Troskie et al. Genome Biology (2021) 22:146
https://doi.org/10.1186/s13059-021-02369-0

Genome Biology

SHORT REPORT Open Access

Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome

Robin-Lee Troskie[1], Yohaann Jafrani[1], Tim R. Mercer[2], Adam D. Ewing[1*], Geoffrey J. Faulkner[1,3*] and Seth W. Cheetham[1*]

* Correspondence: adam.ewing@mater.uq.edu.au; faulknergj@gmail.com; seth.cheetham@mater.uq.edu.au
[1]Mater Research Institute-University of Queensland, TRI Building, QLD 4102 Woolloongabba, Australia
Full list of author information is available at the end of the article

Abstract

Pseudogenes are gene copies presumed to mainly be functionless relics of evolution due to acquired deleterious mutations or transcriptional silencing. Using deep full-length PacBio cDNA sequencing of normal human tissues and cancer cell lines, we identify here hundreds of novel transcribed pseudogenes expressed in tissue-specific patterns. Some pseudogene transcripts have intact open reading frames and are translated in cultured cells, representing unannotated protein-coding genes. To assess the biological impact of noncoding pseudogenes, we CRISPR-Cas9 delete the nucleus-enriched pseudogene PDCL3P4 and observe hundreds of perturbed genes. This study highlights pseudogenes as a complex and dynamic component of the human transcriptional landscape.

Keywords: Pseudogene, PacBio, Long-read, lncRNA, CRISPR

Background

Pseudogenes are gene copies which are thought to be defective due to frame-disrupting mutations or transcriptional silencing [1, 2]. Most human pseudogenes (72%) are derived from retrotransposition of processed mRNAs, mediated by proteins encoded by the LINE-1 retrotransposon [3, 4]. Due to the loss of parental *cis*-regulatory elements, processed pseudogenes were initially presumed to be transcriptionally silent [1] and were excluded from genome-wide functional screens and most transcriptome analyses [2]. Transcriptomic surveys of cancer [5] and normal human tissues [6] by high-throughput short-read sequencing suggest that pseudogene transcription may be widespread. However, studies of pseudogene transcription are hindered by the limited capacity of short-read sequencing, and microarray hybridisation, to discriminate pseudogenes from their highly similar parent genes [2, 7]. Most full-length pseudogene transcripts found to date were identified by relatively low-throughput capillary sequencing of full-length cDNA libraries [8–10]. As a result, the extent of the human pseudogene transcriptome in most spatiotemporal contexts remains largely unresolved.

Pseudogene transcripts can control the expression of their parent genes by acting as competitive endogenous RNAs [11] (ceRNAs), antisense transcripts [12], precursors for small interfering RNAs [13, 14] (siRNAs), and piwi-interacting RNAs [15] (piRNAs). Whilst most pseudogenes are presumed to act by noncoding mechanisms, some retain the capacity to encode full-length or truncated proteins [16–19].

Results and discussion

Long-read cDNA sequencing via Pacific Biosciences Isoform Sequencing (PacBio Iso-Seq) or Oxford Nanopore Technologies is a potentially powerful approach to identify full-length pseudogene transcripts and accurately differentiate pseudogenes and their parent mRNAs. PacBio Iso-Seq is particularly suitable for this application due to the high consensus accuracy enabled by circular consensus reads. To comprehensively survey the human processed pseudogene transcriptome, we sequenced high quality RNA from 20 normal mixed adult and foetal human tissues (Qiagen XpressRef Universal Total RNA) on a Sequel II platform (Fig. 1a) [20]. To further broaden the biological scope of our analysis, we integrated our data with a deep PacBio in-house Sequel II dataset of 6,775,127 full-length reads from a mixture of 10 human cell lines [21]. We aligned the reads to the human reference genome (hg38) at high stringency (q60) and compared the identified transcript isoforms to Gencode [22] annotations using SQANTI2, a bioinformatics QC tool designed to annotate full-length transcript (Iso-Seq) data with respect to a reference transcriptome [23].

We identified 1170 transcripts, each supported by at least two full-length reads, that overlapped 521 processed pseudogenes. Two hundred twenty pseudogenes (318 transcripts) transcribed in sense (the same orientation as their parent gene) were independent (non-intronic and have greater overlap with the pseudogene than other gene models) of known genes, only 43 of which were previously annotated as transcribed pseudogenes in Gencode (Fig. 1b, for examples see Additional file 1: Figure S1; Additional file 2: Table S1). All identified transcripts were poly-adenylated (a requirement of the Iso-Seq library prep and analysis pipeline) and 175/318 (55%) contain a canonical polyA motif within 100 bp of the 3′ terminus. One hundred one of these transcripts were multi-exonic, and the vast majority (84%) did not incorporate splice junctions with known Gencode transcripts. Pseudogenes are typically transcribed in the same orientation as their parent genes [25]. However, we found 78/396 (20%) of independent pseudogene transcripts were produced in antisense with respect to their parent gene (for examples see Additional file 1: Figure S2a-b; Additional file 3: Table S2). In contrast, only 168/2669 (6.3%) of unprocessed pseudogenes are transcribed in antisense (Additional file 1: Figure S2c). This difference may be attributable either to the propensity for novel downstream promoter elements to regulate expression of a retrotransposed pseudogene relative to unprocessed pseudogenes (which retain their parental promoter) or to selection for regulatory potential. Manual inspection of the antisense pseudogene transcription start sites (TSSs) did not reveal an obvious initiation site bias. Antisense pseudogene transcripts have significant potential to regulate their parent genes by antisense-mediated translational inhibition or by processing into siRNAs [12–14]. In support of the pseudogene transcripts identified here being full-length, we intersected our data with an atlas of Cap Analysis Gene Expression 5′ mRNA sequencing (CAGE-seq) data generated by FANTOM5 [24]. CAGE signal was highly enriched at pseudogene TSSs

