Su, L., Burki, E.A., *et al.* (2008). A SNP in a let-7 microRNA complementary site in the

KRAS 3' untranslated region increases non-small cell lung cancer risk. Cancer Res *68*, 8535-8540.

Christensen, B.C., Moyer, B.J., Avissar, M., Ouellet, L.G., Plaza, S.L., McClean, M.D., Marsit, C.J., and Kelsey, K.T. (2009). A let-7 microRNA-binding site polymorphism in the KRAS 3' UTR is associated with reduced survival in oral cancers. Carcinogenesis *30*, 1003-1007.

Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibe, B., Bouix, J., Caiment, F., Elsen, J.M., Eychenne, F., *et al.* (2006). A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat Genet *38*, 813-818.

Davis, E., Caiment, F., Tordoir, X., Cavaille, J., Ferguson-Smith, A., Cockett, N., Georges, M., and Charlier, C. (2005). RNAi-mediated allelic trans-interaction at the imprinted Rtl1/Peg11 locus. Curr Biol *15*, 743-749.

Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494-498.

Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science *310*, 1817-1821.

Fedorov, Y., Anderson, E.M., Birmingham, A., Reynolds, A., Karpilow, J., Robinson, K., Leake, D., Marshall, W.S., and Khvorova, A. (2006). Off-target effects by siRNA can induce toxic phenotype. RNA *12*, 1188-1196.

Filipowicz, W., Bhattacharyya, S.N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet *9*, 102-114.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., *et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

Friedman, R.C., Farh, K.K., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res *19*, 92-105.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., *et al.* (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493-521.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199-202.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell *27*, 91-105.

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B.J., Chiang, H.R., King, N., Degnan, B.M., Rokhsar, D.S., and Bartel, D.P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature *455*, 1193-1197.

Grivna, S.T., Beyret, E., Wang, Z., and Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. Genes Dev *20*, 1709-1714.

Gunawardane, L.S., Saito, K., Nishida, K.M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., and Siomi, M.C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science *315*, 1587-1590.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835-840.

Hammond, S.M., Bernstein, E., Beach, D., and Hannon, G.J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature *404*, 293-296.

Hammond, S.M., Boettcher, S., Caudy, A.A., Kobayashi, R., and Hannon, G.J. (2001). Argonaute2, a link between genetic and biochemical analyses of RNAi. Science *293*, 1146-1150.

Hannon, G.J. (2002). RNA interference. Nature *418*, 244-251.

Hendrickson, D.G., Hogan, D.J., McCullough, H.L., Myers, J.W., Herschlag, D., Ferrell, J.E., and Brown, P.O. (2009). Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. PLoS Biol *7*, e1000238.

Kaneda, M., Tang, F., O'Carroll, D., Lao, K., and Surani, M.A. (2009). Essential role for Argonaute2 protein in mouse oogenesis. Epigenetics Chromatin *2*, 9.

Kim, J., and Bartel, D.P. (2009). Allelic imbalance sequencing reveals that single-nucleotide polymorphisms frequently alter microRNA-directed repression. Nat Biotechnol *27*, 472-477.

Kim, V.N., Han, J., and Siomi, M.C. (2009). Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol *10*, 126-139.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res *39*, D152-157.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., *et al.* (2005). Combinatorial microRNA target predictions. Nat Genet *37*, 495-500.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science *294*, 853-858.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science *294*, 858-862.

Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science *313*, 363-367.

Lee, R.C., and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science *294*, 862-864.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell *75*, 843-854.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15-20.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of mammalian microRNA targets. Cell *115*, 787-798.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769-773.

Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science *305*, 1437-1441.

Lu, J., Shen, Y., Wu, Q., Kumar, S., He, B., Shi, S., Carthew, R.W., Wang, S.M., and Wu, C.I. (2008). The birth and death of microRNA genes in Drosophila. Nat Genet *40*, 351-355.

Ma, J., Flemr, M., Stein, P., Berninger, P., Malik, R., Zavolan, M., Svoboda, P., and Schultz, R.M. (2010). MicroRNA activity is suppressed in mouse oocytes. Curr Biol *20*, 265-270.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol Cell *15*, 185-197.

Murchison, E.P., Stein, P., Xuan, Z., Pan, H., Zhang, M.Q., Schultz, R.M., and Hannon, G.J. (2007). Critical roles for Dicer in the female germline. Genes Dev *21*, 682-693.

Nozawa, M., Miura, S., and Nei, M. (2010). Origins and evolution of microRNA genes in Drosophila species. Genome Biol Evol *2*, 180-189.

Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M., and Lai, E.C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell *130*, 89-100.

Pal-Bhadra, M., Bhadra, U., and Birchler, J.A. (2002). RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in Drosophila. Mol Cell *9*, 315-327.

Piriyapongsa, J., and Jordan, I.K. (2007). A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS One *2*, e203.

Piriyapongsa, J., Marino-Ramirez, L., and Jordan, I.K. (2007). Origin and evolution of human microRNAs from transposable elements. Genetics *176*, 1323-1337.

Ratner, E., Lu, L., Boeke, M., Barnett, R., Nallur, S., Chin, L.J., Pelletier, C., Blitzblau, R., Tassi, R., Paranjape, T., *et al.* (2010). A KRAS-variant in ovarian cancer acts as a genetic marker of cancer risk. Cancer Res *70*, 6509-6515.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature *403*, 901-906.

Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007a). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83-86.

Ruby, J.G., Stark, A., Johnston, W.K., Kellis, M., Bartel, D.P., and Lai, E.C. (2007b). Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome Res *17*, 1850-1864.

Ryan, B.M., Robles, A.I., and Harris, C.C. (2010). Genetic variation in microRNA networks: the implications for cancer research. Nat Rev Cancer *10*, 389-402.

Saunders, M.A., Liang, H., and Li, W.H. (2007). Human polymorphism at microRNAs and microRNA target sites. Proc Natl Acad Sci U S A *104*, 3300-3305.

Schmidt, A., Palumbo, G., Bozzetti, M.P., Tritto, P., Pimpinelli, S., and Schafer, U. (1999). Genetic and molecular characterization of sting, a gene involved in crystal formation and meiotic drive in the male germ line of Drosophila melanogaster. Genetics *151*, 749-760.

Sethupathy, P., and Collins, F.S. (2008). MicroRNA target site polymorphisms and human disease. Trends Genet *24*, 489-497.

Shin, C., Nam, J.W., Farh, K.K., Chiang, H.R., Shkumatava, A., and Bartel, D.P. (2010). Expanding the microRNA targeting code: functional sites with centered pairing. Mol Cell *38*, 789-802.

Smalheiser, N.R., and Torvik, V.I. (2005). Mammalian microRNAs derived from genomic repeats. Trends Genet *21*, 322-326.

Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. Science *305*, 1434-1437.

Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell *123*, 1133-1146.

Suh, N., Baehner, L., Moltzahn, F., Melton, C., Shenoy, A., Chen, J., and Blelloch, R. (2010). MicroRNA function is globally suppressed in mouse oocytes and early embryos. Curr Biol *20*, 271-277.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861-872.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M., *et al.* (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature *453*, 534-538.

Tanzer, A., and Stadler, P.F. (2004). Molecular evolution of a microRNA cluster. J Mol Biol *339*, 327-335.

Tolia, N.H., and Joshua-Tor, L. (2007). Slicer and the argonautes. Nat Chem Biol *3*, 36-43.

Wang, G., van der Walt, J.M., Mayhew, G., Li, Y.J., Zuchner, S., Scott, W.K., Martin, E.R., and Vance, J.M. (2008). Variation in the miRNA-433 binding site of FGF20 confers risk for Parkinson disease by overexpression of alpha-synuclein. Am J Hum Genet *82*, 283-289.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., *et al.* (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature *453*, 539-543.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.

Yekta, S., Shih, I.H., and Bartel, D.P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. Science *304*, 594-596.

Zamore, P.D., and Haley, B. (2005). Ribo-gnome: the big world of small RNAs. Science *309*, 1519-1524.

CHAPTER II

Characterization of the piRNA Complex from Rat Testes

The work presented in this chapter was collaboration between Nelson Lau, Anita Seto, and myself. Specifically, I did all the computational analysis of high-throughput sequencing data.

This work has been published previously as:

Nelson C. Lau, Anita G. Seto, Jinkuk Kim, Satomi Kuramochi-Miyagawa, Toru Nakano, David P. Bartel, Robert E. Kingston, Characterization of the piRNA Complex from Rat Testes, *Science* 2006, 313(5785):363-367

ABSTRACT: Small noncoding RNAs regulate processes essential for cell growth and development, including mRNA degradation, translational repression, and transcriptional gene silencing (TGS). During a search for candidate mammalian factors for TGS, we purified a complex that contains small RNAs and Riwi, the rat homolog to human Piwi. The RNAs, frequently 29 to 30 nucleotides in length, are called Piwi-interacting RNAs (piRNAs), 94% of which map to 100 defined ($\leq$101 kb) genomic regions. Within these regions, the piRNAs generally distribute across only one genomic strand or distribute on two strands but in a divergent, nonoverlapping manner. Preparations of piRNA complex (piRC) contain rRecQ1, which is homologous to qde-3 from Neurospora, a gene implicated in silencing pathways. Piwi has been genetically linked to TGS in flies, and slicer activity cofractionates with the purified complex. These results are consistent with a gene-silencing role for piRC in mammals.

Gene-silencing pathways guided by small RNAs, essential for maintaining proper cell growth and differentiation, operate at either the transcriptional or posttranscriptional level (Zamore and Haley, 2005). Posttranscriptional gene silencing acts through mRNA destabilization or inhibition of mRNA translation (Zamore and Haley, 2005), whereas TGS represses gene expression by altering chromatin conformation (Matzke and Birchler, 2005). Each pathway uses a core complex containing small RNA associated with a member of the Argonaute (Ago) protein family; however, the different mechanistic needs of each pathway require differences in complex composition. Although RNA-mediated TGS has been studied in fission yeast and other eukaryotes (Hall et al., 2002; Volpe et al., 2002; Verdel et al., 2004; Matzke and Birchler, 2005), the mechanism of this process in mammals remains elusive.

To identify candidate complexes for TGS in mammals, we exploited the previous observations that TGS might use small RNAs longer than the 21- to 23-nucleotide (nt) microRNAs (miRNAs). In Arabidopsis, Tetrahymena, Drosophila, and zebrafish, RNAs that are 24 nt and longer have been associated with TGS and/or genomic repeats, which are often silenced (Hamilton et al., 2002; Mochizuki et al., 2002; Aravin et al., 2003; Zilberman et al., 2003; Chen et al., 2005; Lee and Collins, 2006). In Drosophila, these

repeat-associated small interfering RNAs (rasiRNAs) are enriched in the testis (Aravin et al., 2003; Aravin et al., 2004). Therefore, we prepared extract from rat testes and fractionated it on an ion-exchange Q column, monitoring the small RNAs. A peak of small RNAs longer than a 22-nt marker eluted in mild salt conditions, which suggested the presence of a novel ribonucleoprotein complex (Fig. 1A).

To characterize the small RNAs, we sequenced cDNA libraries made from flowthrough and eluate fractions, obtaining 61,581 reads from the eluate that matched perfectly to the *Rattus norvegicus* genome (Gibbs et al., 2004). In contrast to the flowthrough RNAs, which were mostly miRNAs (69%), the eluate RNAs derived primarily from regions of the genome not previously thought to be expressed (Fig. 1A). Some eluate reads matched expressed sequence tags (EST) (11%), but only a small fraction matched annotated mRNAs (< 1.1%). Some also matched repeats (20%), but when considering that ~40% of the genome is annotated as repeats (Gibbs et al., 2004), the eluate reads were depleted in repeat sequences and thus, as a class, did not represent rasiRNAs.

The eluate RNAs were mostly 25 to 31 nt in length (Fig. 1A), and Northern blot analysis indicated a testis-specific expression pattern (Fig. 1B). Most eluate RNAs began with a 5′ uridine (~84%), but no other sequence features or motifs were detected. A

dominant subpopulation at 29 to 30 nt was observed (Fig. 1C); however, these 29- to 30-nt oligomers could not be distinguished from most of the remaining eluate reads by other criteria, including 5′ nucleotide, genomic locus, and annotation. Thus, all the eluate RNAs that did not match annotated noncoding RNAs (miRNA, tRNA, rRNA, and snRNA) were considered together as representing a single newly identified class of small RNAs.

To understand potential functions for these RNAs, we purified the associated proteins. By monitoring the RNAs, we developed a five-step scheme to purify the native complex to near homogeneity (Fig. 2, A and B). Mass spectrometry of the purified complex identified the rat homologs to Piwi (Riwi) and the human RecQ1 protein (Fig. 2). Western blotting confirmed the copurification of Riwi and rRecQ1 with the small RNAs (Fig. 2C, but see also independent purification described below). We designate RNAs found with rat Piwi to be Piwi-interacting RNAs (piRNAs) and the complex to be the piRNA complex (piRC).

To gain insight into the origins of piRNAs, we examined the genomic loci from which they presumably derived. About two-thirds of the piRNA sequences each perfectly matched a single locus, and in some cases that specific locus was matched by multiple reads (up to 149). For the remaining one-third of the reads, which each mapped to multiple

loci (up to 25,044 loci), we normalized the number of reads by the number of genomic hits and assigned this normalized hit count equally to all the loci; thus, a piRNA read with four perfect genomic hits contributed a quarter of a count to each of its four loci. Counts were integrated into bins and plotted across each chromosome. The majority of counts (94%) fell into 100 genomic clusters that each contained at least 20 uniquely mapping reads. As exemplified by four clusters on chromosome 20 (Fig. 3A), the clusters distributed across the genome; however, some chromosomes were underrepresented in piRNA hits and clusters (data not shown). These clusters spanned 1 to 101 kb and in aggregate made up less than 0.1% of the rat genome.

Known silencing RNAs (siRNAs and miRNAs) derive from double-stranded RNA precursors or foldback structures(Zamore and Haley, 2005). In contrast, piRNAs of most clusters mapped exclusively to either the plus or minus genomic strand in irregular, sometimes overlapping, patterns, with no evidence of extensive foldback structures or double-stranded origins (Fig. 3B). Sixteen clusters, such as cluster 1 (Fig. 3, B and C), contained regions of minus- and plus-strand hits that were juxtaposed with each other but separated by a gap of ~100 to 800 base pairs (bps), an orientation that suggests divergent, bidirectional transcription, starting within the gap that separated the two distributions. Only

two clusters had hits that suggested convergent or overlapping transcription (clusters 31 and 38). Northern blot analysis confirmed that piRNAs derived predominantly from one of the two genomic strands (Fig. 3D). Reverse transcription polymerase chain reaction results suggested that longer transcripts of the same polarity, perhaps piRNA precursors, also derived from these regions.

Analogous production of piRNAs from at least 94 clusters occurred in the mouse, as indicated by the analysis of 68,794 piRNA reads generated in the same manner as those of the rat (Fig. 3B). Most of the mouse clusters were homologous to rat clusters, with strikingly similar strand specificity and abundance profiles (Fig. 3B). Nonetheless, their sequence conservation was low. Probes against rat clusters 4 and 6 hybridized only weakly to mouse piRNAs, as expected by the numerous point substitutions in the orthologous mouse piRNAs (Fig. 3D). Overall, the single-nucleotide substitution rate of the piRNA clusters was within the 15 to 20% expected for neutral residues (Gibbs et al., 2004). Nevertheless, residues represented by more reads had lower substitution and insertion/deletion rates, indicating detectable evolutionary pressure to conserve the sequence of the abundant piRNAs (Fig. 3E). We conclude that the production of piRNAs is highly conserved, but the sequence identities of the piRNAs are only weakly conserved.

The weak conservation favors models in which piRNAs target the loci/transcripts that correspond to the same loci from which they derive.

We characterized two potential biochemical functions of piRC suggested by activities previously attributed to RecQ and Ago family members. Human RecQ1 is an adenosine triphosphate (ATP)–dependent DNA helicase (Cui et al., 2003). Both adenosine triphosphatase (ATPase) and DNA unwinding activities followed the rRecQ1 protein of piRC (Fig. 4, A and B). Riwi contains the catalytic residues that other Ago proteins use for RNA-guided cleavage of target RNAs (Rivas et al., 2005). Using a substrate complementary to a piRNA, we detected cleavage activity, peaking with fractions containing Riwi and piRNAs (Fig. 4D). However, it was not robust, perhaps because of the small representation of the cognate piRNA in the diverse population of piRNAs (<0.2%).

Our purification of piRC uncovered a novel class of small RNAs and identified as copurifying factors Riwi and rRecQ1, two proteins with intriguing functions genetically determined in other species. Piwi represents a subclade of the Ago family of proteins (Carmell et al., 2002) and was first discovered to regulate germ stem cell maintenance in *Drosophila* (Cox et al., 1998). Subsequently, mammalian Piwi members were found to regulate germ cell maturation (Deng and Lin, 2002; Kuramochi-Miyagawa et al., 2004).

Drosophila *piwi* mutants are also defective in small RNA-dependent transgene and retrotransposon silencing (Pal-Bhadra et al., 2002; Kalmykova et al., 2005) and lose the inability to localize heterochromatic proteins, including the repressive Polycomb-group proteins (Pal-Bhadra et al., 2004; Grimaud et al., 2006). *Tetrahymena* Piwi (*TIWI*) is needed for siRNA-mediated DNA elimination (Mochizuki et al., 2002).

In *Neurospora*, a screen for mutants in quelling (gene silencing during vegetative growth) identified both QDE-2, an Ago-family protein, and QDE-3, a RecQ1 homolog (Cogoni and Macino, 1999; Catalanotto et al., 2002). When compared with RecQ homology in mammals and other organisms, *Neurospora* QDE-3 resided in the same clade as rRecQ1 (Fig. 2D). rRecQ1 did not always precisely cofractionate with Riwi and the piRNAs during our final purification step (Fig. 4A). The lack of tight association of rRecQ1 might have reflected conditions specific to this step or might indicate that rRecQ1 is generally less tightly associated with piRNAs than is Riwi. Perhaps rRecQ1 is not critical for piRC function. However, the genetic links between the QDE-2 and QDE-3 silencing factors suggest that the biochemical association between Riwi and rRecQ1 has biological importance and, furthermore, implies a gene-silencing function for piRC. Addressing the

functions of piRC and the biogenesis and localization of the piRNAs will be important questions for elucidating the potential for piRC to regulate the genome.

METHODS

Preparation of rat testes extract

Frozen trimmed rat testicles from adult Sprague-Dawley rats were purchased from Pel-Freez and were quick-thawed in washes of ice-cold 1X PBS. All subsequent steps were carried out on ice or at 4°C. Thawed testicles were washed twice in Buffer A, and then minced in a Waring blender with 3 ml / 1 g tissue of Buffer A with six bursts of 6 seconds each. The homogenate was filtered through 3 layers of cheesecloth, and then centrifuged at 10,000 g for 10 minutes. The supernatant was set aside, and the pellet was resuspended in 0.5 volume / g tissue in Buffer B. The suspension was then dounced with 6 strokes of a type B pestle, and 0.5 volume / g tissue of Buffer C was added dropwise. The suspension was incubated with gentle rocking for 1 hr, dounced periodically to homogenize extract, and then centrifuged at 30,000 g for 1 hr. The supernatant was then dialyzed in Buffer D, and then centrifuged again at 16,000 g for 30 min. The final supernatant represented the rat testes extract and was flash frozen.

Sixty grams of testicles yielded ~50 ml of extract with a concentration ranging from 7-10 mg/ml of protein. Buffer D: 20 mM HEPES KOH, pH 7.9, 20% glycerol, 100 mM KCl, 0.2 mM EDTA, 1.5 mM MgCl2, 0.2 mM PMSF, and 1.0 mM DTT. Buffer A: same

as Buffer D but with 0.5 μg/ml leupeptin and 0.5 μg/ml aprotinin. Buffer B: same as Buffer A but with 20 mM KCl. Buffer C: same as Buffer A but with 1.2 M KCl.

Analysis of Small RNAs

Chromatography fractions were deproteinized by phenol/chloroform extraction, pH 6.6, and RNAs were precipitated with ethanol and glycogen carrier. RNAs were endlabeled with 32P-cordycepin triphosphate and yeast poly-A polymerase (USB) according to the manufacturer's instructions. Small RNAs for cDNA libraries were extracted from column fractions with Tri-Reagent (Sigma). Northern blots were performed and cDNA libraries were constructed essentially as previously described (Lau et al., 2001). To prepare Q column eluate cDNA libraries for high-throughput pyrosequencing, the original cDNA libraries were diluted 200 fold into a PCR that was thermocycled for 12-16 cycles and that contained the forward primer:

(GCCTCCCTCGCGCCATCAGTATCGTAGGCACCTGAGA) and the reverse primer:

(AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA/iSp18/

GCCTTGCCAGCCCGCTCAGTATTGATGGTGCCTACAG). The asymmetric PCR products were then resolved on a 90% formamide, 8% polyacrylamide gel, and the longer

DNA strands are gel purified and submitted to the 454 Life Sciences' Genome Sequencer 20 system (Margulies et al., 2005).

Protein Chromatography and Analysis

Purification of piRC

All steps were performed on ice or at 4°C. Approximately 350 mg of protein from rat testes extract was brought to 325mM KOAc by the addition of HKA1000. 25 ml DEAE FF Sepharose resin (GE Healthcare) was equilibrated in HKA100. Rat testes extract was added to resin and bound with stirring on ice under vacuum for 2-3 hours. Slurry was poured into a column (20x3 cm) and supernatant flowed through by gravity. Buffer HKA100 was added to keep column wet. 10 ml flowthrough fractions were collected until protein was no longer detectable by Bradford protein assay (Bio-Rad). To pooled DEAE flowthrough fractions, one volume of HKA0 buffer was slowly added while swirling. Flowthrough material was loaded onto HiLoad 16/10 Q Sepharose HP (GE Healthcare), equilibrated in HKA100. The bound material was eluted with a salt gradient from 0.1–1 M KOAc. Fractions were assayed for piRNAs by 3' end-labeling with polyA polymerase and 32P-cordycepin triphosphate. The peak of RNA was pooled (eluted at approximately 215-350 mM KOAc).

Pooled fractions were split in half and each was loaded onto a HiPrep 16/10 Heparin FF column (GE Healthcare). The bound material was eluted with a salt gradient from 0.1–1 M KOAc. The piRNAs eluted in a peak of approximately 235–300 mM KOAc. Peak fractions from both heparin columns were pooled and diluted with one volume of HKA0. This material was then loaded onto a Mono S 5/50 GL column (GE Healthcare), equilibrated in 100 mM KOAc. The bound material was eluted with a salt gradient from 0.1–1 M KOAc. The piRNAs eluted in 2 fractions of approximately 185–275 mM salt. The lower salt fraction was loaded onto a Superdex-200 10/300 GL column equilibrated and eluted in HKA200. 500 ul fractions were collected from Superdex-200 column immediately upon sample injection. HKA buffer: 30 mM HEPEs pH 7.5, 2 mM MgOAc, 10% glycerol, and 0–1 M KOAc.

