Université de Montréal

# Novel Bioinformatics Programs for Taxonomical Classification and Functional Analysis of the Whole Genome Sequencing Data of Arbuscular Mycorrhizal Fungi

par Jee Eun Kang

Département de Biochimie
Faculté de Médecine

Thèse présentée
en vue de l'obtention du grade de Doctorat
en Bio-informatique

Octobre, 2018

# Résumé

[TITRE] Classification taxonomique et analyse fonctionnelle spécifique àla position des séquences génomique des champignons mycorhiziens arbusculaires et les microorganismes qui leurs sont associés [PROBLÉMATIQUE ET CADRE CONCEPTUEL] Les champignons mycorhiziens arbusculaires (CMA) sont des symbiotes obligatoires des racines de la majoritédes plantes vasculaires. Les CMA appartiennent au phylum Glomeromycota et ils sont considérés comme une lignée fongique primitive qui a conservé la structure coenocytique des hyphes et la production des spores asexuées multinucléées. De nombeuses études ont démontréque plusieurs microorganismes sont associés avec les mycélia des CMA soit àla surface des hyphes et des spores mais aussi àl'intérieurs de celles-ci. Le séquençage des génomes des CMA cultivés *in-vivo* représente un défi considérable car il s'agit d'un métagénome constituédu génome du CMA lui-même et les génomes des microbes qui lui sont associés. Par conséquence, l'identification de l'origine taxonomique de chaque séquence représente une tâche extrêmement ardue. Dans mon projet, j'ai développédeux nouveaux programmes bioinformatiques qui permettent de classer les séquences selon groupe taxonomique et d'identifier les fonctions de celles-ci. J'ai crééune base de données avec 444 génomes d'espèces appartenant à54 genres. Le choix de ces espèces des bactéries et des champignons a étébasésur leur abondance dans les sols). [MÉTHODOLOGIE] Le programme bioinformatique utilise le tableau des références des microorganismes et des méthodes statistiques pour la classification taxonomique des séquences. Par la suite, des tableaux des codons synonymes étaient créés àpartir des structures secondaires (SS) des bases de données de protéines (PDB) pour les séquences codantes (SC) et des motifs de composition pour les séquences non-codantes (SNC). Chaque tableau est composéde 3 niveaux - les caractéristiques d'acides aminés; l'utilisation des acides aminés synonymes correspondants, et l'utilisation des codons synonymes correspondants. En comparant les méthodes existantes qui utilisent les taux de substitution moyenne globale quelle que soit les spécificités des acides aminés dans diverses structures, mon programme fournit une classification àhaute résolution pour des séquences courtes (150-300 pb) parce que les biais dans l'utilisation des codons synonymes àpartir d'environ 8000 trimères d'acides aminés spécifiques des sous-unités de structure secondaire, ont étéextraits avec des substitutions d'acides aminés pris en considération dans chaque trimère spécifique. Pour l'analyse fonctionnelle, le programme crée dynamiquement des données comparatives de 54 genres microbiens basés sur leurs biais dans l'utilisation des codons synonymes d'appariement de trois codons d'ADN (9-mères) identifiés dans une séquence de requête. Le programme applique une analyse en composantes principales basée sur la matrice de corrélation en association avec le partitionnement en *k*-moyennes aux données comparatives. [RETOMBÉES] Les taux de prédiction correcte de la CDS et les non-CDS étaient de 50 à71% pour les bactéries, et 65 à73% pour les champignons, respectivement. Pour les CMA, 49% des CDS et 72% des non-CDS ont étécorrectement classés. Ce programme nous permet d'estimer les abondances approximatives des communautés microbiennes associées au CMA. Les résultats de l'analyse fonctionnelle peuvent fournir des informations sur des sites d'interaction moléculaire importants impliqués dans la diversification des séquences et l'évolution des gènes. Les programmes sont disponibles gratuitement sur **www.fungalsesame.org**.

**Mots-clés**: sesame, sesame PS function, les caractéristiques d'acides aminés, trois codons ADN 9-mères, structure secondaire, classification taxonomique, analyse fonctionnelle spécifique àla position; Code génétique; Post; Étude Comparative; Génome Mitochondrial

# Abstract

Arbuscular Mycorrhizal Fungi (AMF) are obligate plant-root symbionts belonging to the phylum Glomeromycota. They form coenocytic hyphae and reproduce through large multinucleated asexual spores. Numerous studies have shown that AMF interact closely or loosely with a myriad of microorganisms, particularly bacteria and fungi that live on the surface of or inside of their mycelia and spores. Whole genome sequencing (WGS) data of the AMF grown *in-vivo* (typically grown in root of a host plant in pot filled with soil) contain a large amount of sequences from microorganisms inhabiting in their spore along with their own genome sequences, resulting in a metagenome.

The goal of my study was to develop bioinformatics programs for taxonomical classification and for functional analysis of the WGS data of the AMF. In the area of metagenomics, there are mainly two approaches for taxonomical classification: similarity-based (i.e., homology search) and composition-based (i.e., k-mers) methods. Similarity-based method solely depends on bioinformatics sequence databases and homology search programs such as BLAST program. The similarity-based method may not be suitable for ancient fungi AMF, because bioinformatics databases represent only a small fraction of the diversity of existing microorganisms, and gene prediction programs are highly biased towards intensively studied microorganisms. Considering that AMF have high inter/ intra genome variations, in addition to coenocytic and multi-genomic characteristics, probably due to their adaptation via various kinds of symbioses, composition-based method alone is not an effective solution for AMF, because it relies on base composition biases and focuses on taxonomical classification for prokaryotic organisms.

In the first project, I a developed novel bioinformatics program, called SeSaMe (Spore associated Symbiotic Microbes), for taxonomical classification of the WGS data of the AMF. I selected microorganisms that were dominant in soil environment and grouped them into 54 genera which were used as references. I created a reference sequence database with a variable called *Three codon DNA 9-mer.* They were created based on a large number of structure files from Protein Data Bank (PDB): approx. 224,000 *Three codon DNA 9-mers* encoding for subunits of protein secondary structures. Based on the reference sequence database, I created genus specific usage databases containing codon usage and amino acid usage per taxonomic rank- genus. The program distinguishes between coding sequence (CDS) and non-CDS, detects an open reading frame, and classifies a query sequence into a genus group out of 54 genera used as reference. The developed program enables us to estimate relative abundances of taxonomic groups and to assess symbiotic roles of taxonomic groups associated with AMF. The program can be applied to other microorganisms as well as soil metagenome data. The program has applications in applied environmental microbiology. The developed program is available for free of charge at www.fungalsesame.org.

In the second project, I developed another bioinformatics program, called SeSaMe PS Function, for position specific functional analysis of the WGS data of the AMF. AMF may contain a large portion of genes with unknown functions for which we may not be able to find homologues in existing sequence databases. While existing motif annotation programs rely on sequence alignment and have limitations for inferring functionality of novel genes, the developed program identifies potentially important interaction sites that are structurally and functionally distinctive from other subsequences, within a query sequence with exploratory data analysis. The program identifies matching *Three codon DNA 9-mers* in a query sequence, and dynamically creates comparative dataset of 54 genera, based on codon usage bias information retrieved from the genus specific usage databases. The program applies correlation Principal Component Analysis in conjunction with K-means clustering method to the comparative dataset. The program identifies outliers; *Three codon DNA 9-mers,* assigned into a cluster with a single member or with only a few members, are often outliers with important structures that may play roles in molecular interaction.

In the third project, I developed a novel bioinformatics program called Posts (POsition Specific genetic code Tables) that assigns a codon into an amino acid group according to the codon position. The standard genetic code table may be more readily applicable to the genes whose genetic codes comply with the standard biological coding rules obtained from model organisms grown under laboratory condition. However, it may be insufficient for studying evolutions of genetic codes that may provide important information about codon properties. The mainstream hypotheses of genetic code origin suggested that codon position played important roles in the evolution of genetic codes. As a case study, we investigated irregular codons in 187 mitochondrial genomes of plants, lichen-forming fungi, endophytic fungi, and AMF. Each column of the Post contains 16 codons and the amino acids encoded by these are called an amino acid characteristics group (A.A. Char Group). Based on A.A. Char Group, an irregular codon can be classified into within-column type or trans-column type. The majority of the identified irregular codons belonged to the within-column type. The Post may offer new perspectives on codon property and codon assignment. The developed program is freely available at www.codon.kr. Taken together, the developed programs, the SeSaMe, the SeSaMe PS Function, and the Post, provide important research tools for advancing our knowledge of AMF genomics and for studying their symbiotic relations with associated microorganisms.

**Keywords**: Sesame; Spore associated Symbiotic Microbes; Symbiosis; Sesame PS function; Arbuscular mycorrhizal fungi; Three codon DNA 9-mer; Amino acid characteristics; Secondary structure; Taxonomical classification; Position specific functional analysis; Position specific genetic code tables; Post; Comparative study; Mitochondrial genome

# Table of Contents

# List of Tables

**Chapter 1**

**Chapter 2**

**Chapter 3**

**Chapter 4**

# List of Figures

**Chapter 1**

**Chapter 2**

Chapter 3

Chapter 4

# Definitions

In the introduction section, a word whose definition is given has a superscript indicating a number in this definition section. All definitions in this section are cited as direct quotation according to online wikipedia- https://www.wikipedia.org or https://www.wiktionary.org unless another reference is indicated inside parentheses. These references, other than online wikipedia or wiktionary, are provided in the reference section. Use of quotation marks “” in the beginning and in the end of each definition has been omitted.

1) A mycorrhiza (pl. mycorrhizae or mycorrhizas) is a symbiotic association between a fungus and the roots of a vascular host plant.

2) Phytoremediation refers to the technologies that use living plants to clean up soil, air, and water contaminated with hazardous chemicals.

3) Biodegradation is the disintegration of materials by bacteria, fungi, or other biological means.

4) Biotrophic describes a parasite or symbiont that needs its host to stay alive.

5) Mycelium is the vegetative part of a fungus or fungus-like bacterial colony, consisting of a mass of branching, thread-like hyphae.

6) Fungal mycelia in which hyphae lack septa are known as "aseptate" or "coenocytic".

7) Multinucleate cells (also called multinucleated or polynuclear cells) are eukaryotic cells that have more than one nucleus per cell, i.e., multiple nuclei share one common cytoplasm.

8) A hypha (plural hyphae) is a long, branching filamentous structure of a fungus.

9) A heterokaryon is a multinucleate cell that contains genetically different nuclei.

10) A telomere is a region of repetitive nucleotide sequences at each end of a chromosome, which protects the end of the chromosome from deterioration or from fusion with neighboring chromosomes.

11) Protoplast, in biology, it refers to the entire cell, excluding the cell wall, but currently has several definitions: a plant, bacterial or fungal cell that had its cell wall completely or partially removed using either mechanical or enzymatic means.

12) An endophyte is an endosymbiont, often a bacterium or fungus that lives within a plant for at least part of its life cycle without causing apparent disease.

13) QR code (abbreviated from Quick Response Code) is the trademark for a type of matrix barcode (or two-dimensional barcode) first designed for the automotive industry in Japan.

14) Homoplasmy is a term used in genetics to describe a eukaryotic cell whose copies of mitochondrial DNA are all identical.

15) Compositional biases are local shifts in amino acid or nucleotide frequencies that can occur as an adaptation of an organism to an extreme ecological niche, or as the signature of a specific function or localization of the corresponding protein (Antonets KS et al 2013)

16) The term k-mer typically refers to all the possible substrings of length k that are contained in a

string. In computational genomics, k-mers refer to all the possible subsequences (of length k) from a read obtained through DNA Sequencing.

17) The mycorrhizosphere is the region around a mycorrhizal fungus in which nutrients released from the fungus increase the microbial population and its activities.

18) The English-language neologism omics informally refers to a field of study in biology ending in -omics, such as genomics, proteomics or metabolomics.

19) Codon usage is a phenomenon of non-uniform usage of codons (Behura SK et al 2012)

Codon usage bias refers to differences in the frequency of occurrence of synonymous codons in coding DNA. A codon is a series of three nucleotides (a triplet) that encodes a specific amino acid residue in a polypeptide chain or for the termination of translation (stop codons).

There are 64 different codons (61 codons encoding for amino acids plus 3 stop codons) but only 20 different translated amino acids. The overabundance in the number of codons allows many amino acids to be encoded by more than one codon. Because of such redundancy it is said that the genetic code is degenerate.

20) Transcription

Transcription proceeds in the following general steps:

- RNA polymerase, together with one or more general transcription factors, binds to promoter DNA.
- RNA polymerase creates a transcription bubble, which separates the two strands of the DNA helix. This is done by breaking the hydrogen bonds between complementary DNA nucleotides.
- RNA polymerase adds RNA nucleotides (which are complementary to the nucleotides of one DNA strand).
- RNA sugar-phosphate backbone forms with assistance from RNA polymerase to form an RNA strand.
- Hydrogen bonds of the RNA–DNA helix break, freeing the newly synthesized RNA strand.
- If the cell has a nucleus, the RNA may be further processed. This may include polyadenylation, capping, and splicing.
- The RNA may remain in the nucleus or exit to the cytoplasm through the nuclear pore complex.

21) Translation

Messenger RNA (mRNA) carries information about a protein sequence to the ribosomes, the protein synthesis factories in the cell. It is coded so that every three nucleotides (a codon) corresponds to one amino acid. In eukaryotic cells, once precursor mRNA (pre-mRNA) has been transcribed from DNA, it is processed to mature mRNA. This removes its introns—non-coding sections of the pre-mRNA. Ribosomes link amino acids together in the order specified by messenger RNA(mRNA) molecules.

22) The central dogma of molecular biology is an explanation of the flow of genetic information within a biological system. It was first stated by Francis Crick in 1958

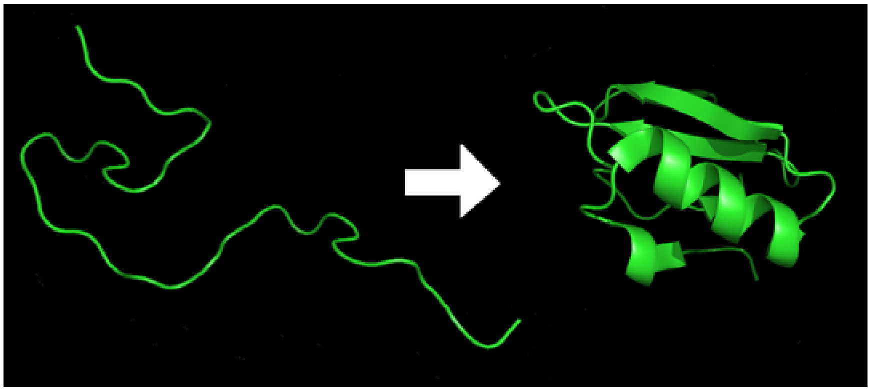
“The Central Dogma. This states that once ‘information’ has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein."

Definition Table 1 Three classes of information transfer suggested by the dogma

| General | Special | Unknown |
|---|---|---|
| DNA → DNA | RNA → DNA | protein → DNA |
| DNA → RNA | RNA → RNA | protein → RNA |
| RNA → protein | DNA → protein | protein → protein |

(source: https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology)

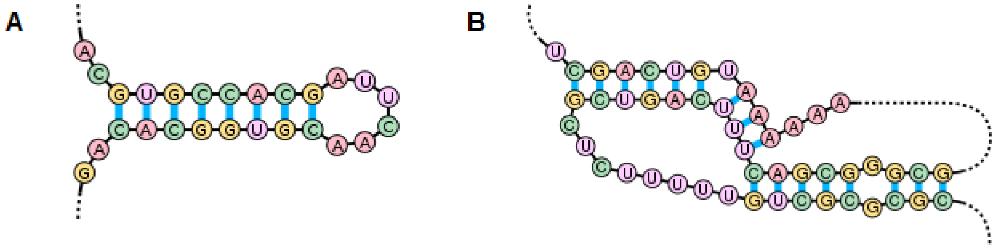
23) Protein folding is the physical process by which a protein chain acquires its native 3-dimensional structure, a conformation that is usually biologically functional, in an expeditious and reproducible manner. It is the physical process by which a polypeptide folds into its characteristic and functional three-dimensional structure from random coil.



**Definition Figure 1 Protein before and after folding.**
**(source: https://en.wikipedia.org/wiki/Protein_folding)**

24) Codon context generally refers to sequential pair of codons in a gene (Behura SK et al 2012)

25) Nucleic acid secondary structure is the base pairing interactions within a single nucleic acid polymer or between two polymers. It can be represented as a list of bases which are paired in a nucleic acid molecule. The secondary structures of biological DNA's and RNA's tend to be different: biological DNA mostly exists as fully base paired double helices, while biological RNA is single stranded and often forms complex and intricate base-pairing interactions due to its increased ability to form hydrogen bonds stemming from the extra hydroxyl group in the ribose sugar.



**Definition Figure 2 Nucleic acid secondary structure**
**A. An RNA stem-loop secondary structure. B. An RNA pseudoknot structure. For example, the RNA component of human telomerase.**

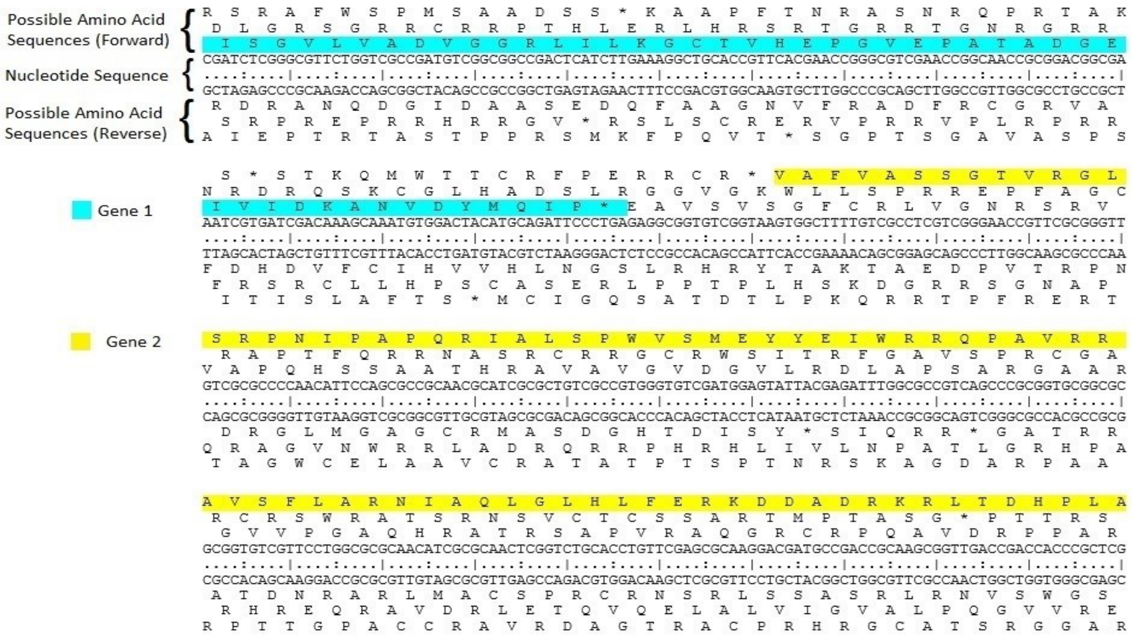
**(source: https://en.wikipedia.org/wiki/Nucleic_acid_secondary_structure)**

26) A thermophile is an organism—a type of extremophile—that thrives at relatively high temperatures, between 41 and 122 °C (106 and 252 °F).

27) A mesophile is an organism that grows best in moderate temperature, neither too hot nor too cold, typically between 20 and 45 °C (68 and 113 °F).

28) Psychrophiles or cryophiles (adj. psychrophilic or cryophilic) are extremophilic organisms that are capable of growth and reproduction in low temperatures, ranging from −20 °C to +10 °C. They are found in places that are permanently cold, such as the polar regions and the deep sea.

29) In molecular biology, a reading frame is a way of dividing the sequence of nucleotides in a nucleic acid (DNA or RNA) molecule into a set of consecutive, non-overlapping triplets. Where these triplets equate to amino acids or stop signals during translation, they are called codons.


Possible Amino Acid Sequences (Forward)
Nucleotide Sequence
Possible Amino Acid Sequences (Reverse)
Gene 1
Gene 2
R S R A F W S P M S A A D S S * K A A P F T N R A S N R Q P R T A K
D L G R S G R R C R R P T H L E R L H R S R T G R R T G N R G R R
I S G V L V A D V G G R L I L K G C T V H E P G V E P A T A D G E
CGATCTCGGGCGTTCTGGTCGCCGATGTCGGCGGCCGACTCATCTTGAAAGGCTGCACCGTTCACGAACCGGGCGTCGAACCGGCAACCGCGGACGGCGA
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
GCTAGAGCCCGCAAGACCAGCGGCTACAGCCGCCGGCTGAGTAGAACTTTCCGACGTGGCAAGTGCTTGGCCCGCAGCTTGGCCGTTGGCGCCTGCCGCT
R D R A N Q D G I D A A S E D Q F A A G N V F R A D F R C G R V A
S R P R E P R R H R R G V * R S L S C R E R V P R R V P L R P R R
A I E P T R T A S T P P R S M K F P Q V T * S G P T S G A V A S P S
S * S T K Q M W T T C R F P E R R C R * V A F V A S S G T V R G L
N R D R Q S K C G L H A D S L R G G V G K W L L S P R R E P F A G C
I V I D K A N V D Y M Q I P * E A V S V S G F C R L V G N R S R V
AATCGTGATCGACAAAGCAAATGTGGACTACATGCAGATTCCCTGAGAGGCGGTGTCGGTAAGTGGCTTTTGTCGCCTCGTCGGGAACCGTTCGCGGGTT
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
TTAGCACTAGCTGTTTCGTTTACACCTGATGTACGTCTAAGGGACTCTCCGCCACAGCCATTCACCGAAAACAGCGGAGCAGCCCTTGGCAAGCGCCCAA
F D H D V F C I H V V H L N G S L R H R Y T A K T A E D P V T R P N
F R S R C L L H P S C A S E R L P P T P L H S K D G R R S G N A P
I T I S L A F T S * M C I G Q S A T D T L P K Q R R T P F R E R T
S R P N I P A P Q R I A L S P W V S M E Y Y E I W R R Q P A V R R
R A P T F Q R R N A S R C R R G C R W S I T R F G A V S P R C G A
V A P Q H S S A A T H R A V A V G V D G V L R D L A P S A R G A A R
GTCGCGCCCCAACATTCCAGCGCCGCAACGCATCGCGCTGTCGCCGTGGGTGTCGATGGAGTATTACGAGATTTGGCGCCGTCAGCCCGCGGTGCGGCGC
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
CAGCGCGGGGTTGTAAGGTCGCGGCGTTGCGTAGCGCGACAGCGGCACCCACAGCTACCTCATAATGCTCTAAACCGCGGCAGTCGGGCGCCACGCCGCG
D R G L M G A G C R M A S D G H T D I S Y * S I Q R R * G A T R R
Q R A G V N W R R L A D R Q R R P H R H L I V L N P A T L G R H P A
T A G W C E L A A V C R A T A T P T S P T N R S K A G D A R P A A
A V S F L A R N I A Q L G L H L F E R K D D A D R K R L T D H P L A
R C R S W R A T S R N S V C T C S S A R T M P T A S G * P T T R S
G V V P G A Q H R A T R S A P V R A Q G R C R P Q A V D R P P A R
GCGGTGTCGTTCCTGGCGCGCAACATCGCGCAACTCGGTCTGCACCTGTTCGAGCGCAAGGACGATGCCGACCGCAAGCGGTTGACCGACCACCCGCTCG
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
CGCCACAGCAAGGACCGCGCGTTGTAGCGCGTTGAGCCAGACGTGGACAAGCTCGCGTTCCTGCTACGGCTGGCGTTCGCCAACTGGCTGGTGGGCGAGC
A T D N R A R L M A C S P R C R N S R L S S A S R L R N V S W G S
R H R E Q R A V D R L E T Q V Q E L A L V I G V A L P Q G V V R E
R P T T G P A C C R A V R D A G T R A C P R H R G C A T S R G G A R

**Definition Figure 3 An example of a six-frame translation**

**(source: https://en.wikipedia.org/wiki/Reading_frame)**

30) In molecular genetics, an open reading frame (ORF) is the part of a reading frame that has the potential to be translated. An ORF is a continuous stretch of codons that contain a start codon (usually AUG) and a stop codon (usually UAA, UAG or UGA).

31) In molecular biology, DNA polymerases are enzymes that synthesize DNA molecules from deoxyribonucleotides, the building blocks of DNA. These enzymes are essential for DNA replication and usually work in pairs to create two identical DNA strands from a single original DNA molecule. During this process, DNA polymerase "reads" the existing DNA strands to create two new strands that match the existing ones.

32) RNA polymerase (ribonucleic acid polymerase), both abbreviated RNAP or RNApol, official name DNA-directed RNA polymerase, is a member of a family of enzymes that are essential to life: they are found in all organisms (species) and many viruses. RNAP locally opens the double-stranded DNA (usually about four turns of the double helix) so that one strand of the exposed nucleotides can be used as a template for the synthesis of RNA, a process called transcription. A transcription factor and its associated transcription mediator complex must be attached to a DNA binding site called a promoter region before RNAP can initiate the DNA unwinding at that position.

33) In bacteria, all transcription is performed by a single type of RNA polymerase. This polymerase contains four catalytic subunits and a single regulatory subunit known as sigma (s). Interestingly, several distinct sigma factors have been identified, and each of these oversees transcription of a unique set of genes. Sigma factors are thus discriminatory, as each binds a distinct set of promoter sequences.

Sigma factors are subunits of all bacterial RNA polymerases. They are responsible for determining the specificity of promoter DNA binding and control how efficiently RNA synthesis (transcription) is initiated (R.R. Burgess Encyclopedia of Genetics).

34) In genetics, an enhancer is a short (50-1500 bp) region of DNA that can be bound by proteins (activators) to increase the likelihood that transcription of a particular gene will occur. These proteins are usually referred to as transcription factors.

35) In molecular biology, a transcription factor (TF) (or sequence-specific DNA-binding factor) is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence. TFs work alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase (the enzyme that performs the transcription of genetic information from DNA to RNA) to specific genes.

36) An operon is a functioning unit of genomic DNA containing a cluster of genes under the control of a single promoter. A promoter is a region of DNA that initiates transcription of a particular gene. Promoters are located near the transcription start sites of genes, on the same strand and upstream on the DNA (towards the 5' region of the sense strand).

# List of Abbreviations and Acronyms

80% components: Components whose sum accounts for 80% of inertia

*A*

A.A. Char: Amino Acid Characteristic

A.A. Char Group: Amino Acid Characteristics Group

A.A. Char Trimer: Amino Acid Characteristic Trimer

A.A. Group: Amino Acid Group

aaRs: aminoacyl-tRNA synthetases

AFE: Éducation et de l'Enseignement supérieur Quebec

AM: Arbuscular Mycorrhizal

AMF: Arbuscular mycorrhizal fungi

ASD: Allosteric Site

*B*

BLAST: Basic Local Alignment Search Tool

BLOSUM: BLOcks SUbstitution Matrix

BothCA: Both CSA and ASD

*C*

CDS: Coding sequence

CSA: Catalytic Site

*D*

DNA: Deoxyribonucleic acid

*F*

FESP: Faculté des études supérieures et postdoctorales de l'UdeM

FS: Functional Segment

*G*

Genus Specific DB: Genus-Specific usage bias Database

GPB: Genomics Proteomics Bioinformatics

*H*

HMG: High Mobility Group

*I*

IRBV: Institut de Recherche en Biologie Végétale de l'Université de Montréal

ITS: Internal transcribed spacer

*K*

KW: Kruskal Wallis

*L*

lncRNAs: Long non-coding RNAs

$\log_{10}$ (inverse of *P* value score): Base 10 logarithm of an approximated inverse of a rank sum based *P* value score.

*M*

MDR1: Multidrug Resistance 1

MDS*:* multidimensional scaling

MRE: Mycoplasma-Related Endobacteria

*N*

NADH: Reduced form of Nicotinamide adenine dinucleotide

None: None of CSA nor ASD

*O*

ORF: Open reading frame

*P*

PAM: Point Accepted Mutation

PCA: Principal Component Analysis

PCA-Kmeans: PCA in conjunction with K-means clustering method

PDB: Protein Data Bank

Post: Position specific genetic code table

*R*

*R. irregularis: Rhizophagus irregularis*

RNA: Ribonucleic acid

rRNA: ribosomal RNA

*S*

SEM: Scanning electron microscopy

SeSaMe: Spore associated Symbiotic Microbes

SeSaMe PS Function: SeSaMe Position Specific Function

SSU: Small subunit

*T*

The First/Second components: The first principal component and the second component

Trimer Ref. DB: Trimer Reference sequence Database

*W*

WGS: Whole genome sequencing

WMS: whole metagenome sequencing

Abbreviations of DNA, RNA

C: Cytosine

T: Thymine

G: Guanine

A: Adenine

U: Uracil

R: Purine

Y: Pyrimidine

Abbreviations of 20 amino acids and stop codon

| | |
|---|---|
| A: | Alanine/ Ala |
| R: | Arginine/ Arg |
| N: | Asparagine/ Asn |
| D: | Aspartic acid/ Asp |
| C: | Cysteine/ Cys |
| E: | Glutamic acid/ Glu |
| Q: | Glutamine/ Gln |
| G: | Glycine/ Gly |
| H: | Histidine/ His |
| I: | Isoleucine/ Ile |
| L: | Leucine/ Leu |
| K: | Lysine/ Lys |
| M: | Methionine/ Met |
| F: | Phenylalanine/ Phe |
| P: | Proline/ Pro |
| S: | Serine/ Ser |
| T: | Threonine/ Thr |

| | |
|---|---|
| W: | Tryptophan/ Trp |
| Y: | Tyrosine/ Tyr |
| V: | Valine/ Val |
| * : | Stop codons |

Mitochondrial genes

cox1: mitochondrial cytochrome c oxidase subunit I gene

rnl: mitochondrial large ribosomal subunit RNA gene

nad2: mitochondrial NADH dehydrogenase subunit 2 gene

atp6: mitochondrial gene encoding the ATP synthase Fo subunit 6

atp9: mitochondrial gene encoding the ATP synthase subunit 9

cox2: mitochondrial gene encoding the Cytochrome c oxidase subunit 2

cox3: mitochondrial gene encoding the Cytochrome c oxidase subunit 3

cob: mitochondrial gene encoding cytochrome b- the component of the ubiquinol-cytochrome c reductase complex

cytb: mitochondrial gene encoding subunit from the bc1 complex

nad1: mitochondrial gene encoding NADH-ubiquinone oxidoreductase chain 1

nad3: mitochondrial gene encoding NADH-ubiquinone oxidoreductase chain 3

nad4: mitochondrial gene encoding NADH-ubiquinone oxidoreductase chain 4

nad4L: mitochondrial gene encoding NADH-ubiquinone oxidoreductase chain 4L

nad5: mitochondrial gene encoding NADH-ubiquinone oxidoreductase chain 5

nad6: mitochondrial gene encoding NADH-ubiquinone oxidoreductase chain 6

We cannot see microorganisms nor air with naked eyes. However they have existed long before us.
Just because other beings cannot be seen, it does not mean that they do not exist.
Although I cannot recognize their presence, I hope to become a better being, forming symbiosis with other beings.

To other beings

within me, among us, somewhere universe
no worship but friendship
no rejection but co-existence
no greedy but fair
seeking approval not of others but of true self

# Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Prof. Mohamed Hijri at the IRBV at the Universitéde Montréal for his continuous support of my Ph.D. study. He helped me to understand AMF research with his immense knowledge. His guidance helped me in all the time of research and of writing scientific articles. He has supported me throughout my research and helped me focus on my projects. He is patient, prudent, and considerate: the virtues required for marathoner. Above all, I have found the thesis topic most interesting throughout my study. The subject was my first choice when I was looking for research advisor, and I am so grateful to him that he provided me an opportunity to pursue my Ph.D. degree under his supervision. I am thankful to Prof. Nicholas Lartillot and Ms. Elaine Meunier for helping me completing the process of searching for research advisor and to the Dept. of Biochemistry for allowing me to study under his direction.

I would like to show my appreciation to my co-advisor, Prof. Antonio Ciampi in the EBOH at McGill University. I am thankful for his aspiring guidance, invaluable criticism, and helpful advice during the project and for writing scientific articles. He provided me with practical guidance about statistical methods and writing scientific articles about them.

Besides my advisors, I would like to thank the rest of my thesis committee members: Prof. Jesse Shapiro and Prof. David Walsh for their insightful comments, encouragement, and invaluably constructive criticisms which inspired me throughout my study. Especially, they have encouraged me to add a new function- p-value scoring method. I am thankful to Prof. Sylvie Hamel and Dr. Etienne Yergeau for sharing their expertise during my pre-doc exam. I would also like to take this opportunity to express my gratitude to Prof. Pascal Chartrand, Prof. Sylvie Hamel, Prof. Gertraud Burger, Prof. Marlene Oeffinger, Prof. Sylvie Mader, and other instructors for being a great teacher. I would like to thank the former and current lab members, especially Dr. Bachir Iffis, Dr. Maryam Nadimi, Dr. Ivan de la Providencia, Dr. Franck Stefani, Dr. Yves Terrat, and Soon-Jae Lee. I am very grateful to the rest of the jury members of my thesis- Prof. Adrian Serohijos, Prof. Anne-Lise Routier, and Prof. Vladimir Makarenkov for taking time out of their busy schedules to read and evaluate my work. Their expertise and insights are truly valuable.

I would like to express my special thanks to Ms. Elaine Meunier for her continuous support in every important step throughout my doctoral study. Especially, I would not have been able to resolve time critical issues without her help- admission application, pre-doc exam, registration, and thesis preparation. I wish to express my sincere thanks to AFE (Éducation et de l'Enseignement supérieur Quebec), FESP (Facultédes études supérieures et postdoctorales de l'UdeM), and IRBV (Institut de Recherche en Biologie Végétale de l'Universitéde Montréal) for awarding me scholarships.

I would like to thank my family: my parents In Wee Kang and Moo Suk Lim, my brother Sung Joon Kang, and my sister Jung Eun Kang for supporting me throughout my life. Especially, I am very thankful to Jung Eun Kang for her speedy and professional assistance in the preparation of the figures and the tables in my articles. I would like to thank my friend Eun Joo Ha and her family for the good times filled with funny jokes, laughs, and her children's rejoicing.

# Thesis outline

During my doctoral study, I developed three different types of bioinformatics programs for AMF research. I introduce them in this thesis. The thesis consists of five chapters. In the first chapter, I address the importance of AMF- effectiveness of AMF inoculants for sustainable agriculture and phytoremediation. I review literature to explain why the WGS data of the AMF result in metagenome: the nature of AMF harboring a large number of symbiotic microorganisms inside of their spores and mycelia. I summarize research articles about the complex genomic properties of AMF and explain the limitation of the available genome sequencing data for serving as a reference genome due to high inter/intra genome variations of AMF isolates. I survey the existing approaches for taxonomical classification of metagenome sequencing data and discuss why they are ineffective for analyzing the WGS data of the AMF.

I summarize recent studies that have documented the important roles of codon usage and codon context in co-translational process and regulation of gene expression and gene products and their contribution to microbial adaptation in order to show the biological importance of the main variable of the developed programs. In the last section of this chapter, I state research problem, research objectives, and the contribution of the developed programs to AMF research.

In the second chapter, I introduce the developed programs, the SeSaMe, for taxonomical classification of the WGS data of the AMF. One program classifies a query sequence into one of 54 genera used as references and the other into one of 13 taxonomic groups with the higher taxonomic rank. In this chapter, I provide users with the details of the components of the programs, guidelines of how to interpret program results, and P value score tables for assessing the statistical significance of predicted outcome. In the third chapter, I introduce another developed program, the SeSaMe PS Function, for position specific functional analysis of the WGS data of the AMF. I provide users with the details of the methods and a case study for demonstrating how to detect outliers in a query sequence which is just one of many applications of the program. I used existing bioinformatics programs for inferring functions of the outliers that may play roles in undiscovered mechanisms. In the fourth chapter, I introduce the developed method, the Post, that assigns a codon with respect to codon position. It may provide systematic tools for studying novel properties of genetic codes and codon assignment. I discuss the method in detail, and provide a case study- identification of irregular codons in 187 mitochondrial genomes of various plants and fungi.

The fifth chapter consists of conclusions, discussions, and future work.

The articles in the second and the third chapters have been accepted by the journal, Genomics Proteomics Bioinformatics (GPB). The article in the fourth chapter is in preparation for submission.

# Introduction

## 1.1 Importance of AMF

Arbuscular mycorrhizal fungi (AMF) are root colonizing symbiotic microorganisms that stimulate plant growth and improve soil structure (Hijri 2016, Roy-Bolduc and Hijri, 2011, Zarik et al 2016). They supply plants with essential mineral nutrients, protect them against soil-borne pathogens, and reduce their environmental stresses (Bunn et al 2009).

