

Original Research Paper

# The Relevance of Bioinformatic Tools in the Study of Polymorphisms of the B4GALNT2 Gene and its Association with Cancer

<sup>1,2</sup>Eddaikra Atika, <sup>1</sup>Haddouche Hayet and <sup>2</sup>Touil-Boukoffa Chafia

<sup>1</sup>Department of Cell Biology and Physiology, Faculty of Nature and Life, SAAD DAHLEB University Blida1, Algeria, Algiers

<sup>2</sup>Laboratory of Cellular and Molecular Research, Cytokines and NO Synthase, Faculty of Biological Sciences (FSB), USTHB, Algiers

## Article history

Received: 25-12-2019

Revised: 11-02-2020

Accepted: 25-02-2020

## Corresponding Author:

Atika Eddaikra

Department of Cell Biology and

Physiology, Faculty of Nature

and Life, SAAD DAHLEB

University Blida1, Algeria,

Algiers; B.P.:270 Route de

Soumâa – Blida, Algeria

Email: aeddaikra@yahoo.fr

**Abstract:** The human gene B4GALNT2 encodes an enzyme ( $\beta$ 1,4-N-acetylgalactosaminyltransferase II) that controls the expression of the blood group Sd<sup>a</sup> carbohydrate's antigen. This gene is located in position 17q21,32 and consists of 11 exons. The characterization and understanding of genetic variation is a real challenge in human genetics, both for healthy individuals and diseased ones. The in silico study of the B4GALNT2 gene's polymorphism using a bioinformatic methodology by means of analyzing various databases and open source web browsers has shown that this gene is characterized by a polymorphic profile that has a very large number of Cosmic SNPs associated with different types of cancer. The prediction of the 3D structure in silico is an important step to better understand the overall architecture of the B4GALNT2 protein. The chosen model this study is one of chondroitin synthase with a recovery percentage of 20.10% relative to the target sequence. Our findings suggest that these cosmic polymorphisms are at the origin of a cellular disorder responsible for the initiation, birth and proliferation of tumors. Bioinformatics has become an indispensable tool in identifying and predicting the function of the B4GALNT2 gene and its relation to cancer.

**Keywords:** B4GALNT2, Cancer, SNP, Polymorphism, Bioinformatics

## Introduction

The human gene Beta-1,4-N-Acetyl-Galactosaminyltransferase II (B4GALNT2) (ID: 124872, OMIM: 111730) is located on the 17q21.32 chromosome (NC\_00017.11). It consists of 11 exons and encodes a glycosyltransferase that catalyzes the last step in the biosynthesis of the human Sd<sup>a</sup> antigen (<https://www.ncbi.nlm.nih.gov/>). GalNAc is the immunodominant glycan of the Sd<sup>a</sup> antigen that is present in more than 90% of red blood cells (Montiel *et al.*, 2003). The B4GALNT2 gene is primarily expressed in the Gastrointestinal (GI) epithelium of the colon, though lower levels of expression can be found in the kidneys, ileum, stomach and rectum (Kawamura *et al.*, 2005).

Single Nucleotide Polymorphisms (SNPs) are one of the most common types of genetic variations in the human genome. Single Nucleotide Polymorphisms

(SNPs) in genes that regulate DNA mismatch repair, cell cycle regulation, metabolism and immunity, are associated with genetic susceptibility to cancer (Deng *et al.*, 2017). The characterization and understanding of genetic variation is a real challenge in human genetics, both for healthy individuals and for patients. SNPs are point variations of a single nucleotide. They are the smallest form of polymorphism because they affect only one base pair. They are distributed throughout the human genome and are the most common form of genetic variation. Indeed, they account for more than 90% of the differences between individuals. In addition, cancer of the digestive tract is a real public health problem. This particular type of cancer combines the involvement of genetic and environmental factors of many genes.

The present in silico study aims to identify the B4GALNT2 gene polymorphisms associated with cancer.