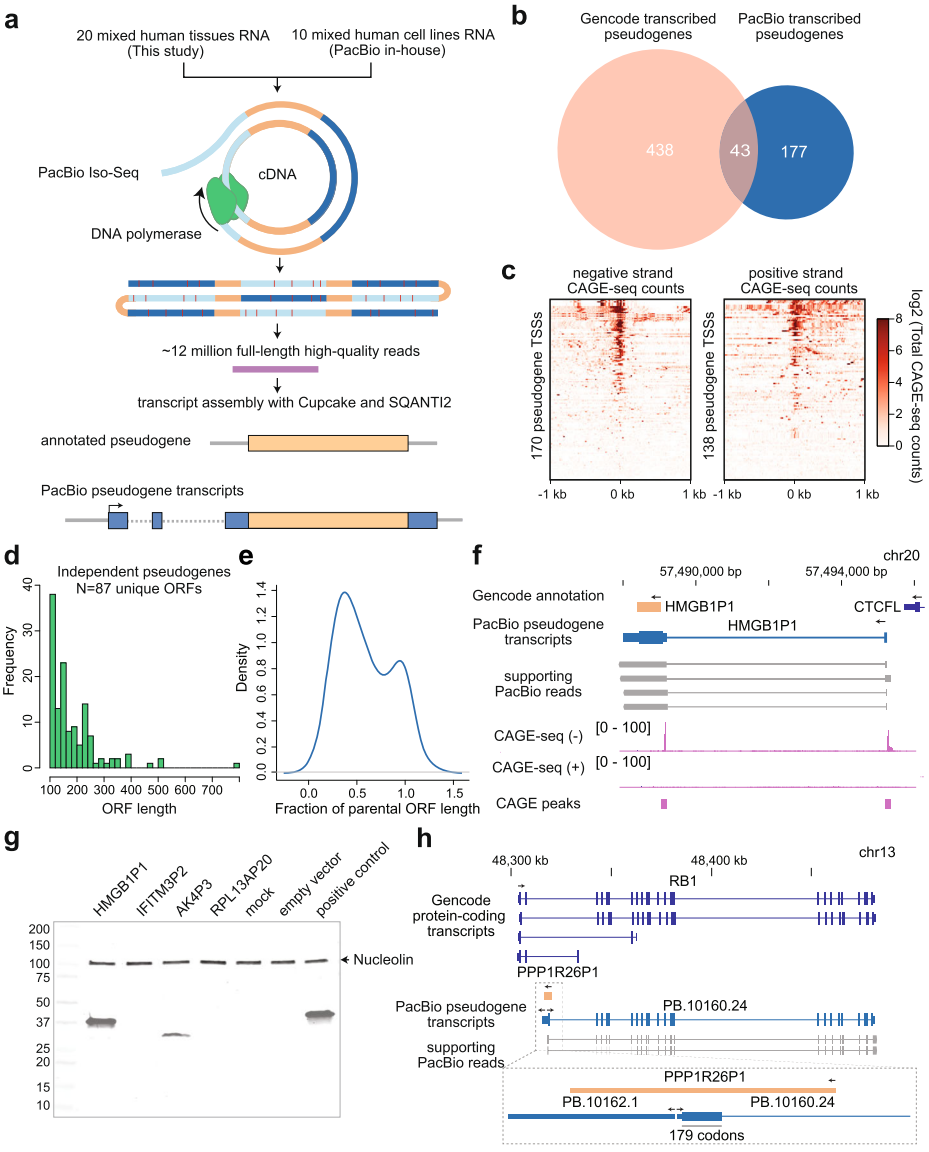

Fig. 1 Long-read cDNA sequencing elucidates the human pseudogene transcriptome. **a** Full-length consensus PacBio cDNA reads from normal tissues and cell lines were compared to Gencode annotations to generate a pseudogene transcriptome. **b** Most transcribed pseudogenes identified here were absent from Gencode. **c** The transcription start sites (TSSs) of full-length pseudogene transcripts are enriched for CAGE-seq signal (data from FANTOM5 [24]). **d** Open Reading Frame (ORF) lengths of potentially coding-independent pseudogene transcripts. **e** Fraction of parental ORF length found intact in transcribed potentially coding-independent pseudogenes. **f** HMGB1P1 has a novel 5′ exon and is transcribed from an upstream CAGE-confirmed TSS. **g** Expression of 3XHA-tagged pseudogene ORFs in HEK293T cells detected by Western blot. HMGB1P1 and AK4P3 are translated when expressed in cultured cells. **h** A novel isoform of retinoblastoma (RB1) is transcribed from a TSS located within the pseudogene PPP1R26P1. The pseudogene sequence adds 179 codons to the RB1 ORF

(Fig. 1c) and 41% of pseudogene TSSs were within 100 bp of a FANTOM5 [24] CAGE peak, indicating that a large fraction of pseudogene transcripts have accurate 5′ ends. The proportion of pseudogene transcripts that overlap CAGE peaks is lower than for protein-coding transcripts (69%) and comparable to lincRNA transcripts (44%) (Additional file 1: Figure S3). Fifty-one percent of independent pseudogene transcripts were supported by the cutoff of two full-length reads (Additional file 1: Figure S4), and our datasets do not

comprehensively capture diversity of human cell-types and developmental stages. Therefore, it is probable that our data still significantly underestimate pseudogene transcription; further transcripts would very likely be identified by increased sequencing depth applied to individual tissues or cell types.

Next, we annotated the coding potential of independent pseudogene transcripts using SQANTI2. One hundred sixty of 318 pseudogene transcripts (50%) encode putative proteins that are > 100 amino acids in length (Fig. 1d) and, strikingly, 53 of the pseudogene open reading frames (ORFs) were > 90% of the length of the parent gene ORF (Fig. 1e). An illustrative example of a potentially coding pseudogene transcript is the processed pseudogene of the high mobility group box 1 on chromosome 20 (HMGB1P1). The Gencode HMGB1P1 annotation is a single contiguous region of 98% identity to the HMGB1 ORF, which has no introns (Fig. 1f). Iso-Seq revealed that HMGB1P1 was transcribed from an upstream promoter, which yields a novel 5′ exon, and was supported by a FANTOM5 CAGE peak. HMGB1P1 contained no frameshift mutations and encoded a protein of the same length as HMGB1, with an intact HMG domain. To assess the coding potential of HMGB1P1 and other pseudogene transcripts, we amplified the 5′ exons and coding sequence of four spliced pseudogenes with intact ORFs (HMGB1P1, AK4P3, IFITM3P2 and RPL13AP20) and cloned them into a vector with a C-terminus 3XHA tag. Transfection into HEK293T cells resulted in clear translation of the HMGB1P1 and AK4P3 transcripts (Fig. 1g, Additional file 1: Figure S5; Additional file 4). To further substantiate that pseudogenes can be translated in vivo we interrogated the neXtprot human proteomics database [26]. Eleven potentially coding pseudogenes have entries in neXtprot of which four, HMGB1P1, SUMO1P1, MSL3P1, and PLEKHA8P1, have matched unique peptides (Additional file 1: Figure S6). Thus, pseudogene transcripts can encode intact proteins that are translated in human cells.

