OPEN ACCESS Freely available online

PLoS one

Human Cancer Long Non-Coding RNA Transcriptomes

Ewan A. Gibb[1,2]*, Emily A. Vucic[1,2], Katey S. S. Enfield[1,3], Greg L. Stewart[1,3], Kim M. Lonergan[1], Jennifer Y. Kennett[1,2], Daiana D. Becker-Santos[1,3], Calum E. MacAulay[1,2,3], Stephen Lam[1], Carolyn J. Brown[1,4], Wan L. Lam[1,2,3]

1 British Columbia Cancer Agency Research Centre, Vancouver, British Columbia, Canada, **2** Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada, **3** Interdisciplinary Oncology Program, University of British Columbia, Vancouver, British Columbia, Canada, **4** Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada

Abstract

Once thought to be a part of the 'dark matter' of the genome, long non-coding RNAs (lncRNAs) are emerging as an integral functional component of the mammalian transcriptome. LncRNAs are a novel class of mRNA-like transcripts which, despite no known protein-coding potential, demonstrate a wide range of structural and functional roles in cellular biology. However, the magnitude of the contribution of lncRNA expression to normal human tissues and cancers has not been investigated in a comprehensive manner. In this study, we compiled 272 human serial analysis of gene expression (SAGE) libraries to delineate lncRNA transcription patterns across a broad spectrum of normal human tissues and cancers. Using a novel lncRNA discovery pipeline we parsed over 24 million SAGE tags and report lncRNA expression profiles across a panel of 26 different normal human tissues and 19 human cancers. Our findings show extensive, tissue-specific lncRNA expression in normal tissues and highly aberrant lncRNA expression in human cancers. Here, we present a first generation atlas for lncRNA profiling in cancer.

Citation: Gibb EA, Vucic EA, Enfield KSS, Stewart GL, Lonergan KM, et al. (2011) Human Cancer Long Non-Coding RNA Transcriptomes. PLoS ONE 6(10): e25915. doi:10.1371/journal.pone.0025915

Editor: Eric J. Bernhard, National Cancer Institute, United States of America

Received August 1, 2011; **Accepted** September 13, 2011; **Published** October 3, 2011

Funding: This work was supported by grants from the Canadian Institutes for Health Research (CIHR) [MOP 86731, MOP 77903 to W.L.L., MOP 13690 to C.J.B.]; National Institutes of Health [NIH 2R01 CA103830 – 6A1]; Department of Defense [CDMRP W81XWH-10-1-0634]; CIHR and Michael Smith Foundation for Health Research (MSFHR) Postdoctoral Fellowships [to E.A.G.]; and CIHR Frederick Banting and Charles Best Canada Graduate Scholarship [to E.A.V.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: egibb@bccrc.ca

Introduction

Genome instability and mutation are a hallmark of cancer [1]. Genetic and epigenetic changes result in aberrant expression of protein-coding genes and many classes of non-coding RNAs (ncRNAs), including microRNAs (miRNAs). MiRNAs have proven to be major players in human carcinogenesis, despite comprising only a small fraction of ncRNAs [2].

Once thought to be the 'dark matter' of the genome, ncRNAs have emerged as an integral component of the mammalian transcriptome [3,4,5]. These enigmatic molecules are defined by lack of protein-coding sequence, yet can play both structural and functional roles in the cell [6,7]. NcRNAs can been grouped into two major classes, the small ncRNAs, which include miRNAs and other non-coding transcripts of less than 200 nucleotides (nt), and the more recently described lncRNAs, which range from 200 nt to >100 kilobases (kb) [8].

LncRNAs can be intergenic, intronic, antisense or overlapping with protein-coding genes or other ncRNAs [9,10,11,12]. The known repertoire of lncRNA functions is rapidly expanding – with demonstrated roles as mediators of mRNA decay [13], structural scaffolds for nuclear substructures [14,15], as host genes for miRNAs [16,17], and as regulators of chromatin remodeling [18,19,20,21] – even though the functional identities of many lncRNAs have yet to be uncovered [6,7,22]. Recently, human cancers have been described to have altered expression of satellite repeats [23], transcribed ultra conserved regions (T-UCRs) [24], and antisense transcripts [25]. Beyond expression changes, accumulating evidence indicates aberrant expression of lncRNAs may play an important functional role in cancer biology [26,27,28]. The well-studied HOX antisense intergenic RNA (*HOTAIR*), for example, is highly expressed in breast cancers and breast cancer metastases and plays a role in retargeting chromatin remodeling complexes [29]. Similarly, high expression of the nuclear speckle associated lncRNA metastasis-associated lung adenocarcinoma transcript 1 (*MALAT1*) modulates alternative splicing and has been associated with metastasis and poor outcome in patients with lung cancer [30,31]. While these examples are intriguing, the extent of the contribution of differential lncRNA expression to human cancer is currently unknown.

With a conservative estimate of 23,000 lncRNAs in the human genome, these transcripts rival the ~20,000 protein-coding genes [5,11,32,33]. Over the past two decades, microarray profiling has generated a wealth of information on protein-coding gene expression patterns in human cancers. However, as lncRNA specific probes are underrepresented on commercial microarrays used in cancer transcriptome profiling, these data do not apply to ncRNAs. Global sequencing of RNA populations is a new approach used to profile RNA expression levels that will capture the extent of lncRNA expression. Recently, genome-wide ncRNA expression profiles were determined in 11 samples representing different types of human tissues [34].

One sequence-based method for enumerating the abundance of polyadenylated transcripts is SAGE [35]. As many lncRNAs

themselves are polyadenylated, lncRNA transcript levels can be deduced by way of direct enumeration of corresponding sequence tags using SAGE technology. In fact, two antisense lncRNAs were discovered using a SAGE-based method [25]. Since the invention of SAGE technology in the mid 1990s, numerous SAGE libraries representing a diversity of human and mouse, normal and malignant tissues and cell lines have become publically available [36]. Of the 755 human SAGE libraries in the Gene Expression Omnibus (GEO) database, ~276 include SAGE libraries derived from human cancers or dysplasias [37].

In this study, we compiled 272 human SAGE libraries to delineate lncRNA transcription patterns across a broad spectrum of human tissues and cancers. Using a custom lncRNA discovery pipeline, we parsed over 24 million SAGE sequence tags to deduce (1) the specific lncRNA expression patterns in 26 human tissues and discovered ubiquitously expressed as well as tissue specific lncRNAs, and (2) the aberrant expression patterns of lncRNAs in 19 human cancers.

Results

Assembling human SAGE libraries of normal and cancer tissues

A total of 1,824 SAGE libraries (in short SAGE, long SAGE and SAGE-seq format) of human and non-human origins are publically available via GEO. To explore lncRNA expression in the broadest range of human tissue types and cancer types, we downloaded 360 GEO accessioned human short SAGE libraries comprised of libraries curated by the Cancer Genome Anatomy Project (324 libraries) and lung tissue and cancer datasets (36 libraries) (Table S1). Individual libraries were filtered for sequence depth, retaining only those libraries with >50,000 raw tags, to provide 272 SAGE libraries for analysis using our lncRNA discovery pipeline (Table S2). The 272 SAGE libraries are comprised of a total of 24,436,076 raw sequence tags with an average raw tag count of 90,212 per library. Collectively, the libraries spanned 26 normal human tissue types, including 19 human cancer types, and 9 tissue types derived from cell line libraries (Figure 1, Table S3).

