Analysis of fusion transcripts indicates widespread deregulation of snoRNAs and their host genes in breast cancer

Helena Persson 1, Rolf Søkilde 1,2, Jari Häkkinen 1, Johan Vallon-Christersson 1, Felix Mitelman 3, Åke Borg 1,2,4, Mattias Höglund 1 and Carlos Rovira 1,2

1 Department of Clinical Sciences Lund, Oncology, Lund University Cancer Center, Lund, Sweden
2 BioCARE, Strategic Cancer Research Program, Lund, Sweden
3 Department of Laboratory Medicine, Clinical Genetics, Lund University, Skåne University Hospital, Lund, Sweden
4 CREATE Health, Strategic Centre for Translational Cancer Research, Lund, Sweden

Genomic rearrangements in cancer can join the sequences of two separate genes. Studies of such gene fusion events have mainly focused on identification of fusion proteins from the chimeric transcripts. We have previously investigated how fusions instead can affect the expression of intronic microRNA (miRNA) genes that are encoded within fusion gene partners. Here, we extend our analysis to small nucleolar RNAs (snoRNAs) that also are embedded within protein-coding or noncoding host genes. We found that snoRNA hosts are selectively enriched in fusion transcripts, like miRNA host genes, and that this enrichment is associated with all snoRNA classes. These structural changes may have functional consequences for the cell; proteins involved in the protein translation machinery are overrepresented among snoRNA host genes, a gene architecture assumed to be needed for closely coordinated expression of snoRNAs and host proteins. Our data indicate that this structure is frequently disrupted in cancer. We furthermore observed that snoRNA genes involved in fusions tend to associate with stronger promoters than the natural host, suggesting a mechanism that selects for snoRNA overexpression. In summary, we highlight a previously unexplored frequent structural change in cancer that affects important components of cellular physiology.

Introduction

Small nucleolar RNAs (snoRNAs) are one of the most abundant noncoding RNA classes.1 The human genome encodes several hundred snoRNA genes that are involved in a variety of cellular functions.2 They mainly guide posttranscriptional chemical modification of targeted RNAs, resulting in 2′-O-ribose methylation or pseudouridylation of ribosomal RNA (rRNA) or spliceosomal small nuclear RNA (snRNA). These modifications are important for correct folding and processing of rRNA, efficient ribosome assembly, accurate protein translation and regulation of alternative mRNA splicing. In addition, some snoRNAs may be processed into ~22 nt small RNAs that can be loaded into Argonaute complexes in a Drosha/DGCR8-independent, Dicer-dependent pathway for the production of miRNA-like molecules,3 thus regulating the expression of targeted protein genes without the mediation of any chemical modification.

SnoRNAs can be divided into two major subclasses depending on the presence of different evolutionarily conserved sequence elements; the C/D box class and H/ACA box class. C/D box snoRNAs guide the 2′-O-methyltransferase fibrillarin to RNA methylation and H/ACA box snoRNAs guide the pseudouridine synthase dyskerin to convert uracil into pseudouridine. Most snoRNAs localize to the nucleoli, but a special subtype of H/ACA snoRNAs called small Cajal
body-specific RNAs (scARNAs) localize to the Cajal bodies, nuclear organelles involved in RNA processing and modification of snRNAs. An additional type of H/ACA box RNA is produced in human cells from intron-encoded Alu elements and form a large H/ACA snoRNA subset called AluACA RNAs. A class of C/D box-like snoRNAs of unknown functional relevance was recently identified by photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) experiments with snoRNA-associated proteins in human cells.5

Some snoRNAs are transcribed from independent promoters, including the C/D box snoRNAs SNORD3A (U3), SNORD118 (U8) and SNORD13 (U13), but more than 90% of all human snoRNA genes reside inside host genes. The majority of these are protein-coding and the snoRNAs originate from debranching of spliced introns. Notably, genes hosting snoRNAs are often involved in nucleolar function, ribosome structure or protein synthesis. This structural bias is assumed to reflect an expression strategy that preserves a balance for critical components of the protein translation machinery.6

SnoRNAs were long regarded as static components of the housekeeping machinery of the cell, but recent studies have revealed that their plastic expression contributes to the dynamic regulation of essential macromolecular complexes such as ribosomes through control of rRNA modifications.7,8 Although altered expression of snoRNAs has been repeatedly observed in many cancers, the reason for their aberrant expression remains not completely understood. Their dysfunction could promote or, alternatively, suppress oncogenesis by modulating global protein synthesis, a hallmark of many cancers.9 The role of snoRNas in cancer may also extend beyond their primary function in ribosome biosynthesis through direct control of cellular pathways strongly associated with the disease. For example, SNORD126 modulates PI3K-AKT and K-Ras to regulate proliferation and apoptosis.10 Downregulation of snoRNAs has been shown to facilitate p53-dependent apoptosis of tumor cells,11 while reduced expression of a single H/ACA snoRNA, SNORA24, has been observed to promote liver cancer in cells expressing pro-proliferative signals.12 Furthermore, high expression of SNORA42 showed significant correlation with poor survival in lung cancer and deeper studies suggested that it may act as an oncogene.13

In a few cases, snoRNA expression has been reported to be affected by genomic rearrangements. Genetic studies showed that genomic alterations that deregulate the expression of C/D box snoRNAs encoded in the 15q11-q13 locus affect the serotonin receptor HTR2C in patients with Prader-Willi syndrome.14,15 Evidence is now emerging that implicates snoRNAs in cancer and genomic aberrations that alter snoRNA expression have been associated with tumorigenesis. For example, SNORD50A and SNORD50B map to 6q14-q16, a region frequently deleted in prostate and breast cancer.16 These snoRNAs have been observed to bind directly to K-Ras and their loss increases the amount of GTP-bound, active K-Ras and hyperactivates Ras-ERK1/ERK2 signaling.17 This deletion has also been found associated with a chromosomal breakpoint in a human B-cell lymphoma.18 Furthermore, the (4;14) chromosomal translocation leads to overexpression of SCARNA22 (also known as ACA11) in multiple myeloma, resulting in increased oxidative stress and cell proliferation.19

