Preprints.org

Article

Not peer-reviewed version

Impacts of Natural Selection on Evolution of Core and Symbiotically Specialized Genes in the Polytypic Species *Neorhizobium galegae*

Evgeny S. Karasev , Hosid L. Sergey , Tatiana S. Aksenova , Olga P. Onishchuk , Oksana N. Kurchak , Nikolay I. Dzyubenko , Evgeny E. Andronov * , Nikolay A. Provorov

Posted Date: 21 August 2023

doi: 10.20944/preprints202308.1417.v1

Keywords: Neorhizobium galegae biovars orientalis and officinalis; polytypic rhizobia species; evolution of symbiosis; core and symbiotically specialized (sym) genes; nucleotide polymorphism of genes; driving and purifying selection; p-distance; dN/dS statistics; goats' rue (Galega); Illumina

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2023 doi:10.20944/preprints202308.1417.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Impacts of Natural Selection on Evolution of Core and Symbiotically Specialized Genes in the Polytypic Species *Neorhizobium galegae*

Evgeny S. Karasev [1], Hosid L. Sergey [1,], Tatiana S. Aksenova [1], Olga P. Onishchuk [1], Oksana N. Kurchak [1], Nikolay I. Dzyubenko [2], Evgeny E. Andronov [1,3,*] and Nikolay A. Provorov [1,]

[1] All-Russia Research Institute for Agricultural Microbiology, e-mail: arriam@arriam.spb.ru
[2] All-Russia Research Institute of Plant Genetic Resources, email: secretary@vir.nw.ru
[3] Dokuchaev Soil Science Institute, Moscow, Russia
* Correspondence: eeandr@gmail.com

Abstract: Nodule bacteria (rhizobia) represent a suitable model to address a range of evolutionary problems including the tradeoff between genetic polymorphism and natural selection. Rhizobia possess the complicated genomes in which symbiotically specialized (*sym*) genes differ in their natural histories from core genes encoding for housekeeping functions. Diversification of *sym* genes is responsible for the rhizobia microevolution which depends on the host-induced natural selection. For the rhizobia speciation, diversification of core genes is responsible for which the selective factors are unclear. In this paper we demonstrate that in the goats' rue rhizobia (*Neorhizobium galegae*) populations collected at North Caucasus and composed of two host-specific biovars *orientalis* and *officianalis* (N_2-fixing symbionts of *Galega orientalis* and *G. officinalis*, respectively), evolutionary mechanisms are different for core and *sym* genes. In both *N. galegae* biovars, core genes are more polymorphic than *sym* genes. In bv. *orientalis*, evolution of core genes occurs under the impacts of driving selection (dN/dS > 1), while evolution of *sym* genes is close to neutral (dN/dS ≈ 1). In bv. *officinalis*, evolution of core genes is neutral, while for *sym* genes, it is dependent on purifying selection (dN/dS < 1). A marked phylogenetic congruence of core and *sym* genes revealed using the ANI analysis may be due to a low intensity of gene transfer within and between *N. galegae* biovars. Polymorphism of both gene groups and the impacts of driving selection on the core gene evolution are more pronounced in bv. *orientalis* than in bv. *officianalis* reflecting the diversities of respective host plant species. In bv. *orientalis*, highly *significant* ($P_0 < 0{,}001$) positive correlation was revealed between the p-distance and dN/dS values for core genes, while in bv. *officinalis* this correlation is lowly significant ($0{,}05 < P_0 < 0{,}10$). For *sym* genes, correlation between the p-distance and dN/dS values is negative in bv. *officinalis* but is not revealed in bv. *orientalis*. These data along with the functional annotation of core genes implemented using the Gene Ontology tools suggests that evolution of bv. *officinalis* is based mostly on adaptation for *in planta* niches while in bv. *orientalis* evolution presumable depends on adaptation for soil niches. New insights into the tradeoff between natural selection and genetic diversity are presented suggesting that the gene polymorphism may be extended by driving selection only in the ecologically versatile organisms capable to support a broad spectrum of gene alleles in their genepools.

Keywords: *Neorhizobium galegae* biovars *orientalis* and *officinalis*; polytypic rhizobia species; evolution of symbiosis; core and symbiotically specialized (*sym*) genes; nucleotide polymorphism of genes; driving and purifying selection; p-distance; dN/dS statistics; goats' rue (*Galega*); Illumina

1. Introduction

Root nodule bacteria (rhizobia), N_2-fixing symbionts of leguminous plants, represent a convenient model for the evolutionary genetics of symbiotic organisms. As in other bacteria, rhizobia genomes are composed of conservative core and variable accessory parts [1]. Core genes are responsible for housekeeping functions (basic metabolism, cell development and reproduction, template processes) while the accessory genes encode for various adaptive functions including symbiotic interactions with the leguminous plants. Rhizobia genomes are subjected to the multilevel evolution based on modifications of: (i) symbiotically specialized (*sym*) genes resulted in formation

© 2023 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2023 doi:10.20944/preprints202308.1417.v1

of polytypic species composed of host-specific biovars; (ii) core genes resulted in formation of cryptic (twin) species [2].

Previously we studied the genome diversification in *Rhizobium leguminosarum*, a polytypic species which includes two biovars: *viciae* (symbionts of legumes from tribe *Fabeae*, genera *Lathyrus*, *Lens*, *Pisum*, *Vavilovia*, *Vicia*) and *trifolii* (symbionts of genus *Trifolium* from the tribe Trifolieae) [3]. Cross-inoculation between these biovars is limited and results in the non-N_2-fixing nodules, which are usually underdeveloped and morphologically abnormal [4]. In *R. leguminosarum*, genomes are composed of circular chromosomes and several plasmids, one of them (pSym) having size 200-500 kb harbors the majority of *sym* genes. They include *nod* genes encoding for synthesis of nodulation-inducing lipo-chito-oligosaccharidic Nod factors, and *nif*/*fix* genes encoding for the nitrogenase synthesis and operation. We suggested that in *R. leguminosarum*, evolution of *sym* genes is implemented under impacts of the host-induced natural selection [3,5] while mechanisms for the core gene evolution remain obscure.

In the presented paper we compare the evolutionary dynamics of *sym* and core genes in goats' rue rhizobia (*Neorhizobium galegae*), a polytypic species differentiated into host-specific biovars *orientalis* and *officinalis* – N_2-fixing symbionts of *Galega orientalis* and *G. officinalis*, respectively. In contrast to *R. leguminosarum*, *N. galegae* biovars cross-inoculate their hosts readily resulting in the morphologically normal although non-N_2-fixing nodules. Majority of *N. galegae sym* genes are located on chromids having sizes over 1600 kb. These circular replicons have a plasmid type *rep*ABC system combined with the core genes which are typically located on bacterial chromosomes, including tRNA and rRNA encoding genes [6].