Long non-coding RNA discovery pipeline

To generate lncRNA expression profiles, we developed a lncRNA discovery pipeline to map tag-to-lncRNA matches (Figure 2). A SAGE tag expression matrix was constructed from all unique tags (n = 716,330) identified within the dataset of 272 libraries. Unigene mapped and unmapped SAGE tags (n = 269,785 and n = 446,545, respectively) were separated into distinct expression matrices which were subsequently filtered to retain only those tags with at least 2 raw tag counts in 3 or more SAGE libraries. Using SAGE Genie to assign gene identifiers to the Unigene IDs, 263 of the 61,054 filtered tags with corresponding Unigene IDs mapped to known lncRNAs, and 15,773 tags either lacked gene names or had ambiguous annotations (e.g. transcribed loci, cDNAs, hypothetical genes). Based on the absence of confirmed association with known genes, these 15,773 tag-to-Unigene ID matches were considered as candidate lncRNA tags.

The 15,773 Unigene tags with ambiguous gene identifiers were combined with the 17,816 unmapped, filtered tags for a total of 33,589 SAGE tags with the potential to generate tag-to-lncRNA matches. Using SeqMap, we mapped 7,040 of the 33,589 tags to lncRNA sequences from the reference lncRNA list (Table S4). The proportion of tag-to-lncRNA matches is consistent with the fact that our reference list of 9,891 lncRNAs represents only a portion

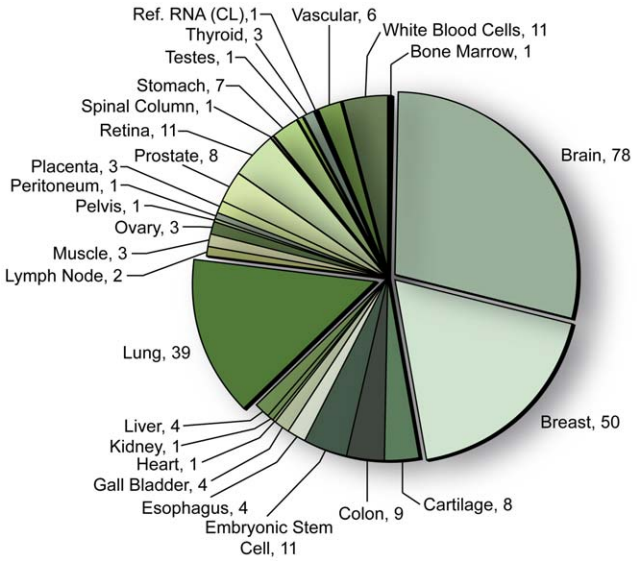

Figure 1. Tissue-type distribution of the 272 SAGE libraries with a minimum raw tag count of 50,000. (CL) indicates one SAGE library that was generated from a mixture of human cell lines.
doi:10.1371/journal.pone.0025915.g001

of the estimated 23,000 lncRNAs in the genome [33]. The remaining tags that do not map to lncRNAs from our reference list may represent antisense transcripts to protein-coding genes or other ncRNAs which were filtered.

Of the 7,040 lncRNA tag matches, 3,831 mapped in the forward orientation, while 3,209 mapped in the reverse direction. In SAGE, tags matching transcript in the forward orientation are likely derived from that transcript, while tags matching in the reverse orientation are not. This is true regardless of whether the gene is normally transcribed from the plus or minus DNA strand. In this study, we were interested in the expression profiles of a curated set of lncRNAs, rather than novel gene discovery. As reverse tag matches do not corroborate the expression of the lncRNAs described herein, these tags were excluded from further analysis.

The 3,831 tags newly mapped to lncRNAs were combined with the 263 tags identified from Unigene mapping for a total of 4,094 tags uniquely mapping to lncRNAs. Where multiple tags mapped to a distinct lncRNA, the tags were collapsed by summing the tag counts to capture all transcript variants and isoforms. The end result was a lncRNA expression matrix consisting of 2,649 distinct lncRNAs (Tables S5 and S6). The lncRNAs with the highest expression were detectable in the majority (>90%) of the 272 libraries (Table 1). These included characterized examples such as nuclear paraspeckle assembly transcript 1 (*NEAT1*) and growth arrest-specific 5 (*GAS5*).

Long non-coding RNA expression profiles in normal human tissues

Of the 272 SAGE libraries, 72 represented normal human tissues. Expression of lncRNAs was detected in all tissue types, although the number of unique lncRNAs detected varied considerably (Figure 3A). On average, there were 145 distinct lncRNAs with a mean tags per million (TPM) of 20 detected in each tissue. Tissues such as lymph node and gall bladder showed the highest number of distinct lncRNAs, while the lowest numbers of distinct lncRNAs were found in the muscle and liver.

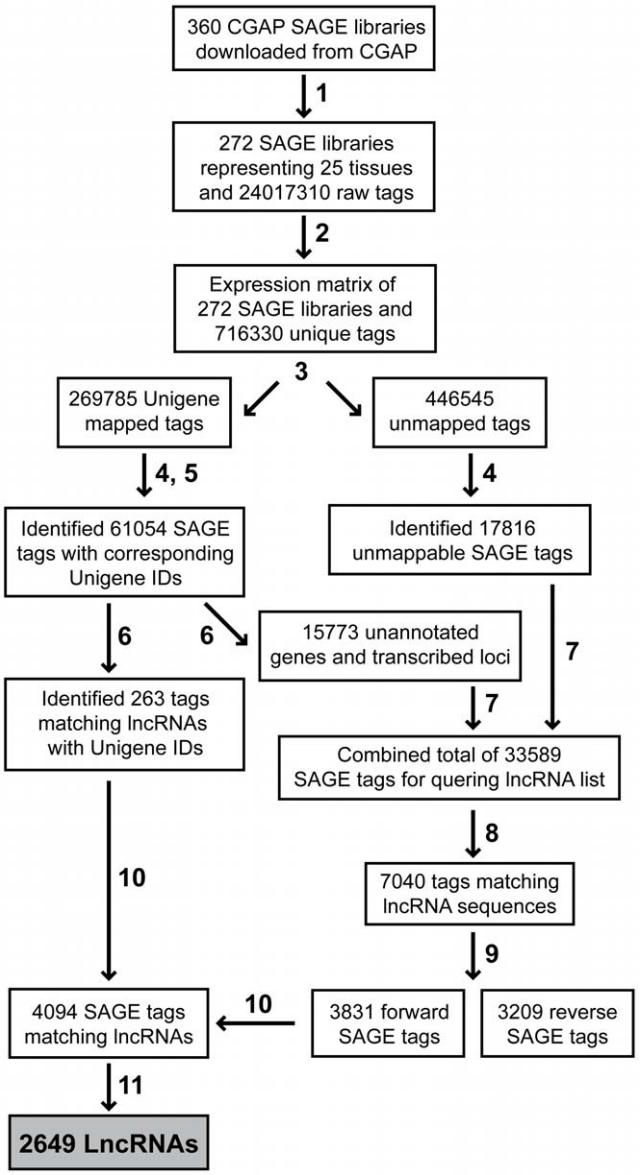

Figure 2. LncRNA discovery pipeline using SAGE analysis. Numbers indicate programs or filtering steps as follows: (1) filtering to retain only those libraries with a minimum of 50,000 raw tag counts, (2) identifying unique SAGE tags and constructing SAGE tag expression matrix, (3) mapping SAGE tags to Unigene IDs using SAGE Genie mapping files, (4) filtering lists to retain only tags with $\geq$2 raw counts in a $\geq$3 of 272 libraries, (5) determining gene identity using SAGE Genie, (6) separating Unigene tags mapping to lncRNAs and ambiguous transcripts, (7) pooling ambiguous tags and unmapped tags, (8) mapping sequence tags to the reference list of 9,891 lncRNAs using SeqMap, a tag-to-gene mapping program, (remaining tags may map to unannotated lncRNAs or antisense transcripts not included in our reference list) (9) filtering tag matches for strand sense, (10) pooling forward mapping tags and tags determined from Unigene, and (11) confirming tag-to-lncRNA matches and summing tag counts for lncRNAs with multiple tag matches. A complete list of lncRNAs is provided as Table S5 and tag-to-lncRNA matches are provided as Table S6.
doi:10.1371/journal.pone.0025915.g002

We next focused on these libraries to determine whether tissue-specific lncRNA expression profiles could be generated (Table S7). Figure 4A shows the top 20 most highly expressed lncRNAs

Table 1. The eleven most highly expressed lncRNAs detected in >90% of the 272 SAGE libraries.

