AcademicPres

Available online at www.notulaebotanicae.ro

Print ISSN 0255-965X; Electronic 1842-4309

Not. Bot. Hort. Agrobot. Cluj 38 (2) 2010, Special Issue, 09-15

Notulae Botanicae Horti Agrobotanici Cluj-Napoca

Plant Epigenetics: From Genomes to Epigenomes

Alain RIVAL, Thierry BEULÉ, Frédérique ABERLENC BERTOSSI, James TREGEAR, Estelle JALIGOT

UMR DIAPC, CIRAD, IRD, BP 64501, Montpellier, Centre IRD, F-34394 Montpellier Cedex5, France; alain.rival@cirad.fr

Abstract

Epigenetics is the study of heritable changes in gene expression that occur without a change in the DNA sequence. In recent years, this field has attracted increasing attention as more epigenetic mechanisms affecting gene activity are being discovered. Such processes involve a complex interplay between DNA methylation, histone modifications, and non-coding RNAs, notably small interfering RNAs (siRNAs) and micro RNAs (miRNAs). Epigenetic regulation is not only important for generating differentiated cell types during plant development, but also in maintaining the stability and integrity of their respective gene expression profiles. Although epigenetic processes are essential for normal development, they can become misdirected which leads to abnormal phenotypes and diseases, especially cancer. Sensing environmental changes and initiating a quick, reversible and appropriate response in terms of modified gene expression is of paramount importance for plants which are sessile autotrophs. Although epigenetic mechanisms help to protect plant cells from the activity of parasitic sequences such as transposable elements, this defense can complicate the genetic engineering process through transcriptional gene silencing. Epigenetic phenomena have economic relevance in the case of somaclonal variation: a genetic and phenotypic variation among clonally propagated plants from a single donor genotype. The success of sequencing projects on model plants has created widespread interest in exploring the epigenome in order to elucidate how plant cell decipher and execute the information stored and encoded in the genome. New high-throughput techniques are making it easier to map DNA methylation patterns on a large scale and results have already provided surprises.

Keywords: epigenetics, gene expression, chromatin, DNA methylation, silencing

Introduction

Epigenetics is the study of heritable changes in gene expression that occur without a change in DNA sequence. In all living organisms, epigenetic controls are essential for normal development and they become misdirected in cancer cells and other human disease syndromes.

In plants, although epigenetic mechanisms help to protect cells from parasitic elements, this defence can complicate the genetic engineering process through transcriptional gene silencing. Furthermore, these phenomena have economic relevance, for example, in somaclonal variation: a genetic and phenotypic variation among clonally propagated plants from a single donor genotype. The loss of phenotypic fidelity is now a major impediment to the development of large scale propagation of plants through in vitro processes such as somatic embryogenesis. Examples of aberrant phenotypes in regenerated plants include abnormal leaf structures and variant floral morphology, both organs being of paramount importance for applications in horticulture and/or agriculture.

Changes in DNA methylation have been hypothesised to play a key role in the mechanism underlying tissue-culture induced changes. Indeed, studies of both global methylation levels and the methylation of specific sites show that variation in DNA methylation occurs frequently during growth in tissue culture. In vitro plant regeneration, like somatic embryogenesis, bypasses the normal developmental process of fertilisation and plant development, thus potentially resulting in the instability of epigenetic patterns.

Next Generation Sequencing and the sequence explosion

The sequencing of the 3 billion base pairs of the human genome took 3 to 4 years using conventional Sanger sequencing and generated a bill of about $300 million. The commercial availability of next-generation sequencing (NGS) technologies that are up to 200 times faster and cheaper than conventional Sanger machines is spawning a flurry of ambitious new sequencing projects (Fig. 1). Over the past five years or so, massively parallel DNA sequencing platforms have become widely available, reducing the cost of DNA sequencing by several orders of magnitude, and democratizing the field by putting the sequencing capacity of a major genome center in the hands of individual investigators. Next-generation DNA sequencing has the potential to dramatically accelerate biological and biomedical research, by enabling the comprehensive analysis of genomes, transcriptomes and interactomes to become inexpensive, routine and widespread, rather than requiring significant production-scale efforts (Shendure and JI,