Protein analysis

Five-20 µl of Superdex-200 column fractions were resolved on a 4-12% Bis-Tris PAGE gel run in MOPS buffer (Invitrogen). Gels were either stained with SilverQuest silver staining kit (Invitrogen) or transferred to PVDF membrane (Bio-Rad) for Western blotting. Blots were probed with anti-Miwi antibody (Kuramochi-Miyagawa et al., 2004) or anti-human

RecQ1 antibody (Bethyl BL2071) followed by anti-rabbit IgG (GE Healthcare), then visualized by ECL Plus (GE Healthcare). LC/MS/MS of bands excised from silver-stained gels was performed by the Taplin Biological Mass Spectrometry Facility (Harvard Medical School).

Activity assays

ATPase assays

γ-[32P]-ATP was incubated with 4 mM MgCl2, 4 μg plasmid DNA, and 3 μl of Superdex-200 fraction in a 5 μl reaction for 30 min. at 30 C. Reactions were quenched with 12.5 μl of stop solution (3% SDS, 100 mM EDTA, 50 mM Tris [pH 7.7]). Inorganic phosphate and ATP were separated on PEI-cellulose TLC plates using 0.5M LiCl and 1 M formic acid, and visualized on a PhosphorImager (GE Healthcare).

Helicase assays

[32P]-5' end labeled 17-mer (5'-GTTTTCCCAGTCACGAC-3') was annealed to M13mp18 ssDNA as described (Cui et al., 2003). The dsDNA substrate was purified away from the excess 17-mer by successive purifications over four MicroSpin S-400 HR columns

(GE Healthcare). Duplexed DNA substrate was incubated with 0.67 µl of Superdex-200 column fractions in a 10 µl reaction at 37°C as described (Cui et al., 2003). Reactions were stopped at indicated timepoints with 0.3% SDS, 10 mM EDTA, 5% glycerol and then resolved on a 12% nondenaturing polyacyramide gel. Gels were dried, visualized, and quantitated on a PhosphorImager (GE Healthcare).

Slicer assays

RNA cleavage substrates were transcribed from DNA oligos to form the following RNAs: 5'-GGAACCGAGCUC-[antisense sequence to piRNA]-AGCUAGCAACC-3'. Cleavage substrates were cap-labeled and utilized in slicer reactions essentially as described (Martinez et al., 2002), except that 10 µl of purified fractions were mixed 1:1 with 10 µl 2X reaction components containing 8 mM MgCl2, and incubations were performed at 35°C.

Computational analysis of piRNAs

Processing and annotation of large-scale sequencing reads

High-throughput sequencing of the eluate cDNA library from rat and mouse testes extract yielded 99,753 and 105,793 raw sequencing reads, respectively. After filtering out reads

that did not match 5' and 3' linker sequences, reads that contained an ambiguous base ('N'), reads with lengths outside of the gel-purification size range (18-32 nt), or reads matching size marker RNAs used in library construction, 85,489 rat and 95,423 mouse reads remained. Because some sequences represented more than one read, these corresponded to 61,293 unique small RNA sequences in rat, and 65,681 in mouse.

For the sequences of each species, WU-BLAST (parameters: nogaps, E=0.01, W=[length of a sequencing read], hspsepSmax=0, hspmax=60000, B=60000) was used to find matches to: (a) the mammalian (human/mouse/rat/dog) miRNA hairpin sequences registered at miRBase (Griffiths-Jones et al., 2006), (b) the cluster of rat or mouse 18S, 5.8S, and 28S rRNA sequences (accession number: V01270 for rat, J01871, X00686 and X00525 for mouse), and (c) the Rattus norvegicus genome (build rn3) or Mus musculus genome (build mm7) (Gibbs et al., 2004). As a result, 40,698 unique sequences representing 61,581 reads of rat and 43,332 sequences representing 68,794 reads of mouse were confirmed to derive from the rat and mouse genome respectively. For each unique sequence, the number of genomic hits was counted.

We annotated the unique sequences with determinable genomic coordinates according to a hierarchical manner that classified RNAs into specific groups. First,

sequences that perfectly matched to microRNA hairpin sequences were classified as 'miRNA'. Second, the remaining unique sequences that perfectly matched rRNA cluster sequences were classified as 'rRNA'. Third, the remaining unique sequences not classified as miRNA nor rRNA were classified as 'rmsk' if they mapped to at least one locus from the RepeatMasker annotation tracks that were downloaded from the UCSC genome browser.

To analyze the repeat-associated sequences (classified as 'rmsk') in greater detail, we examined the classifications of this sequence group amongst the 15 subclasses in RepeatMasker annotations (5S rRNA, tRNA, snRNA, scRNA, srpRNA, LTR, LINE, SINE, Satellite, Low_complexity, Simple_repeat, DNA, RNA, Other and Unknown), which include not only repeat elements, but also non-coding RNA species. Many of these sequences hit multiple loci, of which some loci carry a RepeatMasker annotation, while other loci may lack any annotation. For example, a sequence can hit one locus that has no annotation, and it can hit a second locus that is annotated as both a tRNA and a LTR. To minimize bias and improve consistency in classification, we would assign classifications in such an example by first evenly distributing an identity score (100%) to each of the two loci, giving a "score" of 50% to each locus. For the locus that was annotated as a tRNA and LTR, the identity score was further divided, so that the tRNA and LTR classifications each

would receive a 25% identity score. So, if this sequence was read 10 times, a score of 5 (= 10*25/50) was assigned to "tRNA", and another score of 5 was assigned to "LTR". Annotations on both sense and antisense strands were considered.

Finally, the remaining sequences that were not yet classified were examined for coordinate overlap with the coordinates of mRNAs and ESTs. The genomic coordinates of mRNAs and ESTs were downloaded from UCSC genome browser, and sequences yet containing at least one locus annotated by mRNA or EST were classified as 'mRNA' and 'EST', respectively. However, rat mRNA sequences whose accession number started with 'DQ' were disregarded because these represented a large group of annotation in which the mRNAs did not derive from the annotated loci (their annotated exons had multiple mismatches to the cDNA sequences and their annotated introns did not have canonical splicing sites).

Detecting motifs

The 40,698 unique rat sequences that did not represent fragments of rRNA or other annotated non-coding RNAs were divided into subset, based on a common length or annotation. Each subset was examined for the presence of sequence motifs. Every subset

strongly exhibited a 5' U at the first nucleotide, but no other significant enrichment of motifs was detected from any subsets. The 5-nt long sequences immediately upstream of the 5' end of the mapped loci were collected and examined for the presence of sequence motifs for piRNA processing. This search did not result in any significant enrichment of motifs regardless of the density of reads.

Calculating genomic proportions of each annotation class

The aggregate proportions of the genome that are comprised of RepeatMasker annotations, mRNAs, and ESTs were calculated in a hierarchical manner. Genomic regions annotated by RepeatMasker were determined first, and proportions of regions covered by different RepeatMasker subclasses were further analyzed in a manner analogous to the repeat subclass analysis of the small RNA sequences. After determining the repetitive proportion of the genome, the proportions of the genome comprised of mRNAs and ESTs were determined in succession. Calculations were based on number of bases annotated divided by number of bases in the genome.

Detecting piRNA clusters

Clusters were identified by scanning a 20 kb window off-set by 1 kb across chromosomes and detecting genomic regions where at least 20 or more normalized hit counts were mapped. When qualified genomic regions were first encountered, the right hand side boundary of the window was extended progressively further in 1 kb intervals until the total normalized hit counts within 20 kb dropped below 20.

The identities and boundaries of some putative and bona fide clusters were confounded by piRNAs that mapped to several genomic loci, making it difficult to assess the defined origin of these particular sequences. Thus, we disregarded genomic regions in boundaries or putative small cluster where the total number of unique reads was less than 20. Then, both the left and right hand side boundaries were, independently, fine-tuned by trimming them progressively by 1 kb at a time toward the center of the cluster until the normalized hit counts within 1 kb became at least 1.

piRNA clusters were defined into four types (divergent, plus-strand, minus-strand, and mixed) by the following algorithm. Each clusters was scanned first on the plus strand (from the left boundary to the right boundary) and sequentially on the minus-strand (from the right boundary to the left boundary) for 5 consecutive loci where reads were mapped uniquely. Searches that identified 5 consecutive loci only from one strand in a cluster

logically classified the cluster as either a plus-strand or minus-strand type. If 5 consecutive loci were identified on both the plus- and minus-strand searches, the algorithm determined whether plus-strand loci were located downstream of minus-strand loci. Such a cluster would then be classified as a divergent type and the distance between the two plus- and minus-strand loci found by each search was calculated as a gap. In other cases, the cluster is classified as mixed type. This procedure identified exactly 100 clusters from rat and 94 from mouse and the clusters were ranked by the total normalized hit counts within each cluster. In total, the 100 rat clusters contained 56,738 normalized hit counts, which was 94% of total reads, and they covered 2,733 kb, which was less than 0.1% of the genome. Similarly, the 94 mouse clusters contained 62,466 normalized hit counts, which was 93% of total reads, and they covered 2,489 kb, which was less than 0.1 % of the genome.

Displaying the spatial distribution of sequencing reads

For each unique sequence, we normalized the number of reads by the number of genomic hits and assigned this normalized hit count equally to all the loci. For the whole chromosome view plot (Fig. 3A), the hit counts were integrated into 1 Mb bins based on the start position of their loci and plotted across each chromosome. For the cluster view plot,

the cluster region was divided into 150 equal-sized bins, the hit counts were integrated into the bins, and plotted across the cluster region. However, for the three cluster view plot in Fig. 3B, in order to compare the number of reads in a bin across three clusters, fixed-size (600 nt) bins are used regardless of size of clusters. For divergent type clusters, bins were defined so that a bin boundary can lie within the gap. For the 1 nt resolution plot, the hit counts were rounded up to the closest integer and this number was used to determine the number of horizontal bars, which were duplicated accordingly and plotted in a stacked representation on the corresponding locus (Fig. 3C).

Conservation analysis of the piRNAs in rat and mouse

In order to determine the mouse chromosomal regions orthologous to rat chromosomal regions and to estimate the primary sequence conservation between the two regions, we downloaded rat centric rat-mouse pairwise alignments from UCSC genome browser. We then used a 30-nt window to scan along a rat piRNA cluster region of interest by 1nt offsets and counted the number of conserved residues. The conservation levels ([# conserved residues] / 30) were plotted across rat piRNA cluster regions.

Estimation of substitution and insertion/deletion rate was performed as follows. For each residue within 100 rat clusters, we calculated the number of reads that uniquely mapped to that residue (and disregarded reads that mapped to more than one locus). The residues were divided into 5 groups based on the calculated number of reads: (1) 0 reads (2) 1 reads (3) 2-4 reads (4) 5-14 reads (5) more than 14 reads. For each group, the substitution rate was calculated as (total number of substituted residues) divided by (total number of residues aligned without gap). The insertion/deletion rate was calculated as (total number of residues missing in mouse) divided by (total number of residues). The calculated substitution rate was adjusted with Jukes-Cantor multiple hit correction.

To examine the conservation of piRNAs between rat and mouse, for all the mouse reads which uniquely mapped to the genome, we determined the rat loci that were orthologous to the mouse loci where mouse piRNA reads were mapped by parsing mouse centric mouse-rat pairwise alignments. The mouse reads were binned by their calculated rat loci and such bins that matched more than 2 mouse reads were depicted above and below the histograms in rat cluster view plots (Fig. 3B).

ACCESSION NUMBER GEO, GSE5026

ACKNOWLEDGMENTS

We thank N. Francis for initial work on this project; W. Johnston for technical assistance; L. Davidow, J. Morris, L. Lim, and J. Ruby for bioinformatics assistance; Z. Zhang and J. Goldman for performing preliminary assays; D. Schwarz for advice on slicer assays; and C. Woo, E. Troemel, J. Song, and S. Aigner for comments on the manuscript.

REFERENCES

Aravin, A.A., Klenov, M.S., Vagin, V.V., Bantignies, F., Cavalli, G., and Gvozdev, V.A. (2004). Dissection of a natural RNA silencing process in the Drosophila melanogaster germ line. Mol Cell Biol *24*, 6742-6750.

Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The small RNA profile during Drosophila melanogaster development. Dev Cell *5*, 337-350.

Carmell, M.A., Xuan, Z., Zhang, M.Q., and Hannon, G.J. (2002). The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. Genes Dev *16*, 2733-2742.

Catalanotto, C., Azzalin, G., Macino, G., and Cogoni, C. (2002). Involvement of small RNAs and role of the qde genes in the gene silencing pathway in Neurospora. Genes Dev *16*, 790-795.

Chen, P.Y., Manninga, H., Slanchev, K., Chien, M., Russo, J.J., Ju, J., Sheridan, R., John, B., Marks, D.S., Gaidatzis, D., *et al.* (2005). The developmental miRNA profiles of zebrafish as determined by small RNA cloning. Genes Dev *19*, 1288-1293.

Cogoni, C., and Macino, G. (1999). Posttranscriptional gene silencing in Neurospora by a RecQ DNA helicase. Science *286*, 2342-2344.

Cox, D.N., Chao, A., Baker, J., Chang, L., Qiao, D., and Lin, H. (1998). A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. Genes Dev *12*, 3715-3727.

Cui, S., Klima, R., Ochem, A., Arosio, D., Falaschi, A., and Vindigni, A. (2003). Characterization of the DNA-unwinding activity of human RECQ1, a helicase specifically stimulated by human replication protein A. J Biol Chem *278*, 1424-1432.

Deng, W., and Lin, H. (2002). miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev Cell *2*, 819-830.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., *et al.* (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493-521.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res *34*, D140-144.

Grimaud, C., Bantignies, F., Pal-Bhadra, M., Ghana, P., Bhadra, U., and Cavalli, G. (2006). RNAi components are required for nuclear clustering of Polycomb group response elements. Cell *124*, 957-971.

Hall, I.M., Shankaranarayana, G.D., Noma, K., Ayoub, N., Cohen, A., and Grewal, S.I. (2002). Establishment and maintenance of a heterochromatin domain. Science *297*, 2232-2237.

Hamilton, A., Voinnet, O., Chappell, L., and Baulcombe, D. (2002). Two classes of short interfering RNA in RNA silencing. EMBO J *21*, 4671-4679.

Kalmykova, A.I., Klenov, M.S., and Gvozdev, V.A. (2005). Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline. Nucleic Acids Res *33*, 2052-2059.

Kuramochi-Miyagawa, S., Kimura, T., Ijiri, T.W., Isobe, T., Asada, N., Fujita, Y., Ikawa, M., Iwai, N., Okabe, M., Deng, W., *et al.* (2004). Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. Development *131*, 839-849.

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science *294*, 858-862.

Lee, S.R., and Collins, K. (2006). Two classes of endogenous small RNAs in Tetrahymena thermophila. Genes Dev *20*, 28-33.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380.

Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell *110*, 563-574.

Matzke, M.A., and Birchler, J.A. (2005). RNAi-mediated pathways in the nucleus. Nat Rev Genet *6*, 24-35.

Mochizuki, K., Fine, N.A., Fujisawa, T., and Gorovsky, M.A. (2002). Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. Cell *110*, 689-699.

Pal-Bhadra, M., Bhadra, U., and Birchler, J.A. (2002). RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in Drosophila. Mol Cell *9*, 315-327.

Pal-Bhadra, M., Leibovitch, B.A., Gandhi, S.G., Rao, M., Bhadra, U., Birchler, J.A., and Elgin, S.C. (2004). Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery. Science *303*, 669-672.

Rivas, F.V., Tolia, N.H., Song, J.J., Aragon, J.P., Liu, J., Hannon, G.J., and Joshua-Tor, L. (2005). Purified Argonaute2 and an siRNA form recombinant human RISC. Nat Struct Mol Biol *12*, 340-349.

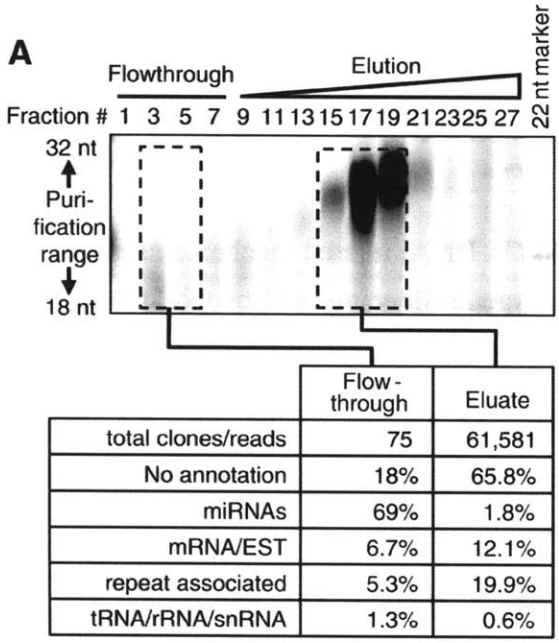
Verdel, A., Jia, S., Gerber, S., Sugiyama, T., Gygi, S., Grewal, S.I., and Moazed, D. (2004). RNAi-mediated targeting of heterochromatin by the RITS complex. Science *303*, 672-676.

Volpe, T.A., Kidner, C., Hall, I.M., Teng, G., Grewal, S.I., and Martienssen, R.A. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science *297*, 1833-1837.

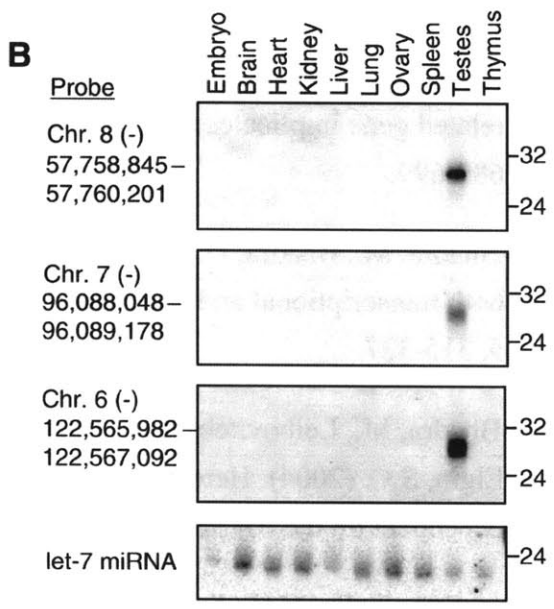
Zamore, P.D., and Haley, B. (2005). Ribo-gnome: the big world of small RNAs. Science *309*, 1519-1524.

Zilberman, D., Cao, X., and Jacobsen, S.E. (2003). ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. Science *299*, 716-719.

Figure 1



	Flow-through	Eluate
total clones/reads	75	61,581
No annotation	18%	65.8%
miRNAs	69%	1.8%
mRNA/EST	6.7%	12.1%
repeat associated	5.3%	19.9%
tRNA/rRNA/snRNA	1.3%	0.6%



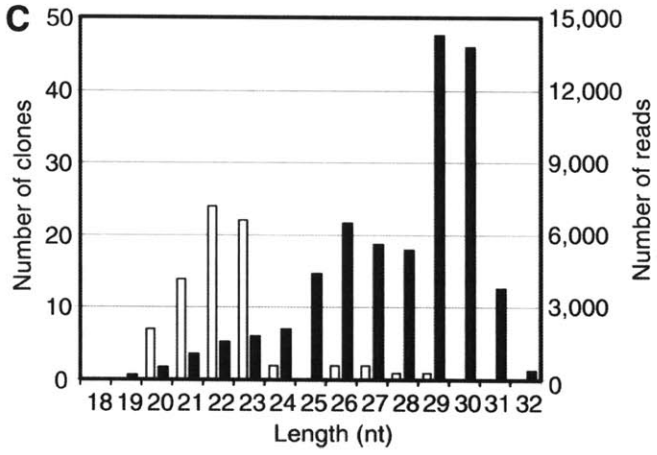


Figure 2

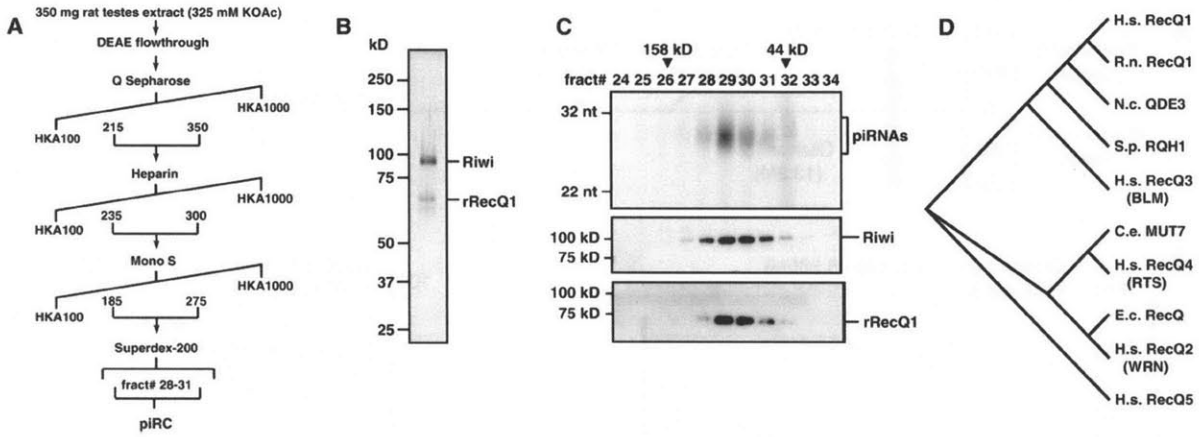

A
350 mg rat testes extract (325 mM KOAc)
DEAE flowthrough
Q Sepharose
HKA100
HKA1000
215
350
Heparin
235
300
Mono S
185
275
Superdex-200
fract# 28-31
piRC
B
kD
250
150
100
75
50
37
25
Riwi
rRecQ1
C
158 kD
44 kD
fract# 24 25 26 27 28 29 30 31 32 33 34
32 nt
22 nt
piRNAs
100 kD
75 kD
Riwi
rRecQ1
D
H.s. RecQ1
R.n. RecQ1
N.c. QDE3
S.p. RQH1
H.s. RecQ3 (BLM)
C.e. MUT7
H.s. RecQ4 (RTS)
E.c. RecQ
H.s. RecQ2 (WRN)
H.s. RecQ5

