Published in final edited form as:

Nature. 2009 September 10; 461(7261): 206–211. doi:10.1038/nature08452.

Implications of chimeric non-collinear transcripts

Thomas R. Gingeras

Cold Spring Harbor Laboratory, 1 Bungtown Rd., Cold Spring Harbor, New York, 11724

Thomas R. Gingeras: gingeras@cshl.edu

Preface

Recent deep sequencing of transcriptomes from worm to human reveals that individual transcripts can be composed of sequence segments that are not collinear — with some mapping great distances apart and others to other chromosomes. Some of these chimeric transcripts are formed by genetic rearrangements but others appear to arise during post-transcriptional events. While in lower eukaryotes, this is accomplished by a well characterized trans-splicing process, in higher eukaryotes the processes leading to their formation remains unclear. While the biological importance of most chimeric RNAs is unclear as yet, the implications of their existence to the potential information content and functional organization of genomes are profound.

Projects like the Encyclopedia of DNA Elements (ENCODE) ¹, model genome ENCODE (modENCODE)², FANTOM ³, NIH Roadmap Epigenome ⁴ as well as efforts from many individual laboratories have contributed to an increasing appreciation of the pervasive transcription occurring in most genomes and the multiple functional roles that RNAs play within cells ^{5,6}. Part of the goals of these consortia has been to map and to characterize comprehensively the transcriptomes found in human and a diverse collection of model organisms. The relatively rapid development of powerful high throughput sequencing ⁷ and *in vivo* screening techniques ⁸ have provided the means to accomplish these goals in a very short time. A consistent finding in all of these studies has been that non-protein coding RNAs comprise most of the transcriptional output of genomes and many of the functional systems operating within cells involve non-coding transcripts ^{9,10}. Additionally, the mapping and sequence characterization of the RNAs involved in many of these cellular processes have resulted in interesting observations concerning the organization of information stored in genomes, the regulation of expression of these RNAs, novel biochemical processes that point to a possible function and fate of new RNA classes and the evolutionary implications highlighted by increased complexity in genome organization. In turn, the involvement of RNAs in each of these areas has prompted a reconsideration of the roles of RNA in inheritance and disease ^{11,12}.

Organization of Information Encoded in Genomes

In eukaryotic cells the central storage units of heritable information are compartmentalized in the nucleus as DNA and as epigenetic marks etched on to the DNA itself and the associated chromatin proteins. RNA's functional role has been seen to be primarily an intermediary that faithfully transfers information from the genome. Not surprisingly, functional roles like assisting in the synthesis of proteins by acting as a messenger (mRNA) or as a scaffold for protein synthesis (rRNAs) and assisting to gather amino acids (tRNA) for

protein synthesis have encouraged the view that information stored in a genome is transferred to RNA in a collinear fashion. This means that the nucleotide sequences found in RNA transcripts are ordered in the same linear fashion as that found in the DNA genome. While the discovery of splicing has provided a more modular and non-contiguous view of this collinear relationship, the order of sequences in both DNA and RNA has been maintained (Figure 1A). This collinear organization seems logical and efficient given the perceived primacy of DNA in the genetic hierarchy. Additionally, underlying this collinear organization of information in the genome and its transfer to RNA is the premise that the sequences that will be joined together in the mature RNAs reside on the same precursor RNA molecule. Indeed this seems to be the primary path of RNA processing from primary to mature transcripts.

However, structural studies of RNAs in several species have revealed that the sequences that are ultimately joined together on the same mature transcript can be encoded in separately transcribed RNAs with multiple distinct genomic origins (Figure 1B). Individual RNAs can be transcribed on separate chromosomes (Figure 1B, Form 1), on the same chromosome but with a different the genomic order from that found in the mature RNA (Figure 1B, Form 2), on the same chromosome but transcribed from different strands (Figure 1B, Form 3), or finally on the same chromosome but from different alleles (Figure 1B, Form 4).

One of the first observations that supported the joining of sequences derived from separate molecules was the discovery of trans-splicing. Discovered first in trypanosomes¹³ the process was later shown to occur also in *Caenorhabditis elegans* and other nematodes¹⁴, in *Euglena*¹⁵, in flatworms^{16; 17} and higher eukaryotes¹⁸. In the cases of the worm and trypanosomes, short stretches of nucleotides (nt) mapping to separate and distal regions in their genomes are trans-spliced onto the 5' ends of many protein coding genes. These short sequences are derived from short leader RNA genes (SL genes) that can be positioned thousands of nucleotides away, upstream or downstream (5' or 3', respectively) or on different chromosomes from the genes to which their sequences are spliced. Additionally, unlike cis-splicing observed in higher eukaryotic cells this trans-splicing involves the cleavage and joining of two separate transcripts.

