

RESEARCH

Open Access



De novo annotation of lncRNA *HOTAIR* transcripts by long-read RNA capture-seq reveals a differentiation-driven isoform switch

Evdokiia Potolitsyna^{1†}, Sarah Hazell Pickering^{1,2†}, Ave Tooming-Klunderud³, Philippe Collas^{1,2*} and Nolwenn Briand^{1,2*}

Abstract

Background: lncRNAs are tissue-specific and emerge as important regulators of various biological processes and as disease biomarkers. *HOTAIR* is a well-established pro-oncogenic lncRNA which has been attributed a variety of functions in cancer and native contexts. However, a lack of an exhaustive, cell type-specific annotation questions whether *HOTAIR* functions are supported by the expression of multiple isoforms.

Results: Using a capture long-read sequencing approach, we characterize *HOTAIR* isoforms expressed in human primary adipose stem cells. We find *HOTAIR* isoforms population displays varied splicing patterns, frequently leading to the exclusion or truncation of canonical LSD1 and PRC2 binding domains. We identify a highly cell type-specific *HOTAIR* isoform pool regulated by distinct promoter usage, and uncover a shift in the *HOTAIR* TSS usage that modulates the balance of *HOTAIR* isoforms at differentiation onset.

Conclusion: Our results highlight the complexity and cell type-specificity of *HOTAIR* isoforms and open perspectives on functional implications of these variants and their balance to key cellular processes.

Keywords: *HOTAIR*, Adipose differentiation, Adipose stem cells, Capture-seq, Long-read sequencing, lncRNA isoform

Background

Long non-coding RNAs (lncRNAs) are increasingly recognized as major regulators of physiological processes and have emerged as biomarkers for disease diagnosis and prognosis [1]. *HOTAIR* (*HOX* Transcript Antisense RNA) is a long antisense transcript of ~500–2200 base pairs (bp) located within, and extending upstream of, the *HOXC11* gene, in the *HOXC* locus on chromosome 12. *HOTAIR* is mostly studied in cancer models where

its overexpression promotes cell migration and metastasis by altering gene expression [2–5]. *HOTAIR* is also the most differentially expressed gene between upper- and lower-body adipose tissue, and its expression is induced during adipose differentiation of gluteofemoral adipose stem cells (hereafter referred to as ASCs) [6, 7]. The function of *HOTAIR* in adipose tissue remains, however, unclear. In contrast to its effect in cancer cell lines [8–11], *HOTAIR* overexpression does not affect adipose progenitors' gene expression or proliferation rates [7, 12]. These observations point to different functions and mechanisms of action of *HOTAIR* in cancer vs. mesenchymal progenitor cells.

lncRNAs interact with proteins, DNA and other RNAs to regulate gene expression at multiple levels. *HOTAIR* has been detected in the cytoplasm where it can promote

[†]Evdokiia Potolitsyna and Sarah Hazell Pickering contributed equally to this work.

*Correspondence: philc@medisin.uio.no; nolwenn.briand@medisin.uio.no

¹ Department of Molecular Medicine, Institute of Basic Medical Sciences, Faculty of Medicine, University of Oslo, PO Box 1112, 0317 Oslo, Blindern, Norway

Full list of author information is available at the end of the article



ubiquitin-mediated proteolysis by associating with E3 ubiquitin ligases [13] or function as a microRNA sponge [14]. *HOTAIR* is also found in the nucleus, where it can bind chromatin [13, 15, 16] and act as scaffold for chromatin-modifying complexes through binding to the Polycomb repressor complex PRC2 subunit EZH2 [17], a histone H3K27 methyltransferase, and to LSD1/KDM1A, the H3K4/K9 demethylase of the REST/CoREST complex [18]. Recent evidence demonstrates that *HOTAIR*-PRC2 binding can be modulated by changes in *HOTAIR* structure mediated by RNA binding protein (RBP)-RNA-lncRNA interactions [19].

lncRNA folding into secondary and tertiary structures dictates their interactome, making their function dependent on structural conservation [20]. While the full-length *HOTAIR* sequence is poorly evolutionarily conserved, folding prediction has identified two well-conserved structures in the 5' and 3' ends of *HOTAIR* [21]. The currently reported primary structure of the *HOTAIR* gene is complex, with 2 predicted promoters, multiple predicted transcription start sites (TSSs) and potential splice sites leading to several isoforms [22–25]. *HOTAIR* transcripts can harbor small exon length variations [26], or alternative splice site usage that eliminates the PRC2 [18] or RBP [19] binding domains. Therefore, distinct *HOTAIR* splice variants likely have distinct functions, warranting an isoform-specific annotation in relevant tissues.

Current reference annotations for lncRNAs are incomplete due to their overall low expression level, weak evolutionary conservation, and high tissue specificity [27]. Identification of lncRNA isoforms using short-read RNA sequencing (RNA-seq) is challenging because almost every exon can be alternatively spliced [28], and short reads cannot resolve the connectivity between distant

exons. Long-read sequencing technologies can address this challenge by covering the entire RNA sequence in a single read, enabling mapping of isoform changes that may impact lncRNA structure and function [29].

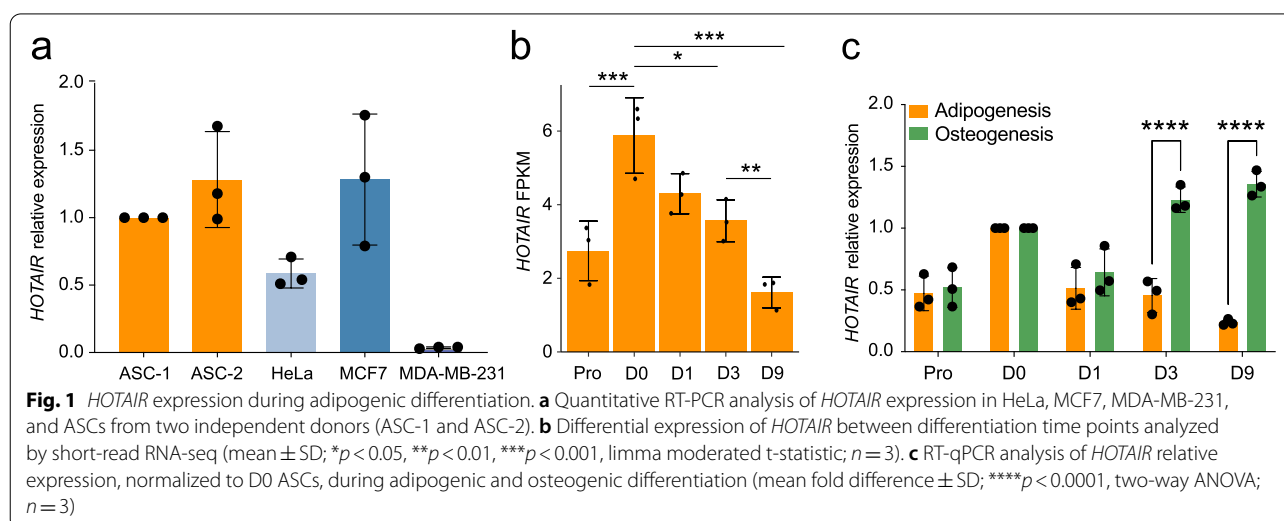
Here, we combine long-read Capture-seq and Illumina short-read RNA-seq to resolve changes in the composition of *HOTAIR* isoforms in a well-characterized adipogenic differentiation system [30]. We uncover a temporal shift in the composition of *HOTAIR* isoforms upon induction of differentiation, regulated by distinct promoter usage and hormonal and nutrient-sensing pathways.

Results

HOTAIR is highly expressed in ASCs and regulated during adipogenesis

We first assessed *HOTAIR* expression level in ASCs versus cancer cell lines where *HOTAIR* function has been previously studied [28]. We find that *HOTAIR* expression level in ASCs from two unrelated donors is higher or comparable to that of cancer cell lines (Fig. 1a), confirming the relevance of primary ASCs as a model system to assess the relative abundance of *HOTAIR* isoforms.

