Journal of Genetic Engineering and Biotechnology 22 (2024) 100427

Contents lists available at ScienceDirect

Journal of Genetic Engineering and Biotechnology

journal homepage: www.elsevier.com/locate/jgeb

Insights into the human cDNA: A descriptive study using library screening in yeast

Zina Alaswad [a,b], Nayera E. Attallah [a,b,1], Basma Aboalazm [a,b,1], Eman S. Elmeslhy [a,b,1], Asmaa S. Mekawy [a,b], Fatma A. Afify [a,b], Hesham K. Mahrous [a,b], Ashrakat Abdalla [a,b], Mai A. Rahmoon [a,c], Ahmed A. Mohamed [a,b], Ahmed H. Shata [a,b], Rana H. Mansour [a], Fareed Aboul-ela [b,d], Mohamed Elhadidy [a,b], Biola M. Javierre [e], Sherif F. El-Khamisy [f,g], Menattallah Elserafy [a,b,*]

[a] Center for Genomics, Helmy Institute for Medical Sciences, Zewail City of Science and Technology, Giza, Egypt
[b] University of Science and Technology, Zewail City of Science and Technology, Giza, Egypt
[c] Department of Pharmaceutical Biology, Faculty of Pharmacy and Biotechnology, German University in Cairo, Egypt
[d] Center for X-Ray Determination of the Structure of Matter, Zewail City of Science and Technology, Giza, Egypt
[e] Josep Carreras Leukaemia Research Institute, Badalona, Barcelona, Spain
[f] The Healthy Lifespan Institute and Institute of Neuroscience, School of Bioscience, University of Sheffield, South Yorkshire, UK
[g] The Institute of Cancer Therapeutics, University of Bradford, West Yorkshire, UK

ARTICLE INFO

Keywords:
cDNA library screens
Yeast two-hybrid
Saccharomyces cerevisiae
Genetic screens

ABSTRACT

The utilization of human cDNA libraries in yeast genetic screens is an approach that has been used to identify novel gene functions and/or genetic and physical interaction partners through forward genetics using yeast two-hybrid (Y2H) and classical cDNA library screens. Here, we summarize several challenges that have been observed during the implementation of human cDNA library screens in *Saccharomyces cerevisiae* (budding yeast). Upon the utilization of DNA repair deficient-yeast strains to identify novel genes that rescue the toxic effect of DNA-damage inducing drugs, we have observed a wide range of transcripts that could rescue the strains. However, after several rounds of screening, most of these hits turned out to be false positives, most likely due to spontaneous mutations in the yeast strains that arise as a rescue mechanism due to exposure to toxic DNA damage inducing-drugs.

The observed transcripts included mitochondrial hits, non-coding RNAs, truncated cDNAs, and transcription products that resulted from the internal priming of genomic regions. We have also noticed that most cDNA transcripts are not fused with the GAL4 activation domain (GAL4AD), rendering them unsuitable for Y2H screening. Consequently, we utilized Sanger sequencing to screen 282 transcripts obtained from either four different yeast screens or through direct fishing from a human kidney cDNA library. The aim was to gain insights into the different transcription products and to highlight the challenges of cDNA screening approaches in the presence of a significant number of undesired transcription products. In summary, this study describes the challenges encountering human cDNA library screening in yeast as a valuable technique that led to the identification of important molecular mechanisms. The results open research venues to further optimize the process and increase its efficiency.

Abbreviations: Y2H, Yeast two-hybrid; GAL4AD, GAL4 activation domain; RT, Reverse transcriptase; GAL4 DNA-BD, GAL4 DNA-Binding Domain; HR, Homologous recombination; TDP1, Tyrosyl phosphodiesterase 1; SSBR, Single-strand break repair; ATM, Ataxia-Telangiectasia Mutated; TDP2, Tyrosyl-DNA phosphodiesterase 2; ncRNAs, Non-coding RNAs; CDS, Coding sequences; PASs, Polyadenylation signals; SC, media, Synthetic complete; MMS, Methyl Methansulfonate; HU, Hydroxyurea; CPT, Camptothecin; ORF, Open reading frame; 5′UTR, 5′ untranslated region; NLS, Nuclear localization signal; ESTs, Expressed sequence tags; eccDNA, Extrachromosomal circular DNA; LNA, Locked nucleic acid; PDD, Probe-directed degradation; NSR, Not-so-random.

* Corresponding author.
E-mail address: melserafy@zewailcity.edu.eg (M. Elserafy).
[1] Equal contribution

https://doi.org/10.1016/j.jgeb.2024.100427
Received 27 May 2024; Received in revised form 29 September 2024; Accepted 30 September 2024

1. Introduction

Yeast stands out as a prominent example of eukaryotic cells, characterized by its fully sequenced genome, remarkable genetic adaptability, rapid replication rate, and cost-effectiveness[1]. Despite being a simple, single-celled organism, yeast plays a crucial role in unravelling the intricate mechanisms of various molecular pathways[1]. Consequently, the invaluable knowledge gained from yeast research serves as a valuable compass for molecular biologists seeking to navigate analogous pathways in the significantly more complex and sophisticated mammalian cell systems.[1,2] Moreover, yeast studies help in deciphering the genetic and cellular disruptions underlying a wide range of diseases. Thus, this has paved the way for tailoring personalized medicine, improving diagnostics, and advancing therapies.[1] Human cDNA library screening in yeast has been utilized by various labs to gain insights into gene functions and unravel molecular pathways or mechanisms [3]. The approaches encompass classical human cDNA library screens and yeast two-hybrid screens (Fig. 1). The cDNA library is constructed through the utilization of the universal poly-A sequence present at the 3′ end of mRNA. An oligo (dT) primer anneals to poly-A mRNA, then the reverse transcriptase (RT) enzyme extends the annealed primer along the mRNA template to synthesize the cDNA.[4] Afterwards, the cDNA is cloned into an appropriate vector, which is ultimately transferred into host bacteria to create the cDNA library[5] (Fig. 1A).

Yeast two-hybrid (Y2H) assay is a powerful method to determine protein–protein interactions (PPIs).[6–8] In this assay, prey and bait, are fused to the activation domain and the DNA binding domain of Gal4, respectively[9]. The reporter transcription is activated upon physical interaction of the two proteins of interest, which are either fused to GAL4 Activation Domain (GAL4 AD) or GAL4 DNA-Binding Domain (GAL4 DNA-BD)[10]. Through screening approaches, multiple preys can be screened against one bait to determine protein–protein interactions[11]

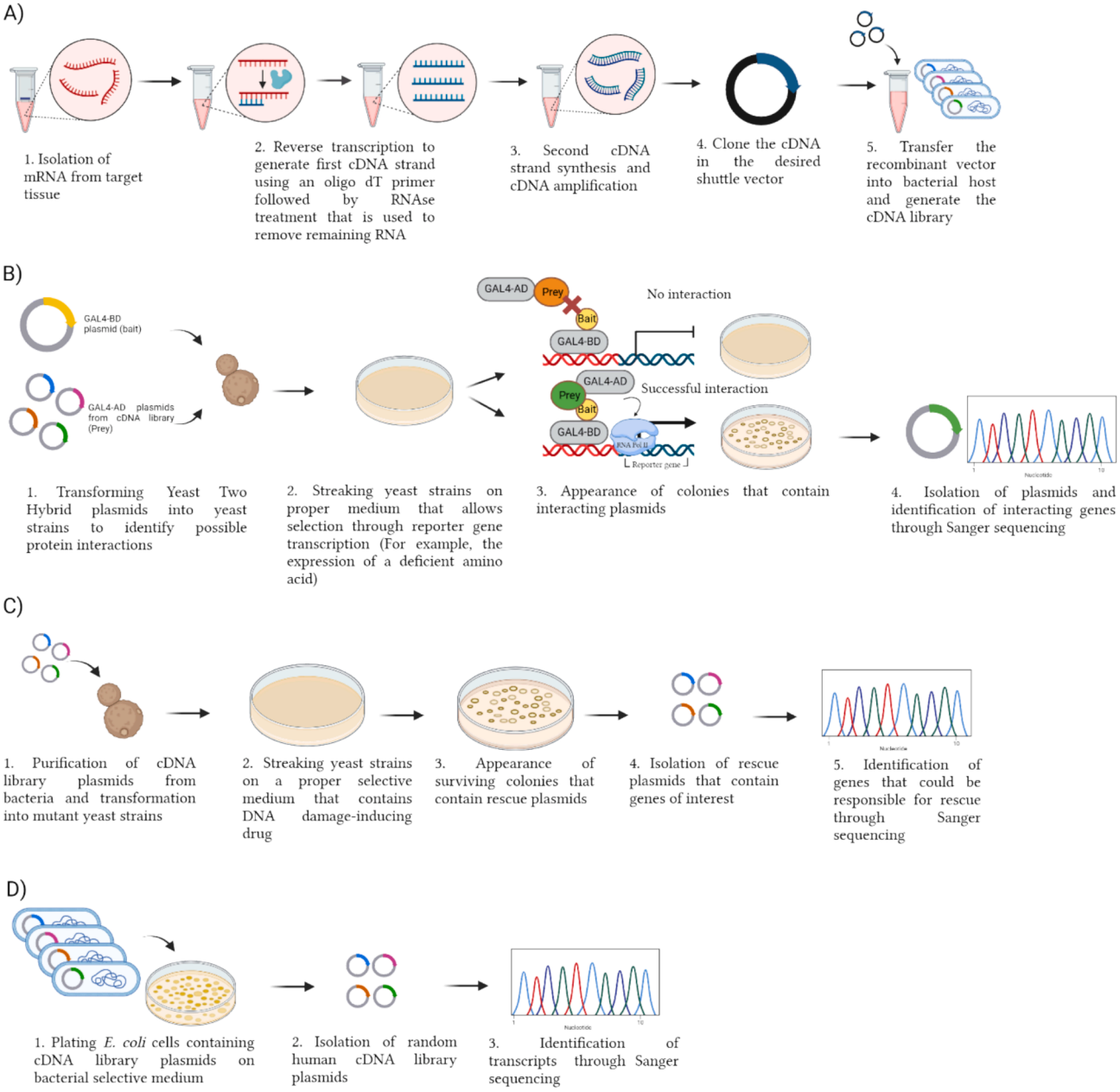

Fig. 1. Y2H cDNA library construction and screening procedure. A) The conventional procedure for cDNA library generation. B) Using the cDNA library in Y2H screens to identify protein–protein interactions. C) Using the cDNA library to screen for rescue genes involved in DNA repair. D) Screening randomly selected plasmids from the cDNA library to screen random samples from the library "Created with Biorender.com.".

