Carcinogenesis vol.32 no.11 pp.1655–1659, 2011
doi:10.1093/carcin/bgr187
Advance Access publication August 19, 2011

Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk

Guangfu Jin[1], Jielin Sun[1,2], Sarah D.Isaacs[3], Kathleen E.Wiley[3], Seong-Tae Kim[1,2], Lisa W.Chu[1], Zheng Zhang[1,2], Hui Zhao[1,2], Siqun Lilly Zheng[1,2], William B.Isaacs[3] and Jianfeng Xu[1,2,*]

[1]Center for Cancer Genomics and [2]Center for Genomics and Personalized Medicine Research, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157, USA and [3]Department of Urology, Johns Hopkins Medical Institutions, Baltimore, MD 21205, USA

*To whom correspondence should be addressed. Tel: +1 336 713 7500; Fax: +1 336 713 7566;
Email: jxu@wfubmc.edu

Long non-coding RNAs (lncRNAs), representing a large proportion of non-coding transcripts across the human genome, are evolutionally conserved and biologically functional. At least one-third of the phenotype-related loci identified by genome-wide association studies (GWAS) are mapped to non-coding intervals. However, the relationships between phenotype-related loci and lncRNAs are largely unknown. Utilizing the 1000 Genomes data, we compared single-nucleotide polymorphisms (SNPs) within the sequences of lncRNA and protein-coding genes as defined in the Ensembl database. We further annotated the phenotype-related SNPs reported by GWAS at lncRNA intervals. Because prostate cancer (PCa) risk-related loci were enriched in lncRNAs, we then performed meta-analysis of two existing GWAS for discovery and an additional sample set for replication, revealing PCa risk-related loci at lncRNA regions. The SNP density in regions of lncRNA was similar to that in protein-coding regions, but they were less polymorphic than surrounding regions. Among the 1998 phenotype-related SNPs identified by GWAS, 52 loci were located directly in lncRNA intervals with a 1.5-fold enrichment compared with the entire genome. More than a 5-fold enrichment was observed for eight PCa risk-related loci in lncRNA genes. We also identified a new PCa risk-related SNP rs3787016 in an lncRNA region at 19q13 (per allele odds ratio = 1.19; 95% confidence interval: 1.11–1.27) with P value of 7.22 × 10^{-7}. lncRNAs may be important for interpreting and mining GWAS data. However, the catalog of lncRNAs needs to be better characterized in order to fully evaluate the relationship of phenotype-related loci with lncRNAs.

Introduction

Transcriptome analysis indicates a major portion of the human genome is transcribed, yet the minority of transcripts is translated into proteins (1–4). The non-protein-coding transcripts [termed non-coding RNAs (ncRNAs)] are generally divided into housekeeping and regulatory ncRNAs (5). Housekeeping ncRNAs include ribosomal, transfer small nuclear and small nucleolar RNAs, which are usually expressed constitutively. Among regulatory ncRNAs, there are at least two types: short ncRNAs, including microRNAs, small interfering RNAs and piwi-interacting RNAs, and long non-coding RNAs (lncRNAs). Although recent studies have revealed the functional importance of short ncRNAs (6–9), less is known about lncRNAs, which make up most of the transcribed ncRNAs (5).

lncRNAs are 100–200 nts or longer transcripts that are similar to transcripts of protein-coding genes but do not contain functional open-reading frames (10). These lncRNA transcripts may be located within the cell's nucleus or cytoplasm, may or may not be polyadenylated, and are often transcribed from either strand within a protein-coding locus (5). In contrast to other transcripts in human genome, such as those coding proteins and microRNAs, the biological function of lncRNAs is the least understood to date. Recent studies have shown that lncRNAs can regulate the expression of genes in close genomic proximity (cis-acting regulation) as well as target distant transcriptional activators or repressors (trans-acting) via a variety of mechanisms, such as transcriptional interference, initiation of chromatin remodeling, promoter inactivation by binding to basal transcriptional factors and activation of an accessory protein (5,9,11).

Over the past few years, genome-wide association studies (GWAS) have revealed a large number of genetic variants related to diseases and/or traits, but at least one-third of the identified variants are not within protein-coding genes and rather map to non-coding intervals (12). Although enhancers in the non-coding regions have been anticipated to contain some of these risk variants (13), another possibility is that these risk variants reside in ncRNAs, which are evolutionally conserved across mammals and are biologically functional as cis- and/or trans-regulators of gene activity (5–7,11,14). For example, recent emerging evidence has indicated the important role of genetic variants of microRNAs in diseases (15). However, to date, little is known about the genetic significance of lncRNAs.