We examined by short-read RNA-seq the transcriptome of ASCs in the proliferating stage (Pro), after cell cycle arrest (day 0; D0), and after 1, 3 and 9 days of adipogenic induction (D1, D3, D9). Hierarchical clustering of differentially expressed genes across time ($\alpha < 0.01$ between at least two consecutive time points) confirms the upregulation of genes pertaining to the hallmark “Adipogenesis”, including the master adipogenic transcription factor *PPARG* (Additional file 1, Fig. S1). Moreover, *HOTAIR* displays a diphasic expression profile in this time course, with increased levels upon cell cycle



arrest on D0, followed by a progressive downregulation (Fig. 1b). *HOTAIR* expression is maintained during osteogenic differentiation (Fig. 1c), indicating a lineage-specific mode of regulation.

Identification of main *HOTAIR* isoforms

To identify *HOTAIR* isoforms, we performed PacBio single-molecule, long-read isoform sequencing of captured polyadenylated *HOTAIR* transcripts (PacBio Capture-seq) in proliferating ASCs and during adipose differentiation. Full-length reads were clustered into non-redundant transcripts and aligned to the hg38 genome assembly, providing excellent coverage over the *HOTAIR* locus with sharp exon boundaries (Fig. 2a,b). The Isoseq3 pipeline yielded ~6000 *HOTAIR* isoforms; these were further filtered and merged both across time points and based on internal junctions using Cupcake ToFU [31, 32] or TAMA [33] (Fig. 2a), resulting in 34 isoforms (Fig. 2a,c; Additional file 2, Table S1, Additional file 3).

To characterize these *HOTAIR* isoforms, we first classified transcripts into four main categories using SQANTI [34]. Out of the 34 aforementioned isoforms, we find (i) 8 full splice matches (FSM) transcripts, (ii) 6 incomplete splice matches (ISM) of an annotated (known) transcript, (iii) 12 novel in catalog (NIC) transcripts containing new combinations of annotated splice sites, and (iv) 8 novel not in catalog (NNC) transcripts using at least one unannotated splice site (Fig. 2d, Additional file 1, Fig. S2). Second, we assessed isoform variations at the 5' and 3' ends. Transcripts start sites supported by Cap analysis of gene expression (CAGE) data are distributed evenly across SQANTI categories, with 15 transcripts starting within 15 bp of a CAGE peak summit (Fig. 2e). Transcripts with a TSS located more than 250 bp away from a CAGE peak likely represent lowly expressed or highly cell type-specific isoforms of *HOTAIR*, explaining the absence of a dedicated CAGE peak [35, 36]. *HOTAIR* full length transcripts also show variable 3' ends (exon 7; E7) corresponding to the alternative usage of 4 different canonical polyA signals for transcription termination (Fig. 2f, Additional file 2, Table S2). Third, we find the highest number of reads for transcripts in FSM, ISM and NIC SQANTI categories (Fig. 2g), indicating that most *HOTAIR* transcripts identified here have a known exon and splice junction composition.

To identify the top isoforms expressed across differentiation, we further filtered candidates based on read counts (Fig. 2h). Only 23 transcripts accumulate more than 100 reads, and these are also detected in at least 2 samples. These 23 high-confidence isoforms arise from multiple TSS usage, alternative splicing and intron retention events, as well as polyA site usage (Fig. 2i). Altogether, our PacBio sequencing analysis identifies with high

confidence known and novel uncharacterized *HOTAIR* transcript isoforms in our adipose differentiation system, with notable variation in their TSS and polyA site usage.

HOTAIR splicing affects LSD1 and PRC2 interacting domains

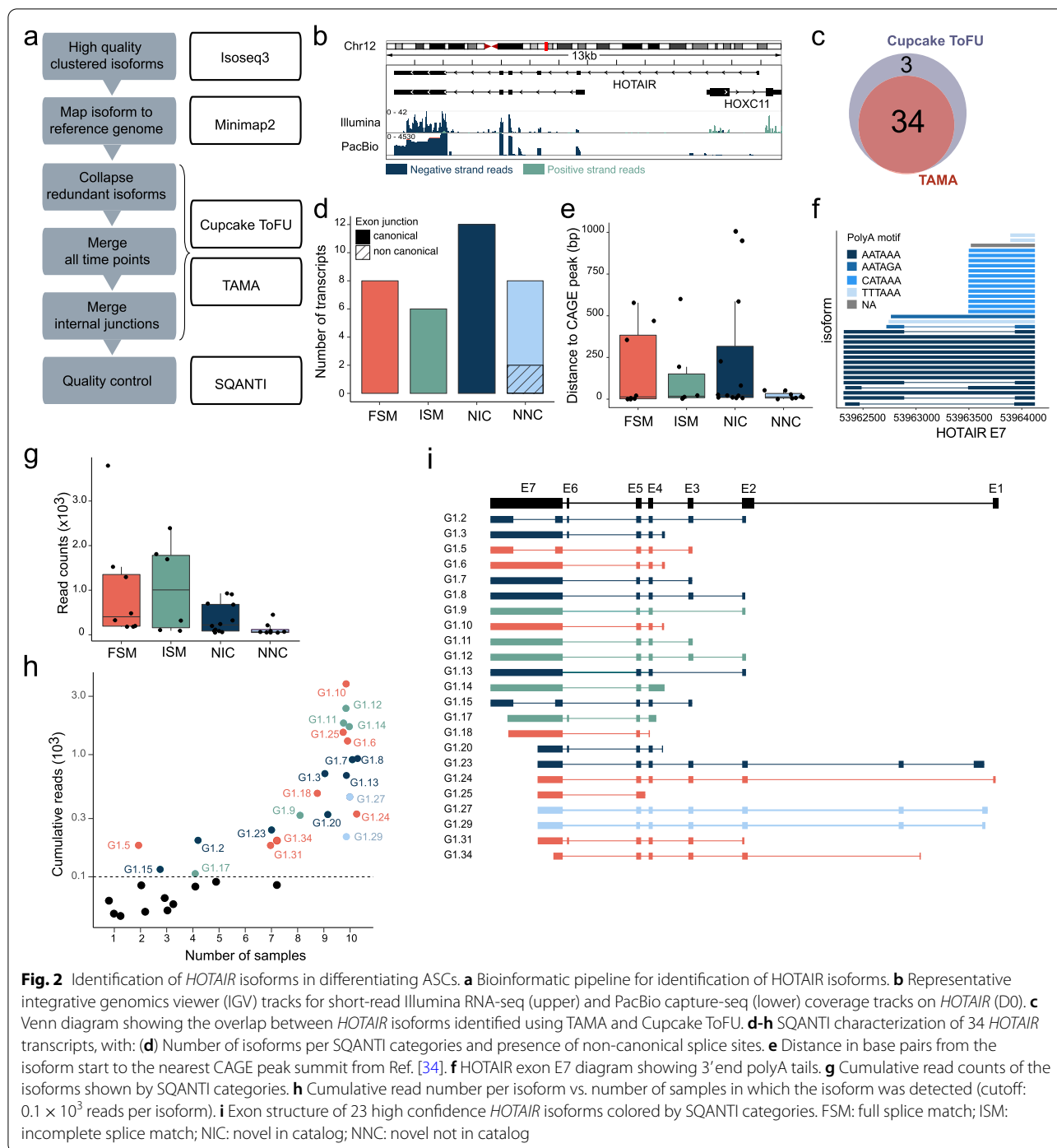
HOTAIR has been described as a scaffold for LSD1 [18] and PRC2 [18], epigenetic modifiers involved in the regulation of adipogenesis [37–39]. The LSD1 binding domain lies in the last 500 bp of *HOTAIR* exon 7 (E7), whereas the PRC2 binding domain spans exons 4 and 5 [40] (Fig. 3a).

Inclusion of the LSD1 binding domain in *HOTAIR* isoforms depends on polyA site usage (see Figs. 3a and 2f,i); thus we examined potential changes in exon E7 length during adipogenesis. The proportion of PacBio reads for the long forms of E7, harboring the LSD1 binding domain, does not significantly vary during differentiation (Fig. 3b). In agreement, the expression profile of LSD1-containing *HOTAIR* isoforms (E7 long and medium) is similar to that of the total isoform pool (E7 short) (Fig. 3c, see Fig. 1bc), confirming that alternative polyadenylation pattern of *HOTAIR* is maintained during differentiation (Fig. 3c). Hence, variations in *HOTAIR* isoforms during adipogenesis do not in principle impact its LSD1 scaffolding function.