We recently reported the existence of common fusion transcripts in cancer that involve miRNA host genes and affect the expression of the included miRNAs.20 Since the only requirement for miRNA expression is transcription, the protein-coding potential of a host gene fusion is unimportant. We therefore pooled all fusion genes involving the same miRNA host as 3’ partner into groups of so-called miRNA-convergent fusions. Considering that snoRNAs are also embedded in host genes we applied the same logic for snoRNAs searching for a cause for their transcriptional general deregulation in cancer. We found that, as for miRNA host genes, snoRNA-encoding hosts are significantly enriched among fusion genes and that these rearrangements are common events that change the transcriptional control of the embedded snoRNAs. These rearrangements can therefore disrupt the transcriptional control of snoRNAs as well as the balance between host genes and their nested snoRNAs. In summary, we present the first global study of how rearrangements in the form of gene fusions affect the expression of central components of the human cell such as snoRNAs and their hosts in cancer.

Materials and Methods

Patient material

The study was conducted in accordance with the Declaration of Helsinki. Tumor material is obtained by SCAN-B (ClinicalTrials.gov NCT02306096),21,22 an ongoing multicenter prospective study that until now have included more than 13,000 breast cancer patients. The study has been approved by the Regional Ethical Review Board of Lund (registration numbers 2007/155, 2009/658, 2009/659, 2014/8), the county governmental biobank center and the Swedish Data Inspection group (registration number 364-2010). Written information was given by trained health professionals and all patients provided written informed consent.

Strand-specific mRNA sequencing

RNA sequencing was performed as described.21 Briefly, mRNA was isolated using the Dynabeads mRNA DIRECT Kit. Isolated mRNA was subjected to zinc-mediated fragmentation. Fragmented mRNA was purified on Zymo-Spin to preserve fragmented mRNA >16 nucleotides (Zymo Research). First-strand cDNA synthesis was performed with SuperScript II reverse transcriptase using random hexamers. Second strand synthesis was performed with DNA Pol I and RNase H before clean-up on Zymo-Spin 1-96 plates and low EtOH (~57%) to preserve DNA > 80 bp. The cDNA was end-repaired and A-tailed and adapter ligation was performed with a 1:5 dilution of adapters from TruSeq DNA LT Sample Prep Kit A and B (Illumina, San Diego, CA). Adapter-ligated cDNA was size-selected on a KingFisher Flex Magnetic Particle Processor (Thermo Fisher Scientific, Waltham, MA).
using polyethylene glycol MW 8000 and carboxylic acid (CA)-beads to remove fragments <200 bp before digestion of the second cDNA strand with uracil-DNA glycosylase. Single-stranded cDNA was amplified by PCR. Fragments >700 bp and <200 bp were excluded by two cycles of size selection. Libraries were pooled and sequenced by 2 x 50 bp paired-end sequencing on a HiSeq 2000 (Illumina).

Small RNA sequencing
Small RNA sequencing was performed as in Persson et al.23 Total RNA was extracted from tumor homogenates using TRIzol LS (Thermo Fisher Scientific). Then, 500 ng total RNA and 20 pmol of pre-adenylated 3' adapter were ligated. 20 pmol of reverse transcription primer was added and annealed by incubation. Then, 20 pmol of 5' adapter was denatured separately then ligated in 25 µl reactions. Reverse transcription was performed in 40 µl reactions with 200 U ProtoScript II reverse transcriptase. Then, 20 µl of cDNA was used for PCR amplification in 50 µl reactions 250 nM each of indexed forward and reverse primers. Pooled libraries were sequenced on a NextSeq 500 with High Output v2 75 cycle-kits (Illumina). Sequences were demultiplexed using Picard and aligned against hg38 using Novoalign with settings -a TGGAATTCTCGG GTGCCAAGG -l 14 -h -l -l -l 90 -g 50 -x 15 -o SAM -o FullNW -r All 51 -s 51. The expression of snoRNAs and miRNAs was calculated using custom Perl scripts, normalized using the R package edgeR version 3.22.124 and exported as counts per million reads (cpm).

Detection of fusion transcripts in mRNA-Seq data
i. FusionCatcher version 0.99.7c was used to search the mRNA-Seq data for chimeric transcripts with options changed from default values as follows: pairs-fusion 2,2,2,2,2.
ii. reads-fusion 1,1,1,1,1.
iii. anchor-fusion 13,14,14,14,14.
iv. anchor-fusion2 22.
v. mismatches-psl 4.
vi. skip-filter-psl True.

Three additional aligners (BLAT, STAR and Bowtie 2) were used. Candidate fusion transcripts with the following descriptions were excluded from further analysis: 1000genomes, 1 K < gap < 10 K, adjacent, ambiguous, duplicates, ensemble_partially_overlapping, gap < 1 K, gencode_fully_overlapping, gencode_partially_overlapping, gencode_same_strand_overlapping, healthy, m0, multi, non_cancer_tissues, non_tumor_cells, refseq_partially_overlapping, tcga-normal, ucsc_partially_overlapping, banned, bodymap2, cagg, conjoing, cta_gene, ctb_gene, ctc_gene, ctd_gene, distance1000bp, ensemble_fully_overlapping, ensemble_same_strand_overlapping, gtx, hpa, mt, pair_pseudo_genes, paralogs, readthrough, refseq_fully_overlapping, refseq_same_strand_overlapping, rp_gene, rp11_gene, rna, similar_reads, similar_symbols, ucsc_fully_overlapping, ucsc_same_strand_overlapping.

