**Database of Putative Quadruplex Sequences in Dysregulated lncRNAs for select cancers and *in silico* Meta-Analysis of %GC-Frequency of the PQS**

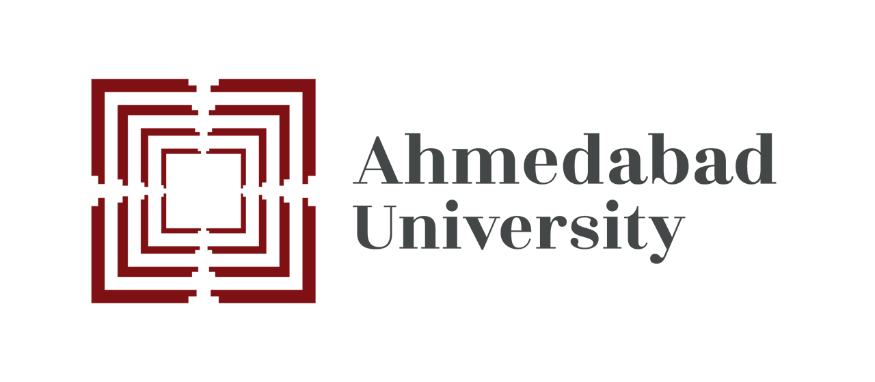
***Submitted by***

**Priya Darshan Gandhi**

**AU1621013**

**Session: 2016-2021**

***for the partial fulfillment of iM.Sc. in Life Sciences***



**Biological and Life Sciences**

**School of Arts and Sciences**

**Ahmedabad University**

**Ahmedabad-380009**

**Gujarat**

**JUNE 2021**

**WORKPLACE CERTIFICATE**

**CERTIFICATE**

This is to certify that the dissertation titled **“Database of Putative Quadruplex Sequences in Dysregulated lncRNAs for select cancers followed by Meta-Analysis of %GC-Frequency of PQS in silico”** submitted by **Ms. Priya Darshan Gandhi, AU1621013 ,** to the Biological and Life Sciences, School of Arts and Sciences, Ahmedabad University, Ahmedabad – 380009, for the degree of **Integrated** **Master of Science** in **Life Sciences** is her original work, based on the results of the experiments and investigations carried out independently by her during the period **January - 2021** to **May - 2021** of study under the supervision of **Dr. Bhaskar Datta** at **Indian Institute of Technology, Gandhinagar.**

This is also to certify that the above said work has not been previously submitted for the award of any degree, diploma or fellowship in any Indian or Foreign University.

**Programme Chair Dean of the school of Arts and Sciences**

**CERTIFICATE OF ORIGINALITY AND CONTRIBUTION**

I hereby declare that the project titled “XYZ” is my own original work and that, to the best of my knowledge and belief, it contains no material previously completed or written by another person. This work has not been submitted as a project report at this or any other institution before, except where due acknowledgment is made in the report. Any contribution made to the report by colleagues, with whom we have worked at SAS or elsewhere, during the project, is fully acknowledged.

I also declare that the intellectual content of this report is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Name: Priya Gandhi

Signature:

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **Acknowledgements** |  |
| **Abstract** |  |
| **Abbreviations** |  |
| **Aim and Objective** |  |
| **Introduction** |  |
| **Review of Literature** |  |
| **Materials and Method** |  |
| **Result** |  |
| **Discussion** |  |
| **Conclusion** |  |
| **Future Prospectives** |  |
| **References** |  |

**LIST OF FIGURES and TABLES**

|  |  |
| --- | --- |
| **Figure 1** | Upregulated and downregulated long non-coding RNAs in respective cancers. |
| **Figure 2** | G-Quadruplex Structure. |
| **Figure 3** | Schematic representation of some G4 topologies. |
| **Figure 4** | Interface of Lnc2Cancer 3.0. |
| **Figure 5** | Layout of .exe file for QGRS Mapper |
| **Figure 6** | Nucleotide sequence in FASTA format for lncRNA LINC02476 |
| **Figure 7** | Interface of QGRS Mapper |
| **Figure 8** | Search results for input sequence with dynamic visualization outputs |
| **Figure 9** | Data view of QGRS sequences found for LINC02476 |
| **Figure 10** | Organization of QGRS Mapper information about lncRNAs in database |
| **Figure 11** | Layout of .exe file for G4 Hunter |
| **Figure 12** | ‘My Sequences’ section in DNA analyser |
| **Figure 13** | Interface for analysis setting in G4Hunter |
| **Figure 14** | Heatmap analysis of lncRNA sequence ABHD11-AS1 (Count) |
| **Figure 15** | Heatmap analysis of lncRNA sequence ABHD11-AS1 (Coverage) |
| **Figure 16** | Interface for analysis output for a typical lncRNA sequence |
| **Figure 17** | Organization of G4 Hunter information about lncRNAs in curated database |
| **Figure 18** | Percentage of 2G, 3G and 4G PQS containing LncRNA in prostate cancer from the total validated and reviewed LncRNAs. |
| **Figure 19** | Percentage of 2G, 3G and 4G PQS containing LncRNA in liver cancer from the total validated and reviewed LncRNAs. |
| **Figure 20** | A graphical representation of analysed data for prostate and liver cancer in G4 Hunter |
| **Figure 21** | Relationship between observed frequency of PQS per 1000 bp and GC content in lncRNAs in prostate cancer. |
| **Figure 22** | Relationship between observed frequency of PQS per 1000 bp and GC content in lncRNAs in liver cancer. |
| **Table 1** | Sequences containing G-groups (2G, 3G, 4G) with respective parameters |
| **Table 2** | Count statistics for prostate cancer and liver cancer |
| **Table 3** | Distribution of unique 2G, 3G and 4G PQS in lncRNAs across all transcript variants in prostate cancer |
| **Table 4** | Distribution of unique 2G, 3G and 4G PQS in lncRNAs across all transcript variants in liver cancer |
| **Table 5** | List of unique lncRNAs with at least 3G or 4G tracts in PQS in prostate cancer. |
| **Table 6** | List of unique lncRNAs with at least 3G or 4G tracts in PQS in liver cancer. |
| **Table 7** | PQS Distribution at 0.9 and 1.4 threshold for prostate and liver cancer |
| **Table 8** | Prostate cancer LncRNA sequences according to their sizes (bp), % GC Content, PQS (total number of predicted PQS), and frequency of predicted PQS (per 1000 bp). |
| **Table 9** | Liver cancer LncRNA sequences according to their sizes (bp), % GC Content, PQS (total number of predicted PQS), and frequency of predicted PQS (per 1000 bp). |

**ACKNOWLEDGEMENTS**

Five years ago, the thought of finishing my Master’s thesis and getting my Master of Science degree felt like a distant dream – now it is coming true. I would like to extend my deepest gratitude to the Biological and Life Sciences, Ahmedabad University and the Indian Institute of Technology, Gandhinagar for providing me with this opportunity to successfully pursue my thesis. I would like to thank my project guide Dr. Bhaskar Datta for his inputs, constant support and encouragement throughout the course of my work. His guidance was truly valuable and made this project a great learning experience. I would like to thank his Ph.D. students Chinmayee Shukla, Shubham Sharma and Nakshi Desai for their invaluable help and mentorship. I would like to thank my fellow colleagues, Janki Thakkar and Drashti Shah for their inputs. Lastly, I would like to thank my friends and family without whom this thesis would not have been completed during the course of the pandemic.

**ABSTRACT**

The ever-growing interest in G4 Quadruplexes increases with the discovery of new G4 structures and knowledge about their biomedical potential. This thesis studied different available bioinformatics resources for RNA G-quadruplexes, including QGRS Mapper and G4 Hunter, and examined their usefulness in G4 lncRNA analysis. The main purpose of this study was to design, analyse and evaluate a new database about putative quadruplex sequences expressed in prostate and liver cancer for the needs of the scientific community. It was needed to find out which database design principles and techniques should be followed to achieve a well-performing and easily maintainable database. We share the results obtained from processing lncRNA datasets with these tools in the database. We illustrate our findings using known and novel cancer-related lncRNAs. Finally, we highlight key advantages and limitations of the in-silico tools investigated, and provide a viewpoint for the future exploitation of lncRNA-oriented databases. In the future it would be beneficial to study how an automated system for retrieving lncRNA data from software to database can be created to meet the needs of the scientific community.

**Keywords**

G-Quadruplex, Long non-coding RNA, PQS, Cancer, QGRS Mapper, G4 Hunter

**Abbreviations**

|  |  |
| --- | --- |
| DNA | Deoxyribonucleic acid |
| RNA | Ribonucleic acid |
| PQS | Putative quadruplex forming sequences |
| lncRNA | Long non-coding ribonucleic acid |
| ncRNA | Non-coding ribonucleic acid |
| mRNA | Messenger ribonucleic acid |
| rRNA | Ribosomal ribonucleic acid |
| tRNA | Transfer ribonucleic acid |
| G4 | G-quadruplex |
| ATP | Adenosine triphosphate |
| FASTA | Fast-all |
| GC | Guanine-cytosine |

**AIMS AND OBJECTIVES**

The major objective of this thesis is to create a database to store, organize and share data about the existence of putative quadruplex sequences in long non-coding RNAs in prostate and liver cancer. It is a part of a larger collection of G-quadruplex forming long non-coding RNAs that are dysregulated in multiple cancers. This study also explores the relationship between %GC content in long non-coding RNAs with the frequency of quadruplex forming sequences seen *in-silico*.

**INTRODUCTION**

G-quadruplexes are secondary structures formed in nucleic acids, especially guanine rich strands of DNA and RNA. Four guanine molecules form Hoogsteen bonds with each other and give rise to a G-tetrad. The role of G4 quadruplexes has been extensively studied in regulatory processes like transcription, translation, post-translation modifications and its effects on cancer. The vast amounts of data generated about the structures lead scientists to store and analyse quadruplexes using recently developed bioinformatic tools. However, the finding that most of the genome in complex organisms is not transcribed, lead to a new class of regulatory non-coding RNAs (ncRNAs), with increasing number of reports about long transcripts that act functionally as RNAs.

Long ncRNAs are considered to be longer than ~200 nucleotides, on the basis of current literature. They have physiological and biochemical significance inside the human body. The database created for this thesis utilises bioinformatic tools QGRS Mapper and G4Hunter to screen lncRNAs known to be overexpressed or under-expressed in select cancers. The web-based tools help in analysing the nucleotide sequences for putative quadruplex forming regions, each with their own unique algorithm. Based on screening via the unique algorithms, the database provides a comprehensive collection of analysed information in a single database. Meant to be a part of a larger database spanning multiple cancers, this thesis contributes to the ease of retrieval of scientific data about quadruplex forming sequences in lncRNAs for future generation of scientists worldwide. Scientists are able to access the no of quadruplex forming sequences in each of hundreds of lncRNAs, retrieve information according to individual transcript variants of lncRNA, obtain the algorithmic scores allotted to individual G4 sequence for each lncRNA and retrieve annotated information from one database.

The list of new lncRNAs and discovery of functions of it grows daily; however, this is likely only the “tip of the iceberg”. This database about quadruplex forming sequences within functionally validated lncRNAs will facilitate a wide array of investigations. This analysis also provided clues about the potential ability of certain segment of sequence to form secondary structures within lncRNAs, and can prompt a study into the otherwise overlooked sequence segments.

**Review of Literature**

**Non-Coding RNA**

Ribonucleic Acid (RNA) is one of the most well-studied polymeric molecules. It is classified under the category of nucleic acids along with deoxyribonucleic acid (DNA), well-known as an essential macromolecule for the sustenance of life. In vivo studies have shown the ability of RNA molecules to form complex secondary and tertiary structures, that have unique functional properties associated with their structures[1]. RNA molecules can be classified into messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA) and other non-coding RNAs. However, the limited availability of structural information about non coding RNAs been one of the major shortcomings in the advancement of RNA research. Initial research in this field devalued the significance of the non-coding parts of the genome, referring to them as junk DNA, which might have contributed to the lack of advancement comparatively [1]. An example for understanding the magnitude of the existence of non-coding RNAs can be between bacteria and humans. While a typical bacterium expresses protein-coding genes in approximately 90% of its genome, the human genome has approximately 95% that does not code for proteins [2]. Non coding RNA consists of both short and long non-coding RNA genes, which out-numbers the protein-coding genes [3]. ncRNAs do not code for a protein, but they exist as sequences that can be further spliced into microRNAs, snoRNAs and long non-coding RNAs [2].

**Long non-coding RNA**

Long ncRNAs are RNA sequences that are longer than ~200 nucleotides [2]. With the advancement of research, it was shown that a large portion of the human transcriptome has little or no protein-coding capacity. The non-coding transcriptome is large and contributes to human beings’ complex physiology and structure. LncRNAs are notorious for being poorly conserved and undergoing differential splicing[3]. The distribution of LncRNAs is often found to be overlapping/interspersed between multiple coding and non-coding transcripts [4]. Trying to classify the lncRNAs based on their genomic proximity to protein-coding genes, including overlapping, cis-antisense, bidirectional or intronic ncRNAs. In reality, many transcripts resist classification into any particular category, and instead exhibit a combination of these qualities [4]. In 1992, an important ncRNA that plays a role in the regulation of chromosome structure, such as Xist, was discovered [5]. Subsequently, databases of curated lncRNA have shown hundreds of sequences to have a significant role in physiological and structural functions.