To determine if the pseudogene ORFs are subject to purifying selection, we identified orthologous positions in non-human primate genomes by aligning the human transcripts with a 1000 bp window on each side to higher primates (chimp, gorilla, orangutan, and rhesus). The alignments were then further refined to identify orthologous cDNA sequences and the resulting ORFs were translated into amino acid sequences (the "Methods" section). The extent of selection on pseudogene ORFs was determined by maximum likelihood estimation of the ratio of substitution rates between two divergent species that result in nonsynonymous vs synonymous changes (dN/dS). A ratio of > 1 suggests diversifying selection whilst a ratio of < 1 is consistent with purifying selection. This index has been used as evidence of conserved function for human processed pseudogenes [27]. Of the pseudogene ORFs which are conserved in rhesus and have sufficient nucleotide diversity, 29/35 (83%) have a dN/dS < 1 (median 0.4483) suggesting that most of the conserved pseudogene ORFs were under purifying selection across 25 M years of evolution (Additional file 5: Table S3).

In addition to independent protein-coding potential, pseudogenes can contribute to the coding sequences of known genes. We found that 93 protein-coding genes contained coding sequences derived from pseudogenes, often adding hundreds of codons (Additional file 1: Figure S7; Additional file 6: Table S4). Notably, the pseudogene PPP1R26P1 constitutes most of a novel 5′ exon fused to the major tumour suppressor gene retinoblastoma (RB1), adding 179 codons to RB1 from the antisense strand of PPP1R26P1 (Fig. 1h). Indeed, PPP1R26P1 was previously shown to constitute an

alternative imprinted RB1 promoter [28]. Gene-pseudogene fusion transcripts can add novel domains to genes, such as the case of HMGN1P18, which adds a HMGN domain to CPED1 (Additional file 1: Figure S8a). Even well-characterised long noncoding RNAs (lncRNAs) may have isoforms that encode pseudogene proteins, including the lncRNA FIRRE [29], which has an isoform that splices into an intact MCRIP2 pseudogene (Additional file 1: Figure S8b).

To better evaluate spatial patterns of pseudogene transcription, we aligned deep RNA-sequencing from 16 adult tissues of the Illumina Body Map 2.0 [30] to our independent pseudogene annotations (Additional file 7: Table S5). PacBio-identified pseudogene transcripts were highly tissue-specific (Additional file 1: Figure S9) and divergent from expression of their parent genes. For example, AMD1P4 and YWHAEP1 expression is, respectively, liver- and testis-specific (Additional File 1: Figure S10a-b) whilst their parent genes are broadly expressed (Additional file 1: Figure S10c-d), indicating that pseudogene expression is controlled by distinct regulatory elements. Short-read sequencing can therefore be leveraged to quantify the expression of pseudogene transcripts discovered by long-read sequencing.

To determine if long-read sequencing outperforms short-read sequencing at assembling pseudogene transcripts, we generated matched PacBio and Illumina datasets from the haploid leukaemia cell line HAP1 [20]. Without a reference transcriptome, 71% of the 163 HAP1 PacBio pseudogene transcripts (Additional file 8: Table S6) were detected by short-read transcript assembly with StringTie [31], whilst with Gencode (v29) as a reference, 91% of pseudogene transcripts were detected. However, the reference-guided short-read assembled transcripts were significantly shorter than the PacBio transcripts (average of 750 bp shorter, $p = 8.5E^{-11}$ Mann-Whitney) indicating that these assemblies do not cover complete transcripts (Additional file 1: Figure S11a-b).

Transcribed pseudogenes can regulate gene expression through coding-independent mechanisms. Haploid cells are ideally suited to genetic manipulation as only a single allele needs to be inactivated for complete loss-of-function [32, 33]. Among 163 independent pseudogene transcripts, we identified PDCL3P4 as being highly expressed from a human endogenous retrovirus-K (HERV-K) long terminal repeat (LTR) promoter on chromosome 3 (Fig. 2a). PDCL3P4 is derived from retrotransposition of phosducin-like 3 (PDCL3), a putative chaperone protein implicated in angiogenesis and proliferation [34]. Unlike the well-characterised pseudogene PTENP1 [11, 35] (Additional file 1: Figure S12a), expression of PDCL3P4 expression does not correlate with that of its parent gene, indicating they are likely regulated independently (Additional file 1: Figure S12b). As a route to test the regulatory impact of the PDCL3P4 locus, we deleted the pseudogene from HAP1 cells with CRISPR-Cas9 genome engineering by directing a Cas9 endonuclease-guide RNA (gRNA) complex to unique genomic regions flanking PDCL3P4. Three independent clonal PDCL3P4 knockout lines were generated with a combination of gRNAs to reduce the risk of off-target mutations. Genotyping of the PDCL3P4 locus in knockout cells revealed that two of the lines, null1 and null3, contained complete deletions (Fig. 2b) whilst the remaining line (null2) contained a complete deletion and a 154 bp insertion at the site of CRISPR-Cas9 mutagenesis. Replicate RNA-seq confirmed PDCL3P4 expression was entirely abrogated in each knockout line (Fig. 2c, null1: $N = 2$, null2: $N = 4$, null3: $N = 4$) [20]. Additionally, 137 differentially expressed genes (DEGs) were detected in the