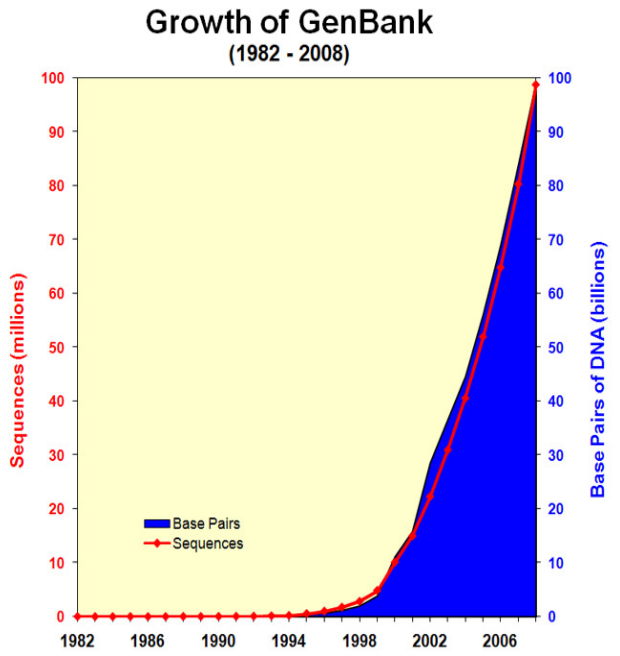

Fig. 1. The Sequence Explosion: changes in the number of base pairs and sequences deposited in Genbank between 1982 and 2008

2008). Several new sequencing technologies are emerging that have the potential to provide tremendous increases in throughput and reductions in cost. Indeed, companies such as 454 Life Sciences, Solexa, and Helicos Biosciences have each implemented competing technologies, vying to be widely adopted for the next generation of sequencing machines (Metzker, 2010).

Despite the recent exciting research advances involvingDespite the recent exciting research advances involving next-generation sequencers, it should be noted that method development is still in its infancy. Dedicated data storage facilities and extensive bioinformatics pipelines are indispensable for the proper exploitation of the considerable amount of sequence data resulting from these sequencing campaigns, so the corresponding investment must be carefully considered beforehand. Also, more studies are needed to address the reliability of these techniques as well as the correspondence between these results and those obtained with previous sequencing methods.

The success of genome sequencing projects has created wide-spread interest in exploring epigenomes in order to elucidate how the genome executes the information it holds.

The DNA double helix is heavily packed

A mammal diploid cell contains in average 6,000 Mb of DNA, corresponding to 1.8 meters in length, which are packed within a nucleus of 6 microns in diameter, following a strongly organized packaging. Indeed, each chromosome contains a single molecule of linear DNA which is folded several times and distributed over the centromere. Within the nucleus, DNA is never encountered under a free form but instead it is associated to other molecules, principally histones. Histones are small basic proteins which are present with the same proportion as DNA. Other proteins associated to DNA are non-histones acidic proteins, which account for 10 to 30% of the whole structure. This DNA-Protein complex is called chromatin (as it is easily stained) and it constitutes the chromosomes (Fig. 2).

The dynamic structure of chromatin reflects gene activity

Heterochromatin refers to regions of the genome that havHeterochromatin refers to regions of the genome that have low gene density, contain satellite repeat elements and are late replicating.

Euchromatin is a lightly packed form of chromatin that is rich in gene concentration, and is often under active transcription. Euchromatin comprises the most active