Symbiosis between plants and AMF is widespread and it is well accepted that AMF form symbiosis with more than 80% of vascular plants worldwide (Smith and Read 2008). Arbuscular Mycorrhizal (AM) symbiosis has been studied in numerous disciplines including plant sciences, microbiology, mycology, ecology, environmental science, and agriculture. Plant scientists employ genomics and transcriptomics tools to study taxonomical classification and functional analysis of mycorrhizae[1] metagenome data and are particularly interested in symbiosis-related genes and their regulatory mechanisms (Vangelisti et al 2018, Handa et al 2015). Recent advances in high throughput sequencing technologies have enabled fungal scientists to study AMF genomics and symbiotic interactions of microbial community inhabiting in spores of AMF (Bianciotto et al 2011). Environmental scientists have applied AMF inoculants for cleaning up contaminated soils- phytoremediation[2] (Iffis et al 2014). Researchers in the area of agriculture have made efforts to develop biofertilizers based on AMF inoculation (Zarik et al 2016, Hijri et al 2016).

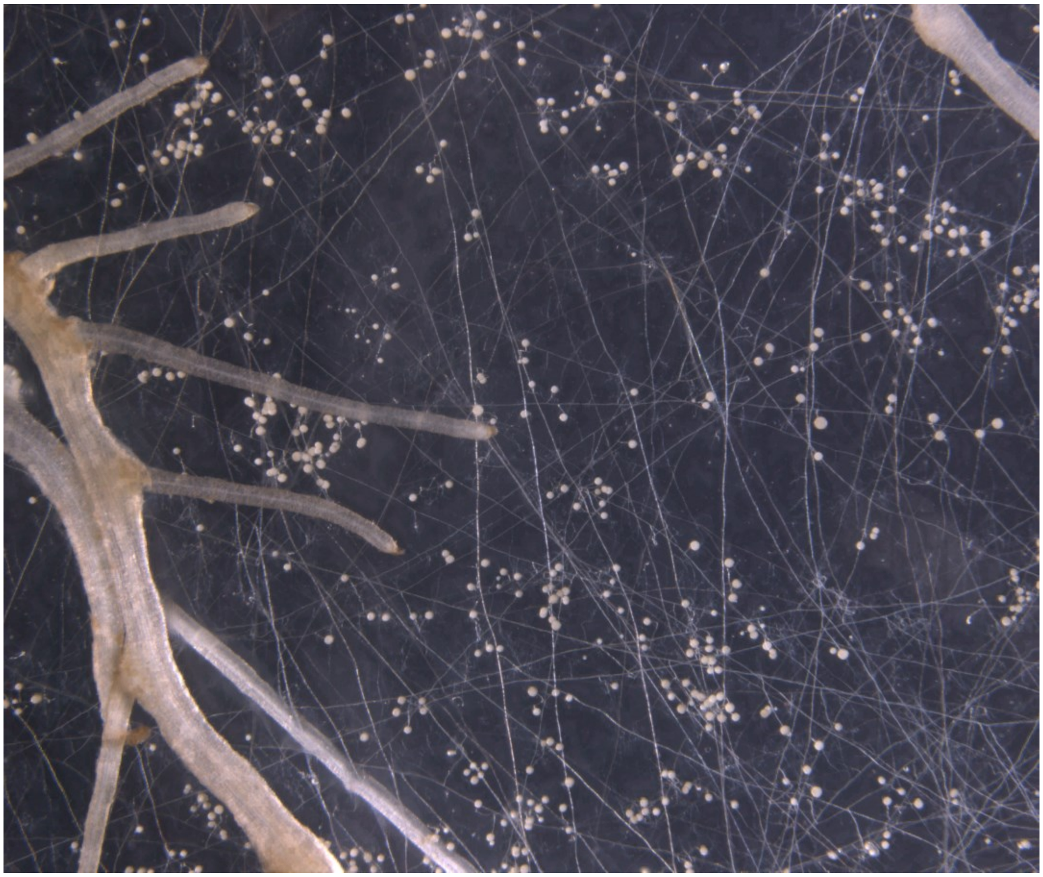


**Figure 1** Carrot roots colonized by AMF. The extensive networks of mycelia increase absorbing surface of water and

nutrients (Source: Hijri's labs).

Chemical fertilizers in modern and intensive agriculture cause leaching of phosphate which is a nonrenewable natural resource (Roy-Bolduc and Hijri, 2011). Moreover, amendment of soil with phosphate has been shown to negatively influence the structure of soil microbial communities (Beauchemin et al 1999). In addition, it is well known that use of pesticide contaminates soil and water and deteriorates ago-ecosystems. A large number of researchers have documented positive contributions of AMF to agriculture (Hassan et al 2013, Zarik et al 2016, Hijri et al 2016). AMF inoculation helps plants uptake phosphorus and nitrogen in low input agriculture (Hassan et al 2013). Likewise, AMF inoculation increased yields of potatoes in large-scale agriculture due to improvement of nutrient uptake via their extensive networks of mycelia$^3$ (Hijri 2016). Figure 1 shows that the carrot root colonized by AMF. Extensive AM fungal mycelia function as extension of root and significantly increase the absorbing area of nutrients (Roy-Bolduc and Hijri, 2011). Plants inoculated with AMF showed higher mineral uptake and hydration status compared to non-mycorrhized plants, which suggests that AMF ameliorate drought stress (Zarik et al 2016).

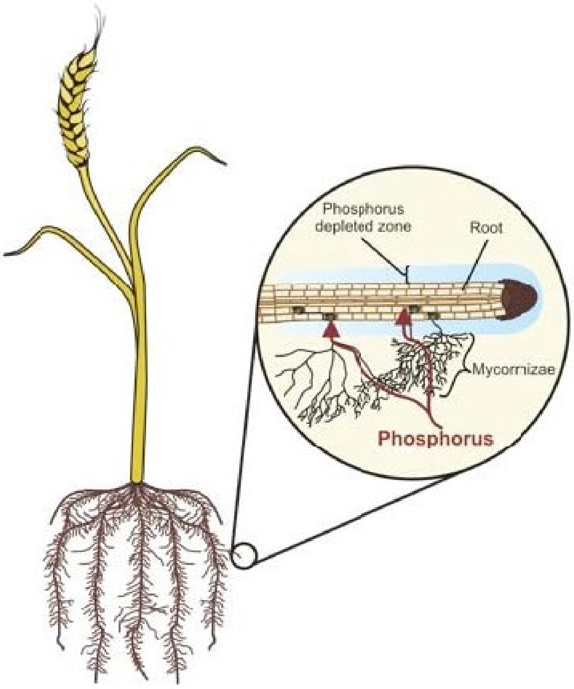


**Figure 2** Typical mycorrhized plant showing extraradical AMF hyphae. One of the principal host benefits is the increased uptake of phosphorus. Phosphate ions in soil are largely unavailable to roots because they form insoluble complexes with naturally occurring metal cations. Fungal hyphae are able to extend beyond the root depletion zone, taking up bioavailable phosphate which is outside the reach of the plant (source: Roy-Bolduc and Hijri 2011).

Moreover, AMF colonization also stimulated bacterial growth in rhizosphere; the total number of bacterial population was higher in mycorrhized plant than in non-mycorrhized plant (Johansson JF et al 2004). AMF inoculation is sustainable green technology based fertilizer as an alternative to chemical fertilizers and pesticides (Figure 2) (Roy-Bolduc and Hijri, 2011).

In addition to the importance of AMF in agriculture, they also play key roles in phytoremediation (Iffis et al 2014). A number of researchers documented that AMF and their associated microbial communities inhabiting in plant's root are key determinants in the effectiveness of phytoremediation (Iffis et al 2014, Hassan et al 2014, Marchand et al 2016, de la Providencia I et al 2015, Chanda et al 2014). Marchand et al investigated microbial capacity for biodegradation[4] of contaminants with respect to soil origin, type of culture media, and strain taxonomy. They documented that only microbial taxonomy had a significant impact on the effectiveness of biodegradation of contaminants (Marchand et al 2017 (a)). In addition, aided phytoremediation in biopiles and co-planting effectively cleaned up co-contaminated soil containing multiple contaminants such as petroleum hydrocarbons and metals (Marchand et al 2017 (b)).

## 1.2 AMF genetic structure

It has never been successful to grow AMF in pure culture without a host plant due to their strict symbiotic interaction- biotrophic[5] life-cycle (Kuhn 2003). Unlike most fungi, it is suggested that AMF may not undergo single-nucleus stage (Marleau et al 2011). Their mycelia are formed by coenocytic[6] and multinucleate[7] hyphae[8] that reproduce through asexual multinucleated spores. Nuclei massively migrate from parental hypha to a child spore during sporulation (Marleau et al 2011). Marleau et al. stained the nuclei with cytogreen and used confocal microscopy with time-lapse sequence monitoring system to observe how the nuclei move from parental hypha to a child spore. The stained nuclei moved unidirectionally into the spore. They discovered that the number of nuclei in a spore varies widely among individual AMF isolates from some hundreds up to some thousands, and it is estimated that a spore can contain up to thousands of nuclei. In addition, they also speculated that AMF may undergo mitosis although it may be a small portion of nuclei. Although the extent of heterogeneity is unknown, nuclei in a spore are believed to be heterogeneous.

Boon et al used genetic markers to estimate the degree of polymorphism, and reported that AMF had a very high level of polymorphism- as high as 103- in some loci (Boon et al 2015). Comparisons of allele distributions among parent isolate and sister spores suggest that smaller genetic diversity passes onto sister spores during sporulation (Boon et al 2013). In addition, they performed the comparative study of the WGS data of AMF with those of other fungal genomes using clustering approach. They concluded that it is possible that AMF may be heterokaryons[9]. Assuming that AMF are heterokaryons, it is important to assess the extent of heterogeneity among nuclei. Hijri et al documented that telomere[10] associated sequences are promising molecular markers for heterogeneous nuclei in AMF (2007). The variations in telomere regions may enable us to make inference about genetic variation among nuclei and to study genome structure of *Glomus intraradices*-

the multigenomic fungus.

## 1.3 Microorganisms associated with AMF mycelia

Numerous studies reported that AMF mycelia harbor a large number of microorganisms that are residing inside or on the surface of their spore and hyphae (Figure 3) (Lecomte et al 2011, Bonfante et al 2003).

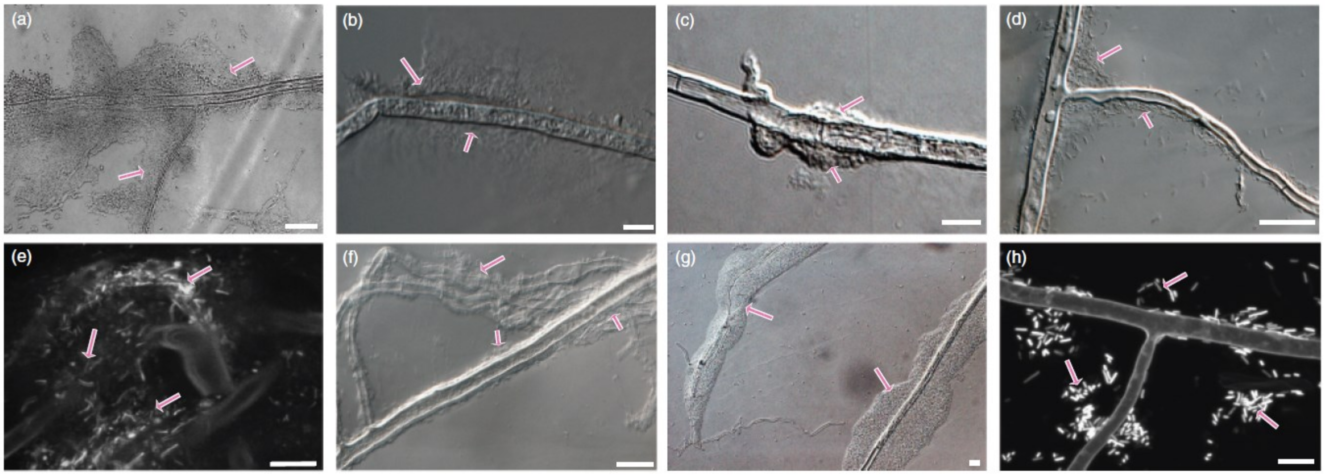


**Figure 3** Bacterial growth patterns on *Glomus sp.* hyphae cultivated *in-vitro* and observed with a DIC microscope using a ×63 objective (a) *Bacillus simplex*; (b) *Kocuria rhyzophila*; (c) *Bacillus megaterium*; (d) *Variovorax paradoxus*; (e) *Sphingomonas sp*.; (f) *Microbacterium ginsengisoli*; (g) *Pseudomonas sp*.; and (h) *Escherichia coli*. (a) and (g) bacteria are shown after 15 days of growth; (b) and (c) bacteria after 30 days of growth while (d)–(f) and (h) bacteria are shown after 45 days of growth. Images in (d), (e) and (h) were acquired using confocal microscopy. Scale bars=10µm. (source: Lecomte et al 2011).

We defined three types of symbiotic interactions between AMF and their associated microorganisms. The first type is the loose interaction between AMF and microorganisms in soils. The second type is the intimate interaction because microorganisms tightly adhere to the surface of the spore or of mycelia of AMF (Iffis et al 2016). The third type is the interaction between AMF and endosymbionts living inside of their spores and mycelia. Since the first report of the occurrence of bacteria-like organelles inside the spore of the AMF by MacDonald et al. (1982), a number of studies have documented different types of endosymbionts (Hijri et al 2002, Cruz et al 2008, Cruz et al 2012, Bonfante 2003, Naito et al 2015, Torres-Cortes et al 2015).

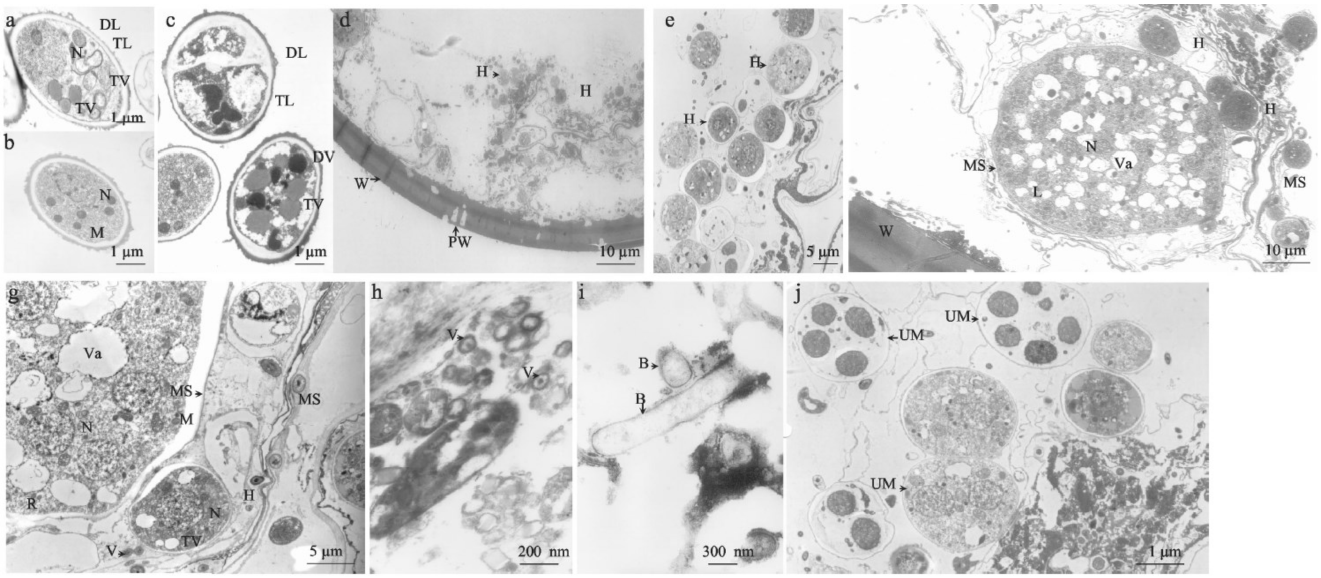


**Figure 4** Electron micrographs shows similarities between the isolated fungi and fungal structures inside the *Scutellospora castanea* spores and also the presence of other microorganisms. Panels a and b show ultrastructural features of the *Nectria sp.* isolated from cloudy spores of *S. castanea*. Panel c shows Leptosphaeria sp. isolated from healthy spores of *S. castanea*. Panels d and e show ultrastructure of cloudy nonviable spores. Panel f and g show healthy *S. castanea* spores. Panel h shows healthy spores of *S. castanea* harboring virus-like particles. Panel i shows bacterium-like organisms present in healthy spores of *S. castanea*. Panel j shows other eukaryotic unidentifiable microorganisms observed inside *S. castanea*. B, bacterium-like organisms; DL, electron-dense layer; DV, electron-dense vacuole; H, hyphae; L, lipids; M, mitochondria; MS, membrane system; N, nucleus; PW, perforated wall; R, ribosomes; TL, electron-transparent layer; TV, electron-transparent vacuole; UM, unidentified microorganisms; V, virus-like particle; Va, vacuole; W, spore wall. (source: Hijri et al 2002).

In studying endosymbiont, distinctions need to be made with respect to inheritability- capable of being inherited- and to the ability of free-living- capable of maintaining its own reproduction and metabolic systems. Some portions of inheritable endosymbiotic microbial organisms that have lost such systems and depended on AMF may be in the process of being endosymbiotic organelles. Endosymbiotic organelles- mitochondria and hydrogenosomes are known to exist in fungi (Hackstein et al 2007).

Cruz et al have discovered presumably free-living bacteria in *Gigaspora margarita* (2008). The bacteria were isolated using osmosis method applied to a protoplast[11] which was derived from a spore by treatment with series of enzymes. They were identified as bacteria close to *Janthinobacterium lividum* and *Paenibacillus polymyxa*. They were cultivated in cell-free media. They showed ability to solubilize phosphorus. *J. lividum* showed an ability to suppress plant pathogen. Further experiments on two bacteria increased a possibility that they were derived from inside of the spore. The same research group used another method- hypodermic ultrathin needle in *Gigaspora margarita*, isolated free-living endosymbionts, and identified them as *Bacillus sp., Bacillus thuringiensis*, and *Paenibacillus rhizospherae* (Cruz et al 2012). They documented that all of the bacteria made more than one of the following contributions: phosphorus solubilization, ethylene production, nitrogenase activity, and stimulation of hyphal growth. In addition to free-living bacteria, two filamentous fungal endosymbiont were found inside spores of *Scutellospora castanea* using transmission electronic microscope (Figure 4), and identified as fungi belonging to *Necteria* and to *Leptosphaeria* (Hijri et al 2002). The discovery supported the presence of fungal sequences other than those of glomalean origin in the WGS data of the AMF.

In contrast, several researchers isolated obligate endosymbionts in major AMF species, and identified them as heritable endobacteria called Mycoplasma-Related Endobacteria (MRE). MRE have experienced genome reduction, especially genes relating metabolism (Naito et al 2015, Torres-Cortes et al 2015). Their genomes were distinct from each other due to a high degree of adaptation to their hosts via genome reduction, genetic recombination, mobile elements, trans-kingdom horizontal gene transfer from the fungal host, and plectroviral invasion. It was also documented that AMF harbored another kind of heritable obligate endosymbiont belonging to *Candidatus Glomeribacter gigasporarum* (Jargeat et al 2004). It is a beta-proteobacterium that has a close phylogenetic relation with *Burkholderia*. In addition, microorganisms associated with AMF were isolated and identified as bacteria belong to a novel unidentified group during the process of production of *in-vitro* AMF. Further experiments showed that the unidentified bacteria-like organisms were presumably obligate endophytes[12] with bacterial origin (Gulbis et al 2013).

## 1.4 Effectiveness of AMF inoculants in sustainable agriculture and phytoremediation

AMF associated bacteria have a great impact not only on the fitness of host plant but also on that of AMF (Hafidi et al 2015, Iffis et al 2016). Johansson et al performed morphological, physiological, and biochemical experiments and showed that bacteria in mycorrhizosphere[17] affect AMF in a number of different ways, such as activation/ inhibition of sporulation and promotion of mycorrhizal development through nitrogen fixation (2004). However, experiments in laboratory setting are restricted to bacteria culturable in laboratory conditions and to those with known functions. Microorganisms associated with AMF grown *in-vivo* provide invaluable information in studying evolution and symbiotic interactions. More than 99% of soil microorganisms are not culturable in laboratory conditions. Recent advances in high-throughput DNA/ RNA sequencing technologies have opened new ways in investigating a large-scale characterization of genes in sequencing data from AMF grown *in-vivo* and enabled researchers to study interactions between AMF and its associated microorganisms (Bonfante et al 2009). Taxonomical classification is essential not only for the WGS of the AMF grown *in-vivo* but also for environmental data from AMF mycorrhizosphere and hyphosphere of agricultural and phytoremediation projects.

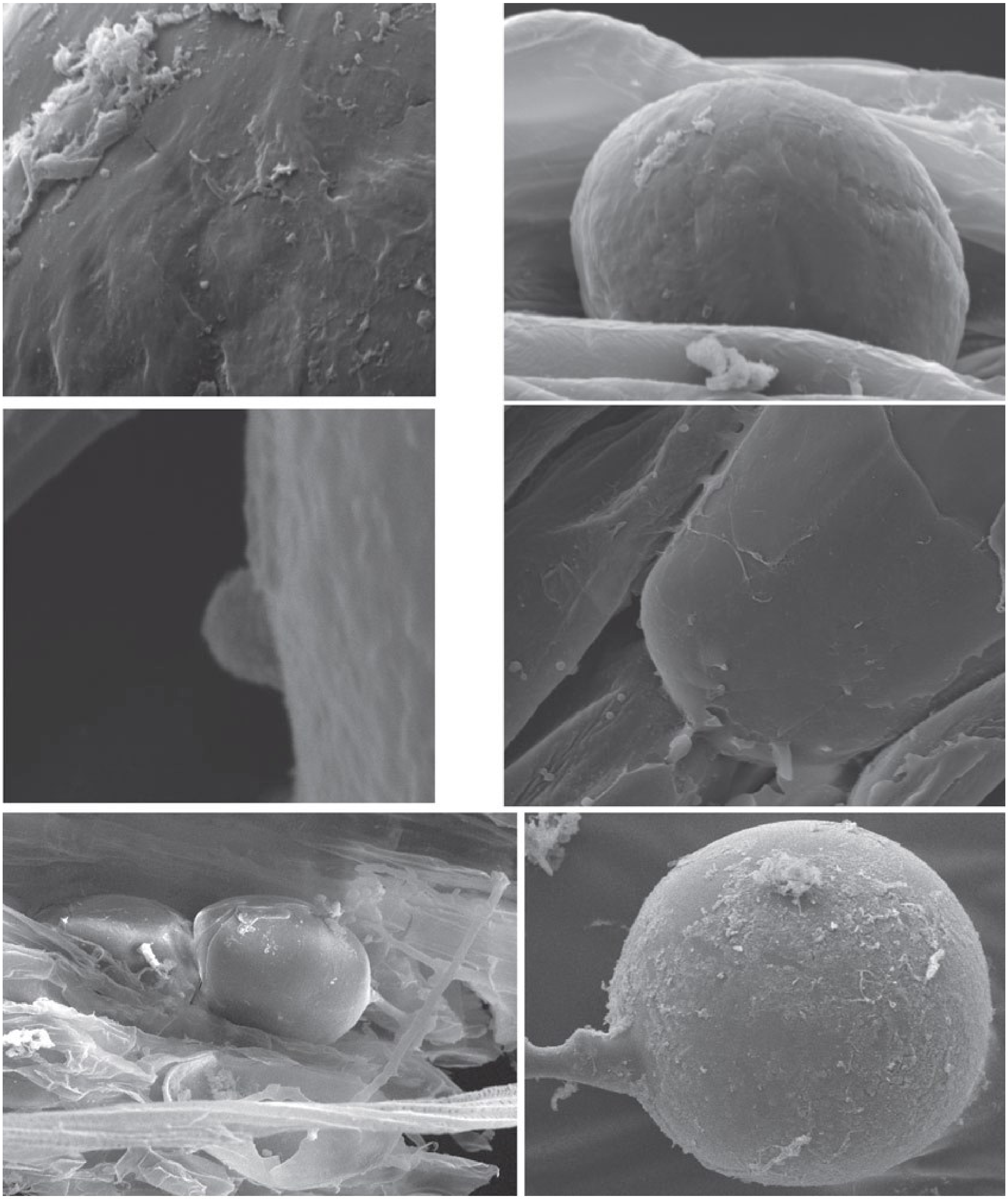


**Figure 5** AMF colonized root of *Solidago rugosa* growing in petroleum contaminated soil was sampled and prepared for scanning electron microscopy (SEM). SEM shows that bacterial cells and bio-film like structures are attached to AMF hyphae (H) and propagules (P). Panel b and c shows the magnification of the selected section in panel a and d, respectively. Panel f shows many microorganisms attached to the cell wall of AMF spore isolated from the rhizospheric soil of *S. rugoda* after it was washed multiple times with sterilized water (Iffis et al 2014).

Iffis et al showed that microorganisms were intimately associated with AMF that colonized roots of plants growing in soils contaminated with petroleum (Figure 5) (2014). Several studies have shown that microbiota in hyphosphere are key determinants in optimizing phytoremediation process (Iffis et al 2017). Bell et al showed the importance of microbiota containing a variety of microorganisms for effective phytoremediation. They assessed the efficiency of bioremediation with three different microcosms (2016). Initial soil was collected from place adjacent to former petroleum refinery in order to sample uncontaminated soil. Three microcosms were derived from the initial soil and prepared with sterilized soil: one reinoculated with initial soil, another with all bacteria isolated in regular media, and the other with all bacteria isolated in media containing crude oil. It showed that the sterilized soil reinoculated with initial soil showed maximum biodegradation capacity after 6 weeks. Soil reinoculated with initial soil showed higher efficiency compared to the other two, probably because the microcosm contains the most diversity of microorganisms and consequently the greatest variety of functional gene pools for dynamically governing the degradation of crude oil.

AMF and their associated microbial communities are proven to be the key factors determining the success of sustainable agriculture as well. Recent discoveries also proved beneficial roles of endosymbionts inside of AMF spore, such as nitrogen-fixing, phosphate solubilizing, and plant growth promoting abilities (Cruz et al 2008, 2012). Considering over-fertilization disrupts the balance of ecosystems and has negative environmental impacts, AMF inoculation may be an efficient approach for replacing the inadequate agricultural practices and for establishing sustainable agroecosystem (Roy-Bolduc and Hijri, 2011, Barea et al 2002).

However, AMF inoculation in agricultural practices has encountered challenges, because interactions between host plants and AMF species cannot be predicted due to high intra isolate variation of AMF and to the difficulty of marker development for AMF taxonomical identification (Zimmerman et al 2009). AMF genetics is poorly understood due to their coenocytic and multinucleate nature. With current molecular biology technology, we have limitations in studying the fundamental genetics of AMF such as genome segregation, a degree of heterogeneity of nuclei, and distribution of essential and functional genes across nuclei. Furthermore, it is challenging to remove and cure AMF from their associated microbes with current molecular biology technology. In contrast, with currently available omics[18] technologies and a vast amount of sequencing data, computational biology and bioinformatics may be able to provide key insights to solve the bottlenecks encountered in AMF research. Therefore, it is important to develop bioinformatics programs for taxonomical classification and for functional analysis of the WGS data of the AMF.

## 1.5 Commonly used methods for taxonomical classification of bacterial and fungal sequences from environmental samples

Taxonomical classification of microorganisms had been solely dependent on phenotypic analyses- morphological, physiological, and biochemical characterization- until genotypic analyses based on DNA information became available several decades ago. Taxonomical classification based on small subunit (SSU) ribosomal RNA (rRNA) has been most commonly used for prokaryotic organisms for past two decades (Stackebrandt et al 1997).

SSU rRNA genes have advantage in making phylogenetic inference because they are much conserved among prokaryotic organisms. However, for the same reason, their discriminating power may not be sufficient for some microorganisms. Therefore, a number of researchers add complementary sequences such as protein coding sequence (CDS) in addition to SSU rRNA genes to improve the resolution. Several researchers have improved the method based on SSU rRNA by incorporating the sequence profiling signatures that have been widely used for taxonomical classification of prokaryotic organisms in the area of metagenomics; for example, nucleotide sequence patterns with high discriminating powers were identified in 16 SSU rRNA genes and incorporated into the method for taxonomical classification of *Bacillus* with different levels (More et al 2016). In addition, barcode system such as Quick Response[13] has been employed for accelerating the process of the taxonomical classification. The strength of the taxonomical classification method based on the SSU rRNA genes includes the improvement made with sequence profiling features, informative databases, and numerous software programs that are convenient to use. However, recent studies have documented the mosaicism due to horizontal gene transfer in 16 SSU rRNA gene and heterogeneity of multiple rRNA genes within a single microorganism (Rajendhran 2010). The discovery has biologically significant importance for studying new mechanisms governing translational processes under environmental stresses. On the other hand, it has complicated the interpretation of taxonomical classification results.

As an alternative to 16 SSU rRNA gene, protein-coding genes such as heat-shock proteins have been also used for prokaryotic taxonomical classification. When assigning a new strain, phylogeny study plays an important role. A large number of researchers have constructed prokaryotic phylogenetic trees based on protein coding genes (Golding et al 1995, Ahmad et al 1999). However, some researchers believe that protein coding genes have higher discriminating power but weaker representation of phylogeny (Glaeser et al 2015). To reduce a bias produced by single gene based method, multiple gene based method has been gaining its popularity. Furthermore, a new approach, multilocus sequence analysis, has been developed; DNA fragments of several protein coding genes are used for prokaryotic taxonomical classification, and concatenation of their fragments are used for phylogeny study (Glaeser et al 2015). In addition, a number of researchers have developed new approaches. For example, Gupta et al extracted entire coding genes per bacterial genome, and identified taxon-specific genes that were unique to each taxonomic group using eggNOG (a database of orthologous groups and functional annotation) and Blast program (2015). They documented that the taxon-specific gene approach provides more accurate method of bacterial taxonomical classification.

Table 1 rRNA : Prokaryotes (*Escherichia coli*) vs. Eukaryotes (human)

| Type | Size | Large subunit | Small subunit |
|---|---|---|---|
| Prokaryotic | 70S | 50S (5S : 120 nt, 23S : 2906 nt) | 30S (16S : 1542 nt) |
| Eukaryotic | 80S | 60S (5S : 121 nt, 5.8S : 156 nt, 28S : 5070 nt) | 40S (18S : 1869 nt) |

Note : S in 16S represents Svedberg units. nt stands for length in nucleotide.

(Source : **https://en.wikipedia.org/wiki/Ribosomal_RNA**)

While SSU rRNA genes have been the most widely used markers for prokaryotic organisms, a variable region of mitochondrial cox1 gene has been formally chosen as barcode marker for animals (Schoch et al 2012). Consortium for the Barcode of Life chose the same region for fungi as well. However, another multi-laboratory consortium found that the region of the gene was too difficult to amplify. They chose six DNA regions as potential candidates, which included three protein coding genes (the largest subunit of RNA polymerase II, the second largest subunit of RNA polymerase II, and minichromosome maintenance protein). The protein coding genes had a higher correct percentage of taxonomical identification. But they were excluded as marker candidates due to the difficulty in amplication. They chose internal transcribed spacer (ITS) region as a marker for fungal taxonomical classification. In fungi, there are two ITSs- one located between 18S rRNA gene and 5.8S rRNA gene and the other located between 5.8S rRNA gene and 28S rRNA gene (Table 1). Two ITSs with the 5.8S rRNA gene between them are referred as ITS region. ITS region is commonly used for taxonomical classification of Dikarya because it has high inter-specific variations but low intra-specific variations (Lindahl 2013). However, ITS region is too variable to make an alignment of sequences from some fungal groups for phylogeny inference. To supplement the weakness, a new approach, the combination of ITS region with its secondary structures, has been developed for phylogeny study (Merget et al 2012).

However, in case of AMF, rRNA genes from nuclei have limitations due to high variations of inter- and intra- specific DNA and RNA sequences of nuclear genes (Sarma et al 2017). For the same reason, ITS region has limitations in AMF taxonomical identification as well. A number of DNA regions have been evaluated for AMF marker development (Sarma et al 2017). For past years, mitochondrial genes have attracted AM fungal researchers because they have been found to be homoplasmic[14] (Sarma et al 2017, Lee et al 2009, Beaudet et al 2013). In mitochondrial genes, especially introns have been considered suitable candidates for marker development (Nadimi et al 2012, Nadimi et al 2015). Mitochondrial genes, cox1 in combination with rnl or nad2, have also showed potentials for phylogeny study in AMF (Nadimi et al 2016). However, insufficient knowledge of mechanisms of the mitochondrial inheritance in addition to unknown genome structure of AMF have posed the challenges in AMF marker development (Sarma et al 2017). In addition to mitochondrial genes, Sokolski et al. documented partial sequences of inorganic phosphate transporter genes may be a good candidate for discriminating morphologically defined glomus species (2011).

## 1.6 Taxonomical classification of the WGS data of the AMF

### 1.6.1 Culturing methods and the WGS data of the AMF

Complex genome structure of AMF has posed limitation not only in developing AMF markers but also in analyzing their WGS data. Additionally, due to the high intra/ inter genomic variations of AMF strains, there is no reference genome for AMF. Furthermore, the WGS data of a single AMF spore contain a considerable portion of non-AMF DNA sequences. Depending on AMF species and culturing methods, diversity and abundance of microbial communities associated with AMF vary widely. We need to define at least two types of culturing method: *in-vivo* culture (typically with a host plant in a pot filled with soil) and *in-vitro* culture. There are several types of *in-vitro* cultures. Axenic system refers only AM fungus grown on agar-like substrates. Monoaxenic system refers AM fungus and a root organ of plant grow in a plate. Dixenic system refers monoaxenic system with another organism grown on a plate (Declerck et al 2005). Because a wide spectrum of antibiotics are used to initiate *in-vitro* cultures (Bécard and Fortin 1988), antibiotic cocktails kill the symbiotic microorganisms on the surface and inside of AMF spore. It has been documented that only few AMF taxa are able to be cured and cultivated in axenic cultivation system, and most successful isolates in such system belong mainly to the genus- *Rhizoglomus* (Declerck et al 2005). It suggests that AMF may be a meta-organism. In other words, symbiotic bacterial and fungal partners are indispensable from AMF.

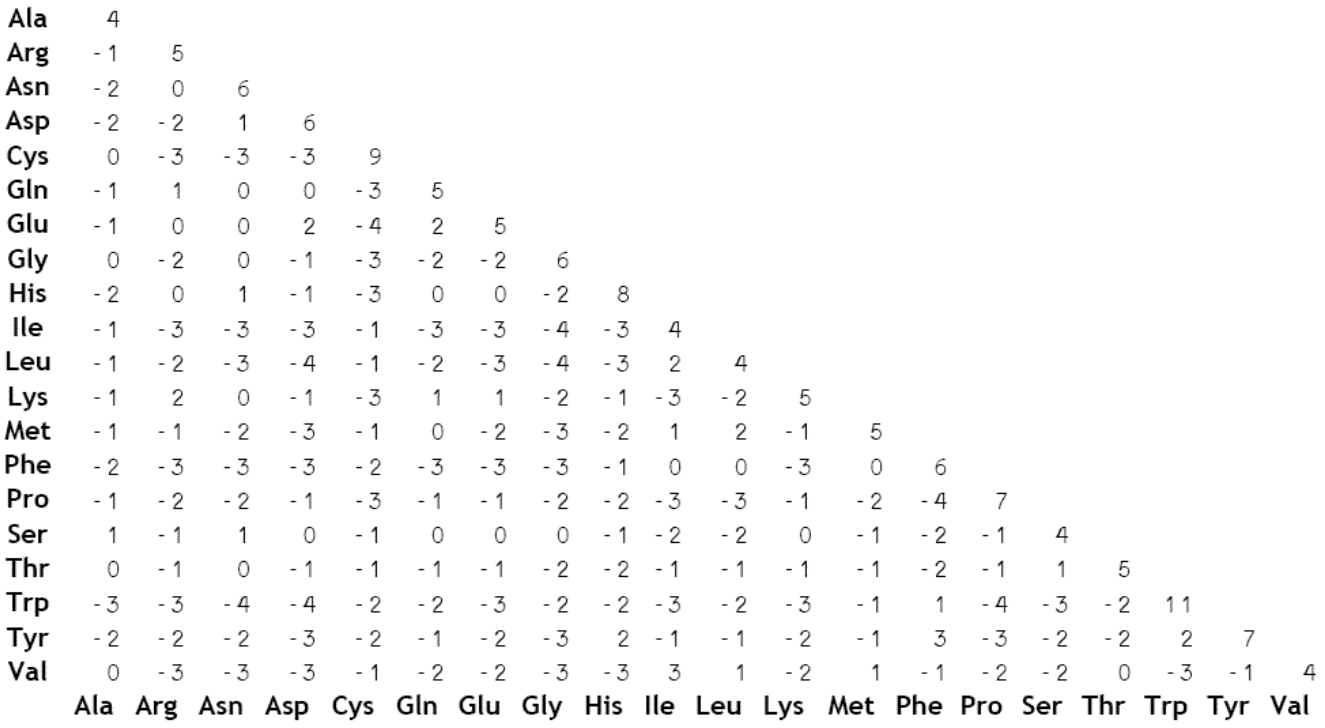
The WGS of the AMF taxa has been achieved exclusively from those grown in monoaxenic system. With *in-vitro* culturing method, the WGS data of the AMF contain less non-AMF sequences. Downside of *in-vitro* cultivation is the use of the wide variety of antibiotics that kill AMF spore associated microorganisms that provide crucial information about AMF. In contrast to *in-vitro* culturing method, the WGS data of the AMF taxa grown *in-vivo* contain a large amount of non-AMF sequences from their associated microorganisms. Considering that a great majority of soil microorganisms are not culturable in laboratory conditions outside of soil and that sequence databases represents a tiny fraction of existing microorganisms (Jeffery et al 2010), sequence information from *in-vivo* culture provides invaluable information on symbiotic interactions of AMF with their associated microbial communities.