## Materials and Methods

In looking for SNPs and in sorting SNs (snip references) of the B4GALNT2 gene, we adopted an in silico methodology based on bioinformatic tools. The method used was based on the querying of open source databases: NCBI (dbSNP) available on (<http://www.ncbi.nlm.nih.gov/SNP>) (Sherry *et al.*, 2001), UCSC genome browser (<https://genome.ucsc.edu/>) and Ensemble genome browser (<https://www.ensembl.org>). Pathological associations were identified from web browsers: *Gene Nomenclature committee* (HUGO) (<http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/>) and Digest.Net (Piñero *et al.*, 2015).

Concerning the 3D modeling, we used two softwares: SWISS-MODEL (<<https://swissmodel.expasy.org/>), SWISSProt (Bairoch and Apweiler., 1996) and Pymol for 3D visualization.

## Results and Discussion

The identifier of our Alignment request is ID: NM001159387.2. Alignment of the FASTA nucleotide sequence for the variant 2 transcript of the B4GANT2 gene using the NCBI BLAST tool generated a length of 8657 with 338 blast hits against 293 sequences. We retained only those with a similarity score of 99% referring to 4 Blastn hits (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>).

The results of the search for SNPs variations of the B4GALNT2 gene using the UCSC navigator enable us to visualize 180 common SNPs. Results in the RS track refer to synonymous coding SNPs (green), non-

synonymous coding (red), intronic regions (black) and untranslated regions (blue) (Fig. 1).

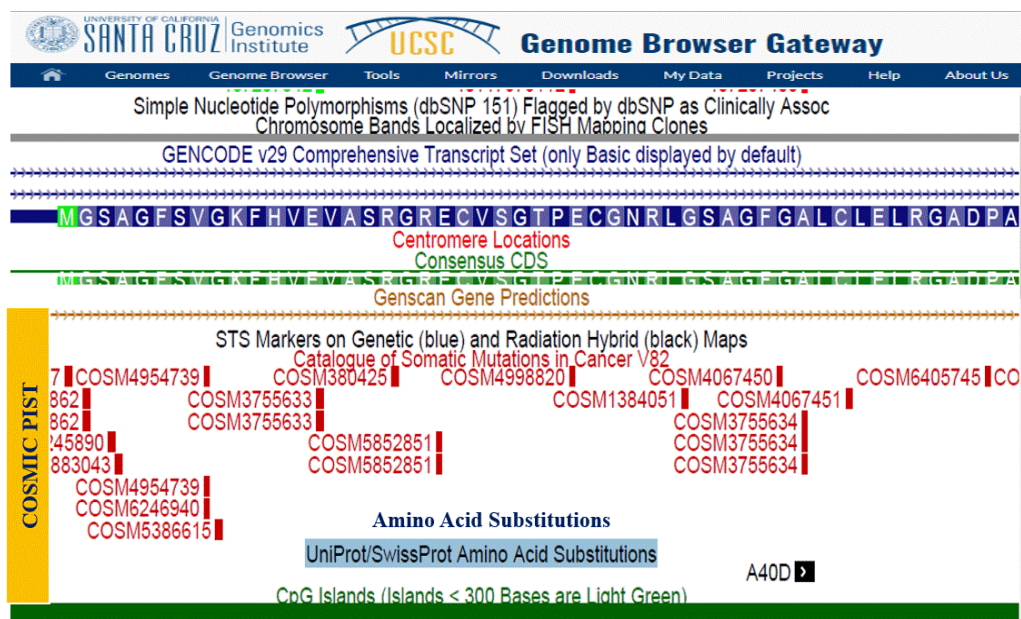
The configuration of the tracks allows us to visualize a cosmic track related to clinical studies on cancer. This track shows that the numbers of accessions in red are distributed over the entire length of the B4GALNT2 gene. Indeed, in recent years, several Genome-Wide Association Studies (GWAS) have identified genetic variants or susceptibility locus associated with various human diseases, including cancer (Stadler *et al.*, 2010).

We also find, at the level of the comic track, a variant A40D which presents the genomic positions of natural and artificial amino acid variants in the UniProt/SwissProt database. Data related to these variants were selected from scientific publications by UniProt.