Gene Name	Ensembl Gene	Chr	Start (bp)	End (bp)	Strand
MALAT1	ENSG00000251562	11	65265233	65273940	1
GAS5	ENSG00000234741	1	173833038	173838020	−1
NEAT1	ENSG00000245532	11	65190245	65213011	1
NCRNA00188	ENSG00000175061	17	16342289	16367300	1
RP11-425M5.7	ENSG00000225759	20	36247700	36251521	−1
SNHG6	ENSG00000245910	8	67833919	67838633	−1
SNHG5	ENSG00000203875	6	86386725	86388451	−1
SCAND2	ENSG00000176700	15	85174682	85185695	1
AC104759.1	ENSG00000246638	15	31685046	31696932	1
AC002472.9	ENSG00000230513	22	21356175	21364631	1
AC090937.2	ENSG00000225733	3	14961854	14989931	−1

Also see Table S5.
doi:10.1371/journal.pone.0025915.t001

detected in the panel of normal tissues. Distinct lncRNAs detected at high expression levels in normal tissues included those characterized in the literature such as *NEAT1*, *GAS5* and X-inactive-specific transcript (*XIST*). However, at least half of the highly expressed lncRNAs are novel and currently uncharacterized. To confirm the lncRNA expression profiles, we queried the expression patterns of the most highly expressed lncRNAs using RNASeq data from the Illumina Human BodyMap 2.0 project. This data was recently added to Ensembl release 62 and is presented as an optional track. Of our most highly expressed lncRNAs, the majority were widely expressed in the tissue samples from the Illumina dataset, consistent with our findings (Table S8, Figures S1 and S2). Concurrently, lncRNA expression was also found to be highly variable, with each human tissue having a unique lncRNA expression pattern (Figure 4B). Intriguingly, a number of lncRNAs were expressed in a tissue-exclusive manner (Figure 3B).

Long non-coding RNA expression profiles in human cancers

Aberrant protein-coding gene expression is well described in cancer. However, aberrant expression of ncRNAs, including miRNAs and lncRNAs, has only recently been associated with this disease [2,26,27,38]. To delineate lncRNA expression profiles associated with human cancers, we created a human cancer expression matrix based on 167 cancer SAGE libraries included in our dataset (Table S9). For the lung cancer dataset, metaplasia, dysplasia and inflammatory tissues were excluded from analysis as these represent precancerous stages [39,40]. Figure 5A shows the top 20 most highly expressed lncRNAs across the profiled cancers. Like the normal tissues, lncRNA expression in human cancer was also found to be highly variable (Figure 5B).

Human cancers demonstrate significantly altered lncRNA expression patterns

To determine the extent of differential lncRNA expression in human cancer, we created three expression matrices for each breast, brain and lung cancer which included a minimum of five normal and five cancer SAGE libraries (Table S10). The breast, brain and lung lncRNA expression matrices were independently sorted for significant and differentially expressed lncRNAs (p-

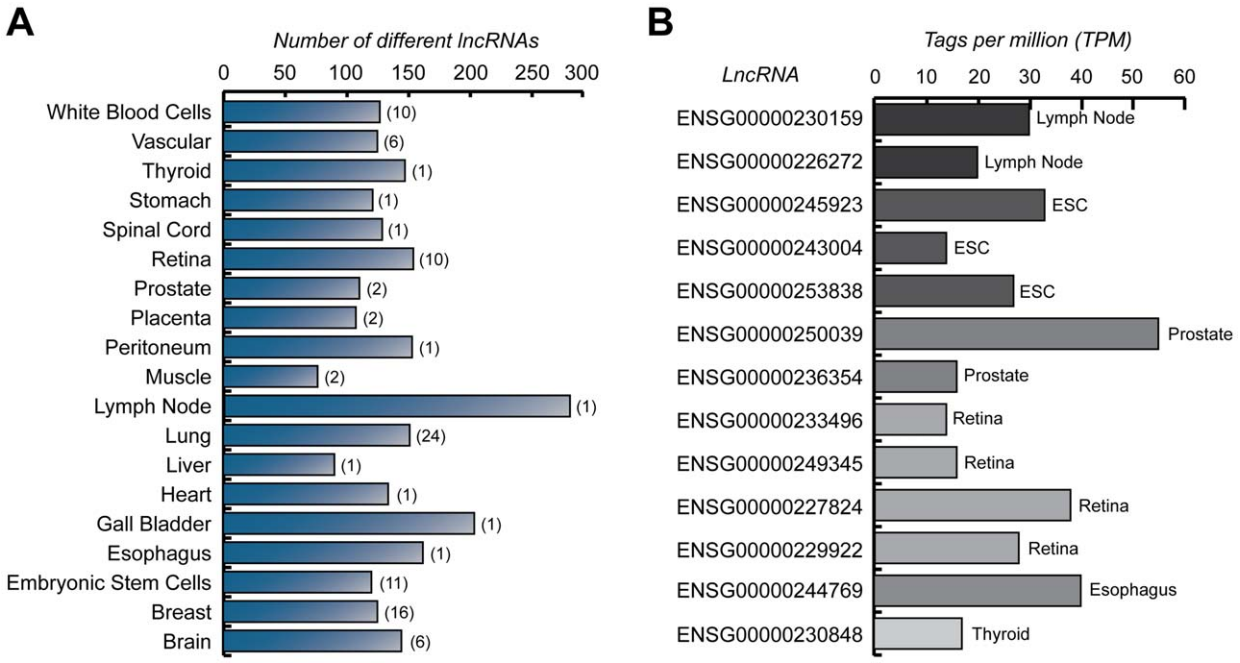

Figure 3. Distribution and levels of lncRNA expression in normal human tissues. (A) Number of distinct lncRNAs expressed in normal human tissues, white blood cells and embryonic stem cells with a minimum average TPM of 20. The values in brackets indicate the number of SAGE libraries for each tissue. (B) Examples of lncRNAs detected exclusively in a single normal human tissue or in embryonic stem cells (ESC) with a minimum expression level of 10 TPM. For tissues with two or more libraries, the TPM values were averaged. LncRNAs without names are labeled with an Ensembl ID.
doi:10.1371/journal.pone.0025915.g003

value$<$0.05, $\geq$2-fold expression change based on a non-parametric permutation test [41]). In each type of cancer, we found at least 200 lncRNAs to have significant differential expression based on these criteria (Figure 6A). Intriguingly, there was overlap between the lncRNAs that were differentially expressed in each tissue (Figure 6B), including 8 lncRNAs that were differentially expressed in all three cancers (Table 2). The ten most up- and down-regulated lncRNAs for each cancer are found in Table S11.

Chromosomal distribution of long non-coding RNAs

We constructed a distribution plot to determine the chromosomal distribution of the 9,891 lncRNA genes in our lncRNA reference list (Table S3). The lncRNAs are distributed throughout the genome and are present on every chromosome (Figure 7). Protein-coding genes and miRNAs appear to share a similar chromosome distribution (Spearman correlation $p>0.05$, Figure S3A). However, the chromosome distribution of lncRNAs did not correlate with either protein-coding genes or miRNAs (Spearman correlation $p<0.05$, Figures S3B, S3C).