Identification of host genes and snoRNAs in fusions
Coordinates from snoRNA atlas2 were compared to gene annotation from GENCODE25 release27 to identify overlapping genes as hosts or if no host was identified, genes located within 2 kb upstream of a snoRNA. Breakpoints from FusionCatcher in fusion transcripts involving snoRNA hosts were then compared to snoRNA coordinates to identify 5' and 3' host gene fusions that include and exclude snoRNAs. Primary miRNA coordinates from miRBase release 2226 were used to identify miRNA host genes and miRNAs in fusion transcripts in a similar manner. FusionCatcher classification was used to divide fusion transcripts into the three classes “in-frame,” “out-of-frame” or “other” (for all other categories). For TCGA samples, fusion transcript breakpoints from Hu et al.27 were converted to the hg38 genome assembly using UCSC LiftOver and combined with Entrez annotation to determine the ends of partner genes. Since no breakpoint coordinates were available for the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer,28 snoRNA host genes in fusions were identified through matching by gene symbol.

Statistical analysis
Logistic regression to model the log odds of a gene being involved in fusion events was performed with the two models fusion ~ snoRNA host gene + log2(size) and fusion ~ snoRNA host gene + log2(size) + snoRNA host gene log2(size) for protein-coding genes expressed in the mRNA sequencing data. For comparison, RepeaMask annotation was obtained from the UCSC Table Browser and analyzed using custom Perl scripts to identify genetic elements that occur within protein-coding genes at a rate similar to that of snoRNAs. GO enrichment analyses were done by Gene Set Enrichment Analysis (GSEA) using clusterProfiler30 with Benjamini-Hochberg correction to control the false discovery rate. Student’s t-test was used to test for differences in the log2-transformed expression in fpkm of 5 partners as a group for all snoRNAs and for separate snoRNA classes. Spearman rank correlation was used to compare the log2-transformed expression of snoRNAs in cpm (counts per million reads) and host genes in fpkm. The R package limma version 3.36.531 was used for differential expression analysis of host genes using the function eBayes with trend = TRUE for log2-transformed fpkm values. The R package edgeR version 3.22.124 was used for snoRNA differential expression analysis. Fisher’s exact test was used to calculate the association between snoRNA fusion status and clinical parameters with Benjamini–Hochberg correction to control the false discovery rate. All statistical analyses were performed in R.

Data availability
Due to Swedish law, the patient consent, and the risk that the sequence data contain person-identifiable information and hereditary mutations, we cannot deposit the raw sequence data in a repository. The datasets supporting the conclusions of this article are available in the NCBI Gene Expression
Omnibus (GEO) repository, accession numbers GSE96058 and GSE131599, and from the corresponding author upon request. Supporting Information Table S1 contains a list of matched tumor sample identifiers for the two GEO datasets.

Results
snoRNAs and their host genes
We started our analysis by studying the distribution of snoRNAs and their host genes in the human genome using the full set of 1,117 snoRNAs and snoRNA-related RNAs cataloged in the snoRNA Atlas. Overlapping candidate host genes could be assigned to 1,043 (93%) of the snoRNAs and genes located less than 2 kb upstream were assigned as hosts for an additional 9 (1%; Supporting Information Table S2). Among the assigned host genes, 723 (91%) are protein-coding and 74 (9%) are noncoding genes or pseudogenes (Supporting Information Table S3). Only 65 (6%) of all snoRNAs lacked overlapping transcript annotation and are presumably transcribed from independent promoters. The distribution varies between snoRNA classes; while 6 and 11% of the C/D- and H/ACA-box snoRNAs were identified as independent transcriptional units, all sno-IncRNAs and scaRNA were localized inside another gene (see Supporting Information Table S2). Multiple snoRNAs of different classes were present in 43 host genes. Since most miRNAs are also intronically encoded we searched our annotated snoRNA host genes for the presence of miRNAs. Among all host genes 10% also contained miRNAs, demonstrating the common polyfunctionality of genetic loci that act as expression modules to encode proteins, regulate translation and exert post-transcriptional control of gene expression (Supporting Information Tables S2 and S3).

Detection of snoRNA hosts involved in fusion genes
To study the frequency of fusion transcripts that could affect the expression of snoRNA hosts, we used FusionCatcher to detect candidate fusions in RNA sequencing (RNA-Seq) data from a set of 1,553 breast tumors. The default settings in FusionCatcher exclude a number of noncoding transcripts from further analysis, among them at least 22 host genes containing 130 snoRNAs. We therefore changed several parameters from the default values (see Materials and Methods for a detailed description) and identified a total of 215,416 fusion transcripts (not limited to unique transcripts or genes) across the 1,545 tumors that had fusion transcripts. Then, 14,684 (7%) of these fusions involved host genes and their encoded snoRNAs. Among these, 8,636 fusion transcripts included one or more snoRNAs within the 5’ partner gene while 6,388 fusions involved snoRNAs located in the 3’ partner (Supporting Information Table S4). For the latter cases, where the snoRNA is located within the 3’ partner gene and downstream of the breakpoint, the transcriptional control will be altered also for the included snoRNA by the introduction of a new promoter. A small number of fusion transcripts (340) included snoRNA hosts on both sides of the breakpoint. Only 25,987 (12%) of all detected fusion transcripts were classified as in-frame fusions and 7% of these involved snoRNA hosts. The vast majority (87%) of fusions containing snoRNA hosts were classified as out-of-frame, truncated or joining nonprotein coding regions (UTR, intronic, intergenic, etc.; Supporting Information Table S4). Detailed information about snoRNAs, host genes and identified fusion transcripts is included in Supporting Information Table S5.

