NEOPLASIA
www.neoplasia.com

Volume 14 Number 11 November 2012 pp. 1075–1086 1075

The Transcriptional Consequences of Somatic Amplifications, Deletions, and Rearrangements in a Human Lung Squamous Cell Carcinoma[1,2]

Lucy F. Stead*, Stefano Berri*, Henry M. Wood*, Philip Egan*, Caroline Conway*, Catherine Daly*, Kostas Papagiannopoulos† and Pamela Rabbitts*

*Leeds Institute of Molecular Medicine, University of Leeds, Leeds, United Kingdom; †Department of Thoracic Surgery, St. James's University Hospital, Leeds, United Kingdom

Abstract

Lung cancer causes more deaths, worldwide, than any other cancer. Several histologic subtypes exist. Currently, there is a dearth of targeted therapies for treating one of the main subtypes: squamous cell carcinoma (SCC). As for many cancers, lung SCC karyotypes are often highly anomalous owing to large somatic structural variants, some of which are seen repeatedly in lung SCC, indicating a potential causal association for genes therein. We chose to characterize a lung SCC genome to unprecedented detail and integrate our findings with the concurrently characterized transcriptome. We aimed to ascertain how somatic structural changes affected gene expression within the cell in ways that could confer a pathogenic phenotype. We sequenced the genomes of a lung SCC cell line (LUDLU-1) and its matched lymphocyte cell line (AGLCL) to more than 50× coverage. We also sequenced the transcriptomes of LUDLU-1 and a normal bronchial epithelium cell line (LIMM-NBE1), resulting in more than 600 million aligned reads per sample, including both coding and non-coding RNA (ncRNA), in a strand-directional manner. We also captured small RNA (<30 bp). We discovered significant, but weak, correlations between copy number and expression for protein-coding genes, antisense transcripts, long intergenic ncRNA, and microRNA (miRNA). We found that miRNA undergo the largest change in overall expression pattern between the normal bronchial epithelium and the tumor cell line. We found evidence of transcription across the novel genomic sequence created from six somatic structural variants. For each part of our integrated analysis, we highlight candidate genes that have undergone the largest expression changes.

Neoplasia (2012) 14, 1075–1086

Introduction

In the United Kingdom, there is a death from lung cancer every 15 minutes. This is due to the high prevalence of the disease, late stage at presentation, and inadequate treatment options. As for other tumor types, molecular genetic analysis has identified specific drug targets that allow more individualized treatments, but in lung cancer, only for the adenocarcinoma histologic subtype, this has been translated into novel therapies [1–5]. The squamous cell carcinoma (SCC) subtype, although reducing in prevalence in developed countries [6], predominates in developing countries such as India and Indonesia and is predicted to be responsible for millions of deaths as the smoking epidemic sweeps through Southeast Asia. Although surgery is the preferred option for early stage disease, many patients present with

Abbreviations: lincRNA, long intergenic non-coding RNA; miRNA, microRNA; ncRNA, non-coding RNA; PI3K, phosphatidylinositol 3-kinase; SCC, squamous cell carcinoma; snoRNA, small nucleolar RNA; RNA; TCGA, The Cancer Genome Atlas

Address all correspondence to: Pamela Rabbitts, PhD, Leeds Institute of Molecular Medicine, University of Leeds, Leeds LS9 7TF, United Kingdom. E-mail: P.Rabbitts@leeds.ac.uk

[1]This work was supported by Yorkshire Cancer Research (grant number L341PG to P.R.), the Leeds Teaching Hospitals Charitable Foundation, and the Betty Woolsey Bequest for Thoracic Research.

[2]This article refers to supplementary materials, which are designated by Tables W1 and W2 and Figures W1 and W2 and are available online at www.neoplasia.com.

Received 24 August 2012; Revised 25 September 2012; Accepted 28 September 2012

 1522-8002/12/$25.00
DOI 10.1593/neo.121380

disseminated disease or with comorbidities that preclude surgery. Thus, the discovery of new drugs is urgent, requiring a search for effective drug targets; following success in other cancers, identification of tumor-specific genomic abnormalities is expected to be a valuable approach. Accordingly, lung SCC was named as one of the three tumors to be analyzed by high-throughput methods in the pilot study of The Cancer Genome Atlas (TCGA). The traditional approach to the task of identifying genomic drivers of tumor development and progression has involved comparing large numbers of tumor samples looking for common features, an effective strategy when the objective is to identify tumor markers. However, comparing multiple genomes, while obviously an important starting point, has revealed a dearth of significantly mutated genes that could constitute potential drug targets for specific cancer types [7]. It is likely that there is a requirement for a more functional understanding of the malignant phenotype based on recognition of the multiplicity of affected genes and signal transduction pathways acting in concert.

Application of high-coverage, high-throughput sequencing to the whole genomes of single samples of small cell and adenocarcinoma, the other main lung cancer histotypes, was able to reveal insights into disease etiology and selective pressures within the tumor microenvironment [8,9]. Here, we have questioned whether the same approach can yield further insight into the biology of, and reveal candidate carcinogenic mechanisms for, lung SCC by next-generation sequencing of an appropriate tumor cell line (LUDLU-1) and a matched lymphocyte cell line (AGLCL). However, to further understand the cellular consequences of somatic events and allow us to prioritize them functionally, we integrated our genomic data with that of the transcriptome by also undertaking high-coverage RNAseq of both LUDLU-1 and a normal bronchial epithelial cell line (LIMM-NBE1). We captured coding, non-coding, and small transcripts in a strand-directional manner, making ours the first study, to our knowledge, to provide an integrated whole-genome and whole-transcriptome analysis to this level of transcriptional detail. Our aim was not only to catalogue genomic abnormalities but also to determine if they are functionally active by revealing the consequences for gene expression both for coding and non-coding genes and also to consider how their combined effect could contribute to malignancy in this tumor specimen. We provide general results from our genome-wide analysis as well as a list of genes and events that were highlighted by our integrated approach.

Materials and Methods

Cell Lines

LUDLU-1 is a cell line derived from a lung SCC, and AGLCL is its matching EBV-transformed lymphoblast cell line [10]. Both were cultured using standard cell culture techniques. The LUDLU-1 cell line was shown by immunohistochemistry, after being formalin fixed and paraffin-embedded, to be positive for P63 and negative for TTF-1, confirming it to be a lung SCC (results not shown). LIMM-NBE1 is a transformed bronchial epithelial cell line established for this study. Briefly, a primary culture derived from normal lung tissue was exposed to a *TERT* gene lentiviral vector after passage 5. Its epithelial origin was confirmed using antibodies to cytokeratin (results not shown). The cells were cultured in serum-free defined medium using standard techniques and DNA/RNA was prepared after passage 39 from cells harvested in the exponential growth phase. Written consent was obtained from the patient and ethics approval was given by the Leeds West Research Ethics Committee (07/H13071/146).

Multicolor fluorescence in situ *hybridization*

Chromosomes were prepared from LUDLU-1 cell cultures following standard procedures and labeled using the Metasystems 24XCyte mFISH probe kit together with a 4′,6-diamino-2-phenylindole counterstain. Images were recorded and merged using a Zeiss Axioskop microscope with Smart Capture 3.1.0 and pseudocolored and identified using Smart Type software (both Digital Scientific Ltd, Cambridge, United Kingdom).

DNA/RNA Extraction

DNA and total RNA were co-extracted from cell lines using TRIzol Reagent (Life Technologies, Paisley, United Kingdom), following the manufacturer's instructions. To remove phenol contamination, total RNA was repurified using the RNeasy Kit (Qiagen, Crawley, United Kingdom). Quality control consisted of quantification using Qubit Assay Kits and the Qubit fluorometer (Life Technologies) and measurement of DNA/RNA integrity using gel electrophoresis and the Agilent Bioanalyzer (Agilent Technologies, Stockport, United Kingdom). Ribosomal RNA was removed from total RNA, by LGC Genomics, using Ribo-Zero Kit (Epicentre Biotechnologies, Madison, WI).

Sequencing

DNA was sequenced by Complete Genomics for LUDLU-1 and AGLCL as described in Drmanac et al. [11]. LIMM-NBE1 DNA was sequenced to low coverage (0.06×), for the purposes of copy number estimation, on an Illumina GAII according to Wood et al. [12].

Two RNAseq libraries were prepared using Illumina protocols that were optimized, in-house, by LGC Genomics: one strand-specific and one small RNA library. Single-end reads (50 bp) were sequenced using an Illumina HiSeq 2000.

Sequencing Read Archive accession numbers are ERP001771 (high-coverage LUDLU-1 and AGLCL DNA sequencing), ERP001814 (low-coverage LIMM-NBE1 DNA sequencing), and ERP001465 (LUDLU-1 and LIMM-NBE1 RNA sequencing).

Alignment

All sequenced reads were aligned to National Center for Biotechnology Information human genome reference build 37, except for the small RNA reads, which were aligned to build 36 known microRNA (miRNA) using miRanalyzer [13]. DNA reads were aligned as part of the Complete Genomic pipeline [11]. Adaptor sequence was removed from RNA reads, which were then trimmed from the 3′ end to ensure an average Phred-scaled quality of 20. Strand-specific RNA reads >19 bp were aligned using Tophat [14], with a maximum of 4% mismatches and segment lengths set to half the read length for reads >33 bp. Reads that aligned to more than five genomic locations were discarded. Strand-specific RNA reads <20 bp were aligned to National Center for Biotechnology Information human reference genome build 36 using miRanalyzer [13].

Mutation Detection

Structural variants were extracted from the appropriate Junctions files, supplied as standard with Complete Genomics sequencing data [15].

Mutation Validation

Sixty somatic structural variants were selected for validation by polymerase chain reaction and Sanger sequencing, based on their predicted ability to interfere with the transcription of genes. Primer design failed for three, but of the 57 that we were able to successfully investigate, there were two false positives (germ line) and 55 true positives.

Copy Number Data

For LUDLU-1 and AGLCL, we used an adapted version of the script bam2windows.pl from CNAnorm [16] to count reads from the raw data provided by Complete Genomics (the mapping_* tsv.bz2 files) across 50-kb windows. We excluded reads not mapping uniquely, those with quality score < 5, and those where the two reads in a pair are mapping on different chromosomes, or more than 1000 bp apart. Other parameters were left as default.

