Published in final edited form as:
Genes (Basel). ; 3(3): 344–360. doi:10.3390/genes3030344.

The Human Transcriptome: An Unfinished Story

Mihaela Pertea
McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

Abstract

Despite recent technological advances, the study of the human transcriptome is still in its early stages. Here we provide an overview of the complex human transcriptomic landscape, present the bioinformatics challenges posed by the vast quantities of transcriptomic data, and discuss some of the studies that have tried to determine how much of the human genome is transcribed. Recent evidence has suggested that more than 90% of the human genome is transcribed into RNA. However, this view has been strongly contested by groups of scientists who argued that many of the observed transcripts are simply the result of transcriptional noise. In this review, we conclude that the full extent of transcription remains an open question that will not be fully addressed until we decipher the complete range and biological diversity of the transcribed genomic sequences.

Keywords

transcriptome; pervasive transcription; RNA-seq; mRNA; ncRNA

1. Background and Introduction

The transcriptome of a cell is the collection of all the RNA molecules, or transcripts, present in that cell. To generate the transcriptome, the DNA of an organism is first transcribed by RNA polymerase to create complementary RNA strands, which in turn are spliced to remove introns, producing mature transcripts that contain only exons. For many years, it was assumed that these RNA transcripts were primarily used as templates for translation to proteins. The vast majority of the remaining human genome, which is not protein coding, was thought to be non-functional and therefore considered “junk” DNA [1]. Soon after the publication of the human genome sequence in 2001 [2,3], a new view emerged, holding that only a small percentage of the human transcriptome is clearly translated into proteins [4–6], and most of the remaining transcripts have unknown purposes. In recent years, the number and variety of known RNA genes has grown dramatically, and in addition to protein-coding messenger RNAs (mRNAs), the catalog of transcribed elements now includes a myriad of non-coding RNAs (ncRNAs) that play multiple structural and regulatory roles in the molecular biology of the cell [7].

Ever since the discovery of the genetic code, scientists have labored to decipher the complete human transcriptome. It was only with the emergence of automated DNA sequencing in the 1980s that real progress was made in this direction [8]. In the 1990s, scientists realized the value of using expressed sequence tag (EST) sequencing to rapidly identify expressed genes, or at least fragments of those genes, in many human tissues [9,10]. Although at the time EST sequencing was considered a very high-throughput technique,

mpertea@jhu.edu; Tel.: +1-443-287-0972 .

both costs and technical limitations prevented it from producing a complete transcript catalog. As a consequence, much of our knowledge of the protein-coding portion of the human transcriptome relied on different computational gene prediction methods [11,12].

Various other technologies were developed to complement the traditional EST approach. These include tag-based methods such as serial analysis of gene expression (SAGE) [13], cap analysis of gene expression (CAGE) [14], and massively parallel signature sequencing (MPSS) [15]. Unlike the EST approach, the tag methods uniquely identify each transcript to achieve gene-level expression quantification. However they are generally unable to distinguish specific isoforms. In addition, most of them are based on traditional Sanger sequencing technology, making them very expensive to apply on a large scale.

Hybridization-based microarrays provided the first relatively inexpensive way to detect and quantify transcripts on a large scale [16–18]. These include transcription tiling arrays, which allow the mapping of transcribed regions to a very high resolution, from 5 to 50 base pairs (bp), depending on probe density [19,20]. They have several advantages over previous methods, including their high throughput and their ability, with some designs, to quantify distinct spliced isoforms [21]. However, because of differences in hybridization strength, cross-hybridization, and other experimental variables, microarrays provide a noisy output signal. In addition, they can only measure genes for which the sequence and the precise exon-intron boundaries are known, making them unable to identify novel genes or novel splicing events [22,23].

Recently, RNA-seq methods technologies provide unprecedented opportunities for characterizing the set of RNA transcripts produced in a cell [24–28]. Called a “revolutionary tool for transcriptomics”, RNA-seq is the first sequencing-based method that allows the entire transcriptome to be surveyed in a very high-throughput and quantitative manner [29]. Unlike hybridization-based methods, it is not limited to the detection of known transcripts, and it can measure a much larger range of expression levels. Among its other advantages, RNA-seq data has relatively low background noise; it achieves base-pair resolution, allowing precise identification of exon and intron boundaries; and it can detect single nucleotide polymorphisms (SNPs) and other variants within transcripts. Although RNA-seq has already dramatically changed the landscape of genetic studies, it is clear that many years remain before we will have a complete catalogue of human genes and their expressed isoforms.

2. The Diversity of the Transcriptome

2.1. Various Classes of ncRNAs

Over the past decade, many studies have revealed an unexpected level of diversity in the human transcriptome, which in turn has required scientists to expand their definition of a gene. The traditional definition of a gene—a DNA sequence that is transcribed to produce a functional product—has been expanded to include not only to the ~22,000 protein-coding genes present in the human genome [11], but also a myriad of non-protein coding sequences. These set of transcribed non-protein coding DNA sequences show complex patterns of expression and regulation [30], and they are no longer restricted to the well known ribosomal and transfer RNAs (rRNAs and tRNAs, respectively). Furthermore, when we introduce these new and growing functional RNAs into our gene counts, the number of genes in the human genome increases from ~22,000 (which includes only protein-coding genes) to the 2001 estimates of about 30,000–40,000 genes [31].