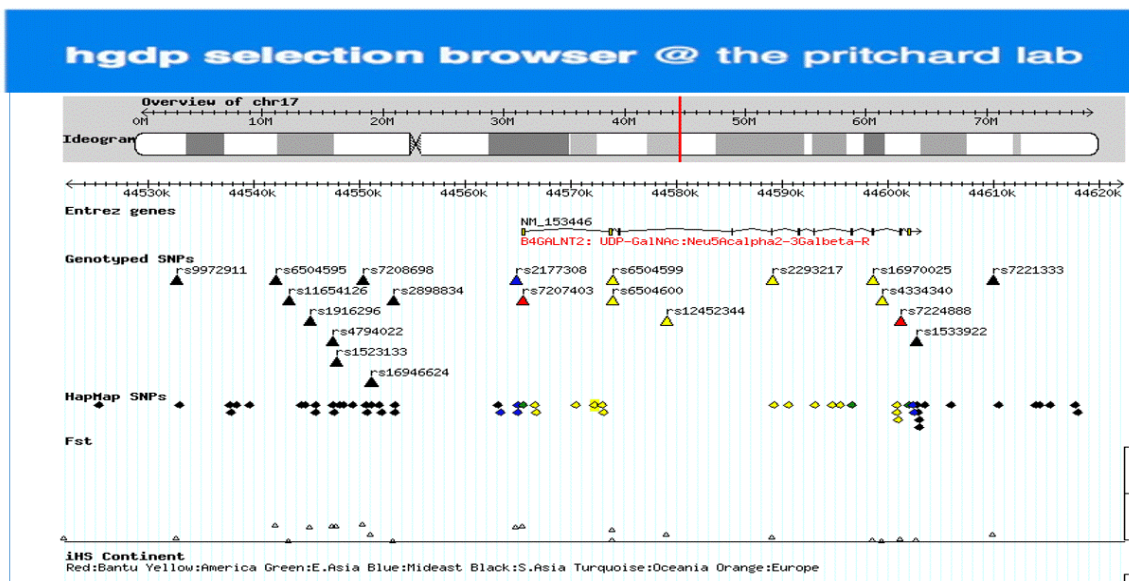
UCSC results show that the B4GALNT2 gene has 18.7% CpG islets. These are presented by a sequence rich in CG dinucleotides.

Our results suggest that CG richness will prevent transcript factors from binding to DNA and thus from transcription. CpG islets are generally common near transcription initiation sites and may be associated with promoter regions. Normally a C base (cytosine) followed immediately by a G base (guanine) (CpG) is rare in vertebrate DNA, since C in this arrangement tends to be methylated and linked to the epigenetic regulation of gene expression (Hanna *et al.*, 2005).