We next investigated alternative splicing events affecting the PRC2 binding domain. Analysis of exon E5 alternative splicing reveals two major alternative splice sites (site 1 and site 3) (Fig. 3d). Induction of differentiation promotes splicing at site 3, leading to a slight increase in proportion of the shorter E5 variant (Fig. 3e). This splicing event can be readily detected by semi-quantitative RT-PCR analysis of exon E3-E5 expression in ASCs from two independent individuals (Fig. 3f, Additional file 1, Fig. S3a,b,c), emphasizing the robustness of the PacBio approach in capturing relative variations in *HOTAIR* transcript levels during differentiation. Overall, PacBio Capture-seq analysis reveals that *HOTAIR* is expressed as a pool of isoforms with differential ability to bind LSD1 and PRC2. However, adipogenic induction does not significantly affect the proportion of isoforms with LSD1 or PRC2 binding capacity.

Adipogenic induction triggers a switch in *HOTAIR* isoform start sites

One noticeable feature of *HOTAIR* isoform pool is the presence of 9 distinct start sites (Fig. 4a, see Fig. 2i). To assess TSS usage, we first quantified the total number of reads for each exon start category in PacBio Capture-Seq dataset (Fig. 4b). Only transcripts starting from exons E2, E3, E3.1 and E5 cumulated more than 500 reads during differentiation. We next quantified the proportion



of each of *HOTAIR* 23 high-confidence isoform at each time point (Fig. 4c). Strikingly, while E3.1-starting isoforms contribute to the majority of reads prior to differentiation onset (Pro, D0), adipogenic induction triggers a decrease in the E3.1-starting isoform pool which is compensated by an increase in E3-starting isoform expression (Fig. 4c,d,e). Short-read RNA-seq analysis confirms

this switch in TSS usage, with exon E3 becoming significantly more expressed than E3.1 upon differentiation onset (D1; Fig. 4f). The sharp drop in E3.1 starting isoforms is readily observed by semi-quantitative RT-PCR using primers spanning exons E3.1-E4, while expression of E3-E5 containing isoforms is maintained during early adipogenesis (Fig. 4g, left; Additional file 1, Fig. S3b).

Importantly, osteogenic differentiation of ASCs maintains E3.1 isoforms expression (Fig. 4g, right; Additional file 1, Fig. S3a), indicating an adipose-specific regulation of *HOTAIR* isoforms. We conclude that adipogenic commitment triggers a switch in *HOTAIR* TSS usage, potentially impacting functional binding domains located in 5' exons [19].

ASCs express a cell type-specific *HOTAIR* isoform

To assess cell type- and tissue-specificity of E3.1 starting *HOTAIR* isoforms, we used Snaptron [41], a search engine for querying splicing patterns in publicly available RNA-seq datasets (Additional file 2, Table S3). We found only 24 cell samples with more than 10 reads containing the E3.1-E4 junction, while the E3-E4 junction was detected in 672 samples (Fig. 5a; Additional file 2, Table S4). We confirmed by semi-quantitative RT-PCR the expression of exon E3.1 in cultured primary myoblasts, BJ fibroblasts and HEK 293 T cells – albeit to lower levels than in ASCs – and its absence in HeLa cells (Fig. 5b).

To gain insight into the regulation of *HOTAIR* isoform expression, we examined the chromatin accessibility landscape of the *HOTAIR* locus in published Assay for Transposable Accessible Chromatin (ATAC)-seq data [44–47] (Fig. 5c). We find two regions of accessible ('open') chromatin in ASCs (R3, R4), which our chromatin immunoprecipitation (ChIP)-seq data show are also enriched in H3K4me3 and H3K27ac, histone modifications characterizing active regulatory sites (Fig. 5c, d). Region R3 coincides with both activating and repressing regulatory element (REM) annotations by EpiRegio [42], suggesting it constitutes a bivalent regulator for *HOTAIR* or nearby genes expression. Region R4, located immediately upstream of exon E3.1 is also in an 'open' chromatin in myoblasts but not in HeLa cells, which respectively do and do not express E3.1-starting transcripts (Fig. 5b,c), and likely represents the active promoter for E3.1-starting isoforms. In contrast, these regions are in a 'closed' configuration in mature adipocytes, consistent with *HOTAIR* downregulation during adipogenesis (see Fig. 1b). Thus, *HOTAIR* displays active regulatory sites in ASCs, in line with the expression of specific isoforms. However, *HOTAIR* upregulation upon cell cycle arrest is

not accompanied by changes of H3K4me3 and H3K27ac levels, and loss of H3K4me3 at region 4 occurs only at later differentiation time points (D3 and D9). Thus, modulation of *HOTAIR* levels during adipogenesis is not mediated by epigenetic regulation (Fig. 5d).

HOTAIR stability increases upon cell cycle arrest

HOTAIR half-life varies according to cell type, from 4 h in HeLa cells [13] to more than 7 h in primary trophoblast cells [48]. We therefore asked whether increased *HOTAIR* levels upon cell cycle arrest (D0; see Fig. 1b) could relate to a change in lncRNA stability. Treatment with Flavopiridol to inhibit RNA polymerase II (Pol II) transcription reveals an increase in *HOTAIR* half-life from 3 h in proliferating ASCs to >4 h in D0 cells (Fig. 6a), while cell cycle arrest does not impact the stability of control short-lived *CEBPD* or long-lived *GAPDH* mRNA (Fig. 6b,c). Thus, while ATAC-seq data indicate that the *HOTAIR* isoform balance is likely regulated by cell type-specific transcription factors, the global increase in *HOTAIR* levels observed at D0 results at least in part from an increase in its stability.

Discussion

lncRNAs can modulate important biological and pathophysiological processes through their interaction with multiple partners. Our long-read Capture-seq of *HOTAIR* unveils a complex and dynamic pool of isoforms in differentiating ASCs. Adipogenic induction triggers a switch in the expression of main *HOTAIR* isoforms which likely impacts on its structure and interactome. Our results emphasize the importance of robust lncRNA annotation in the tissue of interest prior to functional characterization.

A single lncRNA locus can generate many transcript variants through alternative TSS usage, polyadenylation sites and splicing events [49]. We find that previously undescribed *HOTAIR* variants constitutes the major isoforms in human adipose stem cells, and show that not only lncRNA expression level, but also the isoform pool expressed can be highly cell type-specific, adding another layer of complexity to the regulation of biological processes by lncRNAs.

(See figure on next page.)

Fig. 3 *HOTAIR* splicing across functional domains during adipogenesis. **a** Schematic representation of *HOTAIR* 3' exons (exons E4 to E7) containing LSD1 and PRC2 binding domains. Alternative splicing events and polyA (pA) sites usage detected in 5 representative *HOTAIR* high confidence isoforms are represented. **b** Proportion of long-read Capture-Seq reads for each E7 length across differentiation time points. **c** Schematic representation of RT-qPCR amplicons (upper panel) and qRT-PCR analysis of *HOTAIR* E7 length variation during adipose differentiation (lower panel). **d** PacBio Capture-Seq analysis of *HOTAIR* exon E5 alternative splice site usage and **e** of the proportion of major exon E5 splice variants during adipogenic differentiation. **f** Schematic representation of PCR amplicons (upper panel) and semi-quantitative RT-PCR analysis of *HOTAIR* expression using primers located in exons E3 and E5 (E3-E5; lower panel). *SF3A1* is shown as a loading control. RT-: no reverse transcriptase control; NTC: no template control. Full-length gels are presented in Additional file 4, Fig. S4

