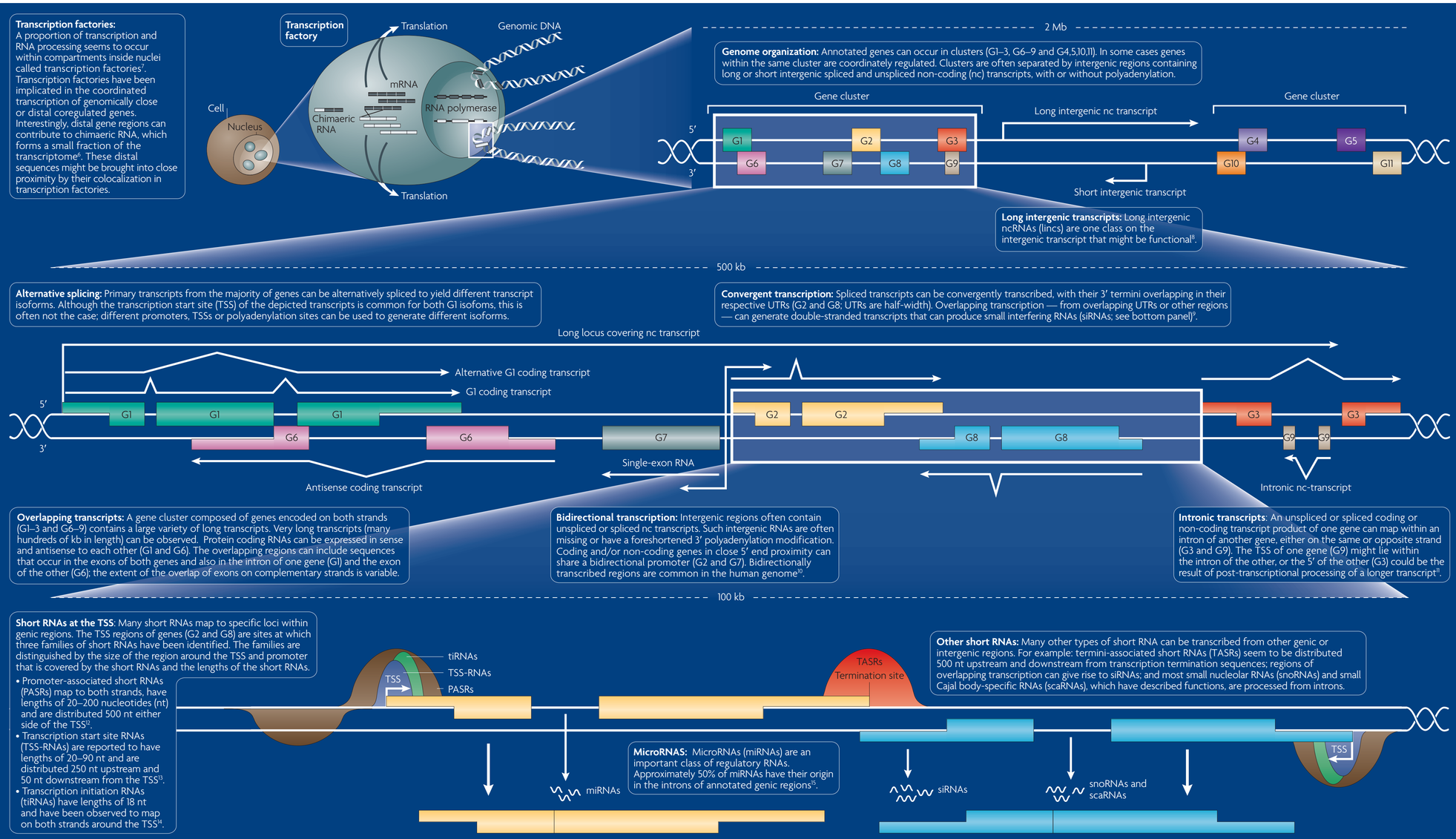
nature REVIEWS
GENETICS

The pervasive and interleaved transcriptome

Thomas R. Gingeras

Our image of the eukaryotic transcriptome has been transformed in recent years: it is now known that — from yeast to humans — almost the entire non-repetitive content of the genome can be transcribed. Not only is the extent of transcription greater than expected but evidence of the complexity and heterogeneity of the transcriptome — in terms of the types of RNA and their regulation — has increased. This new understanding is in large part thanks to technological developments, particularly deep sequencing and tiling arrays, that allow the products of the genome to be characterized at high resolution and without reliance on genome annotation.

However, this complexity is organized. Viewing the composition and organization of the eukaryotic transcriptome at multiple levels of resolution highlights the range of types and sizes of RNAs that are produced and how transcription relates to genome structure. These scales can range from the full genome, which reveals that RNAs can be produced that contain sequences from different chromosomes, to the scale at which long RNAs that cover hundreds of thousands of nucleotides can be seen, down to the level of transcripts made within individual gene regions, including RNAs that are only a few nucleotides in length.



