9-24-2013

Quantifying Mutational Impacts on Intrinsic DNA Flexibility in Prokaryotic Genomes

Mohammed Alawad

Quantifying Mutational Impacts on Intrinsic DNA Flexibility in Prokaryotic Genomes

Author: *Mohammed Alawad*
Advisor: *Gregory Babbitt*

Rochester Institute of Technology
Thomas H. Gosnell School of Life Sciences
College of science
Bioinformatics department

Approval date: 9-24-2013

LIST OF CONTENT:

List of Figures:

Thesis Committee Members:

Advisor:
Dr. Gregory Babbitt
Assistant Professor,
Department of Biological Sciences,
Rochester Institute of Technology

Dr. Gary Skuse
Professor,
Department of Biological Sciences,
Rochester Institute of Technology

Dr. Feng Cui
Assistant Professor
Department of Biological Sciences,
Rochester Institute of Technology

Acknowledgment

The author wishes to thank several people. I would like to thank my parents for their endless love and support while I am far away. I would like to show my greatest appreciation to Prof. Gregory Babbitt. I can’t say thank you enough for his tremendous support and help. Without his encouragement and guidance this project would not have be materialized.

Abstract

The existence of synonymous codon biases across all taxonomic groups is a long standing problem in biology. While codon bias seems to be adequately explained by the maintenance of translation efficiency and accuracy in some organisms, there is still no adequate explanation of why codon biases universally track the intergenic gc content, as these regions of the genome would not be under selection pressures affecting translation. One part of the story may come from the triplet nature of codon in which each third position defines the minor groove width and thus affects the basic structure of the DNA by altering the intrinsic flexibility. In addition, this intrinsic flexibility, which is also GC dependent, play a major role on defining the phosphate linkages of the backbone conformation as well as participating with other binding molecules. Packaging such a type of information within the DNA sequence seems to be essential especially when observing such a variation of codon bias among organism. The potential existence of this form of 'architectural' information in the genome might also predict that evolutionary processes at the synonymous sites are not simply an accident, but it might indicate a fundamental connection between the biophysical aspects of DNA and usage of codons. In this thesis, I present a broad taxonomical analysis of the mutational impacts on the intrinsic flexibility of DNA among 26 prokaryotic genomes and investigate its relationship to entropy based codon bias gc content and protein conservation . I conclude that codon bias appears universally connected to the intrinsic flexibility of the genome especially for genomes with extreme GC contents. In all genomes, genes under strong purifying selection at the level of the protein appear to have constraints in the mutational impacts on DNA flexibility. This may reflect a fundamental limitation in ability of DNA to multiplex information at the levels of protein and nuclear architecture.

1. Introduction:

The triplet nucleotides that define the amino acid, or codons, are most well known as the represented as three letter base combinations assigned to independent cells in a codon lookup table. However, Codons do not actually ever exist in this apparent isolated state in the DNA sequence; instead they always exist in the linear context of a relatively stiff molecular polymer. Looking at the DNA sequence from a biophysical prospective has led us to understand more in drug discovery such as understanding the mechanisms of drug binding to a target protein or defining the structure-function relationships in proteins. Additionally, genomes also might include some biophysical attributes when observing the synonymous variation of codon usage. It is commonly known that 61 different codons encode only 20 amino acids in the translational process. Codon-bias, which is a dynamic and multi-scaled context in the genome architecture, is traditionally defined by the various frequencies of which a synonymous codon is observed to occur. There are multiple different ways to measure codon usage; the simplest is counting each specific codon frequency. To clearly and simply quantify codon-bias, Shannon Information Theory can be applied to count the weighted sum of relative entropy [3][24]. This phenomena has recently been linked to the mutational impacts of the intrinsic DNA flexibility in a yeast genome [4]. Ultimately, codons are as much defined by their phosphate linkages as by their nucleobase assignment. These linkages structurally also are very active in defining the genome architecture; therefore one possible functions of synonymous codon-biases are to specify the flexibility of the nucleotide sequence on top of genetic information in protein coding regions. By choosing a particular codon from another, genomes may control the accessibility of genes and whole-genome folding status through intrinsic flexibility; a level of structurally-encoded information that must be overlapped or multiplexed with genetic information. This introduces a fundamental problem of how genomes may multiplex the genetic information and this structural information defined through intrinsic DNA into the same molecular context of the DNA.

Intrinsic flexibility of the DNA is an essential characteristic of the double helix. In fact, flexibility is a regional quality of a genome, as some part of the genome tends to be stiffer and hence more accessible than others[5]. Every third nucleobase in DNA contributes significantly in these variations, and also defines the minor groove width of DNA structure [4]. All genomes experience multiple compacting processes in order to fit inside the small space of the bacterial cell or eukaryotic nucleus. In theory, flexibility probably should play crucial roles in these packaging processes since long stretch of DNA should have some resistance to molecular deformation due to the proximity of negative phosphate charges on the DNA backbone. Experimental data suggested that some stretches are more flexible than the other based on the sequence composition caused by electrostatics. Heddi et al. [12] proposed a widely accepted experimental scale that quantifies the intrinsic flexibility of the ten-dinucleotide conformations in terms of Twist, Roll and base pair displacement. In other words, the TRX (Twist, roll and x-displacement) scale, based on the reflection of BI/BII conformations, measures the average percentage of time that specific phosphate linkage (connecting two bases) resides in the BII conformation. To study DNA-protein interactions via the intrinsic flexibility; Heddi et al[12] probed the DNA backbone in a solution with absence of protein and observed the phosphate group's conformations in B-DNA using large Nuclear Magnetic Resonance (NMR) p31 chemical shifts then studied the structure. As a result, TRX provides a scale range from 0 (stiff dimer) to 42 (flexible dimer) for all 10 dimers. The most notable thing in this scale is the effect of GC base pairs, which relates also to DNA helical shape. Guanine-cytosine dinucleotide has a wider minor groove than other dimers which indicates more separations between phosphate groups therapy high flexible polymers. In addition, when either Guanine or cytosine exists in the dimer, its score tends to be higher on the scale. This scale, which contains noteworthy variations, can be used to understand the structural information of a genome; since flexibility has shown to play a major role in gene regulation and nucleosome positioning [5].

The discovery of the genetic code has shown that 61 possible codons can be used to express only 20 amino acids. This redundancy of expression allowed for most amino acids to be encoded by two to six different codons, known as synonymous codons. A wide variety of organisms uses different synonymous codons with different frequencies, a phenomenon which has been termed codon bias [13]. In addition, there is a wide variation on how bias codons are among organisms; some species tend to have very strong bias where as others use different synonymous codons with similar frequencies [13]. Surprisingly, there is a long line of evidence that synonymous codon usage is under weak selection and thereby indicates a type of selection that is independent of the protein level. Even more startling, this variation occurs non-uniformly within a genome and/or from gene to gene. In 1982, M.Gouy and C.Gautier [10] speculated that natural selection contribute to that bias by presenting a correlation between codon usage and the gene expression level in Escherichia coli. Later on, many scientists believed that codon bias enriches both efficiency and accuracy of the protein expression; driven eventually by selection. These translational efficiencies are well known to be important, however none of that explains why codon bias tracks the intragenic GC content [13]. Using a complete genomic set across all organisms (Prokaryotes, archaea, and eukaryote), [13] presented a strong correlation between GC content and codon bias.

Although GC content has been linked to many biological processes such as determination of coding regions, the complete nature of this variation has not been completely understood. Given the recent discoveries by Heddi et al. it would seem that bending and twisting the DNA seems to be largely defined by the level of guanine-cytosine, which eventually affects the backbone conformation. So this may help to explain the very nature of selection on GC content and codon bias. Intrinsic flexibility plays a curial role in protein interaction and packaging DNA (supercoil) which indicates a need to a potential control this type of information. This information may have coincided with the genetic information contained in the code and thus can be inferred from the triplet nature of the codon. (Itzkovitz and Alon)[15] suggested that a DNA sequence (or more specifically the universal genetic code) carries arbitrary parallel codes within it by studying alternative genetic codes. Unfortunately, the natures of these parallel codes are still ambiguous and unknown. The goal of this research is to understand the structural information (intrinsic flexibility) specifically on the synonymous sites that are encoded in a DNA sequence to demonstrate the nature of these multiplexed data. We used 24 Bacterial and Archaean genomes that given as multiple alignments of orthologs genes to detect the mutational impacts in the flexibility using an evolutionary timescale. We hypothesis that the DNA sequence encompasses structural information in a multiplexed form by maintaining the genome flexibility at synonymous sites using codon bias. Specifically, we investigated whether selection on the protein-coding level (i.e. dN/dS) is interacting with the mutational impacts on DNA flexibility, especially at synonymous sites where the third base position may actively define the width of the DNA's minor groove.

2.Material and Methods

2.1 Overview

By placing pure B-DNA in solution, the phosphate (p) linkages of a backbone can be one of two molecular conformations, BI or BII. These two conformations differ structurally only in the torsion angles identified as ε and ζ. Figure 1 illustrates the disparity of BI and BII with CpA dinucleotide where ε – ζ = -90 in BI while in BII ε – ζ = +90. Based on these properties of B-DNA, TRX scale [4] quantifies the intrinsic flexibility of ten dinucleotide conformations in terms of twist, roll and base pair displacement. Moreover, TRX measures the average percentage of time that specific p linkages, which connect two bases, remain in the infrequent BII conformation. This scale constructed using large nuclear magnetic resonance dataset based on p31 chemical shifts. Higher the score is, more flexible the dimer. For example, CpG dinucleotide is the most flexible dimer with score of 43 while the average score of all the ten dimers is 21. Table1 show the scores of the ten different dimers with pyrimidine-purine (YR) order for a given dimer. TRX illustrates the effect of the base composition towards DNA flexibility (higher GC content accompanied with more flexibility). This analysis takes the advantages of this flexibility scale in order to examine mutational impacts in Bacterial and Archaea organisms.

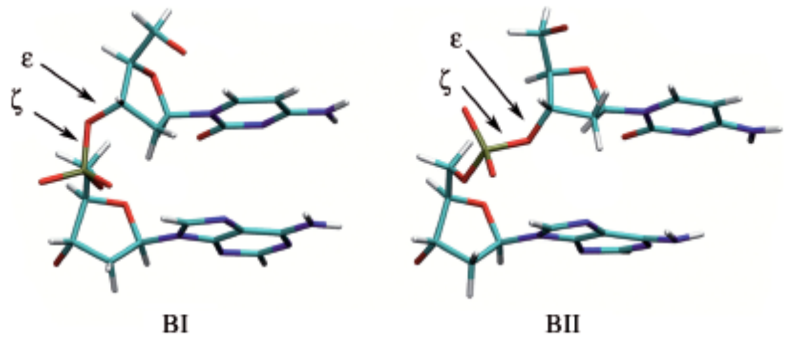


Figure 1 : BI and BII conformations of CpA dinucleotide [12]

Dimer	**pyrimidine-purine**	**TRX**
CpG:CpG	YR:YR	43
CpA:TpG	YR:YR	42
GpG:CpC	RR:YY	42
GpC:GpC	RY:RY	25
GpA:TpC	RR:YY	22
TpA:TpA	YR:YR	14
ApG:CpT	RR:YY	9
ApA:TpT	RR:YY	5
ApC:GpT	RY:RY	4
ApT:ApT	RY:RY	0

Table1: DNA flexibility measured by TRX, pyrimidine-purine (YR) description of a given dimer [12],[4]

By looking at the changes in DNA flexibility (dTRX) that occur over time, analysing mutational impact for a certain genome can be applied. Evolution as a concept is the key to use time as a tool to detect changes in DNA sequences in any organism. PAML [26] is a package of programs that use Maximum Likelihood algorithm to apply multiple evolutionary analyses. One of these programs is Basmel.exe which allows us to assemble ancestral sequences using model-based likelihood approach (Joint Reconstruction) [26] for specific aligned genes of a given organism. ATGC [20], which is a database for closely related Prokaryotic and Archaea genomes, is the data source of the extant sequences in this analysis. To assemble the ancestral sequences from Basmel using ATGC data structure, phylogenetic trees for specific clusters of genes for particular organism is required. MEGA-CC [17] is the other evolutionary software is used to obtain trees by applying multiple algorithms such as Neighbor Joining (NJ) [9]. It allows batch processing of multiple clusters of orthologous genes that represent a cluster of genomes for specific Prokaryotes or Archaea.

Flexibility of any given codon is defined as the sum of the scores for four phosphate linkages. A base substitution in any position of a given codon often causes an alteration in flexibility. Specifically, when the substitution occurs in the first or third position of a codon, both external and internal linkages change. Therefore, to calculate the overall TRX for a codon, both internal and external linkages scores are required. In this comparative genome analysis, we define the mutational impact on the codon as the following:

Formula(1):

$$dTrx_{Codon} = (E1^{trx} + E2^{trx} + I1^{trx} + I2^{trx})_{ancestral} - (E1^{trx} + E2^{trx} + I1^{trx} + I2^{trx})_{extant}$$

Where:

E^{trx} : *TRX score for external linkages.*

I^{trx} : *TRX score for internal linkages.*

A Java code has been implemented to categorize six classes of substitution: synonymous, non-synonymous, synonymous/transition, synonymous/transversion, non-synonymous/transition and non-synonymous/transversion. Additionally, it averages the mutational impacts across multiple Prokaryotes and Archaea genomes. Multiple Perl/Java scripts have been implemented to batch process and control the three major programs in a single run as showen in figure 2.

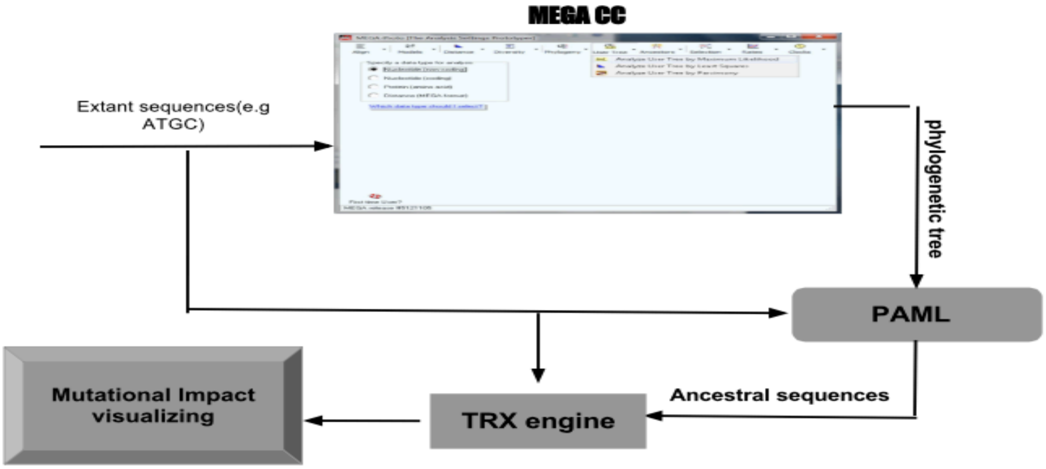
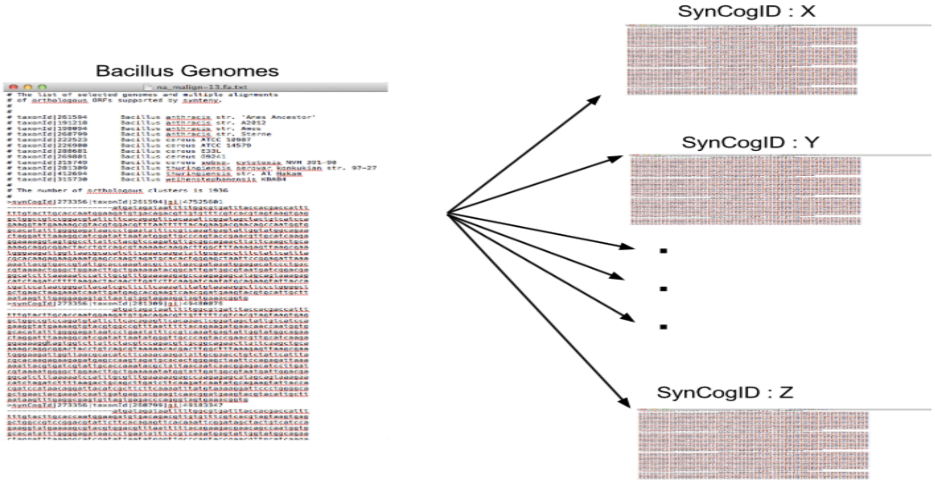


Figure 2 : Flow chart for method implementation

2.2 Data Set

“ATGC (Alignable Tight Genomic Clusters) is a database of closely related microbial genomes optimized for micro evolutionary research [20]”. This database includes more than 30 Prokaryotes and Archaea genomes, which vary in size from 9MB to less than 1MB for a file. There are multiple ways for ATGC to display their genomes’ clusters. One format is pre computed multiple alignments of orthologous Open Reading Frames (ORF) of a specific genome. The objective of this analysis is to obtain the multiple alignments for each orthologous gene in a specific taxon; then these are utilized in the construction the ancestral sequences. Unfortunately, ATGC only provides the data for the whole clusters of genes for a specific taxon. Another Java code was devised and implemented in order to divide ATGC raw clusters of genes from one file into separate files that contain specific gene alignments. For example, Bacillus has 1940 different files of orthologous genes after the separation using this script. ATGC uses SynCogID as pointer for each gene cluster, which is the same pointer used to assign genes to each file. In order to analyze any genome, it is crucial that this preprocessing step be performed before any subsequent steps are attempted.