Discussion

In recent years, the concept of the functional genome has been re-written to include a multitude of newly discovered classes of ncRNA transcripts [42,43,44,45]. Although the functional significance of long non-coding RNAs has long been recognized [46,47], the abundance and scale of lncRNA expression changes in cancer is just beginning to come to light. For this reason, charting the transcriptional landscape of lncRNAs across human tissue and cancer types is a key step in understanding lncRNA functional significance in cancer.

Here, we present the first multi-tissue, cross-cancer lncRNA expression profiling study. Large-scale expression profiling datasets, such as SAGE, represent a valuable resource for investigating the expression pattern of polyadenylated lncRNAs. While this approach excludes the profiling of non-polyadenylated lncRNAs, it nonetheless facilitates the simultaneous profiling of thousands of polyadenylated lncRNAs in a wide range of human tissues and cancers. Using 272 SAGE libraries, representing 26 non-malignant human tissues, 19 human cancer types and 9 cancer cell lines, we have produced a first generation atlas of cross-cancer lncRNA expression profiles as a resource for this fast growing area of cancer research. Current estimates of the number of lncRNAs encoded in the human genome vary widely, ranging from ~7,000 to 23,000 or more [7]. These estimates rival the abundance of the estimated 20,000+ protein-coding genes. Our analysis showed that lncRNAs are distributed on all 22 autosomes and sex chromosomes, yet the distribution pattern did not correlate with either protein-coding genes or miRNAs (Figure 7, Figure S3).

Examination of 72 SAGE libraries of normal human tissues revealed lncRNA expression in brain, breast, esophagus, gall bladder, heart, liver, lung, lymph node, muscle, peritoneum, placenta, prostate, retina, spinal cord, stomach, thyroid, vascular tissue, embryonic stem cells and white blood cells. We find extensive and highly differential patterns of lncRNA expression in normal human tissues (Figures 3 and 4), corroborating a previous report of tissue-specific ncRNA patterns [34]. For example, the lncRNA NCRNA00116 was highly expressed in the contractile tissues, namely heart (TPM = 349) and muscle (TPM = 399). LncRNAs ENSG00000230658 and ENSG00000235621 showed very high expression (TPM = 888) in placenta and esophagus (TPM = 820) respectively, but low or undetectable expression in other tissues, which may indicate a tissue-specific role for these transcripts. The brain-associated and putative tumor suppressor lncRNA maternally

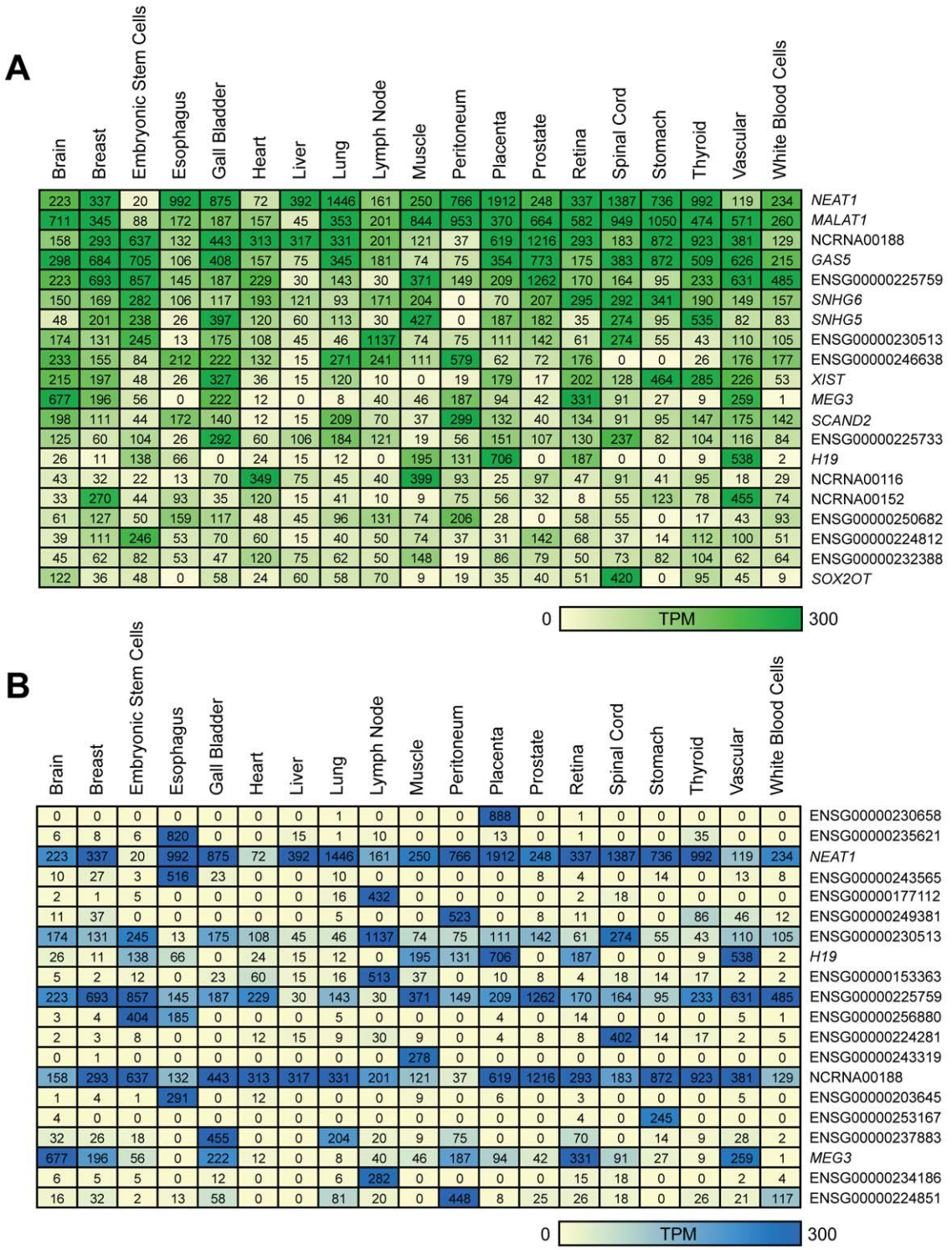

Figure 4. Expression patterns of lncRNAs in normal human tissues. (A) LncRNAs with the highest overall expression (B) LncRNAs with the highest variance by a coefficient of variation (CV) test. Heatmaps indicate the relative intensity (normalized TPM) of each lncRNA across seventeen human tissues, white blood cells and human embryonic stem cells. Where more than one SAGE library was available, the TPM values were averaged. For the heatmap, the maximum threshold was set at 300 TPM. LncRNAs without names are labeled with an Ensembl ID.
doi:10.1371/journal.pone.0025915.g004

expressed 3 (*MEG3*) [48], displayed the highest expression in brain in our dataset (TPM = 677), but showed low level expression in other tissue types (Figure 4). Collectively, these data suggest some lncRNAs may function in a tissue-specific manner.

