Analysis of putative G-quadruplex forming sequences in inflammatory mediators and their potential as targets for treating inflammatory disorders.

Stefan Bidula

School of Biological Sciences, University of East Anglia, Norwich, United Kingdom

Corresponding author: Stefan Bidula; s.bidula@uea.ac.uk

Abstract

G-quadruplexes (G4s) are non-canonical secondary structures located in DNA and RNA which have demonstrable roles in the regulation of transcription and translation. G4s have received considerable interest as a drug target in cancer, given their ability to regulate the expression of proto-oncogenes and inhibit growth of cancer cells. However, their presence in the genes of inflammatory mediators has not been discussed to date. Therefore, we computationally investigated putative quadruplex-forming sequences (PQS) in the promoters and gene bodies of cytokines and chemokines. Here, we demonstrated that the promoters of IL-6, IL-12, IL-17, TGF-β, TNF, and β-chain family cytokines and XC and TAFA family chemokines display high PQS frequencies comparable to those observed in proto-oncogenes. Moreover, 47.82% of the gene promoters contained sequences with high propensity to form G4s. Furthermore, G4s can primarily be found within the GC-boxes and binding sites for specificity protein and Krüppel-like transcription factors. However, they can also be found located in a further 59 sites involved in the binding of transcription factors involved in inflammation and immunity such as NF-κB1, RelA, RelB, IRF5, and NFAT5. We also identified that 72.17% and 70.43% of genes investigated contained sequences highly likely to form G4s in their coding and template strands, respectively. Exploring the regulatory roles of G4s in genes encoding inflammatory mediators could provide novel drug targets to modulate inflammation and treat inflammatory diseases.

Keywords: G-quadruplex, transcription, transcription factors, cytokine storm, therapeutics

Abbreviations: G4 = G-quadruplex, PQS = putative G-quadruplex-forming sequences

1. Introduction

Chronic or uncontrolled inflammation can result in the manifestation of numerous debilitating or life-threatening conditions such as arthritis, asthma, atherosclerosis, blindness, cancer, diabetes, and sepsis. Conversely, deficient production of cytokines and chemokines can result in predisposition to infection. The ability to fine tune the expression of cytokines and chemokines at the gene and protein levels could therefore be an interesting approach to recalibrate the immune response in inflammatory disorders and infection.

G4s are non-canonical secondary structures located in G-rich regions of RNA and DNA. They are formed when four guanine bases associate via Hoogsteen hydrogen bonding to form G-tetrads, with 2 or more tetrads stacking upon one another to form parallel, antiparallel, or hybrid quadruplexes based upon their topology (Figure 1A). G4s are generally defined by the sequence $G_{\geq 2}N_{1\text{-}7}G_{\geq 2}N_{1\text{-}7}G_{\geq 2}N_{1\text{-}7}G_{\geq 2}$; whereby G represents guanines in the tract and N represents the nucleotide loop length. There is growing interest in targeting G4s therapeutically due to their emerging roles in important biological processes such as replication, mRNA splicing, and epigenetic regulation [1]. However, what makes G4s particularly interesting drug targets are their enrichment in regulatory regions such as the promoters and telomeric regions of eukaryotic genomes, their presence in RNAs, and their involvement in regulating transcription and translation [1].

It has been noted that G4s can either enhance or inhibit expression of genes depending upon their location along the gene (e.g. promoter/gene body), or which strand the G4 is found on (template/coding). DNA G4s located upstream of the transcription start site (TSS) in transcription factor (TF) binding sites have been shown to either enhance the binding of TFs (as observed for specificity protein [SP] 1) or displace TFs, which can result in either an upregulation or downregulation of transcription. Similarly, G4s in the gene body of the coding strand can potentially facilitate transcription re-initiation, or form DNA-RNA hybrids which can result in RNA polymerase II stalling or early termination of transcription. Moreover, G4s in the coding strand can also result in RNAs that can form G4 structures, thus inhibiting translation.

Alternatively, G4s in the gene body of the template strand can block the progression of RNA polymerase II, which results in inhibited transcription (reviewed in [2]).