(Fig. 1B). Remarkably, this robust system has facilitated the development of maps illustrating binary protein interactomes in both humans[12] and yeast[13,14]. However, previous studies have highlighted the shortcomings of Y2H screening, which can yield both false negative and false positive rescuers[15]. Important discoveries related to DNA repair mechanisms were made using this approach, including the interactions between Rad55 and Rad57, Mre11 and Rad50, as well as Rad51 and both Rad55 and Rad54, occurring in both mitotic and meiotic homologous recombination (HR).[16–18] Additionally, thanks to this approach, the functional interaction between tyrosyl phosphodiesterase 1 (TDP1) and DNA ligase IIIa in single-strand break repair (SSBR) was discovered[19]. It is worth noting that the human Mre11 was also identified using this screen.[20].

Screening of human cDNA libraries in yeast was also used to identify genetic rescue by human genes for mutant yeast strains through functional complementation[21]. The cDNA library vectors are transformed into the desired mutant yeast strain and positive hits capable of rescuing the mutant strains in the presence of stressors are identified through Sanger sequencing[22] (Fig. 1C). This approach led to discoveries in different molecular pathways including steroid receptor signaling, apoptosis, the identification of Ataxia-Telangiectasia Mutated (ATM),[23] and Tyrosyl-DNA phosphodiesterase 2 (TDP2).[22,24] .

In our lab, we have conducted several human cDNA library screens in yeast to identify novel genes that rescue yeast mutant strains deficient in DNA damage tolerance pathways through exploiting their sensitivity to cytotoxic drugs including Methyl Methansulfonate (MMS), Hydroxyurea (HU), and Camptothecin (CPT) (Fig. 1C). We have noticed a significant challenge that hindered the fishing of the correct rescuer. In addition most transformants exhibited resistance to the applied cytotoxic stressors during the screening, but the hits were unable to rescue in subsequent rounds of confirmation. A significant number of the transcripts identified were non-coding, represented by non-coding RNAs (ncRNAs) and Sanger sequencing reads that map to genomic loci. We have also observed a large number of mitochondrial coding and non-coding transcripts. Furthermore, among the coding sequences (CDS) pool, most of the transcripts were truncated. Moreover, most CDS were not fused to GAL4 AD, rendering these genes unsuitable for Y2H screens (Fig. 1B), despite their suitability for genetic screens (Fig. 1C).

Therefore, we decided to further investigate the cDNA library to report the different pools of transcripts observed and highlight the issue of classical fishing of mRNA using oligo (dT) priming. As the diversity in transcript types observed not only impacts cDNA library screening but also in all techniques relying on poly-A priming such as qRT-PCR and Poly-A RNA-Seq.[25–28] For this purpose, we combined a set of approximately 133 transcripts (positives (full length and truncated sequences), false positives or undesired non-coding and mitochondrial sequences from four different screens) (Fig. 1C) with 149 transcripts randomly selected from the library without going through the screening process (Fig. 1D), to provide an overview of different transcript types in the library.

2. Materials and methods

2.1. The cDNA library

The Human Kidney Matchmaker™ cDNA Library was purchased from Clontech (Cat. No 638816). The mRNA was extracted from the kidneys of eight normal Caucasians male or female who were 24–55 years old and died with trauma. The purchased library comprising cDNA library plasmids in one ml *E. coli* was amplified and the plasmids were isolated from *E. coli* using QIAGEN® Plasmid Plus Giga Kits (Cat. No. / ID: 12191) according to the manufacturer's instructions.

2.2. Yeast screening

The following yeast strains were used in four different screens: W303 *rad5Δ::TRP1* was used in two screens, one against HU and one against MMS, W303 *RAD5+ mms2Δ::KANMX4* was used in one screen against MMS and the W303 *rad5Δ::TRP1 tdp1Δ::hphNT1 rad1Δ::HIS3MX6* was used in one screen against CPT.

Yeast competent cells were transformed with 1 μg of cDNA, vigorously vortexed, and incubated at room temperature for 15 min. Afterwards, a PEG/LiAc was added, mixed thoroughly by gentle vortexing, and incubated on a thermoshaker at 30 °C, 700 rpm for 30 min. The mixture was gently vortexed every 10 min. Then, DMSO was added, mixed, and the tubes were placed in a 42 °C thermoblock for 15 minThe cells were subsequently centrifuged at 3,000 rpm for 3 min, and the supernatant was discarded. The pellet was re-suspended in one ml of autoclaved water by gentle pipetting and then spread onto synthetic complete (SC) media lacking Leucine (LEU) plates containing lethal doses of different drugs; CPT, HU, and MMS utilized according to the purpose of the screen. The plates were then incubated at 30 °C for 2–5 days. Surviving colonies were isolated from the plates and the plasmids were extracted. For plasmid extraction, the surviving yeast colonies were inoculated in SC-LEU overnight cultures. On the following day, the cell pellets were collected and re-suspended in lysis buffer (2.5 M NaCl, 0.1 M EDTA, 10 mM Tris base (pH 10), NaOH (0.8 g), 0.5 % Triton X-100, 10 % DMSO). This step was optimized according to the size of the pellet. Afterwards, 30 μl lyticase and 1.6 μl beta-mercaptoethanol were added and the suspension was incubated overnight at 37 °C. On the third day, the suspension was centrifuged, and plasmids were isolated using a mini-prep method. The extracted plasmids were then transformed into competent *E. coli* DH5α cells for amplification. The plasmids were further purified and retransformed into the mutant strains to test their rescue capability. During this validation step, false-positive plasmids that did not result in rescue for the mutant same strains were excluded as the rescue mostly arise from yeast spontaneous mutations. The promising plasmids that resulted in rescue were sent for Sanger sequencing using pACT2-F and pACT2-R. To assess the transformation efficiency, 100 μl of 1:10 and 1:100 dilutions of the mutant strains transformed with one μg of the library were spread onto control SC–LEU plates. The transformation efficiency for all screens were in the normal range.

2.3. Random selection of cDNA library plasmids

E. coli DH5α cells were transformed with 1 μl of the cDNA library, then incubated on ice for 30 min, followed by 90 sec heat shock at 42 °C. After 15 min of incubation on ice, one ml LB medium was added and cells were incubated for 1 h at 37 °C. The cells were spun down and 200 μl LB medium were kept to re-suspend the cells. Finally, the cells were plated out on LB + ampicillin selective plates and incubated overnight at 37 °C. The surviving colonies were isolated and the plasmids were extracted. The plasmids were screened via PCR using pACT2-F and pACT2-R to select plasmids of different sizes for further screening, reducing the possibility of sequencing the same gene. Finally, to identify the gene sequences of the randomly isolated plasmids, the plasmids were sent to Eurofins Genomics for Sanger sequencing.

2.4. Data analysis

pACT2-F (GATGATGAAGATACCCCAC) and pACT2-R (CAGTTGAAGTGAACTTGC) were used to sequence the cDNA on the pACT2 library vectors. The quality of the reads for the forward and reverse primers was assessed using A plasmid Editor (ApE) software (https://jorgensen.biology.utah.edu/wayned/ape/). The sequences were subsequently compared to a reference human genome through the NCBI Blastn alignment software and GenBank. To search for polyadenylation signals (PASs) of genomic fragments amplified via internal priming, Human blast search of Santa Cruz Genomics Institute (https://genome.ucsc.edu/cgi-bin/hgBlat) was used to identify the chromosomal locus needed for search on the PolyASite database (https://polyasite.unibas.ch/). Furthermore, to explore GAL4 fusion status, the expasy translate

tool was utilized (https://web.expasy.org/translate/), and the pie chart was designed using Graph Pad Prism 7.

3. Results

Our lab has employed the human cDNA library screening approach to discover new functions for coding genes via functional complementation in yeast. However, the outcome of the screens revealed the presence of a significant number of undesired transcripts and false positives. Therefore, we investigated further analyze the nature of the transcripts within the cDNA library. We analyzed the sequencees of 282 transcripts obtained either from the output of four different yeast screens or through random selection from the cDNA library to provide insights that could be useful for optimizing the library and other techniques that depend on poly-A priming for selecting mRNA transcripts.[28,29] .

The human cDNA kidney library utilized was constructed using the pACT2 plasmid (Fig. 2). To determine the sequence of the DNA fragments carried by the pACT2 vectors, Sanger sequencing was performed using the universal pACT2-F and pACT2-R primers (Fig. 2). Here we describe the different sequences observed from our analysis.

3.1. 5′ end truncated cDNAs are enriched in the library

We have observed that 18.15 % of the analyzed hits comprised 5′end truncated cDNAs as presented in Table 1 and Table S1. To reach such a conclusion, we made sure that the quality of Sanger sequencing covering the adaptor sequence and the sequence of the transcript after the truncation were of very high quality. Interestingly, some Sanger reads mapped only to the 3′ UTRs of the mRNA (Table 2 and Table S2), indicating that the reverse transcriptase was not able to reach the open reading frame (ORF) of the genes (Fig. 3).

3.2. Internal priming for mRNA sequences generates truncated cDNAs among the library transcripts

Since conversion of the mRNA into cDNA is achieved through oligo (dT) priming[30,31] (Fig. 4A), the binding of an oligo (dT) primer to an adenosine-rich (A-rich) region can simultaneously prevent the reverse transcriptase from extending the transcripts if it was previously primed to the downstream poly-A tail[4] (Fig. 4B). Using the PolyASite database, we searched for (PASs) in different cDNA sequences suspected to be amplified through internal priming. For all reads, high-quality sequences were analyzed. For SF3B1, no real poly A-sequence was recognized and the presence of an A-rich sequence was detected in Exon 11, confirming its amplification from an A-rich internal region in the transcript (Table 3 and Table S3).