Figure 3

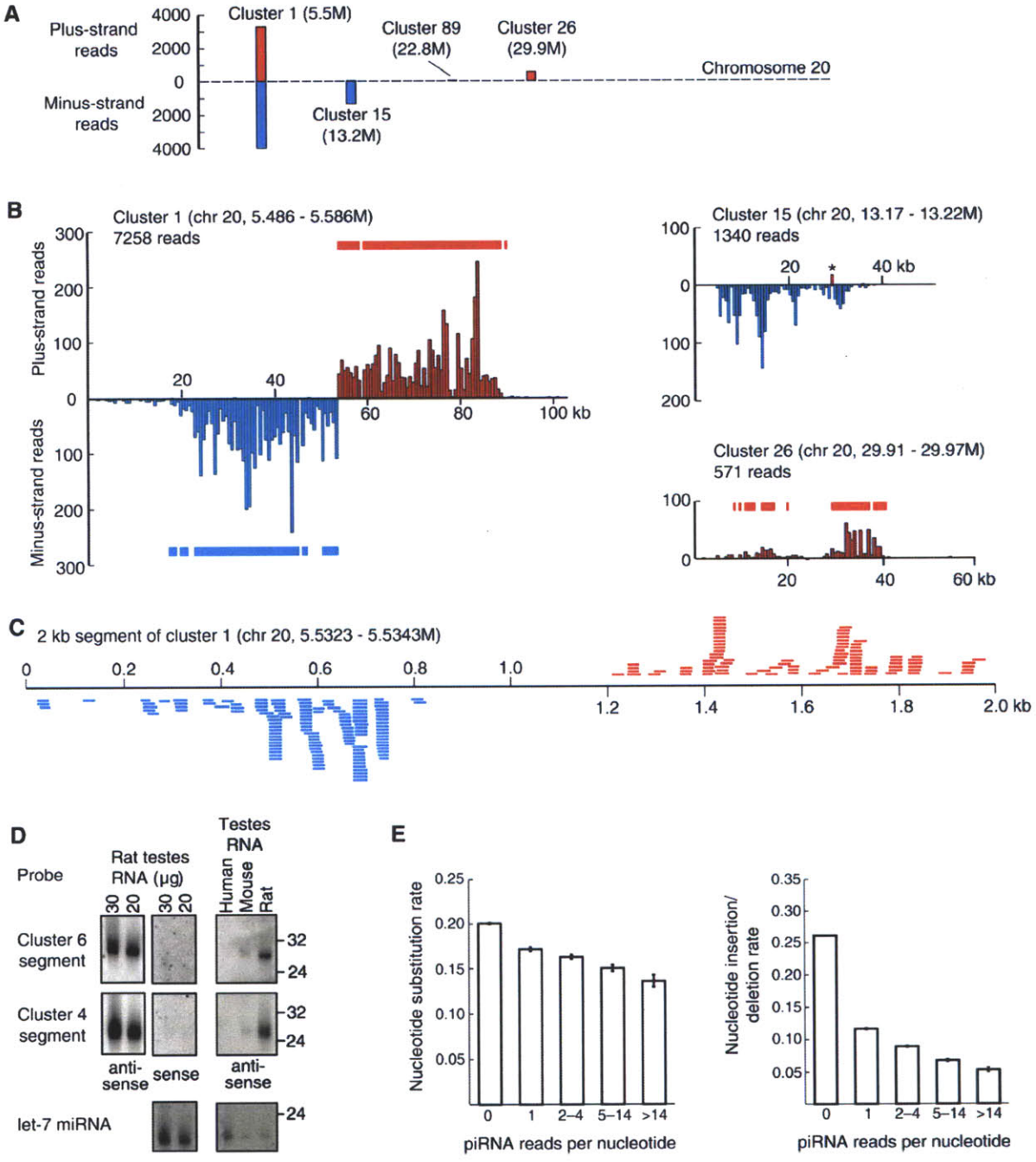

A
Plus-strand reads
Minus-strand reads
4000
2000
0
2000
4000
Cluster 1 (5.5M)
Cluster 15 (13.2M)
Cluster 89 (22.8M)
Cluster 26 (29.9M)
Chromosome 20
B
Cluster 1 (chr 20, 5.486 - 5.586M)
7258 reads
Plus-strand reads
Minus-strand reads
300
200
100
0
20
40
60
80
100 kb
Cluster 15 (chr 20, 13.17 - 13.22M)
1340 reads
20
40 kb
Cluster 26 (chr 20, 29.91 - 29.97M)
571 reads
20
40
60 kb
C
2 kb segment of cluster 1 (chr 20, 5.5323 - 5.5343M)
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0 kb
D
Probe
Rat testes RNA (μg)
30
20
Testes RNA
Human
Mouse
Rat
Cluster 6 segment
Cluster 4 segment
anti-sense
sense
let-7 miRNA
32
24
E
Nucleotide substitution rate
0.25
0.20
0.15
0.10
0.05
0
1
2–4
5–14
>14
piRNA reads per nucleotide
Nucleotide insertion/deletion rate
0.35
0.30
0.25
0.20
0.15
0.10
0.05

Figure 4

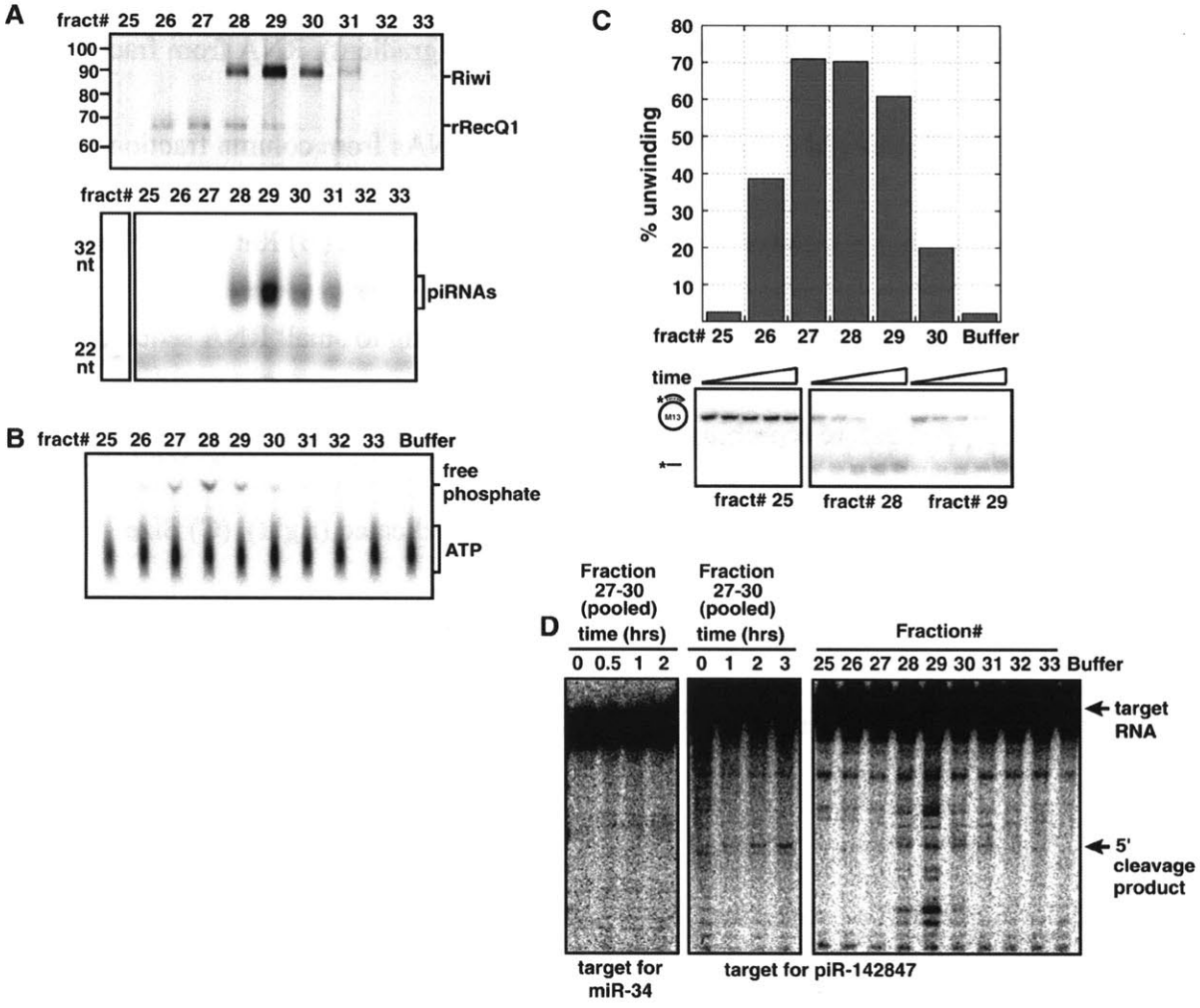

A
fract# 25 26 27 28 29 30 31 32 33
100
90
80
70
60
Riwi
rRecQ1
fract# 25 26 27 28 29 30 31 32 33
32 nt
22 nt
piRNAs
B
fract# 25 26 27 28 29 30 31 32 33 Buffer
free phosphate
ATP
C
% unwinding
80
70
60
50
40
30
20
10
fract# 25 26 27 28 29 30 Buffer
time
M13
fract# 25
fract# 28
fract# 29
D
Fraction 27-30 (pooled)
time (hrs)
0 0.5 1 2
Fraction 27-30 (pooled)
time (hrs)
0 1 2 3
Fraction#
25 26 27 28 29 30 31 32 33 Buffer
target RNA
5' cleavage product
target for miR-34
target for piR-142847

FIGURE LEGENDS

Figure 1 Testes contain a longer class of small RNAs. (A) (Top) Rat testes extract was fractionated on a Q column (0.1 to 1 M potassium acetate gradient). RNA from fractions was end-labeled and resolved on a gel. (Bottom) Small RNAs from column fractions were gel-purified (dashed boxes), converted to cDNAs, and sequenced. (B) Rat tissue Northern blot hybridized with body-labeled RNA probes corresponding to small RNA sequences. The blot was stripped before reprobing for the indicated chromosomal regions or let-7 miRNA (loading control). Migration of RNA markers is indicated (right). (C) Size distribution of small RNAs from flowthrough (white bars), and eluate (black bars). Y-axis scales are different for flowthrough RNAs (left) and eluate RNAs (right).

Figure 2 Purification of a native small RNA-containing complex revealed Riwi and rRecQ1. (A) Schematic of piRC fractionation steps from rat testes extract. Numbers represent potassium acetate (KOAc) concentration (mM). (B) Proteins from the peak fraction of piRC (Superdex-200 column) were resolved on a gel and silver stained. Bands were excised and identified by mass spectrometry to be Riwi and rRecQ1. (C) Small RNAs, Riwi, and rRecQ copurify after five steps of chromatographic separation. Final Superdex-

200 column fractions were assayed for the presence of small RNAs, Riwi, or rRecQ1. (Top) Small RNAs were end-labeled and resolved on a gel. (Middle and bottom) Western blots probed with antibodies to Miwi (mouse Piwi) and to hRecQ1. Elution profile of protein size markers from Superdex-200 column indicated above. (D) Phylogenetic comparison of RecQ DNA helicase family members revealed Neurospora QDE-3 to be a close homolog to rRecQ1.

Figure 3 Genomic characteristics of rat piRNAs. (A) Chromosomal view indicating the number of piRNA reads mapping to clusters on chromosome (chr) 20 [(27)]. (B) Medium-resolution view of clusters 1, 15, and 26. Horizontal bars above and below the histograms indicate regions orthologous to mouse regions that also produce piRNAs (indicated if the bin matches >2 uniquely mapping mouse piRNAs). An asterisk denotes a group of piRNAs from cluster 15 that perfectly mapped to more than one locus in the genome. (C) High-resolution view centered on the gap region that separates minus- and plus-strand piRNA hits in cluster 1. Horizontal bars represent individual piRNAs. (D) Northern blot analysis with probes to the indicated clusters, testing strand-specific piRNA expression (left and middle) and cross-hybridization to mouse and human testes RNAs (right). Migration of

RNA markers is indicated (far right). Blots were stripped and reprobed to let-7 miRNA (loading control). (E) piRNA conservation analysis. Orthologous rat and mouse clusters were identified, and rat residues were binned based on the number of matching rat piRNA reads. The estimated substitution rate per residue (left) and estimated insertion/deletion rate (right), comparing rat to mouse, was calculated for each bin. Error bars indicate 95% confidence intervals of the estimates (27).

Figure 4 piRC fractions contained ATP-dependent DNA helicase and slicer activities. (A) Visualization of proteins (top, silver stain) and end-labeled small RNAs (bottom) in fractions from final Superdex-200 column (independent purification from that shown in Fig. 2C). Size standards indicated on left. (B) Fractions containing rRecQ1 exhibit ATPase activity. Fractions were incubated with radiolabeled ATP. Free phosphate generated by ATPase activity was separated from unhydrolyzed ATP on thin-layer chromatography. (C) Fractions containing rRecQ1 exhibited DNA unwinding activity. Fractions were incubated with a DNA substrate containing a 17-bp duplex. Reactions were resolved on a native gel. Appearance of the faster migrating radiolabeled 17-nt oligomers indicated DNA duplex unwinding. (D) Fractions containing Riwi and piRNAs exhibit slicer activity. Cap-labeled

substrates, complementary to piRNA or miRNA (negative control) sequences, and cleavage products were resolved on a gel. The GenBank accession number for piR-142847 is DQ727525.

CHAPTER III

Allelic Imbalance Sequencing Reveals That Single-nucleotide Polymorphisms Frequently Alter MicroRNA-directed Repression

All experiments and analyses regarding the work presented in this chapter were conducted by myself.

This work has been published previously as:

Jinkuk Kim, David P. Bartel, Allelic imbalance sequencing reveals that single-nucleotide polymorphisms frequently alter microRNA-directed repression, *Nature Biotechnology* 2009, 27(5):472-7

ABSTRACT: Genetic changes that help explain the differences between two individuals might create or disrupt sites complementary to microRNAs (miRNAs), but the extent to which such polymorphic sites influence miRNA-mediated repression is unknown. Here, we describe a method to measure mRNA allelic imbalances associated with a regulatory site found in mRNA transcribed from one allele but not found in that transcribed from the other. Applying this method, called allelic imbalance sequencing, to sites for three miRNAs (miR-1, miR-133 and miR-122) provided quantitative measurements of repression in vivo without altering either the miRNAs or their targets. A substantial fraction of polymorphic sites mediated repression in tissues that expressed the cognate miRNA, and downregulation was correlated with site type and site context. Extrapolating these results to the other broadly conserved miRNAs suggests that when comparing two mouse strains (or two human individuals), polymorphic miRNA sites cause expression of many genes (often hundreds) to differ.

MicroRNAs are ~23-nucleotide endogenous RNAs that pair to mRNAs to direct their post-transcriptional repression (Bartel, 2009). To explore miRNA regulatory diversity within a single species, we considered miRNA complementary sites that are created or disrupted by single-nucleotide polymorphisms (SNPs) in mice. Our study centered on three types of complementary sites that previous computational and experimental results indicated can mediate miRNA recognition3. Each of these three sites includes perfect Watson-Crick pairing to the miRNA seed (miRNA positions 2–7; Fig. 1a). One is a 7-nucleotide site, referred to here as the 7mer-m8 site, for which seed pairing is supplemented by a Watson-Crick match to miRNA nucleotide 8 (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005). Another is the 7mer-A1 site, for which seed pairing is supplemented by an adenine nucleotide across from miRNA nucleotide 1 (Lewis et al., 2005). And the third is the 8-nucleotide or 8mer site, which has both the m8 match and the A across from position 1 (Lewis et al., 2005). We focused on sites recognized by three miRNAs—miR-1 (which for our purposes is synonymous with its paralog, miR-206), miR-133 and miR-122—because these miRNAs show strong, tissue-specific expression in relatively homogenous and accessible tissues, muscle (miR-1 and miR-133) or liver (miR-122) (Lagos-Quintana et al., 2002). In agreement with previous reports) (Clop et al., 2006; Sethupathy and Collins,

2008), searching SNP databases (Sherry et al., 2001; Frazer et al., 2007b) for polymorphisms within mRNA 3' untranslated regions (UTRs), which are the regions most likely to be targeted by miRNAs (Grimson et al., 2007), revealed many SNPs that create or disrupt sites for one of the three miRNAs (with gain or loss considered relative to an outgroup sequence). Because miRNAs often destabilize their target mRNAs (Lim et al., 2005), we reasoned that if these sites were functional in the tissue expressing the cognate miRNA, then less RNA might accumulate from the allele with the site. Moreover, in mice heterozygous for the SNPs, destabilization of mRNA from the target allele, but not from the nontarget allele, would contribute to allelic imbalance in mRNA steady-state levels. Hence, we developed allelic imbalance sequencing (AI-Seq) to measure such imbalances, reasoning that any imbalances would identify and quantify miRNA regulatory diversity within a species, and provide a unique opportunity to examine the molecular consequences of miRNA-mediated repression in vivo without perturbing either the miRNA or its targets.

Because lab strains lack the heterozygosity found in natural populations, we performed five inter-strain crosses to generate mice heterozygous for the parental alleles. Approximately 300 annotated SNPs that create or disrupt target sites for one of the three miRNAs were heterozygous in F_1 progeny from at least one of the five crosses. We chose a

subset of these, preferring those that create or disrupt 8mer sites, those in messages with evidence of expression in the tissues expressing the miRNAs and those that were not linked to many nearby polymorphisms. Allelic imbalance was measured for 67 target sites (28 for miR-122, 28 for miR-1 and 11 for miR-133) in the tissue expressing the cognate miRNA.

For AI-Seq, mRNA fragments containing the SNPs were first reverse transcribed and amplified (PCR), and then the amplicon was subjected to high-throughput pyrosequencing (Margulies et al., 2005) (Fig. 1b). To economize on sequencing, we pooled amplicons derived from different primers. Because the primers flanking the SNPs used for RT-PCR were gene specific but not allele specific, both alleles of the same gene were amplified by the same reaction, and their relative abundance could be inferred from the number of sequencing reads representing each allele. We quantified these relative abundances using the allelic ratio, defined here as the $\log_2$ of the number of reads representing the target allele divided by the number of reads representing the non-target allele, after normalizing to the ratio obtained using genomic DNA (Fig. 1b).

If none of the intact miRNA sites directed repression, the allelic ratios would be expected to center on zero, with individual ratios deviating from zero because of experimental noise. However, contrary to this null hypothesis, when liver tissues were

assayed using AI-Seq, the distribution of the allelic ratios for the 28 miR-122 sites centered below zero (Fig. 2a), consistent with the hypothesis that mRNA from some of the alleles with target sites was destabilized. If miR-122 caused this destabilization, then the shift from zero should depend on the presence of this miRNA. To test this dependency, we measured ratios for 22 of the 28 miR-122 sites in muscle, which does not express miR-122. Ratios for the remaining six sites were not considered because four were in messages not expressed in muscle, and the other two were unusual in that the SNP disrupting the miR-122 site (CACTCCA and ACACTCC, SNP underlined) simultaneously created a site for miR-1 (CATTCCA and ACATTCC), which is expressed in muscle, thereby precluding the use of these as negative controls. As expected for a miRNA-mediated effect, the shift from zero disappeared in muscle (Fig. 2a). When analyzing the ratios for the 39 sites for miR-1 or miR-133, which are expressed in the muscle but not in the liver, the reciprocal pattern was observed—namely, the distribution of the ratios measured in liver centered on zero and that of the ratios measured in muscle was skewed toward lower values (Fig. 2b).

To increase sample size and thereby achieve statistical significance, we combined data sets such that the ratios measured in the presence of the cognate miRNA were analyzed together and compared to those measured in the absence of the miRNA (Fig. 2c). A

significantly large fraction of the allelic ratios were <0 in the presence of the miRNA ($P <$ 0.01, one-sided exact binomial test), but not in the absence of the miRNA ($P = 0.6$), and the difference between the two distributions also was significant ($P = 0.02$, one-sided Kolmogorov-Smirnov (KS) test). Thus, we concluded that at least a subset of the interrogated target sites mediated repression.

On average, the polymorphic sites were associated with mRNA downregulation of 12% (Fig. 2c; 95% confidence interval of 5–18%, bootstrapping). Actual downregulation was likely greater because the signal could have been diluted by both nuclear mRNA and mRNA from cells that do not express the cognate miRNA, such as those from blood or vasculature. Effects of functional sites also might have been diluted by inclusion of nonfunctional sites. Nonfunctional sites presumably were enriched among the set of sites interrogated in this study because natural selection selects against polymorphisms that either disrupt beneficial functional sites or generate functional sites in messages that should not be repressed (Farh et al., 2005; Stark et al., 2005; Chen and Rajewsky, 2006).

We estimated the lower bound for the fraction of functional sites to be 16% by analyzing the maximal vertical displacement of the cumulative distribution curves (correcting for the bumpiness of the distributions (Grimson et al., 2007)). This estimate is

likely to be conservative because simulations showed that under certain assumptions our analysis may only identify about a third of all sites simulated to be functional (see Methods). These simulations incorporated the variability observed from the tissues lacking the miRNA and assumed that all sites mediated target repression by 20%. If, as in this simulation, only about a third of the active sites were detected, then our lower bound of 16% might be only a third of the actual fraction, in which case about half of the examined polymorphic sites mediated repression.

The variability observed in our experiments can be attributed to multiple sources. One source is stochastic sampling error inherent to counting sequencing reads, which can be modeled by the binomial distribution (Fig. 2d). A second source is PCR variability, which can be estimated as the variability of the allelic ratios measured using gDNA minus the stochastic error (Fig. 2d; difference between gDNA and binomial distributions). A third source is biological noise, which could include differences in the epigenetic states of the two alleles or allelic differences in linked *cis*-regulatory elements. To begin to estimate the biological noise, we examined the distribution of the allelic ratios of control mRNAs that were not predicted to be repressed in an allele-specific manner by the three miRNAs, such as mRNAs from tissues lacking the cognate miRNA. Allelic ratios of these control mRNAs

were substantially more variable than those of gDNA, suggesting frequent allelic imbalance not attributable to the sites under investigation (Fig. 2d). However, we were unable to quantify the frequency or magnitude of this potentially widespread allelic imbalance because of the possibility that reverse transcription variability also contributed to the greater variability of the mRNA controls compared to that of the gDNA controls.

Our quantitative assay of site efficacy *in vivo* did not perturb either the miRNAs or their targets and thereby provided a fresh opportunity to examine the influence of site type and site context on miRNA activity. The 8mer sites performed significantly better than did 7mer-m8 or 7mer-A1 sites (Fig. 3a; $P = 0.005$ and $P = 0.001$ respectively, one-sided KS test), and 7mer-m8 sites tended to perform slightly better than did 7mer-A1 sites, although this difference was not statistically significant ($P = 0.1$, one-sided KS test). The overall rank order of the efficacy of the three types was consistent with previous observations from experiments that ectopically expressed or deleted miRNAs (Grimson et al., 2007; Nielsen et al., 2007; Baek et al., 2008; Selbach et al., 2008).