There has been much effort placed into developing drugs that can bind to G4s to either stabilise or disrupt their formation and modulate the biological functions they regulate. As with most gene targeting strategies (e.g. siRNA, shRNA, CRISPR, TALENs) specificity is still currently an issue. However, in the past few years, major advancements have been made in the field and the ability to target specific G4s is starting to become a reality [3]. Thus, the therapeutic potential of targeting these structures in disease has increased exponentially. To date, much of the research in this field has been focused on exploring the role of G4s in cancer, particularly due to their prevalence in proto-oncogenes and telomeres [4]. However, their importance in regulating the expression of growth factors such as vascular endothelial growth factor (VEGF) and platelet-derived growth factor (PDGF) indicate potential important implications for inflammatory and infectious diseases [5, 6]. Here, we show here that most of the genes encoding cytokines and chemokines also contain high frequencies of putative G4-forming sequences (PQS) and sequences highly likely to form stable G4s. We hypothesise that targeting G4s in genes encoding inflammatory mediators could be a novel means to calibrate inflammatory responses.

2. Materials and Methods

2.1 Sequences analysed and the process of analysis

Experimentally validated promoter sequences were obtained from the Eukaryotic Promoter Database (https://epd.epfl.ch/; last accessed 08/01/21) and the sequences from -900 to +100 base pairs relative to the transcription start site (TSS) were retrieved for analysis. Promoter features (e.g. GC box) and putative TF binding sites were identified via the Eukaryotic Promoter Database Promoter Elements and JASPAR motif libraries (2018 release), respectively. The FASTA for the complete gene sequences were obtained from the NCBI gene database.

The G4Hunter web application was used to identify PQS within promoter regions and complete gene sequences [7]. When analysing the sequences, a window size of 25 nucleotides and a threshold of 1.2 was used for analysis. The algorithm used specifically identifies sequences which display the necessary requirements to form quadruplexes. PQS frequency is reported as the number of PQS per 1000 base pairs (PQS/kbp). The same sequences were analysed using another PQS-prediction tool called QGRS Mapper. The highest scoring sequences (indicating those most likely to form G4s) containing 2≥G and a loop length of 12 were documented. The identity of the genes investigated, and a summary of all the results can be found in Supplementary Figure 1.

2.2 Statistical analysis

Statistical analyses were conducted via One-way ANOVA with Dunnett's multiple comparisons. All analyses were performed using GraphPad Prism software v6.01. $p<0.05$ were considered statistically significant.

3. Results and Discussion

G4s have been demonstrated to have key roles in the regulation of gene expression and the location of these nucleic acid structures within a gene is correlated with their ability to regulate transcription. We first identified the frequency of PQS within the promoter regions and gene bodies of genes encoding cytokines and chemokines. The PQS frequencies in the promoters of genes encoding IL-1, IL-10, interferon (IFN), γ-chain, CC, and CXC family group members were significantly lower than those observed in proto-oncogene promoters (Figure 1B). However, genes encoding IL-6, IL-12, IL-17, β-chain, TNF, TGF-β, XC, and TAFA family members have high PQS frequencies in their promoters which are comparable to proto-oncogenes (Figure 1B). As there was only one member in the CX3C group, comparisons could not be made. The promoters of some genes had particularly high frequencies (PQS/kbp ≥ 10), such as *IL1F5*, *TNF*, *TNFSF13*, *TNFSF4*, *CLCF1*, *IL11*, *LEP*, *OSM*, *TGFB1*, *CCL19*, *CCL27*, *CX3CL1*, *CXCL16*, *FAM19A1*, *IL12A*, *IL32*, and *CSF3* (Supplementary Figure 1). Conversely, there were no significant differences between the PQS frequencies found in the entire

sequences of the genes encoding inflammatory proteins or proto-oncogenes (Figure 1C). Furthermore, the promoters of all IL-12, TGF-β, CXC, and XC family members, the coding strands of all TGF-β, TNF, Other, CXC and TAFA family members, and the template strands of all IL-10, TGF-β, Other, CXC, and TAFA family members contained PQS with a high propensity to form stable G4s comprised of ≥3 G-tetrads (Figure 1D).