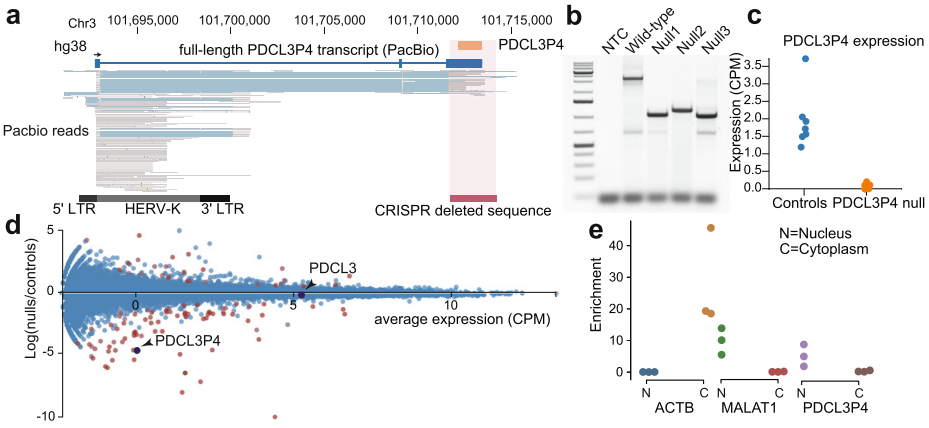

Fig. 2 Deletion of PDCL3P4 impacts the transcriptome of haploid cells. **a** PDCL3P4 is a pseudogene transcribed in HAP1 cells from the canonical long terminal repeat (LTR) promoter of an upstream human endogenous retrovirus-K (HERV-K) sequence. Grey bars within the PacBio reads represent exons and light blue bars represent introns. **b** CRISPR-Cas9 genome engineering removes the retroposed portion of PDCL3P4 from the HAP1 genome in three independent clones. **c** PDCL3P4 expression is ablated in PDCL3P4 mutant clones. **d** PDCL3P4 ablation disrupted the expression of more than 137 genes, while PDCL3 expression was unaffected. **e** PDCL3P4 transcripts were enriched in the nucleus. The nuclear-localised noncoding RNA MALAT1 and mRNA ACTB act as controls

three knockout lines, compared to wild-type cells (N = 4) and a control clone (N = 3) in which PDCL3P4 deletion was unsuccessful (FDR = 0.01, Additional file 9: Table S7) (Fig. 2d). Changes in gene expression were consistent between the independent clonal knockouts, indicating that these expression changes do not represent off-targets. PDCL3 expression was unaffected in the knockout lines, as were any genes within 200 kb of PDCL3P4. Perturbation of unannotated *cis*-regulatory elements within PDCL3P4 was therefore unlikely to drive downstream differential gene expression, although this does not fully exclude the potential caveat of CRISPR-Cas9 genetic manipulation otherwise impacting transcription. The PDCL3P4 ORF was highly disrupted, suggesting the pseudogene could act as a lncRNA. Indeed, PDCL3P4 transcripts were enriched 5.17-fold in the nucleus compared to the cytoplasm in wild-type HAP1 cells (Fig. 2e), consistent with the subcellular localisation of a large fraction of lncRNAs to the nucleus [36–38]. Therefore, transcribed noncoding pseudogenes may impact the transcriptome in a parent gene-independent manner.

Conclusions

Here, we define a complex tissue-specific pseudogene transcriptome using PacBio long-read sequencing, validated by orthogonal CAGE-seq and RNA-seq datasets. This high-quality annotation can be utilised as a resource for transcriptomic analyses and to design functional screens. We contribute to the growing body of evidence that pseudogene translation may be widespread [39–42] and provide proof-of-principle evidence that noncoding pseudogenes can regulate the cellular transcriptome by mechanisms independent of their parent gene. Future work will elucidate the mechanism through which PDCL3P4 affects gene expression. Collectively, these data suggest pseudogenes have prematurely been assumed to be functionless and numerous annotated pseudogenes produce protein-coding or noncoding transcripts. This study is a foundation for the use of long-read transcriptome sequencing to comprehensively identify full-length

pseudogene transcripts and thereby better understand a major and underappreciated component of the transcriptional landscape and its impact on human biology and pathology.

Methods

PacBio sequencing

XpressRef Universal Total RNA (Qiagen cat # 338112) and HAP1 (Horizon) RNA (RIN 10) was sequenced on the Pacific Biosciences Sequel II platform at the University of Maryland Institute for Genome Sciences. Libraries were prepared using the PacBio Iso-Seq library preparation, which amplifies full-length polyadenylated transcripts without size selection. Two 8M flow cells were used to generate the Universal Total RNA data (6,002,282 full-length reads) and one 8M flow cell was used for the HAP1 data (4,098,069 full-length reads).

PacBio data processing

Full-length circular consensus reads were identified with lima v*1.10.0* (https://github.com/PacificBiosciences/barcoding) using the settings lima --isoseq --dump-clips --no-pbi --peek-guess -j 24 removing the primer sequences:

>NEB_5p GCAATGAAGTCGCAGGGTTGGG
>NEB_Clontech_3p GTACTCTGCGTTGATACCACTGCTT

Transcripts were refined and polyA tails removed with isoseq3 v3.2.2 (https://github.com/PacificBiosciences/IsoSeq) *refine* --require-polya. Full-length reads were converted to fastq format using bamtools v2.5.1 [43] *convert*. High-quality clustered reads were aligned with minimap2 [44] v2.17-r941 with the settings -ax splice --secondary=no -C5 -O6,24 -B4 -uf. Redundant isoforms were collapsed with cDNA Cupcake (https://github.com/Magdoll/cDNA_Cupcake) 9.1.1 collapse_isoforms_by_sam.py --dun-merge-5-shorter. 5′ degraded transcripts were removed with filter_away_subset.py and transcript abundance counted with get_abundance_post_collapse.py. SQANTI2 v6.0.0 (https://github.com/Magdoll/SQANTI2) was used to classify the high quality clustered non-redundant reads with respect to Gencode [22] (v29) and FANTOM5 [24] CAGE peaks.

Identification of pseudogene transcripts

Pseudogene transcripts were identified by intersecting the PacBio transcripts with Gencode pseudogenes [25] (v29) using bedtools v2.29.2 [45]. To confirm that these transcripts intersected directly with a retrotransposed copy (rather than with another exon of spliced transcript that is annotated as a transcribed pseudogene), we further intersected these transcripts with the retrogenes.v9 [10] track downloaded from the UCSC genome browser [46]. Independent pseudogene transcripts were classified as those assigned the name of a Gencode pseudogene by SQANTI2. Polyadenylation motifs were determined by extraction the sequences of the 100 bp using bedtools v2.29.2 getfasta and scanning for the motifs AATAAA and ATTAAA using SeqKit [47].

Illumina Human BodyMap quantification

Sequence data in .fastq format was obtained from the Illumina Human BodyMap 2.0 Project [30] (SRA accession PRJNA144517) and aligned to the Ensembl GRCh38 primary assembly with STAR 2.7.3a [48]. The STAR reference was built using Ensembl build 101 [49] with the independent pseudogene models added to the gtf file. Reads were aligned with default parameters with the exception of --outFilterMultimapNmax 1 to limit multi-mapping reads. Reads were counted against the independent pseudogene model using htseq-count 0.11.2 [50]. Single-end and paired-end reads were counted separately across sequence runs and summed into a single count per-tissue. TMM-normalised cpm values were produced using edgeR 3.24.3 [51] and transformed to log2(cpm + 1). Read mappings were visualised in IGV [52].