Copy number analysis for LIMM-NBE1 was performed using CNAnorm. A control sample was made using data from a pool of 20 British individuals downloaded from the 1000 genomes project (PMID 20981092). Aligned LIMM-NBE1 reads were split into windows of 300 kb, averaging 200 reads per window. CNAnorm was used to estimate copy number with normalization method set to "closest."

Expression Analysis

Gene annotation files, including assignment of transcript subtype, were downloaded from Ensembl 60 [17] and transcript records merged to create a single annotation per gene ID. SEQEM [18] was used to assign the most probabilistic location of multireads. The number of reads mapping to exons were summed per gene and normalized by the number of exonic base pairs and total number of mapped reads per sample, where appropriate. We inspected the interquartile ranges of expression in both the tumor and normal and decided that "expression" should be defined as $\geq$10 reads mapped per kilobase per million reads mapped (RPKM). The small RNA sequencing library preparation was separate, so we repeated the above on the resulting data that we used to inspect miRNA, resulting in an expression threshold of $\geq$3.

Functional Annotation

Gene ontology and pathway analysis for protein-coding genes was completed using the DAVID functional annotation tool [19]. Functional analysis of miRNA was completed using TAM (Tool for Analysing miRNA) [20].

Correlation Analysis

Spearman correlation coefficient, and significance thereof, between ploidy and either tumor expression or fold change in expression (tumor/normal) was calculated in R for all genes (or genomic windows) that were expressed in the tumor or in both cell lines, respectively.

Transcription across Breakpoints

A bespoke python script was created to assemble a 600-bp sequence that centered on the each predicted breakpoint and output these in fasta format. The file was then used as a reference set and bowtie alignments identified all reads that aligned to the breakpoints but not anywhere on the human reference genome. To be considered as a transcribed breakpoint, at least three unique reads had to span the breakpoint by more than three bases either side.

Results

The LUDLU-1 genome was sequenced to an average coverage of 61× and the matched lymphocyte cell line to 55×. A total of 600.4 million strand-directional LUDLU-1 RNA reads aligned to the human genome, and 169,000 additional small reads aligned to known miRNA. To provide a baseline from which to quantify tumor-specific changes in expression, we also sequenced the full transcriptome of h-*TERT* transformed normal bronchial epithelium, the cell line for which we named LIMM-NBE1. A total of 700.8 million strand-directional LIMM-NBE1 RNA reads aligned to the genome and 185,000 small reads to known miRNA. We also sequenced the genome of LIMM-NBE1 at low coverage (0.06×) to inspect the copy number profile and found two distinct abnormalities: gain of one copy of chromosome 20 and one copy of 1q (Figure W1). We believe these abnormalities are a consequence of h-*TERT* transformation, based on the observation of similar structural variants, and gain of chromosome 20 specifically, on transformation of a series of other human bronchial epithelium samples [21]. Transcription levels in LIMM-NBE1 in these regions may, therefore, differ from a truly diploid cell line. To assess whether the noted copy number abnormalities in LIMM-NBE1 affected the genome-wide transcriptome research, detailed below, we performed our analysis both with and without those regions included and found there to be no significant difference in the conclusions drawn. One limitation of our approach is a lack of biologic replicates, meaning we are unable to inspect differential gene expression between our samples. This does not inhibit an integrated analysis of the genomic and transcriptomic data but does mean that differences in expression between LUDLU-1 and LIMM-NBE1 can only be denoted as fold change, with no indication of the significance of the difference outside of these samples.

Aneuploidy is a common occurrence in cancer genomes and can be a mechanism for regulating expression of the genes involved in disease progression [22–24]. Our high-resolution copy number analysis in LUDLU-1 enabled comparison with previously characterized lung SCC genomic landscapes, and the RNAseq data enabled concurrent expression quantification in regions of interest.

Candidate Genes in Regions of Common Copy Number Alteration in Lung SCCs

LUDLU-1 is largely tetraploid (Figure 1) with many anomalies previously characterized in lung SCC cohorts [25–30]. Table 1 lists these commonly altered genomic regions and the previously suggested candidates of carcinogenesis, including expression in the normal and the tumor cell line, within them (except for homozygous deletions where all deleted genes are listed). The largest tumor-associated gain in expression is for *SOX2* [ENSG00000181449], a candidate oncogene that can confer stem cell–like qualities to cancer cells and whose amplification has been shown to be a key event in the formation of SCC of the lung [31–33].

LUDLU-1 contains two regions of homozygous deletion, both within 9p. The only involved gene that is expressed in the normal bronchial epithelium is *PTPRD* [ENSG00000153707], which encodes a protein tyrosine phosphatase receptor indicated to have tumor suppressor function in laryngeal SCCs and lung cancer (with homozygous deletions observed in adenocarcinomas and small cell and large cell lung cancer lines) and for which tumor suppressor function has been shown in neuroblastoma [34–36]. This region also contains *CDKN2A* [ENSG00000147889] and *CDKN2B* [ENSG00000147883], which encode p16 and p15 cyclin-dependent kinase inhibitors, respectively. These are known tumor suppressor genes whose products are involved in the onset of cellular senescence [37]. Homozygous deletion of this area has previously been observed in 26% and 36% of lung SCCs in two separate studies and is significantly associated with never smokers compared to current, or previous, smokers [38–40]. The somatic point mutational profile

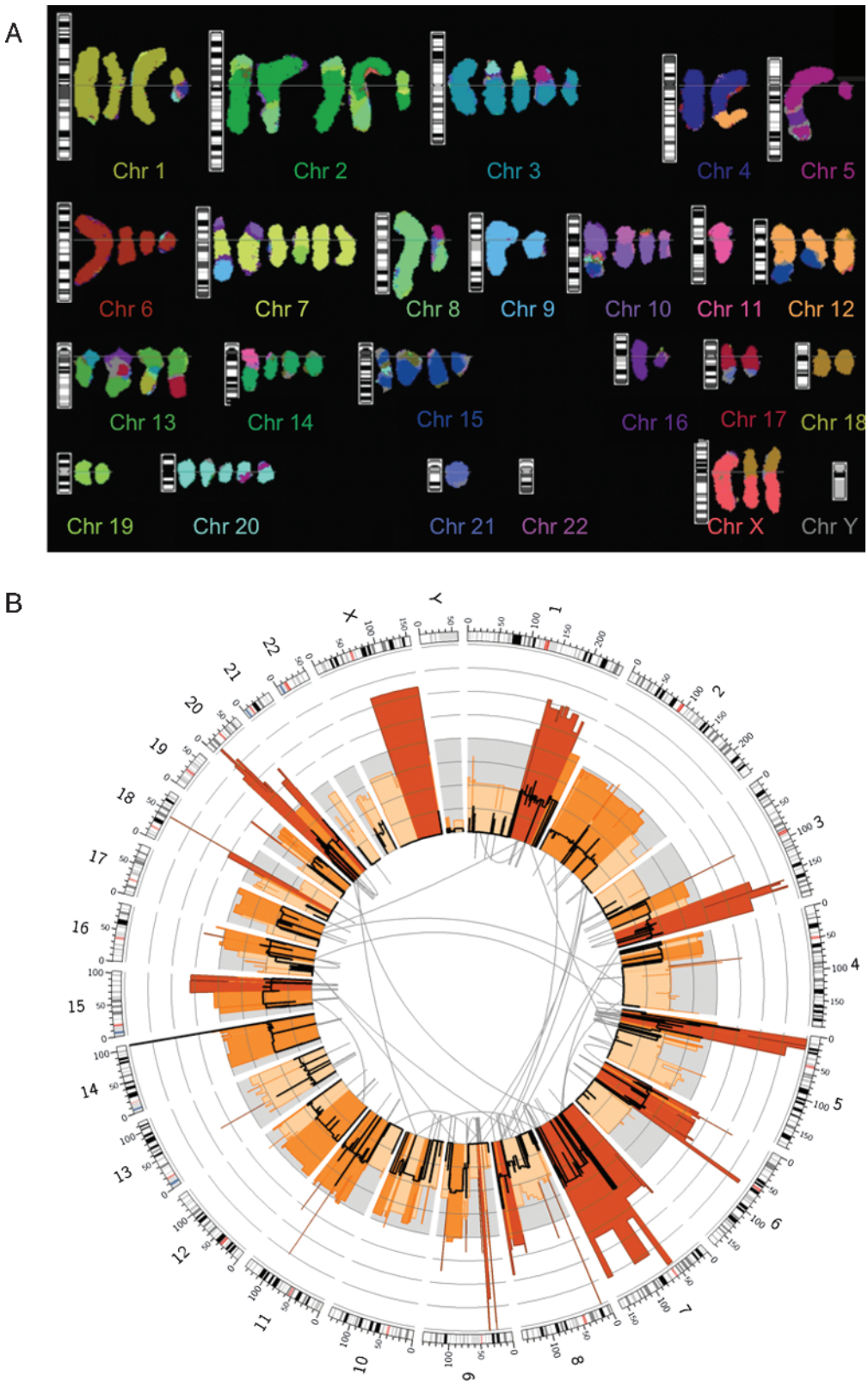

Figure 1. Two depictions of the lung SCC cell line (LUDLU-1) genome. (A) Visualization through multicolor fluorescence *in situ* hybridization. This image is the chromosomes from a single cell. The hybridized probes are colored according to their chromosome of origin, with each chromosome name being written in the corresponding color. (B) The genome averaged over all cells in the sample, visualized through a Circos plot. The outer track consists of chromosome ideograms with centromeres in red. Starting from the inner track, there is sequential depiction of structural rearrangements (intrachromosomal variants appear as short gray lines, whereas interchromosomal variants appear as arcs across the inner circle) and copy number changes (shaded gray is the most common, tetraploid, with dark and light orange showing gain and loss, respectively). Allelic ratios are represented by a black line, superimposed over copy number, with a return to baseline indicating loss of heterozygosity.

Table 1. Copy Number Anomalies in LUDLU-1 that Have Been Previously Observed in Lung SCCs.