We next investigated whether the PQS in these gene promoters could be found located within TF binding sites or promoter motifs. PQS with potential to form G4s comprised of ≥3 G-tetrads were predominantly found within GC boxes, but PQS were also found in initiators (*TNF* and *IL32*), TATA boxes (*IL24* and *IL32*), and CCAAT boxes (*IL13* and *IL32*; Supplementary Figure 1). For all genes, PQS could be found within the binding sites for the zinc finger transcription factors SP 1, 2, 3, 4, 8, Krüppel-like factors (KLF) 1, 4, 5, 9, 12, 13, 14, 16, zinc finger protein (ZNF) 263 and 740, and PLAG1. It is unsurprising to find PQS located within the binding sites for SP/KLF TFs given the prevalence of strong G4-forming sequences within the GC box elements and the affinity of SP1 for binding G4 structures [8]. However, PQS could also be found in the predicted binding sites for an additional 59 TFs, including important TFs associated with the function of the immune system such as NF-κB1, NF-κB2, RelA, RelB, STAT1, STAT3, RORγ, EBF1, ETS1, IRF1, IRF3, IRF5, SPI1, NFAT5, NFATC1, NFATC2, and XBP1 (Figure 1E; Supplementary Figure 1). Interestingly, PQS could be found in the binding sites of 37 TFs in *TNF*, most notably IRF5, SPI1, NFAT5, NFATC1, NFATC2, NF-κB2, and EBF1 (Supplementary Figure 1).

Not only promoters, the presence of G4s within the gene bodies themselves have been demonstrated to modulate transcription and translation [2]. We found that 72.17% (83/115) and 70.43% (81/115) genes contained at least one sequence with the potential to form a G4 compromised of ≥3 G-tetrads in their coding and template strands, respectively (Supplementary Figure 1). Genes with the highest PQS frequencies (PQS/kbp ≥ 5) included *IL18BP, IL1F5, LTB, TNF, TNFSF12, IL29, CLCF1, IL11, LIF, OSM, TGFB1, CCL14, CCL19,*