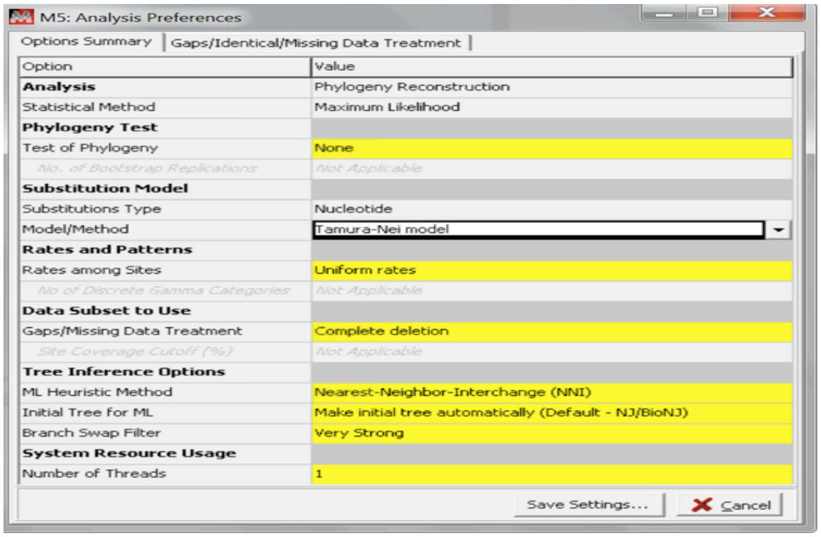
```
# The list of selected genomes and multiple alignments
# of orthologous ORFs supported by synteny.
#
#
# taxonId|261594      Bacillus anthracis str. 'Ames Ancestor'
# taxonId|191218      Bacillus anthracis str. A2012
# taxonId|198094      Bacillus anthracis str. Ames
# taxonId|260799      Bacillus anthracis str. Sterne
# taxonId|222523      Bacillus cereus ATCC 10987
# taxonId|226900      Bacillus cereus ATCC 14579
# taxonId|288681      Bacillus cereus E33L
# taxonId|269801      Bacillus cereus G9241
# taxonId|315749      Bacillus cereus subsp. cytotoxis NVH 391-98
# taxonId|281309      Bacillus thuringiensis serovar konkukian str. 97-27
# taxonId|412694      Bacillus thuringiensis str. Al Hakam
# taxonId|315730      Bacillus weihenstephanensis KBAB4
#
# The number of orthologous clusters is 1936
#
>synCogId|273356|taxonId|261594|gi|47525601
>synCogId|273356|taxonId|281309|gi|49480075
```

Figure 3 : Orthologous gene separation using SynCogID

2.3 Phylogenetic Trees

In order to obtain the ancestral sequences using PAML, phylogenetic trees for each cluster of genes for an organism are required. Molecular Evolutionary Genetic Analysis (MEGA) is a tool that provides multiple evolutionary analyses such as aligning sequences or creating phylogenetic trees. It was designed specifically for any biologist to reconstruct the evolutionary histories of species using the statistical Maximum Likelihood approach. Evolutionary trees in this program are constructed by applying a matrix of pairwise distances using a maximum composite likelihood approach by Neighbor Joining and BIONJ algorithms [9] [26] on nucleotide sequences. MEGA computational core (MEGA-CC) is a newly optimized version of MEGA that enable researchers to funnel the analysis through many kinds of scripts. Most of the important features of MEGA, such as the construction of maximum likelihood trees, are available in the computational core version. A short Perl script (GetTrees.pl) was written and implemented to transport all clusters of genes from the targeted genome into MEGA-CC in order to enable it to construct multiple phylogenetic trees. To batch process MEGA-

CC, the ‘analysis preferences’ dialog box or (MEGA-proto.exe) should be setup with the targeted settings in a .mao file type. (GetTrees.pl) causes MEGA-CC control file ‘.mao’ (which contains the desired settings for certain analysis) to run repeatedly using Windows command prompt for all gene clusters. Figure 4 shows the control file settings that are used to construct Maximum Likelihood trees for all data.



Option	Value
Analysis	Phylogeny Reconstruction
Statistical Method	Maximum Likelihood
Phylogeny Test	
Test of Phylogeny	None
No. of Bootstrap Replications	*Not Applicable*
Substitution Model	
Substitutions Type	Nucleotide
Model/Method	Tamura-Nei model
Rates and Patterns	
Rates among Sites	Uniform rates
No of Discrete Gamma Categories	*Not Applicable*
Data Subset to Use	
Gaps/Missing Data Treatment	Complete deletion
Site Coverage Cutoff (%)	*Not Applicable*
Tree Inference Options	
ML Heuristic Method	Nearest-Neighbor-Interchange (NNI)
Initial Tree for ML	Make initial tree automatically (Default - NJ/BioNJ)
Branch Swap Filter	Very Strong
System Resource Usage	
Number of Threads	1

Figure 4 : Maximum Likelihood tree analysis preferences using MEGA-CC

As with most of the other softwares, MEGA-CC has a formatting requirement for all input alignments. Unfortunately this format (also named MEGA) is slightly different from the targeted data set from ATGC (which is in FASTA format). In order to batch process MEGA-CC, another Java script was devolved to convert the input alignments from FASTA to MEGA format before constructing the phylogenetic trees. This means for execution in this analysis, there are two pre-processing steps: gene separation and format conversion. As a result, multiple trees (.nwk files) can be assembled and utilized as the second required input to PAML in order to create ancestral sequences.

2.4 Ancestral Reconstruction

This project intends to look for the mutational impact that occurs during evolution by comparing ancestral sequences with extant sequences of a certain organism. Basmel is one program from the PAML package that is designed for phylogenetic analyses of DNA or protein sequences. One important feature of this program is that it generates ancestral sequences using the joint reconstruction approach from the extant sequences provided [26], which fits perfectly with this analysis. In addition to the extant sequences, Basmel require the corresponding phylogenetic tree as a second input in order to form the ancestral sequences. According to Zhang [26], the accuracy of generating ancestral sequences using PAML is higher than other methods such as Parsimony method. The previous section described how to generate phylogenetic trees using MEGA-CC. This section is concerns linking each specific tree with its corresponding extant sequences and then run them through Basmel.exe. This has been achieved by devising a Perl script (GetAnces.pl) that manages and maps each .nwk tree with its extant sequences and then executes Basmel.exe to construct their ancestral sequences.

Figure 5 shows a State diagram of the GetAnces.pl script that batch processes basmel.exe and assemble the ancestral sequences for each group of synteny blocks. There are two important files (Phylip.txt and basmeltree.tree) that need to be updated with targeted sequences and trees in each cycle. Basmel.exe stores and creates the ancestral sequences and store them in the .rst control file. GetAnces.pl starts by creating a sub-folder to collect all ancestral sequences. The sequence headers are used to map each extant sequence to its corresponding phylogenetic tree and then the mapped tree kept in basmeltree.tree. In addition, a copy of the aligned genes is transferred to Phylip.txt. In the final step for each cycle, the ancestral sequences are copied from the .rst control file to the sub-folder.

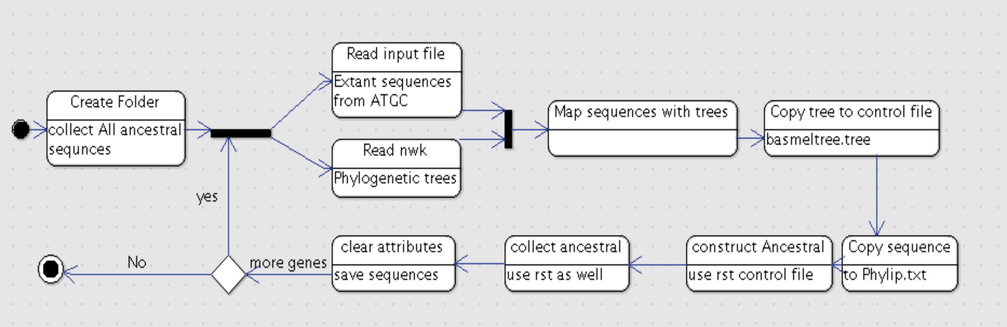


Figure 5 : State diagram of GetAnces.pl

2.5: DNA flexibility engine

As mentioned previously, the main goal of this analysis is to look for the differences in the TRX value (caused by mutations) and occur between the ancestral and extant sequences. The flexibility of a codon is defined as the sum of four TRX scores, which are from two internal and two external phosphate linkages. When mutation occurs in the extant codon, the differences among the TRX scores will indicate gaining or losing flexibility across the codon. Figure 6 shows an example of how a mutation can cause a change in the total TRX value for a codon. Calculating dTRX (see *formula (1))* in this example shows a negative TRX value (-3), which indicates an increase in the flexibility. However, if the difference is a positive value, it will denote stiffening or losing flexibility in the codon. On the other hand, the reason we need to compute the four linkages for each codon is that when a mutation occurs in the first or the third position of the codon, it will cause a change in the total TRX score for both external and internal linkages. As a result, evolutionary constraint could be visualized by averaging the changes in the total TRX across each gene for a whole genome of Prokaryotes or Archaea.

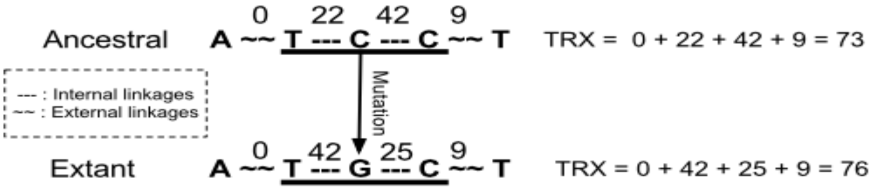


Figure 6 : Example shows the differences that happen in TRX for one codon through a mutation

To analyze evolutionary constraint, dTRX across the genomes has been calculated using six different classes of substitutions (synonymous, non-synonymous, synonymous/transition, synonymous/transversion, non-synonymous/transition and non-synonymous/transversion). Synonymous mutations are “silent changes”, which means the amino acid product is always the same when substitution occurs in the codon, while non-synonymous changes alter the amino acid product. Transition substitutions are the interchange of purines bases (A→G) or pyrimidines bases (C→T) which are less likely to result in amino acid alteration. Conversely, transversion mutations are exchanges of purine to pyrimidine bases or vice versus. Each class of substitutions will assist understanding the evolutionary constraint across the genome in a broad sense (see discussion). In this case, there is a need to develop a code that classifies each type of mutation by looking at the translated amino acid then perform dTRX calculations.

Using Java SE platform and the Integrated Development Environment (IDE) eclipse , an object-oriented based code has been written to calculate the mutational impacts using TRX. To apply the mathematical operations described above and classify each type of mutation for specific sequence, a class (named GeneInfo) has been constructed to store the information needed for any object used in the main class ‘operation’. Each object represents all the results needed for each sequence such as total TRX value or the number of substitutions. The first step when running this code, involves loading the input sequences from a text file into the code variables. Next, AnalyszeSeq() calculates all required analyses for a sequence such as the number of substitutions or their positions. AnalyszeSeq() uses sub-functions that calculate the total TRX score, classify synonymous or non-synonymous mutations and decides whether the mutation is transitional or transversional. Figure 7 shows a UML class diagram for this code while table 2 shows the description of the major functions. To ensure optimal results, gap mutations have been ignored and multiple error detection methodologies have been applied to decrease the margin of error.

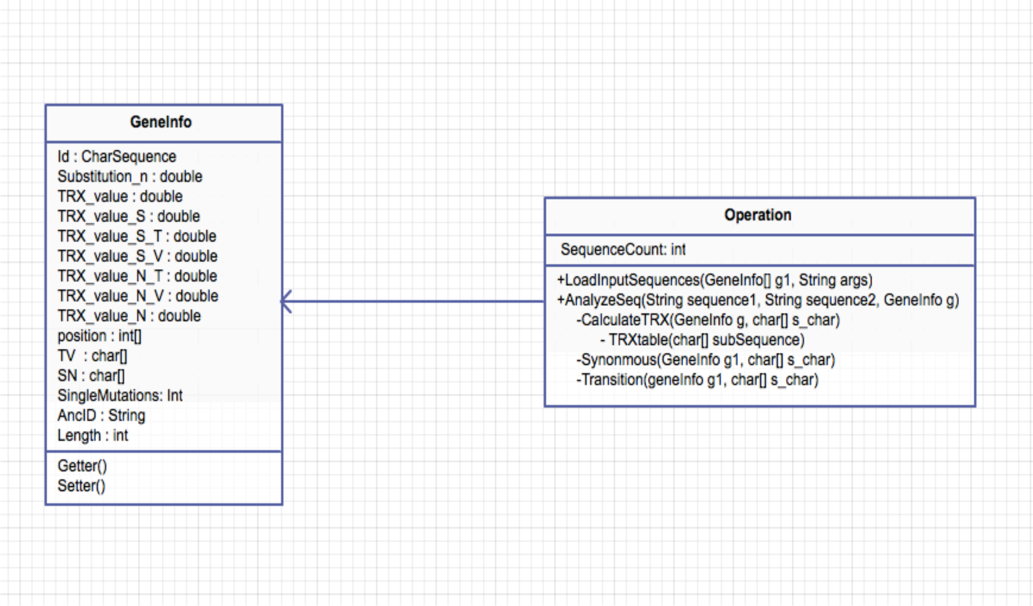


Figure 7 : UML class diagram for TRX engine

Function	***Task***	***Output***
LoadInputSequences (geneInfo)	Load input dataset from txt file and create objects and send each sequence to Analyze seq	null
AnalyzeSeq (string,string,geneInfo)	Receive two string and apply TRX analysis	Fill GeneInfo with required infor
CalculateTRX (GeneInfo, char[])	Receive Subsequence '5 pase long' and calculate TRX for this codon.	Return TRX value for specific mutation
TRXTable(char[])	Used by CalculateTRX() for TRX calculation	TRX value for specific dimer
Synonmous(geneInfo, char[])	Check whether mutation synonymous or non- sysnmous	Set object with correct value
Transition (geneInfo, char[]	Check whether mutation is transition or transverion	Set object with correct value

Table 2: Description for the major functions in this class

3.Results:

To test the hypothesis that DNA flexibility and codon bias potentially interacts with evolutionary processes acting at the protein level, the results were analyzed at three levels across multiple prokaryotes and achaea genomes. To describe the relationship between codon bias, mutational impact on the flexibility and GC content, we needed to look at the gene level first. The main focus of the second part is to examine fundamental constraints on the protein level evolution related to DNA flexibility at a whole genome level. One major concern of this analysis is to explore the variations at genome level of multiple organisms, and look for evidence of a fundamental relationship between codon bias and the averages of the mutational impacts on the flexibility.

3.1 Strong evidence in constraint on the synonymous sites between protein evolution and mutational impacts on intrinsic DNA polymer flexibility

It is commonly known that "the ratio of the number of Non synonymous substitutions per non-synonymous site (dn) to the number of synonymous substitutions per synonymous site (ds),[wiki]" dn/ds is used to infer the direction of natural selection (i.e. selective pressure). A higher ratio indicates selective pressure or positive selection on a specific gene while a lower value indicates functionally conserved genes (i.e. stable selection). We found that there is a fundamental constraint on the genes that have higher selection pressure when looking at the mutational impact on the flexibility in all the data set. Among these data, figure 8 shows example plots of four different genomes that vary in codon bias entropy and gc content. The upper half of each set in figure8 shows the mutational impacts on synonymous site (left) and non-synonmous sites (right) while the lower half of the graphs represent the evolutionary part for both synonymous (left) and non synonymous sites (right). Bacillus (figure8 a), which has low codon bias and low GC content, presents these selections on the mutational impacts; genes with higher dn/ds on synonymous sites tend to be more neutral while non synonymous sites are variable for high dn/ds values. The same selection is presented for the rest of the data as in: 1. Yersinia (figure 8b) low codon bias and average GC content, 2.Pseudomonas (figure 8c) showed extreme codon bias and high GC content and 3.Prochlorococcus (figure 8d) presented with extreme codon bias and low GC content. In addition, genomes with extreme codon bias tend to gain (with high GC content) or lose (with low GC content) flexibility on the synonymous sites through time as it been shown in figures 8c,d.

3.2 Essential relationship between the average deviation in the flexibility with both codon bias and GC content in genomic level

When looking at the final maps of the average mutational impacts for each genome and linking that to both Codon bias and GC content, fundamental correlation is presented. As some genomes tend to gain or lose flexibility, under the sub-optimal value (average TRX) genomes deviate at that point and correlate with codon bias variation. Figure 9a, shows this essential relationship as deviating from the middle point which indicates higher codon bias in both situations with r=0.72. On the other hand, as a reflection of the TRX scale itself, an unblemished relationship between GC content and mutational impact in the intersic flexibility shown in figure 9b.

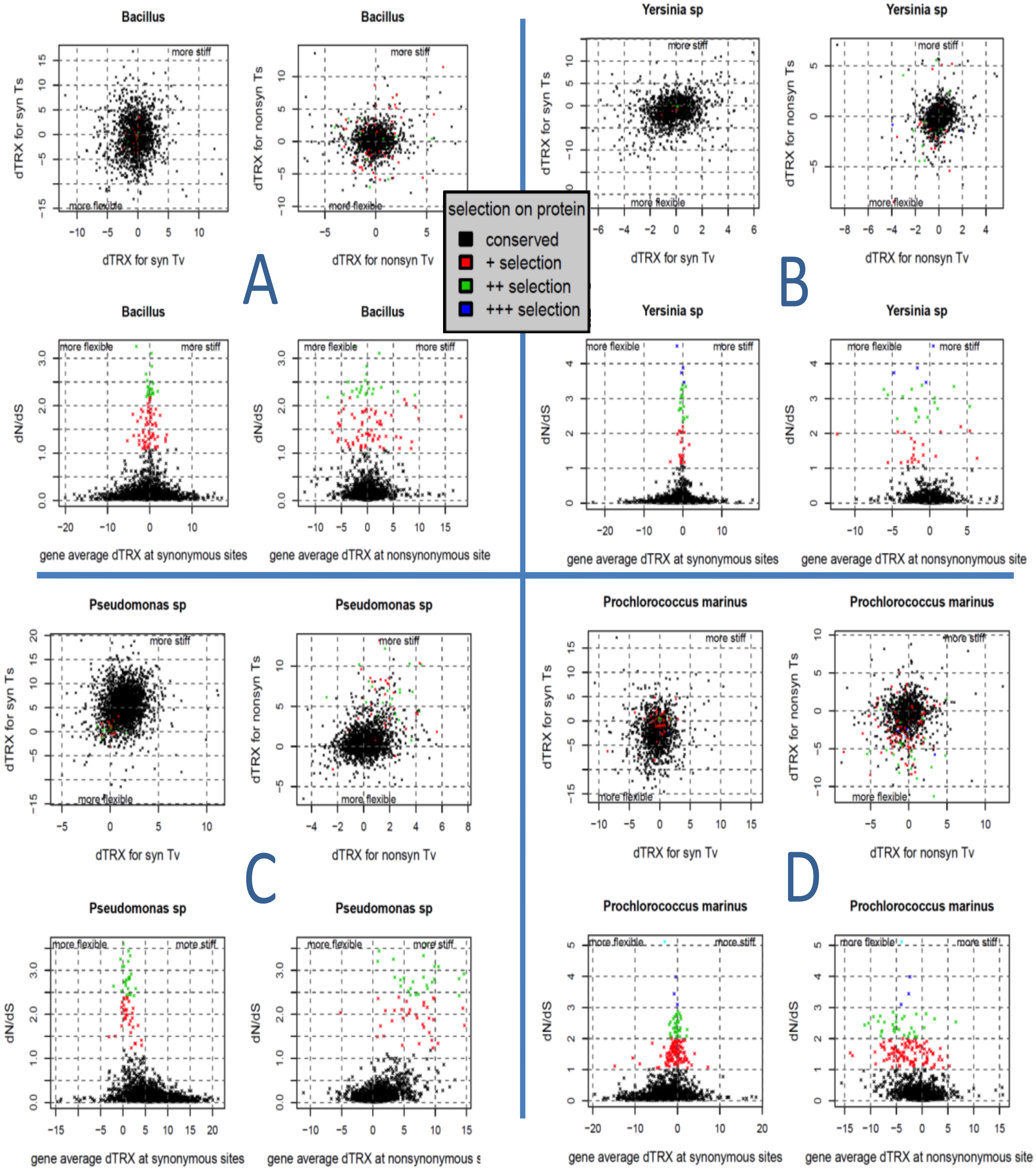


Figure 8 : Fundamental constraints between protein evolution (dN/dS) and mutational impacts on intrinsic DNA flexibility (dTRX) or genome architecture. *Example plot sets are shown for two genomes with low codon bias (A) Bacillus and (B) Yersinia, and two genomes with extreme codon bias; (C) Pseudomonas = high GC and (D) Prochlorococcus = low GC. Within each plot set, genes functionally conserved at the protein-level (i.e. low dN/dS) are shown in black, while genes adaptively altered at protein level are shown colored. Mutational impacts on flexibility (dTRX) are shown separately for synonymous sites (left side of plot set) and non-synonymous sites (right side of plot set). dTRX for transitions and transversion are separated in the upper plots of each set.*

