scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A Dataset Curated for the Assessment of G4s in the LncRNAs Dysregulated in Various Human Cancers

Shubham Sharma[1], Muhammad Yusuf Hassan[2,3], Noman Hanif Barbhuiya[4], Ramolia Harshit Mansukhbhai[2,3], Chinmayee Shukla[1], Deepshikha Singh[1] & Bhaskar Datta[1,5] ✉

Dysregulated expression of long non-coding RNAs (lncRNAs) in cancer contributes to various hallmarks of the disease, presenting novel opportunities for diagnosis and therapy. G-quadruplexes (G4s) within lncRNAs have gained attention recently; however, their systematic evaluation in cancer biology is yet to be performed. In this work, we have formulated a comprehensive dataset integrating experimentally-validated associations between lncRNAs and cancer, and detailed predictions of their G4-forming potential. The dataset categorizes predicted G4-motifs into anticipated G4 types (2G, 3G, and 4G) and provides information about the subcellular localization of the corresponding lncRNAs. It describes lncRNA-RNA and lncRNA-protein interactions, together with the RNA G4-binding capabilities of these proteins. The dataset facilitates the investigation of G4-mediated lncRNA functions in diverse human cancers and provides distinctive leads about G4-mediated lncRNA-protein interactions.

Background & Summary

Only a fraction of the human genome directly codes for proteins, while the majority is composed of non-coding regions. This apparent paradox was partly explained by the discovery of non-coding RNAs (ncRNAs)[1–5]. Long non-coding RNAs (lncRNAs) are a prominent subset of ncRNAs and have been implicated in various functions in cellular physiology and disease progression[6,7]. LncRNAs are dysregulated and mutated across various cancer types and can exhibit oncogenic or tumor-suppressing functions[8–13]. The dysregulation of lncRNA disrupts cellular homeostasis and contributes to cancer cell proliferation, migration, invasion, metastasis, apoptosis, altered cell metabolism, cell cycle regulation, and treatment resistance[14–19]. Key questions persist regarding the precise mechanism of lncRNA activity *vis-à-vis* cancer pathogenesis[13,20].

While dysregulation of lncRNAs has been correlated to cancer growth and progression, the impact of lncRNA secondary structures on disease progression is not entirely understood[21,22]. G-quadruplexes (G4s) are a prominent higher-order structure within the repertoire of structures adopted by lncRNAs, possessing biological significance[5,22,23]. Among the subset of lncRNAs harbouring G4s, *GSEC*, *REG1CP*, *LUCAT1*, *NEAT1*, and *HOTAIR* have emerged as noteworthy contributors to cancer development, underscoring the pivotal role of these structures in fostering cellular malignancy. Such studies have also emphasized the importance of lncRNA G4-protein interactions in regulating gene expression and promoting cancer progression[20–22,24–27].

Several informatics and computational tools have emerged that enable the prediction and assessment of G4-motifs in DNA and RNA. These are based on different methodologies, such as regular expression matching, score-based ranking, and machine learning methods[28–39]. A select number of resources focus specifically on RNA G4s, including the datasets G4RNA, GRSDB2, GRS_UTRdb, and QUADRatlas, which provide

[1]Department of Biological Sciences and Engineering, Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat, 382055, India. [2]Department of Electrical Engineering, Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat, 382055, India. [3]Department of Computer Science and Engineering, Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat, 382055, India. [4]Department of Physics, Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat, 382055, India. [5]Department of Chemistry, Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat, 382055, India. ✉e-mail: bdatta@iitgn.ac.in

information on experimentally-validated and predicted RNA G4s, particularly in mRNAs and untranslated regions (UTRs)[40–42]. Datasets like G4LDB and G4IPDB catalogue RNA G4-ligands and proteins, and resources like ONQUADRO and DSSR-G4DB display RNA and G4 structures[43–46]. A limited number of resources have systematically documented lncRNAs in cancers[47]. The platform Lnc2Cancer focuses on lncRNA-cancer associations, while datasets like NPInter and LncTarD catalogue lncRNA interacting partners and disease associations[48–50]. Another dataset, LncATLAS, provides information on the subcellular (cytoplasmic to nuclear) localization of the lncRNAs across diverse human cell lines[51]. There is an absence of platforms that integrate different data types and enable the correlation of G4s with established lncRNAs and their associations with cancer, along with the exploration of G4-mediated lncRNA-protein interactions. Such platforms are required to facilitate the discovery of cellular mechanisms involving lncRNA G4s.

In this work, we describe a meticulously curated dataset consolidating 17,666 experimentally-validated associations between 6,408 human lncRNAs, encompassing their transcript variants, and 15 distinct types of human cancers. The dataset is named CanLncG4 and offers: (1) an extensive G4-prediction analysis for each lncRNA transcript variant with categorization of predicted G4-motifs into anticipated G4 types (2 G, 3 G, and 4 G), (2) the subcellular localization of catalogued lncRNAs across a diverse range of cell lines, and (3) the meta-analysis of lncRNA interacting partners (RNA and protein) and information on RNA G4-binding proteins (RGBPs). Considering that the majority of catalogued lncRNAs contain putative G4-forming regions, information on proteins interacting with these lncRNAs and their established RNA G4-binding potential can serve as an informed starting point for investigating G4-mediated lncRNA-protein interactions.

Notably, the G4-harbouring lncRNAs are often dysregulated in various diseases[20,24–26,52–55]. The propensity of G4-formation has been observed to differ between diseased and normal states[56–59]. This distinctive G4-forming potential of lncRNAs can serve as a molecular hook for capturing the lncRNAs engaging with specific interacting partners. An unorthodox bottom-up approach can facilitate the experimental investigation into lncRNA biology that is pertinent to cancer growth and progression, and extend to other diseases. Consequently, the detection of such lncRNAs, their cognate G4s, and interacting partners may contribute to early disease diagnosis and offer potential therapeutic avenues.

Over the past decade, G4-specific ligands have emerged as promising tools for reporting, stabilizing, or destabilizing G4s[60–65]. Notably, a few G4-ligands, including CX-5461 and QN-302, have reached clinical investigations for cancer therapy. CX-5461 showed antitumor potential in Phase 1 dose escalation studies for advanced hematologic malignancies and solid tumors enriched for DNA-repair deficiencies[66,67]. QN-302, with preclinical efficacy in PDAC, recently received Food and Drug Administration (FDA) IND clearance to initiate Phase 1a clinical trials for the treatment of Pancreatic Cancer[68–70]. These advancements highlight the therapeutic potential of targeting lncRNA G4s across different cancers.

