Integrating Omics Data

George Tseng, Debashis Ghosh, Xianghong Jasmine Zhou

Book DOI: http://dx.doi.org/10.1017/CBO9781107706484

Online ISBN: 9781107706484

Hardback ISBN: 9781107069114

Paperback ISBN: 9781107697577

Chapter

18 - Data Integration on Noncoding RNA Studies pp. 403-424

Chapter DOI: http://dx.doi.org/10.1017/CBO9781107706484.019

Cambridge University Press

18

Data Integration on Noncoding RNA Studies

ZHOU DU, TENG FEI, MYLES BROWN, X. SHIRLEY LIU, AND YIWEN CHEN

Abstract

Recent genome-wide studies revealed that the human genome encodes over 10,000 long non-coding RNAs (lncRNAs) with little protein-coding capacity. Growing evidence suggests that many lncRNAs may have important functions in complex diseases and are potentially a new class of therapeutic targets for treating complex disease. In contrast to the fast pace of cataloguing lncRNAs in the human genome, the function of the vast majority of lncRNAs remain unknown. In this chapter, we described data integration strategies for identifying lncRNA that are associated with cancer subtypes and clinical prognosis, and predicted those that are potential drivers of cancer progression.

18.1 Introduction

The advancement in high-throughput technologies such as microarray, next-generation sequencing (NGS) has greatly facilitated cost-effective large-scale data generation. As a result, the amount of genomic data deposited into various public data sources such as Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) has grown tremendously in the past several years. Taking NCBI short reads archive database (http://www.ncbi.nlm.nih.gov/sra) as an example, the amount of data in this database went from about 10 terabytes (TB) in 2008 to about 1000 TB in 2012, an around 100-fold increase in only four years. These public data sources not only provide the raw data for the researchers to reproduce the discovery that were reported in the original study but also provided opportunities for using the same data for new discoveries. Moreover, integrating the data across individual studies either horizontally or vertically offers unique opportunities to make novel discoveries that would have been impossible based on the data from a single study. The integration of genomic data from the same individual under a specific disease condition is particularly powerful for disease-relevant

discoveries. In those genomics-based clinical studies, the orthogonal genomic data and corresponding clinical information were systematically collected from the same group of human subjects. These data can be integrated to discover genes that play important roles in the etiology of the disease and those that may serve as diagnostic, prognostic, and predictive biomarkers.

Recent transcriptome profiling in human cells from the ENCODE (http://encodeproject.org/ENCODE/) and GENCODE (http://www.gencodegenes.org/) projects showed that cumulatively ~70% of the human genome [1] can be transcribed, whereas only ~2% of the genome encodes proteins. In contrast to ~20,000 protein encoding genes (PCGs), there are ~35,000 (GENCODE) noncoding RNA genes in the human genome. The noncoding RNAs can be classified as either small noncoding RNAs (sncRNAs), which are shorter than or equal to 200 base-pair (bp), or long noncoding RNAs (lncRNAs), which are longer than 200 bp. Data integration has played a pivotal role in identifying the sncRNAs, especially microRNAs (miRNAs) in different species, and predicting the targets and biological function of miRNAs in physiology and disease [2–7]. Although significant knowledge has been accumulated on the sncRNA biology in the past decade with the joint effort of computational and experimental research, the identity and function of the lncRNAs in human genome are just beginning to be revealed. Data integration has played a critical role in identifying the lncRNA genes from a variety of genomic data in different biological contexts as well as in providing the evidence for lncRNA function [8–10]. Systematic efforts to catalog lncRNAs by traditional cDNA Sanger sequencing [11] and the integration of histone mark chromatin immunoprecipitation sequencing (ChIP-seq) [9, 12] and RNA sequencing (RNA-seq) [8, 13] data have revealed that the human genome encodes more than 10,000 lncRNAs. We refer the interested readers to other published reviews for data integration studies on both sncRNAs [2–4, 7] and lncRNAs [8–10]. This chapter is dedicated to describing the approaches to integrate the data from clinical studies for elucidating lncRNA function and uncovering its potential utility in diagnosis and prognosis in human diseases such as cancer [14].

Given their lower expression level compared with protein-coding genes (PCGs) [8], it has been debated whether the lncRNAs are simply the transcriptional noise in the cell or whether they may have biochemical function. Although we do not know how many of them are functional, growing evidence suggests that lncRNAs, similar to PCGs, may play important roles in both development [15] and human diseases such as cancer [16]. A growing list of lncRNAs has been shown to mediate oncogenic or tumor-suppressing effects in cancer, and they promise to be a new class of cancer therapeutic targets [17]. Although a handful of lncRNAs have been functionally characterized, little

is known about the functions of most lncRNAs in normal physiology or disease [18]. LncRNAs may serve as cancer diagnostic or prognostic biomarkers that are independent of PCGs. A well-known example of a cancer diagnostic biomarker is PCA3 [19], a prostate-specific lncRNA gene that is significantly overexpressed in prostate cancer. Noninvasive monitoring of the ratio of urinary PCA3 and prostate-specific antigen (PCA) transcript level was recently approved by FDA as a diagnostic assay for prostate cancer [20].

In this chapter, we present a case study of data integration in a cancer-related lncRNA study [14], in which we identified lncRNA that are associated with cancer subtypes and clinical prognosis, and predicted those that are potential drivers of cancer progression in multiple cancers, including glioblastoma multiforme (GBM) [21], ovarian cancer (OvCa) [22], lung squamous cell carcinoma (lung SCC) [23], and prostate cancer [24]. We validated our predictions of two tumorgenic lncRNAs by experimentally confirming the prostate cancer cell growth dependence on these two lncRNAs. Our integrative analysis provided a resource of clinically relevant lncRNA for development of lncRNA biomarkers and identification of lncRNA therapeutic targets for human cancer.