Trans-splicing is not limited to worms and trypanosomes. Clear evidence of trans-splicing in mammalian cells is beginning to emerge. One of the most striking examples was reported in human cells by Li et al.¹⁹. Li et al. characterized a developmentally regulated trans-splicing event involving the 5' exons of the transcripts from the *JAZF1* gene on chromosome 7p15 and the 3' exons of *JJAZ1* (also known as *SUZ12*) located on chromosome 17q11. The fate of this chimeric RNA in endometrial stroma cells is to be translated into a chimeric anti-apoptotic protein. Interestingly, neoplastic stroma cells of the endometrium constitutively express the chimeric RNA and protein because of a translocation involving these two chromosomal regions (t(7;17 p15q11). Thus, by two different non-overlapping mechanisms similar chimeric RNA and proteins are made in normal and genetically rearranged neoplastic cells. The expression level of the chimeric RNA in the cells containing the translocation is elevated and this loss of regulation of the expression of the chimeric gene products is similar to oncogene regulatory mutations associated with neoplastic transformation²⁰. These studies also raise the possibility that the trans-splicing events may

be a precondition for chromosomal exchange and further suggest studies to determine if chimeric RNAs, which are often found as result of such structural mutations are present in the same non-transformed cells.

A second example of a clinically important trans-spliced RNA was observed in a study of a novel erythroblast transformation specific (ETS) family fusion transcript *SLC45A3-ELK4* (²¹). In this study, the chimeric *SLC45A3-ELK4* transcript was found to be expressed in normal and benign cancer prostate cells. Characterization of the fusion mRNA revealed a major variant in which *SLC45A3* exon 1 is fused to *ELK4* exon 2 leading to the expression of a novel protein. Based on quantitative PCR analyses of DNA, unlike other ETS fusions described in prostate cancer, the expression of *SLC45A3-ELK4* mRNA is not exclusive to the formation of chromosomal rearrangements. Similar to the *JJAZI* and *JAZF1* case both the chimeric RNA and protein can be formed by either in vivo trans-splicing or a genetic rearrangement event. The *SLC45A3-ELK4* chimeric RNA finding is notable for two clinical reasons. The first is that the *SLC45A3-ELK4* chimeric transcript can be detected at high levels in urine samples from men at risk for prostate cancer. Second, treatment of LNCaP cell line with R1881 synthetic androgen indicated that the fusion transcript was differentially regulated, making treatment with androgen antagonists of potential therapeutic value.

The increased prevalence of chimeric RNAs in other normal cells is supported by both clinical and empirical observations. Cancer geneticists have puzzled over why chimeric RNAs are detected in normal tissues used as controls for transformed cells. As with *JAZF1* and *JJAZI* genes, chimeric RNAs have been observed in spleen tissues of normal individuals involving the immunoglobulin heavy chain gene (*IGH*) and the *BCL2* gene. Interestingly a translocation (t14:18) involving the same genes is seen in neoplastic hematopoietic cells such as lymphomas²². Additionally, in other molecular biology studies designed to identify the 5' termini of genes expressed in many different tissues and cell lines as part of the pilot phase of the ENCODE, approximately 65% of the genes tested were observed to be involved in the formation of chimeric RNAs. These genes had distal unannotated 5' transcription start site (TSS) that were joined to RNAs from genes positioned hundreds of thousands of nucleotides 3' to these sites. In such cases, multiple intervening genes and their corresponding transcripts were not involved in the formation of the chimeric RNAs. While these results could be examples of very long distance collinear splicing, this is unlikely since some of the genes that part-take in the formation of the chimeric RNAs are encoded on the opposite strand. However, unlike the studies conducted by Li et al ¹⁹, these normal tissues were not tested for the presences of low level rearrangement events. Recently, two studies have suggested that detection of chimeric RNAs by deep sequencing of transcriptomes of individuals could provide early detection of chromosomal rearrangements found in cancer patients ²³. While the detection of chimeric RNAs may at times result from a genetic translocation event, the assumption that such rearrangements are present may prove to be misleading given that detection of chimeric RNAs arising from non-chromosomal rearrangements is possible.

