OrthoWeb – web application for macro- and microevolutionary analysis of genes

Zakhar Mustafin[1,2], Alexey Mukhin[1,2], Dmitry Afonnikov[1,2,3], Yury Matushkin[2], Sergey Lashin[1,2,3]

1 – Kurchatov Genomics Center
2 – Institute of Cytology and Genetics
3 – Novosibirsk State University
Novosibirsk, Russia

Study was supported by the RFBR 20-04-00885 grant

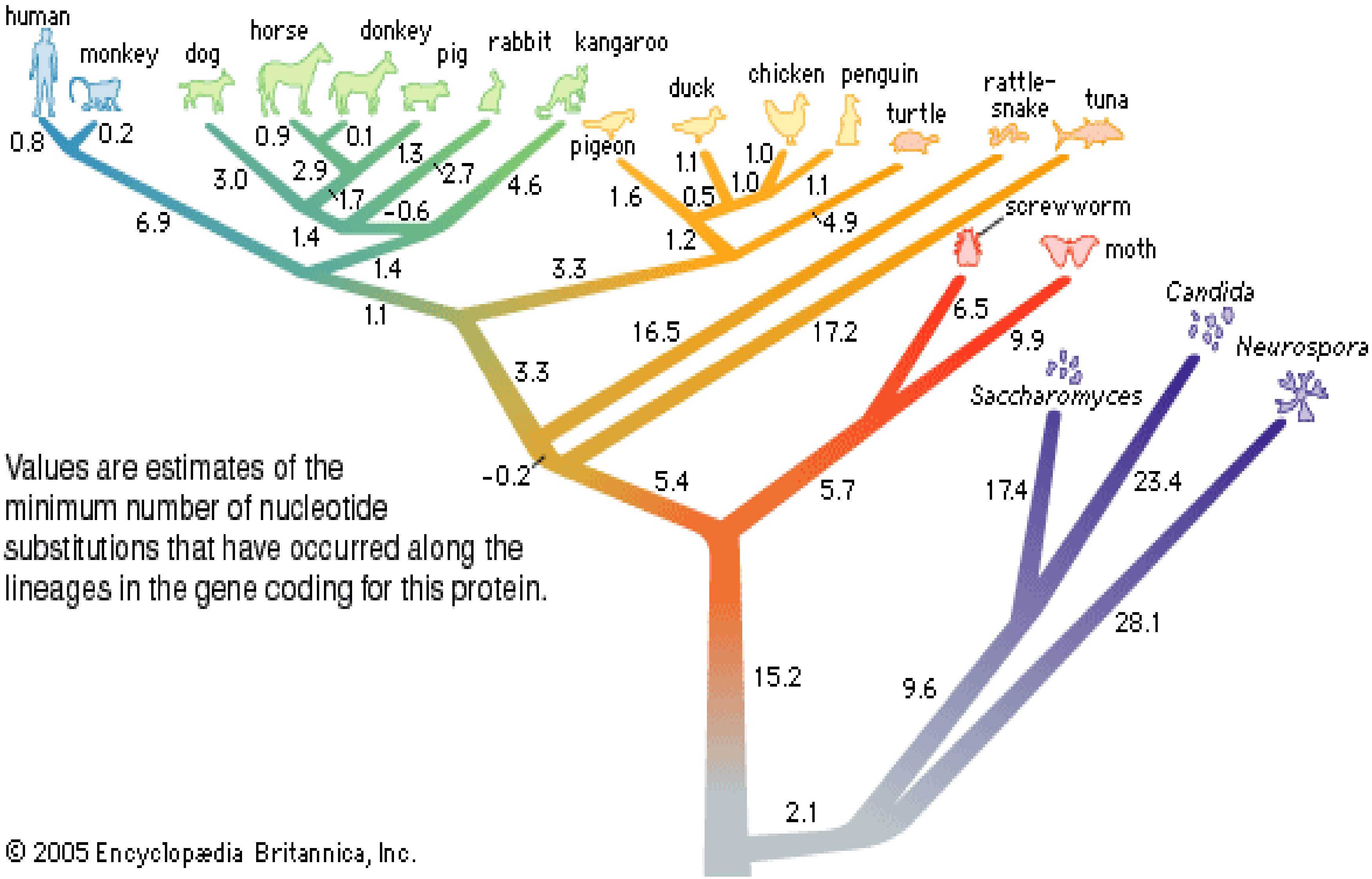
Novosibirsk, 2020

Phylostratigraphic analysis

Tomislav Domazet-Loso and co-authors proposed phylostratigraphic analysis in the beginning of 2000 years[1]. Phylostratigraphic analysis allows to trace evolutionary innovations in genome. The analysis is usually based on two steps:

1) Macroevolutionary analysis. It allows to find "the age of gene" and related characteristics.