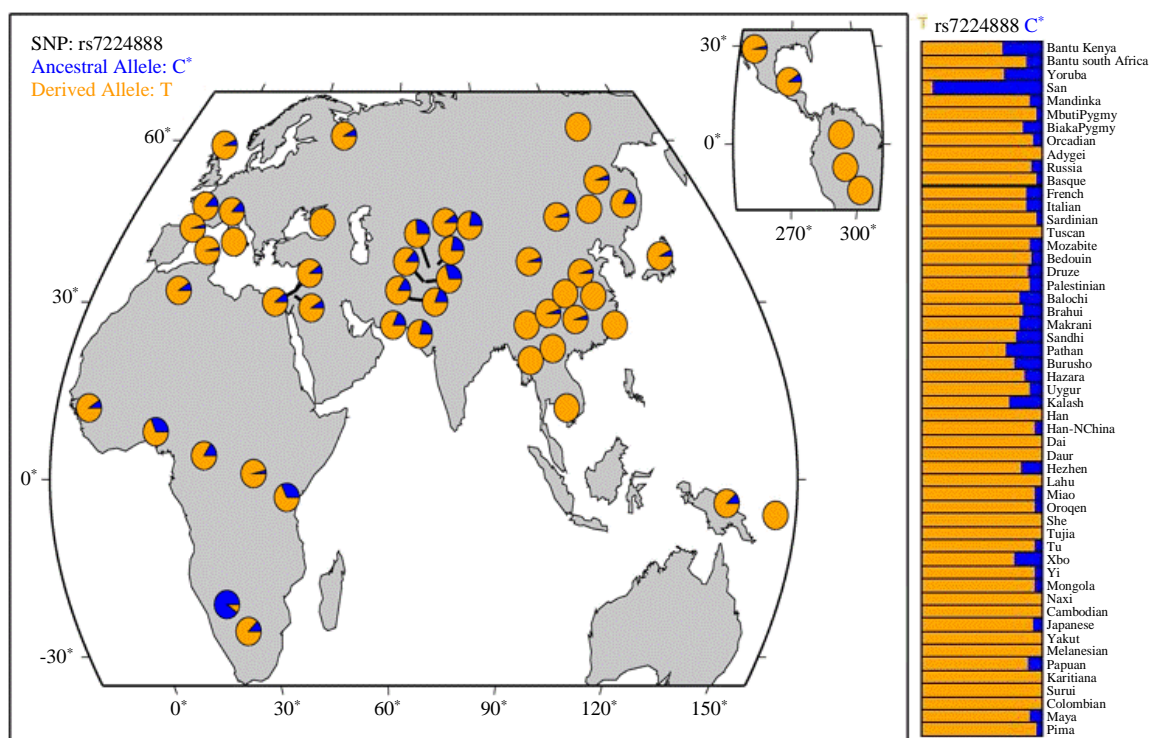
We also note at the level of the tracks generated using UCSC, the existence of a source variant (Uniprot/swissProt) with the identifier: Var\_049238. This variant is a mutation at the coding sequence at position 40 (Ala> Asp): chr17: 49133143-49133145; band17q21.32 genomic size "3" stand +.



**Fig. 1:** Result of the search for SNPs variants using UCSC browser " B4GALNT2 " gene (<http://genome.ucsc.edu/cgi-bin/hgTracks>).



**Fig. 2:** Different SNPs of the HUGO link <http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=44601162&chr=chr17&rs=rs7224888&imp=false>



**Fig. 3:** Distribution of ancestral alleles and alleles derived from rs 7224888

The 40D variant generated the snip reference "RS: 7207403". For this same RS the ensemble navigator allowed us to visualize the distribution of genotypic and allelic frequencies in populations. Similarly, the 40D variant allowed us to switch to the following link (<http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/>) (Fig. 2).

A simple click on RS 7207403 allowed us to visualize the distribution of ancestral alleles and alleles derived from different populations according to geographic location (Fig. 3)

The prediction of the 3D in silico structure is an important step to better understand the overall architecture of this structure. To perform this prediction,

we used the SWISSMODEL browser based on the comparative modeling of B4GALNT2 protein.

In total, 291 models were found to match the target sequence (B4GALT2 protein sequence). This list was filtered by a heuristic method to 50. Our chosen model is that of Chondroitin Synthase with a recovery percentage of 20.10% in relation to the target sequence

(Fig 4 and 5). Variant Q8NHY0: Variant p.Pro459His generated by the SwissModel Navigator ([https://web.expasy.org/variant\\_pages/VAR\\_035990.html](https://web.expasy.org/variant_pages/VAR_035990.html)).

Our findings suggest that this variant was found in a colorectal cancer sample. It is identified as a somatic mutation.