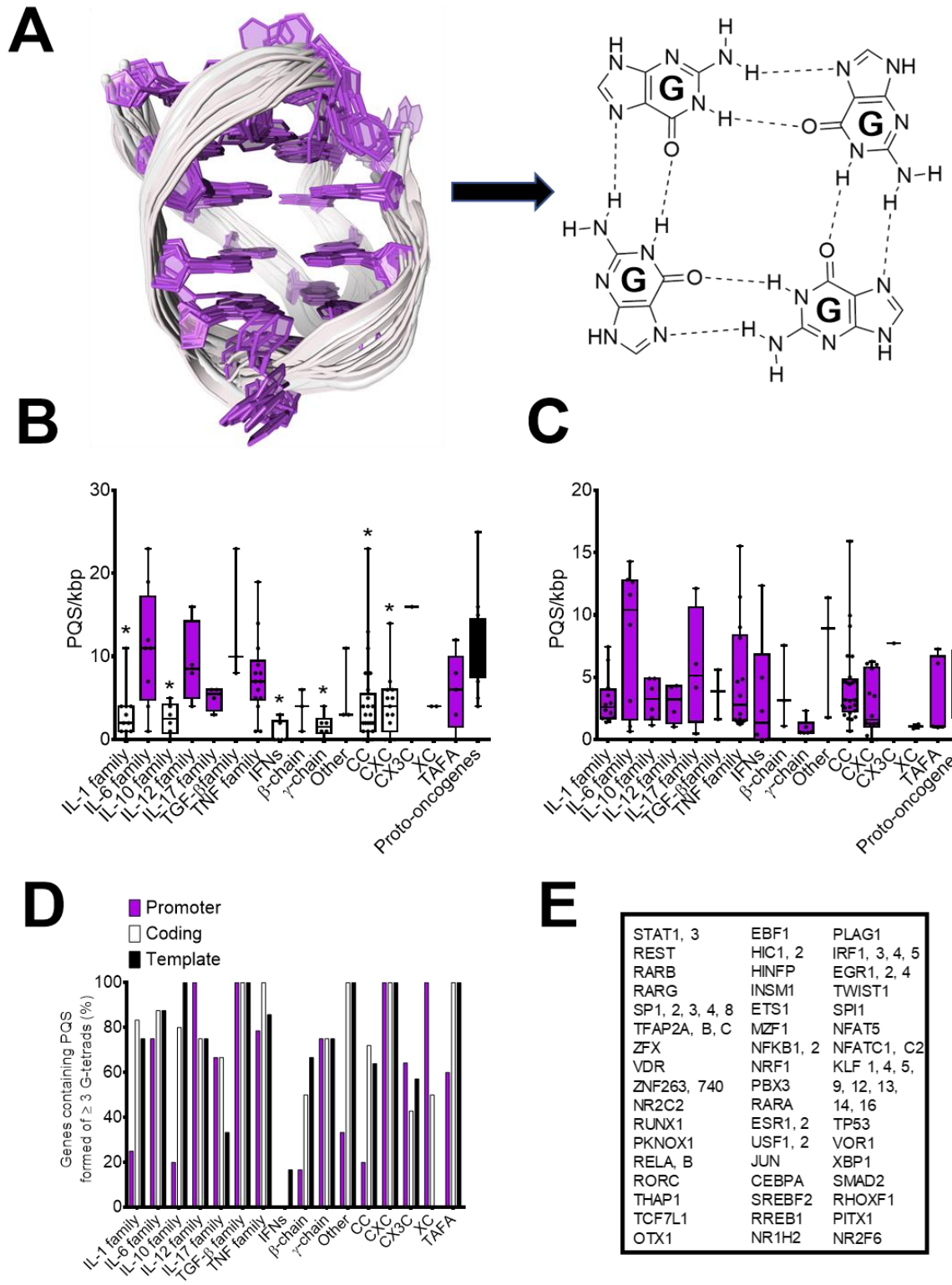

Figure 1. The frequency of PQS in the promoters and entire sequences of genes encoding cytokines and chemokines (A) A representative image of a G4 depicting a 2-tetrad human telomeric DNA quadruplex in K^+ solution (PDB: 2KKA). One of the 2 G-tetrads is highlighted and the Hoogsteen hydrogen bonding interactions between four guanine bases are displayed. The frequency of PQS in the promoters **(B)** or entire gene sequences of

cytokines and chemokines **(C)**. **(D)** The percentage of genes in cytokine and chemokine families containing PQS with high propensity to form G4s composed of ≥ 3 G-tetrads in their promoters, coding strands, or template strands. **(E)** A list of all the TFs which may be affected by stabilising G4s in the promoter regions of cytokines and chemokines. All gene promoters contained PQS in the binding sites for the SP/KLF family members. The identity of gene promoters containing PQS in the binding sites for all other TFs can be found in Supplementary Figure 1. Significance was determined via one-way ANOVA with Dunnett's multiple comparisons. $p < 0.05$ was considered statistically significant.

CCL21, CCL24, CCL27, CCL4, CX3CL1, CXCL1, CXCL14, CXCL16, CXCL3, FAM19A3, FAM19A5, IL13, IL32, CSF2, CSF3, IL17C, and *IL17D.* It is also interesting to note that the sequence **GGGG**ACTGTTGT**GGGG**T**GGGGGG**A**GGGGGG** (or subtle variations to this sequence) appears frequently in the genes of inflammatory mediators (such as *IL1RA, EDA, TNFSF13, TNFSF4, FAM19A1, FAM19A2,* and *IL15*) and their receptors (*IL1RAPL1, IL1RAPL2, FAS, TNFRSF11B, LEPR, IL20RB,* and *ACVR1C*; data not shown). Thus, the prevalence of this sequence suggests it likely has a role to play in the function/regulation of these genes.

