

**Research Article** 

# Explore Selection Patterns in Cancer-Associated Genes in Various Species: The Comparative Analysis.

Sahar Madani<sup>1\*</sup>, Irine Gotsiridze<sup>1</sup>

<sup>1</sup>Faculty of Informatics and Control Systems, Biomedical Engineering department, Georgian Technical University, Tbilisi, Georgia. **Corresponding Author:** Sahar madani, Faculty of Informatics and Control Systems, Biomedical Engineering department, Georgian Technical University, Tbilisi, Georgia.

Received: 🗰 2024 Apl 02

Accepted: 🛱 2024 Apl 15

Published: 🗰 2024 May 23

# Abstract

Cancer has long been a complex and puzzling disease for humans to understand. It is primarily caused by the mutation of two groups of genes known as oncogenes and tumor suppressor genes (TSGs). Certain inherited medical conditions can increase the risk of developing more common forms of thyroid cancer associated with these genes. Uncommon genetic conditions have also been linked to higher rates of thyroid cancer. The genes associated with thyroid cancer are an ideal model for investigating the effects of selection at different stages of thyroid cancer.

A study is currently underway to identify patterns of positive selection in seven highly mutated genes associated with thyroid cancer through a detailed phylogenetic analysis. The nucleotide sequences for these genes were obtained from the Ref Seq database at the National Center for Biotechnology Information (NCBI) and aligned using the Muscle software. The best-fit model for nucleotide substitution was chosen using JMODELTEST, and the corresponding phylogenetic trees were constructed using PAML. The CODMEL software was used to calculate the non-synonymous to synonymous ratio and to select the most plausible evolutionary model for each gene.

The results of the analysis revealed that three genes (BDP1, TG, and TNN) showed strong signatures for positive selection, with p-values of 8.49E-18, 0.002, and 1.75E-87, respectively. The study is ongoing, and the analysis will be extended to the non-coding part to increase the evidence of positive selection on these genes. The aim is to better understand the evolutionary history of thyroid cancer-related genes, enrich the evidence data for the positively selected sites, and improve our understanding of the functionalities of these proteins, domains, and motifs. This could potentially provide valuable insights into drug-targeting sites.

**Keywords:** Positive Selection, Thyroid Cancer, Synonymous, Non-Synonymous, Phylogenetic Tree, Multiple Alignment, Comparative Analysis.

# **1. Introduction**

1. 1 Sequences Data Collection: The detected signals of positive selection within thyroid genes in this study are either determined by ancient mammalian bouts of selection or more recent mammalian group variegation with no clear evidence to suggest any role for sexual selection during the diversification of these genes in mammals. The proteincoding and noncoding sequences for 20 mammalian species used in this study were obtained from NCBI database [The National Center for Biotechnology Information] and the ENSEMBL database. For genes with multiple protein sequences representing spliced forms, we selected the longest sequences as representatives for further analysis. The gene associated with thyroid cancer for mammalian were retrieved from GDC Data Portal (https://portal.gdc. cancer.gov/). The data set included 7 genes including BRAF, TG, BDP1, HMCN1, HRAS, NRAS, and TTN genes for 20 species [1,2].

## 1.2 Phylogenetic Tree Reconstruction.

The multiple sequence alignment (MSA) for all 7 genes was built using the retrieved coding sequences translated to amino acids and further back-translated to nucleotides and muscle implemented in SEAVIEW V4. We reduplicate the multiple sequences alignment in G BLOCKS to reduce the false positive resulting from incorrectly aligned positions. After the filtered MSA we used MrAIC to detect the best evolutionary models from the filtered MSA and we used AICc (Akaike information criterion correction) for the models' comparison. We choose the model with the lowest AIC (2thyroid information criterion) value as it takes into account both the likelihood of the model and the number of parameters contained within a model. PAML is a package of programs for phylogenetic Analyses of DNA and protein sequences using maximum likelihood (ML). PhyML v4 was used to get Phylogenetic gene-based tree reconstructions.

#### **CME Journal of Clinical Case Report**

The phylogenetic tree for the individual thyroid genes was reconstructed using Maximum Likelihood (ML) in MEGA 5.05 and the reliability of the tree branching was assessed using 1,000 replicate bootstraps [4,5].

#### 1.3 Tests of Selection.

Maximum probability is a customary statistical method for estimating unknown parameters of a probability model. A parameter is some descriptor of the model. An acquainted model may be the normal distribution of a population with two parameters: the mean and variance. In phylogenetics, there are many parameters, which include rates, differential transformation costs, and, most importantly, the tree itself. Maximum likelihood is the third method used to build trees. Likelihood provides probabilities of the sequences given a model of their evolution on a particular tree. The more probable the sequences given the tree, the more the tree is preferred. Phylogenetic methods for comparative evaluation of DNA and protein sequences are becoming ever more vital with the rapid accumulation of molecular sequence data, spearheaded by way of numerous genome projects [5-11].

