St. Laurent et al. BMC Medicine 2014, 12:97
http://www.biomedcentral.com/1741-7015/12/97

BMC Medicine

MINIREVIEW Open Access

Dark matter RNA illuminates the puzzle of genome-wide association studies

Georges St. Laurent[1], Yuri Vyatkin[1,2] and Philipp Kapranov[1*]

Abstract

In the past decade, numerous studies have made connections between sequence variants in human genomes and predisposition to complex diseases. However, most of these variants lie outside of the charted regions of the human genome whose function we understand; that is, the sequences that encode proteins. Consequently, the general concept of a mechanism that translates these variants into predisposition to diseases has been lacking, potentially calling into question the validity of these studies. Here we make a connection between the growing class of apparently functional RNAs that do not encode proteins and whose function we do not yet understand (the so-called 'dark matter' RNAs) and the disease-associated variants. We review advances made in a different genomic mapping effort – unbiased profiling of all RNA transcribed from the human genome – and provide arguments that the disease-associated variants exert their effects via perturbation of regulatory properties of non-coding RNAs existing in mammalian cells.

Keywords: Genome-wide association study, Non-coding RNA, vlincRNA, Intronic RNA, lncRNA, RNA scaffold, LincRNA, Long Non-coding RNA, Long intergenic non-coding RNA, Very long intergenic non-coding RNA

Introduction

Connecting variations in DNA sequence with a biological or medical phenotype has long served to map functional elements of a genome. The recent genomics revolution has facilitated the identification of such variants on a massive scale, ushering in the era of genome-wide association studies (GWAS). Since the first pioneering report in 2005 [1], hundreds of such analyses have identified thousands of changes in DNA sequence (primarily single nucleotide polymorphisms (SNPs)) associated with a large number of complex diseases (cancers, heart disease, brain disorders, obesity, and many others; [2,3]. However, most of these variants have accumulated in unannotated, non-coding regions of the genome, whose functions continue to pose an enigma (Figure 1). Therefore, much of the wealth of GWAS information remains unrealized, with the mechanisms of action of the underlying genomic regions unknown, despite their widespread associations with disease.

* Correspondence: philipp@stlaurentinstitute.org
[1]St. Laurent Institute, 317 New Boston St, Suite 201, Woburn, MA 01801, USA
Full list of author information is available at the end of the article

Pervasive transcription: the answer to function behind the non-coding GWAS variants?

Only 2 to 3% of human DNA (genome) encodes proteins, the building blocks of life whose function we understand fairly well. The remaining 97 to 98% represent non-coding sequences, which were long considered 'junk DNA' because they did not fit the protein-centric view that dominated biology for decades. The goal of connecting DNA sequence variants to this protein-coding sliver of the human genome has shaped their interpretation, yet over 90% of GWAS hits lie in the non-coding parts of the genome (Figure 1; see Additional file 2: Supplementary Tables 1–3). (Table 1). Three possible explanations exist for the large preponderance of non-coding GWAS hits. They could arise from methodological errors such as imprecise measurements of phenotypes [4], differences in population structures [5] or DNA quality issues [6,7] between cases and controls. Second, they could affect distal regulatory regions of known (mostly protein-coding) genes. Third, they could represent novel genes or transcripts. However, as the number of GWAS increase, support for the first two arguments continues to weaken. The sheer number of non-coding GWAS hits, their continued accumulation as statistical power has improved, and their consistent

