Cell

The Cell Biology of Genomes: Bringing the Double Helix to Life

Tom Misteli[1],*
[1]National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA
*Correspondence: mistelit@mail.nih.gov

http://dx.doi.org/10.1016/j.cell.2013.02.048

The recent ability to routinely probe genome function at a global scale has revolutionized our view of genomes. One of the most important realizations from these approaches is that the functional output of genomes is affected by the nuclear environment in which they exist. Integration of sequence information with molecular and cellular features of the genome promises a fuller understanding of genome function.

It was a moment of scientific amazement in 1953 when Watson and Crick revealed the structure of DNA. The magnificence of the double helix and its elegant simplicity were awe inspiring. But more than just being beautiful, the double helix immediately paved the way forward; its structure implied fundamental biological processes such as semiconservative replication and the notion that chemical changes in its composition may alter heritable traits. The linear structure of DNA laid the foundation for the concept that a string of chemical entities could encode the information that determines the very essence of every living organism. The beauty of the double helix was the promise that, if the sequence of bases in the genome could be mapped and decoded, the genetic information that underlies all living organisms would be revealed and the secret of biological systems would be unlocked.

The idea of linearly encoded genetic information has been spectacularly successful, culminating in the recent development of powerful high-throughput sequencing methods that now allow the routine reading of entire genomes. The conceptual elegance of the genome is that the information contained in the DNA sequence is absolute. The order of bases can be determined by sequencing, and the result is always unequivocal. The ability to decipher and accurately predict the behavior of genome sequences was appealing to the early molecular biologists, has given rise to the discipline of molecular genetics, and has catalyzed the reductionist thinking that has driven and dominated the field of molecular biology since its inception.

But the apparent simplicity and deterministic nature of genomes can be deceptive. One of the most important lessons learned from our ability to exhaustively sequence DNA and to probe genome behavior at a global scale by mapping chromatin properties and expression profiling is that the sequence is only the first step in genome function. In intact living cells and organisms, the functional output of genomes is modulated, and the hard-wired information contained in the sequence is often amplified or suppressed. While mutations are an extreme case of genome modulation, most commonly occurring changes in genome function are more subtle and consist of fluctuations in gene expression, temporary silencing, or temporary activation of genes. Although not caused by mutations, these genome activity changes are functionally important.

Several mechanisms modulate genome function (Figure 1). At the transcription level, the limited availability of components of the transcription machinery at specific sites in the genome influences the short-term behavior of genes and may make their expression stochastic. Epigenetic modifications are capable of overriding genetically encoded information via chemical modification of chromatin. Similarly, changes in higher-order chromatin organization and gene positioning within the nucleus alter functional properties of genome regions.

The existence of mechanisms that modulate the output of genomes makes it clear that a true understanding of genome function requires integration of what we have learned about genome sequence with what we are still discovering about how genomes are modified and how they are organized in vivo in the cell nucleus.

The Stochastic Genome

The genome is what defines an organism and an individual cell. It is therefore tempting to assume that identical genomes behave identically in a population of cells. We now know that this is not the case. Individual, genetically identical cells can behave very differently even in the same physiological environment. It is rare to find a truly homogeneous population of cells even under controlled laboratory conditions, as anyone who has tried to make a cell line stably expressing a transgene knows. Much of the variability in biological behavior between individual cells comes from stochastic activity of genes (Raj and van Oudenaarden, 2008).

Genes are by definition low-copy-number entities, as each typically only exists in two copies in the cell. Similarly, many transcription factors are present in relatively low numbers in the cell nucleus. The low copy number of genes and transcription factors makes gene expression inherently prone to stochastic effects (Raj and van Oudenaarden, 2008). Numerous observations make it clear that gene expression is stochastic in vivo. For example, dose-dependent increases in gene expression after treatment of cell populations with stimulating