The discoveries of endogenous small interfering RNA (siRNA) [32] and microRNA (miRNA) [33] genes represented dramatic breakthroughs in our understanding of the transcriptome. These two classes of small ncRNAs play a central role in RNA interference by binding to specific mRNA molecules to either increase or decrease their activity. Various other classes of ncRNAs have a now-broadly recognized functional role. These include regulatory RNAs such as PIWI-interacting RNAs (piRNAs), promoter-associated RNAs (PARs), transcription initiation RNAs (tiRNAs), X-inactivation RNAs (xiRNAs), and many others [34,35]. Among them, the long non-coding RNAs (lncRNAs), defined as ncRNAs longer than 200 bp, are probably the least well-understood transcripts. Although few of them have been experimentally studied, a view is emerging that these are key regulators of epigenetic gene regulation in mammalian cells [36].

Large intergenic RNAs (lincRNAs) are a subclass of lncRNAs that do not overlap protein-coding regions. Cabili *et al.* [37] catalogued more than 8,000 lincRNAs (58% of which were novel) using an integrative approach that unifies existing annotation sources with transcripts assembled from RNA-seq data collected from 24 tissues and cell types. Several global properties of lincRNAs were evidenced by this study:

- they are expressed in a highly tissue-specific manner compared to protein-coding genes,
- they are typically co-expressed with their neighboring genes, and
- they only show moderate conservation in other species.

The functional classification of lincRNAs is far from complete, even though Cabili *et al.* assigned putative functions to many predicted lincRNAs based on the functions of protein-coding genes with similar expression patterns.

2.2. Alternative Splicing

Even when considering only protein-coding RNAs, the scientific community still does not have a complete picture of the transcriptome. Not only is there uncertainty about the exact number of human protein-coding genes, but recent evidence has emerged to show that different humans have slightly different individual gene sets [38–40]. The number of mature mRNA transcripts is even less certain, and varies across tissues and different stages during cell differentiation [41,42]. Further complicating matters, we now know that more than 90% of multi-exon protein-coding genes undergo alternative splicing [43,44], which is considered to play a major role in increasing cellular and functional diversity in the transcriptomes of higher eukaryotes [45]. However, we do not yet know the function of the vast majority of alternatively spliced human transcripts, and it is now clear that alternative splicing does not simply act to generate variant protein sequences [46].

Alternative splicing also affects ncRNA genes, about 30% of which produce at least one alternatively spliced transcript [47]. Cabili *et al.* found that lincRNAs, although shorter and with fewer exons than mRNAs, are also alternatively spliced with an average of 2.3 isoforms per locus [37]. New transcripts are continuously being discovered [19,41,48,49], strengthening the observation that we are far from determining all transcript isoforms.

2.3. Estimating the Annotated Human Transcript Count

In an attempt to identify how many human transcripts are currently annotated, I combined all human gene annotations from Ensembl (release 64) [50], NCBI's RefSeq database [51], and the UCSC Genome Browser [52] with the lincRNAs catalogued by Cabili *et al.* [37]. After eliminating redundant transcripts (*i.e.*, transcripts with identical annotation as an already included transcript from one of the databases), I divided the remaining ones into three categories: mRNAs if they were annotated as protein-coding transcripts, long ncRNAs if they were annotated as non-coding and were at least 200 bp long, and small ncRNAs otherwise.

As also observed by others [53], I found a highly complex architecture in the human transcriptome, in which some base pairs could be part of many overlapping transcripts in any of the three categories, and emanating from both strands of the genome. Loci containing all three categories of transcripts were not frequent (see Figure 1a). Not surprisingly, the annotations include more mRNA than ncRNA transcripts, possibly due to a bias towards annotating protein-coding transcripts, although loci with at least one ncRNA are more numerous than loci containing one or more mRNAs (see Table 1). Overall, annotated transcripts today cover 4.62% or 3.85% of the human genome, depending on whether or not we include pseudogenes. Expression of pseudogenes is controversial, with some reports suggesting that they might be transcribed and could play a significant part in gene regulation [54,55]. They cover about 30% of the total base pairs included in all ncRNA transcripts. Figure 1b shows the base pair coverage of the human transcriptome (including pseudogenes) by the three categories of transcripts. I found that 62% of the base pairs in the transcriptome are part of mRNAs, supporting the fact that ncRNAs tend to be smaller in length than mRNAs.

2.4. RNA Editing

RNA editing is another cellular process that contributes to the complex landscape of mammalian transcriptomes. In the RNA editing process, single nucleotide changes occur after DNA has been transcribed into RNA. The resulting RNA transcripts may produce altered proteins, or they may disrupt translation more severely [56]. Two RNA editing mechanisms are known in humans, causing two types of substitutions: adenosine to inosine, and cytosine to uracil. The A-to-I editing, also called A-to-G, is a process mediated by a family of adenosine deaminases (ADARs) that act on RNA and replace certain adenosines (A) with inosines, which then act as guanosines (G) during translation [57,58]. Similarly, the C-to-U switches are mediated by APOBEC1 [59–61].