The Nobel Prize in Physiology or Medicine 2024 (https://www.nobelprize.org/prizes/medicine/2024/press-release) underscores the fundamental and physiological significance of ncRNAs (miRNAs), reinforcing the urgency to decode ncRNA biology. In this context, the CanLncG4 dataset will likely help researchers prioritize *in vitro* and *in cella* investigations of specific lncRNAs, thereby accelerating the discovery of lncRNAs and their interactome with diagnostic and therapeutic potential in cancer research. Given that most catalogued lncRNAs harbour putative G4-forming regions, understanding their interaction with proteins – particularly those with established RNA G4-binding potential can serve as a rational starting point for experimental studies on G4-mediated lncRNA-protein interactions. This approach is based on the hypothesis that experimentally-validated RGBPs, already known to interact with the G4-forming lncRNAs, are highly likely to engage with such lncRNAs by virtue of their G4s. Implementing this targeted strategy will streamline the screening of promising candidates for detailed investigation, making the process more time- and resource-efficient.

Looking ahead, the continual expansion of the dataset to encompass lncRNAs associated with additional cancers will broaden its applicability and contribute to advancing our understanding of the intricate roles played by lncRNAs in cancer biology. CanLncG4 serves as a comprehensive dataset for accessing critical information on cancer-dysregulated lncRNAs, their G4-forming potential, and interacting partners, supporting future research and therapeutic advancements.

Methods

Data Collection and Generation. *In silico G4-prediction in Cancer-Dysregulated LncRNAs.* The comprehensive list of lncRNAs dysregulated in diverse human cancers, their expression patterns, methodologies of their identification, and references to research articles (PubMed ID) used to gather these details were obtained from the Lnc2Cancer 3.0 dataset (http://bio-bigdata.hrbmu.edu.cn/lnc2cancer/download.html) (Fig. 1a)[48]. The aliases of all these lncRNAs were manually compiled from the GeneCards dataset (https://www.genecards.org)[71]. The nucleotide sequences and the corresponding NCBI accession numbers of all the identified lncRNAs, including their functional transcript variants, with "validated" or "reviewed" RefSeq status, were retrieved from NCBI Nucleotide dataset (https://www.ncbi.nlm.nih.gov/nucleotide)[72]. CanLncG4 dataset documents 17,666 entries establishing correlations between 6,408 human lncRNAs, including their transcript variants, and 15 distinct types of human cancers. Incorrect, missing, or duplicate entries were identified and addressed through meticulous examination of the sourced data. The cancers included in the dataset are General: Head and neck, skin, lung, liver, gastric, colorectal, brain, bone, and blood cancer; Male-dominant: prostate and testicular cancer; Female-dominant: breast, ovarian, uterine, and cervical cancer.

For identification of Putative Quadruplex-forming Sequences (PQS) within these lncRNAs, their FASTA sequences are imported to the QGRS mapper (https://bioinformatics.ramapo.edu/QGRS/analyze.php), a tool that presents data on the constitution and distribution of Quadruplex-forming G-rich sequences (QGRS) (Figs. 1a, 2a). The QGRS mapper identifies sequences as PQS based on the alignment with canonical sequence of

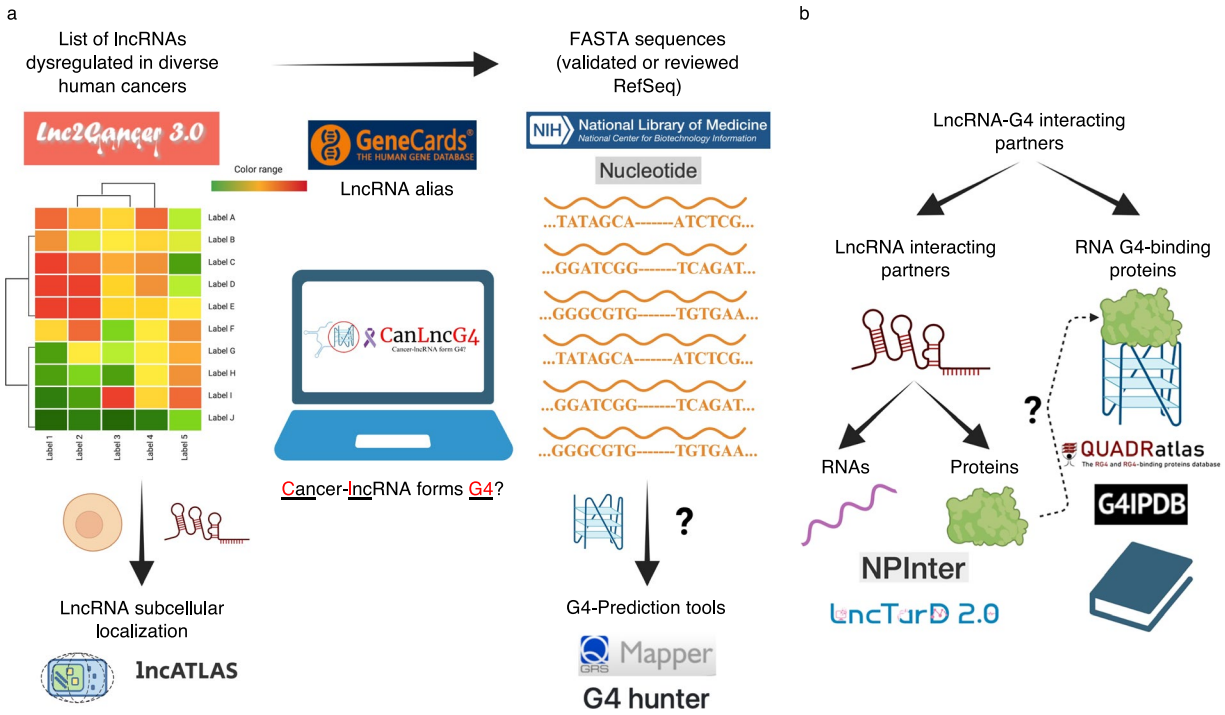

Fig. 1 Schematic of CanLncG4 dataset workflow. (**a**) *In silico* prediction of G4-formation in the lncRNAs dysregulated in human cancers using Lnc2Cancer 3.0 dataset: lncRNA-cancer association, GeneCards dataset: lncRNA aliases, NCBI Nucleotide dataset: lncRNA FASTA sequence; G4-prediction tool: QGRS mapper and G4Hunter; LncATLAS dataset: lncRNA subcellular localization. (**b**) Identification of lncRNA-G4 interacting partners using NPInter v4.0 and LncTarD 2.0 datasets: LncRNA-RNA and LncRNA-protein interactions; QUARRatlas and G4IPDB datasets, and scientific literature mining: RNA G4-binding proteins (RGBPs).

the format: $G_xN_{y1}G_xN_{y2}G_xN_{y3}G_x$, where x = number of G-quartets in a G4 or length of G-tract (tandem repeats of guanines), and y1, y2, y3 = lengths of loops, which collectively define the G-Score of each PQS[33]. All the PQS possible in catalogued lncRNAs, encompassing their transcript variants, are identified using the following parameters: max length: 45; min G-group: 2; loop size: 0 to 36. Only the highest-scoring PQS amongst all the overlapping candidates are presented to ease the PQS selection process for the user.

