Systems genomics analysis centered on epigenetic inheritance supports development of a unified theory of biology

Abhay Sharma
CSIR-Institute of Genomics and Integrative Biology
Council of Scientific and Industrial Research
Sukhdev Vihar, Mathura Road
New Delhi 110025
Telephone: +91-11-26932421
Fax: +91-11-27662407
Email: abhaysharma@igib.res.in
Abstract

New discoveries are increasingly demanding integration of epigenetics, molecular biology, genomic networks, and physiology with evolution. This article provides a proof of concept for evolutionary transgenerational systems biology, proposed recently in the context of epigenetic inheritance in mammals. Gene set enrichment analysis of available genome level mammalian data presented here seems consistent with the concept that (1) heritable information about environmental effects in somatic cells is communicated to the germline by circulating microRNAs (miRNAs) or other RNAs released in physiological fluids, (2) epigenetic factors including miRNA-like small RNAs, DNA methylation and histone modifications are propagated across generations via gene networks, and (3) inherited epigenetic variations in the form of methylated cytosines are fixed in the population as thymines in evolutionary time course. The analysis supports integration of physiology and epigenetics with inheritance and evolution. This may catalyze efforts to develop a unified theory of biology.

Keywords

epigenetic inheritance, microRNA, gene expression and networks, DNA methylation, histone modification
Introduction

With the inability of the contemporary gene and natural selection centric evolutionary theory in fully explaining heritability of phenotypic traits based solely on DNA sequence variation in one hand, and the increasing evidence of non-genetic inheritance in another, there is a profound interest in integrating epigenetics, molecular biology, genomic networks, and physiology with the theory of evolution (Petronis, 2010; Danchin et al., 2011; Mattick, 2012; Ball, 2013; Noble, 2013; Noble et al., 2014; Noble, 2015). A top down approach toward this unification may begin with developing a broad conceptual mechanistic framework and testing that in a proof of concept analysis using empirical data. Notably, a recently proposed framework that is supported by observations reported in the literature (Sharma, 2014; Sharma, 2015a-c) provides an opportunity to analytically test the concept of a unified theory of biology. The proposed model explains transgenerational epigenetic inheritance and its evolutionary significance by integrating gene expression and gene networks, miRNA or other RNA, DNA methylation, histone modifications, and DNA methylation induced mutation, in three mechanistic steps (Supplementary Fig. S1). First, heritable information about environmental effects in somatic cells is communicated to germline by circulating RNAs, representing physiological conditions. Second, epigenetic modifications, in the form of RNA, DNA methylation and histone modifications, are propagated across generations through gene expression and gene networks. Third, inherited epigenetic variations, represented by methylated cytosines, are fixed in the population as thymines, in evolutionary time scale. In the present analysis, these three principles are tested using available large scale data related to gene expression, DNA methylation, histone modification, cytosine methylation induced mutation, chemical gene interaction, and genetic association in diseases. The analysis is based on gene set enrichment as a measure of association. Given that the original framework
Results and discussion

**Figure 1** illustrates evidence supporting the first principle. As shown, the circulating miRNA profiles of exosomes and body fluids closely resemble that of gonads and germ cells, in both males and females, and the miRNA profiles of germ cells resemble that of embryonic stages (Fig. 1A). The miRNA based similarity observed between circulating factors, germ cells, and developing embryo is found statistically significant (Fig. 1B). The exosome and body fluid miRNAs were found to be overrepresented in germ cells, and that of gametes in developing embryo. In contrast, number matched control miRNAs, selected randomly from the combined set of miRNAs representing 461 organs, tissues and various other samples in the database, were not overrepresented in general. These results clearly suggest that gametogenesis and development are strongly related to circulating factors in terms of miRNA profile.