Previously we demonstrated [7] that populations of *N. galegae* bv. *orientalis* collected at the North Caucasian region are more polymorphic for *sym* and core genes than *N. galegae* bv. *officinalis* populations. This difference apparently reflects the diversity of respective host plant species which is sufficiently higher in *G. orientalis* than in *G. officinalis*. Difference between two *N. galegae* biovars for *nif*/*fix* genes was much more pronounced than for *nod* genes since the host specificity of compared biovars pertains N_2 fixation, not nodulation activity [7].

The presented paper demonstrates that in *N. galegae*, impacts of driving/purifying selection (dN/dS) on gene evolution differ sufficiently in the core and *sym* genes and are biovar-specific. Strict phylogenetic congruence of core and *sym* genes was revealed in *N. galegae*, reflecting the location of *sym* genes on non-transmissible chromids. In spite of this congruence, mechanisms of evolution are different in core and *sym* genes, as it was demonstrated by analysis of correlations between p-distance and dN/dS values. These correlations as well as analysis of core gene ontology groups allow us to suppose that in bv. *officinalis* evolution is more dependent on adaptation to endosymbiotic niches than in bv. *orientalis*.

Analyses of the presented data completes some important gaps in our knowledge on the tradeoff between polymorphism of genes and impacts of natural selection on their evolution. Specifically, we suggest that in ecologically versatile organisms, such as *N. galegae* bv. *orientalis*, driving selection enhances the nucleotide gene polymorphism since the newly emerged alleles may coexist stabile with the preexisted ones. However, in ecologically restricted organisms such as *N. galegae* bv. *officinalis*, driving selection does not influence or even decreases the gene polymorphism since new alleles supported by this selection replace (push away) the preexisted alleles from the genepool.

2. Results

2.1. Gene polymorphism and natural selection

We demonstrat that nucleotide polymorphism in polytypic species *N. galegae* depends on driving/purifying (dN/dS-measured) natural selection which operates in a biovar-specific manner and differs in the core and *sym* genes (Table 1). The maximal impacts of driving selection (dN/dS > 1) were revealed in the high-polymorphic core genes of bv. *orientalis*, the minimal impacts – in the low-polymorphic *sym* genes of bv. *officinalis*. A complete phylogenetic congruence between *sym* and core genes was revealed in *N. galegae* using the Average Nucleotide Identity (ANI) analysis (Figure 1) reflecting the parallel and possibly inter-dependent processes of microevolution and cryptic speciation dependent on modifications of *sym* and core genes, respectively.

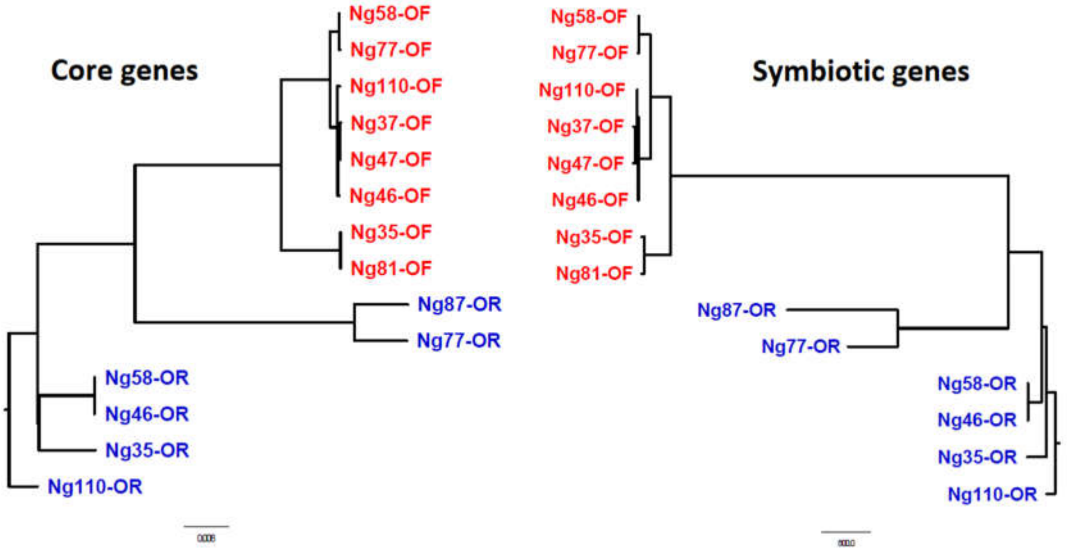

Figure 1. Phylogenetic congruence of core (left) and sym (right) genes according to Average Nucleotide Identity (ANI) analysis. Strains of *Neorhizobium galegae* bv. *officinalis* (OF) are represented in red, of *N. galegae* bv. *orientalis* (OR) – in blue.

Table 1. Nucleotide polymorphism (p-distance) and driving/purifying selection (dN/dS) impacts on the core and *sym* gene evolution in of the host-specific *Neorhizobium galegae* biovars.

Genes*	**Means ± standard errors**		
	bv. *orientalis*	**bv. *officinalis***	**t_{St} (P_0)**
	p-distance		
core	0.048±0.001	0.010±0.001	106.4 (< 0.001)
sym	0.028±0.008	0.005±0.001	2.84 (< 0.05)
t_{St} (P_0)	2.47 (< 0,05)	3.57 (< 0,01)	-
	dN/dS**		
core	1.571±0.050 (D)	1.013±0.026 (N)	9.91 (< 0.001)
sym	1.009±0.142 (N)	0.272±0.111 (P)	4.09 (< 0.001)
t_{St} (P_0)	3.72 (< 0,01)	6,50 (< 0.001)	-

* Numbers of studied core genes are 3840 for bv. *orientalis* and 2734 for bv. *officinalis*; number of studied *sym* genes for both biovars is 39 (16 *nod*, 8 *nif*, 15 *fix*). The Student t-test (t_{St}) was used to assess the probability of null-hypothesis (P_0) suggesting no difference between gene groups or *N. galegae* biovars. ** Natural selection is: D – driving (dN > dS), P – purifying (dN < dS); N – no selection (dN ≈ dS; neutral evolution occurs).

Analysis of correlations between nucleotide polymorphism (p-distance) and the driving/purifying selection (dN/dS) impacts suggests (Table 2) that this selection implements different roles in evolution of core and *sym* genes which depend on the *N. galegae* biovars. For core genes, driving selection results in a marked increase of polymorphism in bv. *oreintalis* (indicated by a highly significant positive correlation between p-distance and dN/dS), but this increase is much less evident in bv. *officinalis* (indicated by a significantly lower although positive correlation between p-distance and dN/dS). For *sym* genes, natural selection does not influence polymorphism in bv. *orientalis* (no correlation between p-distance and dN/dS values) but results in a decreased polymorphism in bv. *officinalis* (negative correlation between these values).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2023 doi:10.20944/preprints202308.1417.v1

Table 2. Correlations between nucleotide polymorphism (p-distance) and natural selection (dN/dS) impacts in core and *sym* genes of the *Neorhizobium galegae* biovars.