It is now frequent for phylogeny reconstruction to be carried out with the usage of massive record sets involving hundreds or even thousands of genes. Similarly, phylogenetic strategies are widely used to estimate the evolutionary quotes of genes and genomes to observe footprints of natural selection, and the evolutionary information is used to interpret genomic data. For example, both evolutionary conservations indicating terrible purifying selection and accelerated evolution driven via nice Darwinian resolution have been employed to discover functionally good-sized regions of the genome Tests of selection were executed using CODEML within Phylogenetic Analysis by Maximum Likelihood (PAML; v 4.4e). For each thyroid gene, the Likelihoods of the M7 and M8 models were calculated to recognize phylogenetic groups that had experienced positive selection. The M7 model postulates that  $\omega$  values lie between 0 and 1 across the sequences, which is indicative of purifying selection or neutral evolution. The M8 model lets  $\omega$  to override1, which is distinguishing of positive selection. To ensure that relaxed selection was not misinterpreted as positive selection, the likelihood of the null model M8a (where  $\omega$  is fixed at 1) was also compared to that of M8. When codons within an alignment were identified as being under positive selection, the Bayes Empirical Bayes (BEB) method was used to identify the specific codon sites under positive selection. Codon sites with a posterior probability of positive selection higher than 90% were further evaluated based on the nature of the amino acid substitution occurring between species (e.g., cysteine residues influencing the secondary structure and phosphorylation sites that can change the function or localization of a protein). For each gene, the number of nonsynonymous substitutions per nonsynonymous site (dN) and the number of synonymous

#### Copyright © Sahar madani

substitutions per synonymous site (dS) were calculated using a maximum-likelihood method CODEML implemented in PAML v4.6 [7-27].

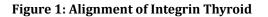
The site model was used to identify positive selection. Which is assessed based on the nonsynonymous/synonymous substitution ratio  $(dn/ds = \omega)$ . If there is no election, the nonsynonymous and synonymous substitution ratio are fixed with the same probability ( $\omega$  =1). Furthermore, because of selective advantage, the selection can increase fixation probabilities for nonsynonymous mutations (positive selection,  $\omega > 1$ ), or because of selective coerce the probabilities are decreased (purifying/negative selection,  $\omega$ < 1). Tests of selection were executed using CODEML within Phylogenetic Analysis by Maximum Likelihood (PAML; v 4). For each thyroid gene, the Likelihoods of the M7 and M8 models were calculated to recognize phylogenetic groups that had experienced positive selection [8]. The M7 model postulates that  $\omega$  values lie between 0 and 1 across the sequences, which is indicative of purifying selection or neutral evolution. The M8 model lets  $\omega$  to override1, which is distinguishing of positive selection. To ensure that relaxed selection was not misinterpreted as positive selection, the likelihood of the null model M8a (where  $\omega$  is fixed at 1) was also compared to that of M8. When codons within an alignment were identified as being under positive selection, the Bayes Empirical Bayes (BEB) method was used to identify the specific codon sites under positive selection [11-25].

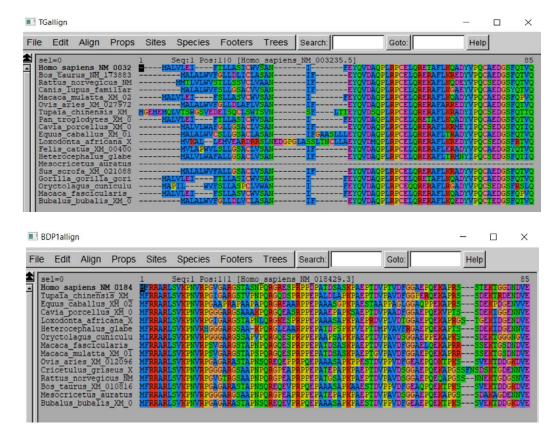
For each gene the number of nonsynonymous substitutions per nonsynonymous site (dN) and the number of synonymous substitutions per synonymous site (dS) were calculated using a maximum-likelihood method CODEML implemented in PAML v4.6. Estimations of dN, dS and dN/ dS, were obtained using six different models (Model 0, 7, 8 and 8a. Model 0 (M0, one-ratio) was used to estimate global dN/dS, dN and dS. Models 7 (M7, beta) and 8 (M8, beta +o > 1), approximate the dN/dS variation over sites through a beta distribution, estimating the proportion and the dN/ dS ratio of the positively selected sites, whereas M8 only includes site-classes above neutrality. Comparisons between models M8 and M8a were used to identify deviations from neutrality. This pairwise comparison focuses on testing whether sites belonging to a site-class with a dN/dS> 1 are evolving differently from near neutrality (dN/dS&1). For each pairwise comparison M7 versus M8, M8 versus M8a, the LRT obtained were compared against a 2 distribution. The degrees of freedom, used to obtain the 2 critical values, were the difference in the number of parameters in the null and alternate model for each pairwise test. For each P value, we also estimated the corresponding q value. When the q value was below, the P value obtained for the LRT value the gene was considered to be under positive selection (1), and when above, the gene was considered negatively selected (0).