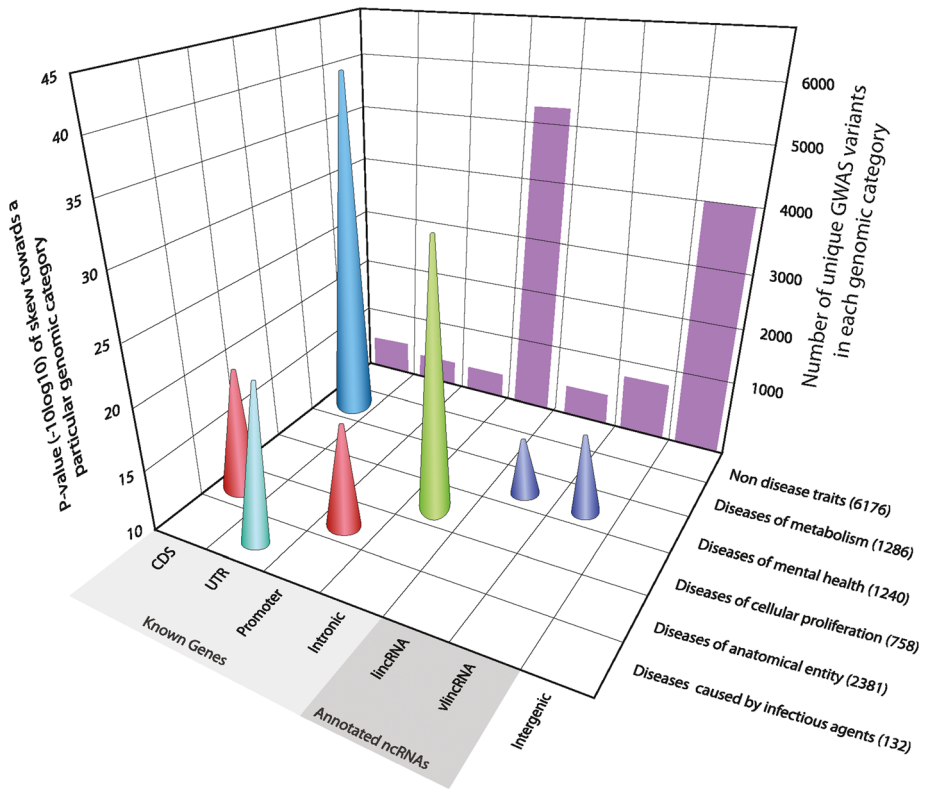

Figure 1 Discovery of genome-wide association studies (GWAS) variants in different genomic elements and disease types. GWAS variants were assigned to a disease type (*y* axis) or non-disease traits. For each disease type, the *P*-value (*x* axis) for a skew towards a particular genomic category (*x* xis) was calculated (see Additional file 1: Supplementary Text). Numbers of unique GWAS variants for each of the genomic categories are shown as the purple bars; the corresponding numbers for each of the disease types are shown in parenthesis. Only disease types with >100 GWAS variants are shown. CDS, coding DNA sequence (coding regions of known genes); UTR, untranslated region (non-coding regions of known genes; promoter and intronic regions are those of known genes). See Additional file 1: Supplementary Text for details of the analysis.

discovery across different diseases and different studies (Figure 1), argues against a widespread pattern of errors. Although errors must exist, they are unlikely to represent such a large proportion of GWAS events. Similarly, distal enhancers and regulatory regions will eventually explain some fraction of non-coding GWAS hits. However, these variants must interrupt fairly small sites of transcription factor binding and chromatin signaling within the regulatory regions, which represent a small minority of the genome. For example, Khurana *et al*. [8] reported that conserved transcription factor binding motifs and DNAse I hypersensitive regions make up only 0.4% of the genome. As expected, only 88 out of approximately 12,000 variants in the National Human Genome Research Institute GWAS catalog currently map to these regions. Therefore, the third explanation, that non-coding SNPs affect novel genes or transcripts, has begun to take center stage. In effect, the broad and continued accumulation of GWAS data, with the same pattern of distribution in non-coding regions, highlights the importance of pervasive transcription and dark matter RNA.

In 2002, the first report of pervasive transcription [9] subsequently triggered a series of genome-mapping endeavors that have discovered large numbers of dark matter RNAs transcribed from much of the non-coding space of the human genome [10-12]. Although an object of considerable debate over the last decade [13], an increasing number of independent observations in different species [14-19] have confirmed these results by continually increasing the annotation of the dark matter transcriptome. Presently, little doubt remains that the human genome produces large amounts of RNA whose function we still do not understand [10-12]. In fact, non-coding (nc)RNA represents at least 75% of the human genome [20], and its relative mass outweighs that of protein-coding mRNA [21].