Genes	**Pearson correlation coefficients (r)***		t_{St} (P_0)
	bv. *orientalis*	**bv. *officinalis***	
core	+ 0,346 (P_0 < 0,001)	+ 0,066 (0,05 < P_0 < 0,10)	12,73 (< 0,001)
*sym***	+ 0,078 (P_0 > 0,10)	– 0,991 (0,05 < P_0 < 0,10)	4,18 (< 0,001)
t_{St} (P_0)	0,99 (> 0,05)	50,03 (< 0,001)	-

* Probabilities of the null-hypothesis suggesting no correlation between p-distance and dN/dS are given in parentheses after r values; t_{St} (P_0) used to compare the r values is introduced in the footnote for Table 1. **Numbers of studied core genes are 2864 for biovar *orientalis* and 2076 for biovar *officinalis*; numbers of studied *sym* genes are: 15 for bv. *orientalis* and 3 for bv. *officinalis* (only the genes polymorphic in both biovars were used for the correlation analysis).

Importantly, fraction of polymorphic genes in the total genepools is higher in bv. *orientalis* than in bv. *officinalis* for *sym* genes: 38.5±7,8% and 7.7±4,3%, respectively (t_{St} = 3.46; P_0 < 0,01). However, for core genes these fractions do not differ: 74.6±0.70% and 75.9±0.82%, respectively (t_{St} = 1,08; P_0 > 0,10). Importantly, fractions of polymorphic genes is higher for core than for *sym* genes in both biovars suggesting a strong purifying selection for *sym* genes.

Differential impacts of natural selection on polymorphism of core and *sym* genes are evident in bv. *officinalis* (r values differ significantly), not in bv. *orientalis* (r values do not differ) suggesting that adaptive impacts of these genes are biovar-dependent (Table 2).

2.2. Gene ontology analysis

Previously we suggested that in *R. leguminosarum*, evolution of *sym* genes occurs under the impacts of host-controlled natural selection, while the factors responsible for evolution of core genes remain unclear [3,8]. For addressing these factors, we used the Gene Ontology tools providing the functional annotation of core genes that vary sufficiently for nucleotide polymorphism (p-distance) and for susceptibility to driving/purifying selection (dN/dS) [9]. The polymorphic core genes were distributed into a range of Gene Ontology Groups (GOGs) contrasting for p-distance or for dN/dS in which deviations of these parameters from their average values of GO-enrichment exceed the standard deviation (1.509 for dN/dS and 1.519 for p-distance). This approach allowed us to distribute 782 core genes which are polymorphic in both *N. galegae* biovars into 76 GOGs comprising four clusters with contrasting p-distance and dN/dS values: (i) GOGs are elevated over average in both biovars, *orientalis* and *officinalis* (Ori+Off+); (ii) GOGs are elevated over average in bv. *orientalis* but are below average in bv. *officinalis* (Ori+Off–); (iii) GOGs are elevated over average in bv. *officinalis* but are below average in bv. *orientalis* (Ori–Off+); (iv) GOGs are below average in both biovars (Ori–Off–). For the statistical analysis, these four clusters were established independently for p-distance and dN/dS values and are presented as CP-I…CP-IV and CS-I…CS-IV, respectively.

We demonstrated that clusters CP-IV and CS-IV in which both analyzed parameters (p-distance and dN/dS) are below average in both *N. galegae* biovars (Ori–Off–) are most numerous suggesting that purifying selection (dN/dS < 1) resulting in a decreased nucleotide polymorphism represents an important factor of core gene evolution (Table 3). However, tradeoff between gene polymorphism and natural selection differs greatly in the analyzed clusters: representations (%) in the total pool of 782 polymorphic genes for Ori+Off+ and Ori–Off+ clusters are higher for p-distance than for dN/dS (CP-I > CS-I; CP-III > CS-III), while in Ori+Off– and Ori–Off– clusters, gene representations are higher for dN/dS than for p-distance (CS-II > CP-II; CS-IV > CP-IV) (Figure 2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2023 doi:10.20944/preprints202308.1417.v1

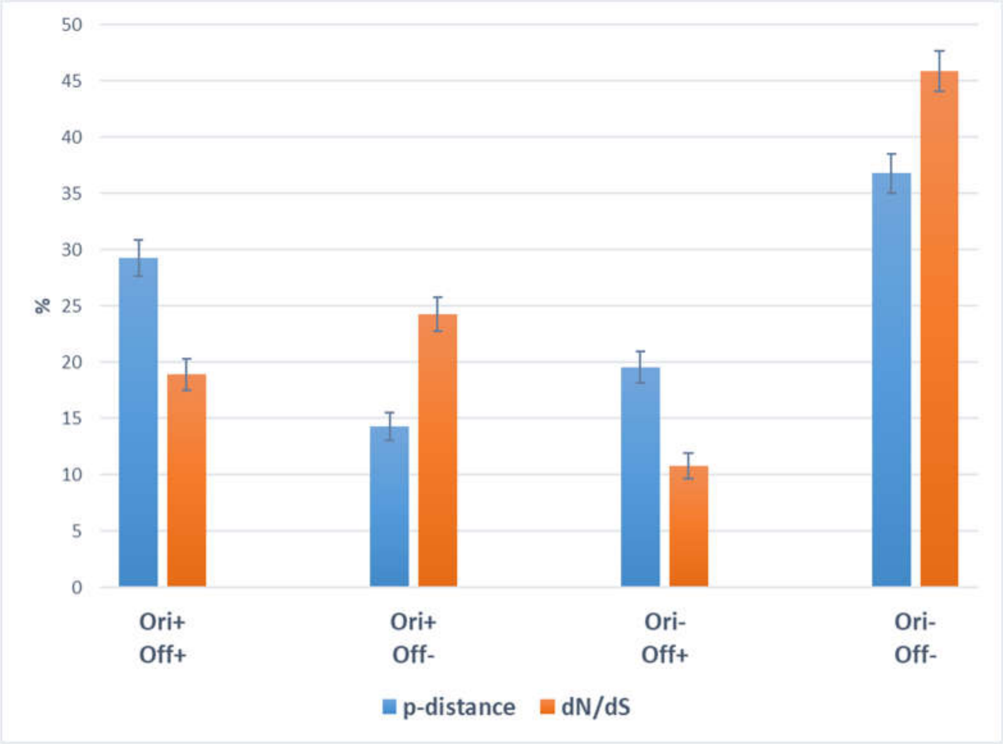

Figure 2. Distribution of 782 polymorphic core genes into the clusters (introduced in text) contrasting for p-distance (CP-I…CP-IV are in blue) and for dN/dS (CS-I…CS-IV are in orange) values in the *Neorhizobium galegae* biovars *orientalis* (Ori) and *officinalis* (Off).