**Long non-coding RNAs in Transcript Variants**

Alternative splicing is seen as a post-transcriptional modification of most mammalian organisms [6]. The process of alternative splicing ensures that noncoding transcriptional diversity is generated and genetic evolution is consistent [7]. However, most human genes express only a single major transcript variant despite the presence of multiple splice variants [8]. Non-coding variants are formed by the skipping of the first or last exon, which might result in the loss of start or stop codons, respectively [8]. GenBank is a repository of sequenced genes with information such as expression levels in long non-coding transcript variants displayed.

**Function of Long Non-Coding RNA**

LncRNAs are touted by scientists for their functional versatility, which means that they have the ability to conform to different structures and molecular interactions with proteins, RNA and DNA to cause histone modification, transcription, and/or posttranscriptional regulation[9]. Many examples have been found experimentally to corroborate the data. MANTIS, a lncRNA found in endothelial cells interacts with ATPase catalytic subunits and confers specificity to the switch/sucrose non-ferentable (SWI/SNF) chromatin remodeling complex. This structure further initiates transcription and nucleosome remodeling [10, 11]. The MEG3 lncRNA gene contains different motifs that are important for p53 activation [12]. CHAER lncRNA is found in the heart, and acts as an epigenetic switch by interfering with the repressive complex 2 (PRC2) and inhibiting H3K27m3 at genes involved in cardiac hypertrophy [12]. Thus, the interaction of lncRNA with protein, RNA, DNA by forming duplexes and triplexes holds significance in understanding the function of these molecules in biological pathways.

**Role of Long Non-Coding RNA in Cancer**

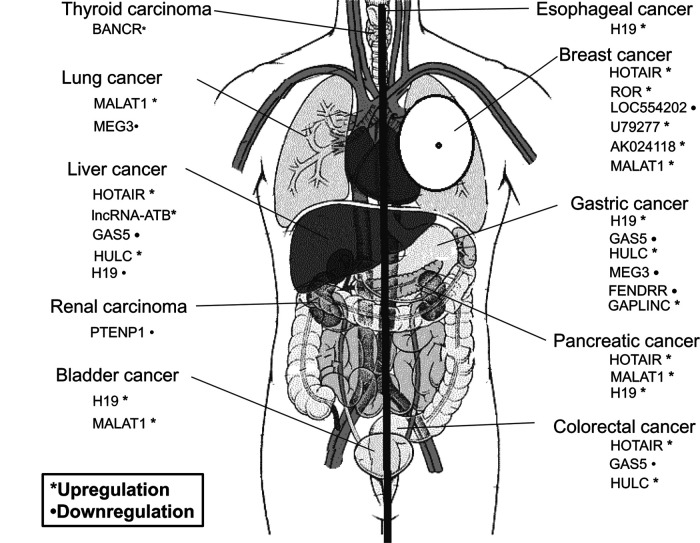
Apart from their role as activators, decoys, guides, or scaffolds for their interacting proteins, DNA and RNA [13], lncRNAs are increasingly being found to play a role in stem cell maintenance, cell apoptosis, cell invasion, cell proliferation, and metastasis [14, 15].

Many lncRNAs have been significantly overexpressed in select cancers, such as prostate, cervical, lung, liver etc. Looking at prostate cancer in depth, early cancer literature has shown the expression of highly specific Prostate cancer-associated transcript 3 (PCA3) in prostate cancer cells[16]. The overexpression of lncRNA PCAT1 caused the inhibition of tumour suppressor BRCA2, which promotes prostate cancer growth [17]. Other lncRNAs found to be overexpressed in prostate cancer cells are PCGEM1 and PRNCR1 [18], H19 [19], PCAT6[20], PCAT7 and PVT1 [21], CTBP1-AS [22], SchLAP1[15] and others.

A 2013 paper by Takayama K et al, showed that CTBP1-AS might be associated with prostate cancer by sense-antisense transcriptional regulation. With the help of laser capture microdissection (LCM) and quantitative reverse transcriptase–PCR (qRT–PCR) analysis, they compared the expression levels of both CTBP1 and CTBP1-AS in benign and cancerous regions. An upregulation of CTBP1-AS and downregulation of CTBP1 in prostate cancer was found [22].

In hepatocellular carcinoma, the expression pattern of lncRNA HOTAIR using RT-PCR showed a high expression as compared to non-cancerous surrounding cells. This can be used as an early prognosis marker [23]. HOTAIR is also involved as an early predictor and prognostic marker in gastric cancer cells, cervical cancer and breast cancer [23].

The following figure depicts commonly upregulated and downregulated long non-coding RNAs in respective cancers in the human body.

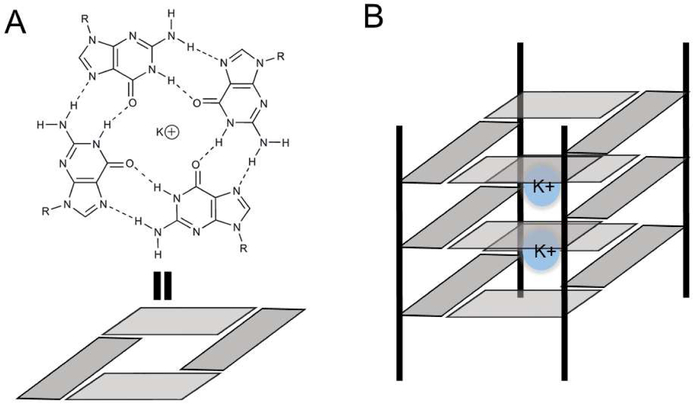


**Figure 1:** Upregulated and downregulated long non-coding RNAs in respective cancers. **Source:** Li, J., Meng, H., Bai, Y., & Wang, K. (2016). Regulation of lncRNA and Its Role in Cancer Metastasis. Oncology research, 23(5), 205–217. <https://doi.org/10.3727/096504016X14549667334007>

It is interesting to note that while many lncRNAs are unique to a cancer type, there are numerous lncRNAs that are expressed in multiple cancer types, indicating its role in dysregulation of essential cell-regulation pathways.

**G-Quadruplex Formation**

Non-canonical structures are those structures adopted by nucleic acids that do not conform to the Watson-Crick model. Guanine quadruplexes (G4) are a family of non-canonical nucleic acid structures that have building blocks of guanine quartet [24]. They have high structural stability for G4 formation under specific physiological conditions [24]. With the help of Hoogsteen hydrogen bonding, the stabilisation of G4s contain two or more stacks of four coplanar guanine residues occurs [25]. The coordination of monovalent cations in the centre is responsible for imparting the stability to the G4 stacks.



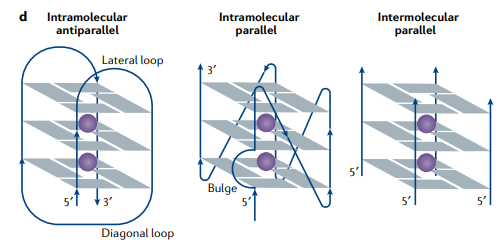
**Figure 2:** G-Quadruplex Structure. (A) Orientation of guanines in a G-quadruplex quartet. (B) Central occupancy by monovalent cations, provides stability via negating the partial negative charge of O6 oxygen of adjacent quartets. Vertical black lines depict phosphate backbone. **Source:** Monsen, R. C., & Trent, J. O. (2018). G-quadruplex virtual drug screening: A review. Biochimie, 152, 134–148. https://doi.org/10.1016/j.biochi.2018.06.024

As seen in Figure 2, adjacent location of guanine nucleotides is a requirement for G4 formation. These adjacent nucleotides form a run, and it is with the help of these runs that quadruplex forming sequences are identified [24]. The presence of monovalent cations, specifically K+ and Na+ impart stability to the structure [26], where K+ is considered to be more physiologically relevant due to its high intracellular concentration (~140 mM) whereas Na+ (~10 mM) [28]. The stability is imparted by cations by negating the partial negative charge of O6 oxygen atoms in the centre channel of the adjacent stacked G-tetrads [26].

**Long non-coding RNA G-quadruplexes**

G-quadruplexes that are formed in G-rich RNA sequences are more likely to occur as compared to DNA G-quadruplexes because of the single-stranded structure [27]. An important component of RNA function is the formation of secondary structures. Seeing the physiological role of lncRNAs in the cell, the possibility of lncRNAs harbouring G-quadruplexes imparting them specific functional characteristics has been explored by scientists [28]. A 2012 study, the first to analyse lncRNA for G-quadruplexes, analysed the number of ncRNAs with at least one G-quadruplex motif. They found that 200–300 nucleotide long ncRNA harboured a significantly higher percentage of G-quadruplexes [28].

The stacking of G-tetrads forms quadruplex structures in three distinct topologies, i.e., parallel (all four guanine chains are parallel to each other [29]); anti-parallel (two G-tracts are anti-parallel [30]); and hybrid (One G-tract is antiparallel [29]). The type of topology structure formed is further connected to G4-related biological functions [29].



**Figure 3:** Schematic representation of some G4 topologies. **Source**: Varshney, D., Spiegel, J., Zyner, K., Tannahill, D., & Balasubramanian, S. (2020). The regulation and functions of DNA and RNA G-quadruplexes. Nature Reviews Molecular Cell Biology. doi:10.1038/s41580-020-0236-x

Loops are another point of importance, they are sequences connecting the G-tracts in G4s. The propeller loop connects adjacent parallel G-main chains [29], lateral loops connects adjacent anti-parallel G-backbones [29], and diagonal loop connect the G-main chain diagonally to the G-quartet [29].

**2G, 3G and 4G G-tracts and its Significance**

A key feature in the formation of G-quadruplex is a requirement of adjacent runs of G-tracts [24]. The G-tracts must be at least four stretches of G residues, where each stretch is comprised of at least two Gs [30]. There is a known gap of 1-7 nucleotides between two adjacent G runs [31]. For intramolecular GQ formation, the quadruplex structure usually forms loops that impart stability to the overall structure. The stability of the structure depends upon length and nucleotide composition of the loops [32, 33]. G-tracts with four consecutive G’s are generally considered most stable.

A lot of kinesiological factors come into play during the formation of G-quadruplexes. Earlier, scientists were sceptical about existence of G2-quadruplexes (having two guanine nucleotides consecutively) because of their low stabilities and slow folding rates [34]. Using bioinformatic tools, researchers found that G2-quadruplexes were enriched under G-quadruplex stabilizing conditions. The activities of both the G2 and G3- quadruplexes were temperature-dependent – up to 80% reduction in G-quadruplex activity was observed at 30 °C compared to that at 37 °C [35]. Thus, such studies prove the presence of G2-quadruplexes in the cell functionally, but this activity may rely on the presence of stabilizing factors.

Out of all the non-canonical PQS’s, one of the most widespread in nature is the bulged G3-PQS (a G3-PQS with at least one nucleotide interruptions in at least one of the G-runs) [36].

**Using Bioinformatic Tools for G-Quadruplex Prediction**

The increased demand for identifying G-quadruplex forming motif regions was met by bioinformatics tools. The GnNmGnNoGnNpGn stereotypical pattern was considered the building block to identify favourable quadruplex formation sequences in first generation algorithms [37]. The combination of *in silico* analysis with *in vitro* studies are successful in characterising G-quadruplex structures [38]. In the context of RNA G-quadruplex sequencing, many tools such as rG4-seq [38], G4RP-seq [38] were developed for detecting RNA quadruplexes on a transcriptome-wide level. Approaches such as regular expression matching, use of score calculations, sliding window algorithms and machine learning has been used for algorithmic development in quadruplex detection [38]. This has been helpful in modelling secondary and tertiary structures, simulate molecular dynamics, calculate free energy or perform molecular docking [38].

​**NCBI – Nucleotide**

Quoting directly from the website of NCBI-Nucleotide database, “The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.”

**Lnc2Cancer 3.0**

The Lnc2Cancer database is a collection of data on experimentally supported lncRNAs and circular RNAs (circRNAs) associated with human cancers [39]. The latest version of the database contains 10 303 entries of associations between 2659 human lncRNAs, 743 circRNAs, and 216 cancer subtypes [39]. Lnc2Cancer 3.0 provides information for the link between lncRNAs and cancer, with increased data than the previous versions. More than 8000 studies were analysed in the PubMed database, out of which more than 6500 (lncRNAs) and 1570 (circRNAs) were found and included [39]. The database contains useful information about experimentally supported regulatory mechanisms, biological functions, clinical applications of lncRNAs in human cancer [39]

Detailed description about interface and content will be described in section Materials and Methods.

**QGRS Mapper**

To address the need for user-friendly bioinformatics tools for studying G-quadraplexes, a web-based server, QGRS Mapper, was developed that predicts quadruplex forming G-rich sequences (QGRS) in nucleotide sequences [40].

With the help of this server, scientists can input NCBI nucleotide sequences to extract information about its composition and distribution of putative quadruplex forming G-rich sequences (QGRS) [40]. Applications of QGRS Mapper:

* Interactive graphic modules to visualize QGRS distribution patterns [40]
* The analysis of mammalian pre-mRNA sequences (including alternatively processed), genomic sequences (promoter, telomeric and RNA regions), oligonucleotides [40].
* Functional relevance of G-quadruplex structure [40].