LUDLU-1 Copy Number Anomaly	Region	Percentage of Lung SCC Samples Observed (*n*)	References	Candidate Genes					
				ID	Name	Copies in LUDLU-1 (Allelic Ratio)	LUDLU-1 RPKM	LIMM-NBE1 RPKM	Fold Change[†]
Gain	3q	76% (258)	[25–30]	ENSG00000114200	*BCHE*	7 (1:6)	13	2	7.83
				ENSG00000163581	*SLC2A2*	7 (1:6)	0	0	0
				ENSG00000085276	*EVI1*	7 (1:6)	127	294	0.43
				ENSG00000136603	*SKIL*	7 (1:6)	35	78	0.45
				ENSG00000114346	*ECT2*	8 (1:7)	1036	390	2.66
				ENSG00000121879	*PIK3CA*	6 (1:5)	374	273	1.37
				ENSG00000113916	*BCL6*	6 (1:5)	63	69	0.92
				ENSG00000114315	*HES1*	6 (1:5)	72	46	1.58
				ENSG00000181449	*SOX2*	6 (1:5)	11	0	108.91
	5p	56% (229)	[25–28,30]	ENSG00000145604	*SKP2*	7 (1:6)	358	235	1.52
				ENSG00000168621	*GDNF*	7 (1:6)	0	0	1.01
	8q	44% (229)	[25–28,30]	ENSG00000136997	*MYC*	5 (1:4)	104	183	0.57
	7p	43% (156)	[25,26,28]	ENSG00000106268	*NUDT1*	5 (0:5)	19	8	2.29
				ENSG00000146648	*EGFR*	8 (2:6)	2120	935	2.27
	1q	49% (51)	[25,30]	ENSG00000213190	*AF1Q*	6 (1:5)	5	6	0.85
				ENSG00000143549	*TPM3*	6 (1:5)	1314	728	1.81
				ENSG00000163131	*CTSS*	6 (1:5)	15	2	7.19
				ENSG00000162924	*REL*	4 (0:4)	53	49	1.09
	20p	14% (135)	[25,27,28]	ENSG00000172264	*MACROD2*	7 (1:6)	21	7	2.80
	7q	35% (144)	[26,28]	ENSG00000105851	*PIK3CG*	7 (1:6)	0	0	0
				ENSG00000106462	*EZH2*	5 (1:4)	228	151	1.51
				ENSG00000105989	*WNT2*	5 (1:4)	0	0	1.02
				ENSG00000105976	*MET*	5 (1:4)	529	971	0.55
	20q	29% (144)	[26,28]	ENSG00000101412	*E2F1*	4 (1:3)	12	25	0.50
Loss	3p	53% (229)	[25–28,30]	ENSG00000189283	*FHIT*	2 (0:2)	4	9	2.10
				ENSG00000068028	*RASSF1*	2 (0:2)	4	16	4.47
				ENSG00000134086	*VHL*	2 (0:2)	9	12	1.26
	5q	34% (195)	[25,26,28,30]	ENSG00000134982	*APC*	2 (0:2)	45	95	2.08
	8p	33% (195)	[25,26,28,30]	ENSG00000149557	*FEZ1*	4 (1:3)	60	334	5.60
				ENSG00000120875	*DUSP4*	2 (0:2)	19	168	8.78
	13q	38% (195)	[25,26,28,30]	ENSG00000139687	*RB1*	3 (1:3)	484	832	1.72
	17p	20% (12)	[25]	ENSG00000141510	*TP53*	2 (0:2)	262	181	0.69
	18q	22% (67)	[25,26]	ENSG00000187323	*DCC*	2 (0:2)	0	0	0.98
	4q	36% (128)	[26,27]	ENSG00000163629	*PTPN13*	2 (0:2)	500	842	1.68
	9p	34% (135)	[25,27,28]	ENSG00000197919	*IFNA1*	1 (0:1)	0	0	0
Loss (homozygous)	chr9:8.65-10.1Mb	N/A	N/A	ENSG00000147889	*CDKN2A*	N/A	0	0	0
				ENSG00000147883	*CDKN2B*		0	0	0
				ENSG00000176399	*DMRTA1*		0	2	0
				ENSG00000234840	*FLJ35282*		0	0	0
	chr9:21.85-23.05Mb	N/A	N/A	ENSG00000153707	*PTPRD*	N/A	0	11	110.00

[†]Fold change is given as tumor/normal for regions of gain and normal/tumor for regions of loss.

for LUDLU-1 is not consistent with tobacco exposure, leading us to propose a non-tobacco exposure–based etiology for this specific SCC (L. Stead, S. Berri, H.M. Wood, P. Egan, C. Conway, C. Daly, R. Chalkley, O. Belvedere, K. Papagiannopoulos, K. MacLennan and P. Rabbitts, submitted for publication) [8].

To identify additional candidates, we inspected the fold change in expression between LIMM-NBE1 and LUDLU-1 for every gene within the commonly altered regions outlined in Table 1 (Tables W1 and W2).

Regions of gain. In total, 10,218 genes lie in the commonly amplified genomic regions, with seven expressed more than 1000-fold higher in the tumor than the normal. Five of these have previously been associated with lung carcinogenesis, although one, *PON1* [ENSG00000005421], an antioxidant paraoxonase that scavenges endogenous free radicals, has previously been observed at decreased levels in tumors [41]. Elevated expression of *PON1* in LUDLU-1 may be a result of increased levels of endogenous reactive oxygen species in this cell line, specifically, a phenomenon for which we have devised a hypothesis (L. Stead, S. Berri, H.M. Wood, P. Egan, C. Conway, C. Daly, R. Chalkley, O. Belvedere, K. Papagiannopoulos, K. MacLennan and P. Rabbitts, submitted for publication). The remaining candidates with previous lung cancer associations are *ELF3* [ENSG00000163435], which can be overexpressed in lung carcinomas and has the ability to regulate lung cell proliferation and differentiation in response to bronchial epithelial cell damage in mice [42,43]; *STEAP4* [ENSG00000127954], homologous to a gene that is overexpressed in lung SCCs and whose protein products are candidates as tumor-associated antigens [44]; *RAB25* [ENSG00000132698], a known oncogene that is part of a 12-gene subset whose expression pattern is able to classify the lung cancer histologic subtypes, with apparent increased protein product levels observed in SCC especially [45]; and *KIAA1324L* [ENSG00000164659], homologous to a gene that is overexpressed in metastatic lung cancers and able to stop cells from undergoing apoptosis in response to stress [46,47]. The remaining genes within commonly amplified regions that exhibit large increased expression in LUDLU-1 are *GRHL2* [ENSG00000083307] and *RIMS2* [ENSG00000176406]. While these genes reside in 8q, a region that is amplified, overall, in LUDLU-1, they are both present at three copies, i.e., lower than the tetraploid state that is considered "normal" for this cell line. *GRHL2* is an oncogenic candidate owing to its ability to

inhibit expression of death receptors, thereby inhibiting apoptosis [48]. *RIMS2*, to our knowledge, has not been associated with carcinogenesis. It is known to encode a protein involved in exocytosis that can interact with Rab3, a member of the Rab family of guanine triphosphatases (GTPases), which have been implicated in cancer but not to the extent of the Ras and Rho GTPase families [49]. Gene ontology analysis of all genes within the specified, commonly amplified regions in lung SCC revealed significant (hypergeometric, false discovery rate [FDR] < 0.05) enrichment of the biologic processes of keratinization and epithelial, epidermal, and keratinocyte differentiation. The presence of keratin "pearls" is one of the markers that pathologists use to assign an SCC histology to lung cancer.

Regions of loss. Cumulatively, the regions that commonly show a reduction in copy number for lung SCC contain 8071 genes, a subset with significant (hypergeometric, FDR < 0.05) enrichment in the cellular component of extracellular space, biologic processes of cell-cell signaling and adhesion, and molecular function of cytokine activity and binding. Cytokines are intercellular signaling molecules, whose expression and activity is often altered during carcinogenesis with pleiotropic effects [50]. Increased expression of cytokines in lung SCC cell lines has been previously recorded [51]. Nine of the 8071 genes are expressed 1000-fold less in the tumor than the normal bronchial epithelium, and in concurrence with the gene ontology analysis, four are associated with the extracellular matrix: *POSTN* [ENSG00000133110], *SPARC* [ENSG00000113140], *VCAN* [ENSG00000038427], and *FBN2* [ENSG00000138829]. The five additional highly underexpressed genes are *FAT4* [ENSG00000196159] that encodes a cadherin that has been associated with tumor suppressor function in breast cancer [52]; *EDNRA* [ENSG00000151617] that is a homologue of a gene observed to undergo hypermethylation in lung cancers [53]; *ODZ2* [ENSG00000145934] that resides in 5q and encodes a protein called Teneurin-2, which has been shown to be a membrane-bound transcriptional regulator with the potential to switch neural cells from a stage of growth to one of differentiation; and both *AREG* [ENSG00000109321] and *AREGB* [ENSG00000205595] that encode amphiregulins. Amphiregulins are epidermal growth factors that bind and activate the receptor epidermal growth factor receptor (EGFR), a signaling pathway that is activated specifically in some lung adenocarcinomas [54]. The apparent down-regulation of these genes in lung SCC again emphasizes the differences in the underlying biology of the two main non–small cell lung cancer histotypes [55].

Finally, using the list of 10,961 protein-coding genes that were expressed in either LUDLU-1 or LIMM-NBE1 as background, we looked for pathway enrichment in the subset of genes exhibiting a fold change in expression of 10 or more, irrespective of whether the gene was upregulated or downregulated in the tumor. Significant enrichment (FDR < 0.05) was observed for pathways involved in extracellular matrix-receptor interactions, focal adhesion, and integrin cell–surface interactions. This is consistent with the functional annotation enrichment in the subset of genes in regions of loss. Additionally, pathways involved in hemostasis were enriched. Hemostasis will be affected during sustained angiogenesis, one of the hallmarks of cancer in which integrin signaling also plays a known role [56]. This implies that angiogenic regulation is a particularly important process in LUDLU-1 cells. Thirteen members of the phosphatidylinositol 3-kinase (PI3K) pathway exhibited more than 10-fold change in expression between the SCC and the bronchial epithelial cell line, but the enrichment in members of this pathway was not significant after multiple testing correction.