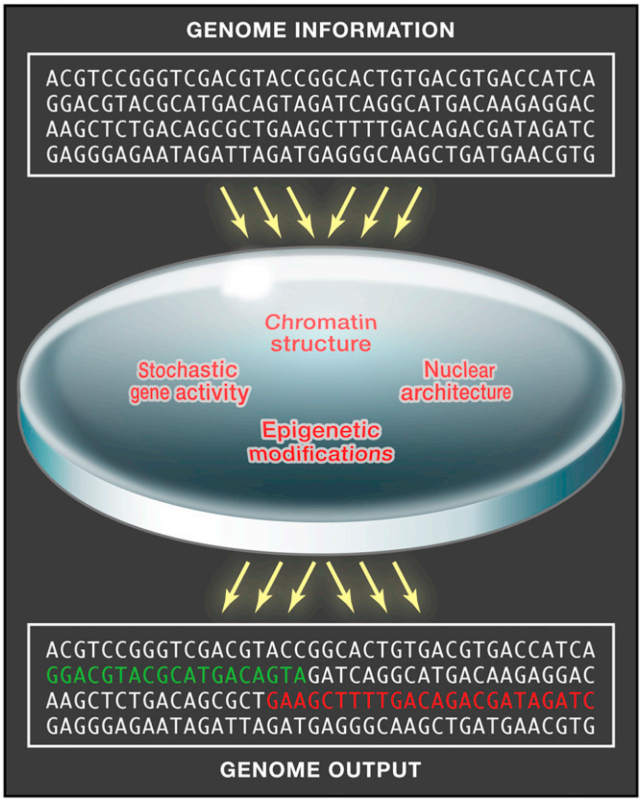
GENOME INFORMATION
ACGTCCGGGTCGACGTACCGGCACTGTGACGTGACCATCA
GGACGTACGCATGACAGTAGATCAGGCATGACAAGAGGAC
AAGCTCTGACAGCGCTGAAGCTTTTGACAGACGATAGATC
GAGGGAGAATAGATTAGATGAGGGCAAGCTGATGAACGTG
Chromatin structure
Stochastic gene activity
Nuclear architecture
Epigenetic modifications
ACGTCCGGGTCGACGTACCGGCACTGTGACGTGACCATCA
GGACGTACGCATGACAGTAGATCAGGCATGACAAGAGGAC
AAGCTCTGACAGCGCTGAAGCTTTTGACAGACGATAGATC
GAGGGAGAATAGATTAGATGAGGGCAAGCTGATGAACGTG
GENOME OUTPUT

Figure 1. From Primary Sequence to Genome Output
The hard-wired primary information contained in the genome sequence is modulated at short or long timescales by several molecular and cellular events. Modulation may lead to activation (green) or silencing (red) of genome regions.

ligands, such as hormones, are often brought about by high expression of target genes in a relatively small number of cells in the population rather than by a uniform increase in the activity in all cells. Stochastic gene behavior is most evident in single-cell imaging approaches, and mapping by fluorescence in situ hybridization of multiple genes, which according to population-based PCR analysis are active in a given cell population, shows that only a few cells transcribe all “constitutively active” genes at any given time. Most cells only express a subset of genes, and the combinations vary considerably between individual cells. These observations suggest that many genes blink on and off and are expressed in bursts rather than in a continuous fashion (Larson et al., 2009).

The molecular basis for stochastic gene expression is unknown. There are several candidate mechanisms, all of which are related to genome or nuclear organization. Most genes require some degree of chromatin remodeling for activity, which is thought to make regulatory regions accessible to the transcription machinery. Several observations suggest that chromatin remodeling contributes to the stochastic bursting of gene expression. Maybe most compelling is the finding that genes located near each other on the same chromosome show correlated blinking behavior, indicating that a local chromosome property, such as chromatin structure, drives stochastic behavior (Becskei et al., 2005). Furthermore, altering chromatin, for example by deletion of chromatin remodeling machinery, affects stochastic variability in yeast. It can be envisioned that the stochastic behavior of genes is caused by the requirement for cyclical opening of chromatin regions. Open chromatin has a limited persistence time, and maintaining chromatin in an open state requires the cyclical action of chromatin remodelers. Whether an “active” gene is transcribed at any given time may thus depend on the transient condensation status of its chromatin at a particular moment.