These and other studies raise the intriguing question of whether chimeric RNAs that contain sequences from multiple non-collinear positions in the genome are more commonly found in normal tissues than previously appreciated. Evidence from the pilot phase of the ENCODE

project^{1,24,25} and other independent studies (see for example REF²⁶) support the prevalence of such chimeric RNA production in both normal tissues and established cell lines¹⁸. The pilot ENCODE studies reported that approximately 65% of the protein coding genes mapping within the boundaries of the ENCODE circumscribed regions contributed to the formation of chimeric RNAs, half of which contained sequences of genes that mapped to the opposite strand of the index gene used in these studies (data not published). It is important to note that the observation of chimeric RNAs involving a large percent of genes does not imply high copy numbers for such RNAs. However, as noted by Denoeud et al. the levels of expression of exons comprising chimeric RNAs are often similar to those of the annotated transcripts²⁴.

An unusual type of chimeric transcripts has been reported and is represented by chimeric RNAs that contained exon sequences that were out of order compared to the order of the exons found in the human genome. This type of non collinear representation of information in the genome is exemplified by a *SEC14L2* isoform in which upstream exons were inserted between downstream exons (Figure 2). Spliced portions of exons 3, 4 and 5 were inserted downstream of exons 6, 7 and 8 and connected to the 3' UTR of the downstream gene *HSPC242*²⁷. This arrangement preserves a large open reading frame while putting amino acids 56–93 of *SEC14L2* after amino acid 218 of this protein.

The molecular mechanisms controlling the joining of sequences from two individual RNA molecules as described in the cases described previously are still uncertain, although not without leads. Bruzik and Maniatis showed that RNA molecules containing a 3' splice site and enhancer sequence are efficiently spliced in *trans*.^{28,29} The products are RNA molecules containing normally *cis*-spliced 5' splice sites or *trans*-spliced as seen with the SL-RNAs from lower eukaryotes. Additionally, Li et al¹⁸ computationally identified chimeric RNAs in the RNA and EST databases of yeast, fly, mouse and human and have confirmed approximately 30% of them by RT-PCR reactions. Interestingly, they note that short homologous sequences (SHSs) at the junction sites of the chimeric RNAs are frequently observed and have curiously proposed a transcriptional slippage model rather than the a copy choice³⁰ or classic trans-splicing model³¹, to explain the generation of those chimeric RNAs synthesized from templates with SHSs.

Whatever the mechanism(s) that control the formation of chimeric RNAs, their utilization in eukaryotic cells is seen from worms to humans. Nevertheless, with the exception of some well studied cases as cited previously, no function has yet been identified for the large majority of observed chimeric RNAs.

Regulation of Expression of Non-Collinear Information

The organization of information in a genome and its non-collinear transfer into RNA has implications concerning the regulation of such transcriptional processes. If information transferred to RNA can be derived from sequences found in at least two different transcripts, then either the regulation of the synthesis of the individual transcripts that are involved in the trans-splicing must be coordinated or the half-life characteristics of at least some transcripts must allow for temporally uncoordinated regulation of their expression. The

coordinated regulation of expression of individual transcripts and the subsequent joining of these individual transcripts raises the possibility that these activities are carried out in close three-dimensional space in the cell.

Studies of nuclear organization in a cell indicate that non-ribosomal RNAs may be transcribed in sub-compartments or foci known as transcription factories³² and that these factories are stably maintained in the absences of active transcription³³. Transcription factories are sites enriched in RNA polymerase II and may contain other processing factors needed to create messenger RNAs^{34,35}. Interestingly, for any single cell there appear to be fewer transcriptional factories than the number of expressed genes³⁶. This suggests that multiple genes share the same transcriptional machinery. Genes located *cis* and *trans* to one another may share the same factory, suggesting that very distal genes migrate to preassembled nuclear sites. Osborne et al. used fluorescent *in situ* hybridization to investigate the genes that were associated with transcription factories during immediate early (IE) gene induction in mouse B lymphocytes.³⁷ They found that the mouse *Myc* proto-oncogene on chromosome 15 is recruited to the same transcription factory as the highly transcribed immunoglobulin heavy chain (*Igh*) gene located on chromosome 12. A similar association is seen in human cells³⁷. Interestingly, human *Myc* and *Igh* are the most frequent translocation partners in plasmacytoma and Burkitt lymphoma. These data suggest a direct link between the nonrandom association of two genes in an interchromosomal organization of transcribed genes located within transcription factories and the incidence of specific chromosomal translocation. While chimeric RNAs have not as yet been observed in normal plasmacytes, the 3 dimensional association of these two genes to the same transcriptional factory offers the possibility of the RNAs made from *Myc* and *Igh* to participate in a *trans*-splicing and possibly chromosomal translocation events, as seen with the *JAZF1* and *JJAZ1* and *SLC45A3* and *ELK4* genes described earlier. Such transcription factories capable of carrying out transcription and the formation of chimeric RNAs may be normal or specialized factories. Evidence for specialized factories has been described³⁶ of which the nucleolus provides an excellent example. Figure 3 shows a model for a hypothetical specialized factory. Such a model predicts a non-random correlation among the genomics sites that can be observed in close proximity in 3 dimensional space as detected by Carbon-Copy Chromosome Conformation Capture (5C)³⁸ and genomic regions encoding transcripts that are involved in the formation of chimeric RNAs.