To consider the influence of site context, we calculated the 'context score' for each polymorphic site. Context scores quantitatively evaluate site type and three features of site context (that is, surrounding AU content, position within the 3' UTR and pairing to the 3'

region of miRNA) to predict site efficacy (Grimson et al., 2007). Context scores significantly correlated with target downregulation in the presence of the cognate miRNA, but not in the absence of the miRNA ($P < 0.001$; Fig. 3b). Significant correlation was retained in the presence of the miRNA even after the contribution of site type had been factored out, thereby indicating that site context, as scored by this model, influences the efficacy of polymorphic sites ($P < 0.01$; Fig. 3c).

Our experiments focused on three of the 87 miRNA families conserved in chicken or more divergent vertebrates (Friedman et al., 2009). Expanding our SNP database search to the other 84 broadly conserved miRNA families and the ~8 million SNPs annotated in 15 mouse strains (Frazer et al., 2007b) showed that any two strains have on average 2,430 distinct polymorphic sites (bottom and top 2.5 percentile, 810–4,600; median, 1,470) and 1,510 genes with at least one polymorphic miRNA site (bottom and top 2.5 percentile, 520–2,790; median, 950). These numbers would increase if sites recognized by the hundreds of additional annotated miRNAs were also considered. However, because species-specific miRNAs and those conserved only within mammals tend to be expressed at lower levels, their 7- to 8-nucleotide sites are thought to be less frequently sufficient for mediating

repression (Bartel, 2009; Friedman et al., 2009). Therefore, to guard against overstating the impact of polymorphic sites, we did not consider these additional miRNAs.

An estimate of the direct impact of polymorphic miRNA sites on gene-expression variation within a species can be extrapolated from our results as follows. First, for the 67 sites examined, we observed average downregulation of 12%, with at least 16% of the sites responsible for the observed downregulation. Correcting for our preference in choosing 8mer sites for analysis slightly lowered the average downregulation to 10% and the percentage of functional sites to 15% as our best estimates for all 7- to 8-nucleotide polymorphic sites in mRNAs expressed in cognate tissues. If we assume that 50% of the genes with these sites are coexpressed with the cognate miRNA in the same cell type, we can use our observed lower limit of 15% functional sites to estimate that at least 7.5% (0.5×0.15) of the genes with polymorphic sites will be differentially regulated between two strains. Thus, between many mouse strains, over a hundred messages are likely to be differentially regulated through polymorphic sites, with average mRNA downregulation for these messages >60% (correcting for dilutive effect of inactive sites by dividing the average downregulation of all sites, 10%, by 0.15). Using a less conservative estimate that 50% of the sites in cognate tissues are functional, proportionally more messages would be

downregulated, with average downregulation of functional sites still ~20%. Because miRNAs also influence translation, effects on the proteome are presumed to be even greater. Overall, it is hard to escape the conclusion that polymorphic miRNA regulatory sites have a substantial impact on gene-expression variation within a species.

Our results in mice, considered with SNP frequencies in humans, indicated that any two unrelated humans probably have more than a hundred genes differentially regulated because of polymorphic miRNA targeting (Supplementary Discussion). Assuming that some of these could explain differences in disease risk among individuals, our results suggest that, as more genome-wide association studies are conducted with improved coverage in 3' UTRs, more miRNA target-site polymorphisms will be associated with clinical conditions and individual traits[2].

Another approach for detecting effects of regulatory SNPs is provided by studies of expression quantitative trait loci (eQTL) (Cookson et al., 2009). In eQTL studies, correlation between the genotype of a polymorphic locus and expression of a gene is calculated for each locus:gene pair. In principle, these studies involving unrelated individuals should preferentially identify polymorphic targets as *cis*-regulated because the SNPs in functional target sites (and other linked SNPs) should be associated with

expression of the targets. However, when we analyzed the results of a large-scale eQTL study that used >400 human liver samples (Schadt et al., 2008), polymorphic miR-122 targets were not enriched among the genes identified as *cis*-regulated any more than were polymorphic miR-1 targets (data not shown). We attribute the greater sensitivity of AI-Seq to the internal reference provided by the nontarget allele, which normalizes for environmental differences, *trans*-acting genetic differences and other sources of sample variability, thereby more effectively isolating the influence of the site on expression. Also important for the success of our approach in detecting the relatively subtle effect of miRNAs was the precision achieved by high-throughput sequencing. Previous studies using heterozygous SNPs to detect allelic expression imbalances rely on allele-specific hybridization or primer extension (Lo et al., 2003; Gimelbrant et al., 2007; Serre et al., 2008; Tan et al., 2008), both of which, when compared at the gDNA level, were substantially noisier than our sequencing-based method (Supplementary Fig. 1).

Despite detecting the effects of polymorphic miRNA sites in mouse tissues, miRNA effects were not detected in a HapMap (Frazer et al., 2007a) panel of lymphoblastoid cell lines when we used AI-Seq to measure the allelic imbalance of 56 heterozygous target sites for nine miRNA families most highly expressed in these cell lines (data not shown). In this

case, the imbalances expected to result from polymorphic miRNA sites might have been overwhelmed by random monoallelic expression present in clonal subpopulations of these lines (Plagnol et al., 2008). Moreover, the process of establishing lymphoblastoid cell lines, which involved Epstein-Barr-virus infection and subsequent transformation of B-cells, might have downregulated miRNA expression (Lu et al., 2005).

Experiments examining the influence of miRNA knockouts on the transcriptome and proteome have been informative for inferring the effects of conserved and nonconserved sites that are not polymorphic (Rodriguez et al., 2007; Baek et al., 2008). Our results complement these studies by revealing the influence of polymorphic sites without perturbing either the miRNAs or their targets. Following miRNA knockout, upregulation of targets can trigger feedback regulation that reduces the observed effect of losing the miRNA. Such a response is not likely to confound our AI-Seq results because feedback regulation is usually not allele specific and therefore is unlikely to change the relative expression of the target compared to the nontarget alleles. Our approach can be extended to characterize other *cis*-regulatory elements that might influence mRNA levels. As the capacity of high-throughput sequencing increases, we anticipate that RNA-Seq coverage will expand so that directed amplification of specific loci will no longer be required to

accurately detect allelic imbalances. Then, our approach of correlating imbalances with predicted regulatory sites can be applied transcriptome-wide to reveal many of the polymorphic regulatory sites contributing to these imbalances.

METHODS

Mouse tissues and preparation of cDNA and gDNA

The study was approved by the MIT Committee on Animal Care. The Jackson laboratory performed five inter-strain crosses (CAST/EIJ × PWD/PHJ, FVB/NJ ×PWD/PHJ, A/J × C57BL/6J, WSB/EiJ × MOLF/EiJ, A/J × DBA/2J) and dissected liver and skeletal muscle from two 4-week-old F_1 littermates of each cross. For each cross and tissue, ~0.6 g tissue (~0.3 g from each littermate) was homogenized for RNA extraction (RNeasy Maxi kit, Qiagen), and cDNA was synthesized from total RNA in reverse transcriptase reactions (Superscript III, Invitrogen) primed with random hexamers. For each cross, gDNA was isolated from ~50 mg of either liver or muscle from either littermate (DNeasy Blood and Tissue kit, Qiagen).

Computational identification of polymorphic sites

Genomic coordinates of known mouse SNPs on the July 2007 genome assembly (mm9) were obtained from NCBI dbSNP build 128 (Sherry et al., 2001). Genotypes of mouse strains were obtained from dbSNP build 128, mm9 genomic sequence (for C57BL/6J strain) and the Perlegen data (http://mouse.perlegen.com/) (Frazer et al., 2007b). Gene annotation

on mm9 was obtained from UCSC genome browser. We identified SNPs that generate heterozygous sites for miR-122, miR-1 or miR-133 in at least one of the five crosses (123 SNPs for miR-122, 109 SNPs for miR-1 and 74 SNPs for miR-133; 7-nucleotide sites to 8-nucleotide sites ratio, 10.3), excluding those that modify the sites, for example, by converting a 7mer site to an 8mer site or vice verse.

Site selection and DNA amplification

Polymorphic sites located <15 nucleotides from the stop codon were excluded (Grimson et al., 2007). Also excluded were those in the genes expressed, according to the mouse expression atlas (Su et al., 2004), at a level lower than that of 90% of all genes in the tissue that expresses the cognate miRNA; sites in genes without an expression measurement were not excluded. Out of the remaining sites, all polymorphic 8mer sites were chosen. A subset of the 7-nucleotide polymorphic sites was chosen somewhat arbitrarily, preferring those with fewer additional SNPs in flanking regions, which could potentially interfere with primer annealing. A total of 138 SNPs were carried forward for primer design, performed with the aid of PRIMER3, and suitable primers were found for 124 of those, which corresponded to 136 polymorphic sites (7-nucleotide sites to 8-nucleotide sites ratio, 5.5).

PCR amplification reactions of each of the SNPs were performed individually (Phusion Hot Start polymerase, New England Biolabs). All PCR reactions done with cDNA were accompanied by a matching no-reverse transcriptase control. Fragments that failed to be amplified at sufficient yield from either gDNA or cDNA were discarded. For each successful amplification, the procedure was repeated using cDNA from the noncognate tissue. As additional controls to examine variability of allelic imbalances that were not attributable to miRNA targeting, SNPs that do not generate polymorphic miRNA sites for the three miRNAs were identified in the open reading frames (ORFs) of 27 genes that had polymorphic sites in the 3' UTRs, and these 27 SNPs were amplified using F_1 hybrids that were heterozygous for the ORF SNP but homozygous for the 3' UTR SNP. In total, 240 amplicons (70 3' UTR SNPs from cDNA of the tissue with miRNA expression, 49 3' UTR SNPs from cDNA of the tissue without miRNA expression, 27 ORF SNPs from cDNA of either tissue and 94 SNPs from gDNA) were prepared for sequencing.

Mixing and purifying PCR products for pyrosequencing

Because gDNA-templated, liver-mRNA-templated and muscle-mRNA-templated amplicons all shared the same primers, they needed to be sequenced in separate pools, so

that they could be distinguished from one another. Because one pyrosequencing plate can be divided into four segments without contamination between segments, the 240 amplicons were mixed into four pools. Each pool had ~60 amplicons, mixed in equimolar amounts after determination of each amplicon concentration (Bioanalyzer, Agilent Technologies). Each pool was deproteinated (phenol, chloroform with iso-amyl-alcohol), purified by native PAGE gel, taking precautions to avoid denaturing the double-stranded PCR products, and submitted to 454 Life Sciences for sequencing. Three sequencing runs were performed.

Analysis of sequencing reads

Of the ~1.194 million reads acquired, ~1.085 million (~91%) correctly mapped to the 3'-terminal 10-nucleotide fragments of the unique primer pairs. Of the 240 amplicons, 9 were excluded for at least one of the following reasons: (i) the number of reads obtained per amplicon was <300, (ii) a SNP was not detected, (iii) a severe allelic bias was observed with gDNA. The remaining 231 amplicons had a median of 3,978 reads (range, 541–18,714) and corresponded to 65 3' UTR SNPs and 25 ORF SNPs. Although most of the sites were polymorphic for only one of the three miRNAs, five exceptional SNPs allowed one allele to have a miR-122 site and the other allele to have a miR-1 or miR-133 site. Three of the five

were in mRNA expressed only in muscle, but the remaining two (NCBI dbSNP IDs: rs36333425, rs30114270) were expressed in both tissues, allowing the measurement taken from liver to report on the miR-122 site and that from muscle to report on miR-1 site. Other exceptions were the two polymorphic sites in the same mRNA (NCBI dbSNP IDs: rs32325030, rs32323893) that exist in the same cross and collectively allow one *Snap29* allele to have two miR-122 sites and the other allele to have none. For these, the allelic ratio was measured separately in liver for each site, but because each ratio was likely to reflect the effect of two target sites, each was analyzed after reduction by half the $\log_2$value.

Evaluation of previous allelic imbalance measurement methods

To evaluate SNP arrays, we downloaded from the HapMap (Frazer et al., 2007a) website the raw signal-intensity data generated by hybridizing gDNA of a HapMap individual (NA19193) on the Affymetrix GeneChip 250K Nsp array. For evaluating the GoldenGate primer-extension assay, we downloaded from the Gene Expression Omnibus the raw signal-intensity data (GSM199494, GSM200074) generated from the Illumina GoldenGate assay with gDNA of a HapMap individual (NA10836) (Tan et al., 2008). For both cases,

SNPs annotated as heterozygous in the individual were identified from the HapMap genotype database, and the allele-specific probe intensity values for the SNPs were used to calculate $\log_2$ ratios of one random allele to the other.

Statistical analyses

MATLAB was used for all statistical analyses. To calculate the statistical significance for the observed number of allelic ratios with values < 0, we used the one-sided exact binomial test, in which the *P* value was the probability that a random variable following binomial distribution (parameters: $P = 0.5$, n = [the total number of sites]) was equal to or larger than the observed number. To calculate the significance of difference between two distributions, we chose the KS test over the Wilcoxon rank sum test (Mann-Whitney U test) or *t*-test, because the KS test is based on fewer assumptions on the data and almost always provided the most conservative *P*-value compared to the other two tests. When estimating the lower bound for the number of active polymorphic sites, the contribution of experimental noise (illustrated by the bumpiness of the cumulative distributions) in increasing the maximal offset between the cognate and control distributions needed to be subtracted. To estimate the contribution of this noise, we merged the two distributions and generated 1,000 pairs of

distributions by random sampling, with replacement, maintaining the sizes of the original distributions. Then we calculated the maximum difference in cumulative fraction for each pair of simulated distributions and subtracted the median of the 1,000 values from the observed maximum offset. When simulating under the assumption that all sites mediate 20% downregulation, we started with the control distribution of 47 allelic ratios measured using the tissue lacking the miRNA and randomly drew (with replacement) 67 samples, which matched the size of the cognate distribution. To simulate 20% downregulation mediated by all sites, the sampled allelic ratios were each adjusted by offsetting them by -0.32 or $\log_2(0.8)$. We generated 1,000 such simulated distributions, each of which was compared to the control distribution to determine the maximum offset in cumulative fraction. The median of the resulting 1,000 values was considered as the representative estimate for the maximum cumulative difference between the simulated and control distributions. This difference was corrected for the bumpiness of the distributions, as explained above, to yield the detectable fraction of functional sites. The observed average downregulation of all examined polymorphic sites was corrected for our preferential choice of 8mer sites for analysis by recalculating the mean allelic ratio of all sites after reducing the contribution from 8mer sites by 1.87 fold (10.3/5.5), which was the enrichment of 8mer sites among the

polymorphic sites analyzed. The observed lower bound for the fraction of functional sites was similarly adjusted.

ACCESSION NUMBER GEO, GSE15675

ACKNOWLEDGMENTS

We thank Tim Harkins and the 454 Sequencing Facility for high-throughput sequencing, members of the laboratory for helpful comments on this manuscript, and T. DiCesare for illustration.

REFERENCES

Baek, D., Villen, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. (2008). The impact of microRNAs on protein output. Nature *455*, 64-71.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215-233.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of microRNA-target recognition. PLoS Biol *3*, e85.

Chen, K., and Rajewsky, N. (2006). Natural selection on human microRNA binding sites inferred from SNP data. Nat Genet *38*, 1452-1456.

Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibe, B., Bouix, J., Caiment, F., Elsen, J.M., Eychenne, F., *et al.* (2006). A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat Genet *38*, 813-818.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. Nat Rev Genet *10*, 184-194.

Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science *310*, 1817-1821.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., *et al.* (2007a). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

Frazer, K.A., Eskin, E., Kang, H.M., Bogue, M.A., Hinds, D.A., Beilharz, E.J., Gupta, R.V., Montgomery, J., Morenzoni, M.M., Nilsen, G.B., *et al.* (2007b). A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature *448*, 1050-1053.

Friedman, R.C., Farh, K.K., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res *19*, 92-105.

Gimelbrant, A., Hutchinson, J.N., Thompson, B.R., and Chess, A. (2007). Widespread monoallelic expression on human autosomes. Science *318*, 1136-1140.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell *27*, 91-105.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., *et al.* (2005). Combinatorial microRNA target predictions. Nat Genet *37*, 495-500.

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. Curr Biol *12*, 735-739.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15-20.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature *433*, 769-773.

Lo, H.S., Wang, Z., Hu, Y., Yang, H.H., Gere, S., Buetow, K.H., and Lee, M.P. (2003). Allelic variation in gene expression is common in the human genome. Genome Res *13*, 1855-1862.

Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., *et al.* (2005). MicroRNA expression profiles classify human cancers. Nature *435*, 834-838.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376-380.

Nielsen, C.B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., and Burge, C.B. (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. RNA *13*, 1894-1910.

Plagnol, V., Uz, E., Wallace, C., Stevens, H., Clayton, D., Ozcelik, T., and Todd, J.A. (2008). Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. PLoS ONE *3*, e2966.

Rodriguez, A., Vigorito, E., Clare, S., Warren, M.V., Couttet, P., Soond, D.R., van Dongen, S., Grocock, R.J., Das, P.P., Miska, E.A., *et al.* (2007). Requirement of bic/microRNA-155 for normal immune function. Science *316*, 608-611.

Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., *et al.* (2008). Mapping the genetic architecture of gene expression in human liver. PLoS Biol *6*, e107.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58-63.

Serre, D., Gurd, S., Ge, B., Sladek, R., Sinnett, D., Harmsen, E., Bibikova, M., Chudin, E., Barker, D.L., Dickinson, T., *et al.* (2008). Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. PLoS Genet *4*, e1000006.

Sethupathy, P., and Collins, F.S. (2008). MicroRNA target site polymorphisms and human disease. Trends Genet *24*, 489-497.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res *29*, 308-311.

Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell *123*, 1133-1146.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., *et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A *101*, 6062-6067.

Tan, A.C., Fan, J.B., Karikari, C., Bibikova, M., Garcia, E.W., Zhou, L., Barker, D., Serre, D., Feldmann, G., Hruban, R.H., *et al.* (2008). Allele-specific expression in the germline of patients with familial pancreatic cancer: an unbiased approach to cancer gene discovery. Cancer Biol Ther *7*, 135-144.

Figure 1

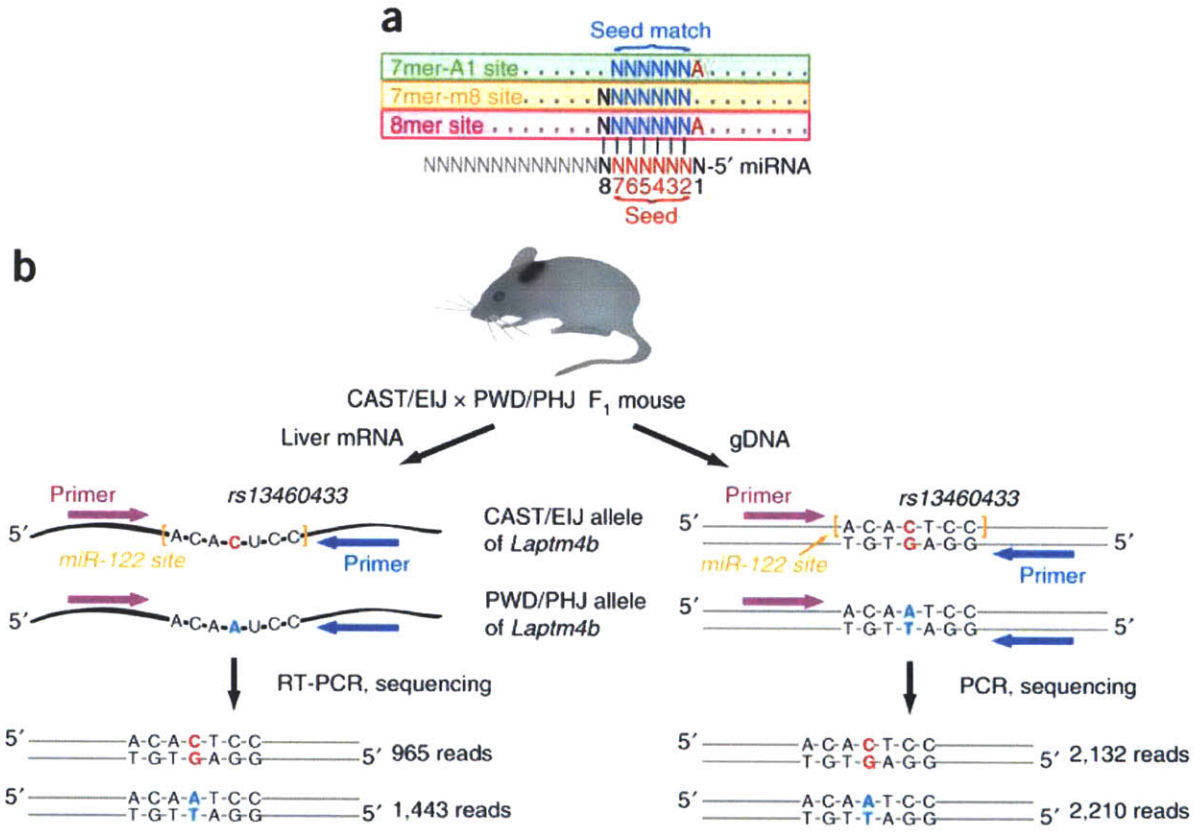

a
Seed match
7mer-A1 site . . . . . NNNNNNA . . . . . .
7mer-m8 site . . . . . NNNNNNN . . . . . .
8mer site . . . . . NNNNNNNA . . . . . .
NNNNNNNNNNNNNNNNNNNNN-5′ miRNA
87654321
Seed
b
CAST/EIJ × PWD/PHJ F1 mouse
Liver mRNA
gDNA
Primer
rs13460433
5′
A-C-A-C-U-C-C
miR-122 site
Primer
CAST/EIJ allele of Laptm4b
PWD/PHJ allele of Laptm4b
A-C-A-A-U-C-C
RT-PCR, sequencing
PCR, sequencing
A-C-A-C-T-C-C
T-G-T-G-A-G-G
5′ 965 reads
A-C-A-A-T-C-C
T-G-T-T-A-G-G
5′ 1,443 reads
5′ 2,132 reads
5′ 2,210 reads