18.2 Methods

18.2.1 Repurposing Microarray Data to Interrogate lncRNA Expression

As lncRNAs do not encode proteins, their functions are closely associated with their transcript abundance. Though RNA-seq is a comprehensive way to profile lncRNA expression, publicly available RNA-seq data sets of tumors are relatively limited compared to array-based expression profiles because of the high cost associated with the adoption of this technique. In addition, RNA-seq data sets with low sequencing coverage or small sample numbers have only limited statistical power to discover clinically relevant lncRNAs. In contrast, there are a large number of data sets that contain array-based gene expression profiles across hundreds of tumor samples. These array-based expression profiles are often accompanied by matched clinical annotation and/or genomic alteration profiles of tumors such as somatic copy number alteration (SCNA). Although lncRNAs are not the intended targets of measurement in the original array design, microarray probes can be reannotated for interrogating lncRNA expression [25–27]. Compared with RNA-seq data of low sequencing coverage, array-based expression data may have lower technical variation and better detection sensitivity for low-abundance transcripts [28, 29], which is a prominent feature of lncRNAs [8]. Moreover, array-based expression data contain strand information and allow for interrogating the expression of antisense single-exon lncRNAs, whereas most current RNA-seq data in clinical

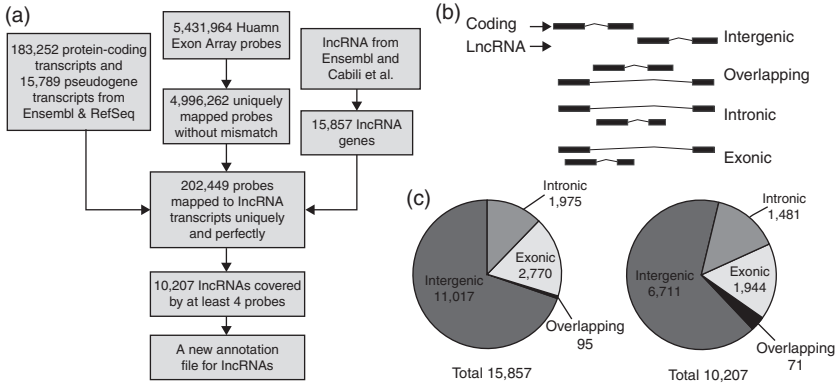

Figure 18.1 (A) Affymetrix Human Exon array probe reannotation pipeline for lncRNA. (B) Adopting the classification scheme from a previous study [34], lncRNA were classified into four categories, intergenic, overlapping, intronic, and exonic, on the basis of their relationship with protein-coding genes. (C) Pie charts showing the number of lncRNA in each category for all collected lncRNA and for those with at least four uniquely mapped exon array probes.

applications do not have strand information and thus are unable to accurately quantify the expression of this class of lncRNAs [30].

Among the different gene expression microarray platforms, we focused on reannotating the probes from the Affymetrix microarrays. These arrays not only have many more short probes that are likely to map to lncRNA genes but also have been the most widely used platforms for gene expression profiling of clinical studies. A computational pipeline was designed as follows to reannotate the probes from five major Affymetrix array types (Figure 18.1A) using the latest annotations of lncRNA and PCG. The lncRNA annotations were derived from two sources: the catalog of lncRNAs from the Ensembl database [31] (*Homo sapiens* GRCh37, release 67) and the catalog of lncRNAs generated on the basis of transcriptome assembly from RNA-seq data [8]. For those lncRNA transcripts with overlap on the same strand between these two sources, we only kept the Ensembl annotation to avoid redundancy. This resulted in a total of 15,857 lncRNA genes. We reannotated probe sets of the affymetrix microarrays for lncRNAs by mapping all probes to the human genome (hg19) by using SeqMap [32]. To avoid potential cross-hybridization of transcribed regions in the genome other than lncRNAs, we only kept those probes that mapped uniquely to the genome with no mismatch and removed all probes that mapped to protein-coding transcripts (183,252) or pseudogene transcripts (15,789) on the basis of the annotations from the Ensembl [31] and UCSC [33] databases.

Table 18.1 *Number of probes corresponding to lncRNAs and number of lncRNAs with at least four probes, coverage in five major Affymetrix array platforms*

	No. of probes corresponding to lncRNAs	No. of lncRNAs with at least four probes
Affymetrix Human Exon array	202,449	10,207
Affymetrix U95Av array	1865	76
Affymetrix U133 plus 2.0 array	43,752	2561
Affymetrix U133B array	21,880	1181
Affymetrix U133A array	2830	143

The preceding strategy was applied to generate the probes that corresponded to lncRNA transcripts for both Affymetrix exon array and the other 3′ IVT Affymetrix array platforms (Table 18.1). Among the five Affymetrix array types, the Affymetrix Human Exon 1.0 ST array has the most comprehensive coverage of the annotated human lncRNAs (Table 18.1), and we used the case of Affymetrix exon array for demonstration. By matching the selected probes to the lncRNA sequences, we obtained 202,449 probes from exon array and 10,207 corresponding lncRNA genes with at least four probes covering their annotated exons (Figure 18.1A), comprising approximately 64% of all 15,857 lncRNA genes (with over 60% coverage in each category [34] of the lncRNA genes) collected in this study (Figures 18.1B and 18.1C). The raw intensity of the exon array probes was corrected with a probe sequence–specific background model, and the expression level of a lncRNA gene was calculated by summarizing the background-corrected intensity of all probes corresponding to this gene [35]. The lncRNA expression was quantile normalized across different biological samples. The gene expression calculation was implemented with Jetta [36]. When batch information was available, Combat [37], an empirical Bayes method, was used to remove potential batch effects.

To gauge the reliability of our approach, we examined the correlation of both lncRNA and PCG expression between exon array and RNA-seq data on the same prostate cancer cell line LNCaP that were generated from two different laboratories [24, 38]. RNA-seq-based gene expression was calculated with Cufflinks1.0.2 [39] (default parameters and the –G option), and the exon array–based gene expression was calculated by the same procedure as was described earlier. The Pearson correlation coefficient was used to quantify the strength of the associations between the exon array–based and RNA-seq-based expression levels. We found that both PCGs ($r = 0.70$, $P < 2.2 \times 10^{-16}$) and lncRNAs ($r = 0.29$, $P < 2.2 \times 10^{-16}$) showed significant concordance of expression between the exon array and RNA-seq data. This observation is consistent with the

previous finding that the correlation between microarray and RNA-seq data is lower in genes with low expression [40], as lncRNAs are generally expressed at lower levels than PCGs [8]. As the level of probe coverage could also influence the accuracy of lncRNA expression derived from a microarray, we further investigated how the correlations of expression between the exon array and RNA-seq data change at different probe coverages by examining those PCGs with expression levels similar to those of lncRNAs. We found that the correlation between exon array- and RNA-seq-based expression showed a moderate increase when all probes (0.28) were used as compared with when only four probes (0.20) were used. The correlations were similar for PCGs (0.28) and lncRNAs (0.29) when we controlled for expression level. These results suggest that although probe coverage may influence the array-based lncRNA expression estimation, the dominant factor that governs the observed difference in correlation between array and RNA-seq data for PCGs and lncRNAs is their expression level. A recent study, in which a 60-mer custom oligonucleotide array was designed to investigate lncRNA expression, showed that the correlation of lncRNA expression between the custom array and RNA-seq data was between 0.24 and 0.31 [34]. Therefore, although the concordance between exon array and RNA-seq data is lower for lncRNA expression than for PCG expression, it may represent the typical performance in comparison of lncRNA expression between an array-based platform and RNA-seq. These examinations demonstrated the reliability of the usage of our reannotated exon array in measuring lncRNAs' expression and laid a foundation for our further study.