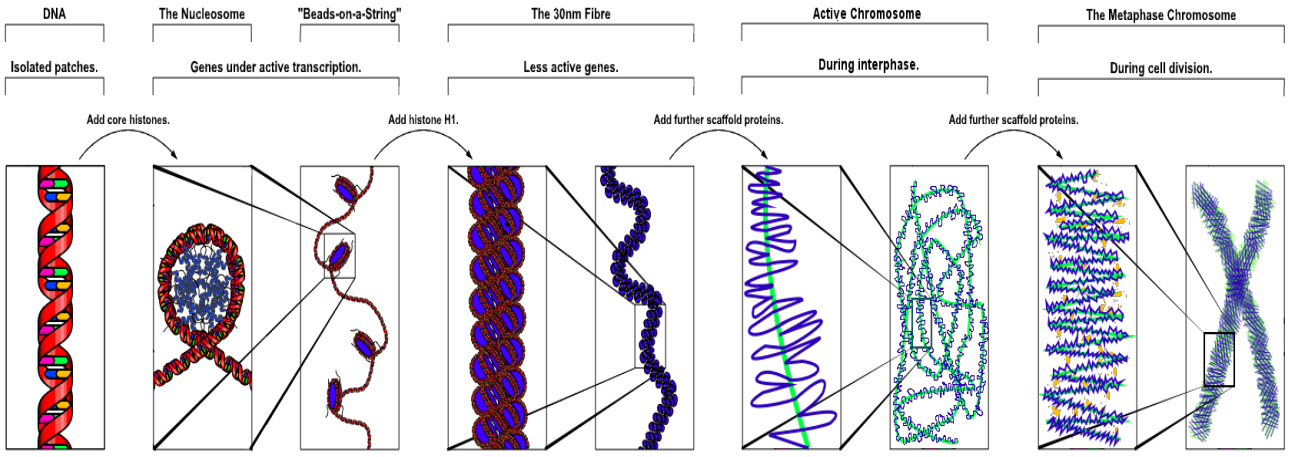

Fig. 2. Increasing orders of DNA packing

portion of the genome within the cell nucleus. Heterochromatin and euchromatin are associated with distinct DNA methylation and histone modification patterns that correlate with particular states of gene activity.

Deciphering the "Da Histone Code"

Covalent histone modifications, including acetylation, methylation, phosphorylation and ubiquitinylation, represent a 'histone code' that controls the transcriptional status of genes. Indeed, the acetylation of histones loosens the chromatin, facilitating replication and transcription. Depending on the position of the affected lysine residue, methylated histones can either have an activating of a repressing effect on transcription through modulations of chromatin condensation (Johnson *et al.*, 2002).

More specifically, methylation at either Lysine 4, Lysine 9 or Lysine 27 on Histone 3 change in the course of plant development and in response to environmental stresses. Lysine 4 methylation acts to promote transcription by recruiting nucleosome remodeling enzymes and histone acetylase whereas methylation at Lysine 9 participates to gene silencing and heterochromatinization. The role of methylation at lysine 27 is more ambiguous since it is abundant in both actively transcribed gene bodies and constitutive heterochromatin.

Transcriptionally active chromatin regions tend to be hyperacetylated and their DNA is globally hypomethylated. If a region of DNA or a gene is destined for silencing, chromatin remodelling enzymes such as histone deacetylases and ATP-dependent chromatin remodelers likely begin the gene silencing process. One or more of these activities may recruit DNA methyltransferase resulting in DNA methylation, followed finally by recruitment of the methyl-CpG binding proteins. The region of DNA will then be heritably maintained in an inactive state.

Small RNAs are key actors of epigenetic processes

Small RNAs have critical roles in most eukaryotes, as they repress gene expression by acting either on DNA to guide sequence elimination and chromatin remodeling, or on RNA to guide cleavage and translation repression.

The recent discovery of an increasing number of large and small non-protein-coding RNAs with specific regulatory roles has changed our view of gene expression. In particular, 20- to 27-nucleotide (nt) small RNAs belonging to two classes, microRNAs (miRNAs) and small interfering RNAs (siRNAs), are known to play essential roles. Small RNAs are involved in a variety of phenomena that are essential for genome stability, development, and adaptive responses to biotic and abiotic stresses. Their mode of action also is diverse. They guide heterochromatin assembly in fungi and plants. They target endogenous mRNAs for cleavage and translational repression in plants and animals, and cells against virus infection through an RNA-based immune system. They also control the movement of transposable elements at the transcriptional and post-transcriptional level (Vaucheret, 2006).

Double-stranded RNAs (dsRNAs) and their `diced' small RNA products can guide key developmental and defense mechanisms in eukaryotes. Some RNA-directed mechanisms act at a post-transcriptional level to degrade target messenger RNAs. However, dsRNA-derived species can also direct changes in the chromatin structure of DNA regions with which they share sequence identity. For example, plants use such RNA species to lay down cytosine methylation imprints on complementary DNA sequences (a mechanism known as RNA-dependent DNA methylation or RdDM), providing a fundamental mark for the formation of transcriptionally silent heterochromatin. Thus, RNA molecules are involved in a feedback mechanism to modulate the accessibility of information stored in the DNA of cognate genes. RNA triggers for DNA methylation can come from different sources, including invasive viral, transgene or transposon sequences, and in some cases are derived from single-stranded RNA precursors by RNA-dependent RNA polymerases. The mechanism by which RNA signals are translated into DNA methylation imprints is currently unknown, but two plant-specific classes of cytosine methyltransferase have been implicated in this process.

DNA methylation profiling by bisulfite DNA sequencing

Cataloguing genome-wide DNA methylation patterns (methylation landscaping) is the most commonly studied epigenetic modification. There are three main approaches to detecting DNA methylation on a large scale, including restriction endonuclease digestion coupled to microarray technology, bisulfite sequencing, and immunoprecipitation of 5′-methylcytosine to separate methylated from unmethylated DNA (Callinan and Feinberg, 2006).