### 1.6.2 Available genome of AMF and its limitation due to high inter- and intra- isolate variations

A spore is the culturing and sequencing unit for AMF WGS, and contains hundreds or thousands of nuclei that are believed to be heterogeneous although the extent of heterogeneity is unknown. In addition, endosymbionts in their spore add extra levels of complexity to the WGS data of the AMF. It was only a few years ago when Tisserant et al. published the genome information of *Rhizophagus irregularis* grown *in-vitro* (2013). They also published transcriptome information of *R. irregularis* and *Rhizophagus diaphanus* (Tisserant et al 2014). Although the published information of the genome and the transcriptome of *Rhizophagus* provide invaluable information for studying AMF genetics, it has limitations in serving as reference genome due to high inter/ intra isolate variations of AMF. The published genome and transcriptome may be a small portion of hundreds or even thousands of heterogeneous nuclei; due to high intra-isolate variations, the genome assembly of *R. irregularis* is challenging and the expression profiles in *R. irregularis* are incomplete. Due to unknown genome structure of AMF, we do not have a template that serves as guidelines for putting sequence puzzles together. In addition, the WGS data of the AMF contain sequences from symbiotic microorganisms inhabiting in the spore of AMF. For such reasons, I have taken the same approach as metagenome data analysis for taxonomical classification of the WGS data of the AMF grown *in-vivo*. In this thesis, I interchangeably use two terms -metagenome data and environmental data.

### 1.6.3 Existing approaches for taxonomical classification of metagenome sequencing data

In the area of metagenomics research, there are mainly two approaches for taxonomical classification: one is composition-based method that employs sequence profiling signatures such as compositional biases[15] of nucleotides (k-mers[16]) and the other is similarity-based method such as homology search method (Kim et al 2013).



| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 4 | | | | | | | | | | | | | | | | | | | |
| Arg | -1 | 5 | | | | | | | | | | | | | | | | | | |
| Asn | -2 | 0 | 6 | | | | | | | | | | | | | | | | | |
| Asp | -2 | -2 | 1 | 6 | | | | | | | | | | | | | | | | |
| Cys | 0 | -3 | -3 | -3 | 9 | | | | | | | | | | | | | | | |
| Gln | -1 | 1 | 0 | 0 | -3 | 5 | | | | | | | | | | | | | | |
| Glu | -1 | 0 | 0 | 2 | -4 | 2 | 5 | | | | | | | | | | | | | |
| Gly | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | | | | | | | | | | | | |
| His | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | | | | | | | | | | | |
| Ile | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | | | | | | | | | | |
| Leu | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | | | | | | | | | |
| Lys | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | | | | | | | | |
| Met | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | | | | | | | |
| Phe | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | | | | | | |
| Pro | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | | | | | |
| Ser | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | | | | |
| Thr | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | | | |
| Trp | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | | |
| Tyr | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | |
| Val | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |
| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |

**Figure 6** The BLOSUM62 matrix used in similarity based method
(source: https://en.wikipedia.org/wiki/BLOSUM)

Homology search program such as BLAST (Basic Local Alignment Search Tool) is the most widely used bioinformatics program in many areas of bioinformatics. BLAST provides a user with different choices of scoring methods, performs pair-wise alignments of a query sequence with sequences in databases, and provides a list of hit sequences of statistical significance that have functional annotations. Scores are computed based on amino acid substitution matrices (20 amino acids in row and 20 amino acids in column) such as BLOSUM (BLOcks SUbstitution Matrix) and PAM (Point Accepted Mutation) (Figure 5).

BLAST is often used for homology search in terms of protein function. When researchers have a sequence with unknown function, they use BLAST program to draw an inference about functions. However, because of lack of taxonomical classification programs for analyzing sequences from metagenomics data, BLAST has been also widely used not only for functional analysis but also for taxonomical classification. However, it is not suitable for taxonomical classification due to weak discriminating power. The amino acid substitution matrices such as BLOSUM were created based on multiple alignments of the most conserved regions of protein families from different species. The 20 by 20 matrix created regardless of function and of taxonomical group provides low resolutions for taxonomical classification of environmental data.

Contrary to the homology search programs, composition-based programs have been developed with the sole purpose- taxonomical classification of metagenome data. Unique sequence signatures have long been used to compare DNA sequences in many areas of bioinformatics. GC content and compositional biases have been used for taxonomic classification as well as for studying evolutions of different regions of the genome (Coenye et al 2003, Karlin et al 1997). Karlin et al documented that phylogenetically closely related groups of prokaryotes have similar patterns of compositional biases of nucleotides (1997). Dozens of metagenomics tools have been developed based on compositional biases for taxonomical classification and they have been gaining popularity.

### 1.6.4 Existing programs for taxonomical classification of metagenome sequencing data

Comprehensive pipelines such as Mothur and QIIME have been widely used for analyzing community sequence data based on 16S rRNA genes (Schloss et al 2009, Caporaso et al 2010). For fungal taxonomical classification, ITS regions of fungal species, that are available at the database UNITE, can be integrated with Mothur and QIIME (Kõljalg et al 2013). These programs have gained popularity for taxonomical classification of the community data from 16S rRNA gene amplicon sequencing. It has been also used for metagenome data that contain many different types of sequences- CDS, non-CDS, rDNA, and etc. In such case, researchers analyze microbial diversity with these pipeline software, and then separately apply homology search programs for inferring functionality of sequences. However, because the software does not have the capacity for taxonomical classification of sequences other than 16S rRNA gene, the approach has limitations in studying interactions of microorganisms. It enables them to infer functionality only of a whole environmental sample but not of taxonomic group. Additionally, methods based on percent sequence similarity for defining operational taxonomic unit in Mothur and QIIME have been criticized for lack of consideration of evolutionary distances (Nguyen et al 2016). Furthermore, multiple copies of rRNA genes within a single isolate may hinder accurate estimation of microbial diversity and quantification of taxonomic groups.

Composition-based approach has advantage over the SSU rRNA based method. It provides a means to classify not only rRNA genes but also other types of sequences such as protein coding genes. It enables users to draw a big picture of symbiotic roles of microorganisms in the community. Furthermore, estimation of taxonomical diversity and quantification of taxonomic groups can be calculated based on a large number of genes, which has less bias compared to single gene based method. Taxonomical classification programs based on composition-based methods include PhyloPythia, NBC, Phymm, Tacoa, and Taxsom (Kim et al 2013). Indus and Twarit are binning methods based on a range of biases of compositional patterns (Reddy et al 2012). In addition, sequence profiles have been used for gene/ exon predictor programs. Multiple gene predictors for eukaryotes have been developed based on sequence profiling features. Augustus is a gene prediction program for eukaryote genomes (Stanke et al 2004). GeneMark-ES employs unique features of fungal coding regions and unsupervised neural network to identify fungal genes (Ter-Hovhannisyan et al 2008). Z-curve method utilized additional factor, frequencies of phase-specific mono-, di-, tri – nucleotides, to detect exons in eukaryotes (Gao et al 2004).

## 1.7 Codon usage and codon context

### 1.7.1 The importance of codon usage and codon context

Over the last several decades, a number of studies have proven wrong the long presumed belief that codon usage[19] serves no biologically meaningful functions in transciption[20] and translation[21,] two main processes of the central dogma[22] (Angov et al 2011). Although the mechanisms remain largely unknown, codon usage bias appears to play important roles in gene regulation (e.g., gene expression, diversification of gene products, translational efficiency and accuracy, mRNA stability, and protein folding[23]). Recent studies have documented the regulatory roles of codon usage and of codon context[24]: regulation of folding dynamics of mRNA and of protein in transcription and translation processes (Angov et al 2011, Bartoszewski et al 2016, Baeza et al 2015, Behura et al 2012, Del Campo et al 2015, Chevance et al 2014, Costafreda et al 2014, Jacobson et al 2016, Khabou et al 2016, Komar 2016, Schieweck et al 2016, Yang 2017, Zhao et al 2017, Zhou et al 2016). Codon usage of multiple consecutive codons within mRNA secondary structures[25] plays critical roles in co-translational protein folding during protein synthesis (Harigaya et al 2017, McCarthy et al 2017). Furthermore, it was documented that non-optimal codons regulate circadian rhythms in response to change in environmental condition, which implies regulatory roles of codon usage managing environmental changes (Xu et al 2013, Zhou et al 2013).

### 1.7.2 Microorganism's survival strategy- codon usage and codon context

In general, mutational bias, genetic drift, and natural selection are believed to contribute to codon usage bias (Hershberg et al 2008). Ermolaeva DM reviewed the important factors that may affect codon usage preference: translational selection, GC composition, strand-specific mutational bias, amino acid conservation, protein hydropathy, transcriptional selection and RNA stability (2001). These factors are believed to make varying contributions to codon usage bias, which is one of microorganism's unique sequence signatures.

A number of researchers have performed comparative analysis with various taxonomic groups to study factors affecting codon bias. Chen et al. found that genome GC content and context-dependent nucleotide bias well discriminated codon bias among different organisms by applying singular value decomposition to codon usage biases of 100 eubacterial and archaeal organisms (2004). They suggested that mutation was the primary cause while the translational selection was the secondary cause of codon usage bias. In the same vein, Suzuki et al. suggested that GC content contributed most while translational selection contributed less to the overall codon usage diversity, respectively (2009). Because translational selection is driving force of codon bias for more efficient and accurate translation, the process is related to codon optimization. They showed that because codon optimization correlates with mRNA levels, it may be able to detect the genes involved in a microorganism's adaptive changes in response to environmental stress in a thermophilic microorganism. Carbone et al. employed statistical methods for dimensionality reduction to study codon bias space (2005). They reported that codon preferences discriminated between thermophiles and mesophiles as well as between aerobic microorganisms and anaerobic microorganisms.

Recent studies documented that a microbe has employed codon usage, codon context, and amino acid composition as its survival strategy to adapt to abiotic stresses (Su et al 2016, Ding et al 2012, Paul et al 2008, Sanjukta et al 2012). A microorganism has a unique range of synonymous codon usage due to individual's evolutionary path (Lee et al 2010, Akashi et al 1994, Grantham et al 1981). Because microorganism’s genome shows the adapted state at present rather than mutational changes over time, researchers have performed comparative study using omics data from extremophiles and non-extremophiles and provided useful insights into roles of codons (Su et al 2016, Ding et al 2012, Paul et al 2008, Sanjukta et al 2012). Comparative analysis on genomics from thermophilic[26], mesophilic[27], and psychrophilic[28] fungi demonstrated that thermophiles and psychrophiles preferred G or C ending codons while mesophiles preferred A or T ending codons (Su et al 2016).

In addition to codon usage bias, amino acid context has played important roles in microorganism's adaptation as well. Protein structural adaptation of extremophiles showed that amino acid composition reflects microorganism's evolution (Ueno et al 2016). Raymond-Bouchard et al. performed comparative analysis on 5 cryophilic permafrost bacteria and their mesophilic relatives and showed that cryophiles had more cold adapted proteins with more serine and fewer prolines/acidic residues

(2018). McDonald investigated what caused amino acid usage bias (2001). With an assumption of that asymmetrical pattern of amino acid substitution on mesophiles and thermophiles reflected selection toward a particular amino acid with respect to a range of temperature, he performed comparative analysis on patterns of the substitutional asymmetry in the mesophilic and thermophilic microorganisms belonging to the same genus or with closely related phylogeny. He concluded that the universal biochemical properties of amino acid and GC content were not sufficient to explain all the asymmetries. He suggested that taxon-specific properties of amino acid must have contributed to the asymmetry as well.

These studies all support our novel strategy; we have completed the development of novel bioinformatics programs for taxonomical classification of metagenome data from the WGS of the AMF based on usage bias and context bias of multiple consecutive codons and amino acids, because we have discovered that they are taxonomically unique sequence property.

## 1.8 Research problem, research objectives, and the contribution of the developed programs to AMF research

### 1.8.1 Research problem and objectives

As I explained in the previous sections, a spore of single AMF isolate contains hundreds or even thousands of nuclei that are speculated to be heterogeneous (Boon et al 2013, Boon et al 2015, Hijri et al 2007). In addition, AMF harbor a large number of symbionts inside of their spores and mycelia, especially if they are grown *in-vivo* (Iffis et al 2016). Therefore, the WGS of the AMF results in metagenome. Due to the inter/ intra genomic variations among isolates, the available genome information has limitation for serving as a reference genome (Sarma et al 2017, Marleau et al 2011, Boon et al 2015, Hijri et al 2007). Moreover, inheritable symbionts inhabiting inside of AMF mycelia have experienced genomic evolution for host-dependent adaptation (Naito et al 2015, Torres-Cortes et al 2015). Due to lack of completely sequenced genomes of inheritable symbionts of the AMF as well as those associated with the AMF, taxonomical classification of the WGS data of the AMF grown *in-vivo* is a complex research problem.

The main theme of my doctoral research was to develop bioinformatics programs for analyzing the WGS data of the AMF.

More specifically,

1) To develop a bioinformatics program for taxonomical classification of the WGS data of the AMF for studying symbiotic relationships of microorganisms associated with AMF

2) To develop a bioinformatics program for studying functions of novel gene candidates from the WGS data of the AMF

3) To develop a bioinformatics tool for studying evolution of genetic codes and undiscovered properties of codons

I developed three different types of bioinformatics programs according to the objectives. In the following section, I will explain the competitive advantages of each program compared to existing methods.

## 1.8.2 The contribution of the developed programs to AMF research

### 1.8.2.1 SeSaMe: A novel program for taxonomical classification of the WGS data of the AMF

In the previous sections, I reviewed several mainstream approaches for taxonomical classification of metagenome sequencing data which I will briefly recall here. Similarity-based method for taxonomical classification is ineffective tool due to low resolution especially for short sequences. Another widely used approach is based on 16S rRNA gene for amplicon sequencing data. In such case, researchers employ two methods: taxonomical classification method for estimating microbial diversity and homology search method for inferring functions of the whole metagenome sequencing data. However, researchers have difficulty in studying symbiotic relationships of microorganisms because the combination of these two methods does not offer a means to infer functionality of sequences within a taxonomic group. Compared to the approaches mentioned above, composition-based method classifies not only rRNA sequences but also CDS/ non-CDS, which enables user to study symbiotic relations of microorganisms in environmental sample. Moreover, this approach may have higher accuracy for estimating diversity and abundances of taxonomic groups, compared to single gene based method.

However, existing programs of composition-based method are designed to classify prokaryote sequences. They calculate observed frequencies of k-mers in entire genome without distinction between CDS and non-CDS. Because bacteria frequently encounter environmental stresses that they need to adapt to, their CDS has unique genomic properties resulting from adaptive mechanisms such overlapping genes. Therefore, compositional biases alone may have enough discriminating power for prokaryotic taxonomical classification. However, composition-based method alone is insufficient for fungal taxonomic groups. Although existing programs had reasonable capacity for taxonomical classification of bacterial sequences, they classified most of fungal sequences into bacterial group. Therefore, I developed a bioinformatics program called SeSaMe for taxonomical classification of the WGS data of the AMF grown *in-vivo*.

The major distinguished features of the developed program, compared to existing composition-based method, include the reference sequence database containing the sequence variables- *A.A. Char Trimer*, *A.A. Trimer*, and *Three Codon DNA 9-mer* that form a three level hierarchy-, and genus specific bias databases containing numeric variables, *trimer usage bias, A.A. Trimer usage,* and *three codon usage*, computed based on observed frequency ratios of *Three codon DNA 9-mers* and their cognate *A.A. Trimers*. The sequence variables have been created based on a large number of amino acid trimers with structural roles- subunits of protein secondary structures from Protein Data Bank (PDB) (Figure 2 of Chapter 2). The main variable, *trimer usage bias*, was computed based on codon usage of *Three codon DNA 9-mer* not within *A.A. Trimer* but within *A.A. Char Trimer* where *A.A. Char* was defined as a group of amino acids whose side chains have similar properties in terms of volume, polarity, and charge. The main variable adopted a broader set of amino acid characteristics, because transcriptional and translational regulators may add additional information to 3D structures of mRNA and protein, which may change biochemical properties of primary sequences. In general, *trimer usage bias* and *three codon usage* were shown to be taxonomically unique sequence property (Chapter 2).

While existing composition-based method relies on frequency of k-mers without taking their biological properties into consideration, the developed program perceives nucleotide subsequences with structural roles differently from those without one. Combination of codon usage with compositional pattern *of Three codon DNA 9-mer* encoding for protein secondary structure not only distinguishes CDS from non-CDS but also identifies an open reading frame. The program draws an inference for fungal group based on six different scores calculated from all reading frames of a query sequence, and therefore has higher resolution for taxonomical classification of short bacterial and fungal sequences in the WGS data of AMF (Figure 1 of Chapter 2). It is freely available at www.fungalsesame.org.

#### 1.8.2.2 SeSaMe PS Function: a novel program for position specific functional analysis of the WGS data of the AMF

Compared to genes of intensively studied organisms or of microorganisms culturable in laboratory conditions, the WGS data of the AMF may include a large portion of novel candidate genes which researchers may not be able to find homologues for in existing sequence databases. Moreover, considering a short history of molecular biology, biological systems of living organisms are so complex that a large portion of mechanisms remain to be unsolved. Even with recent advances in the molecular biology, further investigation may be required for not only novel genes but also genes with known function. Therefore, a majority of existing bioinformatics tools that rely on alignment with annotated sequences have low sensitivity toward undiscovered motifs with new structure and may not be sufficient to study novel properties of genes from the relatively ancient fungi- AMF.

Recent studies documented important regulatory roles of codon usage and codon context in mRNA structure and protein folding (Yang 2017, Harigaya et al 2017, McCarthy et al 2017). DNA sequence encoding for protein secondary structure contains information of mRNA secondary structures under assumption of their canonical base pairing. mRNA secondary structures have been documented to influence transcription initiation, to regulate mRNA splicing, and to guide protein folding by controlling translational elongation speed. A number of studies reported that changes in codon usage and codon context produced altered protein products. For example, Kimchi-Sarfaty et al documented that a synonymous single-nucleotide polymorphism in the gene- Multidrug Resistance 1 (MDR1) produced a protein product with altered substrate specificity. They hypothesized that the change from a frequent codon to a rare codon lengthened the time for co-translational folding, allowing an ion insertion into the structure. And their hypothesis implies the association between rare codon and a high degree of eccentricity from standard protein folding dynamics. For another example, McCarthy et al studied synonymous SNPs in disease related genes and documented that bicodons- synonymous SNP and its neighbor codon- had strong association with alteration of ribosome pause propensity. Codon context and codon usage of multiple consecutive codons also play important roles in protein folding. For example, rare codon and double stranded region of mRNA secondary structure have shown an association with slower velocity of translating ribosome that contributes to optimal folding of protein (Yang 2017). Several studies have made insightful suggestions linking mRNA structure to protein structure (Yang 2017). A role of codon usage may vary widely with respect to mRNA structure and the translated protein structure where the codon is located. Therefore, codon usage and codon context of multiple consecutive codons provide useful insights into their roles in folding of gene products.

I developed a bioinformatics program called SeSaMe Position Specific Function (SeSaMe PS Function) for position specific functional analysis of the WGS data of AMF. In contrast to existing programs that rely on sequence alignment, the developed program uses statistical methods to extract important subsequences based on comparative data created from usage information of subsequences of a query sequence. The main variable of the developed program, that incorporates both codon usage and amino acid usage of multiple consecutive codons, may provide important insights into structural differences among subsequences of a query sequence.

As a case study, we ran the program with 25 AMF CDS. The results showed that the program identifies outliers: subsequences with unique landscape pattern. Landscape pattern was defined as a XY scatter chart belonging to a subsequence where the variable on X-axis is taxonomic group and that on Y-axis is their usage bias value (Supplementary Figure 5 of Chapter 3). Each *Three codon DNA 9-mer* has its own landscape pattern. Landscape pattern is more accurate measurement because it indicates relative extents of usage information of 54 taxonomic groups. Sequences from 454 sequencing had lengths of the range between 100bp and 300bp, and contained several dozens of *Three codon DNA 9-mers* matched to the reference sequence database. While a majority of landscape patterns with similar shape were assigned into one major cluster, outliers with unique landscape pattern were assigned into a cluster with a single member or with only a few members. Considering that codon usage of *three codon DNA 9-mer* may reflect intrinsic property of undiscovered mechanism of a taxonomic group, outliers may play distinctive roles involving in molecular interactions. Because the program identifies outliers based solely on the comparative dataset measured in 54 genera, rather than based on sequence alignment of a query sequence against motifs with known function, it may have high sensitivity for novel motifs of undiscovered mechanisms. The program provides a useful means for studying novel genes from the WGS data of the AMF. SeSaMe PS Function is freely available at www.fungalsesame.org.

#### 1.8.2.3 Post: a novel bioinformatics program for studying codon property and codon assignment

AMF are special in that they harbor a large number of microorganisms inside of their spores and mycelia. Recent studies have identified several obligate endosymbionts of AMF such as MRE and Candidatus Glomeribacter gigasporarum (Naito et al 2015, Torres-Cortes et al 2015, Jargeat et al 2004). It is speculated that some of endosymbionts in AMF may be in the process of being endosymbiotic organelles. Besides, AMF are ancient fungi whose genome contains hundreds or even thousands of nuclei that are believed to be heterogeneous (Marleau et al 2011, Boon et al 2015, Hijri et al 2007). It is possible that some nuclei may result from inheritable symbionts that were acquired during AMF evolution. Inheritable endosymbionts may have retained properties of ancient transcriptional and translational apparatuses that have been lost in many free-living microorganisms. According to recent studies in mitochondria and chloroplasts, the number of genes included in these organellar genomes varies widely from a few genes to hundreds of genes. Genes for rRNA, tRNA, and ribosomal proteins are often encoded by the organellar genome. If some other factors including aminoacyl-tRNA synthetases (aaRs) are encoded by nuclear genome (Brandao et al 2011), nuclear genome often contains different versions of proteins involving in translation, e.g., one for nuclear genome and another for organellar genome. Furthermore, recent studies have documented that non-coding RNAs play important regulatory roles in transcriptional and translational processes (Sun et al 2015, Mathy et al 2017, Herriges et al 2018, Bazin et al 2017). Heterogeneous transcriptional and translational apparatuses tailored to mitochondrial genome may have contributed to the mitochondrial specific codon assignments that are different from the standard genetic code table, for example, UGA for Trp, AUA for Met, AGR for Ser and stop codon, AAA for Asn, CUN for Thr, and UAA for Tyr. Considering that AMF may have a large number of inheritable endosymbionts other than mitochondria, they may contain various transcriptional and translational apparatuses targeting endosymbionts. Consequently, their codon assignments may not comply with the standard genetic code table and vary considerably.

Considering that the current knowledge of genes and proteins has been obtained mostly based on intensively studied model organisms, we may need to study the origin of the genetic code to expand our perspectives in codon property and assignment. Mainstream hypotheses addressing the origin of the genetic code claim that codon position has played important roles in its evolution; the standard genetic codes show an association between the property of the nucleotide either in the first or the second position and that of the cognate amino acid.

We can assess the variation of codon (re)assignment and the diversity of the transcriptional and the translational apparatuses with an immeasurable amount of omics data (Hernández et al 2012). I developed a novel bioinformatics method- POsition Specific genetic code Table (Post) that assigns a codon with respect to nucleotide position in the codon. The developed program may provide researchers with a systematic tool for studying novel genes or new mechanisms in gene organization in context of codon property. The program is versatile and can be used for many types of research objectives. For example, it can be employed to conduct comparative study of irregular codons across different taxonomic groups as shown in a case study in the chapter 4. Or it may be also used to study long non-coding RNAs across different gene types. The Post is freely available at www.codon.kr.

# SeSaMe: Metagenome Sequence Classification of Arbuscular Mycorrhizal Fungi Associated Microorganisms

Jee Eun Kang[1,*], Antonio Ciampi[2], Mohamed Hijri[1]

[1] *Institut de Recherche en Biologie Végétale, Département de Sciences Biologiques, Université de Montréal, Montréal, QC, H1X 2B2, Canada*
[2] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC, H3A 1A2, Canada*

Jee Eun Kang orcid.org/0000-0003-2475-0474

Antonio Ciampi orcid.org/0000-0003-4838-8297

Mohamed Hijri orcid.org/0000-0001-6112-8372

E-mail: jennifer.kang@umontreal.ca (Kang JE) – journal format

Running Title: *Kang JE et al/ SeSaMe: Spore associated Symbiotic Microbes*

| | Words | Figures | Tables | Sup. Fig. | Sup. Tables |
|---|---|---|---|---|---|
| Total Counts | Approx. 7164 | 5 | 2 | 7 | 12 |

## 2.1 Abstract

Arbuscular mycorrhizal fungi (AMF) are plant root symbionts that play key roles in plant growth and soil fertility. They are obligate biotrophic fungi that form coenocytic multinucleated hyphae and spores. Numerous studies have shown that diverse microorganisms live on the surface and inside their mycelia, resulting in a metagenome when whole genome sequencing (WGS) data are obtained from sequencing AMF cultivated *in-vivo.* The metagenome contains not only the AMF sequences, but also those from associated microorganisms. In this article, we introduce a novel bioinformatics program- SeSaMe- designed for taxonomic classification of short sequences obtained by next-generation DNA sequencing. A genus-specific usage bias database was created based on amino acid usage and codon usage of three consecutive codon DNA 9-mers encoding for an amino acid trimer in a protein secondary structure. The program distinguishes between coding sequence (CDS) and non-CDS and classifies a query sequence into a genus group out of 54 genera used as reference. The average correct prediction percentages of the CDS and the non-CDS test sets at the genus level were 71% and 50% for bacteria, 65% and 73% for fungi (excluding AMF), and 49% and 72% for AMF (*Rhizophagus irregularis*), respectively. The program provides a means for estimating not only taxonomic diversity and abundance but also the gene reservoir of the reference taxonomic groups associated with AMF. Therefore, the program enables users to study the symbiotic roles of associated microorganisms. SeSaMe can be applicable to other microorganisms as well as soil metagenomes. It is freely available at www.journal.com and at www.fungalsesame.org.

**KEYWORDS:** SeSaMe; Spore associated Symbiotic Microbes; Arbuscular mycorrhizal fungi; Taxonomic classification; Three Codon DNA 9-mer

## 2.2 Introduction

Arbuscular mycorrhizal fungi (AMF) are plant-root inhabiting fungi, of the subphylum Glomeromycotina, which form symbioses with more than 80% of vascular plants worldwide [1]. They supply plants with essential nutrients particularly phosphorus and nitrogen, protect them against soil borne pathogens, and alleviate their abiotic stresses [1–3]. Therefore, AMF based inoculants have been applied in agriculture as a biofertilizer and in phytoremediation for cleaning up contaminated soil [2,4–7]. Despite the ecological, agricultural, and environmental importance of AMF, their genetics is poorly understood due to their complex genome organization. They form coenocytic hyphae, reproduce through multinucleated asexual spores, and are strict symbionts [8]. Furthermore, it is suggested that AMF are heterokaryons, although this is under debate [9]. In addition, numerous studies reported that bacteria and fungi inhabit the surface and the interior of mycelia and spores [10–14]. In 2012 and 2013, Tisserant et al. published the transcriptome and the genome of the AMF *Rhizophagus irregularis (R. irregularis)* cultivated *in-vitro* [15,16]. However, only a few AMF taxa are able to grow in axenic *in-vitro* systems with transformed roots as a host. Thus, whole genome sequencing (WGS) data from AMF spore DNA originating from *in-vivo* cultures (conventional cultivation method in a pot culture with a host plant), contain a substantial number of non-AMF DNA sequences, but do provide important information on the microbial communities associated with AMF. In contrast, WGS data from *in-vitro* petri-dishes contain fewer non-AMF sequences, because antibiotics are used to initiate axenic cultures [17].

Taxonomic classification of WGS obtained from AMF cultivated *in-vivo* using current bioinformatics approaches is challenging because these data represent a complex metagenome containing sequences of prokaryotic and eukaryotic microorganisms. Two major approaches for taxonomic classification of random whole metagenome sequencing data (e.g., whole metagenome shotgun sequencing data) include composition-based methods and similarity-based search methods [18,19]. The latter ones include BLAST and its sister programs that are adequate for inferring functions of a query sequence [19,20]. Nevertheless, they have limitations in taxonomic classification, because they calculate scores based on a 20 by 20 matrix containing the overall rates of the 20 amino acid substitutions created from the most conserved regions of proteins. The same matrix is applied to all types of query sequences, irrespective of functions, structures, and taxonomic group. However, due to a lack of bioinformatics tools for analyzing random whole metagenome data, similarity-based search methods have been commonly used for taxonomic classification. In addition to similarity-based methods, taxonomic classification pipelines, for analyzing targeted metagenome sequencing data (e.g., 16S rRNA gene-based metagenome sequencing data), have been widely used for analyzing random whole metagenome sequencing data in combination with homology search program. Numerous repository databases and pipelines have been developed based on the 16S rRNA gene. However, recent studies have reported horizontal gene transfer of 16S rRNA genes in prokaryotic organisms and multiple heterogeneous rRNA genes within a single prokaryotic cell [28]. Therefore, they may cause misrepresentation of data if they are not properly dealt with, which may result in erroneous taxonomic classification.

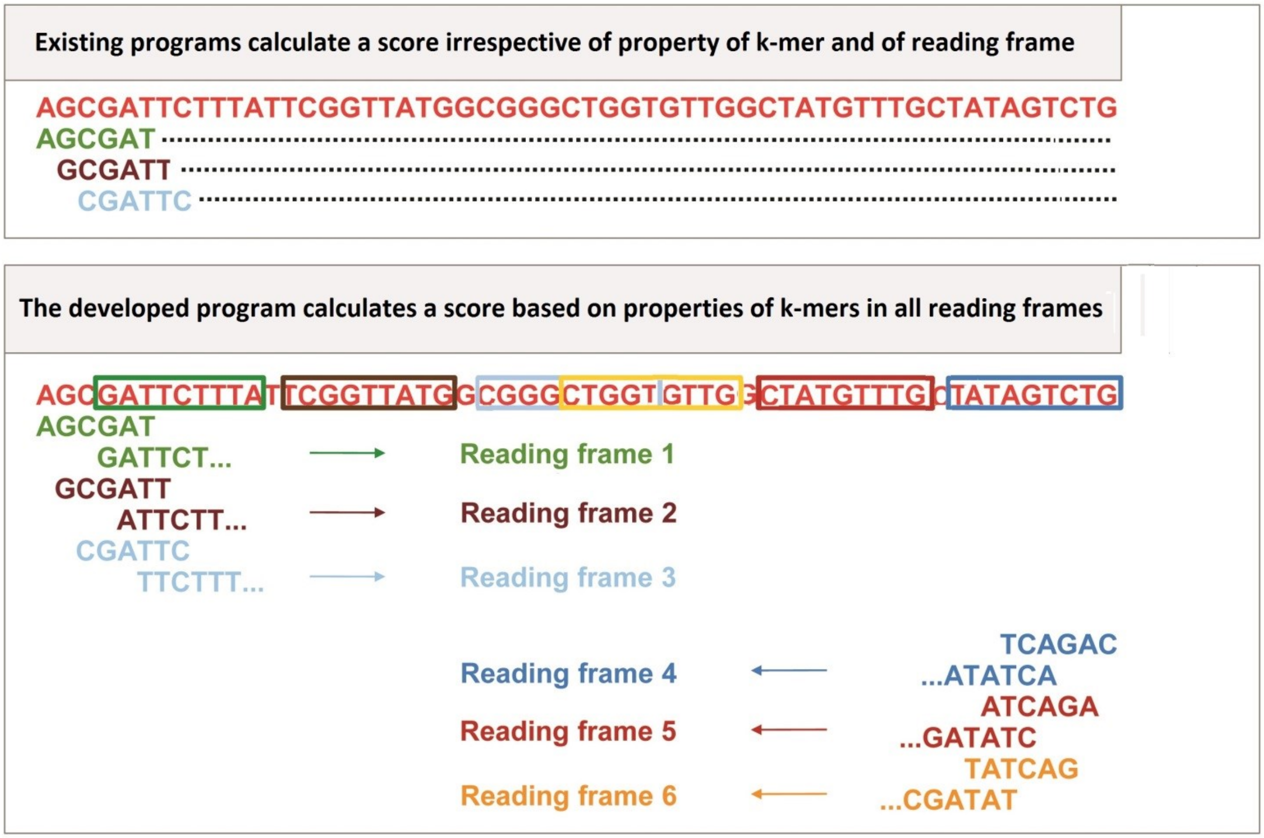

Existing programs calculate a score irrespective of property of k-mer and of reading frame
AGCGATTCTTTATTCGGTTATGGCGGGCTGGTGTTGGCTATGTTTGCTATAGTCTG
AGCGAT
GCGATT
CGATTC
The developed program calculates a score based on properties of k-mers in all reading frames
AGCGATTCTTTATTCGGTTATGGCGGGCTGGTGTTGGCTATGTTTGCTATAGTCTG
AGCGAT
GATTCT...
GCGATT
ATTCTT...
CGATTC
TTCTTT...
Reading frame 1
Reading frame 2
Reading frame 3
TCAGAC
...ATATCA
ATCAGA
...GATATC
TATCAG
...CGATAT
Reading frame 4
Reading frame 5
Reading frame 6

**Figure 1 Unique advantage of SeSaMe over existing programs**

Existing programs calculate a score based on the frequencies of k-mers identified in a query sequence irrespective of properties of the k-mers, or its reading frame. In contrast, SeSaMe identifies k-mers that encode for the amino acids of protein secondary structures in each reading frame. In the figure, matching *Three Codon DNA 9-mers* of the Trimer Ref. DB are marked with a rectangle, where the rectangle's color indicates its reading frame. The program calculates scores based on the *three codon usages* and the *A.A. Trimer usages* of the matching *Three Codon DNA 9-mers* in each reading frame. It classifies a query sequence into a taxonomic group based on the six scores computed from all reading frames.

*Note*: All sequences in this figure are randomly generated for illustration purposes only.

Composition-based methods utilize unique sequence properties such as codon usage bias, compositional patterns in nucleotide sequences (k-mers), and GC content that have been widely used for studying microbial genome evolution in areas of bioinformatics [18,21–25]. K-mers are subsequences of length k in a DNA sequence (e.g., tetramer or 4-mer: ATGT). Composition-based methods using k-mers have been employed in bioinformatics programs for taxonomic classification of random whole metagenome data [26].

They have a number of advantages over similarity-based search methods. It is estimated that more than 99% of existing microorganisms cannot be cultured in laboratory conditions [27] and microbial sequences available in bioinformatics databases represent only a tiny fraction of the diversity of existing microorganisms. Therefore, composition-based methods, that do not require sequence alignments but make predictions based on a microorganism's unique sequence signatures, supposedly excel in taxonomical classification of novel sequences. However, existing bioinformatics programs based on composition-based methods are designed for prokaryotic organisms and their utilization in fungi is inefficient. In this article, we introduce a novel bioinformatics program for random whole metagenome sequence classification, SeSaMe (Spore associated Symbiotic Microbes). It provides a means for estimating taxonomic diversity and abundance, as well as, the reservoir of genes of reference taxonomic groups in AMF metagenome. It therefore enables users to study symbiotic roles of taxonomic groups associated with AMF. In order to filter complex evolutionary signals and obtain comparable evolutionary footprints, we calculated codon usage bias based on the amino acid usage and the codon usage of three codon DNA 9-mer that encodes for three consecutive amino acids located in protein secondary structure. We joined three consecutive codons into one unit, and calculated the unit's relative frequency among synonymous three codon DNA 9-mers, which will be hereafter referred to as *three codon usage*. *Three codon usage* has higher resolution than mono codon usage in assessing the differences among taxonomic groups because evolutionary forces acting on a codon and its encoded amino acid vary widely across protein secondary structures as well as across taxonomic groups. For example, the evolutionary forces acting on the codon AAA, encoding the amino acid Lysine (K) in TGG***AAA***GTG (WKV), will have been different from the evolutionary forces acting on the codon AAA in GAC***AAA***GAA (DKE). We found that *three codon usage* of a three codon DNA 9-mer belonging to protein secondary structure is a taxonomically unique sequence property. SeSaMe calculates a score based on six sets of three codon DNA 9-mers from all reading frames (**Figure 1**), and distinguishes between coding sequence (CDS) and non-CDS. It has an advantage over existing composition-based methods that do not identify nucleotide subsequences with structural roles, or do not consider the biological importance of codon and reading frame. SeSaMe is freely available at www.journal.com and at www.fungalsesame.org.