Vertical axis shows the representations (in %, with standard errors) of each cluster in the total pool of 782 analyzed genes (data from Table 3 are used, sizes of columns are given in Table S1 in Supplement).

Table 3. Distribution of 76 Gene Ontology Groups (GOGs) composed of 782 polymorphic *Neorhizobium galegae* core genes into the clusters with contrasting p-distance (CP-I … CP-IV) or dN/dS (CS-I … CS-IV) values (clusters are introduced in the text).

		Numbers of GOGs in clusters contrasting for p-distance*				
		CP-I (229)	CP-II (153)	CP-III (112)	CP-IV (288)	Total GOGs
The same in clusters for dN/dS*	CS-I (148)	3	0	0	0	3
	CS-II (190)	5	3	1	0	9
	CS-III (85)	4	0	2	4	10
	CS-IV (359)	12	4	20	18	54
	Total GOGs	24	7	23	22	76

* Numbers of genes in each cluster are given in parentheses.

Analysis of COG composition enables us to reveal several regularities in the functional diversity of core genome which pertain its operational (involved in cellular metabolism) and informational (involved in template processes) components. Specifically, analysis of GOGs identified on the basis of p-distance demonstrates (Table S2 in Supplement) that low-polymorphic genes responsible for metabolism of N-compounds (nucleosides, amino acids) are assigned to CP-II, while the low-polymorphic genes for lipid and oligosaccharide metabolism are in CP-III. This functional segregation of core genes may reflect a predominant dependence on symbiotic N_2 fixation typical of bv. *officinalis*: its host, *G. officinalis* grows at the limited locations under unfavorable edaphic conditions restricting the bacteria survival in soil and increasing their dependence on the nodular

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2023 doi:10.20944/preprints202308.1417.v1

niches. This dependence is presumable much lower for bv. *orientalis* which is distributed in favorable locations occupied by its host (*G. orientalis*) and therefore this biovar persist in soil more stably than bv. *officinalis*.

As expected, genes which are less variable in both biovars (Ori–Off– clusters) are associated with highly conservative template processes. Specifically, genes encoding for translation are revealed in CP-IV while genes for replication, transcription, translation and DNA repair are in CS-IV (Tables S2-S4 in Supplement).

In order to address the tradeoff between nucleotide polymorphism and impacts of natural selection on the core gene evolution, we analyzed distributions of GOGs among the clusters identified using p-distance or dN/dS values (Table 3). For assessing statistically the coincidence of these distributions we calculated separately for two *N. galegae* biovars the frequencies of GOGs with elevated (over average) values of dN/dS or of p-distance within the GOGs with elevated or decreased values of p-distance or dN/dS, respectively. Analysis of data on the ecologically versatile bv. *orientalis* demonstrates that: (i) frequency of GOGs with elevated dN/dS values is significantly higher among GOGs with elevated p-distance than among GOGs with decreased p-distance (Figure 3a); (ii) frequency of GOGs with elevated p-distance is higher among GOGs with elevated dN/dS than among groups with decreased dN/dS (Figure 3b). Therefore, for ecologically versatile bv. *orientalis*, in both reciprocal comparisons of GOGs contrasting for p-distance and dN/dS values, "High-in-High" frequencies exceed "Low-in-High" frequencies suggesting that driving selection is responsible for an elevated core gene polymorphism. However, in the ecologically restricted bv. *officinalis*, no such difference was revealed suggesting that tradeoff between the core gene polymorphism and the natural selection impacts on the gene evolution depends on the adaptive strategies of compared *N. galegae* biovars.

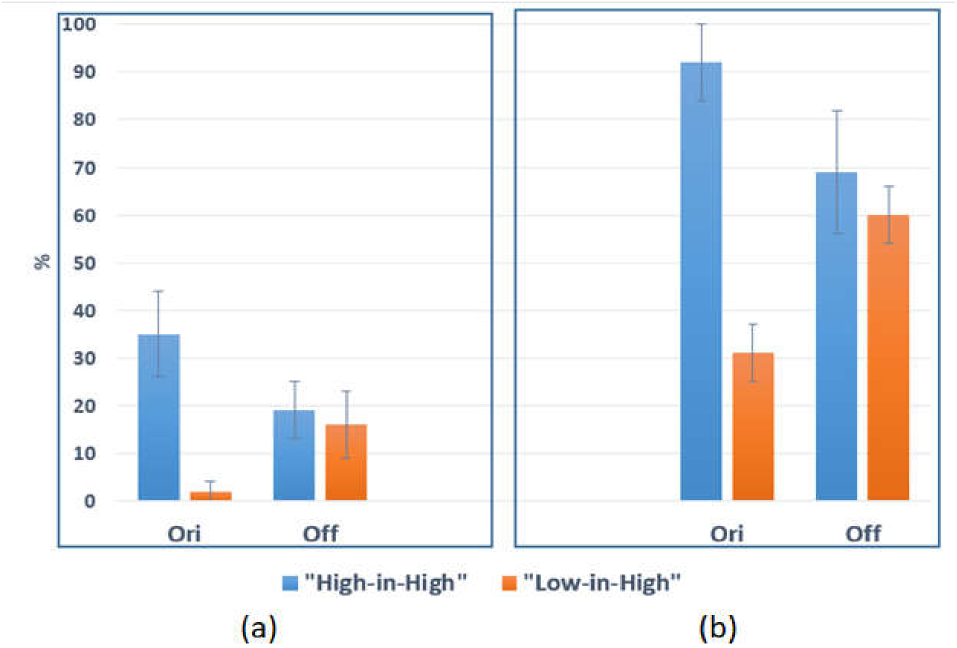

Figure 3. Statistical analysis of clusterization of 76 Gene Ontology Groups (GOGs) contrasting for p-distance and dN/dS values in the *Neorhizobium galegae* biovars *orientalis* (Ori) and *officinalis* (Off) (data from Table 3 are used, sizes of columns are given in Tables S5 and S6 in Supplement). a: frequencies (in % with standard errors) for GOGs with elevated (blue) or decreased (orange) dN/dS values among the GOGs with elevated p-distance values ("High-in-High" and "Low-in-High" freqencies). b: the same for GOGs with elevated or decreased p-distance values among the GOGs with elevated dN/dS values. Significant ($P_0 < 0,01$) differences were reveled in the comparisons of "High-in-High" and "Low-in-High" frequences for bv. *orientalis*, while for bv. *officinalis*, these differences are not significant.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2023 doi:10.20944/preprints202308.1417.v1

3. Discussion

Root nodule bacteria (rhizobia) represent the genetically best studied group of symbiotic microorganisms. Being highly effective producers of N compounds for terrestrial ecosystems, these bacteria are characterized by exclusive ecological and agronomic importance. Moreover, rhizobia are used as a model to address a range of general evolutionary problems including the genomic mechanisms of micro- and macroevolution [8], and the tradeoff between genetic polymorphism and natural selection [10]. As it was demonstrated previously, purifying (stabilizing) selection usually results in a decreased population/gene polymorphism [11] while disruptive and negative frequency-dependent selection result in an extended (balanced) polymorphism [12,13]. The data on influence of driving (directed) selection on the genetic polymorphism are contradictory: it may be increased [14–17], conserved or even decreased [18–20] by this selection.