Bisulfite sequencing is based on the chemical property of bisulfite to induce the conversion of cytosine residues to uracils while leaving 5′-methylcytosines intact. Therefore, sequencing of bisulfite-treated DNA will reveal the positions of methylated cytosines (those positions that remained cytosines following the treatment). PCR amplification of the region of interest in the bisulfite-reacted DNA yield a fragment in which all uracil, formerly cytosine, and thymine residues have been amplified as thymine and only 5-methylcytosines have been amplified as cytosine

Sequences will provide methylation maps of single DNA strands from individual DNA molecules in the original genomic DNA sample. The position of each 5-methylcytosine will be given by a positive band on a sequencing gel.

Taylor *et al.* (2007) improved upon the bisulfite DNA sequencing procedure by using the 454 technology to se-

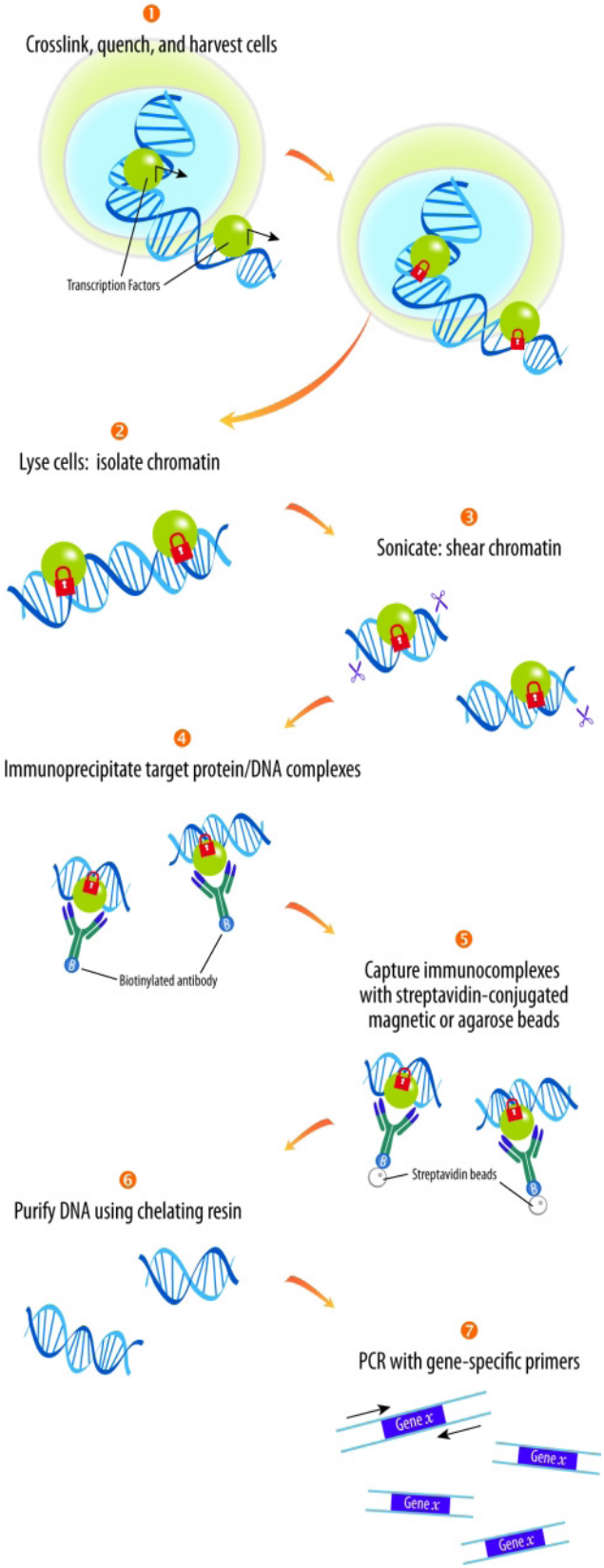

Fig. 3. Principle of chromatin immunoprecipitation (ChIP) technique

quence bisulfite-treated PCR amplicons corresponding to gene-related CpG-rich regions. A similar study using the Illumina technology has been recently reported by Cokus *et al.* (2008) in *Arabidopsis*.

New techniques are making it easier to map DNA methylation patterns on a large scale and the results have already provided surprises. In particular, the conventional view that DNA methylation functions predominantly to irreversibly silence transcription is being challenged. Not only is promoter methylation often highly dynamic during development, but many organisms also seem to target DNA methylation specifically to the bodies of active genes (Suzuki and Bird, 2008).