Conservation analysis

Conservation was assessed by aligning each pseudogene transcript and a 1000 bp flank on both ends to the human (hg38), chimp (panTro6), gorilla (gorGor6), orangutan (ponAbe3), and rhesus (rheMac10) genome assemblies using BLAT [53] (gfServer -stepSize = 5). The human cDNA (derived from "CDS" entries in the input .gff file) was aligned within the larger transcribed region using exonerate [54] in "cdna2genome" mode, and the highest scoring alignment was presumed to be the ortholog of the human cDNA. Where the cDNA sequence length was divisible by 3, the ORFs were translated and compared to the human ORF using exonerate in "ungapped" mode. For each transcript that contained at least one CDS and aligned to two or more species, pairwise dN/dS statistics were obtained as follows. Multiple sequence alignments of the ORF protein sequences were performed via clustal omega [55] with default parameters, cDNAs were codon-aligned using PAL2NAL [56], and pairwise dN/dS was computed via codeml from PAML 4.9j [57]. A python script for carrying out these methods is available at https://gist.github.com/adamewing/3a4cfa8eb1a333ee9c497538ce30b6db.

Cell culture

Low passage (≤ p10) HAP1 cells (Horizon) were cultured in Iscove's Modified Dulbecco's Medium (IMDM) (Gibco cat # 12440-053) supplemented with 10% foetal bovine serum (Sigma Aldrich cat # F2442) and 1% penicillin/streptomycin (Gibco cat # 15140-122) and grown in a tissue culture incubator (37 °C, 5% CO_2). Cells were not maintained above p18. HEK293T cells (ATCC) were cultured in Dulbecco's Modified Eagle Medium (DMEM) (Gibco cat # 21969035) supplemented with 10% foetal bovine serum (Sigma Aldrich cat # F2442) and 1% penicillin/streptomycin (Gibco cat # 15140-122) and grown in a tissue culture incubator (37 °C, 5% CO_2).

Custom guide RNA design

Genomic DNA flanking the PDCL3P4 locus was examined for evidence of enhancer marks or transcriptional activity using the GeneHancer and Layered H3K27Ac tracks on the UCSC Genome Browser. Approximately 700 bp of up- and downstream genomic sequence without evidence of functional activity was selected to design custom CRISPR-Cas9 gRNAs using the IDT Custom Alt-R® CRISPR-Cas9 Guide RNA Design Tool (https://sg.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE). Two

upstream and two downstream gRNAs were chosen based on optimal on- and off-target scores as well as by manual inspection of off-target hits to corresponding gRNA design.

CRISPR-Cas9 genome engineering

PDCL3P4 knockout lines were generated in HAP1 cells (Horizon) following the Alt-R CRISPR-Cas9 System: Cationic lipid delivery of CRISPR ribonucleoprotein complexes into mammalian cells protocol (IDT). Pools of low passage HAP1 cells were individually reverse transfected with alternating combinations of upstream and downstream gRNA (Additional file 10: Table S8): ribonucleoprotein (RNP) complexes labelled with a fluorescent dye (ATTO-550) using Lipofectamine CRISPRMAX Transfection Reagent (Thermo Fisher Scientific cat # CMAX00008).

Cells were incubated with the transfection complexes in a tissue culture incubator (37 °C, 5% CO_2) for 48 h and then prepared for fluorescence-activated cell sorting (FACS). Cells were stained using LIVE/DEAD® Fixable Aqua Dead Cell Stain (Thermo Fisher Scientific cat # L34966) following the manufacturer's instructions and then re-suspended in Hank's Balanced Salt Solution (HBSS) (Thermo Fisher Scientific cat # 14025076). Cell populations were gated on the BD FACSAria™ Fusion Sorter based on viability and a positive signal for ATTO-550. Single cells were sorted into individual wells in a 96-well tissue culture plate containing Iscove's Modified Dulbecco's Medium (IMDM) (Thermo Fisher Scientific cat # 12440053). Single cells were clonally expanded and genomic DNA was extracted from half the clonal population using QuickExtract™ DNA Extraction Solution (Epicentre cat # QE09050) following the manufacturer's instructions. Null1 and null2 were derived using upstream gRNA2 and downstream gRNA1, whilst null3 was generated using upstream gRNA2 and downstream gRNA2. The control clone unsuccessful for PDCL3P4 deletion was treated with upstream gRNA2 and downstream gRNA1.

Sanger sequencing

Individual clones were assessed for PDCL3P4 knockout by performing a genotyping PCR with Q5® High-Fidelity DNA Polymerase (New England BioLabs cat # M0492S) and primers placed outside of the gRNA cut sites. PCR products were run on a 1% agarose gel to inspect for the presence of a wild-type or knockout amplicon. Amplicons indicative of PDCL3P4 knockout were cut from the agarose gel and DNA extracted using the QIAquick Gel Extraction Kit (Qiagen cat # 28704). DNA was capillary sequenced by the Australian Genome Research Facility (AGRF) to validate the knockout.

Quantitative real-time PCR

RNA was extracted from validated clones and wild-type HAP1 cells using the RNeasy Mini Kit (Qiagen cat # 74104) following the manufacturer's instructions and then treated with TURBO DNA-free™ Kit (Life Technologies cat # AM1907) to remove genomic DNA. DNA-free RNA was used for quantitative real-time PCR (qRT-PCR) to validate the PDCL3P4 knockout. Primers for PDCL3P4 were designed to target a SNP-containing region to mitigate off-target binding to the parent gene. Both a standard curve and melt curve were performed using *Power* SYBR® Green RNA-to-CT™ 1-Step

Kit (Thermo Fisher Scientific cat # 4389986) to ensure optimal amplification efficiency and specificity of the primers. qRT-PCR was performed on RNA extracted from validated clones and wild-type HAP1 cells using the abovementioned kit on the Applied Biosystems ViiA™ 7 Real-Time PCR machine, and Ct values were normalised to ACTB expression. PDCL3P4 expression was compared between wild-type HAP1 cells and the knockout clones to validate the absence of expression in knockout clones (data not shown).