Only ~1% of the lncRNAs were ubiquitously expressed across all tissues examined. These constantly expressed lncRNAs are reminiscent of the expression patterns of "housekeeping" protein-coding genes [49]. The eleven lncRNAs in Table 1 were expressed

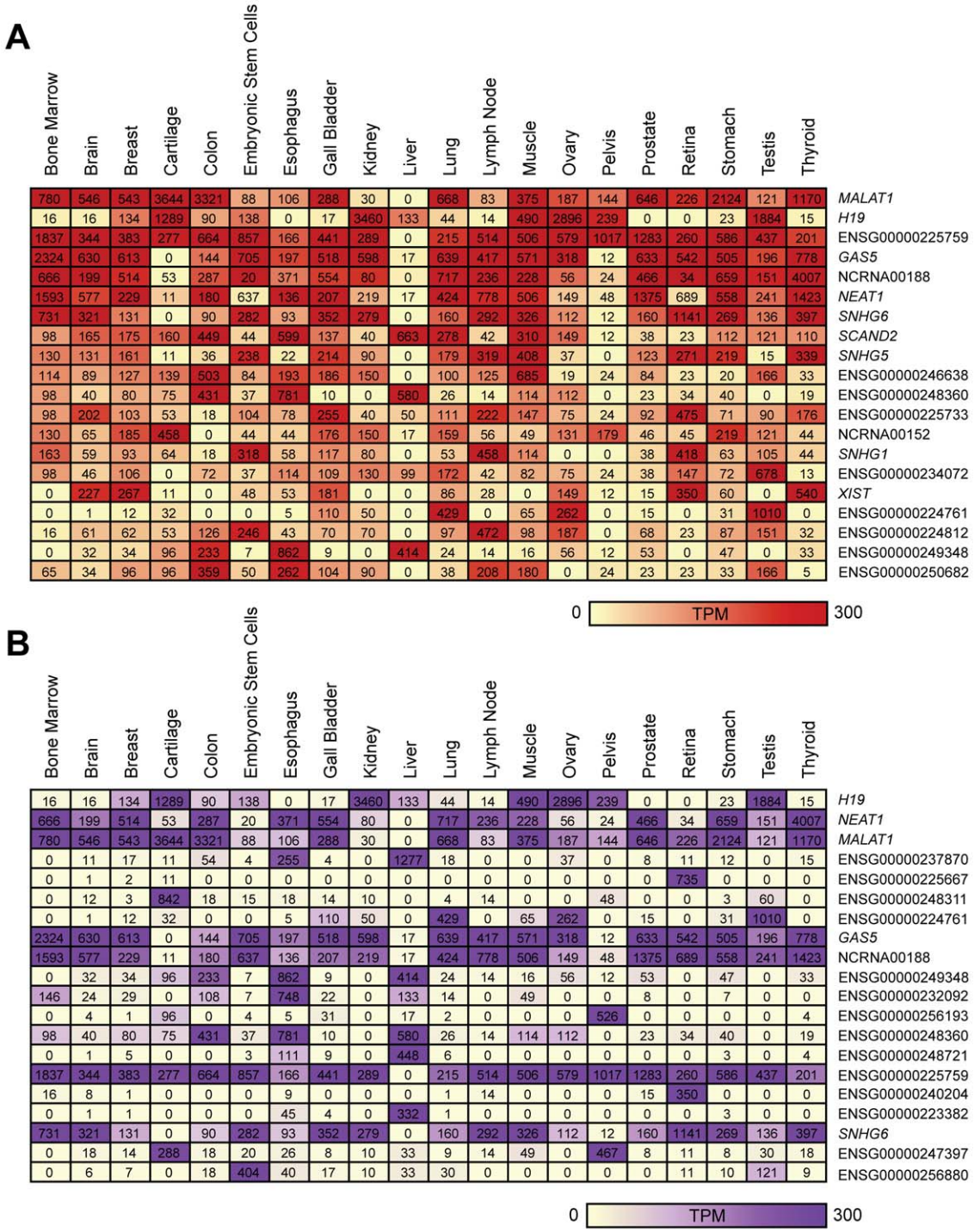

Figure 5. Expression patterns of lncRNAs in human cancers. (A) LncRNAs with the highest overall expression (B) LncRNAs with the highest variance by a coefficient of variation (CV) test. Heatmaps indicate the relative intensity (normalized TPM) of each lncRNA across seventeen human cancers and human embryonic stem cells. Where more than one SAGE library was available, the TPM values were averaged. For the heatmap, the maximum threshold was set at 300 TPM. LncRNAs without names are labeled with an Ensembl ID.
doi:10.1371/journal.pone.0025915.g005

in at least 90% of 272 SAGE libraries in our dataset, implicating that these transcripts may participate in common biological processes. However, the absolute expression level varied for each tissue, sometimes by hundreds of TPM (Figure 4). This suggests certain lncRNAs may be required at different cellular levels in different tissues or under different conditions, much like many constitutively expressed protein-coding genes [50,51,52]. The concept of lncRNAs functioning as constitutively expressed regulators has been previously proposed. For example, the lncRNA *XIST* is critical for female development due to its functional role in X-chromosome inactivation [47,53]. Concordantly, a number of the most highly and frequently expressed

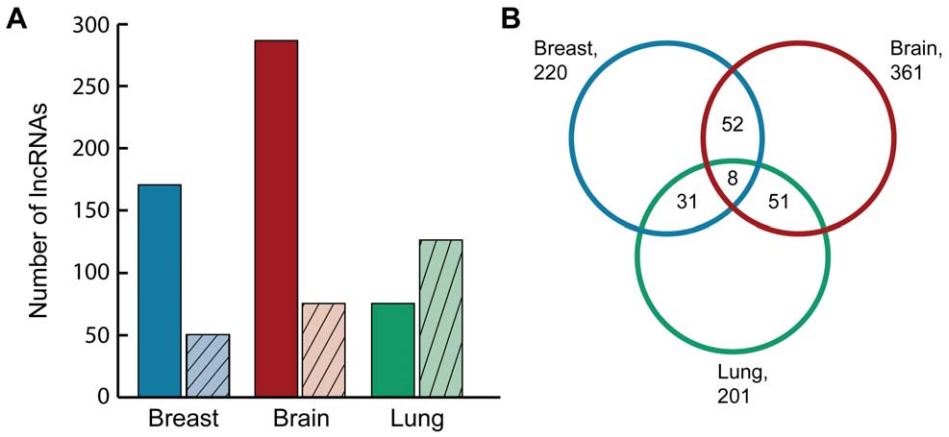

Figure 6. Aberrantly expressed lncRNAs in human cancers. (A) Number of lncRNAs showing significant expression changes. The number of lncRNAs determined to have significant (BH p-value <0.05) differential expression of 2-fold or greater reported. Solid bars indicate upregulated genes, while bars with hatch marks indicate downregulated genes (B) Venn diagram of differentially expressed lncRNAs in human carcinomas.
doi:10.1371/journal.pone.0025915.g006

lncRNAs in our dataset have prior associations with key biological processes, including *NEAT1*, a structural scaffold for paraspeckle formation [14,54], *MALAT1* which regulates alternative splicing [31] and small nucleolar RNA host gene 6 (*SNHG6*) which hosts a snoRNA, which function in RNA modification [55]. These findings suggest that lncRNAs may be critical to normal tissue maintenance and function.

In this cross-cancer type analysis, we found that lncRNAs aberrantly expressed in a specific cancer may also be altered in other cancers. For example, while *MEG3* is highly expressed in normal brain tissues, this lncRNA was strongly decreased in our brain cancer datasets, and strikingly so in gall bladder, retinal and prostate cancers, consistent with the proposed tumor suppressor role for *MEG3* [48,56,57]. In another example, miR155 host gene (*miR155HG*), a lncRNA processed to the miRNA *miR-155*, was highly overexpressed in B-cell lymphoma consistent with previous reports [16], but also was also upregulated in esophageal and gall bladder cancers.