2) Microevolutionary analysis. It allows to detect the type of selection affecting to gene.



Values are estimates of the minimum number of nucleotide substitutions that have occurred along the lineages in the gene coding for this protein.

© 2005 Encyclopædia Britannica, Inc.

https://kids.britannica.com/students/assembly/view/121497

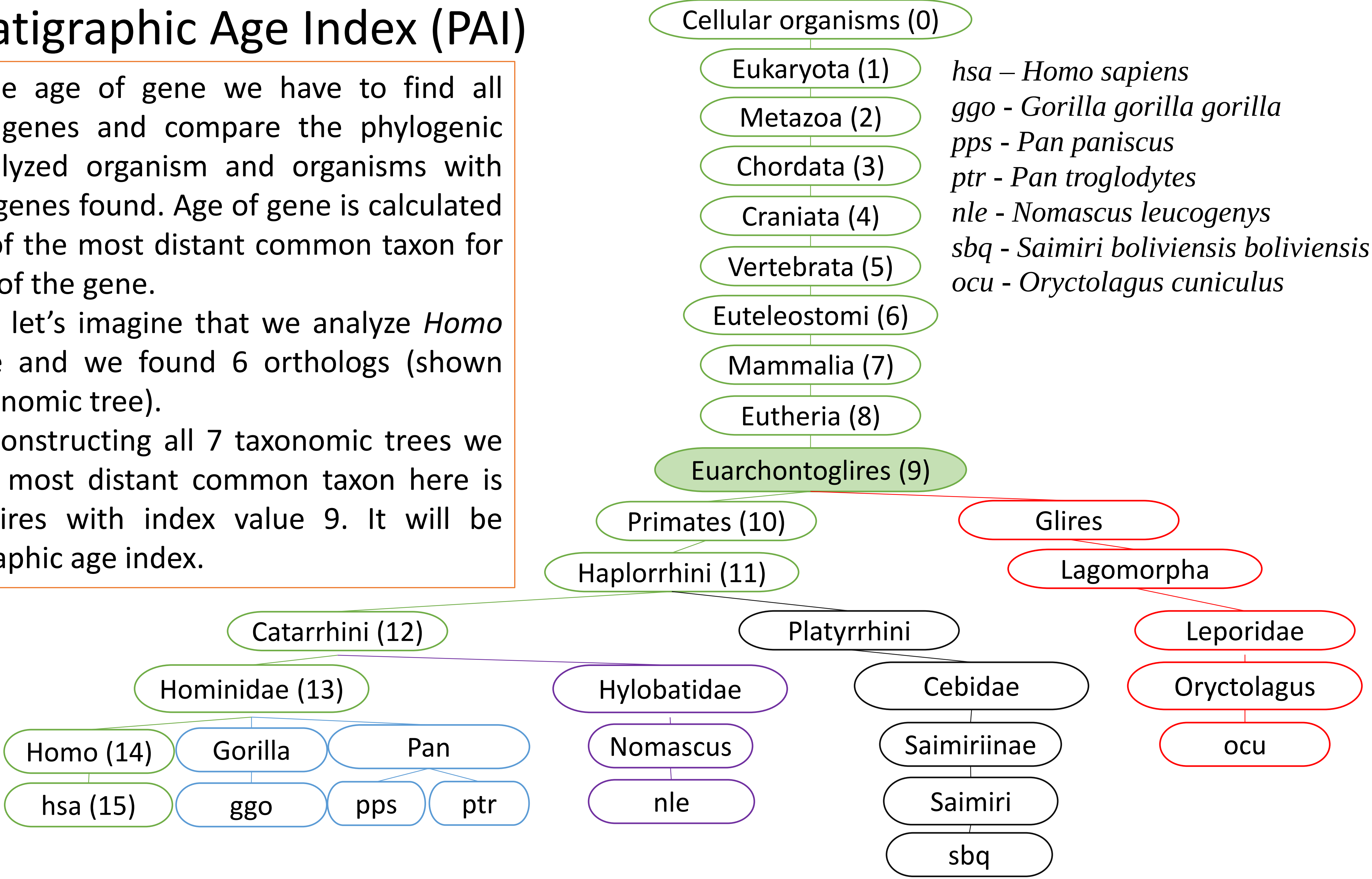
1 - Domazet-Lošo T., Brajković J., Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages // Trends Genet. 2007. T. 23. № 11. C. 533–539.