Until recently considered a rare event, RNA editing is now believed to affect both coding and non-coding sequences of thousands of genes, including ncRNAs [56,62,63]. A 2011 study by Li *et al.* [64] looked at RNA-seq and DNA sequence data from 27 individuals and reported that RNA-DNA differences (RDDs) are not limited to the two previous types of substitutions described above. In their study, Li *et al.*, observed all 12 possible RNA-DNA substitutions at more than 10,000 exonic sites, most of them present in multiple individuals and in different cell types. Their result suggests that previously unknown RNA editing mechanisms may be active in humans. However, this result has been strongly contested by several other groups, who argued that the vast majority of the observed RDDs were technical artifacts, mostly due to read mapping errors or systematic sequencing errors [65–68]. Nevertheless, RNA editing has an important role in molecular biology, and recent studies show that it may produce even more transcriptome diversity than alternative splicing [69].

3. Reconstructing the Transcriptome

As discussed above, high-throughput RNA sequencing surpasses all previous technologies in its ability to profile the extent and complexity of eukaryotic transcriptomes. The latest generation of sequencing machines can generate up to 600 gigabases (Gb) in a single run, equivalent to 200-fold coverage of the human genome. The 600 Gb is produced in the form of 6 billion short reads, each approximately 100 bp in length (using the Illumina HiSeq sequencer), and assembling these reads into chromosomes is a very complex, highly specialized task. Therefore one of the main challenges posed by RNA-seq is a computational one. Here I will briefly mention some of the most common bioinformatics systems for transcriptome assembly, and the challenges faced by these systems. For a more comprehensive review of next-generation transcriptome assembly methods, the interested reader can consult several recent reviews [70–72].

Although many programs have been developed for whole-genome assembly (e.g., [73–75]), these methods cannot be directly applied to transcriptome assembly due to specific characteristics of RNA-seq data sets. Genome assembly programs assume that the DNA sequence's depth of coverage is relatively uniform across the genome. This is not true for transcripts, which have highly variable sequence coverage depending on their expression levels. Sequence depth is used to indicate repeats by genome assemblers, which are designed to take this into account. Another confounding fact for genome assemblers is that alternative transcripts from the same locus typically share exons that are difficult to assemble unambiguously. Specific features of RNA-seq data (e.g., strand-specific sequencing or partially covered gene transcripts from low-abundance genes [48]) can also confound a whole-genome assembly algorithm. Therefore new methods have had to be developed to address the particular characteristics of transcriptome assembly.

There are two main approaches for assembly of a transcriptome: a genome-guided approach when a reference genome is available; or *de novo* assembly, which does not need a genome reference and can theoretically reconstruct transcripts that are transcribed even from parts missing from that genome's assembly. *De novo* transcriptome assembly is far more challenging in higher eukaryotes due to the large number of genes, the great variation in their expression levels, and especially because of the large number of alternatively spliced transcript variants. For this reason, *de novo* methods are primarily used for organisms that lack a sequenced reference genome.

Read mapping is one of the main technical challenges of genome-guided approaches. Alignment of short reads to the reference genome is a challenge in itself, but with RNA-seq data these reads may be sequenced from exons and exon-exon junction regions. Methods such as Bowtie [76] and BWA [77] can be used for the alignment of reads to either a reference genome or directly to the transcriptome, but this strategy will miss novel exons and novel splicing events. Spliced aligners were developed to overcome these limitations. Some of them (e.g., TopHat [78], SpliceMap [79], MapSplice [80]) use an 'exon-first' approach where reads are first mapped to the genome, and then the unmapped reads are split into shorter segments and aligned independently. Other spliced aligners, such as GSNAP [81] or BLAT [82], use a 'seed-and-extend' strategy in which the reads are first divided into small segments (seeds) that are individually aligned to the genome, and then candidate regions are locally aligned to obtain the final spliced alignment of the read. There are different advantages to these strategies, but in general 'exon-first' aligners are usually faster, while 'seed-and-extend' ones may be slightly more sensitive by reducing the bias towards unspliced alignments in the exon-first approach.

After mapping all reads to the reference genome, transcriptome assemblers cluster the overlapping reads at each locus and build a connectivity graph representing all possible isoforms. Different transcriptome assembly programs, such as Cufflinks [41], Scripture [83], IsoInfer [84], and IsoLasso [85], use different criteria to parse the connectivity graph. Cufflinks uses a parsimony principle to generate the minimal number of transcripts that will explain all reads in the graph. If there are multiple ways to assemble a minimal number of transcripts, Cufflinks uses the read coverage across each path to decide which combination is most likely to originate from the same RNA transcript. Scripture reconstructs all possible isoforms by enumerating all possible paths in the connectivity graph that have statistically significant read coverage. While Cufflinks and Scripture estimate the abundance of transcripts after they are assembled, IsoInfer and IsoLasso assemble transcripts at the same time that they estimate their expression levels. They take two different approaches: IsoInfer uses a heuristic approach to reduce the huge search space of all valid isoforms, while IsoLasso uses a multivariate regression method that also minimizes the number of predicted transcripts.