Each gene candidate highlighted in the preceding analysis is protein coding. However, there are several types of transcript besides those with protein-coding capacity [57]. The past decade has brought an increased interest in characterizing functional non-coding RNA (ncRNA), and their associations with carcinogenesis are now well documented [58–60]. The delineation of transcripts outside of the 2% of the genome that encode protein is also still ongoing, with much of the transcriptome remaining to be fully characterized. To overcome these issues and inspect the effect of copy number on transcription in a genome-wide manner, we decided to investigate the correlation between ploidy and expression using 50-kb windows, as well as in a gene-based manner but within subtypes of functional transcript.

Genome-wide Analysis of Copy Number Effects on Transcription

We analyzed the effect of ploidy on transcription in two different ways; we assessed the correlation between copy number in LUDLU-1 and either 1) absolute expression or 2) the fold change in expression between LUDLU-1 and LIMM-NBE1. Significant (Spearman $P < .05$) but weak correlations were observed, using both approaches, when assessing expression within the 50-kb windows, protein-coding genes, antisense transcripts, long intergenic ncRNA (lincRNA), and miRNA but not for small nucleolar RNA (snoRNA), small nuclear RNA, or pseudogenes. Figure 2 displays the relative values of the correlation coefficients. The largest difference in the coefficients obtained by the two expression metrics is for the protein-coding genes, where copy number appears to have a larger effect on fold changes in expression than absolute expression values. Taking proteins as the "analogue output" of the genome [61], this is an interesting observation that implies that copy number affects protein levels in a more complex manner than simple gene dosage, perhaps through the regulatory networks that govern protein production. LincRNAs are intuitive candidates for modulators of this effect owing to their roles in both pre-transcriptional (through chromatin remodeling and gene silencing) and post-transcriptional (by binding to and targeting proteins) regulation of mRNA [62]. We, thus, noted that lincRNAs are the only subtype in which their absolute tumor expression correlates more with copy number than the fold change in expression between tumor and normal does. Boxplots of expression levels and fold changes observed at one to six copies of 50-kb windows, all genes, protein-coding genes, and lincRNA genes (Figure W2) reiterate that copy number correlates more with changes in expression between the normal bronchial epithelium and the tumor cell lines than with the absolute expression in the latter.

To see how the expression levels of each functional subtype of transcript varied, irrespective of copy number, we also inspected our RNAseq data in isolation.

Functional Transcripts Exhibit Variable Expression Patterns

Average expression levels per functional transcript class are given in Table 2 and Figure 3. Only snoRNA exhibited an overall change in the mean expression level with a significant decrease (paired *t* test, $P < .05$) in the tumor compared to the normal. An expression signature of six snoRNA has been proposed as a biomarker for non–small cell lung cancer with all upregulated in tumors [63]. Five of these had two-fold to seven-fold increased expression in LUDLU-1 compared to LIMM-NBE1, despite our overall observed trend of decreased transcription levels for snoRNA. The remaining biomarker snoRNA was expressed at four orders of magnitude lower, in both samples, than the other five.

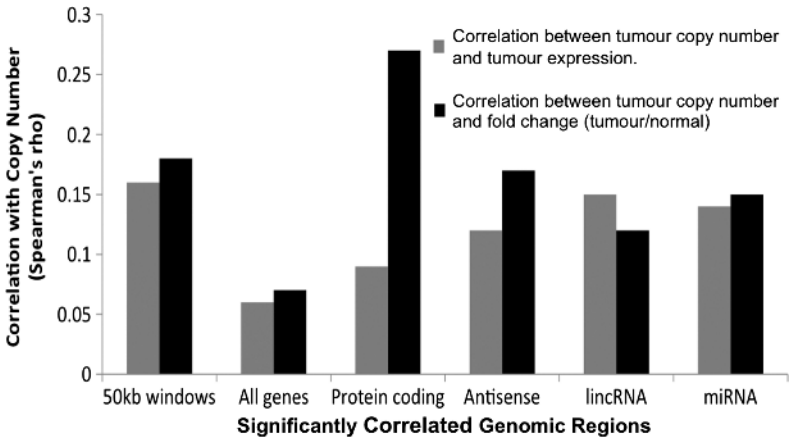

Figure 2. Dosage effects on expression. Correlation coefficients between copy number and either absolute tumor transcript expression (gray bars) or fold change between tumor and normal (black bars) for different functional transcript classes. Each coefficient is significant (Spearman $P < .05$).

The correlation coefficients for expression of the various transcript classes between samples denote whether the genes are ranked in a similar order (Table 2). miRNA have the lowest correlation coefficient at 0.12, indicating that this class of transcript undergoes the largest change in expression pattern. Of the 591 miRNA for which we observed expression in either sample, only 48% were expressed in both, with 32% being uniquely expressed in the bronchial epithelium and 20% unique to the SCC. There are separate sequencing library preparation protocols for the distinct capture of small (<30 bp) RNA molecules. On sequencing the additional libraries that we made for LIMM-NBE1 and LUDLU-1 using these specific protocols, the coefficient increased to 0.37 but remains the lowest of the subtypes. miRNA have been shown to stabilize cell states, whether they be quiescent or proliferative, prompting researchers to propose that these transcripts confer robustness in response to cellular stresses [64]. An overhaul in miRNA expression pattern, as indicated by the low correlation coefficient between our cell lines, may have been causal (potentially offering therapeutic targets) or consequential (potentially offering diagnostic biomarkers) to the tumorigenic progression of LUDLU-1 by facilitating or stabilizing transitory cell states. The *let-7* family of miRNA is believed to act as tumor suppressors in non–small cell lung cancer, with decreased expression correlating with poor prognosis [65,66]. Our data show that, of the 13 *let-7* miRNA, the normalized read count in our normal epithelial cell line was greater than that in tumor with three members exhibiting a greater than 10-fold decrease in the SCC sample (Table 3). miRNA has also been successfully used to distinguish lung SCC from adenocarcinoma, implying a role in the differing biologic processes that underlie these subtypes [67,68]. To expand our analysis, we inspected the enrichment of clusters, families, and functions within the 55 miRNA that exhibited an expression fold change of 10 or more between samples. While no findings were significant after multiple testing corrections were applied, we note that the hsa-let-7e cluster and miR-99 family both had unadjusted *P* values of <.05.

Table 2. Average Expression Levels for Different Functional Transcripts.

Type of RNA	Number of Transcripts	Mean RPKM (Median)		Correlation Coefficient
		Normal	Tumor	
Protein coding	17830	107 (14)	100 (12)	0.67
Pseudogene	410	12.7 (0.3)	12.2 (0.07)	0.87
Antisense	2254	12.7 (0.034)	21.7 (0.01)	0.72
lincRNA	2290	0.7 (0.016)	0.9 (0.02)	0.62
miRNA	591	4.4 (0.09)	17 (0.08)	0.12
snoRNA	1002	2215* (0.74)	1530* (0.54)	0.87
Small nuclear RNA	899	86 (0.025)	147 (0.08)	0.73

*Significant (paired *t* test, $P < .05$) difference in the mean expression level.

The above quantification of expression using RNAseq data involves counting the number of reads that map to annotated regions of transcription within the reference genome. However, structural rearrangements, which commonly occur during carcinogenesis and constitute part of the somatic mutational profile of each tumor, can be identified through whole-genome sequencing, and concurrent RNAseq can then inspect and quantify transcription across newly fused genomic regions. Such structural rearrangements can be key pathogenic drivers of cancer and, thus, provide therapeutic targets, as has been observed for lung adenocarcinoma [2,69].

Transcribed Structural Rearrangements

We identified 171 large (>100 bp) somatic structural variants in LUDLU-1: 66 deletions, 47 inversions, 29 translocations, and 29 duplications. More than half interrupt annotated genes. This is consistent with observations in whole-genome sequencing of both a small cell lung cancer and a lung adenocarcinoma cell line, where 64% and 63% of structural variants interrupted annotated genes, respectively [8,9]. However, we only found evidence of transcription (at least three uniquely aligned tumor RNA reads) across six (Figure 4). Deletion of bases 60,392,103 to 60,650,744 in 3p removes exon 5 of *FHIT* [ENSG00000189283] (Figure 4*A*), creating a frameshift and removing the 5′ untranslated region (UTR) for all known isoforms. Transcription through this breakpoint, while on the same strand as *FHIT*, is, thus, unlikely to be translated. *FHIT* is a tumor suppressor gene and deletion of exon 5 has been observed previously in thyroid, digestive tract, and ovarian tumors [70–72]. RNA reads map throughout the deletion, implying that one of the two copies of *FHIT* in LUDLU-1 retains the region. However, tumor expression is six-fold lower than that in the normal bronchial epithelium.