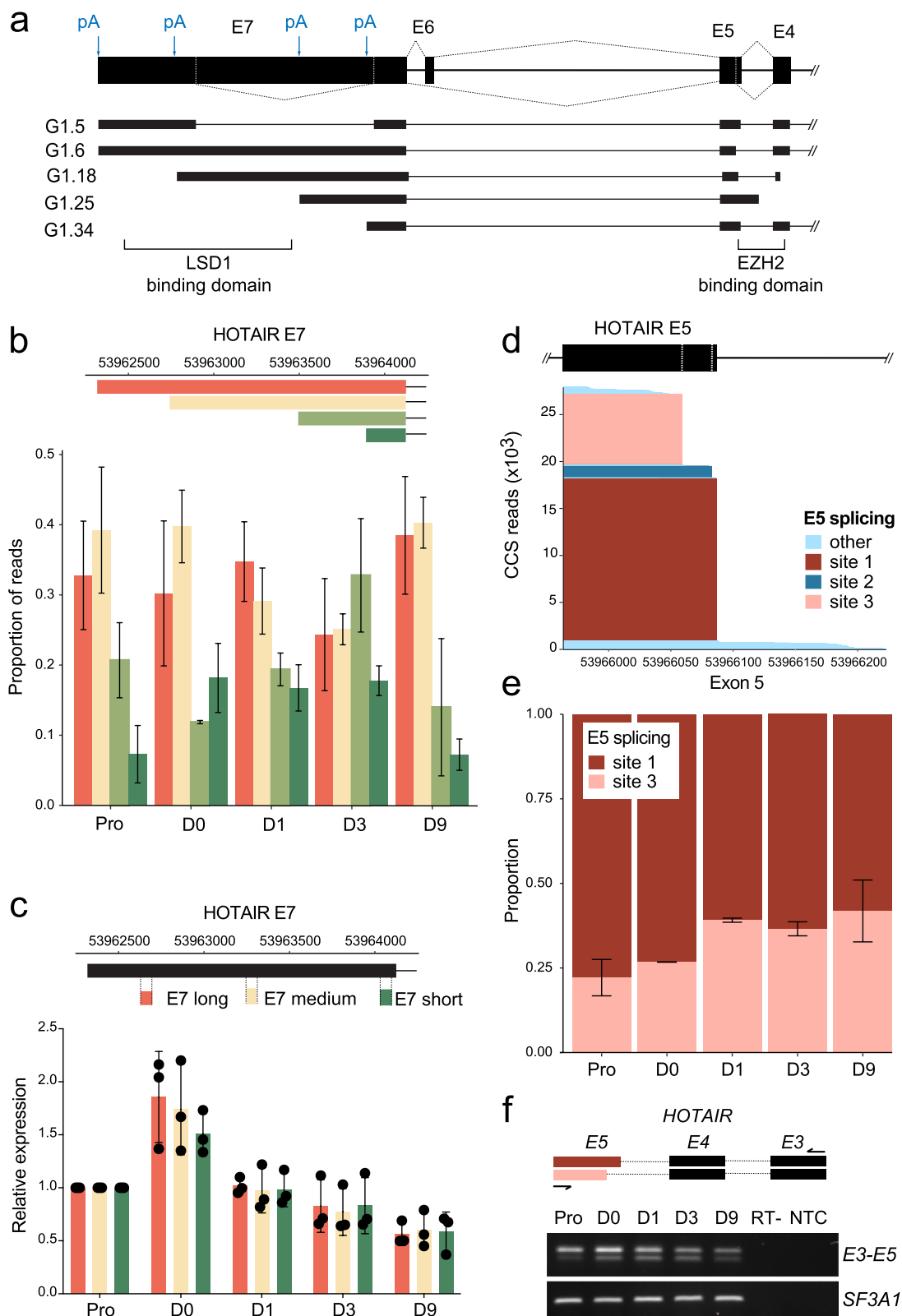
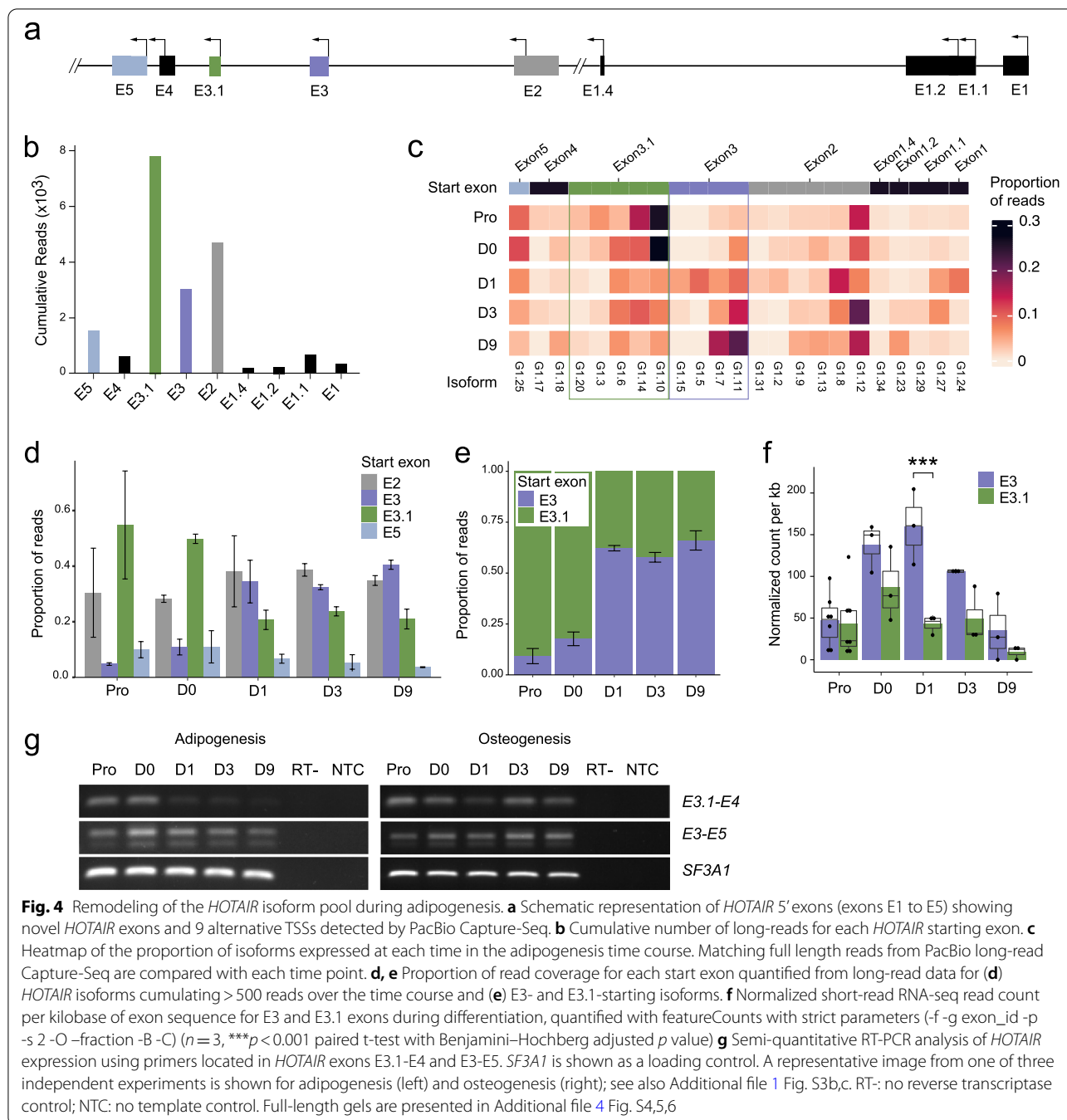
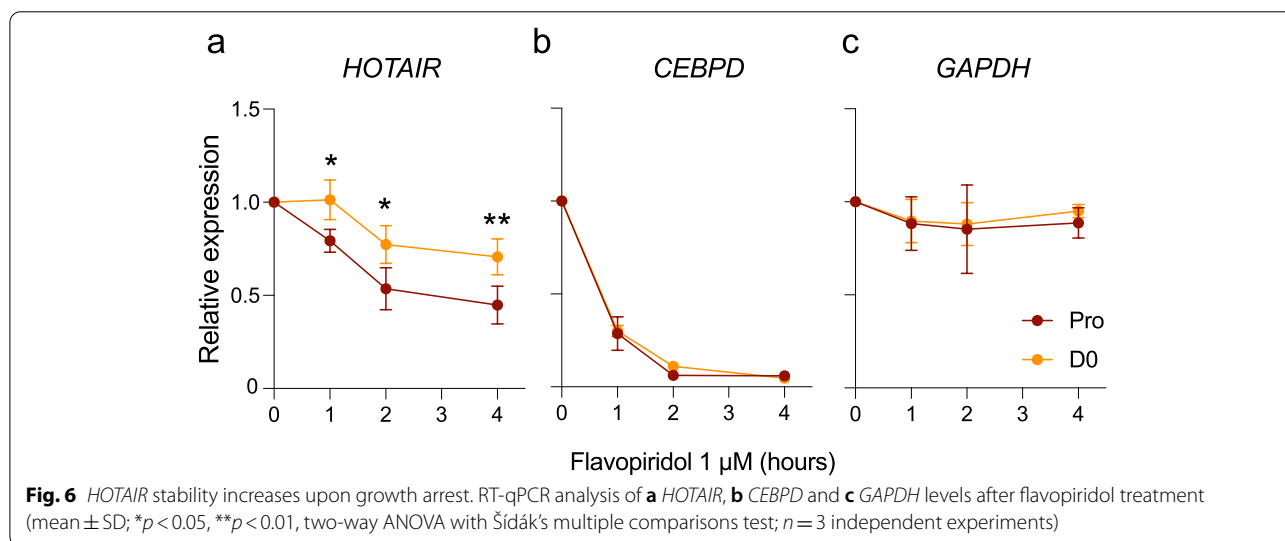
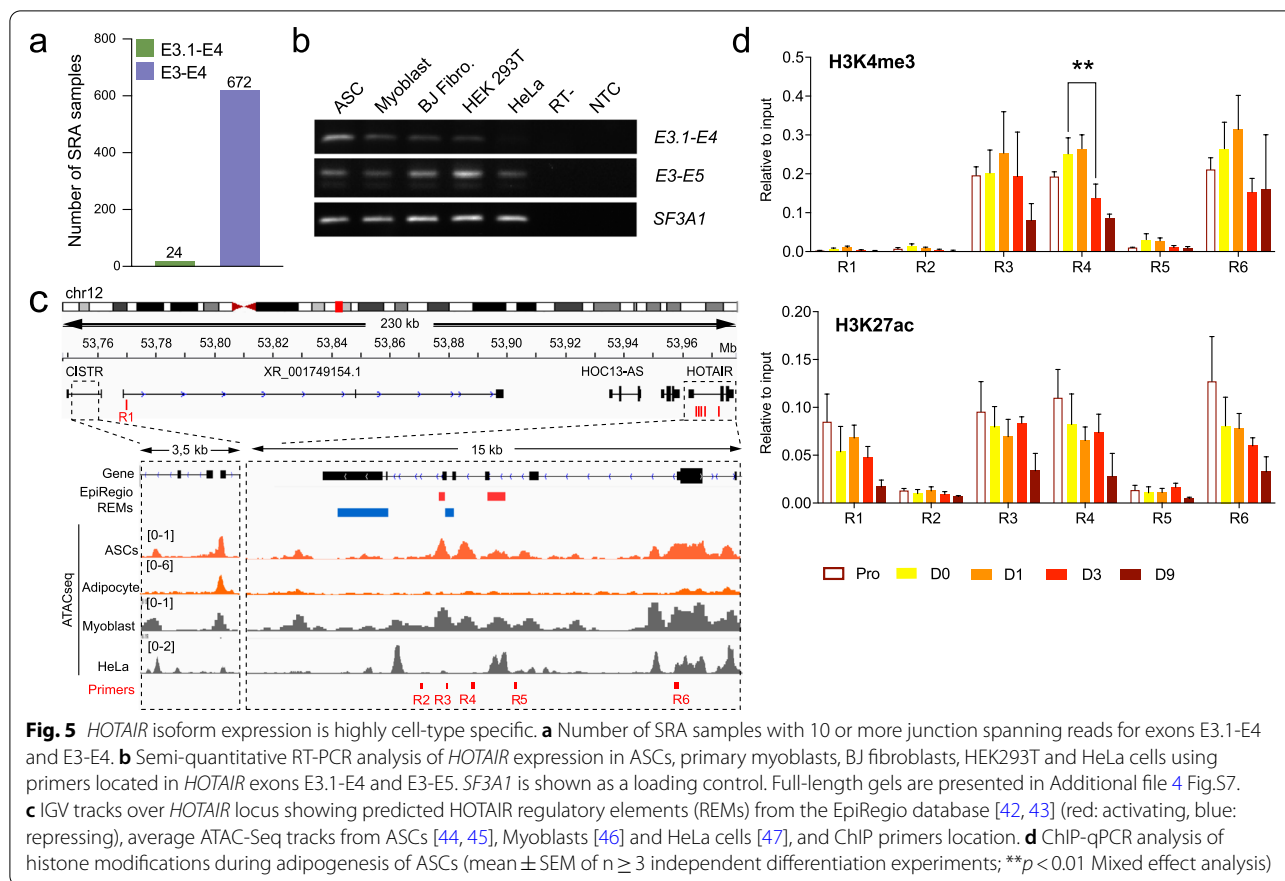