Secreted by all cell types and identified in diverse body fluids, exosomes carry miRNAs that can influence gene expression and cause physiological changes in recipient cells (Chevillet et al., 2014; Melo et al., 2014; Alexander et al., 2015). Exosomal miRNAs are also considered to play a role in male and female reproductive physiology in mammals (Belleannée et al., 2013; Santonocito et al., 2014; Barkalina et al., 2015). Evidence suggests that miRNAs are loaded into exosomes selectively, based on specific miRNA motifs and posttranscriptional modifications, and levels of miRNAs or their targets in the producer cells (Alexander et al., 2015). The preferential sorting is supported by the observation that miRNA signatures of exosomes do not directly mirror miRNA composition of the producer cells (Alexander et al., 2015). These findings have suggested that some miRNAs may have evolved to be parceled...
out in exosomes for performing their biological roles (Alexander et al., 2015). Though separating exosomal miRNAs from other extracellular vesicle associated or vesicle independent circulating miRNAs is technically challenging, it has recently been inferred from quantitative and stoichiometric analysis that most exosomes do not contain many copies of miRNA molecules (Chevillet et al., 2014). This has led to the hypothesis that a given miRNA is either distributed in a few exosomes in the population at low concentration or, alternatively, packaged in rare exosomes at high concentration (Chevillet et al., 2014). Besides, selective loss of certain miRNAs with low GC content during extraction from smaller samples has also been noted (Kim et al., 2012). Given these mechanistic and technical reasons associated with quantitative analysis of extracellular miRNAs, the present qualitative analysis, based simply on presence or absence of a given miRNA, showing profile similarity between a general pool of circulating miRNAs, and gonadal and gamete borne miRNAs may at this time seem sufficient to support the concept that extracellular noncoding RNA can potentially mediate soma to germline transmission of heritable information.

To examine the possibility that certain circulating miRNAs may have evolved to be selectively released into the circulation to mediate epigenetic inheritance, the exosome and body fluid miRNAs present in testis, spermatogonia, spermatocytes, spermatids or spermatozoa, as well as in ovary, oocytes or ovum were examined for de novo discovery of sequence motifs. Interestingly, a motif was discovered in these miRNAs, not in a number matched control set of randomly selected exosome and body fluid miRNAs (Fig. 1C). The motif was present in 19 of the total 131 miRNAs. Startingly, enrichment analysis showed that the motif was significantly overrepresented in promoters, 1000 base pairs upstream to 200 base pairs downstream regions, of genes associated with various gene ontology terms including those that seem consistent with epigenetic inheritance, related to environmental
factors, intercellular communication, gene expression, development, energy metabolism, and nervous system structure and function, for example (Fig. 1D). It is notable that a large proportion of examples of nongenetic inheritance reported so far in mammals relate to carbohydrate and lipid metabolism, and brain and behavior (Choi and Mango, 2014; Bale, 2015; Sharma, 2015b; Szyf, 2015; Dia et al. 2015). Besides, a recent study has found that thousands of genes that escape genome-wide DNA demethylation in human primordial germ cells (PGCs), with complete data set not reported, overrepresent genes expressed in brain, and genes associated with metabolic, and neurological and neuropsychiatric disorders (Tang et al. 2015). Importantly, the potential biological significance of the identified miRNA motif and its presence in promoters was supported by examining nuclear localization of the miRNAs. The raw nuclear read counts for 8 of the 19 motif containing miRNA were available in a reported set of human cell line small RNA deep sequencing data, wherein a count of 10 or more was considered to indicate nuclear localization (Liao et al., 2010). Interestingly, in the data spanning 1307 unique mature miRNA sequences, the motif containing miRNAs were overrepresented in nucleus localized miRNAs (Fig. 1E), with 7 of the 8 motif containing miRNAs figuring within the top 21 miRNAs with highest counts in the nucleus. Further, the top 145 nucleus localized miRNAs, arbitrarily chosen from 1307 sequences, excluding the above 7 miRNAs, were found to be highly enriched for the motif, compared to the bottom 145 sequences with lowest nuclear counts below the threshold of calling nuclear localization (Fig. 1F). Thus, profound nuclear localization of the motif containing miRNAs and prominent presence of the motif in promoters of certain categories of genes together support the possibility that these miRNAs may regulate gene expression at transcriptional level. Indeed, examples for miRNA mediated transcriptional regulation are known (Zhang et al. 2014). In particular, evidence exists to suggest that miRNAs with binding sites in gene promoters, located in the range of around 1000 pase pairs without any unique feature, can
modulate gene expression through epigenetic modifications of the promoter, including histone acetylation and/or methylation (Zhang et al. 2014). This is consistent with the present analysis supporting a role of circulating miRNAs in epigenetic inheritance.

A mechanism of multigenerational epigenetic inheritance mediated by Piwi-interacting RNA (piRNA), a class of small noncoding RNAs that are expressed in male and female germline and play an evolutionarily conserved role in transposon silencing, has previously been demonstrated in the nematode *Caenorhabditis elegans* (Ashe et al. 2012). In this study, it was shown for the first time that a piRNA-dependent foreign RNA response leads to multigenerational gene silencing involving a germline nuclear small RNA/chromatin pathway. In *C. elegans*, the mechanisms underlying piRNA mediated transcriptional gene silencing are considered similar to that involved in nuclear RNAi pathway in somatic tissues, with repressive histone modifications and RNA polymerase II stalling leading to silencing (Weick and Miska, 2014). The present analysis raising the possibility of miRNA, an endogenous small RNA like piRNA, as playing a role in epigenetic inheritance in mammals seems very attractive because it satisfies the requirement of soma to germline communication, as envisaged in the first principle of the model under investigation.

