Commentary

Studying chromosome-wide transcriptional networks: new insights into disease?

Philipp Kapranov

Address: Helicos BioSciences Corporation, One Kendall Square Bldg 700, Cambridge, MA 02139, USA. E-mail: philippk08@gmail.com

Published: 11 May 2009

Genome Medicine 2009, **1**:50 (doi:10.1186/gm50)

The electronic version of this article is the complete one and can be found online at http://genomemedicine.com/content/1/5/50

Abstract

A large amount of experimental data collected over the last decade has shown that genomic organization is very complex and has highlighted the fact that the current set of gene annotations does not fully capture this complexity. Much of the RNA detected in a cell is found to originate from outside the exons of annotated genes. Exons of annotated and unannotated transcripts separated by large genomic distances can be joined together in chimeric transcripts. Any given base-pair in a genome could be traversed by many protein-coding and non-coding RNAs. We discuss the implications of these effects for our understanding of disease.

The interpretation of sequence polymorphism data, such as the data produced in large amounts from genome-wide association studies, is largely based on the concept of a gene as a stand-alone, separate genomic entity with discrete start and end, as defined by the current genomic annotations. The immediate logical corollary of this notion is that the effect of a nucleotide change is most likely to be local, or at least within the locus in which the change was found. However, surveys aimed at an unbiased cataloguing of the transcripts produced by human and other genomes, such as [1-7], challenge the notion of a gene as a separate, discrete genomic unit. This, in turn, may affect the interpretation of any nucleotide change that is found to be associated with a certain phenotype or a disease. Following the results of such surveys of transcriptional output of human and other genomes [1-7], the concept of a gene has expanded in several directions.

First, a multitude of different transcripts are made at any given locus. Analysis of the existing expressed sequence tag (EST) data suggests that a protein-coding locus can produce at least 5.7 different transcripts [1,8]. Although only some of these alternative transcripts seem to have protein-coding capacity, this expands the number of transcripts that a given exon can participate in. Logically, a nucleotide change in a shared exon could affect any of the transcripts that share it, and thus the phenotypic effect of a nucleotide change is likely to be represented as a sum of the effects on the transcripts that express it. It is likely that the profile of expressed transcripts is different in each tissue [9], and the effect of a nucleotide change could thus differ depending on the repertoire of transcripts expressed by the locus in each cell. In a simple example, as shown for the annotated transcripts in Figure 1, the phenotype may show itself in a tissue that expresses an exon overlapping the variant and not in another tissue that expresses transcripts that skip that exon. In a more complex case depicted in Figure 1, a polymorphic nucleotide or stretch of nucleotides could be part of a coding exon in one tissue and a non-coding exon in another; or it could represent both a regulatory region of one group of transcripts and an exon of another group of transcripts. Even more complex scenarios are possible considering that a large number of different isoforms could be expressed in any given cell type.

Second, the annotation of genomic regions that are considered exonic is incomplete. Unbiased studies using rapid amplification of cDNA ends (RACE) on the genes within the 1% of the genome chosen for the ENCODE project have shown that almost half the exons detected in these experiments do not overlap annotated exons [1]. Thus, a nucleotide