3.3. Several identified CDS transcripts were not fused to GAL4 AD

The cDNA should be fused with GAL4 AD to function in the context of the Y2H system.[32] However, the vast majority of our analyzed CDS

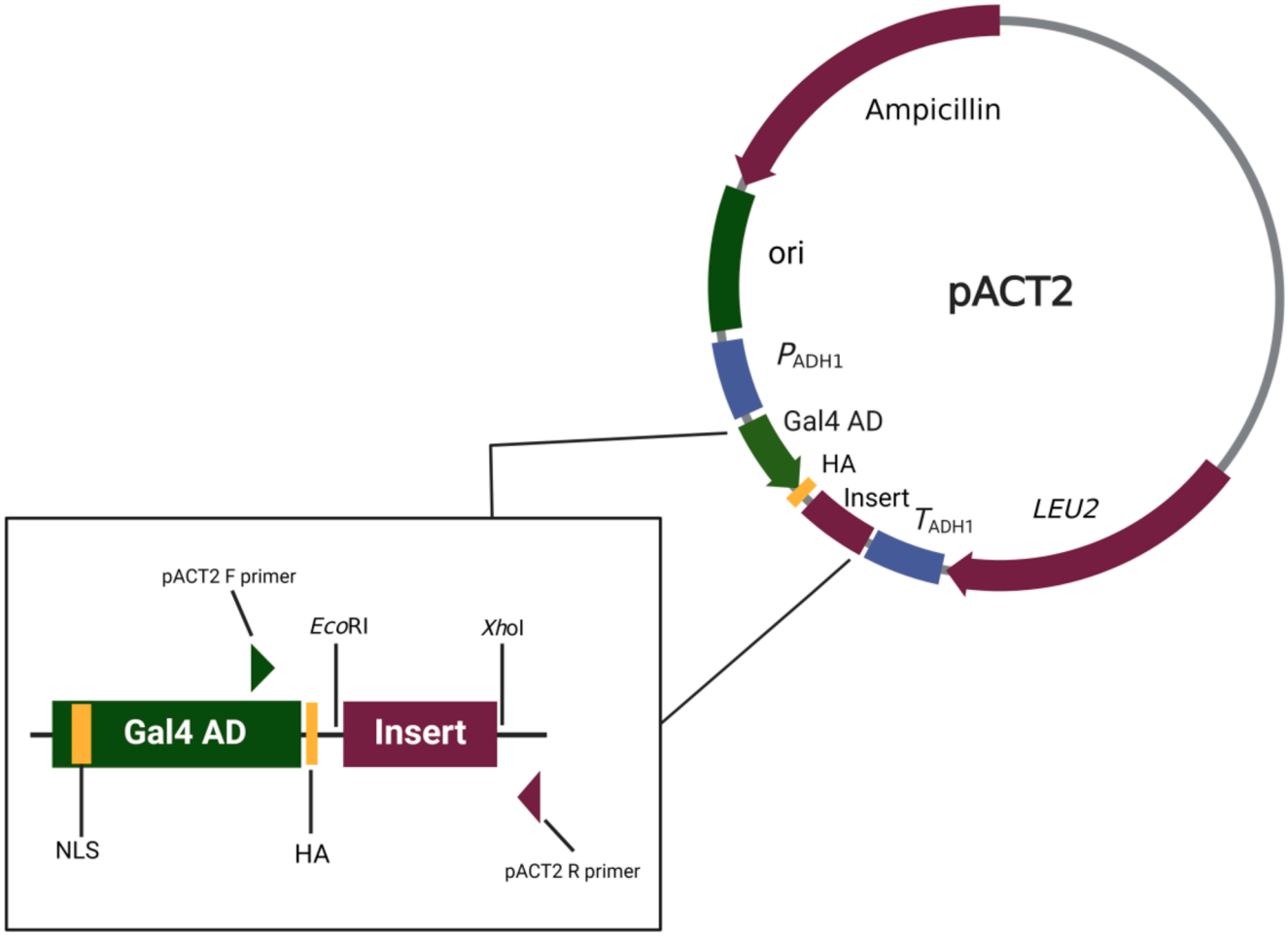

Fig. 2. pACT2 sequence. The pACT2 plasmid map shows the transcript cloned between the two restriction sites (EcoR1 and Xho1). $P_{ADH1;}$ ADH1 promoter, $T_{ADH1;}$ ADH1 transcription termination signal, NLS; Nuclear localization sequence, HA; HA epitope tag. pACT2-F and pACT2-R primers are used for sequencing purposes. "Created with Biorender.com.".

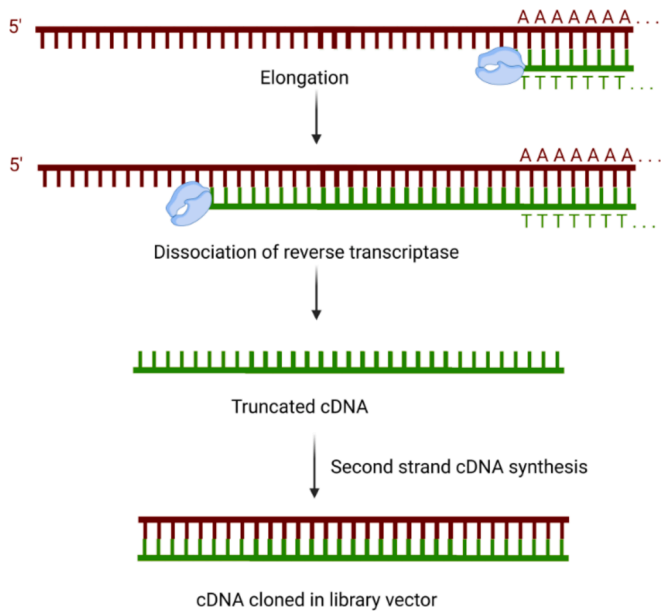

Fig. 3. The inefficiency of the reverse transcriptase on large transcripts. "Created with Biorender.com.".

transcripts were not in frame with GAL4 AD (Tables 4, 5, Table S4, and Table S5). The main reason was that the coding sequence (CDS) was preceded by a 5′ untranslated region (5′UTR), which was present in between the GAL4 AD sequence and the Open Reading Frame (ORF) of the gene. To conclude whether the sequence was in frame or not with the GAL4 AD, we used the HA tag sequence that is cloned in frame with the GAL4 AD sequence as a reference. The sequencing results of the 5′UTR had to be accurate with no noise present in the chromatogram to make a conclusion (Tables 4, 5, Table S4, and Table S5). It is worth mentioning that the only sequence which shows noise in one nucleotide was excluded e.g. Thy-1 cell surface antigen (THY1), transcript variant 1.

As shown in (Table 4) and (Table 5), the vast majority of CDS were not in frame with the GAL4 AD and the HA sequence because the 5′ UTRs either contained stop codons before the start codon of the CDS, or the 5′ UTR sequence was not present in triplets, resulting in a frameshift. Consequently, the unfused hits are not suitable for Y2H experiments.

3.4. Multiple Sanger sequenced reads aligned to non-coding regions

Multiple reported ncRNAs were identified in our analysis. Many studies have indicated that the majority of ncRNAs are polyadenylated.[33,34] As a result, they are picked up by the conventional poly (dT) priming method. Table 6 and Table S6 show the number of non-coding RNAs analyzed.

In addition to the identified non-coding RNAs, we have observed that several sequenced fragments mapped to different genomic/chromosomal regions which were not previously reported to produce transcripts (Table 7). The A-rich region that resulted in fishing these transcripts could not be identified as a poly-A site by the PolyASite database. The 3′ ends of the reads showed no indication of a poly-A cleavage site or upstream poly-A signal (Table S7), suggesting that these genomic sequences were likely picked up from genomic DNA, not from transcription products, through internal priming with poly-A rich regions (Fig. 5). For the analysis, we used sequences that contained both forward and reverse sequencing reads and excluded reads that sequenced with one primer only. We also identified sequences that aligned to clone sequences from different sequencing libraries (Table S8). Moreover, the oligo (dT) can bind to poly-A tracts in genomic DNA, which is expected to be contaminants due to incomplete removal, leading to nonspecific genomic amplification[35] (Fig. 5).

3.5. Mitochondrial transcripts are highly enriched in the library

We observed that the library was enriched with reads that mapped to human mitochondrial sequences; both coding and non-coding (Table 8, Table S9, and Table S10). We summarized, in Table 8, the identified hits and classified them into full- length, 5′ or 3′ truncated. We also checked whether the poly-A site exists in the correct location. Moreover, we analyzed whether they were present as single transcripts or fused with other transcripts. We found that many mitochondrial sequences were

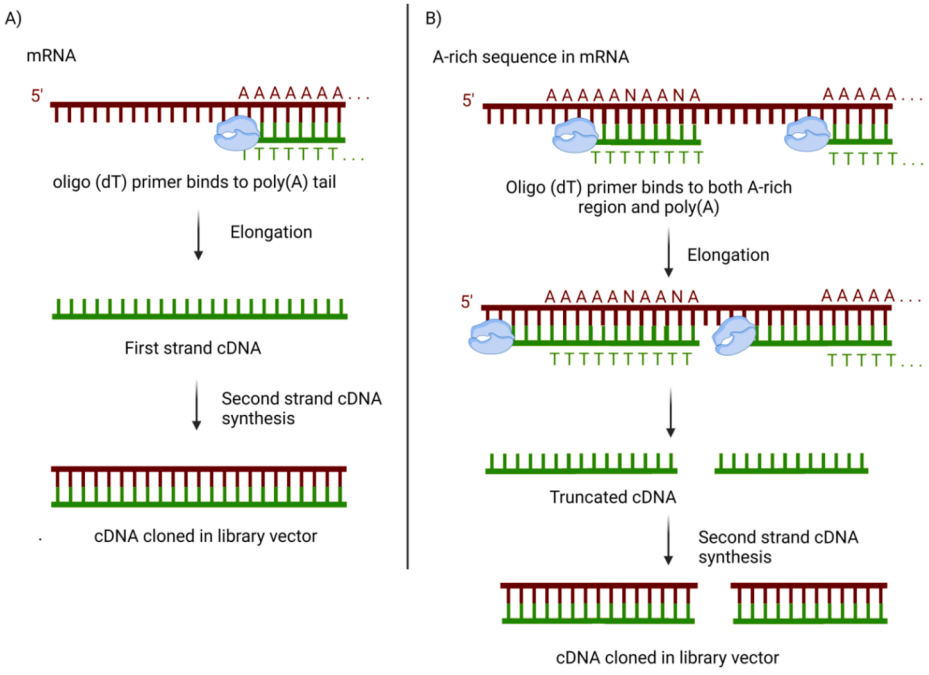

Fig. 4. mRNA priming in cDNA libraries. A) Normal priming through reverse transcriptase elongation of cDNA after oligo (dT) binding to true poly-A tails. **B)** Internal priming that leads to the generation of truncated transcripts due to internal A-rich regions and reverse transcriptase stalling. "Created with Biorender.com.".