**Figure 2** displays evidence in support of the second principle. The common circulating and germline miRNAs are overrepresented among miRNAs identified as differentially expressed in studies examining environmental effects in exposed and unexposed generations (Fig. 2A). Besides, the target genes of these common miRNAs show enrichment for genes that show differential mRNA expression (Fig. 2B) and DNA methylation (Fig. 2C) in gonads, gametes and various other tissues and organs in transgenerational studies. These results suggest a role of gene networks in epigenetic inheritance. Global analysis of data representing normal
conditions further supports this. For example, the targets are found to overrepresent imprinted genes (Fig. 2D), with imprinting representing a mode of epigenetic inheritance. The targets also enrich genes that are known to interact with a broad range of chemicals (Fig. 2E), including environmental factors known to cause transgenerational effects. Genes showing expression, regulation, and differential DNA methylation and histone modifications in gametes and developing embryo under normal condition are also overrepresented in the targets (Fig. 2F-H). Further, the targets enrich genes showing tissue wide expression (Fig. 2I), differential genome level CpG promoter density distribution (Fig. 2J) and histone modification signatures of active promoters and enhancers (Fig. 2K). The targets are also found to overrepresent genes with known function in processes related to gene regulation by noncoding RNA including miRNA, biogenesis and metabolism of these RNAs, gene-specific transcription, response to chemicals and abiotic factors, and embryonic development (Fig. 2L), and to metabolism, and brain development and function (Supplementary Table S7).

Finally, the reported binding regions of BLIMP1 (Magnúsdóttir et al, 2013), a key regulator of PGC specification involved in resetting of epigenome towards a basal state, were found to be highly significantly enriched in the targets (fold change 1.7, \( P < 10^{-16} \)), with the control showing a slight depletion (fold change 0.86, \( P < 0.01 \)). Cumulatively, these results appear consistent with the second principle of the conceptual framework implicating gene networks in epigenetic inheritance.

Figure 3 presents evidence supporting the third principle. The targets show enrichment for transcription factor binding sites created by cytosine methylation induced mutation in low CpG promoter associated genes (Fig. 3A). As CpG dinucleotides in these promoters are constitutively methylated in the germline (Weber et al., 2007), this result supports the evolutionary significance of epigenetic inheritance. Next, the targets were found to
overrepresent genes mutations or polymorphisms in which show association with various diseases (Fig. 3B), potentially connecting epigenetic variations, evolution, and human health. This is consistent with the above mentioned study showing overrepresentation of disease associated genes in DNA demethylation resistant genes in human PGCs (Tang et al., 2015). As proposed (Sharma 2015b,c), the epigenetic modifications may either discontinue to exist, persist as such, or convert to genetic alterations in evolutionary time course (Supplementary Fig. S1). Though transition of 5-methylcytosines to thymines per se could be a passive by-product, the resulting change can become a potential substrate for selection like any other newly arisen genetic variation does in the normal course of evolution. A recently proposed model of RNA-mediated gene evolution has underscored this possibility (Morris, 2015). As regards the overall evolutionary significance of epigenetic inheritance, it has been highlighted in a recently advanced unified theory of molecular aspects of evolution that natural selection may potentially act on acquired traits (Skinner, 2015). Cumulatively, the present results seem consistent with the third principle.