Figure 2

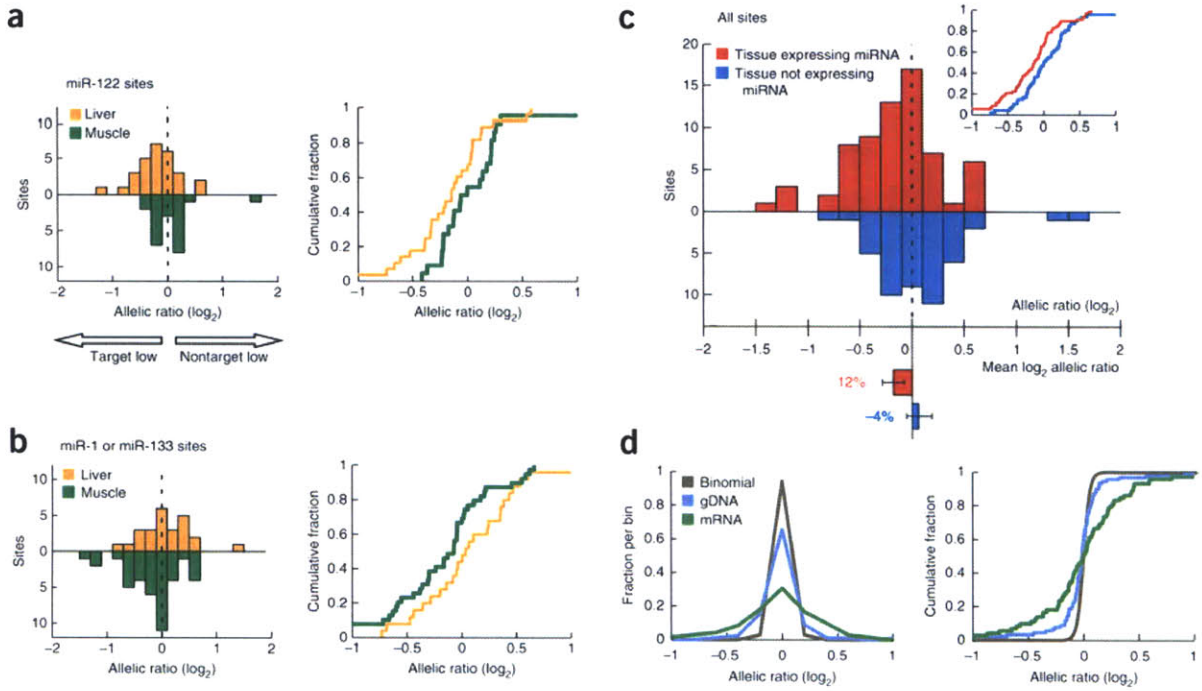

a
miR-122 sites
Liver
Muscle
Sites
Allelic ratio (log2)
Target low
Nontarget low
Cumulative fraction
b
miR-1 or miR-133 sites
c
All sites
Tissue expressing miRNA
Tissue not expressing miRNA
Mean log2 allelic ratio
12%
-4%
d
Binomial
gDNA
mRNA
Fraction per bin

Figure 3

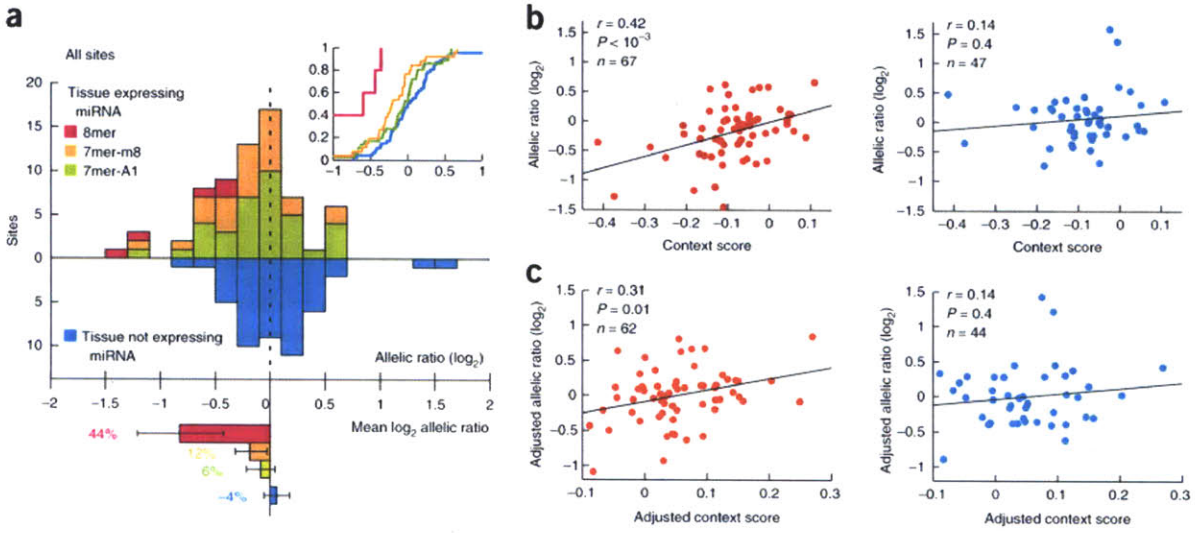

a
All sites
Tissue expressing miRNA
8mer
7mer-m8
7mer-A1
Tissue not expressing miRNA
Sites
Allelic ratio (log2)
Mean log2 allelic ratio
44%
12%
6%
4%
b
r = 0.42
P < 10−3
n = 67
Allelic ratio (log2)
Context score
r = 0.14
P = 0.4
n = 47
Allelic ratio (log2)
Context score
c
r = 0.31
P = 0.01
n = 62
Adjusted allelic ratio (log2)
Adjusted context score
r = 0.14
P = 0.4
n = 44
Adjusted allelic ratio (log2)
Adjusted context score

FIGURE LEGENDS

Figure 1 Measurement of mRNA allelic imbalances associated with heterozygous miRNA target sites. (**a**) Canonical 7–8-nt miRNA target sites. (**b**) AI-Seq, illustrated for SNP rs13460433, which generates a heterozygous miR-122 target site in the *Laptm4b* gene of CAST/EIJ × PWD/PHJ F1 mice. When using liver mRNA, the 965 and 1443 reads obtained from the target and the non-target allele, respectively, imply an allelic imbalance of 0.67 (= 965 / 1443). Because allele-specific PCR bias might have influenced this ratio, amplification and sequencing was performed in parallel with the same primers but using genomic DNA (gDNA) instead of mRNA. The gDNA template produced a target/non-target ratio of 0.96 (= 2132 / 2210), enabling the raw allelic imbalance to be corrected to 0.70 (= 0.67 / 0.96) or -0.51 in $\log_2$ scale. This ratio implied that in liver, a tissue expressing miR-122, the mRNA abundance of the target allele is 70% of that of the non-target allele.

Figure 2 Impact of heterozygous target sites on mRNA allelic imbalance. (**a**) Distribution of allelic ratios, $\log_2$ (target/non-target), measured for miR-122 polymorphic sites using mRNA from either liver (orange, $n = 28$) or muscle (green, $n = 22$), plotted as a histogram

(left, 0.2-unit bins) and cumulative distribution (right). (**b**) Distribution of allelic ratios measured for miR-1 and miR-133 polymorphic sites using mRNA from either liver (orange, $n = 25$) or muscle (green, $n = 39$), plotted as in panel **a**. (**c**) Distribution of allelic ratios, pooling ratios from panels **a** and **b,** measured using either mRNA from the tissue expressing the cognate miRNA (red, $n = 67$), or mRNA from the tissue not expressing the cognate miRNA (blue, $n = 47$). In the inset are the cumulative distributions, plotted as in panels **a** and **b**. Below the histogram is the mean offset from zero for the two distributions, with error bars indicating 95% confidence intervals (bootstrapping) for the mean, and the percentages indicating the average down-regulation of target alleles compared to non-target alleles. (**d**) Sources of variability in allelic ratios, depicted with standard (left, 0.2-unit bins) and cumulative (right) distributions. Total variability not attributable to the cognate miRNAs was measured using mRNA with heterozygous sites not predicted to be regulated by the cognate miRNAs ($n = 72$). Of the 72 ratios determined, 47 were from mRNAs of tissues lacking the cognate miRNA, and 25 were from mRNAs without predicted potential for allele-specific repression mediated by the three miRNAs. (These 72 ratios were not normalized to corresponding gDNA ratios.) PCR variability was measured using gDNA ($n = 90$). Stochastic counting error was simulated using the binomial model ($n = 9{,}000$, 100

simulations per gDNA measurement), with total counts for each simulated amplicon chosen to match those of the gDNA measurements. The differences between each of the three possible pairs of distributions were statistically significant ($P < 0.01$ for each comparison, two-sided KS test).

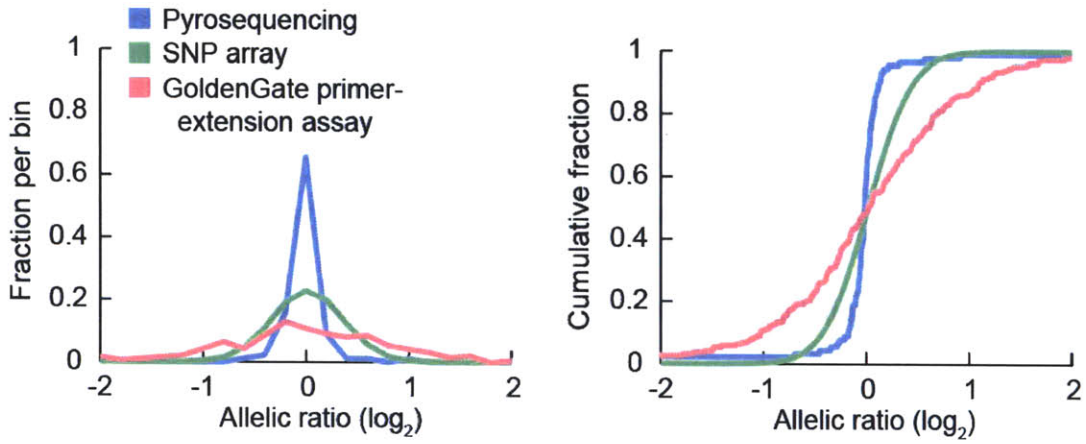
Figure 3 Dependence of target site efficacy on site type and context. (**a**) Efficacy of target sites of different types (8mer, $n = 5$; 7mer-m8, $n = 26$; 7mer-A1, $n = 36$), plotted as in Figure **2c**. (**b**) Relationship between allelic ratio and context score in the tissue expressing (left) and not expressing (right) the miRNA. Lines are the least-square fit to the data (r, Pearson correlation coefficient, with P-value estimated by two-sided Pearson correlation test). When considering only the 47 sites in the left panel that were measured also in the absence of the cognate miRNA (right panel), the correlation remained significant ($r = 0.58$ and $P < 10^{-4}$). (**c**) Relationship between adjusted allelic ratio and adjusted context score for 7-nt sites, plotted as in panel **b**. Context score and log ratio were each adjusted by subtracting the portion contributed by site type and the mean log ratio of sites of the same type, respectively. When considering only the 44 sites in the left panel that were measured

also in the absence of the cognate miRNA (right panel), the correlation remained significant ($r = 0.53$ and $P < 10^{-3}$).

SUPPLEMENTARY DISCUSSION

To start to estimate the impact of polymorphic miRNA target sites on gene-regulatory diversity in human, 15 unrelated individuals were randomly chosen from the HapMap individuals with Utah, USA origin (CEU). Genotype information was available for ~3.9 million common SNPs in these individuals (as of Feb, 2009). Comparing all possible pairs of the 15 individuals revealed that any two individuals have on average 1,225 genes (bottom and top 2.5 percentile, 1,154–1,272; median, 1,226) that are differentially regulated by at least one polymorphic target site for either one of the 87 broadly conserved miRNA families. The estimate would increase if rare SNPs and mammalian- or species-specific miRNA families were also considered. Assuming that properties of polymorphic target sites of human are comparable to those of mouse, between most human individuals, over a hundred genes are likely to be differentially regulated due to polymorphic miRNA targeting.

SUPPLEMENTARY FIGURE



Supplementary Figure 1 Noise of various allelic-imbalance measurement methods. Allelic ratio measurements are compared for gDNA, which has an allelic ratio of 1:1, corresponding to 0.0 on a $\log_2$ scale. Standard (left, 0.2-unit bin) and cumulative (right) distribution of allelic ratios, measured using pyrosequencing (n = 90), the Affymetrix SNP array (n = 73,944), and the Illumina GoldenGate assay (n = 187).

CHAPTER IV

Screening with Comprehensive MicroRNA Library

The work described in this chapter was a collaborative effort between Su Wu, Wenjun Guo, Jen Greiner and myself. Specifically, Su conducted a part of the induced pluripotent stem cell screen experiment. Wenjun conducted a part of the cancer stem cell screen experiment. Jen generated the high-throughput sequencing data from the cells infected with ~200 arbitrary shRNAs. All the other experiments and analyses were conducted by myself.

ABSTRACT: Due to the presence of at least two distinct mode of target recognition by ~22-nt small RNAs, any small interfering RNAs (siRNAs) for the silencing of a specific gene, are bound to have the side effects from the unintended repression of numerous genes that accidentally contain the motifs matching to the seed region of the siRNAs. The side-effects, called the off-target effects, have been frequently found responsible for the phenotypes that were not accounted for by the on-target effects. This motivated us to systematically exploit such off-target effects with the purpose of inducing given phenotypes. We constructed a pooled library of miRNAs covering all possible ~16,000 seed sequences. In constructing the library, the lentiviral short hairpin RNA (shRNA) system—routinely used for siRNA expression—was optimized for miRNA expression, such that the 87% (from baseline of 38%) of the mature products would have the defined seed sequences. Our initial trials of the library for the induced pluripotent stem cells and the cancer stem cells, although not succeeded in yielding any validated hits, revealed technical issues needed to be addressed for the potential success of our strategy. Our yet to be demonstrated strategy might provide insights into the phenotypic spectrums conferrable by the miRNAs spontaneously emerging during the evolutionary history.

Introduction

RNA interference (RNAi) is a phenomenon in which introduction of double-stranded RNAs (dsRNAs) into cells lead to the silencing of the gene that has extensive homology to the RNAs (Fire et al., 1998). Once in the cell, the dsRNAs are subject to Dicer-dependent processing into ~22-nt small RNAs, called small interfering RNAs (siRNAs) (Zamore et al., 2000; Elbashir et al., 2001). The siRNAs are then loaded into the RNA-induced silencing complex (RISC) to direct the endonucleolytic cleavage of the mRNAs with an extensive complementarity to the siRNAs, resulting in the silencing of the corresponding gene (Hannon, 2002). Recognition of this pathway allowed the widespread use of siRNAs as tools to study function of individual genes in cultured cells or other model systems.

Commonly encountered during experiments utilizing siRNAs are the phenotypes that cannot be accounted for by the repression of the intended target gene (Alvarez et al., 2006; Fedorov et al., 2006; Ali et al., 2009). These so-called off-target effects are most likely from the siRNAs acting like typical endogenous microRNAs (miRNAs), repressing numerous genes that accidentally contain the 3'UTR sequences matching to the ~7-nt seed region of the siRNAs (Birmingham et al., 2006; Bartel, 2009). Although the repression for each seed-matched target might be subtle, the subtle repression, coming in a sheer number,

may have a profound impact on the transcriptome. The impact may have adverse consequences, preventing the cells from manifesting the phenotypes of interest. In practices, however, off-target effects in many cases do not impair the cells from manifesting phenotypes, as siRNAs have been successfully used in numerous functional studies. Rather, the off-target effects have often been found responsible for interesting phenotypes (Alvarez et al., 2006; Fedorov et al., 2006; Ali et al., 2009).

This motivated us to systematically exploit the off-target effects with the purpose of inducing useful and/or interesting phenotypes. To do so, we constructed a library of miRNAs covering the entire 4^7 or ~16,000 possible 7-mers as the seed sequences. Under the assumption that the off-target effects are defined by the identity of the seed sequence in the siRNAs, the library would provide us with the unique opportunity to explore all possible off-targeting scenarios. Our strategy might also provide insights into the spectrums of the selective advantages or disadvantages conferrable by the spontaneously emerging miRNAs during the evolutionary history. Here, we describe (1) how the small RNAs expression system was developed, (2) how the library was constructed, and (3) how the library has been used for screening.

Analysis of small RNA expression system

In order to choose a system for the miRNA expression, we started by looking into one of the most commonly used systems for siRNA expression in cultured cells; the system is the lentivirus-based system developed by the RNAi Consortium (Moffat et al., 2006). The use of lentivirus in this system allows an efficient delivery of genetic materials into virtually any types of mammalian cells. Moreover, the stable integration of the viruses into the host genomic DNA (gDNA) allows permanent tagging of the infected cells with the virus, which could come useful for pooled screens. Upon the viral integration, polymerase III (pol III)-driven transcription generates transcripts with the 5' and 3' ends precisely defined by the transcription start and end sites (Figure 1). The transcripts, called the small hairpin RNAs (shRNAs), are designed to mimic the product of Drosha-processing of endogenous miRNAs (Paddison et al., 2002), which allows the shRNAs to bypass the Drosha-processing. Once the shRNAs are exported out of the nucleus through Exportin-5, they become substrates of Dicer, which cuts the loop part away from the hairpin, producing small RNA duplexes (Kim et al., 2009). The selection of the strand to be loaded into RISC is known to be influenced by multiple features, including the balance between the thermodynamic stability at the two ends of the duplex (Khvorova et al., 2003; Schwarz et

al., 2003), as well as the nucleotide identity at the 5' end of the loaded RNA (Mi et al., 2008). Nonetheless, the preferential loading strand of the shRNA-derived duplexes was empirically annotated, by the RNAi consortium, to correspond to the 3' arm of the shRNAs (Figure 1).

Identification of determinants for processing and loading

Because this system was originally developed for siRNA expression, the precise definition of the Dicer cutting site is not of concern, because any shift in the cutting site by a few nucleotides would not change the extensiveness of the complementarity between the mature products and the target genes. However, in adopting this system for expression of miRNAs, the precise definition of the Dicer cutting site is of paramount importance because any shift in the cutting site, and resulting change in 5' end of the mature products will lead to expression of the mature products with unintended seed sequences (Figure 1).

Therefore, we set out to examine the Dicer processing profile by expressing ~200 shRNAs with arbitrary stem sequences in arbitrary human cell lines. High-throughput sequencing of the small RNAs from the infected cells yielded ~170,000 reads mappable to the 200 shRNAs, with the length distribution, peaking at ~22-nt, as expected (Figure 2a).

However, unexpected was the distribution of the reads on the hairpin (Figure 2b), which showed that less than 80% of the mature products were derived from the annotated loading strand, indicating imprecision in the loading strand selection. Dicer cutting site was also found heterogeneous with the most prominent site, as depicted in Figure 2b, being utilized only ~50% of the times.

With the hope that certain features in the stem region of the hairpin might help enhancing the precision of the processing and loading, we examined the 200 shRNAs for any sequence features correlated with the precision of the processing and loading. We first defined the "precision" of loading/processing as the fraction of reads that map to the hairpin position 33, with respect to the 5' of the reads. For example, for the shRNA that showed the best defined processing/loading among the 200 shRNAs, the precision was 96% (Figure 4a). When we examined the relationship between the precision with the nucleotide identity at each of the 21 stem positions (Figure 3a), some positions seemed to have much more influence on the precision. Not surprisingly, the most influential one was the position 33, which corresponds to the 5' end of the intended mature products; nucleotide identity at this position might have affected both the cleavage and loading. Notably, the rank order of nucleotide preference observed at this position (U most preferred, G least preferred)

matched the rank order of the frequency of the nucleotide identity observed at the 5' end of endogenous miRNAs (data not shown). Other influential positions include position 32 and a bit less prominently position 30, 34, and 50. With the exception of position 34, the region that corresponds to the intended seed region of the intended mature product (positions 34 through 40) seemed to have relatively small influence on the precision.

Optimization of shRNA backbones for miRNA expression

Although the analysis of sequence determinants was informative, even after combining the three features—found most highly correlated with the precision—was projected to enhance the precision only up to 64% (Figure 3b). We therefore resorted to an alternative empirical approach where 52 candidate hairpins were designed based on the 13 hairpins that showed the best precision among the 200 tested hairpins; the 13 were modified into 4 different ways in term of the primary sequence and base-pairing status, ending up with the 52. The 52 candidates were tested for their robustness in the precision against the change in the sequence within the seed region. To do so, we constructed a small scale pilot library with complexity 6,500 (52 hairpins, each with 125 arbitrary seed sequences). The library was constructed in a pooled format, starting from the multiplexed in-situ DNA synthesis on

array (*Agilent Technologies*), all the way to the pooled virus production (Figure 5). High-throughput sequencing analysis of the small RNAs derived from the cells infected with our pilot library allowed us to pick three hairpins that demonstrated precise processing of ~88%, on average (Figure 4b).

Construction of the Library

Following the same pooled cloning procedure used for the test library (Figure 5), we constructed the full-scale library, having all ~16,000 possible seed sequences represented in each of the three backbones, which make the total complexity of the library ~48,000. Three aspects of the constructed library were inspected for the quality control purposes. First, the expression level of the small RNAs derived from the library was examined. The expression level matters because too strong expression could be toxic, and too weak expression could be inconsequential. A small RNA sequencing analysis with the library-infected cells showed that the library-derived small RNAs constitute 38% of the total small RNAs in the cells. Considering that the cells were infected at a high MOI, it is unclear how the expression level would be like at MOI of 1. Nonetheless, the fact that the library-derived small RNAs did not swamp the cellular small RNA pool even at the high MOI was

informative. Second, the precision of the processing and loading was examined. Even with the full 16,000 possible 7-mer sequences represented at the seed-region of the hairpins, the precision of processing and loading remained high, as ~87% of the mature products derived from the three hairpin backbones had the intended seed sequences. Third, we inspected whether the complexity of the library was preserved through the multi-step library construction process. Although the theoretical full complexity of the library is ~48,000, the effective complexity comes down to ~46,000 because (1) some hairpins contain restriction sites used for library construction, and (2) some contain premature Pol III termination sites (4 or more consecutive Us). Our small RNA sequencing analysis indicated that among the 46,000, at least 79% were expressed in of the library-infected cells. Moreover, at least 96% of all possible seeds were found to be expressed in context of either one the three backbones. Considering that the sequencing analysis was a diagnostic small-scale sequencing, thus with limited power to robustly capture the complex pool of the small RNAs in the cell, most possible seed sequences were considered fairly represented in the library.

A Screen for the Induction of Pluripotency

The library was applied in a screen for the induction of pluripotency. In the pioneering work by Yamanaka and colleagues, overexpression of four transcription factors—Oct4, Sox2, Klf4, and Myc—into terminally differentiated cells were shown to be able to reprogram them into the induced pluripotent stem (iPS) cells, which are functionally indistinguishable from the embryonic stem (ES) cells (Takahashi and Yamanaka, 2006). The suppression of p53, on top of the overexpression of the four genes, was later shown to enhance the efficiency of the reprogramming process (Hanna et al., 2009; Krizhanovsky and Lowe, 2009). This technology opened up possibilities for human disease modeling and patient-specific tissue engineering, by obliterating the ethical issues associated with the requirement of human eggs for the production of pluripotent stem cells. However, the remaining concerns on the use of iPSCs for therapeutic purposes include the oncogenecity associated with the use of oncogenes, especially Myc, and the gDNA-integrating viruses. The reagents that can at least partially replace Myc, such as valporic acid (Huangfu et al., 2008) or miR-294 (Judson et al., 2009), have been reported. Also reported were methods that can replace integrating viruses, such as methods based on DNA vectors, mRNA, or proteins (Stadtfeld and Hochedlinger, 2010). Since miRNAs are easily deliverable into cells through transfection without concerns on the gDNA-integration, any discovery of

miRNAs that can promote iPS cell generation may have a practical value. Moreover, analyzing target genes of the miRNAs that may come out of the screen might provide some insights on the mechanisms of the reprogramming.