#### **CME Journal of Clinical Case Report**

Copyright © Sahar madani

ile Edit /	Align Pro		Choosion	Footers	Troop	Caarabi		Coto	Links	1
		ps ones	opecies	FOULETS	Tiees	Search.		Goto:	Help	J
sel=0		1	Seg:1 Pos	:1 1 [Homo	sapien:	s NM 0012	267550.21			
Homo sapi	ens NM 001	2 TTOAP	FTOPLOSVV	VLEGSTATE	AHISGFP	PEVSWER	DGQVISTSTL	PGVQISFSDGRA	KLTI PAVT KANS	GRYSLKAT
Macaca_mu								PGVQISFSDGRA		
	s_XM_02796							PGVQISFSDGRA		
	inensis_XM							PGVQISFSDGRA		
	s_griseus_							PGVQISFSDGRA		
	s_XM_02497							PGVQISFSDGRA		
	odytes_XM							PGVQISFSDGRA		
Equus_cab	allus XM (	2 MATQAP						PGVQISFSDGRA		
Cavia_por	cellus_XM	0 MTTQAP						PGVQISFSDGRA		
Felis_cat	us_XM_0232	5 MITQAP						PGVQISFSDGRA		
Canis_lup								PGVQISFSEGRA		
Heterocepi								PGVQISFSDGRA		
Sus_scrof			TETQPLQSVV	VLEGSTATE	AHVSGEP	PEVSWER	DGQVISISIL	PGVQISFSDGRA	RUITPAVIRANS	GRISLRAI
Mesocrice				T POOL TO						
Rattus no:	orilla gon							PGVQISFSDGRA PGVOISFSDGRA		CRISLAN
	us cunicul							PGVOISFSDGRA		
Macaca fa								PGVOISFSDGRA		
	ubalis XM							PGVOISFSDGRA		





genes for 20 mammalian species. GenBank accession numbers: Human (Homo sapiens), NM\_018429.3; Chimpanzee (Pan troglodytes), XM\_024354641.1 ; Rhesus monkey (Macaca mulata), XM\_028830889.1; Mouse (Mus musculus), NM\_010575.2; Rat (Ratus norvegicus), XM\_017600986; Rabbit (Oryctolagus cuniculus), XM\_01734; Dog (Canis lupus familiaris), XM\_022414800.1; Horse (Equus caballus),XM\_023622070.1; Pig (Sus scrofa), XM\_021075901.1; Cattle (Bos taurus),

XM\_024979882.1; sheep (Ovis\_aries),XM\_027964793.1; (Tupaia chinensis),XM\_027773115; Chinese hamster (Cricetulus griseus),XM\_02742010; horse (Equus caballus),XM\_023622070.1;guinea pig (Cavia porcellus), XM\_023561141.1; cat (Felis catus), XM\_023259509.1;molerat (Heterocephalusglaber), XM\_02124; golden hamster (Mesocricetus auratus), XM\_021233; (Gorilla gorilla), XM\_019; monkey (Macaca fascicularis), XM\_01543241; buffalo (Bubalus bubalis), XM\_025274121.1

#### **1.4 How to Calculate Positive Selection in Genes?**

#### SeaView - Multiplatform GUI for molecular phylogeny

#### Version 5.0

NEW: seaview performs reconcilation between gene and species trees using **Treerecs** NEW: bootstrap support optionally with the "Transfer Bootstrap Expectation" method NEW: trimming-rule to shorten long sequence names in phylogenetic trees NEW: 64-bit version for the MS Windows platform NEW: multiple-tree wind NEW: multiple-tree windows NEW: seaview uses **PHYLIP v3.696** to compute parsimony trees NEW: seaview can be run without GUI using a command line NEW: seaview drives the **PhyML v3.1** program to compute maximum likelihood phylogenetic trees. NEW: seaview drives the **Gblocks** program to select blocks of conserved sites. SeaView is a multiplatform, graphical user interface for multiple sequence alignment and molecular phylogeny. · SeaView reads and writes various file formats (NEXUS, MSF, CLUSTAL, FASTA, PHYLIP, MASE, Newick) of DNA and protein sequences and of phylogenetic trees Seaview drives programs <u>muscle</u> or <u>Clustal Omega</u> for multiple sequence alignment, and also allows to use any external alignment algorithm able to read and write FASTA-formatted files.
 Seaview drives the <u>Gblocks</u> program to select blocks of evolutionarily conserved sites. Seaview computes phylogenetic trees by

 parsimony, using PHYLIP's <u>dnapars/protpars</u> algorithm,
 distance, with N1 or <u>BioN1</u> algorithms on a variety of evolutionary distances,
 maximum likelihood, driving program <u>PhyM1</u> 3.1.

 Seaview can use the <u>Transfer Bootstrap Expectation</u> method to compute the bootstrap support of PhyML and distance trees.
 Seaview uses the <u>Transfer Bootstrap Expectation</u> method to compute the bootstrap support of PhyML and distance trees.
 Seaview uses the <u>Transfer Bootstrap Expectation</u> method to compute the bootstrap support of PhyML and distance trees. SeaView prints and draws phylogenetic trees on screen, SVG, PDF or PostScript files. SeaView allows to download sequences from EMBL/GenBank/UniProt using the Internet.

Screen shots of the main <u>alignment</u> and <u>tree</u> windows. Dialog window to perform <u>Maximum-Likelihood</u> tree-building. On-line <u>help</u> document.Old <u>seaview version 3.2</u>

#### Figure 2 : Aligned Using The Muscle Software As Implemented In The Seaview Package.

We collected nucleotide sequences for seven genes (BDP1, BRAF, HMCN1, HRAS, NRAS, TG and TTN) from NCBI, Ref Seq database. The sequences were then aligned using the Muscle software as implemented in the SEAVEIW package. We selected the best-fit model for nucleotide substitution using JMODELTEST, and then we constructed the corresponding phylogenetic trees using PAML. CODMEL was used for calculating non-synonymous to synonymous ratio and for selecting the most plausible evolutionary model for each gene (0, 7, 8 and 8a).