Many cytokines and chemokines with high PQS frequencies participate in sepsis and cytokine storm, classic examples of when inflammatory regulation goes awry. Cytokine storm has been attributed to increased production of IL-1β, IL-6, IL-7, IL-8, IL-9, IL-10, IL-17A, IL-18, FGF, G-CSF, GM-CSF, CXCL10, CCL2, CCL3, CCL4, PDGF, VEGF, and TNF-α [9]. It is already well established that G4s can form in the promoters of *VEGFA* and *PDGFA* and stabilising these G4s can result in decreased transcription [5, 6]. We found here that many of the genes involved in cytokine storm contained sequences with the potential to form G4s formed of ≥3 G-tetrads in their promoter regions (*IL6, IL9, CSF2, CSF3, TNF*), coding strands (*IL6, IL7, IL18, CSF2, CSF3, CCL2, CCL3, CCL4, TNF*), and template strands (*IL1B, IL6, IL7, IL9, IL10, CSF2, CSF3, CCL3, TNF;* Figure 2). Therefore, targeting G4s in these genes could provide a novel therapeutic approach to treating cytokine storm. However, in depth analysis of their roles in transcriptional regulation in inflammatory cells must first be conducted to identify whether stabilising G4s can regulate gene expression.

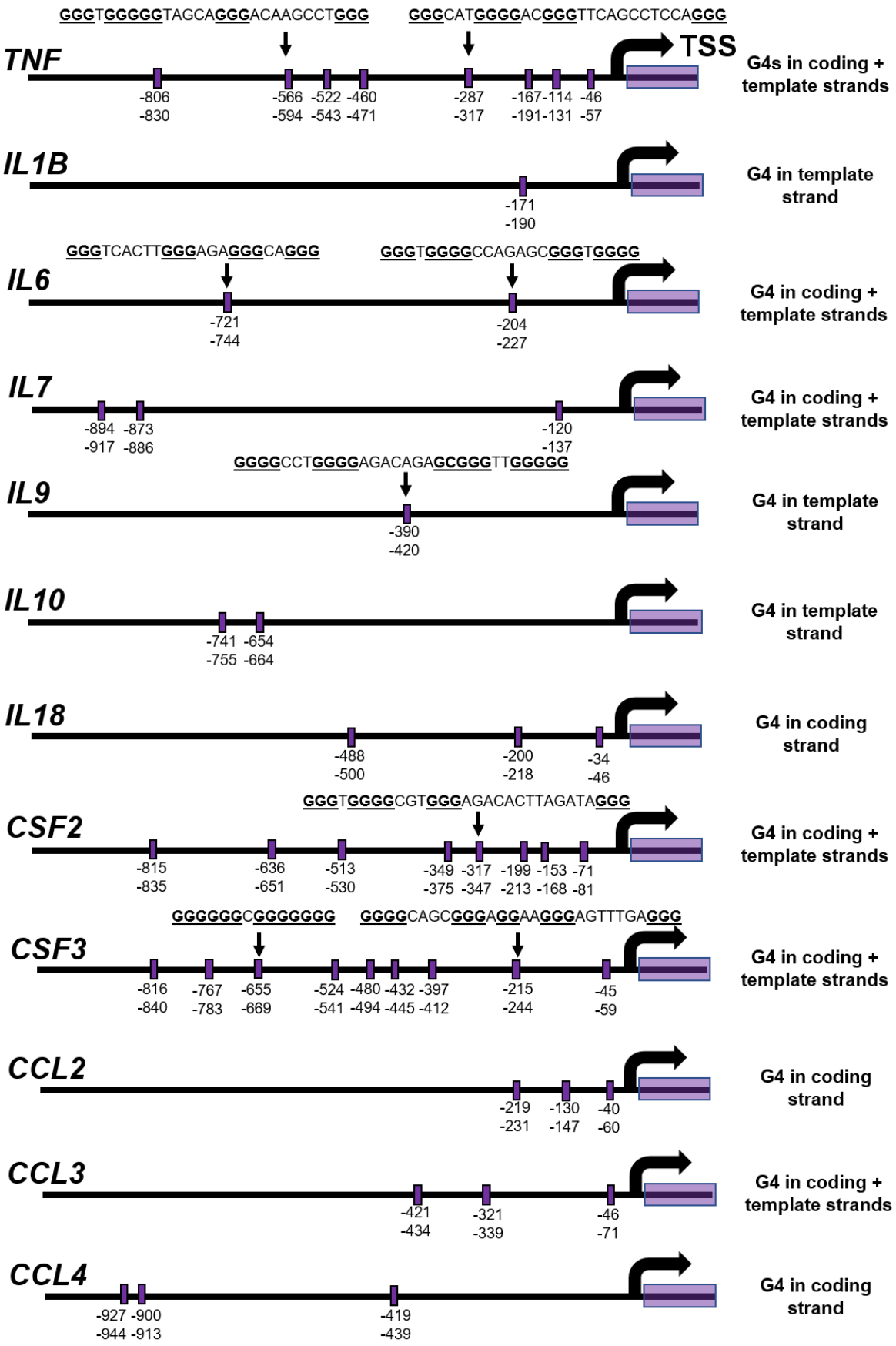
GGGTGGGGGTAGCAGGGACAAGCCTGGG
GGGCATGGGGACGGGTTCAGCCTCCAGGG
TNF
TSS
-806
-830
-566
-594
-522
-543
-460
-471
-287
-317
-167
-191
-114
-131
-46
-57
G4s in coding + template strands
IL1B
-171
-190
G4 in template strand
GGGTCACTTGGGAGAGGGCAGGG
GGGTGGGGCCAGAGCGGGTGGGG
IL6
-721
-744
-204
-227
G4 in coding + template strands
IL7
-894
-917
-873
-886
-120
-137
G4 in coding + template strands
GGGGCCTGGGGAGACAGAGCGGGTTGGGGG
IL9
-390
-420
G4 in template strand
IL10
-741
-755
-654
-664
G4 in template strand
IL18
-488
-500
-200
-218
-34
-46
G4 in coding strand
GGGTGGGGCGTGGGAGACACTTAGATAGGG
CSF2
-815
-835
-636
-651
-513
-530
-349
-375
-317
-347
-199
-213
-153
-168
-71
-81
G4 in coding + template strands
GGGGGGCGGGGGGGG
GGGGCAGCGGGAGGAAGGGAGTTTGAGGG
CSF3
-816
-840
-767
-783
-655
-669
-524
-541
-480
-494
-432
-445
-397
-412
-215
-244
-45
-59
G4 in coding + template strands
CCL2
-219
-231
-130
-147
-40
-60
G4 in coding strand
CCL3
-421
-434
-321
-339
-46
-71
G4 in coding + template strands
CCL4
-927
-944
-900
-913
-419
-439
G4 in coding strand