Fig. 3 (See legend on previous page.)



The chromatin landscape of the *HOTAIR* locus in ASCs is consistent with a promoter upstream of *HOTAIR* exon E3.1, which is in an ‘open’ and active epigenetic configuration in ASCs and myoblasts, where *HOTAIR* E3.1 starting isoforms are expressed, but not in HeLa cells. However, adipogenic induction results in only mild and delayed changes in active histone modification at this site, suggesting that expression of various *HOTAIR* isoforms is rather regulated by differentiation-stage specific

transcription factors. In line, alternative promoter usage is elicited downstream of adipogenic, but not osteogenic, signaling pathways. Additionally, increased *HOTAIR* levels upon cell cycle arrest (D0 cells) also correlates with an increase in its stability. Interestingly, interaction with the RNA-binding protein HuR, a negative regulator of adipogenesis [50], reduces *HOTAIR* stability [13]. Alternatively, increased *HOTAIR* stability could result from its binding to a cell stage-specific protein partner. Collectively, our



results indicate a tight, multifactorial regulation of isoform pool during adipogenesis, arguing for a functional importance of the isoform switch.

In cancer cell lines, *HOTAIR* has been reported to scaffold for chromatin regulators with broad effects on

gene expression [51–53]. Our long-read sequencing data reveal multiple polyA site usage leading to the exclusion of the LSD1 binding domain at the 3' end, or alternative splicing events affecting the PRC2 minimal binding domain. However, the proportion of *HOTAIR* isoforms

containing LSD1 or PRC2 domains does not vary during adipose differentiation, suggesting that *HOTAIR*'s role in adipogenesis is independent from its epigenetic scaffold function. Supporting this idea, *HOTAIR* depletion does not significantly affect PRC2-mediated gene regulation in adipose progenitors [12].

Recent studies have confirmed that a non-protein-coding locus can give rise to functionally distinct transcript isoforms [49, 54–56]. Of note, the switch in *HOTAIR* start site upon differentiation induction leads to the inclusion of *HOTAIR* exon 3 containing a protein binding domain [19], which likely alters *HOTAIR* function. Another intriguing possibility is that short sequence variations at the 5' end impact *HOTAIR*'s secondary structure and thus the folding of functional domains. Observation of *HOTAIR* structure using atomic force microscopy reveals multiple dynamic conformations [57]. It is therefore conceivable that *HOTAIR* functions via conformational changes, induced by or resulting from interactions with protein partners. Hence, variations in isoform composition likely results in cell type specific structures and interactomes, which could account for the divergent roles of *HOTAIR* in primary cells and cancer cell lines.

Conclusions

We generate the first cell type-specific, comprehensive catalog of *HOTAIR* isoforms in a physiological context and describe novel *HOTAIR* isoforms, alternative splicing events, and multiple start site usage. We find that *HOTAIR* splicing in ASCs often leads to the exclusion or truncation of canonical LSD1 and PRC2 binding domains. We uncover a shift in *HOTAIR* TSS usage that controls the balance of *HOTAIR* isoforms during early adipogenesis. The variability of *HOTAIR* isoforms opens new perspectives for studies in (patho)physiological contexts.

Methods

All methods were performed in accordance with the guidelines and regulations of the University of Oslo.