These large and well-validated datasets now provide a strong basis for an in-depth look at ncRNA as a possible answer to the mystery of the many GWAS hits that do not fall neatly into protein-coding regions of the human genome. The non-coding disease-associated polymorphisms from the GWAS studies may have uncovered a

Table 1 Glossary of technical terms

Term	Meaning
Chromatin signaling	A system of regulation of gene activity in a cell that works by affecting the immediate surroundings of DNA, for example, by modifying various proteins that coat DNA inside the nucleus. Depending on the exact nature of the modification, DNA becomes either more or less accessible to cellular machinery that activates genes
Enhancer	A sequence of DNA that can regulate a target gene or genes over long distances
DNAse I hypersensitivity region	A region of DNA identified in an assay where chromatin is digested with DNAse I, an enzyme that degrades DNA. More accessible regions of chromatin, typically containing regulatory elements such as promoters and enhancers, are more susceptible to DNAse digestion and thus are enriched in DNAse I hypersensitivity regions
Gene Ontology (GO) term	GO is a an international initiative aimed at assigning controlled vocabulary, consisting of *terms* such as 'regulation of apoptosis' that define the functional property of each gene. This vocabulary is often very useful in understanding the biological meaning of a genomics experiment. For example, a list of genes activated during a disease would have a list of specific terms associated with each gene. Enrichment of specific terms in the list would suggest general cellular functions in which these genes participatem and give clues to the molecular functions underlying the disease
H1 embryonic stem cells	A line of human embryonic stem cells maintained in culture
H3K27 trimethylation	A certain type of chemical modification of a protein that binds DNA. Important for reversible deactivation oftargeted portions of the genome
Intron	Part of an RNA molecule that is included immediately after transcription and removed during maturation of that molecule
Intronic RNA	RNA encoded by a DNA sequence that also encodes an intron of another transcript
lincRNA-p21	A non-coding RNA activated upon DNA damage and in various tumor cell lines
MYC gene	A gene encoding an important regulator controlling activity of many genes. This gene has been associated with many cancers
Normal human epidermal keratinocytes (NHEK)	A line of primary keratinocytes maintained in culture
Non-coding RNA	RNA that is not used as a template for protein synthesis
Pervasive transcription	Massive transcription from unannotated regions of the genome
PolyA+ RNA	A molecule of RNA containing a long stretch of adenosine residues at the end
PRC2 chromatin signaling complex	A complex composed of multiple protein molecules that reversibly modifies chromatin and silences target genes
Promoter	A sequence of DNA that is located immediately adjacent to a target gene and regulates its activity
Pseudogene	A copy of a gene, presumed to be non-functional, although a number of recent examples describe both non-coding functions and occasionally coding functions for some of these loci
Regulation in *trans*	Regulation via interaction with molecules encoded by distal regions of the genome
RNA Pol II	A complex composed of multiple protein molecules responsible for synthesis of RNA, which is used as template for protein synthesis
Transcript	A molecule of RNA produced by transcription, that is, copying of RNA from the DNA template
Transcription factor	A protein that regulates expression of genes by binding to their promoters and/or enhancers
Transcription factor motif	A short DNA sequence recognized by a transcription factor or group of transcription factors, typically found in promoters and enhancers
Transcriptome	A collection of all the RNA molecules (transcripts) in a cell or a tissue
Transcriptomics	Study of the transcriptome
Xenopus oocytes	Oocytes from frogs of genus *Xenopus*, an important model system for study of developmental biology, cell biology, molecular biology, toxicology, and neuroscience

vast hidden regulatory layer composed of ncRNA transcripts and their network of interactions in the cell. Below we describe the evidence supporting this view, and explain how this perspective can solve a number of outstanding questions.

Undiscovered transcripts may underlie non-coding GWAS variants

As discussed by Mudge *et al.* [12] in their excellent review on functional transcriptomics, we have only begun to annotate the full complexity of RNAs encoded by the human genome. First, the database of complete, full-length cDNAs – the basis for gene annotations – still has surprisingly shallow coverage [10]. In fact, for many gene loci, the database contains only a single complete cDNA. This implies that most protein-coding genes, even well-characterized ones, have yet undiscovered exons that would require highly sensitive in-depth profiling methods to reveal [22,23]. With much less coverage than coding genes, the annotation situation for ncRNAs remains far more incomplete. The main reasons for this include the over-reliance on methods designed for protein-coding mRNAs, such as the use of polyA+ RNA for transcriptome profiling, the analytical focus on spliced RNAs, and the avoidance of intronic RNAs. The vast majority of RNA sequencing (RNAseq) experiments so far have profiled the polyA+ RNA fraction. Although this is informative for mRNAs, it leads to loss of significant complexity of ncRNAs [21], and a bias against the discovery of their unspliced versions. As a consequence, spliced versions of ncRNAs dominate the current annotated lists, which has resulted in an underestimate of their genomic coverage. It is very possible that, unlike protein-coding mRNAs, the longer, unspliced versions of the ncRNAs represent the functional forms. The abundance of ncRNAs in the nucleus compared to the cytosol [18] makes this a likely scenario.