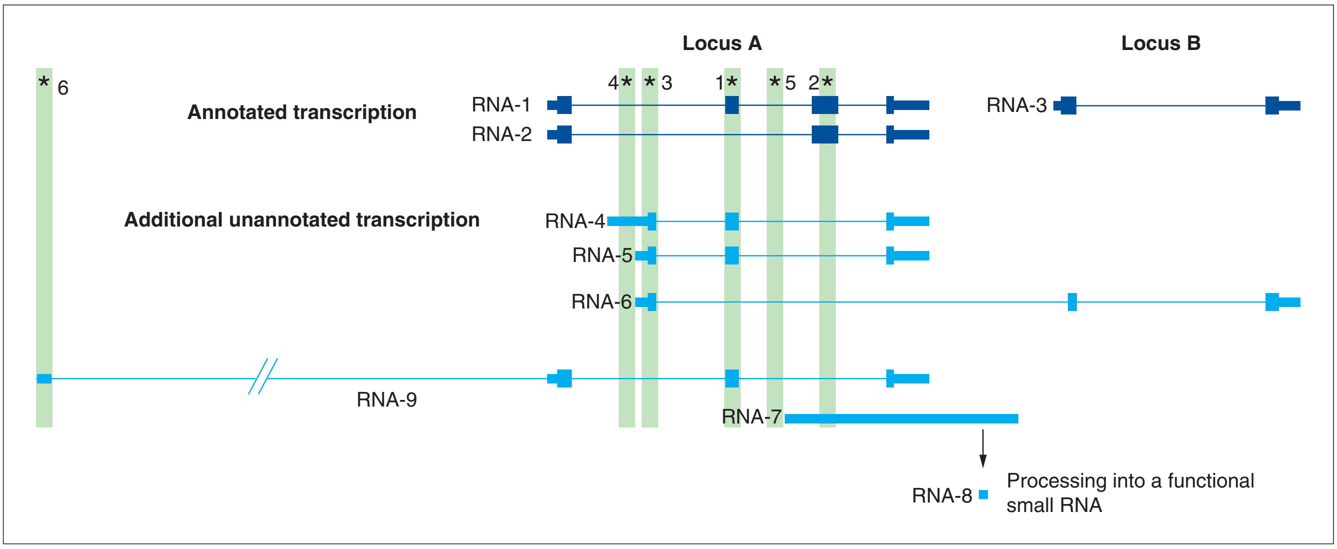

Figure 1

Examples of the potential effects of different sequence polymorphisms on two hypothetical loci, A and B. In this scenario, locus A has two annotated transcripts (RNA-1 and RNA-2, dark blue), expressed in different tissues. Sequence polymorphism 1 would affect an annotated exon of locus A that occurs in the annotated transcript RNA-1 and in unannotated transcripts (RNA-4 and RNA-5, cyan) and not in RNA-2. Variant 2 would affect a coding exon that is present in both the annotated coding transcripts and also in the non-coding transcript RNA-7. These are examples of polymorphisms that would currently be considered to be the only likely 'functional' polymorphisms in locus A, as they are the only ones to affect the annotated transcripts, RNA-1 and RNA-2. Polymorphisms 3-6 are 'non-coding' polymorphisms, with polymorphism 6 being relatively distant from locus A. However, in this example, these polymorphisms in fact overlap unannotated transcripts (cyan) within locus A, some of which extend outside locus A or encode regulatory small RNA molecules that act *in trans* on other loci. Polymorphism 3 overlaps a novel exon that is a part of unannotated transcripts RNA-4, RNA-5 and RNA-6. It could thus affect transcripts derived from both locus A and locus B, whether the two loci are nearby or distant in the genome. Polymorphism 4 overlaps a regulatory region for unannotated transcripts RNA-5 and RNA-6 and the 5' untranslated region of RNA-4. It could thus also affect expression of transcripts from both locus A and locus B. Polymorphism 5 overlaps a regulatory region for a non-coding RNA (transcript RNA-7) that is a precursor for a small RNA, a miRNA (RNA-8). Thus, this polymorphism and polymorphism 2, which also overlaps this non-coding RNA, could affect expression of other loci regulated by this small RNA *in trans*. Polymorphism 6 affects a more distant region in the genome that is connected to locus A by transcript RNA-9. All transcripts are shown transcribed from left to right; non-coding portions of transcripts are represented as thin boxes; coding portions are represented as thicker boxes; introns are shown as thin lines; asterisks indicated polymorphisms.

change in a 'non-coding' region may in fact underlie an as-yet undiscovered exon. Overall, 90% of all genes have been shown to have either a novel internal exon or a novel 5' exon in at least one of the 12 tissues tested [3].