18.2.2 Integrating lncRNA Expression, Somatic Copy Number Alteration Data, and Clinical Information

One of the most important goals of disease research, especially in cancer research, is to identify driver genes that causally contribute to the disease initiation, progression, and maintenance, as these driver genes can potentially serve as targets for therapeutic interventions. Reliable identification of driver genes is challenging. The emergence of genomic technologies such as microarray and next-generation sequencing has greatly facilitated the identification of driver genes with the aid of computational methods. The expression data alone are insufficient for indentifying driver genes because the aberrant gene expression during the course of disease progression could be attributed to an indirect effect that is secondary to the major disease-causing events. Therefore it is important to integrate genomic data from different sources to enhance the specificity to indentify genes that may play a causal function in disease etiology. Aside from

expression data, an important data source that is informative for identifying driver genes is genetic alteration data. For instance, in cancer, a disease with the hallmark of genomic instability [41, 42], many types of somatic genetic alterations are specific to the cancer genome but not to the genome of the normal tissue. These somatic genetic alterations include nucleotide substitution mutations and small insertion/deletions (indels), copy number gains and losses, and chromosomal rearrangements. The copy number gains and losses is a particularly interesting type of somatic genetic alteration because it can often be linked to aberrant gene expression, which makes it a powerful data source in combination with expression profile to identify concordant genetic and gene expression abnormality. The joint analysis of genome-wide somatic copy number alteration profile can lead to the discovery of driver genes by narrowing the vast number of genomic and expression changes in cancer to a small subset that may be more functionally relevant [43, 44]. It can also lead to improvements in cancer diagnosis by utilizing copy number alteration as additional biomarkers [43, 45].

The high-resolution characterization of the SCNA profile in the cancer genome has been made possible by the emergence of both array-based and NGS-based genomic technologies. Array comparative genomic hybridization (aCGH) is among the earliest techniques for characterizing genome-wide somatic copy number alternation in cancer genome. All aCGH arrays are two channel, and they work by first differentially labeling and hybridizing tumor genomic DNA and normal genomic DNA on a microarray that contains hundreds of thousands of probes [46–48]. The ratio between a tumor and the matched normal sample is then calculated for each probe. To quantify the change of copy number difference, the log of base 2 is usually used so that the log-ratio of 1 and -1 corresponds to double or half as many copies, respectively. The log-ratio of 0 corresponds to no change in the copy number in tumor sample compared to the normal sample at that genomic location. Using the ratio values from all the probes that correspond to different genomic locations, the copy number alteration profile along the chromosome can be inferred. There are two major types of aCGH. The first type of aCGH utilizes bacterial artificial chromosome (BAC) probes, which are typically several hundred bp in length [46]. The BAC aCGH has a median genomic resolution of several mega-bases [46]. The second type of aCGH is the oligonucleotide platform. Such oligonucleotide platforms as those from Agilent and Nimblegen have probes shorter than 100 bp, and each array has from hundreds of thousands to more than 1 million probes. Given the difference in design and manufacturing of the aCGHs (probe length, hybridization chemistry, etc.), BAC and oligonucleotide aCGH have their own technical characteristics and may serve for different applications.

With the longer probe, the BAC aCGH in general has higher specificity in the hybridization signal of each probe, and each probe gives more accurate measurement, but it has lower resolution than oligonucleotide aCGH. However, for many applications, in which the aberration of interest is large, the resolution BAC aCGH is rather sufficient. In contrast, the oligonucleotide array has shorter probes and gives more noisy measurement on the individual probe level but provides higher genomic resolution.

In addition to the aCGH platforms, single nucleotide polymorphism (SNP) arrays can also be used to infer somatic copy number alterations in the cancer genome. The SNP arrays are mostly single-color arrays, in which only a tumor or a normal sample is hybridized on a microarray that contains oligonucleotide probes (25–50 bp). The two most popular SNP array platforms are the Affymetrix [49] and Illumina [50] SNP arrays. These arrays contained from hundreds of thousands to more than 1 million probes for inferring SNPs and/or copy number variations. The SNP arrays have the important advantage of measuring copy number alterations and loss of heterozygosity (LOH) simultaneously [51], but they have the disadvantage that the probe design and positioning are not optimal for the estimation of copy number. The advent of next-generation sequencing and the rapid increase in its throughput have made it possible to characterize copy number alteration with a much higher resolution ($<$10 kb) than aCGH or SNP arrays via whole-genome or whole-exome sequencing [52].

Characterizing somatic copy number amplifications and deletions in cancer genome with high resolution is only the first step in inferring genomic regions, the alteration of which are functionally important for the etiology of cancer. Once the genomic alterations have been detected, the next challenge is to distinguish between driver genomic alterations that confer a selective advantage for the tumor to initiate, grow, or persist and passenger genomic alterations that confer no selective advantages. To address this challenge, it is important to perform joint analysis of the somatic genomic alteration profiles across many tumors. Several algorithms [53] are designed for identifying those regions with aberrations that occur significantly more often than would be expected by chance, using permutation tests that are based on the overall pattern of aberrations seen across the genome. In the current study, we used two well-established algorithms, GISTIC [54, 55] and RAE [56], to identify the regions that harbor recurrent SCNAs by using SNP and aCGH data, respectively. As those regions with recurrent SCNAs often contain many lncRNA genes, we further integrate SCNA data and expression data to identify potential driver lncRNA genes based on the reasoning that functional SCNA should cause gene expression change and the driver lncRNAs should show higher or lower gene

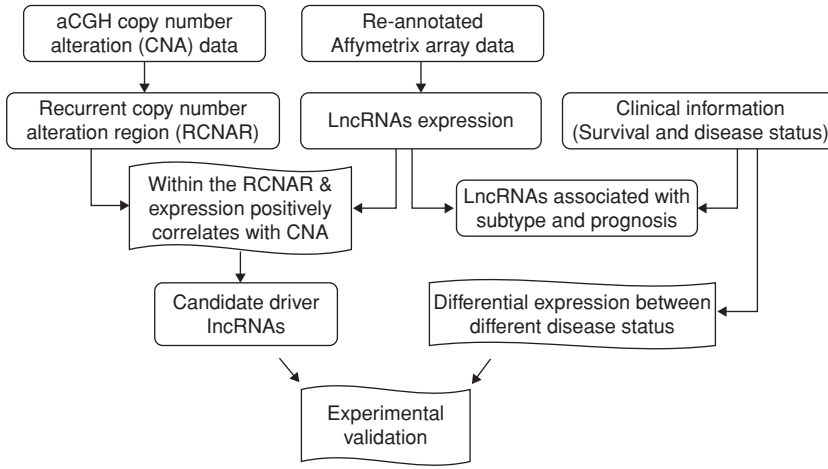

Figure 18.2 The workflow of integrating SCNA data, lncRNA expression data, and clinical information to identify the lncRNAs that are associated with cancer subtypes and clinical prognosis and/or those that are potential drivers of cancer progression.

expression in tumors with the corresponding genomic amplification or deletion compared with the rest tumors (Figure 18.2).