In this study, based on the 1000 Genomes data (16), we summarized the single-nucleotide polymorphisms (SNPs) within the sequences of 1420 lncRNAs, as defined in the Ensembl database. Furthermore, according to the National Human Genome Research Institute (NHGRI) GWAS Catalog (17), we annotated SNPs identified by GWAS as associated with human diseases and/or traits at the lncRNA intervals. Finally, given the enrichment of prostate cancer (PCa)-related loci in lncRNAs, we sought to identify PCa risk-related loci at lncRNA regions using two existing GWAS.

Materials and methods

Identification of SNPs in lncRNA intervals

Data on lncRNA genes (n = 1420) and protein-coding genes (n = 34 627) across the human autosome genome was downloaded from the publicly available Ensembl database using the BioMart data-mining tool (18). All SNPs in these lncRNA genes, protein-coding genes or surrounding intervals were identified based on the 1000 Genomes pilot project releases (16).

Phenotype-related SNPs reported by GWAS

According to the NHGRI GWAS Catalog (17), phenotype-related SNPs reported by GWAS were defined according to following criteria: (i) at least 100 000 SNPs were genotyped in the initial stage, (ii) SNPs were selected in absence of the candidate gene approach, (iii) at least one replication stage was included, (iv) significance level was $<10^{-5}$ for a single SNP and (v) the last reported date was 31 December 2010. Considering that additional SNPs in linkage disequilibrium (LD) with reported phenotype-related loci may also map to lncRNA intervals; we performed LD analysis and detected the overlap of high LD SNPs with lncRNAs using an r^2 value of 0.5 as the threshold, based on European ancestry in Utah (CEU) genotype data of the 1000 Genomes project. For PCa risk-related loci, we selected all 33 PCa risk-associated SNPs exceeding genome-wide significance levels in initial reports ($P < 10^{-7}$) from GWAS reported before December 2010; these 33 SNPs have been replicated in several independent study populations (19–33).

Study populations of PCa studies

To test if any unreported SNPs in lncRNA intervals were potentially related to PCa risk, we performed a meta-analysis of two existing PCa GWAS, Johns Hopkins Hospital (JHH) and Cancer Genetic Markers of Susceptibility

Abbreviations: CGEMS, Cancer Genetic Markers of Susceptibility; GWAS, genome-wide association studies; LD, linkage disequilibrium; lncRNA, long non-coding RNA; ncRNA, non-coding RNA; PCa, prostate cancer; SNP, single-nucleotide polymorphism.

(CGEMS) followed by an additional replication (supplementary Figure 1 is available at *Carcinogenesis* Online). The first population was derived from a PCa GWAS study at JHH, which included 1964 Caucasian men with PCa undergoing radical prostatectomy from 1 January 1999 through 31 December 2008 (34). The clinical characteristics of these patients are presented in supplementary Table I (available at *Carcinogenesis* Online). The control subjects for this population were an independent group of 3172 Caucasian individuals from the Illumina iControlDB (iControls) dataset (35).

The second GWAS population was from Stage 1 of the National Cancer Institute CGEMS study (21). It included 1176 PCa cases and 1157 control subjects, selected from the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. The genotype and phenotype data of this study are publicly available, and our use of the data has been approved by CGEMS.

The replication population included an additional 1114 cases and 822 controls, which were also recruited from JHH but were not scanned by genome-wide SNPs chips. These subjects were used for replication of the selected loci from the meta-analysis of JHH and CGEMS GWAS. The characteristics are presented in supplementary Table I (available at *Carcinogenesis* Online).

Genotyping, imputation and quality control

GWAS in the JHH PCa cases was performed using the Illumina 610 K chip and the GWAS of the iControl population (http://www.illumina.com/science/icontroldb.ilmn) was performed using Illumina Hap300 and Hap550 Chips. Imputation was performed to call genotypes for untyped loci on the basis of HapMap Phase II using the program IMPUTE (36) with a posterior probability of 0.9 as a threshold. The quality control criteria used to filter SNPs included minor allele frequency <0.01, Hardy-Weinberg equilibrium <0.001 and call rate <0.95. In total, data on 41 017 SNPs in lncRNA regions were available for 1909 cases and 3085 controls from JHH, whereas data on 40 300 SNPs in lncRNA regions were available for 1176 cases and 1101 controls from CGEMS. After pooling by meta-analysis, data from 39 320 SNPs that were in common between the two GWAS were used to evaluate their association with PCa (supplementary Figure 2 is available at *Carcinogenesis* Online).