Long non-coding RNAs are also implicated in the regulation of embryogenesis [58,59,60]. Fetal lncRNAs reactivated in cancers may represent critical regulators of pluripotency or cellular growth. For example, the lncRNA urothelial cancer associated 1 (*UCA1*) has demonstrated roles in both embryonic development and is implicated in bladder cancer, supporting this concept [61]. In our datasets, we found several lncRNAs with low expression in normal tissues, but with high expression in both embryonic stem cells and cancer (Table S12). While these reactivated fetal lncRNAs represented mostly uncharacterized examples, *H19*, a well-studied lncRNA with associations in both mammalian development and cancer [53], was also detected in our dataset. Interestingly, *NEAT1*, which is constitutively and highly expressed in normal tissues [34,62], with the exception of embryonic stem cells, was downregulated in lung, liver, esophageal and retinal cancers (retinoblastoma).

Since genomic amplifications and deletions are key mechanisms of gene deregulation in cancer, we investigated changes in lncRNA expression in genomic regions frequently altered in breast, brain and lung cancer. Comparison of the significantly ($p<0.05$) deregulated lncRNAs common between brain, breast and lung cancer tissues revealed eight lncRNAs were differentially regulated (≥2-fold) compared to normal tissue. Intriguingly, three of these lncRNAs - ENSG00000226380, ENSG00000230937 and ENSG00000253288 - were located on 7q32.3, 1q32.2, and 8q24.23, respectively, in regions completely devoid of protein-coding genes. Like protein-coding genes and miRNAs, it is possible that differential lncRNA expression is driven by similar mechanisms of disruption, including copy number gain/loss or aberrant methylation patterns. Indeed, high level amplification of lncRNA containing loci such as cytoband 19p12

Table 2. Aberrantly expressed lncRNAs common to brain, breast and lung cancers.

						Fold Change			Corrected p-value		
lncRNA	Ensembl Gene ID	Chr	Start	End	Strand	Brain	Breast	Lung	Brain	Breast	Lung
AC058791.1	ENSG00000230937	7	130565751	130598069	−1	7.00	−3.00	3.59	0.00122	0.02373	0.00000
CTA-55I10.1	ENSG00000255717	1	209602165	209606183	1	3.37	−2.05	2.72	0.00041	0.00190	0.00000
NCRNA00263	ENSG00000247556	10	102133372	102143125	1	12.37	2.10	2.46	0.00004	0.00056	0.00000
AC080037.2	ENSG00000245411	17	70594180	70636611	−1	3.08	2.76	−2.14	0.00009	0.00027	0.00141
AC012652.1	ENSG00000226380	15	41576203	41601901	1	6.45	3.53	−2.33	0.00026	0.00018	0.03657
RP11-18C24.6	ENSG00000253288	12	120928131	120933743	−1	−2.48	−3.13	−2.86	0.00405	0.00639	0.01311
RP11-238K6.1	ENSG00000235823	8	138821687	139095813	−1	7.07	4.18	−4.35	0.00529	0.00012	0.00037
SNHG1	ENSG00000248008	11	62619460	62623386	−1	3.04	3.27	−5.03	0.00403	0.00043	0.00003

doi:10.1371/journal.pone.0025915.t002

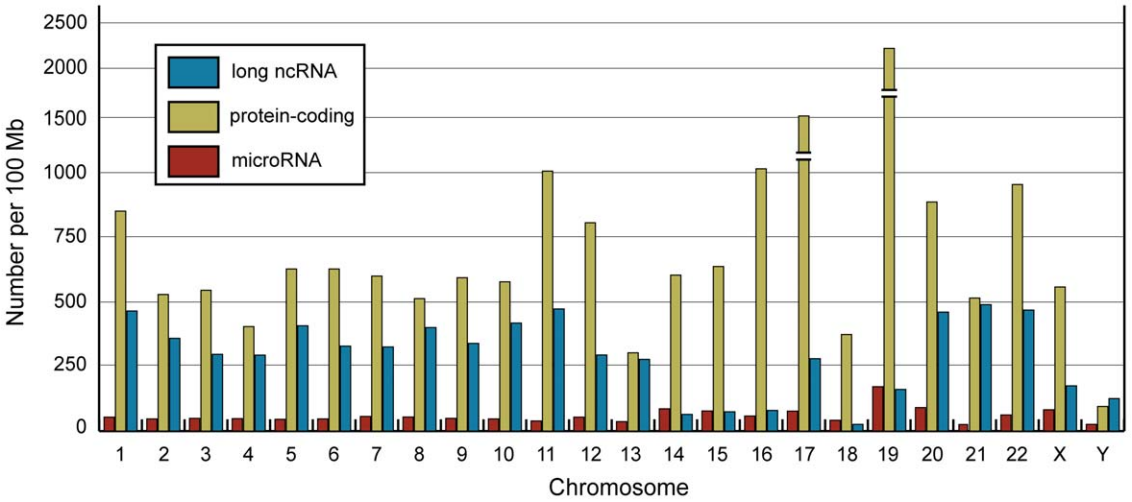

Figure 7. Chromosomal distribution of protein-coding genes, microRNAs and long non-coding RNAs in the human genome. Protein-coding gene (n = 20,655), microRNA (n = 1,746) and long non-coding RNA (n = 9,891) coordinates were downloaded from Ensembl v62 using BioMart.
doi:10.1371/journal.pone.0025915.g007

has been reported in breast cancer [63], while high level amplification of 12p13.2 (which contains a number of lncRNA loci) has been reported in breast cancer, glioblastoma, astrocytoma, and squamous cell lung cancer [64,65,66,67]. Likewise, aberrant expression of a number of lncRNAs has been tied to altered methylation patterns [68,69]. However, the mechanism(s) driving aberrant lncRNA expression remains mostly unknown.

While lncRNAs have been documented for nearly three decades, the magnitude and diversity of lncRNA expression has only recently been appreciated. It is estimated that lncRNAs in the human genome number into the tens of thousands, effectively doubling the number of potential gene targets in cancer gene expression networks. Large scale, cross-tissue and cancer studies are crucial to understanding the regulation of lncRNA expression and how these novel transcripts integrate with our current understanding of the mammalian transcriptome. Moreover, a deeper understanding of lncRNA expression will not only expand the number of potential target cancer genes, but also facilitate development of novel anti-cancer therapies, such as gene regulation mediated by antisense RNAs [70] or targeting lncRNA-protein interactions [28].

Materials and Methods

SAGE Libraries

This study uses publically available SAGE libraries for data analysis. A total of 360 SAGE libraries, including 324 from the Cancer Genome Anatomy Project (CGAP) SAGE library collection (GSE15309), 19 lung bronchial epithelium libraries (GSE3707), 13 lung cancer libraries (GSE7898) and 4 never smoker bronchial epithelium libraries (GSE5473), were downloaded from GEO (Table S1). Libraries constructed from non-human samples, as well as long SAGE and SAGE-seq libraries were not used in this study. To facilitate direct comparison the SAGE libraries were filtered to retain only those libraries with >50,000 raw tag counts resulting in 272 libraries suitable for analysis (Table S2).

Long non-coding RNA reference list

The lncRNA discovery pipeline is based on a reference list of human lncRNAs curated by the online genomic database Ensembl release 62, built on the Genome Reference Consortium release GRCh37 [71]. The lncRNA reference list was compiled from 1,239 Ensembl (v62) IDs designated as 'lincRNAs' (long intergenic non-coding RNAs, a subclass of lncRNAs) and 8,652 Ensembl IDs (v62) designated as 'processed transcripts' for a total of 9,891 lncRNAs (Table S4). All the lncRNAs used to query the SAGE libraries were Ensembl curated transcripts without a predicted open reading frame. The sequences of all lncRNA transcripts were retrieved from Ensembl (v62) using the Biomart data management system.