Table 1
The 5′ end truncated cDNAs identified, bp; base pairs.

cDNA Name	Number of bp missing from 5′ end	Status of the 3′
Cathepsin D (CTSD)	731	Complete
Ferri [ME2] tin heavy chain 1 (FTH1)	21	Complete
Uromodulin (UMOD), transcript variant 5	1122	Complete
Trafficking from ER to Golgi regulator (TFG), transcript variant 2	924	Complete
Protein phosphatase 1 catalytic subunit alpha (PPP1CA), transcript variant 2	74	Complete
Dipeptidyl peptidase 9 (DPP9), transcript variant 14	2146	Complete
Aldolase, fructose-bisphosphate B (ALDOB)	94	Complete
Glutathione peroxidase 3 (GPX3), transcript variant 1	75	Complete
Actinin alpha 2 (ACTN2), transcript variant 1	1833	Complete
Apolipoprotein E (APOE), transcript variant 5	668	Complete
Cervical cancer oncogene 3 mRNA	346	Complete
COPI coat complex subunit alpha (COPA), transcript variant1	3358	Complete
Protein kinase AMP-activated non-catalytic subunit beta 1(PRKAB1)	367	Complete
ADP ribosylation factor like GTPase 6 interacting protein 1 (ARL6IP1), transcript variant 1	368	Complete
Insulin like growth factor binding protein 7 (IGFBP7), transcript variant 1	715	Complete
Transmembrane BAX inhibitor motif containing 6 (TMBIM6), transcript variant 1	113	Complete
Transmembrane BAX inhibitor motif containing 6 (TMBIM6), transcript variant 1	113	Complete
Acyl-CoA dehydrogenase short chain (ACADS), transcript variant 2	667	Complete
Plexin D1 (PLXND1)	5290	Complete
Actin gamma 1 (ACTG1), transcript variant 2	695	Complete
Palladin cytoskeletal associated protein (PALLD), transcript variant 4	1171	Complete
Glutathione peroxidase 3 (GPX3), transcript variant 2	536	Complete
Glycine amidinotransferase (GATM), transcript variant 2	493	Complete
RNA binding motif protein 25 (RBM25)	1846	Complete
Vascular cell adhesion molecule 1 (VCAM1)	1423	Complete
Albumin (ALB)	842	Complete
Exocyst complex component 3 (EXOC3)	1651	Complete
Betaine–homocysteine S-methyltransferase (BHMT)	1142	Complete
UDP-glucose pyrophosphorylase 2 (UGP2), transcript variant 1	1370	Complete
RAP interaction protein gamma 2	52	Complete
Ribosomal protein L3 (RPL3), transcript variant 1	225	Complete
Fibulin-5	212	Complete
General transcription factor IIB	128	Complete
Threonyl-tRNA synthetase 3 (TARS3)	1479	Complete
Methylenetetrahydrofolate dehydrogenase (NADP + dependent) 2, methenyltetrahydrofolate cyclohydrolase (MTHFD2), transcript variant 1	723	Complete
Lysine acetyltransferase 2A (KAT2A), transcript variant 1	1770	Complete
Profilin 2 (PFN2), transcript variant 2	174	Complete
Uromodulin (UMOD), transcript variant 5	1120	Complete
mRNA for KIAA0264 gene	8	Undetectable
Fibulin-5	212	Undetectable
Ornithine decarboxylase 1 (ODC1), transcript variant 3	213	Undetectable
ATP synthase F1 subunit alpha (ATP5F1A), transcript variant 4	640	Undetectable
Proteasome 26S subunit, ATPase 5 (PSMC5), transcript variant 1	19	Undetectable
Heat shock protein 90 alpha family class B member 1 (HSP90AB1), transcript variant 6	914	Undetectable
YY1 transcription factor (YY1)	113	Undetectable
General transcription factor IIB	128	Undetectable
Complement C1s (C1S), transcript variant 3	109	Undetectable
Proteasome (prosome, macropain) 26S subunit, ATPase, 5 (PSMC5)	19	Undetectable

Table 2
Sequences identified as 3′UTRs fragments.

Name of the transcript	Sequence detected
Thiol methyltransferase 1A (TMT1A)	3′UTR only
actin beta (ACTB)	3′UTR only
Adhesion G protein-coupled receptor G1 (ADGRG1), transcript variant 13	3′UTR only
Sarcoglycan beta (SGCB)	3′UTR only
Ras homolog family member A (RHOA), transcript variant 2	3′UTR only
Interactor of little elongation complex ELL subunit 2 (ICE2), transcript variant 3	3′UTR only
KIAA1018 mRNA for KIAA1018 protein	3′UTR only
Zinc finger protein 395 (ZNF395)	3′UTR only
mRNA for BDG-29	3′UTR only
Calnexin (CANX), transcript variant 2	3′UTR only
OTU deubiquitinase 4 (OTUD4), transcript variant 4	3′UTR only
Cingulin like 1 (CGNL1), transcript variant 1	3′UTR only
Tumor protein, translationally-controlled 1 (TPT1), transcript variant 3	3′UTR only
Ubiquilin 4 (UBQLN4), transcript variant 2	3′UTR only
SEC14 and spectrin domain containing 1 (SESTD1)	3′UTR only
WASP like actin nucleation promoting factor (WASL)	3′UTR only

Table 3
Genes confirmed to be fished through internal priming.

Gene name	Internal priming site	A-rich sequence localization	Sequence identified
Splicing Factor 3b Subunit 1 (*SF3B1*)	@197419221	Exon 11	Chr2:197405136–197419220

Table 4
GAL4 AD fusion status in complete CDS.

Gene	GAL4 AD fusion	3′ end status
TMEM9 domain family member B (*TMEM9B*), transcript variant 1	Fused	Complete
Ribosomal protein S4 X-linked (*RPS4X*)		
Beta-2-microglobulin (*B2M*)		
OK/SW-CL.16		
Glutathione peroxidase 3 (*GPX3*), transcript variant 1	Not Fused	Complete
Cytochrome *c* oxidase subunit IV isoform 1		
MAGE family member H1 (*MAGEH1*)		
Glycine cleavage system protein H(GCSH)		
Nascent polypeptide associated complex subunit alpha (NACA), transcript variant 4		
Oxidative stress responsive serine rich 1 (*OSER1*)		
Humanin (*HN1*)		
v-Jun sarcoma virus 17 oncogene homolog (avian) (*JUN*)		
Ribosomal protein *L*10 (*RPL10*), transcript variant 5		
Aspartate dehydrogenase domain containing, transcript variant 1 (ASPDH)		
Signal recognition particle 14, transcript variant 1, (SRP14)		
Phosphomannomutase 1 (PMM1)		

Table 5
Gal4 AD fusion status in CDS with 'complete 5′ end' and 'truncated or undetectable 3′ end'.

Gene	GAL4 AD fusion	3′ end status
Eukaryotic translation elongation factor 1 alpha 1 (*EEF1A1*)	Fused	Undetectable
Splicing factor 3b subunit 1 (*SF3B1*), transcript variant X1	Not Fused	Truncated
Thy-1 cell surface antigen (THY1), transcript variant 1	NA	Undetectable
ER lipid raft associated 1 (*ERLIN1*), transcript variant 2	Not Fused	
Aldolase, fructose-bisphosphate B (*ALDOB*)		
ATP synthase F1 subunit alpha (*ATP5F1A*), transcript variant 4		
Growth arrest specific 6 (*GAS6*)		
Ribosomal protein L4 (*RPL4*)		
Phosphomannomutase 1 (*PMM1*)		
WD repeat containing antisense to TP53 (WRAP53), transcript variant 4		
DAB adaptor protein 2 (DAB2), transcript variant 1		

Table 6
ncRNAs identified in our analysis.

NcRNA name	Sequence identified
Cold inducible RNA binding protein (*CIRBP*), transcript variant 5, non-coding RNA	Partial
Long intergenic non-protein coding RNA 622 (*LINC00622*), long non-coding RNA	Partial
LncNR4A1-AS lncRNA	Partial
Aftiphilin (AFTPH), transcript variant 10, non-coding RNA	Partial
Mitochondrial ribosomal protein L42 (MRPL42), transcript variant 5, non-coding RNA	Partial
Transforming acidic coiled-coil containing protein 1 (TACC1), transcript variant 38, non-coding RNA	Partial
Carbonic anhydrase 12 (CA12), transcript variant 4, non-coding RNA	Partial
Fibronectin type III domain containing 3A (FNDC3A), transcript variant 4, non-coding RNA	Partial

Table 7
Sanger reads mapped to genomic locus. Analysis performed using Human BLAT Search.