**Algorithm and Functions**

QGRS Mapper was built on the motif ‘GxNy1GxNy2GxNy3Gx’, where x = number of guanine tetrads in the G-quadruplex and y1, y2, y3 = length of gaps (i.e., the length of the loops connecting the guanine tetrads) [40].

There are a few restrictions placed on the sequence:

1. The sequence must contain at least two tetrads (i.e., x ≥ 2) (on the basis of literature supporting the existence of stable 2G quadruplex structures *in vivo*) [40]
2. Default length of input sequence is 30 nucleotide bases. This can be extended up to 45 bases. Thus, in accordance with the algorithm, there can be maximum of 6 G-groups [40]
3. Loop length must be within specified length range (1 and 4). A string can also be searched for inside a loop using regular expressions [40]
4. Maximum of one gap is allowed to be of zero length [40]

The following table shows different types of G-groups in QGRS Mapper:

|  |  |
| --- | --- |
| Sequence | Parameters |
| GGAACCATATTGCGGTTCCAGGGCATTAGG | x = 2, y1 = 11, y2 = 5, y3 = 6 |
| GGGAGTGAGGGAATCCGGGGATGCTTGGG | x = 3, y1 = 5, y2 = 5, y3 = 7 |
| GGGGATTAGGGGCTGAGGGGCCAATGGGG | x = 4, y1 = 5, y2 = 6, y3 = 2 |

**Table 1**: Sequences containing G-groups (2G, 3G, 4G) with respective parameters

Based on the output from the algorithm, the QGRS Mapper assigns a score to each sequence based on its likelihood to form a stable G-quadruplex [40]. The scores are given as follows:

* Shorter loops are more common than longer loops [40].
* G-quadruplexes tend to have loops roughly equal in size [40].
* The greater the number of guanine tetrads, the more stable the quadruplex [40].

QGRS Mapper has interactive features to decide the minimum number of tetrads, maximum length of the G-quadruplex motif, size and composition of the loops. Inputs in the form of raw or FASTA format are taken [40]. In case of overlapping sequences, the software will produce a non-overlapping set of QGRS, by selecting the higher scoring QGRS [40]. An intuitive feature is displaying both overlapping and non-overlapping sets side by side [40].

Detailed description about searching, analysis and output will be mentioned in the ‘Materials and Methods’ section. QGRS Mapper is considered revolutionary due to its dynamic and flexible user interface that can serve a large section of non-bioinformatic skilled scientists in the future.

**G4Hunter**

G4Hunter is a web-based software that was built to overcome many limitations of its contemporary software. Algorithms built on [GnNmGnNoGnNpGn] pattern of G-quadruplexes have shown limitations in experimental research [24]. A number of ‘false negative’ sequences have been found that form G-quadruplex structure *in vivo*, but evade the *in-silico* algorithm [41]. Similarly, ‘false positive’ sequences are found that do not form G-quadruplex structures *in vivo* although it obeys the algorithm [42]. The team of researchers behind G4Hunter used G-richness and G-skewness factors to develop a quantitative analysis that provides a score, indicative of quadruplex ‘strength’ [24]. The score i.e., quadruplex propensity, is based on the fraction of Gs in the sequence (richness) and G/C asymmetry (skewness) between the complementary strands [24].

**Algorithm and Functions**

The algorithm is unique in its scoring system, which is done on the basis of both G richness and G skewness. Key points for the scoring system are:

* Scoring is done between −4 and 4, assigned to each nucleotide. The score is 0 for A and T (i.e., neutral or indifferent), positive for G and negative for C [24].
* For E.g.: ‘G’ is given a score of 1; ‘GG’ in a sequence is given a score of 2 each (2+2), ‘GGG’ in a sequence is given a score of 3 (3+3+3), similarly for 4 or more Gs each G is given a score of 4. Similarly, ‘Cs’ are scored negatively [24].
* Simultaneous scoring of the complementary strand also occurs in G4Hunter [24].
* G4Hunter score (G4Hscore) is the mean of sequence of numbers assigned to nucleotides [24]

Output of the algorithm is a user-friendly GUI that provides dynamic information. A heatmap of quadruplex forming sequences is displayed and provides statistical information [43]. The software allows import from NCBI database, and also analyses multiple sequences at once. Parameters such as window size (default: 25), threshold (minimum score, default: 1.4) can be adjusted according to user preference [43]. The results can be exported as a text file for further processing [43].

**Next-Generation Clustered Heatmaps**

To easily visualise large amount of data, scientists in every field use heatmaps. Heatmaps are colour coded panels that are used as data visualisation tools, similar to bar graphs. The MD Anderson Cancer Center at the University of Texas have created a ‘Next Generation-Clustered Heat Viewer. It is a web browser-based software that allows the visualisation of clustered and non-clustered heatmap data [44]. Next-generation clustered heatmaps (NG-CHM) have link-outs to additional information sources, zooming and navigation features, and visualisation of the biology behind the image [44].

**Significance of %GC Content in Nucleotide Composition**

In nucleotide sequences, the GC base pair is found to impart higher thermal stability compared with the AT base pair. This is due to stronger stacking interaction between GC bases, and the existence of a triple hydrogen bond between the paired bases instead of a double [45]. Studies in bacteria and some vertebrates have the evolutionary significance of GC content. Because of its stability, it provides larger heat tolerance [45] GC content of has been used as a rough measure for activities such as protein-producing activity [46], amino acid compositions for energy efficiency [46].

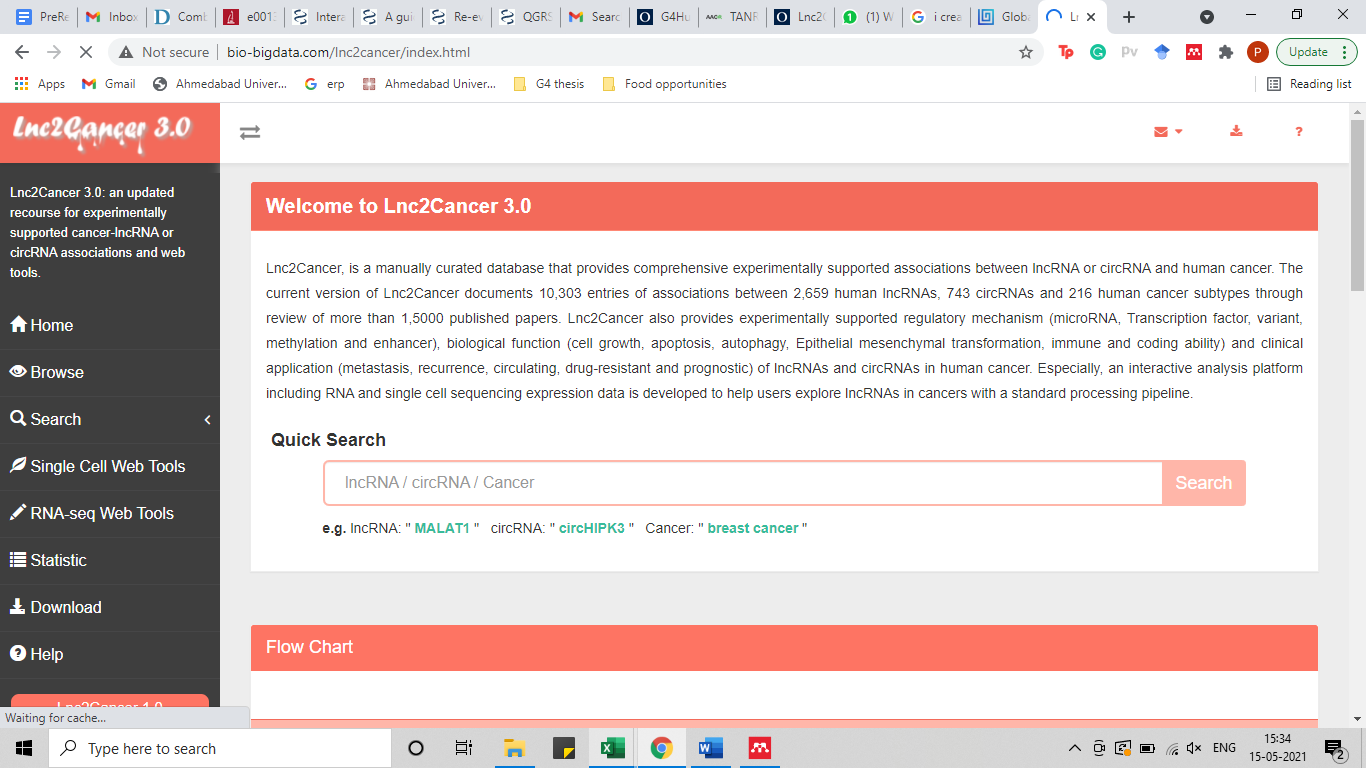
In the case of long non-coding RNAs, a study has shown that lncRNAs that are expressed in a low amount have significantly low GC content in them. They showed that that a higher GC content correlates with a higher RNA steady-state level, which can provide an explanation about the lower cellular level of lncRNAs compared with the protein-coding RNAs. A higher GC content level brings higher rate of transcription or processing [47].

**MATERIALS & METHODS**

For the purpose of the project, the cancers in focus were prostate cancer and liver cancer. The workflow used to generate the data is described below. The workflow utilizes five major steps:

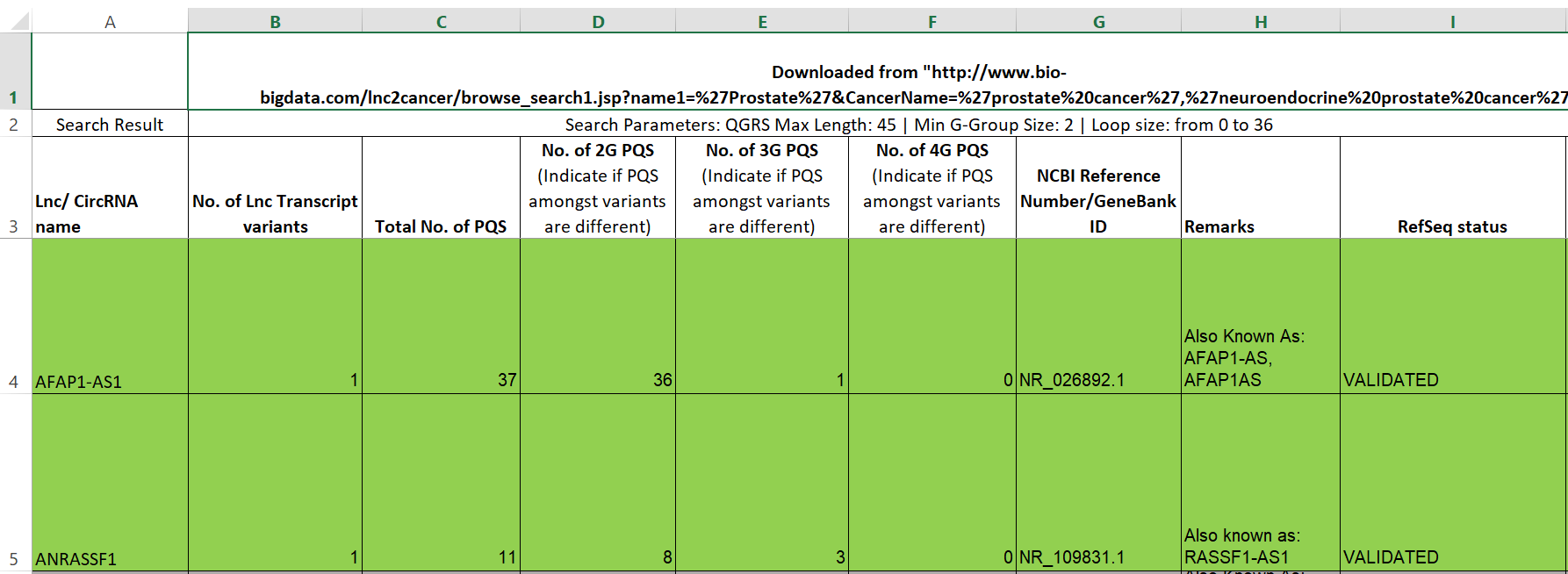
1. **Data Source**

Collection of data was done from the Lnc2Cancer 3.0 database (<http://www.bio-bigdata.net/lnc2cancer>). In Step 1, we access the most recent release of the Lnc2Cancer database, as seen in Figure 4. The interface of Lnc2Cancer 3.0 is as seen below.

**Figure 4:** Interface of Lnc2Cancer 3.0. The left panel shows options to browse by tissue, search, visualize and analyze lncRNAs at single cell and RNA-seq levels.

Subsequently, current list of lncRNAs and circRNAs for both (prostate and liver) cancers were extracted in the form of .exe files.

The .exe files for both cancers are maintained with the following entry categories:



**Figure 5:** Layout of .exe files for QGRS Mapper

* LncRNA name: the name of the lncRNA obtained from the lnc2cancer 3.0 database
* No. of Lnc. transcript variants: The process of alternative processing gives rise to transcript variants in some lncRNAs. Each of the variants were individually assessed.
* No. of 2G PQS: G-tracts with two consecutive guanine nucleotides in a putative quadruplex forming sequence.
* No. of 3G PQS: G-tracts with three consecutive guanine nucleotides in a putative quadruplex forming sequence.
* No. of 4G PQS: G-tracts with four consecutive guanine nucleotides in a putative quadruplex forming sequence.
* NCBI Reference number/GeneBank ID
* Remarks: Notes about alternate names of the lncRNA and other observations
* RefSeq Status: this information was also obtained from the NCBI Nucleotide website. The different input options were: Validated, Reviewed or Predicted.

