Genome Biology

RESEARCH HIGHLIGHT

Annotating conserved and novel features of primate transcriptomes using sequencing

Philipp Khaitovich[1,2]*

See research article: http://genomebiology.com/2010/11/7/R78

Abstract

Recent high-throughput sequencing of chimpanzee brain and liver transcriptomes published in *Genome Biology* reveals multiple transcripts lost in the human genome and highlights the incompleteness of primate genome annotations.

The completion of the human genome was followed by sequencing of the genomes of closely related primate species, such as the chimpanzee and the rhesus macaque. The motivation was simple: as the genome provided the blueprint of an organism, comparisons between the human genome and the genomes of non-human primates should reveal genomic features underlying the human phenotype.

One problem with this approach, however, is that a genome is not really a blueprint of a phenotype, but rather a well-scrambled message, in which functionally relevant sequences are lost in a sea of phenotypically neutral information. A seemingly straightforward way to identify functional sequences is to determine transcribed regions. This is not a simple task, however, as the transcriptome varies greatly across cell types and changes dramatically across an organism's lifespan. Thus, in the past decade, a large effort was put into annotating the human transcriptome, mainly by sequencing transcripts converted into cDNA libraries by conventional Sanger sequencing. As a result, it became clear that given enough sequencing coverage, almost any genomic sequence can be detected on the transcriptome level [1]. This is not entirely surprising, as human genes frequently contain long introns; moreover, RNA polymerase can generate spontaneous transcripts of no functional relevance. Still, this result indicated that dividing the genome into transcribed and non-transcribed parts to determine functionality was largely futile.

These cDNA sequencing projects also showed that the boundaries of most human genes, including transcription start and termination sites and the splicing patterns of internal exons, are rather fuzzy [2-6]. In addition, many of the identified transcripts and gene isoforms turned out to be rare. This does not, however, mean that they are functionally irrelevant, as such transcripts may have important roles in a limited number of cells in a tissue or at a specific stage of development. Further, many important regulators, such as transcription factors, are expressed at low levels. As a result, the current human transcriptome annotation represents a certain trade-off between confidence and comprehensiveness and contains transcripts identified with different degrees of confidence. The difficulty in compiling such an annotation is best illustrated by the differences that exist between RefSeq, Ensembl, the University of California Santa Cruz (UCSC) Genome Browser, the Vega Genome Browser and an integrated database of human genes and transcripts (H-Invitational Database): one finds an average overlap of 60 to 70% comparing any two of these annotation databases.

Another way to determine functionally relevant transcripts is to require that the expression of a given transcript is conserved across species. Alternatively, if one is interested in loci important for the human phenotype, one could identify regions with human-specific transcription profiles. However, the transcriptome annotation of non-human primates is basically non-existent and what is present is entirely based on mapping the human annotation to the respective primate genomes. As human transcriptome annotation itself is far from being comprehensive and the quality of the non-human primate genomes is far worse than the quality of the human genome, such mapping-based annotation is not problem-free. But, most importantly, even though this method might allow identification of transcripts present in humans and absent in the other primates, it does not allow identification of transcripts lost from the human lineage.

*Correspondence: khaitovich@eva.mpg.de
[1]Partner Institute for Computational Biology, Chinese Academy of Sciences, 320 Yue Yang Road, 200031 Shanghai, China. [2]Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany.

In this issue, Lucia Cavelier and colleagues [7] present a study that takes advantage of high-throughput sequencing technology to annotate genomic regions transcribed in the chimpanzee brain cortex and liver. High-throughput sequencing technology, introduced just a few years ago and increasing rapidly in its capacity, allows the sequencing of millions of short reads in a single run. In their study, Cavelier and colleagues [7] used the ABI/SOLiD sequencing platform to generate over 500 million reads of length 35 and 50 nucleotides from poly(A)$^+$ RNA expressed in brain and liver tissue from two chimpanzees. Mapping the obtained sequences to the chimpanzee genome enabled them to identify transcribed regions independently of the existing annotation. Consequently, they found that only about a third of the obtained reads mapped to known chimpanzee exons. This proportion is much lower than that found in human RNA sequencing studies, reflecting the poor quality of the chimpanzee genome annotation. Importantly, they were able to identify in the order of 350 genomic regions that are highly transcribed in the chimpanzee genome but completely absent in the current human genome assembly. Using the rhesus macaque genome as an outgroup, they found that approximately half of these regions were lost from the human lineage. In addition to these transcribed regions of as-yet unknown function, Cavelier and colleagues [7] identified several novel gene isoforms not annotated in humans and a putative novel gene from the ATP-cassette-transporter family that is conserved between chimpanzee and mouse but lost from the human lineage.