De novo transcriptome assembly methods, generally based on de Brujin graphs, are less efficient and less sensitive than genome-guided methods for the human genome. Despite that, running a *de novo* assembler in addition to a genome-guided method may produce a more comprehensive transcriptome. Because *de novo* assemblers do not need a reference genome, they can identify genes that are missing from the reference genome, such as trans-spliced transcripts and similar transcripts originating from chromosomal rearrangements. Trinity [86], Oases [87], SOAPdenovo [88], and Trans-ABySS [89] are some of the programs used for *de novo* transcriptome assembly. A recent comparative study [90] evaluated the performance of different *de novo* transcriptome assembly programs and found that Trinity performed well across various conditions, but took the longest running time; Oases consumed the most memory; SOAPdenovo required the shortest runtime but performed poorly at reconstructing full-length transcripts; and Trans-ABySS showed a good balance between resource usage and quality of assemblies. Although it would undoubtedly prove useful, there is no automated software pipeline to carry out a combined assembly strategy to bring together the high sensitivity of genome-guided assemblers with the ability of *de novo* methods to detect novel and trans-spliced transcripts.

4. The Size of the Transcriptome

Less than 2% of the human genome codes for proteins [91]. As described above, if we add to this fraction the DNA sequences that correspond to annotated ncRNAs, we are still left with less than 5% of the human genome covered by known transcripts. Other reports have found that only ~5–10% of the genome is stably transcribed in cell lines [19,20,92]. My own independent analysis (Figure 2) shows that it is rare to see more than 5% of the total base pairs in the genome covered by assembled transcripts in normal human tissue. While these studies don't capture the expression of the transcriptome at all stages in the cell development, they suggest that only a small portion of the human genome is transcribed. And yet a mounting number of studies suggest that the vast majority of the genome is transcribed at some time or other. Beginning in the early 2000s, full length cDNAs from various mouse tissues at different developmental stages, and genome-wide tiling arrays in different human tissues and cell lines revealed that much more of the mammalian genomes is transcribed than what is annotated in public databases [5,19,20,49,93–95]. These studies culminated with the publication in 2007 of the results from the pilot phase of the ENCODE Project [96], which estimated that as much as 93% of the human genome is transcribed in at least one cell type. Does this broad pattern of transcription mean simply that the cell creates a great deal of transcriptional noise by RNA polymerase binding accidentally (or randomly) to many sites in the genome? Or does this result challenge the long-standing view that most of the human genome is not biologically active? Scientists have conflicting opinions on the answer to this question.

A recent study published by van Bakel *et al.* [97] claims that most 'dark matter' transcripts — defined as ncRNAs of unknown function - are associated with known genes. In this paper, van Bakel *et al.* argue that there is a high false-positive rate associated with the tiling array technology that was the basis of most analyses that suggested the pervasiveness of transcription. When compared to RNA-seq data, tiling arrays produce a larger proportion of low-abundance transcripts originating from intergenic and intronic regions, although tiling arrays and RNA-seq data generally agree on the location of the greatest transcript "mass."

The low coverage of intronic transcripts suggests that they might in fact represent random sampling from partially processed or unprocessed RNAs. Supporting this idea is also the observation that the transcription mass in intergenic regions increases at much lower rates than in intronic regions as the number of reads is increased. Van Bakel *et al.*, also identified several thousand small transcripts that map outside known genes, however most of them could be explained as accidental by-products of enhancer activity. Overall, the authors conclude that most of the genome is not appreciably transcribed, and the majority of intergenic and intronic transcripts observed in previous studies may be attributed to biological and/or technical background noise.

Clark *et al*. [99] acknowledge that indeed most dark matter transcripts are associated with known genes, but they strongly disagree with van Bakel *et al*.'s conclusion that the genome is not as pervasively transcribed as previously reported. In their study, Clark *et al*., argue that we cannot dismiss the observations from multiple independent techniques, including RT-PCR, RACE, and Northern blot analyses, which together validated more than 90% of the identified transcripts [100,101]. They also argue that van Bakel *et al*.'s RNA-seq data suffers from insufficient sequencing depth and poor assembly, and is biased towards polyadenylated RNA, which selectively omits significant amounts of RNA as has been shown earlier [102]. Overall, similarly to other studies [103,104], Clark *et al*. find that the detection accuracy of tiling arrays is not significantly lower than that of RNA-seq, and they conclude that a significant fraction of dark matter RNA comes from very long, intergenic transcribed regions.

In a subsequent paper [105], van Bakel *et al*. agree with the fact that most of the genome appears to be transcribed. But given the various sources of extraneous reads, both biological and laboratory-derived, they expect that given sufficient sequencing depth the whole genome may be covered with transcripts. A recent study that sequenced total RNA from human brain and liver supports van Bakel *et al*.'s suggestion that unannotated transcripts within introns represent unspliced introns rather than unique independent transcriptional units [106]. And yet another study found that sequenced reads observed in conventional RNA sequencing data sets, previously dismissed as noise, are in fact indicative of unassembled rare transcripts [107]. Therefore the debate about the pervasiveness of transcription continues, but as van Bakel *et al*., and others [30,108] point out, it is time to stop arguing over the content of the transcriptome, and focus on finding evidence for dark matter functions.

5. Discussion and Conclusions

The unprecedented depth of sequence coverage achieved by RNA-seq has revealed how much of the human transcriptome is still uncharacterized. Many novel transcripts are still being discovered, stimulating the debate as to the extent to which the genome is transcribed. Non-coding RNAs represent the majority of the human transcripts, and there is no doubt that many of them, initially considered to be transcriptional artifacts, are in fact functional. They play important roles in transcriptional and post-transcriptional gene regulation via both *cis*- and *trans*-acting mechanisms, chromatin modification, control of transcription factor binding, regulation of alternative splicing. These functions have important consequences for development and for diseases, including cancer [30,36,109,110].