The PQS within these lncRNAs are also identified using G4Hunter (https://bioinformatics.ibp.cz/#/analyse/quadruplex), a tool for identifying putative G4-forming motifs based on the G-richness and G-skewness of the query sequence (Figs. 1a, 2a)[37]. The FASTA sequences of each catalogued lncRNA, including their transcript variants, are fed into the G4Hunter algorithm, and all the possible PQS are identified using the following parameters: window size: 45; threshold: 0.9. The output G4H Score conveys the probability of G4-formation by the query sequence. The G4Hunter algorithm was slightly modified to present only the highest-scoring PQS amongst the overlapping ones and to preclude the generation of consensus sequences containing all the overlapping PQS, albeit with different scores. Multiple parameter combinations were employed to identify PQS within these lncRNAs using QGRS mapper and G4Hunter. The above-mentioned parameters yielded more relevant and accurate predictions of the G4-forming potential.

The plots: (1) cytoplasmic to nuclear localization: relative concentration index (RCI) and expression values, and (2) cytoplasmic to nuclear localization: RCI distribution, for the catalogued lncRNAs across diverse human cell lines, were sourced from the LncATLAS dataset (https://lncatlas.crg.eu) to describe their subcellular localization (Fig. 1a)[51].

Identification of LncRNA-G4 Interacting Partners. The details of experimentally-validated RNA and protein interacting partners of catalogued lncRNAs were sourced from NPInter v4.0 (http://bigdata.ibp.ac.cn/npinter4/download) and LncTarD 2.0 (http://lnctard.bio-database.com/Download) datasets[49,50]. The data obtained from LncTarD 2.0 was filtered to present the data exclusively in the context of human cancers. The information on the experimentally-validated RNA G4-binding proteins (RGBPs) interacting with the catalogued lncRNAs was obtained from QUADRatlas (https://rg4db.cibio.unitn.it/download) and G4IPDB (https://people.iiti.ac.in/~amitk/bsbe/ipdb/g4rna.php) datasets, supplemented by scientific literature mining (Figs. 1b, 2b)[42,44,73]. For literature mining, a web search using the keyword strings ["RNA" "G4" "binding" "protein"] and ["RNA G4 binding protein"] was conducted to identify relevant research articles. These articles were then manually assessed for RGBP-related data. Additionally, information from one dissertation obtained through web searching was also included[73].