Besides the identification of potential driver lncRNAs, we integrated lncRNA expression with clinical information of individual patient samples including disease status (normal tissue vs. primary or metastatic tumor), subtype, and overall or progression-free survival information of the corresponding patient to predict those lncRNAs that showed different expression between disease status, subtype-specific expression, and/or associations with disease prognosis (Figure 18.2). The significance of differential expression between different statuses was assessed by Mann-Whitney U-test. To identify the lncRNAs that are associated with prognosis, the expression of which is associated with prognosis, we performed multivariate Cox proportional hazard (Cox regression) analyses to assess the associations between lncRNA expression with overall and progression-free survival while controlling for potentially cofounding clinical variables, including ethnicity, age, and gender.

18.3 Application

Using the earlier described approaches, we performed integrative analyses of lncRNA expression profiles, clinical information, and SCNA profiles of tumors in four different cancer types, including 150 tumor samples of prostate cancer from the Memorial Sloan-Kettering Cancer Center (MSKCC) Prostate

Oncogenome Project [24] and 451 tumor samples of glioblastoma multiforme (GBM) [21], 585 tumor samples of ovarian cancer (OvCa) [22], and 113 tumor samples of lung squamous cell carcinoma (lung SCC) [23] from the Cancer Genome Atlas Research Network (TCGA) project [21]. For prostate cancer, the data set includes exon array data, clinical annotation and SCNA data from the Gene expression Omnibus (GEO) (GSE21034). The SCNA regions were determined as the union of SCNA regions from two different studies [24, 57]. Recurrent SCNA regions across different tumors were identified by the algorithms, GISTIC [54] and RAE [56]. The magnitude of the SCNAs was estimated as the log2 ratios of segmented copy numbers between cancer and control DNAs. The exon array data, clinical annotations, and SCNA data of GBM, OvCa, and lung SCC were downloaded from TCGA (https://tcga-data.nci.nih.gov). We further obtained exon array data of 11 human normal tissues from Affymetrix (http://www.affymetrix.com/).

To validate the utility of exon array data in combination with clinical annotation to identify cancer-related lncRNAs, we examined the expression patterns of 13 literature-curated cancer-related lncRNAs [17] that have corresponding exon array probes in a prostate cancer data set [24]. This data set consists of 29 normal prostate samples, 131 primary prostate tumor samples, and 19 metastatic prostate tumor samples with exon array data [24] (Figure 18.3A). Notably, 9 out of these 13 known cancer-related lncRNAs showed significantly different expression between the tumor and normal prostate samples (Mann-Whitney U-test, $p < 0.05$). Three out of these nine lncRNAs were directly related to prostate cancer, including one known prostate cancer diagnostic biomarker, PCA3 [19, 20], and two lncRNAs, PCAT-1 [38] and PCGEM1 [58], that have been functionally implicated in prostate cancer progression. GAS5, a tumor-suppressive lncRNA known to be down-regulated in breast cancer [59], showed increased expression in prostate cancer (Table 18.2), a result suggesting complex and context-dependent functions of lncRNAs in different cancer types. Notably, several lncRNAs, such as NEAT1 [60], DANCR [61], HOTTIP [62], PRINS [63], and EGOT [64], that have established functions in forming nuclear speckles [60], in development [61] and in autoimmune disease [63], but were not previously known to be related to cancer, showed differential expression between tumor and normal prostate samples (Table 18.2), and this suggests their potential function in prostate cancer.

We next sought to identify lncRNAs that showed significant expression differences between tumors and normal prostate tissues and found 109 up-regulated and 104 down-regulated lncRNAs (Mann-Whitney U-test, false discovery rate < 0.05, fold change > 1.5) (Figure 18.3A). Notably, among the lncRNAs with sufficient exon array probe coverage, we rediscovered seven out

Table 18.2 *Known cancer-related lncRNAs or lncRNAs with established function in noncancer context and their regulation in cancer compared with normal prostate tissue*

Ensembl ID	Gene name	MW-*U* test *p*-value	Cancer vs. normal	Function annotation
ENSG00000225937	*PCA3*	9.50E-12	Up	Prostate cancer
ENSG00000234741	*GAS5*	1.77E-06	Up	Breast cancer
ENSG00000249859	*PVT1*	4.93E-11	Up	Multiple cancers
ENSG00000226950	*DANCR*	3.03E-08	Up	Development
ENSG00000253438	*PCAT1*	1.12E-05	Up	Prostate cancer
ENSG00000227418	*PCGEM1*	4.49E-04	Up	Prostate cancer
ENSG00000245532	*NEAT1*	0.00642	Up	Nuclear speckle
	KCNQ1O			
ENSG00000258492	*T1*	0.0103	Up	Colon cancer
ENSG00000251164	*HULC*	0.0311	Up	Multiple cancers
ENSG00000251562	*MALAT1*	0.285	–	Multiple cancers
ENSG00000214548	*MEG3*	3.92E-08	Down	Multiple cancers
ENSG00000238115	*PRINS*	1.37E-07	Down	Autoimmune disease
ENSG00000243766	*HOTTIP*	1.95E-06	Down	Development
ENSG00000235947	*EGOT*	2.48E-05	Down	Development
ENSG00000214049	*UCA1*	2.11E-02	Down	Bladder cancer
ENSG00000228630	*HOTAIR*	0.0573	–	Multiple cancers
ENSG00000130600	*H19*	0.0842	–	Multiple cancers
ENSG00000240498	*ANRIL*	0.699	–	Prostate cancer

Note: The statistical significance of the expression difference between cancer and normal prostate tissue was evaluated by Mann-Whitney *U*-test (MW-*U* test)

of eight lncRNAs that were reported to show higher expression in prostate cancer from an independent study based on RNA-seq data [38]. Furthermore, we identified an additional 102 lncRNA genes that were up-regulated in prostate cancer but were missed by the other study [38], and this suggests that arrays and RNA-seq may be complementary methods to identify clinically relevant lncRNAs.