1. **Sequence Retrieval**

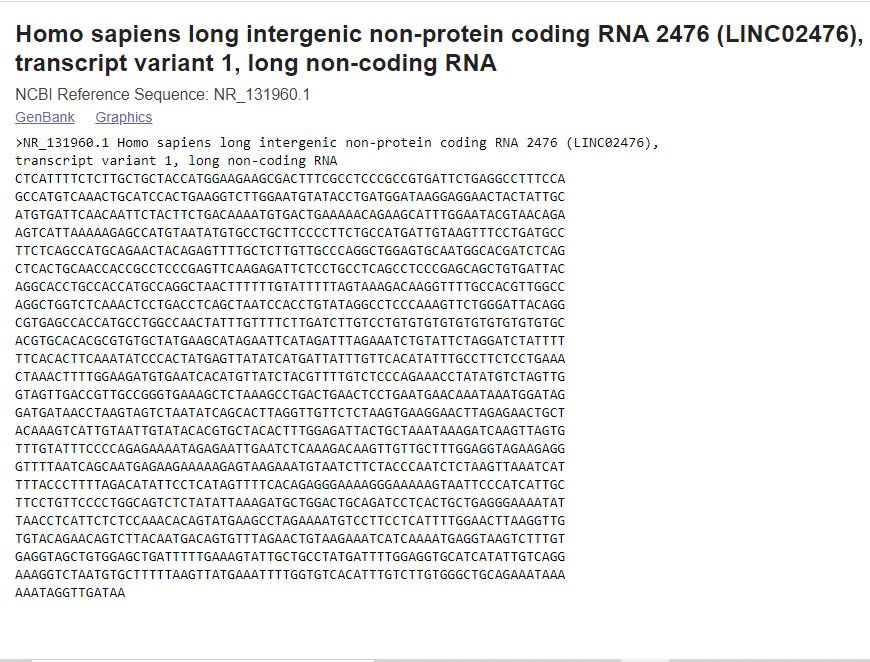
For the purpose of this project, the circRNAs in the database were eliminated from any further analysis. All other types such as miRNA, genomic DNA (all excluding lncRNAs) were noted for discontinuing of further analysis.

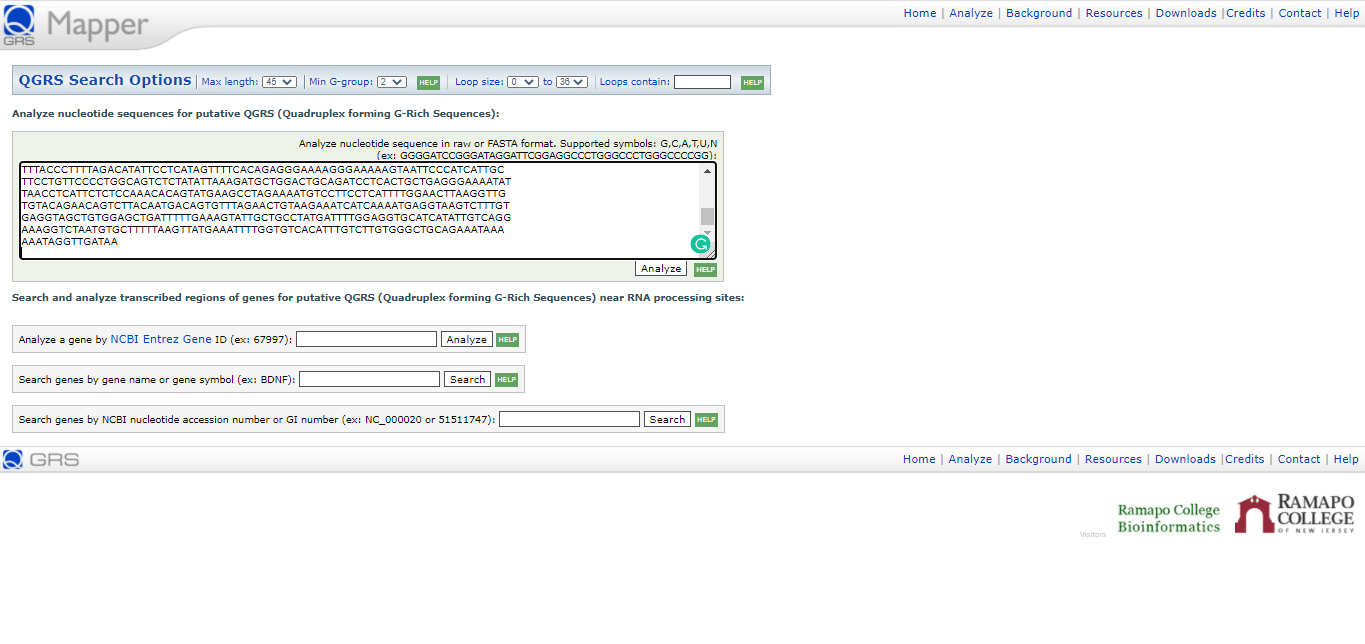
Using NCBI-Nucleotide database, each sequence from the .exe file was searched for information about NCBI Reference number/GeneBank ID, no. of transcript variants, alternate names of lncRNA and RefSeq Status.

1. ***In-silico* Analysis of G-quadruplexes forming sequences**

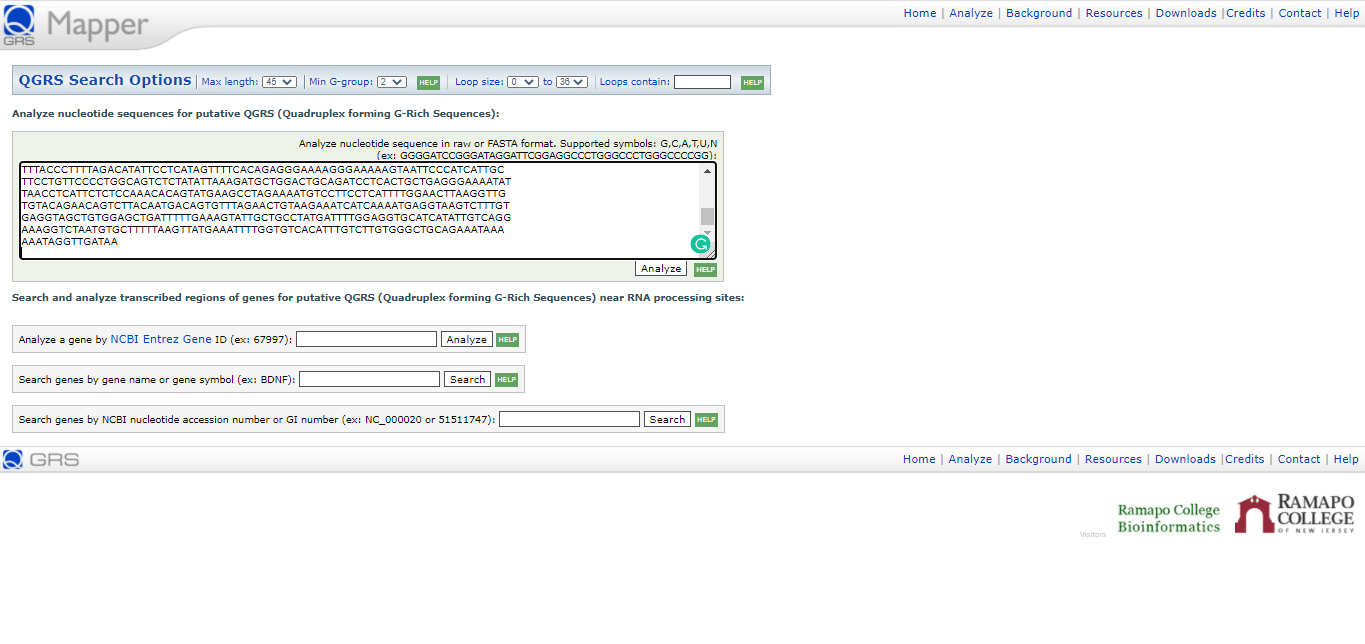
To survey the presence of PQS in lncRNAs, we carried out a computational search for PQS using *in silico* tool, QGRS Mapper. The search algorithm found all G-core motifs that matched the pattern GxNy1GxNy2GxNy3Gx; that is, two or more G-tracts of three or more guanines connected by a loop of 1–7 nt.

Firstly, the sequence was extracted in FASTA format from NCBI-nucleotide, as shown in Figure 6.



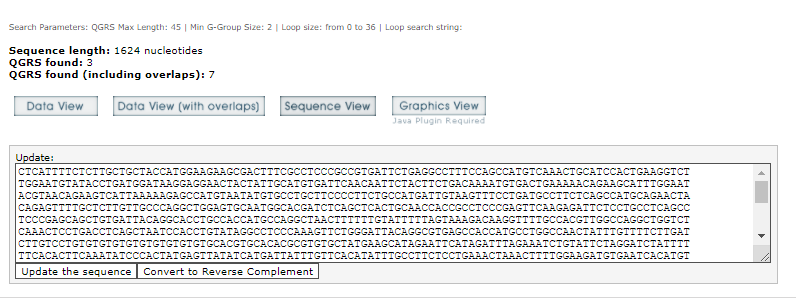
**Figure 6:** Nucleotide sequence in FASTA format for lncRNA LINC02476

Next, the sequence was input in QGRS Mapper, as seen in Figure 7. The following search parameters were used to analyse each sequence; Max length of sequence: 45, Min G-group: 2, Loop size: 0 and Loop contain: blank.



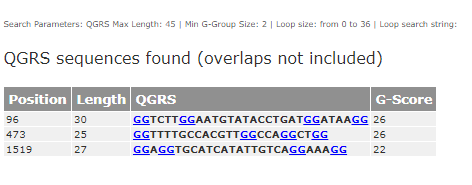
**Figure 7**: Interface of QGRS Mapper, an *in-silico* tool for the analysis of PQS. Top panel shows the interactive search options. An input box is seen for entering sequence in raw or FASTA format. User can also search using NCBI Entrez Gene ID, gene name/gene symbol or NCBI nucleotide accession number.

The search results yielded information about sequence length, QGRS found (without overlaps) and QGRS found (including overlaps), as seen in Figure 8. A non-overlapping set of QGRS was selected using the higher scoring QGRS by sliding window method.



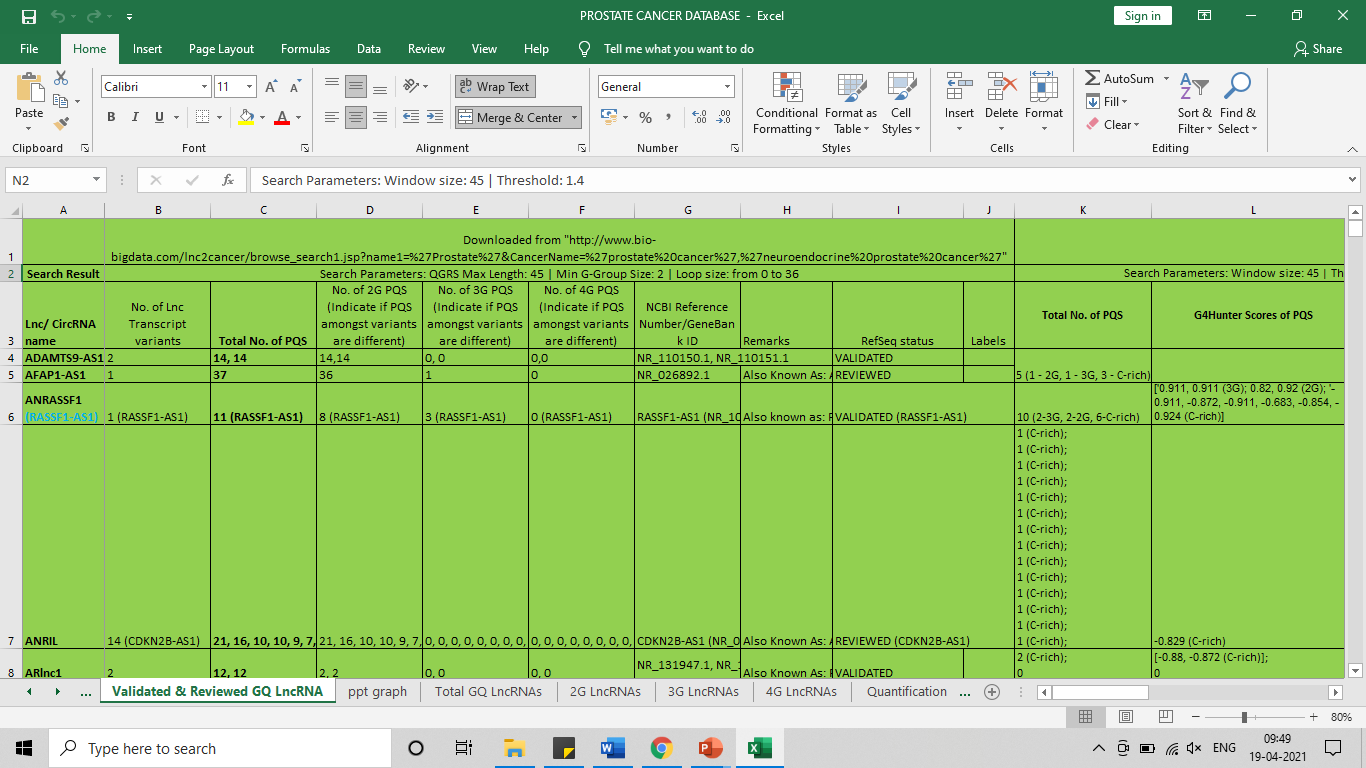
**Figure 8**: Search results for input sequence with dynamic visualisation outputs

User had the option to select from multiple data view options. For the purpose of this project, QGRS found in ‘Data View’ were taken. In Figure 9, it is seen that three QGRS sequences were found, each forming 2G PQS motifs.