To confirm the results from GWAS, significant SNPs were selected for replication in the aforementioned additional set of JHH subjects according to following criterion: (i) $P < 0.001$, (ii) the locus was not reported previously, (iii) the most significant locus was selected for each locus. Ten SNPs were finally selected to be genotyped using the MassARRAY iPLEX system (Sequenom, San Diego, CA) at the Center for Cancer Genomics, Wake Forest University. Duplicates and water samples (negative control) were included in each 96-well plate for genotyping quality control. Genotyping was performed by technicians that were blinded to sample status.

Statistical analysis

The SNP density between lncRNAs and surrounding regions was compared using a paired sample *t*-test. The enrichment was assessed by comparing the density of phenotype-related loci across the genome with two measures: average chromosome length (kb) required for one locus and average number of SNPs containing one locus. The strength of enrichment is high when the measures are small. Association analysis between PCa risk and each SNP in regions of lncRNAs was tested using unconditional logistic regression with one degree of freedom. Per allele odds ratio and 95% confidence interval were estimated based on a log-additive genetic model. Meta-analysis was performed for each SNP between two GWAS or between two GWAS and the replication study based on a random effect model, which presents the pooled result in a conservative manner. All the analyzes were two sided and performed using SAS (v.9.2) and PLINK package (v.1.07) (37).

Results

Sequence variants in lncRNA intervals

A total of 1420 autosomal lncRNA genes were defined by the Ensembl database, with a median length of 6379 bps (range: 77–1 015 961 bps). As shown in Table I, 137 334 SNPs, 107 122 SNPs and 185 737 SNPs in lncRNAs were identified in populations of CEU, Han Chinese in Beijing and Japanese in Tokyo (CHBJPT) and Yoruba from Ibadan (YRI), Nigeria, respectively. In CEU, the density of SNPs in lncRNA regions was 2.685 SNPs/kb, which was similar to the average density across the whole genome (2.694) and in coding protein genes (2.687). Also in CEU, the average density for 1420 lncRNAs (2.53 ± 1.80 SNPs/kb) was significantly lower than flanking regions (2.61 ± 1.66 SNPs/kb; $P = 0.029$) (supplementary Figure 1 is available at *Carcinogenesis* Online). Similar trends were observed in populations of CHBJPT and YRI.

Phenotype-related SNPs and lncRNAs

Among 1998 unique SNPs that were related to phenotypes in reported GWAS, 1242 loci were in protein-coding genes, whereas 52 loci were mapped to the lncRNA intervals (Table I). The enrichment of phenotype-related loci was similar in lncRNA and protein-coding regions, ~1.5-fold of average levels in the whole genome. The 52 phenotype-related SNPs located in lncRNAs are associated with 30 phenotypes (Supplementary Table 2 is available at *Carcinogenesis* Online). After LD analysis of 1998 SNPs based on data from the CEU population, a total of 119 SNPs or SNPs in high LD ($r^2 > 0.50$) overlapped with the 1420 lncRNAs.

PCa risk-related SNPs are enriched in lncRNA intervals

To date, 33 SNPs have been independently associated with PCa risk in populations of European descent (Table II). Of note, eight PCa-related SNPs fall into the intervals of lncRNA. Compared with the average density of PCa risk-related SNPs in the human genome (33/3.02 billion bps) or among all SNPs in the genome (33/7.95 million SNPs), the identified PCa risk SNPs were enriched in intervals of lncRNA (genome: 8/53.4 million bps; variation: 8/0.34 million SNPs) by >5-fold.