lelTest 2.1.10 v20160303				- 6
Analysis Results Tool	ls Help About			
v/ML-log				
Dii-onwards D. De spartment of Bios iversity of Vige spartment of Elee iversity of A Co 1: ddarriba8udc.e	arriba, G.L. Taboau chemistry, Genetics 5, 36310 Vigo, Spar ctronics and System pruma, 15071 A Corr es, dposada@uvigo.o	n. IS Ina, Spain.	ada,	
n 29 04:29:44 PS s 10 10.0. arch:	5T 2020 : amd64, bits: 64,	numcores: 20		
Test 2.1.10 v201 ght (C) 2011 D. rogram comes wit	L60303 Darriba, G.L. Tab th ABSOLUTELY NO W	ada. R. Doallo and		
: This program m	may contain errors.	Please inspect res	: carefully.	
"jModelTest Nature Metho	2: more models, no ods 9(8), 772.	R and Posada D. 201 m heuristics and po		
ers\Building9\Des tible versions:  Test will try to	sktop\jmodeltest-2. 20130103 20131022 ;	of compatibility: 1.10\exe\phyml\PhyM 0141009 20141029 20 m anyway, but it mi riled.	501 20151222	
g data file "TTP er of sequences er of sites: 19	8			
	OF LIKELIHOOD SCORE	S WITH PHYML		
ings::				
idate models = 88 mber of substitut cluding models wi cluding models wi cluding models wi mized free parame	s\Building9\Desktop B tion schemes = 11 ith equal/unequal I ith/without a prop ith/without rate va	/jmodeltest-2.1.10/ ase frequencies (+f rtion of invariable riation among sites ution parameters + BIONJ tree	tes (+1) s) (m(at = 4)	
ress::				
	Exec. Time	Total Time	- 1ei.	
-G 0 -G 0 -G 0 	0h:01:01:04 0h:01:53:07 0h:02:06:04 0h:02:09:03 0h:02:17:03 0h:07:74:08	00h:01:03:01 00h:01:55:05 00h:02:08:02 00h:02:10:09 00h:02:19:00 00h:02:26:04	189335.7941 189336.707 189356.707 189356.733 189359.752 189359.7942	Activate Windows
ed another landed for th	8 models (optimized trees			Go to Settings to activate Windows.
	o moders (optimized trees			пх

Figure 3: Selected The Best-Fit Model For Nucleotide Substitution Using Model test

(Windows\system32\o	Completed 2/88 Elapsed time: 1h:10:58	×	Model Test 2.1.10 v20	1160303	Completed 84/88 Elec	used time: 1h:05:41	×	
ers\Building9\D	Reveal activity		File Edit Analysis Resul	ts Tools Help About	r Thread activity			elect all
Testing MPI en Proceed withou	Computing F&t+T+G	Tymorphism (68,37%),			Thread 0:			elect none
Proceed withou	Thread Or	viymorphiish (ob.s/w).	Man PhyML log		Thread 0:			
	Computing TIM3ef+I+G	ances	TIM2eT+G	00011113/105	Computing	TVNef+G	-	wert selection
	Thread 1:		GTR+I TPH luf+T	00h:03:46:00 00h:07:43:04	Thread 1:			Select
	Computing TIM2ef+I+G		TPM2+I TrNef+I	00h:02:42:03 00h:03:30:08	idle			71_3.0
	Thread 2:	requires at least 263 Mo of memory space.	TPN3+I	00h:02:56:01	Thread 2:			YDC
	Computing SYM+1+G	and the second sec	SYM+I F01	00h:01:03:08 00h:01:03:05	idle			Pre-
	Thread 3:	continue ? [Y/n] Y	TIN3ef+G TPM1+I	00h:14:14:05	Thread 3:			le
	Computing IPMS+1+6	using NHI moves)	K80+G	00h:04:56:02 00h:16:10:00	1d1e			FA File
	Thread 4:	a second s	TINLef+I TPN2ut	00h:04:26:07 00h:01:25:08	Thread 4t			ext Document
	Computing TPM2+I+G	0.9851] [Yopology ] 0.9488] [Topology ]	TrNef+G	00h;16;20;03	ıdle			est Document
	Thread Sc	1.9719] [Topology ]	TIN1+I TPM3ut	00h:10:15:09 00h:01:35:04	Thread S:			ent Document
		.3460] [Topology ]	C TIML	00h:00:26:05 00h:01:59:00				pit Document
	Computing TPM1:IIG	1.1792] (Topology ) 1.6678] [Topology ]	HKY	000103131104	Thread b:			ie.
		.0570] [Topology ]	HKY+I MKY+I+G	00h:14:41:03 00h:22:33:06				Andows Batch File
	Computing TrNef+I+G+++	.8/30] [ropology ]	TD92	00h:02:53:02	Ibread /:	TPM2+G		LINLIG2 File
		8292] [Topology ]	KBD SYM+I+G	00h:02:19:02 00h:22:20:04				UNU0054 File
	Computing K80+I+G	mixing five branches	TPHLut TINJef	00h:06:17:01 00h:02:35:01	idle			MACUS_386 File
	thread as		TPH2	ODh:05:53:08	Thread S:			DF File
	Computing JC+I+G	8292] [Topology ]	TVM	00h:07:42:01 00h:09:46:08	idle			pplication
	Thread 9:	.8163) [Topology ] .8124] [Topology ]	TTHZef	00h:04:48:05	Thread 91			-
	Computing GTR+I+G	[ [ PILS. ]	TrNet TDH3	00h106157108 00h108133102	Computing	TDMLef+G		
	Thread 10:	1.8110) [Topology ]	IVNef TVMcf+I	O0h:04:28:07 O0h:12:53:01	Thread 10:			
	Computing TWH+I+G	inizing five branches	TIM2ef+I	00h:13:48:00	idle			- D ×
	Thread 11:		SYM TFM1	00h:02:47:04 00h:09:40:01	Thread 11:			~ (
	Computing TIM3+I+G	.8110] [Topology ]	TIN1ef TINJef+1	00b:09:11:05 00b:16:03:09	Idle			
	Thread 12:	8109] [Topology ]	GTR	00h:12:13:05	Thread 12:			intron ,0
	Computing TIM2+I+G	mizing five branches	TPH 5 TVMet+I+G	00h:15:49:05 01h:04:35:03	Computing	Charles C.		Size
	Thread 13:			200.0010.0002000.0	Thread 13:	31019111	~	3,145 KB
		mizing five branches	Likelihood scores not	available			entrop	5,058 KB
	Computing TPM2uf+1+G				Thread 141		-	2.202 KD
		urrent tree: -895851.810876.		_			-	1.341 KB
	Computing TPMJuf+I+G	COMPANY CREW: -B95851.610870.			idle Thread 151			1,241,02
		is on the most likely tree						
	Computing TIM1+I+G	[1] S. Saman, A. S. S. Saman, A. S. S. Saman, M. Saman, M. Saman, M. S. Saman, and S. Saman, S. S Saman, Saman, Sama			Thread 161			
	And a second sec	op analysis			and the second s			
	Computing MKY+T+G				idle			
	Thread 17:	20/ 500			Thread 17:			
	Computing TPMLuf+I+G				idle			
	Thread 18:	<b>*</b>			Thread 18:			
	Computing IVMef+1+6				1dle			
	Thread 19:	10-c+_ <b>M</b> II			Thread 19:			
				Desktop				
	1			B Documents			1	
	2%			4 items	059			(100 × 100