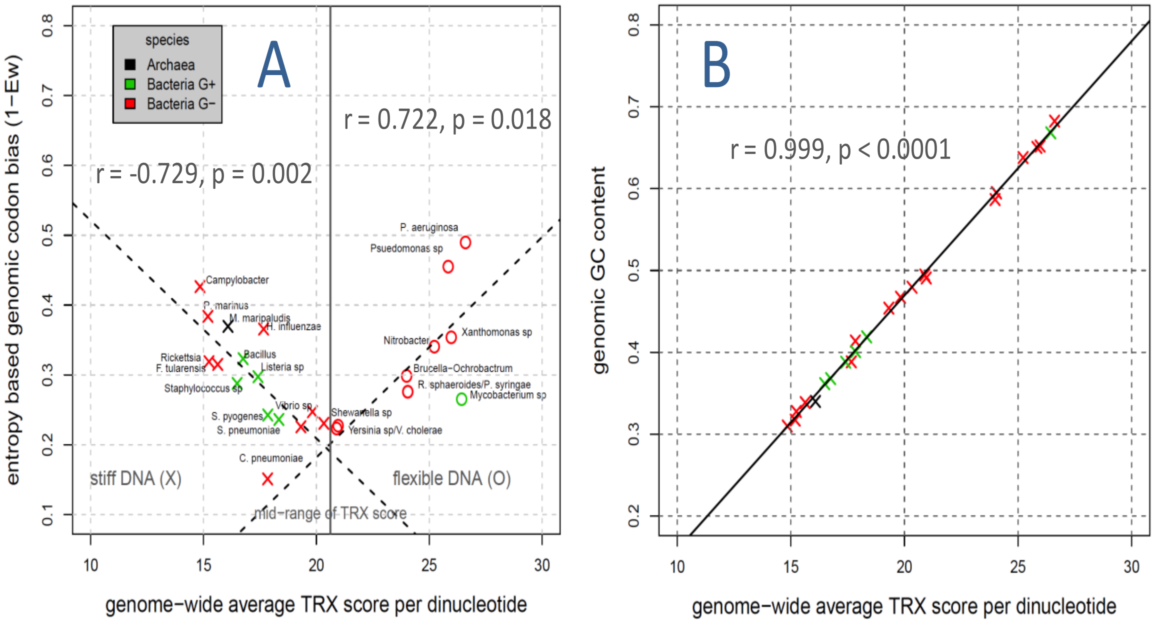


Figure 9 : A fundamental relationship between intrinsic DNA flexibility (TRX score), genomic GC content and entropy-based codon bias*. (A) Prokaryotic genomes with uncharacteristically stiff or flexible genome architecture, and thus deviating from the middle of the TRX scale, demonstrate increased codon bias. (B) The relationship between GC content and intrinsic DNA flexibility at the genomic level is particularly pronounced, reflecting the trends easily observed in the TRX scale itself, where flexibility increases with GC containing dinucleotides***.

4.Discussion:

In recent years, an understanding of the process of protein translation has developed dramatically; specifically by showing codon usage pattern is related to protein synthesis efficiency and caused by selection. However, the exact relationship between natural selection and codon bias usage is not visible yet. One study [7] suggested that Codon usage in prokaryotes is associated strongly with the bacteria's lifestyle. They used 699 different types of bacteria to study the variation of usage in codon bias and concluded that "organisms living in multiple habitats, including facultative organisms, mesophiles and pathogenic bacteria, exhibit high extents of codon usage bias[7]". Some types of bacteria vary from others in codon bias and that is supported in another study by (Singer and Hickey)[23] who found some pattern of codon usage at the synonymous sites in Thermophilic prokaryotes. Indeed, this variation in codon bias among these organisms might indicate some structural information encoded within the DNA sequence other than genetic code and that these structural characteristics of genes and even whole genomes may in some ways relate to the thermodynamics and chemistry of the organisms environment. Taken as a whole, our results suggest that Prokaryotes and some Achaea encode specific structural information within their genomes and rely upon codon bias to maintain the flexibility of their genome. The results indicated that there is clear evidence that codon bias appears as a selective force that drives the shape of the genome by maintaining its flexibility, which seems to be correspondent overall to (Botzman and Margalit)[7] conclusion. Although some strains of bacteria are highly variable on the way they live within each group, the type of data we used in this study was a cluster of genomes for each type, which allowed us only to test these bacteria as a whole. As a general trend, bacteria that live in multiple environments tend to have higher codon bias; and although we did not test this directly, we believe they do that to manage certain flexibility to facilitate backing and folding their DNA within certain environment.

4.1 Genes that are functionally conserved at the protein level encounter evolutionary constraints that limit the evolution of DNA flexibility

One long debate in the scientific community is whether synonymous mutations are neutral or not. Multiple studies [13] indicated that "silent changes" disturb the efficacy of protein translation; while some codon usage patterns translated faster and more accurately than others. However, these findings do not explain the general trends reported by Hershberg and Petrov [13] indicating that codon biases in all organisms trend to strongly track intergene GC content. If selection at the level of translation was solely responsible for codon bias evolution, then why would any property of non-coding regions be related so strongly and uniformly to existing codon bias. Our findings illustrate some important properties of these synonymous changes by observing the average mutational impacts on DNA flexibility for each gene in broad scale. First, for each genome, we plotted the average mutational impacts for transition and transversion mutations to observe the higher scale of mutational impact. No clear patterns clearly appeared for non- synonymous sites; however, synonymous sites showed a strikingly common trend of having more much variation in dTRX, the mutational impact on flexibility, in genes functionally conserved at the protein level (i.e. low dN/dS). In genomes with highly skewed GC content and strong codon biases, genes with low functional constraint at the protein level (i.e. high dN/dS) always clustered towards zero dTRX, even when the genomic average dTRX was strongly positive or negative. For example, the high GC Pseudomonas genome (figure 9c) showed some overall shifts toward gaining more flexibility in the synonymous sites of these genes while the high AT genome of Prochlorococcus marinus(figure 9d) had an opposite overall shift toward losing flexibility. These general mutational shifts in the flexibility clearly indicate some basic constraints between evolution occurring at the levels of protein and genome architecture (i.e. flexibility) which have taken place in these genomes over time. Non-synonymous sites do not have the same property; there is almost no difference between conserved genes and selected genes. We assume that synonymous sites, which fall largely at third base positions, are more affected because of their general involvement in influencing flexibility through their defining of the minor groove width...previously noted. The possibility that synonymous sites may actually have some function in this process provides a possible explanation for why synonymous sites appear to be subject to weak natural selection in most genomes.

4.2 Codon bias is what allows genomes to obtain specific general levels of flexibility

Lastly, we demonstrated a fundamental relationship between intrinsic flexibility and codon bias entropy. Observing the scatter plot of the final averages of flexibility for each genome and linking that to their genome average codon bias, the strong associations of TRX score and codon bias actually reflect around the mid-point of the TRX scale. Genomes with uncharacteristically stiff or flexible genome architecture deviate from the middle point of the TRX scale (average scores of 10 different types of dinucleotide) and correlate well with codon bias entropy on both sides. This clearly demonstrates the strong association in codon bias with flexibility and supports previous single gene findings. In addition, genome-wide GC content and DNA flexibility are extremely well correlated indicating that codon biases are probably designed towards GC or AT preferences at the third codon positions [13]. We have extended the interpretation of the significance of genome-wide GC content by linking it directly to dynamical properties of the DNA itself. These strong correlations allowed us to demonstrate that a codon bias variation in prokaryotes is highly related to genome flexibility. Some bacteria tends to have either stiffer or flexible genomes which is probably based on the environment that needs a specific genome structure; thus codon bias would be the optimal tool to acquire that.

5.Conclusion:

In recent years, we have been able to observe and sequence many genomes with a wide variety of structural nuclear architectures across a broad array of life on this Earth. DNA sequence, as we know, is the fundamental basis for life and yet we are still learning very important things about it. Moreover, understanding the structural variations in the DNA double helix among organisms, and how it is relates to Codon bias provide valuable keys to discovering the extra information encoded within each sequence; thus opening new areas of research aimed toward unlocking more mysteries regarding the many patterns discovered within genomes by modern bioinformaticists. One way we can define these structural variations is through flexibility or how much we can bend or twist this tiny and important molecule. In this study, we tracked the mutational impacts in the flexibility for a large spectrum of 22 prokaryotes and 2 Achaea clusters of genomes and observed a fundamental relationship with codon bias. We conclude as has been noted by others working recently along similar lines, that DNA actually has a tremendous capicity to encode for its own packaging and regulation in the cell[2]. Furthermore, and by using codon bias to affect its flexibility DNA can directly control how easily this relative stiff molecular polymer can be packed within the prokaryotic cell and probably eukayotic chromatin as well.

References:

1. Akashi, H. "Synonymous Codon Usage in Drosophila Melanogaster: Natural Selection and Translational Accuracy." *Genetics* 136, no. 3 (Mar 1994): 927-35.
2. Alberts, Bruce. *Molecular Biology of the Cell*. 5th ed. 1 vols New York: Garland Science, 2008.
3. Angellotti, M. C., S. B. Bhuiyan, G. Chen, and X. F. Wan. "Codono: Codon Usage Bias Analysis within and across Genomes." *Nucleic Acids Res* 35, no. Web Server issue (Jul 2007): W132-6.
4. Babbitt, G. A., and K. V. Schulze. "Codons Support the Maintenance of Intrinsic DNA Polymer Flexibility over Evolutionary Timescales." [In eng]. *Genome Biol Evol* 4, no. 9 (2012): 954-65.
5. Baisnée, P. F., P. Baldi, S. Brunak, and A. G. Pedersen. "Flexibility of the Genetic Code with Respect to DNA Structure." [In eng]. *Bioinformatics* 17, no. 3 (Mar 2001): 237-48.
6. Behura, S. K., and D. W. Severson. "Codon Usage Bias: Causative Factors, Quantification Methods and Genome-Wide Patterns: With Emphasis on Insect Genomes." [In English]. *Biological Reviews* 88, no. 1 (Feb 2013): 49-61.
7. Botzman, M., and H. Margalit. "Variation in Global Codon Usage Bias among Prokaryotic Organisms Is Associated with Their Lifestyles." *Genome Biol* 12, no. 10 (2011): R109.
8. Fraser, H. B., A. E. Hirsh, D. P. Wall, and M. B. Eisen. "Coevolution of Gene Expression among Interacting Proteins." [In eng]. *Proc Natl Acad Sci U S A* 101, no. 24 (Jun 2004): 9033-8.
9. Gascuel, O. "Bionj: An Improved Version of the Nj Algorithm Based on a Simple Model of Sequence Data." [In eng]. *Mol Biol Evol* 14, no. 7 (Jul 1997): 685-95.
10. Gouy, M., and C. Gautier. "Codon Usage in Bacteria: Correlation with Gene Expressivity." *Nucleic Acids Res* 10, no. 22 (Nov 25 1982): 7055-74.
11. Guo, F. B., Y. N. Ye, H. L. Zhao, D. Lin, and W. Wei. "Universal Pattern and Diverse Strengths of Successive Synonymous Codon Bias in Three Domains of Life, Particularly among Prokaryotic Genomes." [In English]. *DNA Research* 19, no. 6 (Dec 2012): 477-85.
12. Heddi, B., C. Oguey, C. Lavelle, N. Foloppe, and B. Hartmann. "Intrinsic Flexibility of B-DNA: The Experimental Trx Scale." [In eng]. *Nucleic Acids Res* 38, no. 3 (Jan 2010): 1034-47.

13. Hershberg, R., and D. A. Petrov. "Selection on Codon Bias." [In eng]. *Annu Rev Genet* 42 (2008): 287-99.

14. Ikemura, T. "Codon Usage and Trna Content in Unicellular and Multicellular Organisms." *Mol Biol Evol* 2, no. 1 (Jan 1985): 13-34.

15. Itzkovitz, S., and U. Alon. "The Genetic Code Is Nearly Optimal for Allowing Additional Information within Protein-Coding Sequences." *Genome Res* 17, no. 4 (Apr 2007): 405-12.

16. Keeling, P. J., and W. F. Doolittle. "Archaea: Narrowing the Gap between Prokaryotes and Eukaryotes." [In eng]. *Proc Natl Acad Sci U S A* 92, no. 13 (Jun 1995): 5761-4.

17. Kumar, S., G. Stecher, D. Peterson, and K. Tamura. "Mega-Cc: Computing Core of Molecular Evolutionary Genetics Analysis Program for Automated and Iterative Data Analysis." [In eng]. *Bioinformatics* 28, no. 20 (Oct 2012): 2685-6.

18. Ma, J., L. Bai, and M. D. Wang. "Transcription under Torsion." *Science* 340, no. 6140 (Jun 28 2013): 1580-3.

19. Nabiyouni, M., A. Prakash, and A. Fedorov. "Vertebrate Codon Bias Indicates a Highly Gc-Rich Ancestral Genome." [In English]. *Gene* 519, no. 1 (Apr 2013): 113-19.

20. Novichkov, P. S., I. Ratnere, Y. I. Wolf, E. V. Koonin, and I. Dubchak. "Atgc: A Database of Orthologous Genes from Closely Related Prokaryotic Genomes and a Research Platform for Microevolution of Prokaryotes." [In eng]. *Nucleic Acids Res* 37, no. Database issue (Jan 2009): D448-54.

21. Pandit, A., and S. Sinha. "Differential Trends in the Codon Usage Patterns in Hiv-1 Genes." [In English]. *Plos One* 6, no. 12 (Dec 2011): 10.

22. Pascal Carrivain, Axel Cournac, Christophe Lavelle, Annick Lesne,abd Julien Mozziconacci, Fabien Paillusson, Laurence Signon, Jean-Marc Victor and Maria Barbi. "Electrostatics of DNA Compaction in Viruses, Bacteria and Eukaryotes: Functional Insights and Evolutionary Perspective." 2012.

23. Singer, Gregory A. C., and Donal A. Hickey. "Thermophilic Prokaryotes Have Characteristic Patterns of Codon Usage, Amino Acid Composition and Nucleotide Content." *Gene* 317, no. 0 (10/23/ 2003): 39-47. Suzuki, H., R. Saito, and M. Tomita. "The 'Weighted Sum of Relative Entropy': A New Index for Synonymous Codon Usage Bias." *Gene* 335 (Jun 23 2004): 19-23.

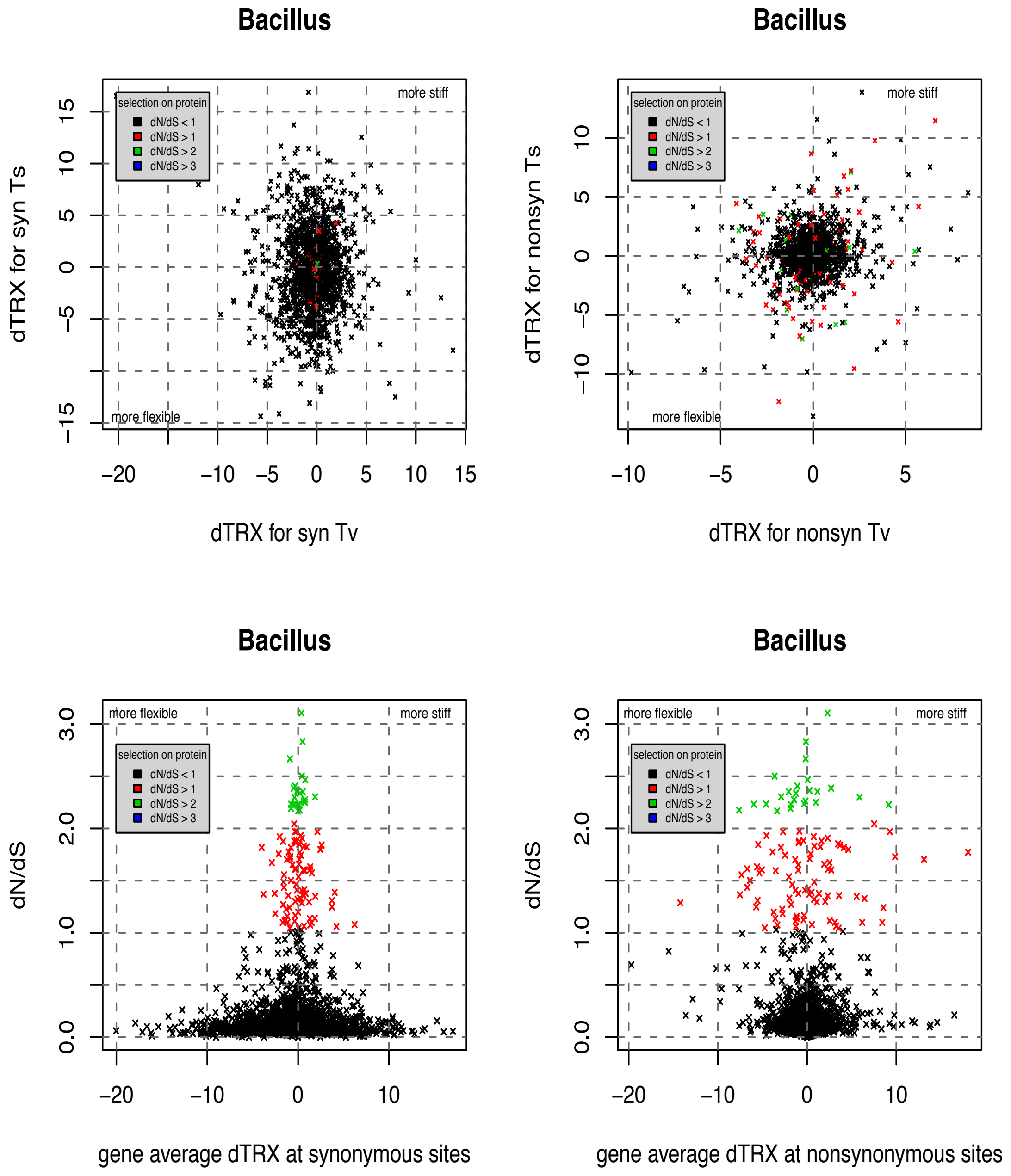
24. Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. "Mega5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary

Distance, and Maximum Parsimony Methods." [In eng]. *Mol Biol Evol* 28, no. 10 (Oct 2011): 2731-9.