RNA-seq

The three validated PDCL3P4 knockout clones, wild-type HAP1 cells and a clone unsuccessful for the knockout were then prepared for RNA-seq (20 M PE150 reads) (Novogene). Cells from each condition (knockout, wild-type, unsuccessful knockout) were seeded in five replicates in a 6-well tissue culture plate and left to proliferate for 24 h. RNA was extracted from each well and genomic DNA removed using the abovementioned kits. A DNA contamination PCR was performed using MyTaq™ DNA Polymerase (Bioline cat # BIO-21105) and primers for ACTB to check for the presence of genomic DNA. RNA quality from DNA-free samples was measured on the Agilent 2100 Bioanalyser (Agilent cat # G2939BA) using the Agilent RNA 6000 Nano Kit (Agilent cat # 5067-1511). RNA concentration and purity were measured on the NanoDrop™ Lite Spectrophotometer (Thermo Fisher Scientific cat # ND-LITE-PR). The four RNA samples per condition with the highest RIN scores and purity (A260/280 ratio) were prepared for shipment. Two micrograms of RNA was dried in RNAstable® tubes (Sigma Aldrich cat # 93221-001-1KT) following manufacturer's instructions and heat sealed in a desiccant bag. Two out of four replicates for null1 and one out of four replicates for the control clone unsuccessful for PDCL3P4 knockout did not pass QC before library preparation. All other samples and replicates passed QC and underwent library preparation for sequencing. RNA-seq data was analysed using STAR 2.7.3a [47] to align RNA-seq reads to GRCh38, htseq-count 0.11.2 [50] to quantify read counts against Ensembl genes GRCh38.83 [49], and EdgeR [50] for DEG analysis via Degust [58]. Wild-type and control clone replicates were compared against the three null line replicates collectively. Normalised read counts for each sample are in Additional file 9: Table S7. The most differentially expressed genes are consistently different between nulls compared to either wild-type cells or to the clone in which excision was unsuccessful.

Cloning

cDNA was generated from 5 μg of Human XpressRef Universal Total RNA (Qiagen cat # 338112) using Superscript III Reverse Transcriptase (Invitrogen cat # 18080093) following manufacturer's instructions. HMGB1P1, IFITM3P2, AK4P3, and RPL13AP20 were amplified from cDNA with Q5® High-Fidelity DNA Polymerase (New England Biolabs cat # M0492S) using primers that amplify the novel full-length transcript and contain HindIII and NotI restriction sites (Additional file 10: Table S8). PCR amplicons were run on a 1% agarose gel and then purified using the QIAquick Gel Extraction kit (Qiagen cat # 28706). One microgram of DNA was digested with HindIII (New England Biolabs cat # R0104S) and

NotI (New England Biolabs cat # R0189S) for 2 h at 37 °C. Digested PCR products were cleaned up using the QIAquick PCR Purification kit (Qiagen cat # 28104) and then ligated into pcDNA 3.1 3xHA cut with HindIII and NotI using the Quick Ligation kit (New England Biolabs cat # M2200L) following manufacturer's instructions. Plasmid constructs were transformed into One Shot™ TOP10 Chemically Competent *E. coli* (Invitrogen cat # C404003) and plated on agar plates containing ampicillin 100 mg/mL (Sigma-Aldrich cat # A1593). Several colonies were cultured overnight in 5 mL of luria broth and plasmid DNA was extracted using the QIAprep Spin MiniPrep kit (Qiagen cat # 27106). Plasmid DNA was sent for sequencing at the Australian Genome Research Facility (AGRF) to confirm the presence of the full-length pseudogene transcripts. Confirmed plasmids were cultured overnight in 50 mL of luria broth and then DNA extracted using the QIAGEN Plasmid *Plus* Midi kit (Qiagen cat # 12945).

Western blot

HEK293T cells (ATCC® CRL-3216™) were seeded at a density of 5×10^5 cells/well in a 6-well culture plate (Sigma Aldrich, cat # CLS3516). The following day, cells were transfected in a 3:1 ratio of FuGENE® HD Transfection Reagent (Promega cat # E2311) and plasmid DNA (pcDNA3.1-HMGB1P1-3xHA, pcDNA3.1-IFITM3P2-3xHA, pcDNA3.1-AK4P3-3xHA, pcDNA3.1-RPL13AP20-3xHA) in OptiMEM (Thermo Fisher cat # 31985062). Empty pcDNA3.1-3xHA was used as a negative control, whilst pcDNA3.1-3xHA-TurboID [59] served as a positive control. Cells were incubated with the transfection complex in a tissue culture incubator (37 °C, 5% CO_2), and after 24 h protein lysate was extracted using RIPA Lysis and Extraction Buffer (Thermo Fisher cat # 89900) containing cOmplete™ Protease Inhibitor Cocktail (Sigma Aldrich cat # 4693116001). Protein concentration was measured using the Pierce™ BCA Protein Assay Kit on the POLARstar® Omega microplate reader (BMG Labtech). Seven micrograms of protein lysate was diluted with 4x Laemmli Sample Buffer (BioRad cat # 1610747) containing 10% 2-Mercaptoethanol (Sigma Aldrich cat # M6250) and reduced for 5 min at 98 °C. Samples were loaded into a 4–20% Mini-PROTEAN® TGX™ Precast Protein Gel (BioRad cat # 4561094) and run in a Mini-PROTEAN Tetra Vertical Electrophoresis Cell for 35 min at 200 V. Proteins were transferred onto the iBlot™ Transfer Stack (Thermo Fisher cat # IB301001) using the iBlot™ Gel Transfer Device (Thermo Fisher) 7 min program. The membrane was dried overnight and activated in 1xTBS for 5 min and then blocked in Odyssey Blocking Buffer TBS (LI-COR cat # 927-50000) for 1 h at room temp. The membrane was incubated with purified anti-HA.11 epitope tag antibody (BioLegend cat # 901503) and nucleolin (D4C7O) rabbit mAb (Cell Signalling cat # 14574) appropriately diluted in Odyssey® Blocking Buffer (TBS) 0.1% TWEEN® 20 (Sigma Aldrich cat # P1379) overnight at 4 °C. The following day, the membrane was washed four times with 1xTBS 0.1% Tween-20 for 5 min and then incubated with goat-anti mouse IgG IRDye680® (Rockland cat # 610144002) and goat anti-rabbit IgG IRDye800® (Rockland cat # 611132122) in Odyssey Blocking Buffer 0.1% Tween-20 for 1.5 h at room temp. The membrane was washed as previously described and then dried completely before being scanned on Odyssey® CLx Imaging System (LI-COR). Fluorescence was quantified using the Image Studio™ Lite (LI-COR) software.