A trial round of the screen was aimed for replacement of Myc in the induction of pluripotency from mouse embryonic fibroblasts (MEFs). The first step in our screen was to provide MEFs with a condition that would be suboptimal for the reprogramming. The condition was to provide the three out of the four factors other than Myc (Oct4, Sox2, and Klf4; OSK), in the background of the p53 suppression for the enhanced kinetics of the reprogramming process (Hanna et al., 2009). The subsequent infection of our pooled library to these MEFs is supposed to provide the cells with the variation in gene expression, through the expression of miRNAs with arbitrary seed sequences. If any of the miRNAs could rescue the lack of Myc, the cells infected with the miRNA will be able to form iPSC colonies within 4—6 weeks. The colonies could be isolated away from MEFs, by physical methods (such as pipeting) or by using the inherent difference in the surface attachment properties between iPSCs and MEFs (Methods). In the case the MEFs had fluorescence- or drug-resistance-reporter trangenes that get turned on in the pluripotent state, FACS or drug selection could also be used for the separation of the iPSCs.

This strategy was implemented in two different systems (Figure 6a). The first one is called "secondary" system (Markoulaki et al., 2009), in which the MEFs with the doxycycline (dox)-inducible transgenic OSK were infected with p53-shRNA (Ventura et al., 2004) and the library. The second one is called "single-cassette" system (Sommer et al., 2009), in which p53-null MEFs were infected with one kind of virus with all three dox-inducible OSK in the same polycistronic transcript. Even though the two systems implement a similar genetic condition, it was found that the secondary system has a low background and the single-cassette system has a high background, meaning that in a negative control condition where an empty virus was infected instead of library, the secondary system barely gives rise to any colony, whereas the single-cassette system gives rise to many colonies. The reasons why the single-cassette system is more efficient than the secondary system could include (1) compared to shRNA-mediated p53 knock-down, genetic p53 knock-out provides not only a clean ablation of p53, but also longer-history of p53 loss, which might be epigenetically reflected in the cells, and (2) compared to OSK transgenes, the OSK single-cassette virus can provide superior overexpression through the integration of multiple copies.

In the high background system, numerous emerged iPSC colonies were separated from MEFs by using their inherent difference in surface attachment properties (Methods). Using the gDNA purified from the separated iPSCs as template, miRNA-viruses were PCR-amplified and subject to high-throughput sequencing, in order to measure the representation of the library in the iPSCs. The library representation in the iPSCs was compared to the representation in the fresh infected MEFs measured in the same way, with the hope that the library miRNAs that promote reprogramming, if there is any, would have been increase in representation in the iPSCs compared to the MEFs. A lot of miRNAs were found to have been increased their representation during the process. However, the observed fluctuation in representation might have originated not from biological selection, but from the experimental variability. To estimate how much of the fluctuation is from the experimental variability, we took advantage of the non-functional shRNAs in the library. The non-functional shRNAs are shRNAs in the library that happen to contain the pre-mature pol-III transcriptional termination site, which leads to abortive transcription of these shRNAs. In fact, the shRNAs with 4 or more Us produce at least 4 fold less mature products than the other shRNAs (Figure 6b). Considering that some of the shRNAs with 4 or more Us still give rise to a substantial amount of mature products, a shRNA was

categorized as non-functional only when the shRNA qualify both of the following conditions: (1) the sequence contain 4 or more Us, (2) the small RNA sequencing detect only 1 or less read mapping to the hairpin. When the representational changes were plotted for the non-functional shRNAs, the distribution spanned >1,000 folds, indicating a high degree of the experimental variability (Figure 7a). When the representational changes were similarly plotted for the functional shRNAs, the distribution was markedly skewed to the left side compared to that of the non-functional shRNAs; the left-skewing suggested that the functional shRNA might have been negatively selected during the reprogramming process. Interestingly, despite the general left-skewing of the functional shRNA distribution, the distribution had a heavier right-tail to the right, suggesting possible positive selection of some functional shRNAs. The quantile-quantile plot plotted between the two distributions showed that the top two percent most enriched functional shRNAs were more so than were the corresponding top two percent of the non-functional shRNAs (Figure 7c). Despite the promising signs suggestive of the positive and negative selection, we could not find any significant statistical evidence of the seed-dependence of selection: the selections on the functional shRNAs sharing the same seed sequence were not found significantly correlated compared to those on the non-functional shRNAs (data not shown).

In the low-background system, only five iPSC colonies emerged; they were physically isolated by pipeting. The viral-inserts in the gDNAs were PCR-amplified, barcoded, and subjected to high-throughput sequencing (Figure 8). Many shRNAs were recovered from each colony, presumably due to reasons including (1) each cell was infected with more than one shRNA-virus, and (2) during the colony picking process, surrounding MEFs might have been contaminated. For example, a seed CTTCAAA accounted for 72% of all the sequencing reads obtained from the first colony, and this seed, when cross-referenced against the results from the high-background system, corresponded to the top 6 percentile in terms of the enrichment. Substantially many of the recovered seeds corresponded to the top 2 percentile in the high-background system results, indicating conformity of the two results. Most remarkably, one seed GTAAATT was recovered in three colonies, making the shRNAs with this seed the top candidate for validation.

The presence of the recurrently recovered seed in the low background experiment and the conformity between the high- and low-background experiments looked promising. On the other hand, concerning were the large experimental noise and the lack of detectable seed dependence in enrichment/depletion observed in the high background experiment. Nonetheless, we decided to proceed to the individual validation of nine shRNAs that stood

out in both experiments. However, in two rounds of validation experiments—where each of the selected candidates were individually infected and monitored for the efficiency of iPSC colony formation, none showed reproducible effects (data not shown).

Discussion

The fact that the iPSC screen failed to yield any validated hit is disappointing. A potential positive control, miR-294—the miRNA that was previously reported to be able to partially replace Myc in the induction of pluripotency (Judson et al., 2009)—also failed to stand out in the screen. One possible explanation could be that the miR-294 overexpression is redundant with the p53 suppression in our experiment. This explanation is supported by the study that showed the human counterparts of miR-294, namely miR-372/373, suppress the p53 pathway in the testes (Voorhoeve et al., 2006).

Beside the iPSC screen, another screen for the induction of cancer-stem-cell-like phenotype also failed to yield any hit (data not shown). Taken together, the most likely explanation for the failures of the screens is that potential technical problems in the implementation of the screen have diminished the sensitivity and the specificity of the screens. The possible technical problems include (1) the weak efficacy of the shRNA-

derived small RNAs in mediating seed-based repression, (2) toxicity, which could be associated with the high MOI for the library-viruses, (3) the high frequency of the background-phenotype formation, (4) a high degree of noise, which is introduced during the manipulation and measurements.

After these technical issues are carefully addressed, the library should be applied to a variety of different phenotypes in a diverse range of settings, in order to hedge against the bad luck of choosing a phenotype or an experimental setting that is not poised to exploit the repertoire of potential biological activities conferred by this library. The detailed discussion on how to follow-up this project in the future is described in Chapter V.

Methods

Sequencing analysis of the small RNAs from the cells infected with ~200 arbitrary shRNAs

The lentiviral vector with the shRNA-cassette (pLKO.1) and other vectors necessary for viral packaging (psPAX2, pMD2.G) were obtained from the RNAi Consortium (TRC) and *addgene.org*. Subcloning was conducted to generate ~200 pLKO.1-based vectors encoding shRNAs with the gene specific sequences for an arbitrarily chosen set of genes. The 200 vectors were used to generate 200 lentiviruses, each of which was infected into four cancer cell lines (AS49, MCF7, Jurkat, U937). The viral production and infection were conducted according to the protocol available at the TRC website. The infected cells were combined for each cell line to generate 4 pools. The total RNAs isolated from each pool were subjected to the library preparation procedure for the *Illumina* small RNA sequencing described previously (Chiang et al., 2010). Four runs of sequencing, one for each pool, produced ~11 million reads, ~6 million of which passed our quality control. Of the 6 million, 170K mapped to the original 200 shRNA sequences.

MicroRNA library construction

The pLKO.1 vector DNA was digested using the following condition: 200 µl reaction, 15 µg DNA, AgeI 8 µl (40 units), EcoRI 8 µl (160 units), 37°C 2 hrs. After the agarose-gel purification of the cut fragment (taking a special care not to expose DNA to UV light), the purified DNA fragment was subjected to dephosphorylation reaction using the following condition: 200 µl reaction, 1.5 µg DNA, 0.5 µl CIP (5 units), 37 °C 5 min. Immediately after the completion of the reaction, the DNA was cleaned-up (*QIAgen* QIAquick PCR purification kit, referred to as QIAquick kit hereafter).

For the preparation of the inserts (the DNAs to be inserted into the pLKO.1 vector), an order was submitted to *Agilent Technologies* for the pooled synthesis of 55,000 93-nt oligonucleotides with the hairpin structure, flanked by Age I and EcoRI sites (Figure 1a), as well as the primer binding sites for PCR amplification. (An example sequence: CTCTGCCAGG CAAACACCGG TAAGCGTCTA ACGATCTGCT TCTCGAGAAG TAGATCGTTAG ACGCTTATTT TTGAATTCTG GTCCATGCGCAG.) 10 pmole of the pooled oligonucleotides, delivered as lyophilized powder, was dissolved in 200 µl of EB buffer, to final concentration of 0.05 µM. The DNA was PCR-amplified using the following condition: forward, CACCT CTGCC AGGCA AAC, reverse, AATCT GCGCA TGGAC CAG, 60 µl reaction, 3% DMSO, 0.2 pmole template, final 6 µM primer, Phusion

(*NEB*) polymerase, 55°C annealing, 7 cycles. After clean-up (*QIAgen* MinElute PCR purification kit, referred MinElute kit hereafter), the PCR products were digested using the following condition: 100 µl reaction, the entire purified PCR products as the template, AgeI 5 µl (25 units), EcoRI 2.5 µl (50 units), 37°C 10 hrs. The digested PCR products were subjected to a final round of clean up (MinElute kit).

The prepared vector and inserts were ligated using the following condition: 100 µl reaction, vector DNA 250 ng, insert DNA 8.5 ng, T4 DNA ligase 5 µl (2000 units), 22.5 °C 30 min. The ligation reaction was subjected to EtOH precipitation using the following condition: 20 µg glycogen, 0.3M (final) NaCl, -20 °C 4 hr, 5 µl H_2O for resuspension. The 5 µl resuspended ligation product was subjected the bacterial transformation with 120 µl of electroporation competent cells (Stbl4, *Invitrogen*) using 6 cuvettes. A small scale plating of the transformed cells at varying dilutions showed ~27 million transformants were obtained in total. In order to amplify the library with minimal representational bias, the bacterial cells were plated on 16 extra-large LB-agarose dishes (500 cm^2, 100 µg/ml ampicillin) and incubated for ~20 hrs at 30 °C. The bacterial colonies were scraped off the plates for plasmid isolation using *QIAgen* Maxiprep.

iPSC experiments

The p53-null MEFs and the secondary MEFs (with dox-inducible OSK transgenes, rtTA transgene, Nanog-GFP reporter knock-in) (Markoulaki et al., 2009) were provided by the laboratory of Rudolf Jaenisch. Dox-inducible OSK-single-cassette lentivirus (Sommer et al., 2009), dox-inducible myc lentivirus (Wernig et al., 2008), rtTA lentivirus, and p53-shRNA lentivirus (Ventura et al., 2004) were also provided by Jaenisch lab. For the low background experiment, ~10 million secondary MEFs were first infected with the p53-shRNA viruses. For the high background experiment, ~10 million p53-null MEFs were first infected with rtTA viruses and OSK-single-cassette viruses. For both experiments, the library infection and dox-addition were started in ~1 week. The MEFs were cultured as previously described (Wernig et al., 2008; Sommer et al., 2009). Upon the observation of iPSC-like colonies, the colonies were isolated by pipeting for the low background experiments. For the high background system, colonies were isolated in a method based on the difference in the surface attachment property between MEFs and iPSCs. Specifically, the MEFs mixed with iPSC colonies were trypsinized and transferred to a new culture dish. 1—2 hr afterward, the cells in the supernatants—highly enriched with iPSCs due to their slow surface attachment kinetics—were harvested. In order to minimize contamination of

MEFs, the harvested cells were plated on irradiated MEF feeder cells and incubated for a few days. Once the plated iPSCs form colonies on the feeder cells, the cells were subjected to another one or two more rounds of similar separation, before the cells were taken for gDNA purification. Any MEFs mixed with iPSCs at this stage are likely irradiated feeder cells, which are not infected with the library viruses, and therefore the MEFs are supposed to be invisible to our gDNA-PCR reaction targeted toward the gDNA-inserts of the library viruses.

Sequencing analysis of gDNA-inserts of the library viruses

3' sequencing adapters (later to be ligated to gDNA PCR products) were prepared as follows. First, dsDNAs were generated by overlapping PCR (generation of primer di-mers) between the following two oligos: forward, GCACTCGAGG ACCATGATCG TCGGACTGTA GAACTCTGAACC, reverse, GCACTCGAGG ACGATCGTCG GACTGTAGAA CTCTGAACCT. The forward oligos contain the 3-nucleotide barcode "ATG", which prefix all sequencing reads with the code; the barcode was varied as needed. The PCR products were cleaned-up (MinElute kit), digested with XhoI, and cleaned-up (MinElute kit) again.

For gDNA-PCR, the forward primer was designed to contain 5' sequencing adapter, so that the gDNA-PCR products would have the 5' adapters, right out of the PCR reaction. The gDNA-PCR was conducted using the following condition: forward, CAAGCAGAAG ACGGCATACG AGGACTATCA TATGCTTAC, reverse, TTTGTCTCGA GGTCGAGAATTC, 400 µl reaction, 3% DMSO, 2 µg template, final 6 µM primer, Phusion (*NEB*) polymerase, 55°C annealing, 30 cycles. After clean-up (QIAquick kit), the PCR products were digested with XhoI and EcoRI for 3 hours, in order to disrupt the hairpin structure and to create a XhoI sticky site for 3' adapter ligation. After another clean-up (QIAquick kit), the products were ligated with the 3' adaptors using the following condition: 10 µl reaction, 7 ng 5' DNA fragment, 3 ng 3' adapter fragment, 0.5 µl (200 units) T4 DNA ligase, 22.5°C 30 min. The ligation reaction was subjected to the PCR amplification using the following primers: forward, CAAGCAGAAG ACGGCATA, reverse, AATGATACGG CGACCACC. The PCR products were agarose-gel purified before submitted for sequencing.

ACKNOWLEDGMENTS

We thank Rudolf Jaenisch, Robert Weinberg, and David Sabatini for helpful advices, Andrew Grimson and Wendy Johnston for technical assistances in generating the small RNA sequencing data from the cells infected with ~200 arbitrary shRNAs.

REFERENCES

Ali, N., Karlsson, C., Aspling, M., Hu, G., Hacohen, N., Scadden, D.T., and Larsson, J. (2009). Forward RNAi screens in primary human hematopoietic stem/progenitor cells. Blood *113*, 3690-3695.

Alvarez, V.A., Ridenour, D.A., and Sabatini, B.L. (2006). Retraction of synapses and dendritic spines induced by off-target effects of RNA interference. J Neurosci *26*, 7820-7825.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215-233.

Birmingham, A., Anderson, E.M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., Karpilow, J., *et al.* (2006). 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. Nat Methods *3*, 199-204.

Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., Auyeung, V.C., Spies, N., Baek, D., Johnston, W.K., Russ, C., Luo, S., Babiarz, J.E., *et al.* (2010). Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes Dev *24*, 992-1009.

Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494-498.

Fedorov, Y., Anderson, E.M., Birmingham, A., Reynolds, A., Karpilow, J., Robinson, K., Leake, D., Marshall, W.S., and Khvorova, A. (2006). Off-target effects by siRNA can induce toxic phenotype. RNA *12*, 1188-1196.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C.J., Creyghton, M.P., van Oudenaarden, A., and Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. Nature *462*, 595-601.

Hannon, G.J. (2002). RNA interference. Nature *418*, 244-251.

Huangfu, D., Maehr, R., Guo, W., Eijkelenboom, A., Snitow, M., Chen, A.E., and Melton, D.A. (2008). Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds. Nat Biotechnol *26*, 795-797.

Judson, R.L., Babiarz, J.E., Venere, M., and Blelloch, R. (2009). Embryonic stem cell-specific microRNAs promote induced pluripotency. Nat Biotechnol *27*, 459-461.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell *115*, 209-216.

Kim, V.N., Han, J., and Siomi, M.C. (2009). Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol *10*, 126-139.

Krizhanovsky, V., and Lowe, S.W. (2009). Stem cells: The promises and perils of p53. Nature *460*, 1085-1086.

Markoulaki, S., Hanna, J., Beard, C., Carey, B.W., Cheng, A.W., Lengner, C.J., Dausman, J.A., Fu, D., Gao, Q., Wu, S., *et al.* (2009). Transgenic mice with defined combinations of drug-inducible reprogramming factors. Nat Biotechnol *27*, 169-171.

Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Wu, L., Li, S., Zhou, H., Long, C., *et al.* (2008). Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. Cell *133*, 116-127.

Moffat, J., Grueneberg, D.A., Yang, X., Kim, S.Y., Kloepfer, A.M., Hinkle, G., Piqani, B., Eisenhaure, T.M., Luo, B., Grenier, J.K., *et al.* (2006). A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. Cell *124*, 1283-1298.

Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J., and Conklin, D.S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. Genes Dev *16*, 948-958.

Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell *115*, 199-208.

Sommer, C.A., Stadtfeld, M., Murphy, G.J., Hochedlinger, K., Kotton, D.N., and Mostoslavsky, G. (2009). Induced pluripotent stem cell generation using a single lentiviral stem cell cassette. Stem Cells *27*, 543-549.

Stadtfeld, M., and Hochedlinger, K. (2010). Induced pluripotency: history, mechanisms, and applications. Genes Dev *24*, 2239-2263.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663-676.

Ventura, A., Meissner, A., Dillon, C.P., McManus, M., Sharp, P.A., Van Parijs, L., Jaenisch, R., and Jacks, T. (2004). Cre-lox-regulated conditional RNA interference from transgenes. Proc Natl Acad Sci U S A *101*, 10380-10385.

Voorhoeve, P.M., le Sage, C., Schrier, M., Gillis, A.J., Stoop, H., Nagel, R., Liu, Y.P., van Duijse, J., Drost, J., Griekspoor, A., *et al.* (2006). A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. Cell *124*, 1169-1181.

Wernig, M., Lengner, C.J., Hanna, J., Lodato, M.A., Steine, E., Foreman, R., Staerk, J., Markoulaki, S., and Jaenisch, R. (2008). A drug-inducible transgenic system for direct reprogramming of multiple somatic cell types. Nat Biotechnol *26*, 916-924.

Zamore, P.D., Tuschl, T., Sharp, P.A., and Bartel, D.P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. Cell *101*, 25-33.

Figure 1

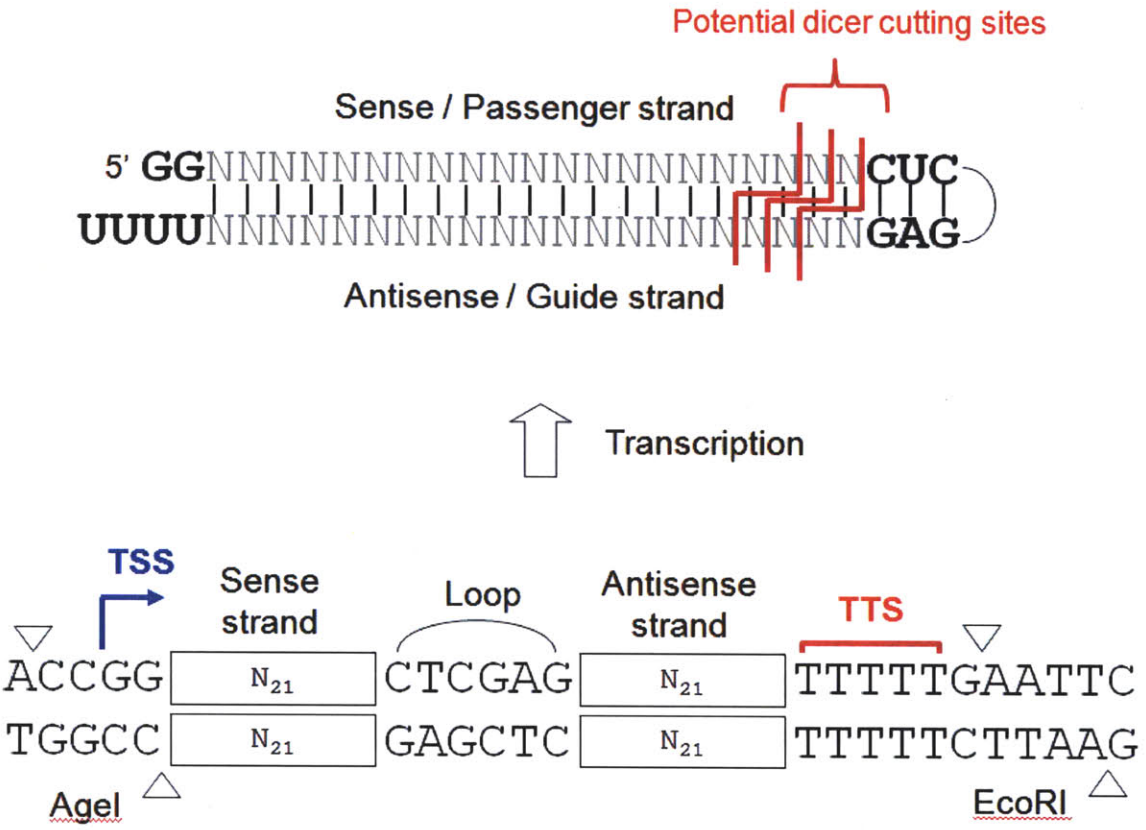

Potential dicer cutting sites
Sense / Passenger strand
5' GGNNNNNNNNNNNNNNNNNNNNNCUC
UUUUNNNNNNNNNNNNNNNNNNNNNGAG
Antisense / Guide strand
Transcription
TSS
Sense strand
Loop
Antisense strand
TTS
ACCGG
N21
CTCGAG
N21
TTTTTGAATTC
TGGCC
N21
GAGCTC
N21
TTTTTCTTAAG
AgeI
EcoRI