Figure 2. The location of PQS in the promoters, coding strands, and template strands of genes involved in cytokine storm. The location of sequences which were most likely to form G4s based upon their QGRS Mapper scores, number of G tracts, and loop length in the promoters of genes involved in cytokine storm are highlighted (purple boxes). This location is noted relative to the transcription start site (TSS). The identity of sequences with the potential to form G4s comprised of ≥ 3 G-tetrads in *TNF, IL6, IL9, CSF2,* and *CSF3* are shown. Additionally, whether there are PQS with the potential to form G4s comprised of ≥ 3 G-tetrads in the coding and template strands are noted.

Interestingly, it was recently identified that the DEAD-box RNA helicases, DDX3 and DDX5, can interact with G4s in RNA [10]. Helicases are enzymes which are vital to processes involving nucleic acids and can unwind G4 secondary structures. There is currently a paucity of information describing the effect of stabilising G4s in immune cells, however, knocking out the function of DDX3 in the THP-1 monocytic cell line resulted in the decreased mRNA and protein expression of CCL1, CCL2, CCL5, CXCL10, and TNF-α following stimulation with LPS or poly(I·C) [11]. Furthermore, genetic deletion of *DDX5* in mice was linked to reduced expression of *IL1B* and *TNF*, and reduced inflammation in an experimental model of colitis [12]. Thus, the inability of helicases to resolve G4s and their presence in inflammatory genes could be associated with reduced inflammation and may also be an interesting hypothesis to explore.