Partly to bring order to this complexity, annotation efforts have classified ncRNAs by their physical characteristics. Typically, they are defined based on length (short, long, or very long), location relative to known genomic features (introns, genes, promoters, enhancers, or intergenic space), and overlap of known genes (sense or antisense) [12]. Of greatest importance to GWAS,long ncRNAs (lncRNAs) include all the classes of ncRNAs greater than 200 nucleotides in length, such as long intergenic ncRNAs (lincRNAs) [24], very long intergenic ncRNAs (vlincRNAs; (>40 kb in length, see below), natural antisense RNAs, and intronic RNAs [19,25-27]. Genomic annotation efforts typically focus on intergenic regions as the logical place to look for novel genes and transcripts, following the general notion that introns of known genes probably represent mere pre-mRNAs. However, this simple assumption has failed in the light of recent RNAseq datasets, as we have shown in a mouse inflammation time-course experiment [28]. In fact, thousands of mouse introns can harbor functional ncRNAs that behave separately from their exonic counterparts [28], resonating with discoveries of independent intronic RNAs in other systems such as *Xenopus* oocytes [29]. These observations support visionary ideas originally conceived by John Mattick almost 2 decades ago [30].

Strikingly, discovery of GWAS variants seem to favor different genomic elements depending on disease type (Figure 1). Only two disease types showed a preference for discovery of GWAS variants in coding regions of known genes (CDSs): metabolic diseases and anatomical diseases. By contrast, diseases affecting mental health favored annotated lncRNAs (lincRNAs and vlincRNAs), while cellular proliferation diseases favored introns (Figure 1). The even distribution in the non-disease trait category (which includes a large number of different phenotypes) provides a perspective for the contrasting results in the disease categories. Although understanding these observations will require additional research, we hope they raise the question of why variants in different diseases favor different categories of genomic elements. For example, it is tempting to speculate that the preference for promoters in anatomical diseases relates to the importance of tight control of gene expression during development.

As alluded to above, most of the existing lists of lncRNAs come from profiling of polyA+ RNA. By contrast, sequencing of total RNA from normal and tumor tissues recently uncovered vlincRNAs), a novel class of lncRNAs that showed statistically significant associations with GWAS variants [31]. Thousands of these vlincsRNAs span at least 10% of the human genome, and probably span much more, once additional tissues are profiled. These RNAs range from 40 kb up to around 1 MB in length, and are controlled by typical RNA Pol II promoters. Interestingly, an intriguing subset of these RNAs – those controlled by promoters within endogenous retroviral elements – characterizes cancerous and pluripotent states.

Figure 2 illustrates a cancer-associated region, for which GWAS has highlighted the potential importance of vlincRNAs. This 8q24 region upstream of the *MYC* gene shows a high level of transcription in two normal human cell lines: H1 embryonic stem cells and normal human epidermal keratinocytes (NHEK), which are primary keratinocytes [20]. vlincRNA transcription occurs on both strands, probably in part from normal RNA Pol II promoters [32] (Figure 2, red arrows). In total, vlincRNAs [31] span 727 kb of that 1.2 MB region and overlap a number of GWAS SNPs. Lower level transcription overlaps additional GWAS SNPs, suggesting the presence of unannotated transcripts in those regions, consistent with