snoRNA host genes are overrepresented among fusion transcripts
We have previously reported that miRNA host genes are overrepresented as fusion partners, suggesting that tumor cells develop under a positive selection for alterations that affect miRNA expression. We investigated if a similar enrichment could be observed for snoRNA host gene fusions. Strikingly, we found that 780 (74%) of all snoRNAs from the snoRNA Atlas were found in fusions in at least one tumor. Furthermore, 622 (59%) were included in 3’ fusions involving their host genes in at least one tumor, indicating that the transcriptional control of snoRNA expression is frequently altered by an aberration. We therefore applied a logistic regression model to assess the relationship between the probability of a gene being involved in a fusion transcript and its status as snoRNA host, taking also gene size into consideration. We noted that noncoding host genes were expressed at low levels, which would make identification of fusion transcripts difficult. Since only a minority of the host genes is noncoding, we decided to limit our analysis to genes expressed in our breast tumor set and annotated by GENCODE as protein-coding. The sno-IncRNAs were also excluded because of their small number (n = 11). Modeled fusion probabilities for recurrent fusions as a function of snoRNA host gene status and gene size for all snoRNAs are shown in Figure 1. Host genes were indeed significantly overrepresented among fusion gene partners for all snoRNAs, as well as for the individual snoRNA classes except scaRNAs, the smallest group with only 25 host genes (Table 1 and Fig. 1a). Although the difference was not significant for scaRNA hosts, the fusion probability was consistently higher compared to the background gene set and the probability increased with gene size. Enrichment of Alu/ACA hosts was low but still significant and also here the fusion probability increased with gene size (contrary to the C/D- and H/ACA-box host genes). In contrast, no enrichment was observed when a number of other sequence elements found at similar frequencies in the introns of protein-coding genes were analyzed (Fig. 1b). The control sequence elements were selected from RepeatMasker annotation and included the simple repeats (AAAT)n and (TC)n, the SINE AluSg4, the DNA repeat Tigger15a, and the long terminal repeats (LTR) MLT1D and THE1B. As noted before, some snoRNA host genes also encode miRNAs. Since we have previously reported that miRNA host genes are enriched among fusion gene partners, we wanted to exclude the possibility that the observed
Figure 1. (a) The host genes of snoRNAs are significantly more likely to be partners of fusion transcripts than nonhosts, both for all snoRNAs and for individual snoRNA classes except scaRNAs. (b) No enrichment was observed for the hosts of a number of randomly chosen genetic elements that occur at similar frequencies in the introns of protein-coding genes compared to nonhost genes. (c) The host genes of snoRNAs and miRNAs are both overrepresented among fusion gene partners. Consequently, genes that are hosts to both miRNAs and snoRNAs have an increased probability of being involved in fusion transcripts compared to snoRNA-only host genes. Fusion probabilities were modeled using logistic regression based on host gene status, gene size, and the interaction between the two. In all panels, solid lines correspond to host genes and dashed lines to nonhosts.
overrepresentation of snoRNAs in fusions was a result of selection for the presence of miRNAs. As shown in Figure 1c, the presence of miRNAs in snoRNA host genes has an additive effect on the overrepresentation of hosts among fusion gene partners, but snoRNA host genes were still significantly enriched when we excluded genes that also encode miRNAs (Table 1).

It has been estimated that a significant fraction of intronic miRNAs target their host transcripts. A fusion transcript that encodes a miRNA within the 3’ partner gene would also retain any miRNA target sites in the 3’ UTR, thereby leading to down-regulation of the newly created protein. We have previously reported that in-frame fusions that create chimeric proteins are depleted among fusion genes that also encode a miRNA in the 3’ partner. To the best of our knowledge, snoRNAs do not regulate protein production from their hosts and, as expected, there was no depletion of in-frame fusions among 3’ snoRNA—including fusion transcripts, with the exception of SNORD-like snoRNAs (p = 0.042, Kolmogorov–Smirnov test). In-frame fraction distributions are shown in Supporting Information Figure S1.

### Fusions affecting snoRNA hosts are frequent events in breast cancer

Out of 1,545 tumor samples with fusions, 1,155 (75%) carried at least one fusion involving a snoRNA host either as 5’ or 3’ partner. In our previous work on miRNA host genes, we suggested a modification of the concept of “recurrent fusions” from a transcriptional standpoint. Different fusion gene partners may involve the same 3’ small RNA host gene and, although replacement of the 5’ part of the host will change the upstream control units, different fusions may still have the same transcriptional effect on the intronically encoded small RNA, even if they are out-of-frame. We refer to this as “recurrent convergent fusions”. Regardless of their protein-coding potential, all of these gene fusions have the potential to alter the expression of any snoRNA encoded downstream of the breakpoint. A total of 53% (821) of all samples had at least one fusion with a snoRNA host as 3’ partner. By applying the concept of convergent fusions for estimating recurrence, we found that 353 (34%) of all snoRNA genes had recurrent 3’ fusions in at least three tumors. Fusion transcripts involving snoRNA host genes as 5’ partners were found in at least one tumor for 697 (66%) of all snoRNAs and were recurrent for 376 (36%) of snoRNAs.

Some snoRNA hosts are particularly promiscuous as fusion partners: At the top of this list with 297 different 5’ partners across all tumors is ZL23/snoID_0714, a snoRNA-like gene encoded inside the ERBB2 locus, a genomic region that is especially unstable in breast cancer. Whether snoRNA-like genes are really functional snoRNA genes is however a matter of debate. The second from the top are SNORA31 and SNORA31B, two snoRNAs encoded inside TPT1 in chromosome arm 13q. We found that as much as 8% of all tumors had fusions involving these two snoRNAs (Supporting Information Figs. S2 and S3). The estimated position of the breakpoint in 90 of these fusions is located between the two snoRNA disrupting their genomic linkage. Only 14 fusions involving TPT1 have been reported in the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. They have been found in multiple myeloma, breast, lung, kidney, stomach, thyroid, uterus and prostate cancer. However, the presence of SNORA31 and SNORA31B inside these fusions was never highlighted. Supporting Information Table S6 lists the snoRNAs whose host genes have the largest number of different 5’ partners.