Next Generation Sequencing is boosting the analysis of epigenetic marks

Epigenetics is the study of heritable gene regulation thaEpigenetics is the study of heritable gene regulation that does not involve the DNA sequence. The two major types of epigenetic modifications regulating gene expression are DNA methylation by covalent modification of cytosine-5' and posttranslational modifications of histone tail (Callinan and Feinberg, 2006). Regulatory RNAs provide another means of epigenetic regulation of gene expression. Recent research has implicated such epigenetic modifications of prime importance in oncogenesis and development, setting the grounds for the Human Epigenome Project (HEP) initiative, which aims to catalog DNA methylation patterns on a genome-wide scale (Esteller, 2006). The next-generation sequencing technologies offer the potential to accelerate epigenomic research substantially. To date, these technologies have been applied in several epigenomic areas, including the characterization of DNA methylation patterns, posttranslational modifications of histones, and nucleosome positioning on a genome-wide scale.

Mapping histone modifications and the locations of DNA-binding proteins

Posttranslational covalent modifications of histone tails, which include methylation, acetylation, phosphorylation, and ADP-ribosylation, are thought to control gene expression by regulating the strength of DNA-histone interactions determining the accessibility of DNA to transcriptional regulators (Bird, 2001; Schones and Zhao, 2008). Historically, histone modifications have been identified by chromatin immunoprecipitation (ChIP; Kuo and Allis, 1999) which, in brief, involves cross-linking proteins to DNA, followed by immunoprecipitation of a protein of interest with a specific antibody, and characterization of the bound DNA by hybridization or PCR amplification (Fig. 3). The genome-wide development of the ChIP method using microarrays, known as ChIP-chip, combined the ChIP procedure with hybridization to a microarray to

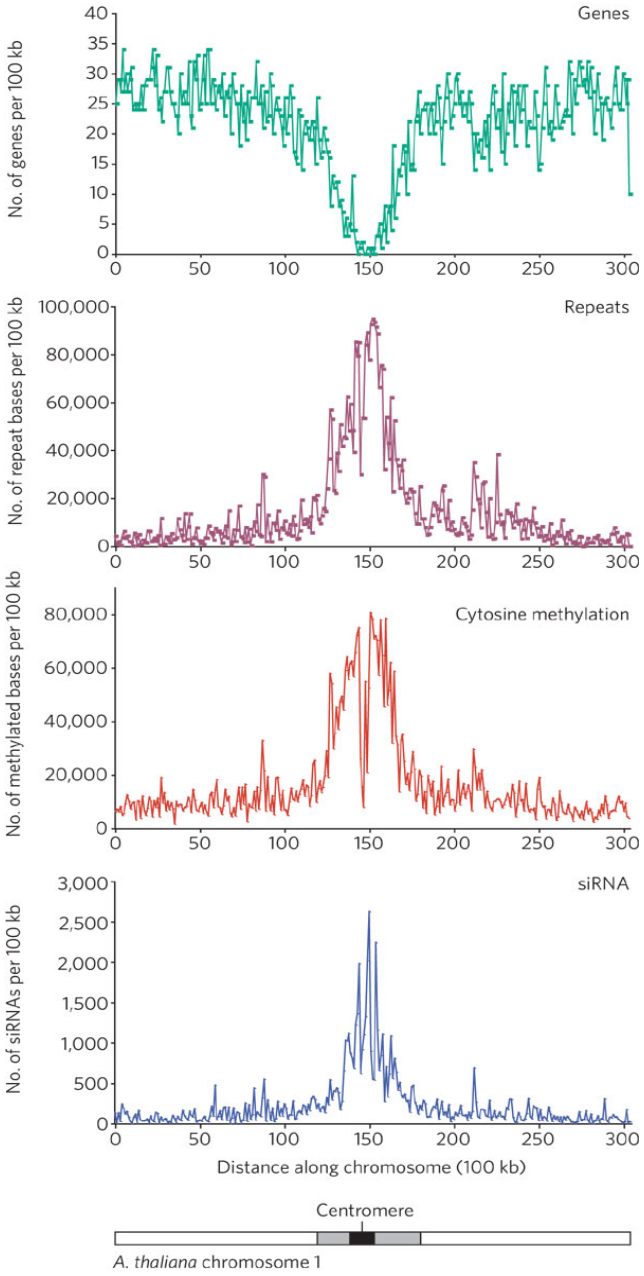

Fig. 4. The epigenetic 'landscape' of *A. thaliana*

The relative abundance of genes (number of annotated genes), repeats (repeat bases per 100 kb), cytosine methylation (methylated bases per 100 kb) and siRNAs (cloned siRNAs per 100 kb) is shown for the length of *A. thaliana* chromosome 1, which is ~30 Mb. Numbers on the *x* axis represent 100-kb windows along the chromosome. A diagram of chromosome 1 is also shown, with white bars indicating euchromatic arms, grey bars indicating pericentromeric heterochromatin and the black bar indicating the centromeric core (Henderson and Jacobsen, doi:10.1038/nature05917).