## 2.3 Methods

### 2.3.1 Bacterial and fungal sequence databases

We selected bacterial genera that were dominant in soil based on a literature review [10,27,31–34]. While NCBI offered a broad selection of more than 2,300 completely sequenced bacterial genomes, we did not have many choices for the majority of fungal phyla. Most of the completely sequenced

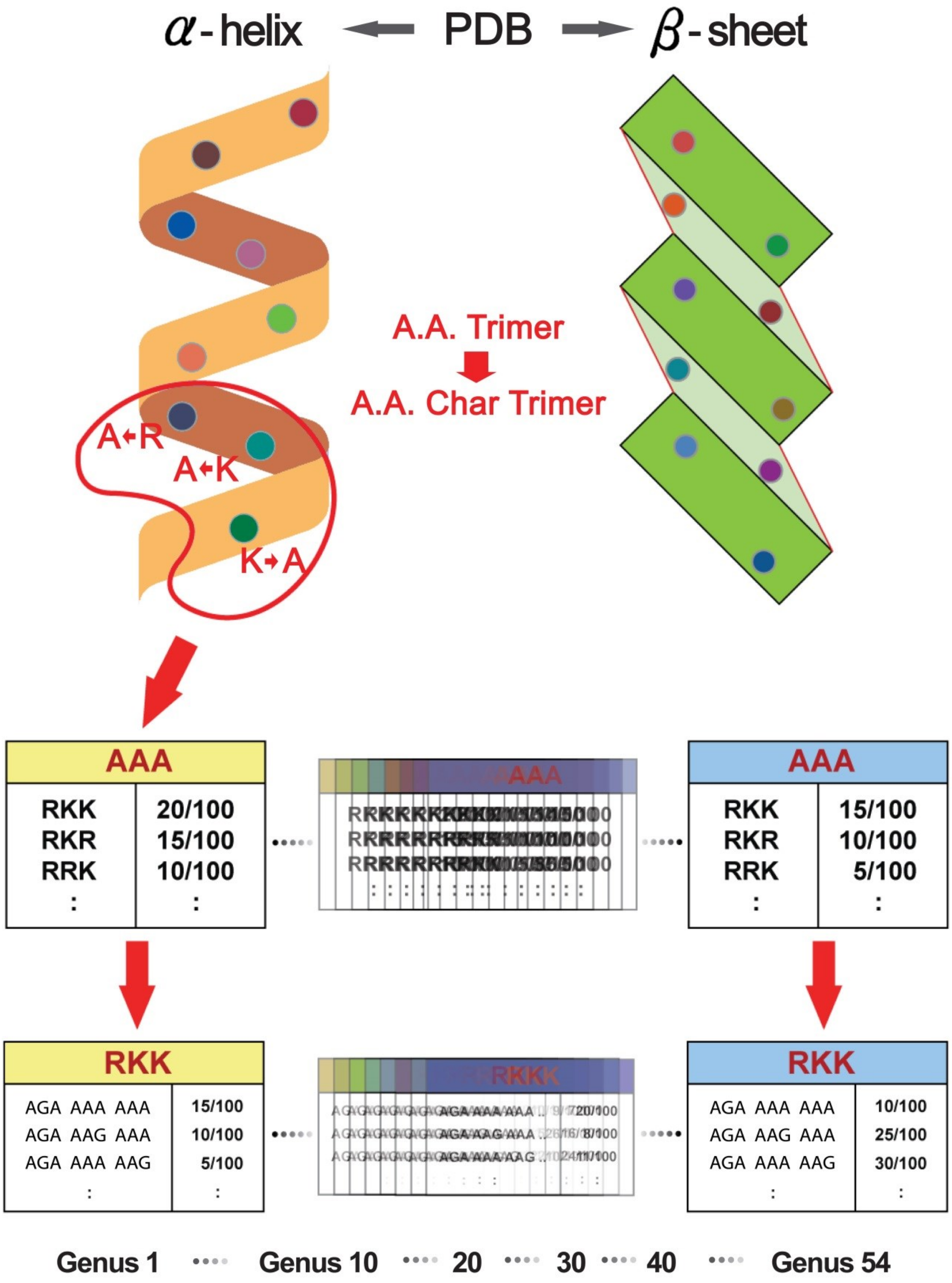


**Figure 2 Database design**

In this figure, A.A. Trimer Usage Table consists of the *A.A. Trimer usages* of the multiple members- RKK, RKR, and RRK belonging to the same *A.A. Char Trimer*- AAA. Three Codon Usage Table consists of the *three codon usages* of the synonymous *Three Codon DNA 9-mers* encoding the *A.A. Trimer*- RKK (e.g., AGA AAA AAA). The *trimer usage bias* of AGA AAA AAA is the multiplication of the *A.A. Trimer usage* of RKK and the *three codon usage* of AGA AAA AAA.

*Note*: All sequences and usage information in this figure are not real, but randomly chosen for illustration purposes only.

fungal genomes in NCBI or JGI were Dikarya, while we needed diverse fungal genomes covering Mucoromycotina, AMF, Blastocladiomycota, Neocallimastigomycota, Microsporidia, and Chytridiomycota. We assigned the completely sequenced genomes of 444 bacteria and of 11 fungi, including *R. irregularis*, to 45 bacterial and 9 fungal genera respectively, and created CDS and non-CDS databases per genus based on CDS lists provided by NCBI, JGI, and Tisserant et al. [16]. The number of genomes per genus varied from 1 to 81, depending on their availability in public databases. The total number of the bacterial, and the fungal, genes and introns, per genus, are shown in Tables S1 and S2. Sequences with an ambiguous nucleotide or with a length shorter than nine—the minimum length of nucleotides required for three codon DNA 9-mers—were excluded. *Cryptococcus* and Agaricomycetes (*Phanerochaete*, *Scleroderma, Sebacina*) belong to the same subdivision, Agaricomycotina, and were grouped together in order to simplify the analysis.

**Table 1 Conversion table from A.A. and stop codon to *A.A. Char***

| ***A.A. Char*** | **A.A.** | **Properties** | ***A.A. Char*** | **A.A.** | **Properties** |
|---|---|---|---|---|---|
| A | K,R | Positively charged | G | G | Special |
| B | H | Special | H | P | Special |
| C | D,E | Negatively charged | I | M | Special |
| D | S,T | Polar uncharged smaller volume | J | A,I,L,V | Hydrophobic smaller volume |
| E | N,Q | Polar uncharged larger volume | K | F,W,Y | Hydrophobic larger volume |
| F | C | Special | L | * | Stop codon |

*Note*: Amino acids were grouped according to their side chain's pKa values and charges at physiological pH (7.4) and their volumes

### 2.3.2 Database design

In selecting a parameter k of k-mer, we chose three codon DNA 9-mer as the length of amino acid and of nucleotide, considering the approximate number of amino acids required to form a turn in helix and a beta-strand. The program consists of two main components- databases and scoring methods. The major distinguishing feature is the trimer reference sequence database (Trimer Ref. DB). 126,093 Protein Data Bank (PDB) entry files were processed with in-house developed parsing programs in order to extract 7,674 amino acid trimers, subunits of protein secondary structures, that were assigned to the sequence variable- *A.A. Trimer* [30]. 224,383 *three codon DNA 9-mer*s, encoding 7,674 *A.A. Trimers*, were assigned to the sequence variable- *Three Codon DNA 9-mer*. In Trimer Ref. DB, the sequence variables- *A.A. Char Trimer*, *A.A. Trimer*, and *Three Codon DNA 9-mer*- form a three level hierarchy where *A.A. Char Trimer* is the highest level (**Figure 2**). To create amino acid characteristic (*A.A. Char*), first, we assigned amino acids with similar properties into one group according to polarity and charge of their side chain, and secondly subdivided each group according to their volume (**Table 1**). Cysteine, Glycine, Histidine, Methionine, and Proline have special properties; Cysteine forms disulfide bonds, Glycine is the simplest amino acid, Histidine can be a proton shuttle, Methionine is often the first amino acid, Proline is an imino acid. Therefore, each of them was assigned as a sole member of *A.A. Char* group. Generally, multiple *A.A. Trimers* with similar properties belong to one *A.A. Char Trimer*. An *A.A. Char Trimer* and an *A.A. Trimer* have A.A. Trimer table and Three Codon DNA 9-mer table containing multiple members, respectively (Figure 2).

Genus-specific usage bias database (Genus Specific DB) contains the main numerical variable, *trimer usage bias*. *Trimer usage bias* represents a *three codon usage bias* of *Three Codon DNA 9-mer*, and is calculated by multiplying the *A.A. Trimer usage* of *A.A. Trimer* by the *three codon usage* of *Three Codon DNA 9-mer* in Trimer Ref. DB (Figure 2). There are 54 CDS Genus Specific DBs and the same number of non-CDS Genus Specific DBs in the program. Each CDS Genus Specific DB contains 1,296 A.A. Trimer Usage Tables and 7,674 Three Codon Usage Tables created based on the CDS database. Each non-CDS Genus Specific DB contains the same number of tables created based on the non-CDS database with the same sequence compositions as those in the CDS Genus Specific DB. We decided to accept inaccuracy in calculating the information frequency in the case of non-CDS in exchange for cost effective CDS and non-CDS classification. Because SeSaMe only needs to compare frequency information of 54 genera calculated based on the same standard genetic code table for the same *Three Codon DNA 9-mers* of a query sequence, inaccuracy in non-CDS is assumed to be insignificant.

### 2.3.3 Scoring methods

We developed two scoring methods, each equipped with a *P* value scoring method. The trimer usage probability scoring method classifies a query sequence into one out of 54 genus references, while the rank probability scoring method classifies a query sequence into one out of 13 taxon groups: Clostridia, Bacilli, Oscillatoriophycideae, Nostocales, Acidobacteriales, Betaproteobacteria, Deltaproteobacteria, Gammaproteobacteria, Alphaproteobacteria, Actinobacteria, AMF (*R. irregularis*), Agaricomycotina, and Pezizomycotina. To avoid repetition, these taxonomic groups will be hereafter referred to as 13 taxon groups, and represented in the same order. We provide users with two different programs, one with the trimer usage probability scoring method and the other with the rank probability scoring method.

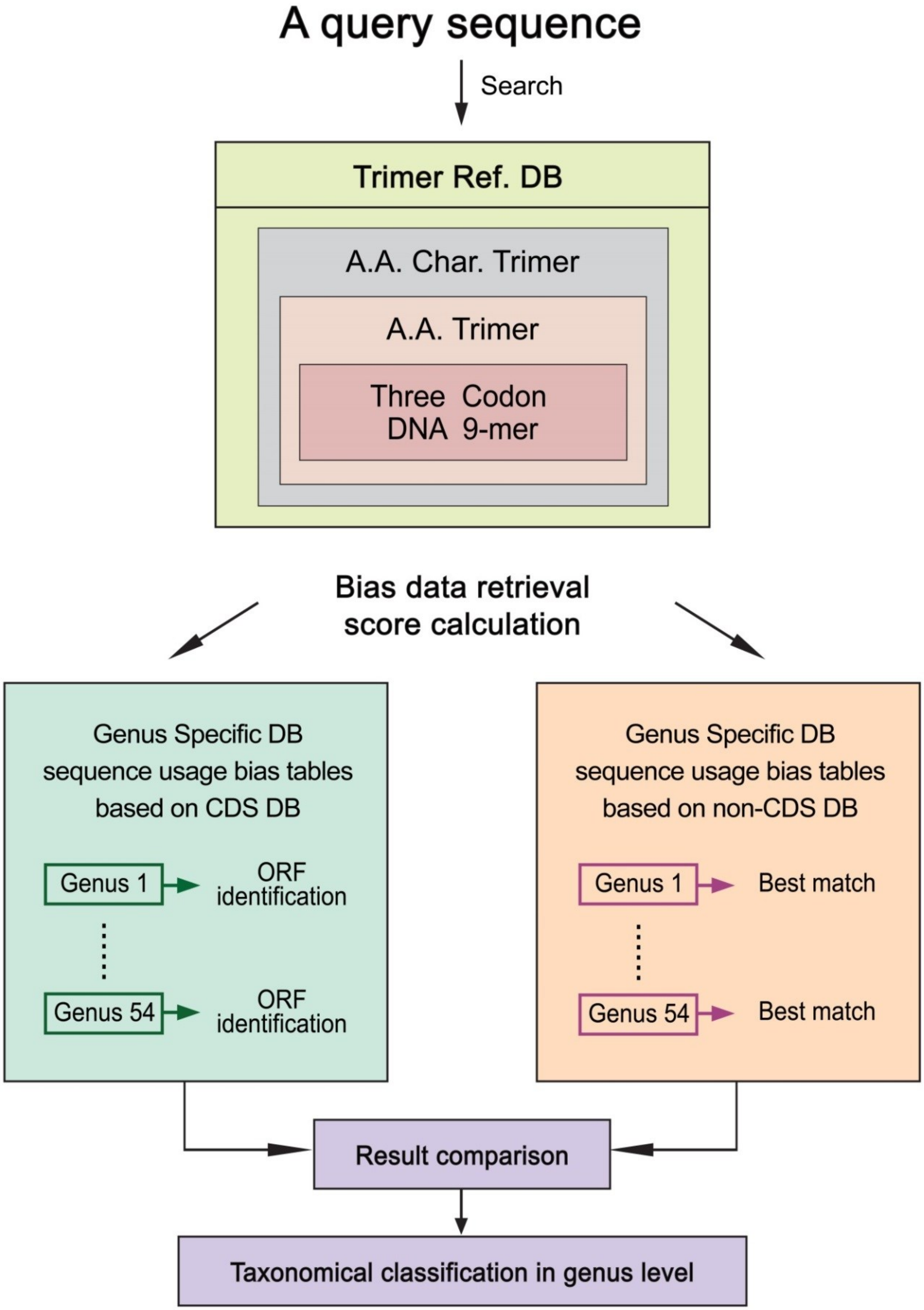


**Figure 3 Flow chart of the program**

#### 2.3.3.1 Trimer usage probability scoring method

This method converts three codon DNA 9-mers in a query sequence into A.A. Char trimers and identifies those with structural roles by searching them against Trimer Ref. DB. For each matching *A.A. Char Trimer*, the method first searches the matching *A.A. Trimer*, and second, the matching *Three Codon DNA 9-mer* in Trimer Ref. DB (**Figure 3**). It retrieves *trimer usage biases* of the matching *Three Codon DNA 9-mers* from CDS Genus Specific DB per genus. It repeats the process in each of 6 reading frames (3 forward reading and 3 reverse reading frames) of a query sequence. It repeats the same process with non-CDS Genus Specific DBs, calculating a trimer usage probability score per genus. It then compares the highest scores from CDS and non-CDS Genus Specific DBs, and selects a genus with the highest score (Figure 3). Users are provided with an option to include genera whose scores have little difference from the highest score calculated.

#### 2.3.3.2 Rank probability scoring method

This method measures a standardized three codon usage relative to an expected three codon usage as computed from three individual mono codon usages. The Average A.A. Usage Table (20 amino acids and stop codons for 12 A.A. Char monomers) and the Average Codon Usage Table (64 codons for 20 amino acid monomers and stop codons) were created based on CDS database per genus. 1,296 Expected A.A. Trimer Usage Tables with the same sequence compositions as the A.A. Trimer Usage Tables were created based on the Average Amino Acid Usage Table. 7,674 Expected Three Codon Usage Tables with the same sequence compositions as the Three Codon Usage Tables were created based on the Average Codon Usage Table (**Figure 4**).

A standardized three codon usage was calculated by dividing a *three codon usage* in a Three Codon Usage Table by an *expected three codon usage* in an Expected Three Codon Usage Table. Based on *trimer usage biases* and standardized three codon usages, we calculated a group mean for each taxon group and Kruskal Wallis (KW) test's h-score on ranks of 13 taxon groups, from which we developed a rank probability score per *Three Codon DNA 9-mer*. The new genus specific score database contains the same number of the rank probability scores as Genus Specific DB, 224,383 scores per genus. Per reading frame of a query sequence, the program retrieves scores for all of the matching *Three Codon DNA 9-mers* from the new database and multiplies the scores to produce a rank probability score per genus. It repeats the process for each of 6 reading frames and classifies a query sequence into one of 13 taxon groups. This method is applicable only to CDS.

#### 2.3.3.3 P value scoring method

We applied the concept of the sum of rolled numbers from a pair of dice to develop the *P* value scoring method (http://www.lucamoroni.it/the-dice-roll-sum-problem/). We drew analogies between the number of faces of a dice and 54 genera and between the number of dices we roll and the number of matching *Three Codon DNA 9-mers* identified in a reading frame of a query sequence. There were 54 possible ranks computed based on *trimer usage biases* per matching *Three Codon DNA 9-mer*. *P* value scores were calculated based on a sum of ranks of matching *Three Codon DNA 9-mers*. Computational costs of *P* values for all possible outcomes, sums of ranks, were too high, however, so to reduce the computational costs we approximated *P* values. We obtained sample data per number of matching *Three Codon DNA 9-mers* based on equation 1.

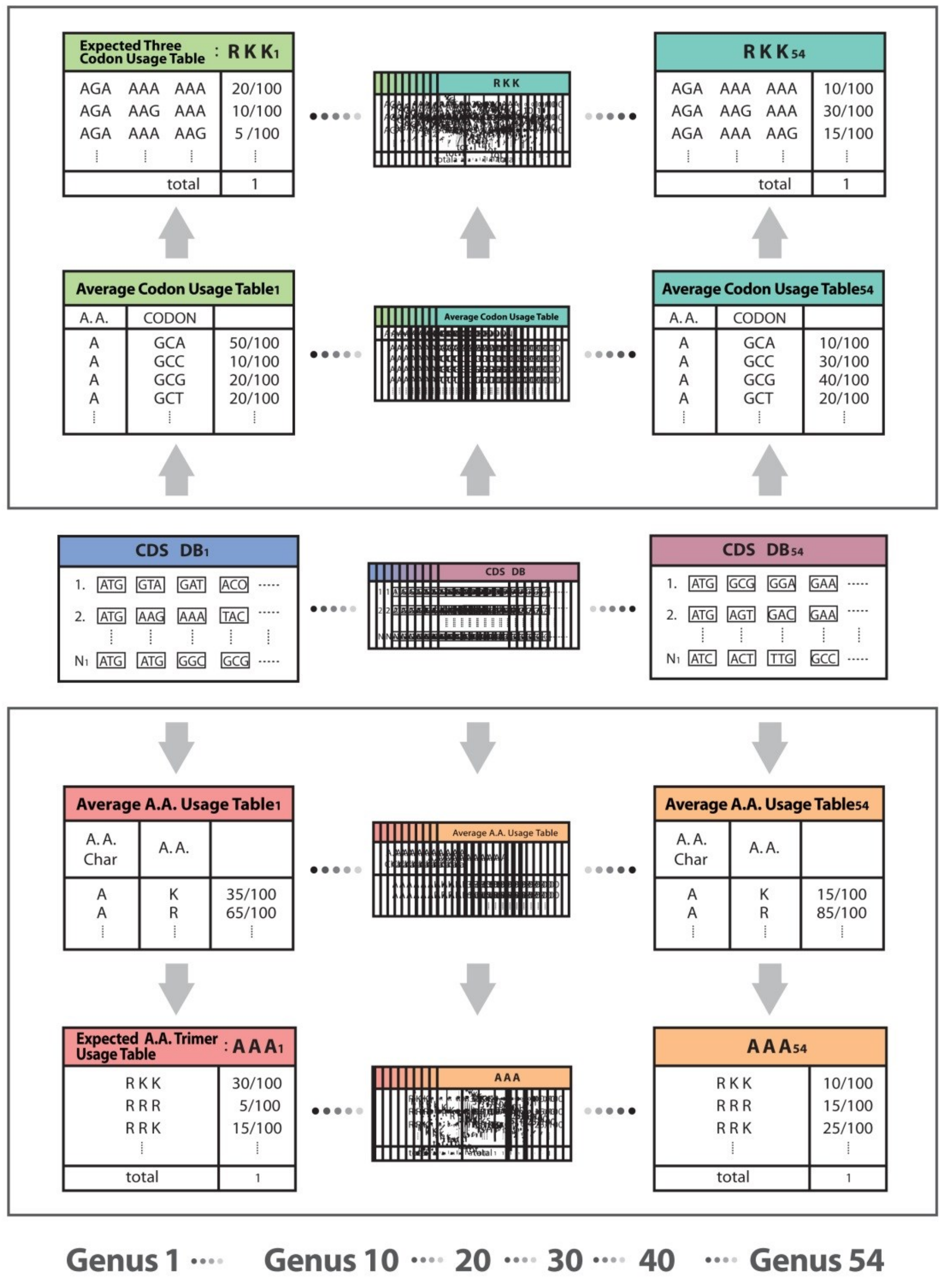


**Figure 4 Creation of expected usage tables for the rank probability scoring method**

Average A.A. Usage Table and Average Codon Usage Table were calculated from the CDS database per genus. Expected A.A. Trimer Usage Tables and Expected Three Codon Usage Tables were created based on the Average A.A. Usage Table and the Average Codon Usage Table, respectively

*Note*: All sequences and expected usage information in this figure are not real, but randomly chosen for illustration purposes only.

Equation. 1:

$$P(p,n,s) \quad = \frac{1}{(s)^n} \sum_{k=0}^{](p-n)/s[} (-1)^k \binom{n}{k} \binom{p-sk-1}{n-1}$$

where p is the sum of ranks, n is the number of dices per roll, s is the number of faces of the dice, 54, and the range of k is between 0 and ](p-n)/s[ where ]x[ is the floor function (e.g., ]7.9[ = 7). We created a table of *P* value scores per number of matching *Three Codon DNA 9-mers*. If a rank sum was less than one with the highest *P* value score, the approximate mean of all of the rank sums in each table, we multiplied the *P* value score with -1, indicating statistically non-significant outcome. In the test sets, the number of matching *Three Codon DNA 9-mers* varied widely, with a minimum of 30 and a maximum of 97. We have 624 tables in the *P* value score database covering 2 – 625 matching *Three Codon DNA 9-mers*. *P* value scores are generated per genus in both the trimer usage probability, and the rank probability, scoring methods to provide users with the statistical significance of predicted outcomes.

### 2.3.4 Implementation and program availability

SeSaMe has been implemented using the Java programming language (www.java.net, www.oracle.com (Java 8)). We have provided two sets of the programs; one requires Apache commons math3 (3.3) and IO (2.4) libraries (www.apache.org), while the other does not. The programs consist of executable Java JAR files and Java class files for Linux/ Unix operating systems. SeSaMe has been tested and confirmed to work on Linux system- CentOS Linux 7 (www.centos.org) and is currently being used at the Biodiversity Center, Institut de Recherche en Biologie Végétale, Département de Sciences Biologiques, Université de Montréal. The trimer usage probability scoring method offered to the public produces output of smaller size, but is sufficient for the purpose of taxonomic classification and is freely available at www.fungalsesame.org. There are no restrictions to use the programs by academic, or non-academic, organizations as long as they comply with the terms and conditions of the license agreements.

### 2.3.5 Input, output, and options

SeSaMe utilizes a command-line interface. Input files should contain DNA sequence(s) in fasta format. The Java JAR files produce detailed output files with sequence information (seq_id, matching *A.A. Char Trimers*, *A.A. Trimers*, and *Three Codon DNA 9-mers*) and genus information (rank, scores, and *P* value score). The output details the information per reading frame per sequence. After processing the output file with Java class files, users are able to obtain a summary file containing one predicted outcome per query sequence. Java JAR files require users to give a mandatory command line argument- input file path. Java JAR files with the trimer probability scoring method may produce multiple genera as an answer if their scores have little differences. A user is given the option with 6 choices to select a cut-off value for the difference: 0.01, 0.05, 0.1, 0.15, 0.2, or 0.3. Users can give the option to the Java class file called compare_result_coding_non_coding.class. The default cut-off value is 0.05. The lower the cut-off value is, the fewer genera will be included in an answer.

**Table 2 Correct prediction percentages at the levels of genus and of higher taxonomic rank of the 13 taxon groups**

| Genus | CDS | | Non-CDS | | Genus | CDS | | Non-CDS | |
|---|---|---|---|---|---|---|---|---|---|
| | Correct genus % | Correct taxon group % | Correct genus % | Correct taxon group % | | Correct genus % | Correct taxon group % | Correct genus % | Correct taxon group % |
| *Acidithiobacillus* | 57 | 78 | 51 | 72 | *Microbacterium* | 87 | 96 | 60 | 89 |
| *Acidobacterium* | 62 | 62 | 40 | 40 | *Micrococcus* | 93 | 97 | 48 | 89 |
| *Agrobacterium* | 65 | 84 | 50 | 65 | *Myxococcus* | 88 | 89 | 27 | 39 |
| *Anabaena* | 41 | 57 | 56 | 78 | *Nitrobacter* | 66 | 90 | 42 | 71 |
| *Azorhizobium* | 87 | 97 | 49 | 80 | *Nitrosococcus* | 51 | 66 | 42 | 69 |
| *Azotobacter* | 75 | 87 | 55 | 71 | *Nitrosomonas* | 45 | 45 | 33 | 33 |
| *Bacillus* | 53 | 53 | 64 | 64 | *Nitrosospira* | 60 | 60 | 70 | 72 |
| *Bdellovibrio* | 61 | 64 | 63 | 66 | *Nocardia* | 79 | 89 | 25 | 44 |
| *Beijerinckia* | 65 | 83 | 56 | 66 | *Nostoc* | 58 | 60 | 62 | 68 |
| *Bradyrhizobium* | 84 | 88 | 41 | 61 | *Oscillatoria* | 58 | 58 | 66 | 66 |
| *Caulobacter* | 79 | 91 | 43 | 59 | *Pseudanabaena* | 76 | 77 | 48 | 53 |
| *Clostridium* | 81 | 85 | 90 | 92 | *Pseudomonas* | 77 | 95 | 52 | 64 |
| *Cyanobacterium* | 72 | 73 | 61 | 61 | *Pseudonocardia* | 88 | 96 | 29 | 74 |
| *Desulfotomaculum* | 49 | 54 | 43 | 69 | *Rhizobium* | 70 | 81 | 48 | 60 |
| *Desulfovibrio* | 43 | 52 | 52 | 55 | *Rhodobacter* | 85 | 94 | 32 | 66 |
| *Erwinia* | 71 | 87 | 47 | 81 | *Rickettsia* | 68 | 68 | 67 | 67 |
| *Frankia* | 72 | 90 | 18 | 50 | *Shewanella* | 75 | 83 | 83 | 86 |
| *Geobacter* | 61 | 67 | 47 | 54 | *Sinorhizobium* | 67 | 83 | 51 | 69 |
| *Klebsiella* | 79 | 95 | 59 | 77 | *Sphingomonas* | 76 | 92 | 31 | 71 |
| *Kocuria* | 88 | 97 | 52 | 89 | *Streptomyces* | 89 | 96 | 55 | 56 |
| *Leuconostoc* | 62 | 72 | 41 | 73 | *Variovorax* | 85 | 85 | 41 | 41 |
| *Mesorhizobium* | 70 | 90 | 40 | 58 | *Xanthomonas* | 91 | 94 | 48 | 60 |
| *Methylococcus* | 76 | 87 | 60 | 82 | **Total** | **3185** | **3587** | **2238** | **2970** |
| | | | | | **Mean (Bacteria)** | **71%** | **80%** | **50%** | **66%** |
| AMF | 49 | 49 | 72 | 72 | *Oidiodendron* | 68 | 68 | 71 | 71 |
| *Aspergillus* | 72 | 72 | 77 | 77 | *Phanerochaete* | 52 | 67 | 66 | 91 |
| *Cenococcum* | 54 | 66 | 88 | 90 | *Scleroderma* | 76 | 87 | 68 | 89 |
| *Cryptococcus* | 72 | 87 | 78 | 89 | *Sebacina* | 58 | 89 | 66 | 90 |
| *Mycosphaerella* | 88 | 93 | 69 | 81 | **Total** | **589** | **678** | **655** | **750** |
| | | | | | **Mean (Fungi)** | **65%** | **75%** | **73%** | **83%** |

*Note*: After genus in an answer was converted to a corresponding taxonomic group in the 13 taxon groups, the mean of the correct prediction percentages (Correct taxon group %) was calculated.

### 2.3.6 Program evaluation

We assessed the accuracy of the classification program by conducting classification experiments. We created metagenome test sets, ran the programs with them, and calculated the correct prediction percentages. We showed the relationship between the correct prediction proportion and the *P* value score in order to provide users with useful examples in assessing the statistical significance of predicted outcomes.

#### 2.3.6.1 Metagenome test sets

We randomly chose 100 sequences from each of the CDS and non-CDS databases for each genus. We randomly selected a starting base pair position in each of the chosen sequences. From the starting position, we randomly selected an ending base pair position so that a sequence length is within the range of 150 ~ 300 bp. Both of the CDS and the non-CDS test sets consisted of 4,500 bacterial and 900 fungal sequences (including AMF).

#### 2.3.6.2 Correct prediction percentages from the trimer usage probability scoring method

The means of the correct prediction percentages of the CDS and the non-CDS test sets at the genus level were 71% and 50% for the bacterial group, 65% and 73% for the fungal group (excluding AMF), and 49% and 72% for AMF, respectively. AMF showed the lowest prediction percentage among the CDS genus test set possibly due to a large number of heterogeneous nuclei and horizontal gene transfers from a variety of endobacteria during their evolution [8,10–14]. The means of correct prediction percentages at the genus level and at higher taxonomic ranks of the 13 taxon groups are shown in **Table 2**.

SeSaMe produced more than one genus as an answer per query sequence when multiple genera had little differences in their scores. We converted each predicted genus into one of the 13 taxon groups and calculated a proportion of the correct taxon group in answer per query sequence. We calculated the mean and the standard deviation of the proportions in each genus test set; 1 represented that answers contained correct taxon groups only, while 0 represented that answers contained incorrect taxon groups only (Tables S3, S4). The mean was 0.9 for the bacterial CDS test set, which indicated that in average 90% of the taxon groups in an answer were correct.

SeSaMe produced only one genus as an answer in 60% and 46% of correctly predicted sequences from the bacterial CDS and non-CDS test sets, respectively (Figure S1, Table S5). A correct taxon group occurred in the first rank in 90% and 76% of correctly predicted sequences in the bacterial CDS and non-CDS test sets, respectively (Figure S2, Table S6). Only 1% ~ 5% of the sequences in the bacterial and the fungal test sets had AMF in an answer (Table S7). Although the trimer usage probability scoring method provides us not with the individual *trimer usage biases* but with the result of multiplying all of the *trimer usage biases* identified in a query sequence, we can often derive general ideas about the query sequence from its answer. Does it contain only one genus in the answer? Or what other genera does it contain in answer? (Figures S3, S4) For example, an AMF test sequence that contains *Clostridium* and AMF in the answer may imply that the query sequence may have been acquired by horizontal gene transfer from a bacterium ancestor to an AMF ancestor during evolution.

#### 2.3.6.3 Correct prediction percentages from the rank probability scoring method

The mean of the correct prediction percentages of the CDS test set was 82% for the bacterial group, 72% for the fungal group (excluding AMF), and 42% for AMF. The mean and the standard deviation of the correct prediction percentages of the CDS test set were 64% ± 4.2%, 71% ± 6.4%, 84% ± 2.5%, 70% ± 2.8%, 73% ± 0%, 83% ± 8%, 74% ± 10%, 81% ± 7.8%, 88% ± 9.2%, 85% ± 5.9%, 42% ± 0%, 65% ± 6.4%, and 79% ± 6.7% for the 13 taxon groups, respectively. Compared to the trimer usage probability scoring method, the rank probability scoring method produced the higher mean and the smaller standard deviation, 82% ± 9.4% for the bacterial group. In general, the rank probability scoring method showed improvement in performance. Although the means for Clostridia and Gammaproteobacteria were lower, their standard deviations were much smaller in the rank probability scoring method than the trimer usage probability scoring method: 4.2% vs 22% and 7.8% vs 9.4%, respectively. The trimer usage probability scoring method showed better performance in Actinobacteria that had low within-group variation of *trimer usage bias*. In contrast, the rank probability scoring method showed better performance in genera that had relatively flat peakness in a frequency distribution curve of synonymous *Three Codon DNA 9-mers*, in addition to genera that had relatively large within-group variation of *trimer usage bias*.

#### 2.3.6.4 Relationship between correct prediction proportion and P value score

The mean of the correct prediction proportions per number of matching *Three Codon DNA 9-mers* calculated based on the result of the trimer usage probability scoring method is shown in Figure S5A and Table S8. The means of correct prediction proportions per base 10 logarithm of an approximated inverse of a rank sum based *P* value score ($\log_{10}$ (inverse of *P* value score)) calculated based on result of the trimer usage probability scoring method and of the rank probability scoring method are shown in Figure S5B and Table S9 and Figure S5C and Table S10, respectively. We divided the results of each genus test set into quartiles and calculated the range of ($\log_{10}$ (inverse of *P* value score)), the mean and the standard deviation of the correct prediction proportions in each quartile. They are shown in Tables S11 and S12 for the trimer usage probability scoring method and the rank probability scoring method, respectively. The first ranked genus with the highest probability score that was selected as an answer of a test sequence always had positive *P* value score. In general, as ($\log_{10}$ (inverse of *P* value score)) became higher—i.e., as positive *P* value score became lower—the correct prediction proportion increased in all test sets. The frequencies of fungal sequences that had a correct taxon group in the 1st, 2nd, 3rd, 4th, or 5th rank, in an answer were comparable due to similarity of Dikarya (Figure S2, Table S6). Because the data for Figure S5B were generated based only on the

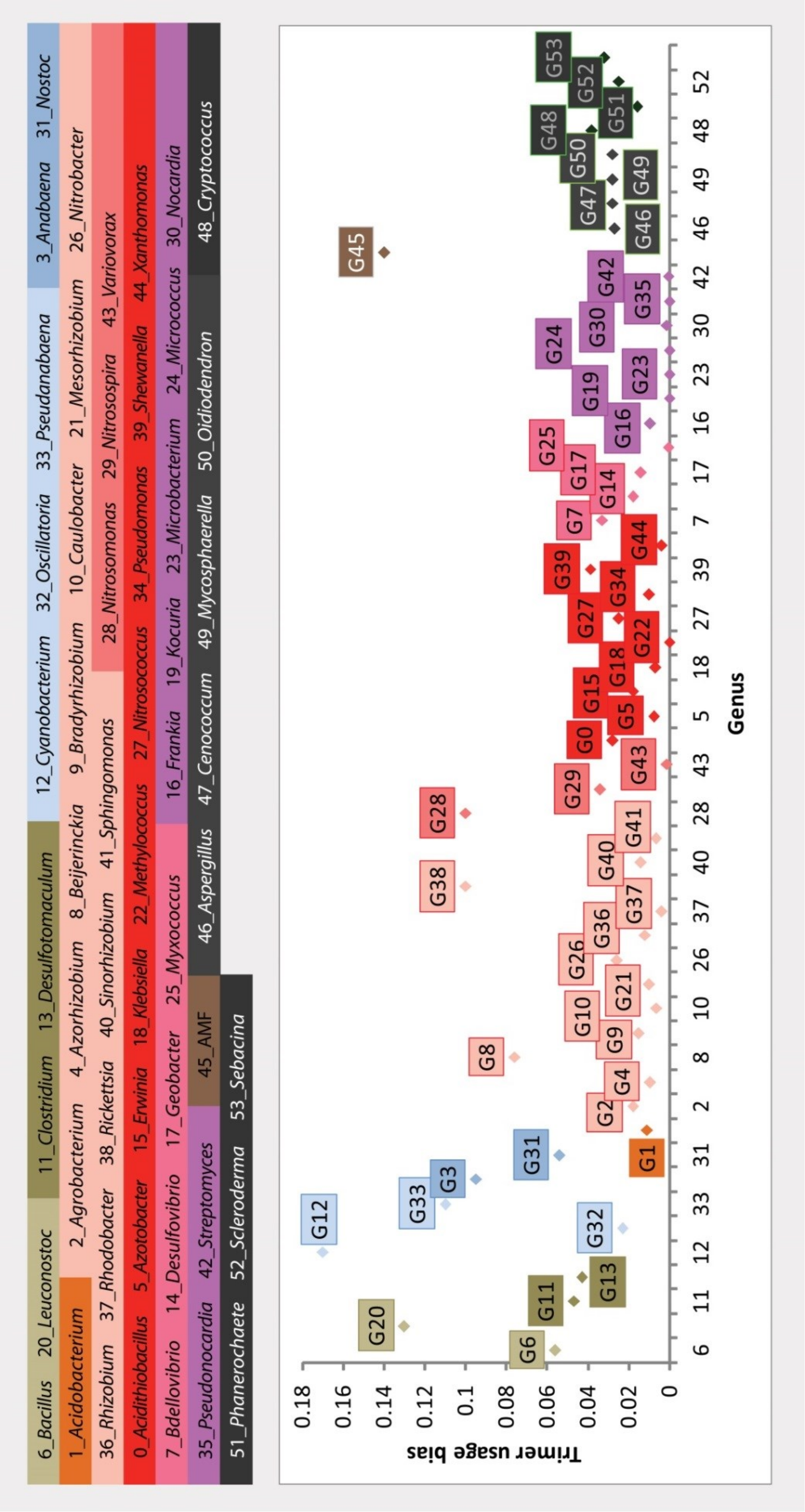


**Figure 5** ***Trimer usage biases*** **of 11th** ***Three Codon DNA 9-mer*****- GATGATCAT in 54 genera**

*Note*: Genera belonging to the same taxonomic group are indicated by the same background color.

first rank, the fungi showed relatively weak correlation between correct prediction proportion and ($\log_{10}$ (inverse of *P* value score)). The AMF database contains only one species, *R. irregularis*, therefore, results from both methods showed little difference.