Phylostratigraphic Age Index (PAI)

To obtain the age of gene we have to find all orthologous genes and compare the phylogenic trees of analyzed organism and organisms with orthologous genes found. Age of gene is calculated as an index of the most distant common taxon for all orthologs of the gene.

For example, let’s imagine that we analyze *Homo sapiens* gene and we found 6 orthologs (shown near the taxonomic tree).

After the reconstructing all 7 taxonomic trees we see that the most distant common taxon here is Euarchontoglires with index value 9. It will be phylostratigraphic age index.



Cellular organisms (0)

Eukaryota (1)

Metazoa (2)

Chordata (3)

Craniata (4)

Vertebrata (5)

Euteleostomi (6)

Mammalia (7)

Eutheria (8)

Euarchontoglires (9)

Primates (10)

Glires

Haplorrhini (11)

Lagomorpha

Catarrhini (12)

Platyrrhini

Leporidae

Hominidae (13)

Hylobatidae

Cebidae

Oryctolagus

Homo (14)

Gorilla

Pan

Nomascus

Saimiriinae

ocu

hsa (15)

ggo

pps

ptr

nle

Saimiri

sbq

hsa – Homo sapiens

ggo - *Gorilla gorilla gorilla*

pps - *Pan paniscus*

ptr - *Pan troglodytes*

nle - *Nomascus leucogenys*

sbq - *Saimiri boliviensis boliviensis*

ocu - *Oryctolagus cuniculus*

Divergence Index (DI)

$$DI = \left(\sum_{ort} dN/dS\right)/N$$

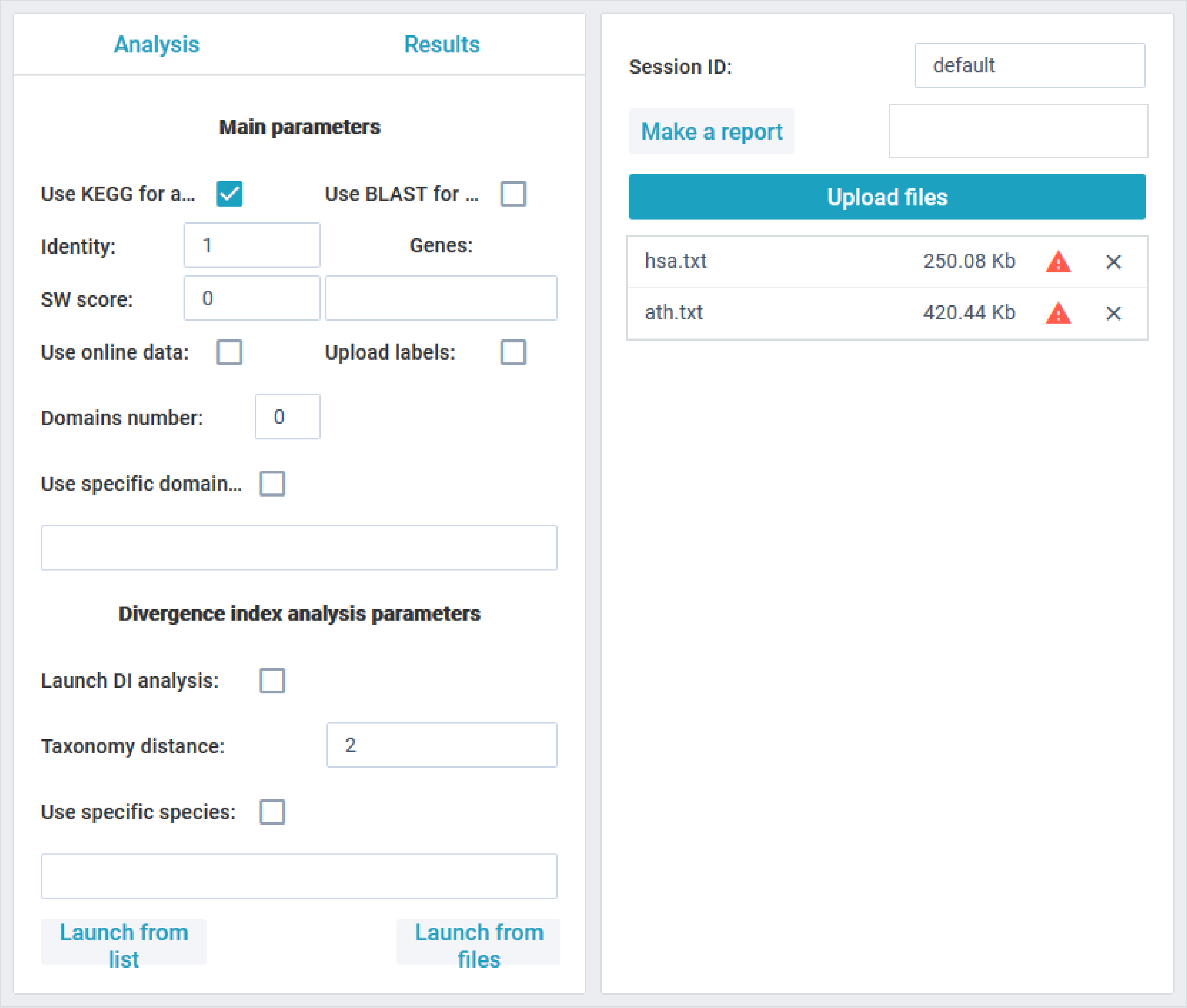
, ort – the closest orthologous gene of current specie,
N – the number of species (with orthologous found)

