Does G-Quadruplexes formation elevate the mutation density in its neighborhoods in the human genome?

Panova V.[1*], Alexeevski A.[2,3,] Zvereva M.[4]

[1] Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia, [2]Belozersky Institute of Moscow State University, Moscow, Russia [3]Department of Mathematics, Scientific Research Institute for System Studies, Russian Academy of Sciences [4] Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia

* *nooroka@fbb.msu.ru*

Introduction

The origin of cancerous processes stems from a diverse range of mutations found in both somatic cells and germ lines. It is well known that promoter mutations in oncogenes play a significant role in the development and progression of cancer. Three G-quadruplexes (G4s) are formed in vitro from the hTERT gene promoter in a 96-nt DNA sample, even in the presence of single or double cancer-specific mutations in G-quadruplex (G4). G4 binds to the MutL component of the MMR repair system (Pavlova et al, 2022). This reduces the efficiency of repair of incorrectly paired bases compared to dsDNA

Aim

The study aims to test the hypothesis that the formation of G4 leads to an increase in the frequency of mutations in the neighborhoods of G4. We have already shown this fact for G4 loop-forming sequences for Primates in the set of mammal TERT promoters in (Panova et al., 2023). Furthermore, it has been demonstrated that in a subset of individuals with multiple myeloma, somatic mutations in genomic areas predicted to create G4 structures are more common in tumor plasma cells (Zhuk et al.,2024). Patients with G4 strong context enrichment in their tumors share certain germline SNPs.

Methods and algorithms

To select experimentally observed G-quadruplexes we used data from Marsico et al, 2019. These data are presented as coordinates of 428624 and 1285463 peaks of DNA polymerase (Observed G-Quadruplexes, OQs) stalling in the whole human genome (assembly GRCh37) and have an identifier GSM3003539 and GSM3003540 in GEO, respectively.

We didn't perform calculations on the assembly GRCh38 due to technical reasons.

Samples from GSM3003539 contain K+, which stabilizes G4 structures, and samples from GSM3003540 additionally contain pyridostatin (PDS) added to K+. PDS is a ligand which specifically stabilizes G4-structures. Nevertheless, the method is not free of overpredictions. We have found 3387 OQs with 0% GC content in GSM3003539 and 9840 OQs in GSM3003540.

Methods and algorithms

We used G-quadruplexes found with canonical G4 pattern G_{3+} L_{1-7} G_{3+} L_{1-7} G_{3+} L_{1-7} G_{3+} and with pattern G_{3+} L_{1-5} G_{3+} L_{1-5} G_{3+} L_{1-5} G_{3+}. in experimental peaks.

To each found G4 sequence, we added 10 nt flanks on 5' and 3' ends and considered them G4 neighborhoods.

We used three controls for canonical quadruplexes and two controls (№ 2 and №3) for quadruplexes with loop length = 1-5.

1)Inter-quadruplex sections in the experimental peaks.

2)Inter-peaks sections

3)Filtered inter-peaks sections

For all controls, in each section, the GC composition was calculated in the subsequent segments equal to the mean G-quadruplex length. Segments with GC composition ≥ 50 % were used. This is performed for each chromosome.

To obtain the third control, the inter-peaks data was filtered as follows:

1.Fasta files with segments were created from second control files.

2.In these fasta files, G-quadruplexes are found by pattern.

3.Segments with G-quadruplexes are excluded from the set.

4.In remained fasta files, segments with at least one CG dinucleotide are excluded from the set.

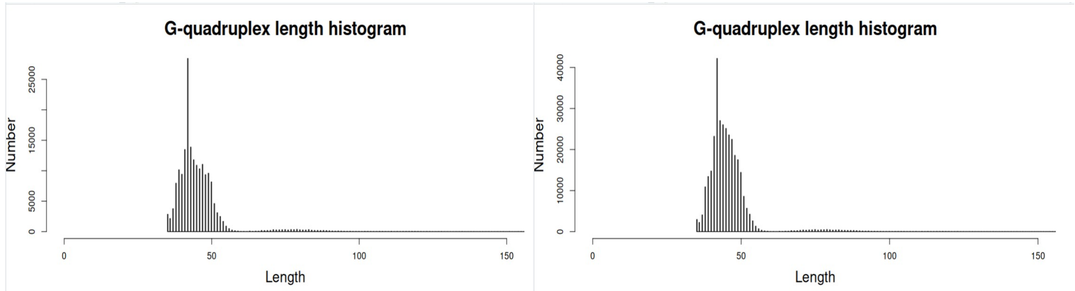
5.For each chromosome, we leave the number of segments with the highest GC-content equal to the number of the quadruplexes for this chromosome.

Methylated CG sites have a high mutation rate. That is why we exclude segments with these sites from the dataset.

Mutations from dbSNP from GRCh37.p13 (all mutations from dbSNP and only one-nucleotide substitutions) and COSMIC non-coding variants (v98) were considered. Mann-Whitney criteria was used to confirm the validity of the results.

Results

The number of considered quadruplexes with loop length 1-7 for assembly GRCh37 was 186490 for GSM3003539 and 325172 for GSM3003540 and with loop length 1-5 was 147956 for GSM3003539 and 223369 for GSM3003540.



GSM3003539 GSM3003540

Distributions of quadruplex lengths (loop length 1-7)

Table 1. Statistical significance that mutation density in the quadruplexes is greater than in control (loop length 1-7)

	Control1			Control2			Control3	
	dbSNP all	dbSNP snp	COSMIC	dbSNP all	dbSNP snp	COSMIC	dbSNP all	dbSNP snp
GSM3003539	1.608e-08	2.273e-08	0.7469	2.276e-10	6.524e-09	0.0001142	9.424e-09	5.716e-10
GSM3003540	2.273e-08	2.087e-07	0.9968	1.608e-08	1.608e-08	9.471e-05	2.276e-10	1.608e-08

Table 2.Statistical significance that mutation density in the quadruplexes is greater than in control (loop length 1-5)

	Control2		Control3	
	dbSNP all	dbSNP snp	dbSNP all	dbSNP snp
GSM3003539	1.787e-10	1.143e-09	1.787e-10	2.882e-10
GSM3003540	1.351e-09	3.217e-08	3.636e-10	9.424e-09

The Snakemake pipeline was developed for automatically processing the experimental data. It takes as an input the experimental and the whole-genome data and outputs the file with mutational density for each chromosome for G4-quadruplexes and control sets, coordinates of G4-quadruplexes and control sequences.

Conclusions

The formation of G4 leads to an increase in the frequency of dbSNP mutations in the neighborhoods of G4. The periodically observed absence of a significant difference between the density of mutations in the first control or a high density of mutations in the inter-quadruplexes is a question for further discussion.

The link to the Snakemake pipeline: https://github.com/nooroka/mutation_density_pipeline_v2

This study was supported by the Russian Science Foundation (project no. 21-14-00161) and by the state budget for scientific research at Lomonosov Moscow State University.