### 2.3.6.5 Classification of an example sequence

Here we demonstrate the analysis of a query sequence selected from the AMF CDS test set. The example sequence was 156 bp (AAATCCCAATGTCAGAATAAAGAAACTACCAGATGATCATCCTGTTTATCCTGGGTATGGATTATTTGCTAACAAAGATCTTAAAAAATTTAATCTAGTCGTTTGTTATACTGGCAAAGTTACAAAAAGAGAAATTGGGGGTGAAGAAGGAAGTGA). The sequence had the highest trimer usage probability score in the second reading frame translation, which was then assumed as the open reading frame. SeSaMe identified 49 matching *Three Codon DNA 9-mers* in the second reading frame that were matched to Trimer Ref DB. The program correctly classified the example sequence into CDS of AMF. Firmicutes, Cyanobacteria, *Rickettsia*, and AMF had higher *trimer usage biases* than Proteobacteria, Actinobacteria, and Dikarya in a majority of *Three Codon DNA 9-mers*. **Figure 5** shows *trimer usage biases* of the 11th *Three Codon DNA 9-mer*- GATGATCAT in 54 genera. GATGATCAT belongs to *A.A. Char trimer*- CCB and to *A.A. Trimer*- DDH. The multidimensional scaling (MDS) method was applied to a matrix containing *trimer usage biases*; it had 54 genera in rows and matching *Three Codon DNA 9-mers* identified in the open reading frame in columns (http://www.inf.uni-konstanz.de/algo/software/mdsj/) [35]. It visualized proximity relationships among 54 genera in XY axis graph (www.jfree.org). It showed that Actinobacteria, Alphaproteobacteria, and Dikarya were compactly clustered, while Betaproteobacteria were spread out in the left side of the graph (Figure S6). Nostocales, Oscillatoriophycideae, Bacilli, and Clostridia were scattered across in the right side. AMF, *Cyanobacterium*, and *Rickettsia* were located in the far-right side.

## 2.4 Future Work

Microorganisms contain a number of heterogeneous alternative sigma factors that are selectively induced in response to environmental stress [36]. They not only provide functionally specialized RNA polymerase subpopulations, but are also involved in regulating the expression of a set of target genes, or regulon [37,38]. In contrast to sigma factors, regulatory systems governing heterogeneous alternative ribosome subpopulations in response to environmental stress remain largely unknown. Since multiple heterogeneous rRNA genes within a single isolate do not necessarily correlate with the extent of heterogeneity of functionally specialized ribosomes, sequence comparison of rRNA genes and ribosomal coding genes within a single isolate, as well as among closely related organisms, will be required in order to study their influence on adaptation of microorganism [39,40].

A codon is an attribute of a set of codes based on which transcriptional and translational regulators produce a gene product from a nucleotide sequence. Codon usage and codon context have been documented to play various important roles in these processes. If there are multiple types of the heterogeneous alternative regulators, there may be multiple sets of codes. *Trimer usage biases* of the Genus Specific DB were calculated based on the CDS database within a genus without considering alternative regulators and regulons. We may need *to* further anatomize evolutionary forces acting on multiple consecutive codons into greater detail, which may increase the accuracy of taxonomic classification. Moreover, comparative studies on alternative regulator subpopulations may provide useful insights into the development of genetic markers with which we can detect changes in microbial community structures in response to environmental stress (Figure S7). It may lead to new perspectives and strategies for improving the analysis of metagenome data, especially AMF inoculant field data sampled from highly stressful environments.

## 2.5 Authors' contributions

KJE designed the program and implemented it using the Java programming language. CA gave advice on developing scoring methods. HM provided knowledge on AMF experiments, the goals of the program, information on recent studies in AMF research, and helped to draft the manuscript. All authors read and approved the final manuscript.

## 2.6 Competing interests

The authors have declared no competing interests.

## 2.7 Acknowledgements

The authors gratefully acknowledge AFE (Éducation et de l'Enseignement supérieur Quebec), FESP (Faculté des études supérieures et postdoctorales de l'UdeM), and IRBV (Institut de Recherche en Biologie Végétale de l'Université de Montréal) for awarding scholarships to KJE. The authors gratefully acknowledge insightful comments from Bachir Iffis, David Walsh, Etienne Yergeau, Franck Stefani, Ivan de la Providencia, Jesse Shapiro, Sylvie Hamel, and Yves Terrat. We also thank Andrew Blakney for English editing.

## 2.8 References

[1] Roy-Bolduc A, Hijri M. The use of mycorrhizae to enhance phosphorus uptake: a way out the phosphorus crisis. J Biofertil Biopestic 2011; 2: 104.

[2] Hijri M. Analysis of a large dataset of mycorrhiza inoculation field trials on potato shows highly significant increases in yield. Mycorrhiza 2016; 26: 209–14.

[3] Zarik L, Meddich A, Hijri M, Hafidi M, Ouhammou A, Ouahmane L, et al. Use of arbuscular mycorrhizal fungi to improve the drought tolerance of *Cupressus atlantica G*. C R Biol 2016;

339: 185–96.

[4] Hassan SE, Bell T, Stefani FOP, Denis D, Hijri M, Yergeau E, et al. Contrasting the community structure of arbuscular mycorrhizal fungi from hydrocarbon-contaminated and uncontaminated soils following willow (*Salix spp. L.*) planting. PLoS One 2014; 9: e102838.

[5] Iffis B, St-Arnaud M, Hijri M. Bacteria associated with arbuscular mycorrhizal fungi within roots of plants growing in a soil highly contaminated with aliphatic and aromatic petroleum hydrocarbons. FEMS Microbiol Lett 2014; 358: 44–54.

[6] de la Providencia I, Stéfani FOP, Labridy M, St-Arnaud M, Hijri M. Arbuscular mycorrhizal fungal diversity associated with *Eleocharis obtusa* and *Panicum capillare* growing in an extreme petroleum hydrocarbon-polluted sedimentation basin. FEMS Microbiol Lett 2015; 362: fnv081.

[7] Chanda D, Sharma GD, Jha DK, Hijri M. Associations of arbuscular mycorrhizal (AM) fungi in the phytoremediation of trace metal (TM) contaminated soils. J Res Biol 2014; 4: 1247–63.

[8] Marleau J, Dalpe Y, St-Arnaud M, Hijri M. Spore development and nuclear inheritance in arbuscular mycorrhizal fungi. BMC Evol Biol 2011; 11: 51.

[9] Boon E, Halary S, Bapteste E, Hijri M. Studying genome heterogeneity within the arbuscular mycorrhizal fungal cytoplasm. Genome Biol Evol 2015; 7: 505–21.

[10] Hijri M, Redecker D, Petetot JAM-C, Voigt K, Wöstemeyer J, Sanders IR. Identification and isolation of two Ascomycete fungi from spores of the arbuscular mycorrhizal fungus. Appl Environ Microbiol 2002; 68: 4567–73.

[11] Cruz AF, Horii S, Ochiai S, Yasuda A, Ishii T. Isolation and analysis of bacteria associated with spores of *Gigaspora margarita*. J Appl Microbiol 2008; 104: 1711–7.

[12] Bonfante P. Plants, mycorrhizal fungi and endobacteria: a dialog among cells and genomes. Biol Bull 2003; 204: 215–20.

[13] Naito M, Morton JB, Pawlowska TE. Minimal genomes of mycoplasma-related endobacteria are plastic and contain host-derived genes for sustained life within Glomeromycota. Proc Natl Acad Sci U S A 2015; 112: 7791–6.

[14] Torres-Cortes G, Ghignone S, Bonfante P, SchuSsler A. Mosaic genome of endobacteria in arbuscular mycorrhizal fungi: transkingdom gene transfer in an ancient mycoplasma-fungus association. Proc Natl Acad Sci U S A 2015; 112: 7785–90.

[15] Tisserant E, Kohler A, Dozolme-Seddas P, Balestrini R, Benabdellah K, Colard A, et al. The transcriptome of the arbuscular mycorrhizal fungus *Glomus Intraradices* (DAOM 197198) reveals functional tradeoffs in an obligate symbiont. New Phytol 2012; 193: 755–69.

[16] Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, et al. Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. Proc Natl Acad Sci U S A 2013; 110: 20117–22.

[17] Bécard G, Fortin JA. Early events of vesicular–arbuscular mycorrhiza formation on Ri T-DNA transformed roots. New Phytol 1988; 108: 211–8.

[18] Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. Evolutionary implications of microbial

genome tetranucleotide frequency biases. Genome Res 2003; 13: 145–58.

[19] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215: 403–10

[20] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25: 3389–402.

[21] Akashi H. Synonymous codon usage in *Drosophila Melanogaster*: natural selection and translational accuracy. Genetics 1994; 136: 927–35.

[22] Gao F, Zhang CT. Comparison of various algorithms for recognizing short coding sequences of human genes. Bioinformatics 2004; 20: 673–81.

[23] Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 1981; 9: r43–74.

[24] Karlin S, Mrázek J, Campbell AM. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 1997; 179: 3899–913.

[25] Sueoka N. On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci U S A 1962; 48: 582–92.

[26] Kim M, Lee KH, Yoon SW, Kim BS, Chun J, and Yi H. Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era. Genomics Inform 2013; 11: 102–113.

[27] Jeffery S, Gardi C, Jones A, Montanarella L, Marmo L, Miko L, et al. European atlas of soil diodiversity. Luxembourg: Publications Office of the European Union; 2010.

[28] Rajendhran J, Gunasekaran P. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res 2011; 166: 99–110.

[29] Sharp P.M., Bailes E., Grocock R.J., Peden J.F., Sockett R.E. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 2005; 33: 1141–53.

[30] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000; 28: 235–42.

[31] Spain AM, Krumholz LR, Elshahed MS. Abundance, composition, diversity and novelty of soil Proteobacteria. ISME J 2009; 3: 992–1000.

[32] Bonfante P, Anca IA. Plants, mycorrhizal fungi, and bacteria: a network of interactions. Annu Rev Microbiol 2009; 63: 363–83.

[33] Lecomte J, St-Arnaud M, Hijri M. Isolation and identification of soil bacteria growing at the expense of arbuscular mycorrhizal fungi. FEMS Microbiol Lett 2011; 317: 43–51.

[34] Tedersoo L, Bahram M, Põlme S, Kõljalg U, Yorou NS, Wijesundera R, et al. Fungal biogeography. global diversity and geography of soil fungi. Science 2014; 346: 1256688.

[35] Algorithmics Group. MDSJ: Java library for multidimensional scaling (Version 0.2) [Internet]. Konstanz: University of Konstanz; 2009, http://www.inf.uni-konstanz.de/algo/software/mdsj/.

[36] Paget MS. Bacterial sigma factors and anti-sigma factors: structure, function and distribution. Biomolecules 2015; 5: 1245–65.

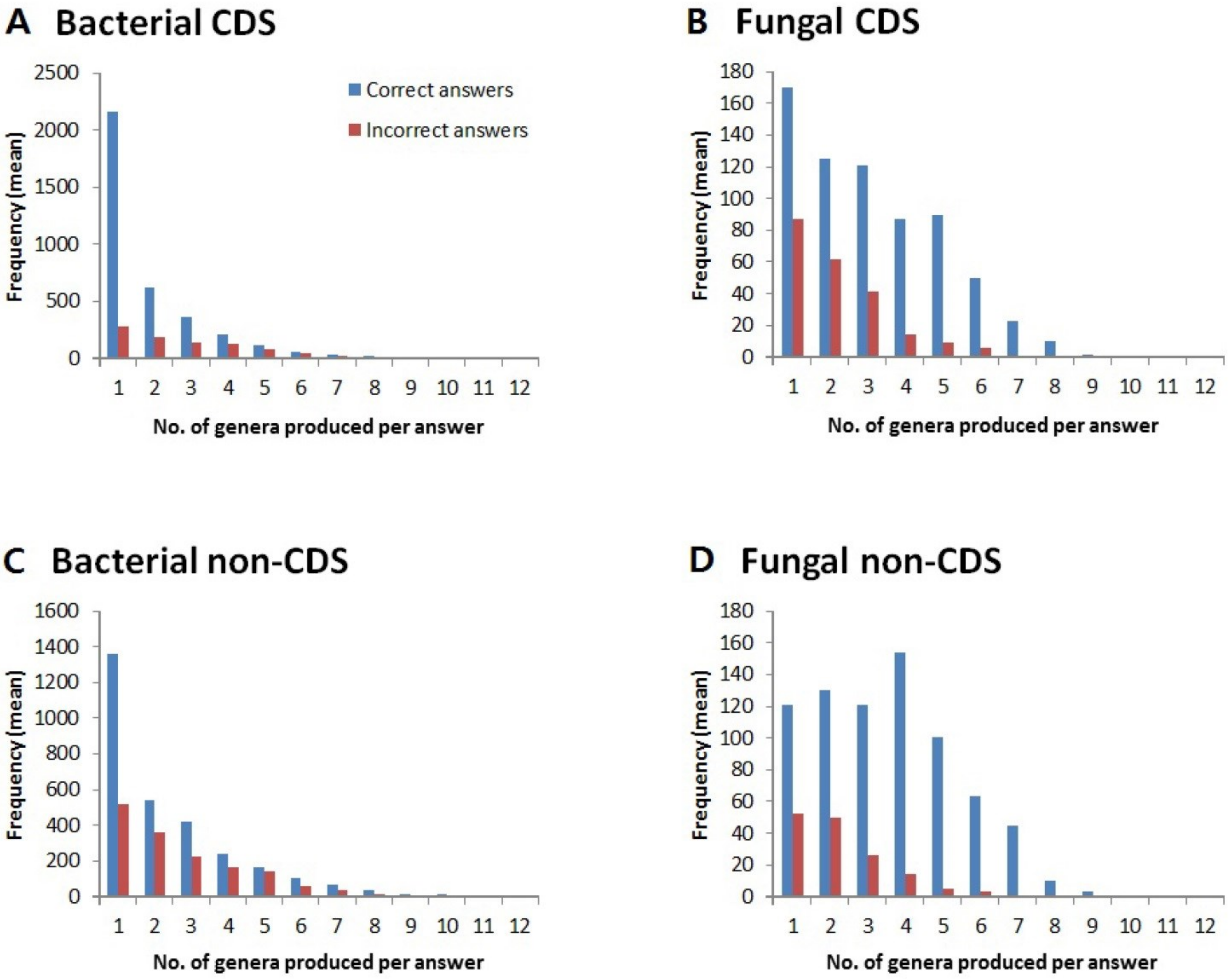
[37] Zhang N, Buck M. A perspective on the enhancer dependent bacterial RNA polymerase. Biomolecules 2015; 5: 1012–9.

[38] Fisher MA, Grimm D, Henion AK, Elias AF, Stewart PE, Rosa PA, et al. Borrelia burgdorferi sigma54 is required for mammalian infection and vector transmission but not for tick colonization. Proc Natl Acad Sci U S A 2005; 102: 5162–7.

[39] Byrgazov K, Vesper O, Moll I. Ribosome heterogeneity: another level of complexity in bacterial translation regulation. Curr Opin Microbiol 2013; 16: 133–9.

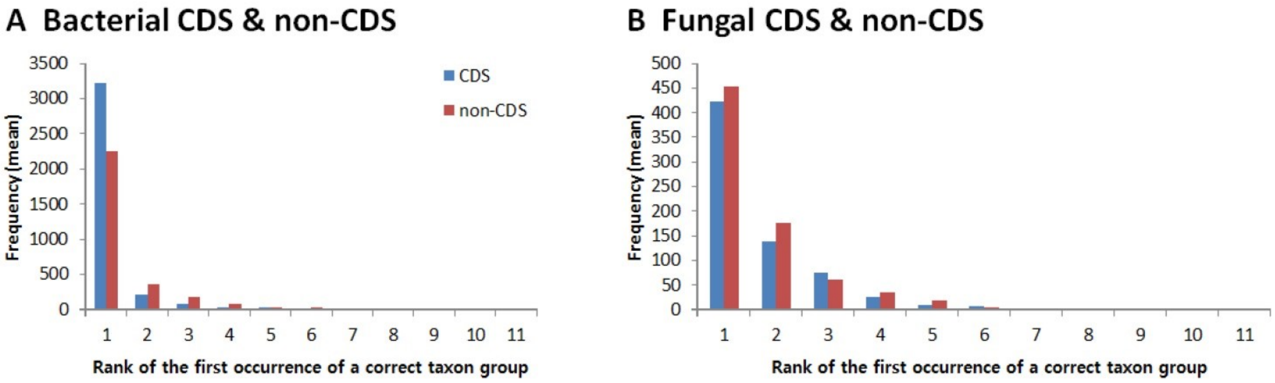
[40] Klappenbach JA, Dunbar JM, Schmidt TM. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol 2000; 66: 1328–33

## 2.9 Supplementary material



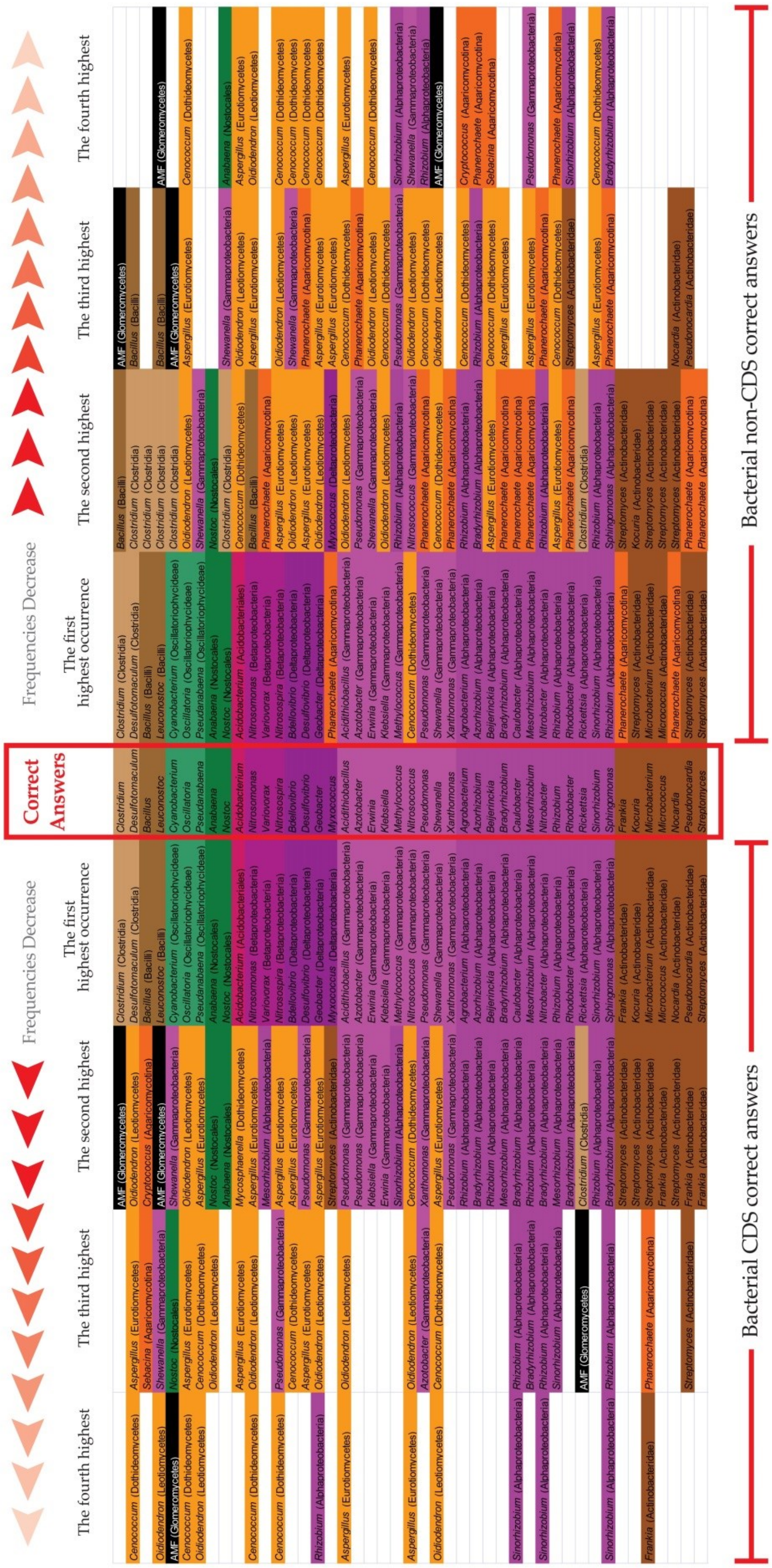
**Supplementary Figure 1 Histogram of the number of genera produced per answer**

X-axis represents how many genera the trimer usage probability scoring method produced in an answer of a query sequence.



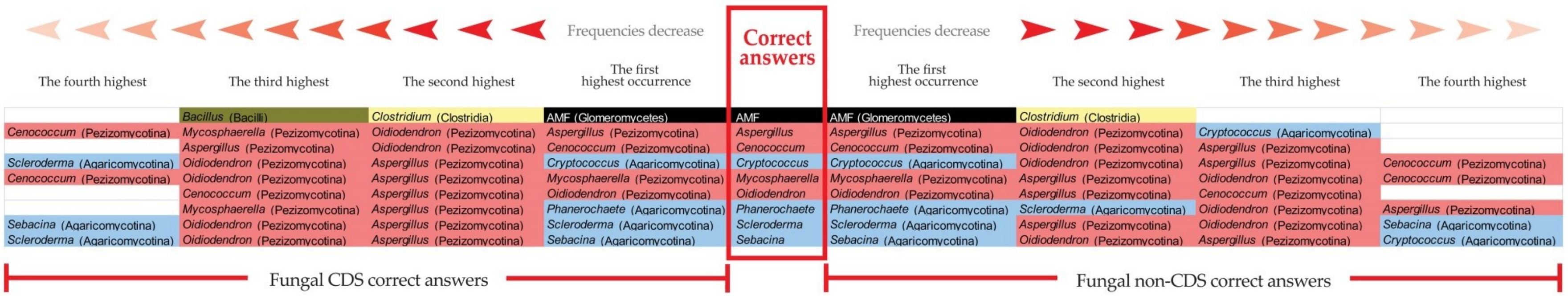
**Supplementary Figure 2 Histogram of the rank of correct taxon group in answer**

X-axis represents at which rank a correct taxon group occurred for the first time in an answer of a query sequence.



**Supplementary Figure 3 Genera with similar trimer usage probability scores (Bacteria)**

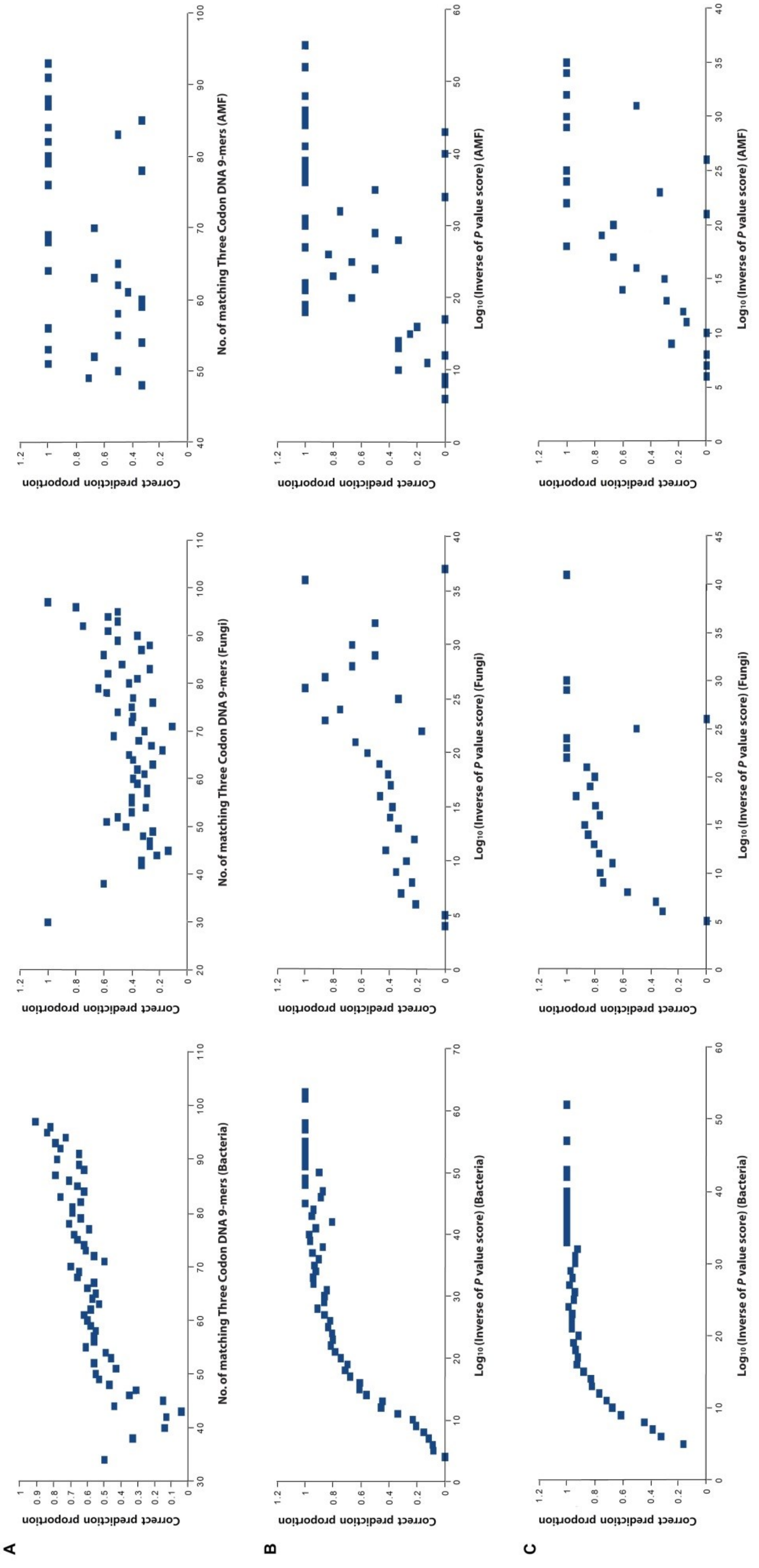
The number of occurrences of each genus included in the correct answers was counted in the bacterial genus CDS and non-CDS test sets. The figure shows the genera with relatively high occurrences. In general, the genera that had little difference in their trimer usage probability scores belonged to the same phylum.



| The fourth highest | The third highest | The second highest | The first highest occurrence | Correct answers | The first highest occurrence | The second highest | The third highest | The fourth highest |
|---|---|---|---|---|---|---|---|---|
|  | *Bacillus* (Bacilli) | *Clostridium* (Clostridia) | AMF (Glomeromycetes) | AMF | AMF (Glomeromycetes) | *Clostridium* (Clostridia) |  |  |
| *Cenococcum* (Pezizomycotina) | *Mycosphaerella* (Pezizomycotina) | *Oidiodendron* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) | *Aspergillus* | *Aspergillus* (Pezizomycotina) | *Oidiodendron* (Pezizomycotina) | *Cryptococcus* (Agaricomycotina) |  |
|  | *Aspergillus* (Pezizomycotina) | *Oidiodendron* (Pezizomycotina) | *Cenococcum* (Pezizomycotina) | *Cenococcum* | *Cenococcum* (Pezizomycotina) | *Oidiodendron* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) |  |
| *Scleroderma* (Agaricomycotina) | *Oidiodendron* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) | *Cryptococcus* (Agaricomycotina) | *Cryptococcus* | *Cryptococcus* (Agaricomycotina) | *Oidiodendron* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) | *Cenococcum* (Pezizomycotina) |
| *Cenococcum* (Pezizomycotina) | *Oidiodendron* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) | *Mycosphaerella* (Pezizomycotina) | *Mycosphaerella* | *Mycosphaerella* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) | *Oidiodendron* (Pezizomycotina) | *Cenococcum* (Pezizomycotina) |
|  | *Cenococcum* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) | *Oidiodendron* (Pezizomycotina) | *Oidiodendron* | *Oidiodendron* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) | *Cenococcum* (Pezizomycotina) |  |
|  | *Mycosphaerella* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) | *Phanerochaete* (Agaricomycotina) | *Phanerochaete* | *Phanerochaete* (Agaricomycotina) | *Scleroderma* (Agaricomycotina) | *Oidiodendron* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) |
| *Sebacina* (Agaricomycotina) | *Oidiodendron* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) | *Scleroderma* (Agaricomycotina) | *Scleroderma* | *Scleroderma* (Agaricomycotina) | *Aspergillus* (Pezizomycotina) | *Oidiodendron* (Pezizomycotina) | *Sebacina* (Agaricomycotina) |
| *Scleroderma* (Agaricomycotina) | *Oidiodendron* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) | *Sebacina* (Agaricomycotina) | *Sebacina* | *Sebacina* (Agaricomycotina) | *Oidiodendron* (Pezizomycotina) | *Aspergillus* (Pezizomycotina) | *Cryptococcus* (Agaricomycotina) |

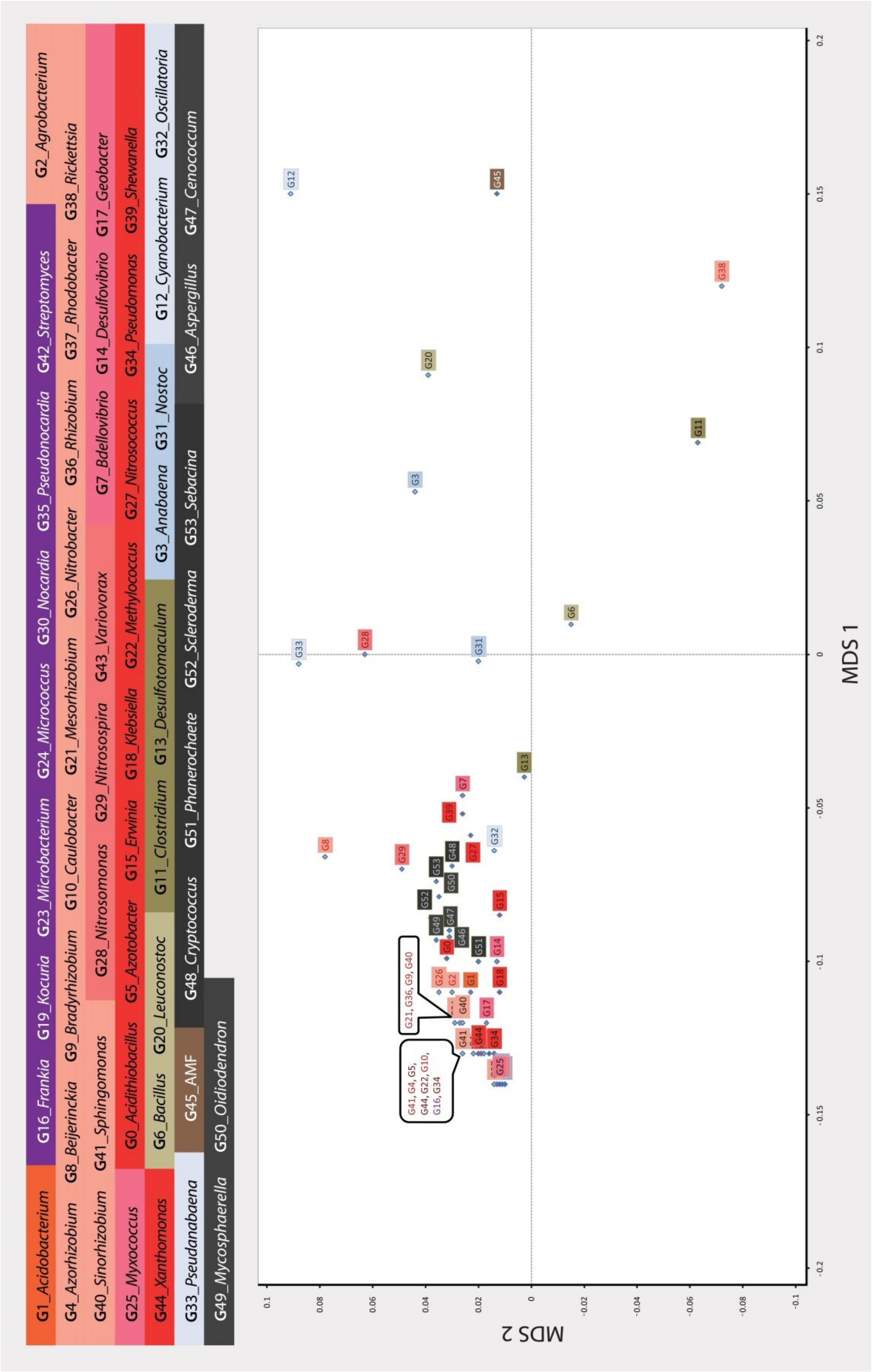
**Supplementary Figure 4 Genera with similar trimer usage probability scores (Fungi)**

The number of occurrence of each genus included in the correct answers was counted in the fungal genus CDS and non-CDS test sets. The figure shows the genera with relatively high occurrences. *Clostridium* occurred the second highest in the AMF test set based on trimer usage probability scoring method. Both Agaricomycotina and Pezizomycotina belong to Dikarya, and the former occurred the second highest in the answers of the latter, and vice versa.



**Supplementary Figure 5 Scatter plot of correct prediction proportion and (log$_{10}$ (inverse of *P* value score))**

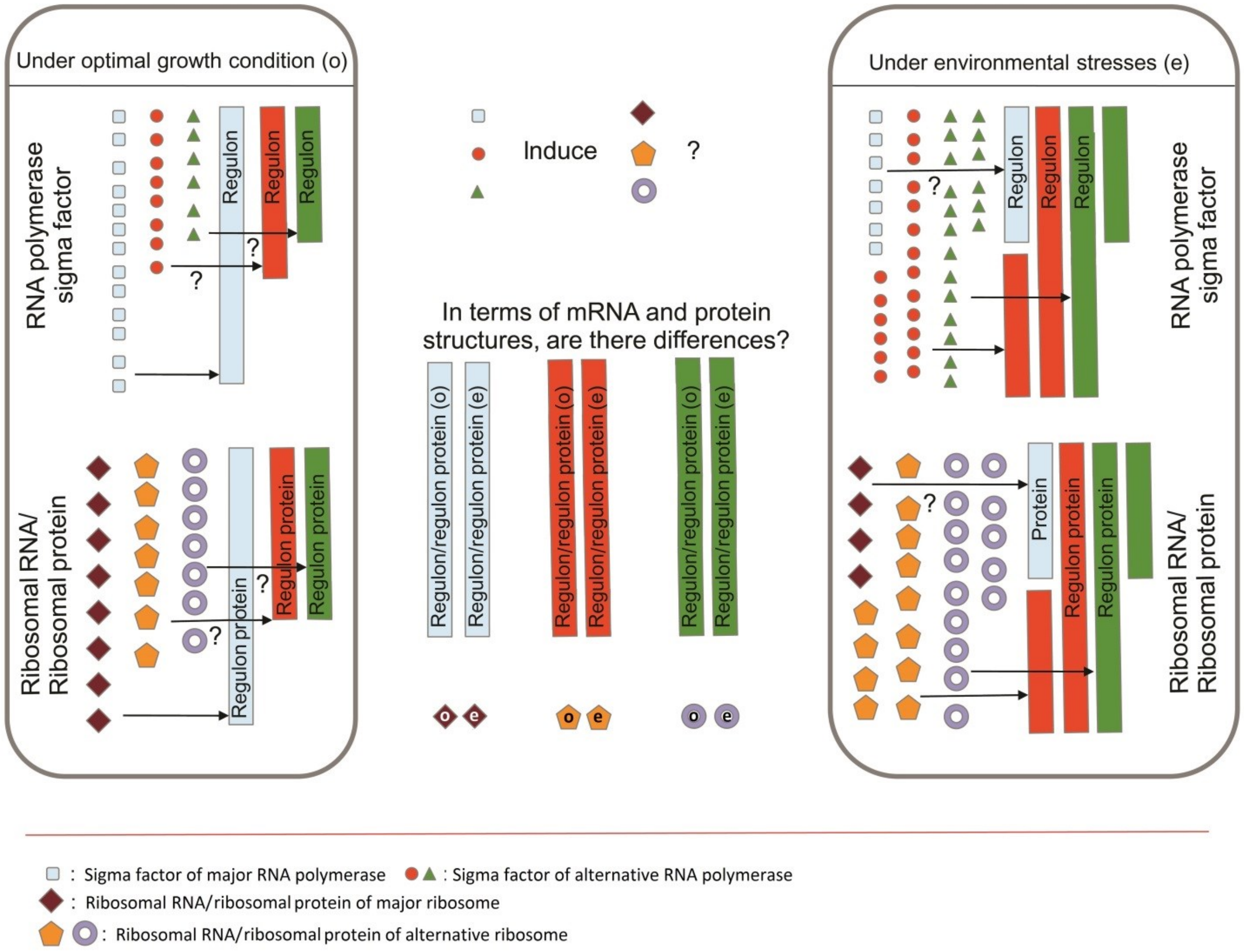
**A.** The mean of the correct prediction proportions was calculated per number of matching *Three Codon DNA 9-mers* in the bacterial, the fungal, and the AMF CDS test sets. For example, 22 sequences had 97 matching *Three Codon DNA 9-mers* and 20 of them were correctly classified in the bacterial test set; the mean was 0.91. **B, C.** We calculated ($\log_{10}$ (inverse of *P* value score)) in order to show a relationship between the correct prediction proportion and *P* value score. For example, the multiplicative inverses of both *P* value scores- 1.0E-10 and 9.0E-10- were approximated to be 1.0E10, and base 10 logarithm of 1.0E10 was 10. **B** was based on the result from the trimer usage probability scoring method, while **C** was based on the result from the rank probability scoring method.