In addition, the boundary of a gene may extend well beyond the current annotation. A gene can have many boundaries and, in fact, exons of different genes can participate in creating chimeric transcripts. The above-mentioned RACE experiments have shown that 68.4% of all genes had a 5' extension in at least one tissue tested [3]. Novel 5' exons were found to be represented both by novel, unannotated regions and by exons of other genes. Indeed, transcripts connecting exons of nearby loci and more distant loci separated by other genes on both strands were commonly found [1-3]. In fact, 57% of loci that were extended at the 5' end had a connection to an exon of an upstream gene [3]. A majority of 5' extensions (87%) reached over an annotated gene [3]. Often 5' extensions were tissue- or cell-line-specific, suggesting that in different tissues the profile of gene-gene connections could be different. Connections in the ENCODE regions could be identified only up to genomic distances of around 0.5 megabases (Mb). A continuation of these studies on human chromosomes 21 and 22 found a wealth of distant connections that span megabases of genomic space [2].

These observations raise several questions. What are the mechanisms responsible for the production of chimeric RNAs encoded by genes separated by very long genomic stretches? What are the functions, if any, of such chimeric RNAs and what are the implications of the uncovered connections (gene to gene or a novel distant exon to known gene) for cell biology and disease? So far, the answers to these questions remain unknown. However, copy number variants can affect the expression of distant genes located megabases away from the bounds of the variable region [10,11]. This shows that the effect of a genomic change does not have to be limited to the immediate vicinity of the change and could in fact result in both local and distant effects.

A third direction in which the concept of a gene has expanded results from the observation that transcripts emanating from any given locus could be carriers of *trans*-

acting non-coding RNAs, such as microRNAs (miRNAs) or small nucleolar RNAs (snoRNAs) [5,12-14]. Thus, a polymorphism affecting either the sequence or the processing of such an RNA molecule [15] could in fact affect the expression of loci regulated by the small RNA *in trans*, with potentially no effect on the locus in which the polymorphism was found, as shown in a hypothetical scenario in Figure 1. Such effects could be prevalent given that we now know the repertoire of the small, non-coding transcripts in a human cell to be far greater than the annotated classes of known small RNAs, and that such novel small RNAs could be carried by long RNA precursors [16,17].

Overall, these observations suggest that the identification of a sequence variant should not be the logical end point that automatically connects the locus that harbors it with a phenotype, but rather a beginning of a set of experimental procedures to unravel the effects of the variant. A necessary prerequisite for such experiments is unraveling the complexity of transcripts that either include the variant or originate nearby, because the variant also could affect a regulatory region of a novel transcriptional unit. Considering the vast number of unannotated transcripts present in a cell, it is important to directly characterize transcript complexity, for example using RACE with oligonucleotides positioned in or around the polymorphism in the biological samples of interest, rather than relying solely on the existing genomic annotations. One can envisage such analysis to be followed by expression profiles to estimate the effects of a sequence variant on all transcripts that it can be associated with, including the ones that could connect it to distant regions in the genome. Such experiments could be followed by direct perturbation of the candidate transcripts by knockdown or overexpression to estimate their contribution to a phenotype.

In addition to aiding our interpretation of sequence polymorphism data, the wealth of novel transcripts found in the human genome, including the chimeric RNAs that connect together distant regions in the genome, is mostly a virgin territory for biomarker discovery. Unannotated transcripts tend to be cell-type-specific [3,18] and thus should be attractive diagnostic molecules. The potential of non-coding RNAs as biomarkers has been shown by Reis *et al.* [19,20]; however, this field remains mostly unexplored because of the emphasis on annotated protein-coding transcripts. Furthermore, novel protein-coding transcript isoforms, specifically those of transcripts encoding proteins amenable to small molecule modulation, could be additional targets for small molecule therapeutics. In this respect, the high cell-type specificity of novel transcripts should provide an advantage: inhibition of a protein encoded by these transcripts is likely to be specific to a tissue or a cell type within a tissue, and thus is less likely to have side effects than the targets designed to the annotated forms of these proteins, which are likely to be the most constitutive isoforms. This calls for a systematic analysis directed at obtaining a full transcript repertoire of such a ‘druggable’ transcriptome in a diverse set of cell types and tissues using highly sensitive technologies, for example RACEarray [2,3,21].