Figure 2

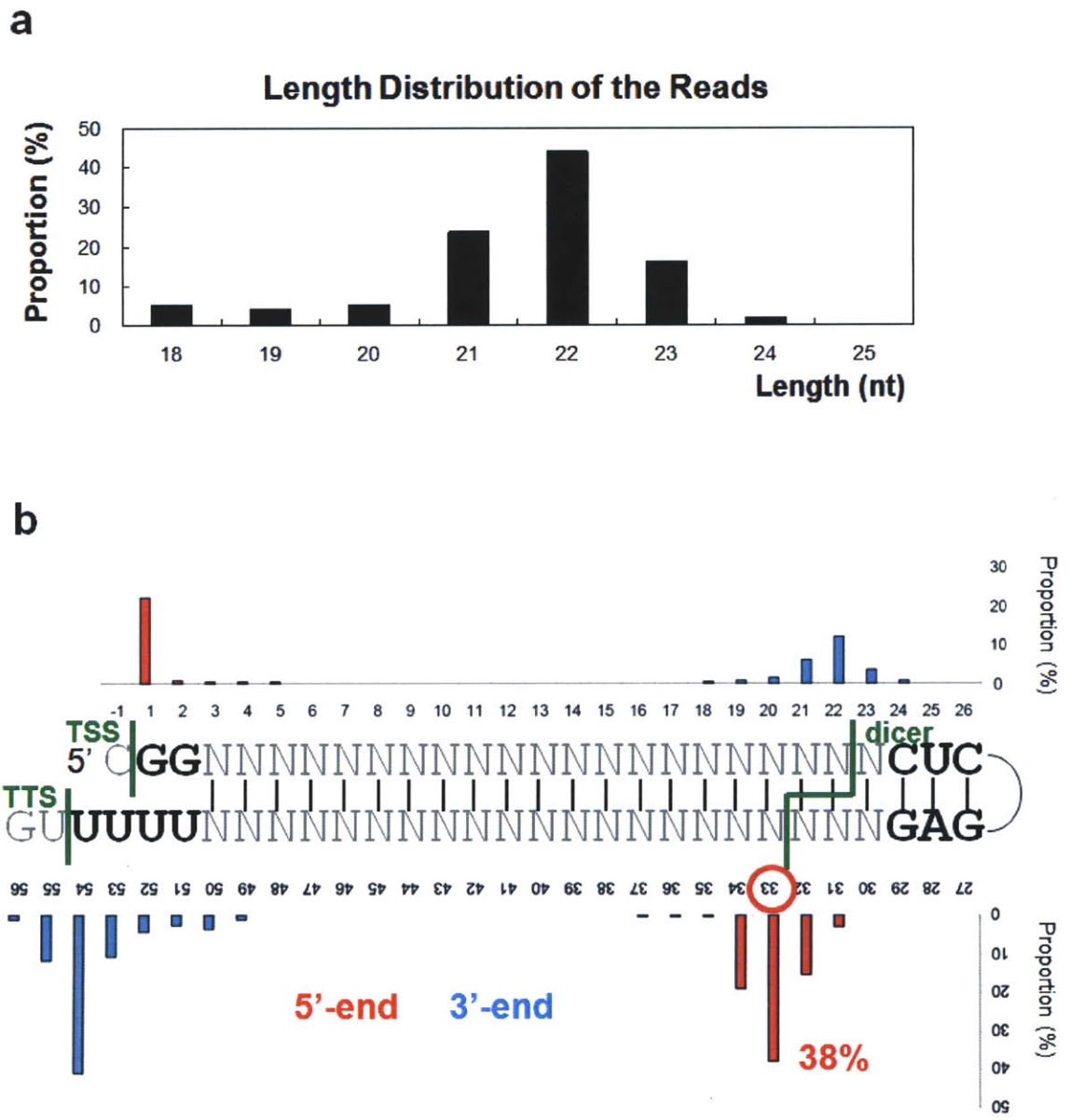
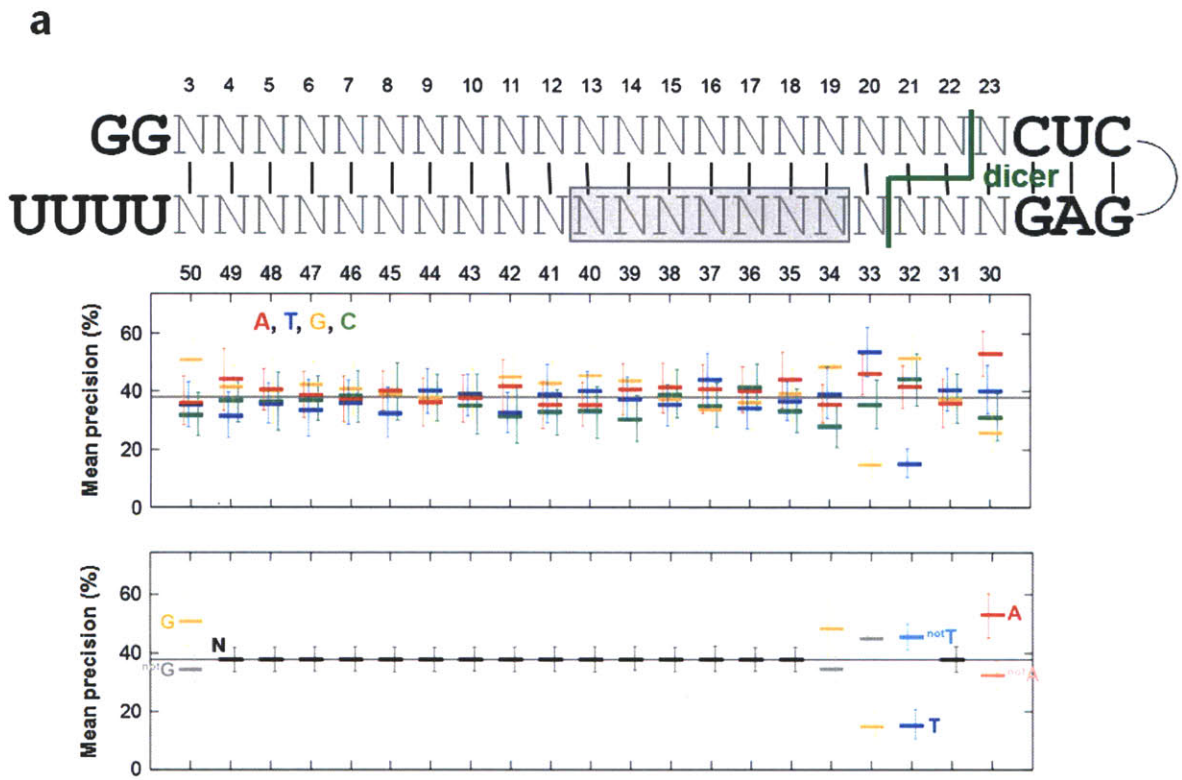

a
Length Distribution of the Reads
Proportion (%)
50
40
30
20
10
0
18
19
20
21
22
23
24
25
Length (nt)
b
Proportion (%)
30
20
10
0
-1 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
TSS
5' CGGNNNNNNNNNNNNNNNNNNNNNNCUC
dicer
TTS
GUUUUUNNNNNNNNNNNNNNNNNNNNNNGAG
56 55 54 53 52 51 50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27
5'-end
3'-end
38%
0
10
20
30
40
50
Proportion (%)

Figure 3


a
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
GGNNNNNNNNNNNNNNNNNNNNNNCUC
dicer
UUUUNNNNNNNNNNNNNNNNNNNGAG
50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30
Mean precision (%)
A, T, G, C
Mean precision (%)
G
N
notG
notT
A
T
notA

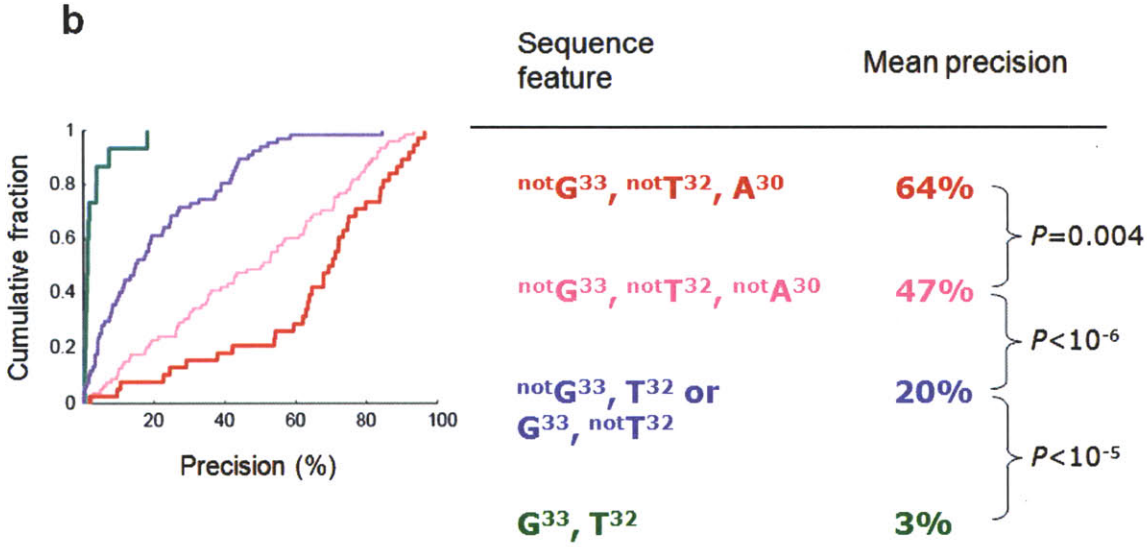

b
Cumulative fraction
Precision (%)
Sequence feature
Mean precision
notG33, notT32, A30
64%
notG33, notT32, notA30
47%
notG33, T32 or G33, notT32
20%
G33, T32
3%
P=0.004
P<10-6
P<10-5

Figure 4

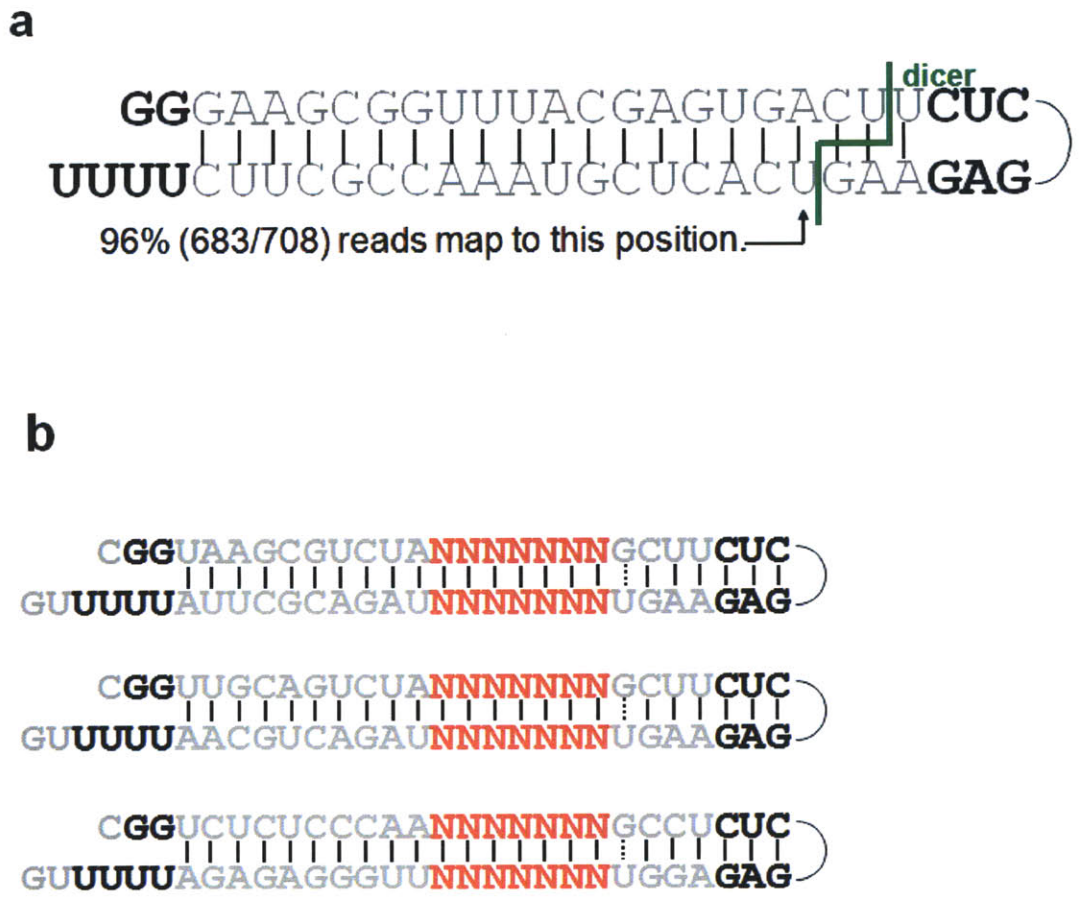

a
dicer
GGGAAGCGGUUUACGAGUGACUUCUC
UUUUCUUCGCCAAAUGCUCACUGAAGAG
96% (683/708) reads map to this position.
b
CGGUAAGCGUCUANNNNNNNGCUUCUC
GUUUUUAUUCGCAGAUNNNNNNNUGAAGAG
CGGUUGCAGUCUANNNNNNNGCUUCUC
GUUUUUAACGUCAGAUNNNNNNNUGAAGAG
CGGUCUCUCCCAANNNNNNNGCCUCUC
GUUUUUAGAGAGGGUUNNNNNNNUGGAGAG

Figure 5

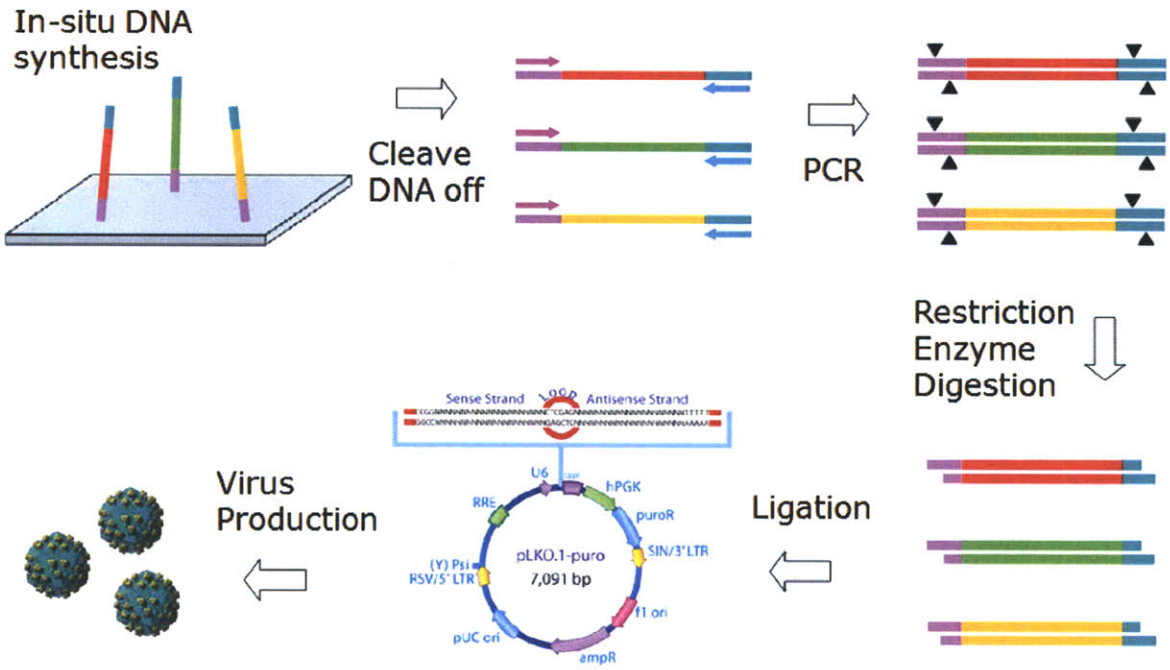

In-situ DNA synthesis
Cleave DNA off
PCR
Restriction Enzyme Digestion
Ligation
Virus Production
Sense Strand
Antisense Strand
U6
hPGK
puroR
SIN/3'LTR
f1 ori
ampR
pUC ori
(Y) Psi RSV/5' LTR
RRE
pLKO.1-puro
7,091 bp

Figure 6

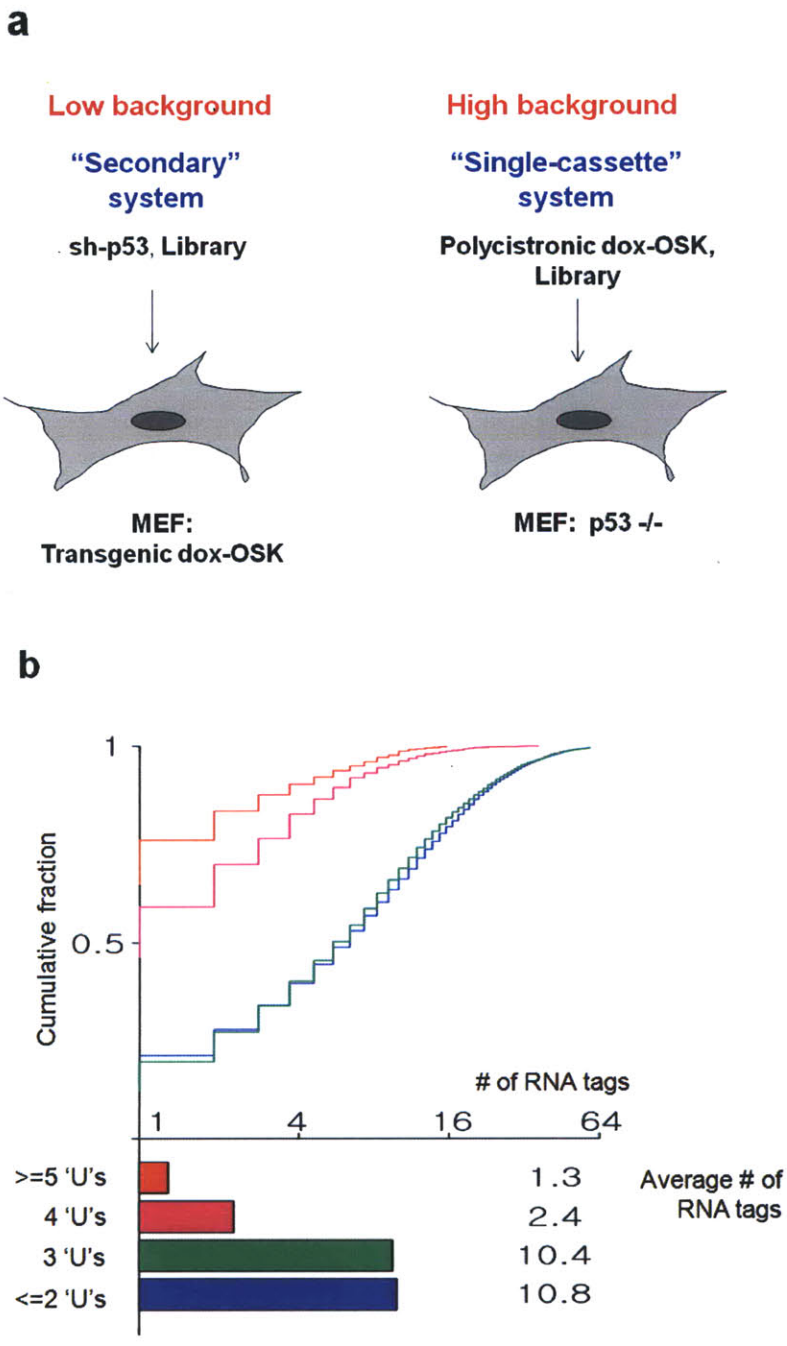
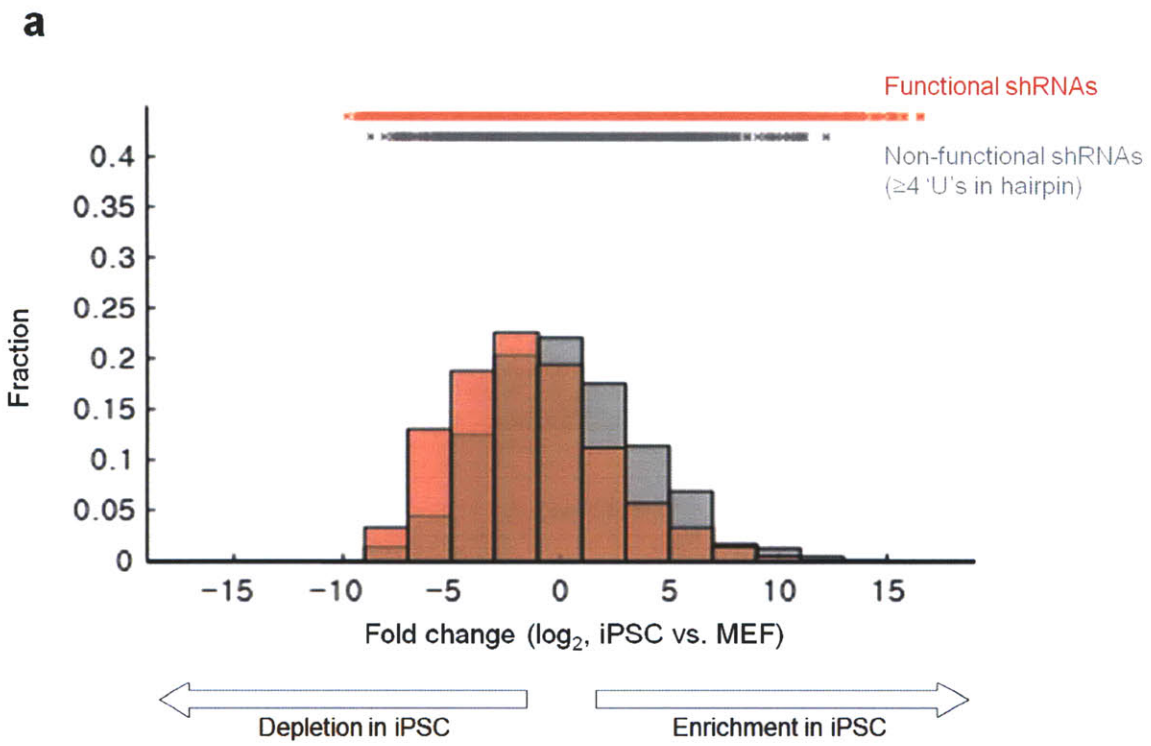

a
Low background
"Secondary" system
sh-p53, Library
MEF: Transgenic dox-OSK
High background
"Single-cassette" system
Polycistronic dox-OSK, Library
MEF: p53 -/-
b
Cumulative fraction
1
0.5
# of RNA tags
1
4
16
64
>=5 'U's
4 'U's
3 'U's
<=2 'U's
1.3
2.4
10.4
10.8
Average # of RNA tags

Figure 7


a
Functional shRNAs
Non-functional shRNAs
(≥4 'U's in hairpin)
Fraction
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−15
−10
−5
0
5
10
15
Fold change (log2, iPSC vs. MEF)
Depletion in iPSC
Enrichment in iPSC

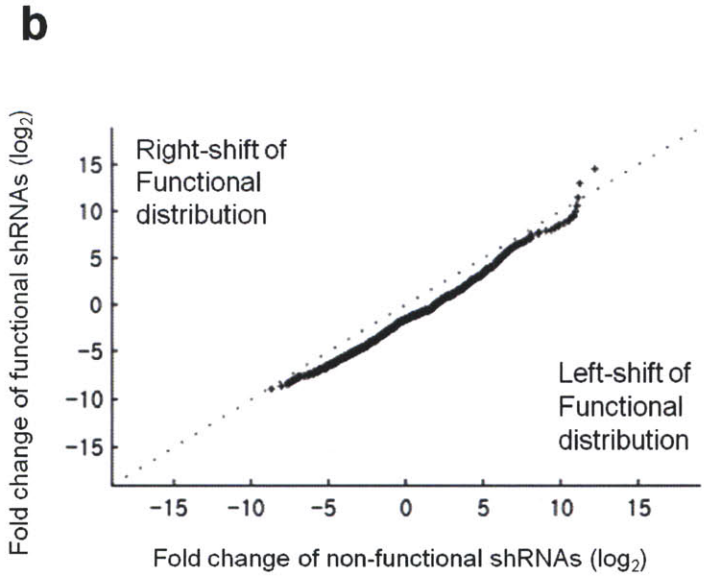

b
Fold change of functional shRNAs (log2)
Right-shift of
Functional
distribution
Left-shift of
Functional
distribution
15
10
5
0
−5
−10
−15
−15
−10
−5
0
5
10
15
Fold change of non-functional shRNAs (log2)

Figure 8

	Seed	Colony composition	Percentile (from top) in enrichment
Colony 1	CTTCAAA	72.1%	5.93
	ATACAAA	17.8%	39.62
	CACTATA	3.0%	13.23
	GTTGTTC	2.5%	99.17
	TAACTAA	1.9%	6.45
Colony 2	CCTATCT	29.8%	35.42
	GTAAATT	12.3%	0.06
	AGTTAAT	4.1%	17.16
	TGTGCCT	3.9%	0.44
	CGATTAA	2.4%	0.04
Colony 3	GTAAATT	22.3%	0.06
	GTATTAG	21.3%	83.27
	TGATGCC	13.2%	0.06
	AATTTTA	10.2%	87.61
	TCAGTAG	8.4%	2.00
Colony 4	CAGTACA	27.1%	96.21
	GATAATT	20.1%	0.23
	ATCCCTA	14.7%	21.60
	GGCCGAA	13.4%	83.69
	AGTTATT	7.8%	11.34
Colony 5	TTTCTAT	47.1%	26.30
	TTCATAA	8.3%	0.18
	CCTCAAA	6.3%	15.82
	GTAAATT	3.8%	0.06
	AACCTTC	3.7%	0.10

FIGURE LEGENDS

Figure 1. Schematics of shRNA expression. At the bottom is the part of pLKO.1 vector sequence encoding shRNA. At the top is the predicted secondary structure of shRNA transcripts. TSS, transcription start site. TTS, transcription termination site.