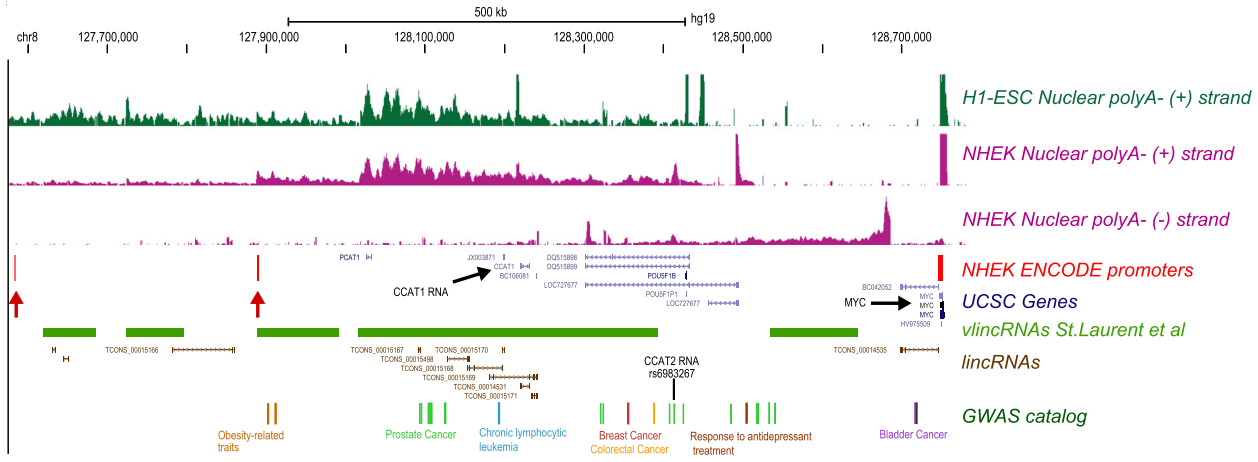

Figure 2 A genomic view of the 8q24 region upstream of the *MYC* gene. For details, see text.

the data in Figure 2. The two previously studied cancer-associated ncRNAs in this region, the 2,613 bp-long CCAT1 [33] and the 340 bp-long CCAT2 [34], represent only a tiny fraction of its transcriptional complexity. Quite possibly, they encompass only segments of much longer transcripts. In fact, the current profile of transcription leaves us with an attractive possibility that a cluster of GWAS SNPs spanning around 500 kb works via its presence in a small set of very long ncRNAs. Obviously, all this shows that we have only begun the exploration of RNAs made in this important region, and by association, many regions throughout the genome, where unexplained non-coding GWAS hits occur. Widespread presence of such very long RNAs suggests that vlincRNAs represent a global property of the human genome, and advances the theory that such RNAs mediate the functions of variants uncovered by GWAS (also see below).

Emerging patterns of lncRNA function in disease

A number of recent examples indicate that ncRNAs can underlie the function of non-coding GWAS SNPs. Perhaps the best-studied example is ANRIL, a lncRNA transcribed from the 9p21.3 locus. the single greatest GWAS risk factor for atherosclerosis that is currently known [35]. Remarkably, the expression level of this ncRNA stands out as the variable most strongly associated with the disease phenotypes linked to 9p21.3 [35]. Recent work has highlighted the importance of the atherogenic SNPs in this locus by demonstrating that ANRIL functions by *trans*-regulating over 900 genes [36]. The study combined various *in vitro* experiments with measurement of gene expression in 2,280 patients with cardiovascular disease from the Leipzig Heart Study to show that these ANRIL-regulated genes can be classified into known atherogenic Gene Ontology terms such as 'cell adhesion' and 'apoptosis' [36].

The study further showed that ANRIL interacts with the PRC2 chromatin signaling complex, and requires an intact Alu sequence (a short repeated sequence present in thousands of copies in the human genome) for its regulatory effects [36]. Delivery of the PRC2 complex, which promotes H3K27 trimethylation and repression of gene expression, to its targets via the Alu-mediated interaction thus provides an attractive model for ANRIL function. Other systems have previously provided similar examples of Alu repeats mediating intermolecular interactions between RNA molecules, and leading to functional consequences [37,38]. All this evidence suggests that many other lncRNAs might function through intermolecular interactions mediated by Alu and other abundant repeated sequences in mammalian genomes.

The molecular mechanism of ANRIL function illustrates a potential general paradigm in gene expression regulation that promises to explain the function of large numbers of lncRNAs. In this model, one type of RNA species could regulate hundreds of targets in *trans* via intermolecular interactions, a mode of regulation previously associated primarily with short RNAs such as micro RNAs. However, it is becoming increasingly clear that lncRNAs can function in this manner, perhaps by providing scaffolds that bring together various protein and RNA molecules [39]. In this regard, the functions of ANRIL parallel those of the lincRNA HOTAIR, which also *trans*-regulates hundreds of genes by changing the chromatin occupancy of PRC2 [40].

A newly discovered vlincRNA associated with Hemolysis, elevated liver enzymes, low platelet count syndrome provides another prominent example of this mode of function [41]. While this syndrome represents a mendelian disorder, mapped using a traditional genetic analysis of affected families, the mutations occur in a non-coding region that harbors an ncRNA approximately 200 kb long.