Identification of novel PCa risk-related SNPs in lncRNA genes

Meta-analyzes of 39 320 SNPs in lncRNAs from JHH and CGEMS populations showed 93 SNPs were associated with PCa risk with *P* value <0.001 (supplementary Figure 3). Of these 93 SNPs, 60 were in the four PCa-related loci that were reported previously, including 8q24 region 1 and region 3, 10q11 and 17q12 (Table II). The remaining 33 SNPs were in 10 LD blocks. One SNP from each of the 10 LD blocks was selected for replication in an additional 1114 cases and 822 controls [Table III and supplementary Table 3 (available at *Carcinogenesis* Online)]. Of the 10 SNPs, 1 SNP (rs3787016 at 19q13) remained significant ($P = 0.011$) with the effect in the same direction as the meta-analysis of the two GWAS studies. After pooling the three populations (Table IV), the A allele of rs3787016 was associated with a 1.19-fold (95% confidence interval: 1.11–1.27) increased PCa risk, and a *P* value that reached 7.22×10^{-7}, which remained significant

Table I. SNPs in lncRNA intervals and phenotype-related loci identified by GWAS

Group	Total length (kb)	Caucasian		Asian		African		Phenotype-related loci enrichment[a]		
		SNPs	Density (SNPs/kb)	SNPs	Density (SNPs/kb)	SNPs	Density (SNPs/kb)	Loci	kb/locus	SNPs/locus
Whole genome	2 866 720	7 723 945	2.694	6 106 220	2.188	10 555 378	3.816	1998	1434.8	3865.8
Protein-coding genes	1 293 146	3 474 536	2.687	2 758 618	2.133	4 791 168	3.705	1242	1041.2	2797.5
lncRNA genes	51 140	137 334	2.685	107 122	2.095	185 737	3.632	52	983.5	2641.0

[a]The enrichment was assessed by comparing the density of phenotype-related loci across the genome with two measures: average chromosome length required for one locus (kb/locus) and average number of SNPs containing one locus (SNPs/locus) in Caucasian population. The strength of enrichment is high when the measures are small.

Table II. Summary of GWAS-identified PCa risk-related SNPs in lncRNAs

Chromosome	SNP	Region	lncRNA gene	Alleles	RA[a]	Reported OR[b]	OR (95% CI)[c]	*P*
2	rs1465618	2p21		A > G	A	1.15		
2	rs721048	2p15		G > A	A	1.18		
2	rs12621278	2q31		A > G	A	1.35		
3	rs2660753	3p12		C > T	T	1.24		
3	rs10934853	3q21		C > A	A	1.12		
4	rs17021918	4q22		C > T	C	1.14		
4	rs7679673	4q24		A > C	C	1.14		
6	rs9364554	6q25		C > T	T	1.17		
7	rs10486567	7p15		T > C	C	1.16		
7	rs6465657	7q21		C > T	C	1.14		
8	rs2928679	8p21		G > A	A	1.13		
8	rs1512268	8p21		G > A	A	1.17		
8	rs10086908	8q24 (Region 5)		T > C	T	1.13		
8	rs16901979	8q24 (Region 2)		C > A	A	1.82		
8	rs16902094	8q24.21	*RP11-382A18.1*	A > G	G	1.20	NA	NA
8	rs620861	8q24 (Region 4)	*RP11-382A18.1*	G > A	G	1.16	1.04 (0.94–1.15)[d]	0.420[d]
8	rs6983267	8q24 (Region 3)	*RP11-382A18.1*	T > G	G	1.20	1.25 (1.17–1.34)	5.54E-11
8	rs1447295	8q24 (Region 1)	*RP11-382A18.1*	C > A	A	1.47	1.40 (1.26–1.56)	1.85E-10
9	rs1571801	9q33		G > T	T	1.17		
10	rs10993994	10q11	AL450342.3	T > C	T	1.25	1.25 (1.17–1.34)	3.91E-11
10	rs4962416	10q26		A > G	G	1.15		
11	rs7127900	11p15		G > A	A	1.25		
11	rs12418451	11q13		G > A	A	1.16		
11	rs10896449	11q13		A > G	G	1.16		
17	rs11649743	17q12	AC091199.1	C > T	C	1.16	1.10 (0.98–1.27)	0.112
17	rs4430796	17q12	AC091199.1	T > C	T	1.22	1.23 (1.12–1.34)	4.69E-06
17	rs1859962	17q24		T > G	G	1.21		
19	rs8102476	19q13		A > G	G	1.12		
19	rs887391	19q13	AC005945.1	T > C	T	1.14	1.09 (0.96–1.25)	0.177
19	rs2735839	19q13		G > A	G	1.30		
22	rs9623117	22q13		T > C	C	1.13		
22	rs5759167	22q13		G > T	G	1.18		
X	rs5945619	Xp11		A > G	G	1.27		

[a] Risk allele (RA) reported in previous studies.
[b] Odds ratios (ORs) were derived from pooled results in reported studies of European descent (33).
[c] Odds ratios (ORs) and 95% confidence interval (95% CI) were presented with pooled results of CGEMS and JHH GWAS.
[d] The results for rs6208961 were not available and were represented by a high LD SNP rs445114 with the T allele as the risk allele.