Evolutionary Implications

Both the information content of genomes and the strategies to transfer this information from DNA to RNA has increased over the course of evolution. The transfer of information in eukaryote cells from worm to human has also been accompanied by increased segmentation of genomic information. This segmentation has allowed for permutations of sequences in an RNA context and the development of increasingly complex RNA processing mechanism(s) to produce permuted transcripts, as demonstrated by RNA splicing and *trans*-splicing. The segmentation of information in a DNA context and mechanisms to permute them in RNA space are potential areas of active selection. The possible use of multiple precursor transcripts and the ability to permute exon combinations found within them to form mature chimeric transcripts would have the advantage of increasing the informational content of

genomes dramatically and likely prompt an increase in the complexity of the processing machinery. In cases of a single primary transcript, this increase in complexity can be expressed by the formula in Box 1 and illustrated in Figure 4A. However, in the case of two or more primary transcripts contributing to a mature processed RNA in a non-collinear and permuted fashion, the total number of the number of possible permuted transcripts can be expressed as seen in formula 2 in Box 1, illustrated Figure 4B and extrapolated seen in the plot depicted in Figure 4C

Box 1

Potential diversity as a consequence of non-collinearity transcripts [Au: OK?]

The total number of possible collinear combinations can be expressed as

$$N(L, m) = \frac{L!}{(L-m)!m!},$$

where L represents the total number of possible exons in the primary transcript, while m is the number of exons that are found in the mature transcript. Thus, from a single primary transcript containing possible 4 exons, ($L = 4$) the number of mature transcripts that have 2 exons is 6 (Figure 4A). If all possible collinear transcripts were to be made containing 1–4 exons, there would be a total of 15 possible transcripts that could be made, or, generally:

$$N(L) \equiv \sum_{m=1}^L N(L, m) = 2^L - 1.$$

If two or more primary transcripts contribute to a mature processed RNA in a non-collinear and permuted fashion, the total number of the number of possible permuted transcripts can be expressed as [Au: OK?]

$$N(L, m) = \frac{L!}{(L-m)!},$$

where L again represents the total number of exons and m represents the number of exons in the mature transcripts. Thus, if two primary transcripts possessing 2 and 3 exons ($L = 5$) are used to create a 3-exon mature transcript ($m = 3$), a total of 60 permuted transcripts can be made (Figure 4B). If all transcripts using from 1–5 exons can be made, a total of 325 permuted transcripts could be observed. The dependence of the total number of mature transcripts on the number of possible exons for the non-collinear permuted case diverges exponentially from the collinear case (Figure 4C).

Using DNA as a reference for evolutionary studies

The potential utilization of multiple precursor transcripts to form a single mature chimeric RNA product and possibility of permuting the order of the genomic sequences in chimeric transcripts raises two points with evolutionary implications. The first concerns the identification of a complete set of both functional sequence segments and the evolutionarily constrained sequences. Constrained genomic sequences are often identified by the alignment of orthologous sequences using one of several sequence analyses programs^{39,40,41,42,43,44}. Based on such alignments functionality is often inferred. However, if functional RNA sequence regions can be made from modules of non-collinear sequences that arise by permutation of genomic sequences that are encoded in the DNA short evolutionary constrained or novel functional regions of genomes may be joined together in RNA space, the product of which is not observable by analysis of genomic sequences. Such novel RNA encoded sequences would likely be missed by programs designed to identify constrained sequences in DNA.