A second mechanism to impose nonuniform stochastic genome activity may be the local availability of the transcription machinery at a gene. Although transcription factors are able to relatively freely diffuse through the nuclear space, and in this way effectively scan the genome for binding sites, their availability and functionality at a given local site may undergo significant temporal fluctuations (Misteli, 2001). The local availability of transcription complexes may affect transcription frequency positive or negatively. On the one hand, it is possible that relatively stable preinitiation complexes persist on a given gene, where they may support multiple rounds of transcription and in this way boost initiation frequency. On the other hand, assembly of the full polymerase is a stochastic and relatively inefficient event itself. In order for a functional polymerase complex to assemble, individual transcription machinery components associate with chromatin in a step-wise fashion, and formation of the mature polymerase complex involves multiple partially assembled intermediates, many of which are unstable and disintegrate before a functionally competent complex is formed (Misteli, 2001). The inefficiency of polymerase assembly may create stochasticity at an individual locus.

A further contributor to stochastic gene expression may be the organization of transcription events in transcription factories. These hubs of transcription consist of accumulations of transcription factors to which multiple genes, often located on distinct chromosomes, are recruited (Edelman and Fraser, 2012). Typically only a few hundred such transcription factories are observed in a mammalian cell nucleus. It is possible that some genes need to physically relocate from nucleoplasmic locations to transcription factories. A nominally “active” gene locus that is not associated with a transcription factory may thus be stochastically silent. The relatively low number of transcription sites makes them a limiting factor in the transcription process and thus a potential mediator of stochastic gene expression.

Epigenetics—And When Epigenetics Is Not Epigenetics

Stochastic effects modulate genome output on short timescales. A mechanism to modulate the hardwired information of genomes on longer timescales is via epigenetics. The Greek-derived “Epi” means “over” or “above,” and epigenetic effects are defined as heritable changes in genome activity caused by mechanisms other than changes in DNA sequence. Epigenetic events are mediated by chemical modifications of DNA or core histones in complex patterns by methylation, acetylation, ubiquitination, phosphorylation, etc. These modifications alter gene expression by changing the chromatin surface and in this way affect the binding of regulatory factors. Well-established examples of such effects include binding of the DNA-methylation-dependent binding of the MeCP2 protein or the binding of PHD-domain-containing proteins to trimethylated histone H3 tails. Prominent biological effects based on epigenetic regulation are phenotypic differences between homozygous twins or imprinted genes that are expressed from only one allele in a diploid organism.

A central tenet in the definition of epigenetic regulation is that its effects are heritable, i.e., transmittable over generations. In fact, the concept of epigenetics was inspired by epidemiological findings that nutrient availability in preadolescents during the 19th century Swedish famine determined life expectance of their grandchildren. The epidemiological studies have recently been complemented by controlled laboratory studies in mice (Rando, 2012), and they have been extended to the molecular level by the findings that loss of the histone H3K4-trimethylation prolongs lifespan in *C. elegans* in a heritable fashion for several generations (Greer et al., 2011).

A complicating aspect of epigenetics is that the same modifications that mediate heritable epigenetic regulation may also bring about nonheritable transient modulations of the genome. In fact, the term "epigenetic" is nowadays often used in a very cavalier manner to refer to any biological effect, heritable or not, that is affected by histone modifications. Even if they are not heritable, histone modifications are biologically relevant modulators of genome function. The system of histone modifications is in many ways akin to the mechanisms by which signal transduction pathways work (Schreiber and Bernstein, 2002). Just as in signal transduction pathways, posttranslational modifications on histone tails create binding sites that are then recognized by adaptor or reader proteins, which in turn elicit downstream effects such as activation of kinases in the case of signaling cascades or recruitment of transcription factors in the case of histone modifications. In further analogy to the reversible events in signaling pathways, histone modifications can be altered or erased by modifying enzymes. Such transient and reversible modulatory effects of histone modifications have been implicated in every step of gene expression, starting from chromatin remodeling to recruitment of transcription machinery and even to downstream events that were thought to be chromatin independent, such as alternative pre-mRNA splicing (Luco et al., 2011). It is often difficult to determine heritability of these histone modification effects, and it therefore remains unclear how many of them are truly epigenetic. Regardless, DNA and histone modifications are an obvious source of modulation of the information contained in the genome sequence.