**Figure 9:** Data view of QGRS sequences found for LINC02476. Data was displayed with position, length, highlighted QGRS sequence and its allotted G-score

The resulting PQS motifs were then grouped into three categories in the .exe file as seen in Figure 10, designated 2G, 3G and 4G, in which the PQS contains 2, 3 and ≥4 G-tracts, respectively. Each lncRNA was individually analysed for multiple transcript variants, and data was collected in similar way.



**Figure 10:** Organisation of QGRS Mapper information about lncRNAs in curated database

1. **Expression quantitation of annotated lncRNAs**

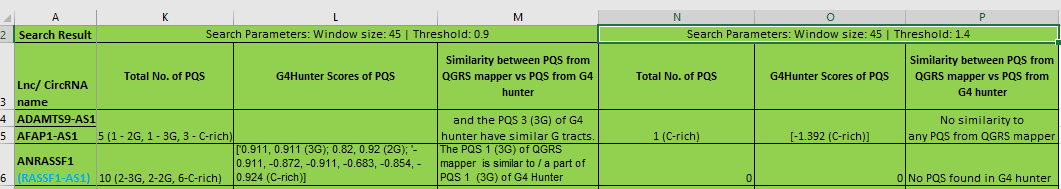
To perform a comprehensive survey of lncRNAs prevalent in prostate cancer and liver cancer, the sequences were screen on the basis of RefSeq Status (taken from NCBI-nucleotide). Only the validated or reviewed lncRNA sequences were filtered from the complete list.

1. **Secondary *in-silico* analysis of G-quadruplex forming sequences**

A primary list of validated and reviewed lncRNAs in prostate cancer and liver cancer was curated by screening through QGRS Mapper. A secondary analysis was done through G4Hunter web-based software to screen via its algorithm for false positives, false negatives.

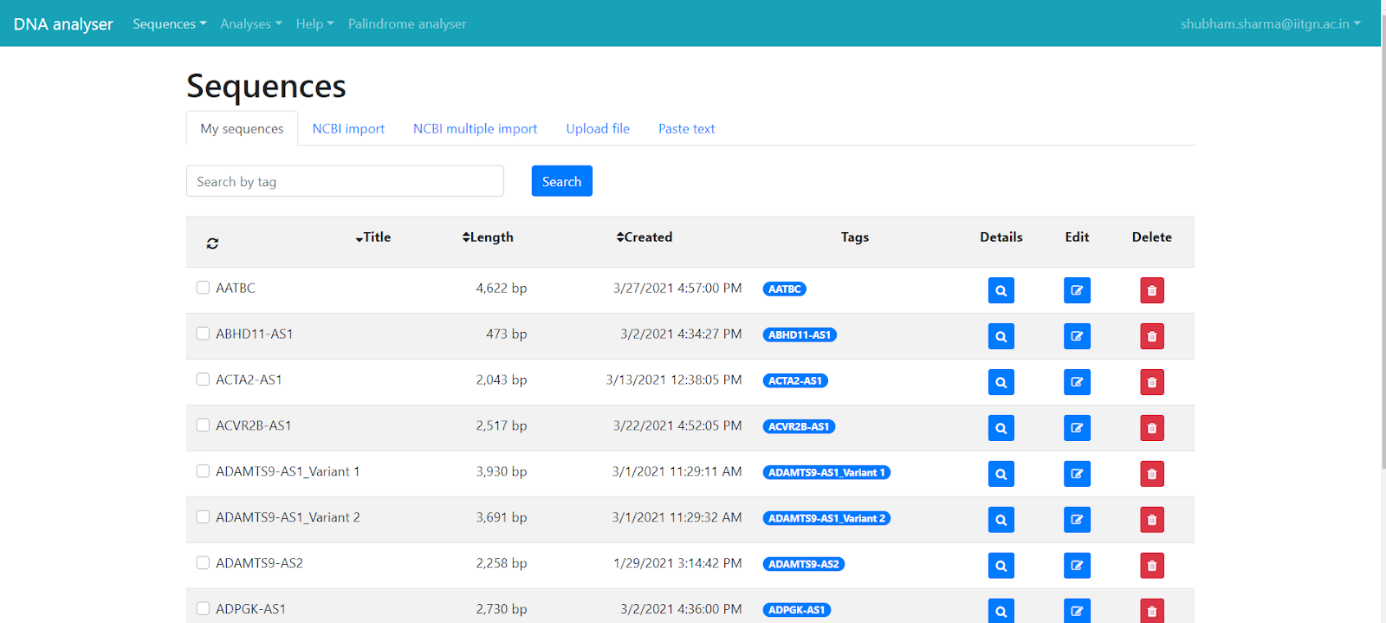
Firstly, the .exe files for both cancers were maintained with the following entry categories as (seen in Figure 11):

* Search parameters: Window size 45, threshold 0.9:
* Total No of PQS
* G4Hunter Scores of PQS
* Similarity between PQS from QGRS Mapper vs PQS from G4Hunter: PQS sequences between QGRS Mapper and G4Hunter were compared to find similarities, frame-shifts and overlaps.
* Search parameters: Window size 45, threshold 1.4:
* Total No of PQS
* G4Hunter Scores of PQS
* Similarity between PQS from QGRS Mapper vs PQS from G4Hunter: PQS sequences between QGRS Mapper and G4Hunter were compared to find similarities, frame-shifts and overlaps.



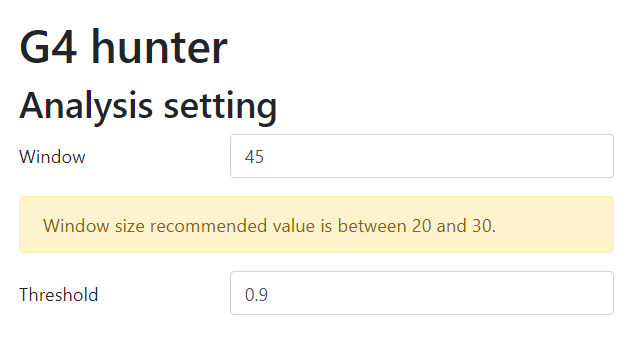
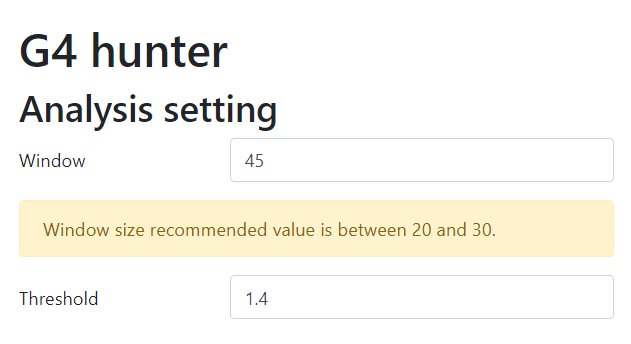
**Figure 11:** .exe file containing categories for data to be filled from G4Hunter

Secondly, sequences were imported using NCBI ID, with the ability to download multiple sequences from NCBI multiple imports, as seen in Figure 12. The software also allows upload of FASTA or txt file and/or directly paste sequences from the clipboard.



**Figure 12:** ‘My Sequences’ section in DNA analyser. It is a collection of uploaded sequences that can be taken for further analysis in G4Hunter

Thirdly, the sequences were analysed at different settings in G4Hunter. A fixed parameter for window size was set at 45. Each sequence (and variants, if any) was observed at thresholds 0.9 to screen for 2G PQS and 1.4 to screen for 3G, 4G PQS respectively, as seen in Figure 13.

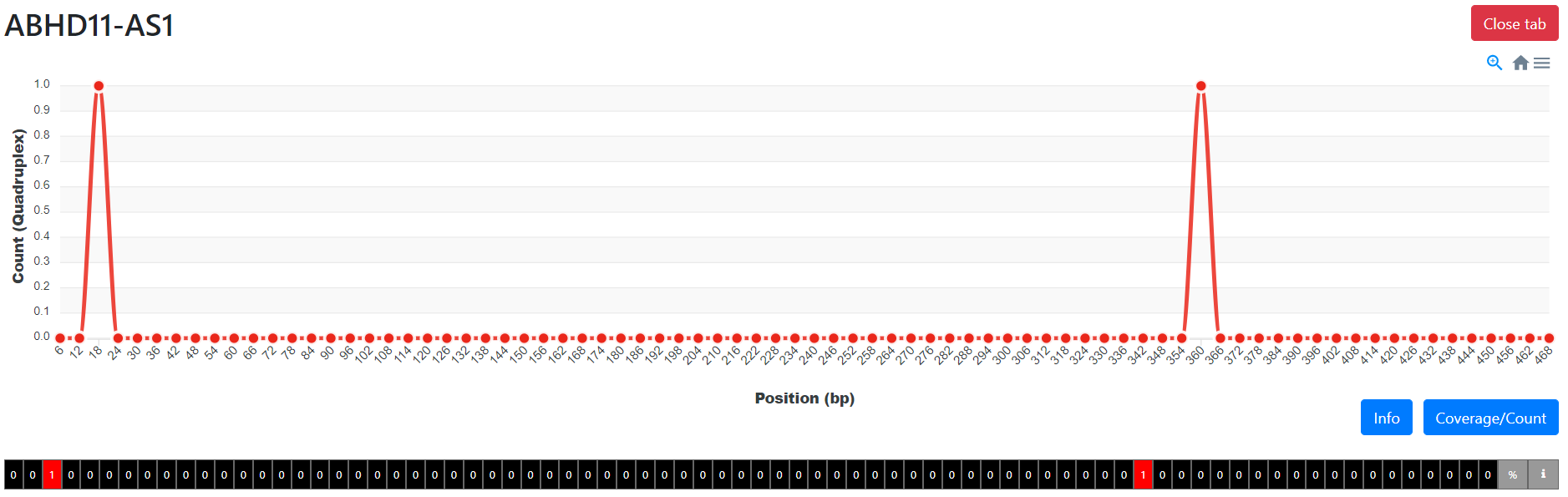
 

**Figure 13**: Interface for analysis setting in G4Hunter

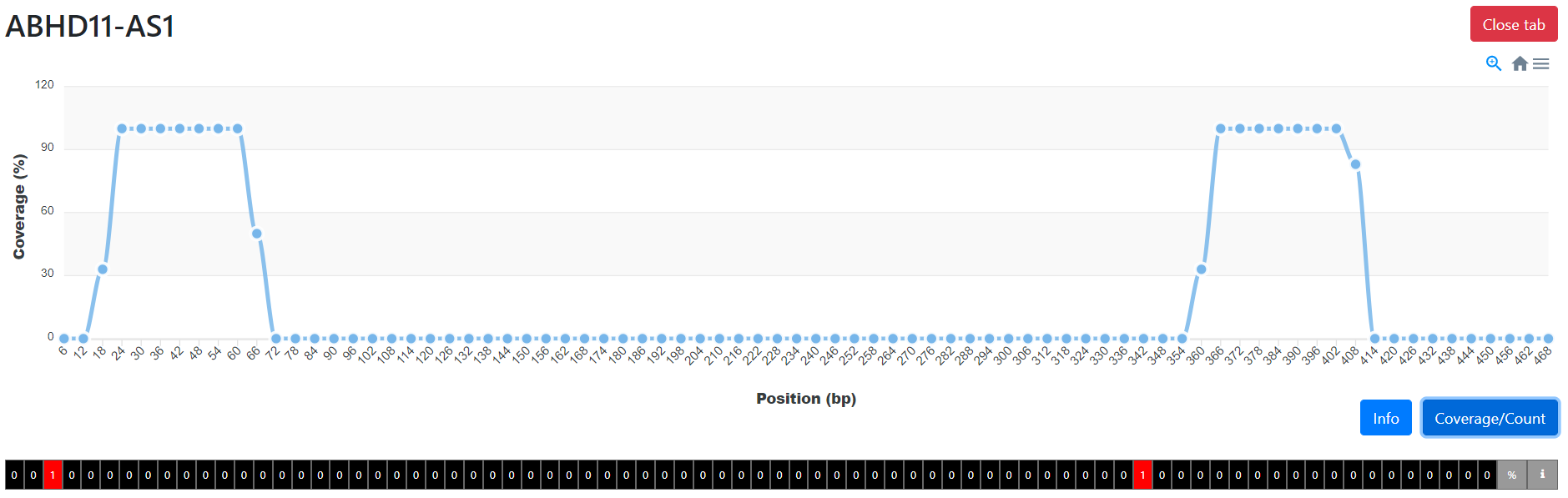
Next, the analysis was done for intuitive visualization and browsing of results. Below the sequence name, a heatmap was analysed that displays the number of G4-forming sequences in each segment. Two options of heatmap are available to view the result:

1. Count (quadruplex) vs Position (bp): This displays the number of quadruplexes with respect to position in the nucleotide sequence (Figure 14).
2. Coverage vs Position: This displays the % coverage with respect to position in the nucleotide sequence (Figure 15).