Table III. Summary results for 10 SNPs in lncRNAs selected for replication with PCa risk

Chromosome	Position	SNP	Alleles	lncRNA gene	lncRNA interval	Ref. genotype	Pooled GWAS[a]		Replication[b]	
							P	OR (95% CI)	*P*	OR (95% CI)
7	23,305,084	rs17729322	G > T	AC005082.3	chr7:23012429-23393750	GG	2.35E-04	1.27 (1.12–1.44)	0.270	1.15 (0.90–1.47)
7	27,144,271	rs6976129	C > T	RP1-170O19.2	chr7:27136121-27160432	CC	5.66E-04	1.22 (1.09–1.36)	0.192	1.16 (0.93–1.44)
8	130,571,112	rs16904092	T > C	RP11-3O20.1	chr8:130433119-130761667	TT	1.49E-04	0.61 (0.48–0.79)	0.167	0.73 (0.47–1.14)
11	17,184,304	rs214901	A > G	AC107956.2	chr11:17171486-17186106	AA	2.46E-04	1.13 (1.06–1.21)	0.135	0.91 (0.80–1.03)
12	125,970,356	rs10773338	A > G	AC078878.1	chr12:125965736-126110895	AA	3.34E-04	0.85 (0.78–0.93)	0.026	1.21 (1.02–1.43)
12	125,994,441	rs10773343	G > T	AC078878.1	chr12:125965736-126110895	GG	7.35E-04	1.12 (1.05–1.20)	0.556	1.04 (0.91–1.18)
16	57,436,122	rs13338289	G > A	AC092378.1	chr16:57341043-57700379	GG	1.06E-04	0.83 (0.75–0.91)	0.423	0.93 (0.78–1.11)
16	57,459,519	rs4784993	C > G	AC092378.1	chr16:57341043-57700379	CC	7.53E-05	0.82 (0.75–0.91)	0.659	0.96 (0.81–1.15)
19	1,041,803	rs3787016	G > A	AC112706.1	chr19:732003-1096404	GG	2.09E-05	1.18 (1.09–1.27)	0.011	1.22 (1.05–1.41)
19	33,768,887	rs11667383	C > T	AC005394.1	chr19:33674613-33815475	CC	6.22E-04	1.12 (1.05–1.20)	0.693	0.97 (0.86–1.11)

[a] The JHH (1909 cases and 3085 controls) and CGEMS (1176 cases and 1101 controls) GWAS were pooled by meta-analysis.
[b] The replication subjects were from an additional 1114 cases and 822 controls in JHH.

after a conservative Bonferroni correction for 39 320 tests (Bonferroni-corrected *P* value: 1.27×10^{-6}).

Discussion

In the current study, we provided some evidence that lncRNAs, a major class of non-coding transcripts, may be important in certain disease etiology. On the basis of 1420 lncRNAs derived from the Ensembl database, we found that the regions of lncRNA had a SNPs density similar to protein-coding regions but were less polymorphic than surrounding regions; this observation is consistent with previous reports that the sequence of lncRNAs are evolutionary conserved (14,38). At least 52 phenotype-related SNPs are within the lncRNA genes and 67 additional loci containing a high LD SNP overlapped with intervals of lncRNAs. Our observations suggest that variation in lncRNA regions may contribute to disease etiology.