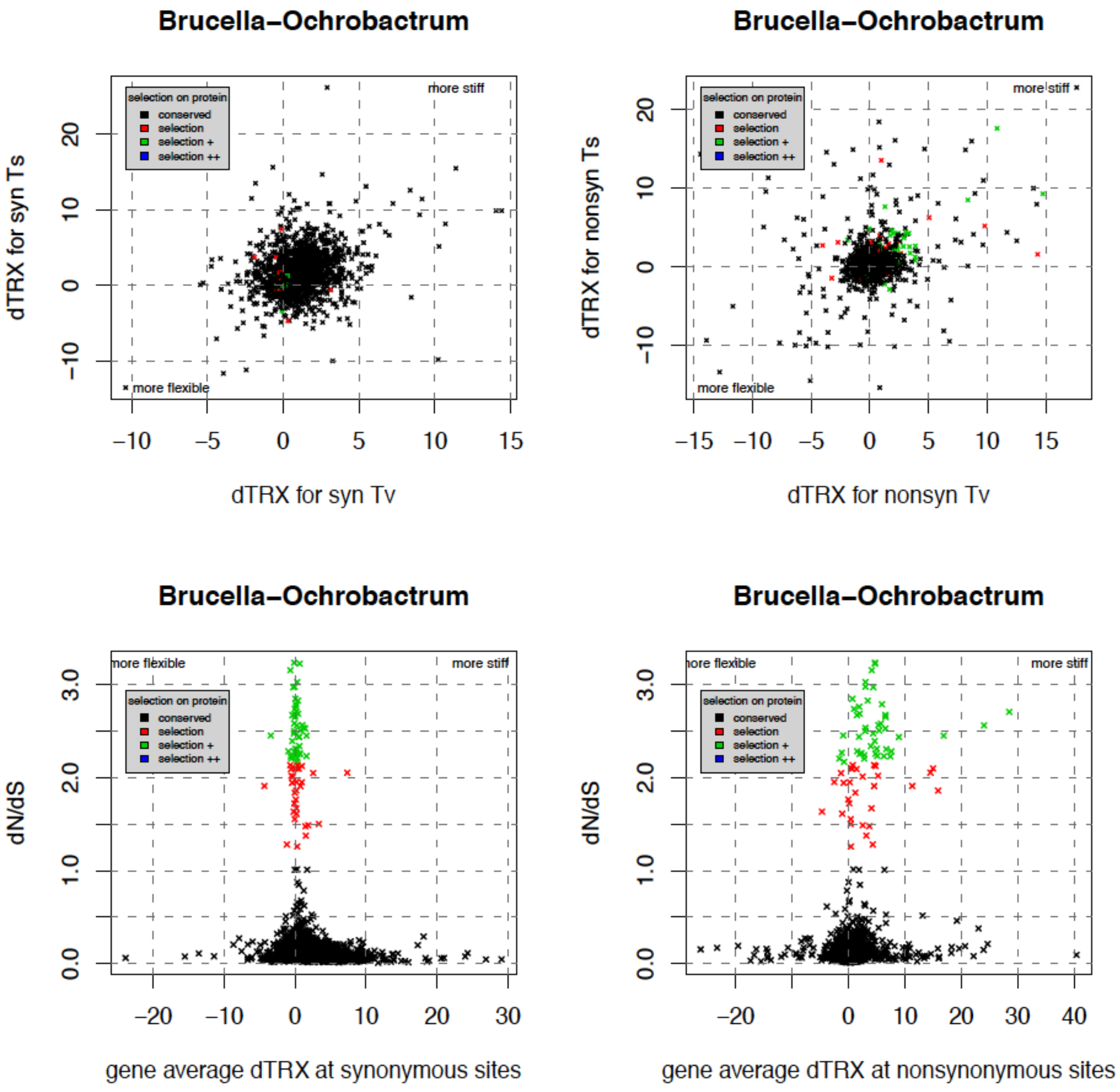
25. Yang, Z. "Paml 4: Phylogenetic Analysis by Maximum Likelihood." [In eng]. *Mol Biol Evol* 24, no. 8 (Aug 2007): 1586-91.
26. Yang, Z., S. Kumar, and M. Nei. "A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences." [In eng]. *Genetics* 141, no. 4 (Dec 1995): 1641-50.

Appendix:

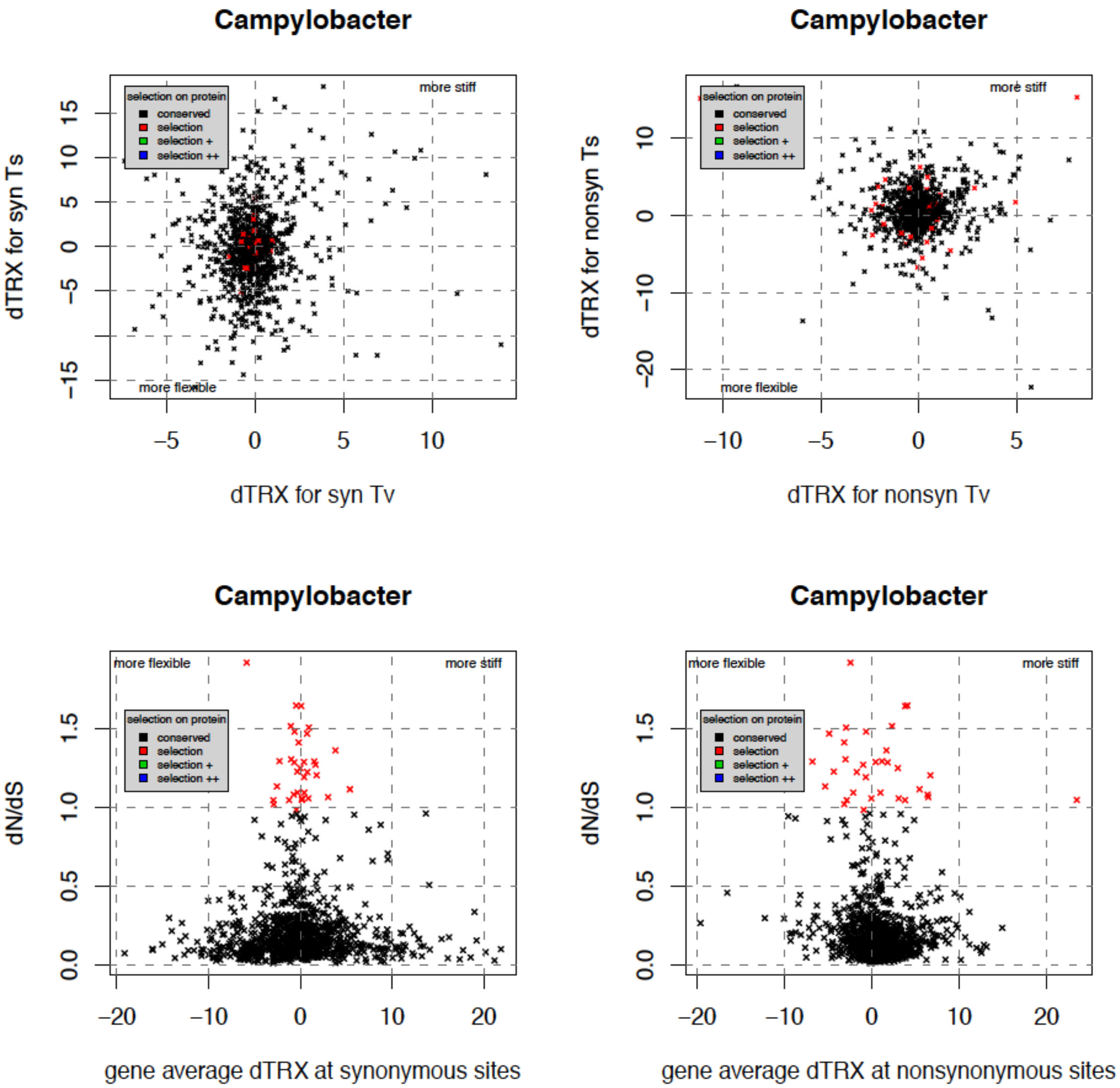
1.Bacillus



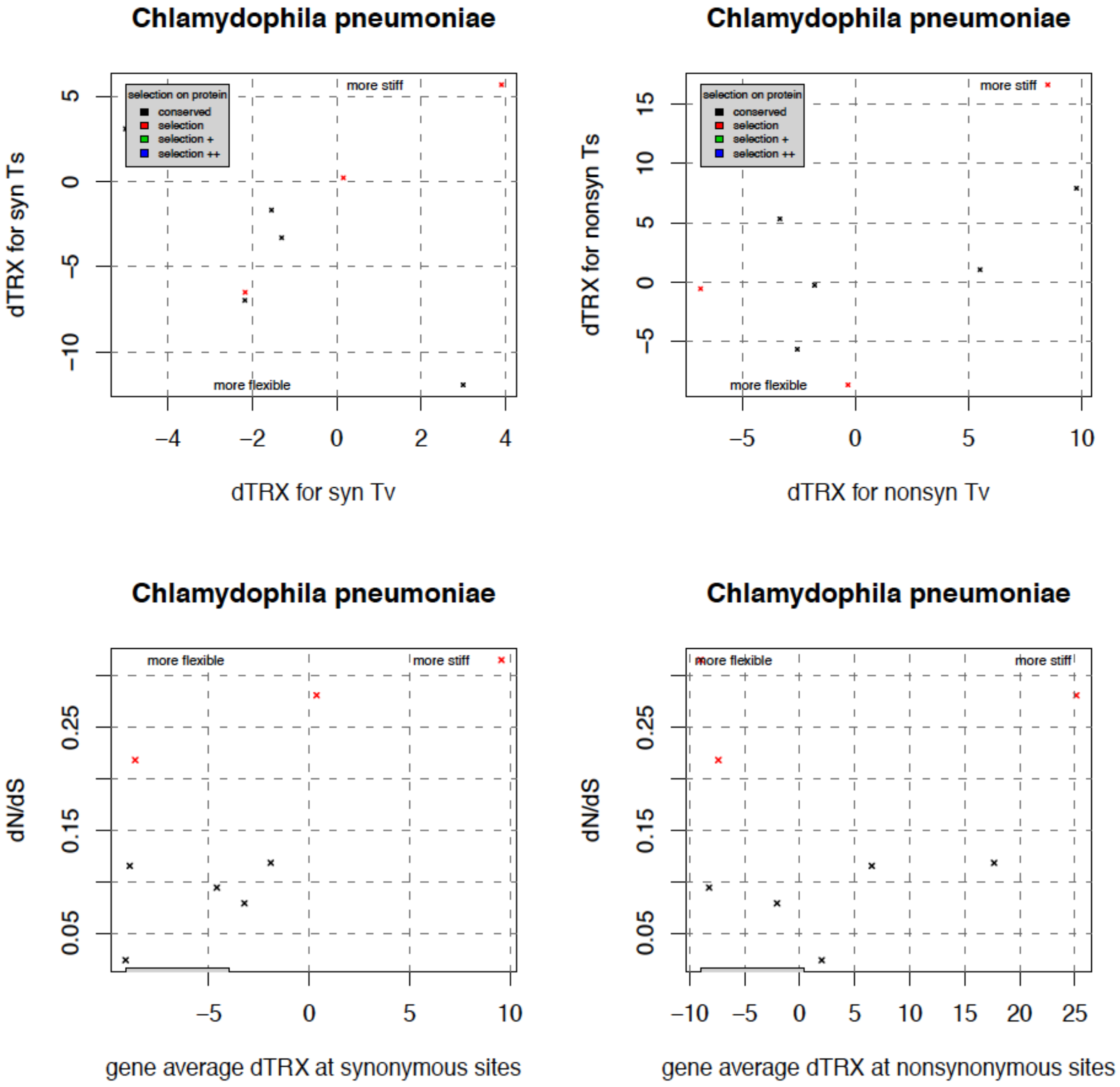
2. Brucella-Ochrobactrum



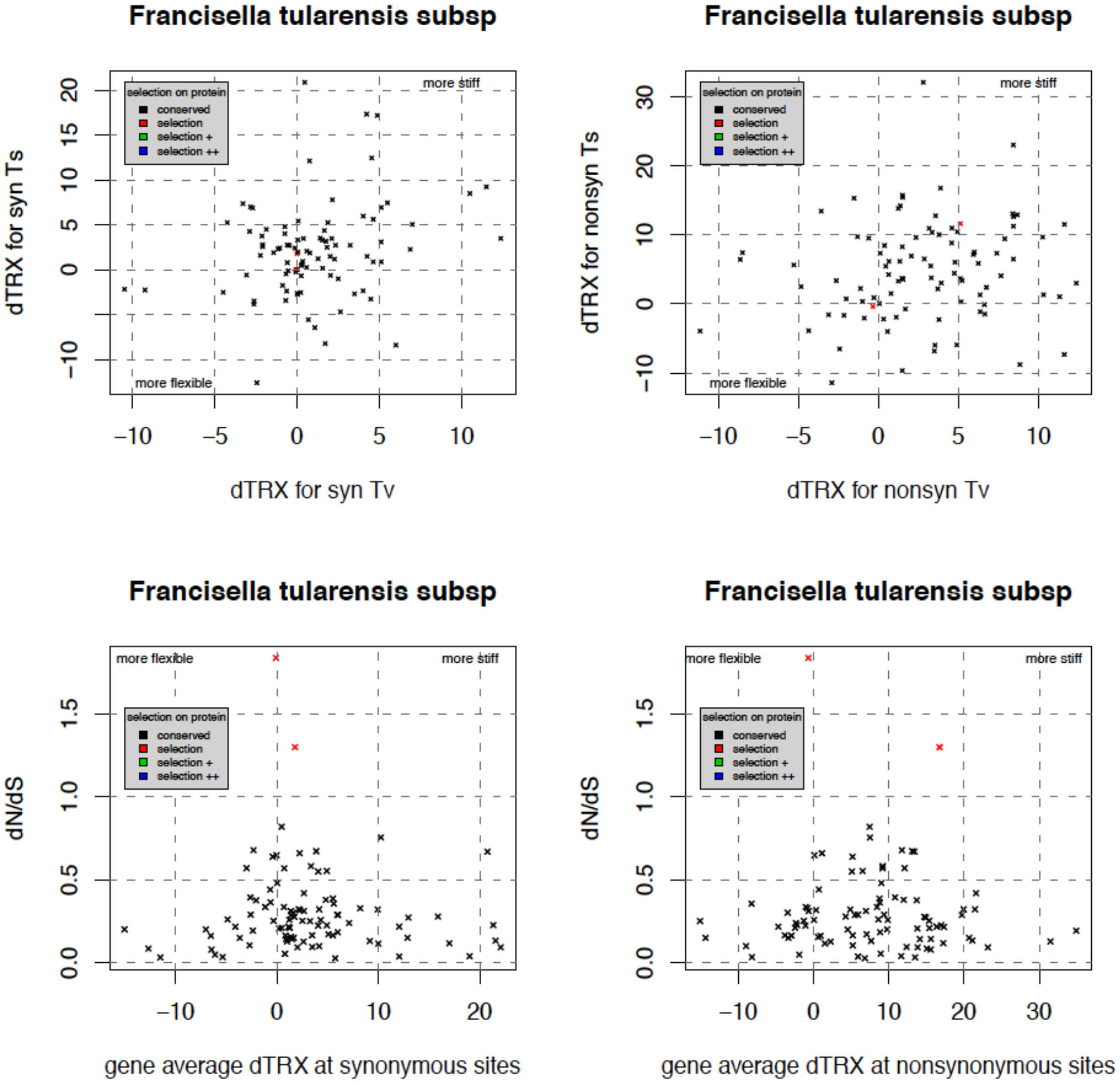
3. Campylobacter



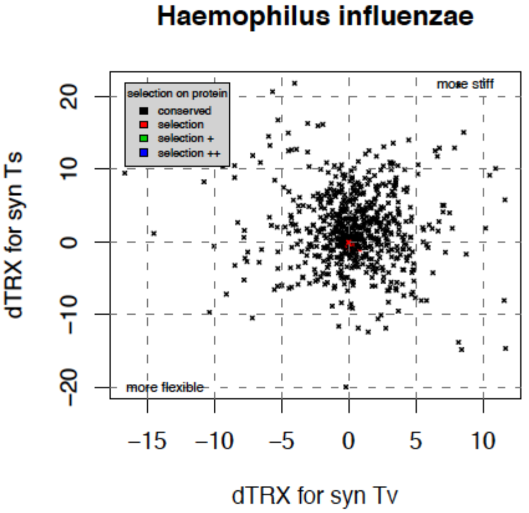
4. Chlamydophila pneumonia



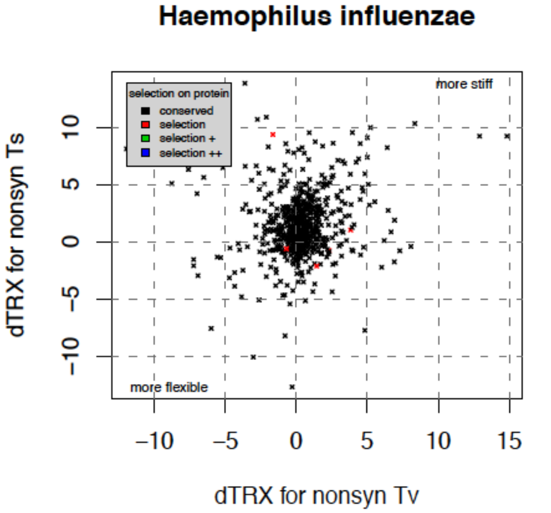
5. Francisella tularensis subsp



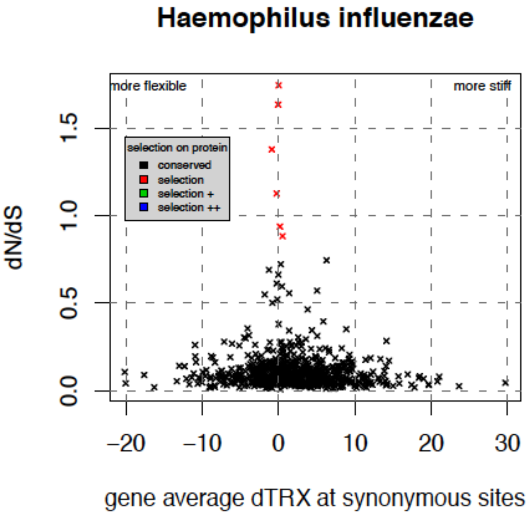
6. Haemophilus influenza



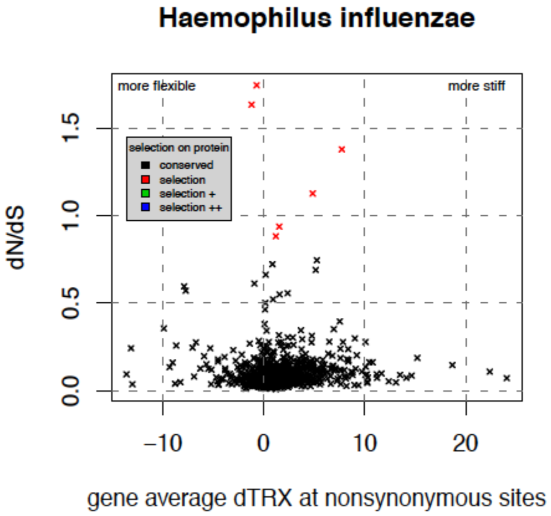
Haemophilus influenzae



Haemophilus influenzae

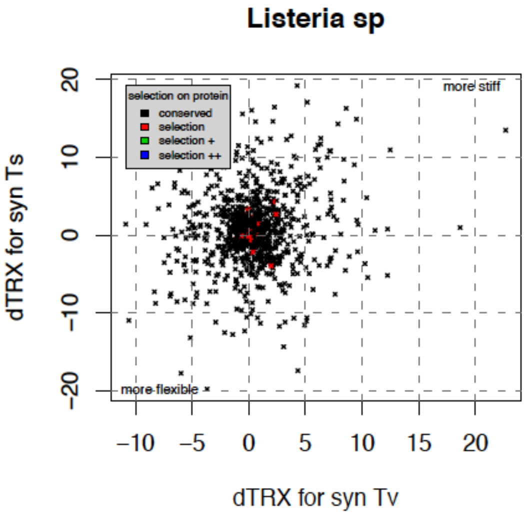


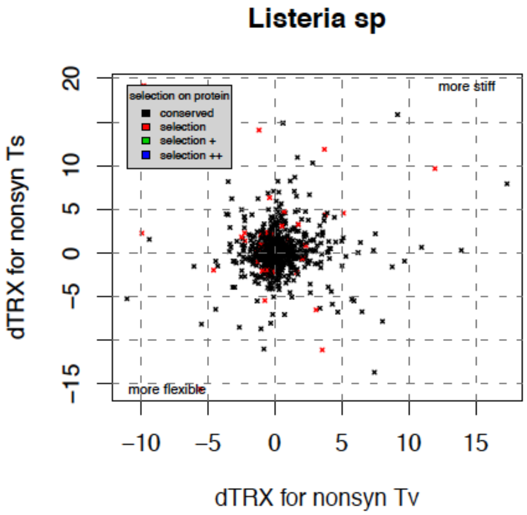
Haemophilus influenzae

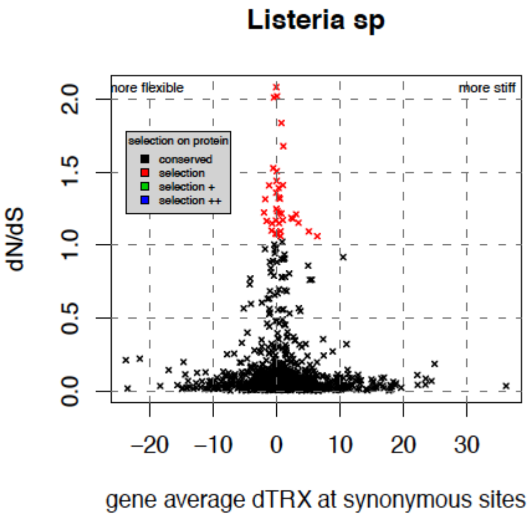


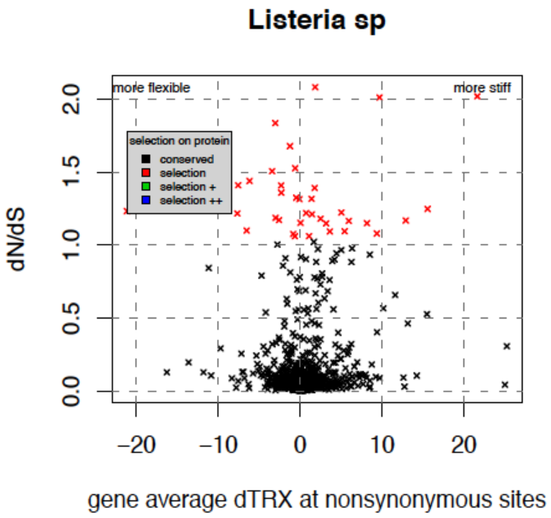
Haemophilus influenzae

7.Listeria sp

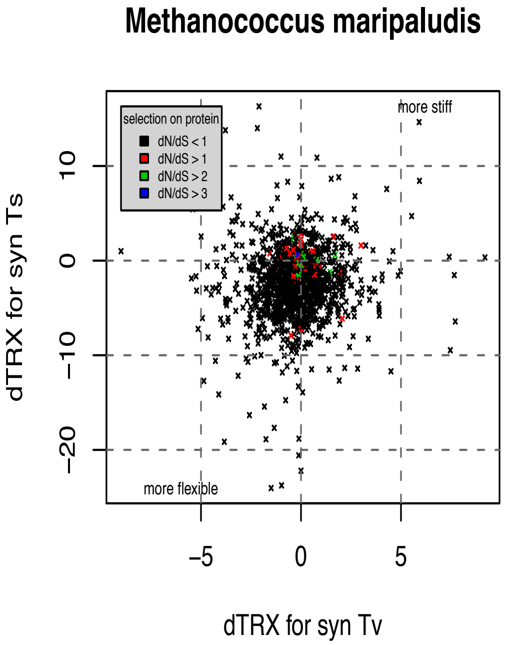


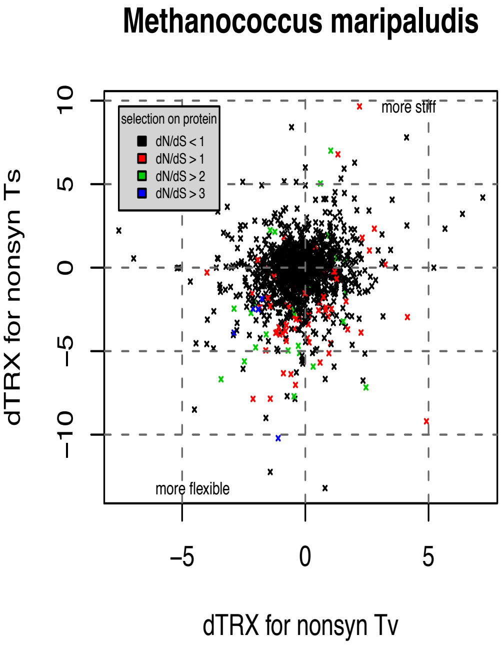


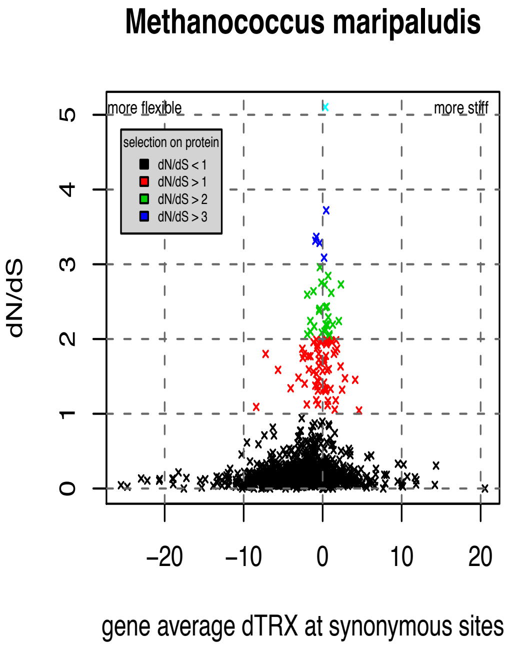


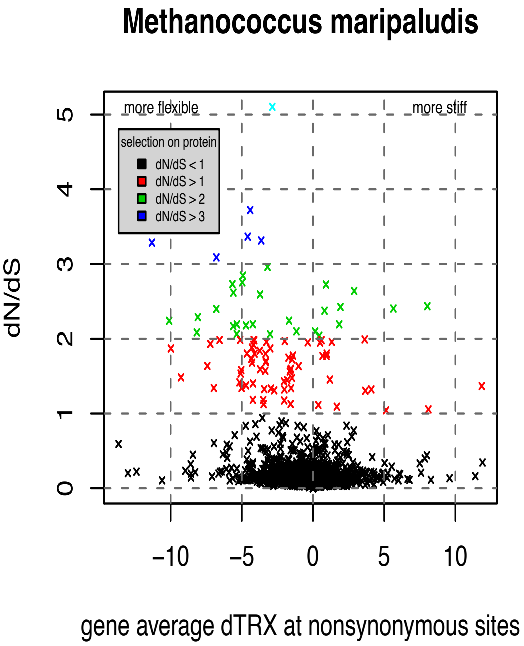


8.Methanocococcus maripaludis

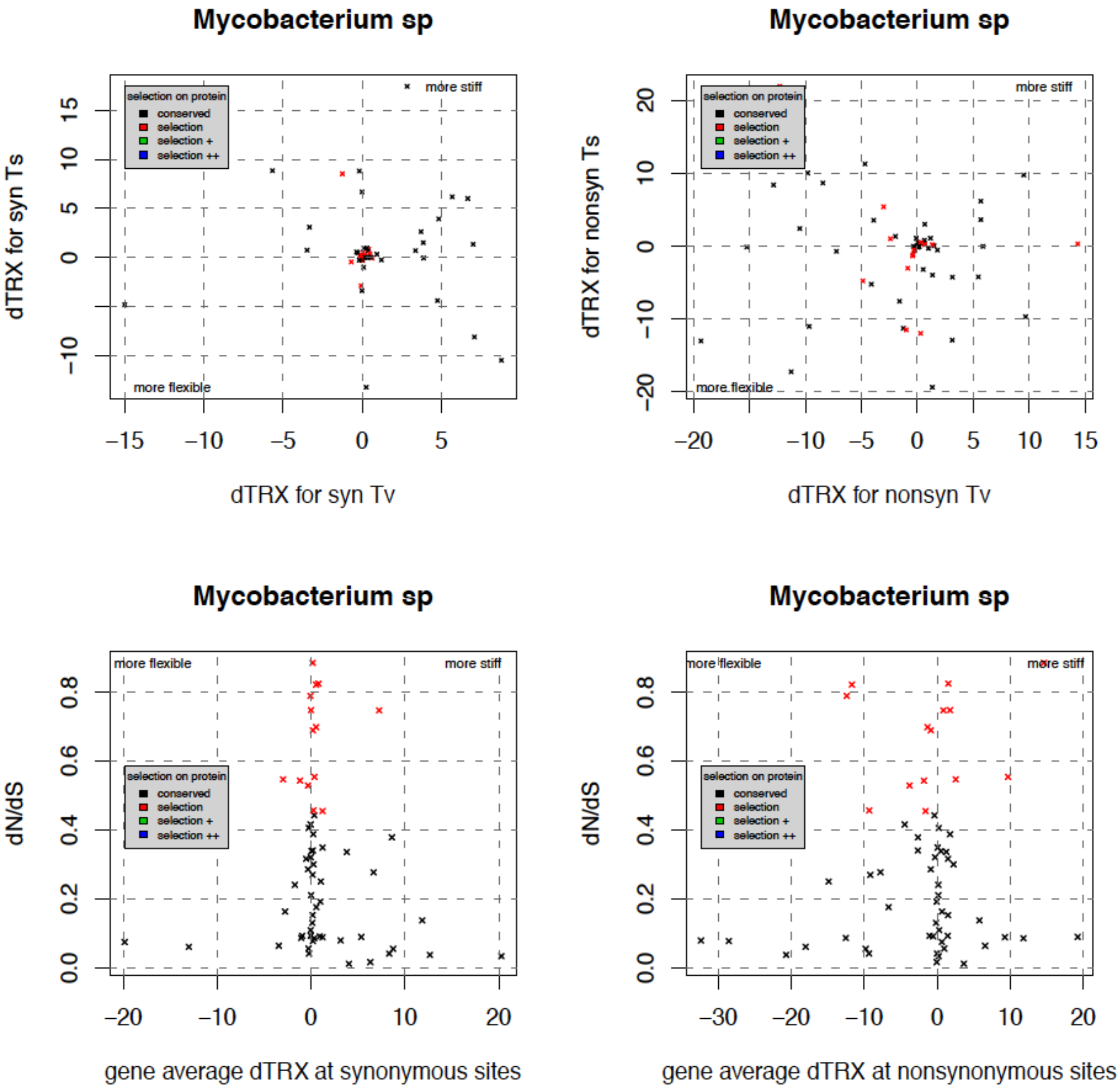




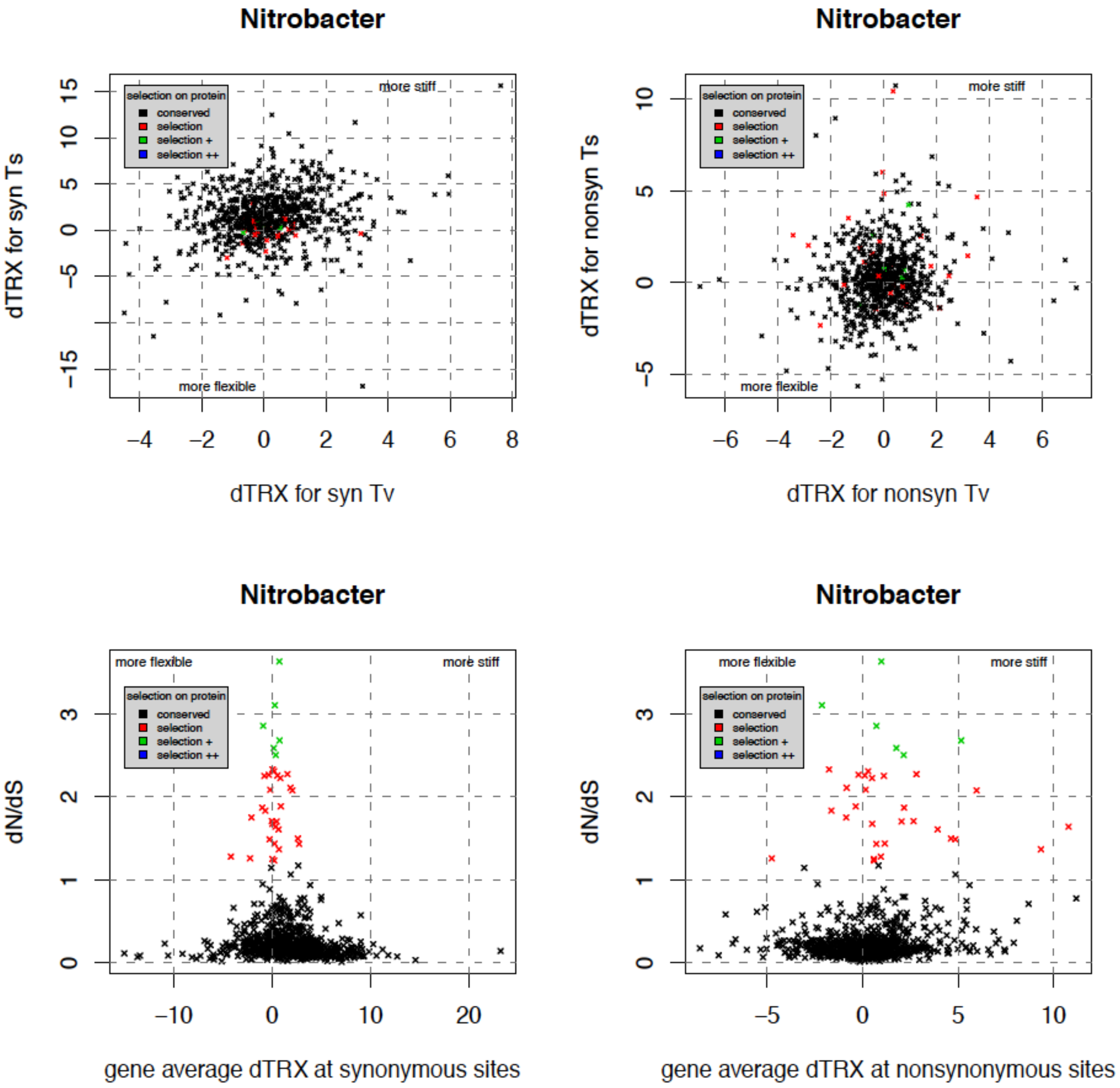




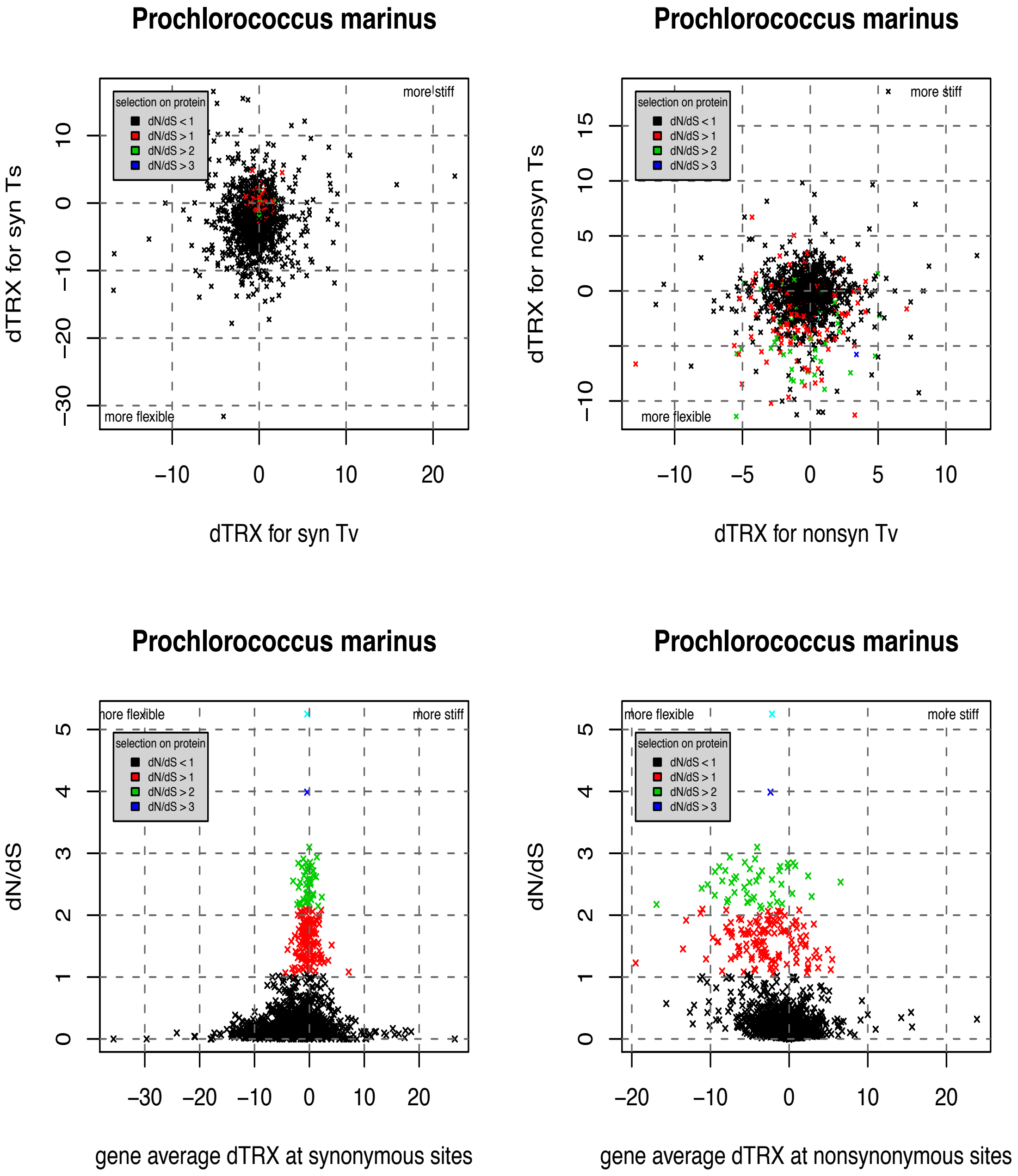