A red or blue coloured highlight in the bottom panel displays the number of results.



**Figure 14**: Heatmap analysis of lncRNA sequence ABHD11-AS1. X-axis shows count (quadruplex), y-axis shows position (bp).



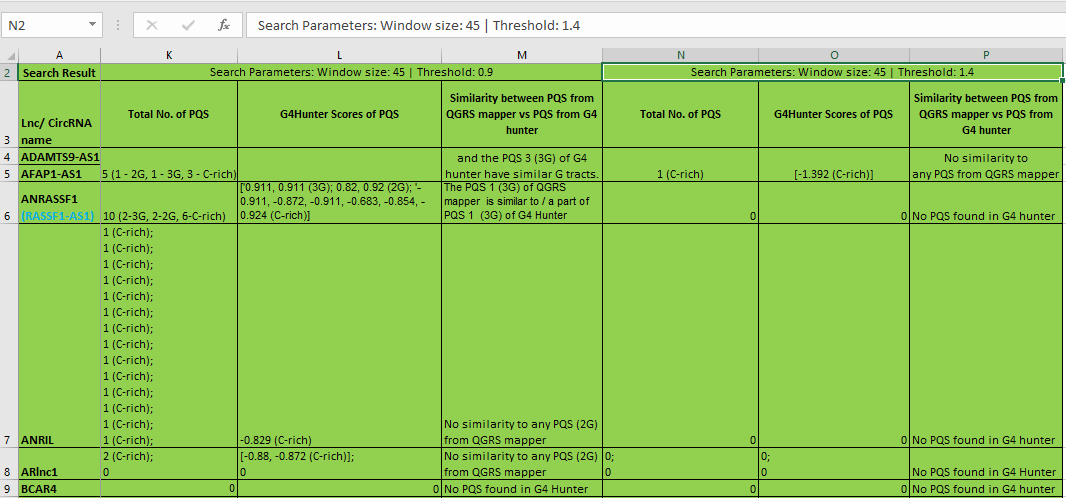
**Figure 15:** Heatmap analysis of lncRNA sequence ABHD11-AS1. X-axis shows coverage %, y-axis shows position (bp).

In Figure 16, the sequences that correspond to the analysis parameters (G in red—the longer the G-track, the brighter the intensity, C-rich in blue) are displayed. The sequence's position, duration, sequence, score map, and G4Hunter score are all shown. C-rich sequences are scored negatively according to the scoring algorithm and can form G-quadruplexes in their complementary strands.



**Figure 16:** Interface for analysis output for a typical lncRNA sequence

Next, the resulting PQS motifs were then grouped into three categories in the .exe file as seen in Figure 17. The total no of PQS found in G4Hunter, as well as type of PQS (2G, 3G, 4G or C-rich) was noted in the .exe file. Each lncRNA was individually analysed for multiple transcript variants, and data was collected in similar way. Data was also exported and collected as a text file for further processing.



**Figure 17**: Organisation of QGRS Mapper information about lncRNAs in curated database

Lastly, the PQS from QGRS Mapper and G4Hunter were compared to find similarities, frame-shifts and overlaps.

**RESULTS**

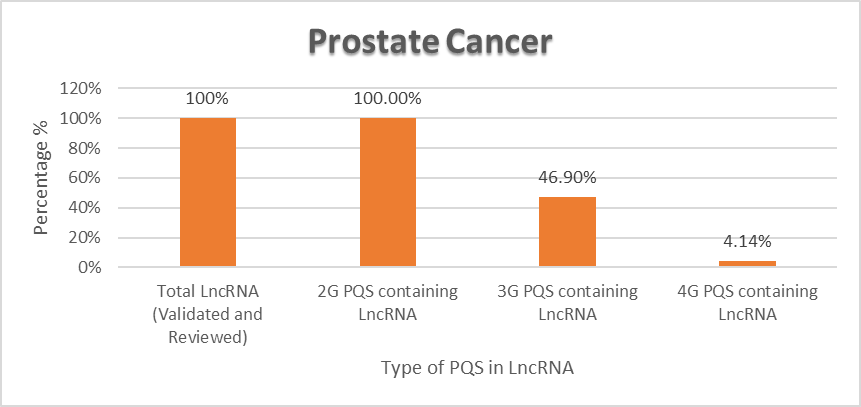
Out of 363 and 1116 sequences obtained from Lnc2Cancer 3.0 database, 145 & 304 validated and reviewed lncRNAs were found in prostate and liver cancer respectively from the NCBI-nucleotide database.

**Percentage distribution of PQS according to G-tract length**

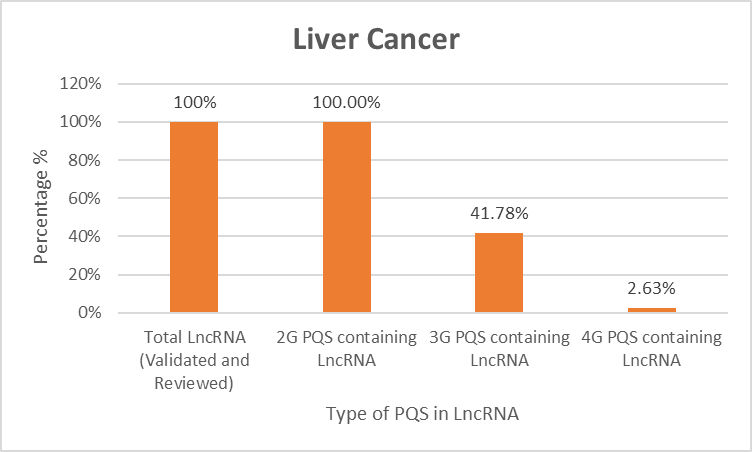
There were three types of quadruplex sequences shortlisted by the QGRS Mapper algorithm, namely sequences having 2G, 3G, and 4G tracts. In prostate cancer, out of 145 validated and reviewed lncRNAs, 100% had 2G tracts, 46.9 % had 3G tracts, and 4.14 % had 4G tracts. The segregation of lncRNAs according to G-tract length in PQS is seen in (Table 2). Whereas in liver cancer, out of 304 validated and reviewed lncRNAs, 100% had 2G tracts, 41.78 % had 3G tracts, and 2.63% had 4G tracts. The percentage of PQS on the basis of G-tract length is seen in Figure 18 for prostate cancer and Figure 19 for liver cancer.



**Table 2:** Count statistics for prostate cancer and liver cancer (in column). Listed from top to bottom: total number of sequences (including mRNA, genomic DNA, lncRNA), number of 2G PQS containing lncRNAs, number of 3G PQS containing lncRNAs, number of 4G PQS containing lncRNAs

****

**Figure 18:** Percentage of 2G, 3G and 4G PQS containing LncRNA in prostate cancer from the total validated and reviewed LncRNAs.

****

**Figure 19**: Percentage of 2G, 3G and 4G PQS containing LncRNA in liver cancer from the total validated and reviewed LncRNAs.

**Distribution of Total Unique PQS Across all LncRNAs**

We calculated the total ‘putative quadruplex forming sequences’ found in each lncRNA, including individual transcript variants. In prostate cancer, we found a total of 3408 unique sequences in 145 lncRNAs that were capable of forming G-quadruplexes, by *in-silico* analysis, as seen in Table 3. A large number of sequences totalling to 3167 PQS had G-tracts of 2G. Subsequently, we found a substantial decrease in PQS that had 3G tracts, totalling to 235. Lastly, the PQS with 4G tracts were merely 6 in number, which accounts to only 0.18% of all PQS. GASL1, H19, KCNQ1OT1 and MIR222HG long non-coding RNAs showed the presence of “GGGG” tracts (4G), with H19 and MIR222HG containing two PQS each.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Type of PQS in 145 LncRNAs** | | | |
|  | **Total No. of Unique PQS across all Transcript variants of LncRNAs** | **Unique 2G PQS across all Transcript Variants** | **Unique 3G PQS across all Transcript Variants** | **Unique 4G PQS across all Transcript Variants** |
| **Total No. of Unique PQS across all Transcript variants of LncRNAs** | 3408 | 3167 | 235 | 6 |
| **Percentage of Unique 2G, 3G or 4G PQS from Total PQS** |  | 92.93% | 6.90% | 0.18% |

**Table 3:** Distribution of unique 2G, 3G and 4G PQS in lncRNAs across all transcript variants in prostate cancer

Similarly in liver cancer, we found a total of 5684 unique sequences in 304 lncRNAs that were capable of forming G-quadruplexes, by in-silico analysis, as seen in Table 4. A large number of sequences totalling to 5285 PQS had G-tracts of 2G. Subsequently, we found a substantial decrease in PQS that had 3G tracts, totalling to 391. Lastly, the PQS with 4G tracts were merely 8 in number, which accounts to only 0.14% of all PQS. BAIAP2-AS1, ZKSCAN2-DT, H19, KCNQ1OT1, LINC00205, lncCAMTA1 long non-coding RNAs showed the presence of “GGGG” tracts (4G), with BAIAP2-AS1 and H19 containing two PQS each.

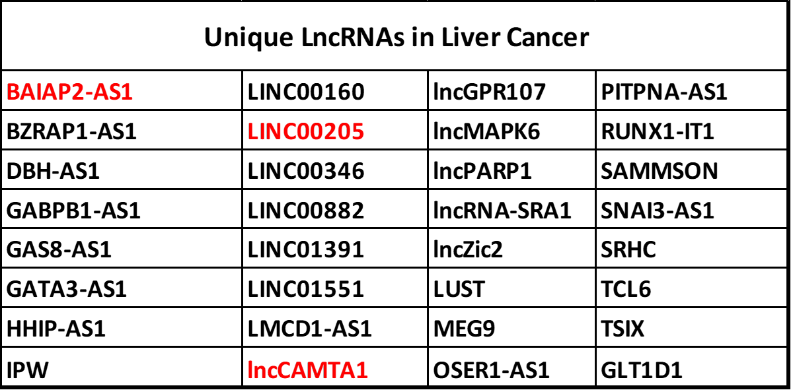
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Type of PQS in 304 LncRNAs** | | | |
|  | **Total No. of Unique PQS across all Transcript variants of LncRNAs** | **Unique 2G PQS across all Transcript Variants** | **Unique 3G PQS across all Transcript Variants** | **Unique 4G PQS across all** **Transcript Variants** |
| **Total No. of Unique PQS across all Transcript variants of LncRNAs** | 5684 | 5285 | 391 | 8 |
| **Percentage** |  | 92.98% | 6.88% | 0.14% |

**Table 4:** Distribution of unique 2G, 3G and 4G PQS in lncRNAs across all transcript variants in liver cancer

**Shortlisted LncRNAs with a Potential to Explore *in vivo***

From a total of 145 and 304 validated and reviewed lncRNAs in prostate and liver cancer respectively, 68 lncRNAs were found to be common in both cancers (excluding same lncRNA with different name).

77 lncRNAs were unique to prostate cancer, whereas 236 lncRNAs were unique to liver cancer. It is to be noted that many of these lncRNAs will be seen in other type of cancers, such as lung cancer, cervical cancer. It is only with a comprehensive evaluation that the true number of unique lncRNAs in each cancer can be found. After including lung, cervical, head & neck and gastric cancer in the comparison, it is seen that there are only 58 lncRNAs in prostate cancer and 161 lncRNAs in liver cancer that are unique to themselves (other cancers have not been included, the real number of unique lncRNAs should be far less). These lncRNAs were further segregated to eliminate those sequences which do not contain any 3G or 4G PQS. Ultimately, we were left with 8 and 32 lncRNAs respectively in prostate and liver cancer, listed in Table 5 and Table 6.

**Table 5:** The list of unique lncRNAs with at least 3G or 4G tracts in PQS in prostate cancer. Red colour is used to depict lncRNA with 4G tracts. **Table 6:** The list of unique lncRNAs with at least 3G or 4G tracts in PQS in liver cancer. Red colour is used to depict lncRNA with 4G tracts.

Out of these lncRNAs, \_\_\_\_\_\_, \_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_ have the least amount of literature available. It is worthy to note these lncRNAs for future explorations in-vivo.

**Screening through G4Hunter**

With the limitations of QGRS Mapper, we used another software G4Hunter that focuses on G/C asymmetry to assign a quadruplex propensity score. We analysed the occurrence of putative G4 sequences (PQS) by G4Hunter in total of 145 and 304 validated and reviewed lncRNAs in prostate and liver cancer respectively at 0.9 and 1.4 threshold. The analysis for each transcript variant of a lncRNA was done individually. For 145 lncRNAs in prostate cancer, a total of 366 lncRNAs were analysed. Similarly for 304 lncRNAs in liver cancer, a total of 533 lncRNAs were analysed. We found that it was easy to screen for 3G or 4G PQS containing lncRNAs at 1.4 threshold, due to their higher quadruplex propensity scores which indicated greater stability. Many C-rich sequences were found in each lncRNA, whose complementary strand would be capable of forming G-quadruplexes in theory.