A smaller deletion in 6p involves bases 26,161,020 to 26,222,257. Transcription through this breakpoint is on the same strand as *HIST1H2BD* [ENSG00000158373], a gene that the deletion affects by removing exon 2 (Figure 4*B*). Seven additional histone genes lie fully

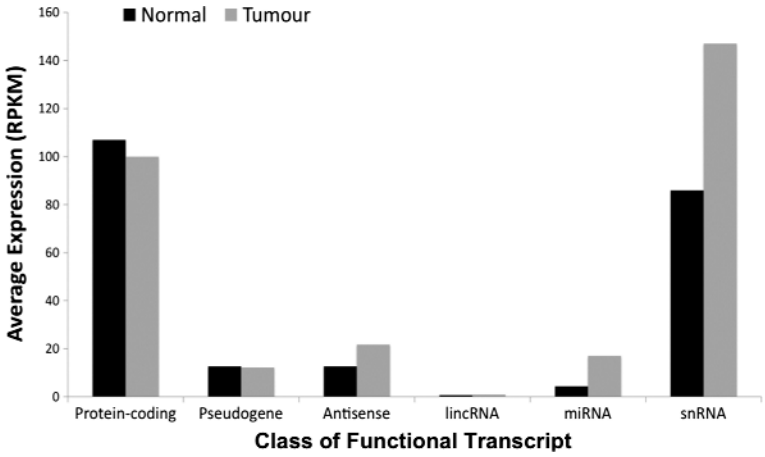

Figure 3. Functional transcript expression. Mean expression levels for different functional transcript classes in both the normal bronchial epithelium cell line (gray bars) and lung squamous cell tumor cell line (black bars). Expression is quantified as RPKM.

within this LUDLU-1 deletion on 6p: *HIST1H2BE*, *HIST1H3D*, *HIST1H2BG*, *HIST1H2AE*, *HIST1H2BF*, *HIST1H4D*, and *HIST1H4E*. Tumor ploidy in this region is four. There are modest decreases in tumor expression, compared to the normal bronchial epithelium, for six of the seven histone genes, but *HIST1H2BE* expression in LUDLU-1 is 40-fold lower than LIMM-NBE1. The 6p histone cluster has not been implicated previously in carcinogenesis. On 7q, deletion of exon 4 of *ATP5J2* [ENSG00000241468] is adjacent to a ~100-bp inversion in a neighboring intron (Figure 4*C*). Transcription through this breakpoint is from the same strand as *ATP5J2* but also that of a known read-though transcript that starts with *ATP5J2* and continues to *PTCD1* [ENSG00000106246], which creates a conjoined gene *ATP5J2-PTCD1* [ENSG00000248919]. Conjoined genes can be functional, with potential regulatory affects on the constituent genes. The somatic rearrangement in LUDLU-1 may favor expression of the read-through owing to the removal of the final exon from *ATPJ52* [73,74]. *ATP5J2* encodes a subunit of complex V of the respiratory chain, while *PTCD1* alters the activity of complex IV [75]. Aberrant, dual regulation of these genes, by the mutated read-through transcript, could affect oxidative phosphorylation in ways that may be carcinogenic, as reviewed in Lemarie and Grimm [76].

A much larger inversion is present in LUDLU-1 on 12q, where base pairs 104,509,980 to 104,715,714 have been duplicated and inverted at position 104,700,066 (Figure 4*D*). This has three main effects: complete disruption of *TXNRD1* [ENSG00000198431], amplification of *EID3* [ENSG00000255150], and aberrant, antisense transcription of *NFYB* [ENSG00000120837]. The latter is owing to transcription through the inversion that results in expression of the complementary strand to the nuclear transcription factor gene. Sense expression of *NFYB* is the same in LUDLU-1 and the normal bronchial epithelium, whereas antisense expression is 25-fold higher in the tumor. This could lead to regulation of the *NFYB* mRNA and, thus, altered protein levels within the tumor [77]. *NFYB* encodes a subunit that forms complexes with, and alters the function of, p53 [78]. Interestingly, we also found a non-synonymous somatic mutation in *TP53* [ENSG00000141510] that results in a lack of a functional copy of p53 in LUDLU-1 (L. Stead, S. Berri, H.M. Wood, P. Egan, C. Conway, C. Daly, R. Chalkley, O. Belvedere, K. Papagiannopoulos, K. MacLennan and P. Rabbitts, submitted for publication). Mutant p53 is able to associate with the NF-Y protein complex, of which *NFYB* encodes a subunit, leading to a variety of cell cycle and apoptotic defects that are not fully characterized [79]. *TXNRD1* expression is approximately halved in the tumor compared to the normal cell line. This gene encodes a pyridine nucleotide oxidoreductase, which can help protect the cell against oxidative stress. Amplification of *EID3*, a potential inhibitor of cell differentiation, in LUDLU-1 is concurrent with a 12-fold increase in its expression [80].

The RNA reads resulting from transcription through a structural mutation on 6q, in which bases 65,693,384 to 71,306,346 are removed, indicate an immediately adjacent inversion up to base 71,306,397 (Figure 4*E*). The latter region is downstream from a predicted protein-coding gene: *C6orf57* [ENSG00000145079]. The deletion includes *BAI3* [ENSG00000135298], *LMBRD1* [ENSG00000168216], *COL19A1* [ENSG00000082293], *FAM135A* [ENSG000000082269], and up to intron 14 of *EYS* [ENSG00000188107], which is predicted to disrupt the codon usage in *EYS* translation if the aberrant transcript was attempted. Despite this deletion, the region is amplified, which may account for modest increases in expression levels of all genes involved in the structural variation; a 2-fold increase is observed for *LMBRD1* and *FAM135A*, 6-fold for *COL19A1*, and 10-fold for *BAI3*.

The final, transcribed structural variant is an inversion that creates a potential fusion gene (Figure 4*F*). It is located on 11q and causes a read-through from intron 3 of *FRMD8* [ENSG00000216391],

Table 3. Reads Mapping to *let-7* Family miRNA Genes, Normalized by Sequencing Depth.

miRBase ID	Normalized Read Count		Fold Change (N/T)
	LUDLU-1	LIMM-NBE1	
hsa-let-7c	51.7	2436.8	47.1
hsa-let-7b-5p	3796.2	67964.9	17.9
hsa-let-7e-5p	100.0	1417.3	14.2
hsa-let-7e-3p	0.9	8.1	9.4
hsa-let-7a-5p	1650.2	14399.0	8.7
hsa-let-7g-5p	544.0	1398.7	2.6
hsa-let-7f-5p	172.4	428.0	2.5
hsa-let-7b-3p	13.8	15.1	1.1
hsa-let-7d-5p	224.2	242.4	1.1
hsa-let-7a-2-3p	0.0	1.2	Inf
hsa-let-7a-3p	0.0	1.2	Inf
hsa-let-7i-5p	2572.8	1018.3	0.4
hsa-let-7d-3p	329.4	114.8	0.3

starting at base pair 65,157,970, into position 65,150,150, which is immediately before the transcription, start site for *SLC25A45* [ENSG00000162241]. These genes are usually on different strands, but the inversion causes them to locate on the same strand, and it is from that which the LUDLU-1 RNAseq reads have originated. It is not possible to predict whether this novel transcript would be capable of producing a chimeric protein using the RNAseq data. *FRMD8* encodes an uncharacterized protein that contains a single FERM domain, usually involved in localizing proteins to the plasma membrane [81], while *SLC25A45* codes for a transmembrane solute carrier that resides within the mitochondrial inner membrane [82].

Discussion

SCCs of the lung have extremely complex karyotypes. Nonetheless, in contrast, particularly to the adenocarcinoma subtype, specific chromosomal abnormalities occur frequently. Deletion or loss of heterozygosity of 3p and gain or amplification of distal 3q are common hallmarks of this subtype [83,84]. Interestingly, and of potential clinical value, the squamous subtype of tumors at other sites such as esophagus, head and neck, and cervix also carry the 3p loss/3q gain motif [85]. These, and other common genomic features, may be due to underlying similarities in etiology, suggesting that possible drug targets identified in one tumor type could be transferable to other organ sites with the same SCC histology. Our integrated approach has allowed us to probe these commonly altered regions in a lung SCC cell line, delineating the breakpoints to base pair resolution and prioritizing a list of candidate genes based on the somatic effects on transcription. Several known lung cancer-associated genes were highlighted in our analysis, such as *SOX2*, *ELF3*, and *RAB25*, indicating that other highlighted genes that do not have previous lung cancer, such as *GRHL2* and *RIMS2*, are worth further inspection using different approaches.

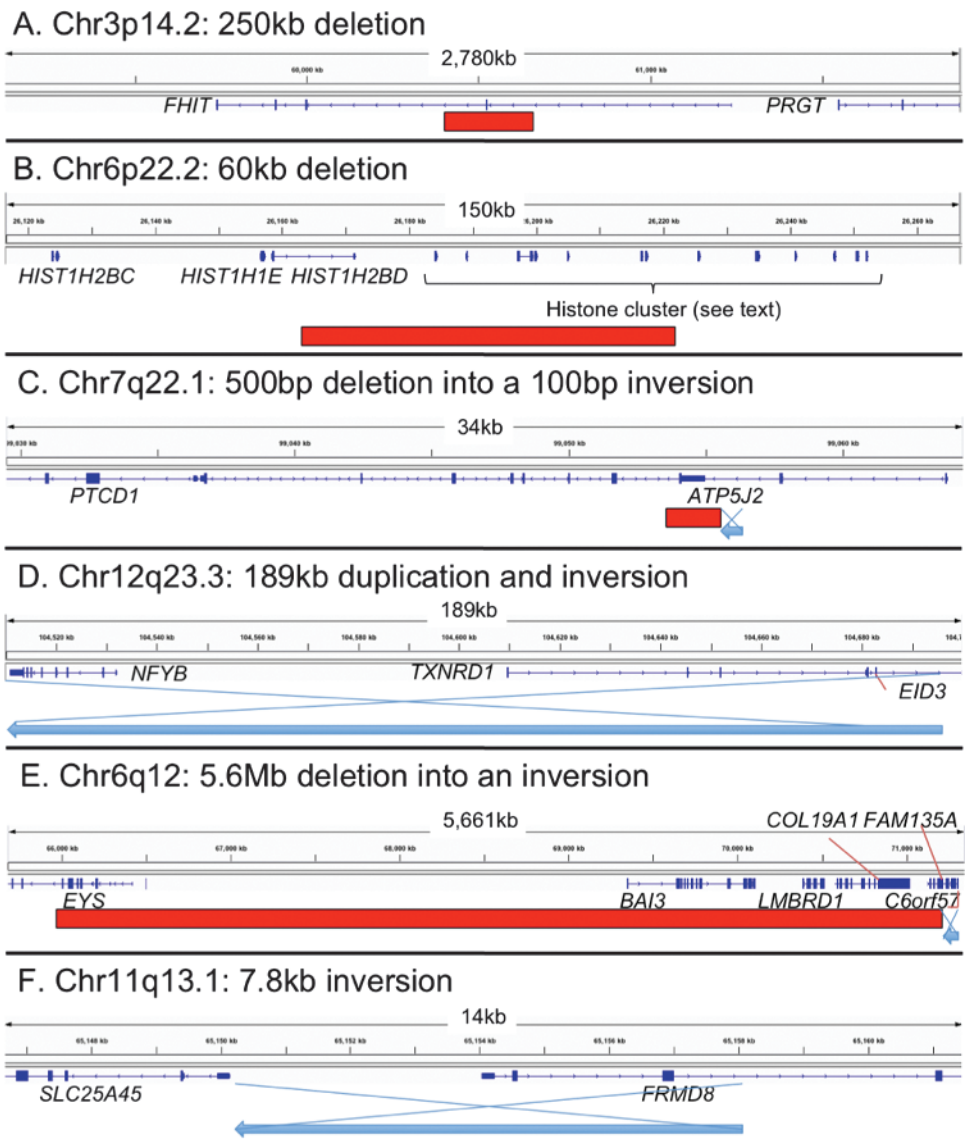

Figure 4. Transcribed breakpoints in the lung SCC cell line (LUDLU-1). Each of A–F depicts a separate event. The reference genome is denoted by a gray, demarcated double line with a double-headed arrow above to indicate scale. Gene regions are indicated below the double line in blue with vertical blocks representing exons and horizontal lines: introns. The name of each gene is given where appropriate. Underneath the genes, the transcribed breakpoints are depicted; deletions are denoted by red blocks and inversions by blue arrows, indicating their reversed direction in the novel genome made by LUDLU-1 structural variants.