reveal the genome-wide distribution of the protein of interest (Ren *et al.*, 2000).

Sequence census methods have been recently coupled to the basic ChIP protocol to provide an alternative method for surveying histone modifications on a genome-wide scale. Roh *et al.* (2004) used ChIP followed by a Sanger sequencing-based SAGE procedure (also referred to as the genome-wide mapping technique) to study the distribution of acetylated histones H3 and H4 in the yeast genome.

The introduction of next-generation sequencing to the field has brought about the development of a new sequencing-based method, named ChIP-Seq, for detecting histone modifications on a genome-wide scale.

In this method, immunoprecipitated DNA is used to construct sequencing libraries for analysis on a next-generation sequencer to generate short sequence reads that, in contrast to ChIP-SAGE or STAGE, could be derived from either end of a ChIP DNA fragment regardless of the presence of a restriction site (Fields, 2007).

The number of reads that map to a particular genomic area can be used to quantify the strength of binding of the protein of interest in this area (or the amount of the assayed histone modification found at the site).

To date, Illumina technology has been most commonly used for the ChIP-Seq application. The 25-to 30-bp read length obtained on an Illumina sequencer suffices to map a typical 150- to 200-bp ChIP DNA fragment that may be sequenced from both ends. ChIP-Seq has been applied to the identification of histone modifications on a genome-wide scale in the human genome (Barski *et al.*, 2007). Of the genome-wide extensions of the ChIP protocol, ChIP-Seq has the potential for the highest resolution as its resolution depends only on the size of the input chromatin fragments and the depth of sequencing. On the other hand, the resolution of ChIP-SAGE (STAGE) also depends on the distribution of the restriction enzyme sites in the input ChIP DNA. The resolution of ChIP.chip depends on the resolution of probes used for the microarray. Both ChIP-SAGE and ChIP-Seq require less PCR amplification than ChIP-chip and, therefore, may provide improved accuracy for quantifying the binding signal (Schones and Zhao, 2008).

Exploring plant methylomes

Zhang *et al.* (2006) reported on the first comprehensive DNA methylation map of an entire genome, at 35 base pair resolution, using Arabidopsis thaliana as a model. Pericentromeric heterochromatin, repetitive sequences, and regions producing small interfering RNAs were found to be heavily methylated. Unexpectedly, over one-third of expressed genes contain methylation within transcribed regions, whereas only approximately 5% of genes show methylation within promoter regions. Interestingly, genes methylated in transcribed regions are highly expressed and constitutively active, whereas promoter-methylated genes show a greater degree of tissue-specific expression. Whole-genome tiling-array transcriptional profiling of DNA methyltransferase null mutants identified hundreds