Sanger reads mapped to chromosomal loci (Last bp indicated is the one before the A-rich region)	Correct Poly A site predicted at the 3′end
chr1:244,844,889–244,844,889	No
chr13:98,207,932–98,208,331	No
chr2:39,465,672–39,467,160	No
chr5:27,131,467–27,132,685	No
chr14:49,490,288–49,491,571	No
chr15:66,499,315–66,504,831	No
chr3:3,325,013–3,325,596	No
chr3:122,224,209–122,225,393	No
chr3:122,224,209–122,225,564	No
chr9:627,969–629,094	No
chrX:9,854,040–9,855,321	No
chr12:74,374,293–74,376,111	No
chr14:75,792,410–75,793,388	No
chr14:75,792,410–75,793,391	No
chr14:49490302–49491571	No
chr8:144,031,721–144,032,233	No
chr4:86,517,270–86,518,620	No
chr15:57,209,896–57,211,427	No
chr22:45,667,231–45,668,771	No
chr22:36,308,432–36,309,919	No

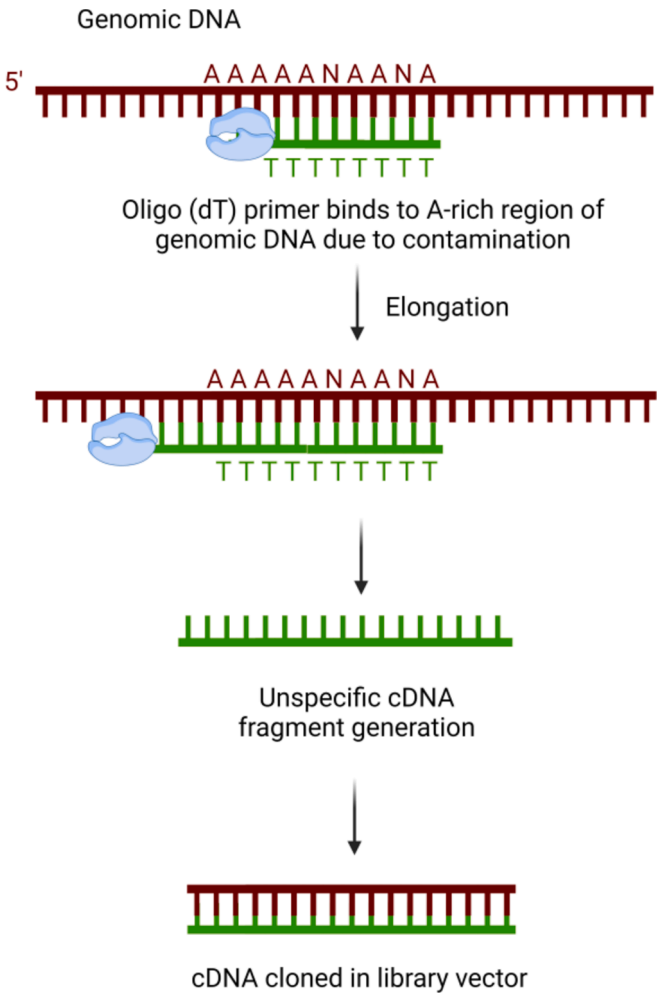

Fig. 5. Oligo (dT) primers bind to A-rich regions in genomic DNA "Created with Biorender.com.".

fused to tRNA. Mitochondrial transcripts are often transcribed as polycistronic primary RNA punctuated by tRNA, then RNases P and Z excise the flanking tRNA and liberate the pre-mRNA.[36,37] However, the mature mt-mRNAs, except for ND6, contain a stretch of poly-A at the 3′ end.[38] This indicates that the oligo (dT) can bind to either unprocessed mt-mRNA or mis-prime to mitochondrial genomic regions.

The ribosomal *16S rRNA* was highly represented among the hits we analyzed (at least 29 transcripts were detected representing 9.96 % of total transcripts). The mitochondrial *16S rRNA* is polyadenylated. In addition, it contains a consecutive A-rich sequence in the middle,[39–41] which explains the reasons behind its frequent identification in the human cDNA library, whether it was amplified via normal or internal priming (Table S10).

A schematic representation for the mitochondrial transcripts we obtained throughout the analysis is represented in Fig. 6.

Our data overall showed that out of the 281 transcripts only 29.55 % of the cDNAs analyzed comprised the desired coding sequences (Fig. 7). However, as indicated in Fig. 7, only 6.41 % were full-length CDS and a smaller percentage was fused with GAL4 AD. Moreover, mitochondrial hits accounted for 27.04 % including transcripts and the 16S rRNA. 6.41 % represented 3′UTR sequences that did not include ORFs. We have also seen other genomic loci-mapped hits, sequences derived from non-coding RNAs and clone sequences mapped-hits (Fig. 7). This, in turn, challenges the identification of positive hits and hinders the process of distinguishing them from false positives and irrelevant transcripts. Thus, we refer to the screening process as 'fishing for a needle in a haystack'.

Table 8
Mitochondrial hits status and their quantification.

Gene Name	Total number of transcripts detected	CDS fused to tRNA	5′ truncated cDNA	3′ truncated cDNA	Correct poly-A site
MT-CYTB	13	0	11	3	6
MT-COX1	4	4	4	0	0
MT-COX2	4	3	3	1	0
MT-COX3	9	0	8	0	7
MT-ND1	6	0	5	0	6
MT-ND2	5	1	3	1	3
MT-ND4	4	0	3	0	4
MT-ATP6	1	0	1	0	0

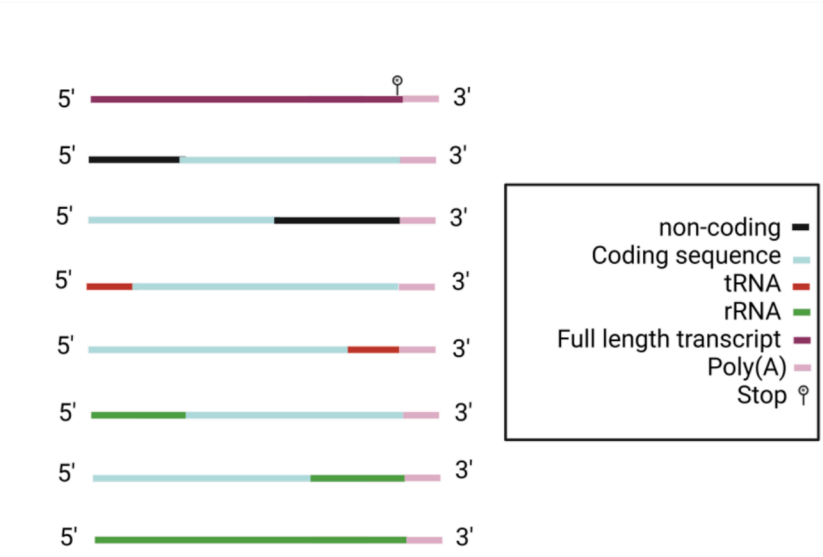

Fig. 6. Schematic view of mitochondrial transcripts observed in our screens "Created with Biorender.com".

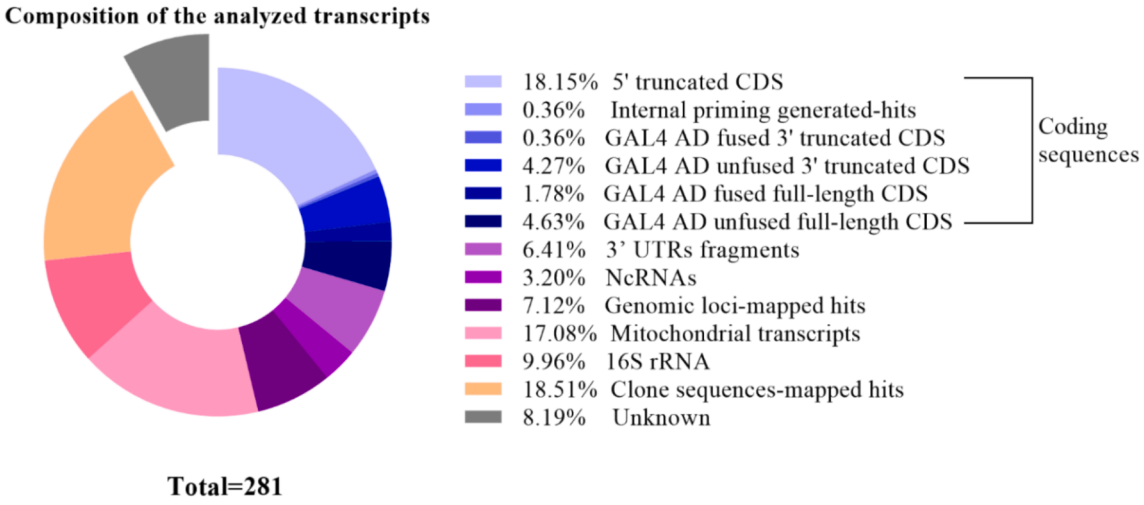

Fig. 7. Pie chart showing the percentage of analyzed cDNAs and other transcripts in this study.

4. Discussion

cDNA library screens are utilized to discover new gene functions and to identify interaction partners. The use of these screens has resulted in the identification of novel or disease-related genes,[22,23,42] as well as providing insights into gene function.[43] Furthermore, the usage of such screens enables the identification of different isoforms produced through alternative polyadenylation.[44] .

Although these libraries are powerful tools to study the complexity and functionality of the transcriptome, several challenges are encountered in the conventional construction methods of cDNA libraries. Consequently, a vast range of transcripts such as truncated cDNAs, ncRNAs, and mitochondrial transcripts exist in the library alongside the full-length cDNAs (Tables 1-6). In addition, genomic sequences are sometimes amplified as a consequence of internal priming[45–47] (Table 7).

In this study, sequencing was performed on 282 plasmids obtained from a human kidney cDNA library to identify the sequence of the DNA

fragment included in each. It is worth mentioning that several transcripts were identified multiple times, but we reported only one transcript for simplicity, except for the mitochondrial hits, for which we showed counts, as significant repetition was observed.