**Supplementary Figure 6 Visualization of proximity relationships among 54 genera using MDS**

The XY axis graph represents proximity relationships among 54 genera based on *trimer usage biases* of *Three Codon DNA 9-mers* identified in the presumably open reading frame of the example sequence.

*Note*: Genera belonging to the same taxonomic group in 13 taxon groups are indicated by the same background color.


Under optimal growth condition (o)
RNA polymerase sigma factor
Regulon
Regulon
Regulon
?
?
Ribosomal RNA/ Ribosomal protein
Regulon protein
Regulon protein
Regulon protein
?
?
Induce
?
In terms of mRNA and protein structures, are there differences?
Regulon/regulon protein (o)
Regulon/regulon protein (e)
Regulon/regulon protein (o)
Regulon/regulon protein (e)
Regulon/regulon protein (o)
Regulon/regulon protein (e)
o
e
o
e
o
e
Under environmental stresses (e)
Regulon
Regulon
Regulon
?
RNA polymerase sigma factor
Protein
Regulon protein
Regulon protein
?
Ribosomal RNA/ Ribosomal protein
□ : Sigma factor of major RNA polymerase
● ▲ : Sigma factor of alternative RNA polymerase
◆ : Ribosomal RNA/ribosomal protein of major ribosome
⬟ ◎ : Ribosomal RNA/ribosomal protein of alternative ribosome

**Supplementary Figure 7 Heterogeneous regulator subpopulations within a single isolate**

The symbols and question marks in the figure indicate the following questions from left to right. Under an optimal growth condition, do alternative regulators transcribe/translate genes? Does sigma factor regulate the expression of functionally specialized ribosomal rRNA and protein coding genes? Do heterogeneous regulator subpopulations produce structurally different gene products? Under environmental stress, do major regulators transcribe/translate genes?

**Supplementary Table 1 Total number of the bacterial genes per genus**

| Bacterial genus | Total number of genes | Bacterial genus | Total number of genes |
|---|---|---|---|
| *Acidithiobacillus* | 12252 | *Microbacterium* | 3676 |
| *Acidobacterium* | 7924 | *Micrococcus* | 2236 |
| *Agrobacterium* | 22773 | *Myxococcus* | 22658 |
| *Anabaena* | 16055 | *Nitrobacter* | 7448 |
| *Azorhizobium* | 4717 | *Nitrosococcus* | 9744 |
| *Azotobacter* | 15105 | *Nitrosomonas* | 11481 |
| *Bacillus* | 392788 | *Nitrosospira* | 2805 |
| *Bdellovibrio* | 9939 | *Nocardia* | 19851 |
| *Beijerinckia* | 3784 | *Nostoc* | 16970 |
| *Bradyrhizobium* | 38418 | *Oscillatoria* | 12156 |
| *Caulobacter* | 17132 | *Pseudanabaena* | 3854 |
| *Clostridium* | 168122 | *Pseudomonas* | 321496 |
| *Cyanobacterium* | 6268 | *Pseudonocardia* | 6797 |
| *Desulfotomaculum* | 21547 | *Rhizobium* | 52249 |
| *Desulfovibrio* | 51438 | *Rhodobacter* | 20963 |
| *Erwinia* | 31931 | *Rickettsia* | 46400 |
| *Frankia* | 29711 | *Shewanella* | 102711 |
| *Geobacter* | 34729 | *Sinorhizobium* | 65703 |
| *Klebsiella* | 69640 | *Sphingomonas* | 15561 |
| *Kocuria* | 2356 | *Streptomyces* | 149826 |
| *Leuconostoc* | 16970 | *Variovorax* | 18984 |
| *Mesorhizobium* | 30388 | *Xanthomonas* | 65650 |
| *Methylococcus* | 2960 | | |

**Supplementary Table 2 Total number of the fungal genes and introns per genus**

| Fungal genus | Total number of genes | Total number of introns |
|---|---|---|
| AMF | 21929 | 52385 |
| *Aspergillus* | 19695 | 38513 |
| *Cenococcum* | 14748 | 27036 |
| *Cryptococcus* | 13174 | 68068 |
| *Mycosphaerella* | 13107 | 33856 |
| *Oidiodendron* | 16703 | 32542 |
| *Phanerochaete* | 10048 | 48688 |
| *Scleroderma* | 21012 | 65184 |
| *Sebacina* | 15312 | 58256 |

**Supplementary Table 3  Correct taxon group proportion in an answer in the bacterial test sets**

| Genus | CDS | | Non-CDS | | Genus | CDS | | Non-CDS | |
|---|---|---|---|---|---|---|---|---|---|
| **Genus** | Mean | SD | Mean | SD | **Genus** | Mean | SD | Mean | SD |
| *Acidithiobacillus* | 0.691 | 0.327 | 0.552 | 0.344 | *Microbacterium* | 0.948 | 0.166 | 0.892 | 0.229 |
| *Acidobacterium* | 0.825 | 0.303 | 0.616 | 0.379 | *Micrococcus* | 0.979 | 0.096 | 0.857 | 0.261 |
| *Agrobacterium* | 0.860 | 0.236 | 0.646 | 0.317 | *Myxococcus* | 0.912 | 0.216 | 0.562 | 0.317 |
| *Anabaena* | 0.762 | 0.311 | 0.745 | 0.299 | *Nitrobacter* | 0.911 | 0.187 | 0.767 | 0.295 |
| *Azorhizobium* | 0.922 | 0.188 | 0.708 | 0.294 | *Nitrosococcus* | 0.632 | 0.344 | 0.577 | 0.351 |
| *Azotobacter* | 0.788 | 0.287 | 0.792 | 0.298 | *Nitrosomonas* | 0.776 | 0.343 | 0.719 | 0.384 |
| *Bacillus* | 0.758 | 0.334 | 0.610 | 0.322 | *Nitrosospira* | 0.826 | 0.297 | 0.722 | 0.351 |
| *Bdellovibrio* | 0.886 | 0.247 | 0.715 | 0.358 | *Nocardia* | 0.850 | 0.254 | 0.755 | 0.285 |
| *Beijerinckia* | 0.834 | 0.279 | 0.662 | 0.341 | *Nostoc* | 0.927 | 0.169 | 0.770 | 0.305 |
| *Bradyrhizobium* | 0.928 | 0.175 | 0.696 | 0.289 | *Oscillatoria* | 0.830 | 0.297 | 0.696 | 0.357 |
| *Caulobacter* | 0.897 | 0.203 | 0.815 | 0.269 | *Pseudanabaena* | 0.939 | 0.184 | 0.643 | 0.344 |
| *Clostridium* | 0.864 | 0.252 | 0.792 | 0.285 | *Pseudomonas* | 0.840 | 0.250 | 0.664 | 0.327 |
| *Cyanobacterium* | 0.960 | 0.150 | 0.736 | 0.328 | *Pseudonocardia* | 0.972 | 0.127 | 0.911 | 0.202 |
| *Desulfotomaculum* | 0.782 | 0.308 | 0.662 | 0.304 | *Rhizobium* | 0.896 | 0.191 | 0.673 | 0.336 |
| *Desulfovibrio* | 0.592 | 0.330 | 0.504 | 0.317 | *Rhodobacter* | 0.934 | 0.190 | 0.687 | 0.267 |
| *Erwinia* | 0.841 | 0.279 | 0.585 | 0.339 | *Rickettsia* | 0.904 | 0.234 | 0.823 | 0.290 |
| *Frankia* | 0.907 | 0.201 | 0.841 | 0.277 | *Shewanella* | 0.773 | 0.336 | 0.715 | 0.337 |
| *Geobacter* | 0.770 | 0.299 | 0.594 | 0.344 | *Sinorhizobium* | 0.855 | 0.239 | 0.776 | 0.304 |
| *Klebsiella* | 0.893 | 0.236 | 0.746 | 0.331 | *Sphingomonas* | 0.917 | 0.181 | 0.757 | 0.267 |
| *Kocuria* | 0.967 | 0.130 | 0.885 | 0.238 | *Streptomyces* | 0.953 | 0.158 | 0.848 | 0.262 |
| *Leuconostoc* | 0.838 | 0.282 | 0.594 | 0.296 | *Variovorax* | 0.938 | 0.195 | 0.684 | 0.337 |
| *Mesorhizobium* | 0.876 | 0.214 | 0.729 | 0.304 | *Xanthomonas* | 0.904 | 0.215 | 0.751 | 0.319 |
| *Methylococcus* | 0.796 | 0.289 | 0.698 | 0.346 | **Mean** | **0.87** | **0.25** | **0.72** | **0.32** |

*Note*: After genera in an answer were converted to the 13 taxon groups, the proportion of the correct taxon group was calculated per sequence in the genus test set. The mean and the standard deviation of the proportions of the correct taxon group are shown in the table.

**Supplementary Table 4 Correct taxon group proportion in an answer in the fungal test sets**

| **Genus** | **CDS** | | **Non-CDS** | | **Genus** | **CDS** | | **Non-CDS** | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | | Mean | SD | Mean | SD |
| AMF | 0.706 | 0.302 | 0.689 | 0.317 | *Oidiodendron* | 0.385 | 0.280 | 0.326 | 0.233 |
| *Aspergillus* | 0.450 | 0.294 | 0.367 | 0.258 | *Phanerochaete* | 0.593 | 0.297 | 0.608 | 0.261 |
| *Cenococcum* | 0.467 | 0.276 | 0.429 | 0.282 | *Scleroderma* | 0.556 | 0.244 | 0.563 | 0.248 |
| *Cryptococcus* | 0.668 | 0.280 | 0.523 | 0.265 | *Sebacina* | 0.605 | 0.272 | 0.499 | 0.226 |
| *Mycosphaerella* | 0.650 | 0.309 | 0.457 | 0.255 | **Mean** | **0.567** | **0.298** | **0.498** | **0.28** |

*Note*: After genera in an answer were converted to the 13 taxon groups, the proportion of the correct taxon group was calculated per sequence in the genus test set. The mean and the standard deviation of the proportions of the correct taxon group are shown in the table.

**Supplementary Table 5 Frequency of the number of genera produced per answer**

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **Sum** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bact. CDS: correct | 2160 | 627 | 361 | 210 | 115 | 56 | 33 | 17 | 2 | 4 | 1 | 1 | 3587 |
| Bact. CDS: incorrect | 281 | 189 | 143 | 126 | 80 | 49 | 23 | 16 | 6 | 0 | 0 | 0 | 913 |
| Fung. CDS: correct | 170 | 125 | 121 | 87 | 90 | 50 | 23 | 10 | 2 | 0 | 0 | 0 | 678 |
| Fung. CDS: incorrect | 87 | 62 | 41 | 14 | 9 | 6 | 1 | 1 | 0 | 1 | 0 | 0 | 222 |
| | | | | | | | | | | | | **Total** | 5400 |
| Bact. non-CDS: correct | 1358 | 539 | 418 | 242 | 163 | 108 | 67 | 37 | 18 | 14 | 5 | 1 | 2970 |
| Bact. non-CDS: incorrect | 521 | 363 | 222 | 165 | 144 | 58 | 38 | 15 | 3 | 0 | 1 | 0 | 1530 |
| Fung. non-CDS: correct | 121 | 130 | 121 | 154 | 101 | 63 | 45 | 10 | 3 | 1 | 1 | 0 | 750 |
| Fung. non-CDS: incorrect | 52 | 50 | 26 | 14 | 5 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 150 |
| | | | | | | | | | | | | **Total** | 5400 |

*Note*: The table shows the frequencies of how many genera the trimer usage probability scoring method produced in an answer of a query sequence in the correct and the incorrect results in the bacterial (bact.) and the fungal (fung.) CDS and non-CDS test sets. Data for Supplementary Figure 1.

**Supplementary Table 6 Frequency of the rank of correct taxon group in answer**

| | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **Sum** | **Total** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bact. CDS | 3218 | 208 | 86 | 33 | 29 | 9 | 1 | 3 | 0 | 0 | 0 | 3587 | 4500 |
| Bact. non-CDS | 2260 | 368 | 178 | 75 | 38 | 28 | 9 | 5 | 5 | 2 | 2 | 2970 | 4500 |
| Fung. CDS | 422 | 138 | 76 | 25 | 10 | 6 | 0 | 1 | 0 | 0 | 0 | 678 | 900 |
| Fung. non-CDS | 453 | 177 | 60 | 36 | 19 | 5 | 0 | 0 | 0 | 0 | 0 | 750 | 900 |

*Note*: The table shows the frequency of at which rank the trimer usage probability scoring method produced a correct taxon group for the first time in an answer of a query sequence in the correct results in the bacterial (bact.) and the fungal (fung.) CDS and non-CDS test sets. Data for Supplementary Figure 2.

**Supplementary Table 7 Percentage of the other group in answers**

| | Correct prediction percentages (genus) | The other group in answers | AMF in answers |
|---|---|---|---|
| Bact. CDS | 3185/4500: 71% | 197/3185: 6% | 119/4500: 3% |
| Bact. non-CDS | 2238/4500: 50% | 489/2238: 22% | 229/4500: 5% |
| Fung. CDS | 589/900: 65% | 93/589: 16% | 6/800: 1% |
| Fung. non-CDS | 655/900: 73% | 157/655: 24% | 10/800: 1% |
| AMF CDS | 49/100: 49% | 24/49: 49% | |
| AMF non-CDS | 72/100: 72% | 35/72: 49% | |

*Note*: The column, The other group in answers, indicates the percentage of the fungal (fung.) and the bacterial (bact.) group in answers of the bacterial and the fungal test sets, respectively, while it indicates the percentage of the bacterial group in case of the AMF test sets. The column, AMF in answers, indicates the percentage of AMF in answers of the bacterial and the fungal test sets.

**Supplementary Table 8 Correlation between the correct prediction proportion and the number of matching *Three Codon DNA 9-mers***

| **Bacteria** | | **Fungi** | | **AMF** | |
|---|---|---|---|---|---|
| **No. of matching Three Codon DNA 9-mers** | **Correct prediction proportion** | **No. of matching Three Codon DNA 9-mers** | **Correct prediction proportion** | **No. of matching Three Codon DNA 9-mers** | **Correct prediction proportion** |
| 34 | 0.5 | 30 | 1 | 48 | 0.33 |
| 38 | 0.33 | 38 | 0.6 | 49 | 0.71 |
| 40 | 0.14 | 42 | 0.33 | 50 | 0.5 |
| 42 | 0.13 | 43 | 0.33 | 51 | 1 |
| 43 | 0.04 | 44 | 0.22 | 52 | 0.67 |
| 44 | 0.44 | 45 | 0.14 | 53 | 1 |
| 45 | 0.15 | 46 | 0.27 | 54 | 0.33 |
| 46 | 0.35 | 47 | 0.27 | 55 | 0.5 |
| 47 | 0.31 | 48 | 0.32 | 56 | 1 |
| 48 | 0.47 | 49 | 0.25 | 58 | 0.5 |
| 49 | 0.53 | 50 | 0.44 | 59 | 0.33 |
| 50 | 0.55 | 51 | 0.58 | 60 | 0.33 |
| 51 | 0.43 | 52 | 0.5 | 61 | 0.43 |
| 52 | 0.56 | 53 | 0.4 | 62 | 0.5 |
| 53 | 0.46 | 54 | 0.3 | 63 | 0.67 |
| 54 | 0.49 | 55 | 0.4 | 64 | 1 |
| 55 | 0.61 | 56 | 0.4 | 65 | 0.5 |
| 56 | 0.56 | 57 | 0.29 | 68 | 1 |
| 57 | 0.56 | 58 | 0.29 | 69 | 1 |
| 58 | 0.55 | 59 | 0.36 | 70 | 0.67 |
| 59 | 0.58 | 60 | 0.39 | 76 | 1 |
| 60 | 0.6 | 61 | 0.31 | 78 | 0.33 |
| 61 | 0.62 | 62 | 0.36 | 79 | 1 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 62 | 0.58 | 63 | 0.25 | 80 | 1 |
| 63 | 0.53 | 64 | 0.39 | 82 | 1 |
| 64 | 0.57 | 65 | 0.42 | 83 | 0.5 |
| 65 | 0.55 | 66 | 0.18 | 84 | 1 |
| 66 | 0.6 | 67 | 0.26 | 85 | 0.33 |
| 67 | 0.56 | 68 | 0.35 | 87 | 1 |
| 68 | 0.66 | 69 | 0.53 | 88 | 1 |
| 69 | 0.65 | 70 | 0.31 | 91 | 1 |
| 70 | 0.7 | 71 | 0.11 | 93 | 1 |
| 71 | 0.5 | 72 | 0.4 |  |  |
| 72 | 0.56 | 73 | 0.39 |  |  |
| 73 | 0.61 | 74 | 0.5 |  |  |
| 74 | 0.62 | 75 | 0.4 |  |  |
| 75 | 0.66 | 76 | 0.25 |  |  |
| 76 | 0.68 | 77 | 0.39 |  |  |
| 77 | 0.59 | 78 | 0.58 |  |  |
| 78 | 0.71 | 79 | 0.64 |  |  |
| 79 | 0.64 | 80 | 0.42 |  |  |
| 80 | 0.69 | 81 | 0.36 |  |  |
| 81 | 0.69 | 82 | 0.57 |  |  |
| 82 | 0.64 | 83 | 0.27 |  |  |
| 83 | 0.76 | 84 | 0.47 |  |  |
| 84 | 0.62 | 86 | 0.6 |  |  |
| 85 | 0.66 | 87 | 0.33 |  |  |
| 86 | 0.71 | 88 | 0.27 |  |  |
| 87 | 0.79 | 89 | 0.5 |  |  |
| 88 | 0.62 | 90 | 0.36 |  |  |
| 89 | 0.65 | 91 | 0.57 |  |  |
| 90 | 0.78 | 92 | 0.75 |  |  |

|  |  |  |  |
|---|---|---|---|
| 91 | 0.65 | 93 | 0.5 |
| 92 | 0.76 | 94 | 0.57 |
| 93 | 0.79 | 95 | 0.5 |
| 94 | 0.73 | 96 | 0.8 |
| 95 | 0.84 | 97 | 1 |
| 96 | 0.82 |  |  |
| 97 | 0.91 |  |  |

*Note*: Data for Supplementary Figure 5.A.

**Supplementary Table 9 Correlation between the correct prediction proportion of the trimer usage probability scoring method and *P* value score**

| Bacteria | | Fungi | | AMF | |
|---|---|---|---|---|---|
| $Log_{10}$ (Inverse of *P* value score) | Correct prediction proportion | $Log_{10}$ (Inverse of *P* value score) | Correct prediction proportion | $Log_{10}$ (Inverse of *P* value score) | Correct prediction proportion |
| 4 | 0 | 4 | 0 | 6 | 0 |
| 5 | 0.0833 | 5 | 0 | 8 | 0 |
| 6 | 0.0879 | 6 | 0.21 | 9 | 0 |
| 7 | 0.116 | 7 | 0.314 | 10 | 0.333 |
| 8 | 0.152 | 8 | 0.236 | 11 | 0.125 |
| 9 | 0.206 | 9 | 0.349 | 12 | 0 |
| 10 | 0.232 | 10 | 0.277 | 13 | 0.333 |
| 11 | 0.338 | 11 | 0.424 | 14 | 0.333 |
| 12 | 0.457 | 12 | 0.22 | 15 | 0.25 |
| 13 | 0.447 | 13 | 0.333 | 16 | 0.2 |
| 14 | 0.562 | 14 | 0.393 | 17 | 0 |
| 15 | 0.613 | 15 | 0.377 | 18 | 1 |
| 16 | 0.609 | 16 | 0.465 | 19 | 1 |
| 17 | 0.678 | 17 | 0.388 | 20 | 0.666 |
| 18 | 0.715 | 18 | 0.406 | 21 | 1 |
| 19 | 0.698 | 19 | 0.47 | 22 | 1 |
| 20 | 0.746 | 20 | 0.555 | 23 | 0.8 |
| 21 | 0.787 | 21 | 0.642 | 24 | 0.5 |
| 22 | 0.818 | 22 | 0.166 | 25 | 0.666 |
| 23 | 0.804 | 23 | 0.857 | 26 | 0.833 |
| 24 | 0.809 | 24 | 0.75 | 27 | 1 |
| 25 | 0.837 | 25 | 0.333 | 28 | 0.333 |
| 26 | 0.822 | 26 | 1 | 29 | 0.5 |

| | | | | | |
|---|---|---|---|---|---|
| 27 | 0.863 | 27 | 0.857 | 30 | 1 |
| 28 | 0.912 | 28 | 0.666 | 31 | 1 |
| 29 | 0.864 | 29 | 0.5 | 32 | 0.75 |
| 30 | 0.862 | 30 | 0.666 | 34 | 0 |
| 31 | 0.845 | 32 | 0.5 | 35 | 0.5 |
| 32 | 0.94 | 36 | 1 | 36 | 1 |
| 33 | 0.944 | 37 | 0 | 37 | 1 |
| 34 | 0.925 | | | 38 | 1 |
| 35 | 0.933 | | | 39 | 1 |
| 36 | 0.903 | | | 40 | 0 |
| 37 | 0.948 | | | 41 | 1 |
| 38 | 0.875 | | | 43 | 0 |
| 39 | 0.966 | | | 44 | 1 |
| 40 | 0.972 | | | 45 | 1 |
| 41 | 0.925 | | | 46 | 1 |
| 42 | 0.809 | | | 48 | 1 |
| 43 | 0.952 | | | 52 | 1 |
| 44 | 0.941 | | | 55 | 1 |
| 45 | 1 | | | | |
| 46 | 0.888 | | | | |
| 47 | 0.875 | | | | |
| 48 | 1 | | | | |
| 49 | 1 | | | | |
| 50 | 0.9 | | | | |
| 51 | 1 | | | | |
| 52 | 1 | | | | |
| 53 | 1 | | | | |
| 54 | 1 | | | | |
| 55 | 1 | | | | |

| | |
|---|---|
| 57 | 1 |
| 58 | 1 |
| 62 | 1 |
| 63 | 1 |

*Note*: The mean of the correct prediction proportions per (log$_{10}$ (Inverse of *P* value score)) was calculated based on the first ranked genus with the highest probability score in the result from the trimer usage probability scoring method applied to the bacterial, the fungal, and the AMF CDS test sets. Data for Supplementary Figure 5.B.

**Supplementary Table 10 Correlation between the correct prediction proportion of the rank probability scoring method and *P* value score**

| **Bacteria** | | **Fungi** | | **AMF** | |
|---|---|---|---|---|---|
| **$Log_{10}$ (Inverse of *P* value score)** | **Correct prediction proportion** | **$Log_{10}$ (Inverse of *P* value score)** | **Correct prediction proportion** | **$Log_{10}$ (Inverse of *P* value score)** | **Correct prediction proportion** |
| 5 | 0.166 | 5 | 0 | 6 | 0 |
| 6 | 0.325 | 6 | 0.315 | 7 | 0 |
| 7 | 0.386 | 7 | 0.363 | 8 | 0 |
| 8 | 0.445 | 8 | 0.565 | 9 | 0.25 |
| 9 | 0.612 | 9 | 0.74 | 10 | 0 |
| 10 | 0.676 | 10 | 0.762 | 11 | 0.142 |
| 11 | 0.717 | 11 | 0.674 | 12 | 0.166 |
| 12 | 0.768 | 12 | 0.767 | 13 | 0.285 |
| 13 | 0.821 | 13 | 0.805 | 14 | 0.6 |
| 14 | 0.828 | 14 | 0.847 | 15 | 0.3 |
| 15 | 0.881 | 15 | 0.871 | 16 | 0.5 |
| 16 | 0.929 | 16 | 0.763 | 17 | 0.666 |
| 17 | 0.922 | 17 | 0.794 | 18 | 1 |
| 18 | 0.939 | 18 | 0.933 | 19 | 0.75 |
| 19 | 0.952 | 19 | 0.833 | 20 | 0.666 |
| 20 | 0.916 | 20 | 0.8 | 21 | 0 |
| 21 | 0.964 | 21 | 0.857 | 22 | 1 |
| 22 | 0.965 | 22 | 1 | 23 | 0.333 |
| 23 | 0.961 | 23 | 1 | 24 | 1 |
| 24 | 0.989 | 24 | 1 | 25 | 1 |
| 25 | 0.95 | 25 | 0.5 | 26 | 0 |
| 26 | 0.942 | 26 | 0 | 29 | 1 |
| 27 | 0.98 | 29 | 1 | 30 | 1 |
| 28 | 0.96 | 30 | 1 | 31 | 0.5 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 29 | 0.975 | 41 | 1 | 32 | 1 |
| 30 | 0.941 |  |  | 34 | 1 |
| 31 | 0.941 |  |  | 35 | 1 |
| 32 | 0.923 |  |  |  |  |
| 33 | 1 |  |  |  |  |
| 34 | 1 |  |  |  |  |
| 35 | 1 |  |  |  |  |
| 36 | 1 |  |  |  |  |
| 37 | 1 |  |  |  |  |
| 38 | 1 |  |  |  |  |
| 39 | 1 |  |  |  |  |
| 40 | 1 |  |  |  |  |
| 42 | 1 |  |  |  |  |
| 43 | 1 |  |  |  |  |
| 47 | 1 |  |  |  |  |
| 52 | 1 |  |  |  |  |

*Note*: The mean of the correct prediction proportions per (log$_{10}$ (Inverse of *P* value score)) was calculated based on the first ranked genus with the highest probability score in the result from the rank probability scoring method applied to the bacterial, the fungal, and the AMF CDS test sets. Data for Supplementary Figure 5.C.

**Supplementary Table 11 Relationship between the correct prediction proportion of the trimer usage probability scoring method and *P* value score in quartiles**

| Genus | 0 - 25th percentile | | | 26 - 50th percentile | | | 51 - 75th percentile | | | 76 - 100th percentile | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Range | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range | Mean | SD |
| *Acidithiobacillus* | 5-8 | 0.151 | 0.18 | 9-13 | 0.41 | 0.217 | 14-19 | 0.461 | 0.402 | 20-32 | 0.6 | 0.547 |
| *Acidobacterium* | 6-10 | 0.276 | 0.18 | 11-15 | 0.564 | 0.223 | 16-20 | 0.62 | 0.073 | 21-28 | 0.875 | 0.306 |
| *Agrobacterium* | 4-9 | 0.045 | 0.0622 | 10-14 | 0.425 | 0.139 | 15-19 | 0.733 | 0.278 | 20-30 | 0.444 | 0.455 |
| *Anabaena* | 5-13 | 0.157 | 0.329 | 14-22 | 0.327 | 0.321 | 23-32 | 0.529 | 0.425 | 33-51 | 0.85 | 0.337 |
| *Azorhizobium* | 7-15 | 0.333 | 0.388 | 16-23 | 0.75 | 0.277 | 24-31 | 0.975 | 0.0707 | 32-43 | 1 | 0 |
| *Azotobacter* | 6-12 | 0.276 | 0.34 | 13-20 | 0.783 | 0.357 | 21-27 | 0.959 | 0.107 | 28-41 | 1 | 0 |
| *Bacillus* | 5-12 | 0.182 | 0.227 | 13-21 | 0.416 | 0.333 | 22-30 | 0.679 | 0.22 | 31-47 | 0.833 | 0.25 |
| *Bdellovibrio* | 6-11 | 0.205 | 0.186 | 12-18 | 0.669 | 0.308 | 19-25 | 0.821 | 0.144 | 26-37 | 0.928 | 0.188 |
| *Beijerinckia* | 5-9 | 0.08 | 0.109 | 10-15 | 0.473 | 0.279 | 16-21 | 0.906 | 0.0924 | 22-34 | 1 | 0 |
| *Bradyrhizobium* | 4-11 | 0.178 | 0.237 | 12-19 | 0.667 | 0.237 | 20-26 | 0.857 | 0.196 | 27-40 | 0.937 | 0.176 |
| *Caulobacter* | 6-13 | 0.328 | 0.37 | 14-22 | 0.781 | 0.213 | 23-31 | 0.982 | 0.0505 | 32-42 | 1 | 0 |
| *Clostridium* | 6-20 | 0.424 | 0.389 | 21-32 | 0.83 | 0.211 | 33-45 | 1 | 0 | 47-63 | 0.972 | 0.0962 |
| *Cyanobacterium* | 7-19 | 0.233 | 0.344 | 20-30 | 0.903 | 0.205 | 31-41 | 0.893 | 0.238 | 42-54 | 1 | 0 |
| *Desulfotomaculum* | 5-10 | 0.201 | 0.178 | 11-16 | 0.42 | 0.291 | 17-22 | 0.48 | 0.43 | 23-38 | 0.69 | 0.365 |
| *Desulfovibrio* | 6-10 | 0.253 | 0.31 | 11-15 | 0.276 | 0.145 | 16-20 | 0.23 | 0.338 | 21-30 | 0.0833 | 0.204 |
| *Erwinia* | 5-10 | 0.163 | 0.146 | 11-16 | 0.488 | 0.289 | 17-22 | 0.877 | 0.142 | 23-34 | 1 | 0 |
| *Frankia* | 5-12 | 0.272 | 0.309 | 13-20 | 0.72 | 0.197 | 21-27 | 0.45 | 0.326 | 28-46 | 0.375 | 0.443 |
| *Geobacter* | 6-9 | 0.243 | 0.204 | 10-14 | 0.464 | 0.232 | 15-18 | 0.69 | 0.359 | 19-28 | 0.85 | 0.223 |
| *Klebsiella* | 5-10 | 0.236 | 0.409 | 11-17 | 0.642 | 0.135 | 18-23 | 0.979 | 0.051 | 24-32 | 1 | 0 |
| *Kocuria* | 6-15 | 0.275 | 0.415 | 17-27 | 0.75 | 0.403 | 29-38 | 0.987 | 0.0395 | 39-54 | 0.977 | 0.0753 |
| *Leuconostoc* | 7-14 | 0.166 | 0.288 | 15-23 | 0.854 | 0.242 | 24-31 | 0.979 | 0.0589 | 32-44 | 1 | 0 |
| *Mesorhizobium* | 5-10 | 0.116 | 0.139 | 11-16 | 0.546 | 0.345 | 17-22 | 0.682 | 0.205 | 23-31 | 0.597 | 0.395 |
| *Methylococcus* | 6-10 | 0.14 | 0.219 | 11-15 | 0.711 | 0.309 | 16-20 | 0.763 | 0.152 | 21-30 | 0.934 | 0.106 |
| *Microbacterium* | 7-16 | 0.0937 | 0.265 | 17-26 | 0.922 | 0.171 | 27-34 | 0.933 | 0.128 | 37-50 | 1 | 0 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Micrococcus* | 10-19 | 0.15 | 0.253 | 20-30 | 0.91 | 0.156 | 31-40 | 0.983 | 0.0527 | 41-55 | 1 | 0 |
| *Myxococcus* | 6-16 | 0.461 | 0.373 | 17-25 | 0.869 | 0.182 | 26-34 | 1 | 0 | 35-49 | 1 | 0 |
| *Nitrobacter* | 6-10 | 0.155 | 0.175 | 11-15 | 0.467 | 0.193 | 16-20 | 0.5 | 0.204 | 21-40 | 0.611 | 0.443 |
| *Nitrosococcus* | 4-8 | 0.0333 | 0.0745 | 9-14 | 0.12 | 0.138 | 15-19 | 0.762 | 0.146 | 20-27 | 0.777 | 0.403 |
| *Nitrosomonas* | 5-10 | 0.203 | 0.211 | 11-16 | 0.449 | 0.192 | 17-22 | 0.875 | 0.209 | 23-48 | 0.833 | 0.408 |
| *Nitrosospira* | 6-9 | 0.106 | 0.093 | 10-14 | 0.644 | 0.328 | 15-19 | 0.883 | 0.111 | 20-25 | 0.8 | 0.447 |
| *Nocardia* | 6-12 | 0.226 | 0.229 | 13-19 | 0.717 | 0.271 | 20-26 | 0.826 | 0.149 | 27-36 | 0.802 | 0.35 |
| *Nostoc* | 5-12 | 0.168 | 0.252 | 13-20 | 0.509 | 0.151 | 21-28 | 0.676 | 0.232 | 29-41 | 0.388 | 0.485 |
| *Oscillatoria* | 6-11 | 0.194 | 0.155 | 12-18 | 0.63 | 0.143 | 19-25 | 0.821 | 0.256 | 26-34 | 1 | 0 |
| *Pseudanabaena* | 6-12 | 0.161 | 0.203 | 13-19 | 0.778 | 0.246 | 20-26 | 0.952 | 0.125 | 27-41 | 0.875 | 0.353 |
| *Pseudomonas* | 7-12 | 0.255 | 0.389 | 13-18 | 0.691 | 0.196 | 19-24 | 0.885 | 0.18 | 25-35 | 0.773 | 0.368 |
| *Pseudonocardia* | 9-19 | 0.323 | 0.404 | 20-30 | 0.815 | 0.229 | 31-40 | 0.966 | 0.105 | 41-55 | 1 | 0 |
| *Rhizobium* | 5-9 | 0.125 | 0.19 | 10-15 | 0.222 | 0.178 | 16-21 | 0.376 | 0.287 | 22-28 | 0.805 | 0.305 |
| *Rhodobacter* | 7-14 | 0.216 | 0.357 | 15-22 | 0.848 | 0.217 | 23-30 | 0.921 | 0.175 | 32-48 | 1 | 0 |
| *Rickettsia* | 5-21 | 0.399 | 0.459 | 22-33 | 0.845 | 0.2 | 34-44 | 0.85 | 0.312 | 45-62 | 1 | 0 |
| *Shewanella* | 5-9 | 0.333 | 0.471 | 10-15 | 0.619 | 0.344 | 16-22 | 0.646 | 0.186 | 23-30 | 0.777 | 0.403 |
| *Sinorhizobium* | 4-10 | 0.16 | 0.158 | 11-15 | 0.288 | 0.208 | 16-20 | 0.616 | 0.273 | 21-30 | 0.7 | 0.447 |
| *Sphingomonas* | 6-13 | 0.135 | 0.274 | 14-21 | 0.689 | 0.327 | 22-29 | 0.907 | 0.202 | 30-40 | 1 | 0 |
| *Streptomyces* | 5-15 | 0.101 | 0.154 | 16-24 | 0.765 | 0.193 | 25-33 | 0.856 | 0.194 | 34-50 | 0.796 | 0.328 |
| *Variovorax* | 6-15 | 0.222 | 0.44 | 16-25 | 0.93 | 0.113 | 26-35 | 1 | 0 | 36-49 | 1 | 0 |
| *Xanthomonas* | 7-13 | 0.454 | 0.252 | 14-20 | 0.83 | 0.187 | 21-28 | 0.94 | 0.104 | 29-42 | 1 | 0 |
| AMF | 6-16 | 0.157 | 0.15 | 17-26 | 0.746 | 0.315 | 27-37 | 0.708 | 0.358 | 38-55 | 0.818 | 0.404 |
| *Aspergillus* | 6-10 | 0.405 | 0.234 | 11-15 | 0.24 | 0.205 | 16-20 | 0.116 | 0.162 | 21-37 | 0.2 | 0.447 |
| *Cenococcum* | 6-10 | 0.165 | 0.205 | 11-15 | 0.331 | 0.158 | 16-20 | 0.133 | 0.217 | 21-32 | 0.25 | 0.418 |
| *Cryptococcus* | 6-10 | 0.228 | 0.435 | 11-16 | 0.292 | 0.285 | 17-23 | 0.475 | 0.404 | 24-32 | 0.833 | 0.408 |
| *Mycosphaerella* | 4-8 | 0.266 | 0.326 | 9-13 | 0.56 | 0.308 | 14-18 | 0.518 | 0.153 | 19-26 | 0.542 | 0.366 |
| *Oidiodendron* | 5-9 | 0.15 | 0.223 | 10-14 | 0.325 | 0.129 | 15-19 | 0.526 | 0.345 | 20-25 | 0.533 | 0.505 |
| *Phanerochaete* | 6-9 | 0.471 | 0.11 | 10-14 | 0.424 | 0.157 | 15-19 | 0.373 | 0.127 | 21-30 | 0.6 | 0.547 |
| *Scleroderma* | 5-9 | 0.04 | 0.0894 | 10-14 | 0.287 | 0.149 | 15-19 | 0.388 | 0.146 | 20-36 | 0.9 | 0.223 |
| *Sebacina* | 5-9 | 0.0833 | 0.117 | 10-14 | 0.357 | 0.153 | 15-19 | 0.558 | 0.275 | 20-28 | 1 | 0 |

*Note*: Range represents a minimum and a maximum of ($\log_{10}$ (Inverse of *P* value score)) values per quartile. After the result from each genus test set was divided into quartiles, the range of ($\log_{10}$ (Inverse *P* value score)) and the mean and the standard deviation of the correct prediction proportions were calculated per quartile. The result was based on the trimer usage probability scoring method. Data for Supplementary Table 9 and Supplementary Figure 5.B.