9.Mycobacterium sp



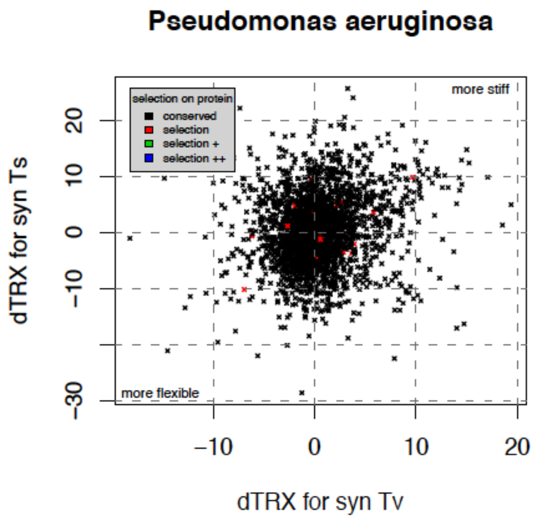
10. Nitrobacter

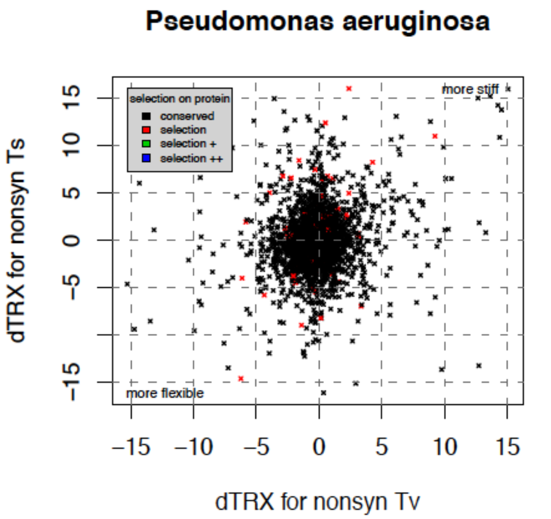


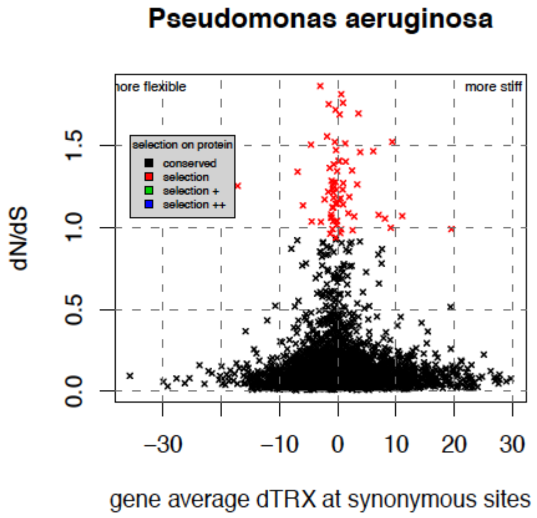
11.Prochlorococcus marinus

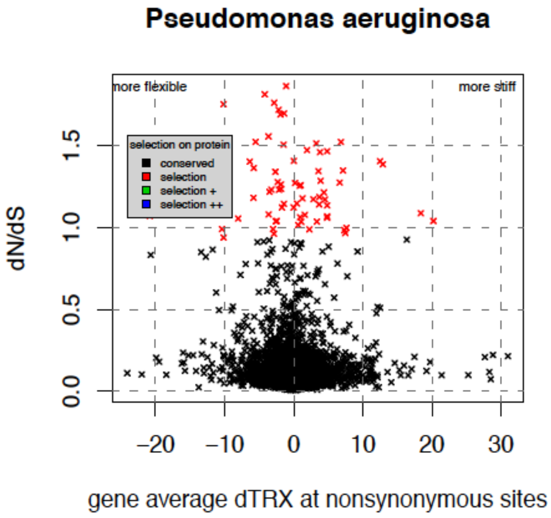


12. Pseudomonas aeruginosa

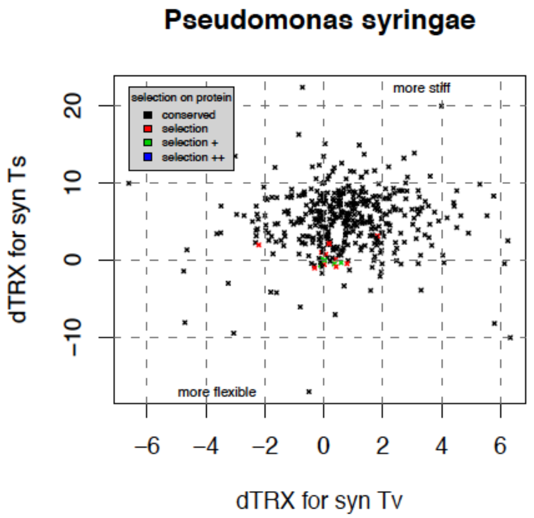


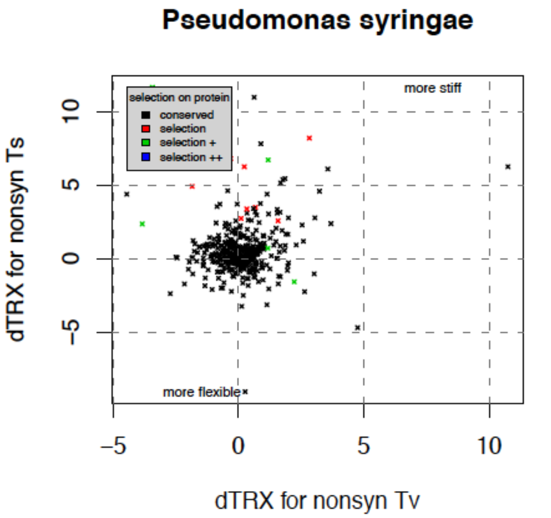


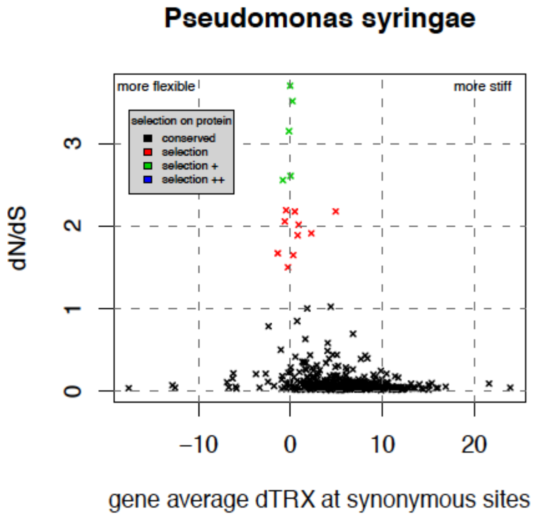


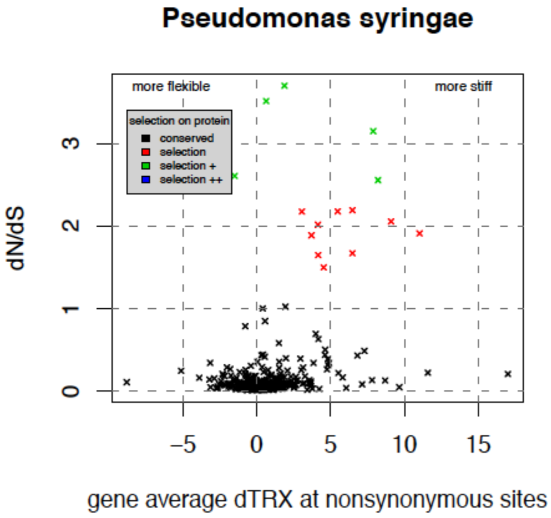


13. Pseudomonas syringae

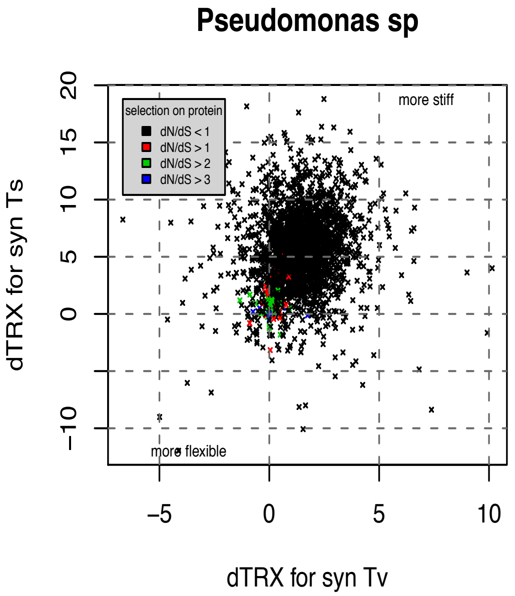


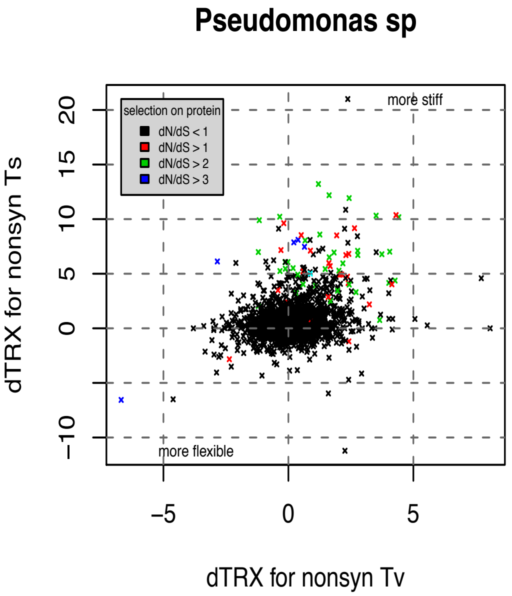


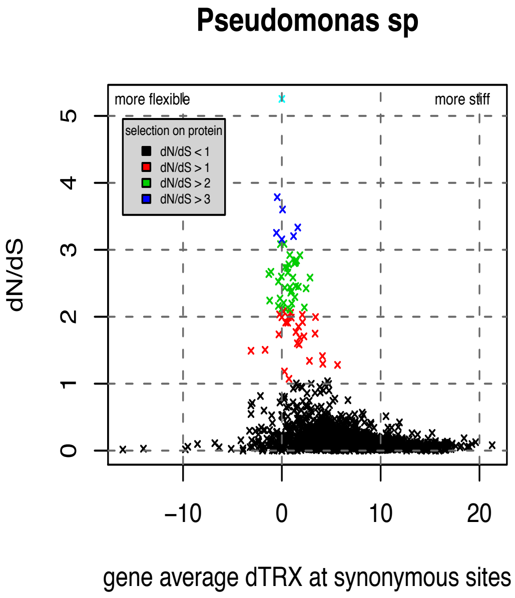


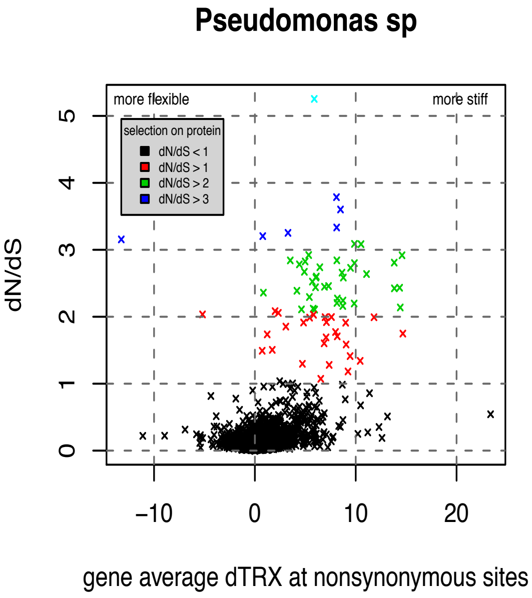


14. Pseudomonas sp

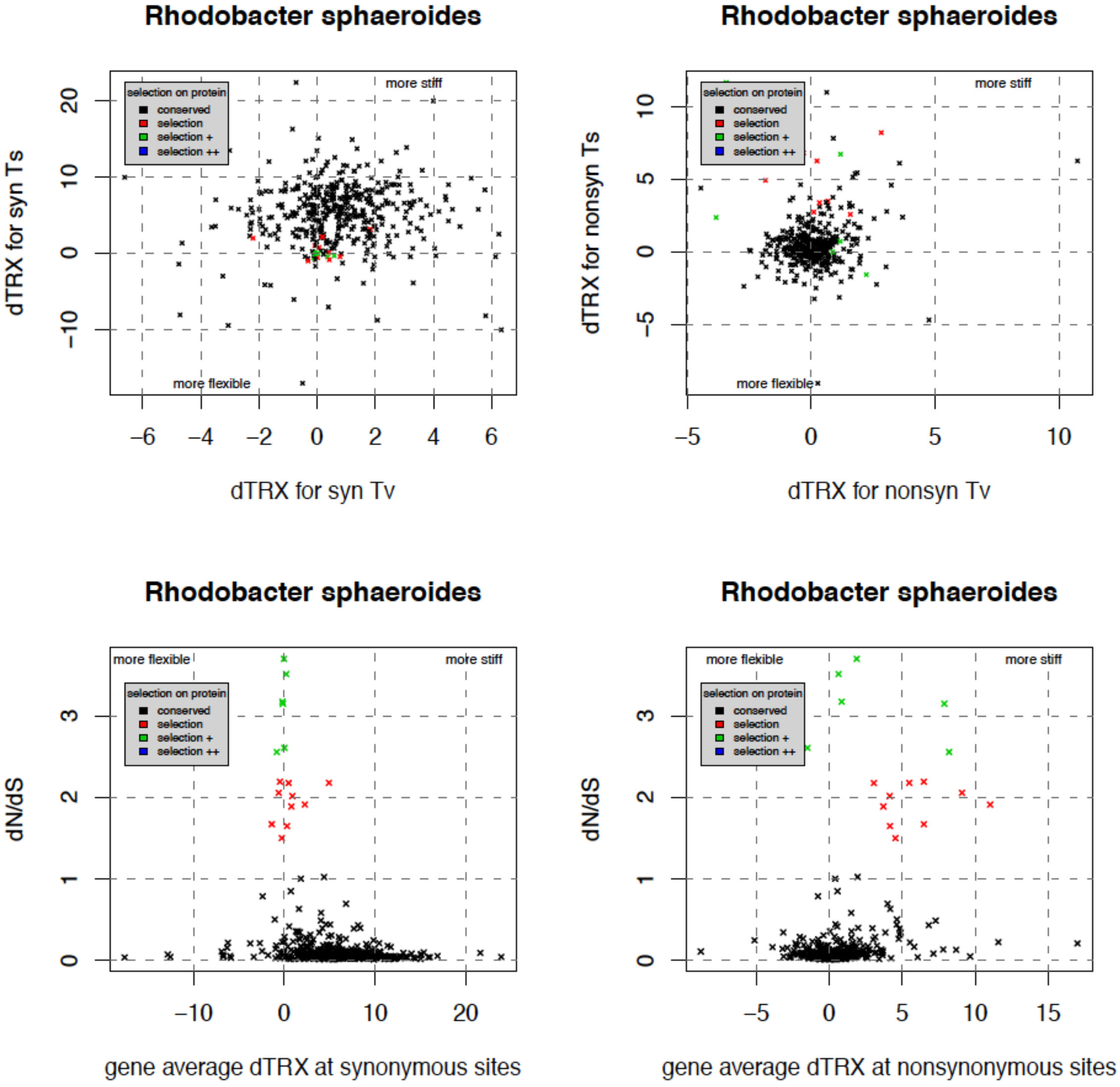




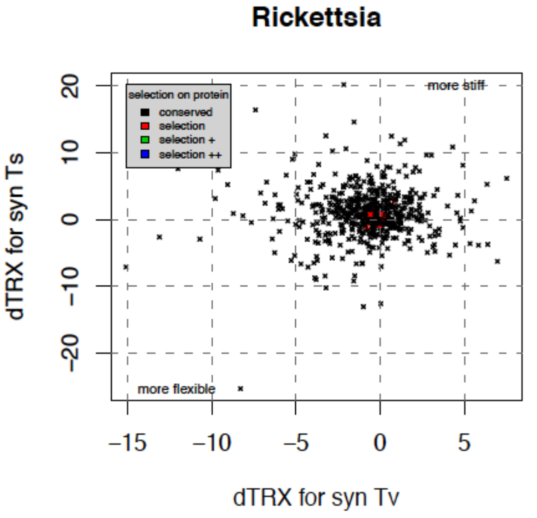


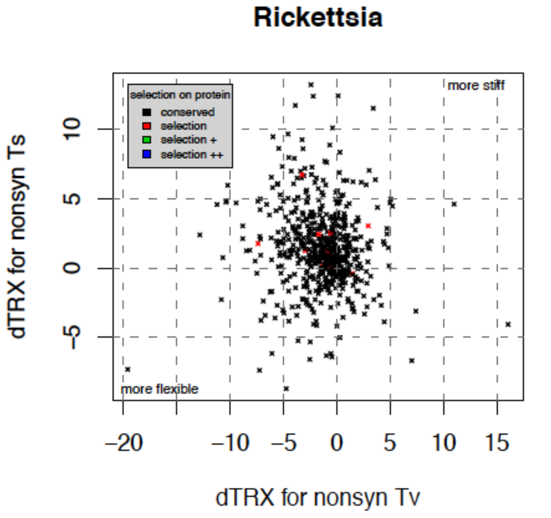


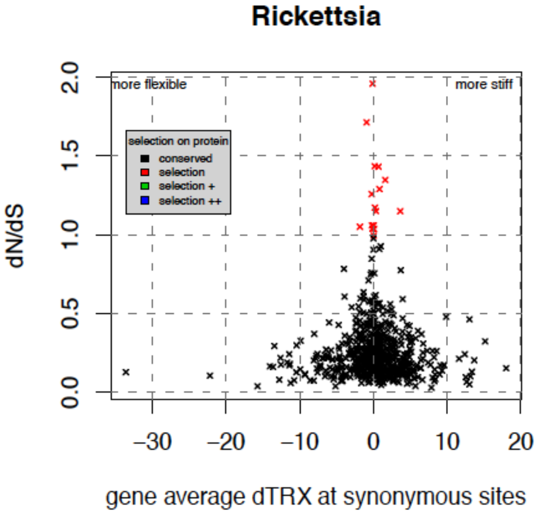
15. Rhodobacter sphaeroides

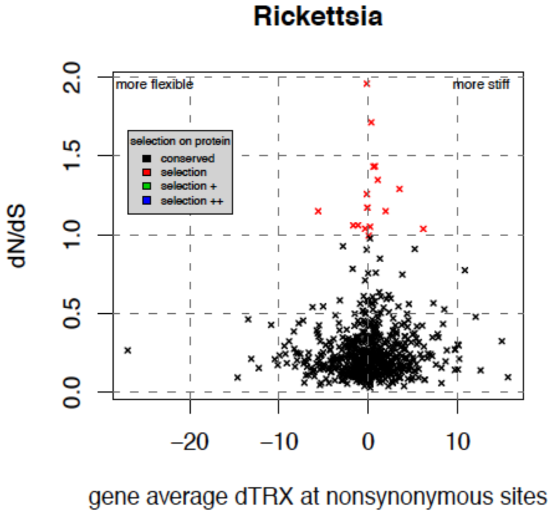


16. Rickettsia

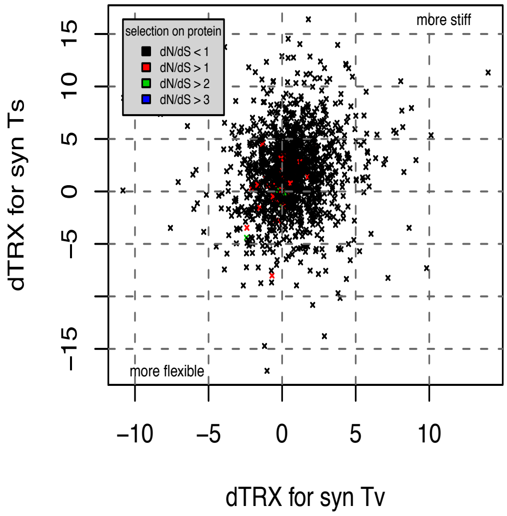

Rickettsia

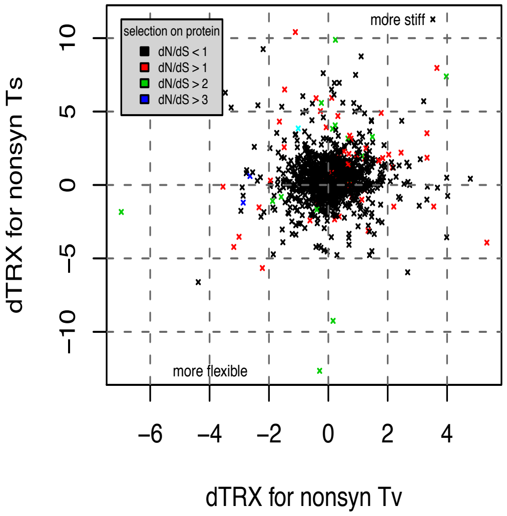

Rickettsia

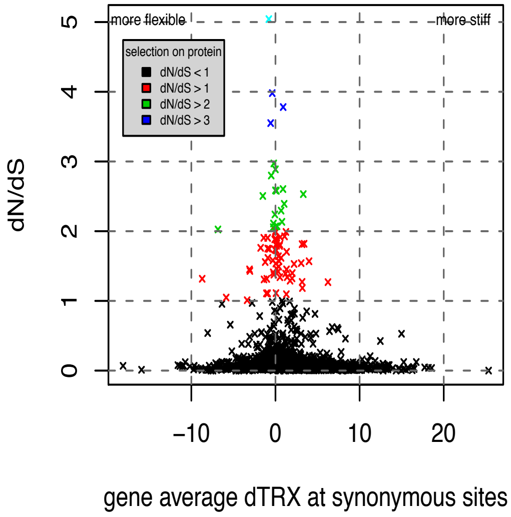