Cancer is a clinically heterogeneous disease, and individual cancer types can be further divided into molecular subtypes, each with specific biological and clinical behaviors. Previous studies established four subtypes of GBM (proneural, neural, classical, and mesenchymal) [21], four subtypes of OvCa (immunoreactive, proliferative, mesenchymal, and differentiated) [22], and four subtypes of lung SCC (basal, classical, primitive, and secretory) [23] on the basis of the expression profiles of PCGs, and six subtypes of prostate cancer on the basis of the SCNA profiles [24]. LncRNAs with subtype-specific expression may have an important function in individual molecular subtypes. We compared

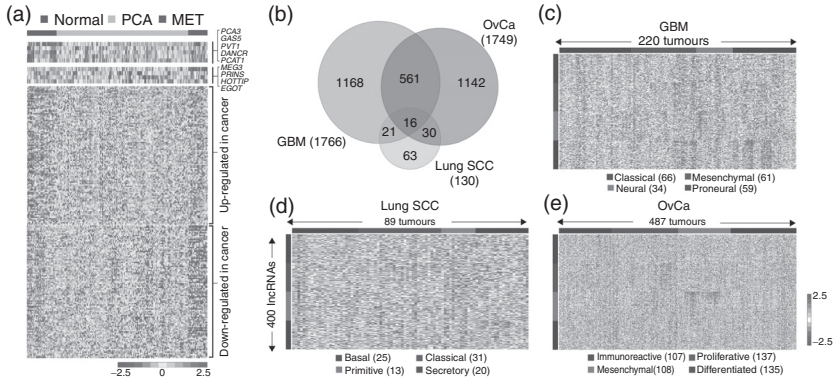

Figure 18.3 (A) The expression level of lncRNA that showed significantly differential expression between cancer and normal prostate tissues shown in heatmap across 29 normal prostate samples and 131 primary and 19 metastatic prostate tumor samples. Several known cancer-related lncRNA or lncRNA with established function in a noncancer context were highlighted. (B) Venn diagram representing the number of subtype-specific lncRNA in three cancers. The expression profile of the top 100 lncRNA that exhibited significantly higher expression in one subtype than the others for (C) GBM, (D) OvCa, and (E) Lung SCC shown in heatmap. (Note: the rank was based on the ascending order of the p-value.) Tumor samples were hierarchically clustered within each subtype.

lncRNA expression across different subtypes and identified hundreds of lncRNAs showing subtype-specific expression patterns in GBM, OvCa, and lung SCC (FDR < 0.05; Figures 18.3B–18.3E). The same approach did not yield any lncRNAs with significant subtype-specific expression in prostate cancer, which was reminiscent of the lack of a robust PCG expression–based subtype of prostate cancer [24]. In addition, 628 lncRNAs showed subtype-specific expression in more than one cancer type (Figure 18.3B), and some of these lncRNAs have been functionally implicated in other physiological or pathological processes. For example, MIAT, a lncRNA that showed specific expression in the mesenchymal subtype of OvCa and the proneural subtype of GBM, is known to confer risk of myocardial infarction [65] and regulate retinal cell fate specification [66]. In addition, RMST, a lncRNA known to be differentially expressed between rhabdomyosarcoma subtypes [67], also showed subtype-specific expression patterns in GBM, OvCa, and lung SCC. The lncRNAs that showed statistically higher expression (false discovery rate < 0.05) in only one subtype were considered to be subtype specific.

A previous study of HOTAIR [16, 68] showed that patients with higher HOTAIR expression had poorer prognosis in colorectal cancer [69]. To identify the lncRNAs that are associated with clinical outcome in prostate cancer,

GBM, OvCa, and lung SCC, we performed multivariate Cox regression analysis to evaluate the significance of the correlations between individual lncRNA expression and overall and progression-free survival in the presence of other confounding factors such as ethnicity, age, and gender. With these data, we are able to identify lncRNAs in prostate cancer, GBM, OvCa, and lung SCC whose expression was significantly correlated with overall or progression-free survival ($p < 0.01$). Notably, nine lncRNAs showed consistent positive or negative correlations between their expression and overall or progression-free survival in different cancer types, and this suggests their potential as more general prognostic biomarkers. The lncRNA gene with the Ensembl ID ENSG00000261582 is an example of a lncRNA that showed negative correlation between its expression and overall survival in both lung SCC and OvCa (Figure 18.4A). This lncRNA also showed subtype-specific expression in OvCa but not in lung SCC. Additionally, five lncRNAs showed marked and consistent positive or negative correlations between both overall and progression-free survival in OvCa (one such example, Ensembl ID ENSG00000225128, is shown in Figure 18.4B).

An important form of somatic genetic alteration in cancer is SCNAs, in which a genomic region is either amplified or deleted. Some of the genes within amplified (or deleted) regions show increased (or decreased) expression levels, leading to altered activity in cancer cells. Studies have suggested that the genes with causal roles in oncogenesis are often located in the SCNAs that are frequently altered across tumors [57, 69, 70]. To reveal the lncRNAs that may have tumor-promoting or -suppressing functions, we identified hundreds of lncRNAs that map to regions of recurrent SCNAs across tumors for prostate cancer, GBM, OvCa, and lung SCC (Figure 18.4C). Some of these lncRNAs also showed marked correlation between overall or progression-free survival [14]. In addition, we identified lncRNAs that were consistently located in regions of SCNAs across different cancers (Figure 18.4C) and found a significant overlap of the lncRNA genes that are located in SCNA gain or loss regions between some of the cancer types [14]. Among the many genes located within regions of SCNAs, probably only a fraction of them are drivers of cancer. To further distinguish driver from passenger lncRNAs in the regions of SCNAs, we integrated SCNA and expression profiles of lncRNAs in tumors. We reasoned that driver lncRNAs with SCNAs should result in corresponding gene expression changes [70, 71], as only those SCNAs that cause changes in transcript abundance could possibly alter lncRNA activity. Therefore, we selected lncRNAs whose SCNAs showed positive correlations with expression level changes as candidate drivers for prostate cancer, GBM, OvCa, and lung SCC. Among the lncRNAs in the SCNA regions, we selected those that showed significant and concordant expression changes (one-tailed Mann-Whitney U-test,