Multiple snoRNA host fusions were also found to co-exist in the same tumor sample. Ten percent of all tumors had five or more different snoRNA host genes as 3’ fusion partners. In one single sample, 126 different snoRNA fusions were identified, involving 11% of all known snoRNAs. These results indicate that, just as for miRNAs, it may be changes to the

### Table 1. Odds ratios, confidence intervals and p values for overrepresentation of snoRNA host genes in fusion transcripts by snoRNA type for all fusions and recurrent fusions

| Class                        | All fusions | Recurrent fusions |
|------------------------------|-------------|-------------------|
|                              | Odds ratio  | 95% CI p          | Odds ratio  | 95% CI p          |
| Any snoRNA                   | 3.54        | 2.83–4.42 <2e-16   | 2.56        | 2.18–3.01 <2e-16   |
| AluACA                       | 2.59        | 1.93–3.47 1.70e–10 | 1.95        | 1.56–2.42 3.03e–09 |
| AluACA excl miRNA hosts      | 2.48        | 1.82–3.37 7.31e–09 | 1.85        | 1.47–2.34 2.36e–07 |
| CD-box                       | 4.91        | 3.21–7.52 2.39e–13 | 3.34        | 2.50–4.47 5.23e–16 |
| CD-box excl miRNA hosts      | 4.62        | 2.90–7.36 1.31e–10 | 3.34        | 2.41–4.64 5.68e–13 |
| H/ACA-box                    | 4.42        | 2.58–7.57 6.37e–08 | 4.06        | 2.72–6.07 8.03e–12 |
| H/ACA-box excl miRNA hosts   | 3.59        | 2.00–6.44 1.81e–05 | 3.40        | 2.16–5.34 1.12e–07 |
| CD- and H/ACA-box excl miRNA hosts | 9.89 | 1.24–78.7 3.04e–02 | 10.53      | 2.24–49.5 2.85e–03 |
| scaRNA                       | 2.37        | 0.88–6.38 8.82e–02 | 1.81        | 0.82–3.98 1.43e–01 |
| scaRNA excl miRNA hosts      | 1.85        | 0.66–5.15 2.40e–01 | 1.40        | 0.58–3.41 4.54e–01 |
| SNORD-like                    | 13.7        | 4.29–43.7 9.94e–06 | 3.53        | 2.14–5.82 7.41e–07 |
| SNORD-like excl miRNA hosts  | 10.2        | 3.17–33.0 1.01e–04 | 2.59        | 1.49–6.50 7.71e–04 |

Only protein-coding host genes expressed in the set of 1,545 breast tumors were included in the analysis and sno-IncRNA hosts were excluded because of their small number (n = 6). n.s. = not significant. Recurrent fusions = fusions found in at least three tumors.
transcriptional control of intron-encoded snoRNAs, rather than the identity of the fused partner, that is under selective pressure.

Disruption of the modular expression of snoRNAs and their hosts

Intronic snoRNAs are often functionally associated with their host genes through participation in related cellular pathways. Notably, snoRNA hosts are enriched in genes associated with ribosome biogenesis, components of the translation machinery and members of the terminal oligo-pyrimidine (TOP) family of genes encoding translation-related proteins.\textsuperscript{1,36,37} It is assumed that this modular architecture places both protein and snoRNA under coordinated transcriptional control to ensure quick expression in response to environmental changes, during development and in disease. However, rearrangements leading to fusions involving snoRNA host genes would disrupt this expression strategy. To investigate the consequences of this change we compared Gene Ontology (GO) enrichment analyses of snoRNA hosts and the 5' partners of snoRNA host genes with recurrent fusions as 3' partners. As previously observed, snoRNA hosts are enriched for genes involved in protein translation and translation-related processes such as nonsense-mediated decay, NMD (Table 2). In contrast, the 5' translation and translation-related processes such as nonsense-snoRNA hosts are enriched for genes involved in protein trans-snoRNA-including fusions than for 3' partners of snoRNA-including fusions. It is assumed that this modular architecture places both protein and snoRNA under coordinated transcriptional control to ensure quick expression in response to environmental changes, during development and in disease. However, rearrangements leading to fusions involving snoRNA host genes would disrupt this expression strategy. To investigate the consequences of this change we compared Gene Ontology (GO) enrichment analyses of snoRNA hosts and the 5' partners of snoRNA host genes with recurrent fusions as 3' partners. As previously observed, snoRNA hosts are enriched for genes involved in protein translation and translation-related processes such as nonsense-mediated decay, NMD (Table 2). In contrast, the 5' fusion partners of snoRNA hosts replace this transcriptional control with genes involved in cell adhesion and extracellular structure organization (Table 2). The results shown are valid for all snoRNAs in snoRNA Atlas except for sno-IncRNAs. CD- and H/ACA-box snoRNAs showed similar enrichment profiles in individual analyses, while the remaining snoRNA classes had little or no significant GO term enrichment.

Fusions are associated with upregulation of snoRNA host gene expression

Interestingly, the 5' partner genes had significantly higher expression for fusions involving snoRNA hosts as 3' partners than for nonhost 3' partners ($p < 2.20 \times 10^{-16}$ for all snoRNAs and each individual snoRNA class, Student's $t$-test). Furthermore, the expression of 5' partners was also significantly higher for 3' snoRNA-including fusions than for 3' snoRNA-excluding fusions for all snoRNA Atlas entries, AluACA, and SNORD-like snoRNAs ($p = 0.040, 7.95 \times 10^{-3}$ and $1.31 \times 10^{-4}$, respectively, Student's $t$-test). The distributions of 5' partner expression for different fusion types and snoRNA classes are shown in Figure 2.

The high expression of 5' partners in fusions involving snoRNA host genes as 3' partners implies that a selective force may increase the expression of specific snoRNAs. We had data from small RNA sequencing (small RNA-Seq) available for 186 of the 1,545 samples\textsuperscript{20,38} and a summary of the snoRNA composition is shown in Supporting Information Table S7. Due to the relatively low number of samples with associated small RNA sequencing data, the set would likely be underpowered for differential expression analysis. To investigate the possibility to instead use RNA-seq data as a proxy, we compared the estimated snoRNA expression from small RNA sequencing data with host gene expression from RNA-Seq (see Supporting Information Fig. S4). Based on Spearman rank correlation 72% of snoRNA-host gene pairs were significantly positively correlated (Benjamini–Hochberg FDR < 0.05). We therefore decided to use host genes as a surrogate for snoRNAs in a differential expression analysis. In total, 79 host genes containing 101 snoRNAs were significantly upregulated in samples with 3' snoRNA-including fusions versus samples with no fusions (Benjamini–Hochberg FDR < 0.05, see Supporting Information Table S8). Seven of these host genes also had significantly higher expression in samples with 3' snoRNA-including fusions compared to 3' snoRNA-excluding fusions, but note that this analysis was limited to host genes that have both types of fusions. Only two host genes containing three snoRNAs were significantly downregulated in samples with 3' snoRNA-including fusions versus samples with no fusions. We then performed a differential expression analysis