1)	AUG	AAC	GGG	GUU	AAC	AAC	UGA
2)	AUG	AAC	GGA	GUG	AAU	AAA	UGA
aminoacid:	start/M	N	A->A	V->V	N->N	N->K	stop

Detecting of the type of selection is based on dN/dS ratio calculation. The main idea is to compare the number of synonymous (dS) and nonsynonymous (dN) substitutions in sequences of gene of analyzed organism and most evolutionary close orthologous genes. The ratio value above 1 indicates the evolution of the gene under positive Darvinian selection. The ratio value close to 1 indicates that a gene evolves under neutral regime. The values close to 0 indicate strong purifying selection acting on a gene.

DI is based on average dN/dS value in comparison of the closest relatives organisms.

OrthoWeb



Analysis

Results

Main parameters

Use KEGG for a... ☑

Use BLAST for ... □

Identity: 1

Genes:

SW score: 0

Use online data: □

Upload labels: □

Domains number: 0

Use specific domain... □

Divergence index analysis parameters

Launch DI analysis: □

Taxonomy distance: 2

Use specific species: □

Launch from list

Launch from files

Session ID: default

Make a report

Upload files

hsa.txt	250.08 Kb	⚠	×
ath.txt	420.44 Kb	⚠	×

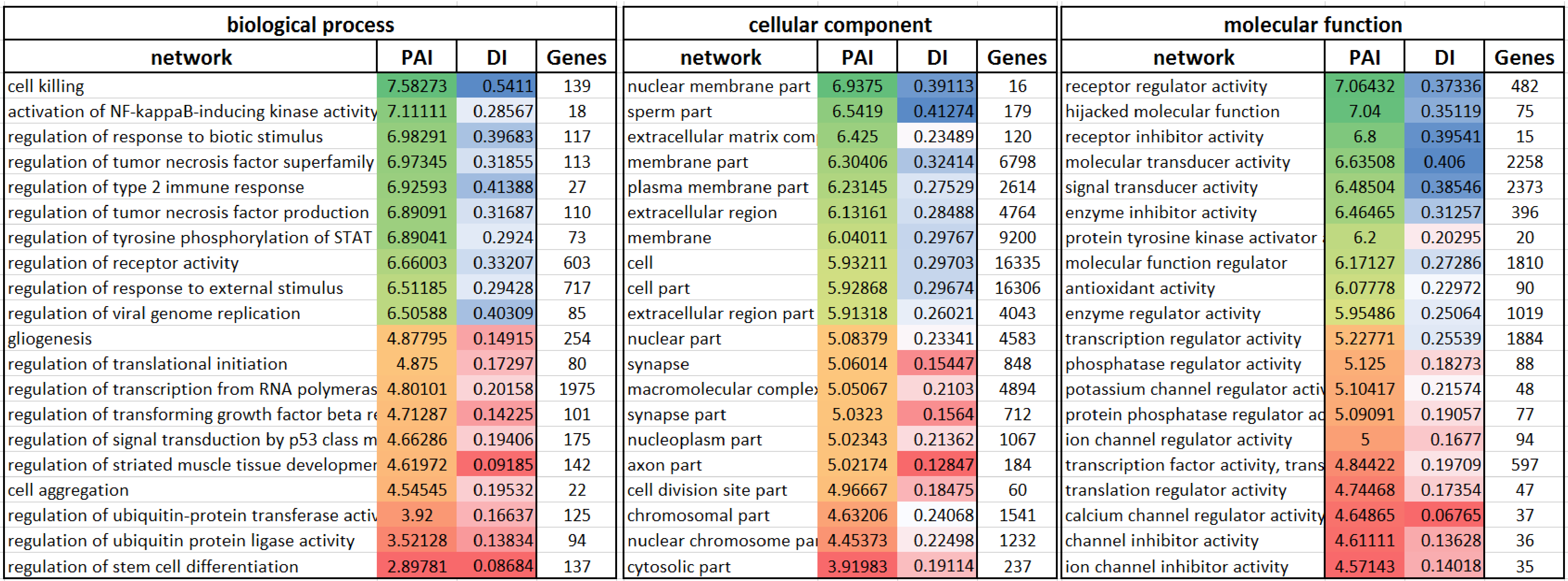
OrthoWeb is a web application, developed by using Java language with Spring and Webix Frameworks.

It allows to count the Phylostratigraphic Age Index and Divergence Index of gene. There are two ways to find the orthologous genes implemented in OrthoWeb: by using KEGG database and by using BLAST software.

It uses MongoDB to store the local data to increase the speed of future analysis.

Available at https://orthoweb.sysbio.cytogen.ru/

Gene ontology analysis



biological process			
network	PAI	DI	Genes
cell killing	7.58273	0.5411	139
activation of NF-kappaB-inducing kinase activity	7.11111	0.28567	18