Figure 4: The Process Of Selecting The Best-Fit Model For Nucleotide Substitution Using Model Test

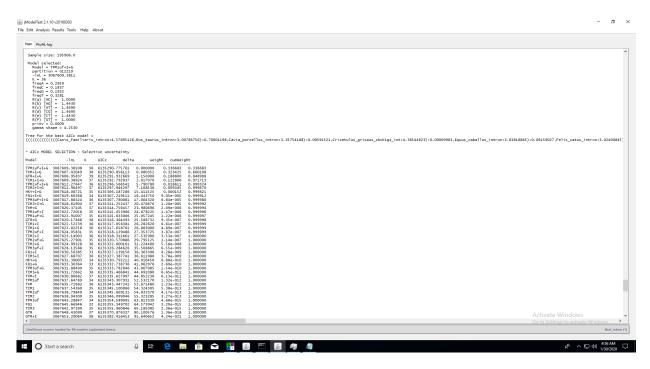


Figure 5: The Process Of Selecting The Best-Fit Model For Nucleotide Substitution Using Model Test.

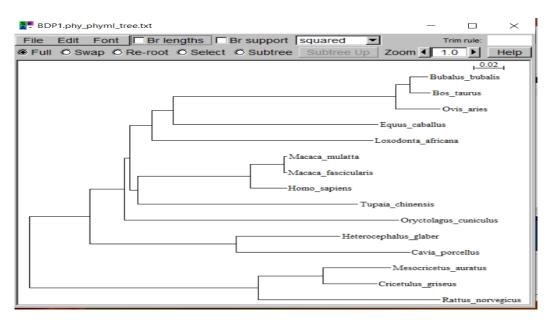
We will try to understand the evolutionary history of thyroid cancer-related genes, and enrich the evidence data for the positively selected site. Understanding the role of positive selection in amino acid substitution will increase our understanding of the functionalities of these proteins, domains and motifs which could provide a recommendation for drug-targeting sites.

Orabity for MAL, splitling for baseder         tog limithods of the covert two: -BRILLBES.         Lanch britting subjets in the math limithy two         me parametric boosting subjets         Image and the second seco
Launch beschrüge statigtet im den meht lättigt bran mer permetrik beschrüge anlights 
Ben parametrik konstatua ankipita 
Ben parametrik konstatua ankipita 
XI         MI           XI         <
Image: Constraint of the second sec
(
19/ 59/ 
[] 149/ 509
[] 39/ 59
VAV 688
300/ 500
. Pointing the most likely tree in file '56gi physi tree.txt'
. Ties und Bh2265s
cruses/suildings/bestspl/hyte_ringsto end
ClDces/BuildegDDextep/ReyR_3.Beeto Execution Histord Foundam Finished
CrusesGuildEgeGeerspile)m (Lispasse Press an Vers Continue
Arts by Africe Contract
Activate Windows
Grito Settingvia activite Windows

Figure 6: The Process Of Selecting The Best-Fit Model For Nucleotide Substitution Using Model Test.

(Windows) system32).cmd.exe		- 0
. Optimise branch lengths : . Optimise substitution model ; . Run ID :		
11         0.0         1         0.0		
O Start a swirch	# 2 # û o # <b>8 = 8 # 4</b>	Activate Windows So is Serreg to activate Windows 요. 고 및 4 2000

Figure 7: Process Of Selected The Best-Fit Model For Nucleotide Substitution Using Model Test.