Application of rhizobia as a model to study impacts of natural selection on genetic diversity is based on the complicated genomic structures in which the core parts encoding for housekeeping functions differ in their natural histories from the accessory parts including the symbiotically specialized (*sym*) genes. Convenient models for analyzing the rhizobia genome dynamics are represented by polytypic species *Neorhizobium galegae* composed of two host-specific biovars – bv. *orientalis* and bv. *officinalis* (symbionts of *Galega orientalis* and *G. officinalis*), and *Rhizobium leguminosarum*, composed of bv. *viciae* (symbionts of plants from tribe Fabeae, genera Lathyrus, Lens, Pisum, Vavilovia, Vicia), bv. *trifolii* (symbionts of genus *Trifolium* from tribe Trifolieae), bv. *phaseoli* (symbionts of *Phaseolus vulgaris* from tribe Phaseoleae).

Previously we demonstrated that in *R. leguminosarum*, *sym* and core genes differ greatly in nucleotide polymorphism (p-distance) and are phylogenetically non-congruent, suggesting that evolution of these genes is independent [3]. This independency may be resulted from intensive horizontal gene transfer in rhizobia populations wherein two gene groups recombine randomly due to location of *sym* genes on mobile plasmids [1].

In order to reveal impacts of natural selection on the rhizobia gene polymorphism we used the set of *N. galegae* strains originated from North Caucasus region. As we demonstrated earlier [7], diversity of nucleotide sequences (measured as p-distance) in *N. galegae* is higher for core genes than for *sym* genes and is biovar-dependent: bv. *orientalis* is more polymorphic than bv. *officinalis* for both gene groups (Table 1). This difference may be due to contrasting levels of diversity in the respective host plant species. Specifically, North Caucasus is the longstanding center of origin for *G. orientalis* while colonization of this area by *G. officinalis* is more recent [21]. Previously we quantified diversity of two *Galega* species in North Caucasus using the nucleotide polymorphism analysis in a range of genes followed by genomic fingerprinting and confirmed the morphological data suggesting a higher *G. orientalis* diversity with respect to *G. officinalis* [22,23]. An important source of genetic diversity in *N. galegae* may be represented by translocations of the Insertion Sequences (IS) which are more abandoned in bv. *orientalis* than in bv. *officinalis* [24].

In this paper, we demonstrated that in *N. galegae*, core and *sym* genes are phylogenetically congruent (Figure 1) apparently due to their restricted recombination based on location of *sym* genes on non-mobile chromids. Nevertheless, some evolutionary important parameters of diversity are different in these genes: tradeoff between nucleotide polymorphism and evolutionary impacts of natural selection depend on the gene group (core or *sym*) and on the *N. galegae* biovar (*orientalis* or *officinalis*) (Table 1). Analyses of the total gene pools (Table 2) as well as of Gene Ontology Groups (GOGs) (Table 3, Figure 3), suggest that driving selection pressures result in an increased polymorphism of core genes in bv. *orientalis*, not in bv. *officinalis*. We suggest that in bv. *orientalis*, maintenance of novel core gene alleles by driving selection (dN/dS > 1) may be combined with preservation of preexisted alleles due to a broad ecological amplitude of this biovar, therefore, its genetic polymorphism is elevated. However, in bv. *officinalis*, the newly emerged gene alleles possibly substitute the preexisted ones (due to a restricted ecological amplitude of this biovar), therefore, gene polymorphism in this biovar is not changed.

In accordance to contrasting ecological affinities of the *Galega* species, a range of differences between their symbionts were revealed: (i) low polymorphic GOGs are affiliated with N metabolism

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2023 doi:10.20944/preprints202308.1417.v1

(apparently responsible for symbiotic adaptations) in bv. *officinalis* and with the synthesis of surface polysaccharides (responsible for adaptations to edaphic stresses) in bv. *orientalis* (Table S1-S3 in Supplement); (ii) *sym* genes evolve under purifying selection (dN/dS < 1) impacts in bv. *officinalis* while a neutral evolution (dN/dS ≈ 1) was revealed for these genes in bv. *orientalis*; (iii) evolution of core genes occurs mostly under impacts of driving selection in bv. *orientalis* while this evolution is neutral in bv. *officinalis* (Tables 1-3, Figure 3). From these data, we can suppose that operation of *sym* genes is most critical for bv. *officinalis* because at North Caucasian region this biovar persists under unfavorable soil conditions and should survive mostly due to colonization of endosymbiotic niches. However, bv. *orientalis* persists under more favorable conditions, as compared to bv. *officinalis*, therefore adaptations to edaphic factors dependent on core genes are highly important for bv. *orientalis*.

Interestingly, *N. galegae* differs in its evolutionary dynamics from the previously studied [3] *R. leguminosarum* species: *sym* and core genes *in N. galegae* are more similar in their diversity parameters than in *R. leguminosarum*. This difference between two polytypic species may be due to a more restricted recombination of *sym* and core genes in *N. galegae* with respect to *R. leguminosarum* (Table 4). Comparative analysis of these species contributes sufficiently to understanding of tradeoff between microevolution, speciation and macroevolution and between genetic polymorphism and natural selection.

Table 4. Comparison of evolutionary sufficient items in *Neorhizobium galegae* and *Rhizobium leguminosarum*.

Items	***Neorhizobium galegae****	***Rhizobium leguminosarum*****
Compared biovars (their hosts)	bv. *orientalis* (*Galega orientalis*), bv. *officinalis* (*G. officinalis*)	bv. *viciae* (*Lathyrus, Lens, Pisum, Vavilovia, Vicia*), bv. *trifolii* (*Trifolium*)
Taxonomic diversity of hosts of the compared biovars	Different species of the same plant genus	Various plant genera and tribes
Replicons harboring *sym* genes	Chromids (> 1600 kb)	Plasmids (200-500 kb)
Phylogenetic congruence of core and *sym* genes	High or complete	Incomplete or absents
Differences between biovars for the diversity parameters of *sym* and core genes	Highly significant for both gene groups	For *sym* genes are more pronounced than for core genes

Variation within biovars:
for core genes
for *sym* genes
	highly significant
	significant but lower than for core genes
		highly significant
		much lower than for core genes or is not revealed

* This research; ** from [3].