Genome Organization as a Modulator of Genome Function

Genomes of course do not exist as linear, naked DNA in the cell nucleus but are organized into higher-order chromatin fibers, chromatin domains, and chromosomes. Many correlations between genome organization and activity have been made—most prominently, the findings that transcriptionally active genes are generally located in decondensed chromatin and that transcriptionally repressed genome regions are often found at the nuclear periphery. These observations point to the possibility that the spatial organization of the genome modulates its functional output.

But in considering the relationship of genome structure with its function, we are faced with a perpetual chicken-and-egg problem. Does structure drive function, or is structure merely a reflection of function? Much of the thinking on this topic has been guided by observations on individual genes. How representative these were for the genome as a whole has been a confounding concern. Recent unbiased genome-wide analysis of structure/function relationships has validated the tight link between structure and function. Large-scale analysis of chromatin structure, histone modifications, and expression profiles shows that genomes are portioned into well-defined domains that closely correlate with their activity status and the presence of active or repressive histone marks (Sexton et al., 2012). The domains are separated by sharp boundaries marked by particular histone modification patterns and binding sites for chromatin insulator proteins such as CTCF. Even stronger evidence comes from the analysis of physical interactions *between* chromatin domains. At least in fruit flies, functionally equivalent domains tend to preferentially interact; that is, domains containing silent regions cluster in three-dimensional space, as do domains containing active regions (Sexton et al., 2012).

But can genome structure drive its function? The best example for structure-mediated gene expression effects is the silencing of genes when they become juxtaposed to heterochromatin domains, be it in the nuclear interior or at the nuclear periphery (Beisel and Paro, 2011). Gene activity has also frequently been linked to the position of a gene within the cell nucleus. The strongest evidence for such a relationship is experiments in which genes are transplanted from the nuclear interior to the lamina, leading to their repression or making them refractory to activation (Geyer et al., 2011). Based on these and similar experiments, it is often quite categorically stated that active genome regions are found in the interior of the nucleus and inactive ones at the periphery. This is a somewhat misleading oversimplification. Although lamina-associated genome regions are generally gene poor and are not transcribed, transcription labeling experiments reveal numerous active transcription sites at the periphery, and genes that are near the periphery, but not physically associated with it, are often active. On the other hand, inactive genes are frequently found in the interior. As far as we can tell, nuclear position per se does not determine activity, but association with repressive regions of the nucleus, be it at the periphery or the interior, does.

So, how then should we think about the chicken-and-egg problem of nuclear structure and function? How can it be that clear evidence exists for both "function-driving-structure" as well for "structure-driving-function"? The likely answer is that both effects are at play and are part of an overarching principle in which the mutual interplay of structure and function at multiple levels influences gene expression. The fact that there are very few known heterochromatic active genes suggests that a structural change in the form of chromatin decondensation is a crucial early step in gene activation. However, because chromatin states are generally unstable, mechanisms that reinforce a decondensed chromatin state must be in force for a gene to remain active. Such reinforcing mechanisms are dependent on gene activity and represent the "activity-drives-function" aspect of gene expression. Reinforcement mechanisms might be mediated by what we consider "active" histone modifications, some of which are known to be deposited during transcription as the polymerases traverse genes. On the flipside, a chromatin domain may also impose its effect on neighboring regions, either in *cis* on the same chromosome by spreading or in *trans* on distinct chromosomes. This effect represents the "structure-drives-function" aspect of genome function. Such a bidirectional, self-enforcing function-structure-function model accounts for most experimental observations on structure-function relationships in gene expression.

Facing the Complexity

Since the discovery of the double helix, we have come to realize that understanding genomes requires more than reading their sequence and that the information contained in the sequence is modulated by the cellular environment. How then do we gain full knowledge of the functional information encoded in genomes?