Our observation that some of the phenotype-related loci identified in non-coding regions (12) actually reside within or in LD with

Table IV. Results for association between rs3787016 at 19p13 and PCa risk in GWAS and replication in 4196 cases and 5007 controls

Population	Case/control	Minor allele frequency		OR (95% CI)[a]	*P*
		Case	Control		
JHH	1906/3084	0.270	0.235	1.20 (1.10–1.32)	9.34E-05
CGEMS	1176/1101	0.265	0.241	1.13 (0.99–1.29)	0.065
Pooled GWAS				1.18 (1.09–1.27)	2.09E-05
Replication	1114/822	0.262	0.226	1.22 (1.05–1.41)	0.011
Combined				1.19 (1.11–1.27)	7.22E-07

[a] Derived from trend test (degree of freedom = 1).

lncRNAs is biologically plausible because lncRNAs are functionally active non-coding transcripts (5,9,11). For example, Chung *et al.* (39) recently identified a lncRNA (PCa ncRNA 1) at the PCa risk-related loci of 8q24 region 2, which was found to be overexpressed in PCa cells and prostatic intraepithelial neoplasia and shown to be involved in prostate carcinogenesis through androgen receptor activity.

Our observation that PCa risk-related loci were enriched in lncRNA intervals suggest that other loci mapping to the lncRNAs may also be related to PCa. This observation prompted us to evaluate additional SNPs within lncRNAs for association with PCa risk. It is of note that the primary aim of our study was not to identify the PCa risk loci but to demonstrate the possibility that genetic variants in lncRNA intervals might be related to diseases. Our results provide a proof-of-principle for a new approach in future GWAS data-mining studies aiming to discover phenotype-related loci by concentrating on lncRNAs. This method can be regarded as a complementary approach to other protein-coding related methods such as pathway analysis (40) or gene-based analysis (41).

Based on our analysis of SNPs in lncRNAs, we identified a new PCa risk-related locus, rs3787016, which is located in AC1127096.1, a lncRNA spanning 364 kb at 19p13. The SNP rs3787016 also localizes to an intron of *POLR2E* gene, which encodes a subunit of RNA polymerase II and is responsible for synthesizing messenger RNA. Two previously published genome-wide linkage studies have identified this same region as a PCa susceptibility region (42,43). However, to date, the causal variants and potential biological mechanism underlying these observations remains unknown. Because the lncRNA AC1127096 was predicted *in silico*, future studies are needed to determine the true function of this lncRNA.

Limitations of this study should be noted. Firstly, our list of lncRNAs may not be comprehensive because we were limited by those that have been identified to date and included in the Ensembl database (1420 lncRNAs). It has been estimated that >5000 lncRNAs, probably equal to or larger than the number of protein-coding genes, might exist (9). Secondly, the catalog of lncRNAs across the genome has not been functionally characterized and thus the biological significance of lncRNAs is also largely unknown, which makes interpreting our results difficult. lncRNAs included in this study were annotated by the Ensembl lncRNA annotation pipeline, most of which have not been validated in experimental models. It is still unclear whether these SNPs reside in functional lncRNAs or if they modify the effects of the lncRNAs. Some insight may be gained by conducting genotype–lncRNA expression correlation analyzes to help establish the relationship between phenotype-related loci and lncRNAs. Findings from our study might guide future functional studies with respect to the phenotype-related loci that we localized to lncRNA regions.

In summary, our results indicate that lncRNAs are less polymorphic and may provide some functional interpretation for some of the phenotype-related loci identified by GWAS. We also identified a new PCa risk-related locus in the intervals of lncRNA, which serves as a proof-of-principle for an approach that can be used for further GWAS data mining, especially for non-coding regions. However, the catalog of lncRNAs is still not well characterized by functional studies and should be the focus of future studies in order to help in the interpretation of the relationship between phenotype-related loci and lncRNAs.

Supplementary material

Supplementary. Tables 1–3 and Figures 1–3 can be found at http://carcin.oxfordjournals.org/

Funding

National Institutes of Health (CA129684 to J.X., CA131338 to S.L.Z.); the Department of Defense Grant (W81XWH-07-1-0088 to J.X.) and the support of Kevin P.Jaffe to W.B.I. is gratefully acknowledged.

Acknowledgements

The authors thank all the study subjects who participated in the JHH study and physicians and researchers in designing and recruiting study subjects, including Drs Bruce J.Trock, Alan W.Partin and Patrick C.Walsh. The authors also thank for the National Cancer Institute CGEMS for making the data available publicly.

Conflict of Interest Statement: None declared.