Rickettsia

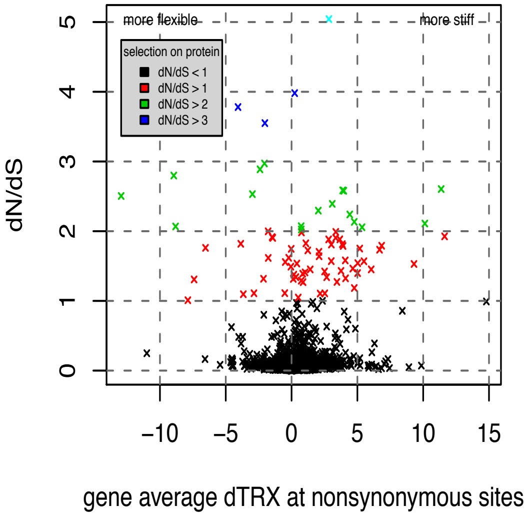

Rickettsia

17.Shewanella sp

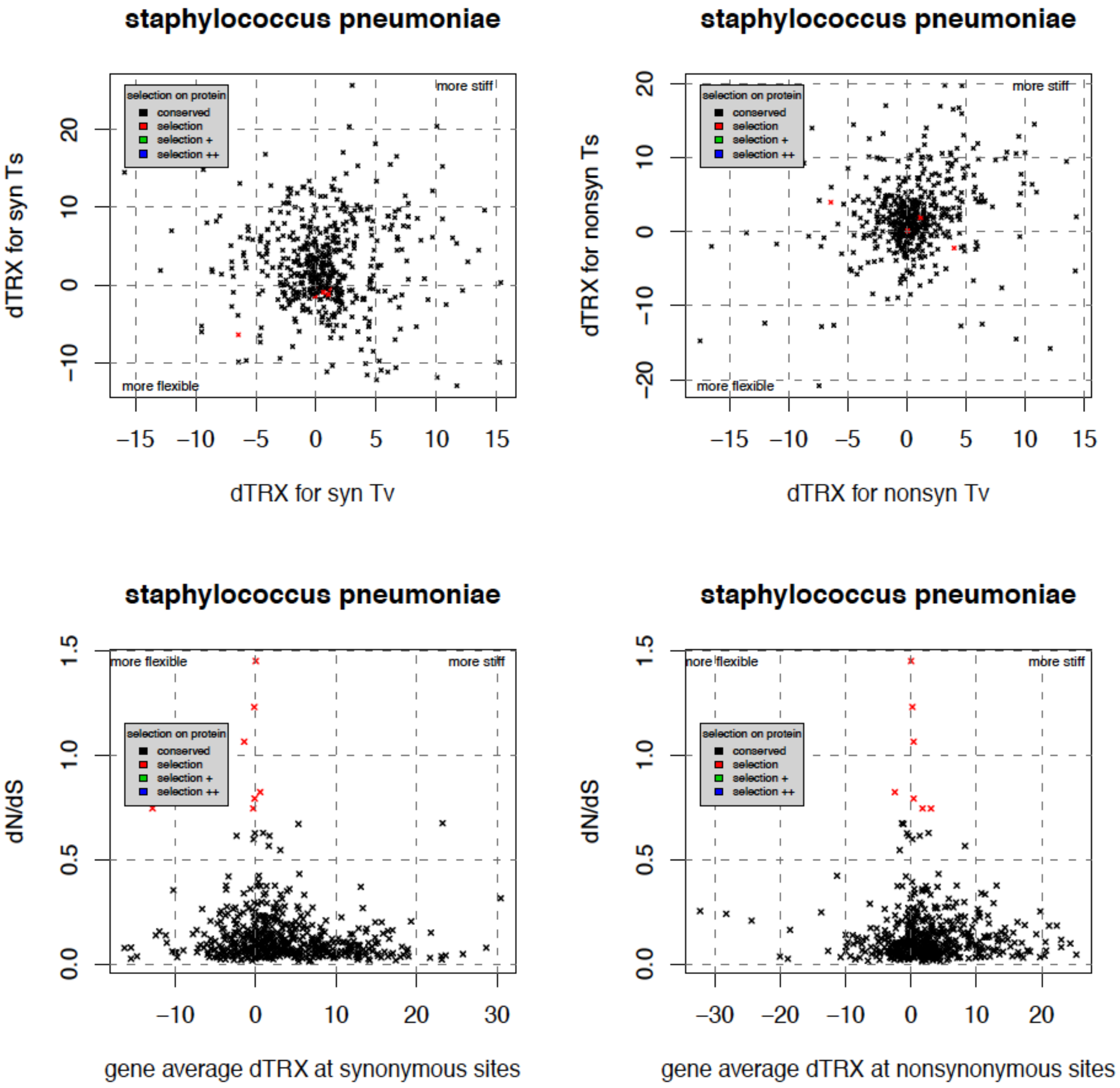

Shewanella sp


Shewanella sp

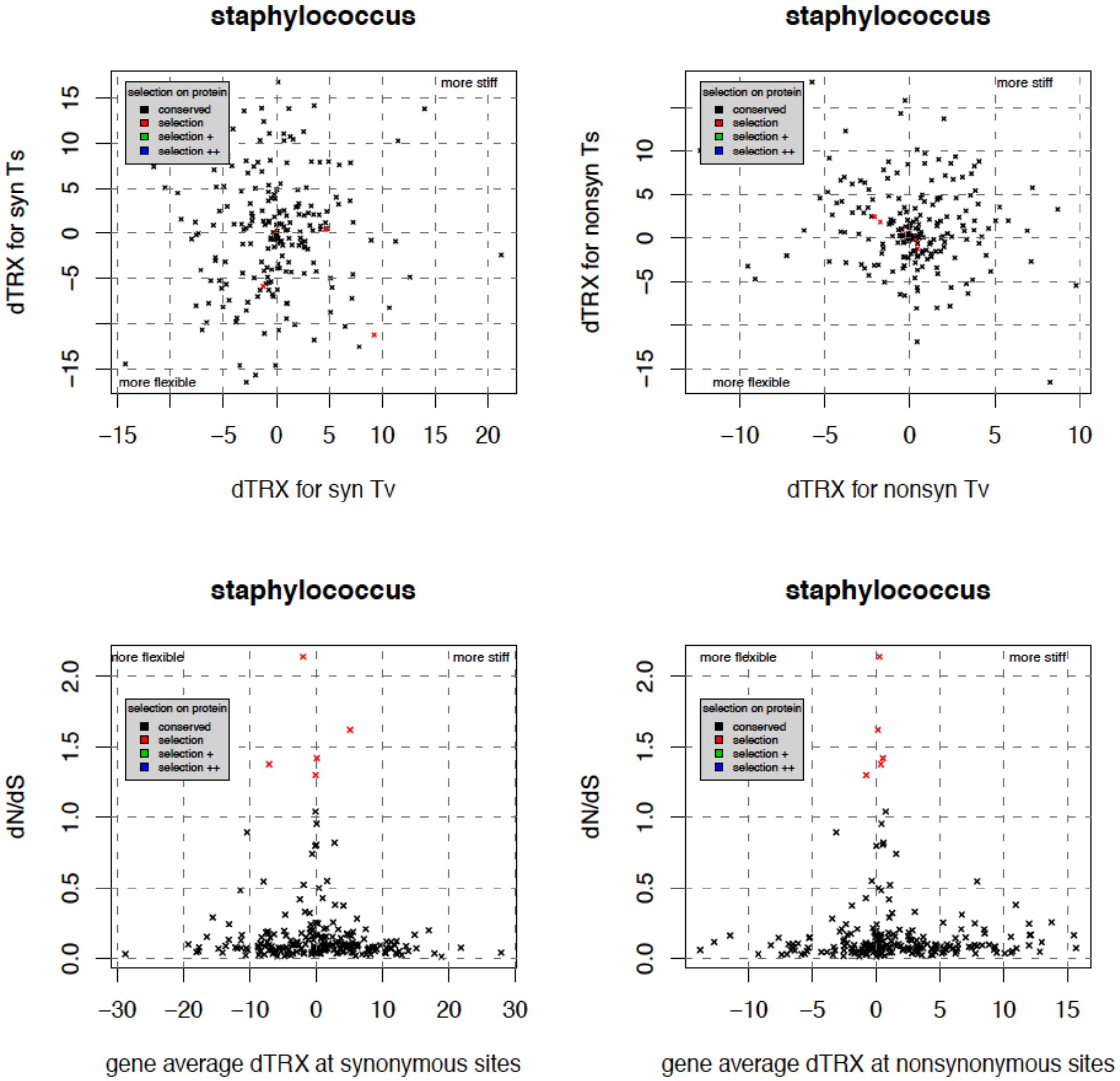

Shewanella sp


Shewanella sp

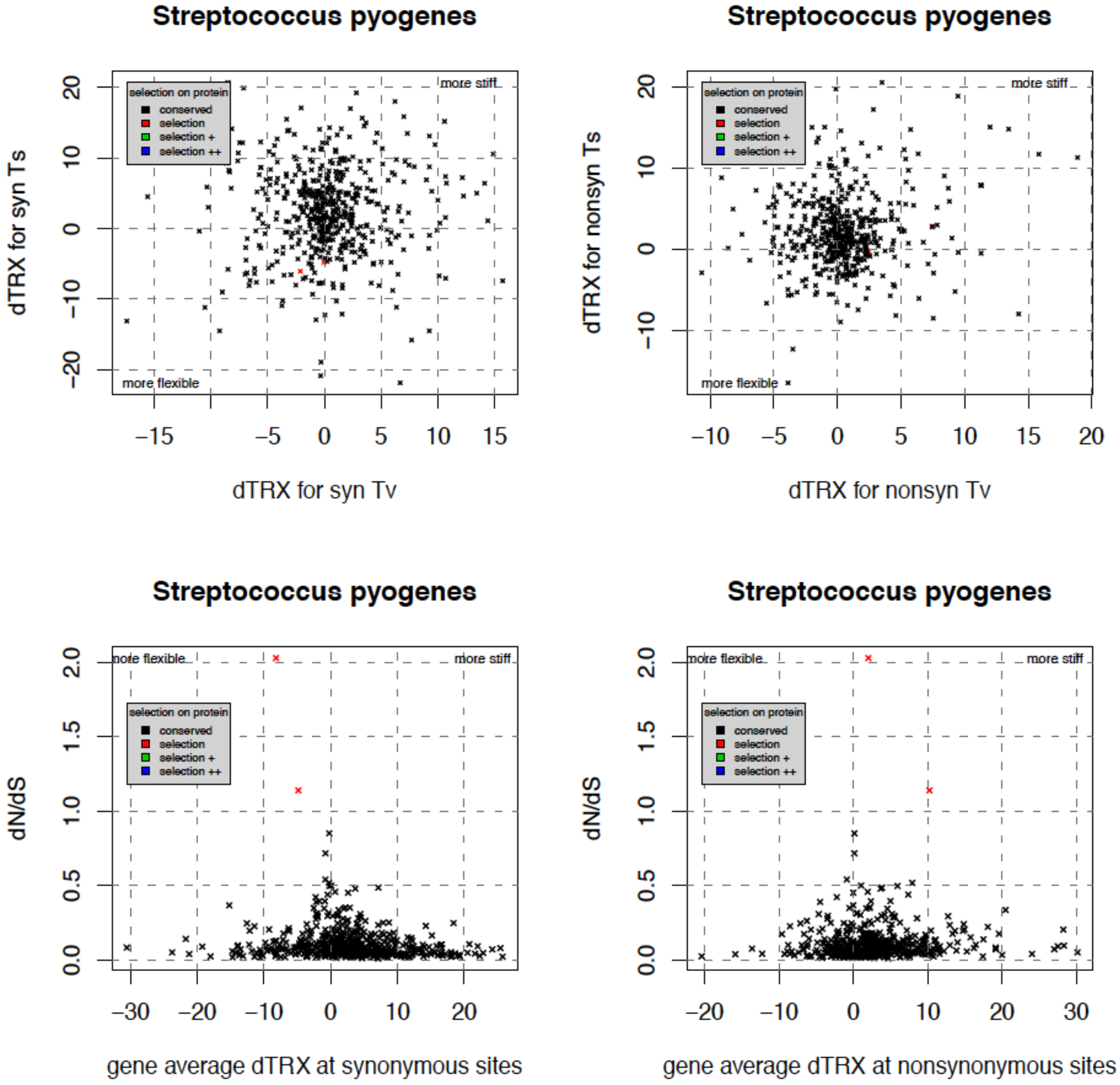
18. staphylococcus pneumonia



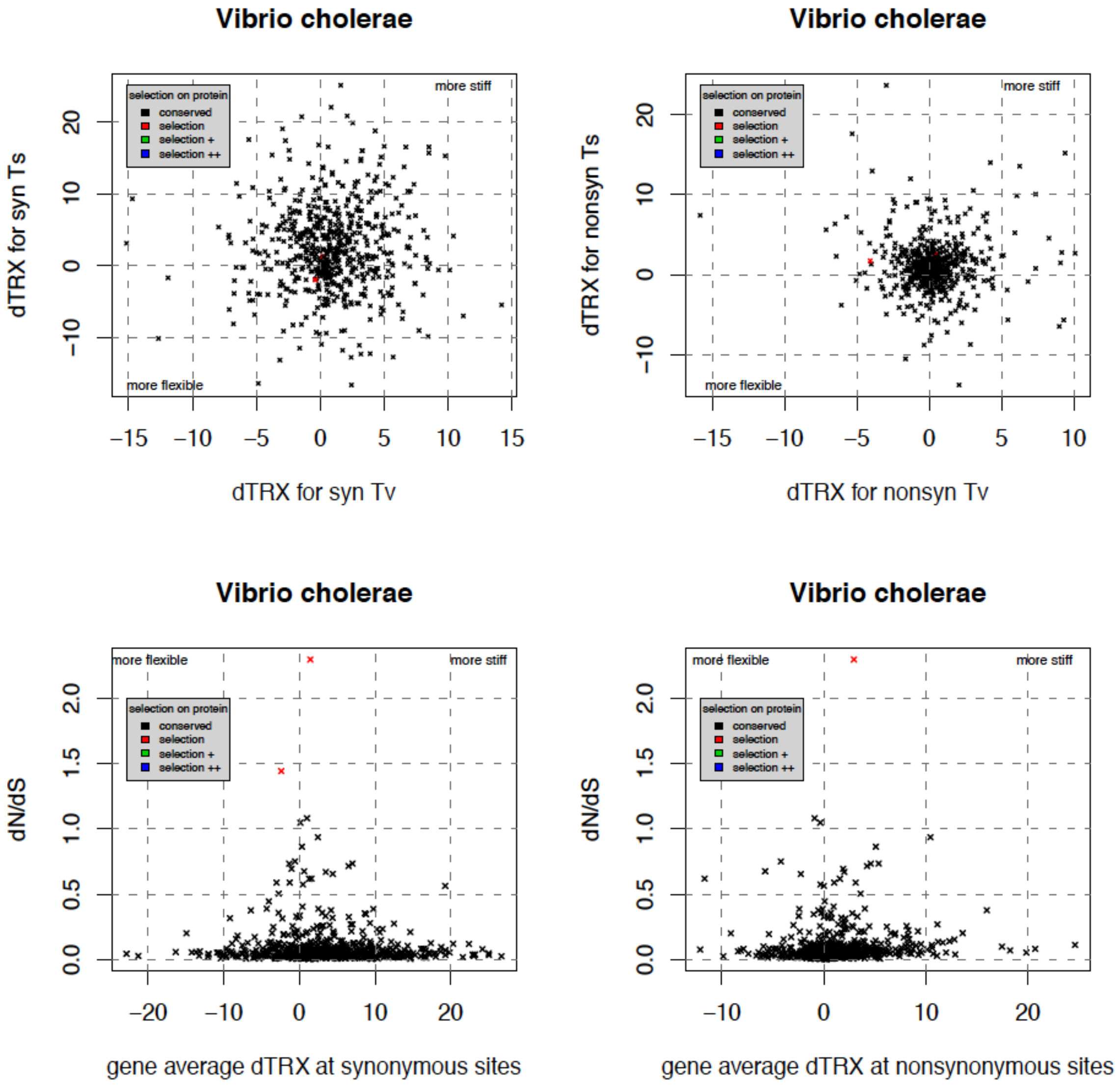
19. staphylococcus



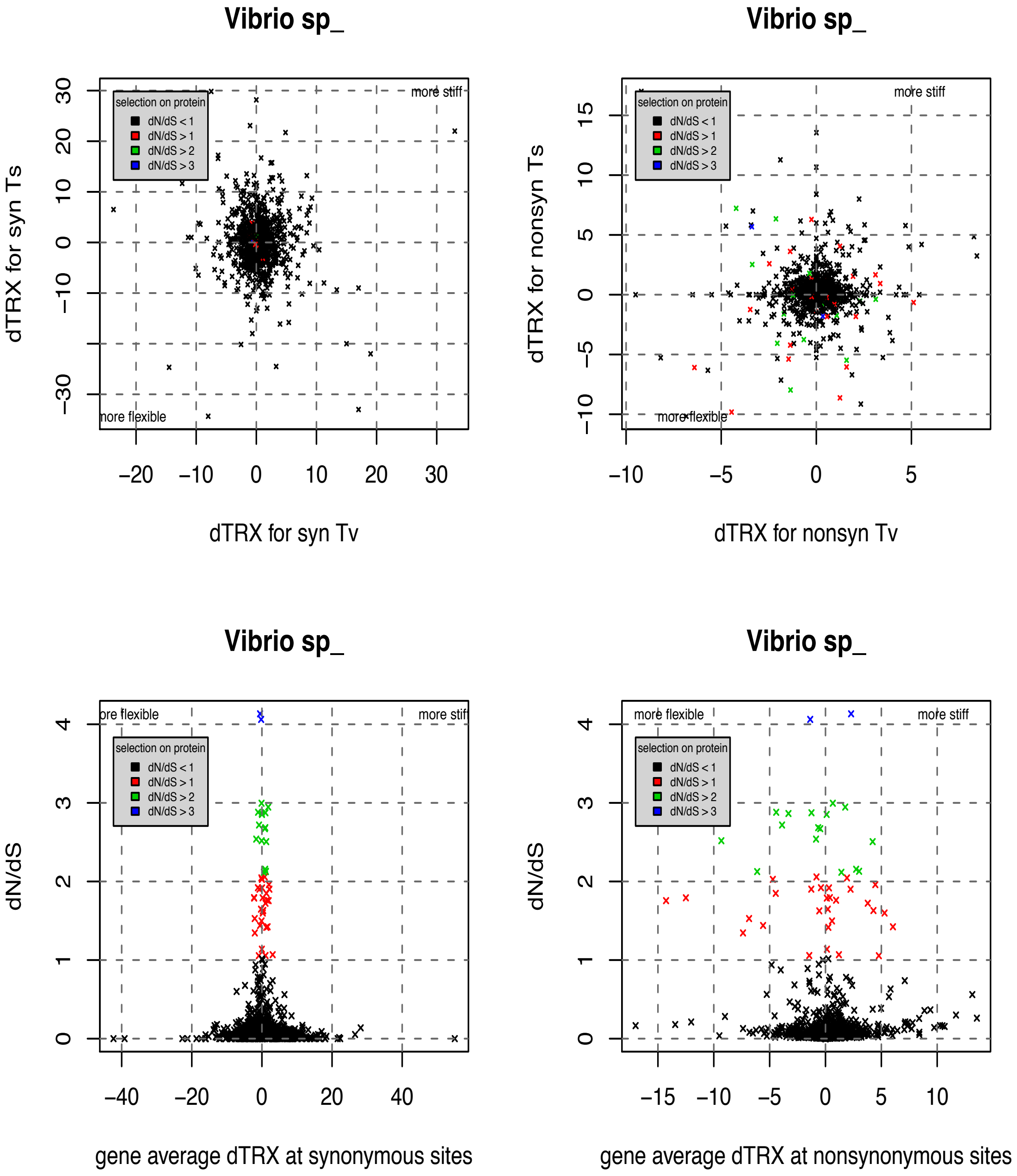
20. Streptococcus pyogenes



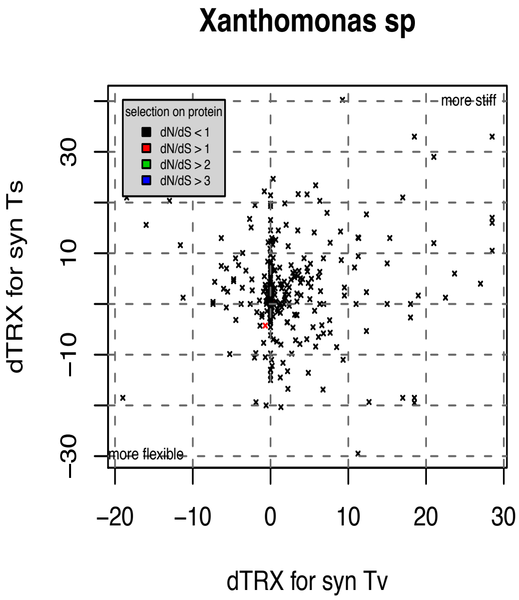
21. Vibrio cholera

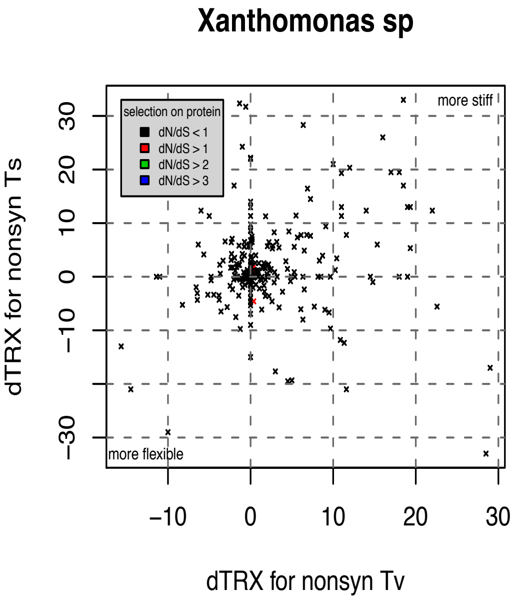


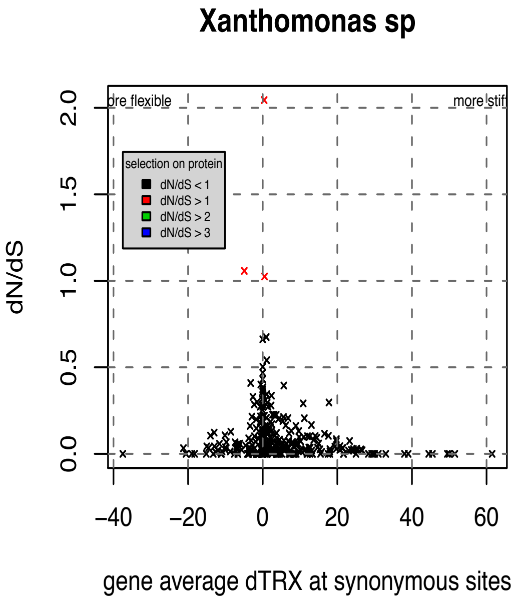
22. Vibrio sp

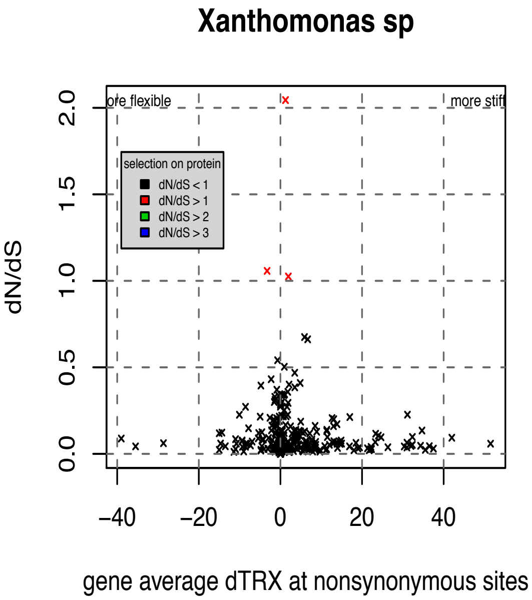

Vibrio sp_
dTRX for syn Ts
dTRX for syn Tv
−30
−10
0
10
20
30
−20
−10
0
10
20
30
more stiff
nore flexible
selection on protein
dN/dS < 1
dN/dS > 1
dN/dS > 2
dN/dS > 3
Vibrio sp_
dTRX for nonsyn Ts
dTRX for nonsyn Tv
−10
−5
0
5
10
15
−10
−5
0
5
more stiff
more flexible
selection on protein
dN/dS < 1
dN/dS > 1
dN/dS > 2
dN/dS > 3
Vibrio sp_
dN/dS
gene average dTRX at synonymous sites
0
1
2
3
4
−40
−20