To get a comprehensive picture of the functional output of genomes, the sequence information needs to be integrated with other information parameters such as epigenetic patterns, higher-order chromatin landscapes, and noncoding RNA profiles. The technology to do this is now available, and intense efforts are currently underway to comprehensively gather these data sets in various biological systems. The first examples of such multilevel mapping analyses are emerging, such as the recent flurry of reports from the ENCODE consortium, which has systematically mapped genome properties ranging from histone modification profiles to regulatory elements and chromatin structure (Ecker et al., 2012). Given the scale and complexity of the generated data, not to mention the technical difficulties in gathering it, this is a challenging undertaking that will require a series of progressively larger studies. Ideally, future studies should be designed to systematically map multiple genome properties for focused biological systems such as specific human diseases.

Large-scale mapping of genome-related parameters and their comparison is a logical and necessary next step in the exploration of genomes and their function. These efforts will create invaluable catalogs of genome properties, and the hope is that, by cross-comparing data sets, insight into the rules that govern genome regulation will be gleaned. One can go one step further and advocate for an even more comprehensive approach in which genome expression data are then compared to other cellular characteristics such as proteomic, metabolomic, morphological, and physiological data to systematically link genome activity to biological behavior. The ultimate version of such an approach was recently described in a report by the US National Academies of Sciences entitled "Toward Precision Medicine," which envisioned a fully minable biomedical data repository that would include information ranging from genomic and epigenetic parameters to physiological features and clinical symptoms.

The elegant simplicity of the DNA structure revealed by Watson and Crick is still stunning. True to its promise when it was first discovered, it opened up the floodgates to understanding heredity. But one of the most profound lessons from the ensuing decades of genome exploration must be that the linear arrangement of bases in the DNA is not an absolute set of instructions but is malleable by the cellular environment. We are just beginning to uncover some of the mechanisms that are responsible for these effects. As is the rule in biology, wherein the whole is often greater than the sum of its parts, we are realizing that the genome is far more complex than the sequence of its DNA.

ACKNOWLEDGMENTS

Due to space limitations, mostly review articles were cited. Work in the author's laboratory is supported by the Intramural Research Program of the National Institutes of Health (NIH), NCI, Center for Cancer Research.

REFERENCES

Becskei, A., Kaufmann, B.B., and van Oudenaarden, A. (2005). Contributions of low molecule number and chromosomal positioning to stochastic gene expression. Nat. Genet. *37*, 937–944.

Beisel, C., and Paro, R. (2011). Silencing chromatin: comparing modes and mechanisms. Nat. Rev. Genet. *12*, 123–135.

Ecker, J.R., Bickmore, W.A., Barroso, I., Pritchard, J.K., Gilad, Y., and Segal, E. (2012). Genomics: ENCODE explained. Nature *489*, 52–55.

Edelman, L.B., and Fraser, P. (2012). Transcription factories: genetic programming in three dimensions. Curr. Opin. Genet. Dev. *22*, 110–114.

Geyer, P.K., Vitalini, M.W., and Wallrath, L.L. (2011). Nuclear organization: taking a position on gene expression. Curr. Opin. Cell Biol. *23*, 354–359.

Greer, E.L., Maures, T.J., Ucar, D., Hauswirth, A.G., Mancini, E., Lim, J.P., Benayoun, B.A., Shi, Y., and Brunet, A. (2011). Transgenerational epigenetic inheritance of longevity in Caenorhabditis elegans. Nature *479*, 365–371.

Larson, D.R., Singer, R.H., and Zenklusen, D. (2009). A single molecule view of gene expression. Trends Cell Biol. *19*, 630–637.

Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., and Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. Cell *144*, 16–26.

Misteli, T. (2001). Protein dynamics: implications for nuclear architecture and gene expression. Science *291*, 843–847.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216–226.

Rando, O.J. (2012). Daddy issues: paternal effects on phenotype. Cell *151*, 702–708.

Schreiber, S.L., and Bernstein, B.E. (2002). Signaling network model of chromatin. Cell *111*, 771–778.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.