Subsequent analysis has shown that mutations can affect stability of this RNA [41], which also functions by *trans*-regulating hundreds of target genes.

Impressive as the aforementioned examples are, even more striking is the trend that has emerged from these and other investigations: more detailed examination of non-coding GWAS loci have increasingly led to the discovery of disease-relevant transcripts in the highlighted region. For example, a careful examination of the GWAS region at 5p14.1 that is implicated in autism led to the discovery of non-coding RNA antisense to a moesin pseudogene. Further experiments determined that this natural antisense RNA probably works in *trans* by lowering the level of moesin protein encoded by a gene on the X chromosome [42]. Supporting this emerging trend, depletion (small interfering RNA or antisense RNA) or overexpression of lncRNAs, even the ultra-long vlincRNAs, now routinely results in apparent phenotypes [31,41]. The growing wealth of examples precludes us from going into details of the studies or even citing all of them. Suffice to say that such functional analysis has begun to connect ncNAs, including those transcribed from GWAS loci, with processes such as cancer, heart disease, degenerative diseases, and senescence [31,36,40,41,43-45].

A complex transcriptome for complex diseases

The magnitude of functional transcripts in non-coding genomic space, many of which still remain either hidden or under-appreciated as functional RNAs, makes it almost certain that these RNAs will explain an important fraction of the non-coding GWAS hits. If so, then further intriguing questions arise. Why are mendelian disorders mostly explained by mutations affecting protein-coding exons, yet complex diseases are explained by mutations in ncRNAs? Does this notion form a pattern consistent with prevailingmechanistic models of how ncRNAs function? Does this pattern tell us something about the underlying systems biology of human cells?

The evidence thus far suggests positive answers to these questions. As described above in the few examples that have been worked out in detail, lncRNAs can act as *trans* regulators, potentially master regulators, of large numbers of known genes. Examples such as ANRIL, HOTAIR, lincRNA-p21, and HELLP highlight this emerging paradigm. A similar model of regulation occurs with transcription factors; however, a mutation in a long ncRNA would generally not have quite the same effect as a mutation in a protein. Considering that the former usually results in a much more flexible phenotype than the latter, a mutation would affect rather than abrogate the interaction affinity of the lncRNA with its partner molecules, such as proteins or other nucleic acids. Moreover, these interactions occur in the context of hundreds or thousands of competing and cooperative interactions with other lncRNAs in a complex ecosystem that controls signaling in the nucleus. The expected result would include small but cooperative and cumulative effects on a large number of downstream targets, thus displaying itself as a relatively subtle contribution to one or possibly many complex phenotypes.

Conclusions

Clearly, we are still at the early stages of understanding the full complexity of functional elements encoded in the human genome. However, recent results paint an emerging picture of a very complex regulatory network composed of numerous ncRNAs and their targets. Each ncRNA molecule in this network could potentially regulate hundreds of other RNAs in *trans*. GWAS variants could function by affecting this overarching layer of ncRNA regulation. In fact, the recent examples of this type of network regulation probably represent the tip of the iceberg of its true significance in complex diseases. Therefore, the time is right to bring the two fields together to fully unravel the underlying relationships.

In addition, the theme of ncRNA in disease brings with it an immediate implication for clinical research. If ncRNAs are as intricately involved in underlying disease mechanisms, as the data reviewed here suggest, then clinical transcriptome sequencing (including ncRNAs) has to come to the forefront of biomedical research. Indeed, first indications suggest that ncRNAs represent excellent biomarkers for cancer diagnostics [46,47]. Considering the much larger complexity of ncRNAs compared with coding RNAs, the former represent a gigantic untapped potential for clinically relevant biomarkers, in addition to expanding our basic knowledge of molecular events leading to disease.

Additional files

Additional file 1: Supplementary Text.
Additional file 2: Supplementary Tables 1-3.

Abbreviations

GWAS: genome-wide association study; lincRNAs: long intergenic non-coding RNAs; lncRNA: long non-coding RNA; ncRNA: non-coding RNA; NHGRI: National Human Genome Research Institute; SNP: single nucleotide polymorphism; vlincRNA: very long intergenic non-coding RNA.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GSL and PK jointly wrote the manuscript, and YV performed bioinformatics analysis. All authors read and approved the final manuscript.