| GO ID     | Description                                      | q       |
|-----------|--------------------------------------------------|---------|
| GO:0000184| Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay | 1.20E–18 |
| GO:0022626| Cytosolic ribosome                                | 4.42E–17 |
| GO:0072599| Establishment of protein localization to endoplasmic reticulum | 6.60E–17 |
| GO:0006614| SRP-dependent cotranslational protein targeting to membrane | 6.60E–17 |
| GO:000956 | Nuclear-transcribed mRNA catabolic process       | 1.20E–16 |
| GO:0006613| Cotranslational protein targeting to membrane     | 2.20E–16 |
| GO:0044391| Ribosomal subunit                                 | 5.11E–15 |
| GO:0044445| Cytosolic part                                    | 1.08E–14 |
| GO:0005840| Ribosome                                         | 4.15E–14 |
| GO:0003735| Structural constituent of ribosome                | 1.64E–13 |

$^5$ partners of snoRNA host gene fusions

Benjamini–Hochberg correction was used to control the false discovery rate.

\textsuperscript{186} of the 1,545 samples\textsuperscript{20,38} and a summary of the snoRNA composition is shown in Supporting Information Table S7. Due to the relatively low number of samples with associated small RNA sequencing data, the set would likely be underpowered for differential expression analysis. To investigate the possibility to instead use RNA-seq data as a proxy, we compared the estimated snoRNA expression from small RNA sequencing data with host gene expression from RNA-Seq (see Supporting Information Fig. S4). Based on Spearman rank correlation 72% of snoRNA-host gene pairs were significantly positively correlated (Benjamini–Hochberg FDR < 0.05). We therefore decided to use host genes as a surrogate for snoRNAs in a differential expression analysis. In total, 79 host genes containing 101 snoRNAs were significantly upregulated in samples with 3' snoRNA-including fusions versus samples with no fusions (Benjamini–Hochberg FDR < 0.05, see Supporting Information Table S8). Seven of these host genes also had significantly higher expression in samples with 3' snoRNA-including fusions compared to 3' snoRNA-excluding fusions, but note that this analysis was limited to host genes that have both types of fusions. Only two host genes containing three snoRNAs were significantly downregulated in samples with 3' snoRNA-including fusions versus samples with no fusions. We then performed a differential expression analysis

Int. J. Cancer: 146, 3343–3353 (2020) © 2020 The Authors. International Journal of Cancer published by John Wiley & Sons Ltd on behalf of UICC
based on small RNA-Seq data for the snoRNAs that had a differentially expressed host gene, at least two sequenced samples with snoRNA-including fusions, and were detected in the small RNA-Seq data. Five out of 56 tested snoRNAs were significantly differentially expressed at a Benjamini–Hochberg FDR < 0.05 (Table 3).

Finally, we investigated if fusions involving any of the 81 differentially expressed host genes were associated with specific tumor characteristics including estrogen receptor α (ESR1) and HER2 (ERBB2) status as determined by pathology examination, as well as molecular subtype derived from the RNA-Seq data. 3′ snoRNA-including fusions involving three host genes located on chromosome arm 17q were significantly associated with HER2-status; ERBB2 itself (ZL23/snoID_0714), MED24 (SNORD124/snoID_0115) and USP32 (SCARNA20/snoID_0592). The corresponding odds ratios were 28.2, 9.0 and 11.6 with $p < 2.20 \times 10^{-16}$, $p = 1.91 \times 10^{-4}$ and $3.13 \times 10^{-2}$, respectively (Benjamini–Hochberg-adjusted Fisher’s exact test). 3′ snoRNA-including fusions involving ERBB2 were also significantly associated with the HER2-enriched subtype (odds ratio 19.3 and $p < 2.20 \times 10^{-16}$) and ER-negativity (odds ratio 0.32 and $p = 1.01 \times 10^{-5}$).

Table 3. Log2 fold change and adjusted $p$ values for snoRNAs differentially expressed in breast tumors with 3′ snoRNA-including fusions versus tumors with no fusions

| snoRNA Host gene(s) | log2fc 3′ incl vs. no fusion | adj 3′ incl vs. no fusion | snoRNA Atlas targets |
|---------------------|-----------------------------|--------------------------|----------------------|
| ZL23/snoID_0714     | ERBB2/ENSG00000141736       | 3.13                     | 1.15E–17             |
| SNORD124/snoID_0115 | MED24/ENSG000000008838      | 1.77                     | 1.82E–05             |
| SNORA30/snoID_0490  | AC106886.5/ENSG000000282034, SRCAP/ENSG00000080603 | 2.85                   | 2.36E–02             |
| SNORD134/snoID_0623 | FASN/ENSG00000169710        | 1.96                     | 2.36E–02             |
| SNORA11/snoID_0529  | MAGED2/ENSG00000102316      | 0.92                     | 2.74E–02             |

Only snoRNAs in differentially expressed host genes were analyzed. Predicted targets listed in snoRNA Atlas were included. 3′ snoRNA-including fusions were defined as fusion transcripts involving the host gene as 3′ partner with the breakpoint upstream of the snoRNA.
Discussion
Understanding the function of novel transcribed sequences created through the fusion of distinct genes is important for both cancer diagnosis and treatment. The study of these fusions has almost exclusively focused on oncogenic hybrid proteins and in consequence, only chimeric transcripts with preserved protein-coding capacity have been selected for further studies. However, advances in genomic sciences have shown that a large portion of the protein-coding genome also encodes small noncoding RNAs in their introns and that these are cotranscribed with their hosts. Hence, rearrangements that affect the hosts will also affect the embedded genes. We recently showed that intronically encoded miRNAs are commonly affected by genomic rearrangements and those miRNA hosts are selectively enriched among fusion transcripts. No selection was detected for the proteins encoded by the host genes, suggesting that it is the encoded miRNAs, rather than the hosts, that provide a functional advantage for the tumor cell. In the present study, we wanted to investigate if this was true also for snoRNA genes. As for miRNAs, the study of gene fusions in cancer has completely disregarded the existence of snoRNAs inside the introns of fusion gene partners. Since the focus of fusion gene analysis typically is preserved protein-coding capacity of the fusion transcripts, analysis pipelines often discard most fusion transcripts that could affect transcription of nested snoRNAs from further analysis. Fusions affecting snoRNA host genes are therefore rare in public databases, an assumption confirmed by a comparison of our results with previously reported data.