Selection using the oligo (dT) strategy allows specific isolation of mRNA subpopulation, since mRNA transcripts typically represent only 1–5 % of the total RNA extract in a given eukaryotic cell.[48] Although this method has been exploited in cDNA library construction, it can introduce bias during data interpretation since some mRNA transcripts have no poly-A sequence.[49,50] Interestingly, a large fraction of postnatal brain transcripts are poly(A)-[51]. Moreover, the expression of some mRNA transcripts is regulated in the cytoplasm by the addition or removal of the poly-A signal. Deadenylation keeps them in a silent state, while re-polyadenylation resumes the expression state[52]. Another issue arises during the identification of genes expressed at low level since they represent a small portion of the transcripts. This leads to a higher probability of identifying highly expressed genes over the low level during the screen process.[53] .

Another challenge involved the limitation of RT enzyme to produce long transcripts, which lead to significant numbers of 3′ and 5′ end truncated cDNAs, especially for mRNAs longer than 1 kb[54] (Tables 1 and 5). As previously reported, we have observed a notable prevalence of 5′ truncated clones.[55] The absence of full-length coding regions of the gene may give rise to the loss of regulatory regions or important functional domains within the gene of interest. To deal with this truncation, several strategies could be employed to mitigate this issue. For instance, previous studies reported that utilizing the thermostable group II intron RTs fusion proteins, in contrast to retroviral RTs, could lead to higher processivity[56,57]. In addition, the thermostable group II intron RTs are capable of consistently profiling the entire transcriptome of RNAs without fragmentation. This conserves the alternative splicing patterns of the long transcripts. They also allow for less-biased profiling of RNA fragments. Furthermore, the high processivity and fidelity of these enzymes enable the analysis of RNAs with secondary structures.[57] .

Even though the use of cDNA libraries has led to the identification of numerous ncRNAs,[58–62] the high abundance of non-coding transcripts complicates the process and analysis when the desired prey genes are coding. The same challenge arises with the use of Y2H screens, which aim to identify protein coding genes. Hence, specialized library construction methods could be employed to deplete ncRNAs, in order to improve the representation of protein-coding transcripts and reduce complexity. It is worth noting that size fractionation is usually implemented to eliminate the ncRNAs ranges 0.5–1 kb.[63] .

Despite the challenge of having a significant number of unfused transcripts for the Y2H approach, the absence of fusion could be advantageous for the utilization of the cDNA library in screening purposes (Fig. 1C). The GAL4 AD is fused to a nuclear localization signal (NLS) to enable the translated proteins to enter the nucleus to employ Y2H.[6] However, the presence of a fusion means a forced nuclear localization for the candidates. If the aim of the screen is to identify novel rescue genes (Fig. 1C), forcing the proteins to localize to the nucleus can prevent them from performing the rescue function if it requires the proteins to localize other cellular compartments. Overall, the absence of a fusion product can be advantageous for functional screens (Fig. 1C) but not for Y2H screens (Fig. 1B). It has been shown that 12 % of the expressed sequence tags (ESTs) in the current dbEST database are 3′ terminal truncated due to internal priming sites in mRNAs. It has been also suggested that the frequency of internal priming is three times higher than terminal priming.[4] In addition, internal priming may cause misidentification of 3′ ends of novel-originated poly-A sites in the studies focusing on alternative polyadenylation[64,65]. Thus, there is a need to decrease the adverse impact of internal priming through library construction. In this context, the use of modified oligo (dT) primers, called anchored primers, was introduced to reduce the higher binding stability of the internal primed oligo (dT).[66]This can be achieved by the addition of 1–2 nucleotides rather than T at the 3′ end of the oligo (dT). The recommended solution could reduce the rate of internal priming, but it does not prevent it.[66].

Although nuclear DNA contamination causes severe consequences during cDNA library synthesis, nuclear genes only exist in an average of two copies in somatic cells compared with hundreds to thousands of mtDNA copies.[67] The large fraction of mitochondrial transcripts in the cDNA library arises from cross-contamination of mtDNA during the RNA purification process or leakage of mitochondrial RNA from damaged cells.[23,24] These sequences get amplified due to the presence of a poly-A sequence at their 3′ end, as polyadenylation is essential in completing their termination codons.[38] Additionally, some mitochondrial genomic regions or tRNAs also get amplified through internal priming. The MssI restriction enzyme has been utilized to degrade the mtDNA in samples containing extrachromosomal circular DNA (eccDNA)[68]. Recently, a CRISPR/Cas9 system has been also implemented to cleave the circular mtDNA with guide RNA followed by digestion of linearized mtDNA by exonucleases[69]. These methods could be tested to determine whether they could decrease mtDNA contamination in RNA extracted for cDNA library construction. To the best of our knowledge, no reports have tested the effect of MssI on RNA integrity. It is worth noting that we observed a significant number of mitochondrial rRNA transcripts due to their abundance in RNA extracts[70]. These rRNAs should be eliminated prior to cDNA library preparation to avoid complicating analysis of obtained data. For instance, a sequence-specific probes approach was used to hybridize the rRNAs with biotinylated DNA or locked nucleic acid (LNA) probes. In addition, probe-directed degradation (PDD) and not-so-random (NSR) primers were two other approaches to deplete rRNA transcripts.[71] .

4.1. General recommendations for cDNA library construction

In our study, we used a pre-constructed library. However, we gathered recommendations from the literature that could be useful for researchers constructing cDNA libraries from scratch. The high abundance of truncated cDNAs is attributed to several reasons, including limitation in reverse transcription as discussed,[46] cDNA size fractionation bias, and mRNA sample degradation. Proper isolation of purified, high quality, and intact full-length mRNA is a critical step for cDNA construction[72,73]. Chong Wang & Hui Liu proposed performing mRNA-dependent experiments within two hours after mRNA extraction due to its relatively short half-life (T1/2) compared to other RNAs types. For instance, in blood samples, its half-life is only 16.4 hours[74]. Thus, cDNA library construction should be improved by proper preparation of fresh mRNA samples within an accelerated timeframe, along with appropriate storage, and incubation duration. The traditional method of mRNA isolation, performed in acidic PH solution,[72] yields a high amount of RNA; but its yield is not totally pure. Moreover, the remaining traces of phenol and chloroform have an adverse impact on the efficacy of RT. Therefore, it is recommended to apply an additional chloroform extraction and an additional 75 % ethanol washing RNA step in order to lower cycle threshold (Ct) values following reverse-transcription and RT-qPCR. In addition, this optimization is able to detect transcripts in low concentration and it is less time consuming compared with the conventional method[73]. Additionally, cDNA size selection is an important step before cDNA cloning into a vector. Following the protocol outlined by Sambrook et al. (1989), gel electrophoresis filtrates non-ligated adaptors, adaptor dimmers, and incomplete cDNA below 400 bp, but it does not efficiently select large full-length cDNA fragments. An optimized sucrose gradient method for cDNA size fractionation was suggested, in which larger cDNA fragments (>1.5 kb) are trapped in higher sucrose concertation.[75] .

4.2. The future of genetic screens

While our utilization of human cDNA library screens in yeast aimed to identify novel genes regulating specific pathways,[5,10,14,74] recently,

CRISPR-Cas9 represents a more efficient platform for genetic screening and has several advantages over cDNA libraries. The well-designed gRNA library used in pooled CRISPR screens, as well as validation of transfected cells using high-throughput sequencing, achieve unbiased phenotypic screening.[76,77] In contrast to cDNA library screens, in pooled CRISPR screens, the targeted cells are introduced to various stressors such as drug treatment, viral infection, and cell competition.[76,77] In Addition, CRISPR screening is applicable in various human cell lines and was successfully conducted in living animals such as transgenic mouse models, considering the entire body's response to physiological and pathological conditions[78]. While in the pooled CRISPR screens, the gRNA library is delivered into cells in bulk, in arrayed CRISPR screens the cells are physically separated in 96-well plate, allowing them to be introduced to multiple perturbations. These two types of CRISPR screens have advantages over other screening approaches, as the pooled screens are used for molecular discoveries and the arrayed screens are employed for follow up investigations through proteomic analysis, imaging technique, and metabolomics profiling.[77,79] Additionally, the CRISPR system is more versatile than cDNAs and can be utilized in multiple screen formats such as loss of function and gain of function screen and CRISPR screening in wild type cells.[80,81] Various Cas9 protein engineering approaches also allow for customization to suit specific experimental designs, facilitating diverse screening strategies.[82].

Furthermore, cDNA libraries are limited by their inability to control gene expression precisely to endogenous levels, potentially compromising the accuracy of gene physiological relevance assessment. Additionally, they face the challenge of accounting for the complexity of the cell's transcriptome[83–85]. In contrast, utilizing the CRISPR system in CRISPR activation screens ensures that every gene has an opportunity for expression, unlike cDNA libraries where longer genes often undergo incomplete transcription during reverse transcription.[54,86].

5. Conclusion

In this study, we have discussed different challenges encountered during cDNA library screens. These include the inefficiency of reverse transcription, resulting in truncated cDNAs, and the utilization of oligo (dT) method results in fishing of non-mRNA transcripts such as mitochondrial transcripts and ncRNA. Furthermore, internal priming results in the amplification of genomic regions. We have suggested some recommendations to decrease the noise of non-specific transcripts and amplification products to provide more efficient genetic and Y2H screens. Nevertheless, we have to acknowledge the limitations and possible weaknesses that may have impacted our findings and conclusion. In many cases, the quality of the reverse primer sequencing was very low, possibly due to the long poly-A sequence[87]. Therefore, the results of the reverse primer were not included for many hits. However, no assumptions were made in this case regarding the 3′UTR or whether the transcript arises from poly A tail or internal priming amplification.

Overall, coding transcripts co-exist with other undesired transcripts/noise and genomic fragments amplified through internal priming, fishing the positive hit could require massive screening processes, including repeated screens until a promising hit is identified. Therefore, we aim to raise awareness on challenges related to 'human cDNA library screening in yeast', which we believe is a valuable technique that have resulted in ground breaking discoveries.