Acknowledgments

We would like to thank Dmitry Shtokalo, Maxim Ri, Denis Antonets, Olga Saik, Tim McCaffrey and Mark Mazaitis for their helpful discussions and help with figure generation.

Author details
[1]St. Laurent Institute, 317 New Boston St, Suite 201, Woburn, MA 01801, USA. [2]AcademGene Ltd., 6, Acad. Lavrentyev ave., Novosibirsk 630090, Russia.

Received: 7 March 2014 Accepted: 22 May 2014
Published: 12 Jun 2014

References

1. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA: **Complement factor H variant increases the risk of age-related macular degeneration.** *Science* 2005, **308:**419–421.
2. *A Catalog of Published Genome-Wide Association Studies.* www.genome.gov/gwastudies.
3. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106:**9362–9367.
4. Barendse W: **The effect of measurement error of phenotypes on genome wide association studies.** *BMC Genomics* 2011, **12:**232.
5. Cardon LR, Palmer LJ: **Population stratification and spurious allelic association.** *Lancet* 2003, **361:**598–604.
6. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA: **Population structure, differential bias and genomic control in a large-scale, case-control association study.** *Nat Genet* 2005, **37:**1243–1246.
7. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661–678.
8. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, Balasubramanian S, Beal K, Chakravarty D, Challis D, Chen Y, Clarke D, Clarke L, Cunningham F, Evani US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gumus ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, *et al*: **Integrative annotation of variants from 1092 humans: application to cancer genomics.** *Science* 2013, **342:**1235587.
9. 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.
10. Kapranov P, St Laurent G: **Dark Matter RNA: Existence, Function, and Controversy.** *Front Genet* 2012, **3:**60.
11. Clark MB, Choudhary A, Smith MA, Taft RJ, Mattick JS: **The dark matter rises: the expanding world of regulatory RNAs.** *Essays Biochem* 2013, **54:**1–16.
12. Mudge JM, Frankish A, Harrow J: **Functional transcriptomics in the post-ENCODE era.** *Genome Res* 2013, **23:**1961–1973.
13. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, Rozowsky JS, Gerstein MB, Wahlestedt C, Hayashizaki Y, Carninci P, Gingeras TR, Mattick JS: **The reality of pervasive transcription.** *PLoS Biol* 2011, **9:**e1000625. discussion e1001102.
14. 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.
15. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo 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, Fiegler H, *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799–816.
16. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schonbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, *et al*: **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002, **420:**563–573.
17. 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.
18. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.** *Science* 2005, **308:**1149–1154.
19. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller 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.
20. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, *et al*: **Landscape of transcription in human cells.** *Nature* 2012, **489:**101–108.
21. Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, Sorensen PH, Reaman G, Milos P, Arceci RJ, Thompson JF, Triche TJ: **The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA.** *BMC Biol* 2010, **8:**149.
22. 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, Guigo 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.
23. 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* 2012, **30:**99–104.
24. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL: **Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression.** *Proc Natl Acad Sci U S A* 2009, **106:**11667–11672.
25. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME: **Widespread transcription at neuronal activity-regulated enhancers.** *Nature* 2010, **465:**182–187.
26. Wahlestedt C: **Targeting long non-coding RNA to therapeutically upregulate gene expression.** *Nat Rev Drug Discov* 2013, **12:**433–446.
27. Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization.** *Nat Rev Genet* 2007, **8:**413–423.
28. St Laurent G 3rd, Shtokalo D, Tackett M, Yang Z, Eremina T, Wahlestedt C, Inchima SU, Seilheimer B, McCaffrey TA, Kapranov P: **Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells.** *BMC Genomics* 2012, **13:**504.
29. Gardner EJ, Nizami ZF, Talbot CC Jr, Gall JG: **Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of Xenopus tropicalis.** *Genes Dev* 2012, **26:**2550–2559.
30. Mattick JS: **Introns: evolution and function.** *Curr Opin Genet Dev* 1994, **4:**823–831.
31. St Laurent G, Shtokalo D, Dong B, Tackett M, Fan X, Lazorthes S, Nicolas E, Sang N, Triche T, McCaffrey T, Xiao W, Kapranov P: **VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer.** *Genome Biology* 2013, **14:**R73.
32. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473:**43–49.
33. Nissan A, Stojadinovic A, Mitrani-Rosenbaum S, Halle D, Grinbaum R, Roistacher M, Bochem A, Dayanc BE, Ritter G, Gomceli I, Bostanci EB, Akoglu M, Chen YT, Old LJ, Gure AO: **Colon cancer associated transcript-1: a novel RNA expressed in malignant and pre-malignant human tissues.** *Int J Cancer* 2012, **130:**1598–1606.