**Supplementary Table 12 Relationship between the correct prediction proportion of the rank probability scoring method and *P* value score in quartiles**

| Genus | 0-25th percentiles | | | 26-50th percentiles | | | 51-75th percentiles | | | 76-100th percentiles | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Range | Mean | SD | Range | Mean | SD | Range | Mean | SD | Range | Mean | SD |
| *Acidithiobacillus* | 6-9 | 0.278 | 0.242 | 10-13 | 0.748 | 0.0505 | 14-17 | 0.968 | 0.0625 | 18-40 | 1 | 0 |
| *Acidobacterium* | 6-10 | 0.277 | 0.277 | 11-15 | 0.82 | 0.106 | 16-20 | 1 | 0 | 21-26 | 1 | 0 |
| *Agrobacterium* | 6-9 | 0.718 | 0.359 | 10-14 | 0.923 | 0.07 | 15-19 | 1 | 0 | 20-27 | 1 | 0 |
| *Anabaena* | 7-12 | 0.189 | 0.244 | 13-18 | 0.721 | 0.37 | 19-24 | 0.901 | 0.113 | 25-31 | 1 | 0 |
| *Azorhizobium* | 7-12 | 0.762 | 0.237 | 13-19 | 1 | 0 | 20-25 | 1 | 0 | 26-34 | 1 | 0 |
| *Azotobacter* | 6-11 | 0.567 | 0.343 | 12-17 | 0.894 | 0.117 | 18-24 | 1 | 0 | 25-43 | 0.833 | 0.408 |
| *Bacillus* | 5-11 | 0.244 | 0.308 | 12-18 | 0.77 | 0.226 | 19-24 | 0.777 | 0.194 | 25-37 | 1 | 0 |
| *Bdellovibrio* | 7-11 | 0.712 | 0.188 | 12-16 | 0.923 | 0.104 | 17-21 | 0.96 | 0.0894 | 22-27 | 1 | 0 |
| *Beijerinckia* | 5-10 | 0.6 | 0.383 | 11-15 | 0.9 | 0.173 | 16-20 | 1 | 0 | 21-34 | 1 | 0 |
| *Bradyrhizobium* | 7-11 | 0.712 | 0.209 | 12-16 | 0.95 | 0.0684 | 17-21 | 1 | 0 | 22-29 | 1 | 0 |
| *Caulobacter* | 7-11 | 0.689 | 0.317 | 12-16 | 0.971 | 0.0638 | 17-21 | 0.92 | 0.178 | 22-29 | 1 | 0 |
| *Clostridium* | 7-13 | 0.122 | 0.19 | 14-20 | 0.439 | 0.255 | 21-26 | 0.891 | 0.174 | 27-40 | 0.857 | 0.377 |
| *Cyanobacterium* | 11-17 | 0.516 | 0.273 | 18-24 | 0.823 | 0.262 | 25-31 | 0.976 | 0.0629 | 32-43 | 1 | 0 |
| *Desulfotomaculum* | 6-10 | 0.133 | 0.217 | 11-15 | 0.617 | 0.0981 | 16-20 | 0.893 | 0.153 | 22-35 | 1 | 0 |
| *Desulfovibrio* | 6-9 | 0.254 | 0.176 | 10-13 | 0.647 | 0.109 | 14-17 | 0.85 | 0.191 | 18-24 | 1 | 0 |
| *Erwinia* | 5-10 | 0.513 | 0.484 | 11-16 | 0.826 | 0.0847 | 17-22 | 1 | 0 | 23-30 | 1 | 0 |
| *Frankia* | 5-8 | 0.107 | 0.214 | 9-13 | 0.729 | 0.219 | 14-18 | 1 | 0 | 19-23 | 0.95 | 0.111 |
| *Geobacter* | 6-9 | 0.502 | 0.413 | 10-13 | 0.697 | 0.0762 | 14-17 | 1 | 0 | 18-22 | 0.9 | 0.223 |
| *Klebsiella* | 5-10 | 0.548 | 0.325 | 11-15 | 0.913 | 0.123 | 16-20 | 0.93 | 0.109 | 21-26 | 1 | 0 |
| *Kocuria* | 7-12 | 0.516 | 0.449 | 13-18 | 0.944 | 0.136 | 19-24 | 0.875 | 0.209 | 25-34 | 1 | 0 |
| *Leuconostoc* | 8-12 | 0.243 | 0.265 | 13-17 | 0.682 | 0.171 | 18-22 | 0.971 | 0.0638 | 23-27 | 1 | 0 |
| *Mesorhizobium* | 6-10 | 0.86 | 0.167 | 11-15 | 0.937 | 0.0908 | 16-20 | 1 | 0 | 21-28 | 1 | 0 |
| *Methylococcus* | 7-10 | 0.5 | 0.408 | 11-15 | 0.853 | 0.123 | 16-19 | 1 | 0 | 20-24 | 1 | 0 |
| *Microbacterium* | 7-11 | 0.4 | 0.418 | 12-17 | 0.877 | 0.113 | 18-23 | 1 | 0 | 24-32 | 1 | 0 |
| *Micrococcus* | 8-13 | 0.549 | 0.389 | 14-19 | 0.933 | 0.0831 | 20-24 | 1 | 0 | 25-35 | 1 | 0 |
| *Myxococcus* | 6-11 | 0.222 | 0.186 | 12-17 | 0.733 | 0.188 | 18-23 | 0.979 | 0.051 | 24-34 | 1 | 0 |
| *Nitrobacter* | 6-10 | 0.671 | 0.393 | 11-15 | 1 | 0 | 16-20 | 1 | 0 | 21-40 | 1 | 0 |
| *Nitrosococcus* | 5-8 | 0.583 | 0.3 | 9-13 | 0.752 | 0.11 | 14-17 | 0.937 | 0.125 | 18-24 | 1 | 0 |
| *Nitrosomonas* | 6-11 | 0.546 | 0.328 | 12-17 | 0.88 | 0.117 | 18-22 | 0.933 | 0.149 | 23-52 | 1 | 0 |
| *Nitrosospira* | 5-9 | 0.253 | 0.347 | 10-14 | 0.956 | 0.0603 | 15-19 | 1 | 0 | 20-26 | 1 | 0 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Nocardia* | 6-10 | 0.283 | 0.298 | 11-16 | 0.767 | 0.183 | 17-21 | 0.98 | 0.0447 | 22-35 | 1 | 0 |
| *Nostoc* | 8-12 | 0.511 | 0.5 | 13-17 | 0.662 | 0.197 | 18-22 | 0.9 | 0.136 | 23-31 | 0.916 | 0.204 |
| *Oscillatoria* | 6-10 | 0.336 | 0.232 | 11-16 | 0.886 | 0.137 | 17-21 | 1 | 0 | 22-29 | 1 | 0 |
| *Pseudanabaena* | 6-12 | 0.236 | 0.29 | 13-19 | 0.927 | 0.0922 | 20-26 | 0.984 | 0.0419 | 27-38 | 1 | 0 |
| *Pseudomonas* | 6-9 | 0.602 | 0.225 | 10-14 | 0.807 | 0.155 | 15-19 | 0.981 | 0.0406 | 20-28 | 1 | 0 |
| *Pseudonocardia* | 7-13 | 0.553 | 0.375 | 14-19 | 1 | 0 | 20-24 | 1 | 0 | 25-32 | 1 | 0 |
| *Rhizobium* | 6-9 | 0.493 | 0.359 | 10-14 | 0.918 | 0.0784 | 15-18 | 0.958 | 0.0833 | 19-25 | 1 | 0 |
| *Rhodobacter* | 6-11 | 0.726 | 0.405 | 12-18 | 0.947 | 0.0899 | 19-24 | 1 | 0 | 25-38 | 1 | 0 |
| *Rickettsia* | 8-17 | 0.199 | 0.247 | 18-25 | 0.67 | 0.242 | 26-33 | 0.651 | 0.342 | 34-47 | 1 | 0 |
| *Shewanella* | 5-9 | 0.25 | 0.204 | 10-14 | 0.708 | 0.136 | 15-18 | 0.947 | 0.0611 | 19-27 | 1 | 0 |
| *Sinorhizobium* | 5-8 | 0.875 | 0.25 | 9-13 | 0.923 | 0.104 | 14-18 | 0.98 | 0.0447 | 19-24 | 1 | 0 |
| *Sphingomonas* | 6-11 | 0.56 | 0.347 | 12-17 | 0.883 | 0.204 | 18-24 | 0.972 | 0.068 | 25-33 | 1 | 0 |
| *Streptomyces* | 5-10 | 0.139 | 0.219 | 11-16 | 0.863 | 0.141 | 17-22 | 1 | 0 | 23-32 | 1 | 0 |
| *Variovorax* | 7-12 | 0.291 | 0.367 | 13-18 | 0.715 | 0.172 | 19-24 | 0.986 | 0.034 | 25-31 | 1 | 0 |
| *Xanthomonas* | 7-11 | 0.516 | 0.207 | 12-17 | 0.839 | 0.202 | 18-23 | 1 | 0 | 25-39 | 1 | 0 |
| AMF | 6-11 | 0.0654 | 0.106 | 12-18 | 0.502 | 0.284 | 19-25 | 0.678 | 0.386 | 26-35 | 0.785 | 0.393 |
| *Aspergillus* | 5-8 | 0.398 | 0.377 | 9-12 | 0.87 | 0.0933 | 13-16 | 0.873 | 0.148 | 17-26 | 0.75 | 0.5 |
| *Cenococcum* | 6-9 | 0.625 | 0.25 | 10-13 | 0.82 | 0.208 | 14-17 | 0.703 | 0.216 | 18-41 | 0.72 | 0.414 |
| *Cryptococcus* | 6-9 | 0.339 | 0.282 | 10-14 | 0.61 | 0.286 | 15-18 | 0.825 | 0.236 | 19-25 | 0.8 | 0.447 |
| *Mycosphaerella* | 6-9 | 0.816 | 0.137 | 10-13 | 0.83 | 0.05 | 14-17 | 0.895 | 0.125 | 18-25 | 1 | 0 |
| *Oidiodendron* | 6-8 | 0.555 | 0.509 | 9-12 | 0.794 | 0.151 | 13-16 | 0.958 | 0.0833 | 17-22 | 0.816 | 0.213 |
| *Phanerochaete* | 5-7 | 0.154 | 0.135 | 8-11 | 0.488 | 0.213 | 12-15 | 0.78 | 0.182 | 16-21 | 0.775 | 0.262 |
| *Scleroderma* | 6-10 | 0.474 | 0.245 | 11-15 | 0.679 | 0.164 | 16-20 | 0.971 | 0.0638 | 21-30 | 1 | 0 |
| *Sebacina* | 7-9 | 0.466 | 0.416 | 10-13 | 0.69 | 0.183 | 14-17 | 0.843 | 0.119 | 18-23 | 1 | 0 |

*Note*: Range represents a minimum and a maximum of ($\log_{10}$ (Inverse of *P* value score)) values per quartile. After the result from each genus test set was divided into quartiles, the range of ($\log_{10}$ (inverse *P* value score)) and the mean and the standard deviation of the proportions of the correct taxon group were calculated per quartile. The result was based on the rank probability scoring method. Data for Supplementary Table 10 and Supplementary Figure 5.C.

# SeSaMe PS Function: Functional Analysis of the Whole Metagenome Sequencing Data of the Arbuscular Mycorrhizal Fungi

Jee Eun Kang[1,*], Antonio Ciampi[2], Mohamed Hijri[1]

[1] *Institut de Recherche en Biologie Végétale, Département de Sciences Biologiques, Université de Montréal, QC, H1X 2B2, Canada*

[2] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC, H3A 1A2, Canada*

Jee Eun Kang orcid.org/0000-0003-2475-0474

Antonio Ciampi orcid.org/0000-0003-4838-8297

Mohamed Hijri orcid.org/0000-0001-6112-8372

E-mail: jennifer.kang@umontreal.ca (Kang JE)

Running Title: *Kang JE et al/ SeSaMe PS Function: Spore associated Symbiotic Microbes Position Specific Function*

| | Words | Figures | Tables | Sup. Figure | Sup. Tables |
|---|---|---|---|---|---|
| Total Counts | Approx. 6963 | 5 | 0 | 5 | 6 |

## 3.1 Abstract

In this article, we introduce a novel bioinformatics program- SeSaMe PS Function (Spore associated Symbiotic Microbes Position Specific Function)- for position-specific functional analysis of short sequences derived from metagenome sequencing data of the arbuscular mycorrhizal fungi. The unique advantage of the program lies in databases created based on genus-specific sequence properties derived from protein secondary structure, namely amino acid usages, codon usages, and codon contexts of three codon DNA 9-mers. SeSaMe PS Function searches a query sequence against reference sequence database, identifies three codon DNA 9-mers with structural roles, and creates comparative dataset containing the codon usage biases of the three codon DNA 9-mers from 54 bacterial and fungal genera. The program applies correlation Principal Component Analysis in conjunction with K-means clustering method to the comparative dataset. Three codon DNA 9-mers clustered as a sole member or with only a few members are often structurally and functionally distinctive sites that provide useful insights into important molecular interactions. The program provides a versatile means for studying functions of short sequences from metagenome sequencing and has a wide spectrum of applications.

**KEYWORDS:** SeSaMe PS Function; Spore associated Symbiotic Microbes Position Specific Function; Outlier; Metagenome

## 3.2 Introduction

Arbuscular mycorrhizal fungi (AMF) are plant root colonizing symbiotic microorganisms that promote plant growth and improve soil quality [1−3]. AMF increase the effectiveness of phytoremediation and improve crop yields in agroecosystems [1,4−10]. Despite the importance of AMF, their genetics is poorly understood, due in large part to their coenocytic multinucleate nature and strict symbiotic partnership with plants [11]. A number of studies reported strong evidence that AMF interact closely- tightly adhering to the surface or in the interior of mycelia and spores- or loosely with a myriad of microorganisms covering major bacterial and fungal taxa [6,12−16]. These microorganisms can be removed from AMF by using cocktails of antibiotics in axenic cultivation systems [17]. Yet, only few AMF taxa are able to be cured and cultivated *in-vitro,* and most successful isolates in such systems mainly belong to the genus *Rhizoglomus* [18]. Given that the majority of AMF have not been successfully cultured axenically, it is possible that AMF may be meta-organisms, inseparable from their bacterial and fungal partners.

Whole genome sequencing (WGS) of AMF taxa has been achieved exclusively from those grown *in-vitro*. Although they provide important insights into AMF genetics, they have limitations in serving as reference genome due to large intra and inter isolate genome variations [19,20]. Furthermore, sequence analysis of the WGS of AMF taxa grown *in vivo*, typically in a pot culture with a host plant, can be challenging because the sequencing data contain a large proportion of sequences belonging to AMF associated microorganisms; the WGS data of AMF represent a complex metagenome [16,21]. However, they provide invaluable information about the associated microbial community because a great majority of the associated microorganisms cannot be cultured in laboratory conditions. Taxonomic classification of the whole metagenome sequencing (WMS) data is essential for studying AMF genomics and their interactions with the associated microorganisms. We introduced the bioinformatics program- SeSaMe (Spore associated Symbiotic Microbes) - for taxonomic classification of the WMS of AMF [22]. In this article, we introduce a novel bioinformatics program- SeSaMe Position Specific Function (SeSaMe PS Function). It predicts important position-specific functional sites in a query sequence, based on amino acid usages, codon usages, and codon contexts of three codon DNA 9-mers derived from protein secondary structures extracted from Protein Data Bank (PDB) (rcsb.org) [23].

Recent studies have documented the multiple regulatory roles of codon usage and of codon context in transcription and translation (e.g., regulation of gene expression, diversification of gene products, translational efficiency and accuracy, and protein degradation efficiency) [24−30]. Several studies have emphasized the regulatory roles of codon usage and codon context of multiple consecutive codons [25,29,30]. In addition, synonymous codons are believed to be a key factor in determining the active folding state of a gene product in response to environmental changes. One recent study showed that a gene with multiple synonymous mutations produced a protein with increased tolerance to abiotic stresses [31]. Moreover, non-optimal codons serve specific roles in regulating circadian rhythms in response to changes of environmental conditions [32,33]. Therefore, codon usage and codon context must have been playing important roles in the adaptation of microorganisms to abiotic stresses [34,35]. We are beginning to scratch the surface of the regulatory roles of codon usage and codon context, and these studies appear to be just a tip of iceberg.

The main variable of the program- *trimer usage bias*- takes usages and contexts of both amino acids and nucleotides into consideration; it is the product of *amino acid usage* and *three codon usage* of a sequence variable called *Three codon DNA 9-mer*. Generally, *trimer usage bias* has a broad range of variations among taxonomic groups but low variations among microorganisms belonging to the same taxonomic group. *Trimer usage bias* reflects the important attributes of multiple consecutive codons. Codon composition- i.e., codon context of three consecutive codons- is an important determinant of properties of mRNA structure that plays key roles in transcription and translation. Codon usage is associated with pauses in translation and determines biochemical properties of gene products. Both of the attributes affect protein folding.

SeSaMe PS Function identifies *three codon DNA 9-mers* with structural roles in a query sequence, and creates comparative dataset based on their *trimer usage biases* that are retrieved from 54 genus-specific bias databases (**Figure 1**). SeSaMe PS Function applies correlation Principal Component Analysis (PCA) in conjunction with K-means clustering method (PCA-Kmeans) to the comparative dataset. It enables users to identify three codon DNA 9-mers with distinctive characteristics: outliers. Outliers are often important position-specific functional sites that provide useful insights into molecular interactions.

In this article, we analyzed one example sequence to demonstrate how to use the program for studying the structure and the function of a query sequence: one of the program's various applications. The program helped to identify the outliers with potentially important functions. Existing bioinformatics programs predicted that most of the outliers belonged to stem-loops, stems, and stem transitions in mRNA structures [36]. Some of the outliers were matched to elements that play roles in promotor regions or in cis-regulatory mechanisms [37−39]. Other bioinformatics programs predicted that the example sequence may bind to DNA/RNA [23,40]. These results suggest that the outliers may contribute to binding activities in undiscovered mechanisms that may have attributes similar to cis-regulatory mechanism.

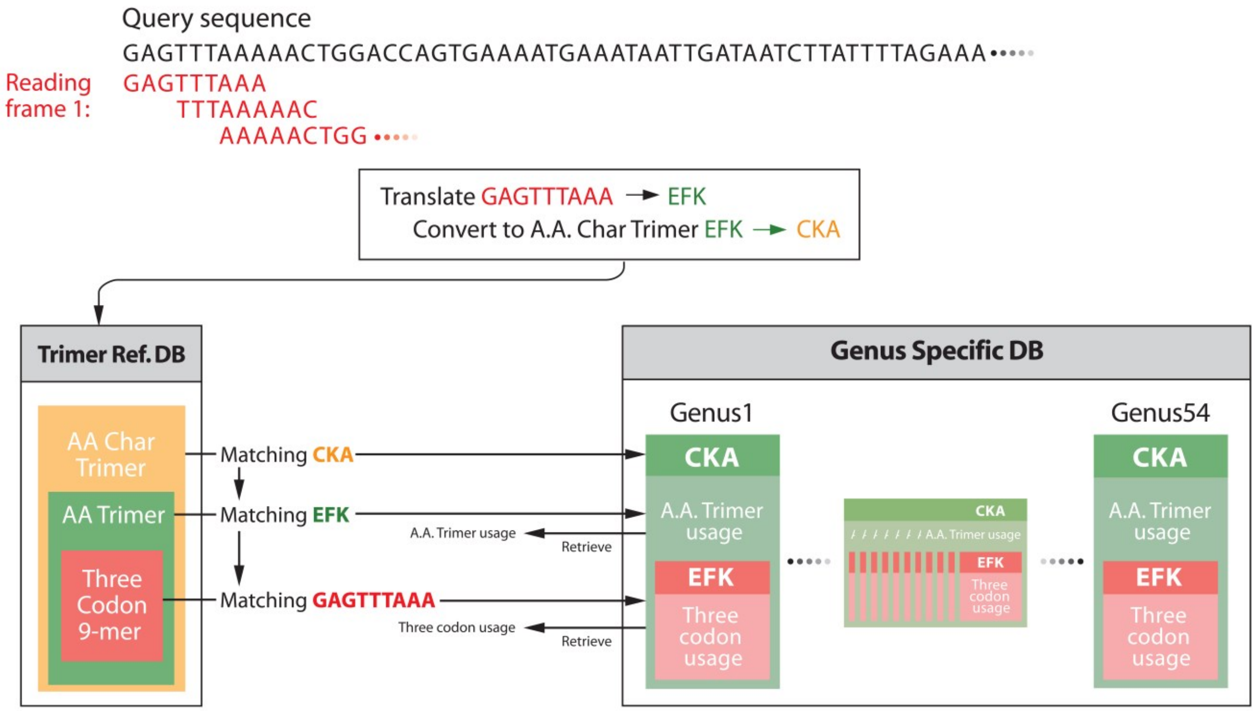


**Figure 1 Dynamic creation of comparative dataset per query sequence**

The program uses a query sequence to search matching *A.A. Char Trimers*, *A.A. Trimers*, and *Three Codon DNA 9-mers* in Trimer Ref. DB, and retrieves the *A.A. Trimer usages* of the matching *A.A. Trimers* and the *three codon usages* of the matching *Three Codon DNA 9-mers* from 54 Genus Specific DBs. It calculates the *trimer usage biases* of the matching *Three Codon DNA 9-mers*, and generates comparative dataset for the query sequence.

A majority of existing bioinformatics tools for position-specific sequence annotation rely on sequence alignments, which have low sensitivity toward hypervariable sequence motifs with flexible structures and various functions. Although they provide important information about a query sequence, their usage is limited to a particular set of motifs with known functions. In contrast, SeSaMe PS Function employs PCA to identify outliers based on internal structure of comparative dataset that contains usage information of structural units of a query sequence measured in 54 genera. Therefore, it may reveal important molecular interaction sites not only in known but also in undiscovered mechanisms. It has been only several decades since advances have been made in molecular biology. Therefore, it is believed that only a small fraction of mechanisms in biological system have been discovered. SeSaMe PS Function provides a useful tool for studying functions of short sequences from metagenome sequencing data. It is available for download free of charge at www.fungalsesame.org.

## 3.3 Methods

### 3.3.1 Database design and comparative dataset creation

The databases were originally created for the metagenome taxonomic classifier- SeSaMe, and then incorporated into SeSaMe PS Function [22]. While NCBI offered a large number of completely sequenced bacterial genomes, only a small number of fungal genomes were completely sequenced. The completely sequenced genomes of 444 bacteria and of 11 fungi, known to be present in soil, were downloaded and assigned into 45 bacterial and 9 fungal genera, respectively. CDS database per genus was created based on CDS lists provided by NCBI, JGI, or Tisserant et al. [19].

The program consists of two types of databases and a PCA-Kmeans method. 126,093 structure files were downloaded from PDB. 7674 amino acid trimers were selected among protein secondary structures from PDB, and then assigned to the sequence variable- *A.A. Trimer* in the trimer reference sequence database (Trimer Ref. DB) (**Figure 2**) [41−44]. Amino acid characteristic (*A.A. Char*) is defined as a group of amino acid(s) with similar property(s), and consists of 12 groups: A (Lysine (K), Arginine (R)), B (Histidine (H)), C (Aspartic acid (D), Glutamic acid (E)), D (Serine (S), Threonine (T)), E (Asparagine (N), Glutamine (Q)), F (Cysteine (C)), G (Glycine (G)), H (Proline (P)), I (Methionine (M)), J (Alanine (A), Isoleucine (I), Leucine (L), Valine (V)), K (Phenylalanine (F), Tryptophan (W), Tyrosine (Y)), and L (stop codons). Trimer Ref. DB consists of three sequence variables that form a three level hierarchy: amino acid characteristic trimer (*A.A. Char Trimer*), *A.A. Trimer*, and *Three Codon DNA 9-mer* (Figure 1).

Genus-specific usage bias database (Genus Specific DB) contains the numerical variables- *A.A. Trimer usage* of *A.A. Trimer* and *three codon usage* of *Three Codon DNA 9-mer*. The main numerical variable, *trimer usage bias*, is calculated by multiplying *A.A. Trimer usage* by *three codon usage*. There are 54 Genus Specific DBs where each Genus Specific DB consists of 1296 A.A. Trimer Usage Tables and 7674 Three Codon Usage Tables created based on the CDS database (Figure 2).

For each reading frame of a query sequence, the program uses a query sequence to search against Trimer Ref. DB, identifying matching *A.A. Char Trimers*, *A.A. Trimers*, and *Three Codon DNA 9-mers*. It retrieves the *trimer usage biases* of the matching *Three Codon DNA 9-mers* from 54 Genus Specific DBs, and creates a comparative dataset of 54 genera (Figure 1). The input matrix to the correlation PCA method is the comparative dataset with 54 genera in rows (observations) and the matching *Three Codon DNA 9-mers* in columns. The input matrix will be called hereafter Z (I x J).

### 3.3.2 Annotation for catalytic and allosteric sites

According to Catalytic Site Atlas and Allosteric Database, *A.A. Trimers* were divided into 4 subgroups based on the property of their second amino acid- catalytic site (CSA), allosteric site (ASD), both CSA and ASD (BothCA), and none of them (None) [45,46]. An *A.A. Trimer* in CSA, ASD, or BothCA groups was annotated with the list of functions of PDB molecules that contained the *A.A. Trimer*. This feature is for making inferences about functionality not of a query sequence but of its *A.A. Trimers*.

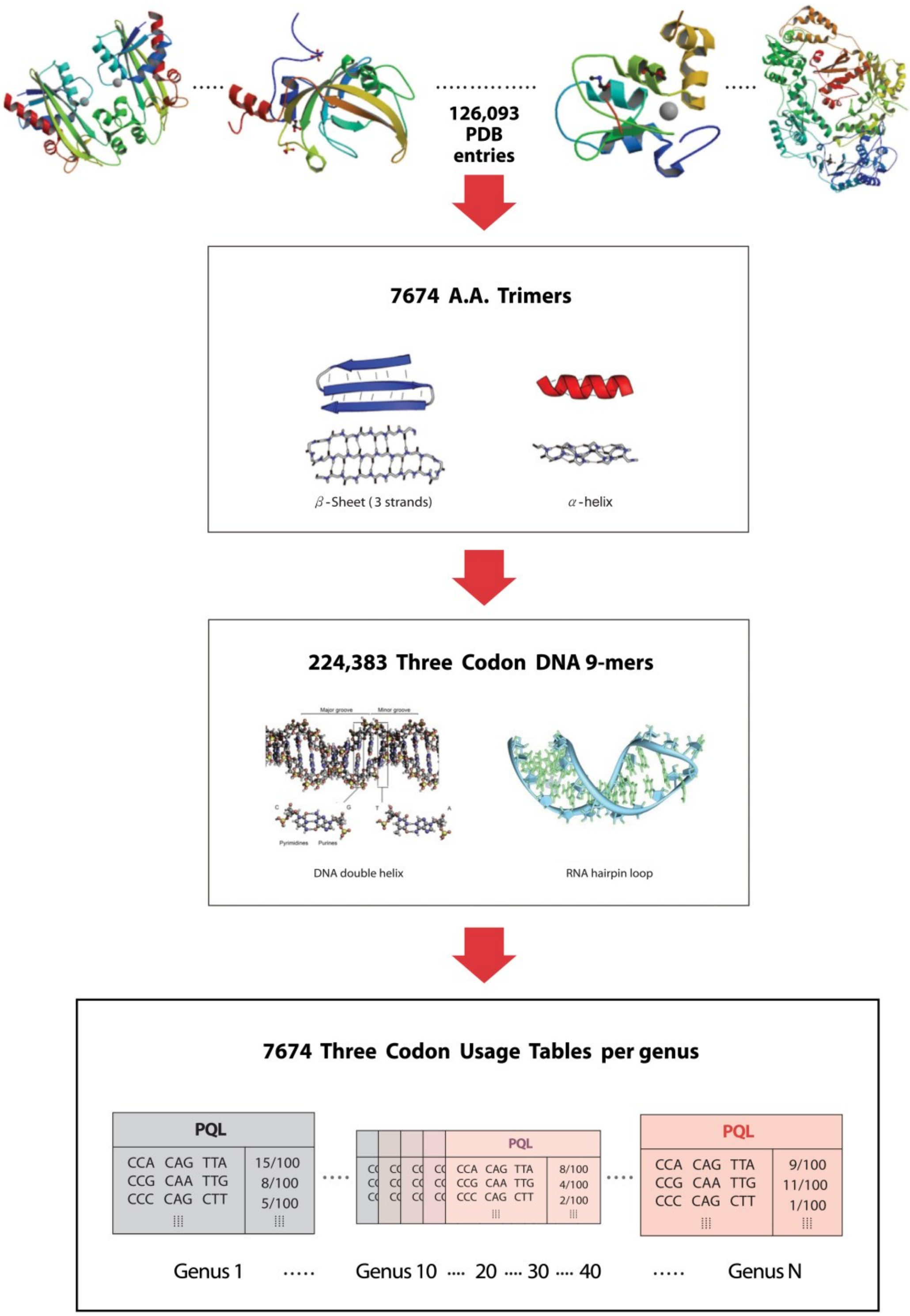


**Figure 2 Database Design**

A large number of PDB entry files were processed to extract 7674 *A.A. Trimers*- subunits of protein secondary structures. A table of *three codon usage* was created per *A.A. Trimer* and per genus in Genus Specific DB.

*Note*: The PDB IDs of the protein structures in the top of the figure from left are 2ZTI, 3DWH, 2VSL, and 3DRP. Their citations are included in the reference section [41–44]. Images of the protein secondary structures and the nucleotide structures in the first box and in the second box are from https://en.wikipedia.org/wiki/Alpha_helix and https://en.wikipedia.org/wiki/RNA, respectively.

### 3.3.3 Implementation of the correlation PCA-Kmeans method

The correlation PCA method was implemented based on the method reported by Abdi et al. (2010) [47], which provides important definitions and multiple examples to help readers understand the concepts underlying PCA [47]. Interpretation of the result from SeSaMe PS Function also relies on Abdi et al. (2010) because eigenvalue decomposition is mathematically closely related to singular value decomposition and has similar underlying concepts. Pearson's correlation method is applied to the centered Z and produces a correlation matrix X (J x J). Eigenvalue decomposition is applied to X and produces components. V is an eigenvector matrix with J x J dimensions and is also called a loading matrix.

#### 3.3.3.1 Loadings: elements of the loading eigenvector matrix V

The program calculates an eigenvector matrix V. Loading is defined as the element of V. V has matching *Three Codon DNA 9-mers* in rows and the same number of components in columns. The program examines loadings on components whose sum accounts for 80% of inertia (80% components) in addition to loadings on the first principal component and the second component (the First/Second components) [47]. The program creates two different input matrices based on V called L1 and L2. They have the same number of *Three Codon DNA 9-mers* in the rows. L1 has 80% components in columns while L2 has the First/Second components in columns. The program separately applies the K-means clustering method (default k = 13) to L1 and L2.

#### 3.3.3.2 Taxon scores of 54 genera in component spaces

The program calculates taxon scores of 54 genera observations. Taxon score matrix (I x J) results from multiplying centered Z by V. Inertia of a component is defined as a sum of squared taxon scores in corresponding component column [47]**.** The program creates two matrices based on taxon score matrix called T1 and T2. They have 54 genera observations in rows. T1 has 80% components in columns, while T2 has the First/Second components in columns. The program separately applies the K-means clustering method (default k = 10) to T1 and T2.

### 3.3.4 Program availability

The program was implemented in Java programming language (www.java.net, www.oracle.com (Java8)). We used the Pearson's correlation, the eigenvalue decomposition, and the K-means clustering methods in the Apache Commons Math3 library (3.3). The program requires the Apache Commons Math3 (3.3) and IO (2.4) libraries (www.apache.org). The program has been made to run on Linux/ Unix operating systems, packaged into an executable Java JAR file, and tested and confirmed to work on Linux system- CentOS Linux 7 (www.centos.org). The program that is being introduced in this article is version 1 and was implemented with the correlation PCA only. The program (version 2) was implemented both with the covariance PCA and with the correlation PCA. They have been used at the Biodiversity Center, Institut de Recherche en Biologie Végétale, Département de Sciences Biologiques, Université de Montréal. They are available for download free of charge at www.journal.com and www.fungalsesame.org. There are no restrictions for using the programs by academic or non-academic organizations as long as a user complies with the license agreement.

### 3.3.5 Input, output, and options

The program has a command-line interface. Input files should contain DNA sequence(s) in fasta format. It requires a command-line argument- input file path. SeSaMe PS Function produces three different types of outputs per query sequence. One is the standard PCA output: the sequence information of matching *Three Codon DNA 9-mers*, the percentage of an explained inertia by a component, and the contribution of an observation to a component [47]. Another is the loading cluster output with the loading information. *Three Codon DNA 9-mers* are annotated with subgroups- CSA/ ASD/ BothCA/ None and the functions of PDB molecules. The other is the genus cluster output with the taxon scores. It should be noted that the cluster result is different for every run, because the K-means clustering method in the Apache Commons Math library randomly chooses initial centers for multiple iterations to decrease chances of poor clustering.

SeSaMe PS Function version 1 and version 2 have an option to specify the k parameter in the K-means clustering method both for genus clusters and for loading clusters (e.g., 11_15). The program version 2 has an additional option called “auto”. If a user wants to run SeSaMe PS Function for a large number of query sequences with varying lengths, he can use the prefix “auto” to set the k parameter for loading clusters according to a simple equation: the number of matching *Three Codon DNA 9-mers* divided by a user specified number. For example, if the user gives the following option “auto_14_8”, it will automatically set one eighth of the number of matching *Three Codon DNA 9-mers* as the k parameter for loading clusters while it will set 14 as the k parameter for genus clusters. A suitable k value may vary widely depending on the length and the complexity of a query sequence. U*ser* can supply the option following the input file path (e.g., /home/input-file auto_14_8).

### 3.3.6 Demonstration of the program usage

#### 3.3.6.1 Selection of the example sequence

We selected 25 correctly predicted sequences out of 100 AMF CDS test sequences that were used for evaluating the accuracy of the metagenome taxonomic classifier, SeSaMe [22]. From 25 sequences, we selected one example sequence that had the largest number of *Three Codon DNA 9-mers* where AMF had the highest *trimer usage bias* among 54 genera.

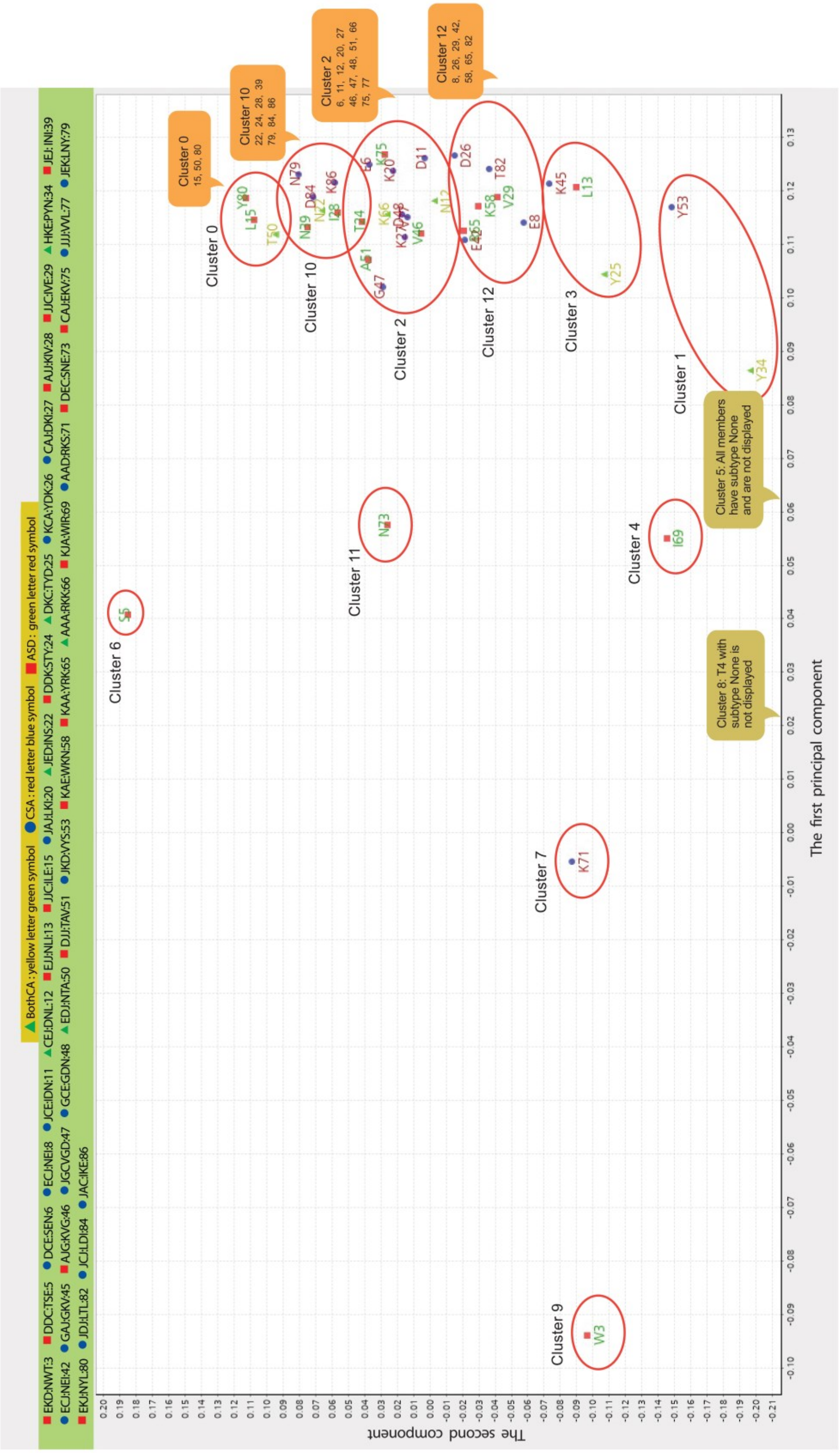


**Figure 3 Loading clusters of the example sequence**

The figure shows elements of the loading matrix V on the space of the First/Second components.