Evidence provided here supports the proposed (Sharma, 2014; Sharma, 2015a-c) model of evolutionary transgenerational systems biology. First, it demonstrates that miRNA based similarity exists between circulating factors, gametes, developmental stages, and adult tissues, under normal condition or under environmental conditions that produce an effect across generations. This is consistent with the hypothesis that RNAs present in physiological fluids hold the potential to mediate soma to germline communication in epigenetic inheritance. Second, the results show an association between the target genes of the convergent miRNAs and regulated gene expression, DNA methylation, and histone modifications in the above conditions. This is in line with the proposition that gene networks may underlie epigenetic memory propagation across generations. Third, the miRNA targets
show association with cytosine methylation induced mutational events, and disease related mutations and polymorphisms. This supports the suggestion that epigenetic inheritance may play an evolutionarily significant role. The above analysis is not exhaustive in the sense that several sets of available data remain to be examined. For example, as mentioned, a recent study has revealed that several thousand genes escape genome-wide DNA demethylation in human PGCs (Tang et al. 2015). Non-availability of the complete data set however prevented its analysis here. Besides, this study and several others (Magnúsdóttir et al. 2013; Smith et al., 2014; Irie et al., 2015) together provide additional data on gene expression and DNA methylation dynamics associated with gametes, and cells, tissues and organs pertaining to embryonic development in mouse and human. It will be interesting to extend the analysis to include these and newer studies in future. Nevertheless, the present results tend to unify environment, physiology, RNA, gene networks, DNA methylation, and histone modifications with inheritance, development, disease, and evolution. A potential role and integration of physiology, epigenetics, RNA, and genetic interactions in inheritance and evolution have previously been suggested and sought for on the basis of theoretical considerations (Richards, 2008; Day & Bonduriansky, 2011; Hunter et al., 2012; Livnat, 2013; Noble, 2013; Noble et al., 2014; Rivoire & Leibler, 2014; Jablonka & Lamb, 2015; Morris, 2015; Noble, 2015; Skinner, 2015). The data analysis presented here offers a proof of principle for a unified theory of biology. The results may prove valuable in directing future efforts toward integration.

Materials and methods

Data available in various public databases and published papers along with associated supplementary materials was used, as provided. The papers were identified through PubMed search using appropriate key words, and the relevant data sets included in the analysis.
without any bias. For miRNA and mRNA clustering or enrichment analysis, all the human miRNAs in the organ-miRNA interaction database (http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk), or all the human genes in gene ontology (http://geneontology.org/) were used as the total miRNA or mRNA population, in that order. To obtain controls for enrichment analysis, matching numbers of miRNAs or mRNAs were randomly selected (https://www.random.org/sequences/) from the aforesaid populations, as appropriate. Documented set of validated miRNA target genes was used (http://zmf.umm.uni-heidelberg.de/apps/zmf/mirwalk2/index.html). The statistical significance of enrichment was computed using hypergeometric distribution probability, with 0.05 as the nominal $P$ value cut-off. Overrepresentation analysis of gene ontology biological processes was carried out using 0.05 as Benjamini-Hochberg adjusted $P$ value cut-off (http://david.abcc.ncifcrf.gov/).

Chi-square test for homogeneity was used to check equality of two different populations. The MEME suite 4.10.1 (Bailey et al., 2009) was used for de novo discovery (http://meme-suite.org/tools/meme) and enrichment (http://meme-suite.org/tools/fimo) of miRNA motif, and its association with promoters of genes linked to gene ontology terms (http://meme-suite.org/tools/gomo), all under default settings. For motif discovery, the complete set of human miRNAs in miRBase (http://www.mirbase.org/) was used as background (Kozomara and Griffiths-Jones, 2014). A 0-order Markov model was assumed for the background, and the minimum motif width set at 6 nucleotides.
Acknowledgements

Work in the laboratory of A.S. was supported by BSC0122 network project of the Council of Scientific and Industrial Research, India.

Conflict of interest

The author declares no competing or financial interests.
References


Figures
Fig. 1. Soma to germline communication of phenotypic information.

(A) Clustering of biological samples based on presence or absence of microRNAs (miRNAs). Exosomal and body fluid profiles match closely with that of gonads and gametes, and the profiles of gametes with that of developmental stages. (B) Overrepresentation of exosomal and body fluid miRNAs combined in gonads and gametes, and of gametic miRNAs in developmental stages. (C) Display of the motif discovered. The motif logo, reverse complement on the right, is shown along with the $E$ value. (D) A subset of enriched gene ontology biological processes, all with $q$-value $< 0.05$, in genes with motif containing promoters. Log transformed nominal $P$ values for individual enrichment are plotted on the y-axis. (E) Venn diagrams showing enrichment of the motif containing miRNAs in the set of nucleus localized miRNAs. Fold enrichment (FE) and enrichment $P$ value are indicated. (F) Pie chart depicting enrichment of the motif in the set of nucleus localized miRNAs excluding the initially discovered (C) motif containing miRNAs. Enrichment $P$ values, hypergeometric distribution, related to data pertaining to bar diagrams are shown above the bars. Values with 16 or more negative exponent of 10 were rounded down to zero. NS, not significant. List of miRNAs representing all the samples indicated in the figure (A, B) are given in Supplementary Table S1. Details of motif discovery analysis (C) are provided in Supplementary Table S2. Complete set of enriched gene ontology biological processes (D) is shown in Supplementary Table S3. Details of motif enrichment (F) are given in Supplementary Table S4.
Fig. 2. Gene networks in germline transmission.

