doi 10.18699/bgrs2024-1.1-24

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*

Key words: G-quadruplex; mutation density; mutations; cancer; pattern

Motivation and Aim: 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 specific mutations in G-quadruplex (G4). G4 binds to the MutL component of the MMR repair system. This reduces the efficiency of repair of incorrectly paired bases compared to dsDNA [1].

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 [2]. 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. Patients with G4 strong context enrichment in their tumors share certain germline SNPs [3].

Methods and Algorithms: To select experimentally observed G-quadruplexes we used data from [4]. 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 [5]. 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 [6].

The method from [4] and [6] is developed for G4 detection because G4 formation is known to stall polymerase. Nevertheless, the method is not free of overpredictions. For example, we have found 3387 OQs with 0 % GC content in GSM3003539 and 9840 OQs in GSM3003540. We searched all peak sequences for the presence of canonical G4 patterns $G_{3+}\ L_{1\text{-}7}\ G_{3+}\ L_{1\text{-}7}\ G_{3+}\ L_{1\text{-}7}\ G_{3+}$. To each found G4 sequence we added 10 nt flanks on 5’ and 3’ ends and considered them G4 neighborhoods. As a control set, we used inter-quadruplex sections in the experimental data. The GC composition was calculated in the subsequent sections equal to the mean G-quadruplex length in each chromosome. Sections with GC-composition ≥ 50 % were considered. The number of considered quadruplexes for assembly GRCh37 was 186490 for GSM3003539 and 325172 for GSM3003540. We didn’t perform calculations on the assembly GRCh38, because the

liftover tool (https://genome.sph.umich.edu/wiki/LiftOver) sometimes translates initial coordinates from GRCh37 to CRGh38 improperly.
Mutations from dbSNP from GRCh37.p13 and COSMIC non-coding variants (v98) were taken into account. Mann-Whitney criteria was used to confirm the validity of the results.
Results: 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.
In G4-quadruplexes the mutation density by dbSNP is statistically significantly higher for GRCh37 (Mann – Whitney test gives a p-value, which is significantly lower than 0.05 for the both experimental data sets)
COSMIC data. which contain somatic mutations in cancer are processed now. An article is currently being written on the results obtained.
Conclusion: We found that the formation of G4 leads to an increase in the frequency of dbSNP mutations in the neighborhoods of G4. It is necessary to check whether the accumulation of somatic mutations represented in COSMIC is related to another process.
The link to the Snakemake pipeline: https://github.com/nooroka/mutation_density_pipeline_v2
Funding: 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.

References

1. Pavlova A.V., Savitskaya V.Y., Dolinnaya N.G. et al. G-Quadruplex Formed by the Promoter Region of the hTERT Gene: Structure-Driven Effects on DNA Mismatch Repair Functions. *Biomedicines*. 2022;10(8):1871. doi 10.3390/biomedicines10081871
2. Panova V.V., Dolinnaya N.G., Novoselov K.A. et al. Conserved G-Quadruplex-Forming Sequences in Mammalian *TERT* Promoters and Their Effect on Mutation Frequency. *Life (Basel)*. 2023;13(7):1478. doi 10.3390/life13071478
3. Zhuk A.S., Stepchenkova E.I., Zotova I.V. et al. G-Quadruplex Forming DNA Sequence Context Is Enriched around Points of Somatic Mutations in a Subset of Multiple Myeloma Patients. *Int J Mol Sci*. 2024;25(10):5269. doi 10.3390/ijms25105269
4. Marsico G., Chambers V.S., Sahakyan A.B. et al. Whole genome experimental maps of DNA G-quadruplexes in multiple species. *Nucleic Acids Res*. 2019;47(8):3862-3874. doi 10.1093/nar/gkz179
5. Clough E., Barrett T. The Gene Expression Omnibus Database. *Methods Mol Biol*. 2016;1418:93-110. doi 10.1007/978-1-4939-3578-9_5
6. Chambers V.S., Marsico G., Boutell J.M. et al. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol*. 2015;33(8):877-881. doi 10.1038/nbt.3295