Despite current intense research efforts, many of the novel transcripts identified thus far have an unknown function. Most of them have been found only in specific cell types, tissues, or developmental stages [37,100,111]. They lack functional ORFs, have lower expression levels, and are only modestly conserved, although conservation is only a week indicator of functionality [96,112,113]. Occasionally, entirely novel protein-coding genes with strong mRNA expression have been identified [114], but most unannotated transcripts that are protein-coding are alternatively spliced isoforms of known mRNAs [41]. However, as of today the vast majority of alternatively spliced transcripts lack described functions, and the role of alternative splicing itself in gene evolution remains largely unexplored [46].

Is low RNA polymerase fidelity the principal cause of the widespread transcription observed in the human genome? We do not have a definite answer to this question [115]. A focus on deciphering the biological functions of transcribed genomic sequences might provide us with a clearer picture. Over the last decade, the estimated proportion of the human genome that might be functional has been constantly adjusted upwards, and today it lies between 10% and 15% [116]. This estimate is still much lower than the ~93% estimate for the transcribed fraction of the genome [96]. In a 2009 review, Ponting *et al.*, argue that a large, but as yet unknown, number of noncoding RNAs cannot be explained solely as the product of transcriptional noise [30]. If ncRNAs were simply transcriptional noise, than their expression levels would not show the wide diversity that is often observed among different tissues. In addition, their nucleotide substitution rates would be very similar to neutrally evolving sequences. Instead, several evolutionary studies suggest that many ncRNAs exhibit signatures of functionality that are more usually associated with protein-coding genes [47,117], or that their low sequence conservation is due to the fact that they are frequently acted upon by positive selection [118,119]. Nevertheless, some percentage of the transcripts observed are very likely the result either of transcriptional noise [120] or of genomic DNA contamination [121]. Even if not functional themselves, these unannotated transcripts might reflect transcriptional processes that facilitate the expression of other genes. Until we can functionally validate these transcripts or gain a better understanding of the range of transcriptional mechanisms involved, the question of how much of the human genome is transcribed will remain an open question.

Acknowledgments

Thanks to Steven Salzberg for helpful comments and suggestions. This work was supported by NIH grants R01 HG006677 and R01 HG006102.

References

1. Ohno S. So much “junk” DNA in our genome. Brookhaven Symp. Biol. 1972; 23:366–370. [PubMed: 5065367]
2. The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. [PubMed: 11237011]
3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. The sequence of the human genome. Science. 2001; 291:1304–1351. [PubMed: 11181995]
4. Chen J, Sun M, Lee S, Zhou G, Rowley JD, Wang SM. Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. Proc. Natl. Acad. Sci. USA. 2002; 99:12257–12262. [PubMed: 12213963]
5. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. Large-scale transcriptional activity in chromosomes 21 and 22. Science. 2002; 296:916–919. [PubMed: 11988577]
6. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE. Using the transcriptome to annotate the genome. Nat. Biotechnol. 2002; 20:508–512. [PubMed: 11981567]
7. Mattick JS. The central role of RNA in human development and cognition. FEBS Lett. 2011; 585:1600–1616. [PubMed: 21557942]
8. Griffin HG, Griffin AM. DNA sequencing. Recent innovations and future trends. Appl. Biochem. Biotechnol. 1993; 38:147–159. [PubMed: 8346903]

9. Adams MD, Kerlavage AR, Fields C, Venter JC. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. Nat. Genet. 1993; 4:256–267. [PubMed: 8358434]
10. Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, Kirkness EF, Weinstock KG, Gocayne JD, White O, et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature. 1995; 377:3–174. [PubMed: 7566098]
11. Pertea M, Salzberg SL. Between a chicken and a grape: Estimating the number of human genes. Genome Biol. 2010; 11:206. [PubMed: 20441615]
12. Strausberg RL, Riggins GJ. Navigating the human transcriptome. Proc. Natl. Acad. Sci. USA. 2001; 98:11837–11838. [PubMed: 11592992]
13. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995; 270:484–487. [PubMed: 7570003]
14. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. USA. 2003; 100:15776–15781. [PubMed: 14663149]
15. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol. 2000; 18:630–634. [PubMed: 10835600]
16. Clark TA, Sugnet CW, Ares M Jr. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. Science. 2002; 296:907–910. [PubMed: 11988574]
17. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995; 270:467–470. [PubMed: 7569999]
18. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc. Natl. Acad. Sci. USA. 1997; 94:13057–13062. [PubMed: 9371799]
19. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. Global identification of human transcribed sequences with genome tiling arrays. Science. 2004; 306:2242–2246. [PubMed: 15539566]
20. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science. 2005; 308:1149–1154. [PubMed: 15790807]
21. Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, Cooper TA, Johnson JM. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat. Genet. 2008; 40:1416–1425. [PubMed: 18978788]
22. Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics. 2006; 7:276. [PubMed: 16749918]
23. Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, Babak T, Siu H, Hughes TR, Morris QD, et al. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Mol. Cell. 2004; 16:929–941. [PubMed: 15610736]
24. Lister R, O’Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008; 133:523–536. [PubMed: 18423832]
25. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. 2008; 5:621–628. [PubMed: 18516045]
26. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008; 320:1344–1349. [PubMed: 18451266]
27. Salzberg SL. Recent advances in RNA sequence analysis. F1000 Biol. Rep. 2010; 2:64. [PubMed: 21173855]
28. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods. 2008; 5:613–619. [PubMed: 18516046]

29. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nat. Rev. Genet. 2009; 10:57–63. [PubMed: 19015660]

30. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell. 2009; 136:629–641. [PubMed: 19239885]

31. Dinger ME. lncRNAs: Finding the forest among the trees? Mol. Ther. 2011; 19:2109–2111. [PubMed: 22134744]

32. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998; 391:806–811. [PubMed: 9486653]

33. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993; 75:843–854. [PubMed: 8252621]

34. Jacquier A. The complex eukaryotic transcriptome: Unexpected pervasive transcription and novel small RNAs. Nat. Rev. Genet. 2009; 10:833–844. [PubMed: 19920851]

35. Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS. Non-coding RNAs: Regulators of disease. J. Pathol. 2010; 220:126–139. [PubMed: 19882673]

36. Derrien T, Guigo R, Johnson R. The long non-coding RNAs: A New (P)layer in the “Dark Matter”. Front Genet. 2011; 2:107. [PubMed: 22303401]

37. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25:1915–1927. [PubMed: 21890647]

38. Iafrate A, Feuk L, Rivera M, Listewnik M, Donahoe P, Qi Y, Scherer S, Lee C. Detection of large-scale variation in the human genome. Nat Genet. 2004; 36:949–951. [PubMed: 15286789]

39. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. Large-scale copy number polymorphism in the human genome. Science. 2004; 305:525–528. [PubMed: 15273396]

40. Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J, et al. Building the sequence map of the human pan-genome. Nat. Biotechnol. 2009; 28:57–63. [PubMed: 19997067]

41. 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. [PubMed: 20436464]

42. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. 2008; 40:1413–1415. [PubMed: 18978789]

43. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, et al. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res. 2004; 14:331–342. [PubMed: 14993201]

44. 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. [PubMed: 18978772]

45. Blencowe BJ. Alternative splicing: New insights from global analyses. Cell. 2006; 126:37–47. [PubMed: 16839875]

46. Mudge JM, Frankish A, Fernandez-Banet J, Alioto T, Derrien T, Howald C, Reymond A, Guigo R, Hubbard T, Harrow J. The origins, evolution, and functional potential of alternative splicing in vertebrates. Mol. Biol. Evol. 2011; 28:2949–2959. [PubMed: 21551269]

47. Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, Fukuda S, Ru K, Frith MC, Gongora MM, et al. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res. 2006; 16:11–19. [PubMed: 16344565]

48. Seok J, Xu W, Jiang H, Davis RW, Xiao W. Knowledge-based reconstruction of mRNA transcripts with short sequencing reads for transcriptome research. PLoS One. 2012; 7:e31440. [PubMed: 22312447]

49. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, et al. The transcriptional landscape of the mammalian genome. Science. 2005; 309:1559–1563. [PubMed: 16141072]

50. [accessed on 5 September 2011] Ensembl Genome Browser. Available online: http://useast.ensembl.org/Homo_sapiens/Info/Index

51. [accessed on 5 September 2011] NCBI's RefSeq Database. Available online: http://www.ncbi.nlm.nih.gov/RefSeq/

52. [accessed on 5 September 2011] UCSC Genome Table Browser. Available online: http://genome.ucsc.edu/cgi-bin/hgTables

53. 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. [PubMed: 15998911]

54. Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M, et al. Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution. Genome Res. 2007; 17:839–851. [PubMed: 17568002]

55. Sasidharan R, Gerstein M. Genomics: Protein fossils live on as RNA. Nature. 2008; 453:729–731. [PubMed: 18528383]

56. Sie CP, Kuchka M. RNA editing adds flavor to complexity. Biochemistry (Mosc). 2011; 76:869–881. [PubMed: 22022960]

57. Bass BL, Weintraub H. An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell. 1988; 55:1089–1098. [PubMed: 3203381]

58. Wagner RW, Smith JE, Cooperman BS, Nishikura K. A double-stranded RNA unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and Xenopus eggs. Proc. Natl. Acad. Sci. USA. 1989; 86:2647–2651. [PubMed: 2704740]

59. Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell. 1987; 50:831–840. [PubMed: 3621347]

60. Chen SH, Habib G, Yang CY, Gu ZW, Lee BR, Weng SA, Silberman SR, Cai SJ, Deslypere JP, Rosseneu M, et al. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. Science. 1987; 238:363–366. [PubMed: 3659919]

61. Teng B, Burant CF, Davidson NO. Molecular cloning of an apolipoprotein B messenger RNA editing protein. Science. 1993; 260:1816–1819. [PubMed: 8511591]

62. Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. 2004; 2:e391. [PubMed: 15534692]

63. Levanon EY, Eisenberg E, Yelin R, Nemzer S, Hallegger M, Shemesh R, Fligelman ZY, Shoshan A, Pollock SR, Sztybel D, et al. Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat. Biotechnol. 2004; 22:1001–1005. [PubMed: 15258596]

64. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. Widespread RNA and DNA sequence differences in the human transcriptome. Science. 2011; 333:53–58. [PubMed: 21596952]

65. Kleinman CL, Majewski J. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science. 2012; 335:1302. [PubMed: 22422962]

66. Lin W, Piskol R, Tan MH, Li JB. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science. 2012; 335:1302–e. [PubMed: 22422964]

67. Pickrell JK, Gilad Y, Pritchard JK. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science. 2012; 335 DOI: 10.1126/science.1210484.