However, many lncRNAs showed 3G and 4G PQS even at a threshold of 0.9. At 0.9 threshold, 192 lncRNAs contained 2G PQS, 61 lncRNAs contained 3G PQS and 6 lncRNAs contained 4G PQS in prostate cancer. At 1.4 threshold, no sequence was found for 2G PQS. 33 lncRNAs were found to contain 3G PQS, while 1 lncRNA was found to contain 4G PQS. Hence, ‘MIR222HG’ was found to contain PQS with 4G tracts. An example of PQS is:



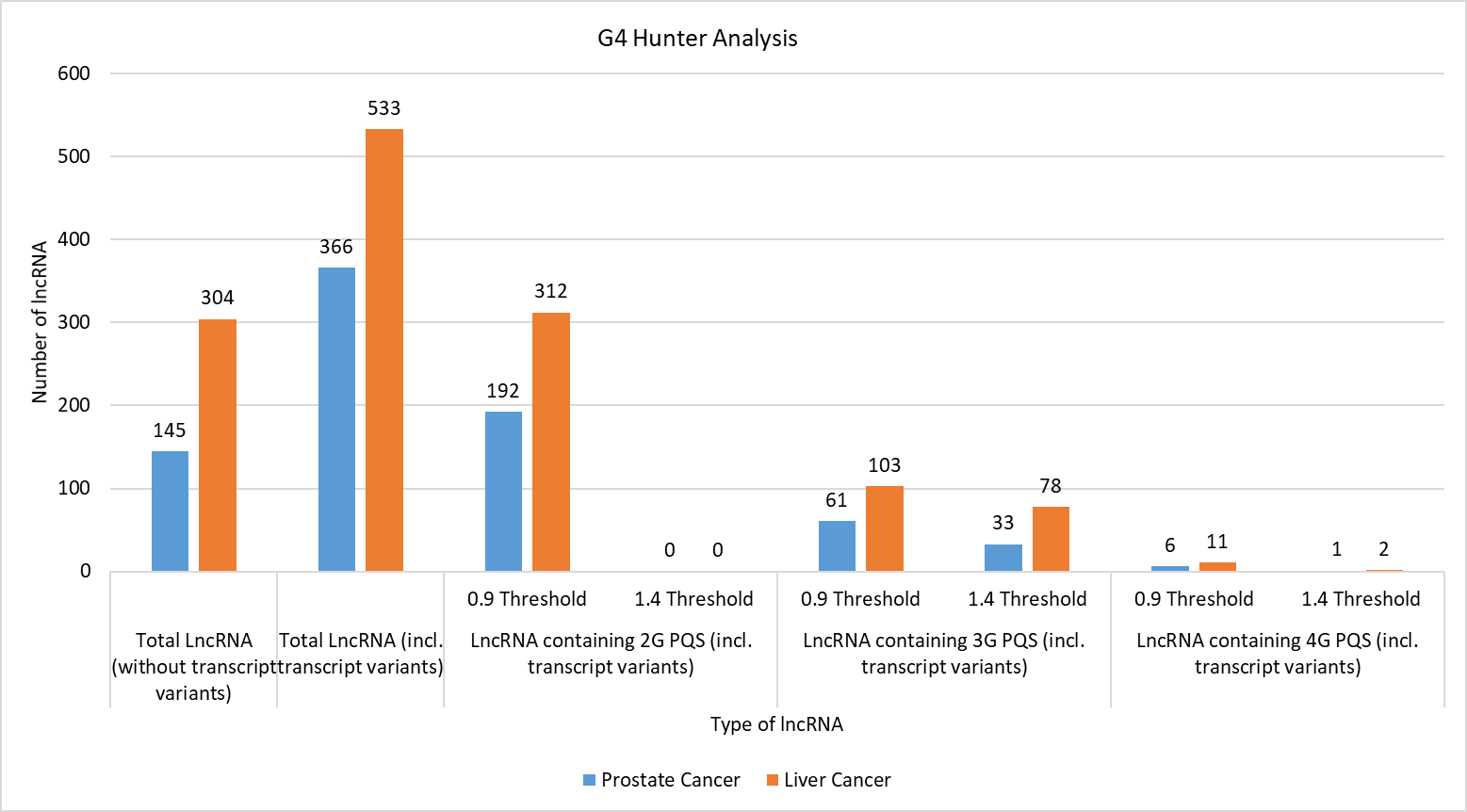
Similarly, at 0.9 threshold, 312 lncRNAs contained 2G PQS, 103 lncRNAs contained 3G PQS and 11 lncRNAs contained 4G PQS in prostate cancer. At 1.4 threshold, no sequence was found for 2G PQS. 78 lncRNAs were found to contain 3G PQS, while 2 lncRNAs were found to contain 4G PQS. Hence, ‘BAIAP2-AS1’ and ‘ZKSCAN2-DT’ was found to contain PQS with 4G tracts.

A comprehensive data of PQS distribution at both 0.9 and 1.4 threshold is seen in Table 7.



**Table 7:** *Columns from left to right:* total number of lncRNAs present in each cancer excluding individual transcript variants; total number of lncRNAs present in each cancer including individual transcript variants; number of 2G PQS containing lncRNAs including individual transcript variants – evaluated for 0.9 threshold and 1.4 threshold; number of 3G PQS containing lncRNAs including individual transcript variants – evaluated for 0.9 threshold and 1.4 threshold; number of 4G PQS containing lncRNAs including individual transcript variants – evaluated for 0.9 threshold and 1.4 threshold.

The above data is graphically represented below in Figure 20. Blue represents data for prostate cancer, orange represents data for liver cancer.



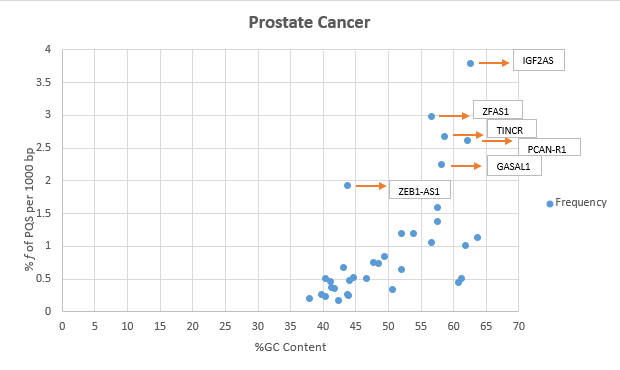
**Figure 20:** A graphical representation of analysed data for prostate and liver cancer in G4 Hunter

**Relationship between %GC Content and Frequency for G4-forming Sequences**

In general, a higher G4Hunter score corelates to a higher probability of G4s forming inside the PQS.The most stable PQS are those with G4Hunter scores above 1.4. We visualized the relationship between %GC content in lncRNA with the frequency of PQS at threshold 1.4. LncRNAs with multiple transcript variants were refined to select a single variant with highest number of PQS. Frequency was calculated by Total PQS/Length in bp \* 1000. %GC was calculated using <http://www.endmemo.com/bio/gc.php>.

Using the following values for G4Hunter algorithm—window size 45 and G4Hunter score above 1.4—we found 36 lncRNAs having a total of 139 PQS for prostate cancer. In prostate cancer, the %GC content average was 50.17%, with minimum 37.96% GC for MIR222HG and maximum 63.71% GC for MYU (VPS9D1-AS1). Mean frequency of predicted PQS per 1000 bp was found to be 1.0105. Highest frequency of predicted PQS per 1000 bp was seen in IGF2AS as 3.7968. Lowest frequency of predicted PQS per 1000 bp was seen in MIR222HG as 0.2006.

The PQS frequencies per 1000 bp for each lncRNA found at 1.4 threshold in prostate cancer is shown in Table 8. The tabular data is represented in graphical form as a scatterplot in Figure 21. With the help of the scatterplot, it is seen that PQS frequencies usually correlate with GC content, however there are many exceptions to this rule. LncRNAs with high PQS frequencies relative to their GC content (over 50% of the maximal observed PQS frequency) are labelled in orange colour, along with lncRNA name.



**Figure 21**: Relationship between observed frequency of PQS per 1000 bp and GC content in lncRNAs in prostate cancer. LncRNA with max. frequency per 1000 bp greater than 50% are described and highlighted in colour.

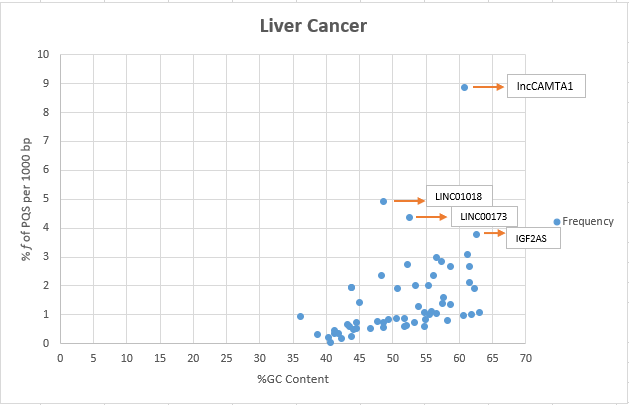


**Table 8:** Prostate cancer LncRNA sequences according to their sizes (bp), % GC Content, PQS (total number of predicted PQS), and frequency of predicted PQS (per 1000 bp).

Hence, IGF2AS, ZFAS1, TINCR, PCAN-R1, GASAL1 and ZEB1-AS1 lncRNAs were found to have highest PQS frequencies according to their GC content. BRE-AS1 was found to be the only lncRNA with high PQS frequency that was unique to prostate cancer.

Similarly, using the following values for G4Hunter algorithm—window size 45 and G4Hunter score above 1.4—we found 60 lncRNAs having a total of 238 PQS for liver cancer. In liver cancer, the %GC content average was 51.39%, with minimum 36.16% GC for WDFY3-AS2 and maximum 62.94% GC for HAR1A. Mean frequency of predicted PQS per 1000 bp was found to be 1.5064. Highest frequency of predicted PQS per 1000 bp was seen in lncCAMTA1 as 8.8809. Lowest frequency of predicted PQS per 1000 bp was seen in TSIX as 0.054.

The PQS frequencies per 1000 bp for each lncRNA found at 1.4 threshold in liver cancer is shown in Table 9. The tabular data is represented in graphical form as a scatterplot in Figure 22. With the help of the scatterplot, the proportional relationship between %GC content and frequency of PQS is maintained. The exceptions with high PQS frequencies relative to their GC content (over 50% of the maximal observed PQS frequency) are labelled in orange colour, along with lncRNA name.



**Figure 22**: Relationship between observed frequency of PQS per 1000 bp and GC content in lncRNAs in liver cancer. LncRNA with max. frequency per 1000 bp greater than 50% are described and highlighted in colour.

Hence, lncCAMTA1, LINC01018, LINC00173, IGF2AS, PCAT6 lncRNAs were found to have highest PQS frequencies according to their GC content. lncCAMTA1 was found to be the only lncRNA with high PQS frequency that was unique to liver cancer.



**Table 9**: Liver cancer LncRNA sequences according to their sizes (bp), % GC Content, PQS (total number of predicted PQS), and frequency of predicted PQS (per 1000 bp).

**PQS Comparison using different algorithms**

For the purpose of the database, each PQS found above 1.4 threshold in G4Hunter was manually compared with PQS from QGRS Mapper for each lncRNA. The results of the comparison were noted in the database under column ‘Similarity between PQS from QGRS mapper vs PQS from G4 hunter’. This comparison of similarity is an essential tool in understanding the accuracy of QGRS Mapper and G4Hunter in finding G quadruplex forming sequences.

**DISCUSSION**

It has been demonstrated that G4 structures could be used as targets for therapeutic drugs. Targeting appropriate dysregulated G4 structures could hold the key to finding the cure for cancer. Through our research, we found large numbers of PQS in particular lncRNAs that are unique to a particular cancer. In prostate cancer, we finalised ‘MIR222HG’ since it was found to contain large number of PQS, form the most hypothetical stable G4 structures based on score and length. Therefore, we hypothesize that this lncRNA must play an important role in the progression of prostate cancer. Studies about role of MIR222HG in regulating the AR function during the castration resistant prostate cancer (CRPC) development [48] back up our investigation. %GC-frequency analysis also revealed BRE-AS1 as having higher frequency (1.1197) according to its GC content (53.9%). With the help of our analysis, the exact site for formation of G-quadruplexes can be known, hence making the process of drug delivery extremely accurate. For liver cancer, we found BAIAP2-AS1, LINC00205 and lncCAMTA1 through QGRS Mapper; BAIAP2-AS1 and ZKSCAN2-DT through G4Hunter and lncCAMTA1 through %GC score-frequency analysis. Therefore, we come down to BAIAP2-AS1 and lncCAMTA1 as having the most potential for stable G-quadruplex formation in liver cancer. Research on hepatocellular carcinoma (HCC) shows its correlation with hepatitis B virus (HBV) infection, and the contributions of BAIAP2-AS1 and lncCAMTA1 to HBV-related HCC are explored [49] Since this area is quite unexplored, this analysis can be used as a preliminary point to shortlist candidate sequences for study of structure, function and ultimately cause of dysregulation in prostate and liver cancer.