regulation of response to biotic stimulus	6.98291	0.39683	117
regulation of tumor necrosis factor superfamily	6.97345	0.31855	113
regulation of type 2 immune response	6.92593	0.41388	27
regulation of tumor necrosis factor production	6.89091	0.31687	110
regulation of tyrosine phosphorylation of STAT	6.89041	0.2924	73
regulation of receptor activity	6.66003	0.33207	603
regulation of response to external stimulus	6.51185	0.29428	717
regulation of viral genome replication	6.50588	0.40309	85
gliogenesis	4.87795	0.14915	254
regulation of translational initiation	4.875	0.17297	80
regulation of transcription from RNA polymeras	4.80101	0.20158	1975
regulation of transforming growth factor beta re	4.71287	0.14225	101
regulation of signal transduction by p53 class m	4.66286	0.19406	175
regulation of striated muscle tissue developmer	4.61972	0.09185	142
cell aggregation	4.54545	0.19532	22
regulation of ubiquitin-protein transferase activ	3.92	0.16637	125
regulation of ubiquitin protein ligase activity	3.52128	0.13834	94
regulation of stem cell differentiation	2.89781	0.08684	137

cellular component			
network	PAI	DI	Genes
nuclear membrane part	6.9375	0.39113	16
sperm part	6.5419	0.41274	179
extracellular matrix com	6.425	0.23489	120
membrane part	6.30406	0.32414	6798
plasma membrane part	6.23145	0.27529	2614
extracellular region	6.13161	0.28488	4764
membrane	6.04011	0.29767	9200
cell	5.93211	0.29703	16335
cell part	5.92868	0.29674	16306
extracellular region part	5.91318	0.26021	4043
nuclear part	5.08379	0.23341	4583
synapse	5.06014	0.15447	848
macromolecular complex	5.05067	0.2103	4894
synapse part	5.0323	0.1564	712
nucleoplasm part	5.02343	0.21362	1067
axon part	5.02174	0.12847	184
cell division site part	4.96667	0.18475	60
chromosomal part	4.63206	0.24068	1541
nuclear chromosome pai	4.45373	0.22498	1232
cytosolic part	3.91983	0.19114	237