Here we found that 780 snoRNA genes, 74% of all snoRNAs identified in human cells, were part of a fusion transcript in at least one breast tumor sample. However, only 15% of all snoRNAs were found among fusion transcripts previously reported in breast cancer samples from the community resource project TCGA. When expanding the search to all snoRNA fusions regardless of tumor tissue type, no more than half of the fusions found here were reported for the TCGA data. The numbers are of course even lower for fusions recurrent in at least three tumor samples: Only 1% of all fusions identified in our study have also been reported in the TCGA Fusion Gene Database (see Supporting Information Table S9). A similar discrepancy was observed when comparing with fusion data deposited in the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer, the major manually curated database for genomic aberrations (see Supporting Information Table S10). Fusion transcripts have also been identified in normal tissues and we have previously reported that some of them involve miRNA host genes, although the frequency and properties of these fusions differed from the fusions we found in our own data for breast tumors. The fusion partners reported for 27 normal tissues by Babiceanu et al. included 94 snoRNA host genes among 2,149 different fusion partners (4.4%), which is slightly lower than in our cancer data (5.1%). Fusions were also found for 76 of these host genes in our breast cancer dataset, but the filtering we perform to remove transcripts flagged by FusionCatcher as likely false positives effectively exclude all fusion gene pairs that overlap between the two datasets. Removing all filters for fusions that have been found in normal tissues increases the total number of fusion gene pairs in our data by 8%, but the overlap remains strikingly small at only 0.16%.

To define recurrent gene fusions we classified groups of fusions that affect the expression of a particular snoRNA as convergent fusions, independently of the identity of the partner gene. A total of 59% of all snoRNAs were located 3’ of a...
breakpoint in at least one breast tumor, indicating that their transcription is regulated by the promoter of the 5′ partner gene. As many as 34% of all snoRNAs were found in 3′ fusions recurrent in at least three breast tumors, but only 0.4% had previously been reported as recurrent in the TCGA data. Based on expression profiling, Gong et al. analyzed the clinical relevance of snoRNAs in more than 10,000 samples across 31 different cancer types. Forty-six snoRNAs were considered to be clinically relevant in at least 12 tumor types due to their association with different parameters such as tumor stage and patient survival. In our analysis, nine of these snoRNAs (SNORD10, SNORD18C, SNORD83B, SNORD82 (hosted by NCL), SNORD34, SNORD96A (hosted by RACK1), SNORD36C, SNORD72 and SNORD35B) were found in fusion transcripts. Six were included in frequent fusions: SNORD10 in 5.6% of all tumors, SNORD83B in 6.0%, SNORD82 in 8.7%, SNORD96A in 6.6%, SNORD72 in 2.5% and SNORD35B in 4.6% of all analyzed tumors.

In our work, we have applied a noncoding RNA-centered logic to highlight the relevance of intron-encoded noncoding genes when analyzing fusion transcripts. But even in cases where snoRNA expression remains intact the host proteins may be affected by the rearrangement (Fig. 3). In fact, since the expression of snoRNAs and their hosts is coordinated, any change that perturbs the expression of one component may affect the protein translation machinery. Among the most frequent genes involved in fusions we found ribosomal components such as RPL3, RPS11, RPL10 and RPS19 (translocated or mutated in almost 25% of all patients with Diamond–Blackfan anemia; see Supporting Information Table S5). This list also includes proteins associated with snoRNA function such as NOP58 that in complex with NOP56 and fibrillarin directs the 2′-O-methylation of ribosomal RNA and RACK1, a multifunctional protein associated with ribosomes with pro-oncogenic or anti-oncogenic actions that are known to regulate translation through PKC-mediated phosphorylation of eIF6. On top, fusions involving TPT1 (mentioned above) and nucleolin (NCL) are found in more than 10% of the tumors analyzed. Both proteins are known to be associated with p53 activity.

Acknowledgements

The authors would like to acknowledge patients and clinicians participating in the SCAN-B study, the staff at the central SCAN-B laboratory at the Division of Oncology, Lund University, the Swedish national breast cancer quality registry (NKBC), Regionalt biobankcentrum Syd, and the South Sweden Breast Cancer Group (SSBCG). Our study was generously supported by the Swedish Cancer Society (Rovira CAN 2016/811, Mitleman CAN 2018/440), the Mrs. Berta Kamprad Foundation (Rovira and Borg-SCAN-B), the L2 Bridge Biltema Foundation (Borg-SCAN-B), the Mats Paulsson Foundation (Borg-SCAN-B), Region Skåne Regional funds, and Swedish governmental funding ALF (Rovira and Borg). We would also like to thank Anders H. Lund and Disa Teller, Biotech Research & Innovation Centre (BRCIC), Copenhagen, Denmark for their constructive comments.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