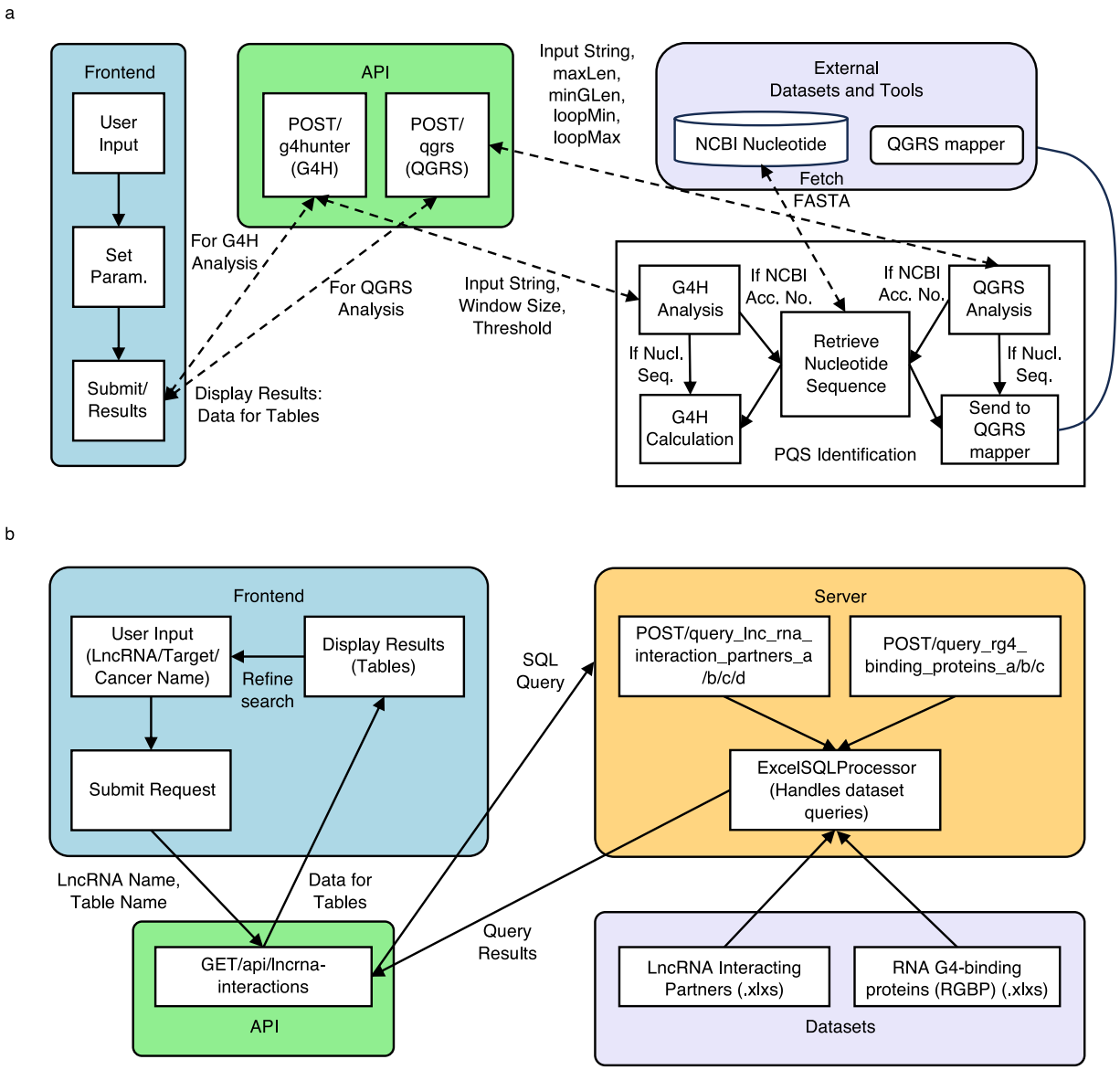

Fig. 2 Flowchart illustrating the frontend-backend workflow for identifying PQS in cancer-dysregulated lncRNAs and lncRNA-G4 interacting partners using datasets and G4-prediction tools. (**a**) The process of PQS identification begins with user input and setting parameter in the frontend, followed by API-based G4-prediction using QGRS Mapper and G4Hunter. PQS identification involves analysing the user-provided nucleotide sequence or retrieving a sequence from the NCBI nucleotide dataset if an accession number is provided. The sequence and parameter information are then sent to QGRS Mapper or subjected to G4Hunter calculations. The results from PQS identification are returned to the API and displayed as tables at the frontend. (**b**) The process of identifying lncRNA-G4 interacting partners begins with user input of a lncRNA or target or cancer name in the frontend, which sends a request to the API. The API then communicates with the server, where an Excel-SQL processor handles dataset queries to retrieve data from Excel sheets containing generated and analysed datasets on lncRNA interacting partners and RNA G4-binding proteins (RGBP). The results from dataset queries are returned to the API and displayed as tables at the frontend. Solid arrows indicate one-way data flow, while double-headed dashed arrows represent bidirectional processes.

Data Records

The dataset is available at Figshare: CanLncG4 dataset (https://figshare.com/collections/CanLncG4_dataset/7510452), with this section being the primary source of information on the availability and content of the data being described[74]. The CanLncG4 dataset is further segregated into several datasets and plots that were generated and analysed using various external datasets and G4-prediction tools, as outlined in the Methods section (Figs. 1, 2). The folder composition and the elements of the CanLncG4 dataset are as follows:

Datasets. The "Datasets" folder includes datasets generated and analysed during the current study. Datasets 1–16 provide information on the experimentally-validated associations between human lncRNAs and 15 human

cancers (Lnc2cancer 3.0 dataset), along with comprehensive G4-prediction (QGRS mapper and G4Hunter) and aliases (GeneCards dataset) for each catalogued lncRNA, including their transcript variants, available as Excel sheets (.xlsx)[33,37,48,71]. Dataset 1: All cancer-LncRNA G4s dataset, contains data concerning all 15 cancers, and datasets 2–16: Cancer Name-LncRNA G4s dataset, contain data related to individual cancers. These sheets include details on: LncRNA Name (Column A) and name of the cancer associated with the lncRNA (Cancer Name: Column B), expression pattern of the lncRNA (Expression Pattern: Column C), experimental techniques used to determine the lncRNA expression or identify the lncRNA-cancer association (Methods of Identification: Column D), reference to research articles linked to the association (PubMed ID: Column E), number of transcript variants of the lncRNA (No. of Transcript Variants (lncRNAs): Column F), LncRNA Aliases (Column G), reference to the lncRNA (NCBI accession number: Column H), status of the RefSeq sequence (RefSeq status: Column I), and categorization of the PQS (numbers) predicted by the G4-prediction tools into anticipated G4 types – 2G, 3G, and 4 G (QGRS Mapper Output, Max length − 45, Min G-group – 2: Columns J – M; G4 Hunter Output, Window - 45, Threshold: 0.9: Columns N – Q; G4 Hunter Output, Window - 45, Threshold: 1.4: Column R – U).

Datasets 17–20 feature a comprehensive meta-analysis of experimentally-validated interacting partners (RNAs and proteins) associated with the catalogued lncRNAs (NPInter v4.0 and LncTarD 2.0 datasets), presented as Excel sheets (.xlxs)[49,50]. Datasets 17: LncRNA-Protein Interactions dataset_NPInter, and dataset 19: LncRNA-RNA Interactions dataset_NPInter, contains data sourced from NPInter v4.0 dataset, including information on: NPInter interaction ID (Column A), name of the lncRNA (Interactor Name: Column B), type of the interactor biomolecule (Interactor Type: Column C), reference to the lncRNA (Interactor ID: Column D), name of the target interacting with the lncRNA (Target Name, Column E), type of the target biomolecule (Target Type, Column F), and reference to the target (Target ID: Column G), mechanism of the lncRNA-protein/RNA interaction (Interaction Mechanism: Column H), Interaction Level (Column J), Interaction Class (Column J), and Interaction Description (Column K), experimental techniques used to identify the interaction (Experimental method for interaction identification: Column L), tissue/cell used for the investigation (Tissue/Cell: Column M), reference to research articles linked to the interaction (PubMed ID: Column N), and source of information on the interaction (Data Source: Column O). Datasets 18: LncRNA-Protein Interactions dataset_LncTarD, and dataset 20: LncRNA-RNA Interactions dataset_ LncTarD, contains data sourced from LncTarD 2.0 dataset, providing details on: interaction ID (Regulation ID: Column A), name of the lncRNA (Regulator Name: Column B), type of the interactor biomolecule (Regulator Type: Column C), reference to the lncRNA (Regulator Ensemble ID: Column D), and aliases of the lncRNA (Regulator Aliases: Column E), name of the target interacting with the lncRNA (Target Name: Column F), type of the target biomolecule (Target Type: Column G), reference to the target (Target Ensemble ID: Column H), and Target Aliases (Column I), mechanism of the lncRNA-protein/RNA interaction (Regulatory Mechanism: Column J), level of the interaction (Level of Regulation: Column K), type of the interaction (Regulatory Type: Column L), direction of the interaction (Regulation Direction: Column M), name of the cancer associated with the interaction (Cancer Name: Column N), function influenced by the interaction (Influenced Function: Column O), evidence for the interaction (Evidence: Column P), cancer characteristics of the interaction (Cancer hallmark: Column Q), expression pattern of the lncRNA (Regulator Expression Pattern: Column R), Experimental method for lncRNA expression (Column S), Experimental method for lncRNA target identification (Column T), occurrence of interaction in cancer stem cell (Cancer Stem Cell: Column U), dysregulation of lncRNA in circulating tumor cells (Regulator dysregulation in circulating tumor cells: Column V), Target dysregulation in circulating tumor cells (Column W), clinical application of the interaction (Clinical application: Column X), name of drugs inhibiting the interaction (Drugs: Column Y), and reference to research articles linked to the interaction (PubMed ID: Column Z).