References

1. Bertone,P. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242–2246.
2. ENCODE Project Consortium *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
3. Cheng,J. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.
4. Kapranov,P. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.
5. Ponting,C.P. *et al.* (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.
6. Carthew,R.W. *et al.* (2009) Origins and mechanisms of miRNAs and siRNAs. *Cell*, **136**, 642–655.
7. Malone,C.D. *et al.* (2009) Small RNAs as guardians of the genome. *Cell*, **136**, 656–668.
8. Baek,D. *et al.* (2008) The impact of microRNAs on protein output. *Nature*, **455**, 64–71.
9. Selbach,M. *et al.* (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature*, **455**, 58–63.
10. Ørom,U.A. *et al.* (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell*, **143**, 46–58.
11. Nagano,T. *et al.* (2011) No-nonsense functions for long noncoding RNAs. *Cell*, **145**, 178–181.
12. Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
13. Visel,A. *et al.* (2009) Genomic views of distant-acting enhancers. *Nature*, **461**, 199–205.
14. Guttman,M. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
15. Mattick,J.S. (2009) The genetic signatures of noncoding RNAs. *PLoS Genet.*, **5**, e1000459.
16. 1000 Genomes Project Consortium *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
17. Hindorff,L.A. *et al.* A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies (31 December 2010, date last accessed).
18. ENSEMBL BioMart [database on the Internet] uswest.ensembl.org/biomart/martview (31 December 2010, date last accessed).
19. Amundadottir,L.T. *et al.* (2006) A common variant associated with prostate cancer in European and African populations. *Nat. Genet.*, **38**, 652–658.
20. Gudmundsson,J. *et al.* (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.*, **39**, 631–637.

21. Yeager,M. *et al*. (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.*, **39**, 645–649.
22. Gudmundsson,J. *et al*. (2007) Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat. Genet.*, **39**, 977–983.
23. Duggan,D. *et al*. (2007) Two genome-wide association studies of aggressive prostate cancer implicate putative prostate tumor suppressor gene DAB2IP. *J. Natl Cancer Inst.*, **99**, 1836–1844.
24. Thomas,G. *et al*. (2008) Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.*, **40**, 310–315.
25. Gudmundsson,J. *et al*. (2008) Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat. Genet.*, **40**, 281–283.
26. Eeles,R.A. *et al*. (2008) Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.*, **40**, 316–321.
27. Yeager,M. *et al*. (2009) Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat. Genet.*, **41**, 1055–1057.
28. Gudmundsson,J. *et al*. (2009) Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat. Genet.*, **41**, 1122–1126.
29. Eeles,R.A. *et al*. (2009) Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.*, **41**, 1116–1121.
30. Sun,J. *et al*. (2008) Evidence for two independent prostate cancer risk-associated loci in the HNF1B gene at 17q12. *Nat. Genet.*, **40**, 1153–1155.
31. Hsu,F.C. *et al*. (2009) A novel prostate cancer susceptibility locus at 19q13. *Cancer Res.*, **69**, 2720–2723.
32. Sun,J. *et al*. (2009) Sequence variants at 22q13 are associated with prostate cancer risk. *Cancer Res.*, **69**, 10–15.
33. Kim,S.T. *et al*. (2010) Prostate cancer risk-associated variants reported from genome-wide association studies: meta-analysis and their contribution to genetic Variation. *Prostate*, **70**, 1729–1738.
34. Xu,J. *et al*. (2010) Inherited genetic variant predisposes to aggressive but not indolent prostate cancer. *Proc. Natl Acad. Sci. USA*, **107**, 2136–2140.
35. Illumina iControlDB [website on the Internet]. Illumina, San Diego, CA, 2010 http://www.illumina.com/science/icontroldb.ilmn (11 September 2009, date last accesssed).
36. Price,A.L. *et al*. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
37. Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
38. Ponjavic,J. *et al*. (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.*, **17**, 556–565.
39. Chung,S. *et al*. (2011) Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci.*, **102**, 245–252.
40. Zhong,H. *et al*. (2010) Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 581–591.
41. Liu,J.Z. *et al*. (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.*, **87**, 139–145.
42. Hsieh,C.L. *et al*. (2001) A genome screen of families with multiple cases of prostate cancer: evidence of genetic heterogeneity. *Am. J. Hum. Genet.*, **69**, 148–158.
43. Wiklund,F. *et al*. (2003) Genome-wide scan of Swedish families with hereditary prostate cancer: suggestive evidence of linkage at 5q11.2 and 19p13.3. *Prostate*, **57**, 290–297.

Received June 23, 2011; revised August 3, 2011; accepted August 4, 2011