0
20
40
ore flexible
more stif
selection on protein
dN/dS < 1
dN/dS > 1
dN/dS > 2
dN/dS > 3
Vibrio sp_
dN/dS
gene average dTRX at nonsynonymous sites
0
1
2
3
4
−15
−10
−5
0
5
10
more flexible
more stiff
selection on protein
dN/dS < 1
dN/dS > 1
dN/dS > 2
dN/dS > 3

23.Xanthomonas sp

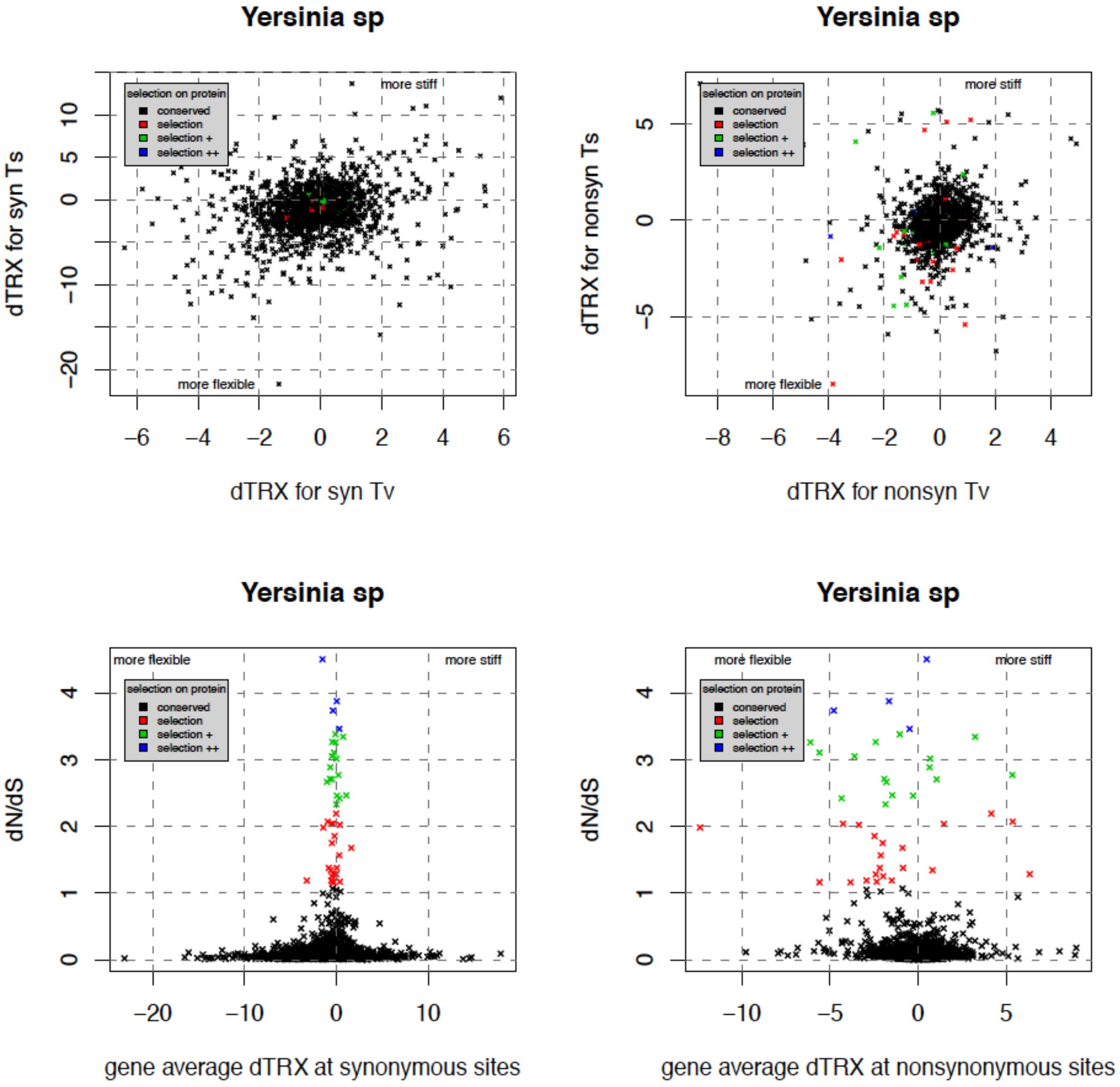








24. Yersinia


Yersinia sp
selection on protein
■ conserved
■ selection
■ selection +
■ selection ++
more stiff
more flexible
dTRX for syn Ts
dTRX for syn Tv
−20
−10
0
5
10
−6
−4
−2
0
2
4
6
Yersinia sp
selection on protein
■ conserved
■ selection
■ selection +
■ selection ++
more stiff
more flexible
dTRX for nonsyn Ts
dTRX for nonsyn Tv
−5
0
5
−8
−6
−4
−2
0
2
4
Yersinia sp
more flexible
more stiff
selection on protein
■ conserved
■ selection
■ selection +
■ selection ++
dN/dS
gene average dTRX at synonymous sites
0
1
2
3
4
−20
−10
0
10
Yersinia sp
more flexible
more stiff
selection on protein
■ conserved
■ selection
■ selection +
■ selection ++
dN/dS
gene average dTRX at nonsynonymous sites
0
1
2
3
4
−10
−5
0
5