Pervasive transcription

Several studies, including recent analyses performed during the ENCODE project, indicate that eukaryotic transcription is pervasive, with almost the full length of non-repeat regions of the genome being transcribed[1]. Although the extent of transcription of repeat regions is less well understood, recent data suggest it is developmentally regulated and functional[2]. Polyadenylated RNAs transcribed by RNA polymerase II are the most intensely studied transcriptional product, but non-polyadenylated RNAs are also made by all three major eukaryotic RNA polymerases.

Non-coding RNAs

Although the majority of the most abundantly expressed RNAs contain protein coding information, most of the newly identified transcription is composed of RNAs with reduced protein coding potential. However, many functional roles for non-coding RNAs (ncRNAs) have now been documented — roles that extend well beyond acting as an information intermediary between a genome and the translational machinery. Several characteristics that underscore their expected biological functionality can be identified for many members of this increasing repertoire of previously unannotated ncRNAs. These features include: the evolutionary conservation of their promoters, splice junction sequences and secondary structures; the regulated patterns of expression during development and perturbation of these patterns in disease states; and localization and association with proteins that are specific to particular subcellular compartments[3].

Complexity and organization

Recent deep RNA sequence characterizations also point to a greater complexity in the organization of information contained within genic and intergenic segments of the genome. The structure of a gene first described by Jacob and Monod has now given way to a genome with multiple transcripts covering each genic region. The ENCODE project has reported 5.4 transcripts per gene locus in the 1% of the human genome analysed by this project[1]. The sequences of these transcripts indicate that more than half of the transcripts at each gene locus seem to be ncRNAs. In addition to there being multiple transcripts in each genic region, the organization of these transcripts in a locus seems to entail a substantial amount of overlap between some RNAs. Other transcripts in the same locus can have transcription start sites (TSSs) that are distinct and distal from one another. Interestingly, the fate of many of the transcripts originating from within genic regions or from intergenic regions is to be processed into shorter RNAs. Many of the most important regulatory short RNAs — including microRNAs (miRNAs), small interfering RNAs (siRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), small Cajal-body specific RNAs (scaRNAs) and piwi-interacting RNAs (piRNAs) — are derived from long coding and non-coding RNAs[4,5].

At the opposite end of the genome scale to the detailed analyses of genic regions, long chimaeric RNAs have been discovered. The observation that these RNAs are sometimes composed of portions of genomic sequences separated by long genomic distances or even on different chromosomes suggests that the information encoded in genomes can be used in a non-colinear fashion[6]. Although the mechanism of formation and the biological importance of chimaeric RNAs remains unclear, their characteristics and prevalence throughout the evolution of eukaryotic cells make them an attractive target for future studies.

The diversity of types and functions of RNAs that transcriptomic studies has now uncovered points to a more elegant, complex and richer reservoir of genetic information being stored and processed from eukaryotic genomes than could previously have been imagined.

Enabling technologies

RNA deep sequencing (RNA–seq): This technology provides unique insights into the composition and characteristics of transcriptomes. The complexity of information encoded by eukaryotic genomes — for example, the products of each strand and non-colinear RNAs — can be captured by specialized application of RNA–seq.

Tiling arrays: Unbiased RNA mapping using high resolution tiling arrays allows the easy and inexpensive surveying of entire genomes. However, as arrays cannot provide some detailed information concerning RNA structure (for example, exon–exon junctions), they are most useful as a complementary technology to RNA–seq for experiments with specific goals.

References

[1]ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007) | [2]Mercer, T. R. *et al.* Long non-coding RNAs: insights into functions. *Nature Rev. Genet.* **10**, 155–159 (2009) | [3]Mattick, J. The genetic signatures of noncoding RNAs. *PLoS Genet.* **5**, 1–12 (2009) | [4]Kim, Y.-K. & Kim, V. N. Processing intronic miRNAs. *EMBO J.* **26**, 775–783 (2007) | [5]Weber, M. J. *et al.* Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet.* **2**, 1984–1997 (2006) | [6]Gingeras, T. R. Implications of chimaeric non-co-linear transcripts. *Nature* **461**, 206–211 (2009) | [7]Iborra, F. J. *et al.* Active RNA polymerases are localized within discrete transcription 'factories' in human nuclei. *J. Cell Sci.* **109**, 1427–1436 (1996) | [8]Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–237 (2009) | [9]Okamura, K. *et al.* Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in *Drosophila melanogaster*. *Nature Struct. Mol. Biol.* **15**, 581–590 (2008) | [10]Trinklein, N. D. *et al.* An abundance of bidirectional promoters in the human genome. *Genome Res.* **14**, 62–66 (2004) | [11]Fejes-Toth, K. *et al.* Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009) | [12]Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007) | [13]Seila, A. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008) | [14]Taft, R. J. *et al.* Tiny RNAs associated with transcription start sites in animals. *Nature Genet.* **41**, 572–578 (2009) | [15]Kim, V. N. *et al.* Biogenesis of small RNAs in animals. *Nature Rev. Mol. Cell. Biol.* **10**, 126–139 (2009)

Acknowledgements

Thomas R. Gingeras leads the Functional Genomics Group at Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. This work is supported by National Human Genome Research Institute grants (U54 HG004557 and U01 HG004271) and the suggestions and efforts of the Gingeras laboratory and collaborators.

Edited by Mary Muers; copy-edited by Lewis Packwood; designed by Patrick Morgan.

http://www.nature.com/nrg/posters/transcriptome