(A) Overrepresentation of miRNAs common to exosomes, body fluids, spermatozoa, ovum, and all other tissues indicated below the bars shown in Fig. 1B except fetus, total 18 in number, in the sets representing differentially expressed miRNAs reported in various studies on epigenetic inheritance. Note combined analysis, represented by the extreme right pair of bars, showing enrichment of common miRNAs in total miRNAs reported in epigenetic studies. Key to miRNA profiles: gender exposed at F0-treatment at F0-generation and sample profiled. ME, male exposure; FE, female exposure; HFD, high fat diet; low protein diet; T, testosterone; F, flutamide; TF, testosterone plus flutamide; B[a]P, benzo(a)pyrene; BA, bisphenol A. Mature miRNA names were used for enrichment analysis. (B) Enrichment of validated target genes of common miRNAs (A), total 708 in number (Supplementary Table S5), in differentially expressed mRNAs reported in studies on epigenetic inheritance in rat. Note combined analysis, represented by the extreme right pair of bars, showing enrichment of the targets in mRNAs reported in epigenetic studies. Key to mRNA expression profiles: gender exposed at F0-treatment at F0-generation, stage, gender and sample profiled. ME, male exposure; FE, female exposure; HFD, high fat diet; VIN, vinclozolin; PRD, protein restricted diet; HCD, high calorie diet. (C) Overrepresentation of targets (B) in differentially methylated genes reported in studies on epigenetic inheritance. Key to DNA methylation profiles: gender and environmental factor used for exposure at F0-generation, gender, sample and DNA elements profiled. DEPH, di-(2-ethylhexyl)phthalate; EE, environmental enrichment; DMRs, differentially methylated regions; CGI, CpG island. (D) Enrichment of targets (B) in imprinted genes. (E) Enrichment of targets (B) in chemical interacting mRNAs and proteins. (F) Gene expression in gametes, zygote and postzygotic stages. Overall, the targets tend to enrich genes expressed in sperm, oocytes, zygote and developing embryo.
ZGA, zygotic gene activation. (G) DNA methylation levels in gametes and developing embryo. The targets in general overrepresent genes with < 50%, and underrepresent genes with > 50% CpG island (CGI) methylation levels. For gene body levels, the targets enrich genes with > 50% methylation, without showing any trend for over- or under-representation of genes with < 50% levels. (H) Histone modifications in gametes and development. Note clear enrichment of the targets in H3K4me2, H3K4me3, and H3K27me3 positive genes in human sperm. The targets also show differential histone modifications in embryonic stem cells (ESCs) and PGCs. (I) Proteomic profiles across human tissues. Overrepresentation of the targets in proteins expressed in all tissues, and depletion in proteins with tissue biased expression, are clearly observed. (J) Genome level CpG density across human promoters. The targets are enriched and depleted in high and low CpG promoters, in that order. (C) Histone modification signatures of active promoters and enhancers in human cell/tissue types. The targets overrepresent active regulatory signatures in cells/tissues other than embryonic stem cells (ESCs). (L) Gene ontology biological process enrichment. A subset of processes enriched in the targets is shown. Control genes did not show enrichment/depletion of any process, as expected. Enrichment P values, hypergeometric distribution, are shown above bars. Values with 16 or more negative exponent of 10 were rounded down to zero. NS, not significant. Missing bars indicate overlapping counts as either nil or excessively low, considered inappropriate for analysis (A-D, F-H). Sources of data used in the analysis (A-L) are indicated in Supplementary Table S6. Complete set of enriched gene ontology biological processes (L) is provided in Supplementary Table S7.
Fig. 3. Evolutionary significance of epigenetic inheritance.

(A) Representation of the targets in gene promoters associated with transcription factor binding sites created by mutation of cytosine to thymine in the context of CpG dinucleotides. Targets are clearly enriched in low CpG promoter associated genes. (B) Enrichment of the targets in disease associated genes. Targets are clearly enriched. NHGRI, National Human Genome Research Institute; GAD, Genetic Association Database; MIM, Mendelian Inheritance in Man; LSDB, Locus Specific Mutation Databases; CTD, Comparative Toxicogenomics Database. Enrichment $P$ values, hypergeometric distribution, are shown above bars. NS, not significant. Values with 16 or more negative exponent of 10 were rounded down to zero. Sources of data used in the analysis (A, B) are indicated in Supplementary Table S8.