These findings [7] add weight to the 'less is more' hypothesis of human evolution, postulating that some of the human-specific features have evolved not through acquisition of novel genetic elements, but through functional loss of previously existing ones. This study from Cavelier and colleagues [7] clearly shows that human-specific loss of transcribed regions is not limited to annotated protein-coding genes, but is common among intergenic transcripts and non-coding RNA. This finding is in good agreement with previous studies of the human and chimpanzee brain transcriptomes carried out using tiling arrays [8], high-throughput sequencing of expressed tags [9] and high-throughput sequencing of the complete transcriptomes [10], which all indicate that a large proportion of human-specific transcription gain and loss originates in as-yet unannotated genomic regions. Thus, the current task is to reveal functional properties of these novel transcripts, if they exist.

Importantly, the study [7] draws attention to the poor state of genome annotation in non-human primates. In humans, the use of high-throughput sequencing technology in transcriptome studies has already revealed much greater variability of the gene transcript isoforms than previously appreciated [2,3]. In non-human primates, such as chimpanzees and rhesus macaques, both the known genome sequence information and, particularly, the genome annotations are in a far worse state than those of humans. The study of Cavelier and colleagues [7] clearly illustrates that careful characterization of human and non-human primate transcriptomes can uncover large numbers of genetic and transcriptional changes specific to humans. Some of these changes will be responsible for the evolution of human-specific features, such as adaptation to a cooked, highly nutritious diet and unique social and cognitive abilities. Finding genetic elements underlying these uniquely human features is important not only for our understanding of human evolution, but also for prevention of their dysfunctions that may result in metabolic and cognitive disorders.

The recent advances in high-throughput sequencing methodology provide us with powerful tools with which to characterize complete transcriptomes in multiple tissues and cell types across primate species, resulting in the comprehensive identification of the transcriptome features specific to humans. The work of Cavelier and colleagues [7] is the first brave step in this direction.

Published: 23 July 2010

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. 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.
3. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.** *Science* 2008, **321:**956-960.
4. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28:**511-515.
5. Tian B, Pan Z, Lee JY: **Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing.** *Genome Res* 2007, **17:**156-165.
6. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, *et al.*: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38:**626-635.
7. Wetterbom A, Ameur A, Feuk L, Gyllensten U, Cavelier L: **Identification of novel exons and transcribed regions by chimpanzee transcriptome sequencing.** *Genome Biol*, **11:**R78.
8. Khaitovich P, Kelso J, Franz H, Visagie J, Giger T, Joerchel S, Petzold E, Green RE, Lachmann M, Paabo S: **Functionality of intergenic transcription: an evolutionary comparison.** *PLoS Genet* 2006, **2:**e171.
9. Babbitt CC, Fedrigo O, Pfefferle AD, Boyle AP, Horvath JE, Furey TS, Wray GA: **Both noncoding and protein-coding RNAs contribute to gene expression evolution in the primate brain.** *Genome Biol Evol* 2010, **2010:**67-79.

10. Xu AG, He L, Li Z, Xu Y, Li M, Fu X, Yan Z, Yuan Y, Menzel C, Li N, Somel M, Hu H, Chen W, Pääbo S, Khaitovich P: **Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq.** *PLoS Comput Biol* 2010, **6:**e1000843.

doi:10.1186/gb-2010-11-7-125
Cite this article as: Khaitovich P: **Annotating conserved and novel features of primate transcriptomes using sequencing.** *Genome Biology* 2010, **11:**125.