We designed the RNAseq for this work to ensure that both coding and ncRNA species were investigated in a strand-specific manner. This allowed the transcriptome to be probed in great detail, including the ability to distinguish sense and antisense transcription. An example of a benefit from this approach was our ability to detect an inversion in 12q that results in a large increase in antisense transcription, which can ultimately affect protein levels, of *NFYB*, a gene that encodes part of a complex that associates with p53. Non–strand-directional methods would have attributed this antisense expression as increased *NFYB* mRNA production, giving a false expression level reading for the gene and potentially leading to a hypothesis about the gene's involvement in carcinogenesis that could, ultimately, be rejected.

Using an additional small RNA sequencing protocol enabled us to capture information on miRNA. While a lack of biologic replicates negates a robust analysis of differential expression, using fold change alone, we were able to highlight the let-7 and miR-99 subfamilies, members of which have been associated with tumor suppressor function in non–small cell lung cancer [65,86]. Furthermore, we showed that, across transcript classes, miRNA expression values are the least correlated between LIMM-NBE1 and LUDLU-1 cell lines. miRNA has been shown to form complex regulatory networks, able to act as either tumor suppressors or oncogenes, depending on their targets [87]. As such they have become an important area in cancer research in recent years, revealing their potential use in diagnostics and as therapeutic targets [68,88,89]. The latter is particularly important for lung SCC. Our analysis show that miRNA undergoes the largest change in expression profile between the normal bronchial epithelial and the tumor cell lines, potentially indicative of the involvement of miRNA early in the disease process, highlighting the importance of continued and expanded research into the involvement of these small RNA in lung carcinogenesis.

While this paper was under review, two large lung cancer studies were reported. TCGA published their findings on 178 lung SCCs, which included analysis of genome, exome, mRNA, and miRNA sequencing data [90]. Their results confirmed the significance in lung SCC of many of the copy number alterations discussed herein, also highlighting both *SOX2* amplification and *CDKN2A* homozygous deletion. Several pathways were highlighted by TCGA publication as yielding potential therapeutic targets: PI3K/AKT, the receptor tyrosine kinases, and RAS. We note that the PI3K pathway was highlighted in our single-sample analysis, although it was not significant following multiple testing. Additional therapeutic targets that were indicated by TCGA study are the fibroblast growth factor receptors (FGFRs). While FGFRs were not highlighted by our analyses, we note that the miRNA that we observed to have undergone the largest fold changes in expression was somewhat enriched for miR-99 family members and that a direct association between *hsa-miR-99b* [ENSG0000007550], acting as a tumor suppressor in non–small cell lung cancer, and *FGFR3* [ENSG00000068078] has been observed [86]. A study of 183 lung adenocarcinomas that reported the findings of genome and exome sequencing data analysis was also published [91]. One of the key similarities between the two studies, and ours, was the involvement of *CDKN2A*. However, no other genes highlighted in the adenocarcinoma publication were specifically extracted during our analysis, further suggesting the biologic differences between adenocarcinoma and SCC of the lung. In their discussion, the authors note that they believe that combined analyses, such as our approach of integrating genomic and transcriptomic data, will improve the annotation of important genes in carcinogenesis in the future.

The cost of next-generation sequencing is decreasing, but there is still choice to be made by researchers as to whether funds are better used to obtain detailed integrated information on single, or few, samples or focused information on a larger number of samples. We believe that the solution depends on whether the research is hypothesis driven or exploratory. This study is an example of the latter and we, herein, have shown that integrating high coverage genome and transcriptome data for a single sample can lead to valuable hypotheses. Both approaches are valuable, perhaps as part of iterative research in which exploratory data drive hypotheses, tested by additional data.

Acknowledgments

We thank Rick Tearle, a senior field applications scientist from Complete Genomics, for useful discussions on data analysis and the use of the company's proprietary tools. We also thank R.A. Weinberg, D. Trono, and I. Verma for supplying the plasmids through Addgene for the lentivirus used in the LIMM-NBE1 transformation.

References

[1] Rosell R, Moran T, Queralt C, Porta R, Cardenal F, Camps C, Majem M, Lopez-Vivanco G, Isla D, Provencio M, et al. (2009). Screening for epidermal growth factor receptor mutations in lung cancer. *N Engl J Med* **361**, 958–967.

[2] Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S-I, Watanabe H, Kurashina K, Hatanaka H, et al. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448**, 561–566.

[3] Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, et al. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *N Engl J Med* **350**, 2129–2139.

[4] Mok TS, Wu Y-L, Thongprasert S, Yang C-H, Chu D-T, Saijo N, Sunpaweravong P, Han B, Margono B, Ichinose Y, et al. (2009). Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* **361**, 947–957.

[5] Mitsudomi T, Morita S, Yatabe Y, Negoro S, Okamoto I, Tsurutani J, Seto T, Satouchi M, Tada H, Hirashima T, et al. (2010). Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial. *Lancet Oncol* **11**, 121–128.

[6] Janssen-Heijnen MLG and Coebergh J-WW (2001). Trends in incidence and prognosis of the histological subtypes of lung cancer in North America, Australia, New Zealand and Europe. *Lung Cancer* **31**, 123–137.

[7] Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113.

[8] Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin M-L, Beare D, Lau KW, Greenman C, et al. (2010). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190.

[9] Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, et al. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477.

[10] Rabbitts P, Douglas J, Daly M, Sundaresan V, Fox B, Haselton P, Wells F, Albertson D, Waters J, and Bergh J (1989). Frequency and extent of allelic loss in the short arm of chromosome 3 in nonsmall-cell lung cancer. *Genes Chromosomes Cancer* **1**, 10.

[11] Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G, et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81.

[12] Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, McKinley C, Egan P, Ross L, Hayward B, et al. (2010). Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res* **38**, e151.

[13] Hackenberg M, Sturm M, Langenberger D, Falcón-Pérez J, and Aransay A (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res* **37**, W68–W76.

[14] Langmead B, Trapnell C, Pop M, and Salzberg S (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.

[15] Complete Genomics Incorporated (2012). Available at: http://cgatools.sourceforge.net/docs/1.3.0/cgatools-methods.pdf. Accessed April 8, 2012.

[16] Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, and Berri S (2012). Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* **28**, 40–47.

[17] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. (2011). Ensembl 2011. *Nucleic Acids Res* **39**, D800–D806.

[18] Paşaniuc B, Zaitlen N, and Halperin E (2011). Accurate estimation of expression levels of homologous genes in RNA-seq experiments. *J Comput Biol* **18**, 459–468.

[19] Huang DW, Sherman BT, and Lempicki RA (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1–13.

[20] Lu M, Shi B, Wang J, Cao Q, and Cui Q (2010). TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs. *BMC Bioinformatics* **11**, 419.

[21] Ramirez RD, Sheridan S, Girard L, Sato M, Kim Y, Pollack J, Peyton M, Zou Y, Kurie JM, DiMaio JM, et al. (2004). Immortalization of human bronchial epithelial cells in the absence of viral oncoproteins. *Cancer Res* **64**, 9027–9034.

[22] Hyman E, Kauraniemi P, Hautaniemi S, Wolf M, Mousses S, Rozenblum E, Ringner M, Sauter G, Monni O, Elkahloun A, et al. (2002). Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res* **62**, 6240–6245.

[23] Järvinen A-K, Autio R, Kilpinen S, Saarela M, Leivo I, Grénman R, Mäkitie AA, and Monni O (2008). High-resolution copy number and gene expression microarray analyses of head and neck squamous cell carcinoma cell lines of tongue and larynx. *Genes Chromosomes Cancer* **47**, 500–509.

[24] Myllykangas S, Junnila S, Kokkola A, Autio R, Scheinin I, Kiviluoto T, Karjalainen-Lindsberg M-L, Hollmén J, Knuutila S, Puolakkainen P, et al. (2008). Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes. *Int J Cancer* **123**, 817–825.

[25] Balsara BR, Sonoda G, du Manoir S, Siegfried JM, Gabrielson E, and Testa JR (1997). Comparative genomic hybridization analysis detects frequent, often high-level, overrepresentation of DNA sequences at 3q, 5p, 7p, and 8q in human non-small cell lung carcinomas. *Cancer Res* **57**, 2116–2120.

[26] Danner BC, Hellms T, Jung K, Gunawan B, Didilis V, Füzesi L, and Schöndube FA (2011). Prognostic value of chromosomal imbalances in squamous cell carcinoma and adenocarcinoma of the lung. *Ann Thorac Surg* **92**, 1038–1043.

[27] Boelens MC, Kok K, van der Vlies P, van der Vries G, Sietsma H, Timens W, Postma DS, Groen HJM, and van den Berg A (2009). Genomic aberrations in squamous cell lung carcinoma related to lymph node or distant metastasis. *Lung Cancer* **66**, 372–378.

[28] Belvedere O, Berri S, Chalkley R, Conway C, Barbone F, Pisa F, Maclennan K, Daly C, Alsop M, Morgan J, et al. (2011). A computational index derived from whole-genome copy number analysis is a novel tool for prognosis in early stage lung squamous cell carcinoma. *Genomics* **99**, 18–24.