A second point is related to the first. It concerns the use of genomic DNA as the sole reference to identify the sequences that manifest the effects of evolutionary pressures. Using DNA as a reference to study evolutionary pressures on sequences naturally stems from its role in inheritance. However, a focus on DNA as the sole subject to catalogue genetically transmitted functional elements and the only molecule to analyze for evolutionary constraints ignores the growing number of acknowledged functional roles that RNA contributes. It also potentially misses the novel RNA encoded sequence elements that are often used to carry out these functions. While the identification of functional regions encoded in a collinear fashion is relatively straightforward for both DNA and RNA, the same can not be said for those functional regions that are created from non-collinear segments. Unless these sequence regions have been catalogued previously (e.g. SL1 and SL2 transcripts in *C. elegans*) or the rules of how distal non-collinear sequences are to be joined in a transcript are known, then analyses of the transcriptome in addition to an organism's genome may be required to catalogue all of the functional regions subject to selection.

Interpreting genetic variation

Genetic variation observed from the single nucleotide to the gross chromosomal rearrangement levels are the basis of evolutionary change. Two biases influence the interpretation of the effects of such genetic variation. The first is that the effect is most often interpreted to be related to the nearest gene annotation. Thus, a variation occurring in an exon of a gene is often interpreted as causing a malfunction of that gene leading to an abnormal phenotype. The second bias is that the effect of the variation is associated in the context of the extended but local genetic locus. If a genetic variation is observed to be proximal but not within an exon of a gene, the effects are most often interpreted to be caused by mis-regulation of a gene located nearby. Such interpretations of genetic variations are common and often well supported. However, current genetic association studies appear to be widening the range of study to include long distance trans effects as is seen with SNPs mapping hundreds of thousands of base pairs associated with urinary bladder and prostate cancers^{45,46}. This is prompted by the recent observation that almost half of all

polymorphisms found statistically correlated with complex diseases or traits are located at a non-annotated and distal genomic site (intron or intergenic region)⁴⁷. The joining of distal genomic regions by the formation of chimeric RNAs provides additional motivation for this broadening of genetic analyses.

Conclusions

Consideration of the non-collinear organization of genomes raises several questions. The first of these questions is why genomes resort to a non-collinear organizational strategy of stored information. While it is clear that the relatively straightforward collinear organization of functional information in genome is likely to be utilized as the most common strategy (Figure 1A) it is also true that utilization of a non-collinear organization appeared early in eukaryotic genome evolution as exemplified by worms and trypanosomes. The advantages of using such a complex strategy include providing a means of increasing the information content of genomes (Figure 4C) and allowing for the possible new combinations of exons, operating in a relatively redundant fashion (motif sharing) and in turn functioning in more significant roles (e.g. formation of new spliced mRNAs allowing for increased protein diversity). Additionally, the use of non-collinear organization leading to the formation of chimeric RNAs might also allow for the real-time monitoring of transcripts of co-regulated RNAs. Such a function would lead to the prediction, which remains to be tested, that chimeric RNAs would be formed from genes that possessed similar expression profiles observed over development or in response to some external stimuli.

A second question concerns the observed relatively low level expression of chimeric transcripts as illustrated by the infrequent reports from the many cDNA library studies concerning their existence. Is this because chimeric RNAs are rarely used by cells or because they are non-functional transcription and comprise a general RNA processing background? The answer to this question is unclear at the moment. The lack of multiple reports of biologically functional chimeric RNAs in additional organisms other than the worm, trypanosomes and a few examples in higher eukaryotes may stem from either the fact they have often been observed in cDNA studies but discarded because of the prevalent genome organizational models or that they are not observed owing to either the shallow depth of sampling of the transcriptome or the method of analyses (e.g. arrays).

It would seem that the non-collinear organization of information in genomes is used more routinely than previously thought. This strategy is seen in organisms in which trans-splicing is integral (lower eukaryotes) and in higher eukaryotes in which their presence is observed but the importance of which is as yet unclear. While the strategic and evolutionary advantages of this organization of information are interesting there are many questions that remain. As with the ongoing discussions concerning the functional importance of the prevalence of non-protein coding transcription, only additional experiments will lead to answers to these questions.

Acknowledgments

Work in my laboratory is supported by NHGRI (U54 HG004557 and U01 HG004271). Thanks to P. Kapranov for discussions and long term collaboration, H. Sussman for helpful discussions and editing of the manuscript and A.

Dobin for assistance in deriving the computational expressions to determine the collinear and non-collinear permutations and the plot depicted in Figure 4C.