SAGE tag-to-gene mapping

Custom Perl scripts were used to create an expression matrix of the unique SAGE tags across the 272 libraries (Perl scripts: getuniquetags.pl and makeTable_April20.pl). The SAGE tags were mapped to Unigene IDs using custom Perl scripts and a short SAGE mapping file (mapping file: Hs_short) downloaded from SAGE Genie (http://cgap.nci.nih.gov/SAGE), to create a matrix of Unigene ID mapped tags and a matrix of unmapped tags (Perl script: extractUnmappedTags_Unigene). The two expression matrices of unmapped tags and Unigene mapped tags were independently filtered to retain only tags with raw tag counts of 2 or more, appearing in at least 3 SAGE libraries.

For the Unigene mapped tags, gene identifiers were assigned to Unigene IDs using SAGE Genie. From this dataset, tags matching known or candidate lncRNAs were extracted manually. Candidate lncRNAs are Unigene IDs with no gene name or matching one or more of the following descriptors: 'non-coding', 'non-protein', 'cDNA', 'transcribed locus', 'clone IMAGE', 'chr(#)orf(#)', 'hypothetical', 'family with sequence similarity', 'FLJ(#)', or 'KIAA(#)'. The candidate lncRNA tags were merged with the unmapped tags and used as a single dataset from which to identify sequence matches to the lncRNA reference list.

The tag-to-gene mapping program SeqMap was used to identify perfect (0 mismatches) tag matches to the transcript sequences from the reference lncRNA list. Tags mapping to lncRNAs were filtered to retain those corresponding to the forward ('sense') strand, while reverse tag matches do not corroborate the expression of the candidate lncRNAs and were not analyzed further. The forward strand tags that mapped to lncRNAs were then combined with the Unigene tags that mapped to lncRNAs to

create an expression matrix of SAGE tags mapping to lncRNAs. This matrix was remapped to the lncRNA reference list to confirm accurate tag-to-lncRNA matches.

Data pre-processing

In cases where multiple tags mapped to the same lncRNA, the tags were compressed by summing the tag counts to capture all lncRNA transcript variants and isoforms (Perl script: sumRows.pl). SAGE tags mapping to more than one lncRNA were discarded. Raw tag counts for each SAGE library were normalized to TPM to facilitate adequate comparison among libraries. Additional expression matrices included only SAGE libraries of interest for a given analysis, while removing any columns with unwanted SAGE libraries. These submatrices were filtered to remove lncRNAs with undetected expression. When a tissue or cancer was represented by more than one SAGE library, the normalized TPM were averaged. Finally, all Ensembl v62 IDs were lifted to Ensembl v63, any missing or reassigned IDs were removed from the final lncRNA list.

Statistical analysis

To ensure statistical significance when comparing normal tissues with cancerous tissues, the lncRNA expression matrix was filtered to retain only those tissues represented by a minimum of 5 normal and 5 cancer SAGE libraries. These SAGE libraries were used to derive cancer specific expression matrices. To compare the expression of lncRNAs between normal libraries and cancer libraries, we performed a normalization of expression by permutation of SAGE (*NEPS*) test as described [41]. LncRNAs with permutation scores of >0.05 were considered to be statistically significant. All fold changes were calculated by dividing the average expression of the cancer SAGE libraries by the average expression of the normal SAGE libraries. Variance calculations were performed by calculating the coefficient of variation (CV) across the averaged normal or cancer SAGE libraries. The lncRNA distribution plots were created by normalizing the number of lncRNAs, miRNAs, or protein-coding genes to 100 megabase (MB) of chromosome and then performing a Spearman correlation.

Supporting Information

Figure S1 Tissue expression profiles of *MALAT1*. Expression was derived from the Human BodyMap 2.0 RNASeq track in Ensembl v62.
(JPG)

Figure S2 Tissue expression profiles of NCRNA00188. Expression was derived from the Human BodyMap 2.0 RNASeq track in Ensembl v62.
(JPG)

Figure S3 Correlation of chromosome distribution between protein-coding genes, miRNAs and lncRNAs. (A) Protein-coding genes compared to miRNAs, (B) Protein-coding genes compared to lncRNAs, (C) lncRNAs compared to miRNAs. The chromosome locations of protein-coding genes (n = 20,655), microRNAs (n = 1746) and long non-coding RNAs (n = 9,891) were downloaded from Ensembl v62. The graphs were generated using GraphPad Prism.
(JPG)

Table S1 GEO Libraries.
(XLSX)

Table S2 Filtered SAGE libraries.
(XLSX)

Table S3 SAGE library information.
(DOC)

Table S4 LncRNA reference list.
(XLSX)

Table S5 LncRNA expression matrix.
(XLSX)

Table S6 Tag-to-lncRNA matches.
(XLSX)

Table S7 Normal tissue lncRNA expression matrix.
(XLSX)

Table S8 Expression validation by BodyMap RNASeq.
(XLSX)

Table S9 Cancer tissue lncRNA expression matrix.
(XLSX)

Table S10 Brain, breast and lung libraries.
(XLS)

Table S11 Top differentially expressed brain, breast and lung lncRNAs.
(XLSX)

Table S12 ESC and cancers.
(XLSX)

Author Contributions

Conceived and designed the experiments: EAG EAV KML CEM SL CJB WLL. Performed the experiments: EAG EAV KSSE GLS JYK. Analyzed the data: EAG EAV KSSE GLS JYK KML DDBS. Contributed reagents/materials/analysis tools: EAG EAV KSSE GLS JYK KML. Wrote the paper: EAG EAV KSSE CEM SL CJB WLL.

References

1. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144: 646–674.
2. Iorio MV, Croce CM (2009) MicroRNAs in cancer: small molecules with a huge impact. J Clin Oncol 27: 5848–5856.
3. Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds CP, et al. (2010) The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. BMC Biol 8: 149.
4. Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL (2011) Genome-wide characterization of non-polyadenylated RNAs. Genome Biol 12: R16.
5. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.
6. Mattick JS (2009) The genetic signatures of noncoding RNAs. PLoS Genet 5: e1000459.
7. Lipovich L, Johnson R, Lin CY (2010) MacroRNA underdogs in a microRNA world: evolutionary, regulatory, and biomedical significance of mammalian long non-protein-coding RNA. Biochim Biophys Acta 1799: 597–615.
8. Costa FF (2010) Non-coding RNAs: Meet thy masters. Bioessays 32: 599–608.
9. Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458: 223–227.
10. Rearick D, Prakash A, McSweeny A, Shepard SS, Fedorova L, et al. (2011) Critical association of ncRNA with introns. Nucleic Acids Res 39: 2357–2366.
11. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559–1563.
12. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, et al. (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res 15: 987–997.
13. Gong C, Maquat LE (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. Nature 470: 284–288.
14. Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, et al. (2009) An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol Cell 33: 717–726.