Subcellular fractionation and quantitative real-time PCR

Three independent wells from a 6-well culture plate (Sigma Aldrich cat # CLS3516) containing 1×10^6 low passage HAP1 cells were lifted using 0.25% Trypsin-EDTA (Gibco cat # 25300096) and washed once with DPBS (Gibco cat # 14190144). Nuclear and cytoplasmic lysates were separated from cells and RNA isolated from both fractions using the PARIS™ Kit (Thermo Fisher cat # AM1921) following manufacturer's instructions. RNA was treated with the TURBO DNA-free™ Kit (Life Technologies cat # AM1907) and concentration was measured on the NanoDrop™ Lite Spectrophotometer (Thermo Fisher Scientific cat # ND-LITE-PR). A qRT-PCR was performed on equal concentrations of DNA-free RNA from both fractions using the *Power* SYBR® Green RNA-to-CT™ 1-Step Kit (Thermo Fisher Scientific cat # 4389986) with primers for MALAT1 (nuclear control), ACTB (cytoplasmic control), and PDCL3P4. Primers for PDCL3P4 were designed to avoid cross-detection of parent gene cDNA. The primers span an exon-exon junction whereby the forward primer is located in a novel upstream exon and the reverse primer in a novel extended sequence of the retroposed portion of PDCL3P4. Samples were run on the ViiA™ 7 Real-Time PCR machine following kit instructions and the ratio of nuclear and cytoplasmic expression was calculated where, nuclear enrichment = $2^{\text{Raw Ct Cytoplasm} - \text{Raw Ct Nucleus}}$ and cytoplasmic enrichment = $2^{\text{Raw Ct Nucleus} - \text{Raw Ct Cytoplasm}}$. Ct values were not normalised to the differential abundance of the MALAT1 and ACTB housekeeping genes between cellular compartments.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02369-0.

Additional file 1. Supplementary figures.

Additional file 2: Table S1. Characterisation of independent sense pseudogene transcripts identified by long-read sequencing.

Additional file 3: Table S2. Characterisation of independent antisense pseudogene transcripts identified by long-read sequencing.

Additional file 4. Uncropped Western blot from Fig. 1g.

Additional file 5: Table S3. Conservation of pseudogene ORFs across primate evolution.

Additional file 6: Table S4. Characterisation of fusion transcripts identified by long-read sequencing.

Additional file 7: Table S5. Tissue-specific expression of pseudogene transcripts in the Illumina Body Map.

Additional file 8: Table S6 Characterisation of independent sense pseudogene transcripts identified in HAP1 cells.

Additional file 9: Table S7. RNA-seq of PDCL3P4 knockout lines.

Additional file 10: Table S8. Oligonucleotides used in this study.

Additional file 11. Review history.

Acknowledgements

The authors would like to thank C. James for technical assistance and acknowledge the Translational Research Institute (TRI) for research space and equipment that enabled this research. We would particularly like to thank the TRI flow cytometry core facility for assistance with this study. The authors thank the University of Queensland Genome Innovation Hub for continuing support.

Review history

The review history is available as Additional file 11.

Peer review information

Barbara Cheifet and Tim Sands were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

G.J.F and S.W.C designed the study. R-.L.T, Y.J, and S.W.C performed the experiments. R.-L.T, T.R.M, A.D.E, and S.W.C performed the analysis. S.W.C and G.J.F funded the study. R-.L.T, A.D.E, G.J.F, and S.W.C wrote the manuscript. All authors read and approved the final manuscript.

Funding

This study was funded by the Australian Department of Health Medical Frontiers Future Fund (MRFF) (MRF1175457 to A.D.E.), the Australian National Health and Medical Research Council (NHMRC) (GNT1173711 to G.J.F. and GNT1161832 to S.W.C.), a CSL Centenary Fellowship to G.J.F., a University of Queensland Early Career Researcher Grant to S.W.C, and by the Mater Foundation (Equity Trustees/AE Hingeley Trust).

Availability of data and materials

All PacBio and RNA-seq data generated by this study is deposited in the Gene Expression Omnibus under accession GSE160383 [20].

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Mater Research Institute-University of Queensland, TRI Building, QLD 4102 Woolloongabba, Australia. [2]Australian Institute for Bioengineering and Nanotechnology, University of Queensland, Brisbane, QLD 4072, Australia. [3]Queensland Brain Institute, University of Queensland, Brisbane, QLD 4072, Australia.

Received: 30 October 2020 Accepted: 28 April 2021
Published online: 10 May 2021

References

1. Vanin EF. Processed pseudogenes: characteristics and evolution. Annu Rev Genet. 1985;19(1):253–72. https://doi.org/10.1146/annurev.ge.19.120185.001345.
2. Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nat Rev Genet. 2019;24:191–201. https://doi.org/10.1038/s41576-019-0196-1.
3. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. Nat Genet. 2000; 24(4):363–7. https://doi.org/10.1038/74184.
4. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, et al. Human L1 retrotransposition: cisPreference versus trans complementation. Mol Cell Biol. 2001;21:1429–39.
5. Kalyana-Sundaram S, Kumar-Sinha C, Shankar S, Robinson DR, Wu Y-M, Cao X, et al. Expressed pseudogenes in the transcriptional landscape of human cancers. Cell. 2012;149(7):1622–34. https://doi.org/10.1016/j.cell.2012.04.041.
6. Guo X, Lin M, Rockowitz S, Lachman HM, Zheng D. Characterization of human pseudogene-derived non-coding RNAs for functional potential. PLoS One. 2014;9:e93972.
7. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. Nat Rev Genet. 2020;21(12):721–36. https://doi.org/10.1038/s41576-020-0251-y.
8. Frith MC, Wilming LG, Forrest A, Kawaji H, Tan SL, Wahlestedt C, et al. Pseudo-messenger RNA: phantoms of the transcriptome. PLoS Genet. 2006;2(4):e23. https://doi.org/10.1371/journal.pgen.0020023.
9. Vinckenbosch N, Dupanloup I, Kaessmann H. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci. 2006;103(9):3220–5. https://doi.org/10.1073/pnas.0511307103.
10. Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. Retrocopy contributions to the evolution of the human genome. BMC Genomics. 2008;9(1):466. https://doi.org/10.1186/1471-2164-9-466.
11. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ, Pandolfi PP. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature. 2010;465(7301):1033–8. https://doi.org/10.1038/nature09144.
12. Korneev SA, Park JH, O'Shea M. Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. J Neurosci. 1999;19(18):7711–20. https://doi.org/10.1523/JNEUROSCI.19-18-07711.1999.
13. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature. 2008;453(7194):539–43. https://doi.org/10.1038/nature06908.
14. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature. 2008;453(7194):534–8. https://doi.org/10.1038/nature06904.
15. Watanabe T, Cheng E-C, Zhong M, Lin H. Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline. Genome Res. 2015;25(3):368–80. https://doi.org/10.1101/gr.180802.114.
16. McCarrey JR, Thomas K. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. Nature. 1987;326(6112):501–5. https://doi.org/10.1038/326501a0.
17. Hayashi H, Arao T, Togashi Y, Kato H, Fujita Y, De Velasco MA, et al. The OCT4 pseudogene POU5F1B is amplified and promotes an aggressive phenotype in gastric cancer. Oncogene. 2015;34(2):199–208. https://doi.org/10.1038/onc.2013.547.
18. Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, et al. Human-specific NOTCH2NL genes expand cortical neurogenesis through Delta/Notch regulation. Cell. 2018;173:1370–84.e16.

19. Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, et al. Human-specific NOTCH2NL genes affect Notch signaling and cortical neurogenesis. Cell. 2018;173:1356–69.e22.
20. Troskie R-L, Jafrani Y, Mercer TR, Ewing AD, Faulkner GJ, Cheetham SW. Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160383.
21. 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(1):751. https://doi.org/10.1186/s12864-020-07123-7.
22. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766–73. https://doi.org/10.1093/nar/gky955.
23. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 2018;28(3):396–411. https://doi.org/10.1101/gr.222976.117.
24. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, et al. A promoter-level mammalian expression atlas. Nature. 2014;507:462–70.
25. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, et al. The GENCODE pseudogene resource. Genome Biol. 2012;13(9):R51. https://doi.org/10.1186/gb-2012-13-9-r51.
26. Zahn-Zabal M, Michel P-A, Gateau A, Nikitin F, Schaeffer M, Audot E, et al. The neXtProt knowledgebase in 2020: data, tools and usability improvements. Nucleic Acids Res. 2020;48(D1):D328–34. https://doi.org/10.1093/nar/gkz995.
27. Xu J, Zhang J. Are human translated pseudogenes functional? Mol Biol Evol. 2016;33(3):755–60. https://doi.org/10.1093/molbev/msv268.
28. Kanber D, Berulava T, Ammerpohl O, Mitter D, Richter J, Siebert R, et al. The human retinoblastoma gene is imprinted. PLoS Genet. 2009;5(12):e1000790. https://doi.org/10.1371/journal.pgen.1000790.
29. Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, et al. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. Nat Struct Mol Biol. 2014;21(2):198–206. https://doi.org/10.1038/nsmb.2764.
30. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011;25(18):1915–27. https://doi.org/10.1101/gad.17446611.
31. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5. https://doi.org/10.1038/nbt.3122.
32. Blomen VA, Májek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, et al. Gene essentiality and synthetic lethality in haploid human cells. Science. 2015;350(6264):1092–6. https://doi.org/10.1126/science.aac7557.
33. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, et al. Identification and characterization of essential genes in the human genome. Science. 2015;350(6264):1096–101. https://doi.org/10.1126/science.aac7041.
34. Srinivasan S, Meyer RD, Lugo R, Rahimi N. Identification of PDCL3 as a novel chaperone protein involved in the generation of functional VEGF receptor 2. J Biol Chem. 2013;288(32):23171–81. https://doi.org/10.1074/jbc.M113.473173.
35. Johnsson P, Ackley A, Vidarsdottir L, Lui W-O, Corcoran M, Grandér D, et al. A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells. Nat Struct Mol Biol. 2013;20(4):440–6. https://doi.org/10.1038/nsmb.2516.
36. Cabili MN, Dunagin MC, McClanahan PD, Biaesch A, Padovan-Merhar O, Regev A, et al. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biol. 2015;16(1):20. https://doi.org/10.1186/s13059-015-0586-4.
37. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A. 2009;106(28):11667–72. https://doi.org/10.1073/pnas.0904715106.
38. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol. 2013;20(3):300–7. https://doi.org/10.1038/nsmb.2480.
39. Brosch M, Saunders GI, Frankish A, Collins MO, Yu L, Wright J, et al. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. Genome Res. 2011;21(5):756–67. https://doi.org/10.1101/gr.114272.110.
40. Gascoigne DK, Cheetham SW, Cattenoz PB, Clark MB, Amaral PP, Taft RJ, et al. Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. Bioinformatics. 2012;28(23):3042–50. https://doi.org/10.1093/bioinformatics/bts582.
41. van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, et al. The Translational Landscape of the Human Heart. Cell. 2019;178:242–60.e29.
42. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife. 2015;4:e08890. https://doi.org/10.7554/eLife.08890.
43. Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011;27(12):1691–2. https://doi.org/10.1093/bioinformatics/btr174.
44. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 2018;34(18):3094–100. https://doi.org/10.1093/bioinformatics/bty191.
45. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033.
46. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. 2019;47(D1):D853–8. https://doi.org/10.1093/nar/gky1095.
47. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One. 2016;11(10):e0163962. https://doi.org/10.1371/journal.pone.0163962.
48. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
49. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic Acids Res. 2019;47(D1):D745–51. https://doi.org/10.1093/nar/gky1113.

50. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9. https://doi.org/10.1093/bioinformatics/btu638.
51. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616.
52. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6. https://doi.org/10.1038/nbt.1754.
53. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64. https://doi.org/10.1101/gr.229202.
54. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005; 6(1):31. https://doi.org/10.1186/1471-2105-6-31.
55. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7(1):539. https://doi.org/10.1038/msb.2011.75.
56. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34(Web Server):W609–12. https://doi.org/10.1093/nar/gkl315.
57. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586–91. https://doi.org/10.1093/molbev/msm088.
58. Powell DR. Degust: interactive RNA-seq analysis [Internet]. Available from: https://doi.org/10.5281/zenodo.3258932.
59. Branon TC, Bosch JA, Sanchez AD, Udeshi ND, Svinkina T, Carr SA, et al. Efficient proximity labeling in living cells and organisms with TurboID. Nat Biotechnol. 2018;36(9):880–7. https://doi.org/10.1038/nbt.4201.
60. Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. Moving beyond P values: data analysis with estimation graphics. Nat Methods. 2019;16(7):565–6. https://doi.org/10.1038/s41592-019-0470-3.

Publisher's Note