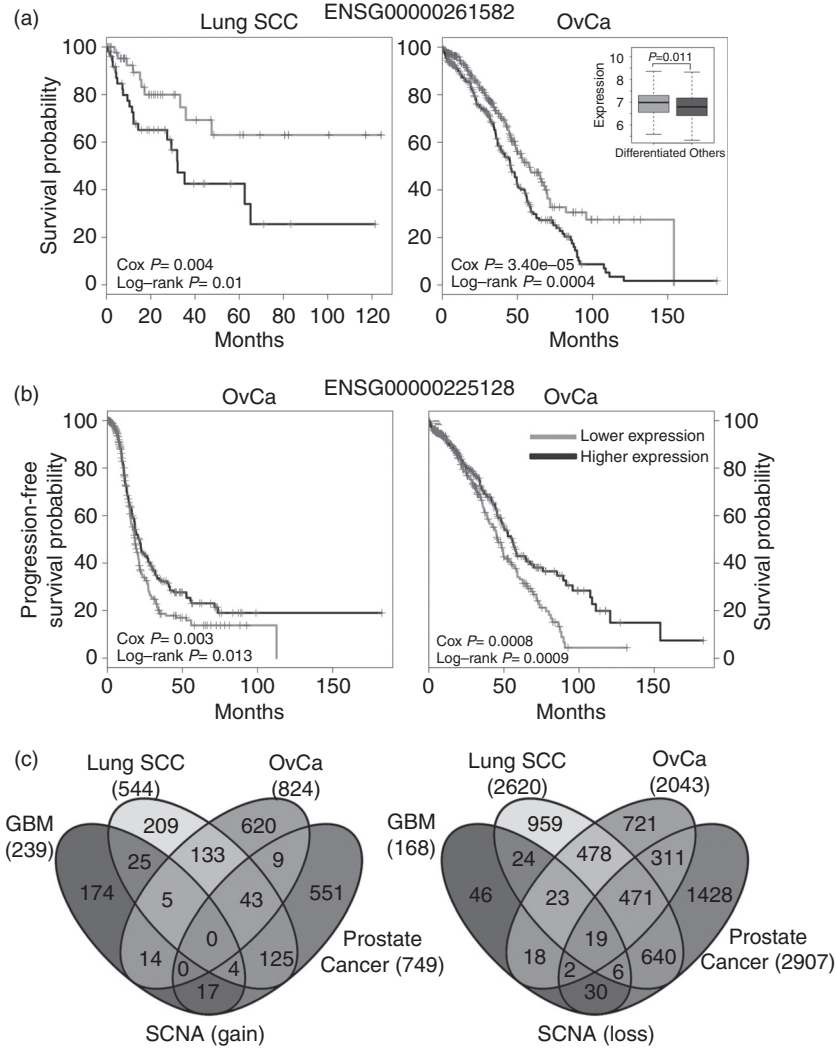

Figure 18.4 (A) Kaplan-Meier curve of two patient groups with higher (top 50%) and lower expression (bottom 50%) of ENSG00000261582 in Lung SCC and OvCa (red, higher expression; blue, lower expression). The box plot demonstrates that ENSG00000261582 was expressed higher in the "differentiated" subtype of OvCa than the other subtypes. Both the *p*-value of the multivariate Cox model for lncRNA expression and the *p*-value of the log-rank test were shown. (B) Kaplan-Meier curve for overall and progression-free survival of two patient groups with higher (top 50%) and lower expression (bottom 50%) of ENSG00000263041 in OvCa. (C) Number of lncRNA located in the SCNA (gain) and SCNA (loss) regions in different cancers shown as Venn diagrams.

$p < 0.05$) in tumor samples with a corresponding somatic copy number gain (log2 ratio > 0.2) or loss (log2 ratio < -0.2) compared to the other samples [14].

To further validate the reliability of the integrative studies, and as it is prohibitive to validate all candidate driver lncRNAs in the four cancer types, we focused our experimental validation and comprehensive annotation on candidate lncRNAs that may have tumor-promoting functions in prostate cancer (i.e., those in recurrent SCNA (gain) regions that showed positive correlations between their SCNAs and expression levels). Among all the candidate driver lncRNAs that showed increasing expression from normal to primary to metastatic prostate cancer, we chose the two that showed the most significant expression difference between tumor and normal prostate tissue (i.e., the two with the smallest p-values calculated by Mann-Whitney U-test) for experimental validation. The criterion of increasing expression from normal to primary to metastatic prostate cancer aimed to uncover lncRNAs that may be important therapeutic targets for both primary and metastatic cancers.

We named these two lncRNAs prostate cancer–associated noncoding RNAs 1 and 2, abbreviated as PCAN-R1 (Ensembl ID ENSG00000228288) and PCAN-R2 (Ensembl ID ENSG00000231806), respectively. Both lncRNAs showed positive correlations between gene expression and the advancement of the disease status and SCNAs (Figures 18.5A and 18.5B). To confirm that the two lncRNAs PCAN-R1 and PCAN-R2 are noncoding, we used two different methods, txCdsPredict from UCSC and phyloCSF [72], to calculate their coding potential. For coding-potential calculations with phyloCSF, we used the multiple sequence alignment of 29 mammalian genomes [73]. We chose the thresholds used previously (txCdsPredict = 800 [38] and phyloCSF = 100 [8]), below which the transcripts were considered to be noncoding. We found that the scores of all possible opening reading frames from the PCAN-R1 and PCAN-R2 transcripts were well below the thresholds (txCdsPredict scores: PCAN-R1, 470 and PCAN-R2, 359; phyloCSF scores: PCAN-R1, –123.1434 and PCAN-R2, –148.5448), supporting that these two lncRNA genes are noncoding.

We chose the prostate cancer cell line LNCaP, in which both lncRNAs have moderate or higher expression levels compared with their expression in other prostate cancer or non–prostate cancer cell lines, for experimental validation. Using 5′ and 3′ rapid amplification of cDNA ends (RACE), we found that for PCAN-R1, although one isoform (PCAN-R1-A) was almost identical to the Ensembl annotated transcript ENST00000425295 (Figure 18.5C), the other isoform (PCAN-R1-B) was a spliced variant of PCAN-R1-A with an intron