Datasets 21–23 contain information on the experimentally-validated RNA G4-binding proteins (RGBPs) interacting with the catalogued lncRNAs (QUADRatlas and G4IPDB datasets, and scientific literature mining), accessible as Excel sheets (.xlxs)[42,44,73]. Dataset 21: RG4BP dataset_QUADRatlas, contains data sourced from QUADRatlas dataset, including details on: Gene name (RGBP) (Column A), type of the RGBP (Biotype: Column B), reference to the RGBP (Ensemble ID: Column C), alias of the RGBP gene (Gene Alias: Column D), chromosome number of the RGBP gene (Chromosome: Column E): start position of the RGBP gene on the chromosome (Start: Column G), end position of the RGBP gene on the chromosome (End: Column H), and location of the RGBP gene on the DNA strand (Strand: Column H), known status of RGBP as RNA binding protein (RBP) (Known RBP: Column I), RBP Type (Column J), link to further information on RGBP (Link to UniProt: Column K, Protein Domain: Column L, PTM: Column M, STRING: Column N, and BioGrid: Column O), name of the RGBP function (Function Name: Column P), type of the RGBP function (Function Type: Column Q), reference to research articles linked to the RGBP, name of the cancer associated with the RGBP (Cancer Name: Column R), and link to source of information on the RGBP (Link to Source Database (DISGENET/OMIM): Column T). Dataset 22: RG4BP dataset_G4IPDB, contains data sourced from the G4IPDB dataset, providing information on: Interaction ID (Column A), RNA G4 Interacting Protein (RGBP) Name (Column B), name of the RGBP in the UniProt entry (UniProt Entry Name: Column C), UniProt ID of the RGBP (UniProt ID: Column D), name of the target RNA of RGBP (Target RNA Name: Column E) and sequence of the target RNA of RGBP (Target RNA Sequence: Column F), and reference to research articles linked to the RGBP (PubMed ID: Column G). Dataset 23: RG4BP dataset_Literature mining, contains data sourced from scientific literature mining, including details on: UniProt ID of the RGBP (UniProt ID: Column A), RNA G4 Binding Protein (RGBP) Name (Column B), Gene Name (RGBP) (Column B), RNA G4 binding domains/motifs in the RGBP (RNA G4 Binding Domains/ Motifs: Column C), type of the target biomolecule (Target Type: Column E), and reference to research articles linked to the RGBP or RGBP-target interaction (PubMed ID: Column E).

Subcellular localization plots. The "Subcellular localization plots" folder includes plots: 1) cytoplasmic to nuclear localization: relative concentration index (RCI) and expression values: LncRNA Name_ratio, and 2) cytoplasmic to nuclear localization: RCI distribution: LncRNA Name_dist, for the catalogued lncRNAs across diverse human cell lines (LncATLAS dataset), presented as static images (.png)[51].

Technical Validation

The data sourced from the external datasets mentioned in the Methods section were meticulously examined to identify discrepancies and cross-validated against available scientific literature. The data curation involved the following steps to ensure the data quality:

Selection of external datasets. Cancer-dysregulated lncRNAs: The Lnc2cancer 3.0 dataset was selected to compile the list of cancer-dysregulated lncRNAs, as it is the most comprehensive repository of experimentally-validated human lncRNA-cancer associations, including cancer subtypes, at a tissue-level[48]. Other similar datasets, such as NONCODEV6, lncRNAdb v2.0, and lncRNome, focus more on the biological characteristics and cellular function of the lncRNAs and present limited information on their dysregulation in cancer[75–77].

Nucleotide sequences: The NCBI nucleotide dataset was used to retrieve the nucleotide sequences and corresponding NCBI accession numbers of the identified lncRNAs, including their functional transcript variants, as it compiles a collection of sequences from widespread sources, including RefSeq, GenBank, TPA, and PDB[72]. Ensembl dataset, while valuable, contains an exhaustive list of transcript variants, most of which are computationally-annotated, with few experimentally-annotated ones[78]. Since the annotation method in Ensembl is available within the summary of each entry, it becomes difficult to manually filter the functional transcript variants amongst the computationally-annotates from the search list. In contrast, the NCBI nucleotide contains a limited and precise list of transcript variants and distinctly displays "PREDICTED" in the search result for computationally-annotated ones, facilitating manual filtering. Additionally, it provides easy access to aliases and relevant scientific literature associated with the searched lncRNA.

LncRNA aliases: The GeneCards dataset was used to compile the available aliases for the identified lncRNAs, as it is the most extensive, better-targeted, and user-friendly repository available for information on human genes (including lncRNAs)[71]. The obtained lncRNA aliases were compared with those catalogued in the NCBI nucleotide to validate their correctness.

Subcellular localization: Being one of the most inclusive repositories of lncRNA subcellular localization in human cells, the LncATLAS dataset was chosen to gather information on the subcellular localization of the catalogued lncRNAs across diverse human cell lines[51]. It surpasses similar datasets like RNALocate v3.0 and lncSLdb in terms of comprehensiveness of localization entries and reliability by sourcing data from RNA-sequencing data sets (ENCODE) from different human cells rather than text mining[79,80]. The obtained plots: (1) relative concentration index (RCI) and expression values, and (2) RCI distribution, for the catalogued lncRNAs across various human cell lines, were manually verified for the accuracy of lncRNA names (including aliases) in the plot header.

LncRNA interacting partners: The NPInter v4.0 and LncTarD 2.0 datasets were used to obtain information on experimentally-validated interactions of RNAs and proteins with the catalogued lncRNAs, as they comprehensively document regulatory interactions between lncRNAs and biomolecules along with their interaction mechanism and level[49,50]. While NPInter v4.0 annotates lncRNAs with disease associations, LncTarD 2.0 links interactions to human diseases. Hence, the LncTarD 2.0 was used to gather interaction information in human cancers. Another similar dataset, LncRNA2Target v2.0, lacks the exhaustiveness of NPInter v4.0 and LncTarD 2.0[81].

RNA G4 interacting partners: QUADRatlas and G4IPDB datasets, along with manual scientific literature mining, were utilized to compile the information on the experimentally-validated RNA G4-binding proteins (RGBPs) interacting with the catalogued lncRNAs, as they are the only available resources in the domain[42,44,73].

After obtaining data from Lnc2cancer 3.0, NPInter v4.0, LncTarD 2.0, QUADRatlas, and G4IPDB datasets, the data was manually screened to identify missing entries. Any absent information was supplemented through a manual scrutiny of the scientific literature associated with the missing entry. Duplicate lncRNA entries (including aliases) were identified and merged accordingly. Since the identification of LncRNA-G4 interacting partners includes meticulous curation of information from well-established datasets, followed by reliability screening and assessment along with the correlation of datasets, no statistical filtering was carried out. This arises from the fact that the information on lncRNA-cancer associations, lncRNA nucleotide sequence, subcellular localization of lncRNA, and lncRNA-G4 interacting partners have already been experimentally-validated in the external datasets and was not acquired and interpreted in the current study. Therefore, the individual entry-level screening and assessment of external datasets as a validatory measure was prioritized over statistical filtering to minimize false positives.