Cell culture and differentiation

ASCs from two non-obese donors (ASC-1 and ASC-2) were cultured in DMEM/F12 with 10% fetal calf serum and 20 ng/ml basic fibroblast growth factor (Pro). Upon confluency, growth factor was removed, and cells were cultured for 72 h before induction of differentiation (D0). For adipose differentiation, ASCs were induced with 0.5 μ M 1-methyl-3 isobutyl xanthine, 1 μ M dexamethasone, 10 μ g/ml insulin and 200 μ M indomethacin. For osteogenic differentiation, ASCs were induced with 0.1 μ M dexamethasone, 10 mM

β -glycerophosphate and 0.05 mM L-ascorbic acid-2 phosphate. Differentiation media was renewed every 3 days, and samples were harvested on D1, D3 and D9 after induction. Differentiation experiments were done in at least biological triplicates. HeLa cells (American Type Culture Collection; CCL-2) were cultured in MEM medium containing Glutamax (Gibco), 1% non-essential amino acids and 10% fetal calf serum. MDA-MB-231 and MCF-7 cells (both from American Type Culture Collection) were cultured in DMEM containing 10% fetal calf serum. Human myoblasts (Lonza) were cultured as described [58]. BJ fibroblasts (American Type Culture Collection) were cultured in DMEM/F12 with 10% fetal calf serum and 20 ng/ml basic fibroblast growth factor. HEK293T (Thermo Scientific) were cultured in DMEM/F12 with 10% fetal calf serum.

RT-qPCR and semi-quantitative PCR

Total RNA was isolated using the RNeasy kit (QIAGEN) and 1 μ g was used for cDNA synthesis using the High-Capacity cDNA Reverse Transcription Kit (ThermoFisher). RT-PCR was done using IQ SYBR green (Biorad) with *SF3A1* as a reference gene. PCR conditions were 95 $^{\circ}$ C for 3 min and 40 cycles of 95 $^{\circ}$ C for 30 s, 60 $^{\circ}$ C for 30 s, and 72 $^{\circ}$ C for 20 s. Semi-quantitative PCR was done using a PCR Master Mix (ThermoFisher) with the following conditions: 95 $^{\circ}$ C for 3 min and 30 cycles of 95 $^{\circ}$ C for 30 s, 60 $^{\circ}$ C for 30 s, and 72 $^{\circ}$ C for 30 s. Products were separated in a 2.5% agarose gel with Tris–Borate–EDTA buffer. PCR primers are listed in Additional file 2 Table S5. Uncropped gels are presented in Additional file 4, Fig. S4 to S7.

Short-read Illumina RNA-sequencing and data analysis.

Differentiation time courses from ASC-1 with 5 time points were sequenced in biological triplicates with short (40 bp), paired end reads on Illumina NextSeq. Reads were aligned to the hg38 genome (ensembl v95 annotation) using hisat2, and counted with featureCounts (`-fraction -M`). To further analyze differentially expressed genes, edgeR and limma packages were used. Low abundance genes were filtered using edgeR's function "filterByExpr". Genes with an adjusted *p*-value < 0.01 (eBayes method, limma package) between two or more consecutive time points were clustered with DPGP software. FDR adjusted *p*-value for Adipogenesis cluster was generated by overrepresentation analysis using Hallmark gene sets from MSigDb (v7.4) with all human genes as background. FeatureCounts with strict parameters (`-f -g exon_id -p -s 2 -O -fraction -B -C`) was used to quantify exon coverage as normalized counts per kilobase of exon sequence.

PacBio capture-seq

PacBio Capture-seq was performed on duplicate differentiation time courses from ASC-1. For each differentiation, RNA samples from 5 differentiation time points were used to synthesize full-length barcoded cDNA libraries using the Template Switching RT Enzyme Mix (NEB). Libraries were prepared using Pacific Biosciences protocol for cDNA Sequence Capture Using IDT xGen® Lockdown® probes (<https://eu.idtdna.com/site/order/ngs>). A pool of 100 probes against all known *HOTAIR* isoform sequences was designed using the IDT web tool. Full-length cDNA was cleaned up using Pronex beads, and 1200 ng was used for each hybridization reactions. Library was sequenced in one 8 M SMRT cell on a Sequel II instrument using Sequel II Binding kit 2.1 and Sequencing chemistry v2.0.

Transcript identification from targeted long-read sequencing

CCS sequences were generated for the entire dataset using the Circular Consensus Sequence pipeline (SMRT Tools v 8.0.0.80502) with minimum number of passes 3 and minimum accuracy 0.99. CCS reads were demultiplexed using the Barcoding pipeline (SMRT Tools v7.0.0.63823). Iso-Seq analysis was performed using the Iso-Seq pipeline (SMRT Link v7.0.0.63985) with default settings. Only clustered isoforms with at least 2 subreads, 0.99 quality score and containing a polyA tail of at least 20 base pairs (bp) were used. Isoforms were aligned to the hg38 genome with minimap2 v.2.17 [59]. Primary alignments to the *HOTAIR* locus (chr12: 53,962,308–53,974,956) were selected as target *HOTAIR* transcripts if they had a mapping quality above 20 and less than 50 clipped nucleotides (samclip; <https://github.com/tseemann/samclip>). Single exon and sense transcripts were also filtered.

HOTAIR transcripts were collapsed with both Cupcake Tofu [31] (https://github.com/Magdoll/cDNA_Cupcake) and TAMA [33] (<https://github.com/GenomeRIK/tama>). Parameters `-dun-merge-5-shorter` and `-x capped` were used for cupcake and TAMA respectively, to prevent shorter transcript models from being merged into longer ones. For TAMA, `-z 100` was also set to increase the allowed 3' variability. To combine transcript lists between timepoints, collapsed transcripts with at least 50 full length reads within one sample were merged using cupcake `chain_samples.py` or tama `merge.py`. Initially this produced ~80 *HOTAIR* transcripts, with much of the variation in the ends of 3' and 5' exons. To achieve a final transcript list, collapsed transcripts were merged on internal junctions only by increasing the allowed 5' and 3' variability (with TAMA options `-a 300 -z 2000`)

and ranking libraries by the number of polyadenylated *HOTAIR* reads. This resulted in 34 TAMA isoforms.

Final cross-validation was conducted by running SQANTI with unclustered ccs reads as the “novel long read-defined transcriptome” and the top 34 TAMA isoforms as the “reference annotation”. This assigned full length reads to their corresponding TAMA isoform and reads annotated as full-splice match were counted for each isoform. SQANTI was used again to classify transcripts against existing reference annotations and to search for polyA motifs near transcript ends from a list of potential human motifs (Additional file 2, Tables S1, S2) [34] (<https://github.com/ConesaLab/SQANTI>). Reference isoforms were collected from four sources: Ensembl (v95), RefSeq, UCSC browser’s “lincRNA and TUCP transcripts” and Fantom CAT (Hon 2015), the last of which included isoforms from encode, hubmap and miTranscriptome (Additional File 1 Fig. S2) SQANTI categories are defined relative to the reference isoforms as follows (i) FSM: the number of exons and all internal junctions are concordant with the reference isoform. (ii) ISM: the isoform has less 5' exons than the reference, and yet the internal junctions are perfectly consistent. (iii) NIC: all donor and acceptor sites exist in the reference isoform list but their combination in a single isoform is novel. (iv) NNC: at least one of the donor or acceptor sites in the isoform is not found in the reference list. For all categories, the exact length of the 5' and 3' ends of first and last exons, respectively, can differ by any amount.

Snaptron

We searched Snaptron’s SRA V2 database [41], which contains ~49,000 public samples, for experiments with reads overlapping *HOTAIR* exon-exon junctions with the web query <http://snaptron.cs.jhu.edu/srav2/snapttron?regions=HOTAIR&rfilter=annotated:1> (Additional File 2, Table S3). Sample IDs for experiments with at least 10 exon spanning reads were extracted and cell type information accessed via a matching metadata query. A simplified version of the exon E3.1 metadata table is presented in Additional file 2 Table S4.