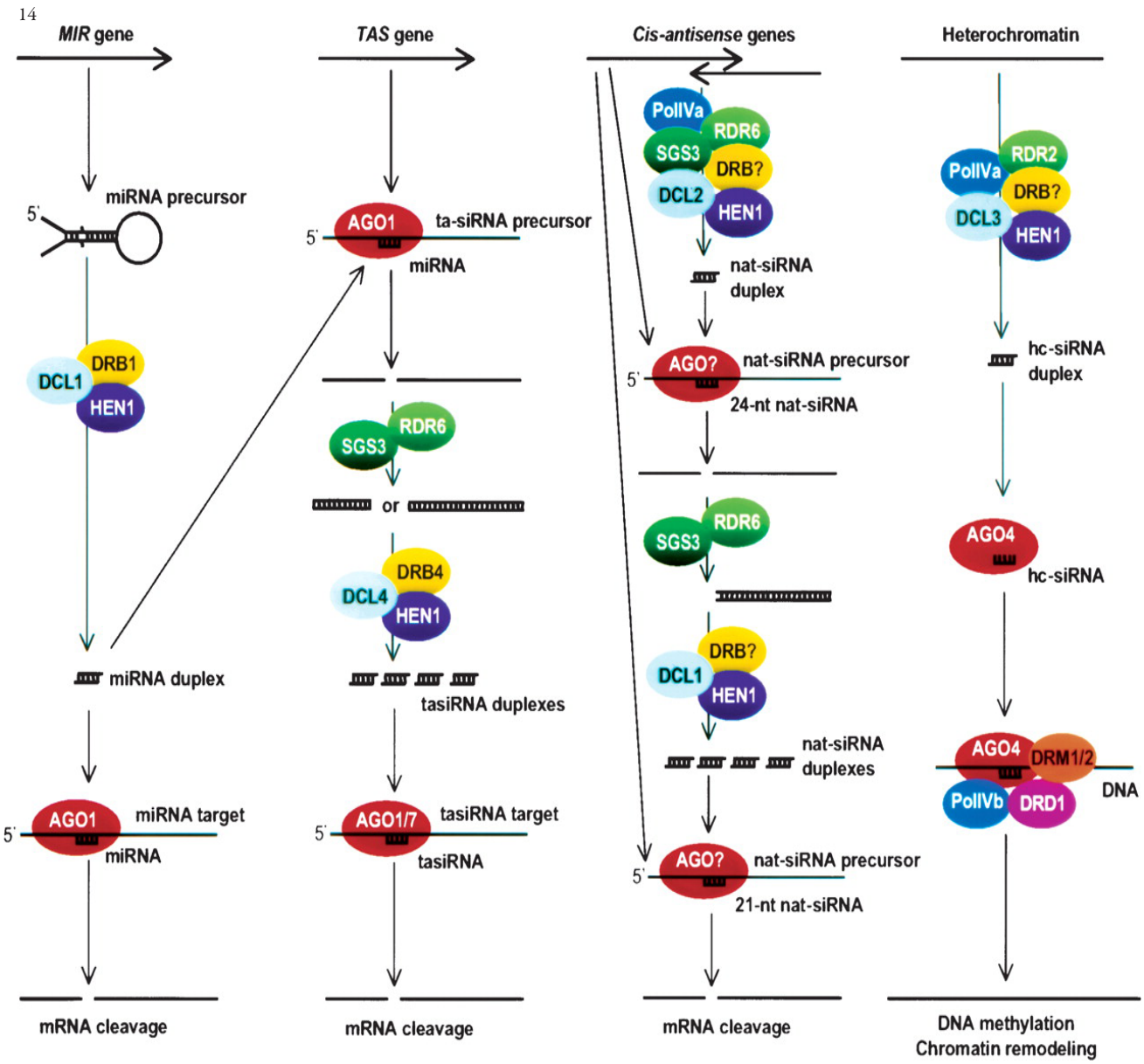

Fig. 5. miRNA, ta-siRNA, nat-siRNA, and hc-siRNA pathways in plants

of genes and intergenic noncoding RNAs with altered expression levels, many of which may be epigenetically controlled by DNA methylation. *A. thaliana* has a compact ~130-megabase (Mb) genome, although it contains considerable amounts of heterochromatin, which is repeat-rich and largely located in the centromeric and pericentromeric regions (Fig. 4). It will be important to describe the 'methylome' of other repeat-rich plant genomes, such as those of the grasses, to test the generality of the patterns observed in *A. thaliana*.

Li *et al.* (2008) provided insights into the rice epigenomes and their effect on gene expression and plant development. High-resolution maps of DNA methylation and H3K4 di- and trimethylation of two entire chromosomes and two fully sequenced centromeres were obtained from rice (*Oryza sativa*) shoots and cultured cells. This analysis reveals combinatorial interactions between these epigenetic modifications and chromatin structure and gene expression. Cytologically densely stained heterochromatin had less H3K4me2 and H3K4me3 and more methylated DNA than the less densely stained euchromatin, whereas centromeres had a unique epigenetic composition. Most transposable elements had highly methylated DNA but no H3K4 methylation, whereas more than half of protein-coding genes had both methylated DNA and di- and/or trimethylated H3K4. Methylation of DNA but not H3K4 was correlated with suppressed transcription. By contrast, when both DNA and H3K4 were methylated, transcription was only slightly reduced. Transcriptional activity was positively correlated with the ratio of H3K4me3/H3K4me2: genes with predominantly H3K4me3 were actively transcribed, whereas genes with predominantly H3K4me2 were transcribed at moderate levels. More protein-coding genes contained all three modifications, and

more transposons contained DNA methylation in shoots than cultured cells. Differential epigenetic modifications correlated to tissue-specific expression between shoots and cultured cells.