Selection of nucleotide sequences. To ensure the accuracy of the retrieved nucleotide sequence for each lncRNA, the search results from the NCBI nucleotide dataset were filtered using Species: Animals and Molecule type: ncRNA. Individual search results were then carefully examined for molecule type: transcribed RNA, gene: searched lncRNA name, ncRNA class: lncRNA, and last update date to ensure the selection of the correct and recent version of the nucleotide sequence. The RefSeq status of individual search results was also verified, selecting only entries labelled "validated" or "reviewed", while those marked "model" were discarded to ensure the inclusion of sequences that have undergone validation or preliminary review.

Selection and modification of G4-prediction tools. Selection of G4-prediction tools: QGRS mapper and G4Hunter, two leading G4-prediction tools, were used for the identification of Putative Quadruplex-forming Sequences (PQS) within catalogued lncRNAs, as these tools are based on score-based ranking of the putative

sequences to enable prediction of the most probable G4-forming sequence[33,37]. While both G4-prediction tools predict the PQS within a sequence, they differ in their approach. The QGRS mapper is more effective at identifying canonical PQS, and the G4Hunter can also identify non-canonical PQS, providing a broader analysis. To ensure comprehensive coverage, both tools were used, and their predictions were compared to validate the G4-forming potential of identified sequences.

Modification of G4Hunter: A key challenge with the G4Hunter tool is its method of presenting PQS. It lists all overlapping PQS with individual scores per the set parameters and generates a consensus sequence containing all overlaps with a different score. This can overwhelm users with excessive data, making identification of the most promising candidates challenging. To address this, the G4Hunter algorithm was slightly modified to: 1) retain only the highest-scoring PQS among overlapping ones, and 2) prevent the generation of consensus sequences that combine multiple overlapping PQS with different scores.

Parameter Optimization for Accurate Predictions: Multiple parameter combinations were tested to identify PQS within the catalogued lncRNAs using: 1) QGRS mapper: maximum PQS length, minimum G-group, and loop size, and 2) G4Hunter: window size and threshold, to ensure relevant and accurate predictions of the G4-forming potential. This approach also ensured the identification of PQS with different anticipated G4 types (2 G, 3 G, and 4 G), while minimizing false positives.

Usage Notes

Web Application of the CanLncG4 Dataset. The datasets generated and analysed during the current study are also compiled as a freely accessible web application named CanLncG4 (https://www.canlncg4.com). These datasets can be downloaded from the downloads section of the web application (https://www.canlncg4.com/downloads). G4-prediction tools, QGRS mapper and G4Hunter, are integrated as standalone tools into the web application of the dataset to facilitate G4-prediction for any uncatalogued or novel nucleotide sequence or NCBI accession number. The G4Hunter standalone tool is made compatible with directly using the nucleotide sequence or NCBI accession number as input, like the QGRS mapper. The web application of the dataset is fully accessible without the need for registration or login. Bug fixes and minor upgrades are carried out periodically.

Experimental Application of the CanLncG4 Dataset in G4-prediction. The CanLncG4 dataset enables efficient selection of promising cancer-dysregulated lncRNAs with high G4-forming potential, as predicted by G4-prediction tools (QGRS Mapper and G4Hunter). This targeted shortlisting helps streamline experimental efforts by guiding the selection of lncRNA candidates most likely to form stable G4s. The shortlisted lncRNAs can then be subjected to *in vitro* validation using established biophysical techniques—such as circular dichroism (CD) and ultraviolet (UV) spectroscopy, CD- and UV-melting analysis, and electrophoretic mobility shift assays (EMSA)—to assess G4-topology and their thermal stability. Complementary biochemical assays, including G4-ligand fluorescence-based and reverse transcriptase (RT) stop assays, can provide further functional insights into G4-formation and stability in a controlled environment. Moreover, the dataset supports rational experimental design by enabling researchers to link G4-predictions with cancer type, expression patterns, and potential interacting partners. This can facilitate the hypothesis-driven investigation of lncRNA G4s in cancer-specific regulatory mechanisms. Ultimately, the dataset streamlines experimental workflows by reducing time, cost, and ambiguity associated with screening a large number of candidates, offering an informed entry point for *in vitro* screening and downstream *in cellulo* studies.

Code availability

All the new and modified codes used in generating and analysing the CanLncG4 dataset and its web application are available on the GitHub repository (https://github.com/BDLab-G4/CanLncG4.git).

Received: 13 November 2024; Accepted: 9 May 2025;
Published online: 23 May 2025

References

1. Matsui, M. & Corey, D. R. Non-coding RNAs as drug targets. *Nat Rev Drug Discov* **16**, 167–179 (2017).
2. Hombach, S. & Kretz, M. Non-coding RNAs: Classification, Biology and Functioning. *Adv Exp Med Biol* **937**, 3–17 (2016).
3. Cech, T. R. & Steitz, J. A. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**, 77–94 (2014).
4. Ransohoff, J. D., Wei, Y. & Khavari, P. A. The functions and unique features of long intergenic non-coding RNA. *Nature Reviews Molecular Cell Biology* **19**, 143–157 (2017).
5. Tassinari, M., Richter, S. N. & Gandellini, P. Biological relevance and therapeutic potential of G-quadruplex structures in the human noncoding transcriptome. *Nucleic Acids Res* **49**, 3617–3633 (2021).
6. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**, 155–159 (2009).
7. Iyer, M. K. *et al*. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**, 199–208 (2015).
8. Acha-Sagredo, A. *et al*. Long non-coding RNA dysregulation is a frequent event in non-small cell lung carcinoma pathogenesis. *British Journal of Cancer* **122**, 1050–1058 (2020).
9. Zhao, H. *et al*. Comprehensive landscape of epigenetic-dysregulated lncRNAs reveals a profound role of enhancers in carcinogenesis in BC subtypes. *Mol Ther Nucleic Acids* **23**, 667–681 (2021).
10. DeSouza, P. A. *et al*. Long, Noncoding RNA Dysregulation in Glioblastoma. *Cancers (Basel)* **13** (2021).
11. Siddiqui, H., Al-Ghafari, A., Choudhry, H. & Al Doghaither, H. Roles of long non-coding RNAs in colorectal cancer tumorigenesis: A review. *Mol Clin Oncol* **11**, 167–172 (2019).
12. Mo, Y. *et al*. Unlocking the predictive potential of long non-coding RNAs: a machine learning approach for precise cancer patient prognosis. *Ann Med* **55**, 2279748 (2023).
13. Chen, S. & Shen, X. Long noncoding RNAs: functions and mechanisms in colon cancer. *Mol Cancer* **19** (2020).
14. Singh, D., Assaraf, Y. G. & Gacche, R. N. Long non-coding RNA mediated drug resistance in breast cancer. *Drug Resistance Updates* **63**, 100851 (2022).