[29] Kim T-M, Yim S-H, Lee J-S, Kwon M-S, Ryu J-W, Kang H-M, Fiegler H, Carter NP, and Chung Y-J (2005). Genome-wide screening of genomic alterations and their clinicopathologic implications in non–small cell lung cancers. *Clin Cancer Res* **11**, 8235–8242.

[30] Yan W, Song L, Liang Q, and Fang Y (2005). Progression analysis of lung squamous cell carcinomas by comparative genomic hybridization. *Tumor Biol* **26**, 158–164.

[31] McCaughan F, Pole JCM, Bankier AT, Konfortov BA, Carroll B, Falzon M, Rabbitts TH, George PJ, Dear PH, and Rabbitts PH (2010). Progressive 3q amplification consistently targets SOX2 in preinvasive squamous lung cancer. *Am J Respir Crit Care Med* **182**, 83–91.

[32] Bass AJ, Watanabe H, Mermel CH, Yu S, Perner S, Verhaak RG, Kim SY, Wardwell L, Tamayo P, Gat-Viks I, et al. (2009). SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nat Genet* **41**, 1238–1242.

[33] Hussenet T, Dali S, Exinger J, Monga B, Jost B, Dembel D, Martinet N, Thibault C, Huelsken J, Brambilla E, et al. (2010). SOX2 is an oncogene activated by recurrent 3q26.3 amplifications in human lung squamous cell carcinomas. *PLoS One* **5**, e8960.

[34] Meehan M, Parthasarathi L, Moran N, Jefferies C, Foley N, Lazzari E, Murphy D, Ryan J, Ortiz B, Fabius AWM, et al. (2012). Protein tyrosine phosphatase receptor delta acts as a neuroblastoma tumor suppressor by destabilizing the aurora kinase a oncogene. *Mol Cancer* **11**, 6.

[35] Giefing M, Zemke N, Brauze D, Kostrzewska-Poczekaj M, Luczak M, Szaumkessel M, Pelinska K, Kiwerska K, Tönnies H, Grenman R, et al. (2011). High resolution ArrayCGH and expression profiling identifies PTPRD and PCDH17/PCH68 as tumor suppressor gene candidates in laryngeal squamous cell carcinoma. *Genes Chromosomes Cancer* **50**, 154–166.

[36] Kohno T, Otsuka A, Girard L, Sato M, Iwakawa R, Ogiwara H, Sanchez-Cespedes M, Minna JD, and Yokota J (2010). A catalog of genes homozygously deleted in human lung cancer and the candidacy of PTPRD as a tumor suppressor gene. *Genes Chromosomes Cancer* **49**, 342–352.

[37] Sclafani R and Schauer I (1996). Cell cycle control and cancer: lessons from lung cancer. *J Investig Dermatol Symp Proc* **1**, 5.

[38] Wiest JS, Franklin WA, Otstot JT, Forbey K, Varella-Garcia M, Rao K, Drabkin H, Gemmill R, Ahrent S, Sidransky D, et al. (1997). Identification of a novel region of homozygous deletion on chromosome 9p in squamous cell carcinoma of the lung: the location of a putative tumor suppressor gene. *Cancer Res* **57**, 1–6.

[39] Panani AD, Maliaga K, Babanaraki A, and Bellenis ION (2009). Numerical abnormalities of chromosome 9 and $p16^{CDKN2A}$ gene deletion detected by FISH in non-small cell lung cancer. *Anticancer Res* **29**, 4483–4487.

[40] Kraunz KS, Nelson HH, Lemos M, Godleski JJ, Wiencke JK, and Kelsey KT (2006). Homozygous deletion of $p16^{INK4a}$ and tobacco carcinogen exposure in nonsmall cell lung cancer. *Int J Cancer* **118**, 1364–1369.

[41] Elkiran E, Mar N, Aygen B, Gursu F, Karaoglu A, and Koca S (2007). Serum paraoxonase and arylesterase activities in patients with lung cancer in a Turkish population. *BMC Cancer* **7**, 48.

[42] Tymms MJ, Ng AY, Thomas RS, Schutte BC, Zhou J, Eyre HJ, Sutherland GR, Seth A, Rosenberg M, Papas T, et al. (1997). A novel epithelial-expressed ETS gene, *ELF3*: human and murine cDNA sequences, murine genomic organization, human mapping to 1q32.2 and expression in tissues and cancer. *Oncogene* **15**, 2449–2462.

[43] Oliver JR, Kushwah R, Wu J, Pan J, Cutz E, Yeger H, Waddell TK, and Hu J (2011). Elf3 plays a role in regulating bronchiolar epithelial repair kinetics following Clara cell-specific injury. *Lab Invest* **91**, 1514–1529.

[44] Hayashi S, Kumai T, Matsuda Y, Aoki N, Sato K, Kimura S, Kitada M, Tateno M, Celis E, and Kobayashi H (2011). Six-transmembrane epithelial antigen of the prostate and enhancer of zeste homolog 2 as immunotherapeutic targets for lung cancer. *J Transl Med* **9**, 191.

[45] Watanabe T, Miura T, Degawa Y, Fujita Y, Inoue M, Kawaguchi M, and Furihata C (2010). Comparison of lung cancer cell lines representing four histopathological subtypes with gene expression profiling using quantitative real-time PCR. *Cancer Cell Int* **10**, 2.

[46] Bauer M, Aust G, and Schumacher U (2004). Different transcriptional expression of KIAA1324 and its splicing variants in human carcinoma cell lines with different metastatic capacity. *Oncology Rep* **11**, 3.

[47] Deng L, Feng J, and Broaddus RR (2010). The novel estrogen-induced gene EIG121 regulates autophagy and promotes cell survival under stress. *Cell Death Dis* **1**, e32.

[48] Dompe N, Rivers CS, Li L, Cordes S, Schwickart M, Punnoose EA, Amler L, Seshagiri S, Tang J, Modrusan Z, et al. (2011). A whole-genome RNAi screen identifies an 8q22 gene cluster that inhibits death receptor-mediated apoptosis. *Proc Natl Acad Sci USA* **108**, E943–E951.

[49] Subramani D and Alahari S (2010). Integrin-mediated function of Rab GTPases in cancer progression. *Mol Cancer* **9**, 312.

[50] Dranoff G (2004). Cytokines in cancer pathogenesis and cancer therapy. *Nat Rev Cancer* **4**, 11–22.

[51] Fukuyama T, Ichiki Y, Yamada S, Shigematsu Y, Baba T, Nagata Y, Mizukami M, Sugaya M, Takenoyama M, Hanagiri T, et al. (2007). Cytokine production of lung cancer cell lines: correlation between their production and the inflammatory/immunological responses both *in vivo* and *in vitro*. *Cancer Sci* **98**, 1048–1054.

[52] Qi C, Zhu YT, Hu L, and Zhu Y-J (2009). Identification of Fat4 as a candidate tumor suppressor gene in breast cancers. *Int J Cancer* **124**, 793–798.

[53] Chen S, Lin C, Chen Y, Fang H, Cheng C, Chang C, Chen R, Tai H, Lee C, Chou M, et al. (2006). Aberrant promoter methylation of EDNRB in lung cancer in Taiwan. *Oncology Rep* **15**, 5.

[54] Ladanyi M and Pao W (2008). Lung adenocarcinoma: guiding EGFR-targeted therapy and beyond. *Mod Pathol* **21**, S16–S22.