34. Ling H, Spizzo R, Atlasi Y, Nicoloso M, Shimizu M, Redis RS, Nishida N, Gafa R, Song J, Guo Z, Ivan C, Barbarotto E, De Vries I, Zhang X, Ferracin M, Churchman M, van Galen JF, Beverloo BH, Shariati M, Haderk F, Estecio MR, Garcia-Manero G, Patijn GA, Gotley DC, Bhardwaj V, Shureiqi I, Sen S, Multani AS, Welsh J, Yamamoto K, *et al*: **CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer.** *Genome Res* 2013, **23:**1446–1461.
35. Pasmant E, Sabbagh A, Vidaud M, Bieche I: **ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS.** *FASEB J* 2011, **25:**444–448.
36. Holdt LM, Hoffmann S, Sass K, Langenberger D, Scholz M, Krohn K, Finstermeier K, Stahringer A, Wilfert W, Beutner F, Gielen S, Schuler G, Gabel G, Bergert H, Bechmann I, Stadler PF, Thiery J, Teupser D: **Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks.** *PLoS Genet* 2013, **9:**e1003588.
37. Gong C, Tang Y, Maquat LE: **mRNA-mRNA duplexes that autoelicit Staufen1-mediated mRNA decay.** *Nat Struct Mol Biol* 2013, **20:**1214–1220.
38. Wang J, Gong C, Maquat LE: **Control of myogenesis by rodent SINE-containing lncRNAs.** *Genes Dev* 2013, **27:**793–804.
39. St Laurent G, Savva YA, Kapranov P: **Dark matter RNA: an intelligent scaffold for the dynamic regulation of the nuclear information landscape.** *Front Genet* 2012, **3:**57.
40. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, Li R, West RB, van de Vijver MJ, Sukumar S, Chang HY: **Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis.** *Nature* 2010, **464:**1071–1076.
41. van Dijk M, Thulluru HK, Mulders J, Michel OJ, Poutsma A, Windhorst S, Kleiverda G, Sie D, Lachmeijer AM, Oudejans CB: **HELLP babies link a novel lincRNA to the trophoblast cell cycle.** *J Clin Invest* 2012, **122:**4003–4011.
42. Kerin T, Ramanathan A, Rivas K, Grepo N, Coetzee GA, Campbell DB: **A noncoding RNA antisense to moesin at 5p14.1 in autism.** *Sci Transl Med* 2012, **4:**128ra140.
43. Askarian-Amiri ME, Crawford J, French JD, Smart CE, Smith MA, Clark MB, Ru K, Mercer TR, Thompson ER, Lakhani SR, Vargas AC, Campbell IG, Brown MA, Dinger ME, Mattick JS: **SNORD-host RNA Zfas1 is a regulator of mammary development and a potential marker for breast cancer.** *RNA* 2011, **17:**878–891.
44. Khaitan D, Dinger ME, Mazar J, Crawford J, Smith MA, Mattick JS, Perera RJ: **The melanoma-upregulated long noncoding RNA SPRY4-IT1 modulates apoptosis and invasion.** *Cancer Res* 2011, **71:**3852–3862.
45. Abdelmohsen K, Panda A, Kang MJ, Xu J, Selimyan R, Yoon JH, Martindale JL, De S, Wood WH 3rd, Becker KG, Gorospe M: **Senescence-associated lncRNAs: senescence-associated long noncoding RNAs.** *Aging Cell* 2013, **12:**890–900.
46. Vergara IA, Erho N, Triche TJ, Ghadessi M, Crisan A, Sierocinski T, Black PC, Buerki C, Davicioni E: **Genomic "Dark Matter" in Prostate Cancer: Exploring the Clinical Utility of ncRNA as Biomarkers.** *Front Genet* 2012, **3:**23.
47. Reis EM, Verjovski-Almeida S: **Perspectives of Long Non-Coding RNAs in Cancer Diagnostics.** *Front Genet* 2012, **3:**32.

10.1186/1741-7015-12-97
Cite this article as: St. Laurent *et al.*: Dark matter RNA illuminates the puzzle of genome-wide association studies. *BMC Medicine* 2014, 12:97