15. Li, Q., Mo, W., Ding, Y. & Ding, X. Study of lncRNA TPA in Promoting Invasion and Metastasis of Breast Cancer Mediated by TGF-β Signaling Pathway. *Front Cell Dev Biol* **9** (2021).
16. Irfan, M. *et al*. Apoptosis evasion via long non-coding RNAs in colorectal cancer. *Cancer Cell Int* **22** (2022).
17. Rupaimoole, R. *et al*. Long Noncoding RNA Ceruloplasmin Promotes Cancer Growth by Altering Glycolysis. *Cell Rep* **13**, 2395–2402 (2015).
18. Zangouei, A. S. *et al*. Cell cycle related long non-coding RNAs as the critical regulators of breast cancer progression and metastasis. *Biol Res* **56** (2023).
19. Ahmad, M., Weiswald, L. B., Poulain, L., Denoyelle, C. & Meryet-Figuiere, M. Involvement of lncRNAs in cancer cells migration, invasion and metastasis: cytoskeleton and ECM crosstalk. *J Exp Clin Cancer Res* **42**, 173 (2023).
20. Matsumura, K. *et al*. The novel G-quadruplex-containing long non-coding RNA GSEC antagonizes DHX36 and modulates colon cancer cell migration. *Oncogene* **36**, 1191–1199 (2016).
21. Ghafouri-Fard, S. *et al*. Interaction between non-coding RNAs, mRNAs and G-quadruplexes. *Cancer Cell Int* **22**, 1–11 (2022).
22. Sahayasheela, V. J. & Sugiyama, H. RNA G-quadruplex in functional regulation of noncoding RNA: Challenges and emerging opportunities. *Cell Chem Biol* https://doi.org/10.1016/J.CHEMBIOL.2023.08.010 (2023).
23. Jayaraj, G. G., Pandey, S., Scaria, V. & Maiti, S. Potential G-quadruplexes in the human long non-coding transcriptome. *RNA Biol* **9**, 81–89 (2012).
24. Yari, H. *et al*. LncRNA REG1CP promotes tumorigenesis through an enhancer complex to recruit FANCJ helicase for REG3A transcription. *Nature Communications* **10**, 1–15 (2019). *2019 10:1*.
25. Wu, R. *et al*. The long noncoding RNA LUCAT1 promotes colorectal cancer cell proliferation by antagonizing Nucleolin to regulate MYC expression. *Cell Death Dis* **11** (2020).
26. Simko, E. A. J. *et al*. G-quadruplexes offer a conserved structural motif for NONO recruitment to NEAT1 architectural lncRNA. *Nucleic Acids Res* **48**, 7421–7438 (2020).
27. Qu, X. *et al*. G-quadruplex is critical to epigenetic activation of the lncRNA HOTAIR in cancer cells. *iScience* **26**, 108559 (2023).
28. Huppert, J. L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res* **33**, 2908–2916 (2005).
29. Todd, A. K., Johnston, M. & Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res* **33**, 2901–2907 (2005).
30. Kudlicki, A. S. G-Quadruplexes Involving Both Strands of Genomic DNA Are Highly Abundant and Colocalize with Functional Sites in the Human Genome. *PLoS One* **11**, e0146174 (2016).
31. Varizhuk, A. *et al*. The expanding repertoire of G4 DNA structures. *Biochimie* **135**, 54–62 (2017).
32. Scaria, V., Hariharan, M., Arora, A. & Maiti, S. Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. *Nucleic Acids Res* **34**, W683 (2006).
33. Kikin, O., D'Antonio, L. & Bagga, P. S. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res* **34**, W676–W682 (2006).
34. Hon, J., Martínek, T., Zendulka, J. & Lexa, M. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics* **33**, 3373–3379 (2017).
35. Eddy, J. & Maizels, N. Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res* **34**, 3887 (2006).
36. Beaudoin, J. D., Jodoin, R. & Perreault, J. P. New scoring system to identify RNA G-quadruplex folding. *Nucleic Acids Res* **42**, 1209 (2014).
37. Brázda, V. *et al*. G4Hunter web application: a web server for G-quadruplex prediction. *Bioinformatics* **35**, 3493–3495 (2019).
38. Garant, J. M., Perreault, J. P. & Scott, M. S. Motif independent identification of potential RNA G-quadruplexes by G4RNA screener. *Bioinformatics* **33**, 3532–3537 (2017).
39. Sahakyan, A. B. *et al*. Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci Rep* **7** (2017).
40. Garant, J. M., Luce, M. J., Scott, M. S. & Perreault, J. P. G4RNA: an RNA G-quadruplex database. *Database* **2015** (2015).
41. Kikin, O., Zappala, Z., D'Antonio, L. & Bagga, P. S. GRSDB2 and GRS_UTRdb: databases of quadruplex forming G-rich sequences in pre-mRNAs and mRNAs. *Nucleic Acids Res* **36** (2008).
42. Bourdon, S. *et al*. QUADRatlas: the RNA G-quadruplex and RG4-binding proteins database. *Nucleic Acids Res* **51**, D240–D247 (2023).
43. Wang, Y. H. *et al*. G4LDB 2.2: a database for discovering and studying G-quadruplex and i-Motif ligands. *Nucleic Acids Res* **50**, D150–D160 (2022).
44. Mishra, S. K., Tawani, A., Mishra, A. & Kumar, A. G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci Rep* **6** (2016).
45. Zok, T. *et al*. ONQUADRO: a database of experimentally determined quadruplex structures. *Nucleic Acids Res* **50**, D253–D258 (2022).
46. Lu, X. J., Bussemaker, H. J. & Olson, W. K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res* **43**, e142–e142 (2015).
47. Li, R. *et al*. A comprehensive checklist of computational resources and methodologies for noncoding RNAs in systems medicine. *Wiley Interdiscip Rev RNA* **15**, e1830 (2024).
48. Gao, Y. *et al*. Lnc2Cancer 3.0: an updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on RNA-seq and scRNA-seq data. *Nucleic Acids Res* **49**, D1251–D1258 (2021).
49. Teng, X. *et al*. NPInter v4.0: an integrated database of ncRNA interactions. *Nucleic Acids Res* **48**, D160–D165 (2020).
50. Zhao, H. *et al*. LncTarD 2.0: an updated comprehensive database for experimentally-supported functional lncRNA-target regulations in human diseases. *Nucleic Acids Res* **51**, D199–D207 (2023).
51. Mas-Ponte, D. *et al*. LncATLAS database for subcellular localization of long noncoding RNAs. *RNA* **23**, 1080–1087 (2017).
52. Ghosh, A. *et al*. Identification of G-quadruplex structures in MALAT1 lncRNA that interact with nucleolin and nucleophosmin. *Nucleic Acids Res* **51**, 9415–9431 (2023).
53. Mou, X., Liew, S. W. & Kwok, C. K. Identification and targeting of G-quadruplex structures in MALAT1 long non-coding RNA. *Nucleic Acids Res* **50**, 397–410 (2022).
54. Su, K. *et al*. The role of a ceRNA regulatory network based on lncRNA MALAT1 site in cancer progression. *Biomedicine & Pharmacotherapy* **137**, 111389 (2021).
55. Gu, J. *et al*. Molecular Interactions of the Long Noncoding RNA NEAT1 in Cancer. *Cancers* **14**, 4009 (2022). *2022, Vol. 14, Page 4009*.
56. Kallweit, L. *et al*. Chronic RNA G-quadruplex accumulation in aging and Alzheimer's disease. *Elife* **14** (2025).
57. Neupane, A., Chariker, J. H. & Rouchka, E. C. Analysis of Nucleotide Variations in Human G-Quadruplex Forming Regions Associated with Disease States. *Genes (Basel)* **14**, 2125 (2023).
58. Maizels, N. G4-associated human diseases. *EMBO Rep* **16**, 910 (2015).
59. Simone, R., Fratta, P., Neidle, S., Parkinson, G. N. & Isaacs, A. M. G-quadruplexes: Emerging roles in neurodegenerative diseases and the non-coding transcriptome. *FEBS Lett* **589**, 1653–1668 (2015).
60. Umar, M. I., Ji, D., Chan, C. Y. & Kwok, C. K. G-Quadruplex-Based Fluorescent Turn-On Ligands and Aptamers: From Development to Applications. *Molecules* **24** (2019).
61. Chen, X. C. *et al*. Tracking the Dynamic Folding and Unfolding of RNA G-Quadruplexes in Live Cells. *Angew Chem Int Ed Engl* **57**, 4702–4706 (2018).