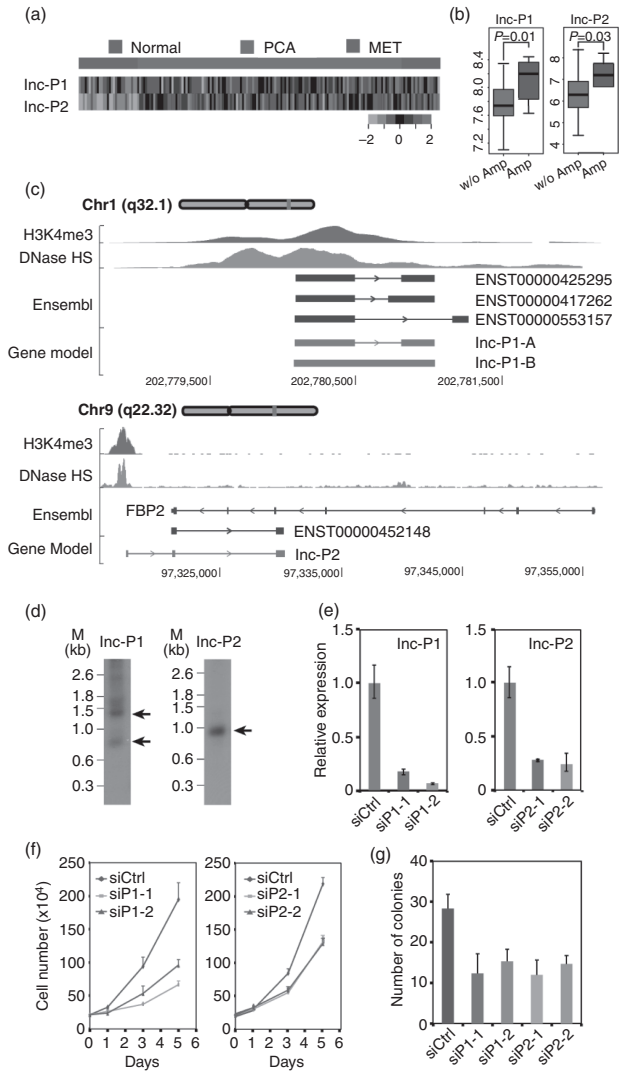

Figure 18.5 Experimental validation of lnc-P1 and lnc-P2 function. (A) Heatmap showing the expression of lnc-P1 and lnc-P2 in normal prostate tissue, primary and metastatic prostate cancer. (B) Box plot of lnc-P1 and lnc-P2 expression in tumors with genomic amplification and in the tumors without genomic amplification. (C) Transcript structure of lnc-P1 and lnc-P2 from Ensembl annotation and determined by 5′ and 3′ RACE experiments in LNCaP cell. In addition, the H3K4me3 and DNase I hypersensitive region profiles in the same cell line are shown. (D) The Northern blot of lnc-P1 and lnc-P2 transcripts. (E) Relative expression level of lnc-P1 and lnc-P2 upon knockdown by two different siRNA (purple and orange) and upon control siRNA treatment (green). (F) Growth curves of LNCaP cell with or without targeted siRNA-mediated knockdown of lnc-P1 or lnc-P2. The growth curves of control siRNA-treated cells and the growth curves of two targeted siRNA-treated cells plotted in purple, orange, and green, respectively. (G) Number of soft-agar colony formation of LNCaP cell with or without targeted siRNA-mediated knockdown of lnc-P1 or lnc-P2.

retention (Figure 18.5C). Notably, for PCAN-R2, the major isoform had an extra exon in the 5′ end, and the remaining two exons also had different lengths from the Ensembl annotation (Figure 18.5C). The new 5′ exon of PCAN-R2 was more consistent with the profile of histone H3 Lys4 trimethylation (H3K4me3), a histone mark of an active promoter and the profile of DNase I hypersensitive regions (i.e., the regions with an open chromatin state) in LNCaP cells.

We confirmed the transcript structures of PCAN-R1 and PCAN-R2 by northern blot and performed short interfering RNAs (siRNAs) knockdown experiments and observed the substantial decreases in cell growth. Additional experiments were further conducted and concordantly proved the influence on cancer cell growth caused by the expression of two lncRNAs. As a lncRNA may act in *cis* and influence the expression of its neighboring PCG, we investigated whether the expression of the neighboring PCG was regulated by PCAN-R1 or PCAN-R2. siRNA knockdown of PCAN-R1 or PCAN-R2 had no effect on the expression of their neighboring PCGs KDM5B and FBP2, respectively, and this suggests that the functional mechanisms of PCAN-R1 and PCAN-R2 are not directly through their neighboring PCGs. Notably, in normal tissues, PCAN-R1 and its neighboring PCG KDM5B showed the highest expression in testis. In contrast, although PCAN-R2 showed similar expression across different tissues, its neighboring PCG FBP2 showed a muscle-specific expression pattern, thus suggesting that the expressions of PCAN-R2 and FBP2 may be differently regulated.

18.4 Discussion

The case study presented in this chapter has demonstrated that integrating the orthogonal genomic data, such as lncRNA expression profiles, and somatic copy number alteration along with clinical information can greatly facilitate the discovery of lncRNA that may serve as therapeutic targets and diagnostic or prognostic biomarkers. Our analyses also indicate that repurposing microarray probes to construct a lncRNA expression profile in a patient sample is a cost-effective approach given the large number of such data sets available in public repositories. The constructed gene expression profiles of both lncRNAs and PCGs from our analyses are a valuable resource for understanding the similarities and differences of transcriptional (e.g., antisense RNA [74]) regulation of PCGs by lncRNAs across different cancer types. In the combination of matched SCNA profile and clinical information, these gene expression profiles also allow network models to be inferred [75, 76], which will help advance the understanding of lncRNA function in cancer etiology.

The experimental validation of two lncRNAs without previous implication in cancer suggests the effectiveness of our integrative analyses in finding functionally important lncRNAs in cancer. Our analyses predicted about 80–300 candidate driver lncRNAs that may have tumor-promoting functions in each of the four cancer types. An intersection of such a list of candidate driver lncRNAs with a list of lncRNAs generated from orthogonal functional genomic data sets, such as that generated by ribonucleoprotein immunoprecipitation followed by sequencing [77] (a genomic technique for identifying lncRNAs physically associated with the protein of interest), would greatly help prioritize their functional valuation in different biological contexts, including epigenetic regulation, and facilitate the discovery of lncRNA therapeutic targets.

In our current study, we only used SCNA and expression data in combination with clinical information for our integrative analysis. It is conceivable that other types of genomic data, such as SNP array [78] and genome sequencing data [52], can be further integrated to reveal the multifaceted relationship between the mutation spectrum and expression of lncRNAs, disease status, and clinical outcome.

In summary, we report a proof-of-principle study for identifying clinically relevant lncRNAs through integrative analyses of orthogonal genomic data sets and clinical information. Our study opens new avenues for leveraging publicly available genomic data to study the functions and mechanisms of lncRNAs in human disease.

References

1 Djebali, S., et al. Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
2 Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
3 Muniategui, A., Pey, J., Planes, F. J., & Rubio, A. Joint analysis of miRNA and mRNA expression data. *Brief Bioinform* **14**, 263–278 (2012).
4 Frampton, A. E., et al. Integrated analysis of miRNA and mRNA profiles enables target acquisition in human cancers. *Expert Rev Anticancer Ther* **12**, 323–330 (2012).
5 Berezikov, E. Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet* **12**, 846–860 (2011).
6 Pritchard, C. C., Cheng, H. H., & Tewari, M. MicroRNA profiling: approaches and considerations. *Nat Rev Genet* **13**, 358–369 (2012).
7 Chen, K., & Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* **8**, 93–103 (2007).
8 Cabili, M. N., et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915–1927 (2011).