References

1. Consortium EP, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*. 2007; 447:799–816. Summary of ENCODE analyses to identify functional regions in 1% of the human genome. [PubMed: 17571346]
2. Celniker SE, et al. Unlocking the secrets of the genome. *Nature*. 2009; 459:927–930. 459927a [pii]. 10.1038/459927a [PubMed: 19536255]
3. Carninci P, et al. The transcriptional landscape of the mammalian genome. *Science*. 2005; 309:1559–1563. 309/5740/1559 [pii]. 10.1126/science.1112014 [PubMed: 16141072]
4. Jones PA, Archer TK, Baylin SB, Beck S, Berger S, Bernstein BE, Carpten JD, Clark SJ, Costello JF, Doerg RE, Esteller ME, Feinberg AP, Gingeras TR, Grealis JM, Henikoff S, Herman JG, Jackson-Grusby L, Jenuwein T, Jirtle RL, Kim Y-J, Laird PW, Lim B, Martienssen R, Polyak K, Stunnenberg H, Tlsty TD, Tycko B, Ushijima T, Zhu J, Pirrotta V, Allis CD, Elgin S, Jones PA, Rine J, Wu C. Moving AHEAD with an international human epigenome project. *Nature*. 2008; 454:711–715. [PubMed: 18685699]
5. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nat Rev Genet*. 2009; 10:155–159. nrg2521 [pii]. 10.1038/nrg2521 [PubMed: 19188922]
6. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. *Cell*. 2009; 136:629–641. S0092-8674(09)00142-1 [pii]. 10.1016/j.cell.2009.02.006 [PubMed: 19239885]
7. Hert DG, Fredlake CP, Barron AE. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*. 2008; 29:4618–4626. 10.1002/elps.200800456 [PubMed: 19053153]
8. Ramadan N, Flockhart I, Booker M, Perrimon N, Mathey-Prevot B. Design and implementation of high-throughput RNAi screens in cultured *Drosophila* cells. *Nat Protoc*. 2007; 2:2245–2264. nprot.2007.250 [pii]. 10.1038/nprot.2007.250 [PubMed: 17853882]
9. Amaral PP, Mattick JS. Noncoding RNA in development. *Mamm Genome*. 2008; 19:454–492. 10.1007/s00335-008-9136-7 [PubMed: 18839252]
10. Mattick JS, Amaral PP, Dinger ME, Mercer TR, Mehler MF. RNA regulation of epigenetic processes. *Bioessays*. 2009; 31:51–59. 10.1002/bies.080099 [PubMed: 19154003]
11. Cooper TA, Wan L, Dreyfuss G. RNA and disease. *Cell*. 2009; 136:777–793. S0092-8674(09)00148-2 [pii]. 10.1016/j.cell.2009.02.011 [PubMed: 19239895]
12. Ashe A, Whitelaw E. Another role for RNA: a messenger across generations. *Trends Genet*. 2007; 23:8–10. S0168-9525(06)00378-7 [pii]. 10.1016/j.tig.2006.11.008 [PubMed: 17126949]
13. Sutton RE, Boothroyd JC. Evidence for trans splicing in trypanosomes. *Cell*. 1986; 47:527–535. 0092-8674(86)90617-3 [pii]. [PubMed: 3022935]
14. Krause M, Hirsh D. A trans-spliced leader sequence on actin mRNA in *C. elegans*. *Cell*. 1987; 49:753–761. 0092-8674(87)90613-1 [pii]. [PubMed: 3581169]
15. Tessier LH, et al. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in *Euglena*. *EMBO J*. 1991; 10:2621–2625. [PubMed: 1868836]
16. Rajkovic A, Davis RE, Simonsen JN, Rottman FM. A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*. *Proc Natl Acad Sci U S A*. 1990; 87:8879–8883. [PubMed: 2247461]
17. Davis RE, et al. RNA trans-splicing in *Fasciola hepatica*. Identification of a spliced leader (SL) RNA and SL sequences on mRNAs. *J Biol Chem*. 1994; 269:20026–20030. [PubMed: 8051087]
18. Li X, Zhao L, Jiang H, Wang W. Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes. *J Mol Evol*. 2009; 68:56–65. Computational analyses of human, mouse and fly RNA databases to identify the presence of chimeric RNAs and the identification of short repeat sequences located at the junction sites forming the chimeric transcripts. 10.1007/s00239-008-9187-0 [PubMed: 19089307]
19. Li H, Wang J, Mor G, Sklar J. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science*. 2008; 321:1357–1361. 321/5894/1357 [pii] Identification of chimeric RNA