Chromatin immunoprecipitation

Cells (2×10^6 /ChIP) were cross-linked with 1% formaldehyde for 10 min and cross-linking was stopped with 125 mM glycine. Cells were lysed for 10 min in ChIP lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl, pH 8.0, proteinase inhibitors, 1 mM PMSE, 20 mM Na Butyrate) and sonicated for 30 s ON/OFF for 10 min in a Bioruptor®Pico (Diagenode) to generate 200–500 bp DNA fragments. After sedimentation at 10,000 g for 10 min, the supernatant was collected and diluted 5 times in RIPA buffer (140 mM NaCl, 10 mM Tris-HCl

pH 8.0, 1 mM EDTA, 0.5 mM EGTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate, protease inhibitors, 1 mM PMSF, 20 mM Na Butyrate). After a 100 μ L sample was removed (input), diluted chromatin was incubated for 2 h with antibodies (2.5 μ g/100 μ L) coupled to magnetic Dynabeads Protein A (Invitrogen). CHIP samples were washed 4 times in ice-cold RIPA buffer, crosslinks were reversed and DNA was eluted for 2 h at 68 °C in 50 mM NaCl, 20 mM Tris-HCl pH 7.5, 5 mM EDTA, 1% SDS and 50 ng/ μ L Proteinase K. DNA was purified and dissolved in H₂O. CHIP DNA was used as template for quantitative (q)PCR using SYBR[®] Green (BioRad), with 95 °C denaturation for 3 min and 40 cycles of 95 °C for 30 s, 60 °C for 30 s, and 72 °C for 30 s. Primers used for CHIP are listed in Additional file 2, Table S5.

Statistical analyses

Statistics were performed with GraphPad Prism 9.2.0 (<https://www.graphpad.com/>).

Abbreviations

ASCs: Adipose stem cells; CHIP: Chromatin immunoprecipitation; GSAT: Glute-ofemoral subcutaneous adipose tissue; HOTAIR: HOX Transcript Antisense RNA; LncRNA: Long non-coding RNA; PRC2: Polycomb repressor complex 2.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-022-08887-w>.

Additional file 1: Figure S1. Validation of adipogenic differentiation efficiency. **Figure S2.** Genome browser view of reference HOTAIR transcripts for SQANTI characterization. **Figure S3.** Semi-quantitative RT-PCR replicates.

Additional file 2: Table S1. SQANTI characterization of HOTAIR isoforms identified with TAMA. **Table S2.** List of human polyA motifs for SQANTI analysis. **Table S3.** Snaptron quantification of exon junctions for the HOTAIR locus. **Table S4.** Snaptron summary of RNA sequencing data containing HOTAIR E3.1-E4 splice junction. **Table S5.** List of primers used in this study.

Additional file 3. BED format file of top 34 HOTAIR isoforms identified in the study. High quality isoforms (black) were further studied while lower quality isoforms (red) were filtered out.

Additional file 4: Figure S4. Uncropped gels for Fig.3f and Additional file 1, FigS3b left panel. **Figure S5.** Uncropped gels for Additional file 1, Fig. S3c. **Figure S6.** Uncropped gels for Fig.4g and Additional file 1, Fig. S3b right panel. **Figure S7.** Uncropped gels for Fig.5b.

Acknowledgements

We thank Anita Løvstad Sørensen for technical assistance and the Norwegian Sequencing Center (Oslo University Hospital) for professional services.

Authors' contributions

EP and NB generated data. SHP performed bioinformatic analysis. ATK performed HOTAIR capture and long-read sequencing. EP, SHP and NB interpreted data and made figures. NB supervised the work. EP, SHP, PC and NB wrote or revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was funded by South-East Health Norway (grant 40040) and the Research Council of Norway (grant No. 249734 and 313508).

Availability of data and materials

The PacBio dataset supporting the conclusions of this article is available in the SRA repository under accession PRJNA730802 (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA730802>) and the filtered transcript list was submitted to GenBank (see accession numbers in Additional file 1, Table S1).

Bulk short-read RNA-Seq data has been deposited in the Gene Expression Omnibus (GEO) under accession GSE176020.

ATAC-seq data were obtained from GEO accession GSE118500 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118500>), GSE139571 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139571>) and GSE157399 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157399>).

CAGE clusters and transcripts from [35] were obtained from FANTOM (https://fantom.gsc.riken.jp/5/suppl/Hon_et_al_2016/data/assembly/lv3_robust/).

Other HOTAIR transcripts were downloaded from ensembl v95 [60] (<http://jan2019.archive.ensembl.org/index.html>) or from the UCSC browser at [61]. Regulatory elements (REMs) [42] associated with the HOTAIR locus were queried from the EpiRegio database (<https://epiregio.de/geneQuery/>; accessed: 23/04/2021).

Declarations

Ethics approval and consent to participate

The study was approved by the Regional Committee for Research Ethics for Southern Norway (REK 2013/2102 and REK 2018–660), and all subjects gave written informed consent.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Molecular Medicine, Institute of Basic Medical Sciences, Faculty of Medicine, University of Oslo, PO Box 1112, 0317 Oslo, Blindern, Norway. ²Department of Immunology and Transfusion Medicine, Oslo University Hospital, 0424 Oslo, Norway. ³Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Oslo, Norway.

Received: 24 March 2022 Accepted: 9 September 2022

Published online: 17 September 2022

References

- Cantile M, Di Bonito M, De TraceyBellis M, Botti G. Functional Interaction among lncRNA HOTAIR and MicroRNAs in Cancer and Other Human Diseases. *Cancers*. 2021;13(3):570.
- Kogo R, Shimamura T, Mimori K, Kawahara K, Imoto S, Sudo T, et al. Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res*. 2011;71:6320–6.
- Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*. 2010;464:1071–6.
- Xue X, Yang YA, Zhang A, Fong K-W, Kim J, Song B, et al. LncRNA HOTAIR enhances ER signaling and confers tamoxifen resistance in breast cancer. *Oncogene*. 2016;35:2746–55.
- Milevskiy MJG, Al-Ejeh F, Saunus JM, Northwood KS, Bailey PJ, Betts JA, et al. Long-range regulators of the lncRNA HOTAIR enhance its prognostic potential in breast cancer. *Hum Mol Genet*. 2016;25:3269–83.
- Pinnick KE, Nicholson G, Manolopoulos KN, McQuaid SE, Valet P, Frayn KN, et al. Distinct developmental profile of lower-body adipose tissue defines resistance against obesity-associated metabolic complications. *Diabetes*. 2014;63:3785–97.