Conclusions: From one single genome to a series of epigenomes

Epigenomics refers to the large scale study of epigenetic marks on the genome, which include covalent modifications of histone tails (acetylation, methylation, phosphorylation, ubiquitination), DNA methylation and the small RNAs machinery.

Epigenetic components are all amenable to genome-wide studies. Integrated approaches that correlate gene expression with DNA methylation and chromatin profiles are being designed. The $1,000 genome may still be years away, but with the arrival of next-generation sequencing (NGS), large-scale sequencing of hundreds or even thousands of genomes is fast becoming reality.

The transcriptomes of an organism are continually changing in response to developmental and environmental cues. Similarly, the epigenome is not static and can be molded by developmental signals, environmental perturbations, and disease states. Therefore, many epigenomes will need to be sequenced for a single organism, making epigenome sequencing perhaps even more challenging than genome sequencing.

References

Barski, A., S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev and K. Zhao (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell 129 (4):823-837.

Beck, S. and V. K. Rakyan (2008). The methylome: approaches for global DNA methylation profiling. Trends in Genetics 24(5):231-237.

Bird, A. (2001). Molecular biology. Methylation talk between histones and DNA. Science 294(5549):2113-2115.

Callinan, P. A. and A. P. Feinberg (2006). The emerging science of epigenomics, Hum. Mol. Genet. 15 Spec. 1:R95-R101.

Chan, S. W. L., I. R. Henderson and S. E. Jacobsen (2005). Gardening the genome: DNA methylation in Arabidopsis thaliana. Nature Reviews Genetics 6:351-360.

Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini and S. E. Jacobsen (2008). Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning, Nature 452:215-219.

Esteller, M. (2006). The necessity of a human epigenome project, Carcinogenesis 27:1121-1125.

Fields, S. (2007). Molecular biology: site-seeing by sequencing, Science 316:1441-1442.

Henderson, I. R. and S. E. Jacobsen. Epigenetic inheritance in plants. Nature 447:418-424. doi:10.1038/nature05917.

Johnson, L. M., X. F. Cao and S. E. Jacobsen (2002). Interplay between two epigenetic marks: DNA methylation and histone H3 lysine methylation. Current Biology 12:1360-1367.

Kuo, M. H. and C. D. Allis (1999). *In vivo* cross-linking and immunoprecipitation for studying dynamic protein:DNA associations in a chromatin environment. Methods 19:425-433.

Mathieu, O. and J. Bender (2004). RNA-directed DNA methylation. Journal of Cell Science 117:4881-4888.

Metzker, M. L. (2010). Sequencing technologies-the next generation. Nature Reviews Genetics 11:31-46.

Morozova, O., M. A. Marra (2008). Applications of next-generation sequencing technologies in functional genomics. Genomics 92(5), 255-264

Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell and R. A. Young (2000). Genome-wide location and function of DNA binding proteins. Science 290:2306-2309.

Roh, T. Y., W. C. Ngau, K. Cui, D. Landsman and K. Zhao (2004). High-resolution genome-wide mapping of histone modifications, Nat. Biotechnol. 22:1013-1016.

Schones, D. E. and K. Zhao (2008). Genome-wide approaches to studying chromatin modifications. Nat. Rev. Genet. 9:179-191.

Shendure, J. and H. Ji (2008). Next-generation DNA sequencing. Nature Biotechnology 26:1135-1145.

Suzuki, M. M., A. Bird (2008). DNA methylation landscapes: provocative insights from epigenomics. Nature Reviews Genetics 9:465-476

Taylor, K. H., R. S. Kramer, J. W. Davis, J. Guo, D. J. Duff, D. Xu, C. W. Caldwell and H. Shi (2007). Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing, Cancer Res. 67: 8511-8518.

Vaucheret, H. (2006). Post-transcriptional small RNA pathways in plants: mechanisms and regulations. Genes Dev. 2006 20: 759-771.

Xueyong, Li, X. Wang, K. He, Y. Ma, N. Su, H. He, V. Stolc, W. Tongprasit, W. Jin, J. Jiang, W. Terzaghi, S. Li and X. W. Deng (2008). High-Resolution Mapping of Epigenetic Modifications of the Rice Genome Uncovers Interplay between DNA Methylation, Histone Methylation, and Gene Expression. The Plant Cell 20:259-276.

Zhang, X., J. Yazaki, A. Sundaresan, S. Cokus, S. W. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini, S. E. Jacobsen and J. R. Ecker (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell 126(6):1189-1201.