Specifically, we demonstrate that in both rhizobia species, *N. galegae* and *R. leguminosarum* core genes responsible for speciation and macroevolution differ greatly in their natural histories from *sym* genes responsible for microevolution. Different genetic mechanisms were proposed for micro- and macro-evolution by J. Philiptschenko (1927) [25] who was the first to define these processes and correlated them with changes of the eukaryotic nuclear and cytoplasmic genes, respectively. Later, R. Goldschmidt (1949) [26,27] suggested that microevolution is based on the minor adaptive changes ("micro-mutations") which can not initiate the speciation and macroevolutionary processes dependent on "macro-mutations" (responsible for generation of "hopeful monsters"). The similar approach was proposed in the punctuated evolution concept [28,29], which may be apparently used to address the symbiosis evolution since hosting of symbiotic microbes by eukaryotic organisms represents the rapid evolutionary bursts in contrast to gradual evolution suggested by the models gradualist evolution based on natural selection [2,30].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2023 doi:10.20944/preprints202308.1417.v1

The other important issue of the rhizobia evolutionary genetics pertains the tradeoff between driving selection and gene polymorphism which may be increased by this selection in an ecologically versatile organism (such as *N. galegae* bv. *orientalis*) allowing a broad allelic diversity in the analyzed genes. However, in an ecologically restricted organism (such as *N. galegae* bv. *officinalis*), gene polymorphism is not changed or is even decreased under impacts of driving selection since co-existence of different gene alleles is presumable blocked. An extended bioinformatics analysis is required to analyze a relationship between adaptive potentials of organisms and impacts of natural selection on their polymorphism expressed in the diverse rhizobia species and in other symbiotic organisms.

4. Materials and Methods

4.1. Collection of strains and DNA sequencing

During expedition to the North Caucasus in 2003, a number of soil samples was collected, from which rhizobia strains of bv. *orientalis* and bv. *officinalis* were isolated [22,23]. A total of 14 rhizobia strains were isolated from soil samples in a microvegetation experiment using nodules of *Galega orientalis* and *G. officinalis* according to standard protocol [31]. They include strains of bv. *officinalis*: NG_35_off (JANFGK000000000), NG_37_off (VYYB00000000), NG_46_off (JANFGL000000000), NG_47_off (JAMQCN000000000), NG_58_off (JANFGM000000000), NG_77_off (JANFGN000000000), NG_81_off (JANFGO000000000), NG_110_off (VZUM00000000)) and of bv. *orientalis*: NG_35_ori (JANFGP000000000), NG_46_ori (JANFGQ000000000), NG_58_ori (VZUN00000000), NG_77_ori (JANFGR000000000), NG_87_ori (VZUL00000000), and NG_110_ori (JANEZU000000000).

Isolates were cultivated at 28°C and 220 rpm for 48 h in modified yeast mannitol broth (YMB) with 1% sucrose [32]. DNA was obtained by the lysozyme–SDS–phenol–chloroform extraction protocol, with minor modifications [33]. Sequencing of strains NG_37_off and NG_87_ori was performed on a Pacbio RSII instrument with P6 in two SMRT cells (Pacific Biosciences of California, Inc., Menlo Park, CA, United States). PacBio sequencing and subsequent error correction analysis and assembly were performed at Arizona Genomics Institute (US). The assembly the strains was carried out de novo using HGAP (https://github.com/jtchien0925/PacBio_HGAP_assembly). Sequencing of other 12 rhizobia strains, 7 strains of biovar *officinalis* and 5 strains of biovar *orientalis*, was performed on a MiSeq genomic sequencer (Illumina Inc., San Diego, CA, United States), according to the manufacturer's protocol, using the MiSeq Reagent Kit, 600 Cycles (Illumina, United States) at the Genomics Core Facility, Siberian Branch, Russian Academy of Sciences (Institute of Chemical Biology and Fundamental Medicine, Novosibirsk). The assembly of the sequences was carried out using the CLC Workbench (https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-clc-genomics-workbench/) by mapping on reference genomes NG_37_off and NG_87_ori each biovar respectively.

4.2. Finding core genes by global alignment

All genes of each genome were matched with the genes of other 13 genomes using the global alignment method. For this, BLAT was used (http://genome.ucsc.edu/cgi-bin/hgBlat). All paired genes were sorted in descending order of identity. The paired genes with a maximal identity of at least 70% in the DNA sequence were selected. After that a table of all genes and their presence in each of the 14 strains was generated. Only genes that were found in all 14 strains were selected for the core genome.

For bv. orientalis, the estimated number of core genes is approximately 5000, for bv. *officinalis* – 4200, while 3900 core genes are common for two biovars. For analyzing the variability and selection indices we used 3840 genes from bv. *orientalis* and 2734 genes from bv. *officinalis* (genes are common for two biovars with non-zero polymorphism and dN/dS).

4.3. Symbiotic genes

For both biovars we analyzed 16 nod genes (encoding for Nod factor synthesis) 8 nif genes (for nitrogenase synthesis) and 15 fix genes (for electron and energy supply of nitrogenase).

4.4. Gene alignment using Muscle

DNA sequences of 14 strain variants of each gene were aligned by MUSCLE (**M**ultiple **S**equence **C**omparison by **L**og-**E**xpectation, https://www.ebi.ac.uk/Tools/msa/muscle/) using standard coding sequence alignment parameters.

4.5. Calculation of nucleotide polymorphism (p-distance)

DNA polymorphism of each gene was calculated based on the number of nucleotide substitutions for each pair of strains using standard metrics (https://www.megasoftware.net/mega1_manual/Distance.html). DNA regions with undetermined sequences (N, non-detected) and gaps were not taken into account. The number of substitutions was normalized by dividing the total length of the compared genes without gaps and undefined nucleotides. The matrixes (sized 14X14 by the number of strains) of p-distances of each gene were calculated. The average polymorphism of each gene was calculated using the average p-distance of all elements of the matrix excluding diagonal elements (distance of a gene with itself is zero).

4.6. Calculation of dN/dS index

The calculation of the dN/dS ratio of nonsynonymous (dN) to synonymous (dS) substitutions was performed according to the Jukes-Cantor (JC) model (https://bioinformatics.cvr.ac.uk/calculating-dnds-for-ngs-datasets/). In JC model, dN/dS for each codon was calculated separately and compared with the theorized ratio of substitutions. For example alanine is encoded by three different codons when there are nine possible single substitution of each codon and consequently its theoretical dN/dS ratio is 3/9 or 1/3. The obtained dN/dS value of each gene was normalized by dividing by the number of coding codons of the compared sequences. Then the matrixes (sized 14X14 by the number of strains) of dN/dS indexes of each gene were calculated. The average dN/dS index of each gene was calculated by the average dN/dS index of all elements of the matrix excluding diagonal elements.