#### Figure 8: Bdp1 Phy-Phyml-Tree.

Copyright © Sahar madani

ile	Edit	Search	Disp	olay Utilities		
-		🖶 🗧	•1	* 🖻 🖪	🛧 🕕 🔍 🤐 🗱 Aa	Σ
DP1	_mlc	*TG_m		N_mlc		
27	163	9924	F	0.864	1.398 +- 0.264	
27	164	9988	I	0.956*	1.470 +- 0.144	
27	165	9993	A	0.553	1.125 +- 0.436	
27	166	10212	н	0.686	1.237 +- 0.406	
27	167	10235	2	0.723	1.271 +- 0.384	
27	768	10382	v	0.575	1.135 +- 0.444	
27	769	10384	A	0.626	1.199 +- 0.405	
27	770	10393	т	0.929	1.448 +- 0.191	
27	71	10417	A	0.535	1.098 +- 0.454	
27	172	10422	A	0.728	1.286 +- 0.361	
27	173	10428	M	0.735	1.292 +- 0.358	
27	174	10430	т	0.873	1.405 +- 0.255	
27	775	10432	-	0.709	1.227 +- 0.449	
27	176	10433	R	0.994**	1.496 +- 0.052	
27	771	10434	E	0.848	1.384 +- 0.281	
27	778	10435	E	0.981*	1.488 +- 0.091	
27	779	10446	бТ	0.694	1.252 +- 0.387	
27	780	10468	M	0.674	1.233 +- 0.399	
27	781	10488	2	0.706	1.259 +- 0.388	
27	782	10491	H	0.509	1.071 +- 0.460	
27	183	10497	F	0.656	1.210 +- 0.417	
27	784	10501	F	0.810	1.348 +- 0.325	
27	785	10533	I	0.705	1.262 +- 0.382	
27	186	10574	Q	0.693	1.251 +- 0.388	
27	787	10586	5 Y	0.618	1.178 +- 0.427	
27	788	10589	S	0.894	1.422 +- 0.233	
27	199	10500	ч	n 557	1 116 ±_ N //R	>

Figure 9: The Process Of Selection Of The Three Genes Was Identified To Be Under Positive Selection.

#### 3. Results and Discussion

The proteins are molecules that play many critical roles in our body. In our study by used a comparative genomics approach to identify genes that are under positive selection in seven genes of thyroid cancer. We find that positive selection targets a wide range of different functions in the thyroid cancer genes, including cell surface proteins such as the TG gene. Three genes were identified to be under positive selection. According to the Bio Cyc database, the three Genes were assigned to Location and functional categories. Three genes were identified to be under positive selection. According to the Bio Cyc database, the three Genes were assigned to Location and functional categories (Table 1). As in previous studies [10], they find that the proteins that typically undergo positive selection are exposed on the cell surface. Many of these proteins are directly involved in interactions with the host, phages, or other bacteria. In our results may suggest that the finding of TG and TTN in intracellular space and membrane respectively may indicate that a substantive of this positive selection may be related to other adaptive processes than immune/defense evasion and interactions of the cell with phages and other microorganisms, possibly metabolic adaptations in response to a changing environment. Thyroglobulin has been shown to interact with Binding immunoglobulin protein. For that, we strongly suggest that the positive selection acting on the TG gene and TTN gene is caused by selection to avoid phage binding and/or binding of defense-related molecules from the host immune/ defense system. In these genes, the Positive selection looks alike to be related to selection to avoid recognition by a host immune system and/or binding of phages and colicins [23,24].

# 3.1 Location of Positively Selected Sites and Sequence Variation

We plotted the genic location of positively selected sites for the 10 groups that had sites detected from whole-sequence analysis Positively selected sites were not homogeneously distributed among regions; 69% (83 of 116) of sites were located in LRRs. The heterogeneous distribution of positively selected sites was clear from the comparison of the proportion of sites under selection within the NBS and LRR regions, the two domains that occur in all proteins. Twoby-two contingency tests revealed that sites under positive selection occur significantly more frequently in the LRR domains ( $P \ll 0.001$ ;  $\chi^2 = 56.13$ ). Nonetheless, 33 positively selected sites were located in non-LRR regions.

We also studied the distribution of indels across regions in the groups in which positive selection was detected. In a two-by-two contingency test, LRRs had a significantly larger proportion of indels than non-LRR domains ( $P \ll 0.001$ ;  $\chi^2 = 145.14$ ). However, the high incidence of indels in the LRR was not unique to proteins under positive selection. Positively selected sites were not detected in groups 1, 5, 7, and 16, but indels were also more frequent in the LRR than the NBS for these sequence groups ( $P \ll 0.001$ ,  $\chi^2 = 91.36$ ). These observations are important for two reasons. First, the high incidence of gaps in the LRR regions provides additional evidence that LRRs are more labile than other domains. Second, gaps alone do not account for the high incidence of positively selected sites in LRRs.

Table 1 : Positively Selected Positions In Genes.
---

Gene	P-value M7:M8	P-value M8:M8a	Positively Selected Positions
BDP1	1.8739E-24	8.49829E-18	443s,702L,778D,1204F,2185R,2528L
BRAF	0.347221283	0.101829993	
HMCN1	7.26898E-14	#NUM!	
HRAS	#NUM!	#NUM!	
NRAS	#NUM!	0.316215894	
TG	5.4209E-07	0.00248918	9T,1866Q,1492V
TTN	3.8448E-166	1.75045E-87	4086D,4128T,4376L,4393M,4732V,4751S,4774L,5337M, 10433R,10435E,10848R,10862L,11158Q,11371L,12291T, 12292I,12296R,12434T,12499L,12500R,12509H,12510P, 12513R,12520K,12522H,12538P,12555E,12556V,12568K, 12590A,12844A,12901I,12903P,12919K,12932A,12939V, 12963P,12967R,12974A,12985K,13207V,13236T,13258P, 13263P,13351A,13356P,133557E,13358A,13363V,13470V, 13512L,13514P,13524K,13568L,13654P,13825L,13888E, 13893I,13895I,14181T,20378L,22476L,23082R,23785V, 28292I,32516T.