1. Bachellerie JP, Cavaille J, Hutterhofer A. The expanding snoRNA world. Biochimie 2002;84:775–90.
2. Jorjani H, Kehr S, Jedlinski DJ, et al. An updated human snoRNAome. Nucleic Acids Res 2016;44:5068–82.
3. Ender C, Krek A, Friedlander MR, et al. A human snoRNA with microRNA-like functions. Mol Cell 2008;32:519–28.
4. Jady BE, Ketela A, Kiss T. Human intron-encoded Alu RNAs are processed and packaged into Wdr79-associated nucleolar pre-mRNA. Genes Dev 2012;26:1897–910.
5. Kishore S, Gruber AE, Jedlinski DJ, et al. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. Genome Biol 2013;14:R45.
6. Perry RP. The architecture of mammalian ribosome. BMC Evol Biol 2005;5:15.
7. Krogh N, Jansson MD, Hafler SJ, et al. Profiling of 2′-O-me in human tRNA reveals a subset of fractionally modified positions and provides evidence for ribosome heterogeneity. Nucleic Acids Res 2016;44:7884–95.
8. Lafontaine DL. Noncoding RNAs in eukaryotic ribosome biogenesis and function. Nat Struct Mol Biol 2015;22:11–9.
9. Truitt ML, Ruggiero D. New frontiers in translational control of the cancer genome. Nat Rev Cancer 2016;16:288–304.
10. Fang X, Yang D, Luo H, et al. SNORD126 promotes HCC and CRC cell growth by activating the PK3-AKT pathway through FGFR2. J Mol Cell Biol 2017;9:243–55.
11. Krefft J, Frampton AE, Mirnezami R, et al. Growth arrest-specific transcript 5 associated snoRNA levels are related to p53 expression and DNA damage in colorectal cancer. PLoS One 2014;9:e89561.
12. McMahon M, Contreras A, Holm M, et al. A single HACA small nucleolar RNA mediates tumor suppression downstream of oncogenic RAS. Elife 2019;8:e48847.
13. Mei YP, Liao JP, Shen J, et al. Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. Oncogene 2012;31:2794–804.
14. Kishore S, Stamm S. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. Science 2006;311:230–2.
15. Sahoo T, del Gaudio D, German JR, et al. Prader-Willi syndrome is caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nat Genet 2008;40:719–21.
16. Dong XY, Rodriguez C, Guo P, et al. snoRNA U50 is a candidate tumor-suppressor gene at 6q14.3 with a mutation associated with clinically significant prostate cancer. Hum Mol Genet 2008;17:1031–42.
17. Siprashvili Z, Webster DE, Johnston D, et al. The noncoding RNAs SNORD50A and SNORD50B bind K-Ras and are recurrently deleted in human cancer. Nat Genet 2016;48:53–8.
18. Tanaka R, Sato T, Moriyama M, et al. Intronic U50 small-nucleolar RNA (snoRNA) host gene of no protein-coding potential is mapped at the chromosome breakpoint (3,6)(q72,q15) of human B-cell lymphoma. Genes Cells 2000;5:277–87.
19. Chu L, Su M, Maggi LB, et al. Multiple myeloma-associated chromosomal translocation activates orphan snoRNA ACA11 to suppress oxidative stress. J Clin Invest 2012;122:2793–806.
20. Persson H, Sokilde R, Hakkinen J, et al. Frequent miRNA-convergent fusion gene events in breast cancer. Nat Commun 2017;8:788.
21. Saal LH, Vallon-Christersson J, Hakkinen J, et al. The Sweden Cancerome analysis network—breast (SCAN-B) initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Med 2015;7:20.
22. Ryden L, Loman N, Larsson C, et al. Minimizing inequality in access to precision medicine in breast cancer by real-time population-based molecular analysis in the SCAN-B initiative. Br J Surg 2018;105:e158–68.
23. Persson H, Sokilde R, Pironta AC, et al. Preparation of highly multiplexed small RNA sequencing libraries. Biotechniques 2017;63:57–64.
24. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2013;29:1064–5.
analysis of digital gene expression data. Bioinformatics 2010;26:139–40.
25. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2019;47:D766–73.
26. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res 2019;47:D155–62.
27. Hu X, Wang Q, Tang M, et al. TumorFusions: an integrative resource for cancer-associated transcript fusions. Nucleic Acids Res 2018;46:D1144–9.
28. Mitelman F, Johansson B, Mertens F. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. Available from http://cgap.nci.nih.gov/Chromosomes/Mitelman 2019.
29. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.
30. Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7.
31. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
32. Nicorici DSM, Edgren H, Kangaspeska S, et al. FusionCatcher—a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv 2014:011650.
33. Hinske LC, Galante PA, Kuo WP, et al. A potential role for intragenic miRNAs on their hosts’ interactome. BMC Genomics 2016;11:533.
34. Cleynen A, Szalat R, Kemal Samur M, et al. Expressed fusion gene landscape and its impact in multiple myeloma. Nat Commun 2017;8:1893.
35. Yoshihara K, Wang Q, Torres-Garcia W, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene 2015;34:4845–54.
36. Maxwell ES, Fournier MJ. The small nucleolar RNAs. Annu Rev Biochem 1995;64:897–934.
37. de Turris V, Di Leva G, Caldarola S, et al. TOP promoter elements control the relative ratio of intron-encoded snoRNA versus spliced mRNA biosynthesis. J Mol Biol 2004;344:383–94.
38. Newey I, Sokilde R, Persson H, et al. HER2-encoded mir-4728 forms a receptor-independent circuit with miR-21-5p through the non-canonical poly(a) polymerase PAPD5. Sci Rep 2016;6:35664.
39. Babiceanu M, Qin F, Xie Z, et al. Recurrent chimeric fusion RNAs in non-cancer tissues and cells. Nucleic Acids Res 2016;44:2859–72.
40. Gong J, Li Y, Liu CJ, et al. A pan-cancer analysis of the expression and clinical relevance of small Nucleolar RNAs in human cancer. Cell Rep 2017;21:1968–81.
41. Attiken S, Semple CA. The circadian dynamics of small nucleolar RNA in the mouse liver. J R Soc Interface 2017;14:20170034.
42. Draptchinskaia N, Gustavsson P, Andersson B, et al. The gene encoding ribosomal protein S19 is mutated in diamond-Blackfan anaemia. Nat Genet 1999;22:75–74.
43. Ceci M, Gaviraghi C, Gorrini C, et al. Release of eIF6 (p27BBP) from the 60S subunit allows 80S ribosome assembly. Nature 2003;426:579–84.
44. Takagi M, Absalon MJ, McLure KG, et al. Regulation of p53 translation and induction after DNA damage by ribosomal protein L26 and nucleolin. Cell 2005;123:49–63.