Recent observations in a liposarcoma cell line demonstrated that global G4 stabilisation predominantly enhanced transcription. The same paper identified that expression of *IL1B, IL6, IL7, CXCL8, FGF1, CSF2, CSF3, CXCL10, CCL2, TNF,* and *VEGFA* were upregulated, whilst *IL18* and *PDGF* were downregulated [13]. Another RNA-seq study conducted in two adenocarcinoma cell lines did not parallel these observations and the effect of G4 stabilisation on transcription would appear to be cell specific [14]. Moreover, the effects on transcription are also likely to be related to the G4 stabilising drug used (e.g. general binder or target specific) and the environmental conditions (e.g. homeostatic vs. inflammatory). An additional limitation to the interpretation of these results is that translation of these genes was not considered, and gene expression doesn't always correlate with protein expression. If indeed G4 stabilisation is instead associated with increased expression of cytokines, developments in the generation of

small molecules that disrupt G4 formation could represent a novel means in which to decrease their expression [15].

5. Conclusions

G4s have demonstrable roles within transcription and may play important roles in the regulation of gene/protein expression of inflammatory mediators. However, there are many questions which must first be asked, such as if the predicted sequences form quadruplexes under physiological/pathophysiological conditions, or whether G4 stabilisation/destabilisation can modulate the gene/protein expression of cytokines and chemokines. If G4s are involved in the transcription/translation of inflammatory mediators, they could represent a novel drug target with significant implications for the treatment of inflammatory and infectious diseases.

Acknowledgements and Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Conflicts of interest

The author discloses no conflicts of interest

References

1. Spiegel, J., S. Adhikari, and S. Balasubramanian, *The Structure and Function of DNA G-Quadruplexes.* Trends Chem, 2020. **2**(2): p. 123-136.
2. Varshney, D., et al., *The regulation and functions of DNA and RNA G-quadruplexes.* Nat Rev Mol Cell Biol, 2020. **21**(8): p. 459-474.
3. Tassinari, M., et al., *Selective targeting of mutually exclusive DNA G-quadruplexes: HIV-1 LTR as paradigmatic model.* Nucleic Acids Res, 2020. **48**(9): p. 4627-4642.
4. Chambers, V.S., et al., *High-throughput sequencing of DNA G-quadruplex structures in the human genome.* Nat Biotechnol, 2015. **33**(8): p. 877-81.

5. Arcondéguy, T., et al., *VEGF-A mRNA processing, stability and translation: a paradigm for intricate regulation of gene expression at the post-transcriptional level.* Nucleic Acids Res, 2013. **41**(17): p. 7997-8010.
6. Qin, Y., et al., *Characterization of the G-quadruplexes in the duplex nuclease hypersensitive element of the PDGF-A promoter and modulation of PDGF-A promoter activity by TMPyP4.* Nucleic Acids Res, 2007. **35**(22): p. 7698-713.
7. Brázda, V., et al., *G4Hunter web application: a web server for G-quadruplex prediction.* Bioinformatics, 2019. **35**(18): p. 3493-3495.
8. Raiber, E.A., et al., *A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro.* Nucleic Acids Res, 2012. **40**(4): p. 1499-508.
9. Tisoncik, J.R., et al., *Into the eye of the cytokine storm.* Microbiol Mol Biol Rev, 2012. **76**(1): p. 16-32.
10. Herdy, B., et al., *Analysis of NRAS RNA G-quadruplex binding proteins reveals DDX3X as a novel interactor of cellular G-quadruplex containing transcripts.* Nucleic Acids Res, 2018. **46**(21): p. 11592-11604.
11. Ku, Y.C., et al., *DDX3 Participates in Translational Control of Inflammation Induced by Infections and Injuries.* Mol Cell Biol, 2019. **39**(1).
12. Abbasi, N., et al., *DDX5 promotes oncogene C3 and FABP1 expressions and drives intestinal inflammation and tumorigenesis.* Life Sci Alliance, 2020. **3**(10).
13. Lago, S., et al., *Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome.* bioRxiv, 2020: p. 2020.08.27.236778.
14. Marchetti, C., et al., *Targeting Multiple Effector Pathways in Pancreatic Ductal Adenocarcinoma with a G-Quadruplex-Binding Small Molecule.* J Med Chem, 2018. **61**(6): p. 2500-2517.
15. Mitteaux, J., et al., *Disclosing the actual efficiency of G-quadruplex-DNA–disrupting small molecules.* bioRxiv, 2020: p. 2020.11.16.382176.