7. Divoux A, Karastergiou K, Xie H, Guo W, Perera RJ, Fried SK, et al. Identification of a novel lncRNA in gluteal adipose tissue and evidence for its positive effect on preadipocyte differentiation. *Obesity*. 2014;22:1781–5.
8. Wu Y, Zhang L, Zhang L, Wang Y, Li H, Ren X, et al. Long non-coding RNA HOTAIR promotes tumor cell invasion and metastasis by recruiting EZH2 and repressing E-cadherin in oral squamous cell carcinoma. *Int J Oncol*. 2015;46:2586–94.
9. Wu Y, Liu J, Zheng Y, You L, Kuang D, Liu T. Suppressed expression of long non-coding RNA HOTAIR inhibits proliferation and tumorigenicity of renal carcinoma cells. *Tumour Biol*. 2014;35:11887–94.
10. Ding W, Ren J, Ren H, Wang D. Long Noncoding RNA HOTAIR Modulates MiR-206-mediated Bcl-w Signaling to Facilitate Cell Proliferation in Breast Cancer. *Sci Rep*. 2017;7:17261.
11. Wu Z-H, Wang X-L, Tang H-M, Jiang T, Chen J, Lu S, et al. Long non-coding RNA HOTAIR is a powerful predictor of metastasis and poor prognosis and is associated with epithelial-mesenchymal transition in colon cancer. *Oncol Rep*. 2014;32:395–402.
12. Kalwa M, Hänzelmann S, Otto S, Kuo C-C, Franzen J, Jousen S, et al. The lncRNA HOTAIR impacts on mesenchymal stem cells via triple helix formation. *Nucleic Acids Res*. 2016;44:10631–43.
13. Yoon J-H, Abdelmohsen K, Kim J, Yang X, Martindale JL, Tominaga-Yamanaka K, et al. Scaffold function of long non-coding RNA HOTAIR in protein ubiquitination. *Nat Commun*. 2013;4:2939.
14. Xu F, Zhang J. Long non-coding RNA HOTAIR functions as miRNA sponge to promote the epithelial to mesenchymal transition in esophageal cancer. *Biomed Pharmacother*. 2017;90:888–96.
15. Yu G-J, Sun Y, Zhang D-W, Zhang P. Long non-coding RNA HOTAIR functions as a competitive endogenous RNA to regulate PRAF2 expression by sponging miR-326 in cutaneous squamous cell carcinoma. *Cancer Cell Int*. 2019;19:270.
16. Zhao Y-H, Liu Y-L, Fei K-L, Li P. Long non-coding RNA HOTAIR modulates the progression of preclampsia through inhibiting miR-106 in an EZH2-dependent manner. *Life Sci*. 2020;253:117668.
17. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Bruggmann SA, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*. 2007;129:1311–23.
18. Tsai M-C, Manor O, Wan Y, Mosammamaparast N, Wang JK, Lan F, et al. Long noncoding RNA as modular scaffold of histone modification complexes. *Science*. 2010;329:689–93.
19. Balas MM, Hartwick EW, Barrington C, Roberts JT, Wu SK, Bettcher R, et al. Establishing RNA-RNA interactions remodels lncRNA structure and promotes PRC2 activity. *Sci Adv*. 2021;7(16):eabc9191.
20. Graf J, Kretz M. From structure to function: route to understanding lncRNA mechanism. *BioEssays*. 2020;42:e2000027.
21. He S, Liu S, Zhu H. The sequence, structure and evolutionary features of *HOTAIR* in mammals. *BMC Evol Biol*. 2011;11:102.
22. Hajjari M, Rahnama S. HOTAIR Long Non-coding RNA: Characterizing the Locus Features by the In Silico Approaches. *Genomics Inform*. 2017;15:170–7.
23. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kocicinski F, et al. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res*. 2012;22:1760–74.
24. O'Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44:D733–45.
25. Schorderet P, Duboule D. Structural and functional differences in the long non-coding RNA hotair in mouse and human. *PLoS Genet*. 2011;7:e1002071.
26. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddloh JA, et al. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol*. 2011;30:99–104.
27. Uszczyńska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet*. 2018;19:535–48.
28. Zampetaki A, Albrecht A, Steinhofel K. Long non-coding RNA structure and function: is there a link? *Front Physiol*. 2018;9:1201.
29. Boquest AC, Collas P. Obtaining freshly isolated and cultured mesenchymal stem cells from human adipose tissue. *Methods Mol Biol*. 2012;879:269–78.
30. Meredith EK, Balas MM, Sindy K, Haislop K, Johnson AM. An RNA matchmaker protein regulates the activity of the long noncoding RNA *HOTAIR*. *RNA*. 2016;22:995–1010.
31. Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One*. 2015;10:e0132628.
32. Tseng E. cDNA_Cupcake Wiki. Github. https://github.com/Magdoll/cDNA_Cupcake.
33. Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, Archibald AL, et al. Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics*. 2020;21:751.
34. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, et al. SQUANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res*. 2018.<https://doi.org/10.1101/gr.222976.117>.
35. Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, et al. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*. 2017;543:199–204.
36. Imada EL, Sanchez DF, Collado-Torres L, Wilks C, Matam T, Dinalankara W, et al. Recounting the FANTOM CAGE-Associated Transcriptome. *Genome Res*. 2020;30:1073–81.
37. Musiri MM, Carmona MC, Hanzu FA, Kaliman P, Gomis R, Párrizas M. Histone demethylase LSD1 regulates adipogenesis. *J Biol Chem*. 2010;285:30034–41.
38. Chen Y, Kim J, Zhang R, Yang X, Zhang Y, Fang J, et al. Histone Demethylase LSD1 Promotes Adipocyte Differentiation through Repressing Wnt Signaling. *Cell Chem Biol*. 2016;23:1228–40.
39. Wang L, Jin Q, Lee J-E, Su I-H, Ge K. Histone H3K27 methyltransferase Ezh2 represses Wnt genes to facilitate adipogenesis. *Proc Natl Acad Sci U S A*. 2010;107:7317–22.
40. Wu L, Murat P, Matak-Vinkovic D, Murrell A, Balasubramanian S. Binding interactions between long noncoding RNA HOTAIR and PRC2 proteins. *Biochemistry*. 2013;52:9519–27.
41. Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics*. 2018;34:114–6.
42. Baumgarten N, Hecker D, Karunanithi S, Schmidt F, List M, Schulz MH. EpiRegio: analysis and retrieval of regulatory elements linked to genes. *Nucleic Acids Res*. 2020;48:W193–9.
43. Schmidt F, Marx A, Baumgarten N, et al. Integrative analysis of epigenetics data identifies gene-specific regulatory elements. *Nucleic Acids Res*. 2021;49(18):10397–10418.
44. Divoux A, Sandor K, Bojcsuk D, Talukder A, Li X, Balint BL, et al. Differential open chromatin profile and transcriptomic signature define depot-specific human subcutaneous preadipocytes: primary outcomes. *Clin Epigenetics*. 2018;10:148.
45. Divoux A, Sandor K, Bojcsuk D, Yi F, Hopf ME, Smith JS, et al. Fat distribution in women is associated with depot-specific transcriptomic signatures and chromatin structure. *J Endocr Soc*. 2020;4:bvaa042.
46. Martone J, Lisi M, Castagnetti F, Rosa A, Di Carlo V, Blanco E, et al. Transgenerational epigenetic regulation associated with the amelioration of Duchenne Muscular Dystrophy. *EMBO Mol Med*. 2020;12:e12063.
47. Fraser LCR, Dikdan RJ, Dey S, Singh A, Tyagi S. Reduction in gene expression noise by targeted increase in accessibility at gene loci. *Proc Natl Acad Sci U S A*. 2021;118:e2018640118.
48. Tian F-J, He X-Y, Wang J, Li X, Ma X-L, Wu F, et al. Elevated tristetraprolin impairs trophoblast invasion in women with recurrent miscarriage by destabilization of *HOTAIR*. *Mol Ther Nucleic Acids*. 2018;12:600–9.
49. Ziegler C, Kretz M. The more the merrier-complexity in long non-coding RNA loci. *Front Endocrinol*. 2017;8:90.
50. Siang DTC, Lim YC, Kyaw AMM, Win KN, Chia SY, Degirmenci U, et al. The RNA-binding protein HuR is a negative regulator in adipogenesis. *Nat Commun*. 2020;11:213.
51. Mozdaran H, Ezzatizadeh V, Rahbar PR. The emerging role of the long non-coding RNA *HOTAIR* in breast cancer development and treatment. *J Transl Med*. 2020;18:152.
52. Hajjari M, Salavaty A. HOTAIR: an oncogenic long non-coding RNA in different cancers. *Cancer Biol Med*. 2015;12:1–9.
53. Bhan A, Mandal SS. lncRNA *HOTAIR*: a master regulator of chromatin dynamics and cancer. *Biochim Biophys Acta*. 2015;1856:151–64.

54. Guo C-J, Ma X-K, Xing Y-H, Zheng C-C, Xu Y-F, Shan L, et al. Distinct processing of lncRNAs contributes to non-conserved functions in stem cells. *Cell*. 2020;181:621–36.e22.
55. Yuan J-H, Liu X-N, Wang T-T, Pan W, Tao Q-F, Zhou W-P, et al. The MBNL3 splicing factor promotes hepatocellular carcinoma by increasing PXN expression through the alternative splicing of lncRNA-PXN-AS1. *Nat Cell Biol*. 2017;19:820–32.
56. Li R, Harvey AR, Hodgetts SJ, Fox AH. Functional dissection of NEAT1 using genome editing reveals substantial localization of the NEAT1_1 isoform outside paraspeckles. *RNA*. 2017;23:872–81.
57. Spokoini-Stern R, Stamoov D, Jessel H, Aharoni L, Haschke H, Giron J, et al. Visualizing the structure and motion of the long noncoding RNA *HOTAIR*. *RNA*. 2020;26:629–36.
58. Oldenburg AR, Delbarre E, Thiede B, Vigouroux C, Collas P. Deregulation of Fragile X-related protein 1 by the lipodystrophic lamin A p.R482W mutation elicits a myogenic gene expression program in preadipocytes. *Hum Mol Genet*. 2014;23:1151–62.
59. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–100.
60. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. *Nucleic Acids Res*. 2018;47:D745–51.
61. Schema for lincRNA TUCP - lincRNA and TUCP transcripts. https://genome.ucsc.edu/cgi-bin/hgTables?db=hg38&hgta_group=genes&hgta_track=lincRNAsTranscripts&hgta_table=lincRNAsTranscripts&hgta_doSchema=describe+table+schema. Accessed 28 Sep 2021.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