- and protein made from *JAZF1* and *JJAZ1* genes in normal and transformed endometrial cells. 10.1126/science.1156725 [PubMed: 18772439]
20. Eychene A, Rocques N, Pouponnot C. A new MAFia in cancer. *Nat Rev Cancer*. 2008; 8:683–693. [PubMed: 19143053]
 21. Rickman DS, et al. SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. *Cancer Res*. 2009; 69:2734–2738. 0008-5472.CAN-08-4926 [pii] Identification of chimeric RNA made from *SLC45A3-ELK4* genes in normal and transformed prostate cells. 10.1158/0008-5472.CAN-08-4926 [PubMed: 19293179]
 22. Janz S, Potter M, Rabkin CS. Lymphoma- and leukemia-associated chromosomal translocations in healthy individuals. *Genes Chromosomes Cancer*. 2003; 36:211–223. 10.1002/gcc.10178 [PubMed: 12557221]
 23. Maher CA, et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature*. 2009; 458:97–101. nature07638 [pii]. 10.1038/nature07638 [PubMed: 19136943]
 24. Denoeud F, et al. Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res*. 2007; 17:746–759. 17/6/746 [pii]. Identification and characterization of chimeric RNAs found in normal tissues and transformed cell lines mapping in the 1% of human genome analyzed by the ENCODE studies. 10.1101/gr.5660607 [PubMed: 17567994]
 25. Djebali S, et al. Efficient targeted transcript discovery via array-based normalization of RACE libraries. *Nat Methods*. 2008; 5:629–635. nmeth.1216 [pii]. 10.1038/nmeth.1216 [PubMed: 18500348]
 26. Parra G, et al. Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res*. 2006; 16:37–44. gr.4145906 [pii]. 10.1101/gr.4145906 [PubMed: 16344564]
 27. Kapranov P, et al. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res*. 2005; 15:987–997. 15/7/987 [pii]. 10.1101/gr.3455305 [PubMed: 15998911]
 28. Bruzik JP, Maniatis T. Spliced leader RNAs from lower eukaryotes are trans-spliced in mammalian cells. *Nature*. 1992; 360:692–695. 10.1038/360692a0 [PubMed: 1465136]
 29. Bruzik JP, Maniatis T. Enhancer-dependent interaction between 5' and 3' splice sites in trans. *Proc Natl Acad Sci U S A*. 1995; 92:7056–7059. [PubMed: 7624368]
 30. Coffin JM. Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses. *J Gen Virol*. 1979; 42:1–26. [PubMed: 215703]
 31. Hastings KE. SL trans-splicing: easy come or easy go? *Trends Genet*. 2005; 21:240–247. S0168-9525(05)00045-4 [pii]. 10.1016/j.tig.2005.02.005 [PubMed: 15797620]
 32. Jackson DA, Hassan AB, Errington RJ, Cook PR. Visualization of focal sites of transcription within human nuclei. *EMBO J*. 1993; 12:1059–1065. First description of nuclear sub-compartments in which RNA transcription occurs. [PubMed: 8458323]
 33. Mitchell JA, Fraser P. Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes Dev*. 2008; 22:20–25. 22/1/20 [pii]. 10.1101/gad.454008 [PubMed: 18172162]
 34. Pombo A, Cook PR. The localization of sites containing nascent RNA and splicing factors. *Exp Cell Res*. 1996; 229:201–203. S0014-4827(96)90360-2 [pii]. 10.1006/excr.1996.0360 [PubMed: 8986598]
 35. Pombo A, et al. Regional and temporal specialization in the nucleus: a transcriptionally-active nuclear domain rich in PTF, Oct1 and PIKA antigens associates with specific chromosomes early in the cell cycle. *EMBO J*. 1998; 17:1768–1778. 10.1093/emboj/17.6.1768 [PubMed: 9501098]
 36. Carter DR, Eskiw C, Cook PR. Transcription factories. *Biochem Soc Trans*. 2008; 36:585–589. BST0360585 [pii]. 10.1042/BST0360585 [PubMed: 18631121]
 37. Osborne CS, et al. Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. *PLoS Biol*. 2007; 5:e192. 06-PLBI-RA-2184 [pii]. 10.1371/journal.pbio.0050192 [PubMed: 17622196]
 38. Dostie J, Zhan Y, Dekker J. Chromosome conformation capture carbon copy technology. *Curr Protoc Mol Biol*. 2007; Chapter 21(Unit 21):14. 10.1002/0471142727.mb2114s80 [PubMed: 18265398]