4.7. Functional annotation of core genes, Gene Ontology (GO)

We used eggnog-mapper (https://github.com/eggnogdb/eggnog-mapper/issues/135) to annotate newly assembled genomes and assign genes to certain functional groups of Gene Ontology. A detailed transcript of each group was performed on the AmiGO 2 website [34]. AmiGO 2 is a project to create the next generation of AmiGO, the current official web-based toolkit for searching and browsing Gene Ontology’s database. The Gene Ontology Consortium (GOC) provides computable knowledge regarding the functions of genes and gene products.

4.8. Determination of the predominance of functional groups of genes GO (gene Ontology) and statistical significance

The prevalence of certain groups of genes was calculated as the ratio of the actual number of genes to their expected number based on the sample size and the total number of genes of the selected group,

$$P_{enrich}=N_{obs}/(S_{sep}*(N_{sep}/N_{genome})),$$

where P_{enrich} is the predominance of a given group of genes, N_{obs} is the number of genes in the sample, S_{sep} is the number of genes in a given group (GO), N_{sep} is the sample size, N_{genome} is the total number of genes found in Gene Ontology.

The statistical significance of the predominance of certain groups of genes was obtained using a permutation test that simulates the same value of the size of the group and the total number of genes.

The permutation test was performed 10,000 times, this was enough to calculate statistical significance 95% (P. val. <0.05).

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Table S1: Representations (%) of *Neorhizobium galegae* core genes in the clusters identified by the gene nucleotide polymorphism (p-distance) and natural selection (dN/dS); Table S2: GO (Gene Ontology) enrichment of functional groups of biovars *orientalis* and *officinalis* by p-distance; Table S3: GO enrichment of functional groups of biovars *orientalis* and *officinalis* by dN/dS (stabilizing selection); Table S4: GO (Gene Ontology) enrichment (predominance of functional groups) of biovars *orientalis* and *officinalis* by dN/dS (driving selection); Table S5: Frequencies (F) of Gene Ontology Groups (GOGs) with high dN/dS values among GOGs with high or low p-distance values; Table S6: Frequencies (F) of Gene Ontology Groups (GOGs) with high p-distance values among GOGs with high or low dN/dS values; Table S7: List of symbiotic *nif*, *fix* and *nod* genes and their allocation on *Rhizobia Galega* genomes bv. *officinalis* and *orientalis***.**

Author Contributions: Conceptualization, P.N. and A.E.; methodology, A.E., K.E. and S.H.; software, K.E. and S.H.; validation, P.N. and S.H.; formal analysis, P.N. A.N. and S.H.; investigation, A.E., K.E. and S.H.; resources A.T., K.O., O.O. and D.N.; data curation, K.E. and S.H.; writing—original draft preparation, P.N., A.E., K.O. and S.H.; writing—review and editing, P.N., A.E., K.O. and S.H.; visualization, A.E. and S.H.; supervision, A,E.; project administration, P.N. and A.E.; funding acquisition, P.N. and A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by by the Russian Science Foundation, grant number 19-16-00081-P.

Data Availability Statement: All 14 genome sequences mentioned in this study can be downloaded by their assesion number from https://www.ncbi.nlm.nih.gov.

Acknowledgments: We are grateful to Olga Hosid for editing the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study, in the collection, analyses and interpretation of data.

References

1. Young, P.W., Crossman, L.C., Johnston, A.WB., Thomson, N.R., Ghazoui, Z.F., Hull, K.H., Wexler, M., Curson, A., Todd, J.D., Poole, P.S., Mauchline, T.H., East, A.K., Quail, M.A., Churcher, C., Arrowsmith, C., Cherevach, I., Chillingworth, T., Clarke, K., Cronin, A., Davis, P., Fraser,A., Hance, Z., Hauser, H., Jagels, K., Moule, S., Mungall, K., Norbertczak, H., Rabbinowitsch, E., Sanders, M., Simmonds, M., Whitehead, S., Parkhill J. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome Biol.* **2006**, *7(4)* R34. (doi: 10.1186/gb-2006-7-4-r34.)
2. Provorov, N.A., Andronov, E.E., Kimeklis, A.K., Onishchuk, O.P., Igolkina, A.A., Karasev, E.S. Microevolution, speciation and macroevolution in rhizobia: genomic mechanisms and selective patterns. *Front. Plant Sci.* **2022**. *13:* 1026943 (doi: 10.3389/fpls.2022.1026943).
3. Kimeklis, A., Chirak, E., Kuznetsova, I., Sazanova, A., Safronova, V., Belimov, A., Onishchuk, O., Kurchak, O., Aksenova, T., Pinaev, A., Andronov, E.E., Provorov N.A. Rhizobia Isolated from the Relict Legume *Vavilovia formosa* Represent a Genetically Specific Group within *Rhizobium leguminosarum* biovar *viciae*. *Genes*, **2019**, *10:* 991. (doi:10.3390/genes10120991)
4. Onischuk O.P., Kurchak O.N., Kimeklis A.K., Aksenova T.S., Andronov E.E., Provorov N.A. Biodiversity of the symbiotic systems formed by nodule bacteria *Rhizobium leguminosarum* with the leguminous plants of galegoidcomplex. Sel'skokhozyaistvennaya Biologiya (Agricultural Biology), **2023**, 58(1), 87-99. (doi: 10.15389/agrobiology.2023.1.87eng)
5. Provorov, N.A., Andronov, E.E., Onishchuk, O.P. Forms of natural selection controlling the genomic evolution in nodule bacteria. *Rus. J. Genet*. **2017**. *53(4)*. 411-419. (doi:10.1134/S1022795417040123)
6. Österman, J., Marsh, J., Laine, P.K., Zeng, Z., Alatalo, E., Sullivan, J.T., Young, P.W., Thomas-Oates, J., Paulin, L., Lindström K. Genome sequencing of two *Neorhizobium galegae* strains reveals a *noe*T gene

responsible for the unusual acetylation of the nodulation factors. *BMC Genomics*. **2014**; *15(1)*: 500. (doi:10.1186/1471-2164-15-500)