# Table2 : Genes That Show Evidence Of Positive Selection.

Gene name	Gene	Likelihood ratio	dN/dS	Function	Location
thyroglobulin	TG	9.14853	0.29289	making thyroglobulin protein, one of the largest proteins in the body. which found only in the thyroid gland [2].	extracellular space
B double prime 1	BDP1	73.833708	0.56531		Nucleus
Titin	TTN	393.099362	0.11445	making a very large protein named titin which plays an important role in skeletal muscles and in cardiac muscle [3].	cytoskeleton, nucleus, membrane

# Table 3 : Genes Models Test.

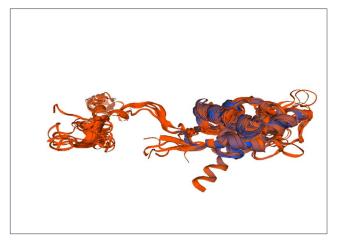
Gene	Model 0 (lnL)	dN/dS	Model 7 (lnL) null model
BDP1	-48574.23142	0.56531	-47934.763
BRAF	-7712.401333	0.05714	-7655.0213
HMCN1	-86441.95578	0.16802	-85331.935
HRAS	-2561.463252	0.01492	-2538.7533
NRAS	-1673.179197	0.0075	-1671.2698
TG	-57284.60195	0.29289	-56128.97
TTN	-512311.6603	0.11445	-501928.08

Gene	Model 7 (Parameters)	Model 8 (lnL) alt model	Model 8 (Parameters)	Model 8a (lnL) null model
BDP1	p = 0.46434 q = -47880.12 0.33840		p0 = 0.92401 p = 0.54738 q = 0.45765 (p1 = 0.07599) w = 2.25355	-47917.04626
BRAF	p = 0.08412 q = -7653.9635 p0 = 0.99778 p = 0.09249 1.15138 q = 1.36260 (p1 = 0.00222) w = 1.90081=		-7655.301804	
HMCN1	CN1 p = 0.33623 q = -85301.6 1.50670		p0 = 0.95879 p = 0.46693 q = 2.76439 (p1 = 0.04121) w = 1.14128	
HRAS	$ \begin{array}{c c} p = & 0.09687 \ q = \\ 4.78581 \end{array} \begin{array}{c} -2538.7551 \\ q = & 4.78581 \end{array} \begin{array}{c} p0 = & 0.99999 \ p = & 0.09687 \\ q = & 4.78581 (p1 = & 0.00001) \\ w = & 7.89021 \end{array} $		q = 4.78581(p1 = 0.00001)	-2537.479488
NRAS	p = 0.01271 q = -1671.2704 $ p = 0.99999 p = 0.01277 $ $ q = 0.48175(p1 = 0.00001) $ $ w = 1.00000$		q = 0.48175(p1 = 0.00001)	-1671.772675
TG	p = 0.37421 q = 0.78368	-56114.542	p0 = 0.95215 p = 0.43836 q = 1.10917(p1 = 0.04785) w = 1.41412	-56119.11664
TTN	p = 0.12951 q = 0.77598	-501547.2	p0 = 0.98656 p = 0.14922 q = 1.04079(p1 = 0.01344) w = 3.79429	-501743.7497

# Table 4 : Genes Models Parameters.

# Table 5 : Lrt (M8 Vs M8a).

Gene	LRT (M7 vs	p-value	LRT (M8 vs	p-value	Positively
	M8)		M8a)		Selected Positions
BDP1	109.268044	1.8739E-24	73.833708	8.49829E-18	
BRAF	2.115586	0.347221283	2.676646	0.101829993	
HMCN1	60.50515	7.26898E-14	-0.683108	#NUM!	
HRAS	-0.0036	#NUM!	-2.551154	#NUM!	
NRAS	-0.001256	#NUM!	1.004534	0.316215894	
TG	28.855668	5.4209E-07	9.14853	0.00248918	
TTN	761.764802	3.8448E-166	393.099362	1.75045E-87	