39. Margulies EH, Program NCS, Green ED. Detecting highly conserved regions of the human genome by multispecies sequence comparisons. *Cold Spring Harb Symp Quant Biol.* 2003; 68:255–263. [PubMed: 15338625]
40. Margulies EH, et al. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci U S A.* 2005; 102:4795–4800. 0409882102 [pii]. 10.1073/pnas.0409882102 [PubMed: 15778292]
41. Margulies EH, et al. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* 2007; 17:760–774. 17/6/760 [pii]. 10.1101/gr.6034307 [PubMed: 17567995]
42. Brudno M, et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 2003; 13:721–731. doi:10.1101/gr.926603 GR-9266R [pii]. [PubMed: 12654723]
43. Cooper GM, et al. Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Res.* 2004; 14:539–548. doi:10.1101/gr.203470414/4/539 [pii]. [PubMed: 15059994]
44. Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 2005; 15:1034–1050. gr.3715005 [pii]. 10.1101/gr.3715005 [PubMed: 16024819]
45. Kiemeny LA, et al. Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat Genet.* 2008; 40:1307–1312. ng.229 [pii]. 10.1038/ng.229 [PubMed: 18794855]
46. Yeager M, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet.* 2007; 39:645–649. ng2022 [pii]. Identification of SNP containing loci mapping to intergenic regions hundreds of thousands of nucleotides upstream from MYC gene with statistically strong association in urinary bladder cancer cases. 10.1038/ng2022 [PubMed: 17401363]
47. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science.* 2008; 322:881–888. 322/5903/881 [pii]. 10.1126/science.1156409 [PubMed: 18988837]

Alignments of Information in DNA and RNA

A. Co-linear Alignments

B. Non-Co-linear Alignments

Figure 1. Models of possible organization of information contained in DNA and its transfer to RNA

(A) Collinear alignments can be categorized in two forms. Directly collinear in which information (sequence) is transferred in an un-interrupted fashion to RNA as is seen in most bacterial mRNAs and modular alignments in which the information is transferred to RNA in a collinear but interrupted (by introns) fashion. (B) Non-collinear alignments can be categorized in 4 forms. The production of precursor RNAs is shown only for Form 1 but is made in all forms. Form 1 represents the formation of chimeric RNAs from two different loci in a genome at two different genic regions. The precursor RNAs are processed (e.g. by trans-splicing or RNA recombination¹⁸) into the chimeric RNA. Form 2 exemplifies the formation of a chimeric RNA from the same gene by rearrangement of exons (see Figure 2 for example). Form 3 illustrates the formation of a chimeric RNA which is made from two precursor RNAs that are transcribed from opposite strands. Form 4 describes the formation of a chimeric RNA that is made from transcripts derived from two alleles of the same gene one of which contains a single nucleotide polymorphism.

Figure 2. Characterization of a chimeric transcript

The example shows a RT-PCR product corresponding to a chimeric transcript derived from SEC14L2 gene where a spliced region corresponding to exons 4, 5 and 6 is inserted downstream of exon 8.²⁷ A product of 3' RACE/array reaction is shown, with a primer positioned in exon 6; the nonsequential RT-PCR product is shown aligned to the genome. Note the RACE/array signal upstream from the RACE primer corresponding to exons 4 and 5. A portion of exon 3, which consisted of only 7 bp, is too small to hybridize to the array.

Figure 3. Model of specialized transcription factory in which transcription and formation of chimeric RNAs are carried out

Based on the observation that there are fewer transcription factories in nuclei than the number of transcribed loci and the existence of specialized transcription factories³⁶, this model hypothesizes that genes A and B encoded in different regions of the genome are collected in a transcription factory and transcribed into primary transcripts by multiple polymerase II transcriptional complexes (white oval). The primary transcripts are then processed to form mature spliced and chimeric RNAs. In this model most of the primary RNAs are involved in cis-splicing and transported to the cytosol for translation. Consistent with steady state estimates of chimeric RNAs levels, a smaller proportion of the primary RNAs are used to create chimeric RNAs. The model can envision a single or multiple isoforms of chimeric RNAs being made from combinations of primary RNA having been transcribed within the same transcription factory. The occurrences of translocation events involving genomic regions that are transcribed to produce chimeric RNAs raises the interesting possibility that such rearrangements may also be facilitated at these specialized factories.

Figure 4. Collinear and non-collinear combinations of information modules of information

(A) Six possible two-exon combinations derived from a four exon genic region are shown.

(B) Five of the possible 350 non-collinear, permuted three-exon transcripts that can be made

from two 2 and 3 exon containing transcripts are shown. (C) A semi-log plot comparing the

number of transcript isoforms that can be created using 1 to 10 exons considering a non-

permuted collinear organization of information (blue) or a permuted non-collinear

organization of information (black). The dependence of the total number of mature

transcripts on the number of possible exons for the non-collinear permuted case diverges

exponentially from the collinear case.