7. Karasev, E.S., Andronov, E.E., Aksenova, T.S., Tupikin, A.E., Provorov, N.A. Evolution of goat's rue rhizobia (*Neorhizobium galegae*): an analysis of the polymorphism of the nitrogen fixation genes and the genes of nodule formation. *Russ. J. Genetics*. **2019**, *55:* 234-238. (doi: 10.1134/S1022795419020078)
8. Provorov, N.A., Andronov, E.E., Kimeklis, A.K., Onishchuk, O.P., Igolkina, A.A., Karasev, E.S. Microevolution, speciation and macroevolution in rhizobia: genomic mechanisms and selective patterns. *Front. Plant Sci*. **2022**. *13*: 1026943 (doi: 10.3389/fpls.2022.1026943).
9. Xin, Z., Cai, Y., Dang, L.T. Burke H.M.S., Revote J., Charitakis N., Bienroth D., Nim H.T., Li H.Y., Ramialison M. MonaGO: a novel gene ontology enrichment analysis visualisation system. *BMC Bioinformatics*, **2022**, *23*: 69. (https://doi.org/10.1186/s12859-022-04594-1)
10. Andrews, C.A. Natural Selection, Genetic Drift, and Gene Flow Do Not Act in Isolation in Natural Populations. *Nature Education Knowledge* **2010**, *3(10):* 5
11. Cheng, C., Kirkpatrick, M. Molecular evolution and the decline of purifying selection with age. *Nat Commun* **2021**, *12*, 2657. (https://doi.org/10.1038/s41467-021-22981-9)
12. Lee C.-R., Mitchell-Olds T. Environmental Adaptation Contributes to Gene Polymorphism across the *Arabidopsis thaliana* Genome, *Molecular Biology and Evolution* **2012**, *29(12)*, 3721–3728 (https://doi.org/10.1093/molbev/mss174)
13. Marchinko, K.B., Matthews, B., Arnegard, M.E., Rogers, S.M., Schluter, D. Maintenance of a Genetic Polymorphism with Disruptive Natural Selection in Stickleback. *Current Biology* **2014** *24(11)* 1289-1292. (https://doi.org/10.1016/j.cub.2014.04.026)
14. Rahman, S., Kosakovsky, P.S.L., Webb, A., He,y J. Weak selection on synonymous codons substantially inflates dN/dS estimates in bacteria. *PNAS* **2021**, *118 (20)* e2023575118 (doi.org/10.1073/pnas.2023575118)
15. Taub, D.R., Page, J. Molecular Signatures of Natural Selection for Polymorphic Genes of the Human Dopaminergic and Serotonergic Systems: A Review. *Front Psychol* **2016**, *8(7)* 857. (doi: 10.3389/fpsyg.2016.00857)
16. Moon, S.U., Na, B.K., Kang, J.M., Kim, J.Y., Cho, S..H, Park, Y.K., Sohn, W.M., Lin, K., Kim, T.S. Genetic polymorphism and effect of natural selection at domain I of apical membrane antigen-1 (AMA-1) in *Plasmodium vivax* isolates from Myanmar. *Acta Trop* **2010**. *114(2)* 71-75. (doi: 10.1016/j.actatropica.2010.01.006)
17. Kang, JM., Ju, HL., Kang, YM. Genetic polymorphism and natural selection in the C-terminal 42 kDa region of merozoite surface protein-1 among *Plasmodium vivax* Korean isolates. *Malar J* **2012**, *11*, 206 (https://doi.org/10.1186/1475-2875-11-206)
18. Barnard-Kubow, K., Sloan, D., Galloway, L. Correlation between sequence divergence and polymorphism reveals similar evolutionary mechanisms acting across multiple timescales in a rapidly evolving plastid genome. *BMC evolutionary biology* **2014**, *14(1)* 1. (doi: 10.1186/s12862-014-0268-y)
19. Vigué, L, Eyre-Walker, A. The comparative population genetics of *Neisseria meningitidis* and *Neisseria gonorrhoeae*. *PeerJ* **2019**, *27(7)* e7216. (doi: 10.7717/peerj.7216)
20. Sunyaev, S., Kondrashov, F.A., Bork, P., Ramensky, V. Impact of selection, mutation rate and genetic drift on human genetic variation, *Human Molecular Genetics* **2003**, 12(24) 3325–3330. (https://doi.org/10.1093/hmg/ddg359)
21. Raig, H., Nõmmsalu, H., Meripõld, H., and Metlitskaja, J. 2001.Fodder Galega. H. Nõmmsalu, ed., Estonian Research Institute of Agriculture, Saku, Estonia. 141 p.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 August 2023 doi:10.20944/preprints202308.1417.v1

22. Andronov, E., Terefework, Z., Roumiantseva, M., Dzyubenko, N., Onichtchouk, O., Kurchak, O., Dresler-Nurmi, A., Young, J. P., Simarov, B., Lindstrom. Symbiotic and Genetic Diversity of *Rhizobium galegae* Isolates Collected from the *Galega orientalis* Gene Center in the Caucasus. *Applied and Environmental Microbiology* **2003**, *69(2)* 1067-1074. (doi: 10.1128/AEM.69.2.1067–1074.2003)
23. Österman, J., Chizhevskaya, E., Andronov, E., Fever, D., Terefework, Z., Roumiantseva, M., Onichuk, O., Dresler-Nurumi, A., Simarov, B., Dzybenko, N., Lindstrom, K. *Galega orientalis* is more diverse than *Galega officinalis* in Caucasus—whole-genome AFLP analysis and phylogenetics of symbiosis-related genes. *Mol. Ecol.* **2011**, *20(22)* 4808-21. (doi: 10.1111/j.1365-294X.2011.05291.x.)
24. Radeva, G., Jurgens, G., Niemi, M., Nick, G., Suominen, L., Lindström, K. Description of two biovars in the *Rhizobium galegae* species: biovar *orientalis* and biovar *officinalis*. *System. Appl. Microbiol.* **2001,** *24(2)* 192-205.
25. Philiptschenko, J. *Varriabilität und Variation*. Berlin Bornträger. 1927.
26. Goldschmidt, R. B. Phenocopies. *Scientific American* **1949**, *181(4)* 46–49.
27. Goldschmidt, R. B. The intersexual males of the beaded minute combination in *Drosophila melanogaster*. *PNAS* **1949**, *35(6)* 314-316.
28. Gould, S.J. Wonderful Life: The Burgess Shale and the Nature of History. W.W. Norton & Company, New York. 1989
29. Gould, S.J. & Eldredge, N. Punctuated equilibrium comes of age. *Nature* **1993**, *366(6452)* 223– 227. (doi: 10.1038/366223a0)
30. Theis, K. R., Dheilly, N. M., Klassen, J. L., Brucker, R. M., Baines, J. F., Bosch, T. C., et al.. Getting the hologenome concept right: an eco-evolutionary framework for hosts and their microbiomes. mSystems. **2016**, 1(2), e00028–e00016. (doi: 10.1128/mSystems.00028-16)
31. Novikova, N., Safronova, V. Transconjugants of *Agrobacterium radiobacter* harbouring sym genes of *Rhizobium galegae* can form an effective symbiosis with *Medicago sativa*. *FEMS Microbiol. Lett.* **1992,** *93* 261–268. (doi: 10.1111/j.1574-6968.1992.tb05107.x.)
32. Allen, O.N. Experiments in soil bacteriology. Minneapolis, Minnesota: Burgess Publishing Co., 1959, 52-59.
33. Somasegaran, P., Hoben, H.J. Isolating and Purifying Genomic DNA of Rhizobia Using a Large-Scale Method. In: Garber R.C., editor. Handbook for Rhizobia. Springer; New York, NY, USA: 1994. pp. 279–283.
34. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.