Funding

This study is funded by Academy of Scientific Research and Technology-Bibliotheca Alexandrina (ASRT-BA 1445) to M Elserafy and SF El-Khamisy, International Centre for Genetic Engineering and Biotechnology (ICGEB)'CRP/EGY18-05_EC', Science and Technology Development Fund (STDF 33376), L'Oreal-UNESCO for women in science fellowship and Zewail City of Science and Technology internal grant (ZC-24) to M Elserafy. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CRediT authorship contribution statement

Zina Alaswad: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Nayera E. Attallah:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Basma Aboalazm:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Eman S. Elmeslhy:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Asmaa S. Mekawy:** Writing – review & editing, Methodology. **Fatma A. Afify:** Visualization, Methodology, Formal analysis. **Hesham K. Mahrous:** Methodology. **Ashrakat Abdalla:** Methodology. **Mai A. Rahmoon:** Writing – review & editing, Methodology. **Ahmed A. Mohamed:** Methodology. **Ahmed H. Shata:** Visualization, Methodology, Formal analysis. **Rana H. Mansour:** Methodology. **Fareed Aboul-ela:** Writing – review & editing, Resources. **Mohamed Elhadidy:** Writing – review & editing, Supervision, Resources. **Biola M. Javierre:** Resources. **Sherif F. El-Khamisy:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization. **Menattallah Elserafy:** Writing – review & editing, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We express our deep appreciation to Reem A. Elghaish for her assistance with PolyASite database.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jgeb.2024.100427.

References

1. Abugable AA, Awwad DA, Fleifel D, Ali MM, El-Khamisy S, Elserafy M. Personalised medicine: genome maintenance lessons learned from studies in yeast as a model organism. *Adv Exp Med Biol*. 2017. https://doi.org/10.1007/978-3-319-60733-7_9.
2. Banik SSR, Counter CM. From bread to bedside: what budding yeast has taught us about the immortalization of cancer cells. *Yeast as a Tool Cancer Res*. 2007:123–139. https://doi.org/10.1007/978-1-4020-5963-6_5.
3. Wan KH, Yu C, George RA, et al. High-throughput plasmid cDNA library screening. *Nat Protoc*. 2006;1:624–632. https://doi.org/10.1038/nprot.2006.90.
4. Nam DK, Lee S, Zhou G, et al. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *PNAS*. 2002;99:6152–6156. https://doi.org/10.1073/PNAS.092140899/ASSET/08839A23-CD85-4252-8E76-5928DB218856/ASSETS/GRAPHIC/PQ0921408005.JPEG.
5. Okayama H. Functional cDNA expression cloning: pushing it to the limit. *Proc Japan Acad Ser B Phys Biol Sci*. 2012;88:102–119. https://doi.org/10.2183/pjab.88.102.
6. Fields S, Song O. A novel genetic system to detect protein–protein interactions. *Nature*. 1989;340:245–246.
7. Mehla J, Caufield JH, Uetz P. Mapping protein-protein interactions using yeast two-hybrid assays. *Cold Spring Harb Protoc*. 2015;2015:442–452. https://doi.org/10.1101/pdb.prot086157.
8. Rajagopala SV. Mapping the protein–protein interactome networks using yeast two-hybrid screens. *Adv Exp Med Bio*. 2015;883:187–214. https://doi.org/10.1007/978-3-319-23603-2_11.
9. Miller J, Stagljar I. Using the yeast two-hybrid system to identify interacting proteins. *Methods Mol Biol*. 2004;261:247–262. https://doi.org/10.1385/1-59259-762-9:247.
10. Van Criekinge W, Beyaert R. Yeast two-hybrid: state of the art. *Biol Proced Online*. 1999;2:1–38.