[55] Lockwood WW, Wilson IM, Coe BP, Chari R, Pikor LA, Thu KL, Solis LM, Nunez MI, Behrens C, Yee J, et al. (2012). Divergent genomic and epigenomic landscapes of lung cancer subtypes underscore the selection of different oncogenic pathways during tumor development. *PLoS One* **7**, e37775.
[56] Hanahan D and Weinberg RA (2000). The hallmarks of cancer. *Cell* **100**, 57–70.
[57] Costa FF (2005). Non-coding RNAs: new players in eukaryotic biology. *Gene* **357**, 83–94.
[58] Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai M-C, Hung T, Argani P, Rinn JL, et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076.
[59] Marshall L and White RJ (2008). Non-coding RNA production by RNA polymerase III is implicated in cancer. *Nat Rev Cancer* **8**, 911–914.
[60] Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, Benjamin H, Shabes N, Tabak S, Levy A, et al. (2008). MicroRNAs accurately identify cancer tissue origin. *Nat Biotechnol* **26**, 462–469.
[61] Mattick J (2007). A new paradigm for developmental biology. *J Exp Biol* **210**, 1526–1547.
[62] Guttman M, Amit I, Garber M, French C, Lin M, Feldser D, Huarte M, Zuk O, Carey B, and Cassady J (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227.
[63] Liao J, Yu L, Mei Y, Guarnera M, Shen J, Li R, Liu Z, and Jiang F (2010). Small nucleolar RNA signatures as biomarkers for non-small-cell lung cancer. *Mol Cancer* **9**, 198.
[64] Ravi A, Gurtan AM, Kumar MS, Bhutkar A, Chin C, Lu V, Lees JA, Jacks T, and Sharp PA (2012). Proliferation and tumorigenesis of a murine sarcoma cell line in the absence of DICER1. *Cancer Cell* **21**, 848–855.
[65] Kumar MS, Erkeland SJ, Pester RE, Chen CY, Ebert MS, Sharp PA, and Jacks T (2008). Suppression of non-small cell lung tumor development by the *let-7* microRNA family. *Proc Natl Acad Sci USA* **105**, 3903–3908.
[66] Takamizawa J, Konishi H, Yanagisawa K, Tomida S, Osada H, Endoh H, Harano T, Yatabe Y, Nagino M, Nimura Y, et al. (2004). Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res* **64**, 3753–3756.
[67] Landi MT, Zhao Y, Rotunno M, Koshiol J, Liu H, Bergen AW, Rubagotti M, Goldstein AM, Linnoila I, Marincola FM, et al. (2010). MicroRNA expression differentiates histology and predicts survival of lung cancer. *Clin Cancer Res* **16**, 430–441.
[68] Bishop JA, Benjamin H, Cholakh H, Chajut A, Clark DP, and Westra WH (2010). Accurate classification of non–small cell lung carcinoma using a novel microRNA-based approach. *Clin Cancer Res* **16**, 610–619.
[69] Kohno T, Ichikawa H, Totoki Y, Yasuda K, Hiramoto M, Nammo T, Sakamoto H, Tsuta K, Furuta K, Shimada Y, et al. (2012). KIF5B-RET fusions in lung adenocarcinoma. *Nat Med* **18**, 375–377.
[70] Yin D-T, Wang L, Sun J, Yin F, Yan Q, Shen R-L, Gao J-X, and He G (2010). Homozygous deletion but not mutation of exons 5 and 8 of the fragile histidine triad (FHIT) gene is associated with features of differentiated thyroid carcinoma. *Ann Clin Lab Sci* **40**, 267–272.
[71] Ohta M, Inoue H, Cotticelli MG, Kastury K, Baffa R, Palazzo J, Siprashvili Z, Mori M, McCue P, Druck T, et al. (1996). The *FHIT* gene, spanning the chromosome 3p14.2 fragile site and renal carcinoma–associated t(3;8) breakpoint, is abnormal in digestive tract cancers. *Cell* **84**, 587–597.
[72] Mandai M, Konishi I, Kuroda H, Nanbu K, Matsushita K, Yura Y, Hamid AA, and Mori T (1998). Expression of abnormal transcripts of the *FHIT* (Fragile Histidine Triad) gene in ovarian carcinoma. *Eur J Cancer* **34**, 745–749.
[73] Prakash T, Sharma VK, Adati N, Ozawa R, Kumar N, Nishida Y, Fujikake T, Takeda T, and Taylor TD (2010). Expression of conjoined genes: another mechanism for gene regulation in eukaryotes. *PLoS One* **5**, e13284.
[74] Kim R, Kim A, Choi S-H, Kim D-S, Nam S-H, Kim D-W, Kim D-W, Kang A, Kim M-Y, Park K-H, et al. (2012). Novel mechanism of conjoined gene formation in the human genome. *Funct Integr Genomics* **12**, 45–61.
[75] Rackham O, Davies SMK, Shearwood A-MJ, Hamilton KL, Whelan J, and Filipovska A (2009). Pentatricopeptide repeat domain protein 1 lowers the levels of mitochondrial leucine tRNAs in cells. *Nucleic Acids Res* **37**, 5859–5867.
[76] Lemarie A and Grimm S (2011). Mitochondrial respiratory chain complexes: apoptosis sensors mutated in cancer? *Oncogene* **30**, 3985–4003.
[77] Faghihi MA and Wahlestedt C (2009). Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol* **10**, 637–643.
[78] Hu Q, Lu J-F, Luo R, Sen S, and Maity SN (2006). Inhibition of CBF/NF-Y mediated transcription activation arrests cells at G2/M phase and suppresses expression of genes activated at G2/M phase of the cell cycle. *Nucleic Acids Res* **34**, 6272–6285.
[79] Imbriano C, Gnesutta N, and Mantovani R (2012). The NF-Y/p53 liaison: well beyond repression. *Biochim Biophys Acta* **1825**, 131–139.
[80] Sasajima Y, Tanaka H, Miyake S, and Yuasa Y (2005). A novel EID family member, EID-3, inhibits differentiation and forms a homodimer or heterodimer with EID-2. *Biochem Biophys Res Commun* **333**, 969–975.
[81] Chishti AH, Kim AC, Marfatia SM, Lutchman M, Hanspal M, Jindal H, Liu S-C, Low PS, Rouleau GA, Mohandas N, et al. (1998). The FERM domain: a unique module involved in the linkage of cytoplasmic proteins to the membrane. *Trends Biochem Sci* **23**, 281–282.
[82] Haitina T, Lindblom J, Renström T, and Fredriksson R (2006). Fourteen novel human members of mitochondrial solute carrier family 25 (SLC25) widely expressed in the central nervous system. *Genomics* **88**, 779–790.
[83] Garnis C, Lockwood WW, Vucic E, Ge Y, Girard L, Minna JD, Gazdar AF, Lam S, MacAulay C, and Lam WL (2006). High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH. *Int J Cancer* **118**, 1556–1564.
[84] Qian J and Massion P (2008). Role of chromosome 3q amplification in lung cancer. *J Thorac Oncol* **3**, 4.
[85] Zabarovsky ER, Lerman MI, and Minna JD (2002). Tumor suppressor genes on chromosome 3p involved in the pathogenesis of lung and other cancers. *Oncogene* **21**, 21.
[86] Kang J, Lee SY, Lee SY, Kim YJ, Park JY, Kwon SJ, Na MJ, Lee EJ, Jeon HS, and Son JW (2012). MicroRNA-99b acts as a tumor suppressor in non-small cell lung cancer by directly targeting fibroblast growth factor receptor 3. *Exp Ther Med* **3**, 149–153.
[87] Visone R and Croce CM (2009). MiRNAs and cancer. *Am J Pathol* **174**, 1131–1138.
[88] Keller A, Backes C, Leidinger P, Kefer N, Boisguerin V, Barbacioru C, Vogel B, Matzas M, Huwer H, Katus HA, et al. (2011). Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients. *Mol BioSyst* **7**, 3187–3199.
[89] Volinia S, Calin GA, Liu C-G, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, et al. (2006). A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci USA* **103**, 2257–2261.
[90] The Cancer Genome Atlas Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525.
[91] Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A, et al. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107–1120.

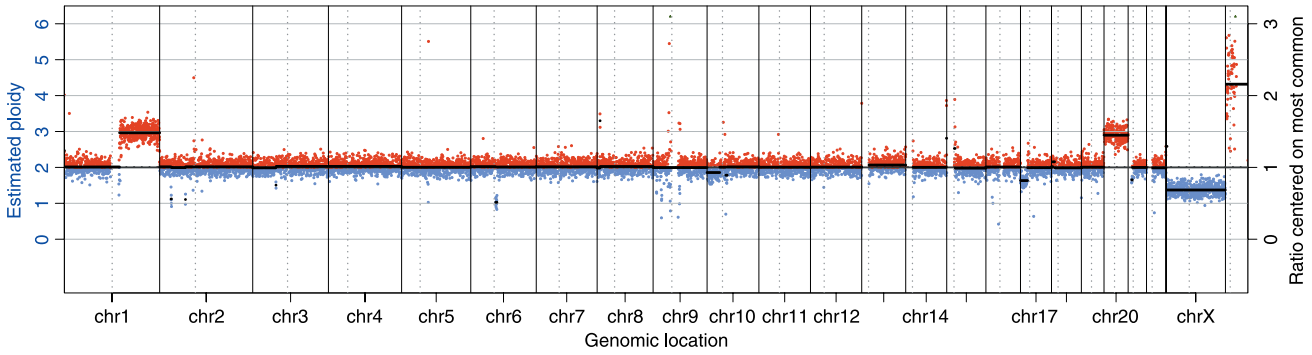

Figure W1. The karyogram for LIMM-NBE1. Estimated ploidy for LIMM-NBE1: a normal bronchial epithelium cell line that was created in-house by h-*TERT* transformation. The red dots show windows in which a copy number gain was inferred (normalized against pooled samples of normal tissue), whereas blue indicates regions of loss. The thick black line is the predicted ploidy.

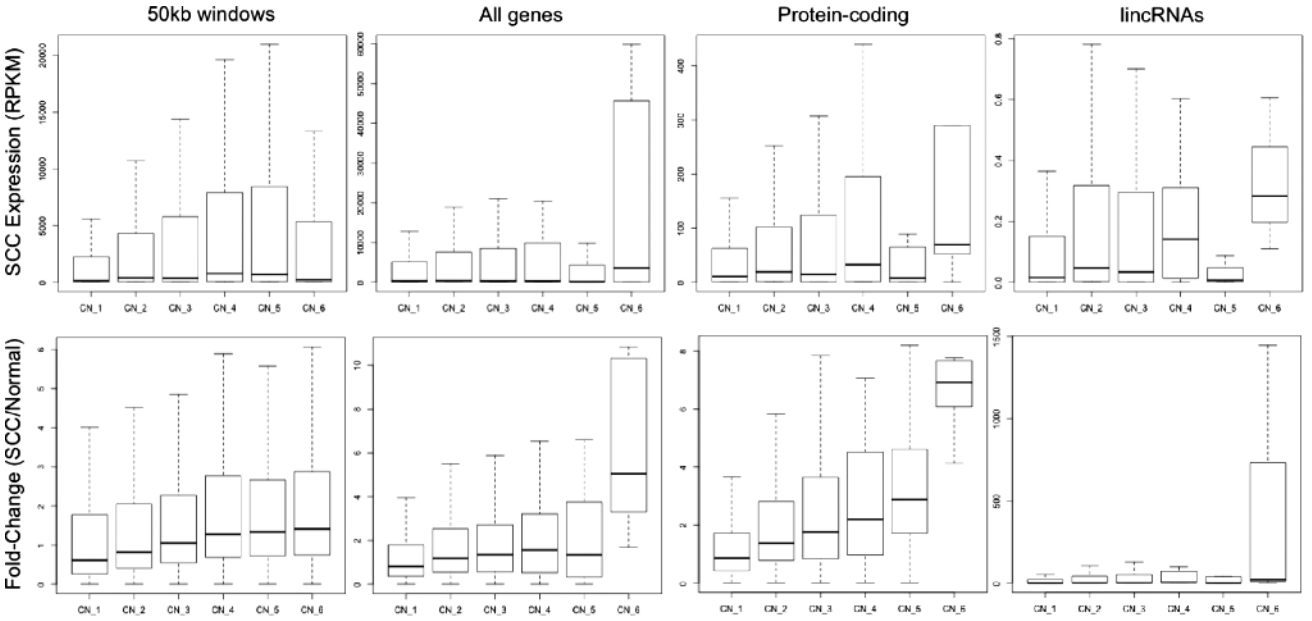

Figure W2. The effects of copy number on expression for different types of transcript or across 50-kb windows genome-wide. Boxplots showing the range of expression values in the tumor (top) or fold change in expression between normal and tumor (bottom) for different copy number, split by functional transcript class. Expression is quantified as RPKM. Ploidy is denoted as copy numbers 1 to 6 (CN_1–CN_6). SCC: squamous cell carcinoma.