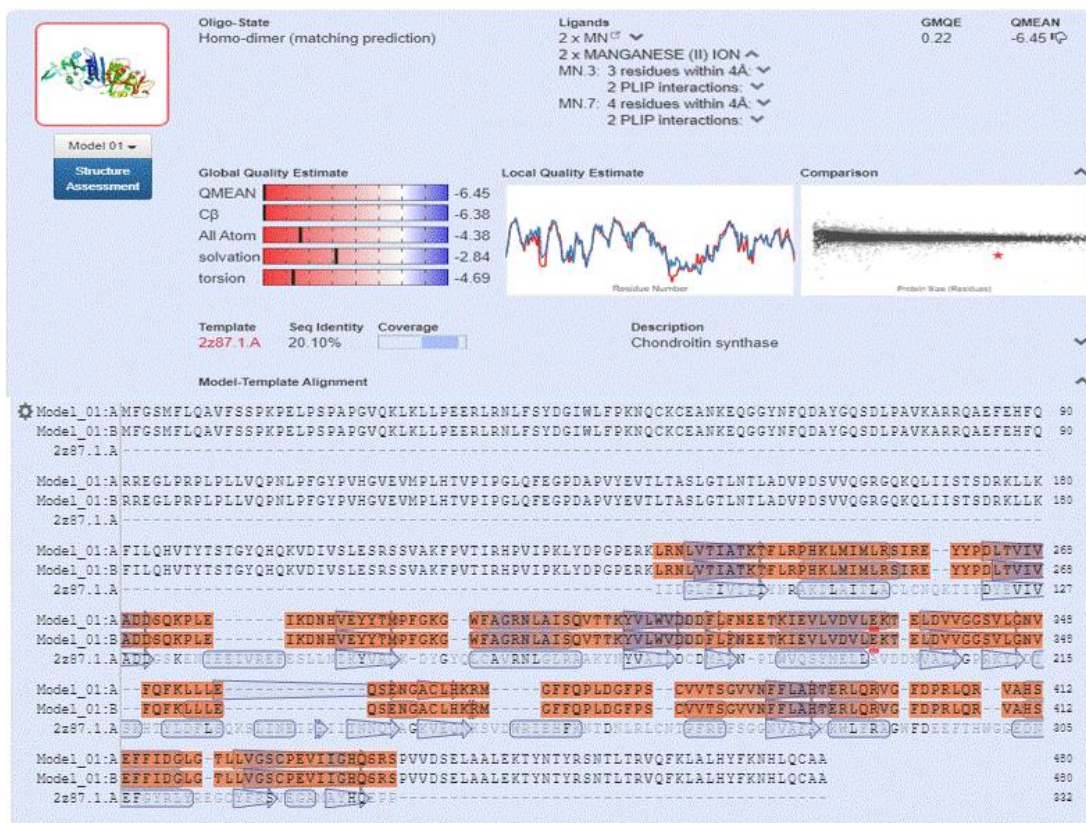


Fig. 4: Comparative modeling of B4GALNT2 protein using SWISSMODEL <https://swissmodel.expasy.org/interactive/6pZRut/models/01>

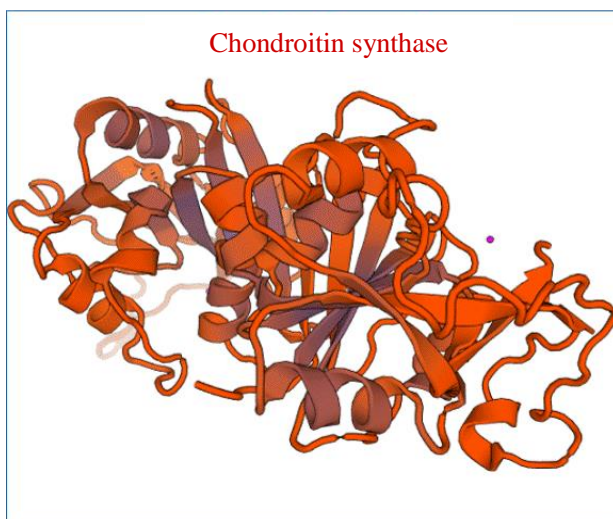


Fig. 5: 3D structure of B4GALNT2 protein generated by PyMol

The examination of the DISGENET database allowed us to identify pathologies associated with B4GALNT2 genes such as colorectal cancer, carcinoma of the stomach, malignant neoplasm and muscular dystrophy. (<https://www.ebi.ac.uk/gwas/search?query=B4GALNT2>).