11. Lin J-S, Lai E-M. Protein–protein interactions: yeast two-hybrid system. *Bact Protein Secret Syst Methods Protoc.*. 2017:177–187.
12. Rolland T, Taşan M, Charloteaux B, et al. A proteome-scale map of the human interactome network. *Cell*. 2014;159:1212–1226.
13. Schwikowski B, Uetz P, Fields S. A network of protein–protein interactions in yeast. *Nat Biotechnol*. 2000;18:1257–1261.
14. Lathouwers T, Wagemans J, Lavigne R. Identification of protein–protein interactions using pool-array-based yeast two-hybrid screening. *Methods Mol Biol*. 2018;1794: 29–48. https://doi.org/10.1007/978-1-4939-7871-7_3.
15. Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U. Yeast two-hybrid, a powerful tool for systems biology. *Int J Mol Sci*. 2009;10:2763–2788. https://doi.org/10.3390/ijms10062763.
16. Johzuka K, Ogawa H. Interaction of Mre11 and Rad50: two proteins required for DNA repair and meiosis-specific double-strand break formation in Saccharomyces cerevisiae. *Genetics*. 1995;139:1521–1532.
17. Hays SL, Firmenich AA, Berg P. Complex formation in yeast double-strand break repair: participation of Rad51, Rad52, Rad55, and Rad57 proteins. *Proc Natl Acad Sci*. 1995;92:6925–6929.
18. Jiang H, Xie Y, Houston P, et al. Direct association between the yeast Rad51 and Rad54 recombination proteins. *J Biol Chem*. 1996;271:33181–33186.
19. El-Khamisy SF, Caldecott KW. DNA single-strand break repair and spinocerebellar ataxia with axonal neuropathy-1. *Neuroscience*. 2007;145:1260–1266. https://doi.org/10.1016/j.neuroscience.2006.08.048.
20. Petrini JHJ, Walsh ME, DiMare C, Chen X-N, Korenberg JR, Weaver DT. Isolation and characterization of the human MRE11 homologue. *Genomics*. 1995;29:80–86.
21. Holz C, Lueking A, Bovekamp L, et al. A human cDNA expression library in yeast enriched for open reading frames. *Genome Res*. 2001;11:1730–1735.
22. Ledesma FC, El Khamisy SF, Zuma MC, Osborn K, Caldecott KW. A human 5′-tyrosyl DNA phosphodiesterase that repairs topoisomerase-mediated DNA damage. *Nature*. 2009;461:674–678. https://doi.org/10.1038/nature08444.
23. Lustig AJ, Petes TD. Identification of yeast mutants with altered telomere structure. *Proc Natl Acad Sci*. 1986;83:1398–1402.
24. Osborn MJ, Miller JR. Rescuing yeast mutants with human genes. *Brief Funct Genomics*. 2007;6:104–111. https://doi.org/10.1093/bfgp/elm017.
25. Macosko EZ, Basu A, Satija R, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*. 2015;161:1202–1214. https://doi.org/10.1016/j.cell.2015.05.002.
26. Massively parallel digital transcriptional profiling of single cells.
27. Klein AM, Mazutis L, Akartuna I, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. 2015;161:1187–1201. https://doi.org/10.1016/j.cell.2015.04.044.
28. Yu B, Van Tol HTA, Stout TAE, Roelen BAJ. Reverse transcription priming methods affect normalisation choices for gene expression levels in oocytes and early embryos. *Mol Hum Reprod*. 2021;27. https://doi.org/10.1093/molehr/gaab040.
29. Eberwine J, Yeh H, Miyashiro K, et al. Analysis of gene expression in single live neurons. *PNAS*. 1992;89:3010–3014. https://doi.org/10.1073/pnas.89.7.3010.
30. Verma IM. Genome organization of RNA tumor viruses. I. In vitro synthesis of full-genome-length single-stranded and double-stranded viral DNA transcripts. *J Virol*. 1978;26:615–629. https://doi.org/10.1128/JVI.26.3.615-629.1978.
31. Weiss GB, Wilson GN, Steggles AW, Anderson WF. Importance of full size complementary DNA in nucleic acid hybridization. *J Biol Chem*. 1976;251: 3425–3431. https://doi.org/10.1016/S0021-9258(17)33455-5.
32. Chang C, Gonzalez F, Rothermel B, Sun L, Johnston SA, Kodadek T. The Gal4 activation domain binds Sug2 protein, a proteasome component, in vivo and in vitro *. *J Biol Chem*. 2001;276:30956–30963. https://doi.org/10.1074/jbc.M102254200.
33. Zhang Y, Yang L, Chen LL. Life without A tail: new formats of long noncoding RNAs. *Int J Biochem Cell Bio. Elsevier Ltd*. 2014:338–349. https://doi.org/10.1016/j.biocel.2013.10.009.
34. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. *Cell*. 2013; 154:26–46. https://doi.org/10.1016/J.CELL.2013.06.020.
35. Murray A, Vollmers C, Schmitz RJ. Smar2C2: a simple and efficient protocol for the identification of transcription start sites. *Curr Protoc*. 2023;3:1–15. https://doi.org/10.1002/cpz1.705.
36. Ojala D, Montoya J, Attardi G. TRNA punctuation model of RNA processing in human mitochondria. *Nature*. 1981;290:470–474. https://doi.org/10.1038/290470a0.
37. Pearce SF, Rebelo-Guiomar P, D'Souza AR, Powell CA, Van Haute L, Minczuk M. Regulation of mammalian mitochondrial gene expression: recent advances. *Trends Biochem Sci*. 2017;42:625–639. https://doi.org/10.1016/j.tibs.2017.02.003.
38. Nagaike T, Suzuki T, Katoh T, Ueda T. Human mitochondrial mRNAs are stabilized with polyadenylation regulated by mitochondria-specific poly(A) polymerase and polynucleotide phosphorylase. *J Biol Chem*. 2005;280:19721–19727. https://doi.org/10.1074/jbc.M500804200.
39. Slomovic S, Laufer D, Geiger D, Schuster G. Polyadenylation and Degradation of Human Mitochondrial RNA: the Prokaryotic Past Leaves Its Mark. 2023 25 6427–6435. doi:10.1128/MCB.25.15.6427-6435.2005.
40. Bratic A, Clemente P, Calvo-Garrido J, et al. Mitochondrial polyadenylation is a one-step process required for mRNA integrity and tRNA maturation. *PLoS Genet*. 2016; 12:1–21. https://doi.org/10.1371/journal.pgen.1006028.
41. Baserga SJ, Linnenbach AJ, Malcolm S, et al. Polyadenylation of a human mitochondrial ribosomal RNA transcript detected by molecular cloning. *Gene*. 1985; 35:305–312. https://doi.org/10.1016/0378-1119(85)90009-5.
42. Xiao-hong C, Zhi C, Hang-ping Y, Feng C, Hai-hong Z, Hong-juan Z. Construction and characterization of a cDNA library from human liver tissue with chronic hepatitis B. *J Zhejiang Univ Sci B*. 2005;6:288–294.
43. Hu W, Liu N, Tian Y, Zhang L. Molecular cloning, expression, purification, and functional characterization of dammarenediol synthase from Panax ginseng. *Biomed Res Int*. 2013.
44. Wang P, Yu P, Gao P, Shi T, Ma D. Discovery of novel human transcript variants by analysis of intronic single-block EST with polyadenylation site. *BMC Genomics*. 2009;10:1–15. https://doi.org/10.1186/1471-2164-10-518.
45. Harrison GP, Mayo MS, Hunter E, Lever AML. Pausing of reverse transcriptase on retroviral RNA templates is influenced by secondary structures both 5' and 3' of the catalytic site. *Nucleic Acids Res*. 1998;26:3433–3442. https://doi.org/10.1093/nar/26.14.3433.
46. Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques*. 2001;30:892–897. https://doi.org/10.2144/01304PF02.
47. Rashighi M, Harris JE. Interpreting reverse transcriptase termination and mutation events for greater insight into the chemical probing of RNA. *Physiol Behav*. 2017; 176:139–148. https://doi.org/10.1053/j.gastro.2016.08.014.CagY.
48. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. From DNA to RNA. *Garland Sci*. 2002.
49. Wang SM, Fears SC, Zhang L, Chen JJ, Rowley JD. Screening poly(dA/dT)- cDNAs for gene identification. *PNAS*. 2000;97:4162–4167. https://doi.org/10.1073/pnas.97.8.4162.
50. Viscardi MJ, Arribere JA. Poly(a) selection introduces bias and undue noise in direct RNA-sequencing. BMC Genomics. 2022 23. doi:10.1186/S12864-022-08762-8.
51. Van Ness J, Maxwell IH, Hahn WE. Complex population of nonpolyadenylated messenger RNA in mouse brain. *Cell*. 1979;18:1341–1349. https://doi.org/10.1016/0092-8674(79)90244-7.
52. Winata CL, Łapiński M, Pryszcz L, et al. Cytoplasmic polyadenylation-mediated translational control of maternal mRNAs directs maternal-to-zygotic transition. *Dev*. 2018;145:1–12. https://doi.org/10.1242/dev.159566.
53. De Fatima BM, Lennon G, Soares MB. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res*. 1996;6:791–806. https://doi.org/10.1101/GR.6.9.791.
54. Malboeuf CM, Isaacs SJ, Tran NH, Kim B. Thermal effects on reverse transcription: Improvement of accuracy and processivity in cDNA synthesis. *Biotechniques*. 2001; 30:1074–1084. https://doi.org/10.2144/01305rr06.
55. Wellenreuther R, Schupp I, Poustka A, Wiemann S. SMART amplification combined with cDNA size fractionation in order to obtain large full-length clones. *BMC Genomics*. 2004;5:1–8. https://doi.org/10.1186/1471-2164-5-36.
56. Blocker FJH, Mohr G, Conlan LH, Qi L, Belfort M, Lambowitz AM. Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase. *RNA*. 2005;11:14–28. https://doi.org/10.1261/rna.7181105.
57. Mohr S, Ghanem E, Smith W, et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA*. 2013;19:958–970.
58. Singh PK, Singh P, Singh RP, Singh RL. From gene to genomics: tools for improvement of animals. *Adv Anim Genomics*. 2021:13–32. https://doi.org/10.1016/B978-0-12-820595-2.00002-3.
59. Peaston AE, Graber JH, Knowles BB, de Vries WN. Interrogating the transcriptome of oocytes and preimplantation embryos. *Methods Enzymol*. 2010;477:481–510. https://doi.org/10.1016/S0076-6879(10)77024-5.
60. Zhang P, Wu W, Chen Q, Chen M. Non-coding RNAs and their integrated networks. *J Integr Bioinform*. 2019;16:1–12. https://doi.org/10.1515/jib-2019-0027.
61. Sun Q, Hao Q, Prasanth KV. Nuclear long noncoding RNAs: key regulators of gene expression. *Trends Genet*. 2018;34:142–157. https://doi.org/10.1016/j.tig.2017.11.005.
62. Rederstorff M, Bernhart SH, Tanzer A, et al. RNPomics: defining the ncRNA transcriptome by cDNA library generation from ribonucleo-protein particles. *Nucleic Acids Res*. 2010:38. https://doi.org/10.1093/nar/gkq057.
63. Yadav RK, Barbi F, Ziller A, et al. Construction of sized eukaryotic cDNA libraries using low input of total environmental metatranscriptomic RNA. *BMC Biotech*. 2014; 14:1–6. https://doi.org/10.1186/1472-6750-14-80.
64. Zhang H, Hu J, Recce M, Tian B. PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res*. 2005:33. https://doi.org/10.1093/NAR/GKI055.
65. Lee JY, Yeh I, Park JY, Tian B. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res*. 2007;35:165–168. https://doi.org/10.1093/nar/gkl870.
66. Svoboda M, Frost HR, Bosco G. Internal oligo(dT) priming introduces systematic bias in bulk and single-cell RNA sequencing count data. *NAR Genomics Bioinforma*. 2022;4:1–12. https://doi.org/10.1093/nargab/lqac035.
67. Wallace DC. Why Do We Still Have a Maternally Inherited Mitochondrial DNA? Insights from Evolutionary Medicine. 2007 76: 781–821. doi:10.1146/ANNUREV.BIOCHEM.76.081205.150955.
68. Møller HD, Mohiyuddin M, Prada-Luengo I, et al. Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. *Nat Commun*. 2018;9:1–12. https://doi.org/10.1038/s41467-018-03369-8.
69. Feng W, Arrey G, Zole E, lv W, Liang X, Han P, et al. Targeted removal of mitochondrial DNA from mouse and human extrachromosomal circular DNA with CRISPR-Cas9. Comput Struct Biotechnol J. 2022;20: 3059–3067. doi:10.1016/j.csbj.2022.06.028.
70. O'Neil D, Glowatz H, Schlumpberge M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol. 2013;Chapter 4. doi:10.1002/0471142727.MB0419S103.
71. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*. 2017;8. https://doi.org/10.1002/wrna.1364.

72. Chomczynski P, Sacchi N. The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. *Nat Protoc*. 2006;1:581–585. https://doi.org/10.1038/nprot.2006.83.
73. Toni LS, Garcia AM, Jeffrey DA, et al. Optimization of Phenol-chloroform RNA extraction. *Methodsx*. 2018;5:599–608. https://doi.org/10.1016/j.mex.2018.05.011.
74. Wang C, Liu H. Factors influencing degradation kinetics of mRNAs and half-lives of microRNAs, circRNAs, lncRNAs in blood in vitro using quantitative PCR. *Sci Rep*. 2022;12. https://doi.org/10.1038/s41598-022-11339-w.
75. Kieffer BL. Short Communications Optimised cDNA size selection and cloning procedure for the construction of representative plasmid cDNA libraries. *Gene*. 1991.
76. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. *Nat Rev Genet*. 2015;16:299–311. https://doi.org/10.1038/nrg3899.
77. Bock C, Datlinger P, Chardon F, et al. High-content CRISPR screening. *Nat Rev Methods Prim*. 2022;2. https://doi.org/10.1038/s43586-021-00093-4.
78. Kuhn M, Santinha AJ, Platt RJ. Moving from in vitro to in vivo CRISPR screens. *Gene Genome Ed*. 2021;2, 100008. https://doi.org/10.1016/j.ggedit.2021.100008.
79. Doench JG. Am i ready for CRISPR? A user's guide to genetic screens. *Nat Rev Genet*. 2018;19:67–80. https://doi.org/10.1038/nrg.2017.97.
80. Ming M, Long H, Ye Z, et al. Highly efficient CRISPR systems for loss-of-function and gain-of-function research in pear calli. *Hortic Res*. 2022:9. https://doi.org/10.1093/hr/uhac148.
81. Condon KJ, Orozco JM, Adelmann CH, et al. Genome-wide CRISPR screens reveal multitiered mechanisms through which mTORC1 senses mitochondrial dysfunction. *PNAS*. 2021:118. https://doi.org/10.1073/pnas.2022120118.
82. Chan Y, Lu Y, Wu J, Zhang C, Tan H, Bian Z, et al. Theranostics CRISPR-Cas9 library screening approach for anti-cancer drug discovery : overview and perspectives. 2022 12. doi:10.7150/thno.71144.
83. Johannessen CM, Johnson LA, Piccioni F, et al. A melanocyte lineage program confers resistance to MAP kinase pathway inhibition. *Nature*. 2013;504:138–142. https://doi.org/10.1038/nature12688.
84. Konermann S, Brigham MD, Trevino AE, et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*. 2015;517:583–588. https://doi.org/10.1038/nature14136.
85. Xue HY, Ji LJ, Gao AM, Liu P, He JD, Lu XJ. CRISPR-Cas9 for medical genetic screens: applications and future perspectives. *J Med Genet*. 2016;53:91–97. https://doi.org/10.1136/jmedgenet-2015-103409.
86. Joung J, Kirchgatterer PC, Singh A, et al. CRISPR activation screen identifies BCL-2 proteins and B3GNT2 as drivers of cancer resistance to T cell-mediated cytotoxicity. *Nat Commun*. 2022;13. https://doi.org/10.1038/s41467-022-29205-8.
87. Liu Q, Zhang P, Wang D, Gu W, Wang K. Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. *Genome Med*. 2017;9:1–16. https://doi.org/10.1186/S13073-017-0456-7/FIGURES/6.