68. Schrider DR, Gout JF, Hahn MW. Very few RNA and DNA sequence differences in the human transcriptome. PLoS One. 2011; 6:e25842. [PubMed: 22022455]

69. Barak M, Levanon EY, Eisenberg E, Paz N, Rechavi G, Church GM, Mehr R. Evidence for large diversity in the human transcriptome created by Alu RNA editing. Nucleic Acids Res. 2009; 37:6905–6915. [PubMed: 19740767]

70. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat. Rev. Genet. 2011; 12:671–682. [PubMed: 21897427]

71. Costa V, Angelini C, de Feis I, Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-Seq. J. Biomed. Biotechnol. 2010; 2010:853916. [PubMed: 20625424]

72. Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods. 2011; 8:469–477. [PubMed: 21623353]

73. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. [PubMed: 18349386]

74. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: A parallel assembler for short read sequence data. Genome Res. 2009; 19:1117–1123. [PubMed: 19251739]

75. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome Res. 2008; 18:810–820. [PubMed: 18340039]

76. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]

77. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

78. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]

79. Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res. 2010; 38:4570–4578. [PubMed: 20371516]

80. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010; 38:e178. [PubMed: 20802226]

81. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010; 26:873–881. [PubMed: 20147302]

82. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002; 12:656–664. [PubMed: 11932250]

83. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat. Biotechnol. 2010; 28:503–510. [PubMed: 20436462]

84. Feng J, Li W, Jiang T. Inference of isoforms from short sequence reads. J. Comput. Biol. 2011; 18:305–321. [PubMed: 21385036]

85. Li W, Feng J, Jiang T. IsoLasso: A LASSO regression approach to RNA-Seq based transcriptome assembly. J. Comput. Biol. 2011; 18:1693–1707. [PubMed: 21951053]

86. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 2011; 29:644–652. [PubMed: 21572440]

87. [accessed on 12 April 2012] Oases: De novo transcriptome assembler for very short reads. Available online: http://www.ebi.ac.uk/~zerbino/oases/

88. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: An improved ultrafast tool for short read alignment. Bioinformatics. 2009; 25:1966–1967. [PubMed: 19497933]

89. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, et al. De novo transcriptome assembly with ABySS. Bioinformatics. 2009; 25:2872–2877. [PubMed: 19528083]

90. Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: A comparative study. BMC Bioinformatics. 2011; 12:S2.

91. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004; 431:931–945. [PubMed: 15496913]

92. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science. 2007; 316:1484–1488. [PubMed: 17510325]

93. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 2002; 420:563–573. [PubMed: 12466851]

94. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, et al. Antisense transcription in the mammalian transcriptome. Science. 2005; 309:1564–1566. [PubMed: 16141073]

95. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, et al. The transcriptional activity of human Chromosome 22. Genes Dev. 2003; 17:529–540. [PubMed: 12600945]

96. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]

97. Van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Most "dark matter" transcripts are associated with known genes. PLoS Biol. 2010; 8:e1000371. [PubMed: 20502517]

98. Asmann YW, Necela BM, Kalari KR, Hossain A, Baker TR, Carr JM, Davis C, Getz JE, Hostetter G, Li X, et al. Detection of redundant fusion transcripts as biomarkers or disease-specific therapeutic targets in breast cancer. Cancer Res. 2012; 72:1921–1928. [PubMed: 22496456]

99. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, et al. The reality of pervasive transcription. PLoS Biol. 2011; 9:e1000625. [PubMed: 21765801]

100. Amaral PP, Mattick JS. Noncoding RNA in development. Mamm. Genome. 2008; 19:454–492. [PubMed: 18839252]

101. Berretta J, Morillon A. Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Rep. 2009; 10:973–982. [PubMed: 19680288]

102. Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, Reaman G, Milos P, Arceci RJ, Thompson JF, et al. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. BMC Biol. 2010; 8:149. [PubMed: 21176148]

103. Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics. 2010; 11:383. [PubMed: 20565764]

104. Malone JH, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. 2011; 9:34. [PubMed: 21627854]

105. Van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Response to "The reality of pervasive transcription". PLoS Biol. 2011; 9:e1001102.

106. Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. Nat. Struct. Mol. Biol. 2011; 18:1435–1440. [PubMed: 22056773]

107. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat. Biotechnol. 2011; 30:99–104. [PubMed: 22081020]

108. Jarvis K, Robertson M. The noncoding universe. BMC Biol. 2011; 9:52. [PubMed: 21798102]

109. Louro R, Smirnova AS, Verjovski-Almeida S. Long intronic noncoding RNA transcription: Expression noise or expression choice? Genomics. 2009; 93:291–298. [PubMed: 19071207]

110. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: Insights into functions. Nat. Rev. Genet. 2009; 10:155–159. [PubMed: 19188922]

111. Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Solda G, Simons C, et al. S. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res. 2008; 18:1433–1445. [PubMed: 18562676]

112. Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. Deletion of ultraconserved elements yields viable mice. PLoS Biol. 2007; 5:e234. [PubMed: 17803355]

113. Monroe D. Genetics. Genomic clues to DNA treasure sometimes lead nowhere. Science. 2009; 325:142–143. [PubMed: 19589978]

114. Knowles DG, McLysaght A. Recent de novo origin of human protein-coding genes. Genome Res. 2009; 19:1752–1759. [PubMed: 19726446]

115. Kaplan CD. The architecture of RNA polymerase fidelity. BMC Biol. 2010; 8:85. [PubMed: 20598112]

116. Ponting CP, Hardison R. What fraction of the human genome is functional? Genome Res. 2011; 21:1769–1776. [PubMed: 21875934]

117. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell. 2004; 116:499–509. [PubMed: 14980218]

118. Wang J, Zhang J, Zheng H, Li J, Liu D, Li H, Samudrala R, Yu J, Wong GK. Mouse transcriptome: Neutral evolution of ‘non-coding’ complementary DNAs. Nature. 2004:431. doi: 10.1038/nature03016.

119. Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. Trends Genet. 2006; 22:1–5. [PubMed: 16290135]

120. Ebisuya M, Yamamoto T, Nakajima M, Nishida E. Ripples from neighbouring transcription. Nat. Cell Biol. 2008; 10:1106–1113. [PubMed: 19160492]

121. Johnson JM, Edwards S, Shoemaker D, Schadt EE. Dark matter in the genome: Evidence of widespread transcription detected by microarray tiling experiments. Trends Genet. 2005; 21:93–102. [PubMed: 15661355]

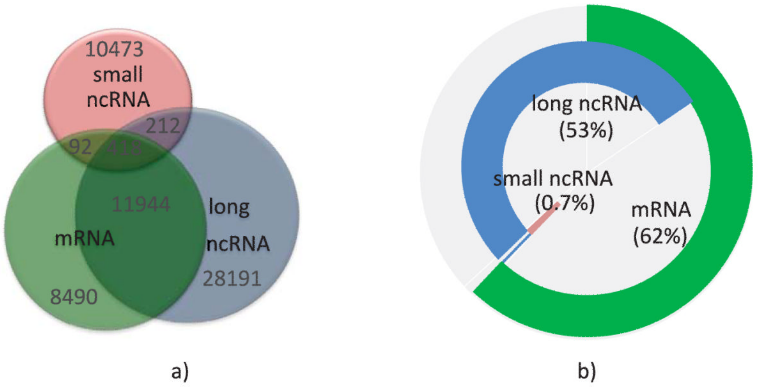

Figure 1.

Composition of the human transcriptome. (**a**) Venn diagram of the number of loci containing mRNA transcripts (green), long ncRNAs (blue), and small ncRNAs (red); (**b**) Base pair coverage of the transcriptome by the three categories of transcripts.

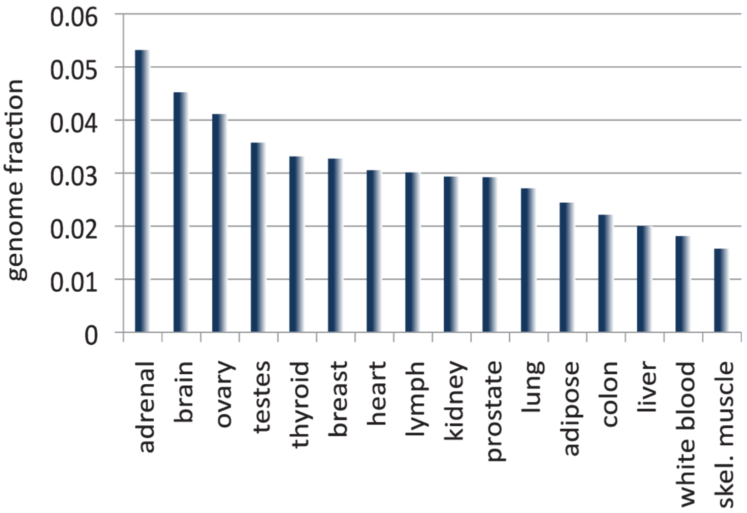

Figure 2.

The size of the transcriptome, computed as the fraction of the total number of base pairs in the human genome covered by the assembled transcripts, for 16 normal human tissues included in the Illumina Body Map [98]. Each RNA-seq data set was mapped to the genome with TopHat [78] and assembled with Cufflinks [41]. Note that except for adrenal tissue, in which transcripts cover 5.3% of the human genome, all other reconstructed transcriptomes are smaller in size than the currently annotated transcriptome.

Table 1

Number of known annotated transcripts and human gene loci collected from Ensembl, NCBI’s RefSeq, UCSC Genome Browser, and Cabili *et al.*’s lincRNA catalog. A single locus typically contains multiple transcripts, particularly for mRNAs.

Annotation	mRNA	Long ncRNA	Small ncRNA
Transcripts	111,451	89,981	11,366
Loci	20,944	40,765	11,195