A study in 2014 reports that, in the absence of cancer, the Sd<sup>a</sup> epitope of the glycan is present. However, the absence of this epitope will upset the balance by causing a new Sd<sup>a</sup> phenotype. In this case, it has been observed that another antigen is overexpressed; sialyl Lewis Antigen (sL<sup>ex</sup>). This antigen has been observed in multiple tumors. In colon cancer, the downregulation of the Sd<sup>a</sup> epitope plays a potential role in the overexpression of Lewis sialyl antigens, increasing the formation of metastases. In addition, it is involved in the lytic function of murine cytotoxic T lymphocytes. and because of this, the expression of Sd<sup>a</sup> antigen has a significant impact on the physiology and pathology of different biological systems (Dall'Olio *et al.*, 2014).

## Conclusion

In light of this work, we can conclude that the cosmic polymorphisms we have dealt with are at the origin of a cellular disorder responsible for the initiation, birth and proliferation of tumors. These polymorphisms can be due to an imbalance of the epigenetic balance (methylation/acetylation). We suggest that the SNPs identified by the web browsers we have used can be considered for case-control studies in order to identify the susceptibility and association of B4GALNT2 gene polymorphism with cancer. They can then be used as genetic markers that will identify the risk exposure to gastrointestinal cancer. Bioinformatics is indeed an essential tool in the identification of SNPs and the prediction of B4GALNT2 gene function and its relation to cancer.

## Author's Contributions

**Eddaikra Atika:** Participated in all experiments, coordinated the data-analysis and contributed to the writing of the manuscript. Designed the research plan and organized the study and coordinated the mouse work.

**Haddouche Hayet:** Participated in all experiments.

**Touil-Boukoffa Chafia:** Coordinated the mouse work.

## Conflict of interest

The authors declare that they have no competitors interests.

## References

- Bairoch, A. and R. Apweiler, 1996. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.*, 24: 21-25.  
DOI: 10.1093/nar/24.1.21
- Dall'Olio, F., N. Malagolini, M. Chiricolo, M. Trinchera and A. Harduin-Lepers, 2014. The expanding roles of the Sda/Cad carbohydrate antigen and its cognate glycosyltransferase B4GALNT2. *Biochim. Biophys. Acta*, 1840: 443-453.  
DOI: 10.1016/j.bbagen.2013.09.036
- Deng, N., H. Zhou, H. Fan and Y. Yuan, 2017. Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget*, 8: 110635-110649.  
DOI: 10.18632/oncotarget.22372
- Piñero, J.P., N. Queralt-Rosinach, A. Bravo, J. Deu-Pons and A. Bauer-Mehren *et al.*, 2015. DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database*.  
DOI: 10.1093/database/bav028
- Hanna, N., B. Parfait, D. Vidaud and M. Vidaud, 2005. Mechanisms and consequences of mutations. *Medecine/Sciences*, 21: 969-80.  
DOI: 10.1051/medsci/20052111969
- Kawamura, Y.I., R. Kawashima, R. Fukunaga, K. Hirai and N. Toyama-Sorimachi *et al.*, 2005. Introduction of Sda carbohydrate antigen in gastrointestinal cancer cells eliminates selectin ligands and inhibits metastasis. *Cancer Res.*, 65: 6220-6227.  
DOI: 10.1158/0008-5472.CAN-05-0639
- Montiel, M.D., M.A. Krzewinski-Recchi, P. Delannoy and A. Harduin-Lepers, 2003. Molecular cloning, gene organization and expression of the human UDP-GalNAc: Neu5Aca2-3Gal $\beta$ -R $\beta$ 1,4-N-acetylgalacto-saminy- transferase responsible for the biosynthesis of the blood group Sda/Cad antigen: Evidence for an unusual extended cytoplasmic domain. *Biochem. J.*, 373: 369-379.  
DOI: 10.1042/bj20021892
- Sherry, S.T., M.H. Ward, M. Kholodov, J. Baker and L. Phan *et al.*, 2001. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.*, 29: 308-311.
- Stadler, Z.K., P. Thom, M.E. Robson, J.N. Weitzel and N.D. Kauff *et al.*, 2010. Genome-wide association studies of cancer. *J. Clin. Oncol.*, 20: 4255-4267.  
DOI: 10.1200/JCO.2009.25.7816