molecular function			
network	PAI	DI	Genes
receptor regulator activity	7.06432	0.37336	482
hijacked molecular function	7.04	0.35119	75
receptor inhibitor activity	6.8	0.39541	15
molecular transducer activity	6.63508	0.406	2258
signal transducer activity	6.48504	0.38546	2373
enzyme inhibitor activity	6.46465	0.31257	396
protein tyrosine kinase activator	6.2	0.20295	20
molecular function regulator	6.17127	0.27286	1810
antioxidant activity	6.07778	0.22972	90
enzyme regulator activity	5.95486	0.25064	1019
transcription regulator activity	5.22771	0.25539	1884
phosphatase regulator activity	5.125	0.18273	88
potassium channel regulator activ	5.10417	0.21574	48
protein phosphatase regulator ac	5.09091	0.19057	77
ion channel regulator activity	5	0.1677	94
transcription factor activity, trans	4.84422	0.19709	597
translation regulator activity	4.74468	0.17354	47
calcium channel regulator activity	4.64865	0.06765	37
channel inhibitor activity	4.61111	0.13628	36
ion channel inhibitor activity	4.57143	0.14018	35

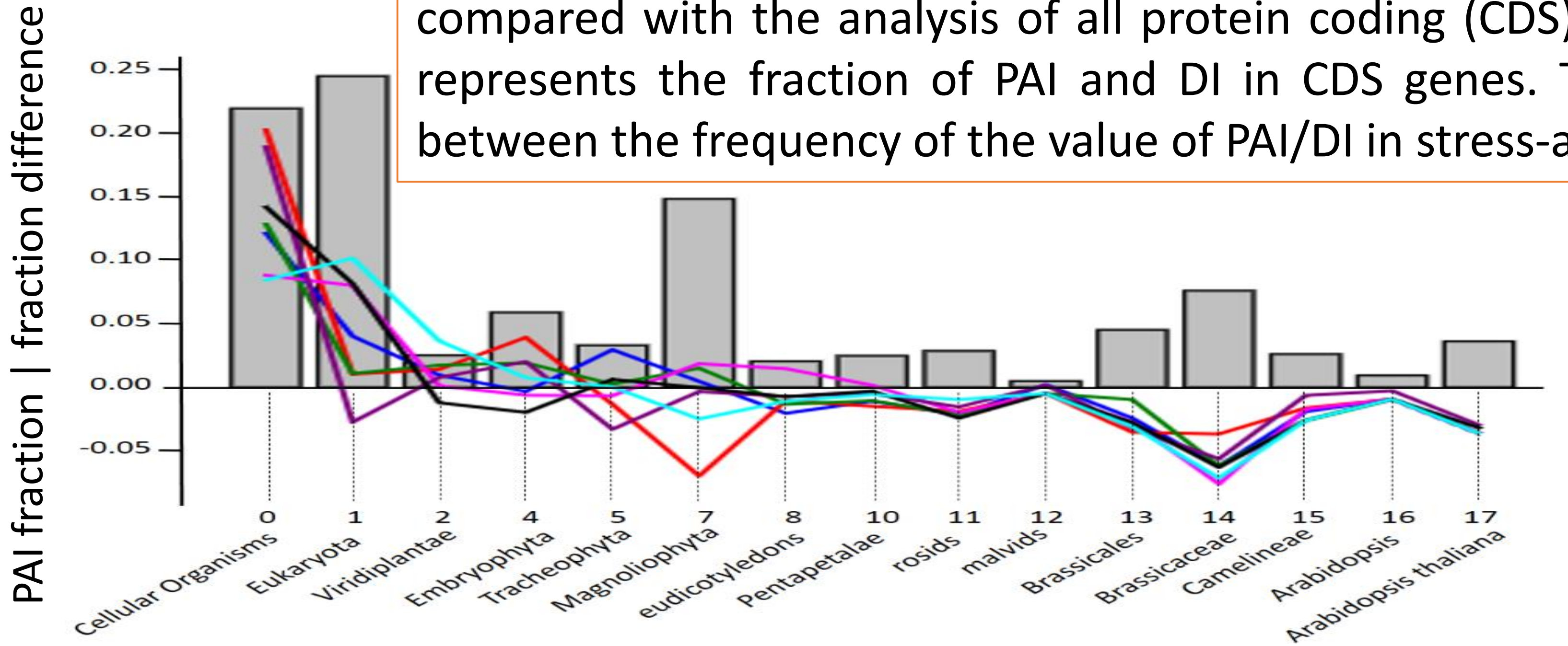
«Young» - immune system, virus response, membrane/receptors