Abbreviations

EST, expressed sequence tag; RACE, rapid amplification of cDNA ends.

Competing interests

The author is an employee and stockholder of Helicos BioSciences Corporation.

References

1. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, *et al.*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**:799-816.
2. Djebali S, Kapranov P, Foissac S, Lagarde J, Reymond A, Ucla C, Wyss C, Drenkow J, Dumais E, Murray RR, Lin C, Szeto D, Denoeud F, Calvo M, Frankish A, Harrow J, Makrythanasis P, Vidal M, Salehi-Ashtiani K, Antonarakis SE, Gingeras TR, Guigó R: **Efficient targeted transcript discovery via array-based normalization of RACE libraries.** *Nat Methods* 2008, **5**:629-635.
3. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, Dike S, Wyss C, Henrichsen CN, Holroyd N, Dickson MC, Taylor R, Hance Z, Foissac S, Myers RM, Rogers J, Hubbard T, Harrow J, Guigó R, Gingeras TR, Antonarakis SE, Reymond A: **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.
4. Gingeras TR: **Origin of phenotypes: genes and transcripts.** *Genome Res* 2007, **17**:682-690.
5. Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization.** *Nat Rev Genet* 2007, **8**:413-423.
6. Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE, Guigo R: **Tandem chimerism as a means to increase protein complexity in the human genome.** *Genome Res* 2006, **16**:37-44.
7. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, *et al.*: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309**:1559-1563.
8. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R: **GENCODE: producing a reference annotation for ENCODE.** *Genome Biol* 2006, **7(Suppl 1)**:S4.
9. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456**:470-476.
10. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET: **Relative impact of nucleotide and copy number variation on gene expression phenotypes.** *Science* 2007, **315**:848-853.
11. Merla G, Howald C, Henrichsen CN, Lyle R, Wyss C, Zabot MT, Antonarakis SE, Reymond A: **Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes.** *Am J Hum Genet* 2006, **79**:332-341.

12. Storz G, Altuvia S, Wassarman KM: **An abundance of RNA regulators.** *Annu Rev Biochem* 2005, **74:**199-217.
13. Baskerville S, Bartel DP: **Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes.** *RNA* 2005, **11:**241-247.
14. Kiss T: **Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions.** *Cell* 2002, **109:**145-148.
15. Borel C, Antonarakis SE: **Functional genetic variation of human miRNAs and phenotypic consequences.** *Mamm Genome* 2008, **19:** 503-509.
16. Affymetrix ENCODE Transcriptome Project; Cold Spring Harbor Laboratory ENCODE Transcriptome Project: **Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs.** *Nature* 2009, **457:**1028-1032.
17. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316:**1484-1488.
18. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22.** *Genome Res* 2004, **14:**331-342.
19. Reis EM, Nakaya HI, Louro R, Canavez FC, Flatschart AV, Almeida GT, Egidio CM, Paquola AC, Machado AA, Festa F, Yamamoto D, Alvarenga R, da Silva CC, Brito GC, Simon SD, Moreira-Filho CA, Leite KR, Camara-Lopes LH, Campos FS, Gimba E, Vignal GM, El-Dorry H, Sogayar MC, Barcinski MA, da Silva AM, Verjovski-Almeida S: **Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer.** *Oncogene* 2004, **23:**6684-6692.
20. Reis EM, Ojopi EP, Alberto FL, Rahal P, Tsukumo F, Mancini UM, Guimarães GS, Thompson GM, Camacho C, Miracca E, Carvalho AL, Machado AA, Paquola AC, Cerutti JM, da Silva AM, Pereira GG, Valentini SR, Nagai MA, Kowalski LP, Verjovski-Almeida S, Tajara EH, Dias-Neto E, Bengtson MH, Canevari RA, Carazzolle MF, Colin C, Costa FF, Costa MC, Estécio MR, Esteves LI, *et al.*: **Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck, and thyroid cancer.** *Cancer Res* 2005, **65:**1693-1699.
21. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR: **Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays.** *Genome Res* 2005, **15:**987-997.