# Figure 10: Bdp1 Protein Structure

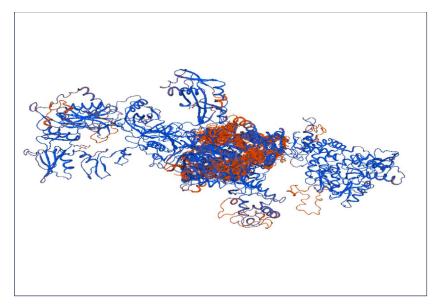


Figure 11:Tg Protein Structure

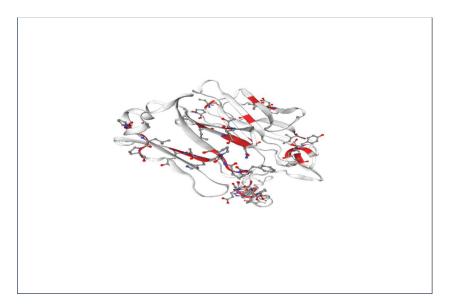


Figure 12: Ttn Protein Structure

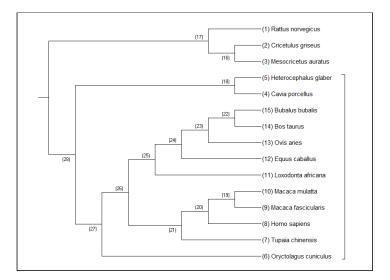
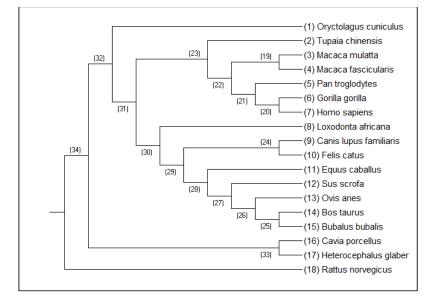
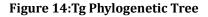
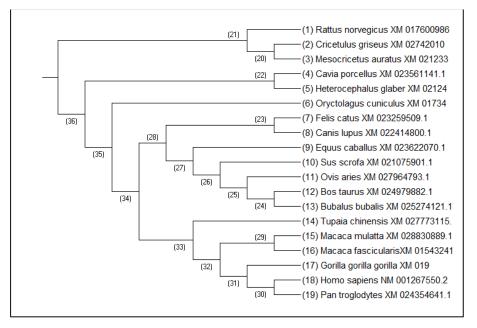
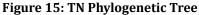


Figure 13:Bdp1 Phylogenetic Tree









#### References

- 1. Yang, Z., Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. Journal of molecular evolution, 46, 409-418.
- 2. Kimura, M. (1983). The neutral theory of molecular evolution. Cambridge University Press.
- 3. Laura, M. (2008). Contrasts in Japanese Society and Culture in Advertisements: 1960s and 1990s (Doctoral dissertation, Gakushuin University).
- 4. W.H.Li, Molecular Evolution, Sinauer Associates, Sunderland, Mass, USA, 1997.
- 5. Koonin, E. V., Wolf, Y. I. (2010). Constraints and plasticity in genome and molecular-phenome evolution. Nature Reviews Genetics, 11(7), 487-498.
- 6. Lynch, M., Conery, J. S. (2000). The evolutionary fate and

consequences of duplicate genes. science, 290(5494), 1151-1155.

- Hughes, A. L., Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature, 335(6186), 167-170.
- Endo, T., Ikeo, K., Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. Molecular biology and evolution, 13(5), 685-690.
- 9. Tsunoyama, K., Gojobori, T. (1998). Evolution of nicotinic acetylcholine receptor subunits. Molecular biology and evolution, 15(5), 518-527.
- 10. Crow, J. F. (2017). An introduction to population genetics theory. Scientific Publishers.
- 11. Onodera, W., Asahi, T., Sawamura, N. (2019). Data for

#### **CME Journal of Clinical Case Report**

positive selection test and co-evolutionary analysis on mammalian cereblon. Data in brief, 26, 104499.

- 12. Ban Y, Greenberg DA, Concepcion E, Skrabanek L, Villanueva R, Tomer Y. Amino acid substitutions in the thyroglobulin gene are associated with susceptibility to human
- 13. and murine autoimmune thyroid disease. Proc Natl Acad Sci U S A. 2003 Dec 9;100(25):15119-24. Epub 2003 Dec 1.
- 14. Targovnik, H. M., Citterio, C. E., Rivolta, C. M. (2011). Thyroglobulin gene mutations in congenital hypothyroidism. Hormone research in paediatrics, 75(5), 311-321.
- 15. Caron, P., Moya, C. M., Malet, D., Gutnisky, V. J., Chabardes, B., et al. (2003). Compound heterozygous mutations in the thyroglobulin gene (1143delC and  $6725G \rightarrow A$ [R2223H]) resulting in fetal goitrous hypothyroidism. The Journal of Clinical Endocrinology & Metabolism, 88(8), 3546-3553.
- González-Sarmiento, R., Corral, J., Mories, M. T., Corrales, J. J., Miguel-Velado, E., et al. (2001). Monoallelic deletion in the 5'region of the thyroglobulin gene as a cause of sporadic nonendemic simple goiter. Thyroid: Official Journal of the American Thyroid Association, 11(8), 789-793.
- Prpić, M., Franceschi, M., Romić, M., Jukić, T., Kusić, Z. (2018). Thyroglobulin as a tumor marker in differentiated thyroid cancer-clinical considerations.

Acta Clinica Croatica, 57(3), 518.

- 18. Yang, Z., Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. Journal of molecular evolution, 46, 409-418.
- 19. Kimura, M. (1983). The neutral theory of molecular evolution. Cambridge University Press.
- 20. Kimura in the 1960s and 70s
- 21. W.H.Li, MolecularEvolution, SinauerAssociates, Sunderland, Mass, USA, 1997.
- 22. Koonin, E. V., Wolf, Y. I. (2010). Constraints and plasticity in genome and molecular-phenome evolution. Nature Reviews Genetics, 11(7), 487-498.
- Lynch, M., Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. science, 290(5494), 1151-1155.
- Hughes, A. L., Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature, 335(6186), 167-170.
- 25. Endo, T., Ikeo, K., & Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. Molecular biology and evolution, 13(5), 685-690.
- 26. Tsunoyama, K., & Gojobori, T. (1998). Evolution of nicotinic acetylcholine receptor subunits. Molecular biology and evolution, 15(5), 518-527.
- 27. Crow, J. F. (2017). An introduction to population genetics theory. Scientific Publishers.