«Old» – stem cells, ubiquitin related, intracellular structures, transcription, translation

OrthoWeb has been used to count PAI and DI of genes associated with different gene ontology terms of *Homo sapiens*. **The main conclusion** is that the intracellular processes mostly associated with old genes and extracellular process mostly associated with young genes.

Stress associated genes analysis

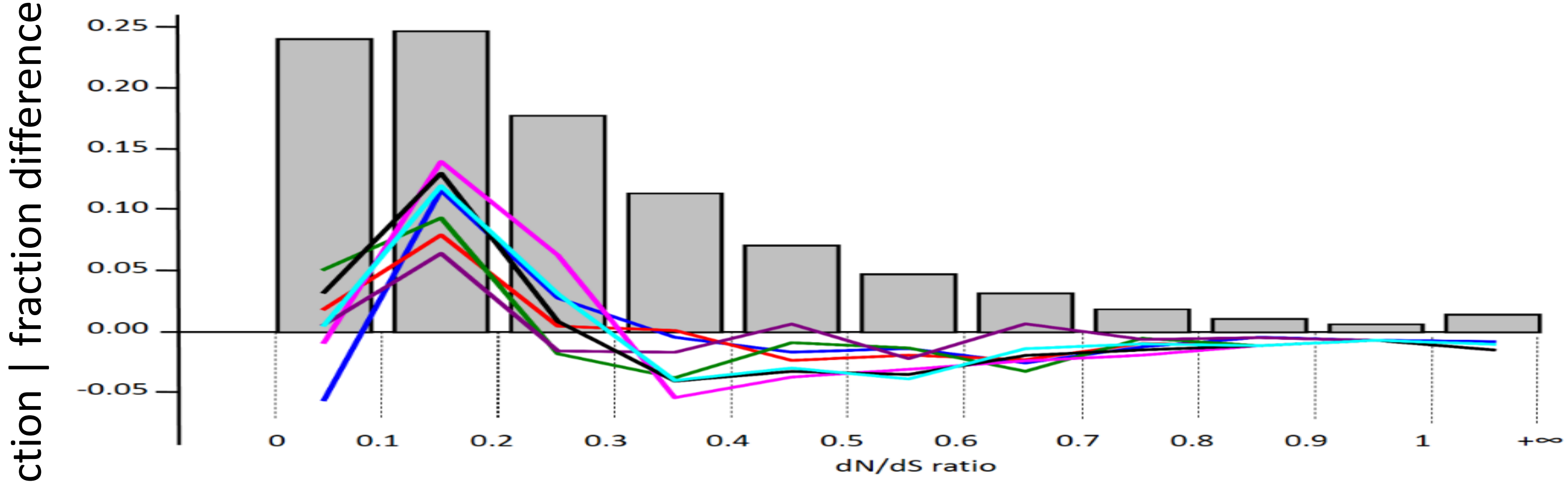
Genes of *Arabidopsis thaliana,* associated with different types of stress, has been analyzed and compared with the analysis of all protein coding (CDS) genes of *A. thaliana*. The gray columns represents the fraction of PAI and DI in CDS genes. The color lines represent the difference between the frequency of the value of PAI/DI in stress-associated genes and CDS genes.



▨ fraction in all CDS *A. thaliana* genes

- cold
- heat
- light
- osmotic
- oxidative
- salt
- water

difference between stress related and CDS fractions



Genes associated with the response to various types of stress:

1) are characterized by elder age than the average for the genome of A. thaliana.

2) are rather under purifying selection than the average for the genome of A. thaliana

Z. S. Mustafin, V. I. Zamyatin, D. K. Konstantinov, A. V. Doroshkov, S. A. Lashin, D. A. Afonnikov, Phylostratigraphic Analysis Shows the Earliest Origination of the Abiotic Stress Associated Genes in A. thaliana // Genes, 2019, 10(12), 963