62. Laguerre, A. *et al*. Visualization of RNA-Quadruplexes in Live Cells. *J Am Chem Soc* **137**, 8521–8525 (2015).
63. Yang, S. Y. *et al*. Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. *Nat Commun* **9** (2018).
64. Mitteaux, J. *et al*. PhpC modulates G-quadruplex-RNA landscapes in human cells. *Chemical Communications* https://doi.org/10.1039/D3CC05155B (2024).
65. Mitteaux, J. *et al*. Identifying G-Quadruplex-DNA-Disrupting Small Molecules. *J Am Chem Soc* **143**, 12567–12577 (2021).
66. Khot, A. *et al*. First-in-Human RNA Polymerase I Transcription Inhibitor CX-5461 in Patients with Advanced Hematologic Cancers: Results of a Phase I Dose-Escalation Study. *Cancer Discov* **9**, 1036–1049 (2019).
67. Hilton, J. *et al*. Results of the phase I CCTG IND.231 trial of CX-5461 in patients with advanced solid tumors enriched for DNA-repair deficiencies. *Nature Communications* **13**, 1–12 (2022). *2022 13:1*.
68. Ahmed, A. A. *et al*. Structure–activity relationships for the G-quadruplex-targeting experimental drug QN-302 and two analogues probed with comparative transcriptome profiling and molecular modeling. *Scientific Reports* **14**, 1–13 (2024). *2024 14:1*.
69. Ahmed, A. A. *et al*. The Potent G-Quadruplex-Binding Compound QN-302 Downregulates S100P Gene Expression in Cells and in an *In Vivo* Model of Pancreatic Cancer. *Molecules* **28**, 2452 (2023).
70. Ahmed, A. A. *et al*. Asymmetrically Substituted Quadruplex-Binding Naphthalene Diimide Showing Potent Activity in Pancreatic Cancer Models. *ACS Med Chem Lett* **11**, 1634–1644 (2020).
71. Stelzer, G. *et al*. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* **54**, 1.30.1–1.30.33 (2016).
72. Sayers, E. W. *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**, D20 (2022).
73. Becker, G., Ivanov, P. & Wank, H. Functional characterization of RNA G-quadruplex-binding proteome. (University of Applied Sciences FH Campus Wien, Wien, 2020).
74. Sharma, S. *et al*. CanLncG4 dataset. *Figshare* https://doi.org/10.6084/m9.figshare.c.7510452 (2025).
75. Zhao, L. *et al*. NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Res* **49**, D165–D171 (2021).
76. Quek, X. C. *et al*. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res* **43**, D168–D173 (2015).
77. Bhartiya, D. *et al*. lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. *Database* **2013** (2013).
78. Dyer, S. C. *et al*. Ensembl 2025. *Nucleic Acids Res* **53**, D948–D957 (2025).
79. Wu, L. *et al*. RNALocate v3.0: Advancing the Repository of RNA Subcellular Localization with Dynamic Analysis and Prediction. *Nucleic Acids Res* **53**, D284–D292 (2025).
80. Wen, X. *et al*. lncSLdb: a resource for long non-coding RNA subcellular localization. *Database* **2018** (2018).
81. Cheng, L. *et al*. LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res* **47**, D140–D144 (2019).

Acknowledgements

The authors thank the Department of Biological Sciences and Engineering, Department of Chemistry, Common Research & Technology Development Hub, and the Central Instrumentation facility at Indian Institute of Technology Gandhinagar, Gandhinagar, Gujarat, India, for providing instrumentation facility. The authors thank Angshuman Mandal for technical assistance with data collection concerning the LncRNA-G4 Interacting Partners. All the illustrations are Created with BioRender.com. The authors also thank the Gujarat State Biotechnology Mission [GSBTM/JD(R&D)/626/22-23/00006262 to B.D.] for supporting this work.

Author contributions

S.S.: Conceptualization, Investigation, Methodology, Data curation, Validation, Formal Analysis, Visualization, Writing – original draft, Writing – review & editing, Funding acquisition, Project administration, Resources. M.Y.H.: Investigation, Data curation, Software. N.H.B.: Investigation, Data curation, Software, Writing – original draft. R.H.M.: Investigation, Data curation, Software. C.S.: Investigation. D.S.: Investigation. B.D.: Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing, Funding acquisition, Project administration, Resources, Supervision.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.D.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025