Figure 2. Analysis of the small RNAs from the cells infected with ~200 arbitrary shRNAs. (**a**) The length distribution of the sequencing reads. (**b**) The positional distribution of the sequencing reads on the hairpin.

Figure 3. Analysis of the determinants for the precision of the Dicer-processing and loading strand selection. (**a**) The influence of each nucleotide at each stem position on the precision. For each position, the 200 hairpins were divided into 4 groups, depending on which nucleotide (A,T,G or C) each hairpin has in the position. Then, for each one of the four subgroups, the mean value of the precision (the fraction of reads that map to the hairpin position 33) was calculated. For example, the 48 hairpins that have T in position 33 had the precision of 54%, whereas the 47 hairpins that have G had mean precision of only 15%. The grey box indicates the intended seed region of the mature products. (**b**) A model

predicting the precision of a given hairpin based on the sequence features. The P-values were calculated by the one-sided KS test.

Figure 4. shRNA-backbones for miRNA expression. (**a**) The shRNA sequence that showed the highest precision in processing/loading among the 200 tested shRNAs. (**b**) The three shRNA backbones used for the construction of the main library.

Figure 5. Schematics on the pooled library construction procedure.

Figure 6. Experimental settings of the iPSC screen. (**a**) Schematics on the two experimental systems used for iPSC experiments. (**b**) Expression of the mature products from the hairpins with a varying number of nucleotide "U"s. Bottom of the panel is the average number of the sequencing reads obtained for the indicated subgroups. Top of the panel is the cumulative distribution of each subgroup.

Figure 7. Enrichment and depletion of shRNAs in the iPSC screen. (**a**) Fold change distribution of the functional and non-functional shRNAs. The crosses at the top of the

panel indicate individual data points. (**b**) A Quantile-Quantile (QQ) plot between the fold changes of the functional and non-function shRNAs.

Figure 8. Analysis of the five iPSC colonies emerged from the low background screen. The library-virus was PCR amplified from the gDNA of each colony. The PCR products, after being uniquely barcoded, were subjected to high-throughput sequencing. Top five seeds recovered from each colony are shown. The top percentile of each seed in terms of the enrichment observed in the high background screen is also shown. Red hollow boxes indicate the recurrently recovered seed. Orange hollow boxes indicate the seeds that fell into the top 2 percentile in terms of the enrichment in high background screen.

CHAPTER V

Future Directions

I hereby discuss how the allelic imbalance study described in Chapter III and the comprehensive microRNA screening study described in Chapter IV can be followed up in the future.

Extension of the Allelic Imbalance Study

In Chapter III, we demonstrated the proof-of-principle that the analysis of the mRNA allelic imbalances with respect to the heterozygous variations in miRNA target sites can reveal in-vivo functionality of the target sites. Our approach can now be further extended into a much larger scale via the genomic technologies that have been developed for the last couple of years; the technologies include the padlock-probe-based sequence capture (Ball et al., 2009; Lee et al., 2009; Li et al., 2009a; Li et al., 2009b; Zhang et al., 2009) and deep mRNA sequencing. The gained scale will provide with the power to assess the functionality of a much wider range of cis-regulatory elements, not just restricted to miRNA target sites. Our approach can be taken into yet another level by adding the procedure of pulling-down RNA-binding proteins, like Argonautes, before subjecting the pulled-down RNAs to the allelic imbalance measurements. The added procedure will

enable more direct measurements of any allelic difference in physical binding between the regulatory elements in mRNAs and the RNA-binding proteins.

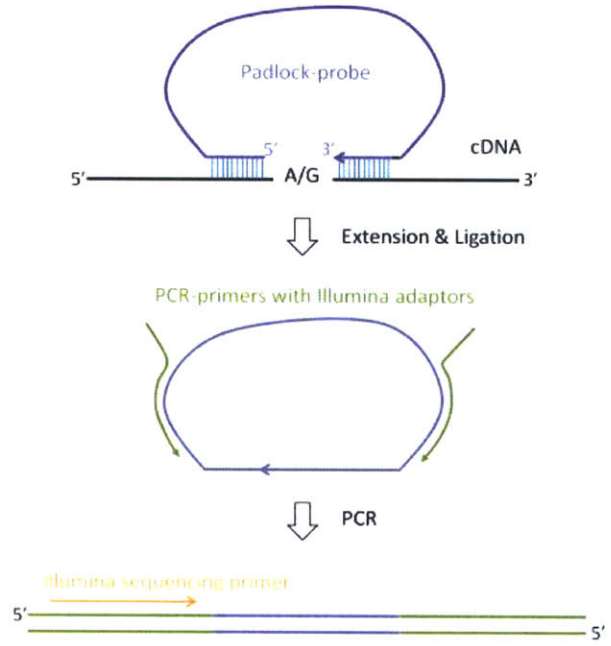


Figure 1. Padlock-probe-based sequence capture.

Scaling up mRNA allelic imbalance measurements by new genomic technologies

One recently developed technology that can increase the scale of the mRNA allelic imbalance measurements is the padlock-probe-based sequence capture (Figure 1), which is a technology for capturing up to ~55,000 specific regions of nucleic acids from a complex pool of nucleic acids, like gDNA or cDNA, for focused analysis of the selected regions. A

padlock-probe is a ~100-nt DNA oligonucleotide that docks at a specific DNA region of interest in a padlock-like shape (Nilsson et al., 1994). Specifically, each of the 5' and 3' ends of the probe can hybridize to each flanking region around the locus of interest. Then, the 3' end of the padlock probe can be made to extend through the addition of nucleotides by polymerases, thus "capturing" the sequence information of the locus of interest, until it reaches the 5' end of the same probe at the opposite side of the locus. Adding ligases after the polymerase reaction links the two ends of the probe, circularizing the probe. This reaction can be multiplexed for up to 55,000 regions in a single test tube, because a pool of 55,000 oligonucleotides can be manufactured by Agilent at a reasonable cost. In order to apply this multiplexed sequence-capture technology to measuring mRNA allelic imbalances, the probes would be designed to specifically target 55,000 selected heterozygous SNPs that fall into cis-regulatory elements of interest. The probes are used with cDNAs to capture not only the genotype information, but also the mRNA allelic imbalance information. The circularized probes can be subject to PCR amplification, followed by high-throughput sequencing to read the captured information. Church and colleagues have recently exploited this method to measure the mRNA allelic imbalances at a large number of loci (Lee et al., 2009; Zhang et al., 2009). However, their data were not

analyzed in a way to correlate the allelic imbalance with the heterozygous variations in cis-regulatory elements. Thus, a reanalysis of their now publicly available data might give us a quick indication on the potential success of this strategy. If looking promising, the strategy could be further expanded by doing the experiments both with our own custom designed set of padlock probes that focus on a specific set of regulatory loci of interests, and with a wider panel of tissues.

Another advance in genomic technologies that can help achieve a higher scale of the mRNA allelic imbalance measurements is the precipitous drop in the cost of sequencing per read. Since our allelic imbalance study using 454 sequencing technology was published, the cost has been reduced by many orders of magnitudes by new sequencing technologies like the Illumina HiSeq (*Illumina*) or SOLiD (*Life Technologies*) technologies. The trend is expected to be maintained for next several years. The decreased cost of sequencing reads is making it increasingly affordable to obtain mRNA-Seq data with a relatively high average coverage. With such a high coverage, a substantial part of the transcriptome would be covered with a large enough number of reads to allow an informative estimation of mRNA allelic imbalances. Considering that a large number of mRNA-Seq data are being deposited

to public databases at a rapid pace, the meta-analysis of the published datasets might also yield informative results.

Systematic analysis of cis-regulatory elements that may affect mRNA level

The gained scale will enable us to make not only more confident functionality assessment of each cis-regulatory motif, but also more comprehensive assessment of a much wider panel of cis-regulatory elements that might affect mRNA level. Such regulatory elements include (1) transcription factor binding sites, (2) RNA-binding protein binding sites, AU-rich elements, and poly-adenylation signals, and (3) splicing signals. For example, the binding sites for a transcription factor, Err-α, (motif: TGACCTTG) could be examined for the functionality in a given tissue. The promoter regions of all genes will be scanned for the SNPs that create or disrupt the motif. Because the site is not supposed to be in a transcribed region, measurement of mRNA allelic imbalance should rely on the "proxy" SNPs that are in the transcribed region and that are in linkage disequilibrium with the promoter SNP; The linkage disequilibrium between SNPs can be looked up on publicly available haplotype maps (Frazer et al., 2007). Analyzing the mRNA allelic imbalance in the individuals heterozygous for both the promoter and proxy SNPs will allow the inference

of the functionality of the promoter in the tissue. Similar analysis of all motifs that are known to affect mRNA level may enable "functionality profiling" of the cis-regulatory elements in a given tissue.

Combining the allelic imbalance approach with the pull-down of Argonaute proteins or other RNA-binding-proteins

Georges and colleagues recently reported a study where they pulled down Argonaute (AGO) proteins in the cells heterozygous for a specific miRNA target site, reverse transcribed the AGO-associated RNAs, and subjected the cDNA to Sanger-sequencing (Takeda et al., 2010). Analysis of the height of the electropherograms revealed the imbalance in the direct AGO-binding between two alleles. They found that the mRNA from the allele with the intact target site was preferentially associated with the AGO, than the other allele without the target site. The pull-down approach demonstrated by this study can be extended to utilize high-throughput sequencing. It can also be extended to other RNA-binding proteins. The additional layer of information on the direct binding will enable more confident assessment of the functionality of regulatory elements.

Follow-up of the Comprehensive MicroRNA Library Screen

In Chapter IV, I describe the construction of the comprehensive miRNA library, which, when applied to the two initial trial screens for the induced pluripotent stem cells (iPSCs) and the cancer stem cells (CSCs), have not yielded any positive hit. The study can be followed-up by trying to address the possible causes of the negative results. There are at least three categories of possibilities: (1) the technical issues at individual miRNA expression, (2) the technical issues related to a high degree of multiplexing in the pooled screen, (3) the issues regarding the intrinsic activities of the library with respect to the phenotypes of choice. These three categories of possibilities can be systematically addressed by a series of experiments.

Potential technical issues regarding individual miRNA expression

What might have caused the unsuccessful screen includes the possibility that the small RNAs generated from our shRNA-based expression system did not mediate the seed-based target repression. Our library, designed to have the maximum complexity at the seed region, would not confer a broad range of functionalities if the small RNAs acted through non-seed-based mechanisms. Back when the two trial screens were designed, it seemed

reasonable to assume that the shRNA-derived small RNAs are capable of the seed-mediated repression, considering that small RNA sequencing experiments showed an abundance of shRNA-derived small RNAs in the size range of typical miRNAs. Nonetheless, in light of the negative results from the initial pilot screens, it seems prudent to experimentally confirm the functionality of the small RNAs by performing luciferase assays with a few shRNAs chosen from our library. For example, we can use the luciferase vectors that were proven to work by previous studies, such as the vector with the 3'UTR of HMGA-2 containing the intact or mutant target sites for let-7 (Mayr et al., 2007). The cell lines without detectable let-7 expression, such as F9, can be used to assess the efficacy of the reporter repression mediated by our shRNA with the let-7 seed.

Potential technical issues regarding a high degree of multiplexing in the pooled screen

Another set of possibilities that might have contributed to the negative results of our trial screens include the stochastic noise that must have been introduced during the experiments. Due to a high complexity of the library, the stochastic noise is bound to be introduced at any manipulation of the library or library-infected cells, or at any measurement processes of the library representation.

The first point of entry for the stochastic noise is at the library infection step. For the iPSC screen, the number of the fibroblasts used at the infection is ~5 million, which is ~100 times bigger than the total number of components in the library; it means that, at a MOI of ~1, each library component would be infected into ~100 cells, which is equivalent to giving each component 100 chances to demonstrate its activity. The 100 chances per component, however, might not be enough considering that even with the use of Myc (the positive control) the efficiency of reprogramming is near low one digit percents (Stadtfeld and Hochedlinger, 2010). Therefore, in order to give more opportunities to each component without increasing the total number of cells, the decision was made to infected the cells with the pooled library at a MOI bigger than 1. However, the results (Figure 7a of Chapter IV) from the iPSC screen suggested that the high MOI infection might have rather decreased the effective scale of the experiments. In the results, a substantial fraction of the library components seemed depleted, presumably because of the diminished viability or proliferation of the host cells. No matter whether the negative effects are from the non-specific toxicity of shRNA-overexpression, or from the specific regulatory activity mediated by the infected miRNAs, the high MOI infection could exacerbate the negative effects, decreasing the number of the cells healthy enough to manifest the phenotype of

interest. The decrease in the effective population of cells would reduce the number of opportunities per each library component to demonstrate its activity. One experiment to check the degree of negative effects associated with shRNA-overexpression would be to infect varying doses of the library viruses into the cells that are already provided with all the necessary conditions for the phenotype formation, monitoring suppression of the phenotype formation at each dose of infection. Alternatively, the issue of shRNA-overexpression could be avoided in the first place by using a substantially larger number of cells for a sufficiently large coverage of the library even at the low MOI.

Second point of entry for the stochastic noise is at the separation of the cells that showed the phenotype of interest from the rest of the cells. In the iPSC screen, iPSCs were separated from MEFs based on the fact that the surface attachment process is slower for iPSCs than for MEFs. However, because the difference in the kinetics is not big enough for the perfect separation of the two populations, many iPSCs must have been lost, and many MEFs must have been contaminated, resulting in many false-positives and false-negatives. In case of the CSC screen, CSCs were separated by the surrogate marker-based cell sorting. The cell gating strategy that prioritized the minimal contamination of the non-CSCs might have resulted in many false-negatives.

Third point of entry for the stochastic noise is during the gDNA-PCR of the separated cells. A big part of the problem was our inability to scale up the gDNA-PCR reaction to the degree enough to accommodate up to 1 mg of the gDNA isolated from tens of millions cells—the number of cells obtained at the end of a typical screen. Because PCR reaction to amplify this much gDNA is too large (up to 50 ml reaction) to be handled efficiently, the reactions of 400 μl was conducted using only about one three hundredth of the harvested gDNA. This undersampling of the gDNA might have introduced noise in our estimation of the library representation. The noise is expected to be especially problematic in estimating the baseline library representation using the gDNA of the cells freshly infected with the library because the library complexity is the largest at this pre-selection stage. One potential solution to this issue is that the representation at the library vector or at the virus is used as the proxy for the representation at the freshly infected cells. However, this solution has disadvantages such as (1) that its use is limited to the pre-selection sample, and (2) that the proxy measurements are blind to any bias introduced during virus production or infection procedures. An alternative solution without these disadvantages would be to reduce the gDNA template complexity before subjecting them to PCR reaction, by purifying the viral-inserts from digested gDNAs. The purification of

viral-inserts could be achieved in two different methods. The first method is to capture the viral-inserts by using probes designed to hybridize with the inserts. The other method is to gel-purify the viral-inserts, based on the fact that the viral inserts generated from the enzyme digestion of gDNA are of a predictable size. The viral-insert purification method, either through the probes or the gel-purification, can reduce both the complexity and mass of the template, thereby enhancing a priming efficiency and also reducing the scale of PCR reactions.

Potential issues regarding intrinsic activities of the library for the phenotypes of choice

While the first two categories of potential issues were relevant to techniques, the last set of issues is more relevant to the inherent property of the library: Even in an ideal scenario in which all technical issues are resolved and the experimental noise is negligible, the screen might not be productive if the library simply has too little phenotype-inducing activity. If only a very minor fraction of the library has the activity, the frequency of background phenotype formation gravely affects the chance of detecting a true positive. The background can be estimated from control experiments. In the case of the assay used for our CSC screen, the empty-virus infected cells (the negative control) showed about 1%

background efficiency in converting non-CSCs into CSCs, whereas the slug-infected cells (the positive control) showed 100% efficiency. So, the slug gives a 100 fold boost over the background in the phenotype induction. Now, even under the optimistic assumption that 500 miRNAs (~1%) in our library are as active as slug, as many inactive miRNAs would end up in CSCs, through background, because the inactive miRNAs in the library outnumber the active miRNAs by 100 fold; the false discovery rate (FDR) in this case would be 50%. Under the more realistic assumption that 50 (~0.1%) or 5 (~0.01%) miRNAs in our library are active, the FDR will go up to 90% or 99%, respectively. The chance of getting a valid hit from a primary round of screen becomes even slimmer when the experimental noises, which are assumed negligible in these FDR calculations, are factored in.

Choosing assays with a low background is not a general solution for the background issue because the low background comes with the decreased chance of getting any hit at all. A more general solution to the background problem would be to do the selection in multiple rounds; in each round, a reduced-complexity library is constructed based on the viral-inserts recovered from the cells that showed the desired phenotype in the previous round selection. A miRNA that gives a 100 fold boost in phenotype formation, for

example, will grow in representation by 100 fold at every round of selection; the growth in representation of the active miRNAs will result in the reduction of FDR also by 100 fold. If there is any miRNA that was previously known to induce the phenotype of interest, checking whether or not the miRNA is growing in representation through each round of selection would give a useful indication whether the selection is working.

Even after addressing all the technical and background issues, the possibility exists that a screen for any given phenotype fails because the library simply does not contain any miRNA for the specific phenotypes of choice. One way to hedge against this possibility is to apply the library to as many phenotypes as possible. Another way is not to restrict the screen into selecting for a few specific phenotypes, but instead to explore all possible phenotypes that can be induced after the library infection. For example, the pooled library can be infected into embryonic stem cells. A few days later, the infected cells can be single-cell-cloned into microtitre plates and expanded. Each expanded clone can be examined for potential differentiation into any specific lineage of development, by using the methods such as morphology examinations or antibody-based staining.

REFERENCES

Ball, M.P., Li, J.B., Gao, Y., Lee, J.H., LeProust, E.M., Park, I.H., Xie, B., Daley, G.Q., and Church, G.M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol *27*, 361-368.

Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., *et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

Lee, J.H., Park, I.H., Gao, Y., Li, J.B., Li, Z., Daley, G.Q., Zhang, K., and Church, G.M. (2009). A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. PLoS Genet *5*, e1000718.

Li, J.B., Gao, Y., Aach, J., Zhang, K., Kryukov, G.V., Xie, B., Ahlford, A., Yoon, J.K., Rosenbaum, A.M., Zaranek, A.W., *et al.* (2009a). Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. Genome Res *19*, 1606-1615.

Li, J.B., Levanon, E.Y., Yoon, J.K., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y., and Church, G.M. (2009b). Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science *324*, 1210-1213.

Mayr, C., Hemann, M.T., and Bartel, D.P. (2007). Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. Science *315*, 1576-1579.

Nilsson, M., Malmgren, H., Samiotaki, M., Kwiatkowski, M., Chowdhary, B.P., and Landegren, U. (1994). Padlock probes: circularizing oligonucleotides for localized DNA detection. Science *265*, 2085-2088.

Stadtfeld, M., and Hochedlinger, K. (2010). Induced pluripotency: history, mechanisms, and applications. Genes Dev *24*, 2239-2263.

Takeda, H., Charlier, C., Farnir, F., and Georges, M. (2010). Demonstrating polymorphic miRNA-mediated gene regulation in vivo: application to the g+6223G->A mutation of Texel sheep. RNA *16*, 1854-1863.

Zhang, K., Li, J.B., Gao, Y., Egli, D., Xie, B., Deng, J., Li, Z., Lee, J.H., Aach, J., Leproust, E.M., *et al.* (2009). Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. Nat Methods *6*, 613-618.