The method of analysing lncRNAs by analysing the relationship between %GC content and frequency of PQS per 1000 bp can also prove to be useful in screening lncRNAs that may evade the algorithmic systems. Contrary to the standard pattern followed by most lncRNAs, the scatterplots revealed a few exceptions that have higher frequency of PQS per 1000 bp compared to %GC content. This enrichment may lead to discovery of unknown stable G4 structures with explorations *in vivo*. Different algorithms in software lead to diverse results in PQS due to various PQS thresholds and algorithms. In our research we used G4Hunter and QGRS Mapper for obtaining the broad variations of PQS frequencies and their scores.

**CONCLUSION**

In this research, we analysed the presence of PQS in long non-coding RNAs in prostate and liver cancer. This is a part of a larger database that will contain quadruplex forming regions in every known lncRNA in prostate, liver, cervical, lung, head & neck, gastric, colorectal, pancreatic and many more cancers.

In prostate cancer, MIR222HG and BRE-AS1 contained the most promising PQS for G4 structures on the basis of multiple analysis methods. In liver cancer, BAIAP2-AS1 and lncCAMTA1 contained the most promising PQS for G4 structures similarly. In conclusion, comprehensive views of the potential secondary structure formation in lncRNAs will provide new insights into its effect on dysregulation of cancers and other multifactorial conditions.

**FUTURE PROSPECTIVES**

Once our collective database is made public, it will have dramatic benefits to the entire scientific community working on G-quadruplex formation in lncRNAs. Many unknown regions that weren’t explored before can be identified based on the scoring systems of QGRS Mapper and G4 Hunter, and % GC content-frequency analysis in our database. Ultimately, the retrieval of information about the exact putative quadruplex forming sequences in each cancer can be done from a single place. Our collective efforts will result in saving huge amounts of time for scientists in the future in analysing hundreds of sequences, and facilitate smooth comparison of sequences, nucleotide-by-nucleotide.

Presently, we have completed the process of creating the database manually. I believe this can be automated to save large amounts of time for updates. In the future, the process can be done on the cloud so we would not require local processing power as well. In my opinion, the automation can be done as the following:

* Firstly, we need to address where we get the sequences from. For this we can either fetch from the API or we can design a crawler that pulls FASTA sequence from the NCBI-nucleotide website via its name or accession ID. This information can either be stored in a warehouse or a database. This can be on the cloud or on a local server.
* The next step would be to run the text in either the G4 Hunter or QGRS Mapper. This can either be done by using a Robotic Process Automation or write a simple OS script which could be scheduled as a job. e.g., Cron job.
* The output of the software would be a picture with the necessary sequences. For this, we could use Google's cloud vision API which will give us the extracted text from the picture.
* Next, we could write lines of code in either Python or Java to read the no of 2G, 3G and 4G sequences from each output text from QGRS Mapper/G4Hunter. For this we could use the FileReader library in Java or the BufferedReader/BufferedWriter libraries in Python. Next, "String" data structures can be used to match the information inside the database.
* With this approach, we can create a user-friendly program in the form of full desktop/mobile operating system. Merely by entering the NCBI name, the output from software can be entered inside the database. In this manner, we can not only save hours of manpower but also make the software accessible and portable.

With the help of clustered heat maps, we can also enable visualisation through graphics and faster interpretation than a tabular data form. Construction of a heat map generally requires the assistance of a biostatistician or bioinformatics analyst; however, we propose using ‘Interactive Heat Map Builder’ such as one developed by M.D Anderson Cancer Center. Thus, a biologist working on G4 structures with no prior bioinformatics experience can use the heatmap to easily move back and forth for comparison of 2G, 3G and 4G PQS forming regions in each lncRNA across cancers. Hence, Next-Generation Clustered Heat Map (NG-CHM) can be used to explore the information in database more dynamically.

**REFERENCES**

1. Conn, G. L., & Draper, D. E. (**1998**). RNA structure. *Current Opinion in Structural Biology*, *8*(3), 278–285.

2. Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., … Gingeras, T. R. (**2007**). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, *316*(5830), 1484–1488.

3. Dhamija, S., & Menon, M. B. (**2018**, August 3). Non-coding transcript variants of protein-coding genes–what are they good for? *RNA Biology*. Taylor and Francis Inc.

4. Mercer, T. R., Dinger, M. E., & Mattick, J. S. (**2009**, March). Long non-coding RNAs: Insights into functions. *Nature Reviews Genetics*. Nat Rev Genet.

5. Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., … Rastan, S. (**1992**). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, *71*(3), 515–526.

6. Kim, E., Magen, A., & Ast, G. (**2007**). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, *35*(1), 125–131.

7. Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., … Frey, B. J. (**2010**). Deciphering the splicing code. *Nature*, *465*(7294), 53–59.

8. Lewis, B. P., Green, R. E., & Brenner, S. E. (**2003**). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(1), 189–192.

9. Guttman, M., & Rinn, J. L. (**2012**, February 16). Modular regulatory principles of large non-coding RNAs. *Nature*. Nature.

10. Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., … Lander, E. S. (**2009**). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, *458*(7235), 223–227.

11. Guttman, M., & Rinn, J. L. (**2012**, February 16). Modular regulatory principles of large non-coding RNAs. *Nature*. Nature.

12. Wang, Z., Zhang, X. J., Ji, Y. X., Zhang, P., Deng, K. Q., Gong, J., … Wang, Y. (**2016**). The long noncoding RNA Chaer defines an epigenetic checkpoint in cardiac hypertrophy. *Nature Medicine*, *22*(10), 1131–1139.

13. Mercer, T. R., & Mattick, J. S. (**2013**, March). Structure and function of long noncoding RNAs in epigenetic regulation. *Nature Structural and Molecular Biology*. Nat Struct Mol Biol.

14. Loewer, S., Cabili, M. N., Guttman, M., Loh, Y. H., Thomas, K., Park, I. H., … Rinn, J. L. (**2010**). Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nature Genetics*, *42*(12), 1113–1117.

15. Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., … Chang, H. Y. (**2010**). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, *464*(7291), 1071–1076.

16. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer - PubMed. (n.d.). Retrieved May 14, 2021, from https://pubmed.ncbi.nlm.nih.gov/10606244/

17. Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., … Chinnaiyan, A. M. (**2011**). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature Biotechnology*, *29*(8), 742–749.

18. Prensner, J. R., Sahu, A., Iyer, M. K., Malik, R., Chandler, B., Asangani, I. A., … Chinnaiyan, A. M. (**2014**). The IncRNAs PCGEM1 and PRNCR1 are not implicated in castration resistant prostate cancer. *Oncotarget*, *5*(6), 1434–1438.

19. Zhu, M., Chen, Q., Liu, X., Sun, Q., Zhao, X., Deng, R., … Yu, J. (**2014**). LncRNA H19/miR-675 axis represses prostate cancer metastasis by targeting TGFBI. *FEBS Journal*, *281*(16), 3766–3775.

20. Walsh, A. L., Tuzova, A. v., Bolton, E. M., Lynch, T. H., & Perry, A. S. (**2014**). Long noncoding RNAs and prostate carcinogenesis: The missing “linc”? *Trends in Molecular Medicine*. Elsevier Ltd.

21. Zhang, A., Zhang, J., Kaipainen, A., Lucas, J. M., & Yang, H. (**2016**, June 1). Long non-coding RNA: A newly deciphered “code” in prostate cancer. *Cancer Letters*. Elsevier Ireland Ltd.

22. Takayama, K. I., Horie-Inoue, K., Katayama, S., Suzuki, T., Tsutsumi, S., Ikeda, K., … Inoue, S. (**2013**). Androgen-responsive long noncoding RNA CTBP1-AS promotes prostate cancer. *EMBO Journal*, *32*(12), 1665–1680.

23. Liu, X. hua, Liu, Z. li, Sun, M., Liu, J., Wang, Z. xia, & De, W. (**2013**). The long non-coding RNA HOTAIR indicates a poor prognosis and promotes metastasis in non-small cell lung cancer. *BMC Cancer*, *13*.

24. Bedrat, A., Lacroix, L., & Mergny, J. L. (**2016**). Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Research*, *44*(4), 1746–1759.

25. Monsen, R. C., & Trent, J. O. (**2018**, September 1). G-quadruplex virtual drug screening: A review. *Biochimie*. Elsevier B.V.

26. Sen, D., & Gilbert, W. (**1990**). A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature*, *344*(6265), 410–414.

27. Jayaraj, G. G., Pandey, S., Scaria, V., & Maiti, S. (**2012**). Potential G-quadruplexes in the human long non-coding transcriptome. *RNA Biology*, *9*(1), 81–89.

28. Fay, M. M., Lyons, S. M., & Ivanov, P. (**2017**, July 7). RNA G-Quadruplexes in Biology: Principles and Molecular Mechanisms. *Journal of Molecular Biology*. Academic Press.

29. Ma, Y., Iida, K., & Nagasawa, K. (**2020**). Topologies of G-quadruplex: Biological functions and regulation by ligands. *Biochemical and Biophysical Research Communications*, *531*(1), 3–17.

30. Bhattacharyya, D., Arachchilage, G. M., & Basu, S. (**2016**). Metal cations in G-quadruplex folding and stability. *Frontiers in Chemistry*. Frontiers Media S. A.

31. Neidle, S., & Balasubramanian, S. (Eds.). (**2007**). *Quadruplex Nucleic Acids*. Cambridge: Royal Society of Chemistry.

32. Risitano, A., & Fox, K. R. (**2004**). Influence of loop size on the stability of intramolecular DNA quadruplexes. *Nucleic Acids Research*, *32*(8), 2598–2606.

33. Rachwal, P. A., Findlow, I. S., Werner, J. M., Brown, T., & Fox, K. R. (**2007**). Intramolecular DNA quadruplexes with different arrangements of short and long loops. *Nucleic Acids Research*, *35*(12), 4214–4222.

34. Rachwal, P. A., Brown, T., & Fox, K. R. (**2007**). Effect of G-tract length on the topology and stability of intramolecular DNA quadruplexes. *Biochemistry*, *46*(11), 3036–3044.

35. Lane, A. N., Chaires, J. B., Gray, R. D., & Trent, J. O. (**2008**, October 1). Stability and kinetics of G-quadruplex structures. *Nucleic Acids Research*. Oxford University Press.

36. Wieland, M., & Hartig, J. S. (**2007**). RNA Quadruplex-Based Modulation of Gene Expression. *Chemistry and Biology*, *14*(7), 757–763.

37. Lightfoot, H. L., Hagen, T., Tatum, N. J., & Hall, J. (**2019**, August 1). The diverse structural landscape of quadruplexes. *FEBS Letters*. Wiley Blackwell.

38. Lombardi, E. P., & No-Vallejo, A. L. (**2020**). Erratum: A guide to computational methods for g-quadruplex prediction (Nucleic Acids Research (2019) (gkz1097) DOI: 10.1093/nar/gkz1097). *Nucleic Acids Research*. Oxford University Press.

39. Gao, Y., Shang, S., Guo, S., Li, X., Zhou, H., Liu, H., … Zhang, Y. (**2021**). 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 Research*, *49*(D1), D1251–D1258.

40. Kikin, O., D’Antonio, L., & Bagga, P. S. (**2006**). QGRS Mapper: A web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Research*, *34*(WEB. SERV. ISS.).

41. Mukundan, V. T., & Phan, A. T. (**2013**). Bulges in G-quadruplexes: Broadening the definition of G-quadruplex-forming sequences. *Journal of the American Chemical Society*, *135*(13), 5017–5028.

42. Guédin, A., Gros, J., Alberti, P., & Mergny, J. L. (**2010**). How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Research*, *38*(21), 7858–7868.

43. Brázda, V., Kolomazník, J., Lýsek, J., Bartas, M., Fojta, M., Šťastný, J., & Mergny, J. L. (**2019**). G4Hunter web application: A web server for G-quadruplex prediction. *Bioinformatics*, *35*(18), 3493–3495.

44. Broom, B. M., Ryan, M. C., Stucky, M., Wakefield, C., Melott, J. M., Akbani, R., & Weinstein, J. N. (**2020**). Interactive Clustered Heat Map Builder: An easy web-based tool for creating sophisticated clustered heat maps. *F1000Research*, *8*, 1750.

45. Šmarda, P., Bureš, P., Horová, L., Leitch, I. J., Mucina, L., Pacini, E., … Rotreklová, O. (**2014**). Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(39), E4096–E4102.

46. Robinson, R. (**2006**). More GC Means More RNA. *PLoS Biology*, *4*(6), e206.

47. Niazi, F., & Valadkhan, S. (**2012**). Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3′ UTRs. *RNA*, *18*(4), 825–843.

48. Qu, F., Sun, T., Wang, X., Lee, G.-S. M., & Kantoff, P. W. (**2016**). Abstract A10: LncRNA MIR222HG transcribed from the miR-221/222 promoter are associated with the development of castration resistant prostate cancer (pp. A10–A10). American Association for Cancer Research (AACR).

49. Gong, X., Wei, W., Chen, L., Xia, Z., & Yu, C. (**2016**). Comprehensive analysis of long non-coding RNA expression profiles in hepatitis B virus-related hepatocellular carcinoma. *Oncotarget*, *7*(27), 42422–42430.