9 Guttman, M., et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
10 Guttman, M., & Rinn, J. L. Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339–346 (2012).
11 Ota, T., et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* **36**, 40–45 (2004).
12 Khalil, A. M., et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667–11672 (2009).
13 Guttman, M., et al. Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503–510 (2010).
14 Du, Z., et al. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat Struct Mol Biol* **20**, 908–913 (2013).
15 Tian, D., Sun, S., & Lee, J. T. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. *Cell* **143**, 390–403 (2010).
16 Gupta, R. A., et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076 (2010).
17 Prensner, J. R., & Chinnaiyan, A. M. The emergence of lncRNAs in cancer biology. *Cancer Discov* **1**, 391–407 (2011).
18 Wapinski, O., & Chang, H. Y. Long noncoding RNAs and human disease. *Trends Cell Biol* **21**, 354–361 (2011).
19 Lee, G. L., Dobi, A., & Srivastava, S. Prostate cancer: diagnostic performance of the PCA3 urine test. *Nat Rev Urol* **8**, 123–124 (2011).
20 Hessels, D., & Schalken, J. A. Urinary biomarkers for prostate cancer: a review. *Asian J Androl* **15**, 333–339 (2013).
21 Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
22 Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
23 Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
24 Taylor, B. S., et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22 (2010).
25 Liao, Q., et al. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res* **39**, 3864–3878 (2011).
26 Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F., & Mattick, J.S. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A* **105**, 716–721 (2008).
27 Michelhaugh, S. K., et al. Mining Affymetrix microarray data for long non-coding RNAs: altered expression in the nucleus accumbens of heroin abusers. *J Neurochem* **116**, 459–466 (2010).
28 Raghavachari, N., et al. A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *BMC Med Genomics* **5**, 28 (2012).
29 Xu, W., et al. Human transcriptome array for high-throughput clinical studies. *Proc Natl Acad Sci U S A* **108**, 3707–3712 (2011).
30 Levin, J. Z., et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**, 709–715 (2010).

31 Flicek, P., et al. Ensembl 2012. *Nucleic Acids Res* **40**, D84–90 (2012).
32 Jiang, H., & Wong, W. H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24**, 2395–2396 (2008).
33 Kuhn, R. M., Haussler, D., & Kent, W. J. The UCSC genome browser and associated tools. *Brief Bioinform* **14**, 144–161 (2012).
34 Derrien, T., et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775–1789 (2012).
35 Kapur, K., Xing, Y., Ouyang, Z., & Wong, W. H. Exon arrays provide accurate assessments of gene expression. *Genome Biol* **8**, R82 (2007).
36 Seok, J., Xu, W., Gao, H., Davis, R. W., & Xiao, W. JETTA: junction and exon toolkits for transcriptome analysis. *Bioinformatics* **28**, 1274–1275 (2012).
37 Johnson, W. E., Li, C., & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
38 Prensner, J. R., et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* **29**, 742–749 (2011).
39 Trapnell, C., et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511–515 (2010).
40 Wang, Z., Gerstein, M., & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).
41 Frohling, S., & Dohner, H. Chromosomal abnormalities in cancer. *N Engl J Med* **359**, 722–734 (2008).
42 Hanahan, D., & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
43 Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *Science* **331**, 1553–1558 (2011).
44 Albertson, D. G., Collins, C., McCormick, F., & Gray, J. W. Chromosome aberrations in solid tumors. *Nat Genet* **34**, 369–376 (2003).
45 Hanash, S. Integrated global profiling of cancer. *Nat Rev Cancer* **4**, 638–644 (2004).
46 Pinkel, D., et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20**, 207–211 (1998).
47 Pinkel, D. & Albertson, D. G. Array comparative genomic hybridization and its applications in cancer. *Nat Genet* **37 Suppl**, S11–S17 (2005).
48 Lee, C., Iafrate, A. J., & Brothman, A. R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet* **39**, S48–S54 (2007).
49 Matsuzaki, H., et al. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* **1**, 109–111 (2004).
50 Shen, R., et al. High-throughput SNP genotyping on universal bead arrays. *Mutat Res* **573**, 70–82 (2005).
51 Beroukhim, R., et al. Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays. *PLoS Comput Biol* **2**, e41 0323–0332 (2006).
52 Meyerson, M., Gabriel, S., & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**, 685–696 (2010).
53 Yuan, X., Zhang, J., Zhang, S., Yu, G., & Wang, Y. Comparative analysis of methods for identifying recurrent copy number alterations in cancer. *PLoS ONE* **7**, e52516 (2013).

54 Beroukhim, R., et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A* **104**, 20007–20012 (2007).
55 Mermel, C. H., et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
56 Taylor, B. S., et al. Functional copy-number alterations in cancer. *PLoS ONE* **3**, e3179 (2008).
57 Beroukhim, R., et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
58 Petrovics, G., et al. Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. *Oncogene* **23**, 605–611 (2004).
59 Mourtada-Maarabouni, M., Pickard, M. R., Hedge, V. L., Farzaneh, F., & Williams, G. T. GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene* **28**, 195–208 (2009).
60 Clemson, C. M., et al. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol Cell* **33**, 717–726 (2009).
61 Kretz, M., et al. Suppression of progenitor differentiation requires the long noncoding RNA ANCR. *Genes Dev* **26**, 338–343 (2012).
62 Wang, K. C., et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120–124 (2011).
63 Szegedi, K., et al. The anti-apoptotic protein G1P3 is overexpressed in psoriasis and regulated by the non-coding RNA, PRINS. *Exp Dermatol* **19**, 269–278 (2010).
64 Wagner, L. A. et al. EGO, a novel, noncoding RNA gene, regulates eosinophil granule protein transcript expression. *Blood* **109**, 5191–5198 (2007).
65 Ishii, N., et al. Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. *J. Human Genet*. **51**, 1087–1099 (2006).
66 Rapicavoli, N. A., Poth, E. M., & Blackshaw, S. The long noncoding RNA RNCR2 directs mouse retinal cell specification. *BMC Dev Biol* **10**, 49 (2010).
67 Chan, A. S., Thorner, P. S., Squire, J. A., & Zielenska, M. Identification of a novel gene NCRMS on chromosome 12q21 with differential expression between rhabdomyosarcoma subtypes. *Oncogene* **21**, 3029–3037 (2002).
68 Rinn, J. L., et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
69 Kogo, R., et al. Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res* **71**, 6320–6326 (2011).
70 Garraway, L. A., et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436**, 117–122 (2005).
71 Akavia, U. D., et al. An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005–1017 (2010).
72 Lin, M. F., Jungreis, I., & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
73 Lindblad-Toh, K., et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
74 Tran, V. G., et al. H19 antisense RNA can up-regulate Igf2 transcription by activation of a novel promoter in mouse myoblasts. *PLoS ONE* **7**, e37923 (2012).

75 Califano, A., Butte, A. J., Friend, S., Ideker, T., & Schadt, E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat Genet* **44**, 841–847 (2012).
76 Pe'er, D., & Hacohen, N. Principles and strategies for developing network models in cancer. *Cell* **144**, 864–873 (2011).
77 Zhao, J., et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40**, 939–953 (2010).
78 Syvanen, A. C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet* **2**, 930–942 (2001).