15. Shevtsov SP, Dundr M (2011) Nucleation of nuclear bodies by RNA. Nat Cell Biol 13: 167–173.
16. Eis PS, Tam W, Sun L, Chadburn A, Li Z, et al. (2005) Accumulation of miR-155 and BIC RNA in human B cell lymphomas. Proc Natl Acad Sci U S A 102: 3627–3632.
17. Mestdagh P, Bostrom AK, Impens F, Fredlund E, Van Peer G, et al. (2010) The miR-17-92 microRNA cluster regulates multiple components of the TGF-beta pathway in neuroblastoma. Mol Cell 40: 762–773.
18. Kanduri C (2011) Kcnq1ot1: A chromatin regulatory RNA. Semin Cell Dev Biol.
19. Kotake Y, Nakagawa T, Kitagawa K, Suzuki S, Liu N, et al. (2011) Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. Oncogene 30: 1956–1962.
20. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, et al. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. Science 329: 689–693.
21. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, et al. (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 142: 409–419.
22. Nagano T, Fraser P (2011) No-Nonsense Functions for Long Noncoding RNAs. Cell 145: 178–181.
23. Ting DT, Lipson D, Paul S, Brannigan BW, Akhavanfard S, et al. (2011) Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers. Science 331: 593–596.
24. Mestdagh P, Fredlund E, Pattyn F, Rihani A, Van Maerken T, et al. (2010) An integrative genomics screen uncovers ncRNA T-UCR functions in neuroblastoma tumours. Oncogene 29: 3583–3592.
25. Maruyama R, Shipitsin M, Choudhury S, Wu Z, Protopopov A, et al. (2010) Breast Cancer Special Feature: Altered antisense-to-sense transcript ratios in breast cancer. Proc Natl Acad Sci U S A.
26. Gibb EA, Brown CJ, Lam WL (2011) The functional role of long non-coding RNA in human carcinomas. Mol Cancer 10: 38.
27. Huarte M, Rinn JL (2010) Large non-coding RNAs: missing links in cancer? Hum Mol Genet 19: R152–161.
28. Tsai MC, Spitale RC, Chang HY (2011) Long intergenic noncoding RNAs: new links in cancer progression. Cancer Res 71: 3–7.
29. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, et al. (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464: 1071–1076.
30. Ji P, Diederichs S, Wang W, Boing S, Metzger R, et al. (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 22: 8031–8041.
31. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, et al. (2010) The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell 39: 925–938.
32. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, et al. (2005) Antisense transcription in the mammalian transcriptome. Science 309: 1564–1566.
33. Carninci P, Hayashizaki Y (2007) Noncoding RNA transcription beyond annotated genes. Curr Opin Genet Dev 17: 139–144.
34. Castle JC, Armour CD, Lower M, Haynor D, Biery M, et al. (2010) Digital genome-wide ncRNA expression, including SnoRNAs, across 11 human tissues using polyA-neutral amplification. PLoS One 5: e11779.
35. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484–487.
36. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al. (2011) NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic Acids Res 39: D1005–1010.
37. Strausberg RL, Buetow KH, Emmert-Buck MR, Klausner RD (2000) The cancer genome anatomy project: building an annotated gene index. Trends Genet 16: 103–106.
38. Farazi TA, Spitzer JI, Morozov P, Tuschl T (2011) miRNAs in human cancer. J Pathol 223: 102–115.
39. Kerr KM (2001) Pulmonary preinvasive neoplasia. J Clin Pathol 54: 257–271.
40. Wistuba II, Gazdar AF (2006) Lung cancer preneoplasia. Annu Rev Pathol 1: 331–348.
41. Chari R, Lonergan KM, Pikor LA, Coe BP, Zhu CQ, et al. (2010) A sequence-based approach to identify reference genes for gene expression analysis. BMC Med Genomics 3: 32.
42. Mattick JS (2011) Genome-sequencing anniversary. The genomic foundation is shifting. Science 331: 874.
43. Brosnan CA, Voinnet O (2009) The long and the short of noncoding RNAs. Curr Opin Cell Biol 21: 416–425.
44. Mattick JS (2001) Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep 2: 986–991.
45. Costa FF (2005) Non-coding RNAs: new players in eukaryotic biology. Gene 357: 83–94.
46. Brannan CI, Dees EC, Ingram RS, Tilghman SM (1990) The product of the H19 gene may function as an RNA. Mol Cell Biol 10: 28–36.
47. Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, et al. (1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature 349: 38–44.
48. Miyoshi N, Wagatsuma H, Wakana S, Shiroishi T, Nomura M, et al. (2000) Identification of an imprinted gene, Meg3/Gtl2 and its human homologue MEG3, first mapped on mouse distal chromosome 12 and human chromosome 14q. Genes Cells 5: 211–220.
49. Eisenberg E, Levanon EY (2003) Human housekeeping genes are compact. Trends Genet 19: 362–365.
50. Rubie C, Kempf K, Hans J, Su T, Tilton B, et al. (2005) Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. Mol Cell Probes 19: 101–109.
51. Steele BK, Meyers C, Ozbun MA (2002) Variable expression of some "housekeeping" genes during human keratinocyte differentiation. Anal Biochem 307: 341–347.
52. Greer S, Honeywell R, Geletu M, Arulanandam R, Raptis L (2010) Housekeeping genes; expression levels may change with density of cultured cells. J Immunol Methods 355: 76–79.
53. Augui S, Nora EP, Heard E (2011) Regulation of X-chromosome inactivation by the X-inactivation centre. Nat Rev Genet 12: 429–442.
54. Chen LL, Carmichael GG (2009) Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. Mol Cell 35: 467–478.
55. Kiss T (2001) Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. EMBO J 20: 3617–3622.
56. Zhang X, Zhou Y, Mehta KR, Danila DC, Scolavino S, et al. (2003) A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells. J Clin Endocrinol Metab 88: 5119–5126.
57. Benetatos L, Vartholomatos G, Hatzimichael E (2011) MEG3 imprinted gene contribution in tumorigenesis. Int J Cancer.
58. Caley DP, Pink RC, Trujillano D, Carter DR (2010) Long noncoding RNAs, chromatin, and development. Scientific World Journal 10: 90–102.
59. van Leeuwen S, Mikkers H (2010) Long non-coding RNAs: Guardians of development. Differentiation 80: 175–183.
60. Pauli A, Rinn JL, Schier AF (2011) Non-coding RNAs as regulators of embryogenesis. Nat Rev Genet 12: 136–149.
61. Wang F, Li X, Xie X, Zhao L, Chen W (2008) UCA1, a non-protein-coding RNA up-regulated in bladder carcinoma and embryo, influencing cell growth and promoting invasion. FEBS Lett 582: 1919–1927.
62. Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB, et al. (2007) A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. BMC Genomics 8: 39.
63. Yu W, Kanaan Y, Bae YK, Gabrielson E (2009) Chromosomal changes in aggressive breast cancers with basal-like features. Cancer Genet Cytogenet 193: 29–37.
64. Letessier A, Sircoulomb F, Ginestier C, Cervera N, Monville F, et al. (2006) Frequency, prognostic impact, and subtype association of 8p12, 8q24, 11q13, 12p13, 17q12, and 20q13 amplifications in breast cancers. BMC Cancer 6: 245.
65. Buschges R, Weber RG, Actor B, Lichter P, Collins VP, et al. (1999) Amplification and expression of cyclin D genes (CCND1, CCND2 and CCND3) in human malignant gliomas. Brain Pathol 9: 435–442; discussion 432–433.
66. Schiffman JD, Hodgson JG, VandenBerg SR, Flaherty P, Polley MY, et al. (2010) Oncogenic BRAF mutation with CDKN2A inactivation is characteristic of a subset of pediatric malignant astrocytomas. Cancer Res 70: 512–519.
67. Kang JU, Koo SH, Kwon KC, Park JW, Kim JM (2009) Identification of novel candidate target genes, including EPHB3, MASP1 and SST at 3q26.2–q29 in squamous cell carcinoma of the lung. BMC Cancer 9: 237.
68. Gejman R, Batista DL, Zhong Y, Zhou Y, Zhang X, et al. (2008) Selective loss of MEG3 expression and intergenic differentially methylated region hypermethylation in the MEG3/DLK1 locus in human clinically nonfunctioning pituitary adenomas. J Clin Endocrinol Metab 93: 4119–4125.
69. Takeuchi S, Hofmann WK, Tsukasaki K, Takeuchi N, Ikezoe T, et al. (2007) Loss of H19 imprinting in adult T-cell leukaemia/lymphoma. Br J Haematol 137: 380–381.
70. Morris KV (2009) RNA-directed transcriptional gene silencing and activation in human cells. Oligonucleotides 19: 299–306.
71. Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2011) Ensembl 2011. Nucleic Acids Res 39: D800–806.