*Note*: Abbreviation: a name of *Three Codon DNA 9-mer* was abbreviated to the second amino acid of its *A.A. Trimer*. For example, *Three Codon DNA 9-mer* (AACTGGACC), encoding for the *A.A. Trimer* NWT, was abbreviated to W (Table S1). A digit next to the abbreviation indicates the order of its position in the example sequence. A digit in the colored box is abbreviation of *Three Codon DNA 9-mer*. For example, 22 under Cluster 10 in the box stands for ATTAATAGT that encodes for the *A.A. Trimer* INS whose order of the position is 22. CSA, ASD, BothCA, and None stand for catalytic site, allosteric site, both catalytic and allosteric site, and none of these sites, respectively.

The example query sequence is TGAGTTTAAAAACTGGACCAGTGAAAATGAAATAATTGATAATCTTATTTTAGAAATGCAATTAAAAATTAATAGTACATATGATAAAATAGTTGAATGGATACCATACAATCAGTTTATTAACATTAACGAAATAGGAAAAGTTGGTGATAATACTGCTGTATATTCAGCAATATGGAAAAATGGTCCACTATATTATAGAAAGAAATGGATAAGGAAATCCAATGAAAAAGTTGTATTAAATTACTTAACATTAGATATTAAGGAATT.

### 3.3.6.2 Outlier's unique pattern of the trimer usage bias and of the three codon usage

Landscape pattern is the comparison of 54 genera based either on the *trimer usage bias* or on the *three codon usage* of a *Three Codon DNA 9-mer*. It provides an accurate way to estimate the relative measure of the usage information across 54 genera. In this article, we abbreviate *Three Codon DNA 9-mer* according to the order of its position in DNA sequence and its *A.A. Trimer* (Table S1). For example*,* AATACTGCT is the 51st matching *Three Codon DNA 9-mer* and encodes for the amino acids NTA. Because the program is zero-based, its abbreviation is 50 NTA. Graphs showing the landscape patterns of *three codon usages* and of *trimer usage biases* retrieved from 54 genera were generated for 17 EMQ and 67 KKW and for 18 MQL and 3 NWT, respectively.

### 3.3.6.3 Comparison of the frequencies of a nucleotide among 13 loading clusters

We counted the frequencies of the nucleotide- adenine (A) in each of the individual *Three Codon DNA 9-mers* and applied a one-way ANOVA test to compare the means among 13 clusters. We repeated the same process for the nucleotides cytosine (C), guanine (G), and thymine (T).

### 3.3.6.4 Comparison between the trimer usage bias and the three codon usage in functional segment

We assigned matching *Three Codon DNA 9-mers* into functional segments (FSs) based on the loading clusters with 80% components and based on the prediction result of the protein secondary structure from a bioinformatics tool- SCRATCH [48].

We created two matrices per FS; one was based on the *three codon usage*, and the other was based on the *trimer usage bias*. Each matrix consisted of the usage information of the matching *Three Codon DNA 9-mers* retrieved from 54 genera; it had the *Three Codon DNA 9-mers* of an FS in rows and the 54 genera in columns. After centering each matrix, we applied Pearson's correlation to the matrix to yield a correlation matrix (I x I), and calculated the mean of the correlations per pair of taxonomic groups- Clostridia, Bacilli, Oscillatoriophycideae, Nostocales, Acidobacteria, Alphaproteobacteria, Betaproteobacteria, Deltaproteobacteria, Gammaproteobacteria, AMF, Agaricomycotina, and Pezizomycotina. From the mean of the correlations of a pair of genera belonging to the same taxonomic group in each FS, we calculated the mean and the standard deviation per taxonomic group. In the same way, we calculated the mean of the correlations for pairs of taxonomic groups- Firmicutes, Cyanobacteria, Proteobacteria, Actinobacteria, AMF, a group of 7 Dikarya, and *Phanerochaete* in each FS.

### 3.3.7 Results of the selected analysis

#### 3.3.7.1 Loading clusters

The example sequence had 270 bp. When we ran the metagenome taxonomic classifier- SeSaMe- with the example sequence, it had the highest trimer usage probability score in the 2nd reading frame translation [22]. It had 87 matching *Three Codon DNA 9-mers* in the 2nd reading frame translation. The PCA method applied to the comparative dataset showed that 51 components represented 80% components, while the First/Second components explained approximately 29% of total inertia.

The K-means clustering method (k = 13) applied to the loadings of 80% components identified outliers, 14 *Three Codon DNA 9-mers* in 12 clusters. Ten clusters had a sole member (50 NTA, 63 LYY, 72 KSN, 4 WTS, 69 WIR, 73 SNE, 24 STY, 30 VEW, 80 NYL, and 51 TAV) while two clusters had two members (33 IPY and 61 GPL and 39 INI and 86 IKE). One major cluster had 73 members.

Structural homology search in PDB and inference of DNA-binding residues in DRNApred suggested that the example sequence may be a DNA/RNA binding protein [23,40]. We used the outliers to search publicly available bioinformatics databases containing DNA motifs with known functions. RSAT indicated that the outlier and its adjacent *Three Codon DNA 9-mer* (4 WTS and 3 NWT) were matched to motifs involved in cis-regulatory mechanisms, one in the + strand and the other in the – strand [37]. BPROM (Prediction of bacterial promoters) predicted that the outliers 30 VEW and 33 IPY were promoter-related elements [38]. GPMiner indicated that three outliers (4 WTS, 33 IPY, and 61 GPL) were matched to statistically significant over-represented oligonucleotides in the promoter region [39]. RNA structure prediction tools predicted that most outliers formed stem-loops, stems, and transition routes to stem in mRNA structure of the example sequence (Figure S1) [36]. A large number of studies have documented stem-loop and stem structures in mRNAs as important regulatory sites and binding sites [49,50]. Considering that we are just beginning to understand the regulatory roles of codon usage and codon context, considerable portions of outliers and their adjacent *Three Codon DNA 9-mers* identified by the program may serve important roles in undiscovered mechanisms.

The loading clusters with the First/Second components based on the *trimer usage bias* are shown in Table S1. It should be noted that Table S1 indicates the *three codon usages* for comparison purpose, which will be discussed in another section. The loadings of *Three Codon DNA 9-mers* with the catalytic or with the allosteric site in the second amino acid were plotted on the space of the First/Second components (**Figure 3)**. A majority of *Three Codon DNA 9-mers* where Firmicutes, Cyanobacteria, *Rickettsia,* or AMF had the highest *three codon usage* were aggregately located on the far-right side (Figure 3). In contrast, those where Deltaproteobacteria, Gammaproteobacteria, or

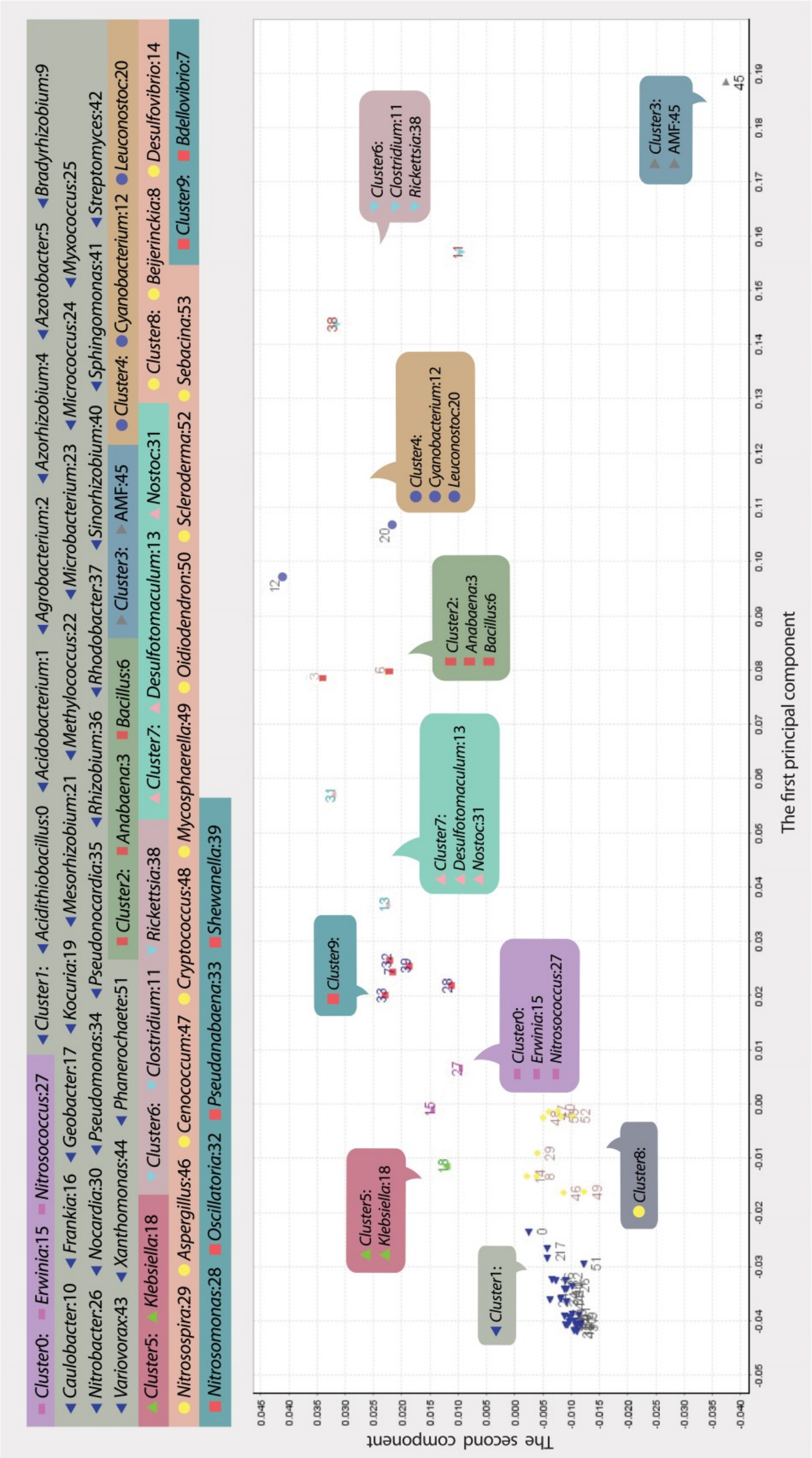


**Figure 4 Genus clusters of the example sequence**

Taxon scores of 54 genera are plotted on the space of the First/Second components.

Actinobacteria had the highest *three codon usage* were dispersed across the left side and the middle of the graph. For example, 3 NWT where *Kocuria* had the highest value was located on the far-left side (Table S1).

### 3.3.7.2 Genus clusters

The genus clusters based on 80% components indicated that genera with close phylogenetic relationships were assigned to the same cluster. In the scatter plot of taxon scorers on the space of First/Second components, Firmicutes, Cyanobacteria, *Rickettsia*, and AMF that frequently had high *trimer usage biases* were located on the right while most members of Actinobacteria and Proteobacteria (cluster 1) that frequently had low values were located on the far-left side (**Figure 4**).

### 3.3.7.3 Outlier's unique landscape pattern of the trimer usage bias and of the three codon usage

For each of the *Three Codon DNA 9-mers* in loading clusters with the First/Second components, we ranked 54 genera in order of decreasing *three codon usage*. We then, ranked the *Three Codon DNA 9-mers* in each subgroup (CSA/ ASD/ BothCA/ None) of the clusters based on a maximum of the *three codon usages* (Table S1). The mean of the maxima was 0.256. AMF, *Clostridium*, and *Rickettsia* frequently had the maximum.

Most of *Three Codon DNA 9-mers* in the major cluster demonstrated similar landscape patterns of the *three codon usage* and of the *trimer usage bias*. For example, 17-EMQ-GAAATGCAA and 18-MQL-ATGCAATTA had the frequently demonstrated landscape pattern (Figures S2 and S3**)**. Outliers had a unique landscape pattern; for example, genera belonging to Dikarya had a higher value than AMF both in 67-KKW-AAGAAATGG and in 3-NWT-AACTGGACC (**Figures 5** and S4).

### 3.3.7.4 Comparison of the frequencies of a nucleotide among 13 loading clusters

One-way ANOVA tests showed that the means of the frequencies of G and C in each of the individual *Three Codon DNA 9-mers* were significantly different among 13 clusters; F-statistics and p-value of A, T, G, and C among 13 clusters were 0.69 (0.76), 1.26 (0.26), 1.91 (0.047), and 3.09 (0.0014), respectively.

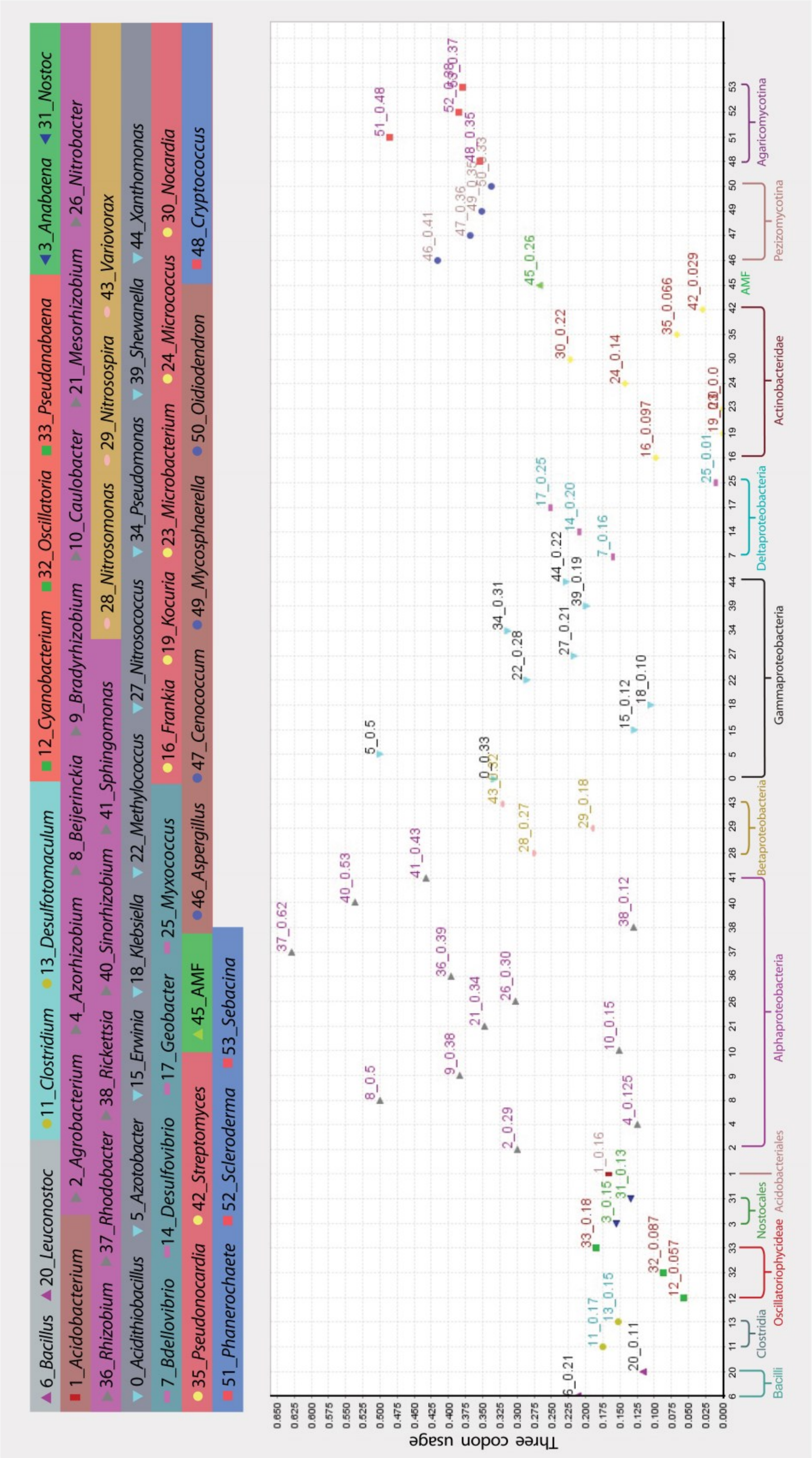


**Figure 5 Landscape pattern of the *three codon usage* of 67-AAK-KKW-AAGAAATGG**

54 genera are arranged into 13 taxonomic groups.

#### 3.3.7.5 Comparison between the trimer usage bias and the three codon usage in *FS*

We merged some of the outliers in 12 clusters according to their proximity in the example sequence, which produced 8 groups. The merged outliers were 50 NTA with 51 TAV, 33 IPY and 61 GPL with 63 LYY, and 69 WIR with 72 KSN and 73 SNE. This was done to simplify the analysis, and is not recommended for real case analyses. Examining the protein tertiary structure predicted by SCRATCH, we added another group (20 LKI), a member of alpha helix, which made a total of 9 groups [48]. We assigned 87 *Three Codon DNA 9-mers* into 9 FSs according to the outliers: FS1: 4 WTS (from *Three Codon DNA 9-mer* 0 – 12); FS2: 20 LKI (alpha helix1: 13 – 21); FS3: 24 STY (22 – 29); FS4: 30 VEW (30 – 32); FS5: 33 IPY, 61 GPL, 63 LYY (33 – 35, 52 – 65); FS6: 39 INI, 86 IKE (36 – 41, 82 – 86); FS7: 50 NTA, 51 TAV (42 – 51); FS8: 69 WIR, 72 KSN, 73 SNE (66 – 73); FS9: 80 NYL (alpha helix2: 74 – 81) (Figure S5).

Generally*,* the mean of the correlations of a pair of genera belonging to the same taxonomic group was the highest in each taxonomic group for all 9 FSs (Tables S2 and S3). Table S4 shows the mean and the standard deviation of 9 FSs calculated from the mean of the correlations of a pair of genera belonging to the same taxonomic group in a FS.

The mean of the correlations of a pair of taxonomic groups based on *three codon usage* (left) and the mean based on *trimer usage bias* (right) are shown in Table S3. Most of them had strong correlations in both alpha helices – FS2 and FS9. This may suggest that roles of amino acids and of codons in alpha helices may be relatively more conserved across taxonomic groups due to functional and structural constraints compared to those in random coils and loops of which flexible structures are equipped for a variety of functions.

#### 3.3.7.6 Comparable properties of 25 selected sequences in AMF CDS test set

In order to show that the program provided outliers of *Three Codon DNA 9-mers* in loading clusters based on 80% components not only in the example sequence but also in all 25 sequences, we included the cluster results of 5 additional query sequences. Genus clusters and loading clusters of the sequences are shown in Supplementary Tables 5 and 6, respectively. The early diverged bacteria and AMF were often clustered as a sole member or with each other. A great majority of *Three Codon DNA 9-mers* were grouped together into one major cluster, while outliers were clustered as a sole member or with only one other member.

## 3.4 Future work

Recent studies have documented that long non-coding RNAs (lncRNAs) play important roles in various cellular processes [51]. Because a large number of lncRNAs contain putative ORF, it is challenging to distinguish them from protein CDS [51]. They have been intensively studied only in mammalian species and other model organisms. AMF CDS list, presumably created based on results from a number of gene prediction programs, may contain lncRNAs. In future, we may make two types of sequence databases, protein CDS and lncRNA, and take a different approach depending on whether a query sequence is classified into lncRNA or into protein CDS.

Recent studies have documented that codon usage and mRNA structure regulate protein folding [25,26,28,30]. For example, some studies showed association between rare codons or double stranded mRNA structures and a decrease of translational speed [26,30]. Other studies have documented relationships between protein secondary structure and mRNA structure; double stranded mRNA regions tend to have an association with alpha helix and beta-strand while single stranded mRNA regions tend to have an association with random coils [52,53]. However, the roles of the codons involved in these rules may vary widely across taxonomic groups. Furthermore, while we need defined structures across various taxa, they are mostly from a small number of model organisms. Therefore, it is challenging to study associations between mRNA structures and their corresponding protein structures in metagenome sequencing data. We may be able to improve SeSaMe PS Function by incorporating a new feature that predicts mRNA single and double stranded regions in a query sequence.

## 3.5 Authors' contributions

KJE designed the program and implemented it using Java programming language. CA gave advice on what needs to be included in result of the correlation PCA-Kmeans method in terms of statistics, and helped to draft the manuscript. HM gave advice on developing the method, and helped to draft the manuscript. All authors read and approved the final manuscript.

## 3.6 Competing interests

The authors have declared no competing interests.

## 3.7 Acknowledgements

The authors gratefully acknowledge AFE (Aide financière aux études), FESP (Faculté des études supérieures et postdoctorales de l'UdeM), and IRBV (*Institut de Recherche en Biologie Végétale de l'Université de Montréal)* for awarding scholarships to Jee Eun Kang. We thank David Morse for editing and commenting on the manuscript.

## 3.8 References

[1] Hijri M. Analysis of a large dataset of mycorrhiza inoculation field trials on potato shows highly significant increases in yield. Mycorrhiza 2016; 26: 209–14.

[2] Roy-Bolduc A, Hijri M. The use of mycorrhizae to enhance phosphorus uptake: a way out the phosphorus crisis. J Biofertil Biopestic 2011; 2: 104.

[3] Zarik L, Meddich A, Hijri M, Hafidi M, Ouhammou A, Ouahmane L, et al. Use of arbuscular

mycorrhizal fungi to improve the drought tolerance of *Cupressus Atlantica* G. C R Biol 2016; 339: 185–96.

[4] Chanda D, Sharma GD, Jha DK, Hijri M. Associations of arbuscular mycorrhizal (AM) fungi in the phytoremediation of trace metal (TM) contaminated soils. J Res Biol 2014; 4: 1247−63.

[5] Lahlali R, Hijri M. Screening, identification and evaluation of potential biocontrol fungal endophytes against *Rhizoctonia Solani* AG3 on potato plants. FEMS Microbiol Lett 2010; 311: 152–9.

[6] Iffis B, St-Arnaud M, Hijri M. Bacteria associated with arbuscular mycorrhizal fungi within roots of plants growing in a soil highly contaminated with aliphatic and aromatic petroleum hydrocarbons. FEMS Microbiol Lett 2014; 358: 44–54.

[7] Iffis B, St-Arnaud M, Hijri M. Petroleum hydrocarbon contamination, plant identity and arbuscular mycorrhizal fungal (AMF) community determine assemblages of the AMF spore-associated microbes. Environ Microbiol 2016; 18: 2689–704.

[8] Hassan SE, St-Arnaud M, Labreque M, Hijri M. Phytoremediation: biotechnological procedures involving plants and arbuscular mycorrhizal fungi. In: Thangadurai D, Busso CA, Hijri M, editors. Mycorrhizal biotechnology. Enfield: Science Publishers; 2010, p. 152−77.

[9] Hassan SE, Hijri M, St-Arnaud M. Effect of arbuscular mycorrhizal fungi on trace metal uptake by sunflower plants grown on cadmium contaminated soil. N Biotechnol 2013; 30: 780−7.

[10] Hassan SE, Bell TH, Stefani FOP, Denis D, Hijri M, St-Arnaud M. Contrasting the community structure of arbuscular mycorrhizal fungi from hydrocarbon-contaminated and uncontaminated soils following willow (*Salix Spp.* L.) planting. PLoS One 2014; 9: e102838.

[11] Marleau J, Dalpe Y, St-Arnaud M, Hijri M. Spore development and nuclear inheritance in arbuscular mycorrhizal fungi. BMC Evol Biol 2011; 11: 51.

[12] Hijri M, Redecker D, Petetot JAM-C, Voigt K, Wöstemeyer J, Sanders IR. Identification and isolation of two ascomycete fungi from spores of the arbuscular mycorrhizal fungus. Appl Environ Microbiol 2002; 68: 4567−73.

[13] Cruz AF, Ishii T. Arbuscular mycorrhizal fungal spores host bacteria that affect nutrient biodynamics and biocontrol of soil-borne plant pathogens. Biol Open 2012; 1: 52–7.

[14] Jargeat P, Cosseau C, Ola'h B, Jauneau A, Bonfante P, Batut J, Becard G. Isolation, free-living capacities, and genome structure of "*Candidatus* Glomeribacter gigasporarum", the endocellular bacterium of the mycorrhizal fungus *Gigaspora margarita*. J Bacteriol 2004; 186: 6876–84.

[15] Gulbis N, Robinson-Boyer L, Robinson G. Studying the microbiome of AMF cultivated in vitro. Asp Appl Biol 2013; 120: 71–6.

[16] Agnolucci M, Battini F, Cristani C, Giovannetti M. Diverse bacterial communities are recruited on spores of different arbuscular mycorrhizal fungal isolates. Biol Fertil Soils 2015; 51: 379–89.

[17] Bécard G, Fortin JA. Early events of vesicular-arbuscular mycorrhiza formation on Ri T-DNA transformed roots. New Phytol 1988; 108: 211−8.

[18] Declerck S, Seguin S, Dalpe Y. The monoxenic culture of arbuscular mycorrhizal fungi as a tool for germplasm collections. In: Declerck S, Strullu DG, Fortin JA, editors. In vitro culture of

mycorrhizas. Berlin Heidelberg: Springer-Verlag; 2005, p. 17–30.

[19] Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, et al. Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. Proc Natl Acad Sci U S A 2013; 110: 20117–22.

[20] Boon E, Halary S, Bapteste E, Hijri M. Studying genome heterogeneity within the arbuscular mycorrhizal fungal cytoplasm. Genome Biol Evol 2015; 7: 505–21.

[21] Lecomte J, St-Arnaud M, Hijri M. Isolation and identification of soil bacteria growing at the expense of arbuscular mycorrhizal fungi. FEMS Microbiol Lett 2011; 317: 43–51.

[22] Kang JE, Ciampi A, Hijri M. SeSaMe: Metagenome Sequence Classification of Arbuscular Mycorrhizal Fungi Associated Microorganisms. Genomics Proteomics Bioinformatics 2018 In Press.

[23] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000; 28: 235−42.

[24] Bartoszewski R, Króliczewski J, Piotrowski A, Jasiecka AJ, Bartoszewska S, Vecchio-Pagan B, et al. Codon bias and the folding dynamics of the cystic fibrosis transmembrane conductance regulator. Cell Mol Biol Lett 2016; 21: 23.

[25] Del Campo C, Bartholomäus A, Fedyunin I, Ignatova Z. Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. PLoS Genet 2015; 11: e1005613.

[26] Costafreda IM, Pérez-Rodriguez FJ, D’Andrea L, Guix S, Ribes E, Bosch A, et al. Hepatitis A virus adaptation to cellular shutoff is driven by dynamic adjustments of codon usage and results in the selection of populations with altered capsids. J Virol 2014; 88: 5029–41.

[27] Komar AA. The yin and yang of codon usage. Hum Mol Genet 2016; 25: R77–85.

[28] Zhao F, Yu C, Liu Y. Codon usage regulates protein structure and function by affecting translation elongation speed in *Drosophila* cells. Nucleic Acids Res 2017; 45: 8484–92.

[29] McCarthy C, Carrea A, Diambra L. Bicodon bias can determine the role of synonymous SNPs in human diseases. BMC Genomics 2017; 18: 227.

[30] Yang J. Does mRNA structure contain genetic information for regulating co-translational protein folding? Zool Res 2017; 38: 36–43.

[31] Kashiwagi A, Sugawara R, Sano-Tsushima F, Kumagai T, Yomo T. Contribution of silent mutations to thermal adaptation of RNA bacteriophage Qß. J Virol 2014; 88: 11459−68.

[32] Xu Y, Ma P, Shah P, Rokas A, Liu Y, Johnson CH. Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. Nature 2013; 495: 116–20.

[33] Zhou M, Guo J, Cha J, Chae M, Chen S, Barral JM, et al. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature 2013; 495: 111–5.

[34] Su Y, Jiang XZ, Wu WP, Wang MM, Hamid MI, Xiang MC, et al. Genomic, transcriptomic and proteomic analysis provide insights into the cold adaptation mechanism of the obligate psychrophilic fungus *Mrakia Psychrophila*. G3 2016; 6: 3603–13.

[35] Sanjukta R, Farooqi MS, Sharma N, Rai A, Mishra DC, Singh DP. Trends in the codon usage patterns of *Chromohalobacter Salexigens* genes. Bioinformation 2012; 8: 1087–95.

[36] Bellaousov S, Reuter JS, Seetin MG, Mathews DH. RNAstructure: web servers for RNA secondary structure prediction and analysis. Nucleic Acids Res 2013; 41: W471−4.

[37] Defrance M, van Helden J. Info-gibbs: a motif discovery algorithm that directly optimizes information content during sampling. Bioinformatics 2009; 25: 2715−22.

[38] Solovyev V, Salamov A. Automatic annotation of microbial genomes and metagenomic sequences. In: Li RW, editor. Metagenomics and its applications in agriculture, biomedicine and environmental studies. New York: Nova Science Publishers; 2011, p. 61−78.

[39] Lee TY, Chang WC, Hsu JB, Chang TH, Shien DM. GPMiner: an integrated system for mining combinatorial cis-regulatory elements in mammalian gene group. BMC Genomics 2012; 13 Suppl 1: S3.

[40] Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res 2017; 2; e84.

[41] Lokanath NK, Pampa KJ, Takio K, Kunishima N. Structures of dimeric nonstandard nucleotide triphosphate pyrophosphatase from *Pyrococcus horikoshii* OT3: functional significance of interprotomer conformational changes. J Mol Biol 2008; 375: 1013–25.

[42] Qian C, Li S, Jakoncic J, Zeng L, Walsh MJ, Zhou MM. Structure and hemimethylated CpG binding of the SRA domain from human UHRF1. J Biol Chem 2008; 283: 34490–4.

[43] Nikolovska-Coleska Z, Meagher JL, Jiang S, Yang CY, Qiu S, Roller PP, et al. Interaction of a cyclic, bivalent Smac mimetic with the X-linked inhibitor of apoptosis protein. Biochemistry 2008; 47: 9811.

[44] Tucker TJ, Sisko JT, Tynebor RM, Williams TM, Felock PJ, Flynn JA, et al. Discovery of 3-{5-[(6-Amino-1H-pyrazolo[3,4-b]pyridine-3-yl)methoxy]-2-chlorophenoxy}-5-chlorobenzonitrile (MK-4965): a potent, orally bioavailable HIV-1 non-nucleoside reverse transcriptase inhibitor with improved potency against key mutant viruses. J Med Chem 2008; 51: 6503–11.

[45] Furnham N, Holliday GL, de Beer TA, Jacobsen JO, Pearson WR, Thornton JM. The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res 2014; 42: D485−9.

[46] Huang Z, Zhu L, Cao Y, Wu G, Liu X, Chen Y, et al. ASD: a comprehensive database of allosteric proteins and modulators. Nucleic Acids Res 2011; 39: D663−9.

[47] Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat 2010; 2: 433–59.

[48] Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005; 33: W72–6.

[49] D'Souza DJ, Kool ET. Strong binding of single-stranded DNA by stem-loop oligonucleotides. J Biomol Struct Dyn 1992; 10: 141−52.

[50] Tan D, Marzluff WF, Dominski Z, Tong L. Structure of histone mRNA stem-loop, human stem-loop

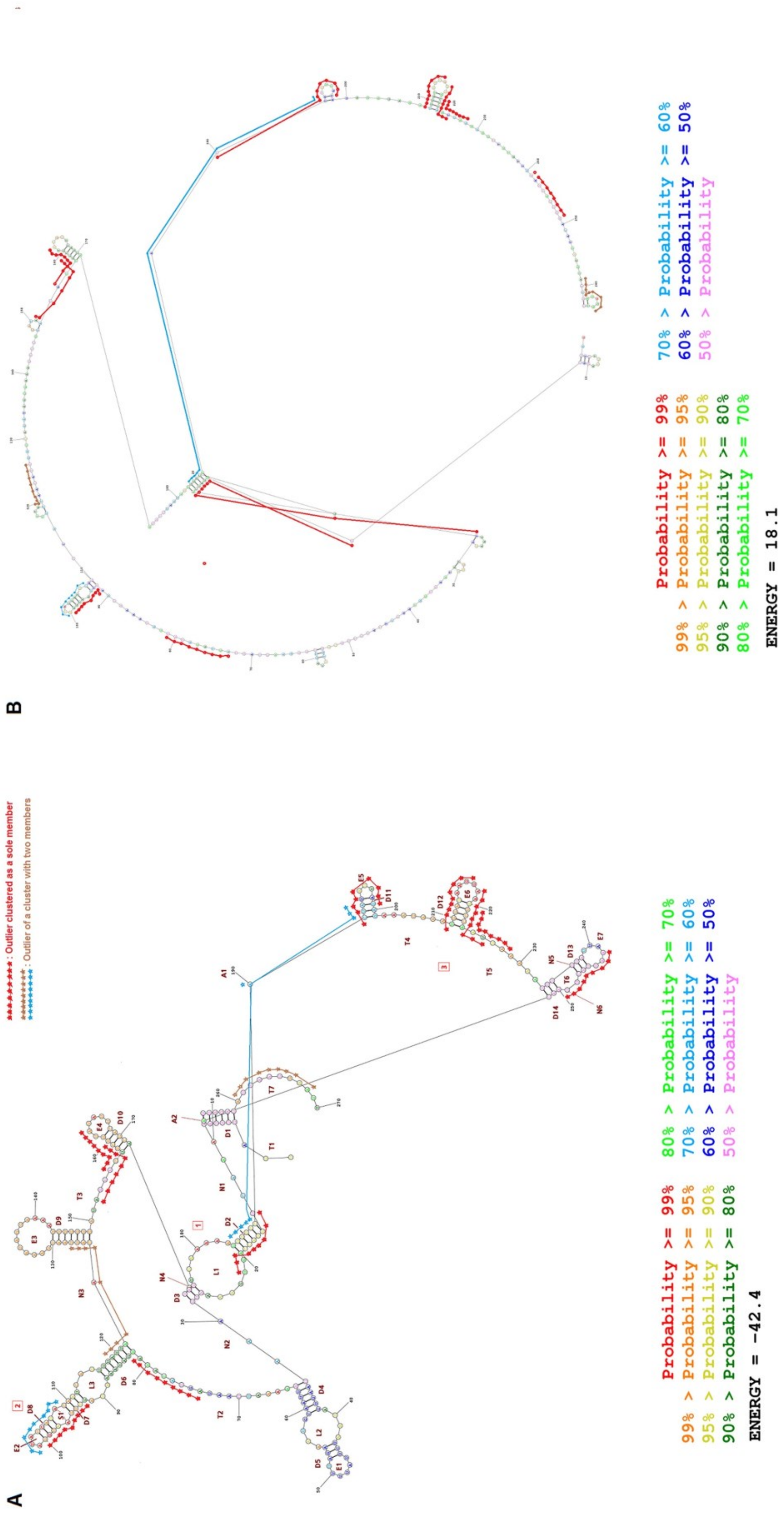
binding protein and 3'hExo ternary complex. Science 2013; 339: 318−21.

[51] Achawanantakun R, Chen J, Sun Y, Zhang Y. LncRNA-ID: Long non-coding RNA IDentification using balanced random forests. Bioinformatics 2015; 31: 3897-905.

[52] Jia MW, Luo LF, Liu CQ. Statistical correlation between protein secondary structure and messenger RNA stem-loop structure. Biopolymers 2004; 73: 16–26

[53] Zhang J, Gu BH, Peng SL, Liu CQ. Distributions of triplet codons in messenger RNA secondary structures. Zool Res 1998; 19: 350–8.

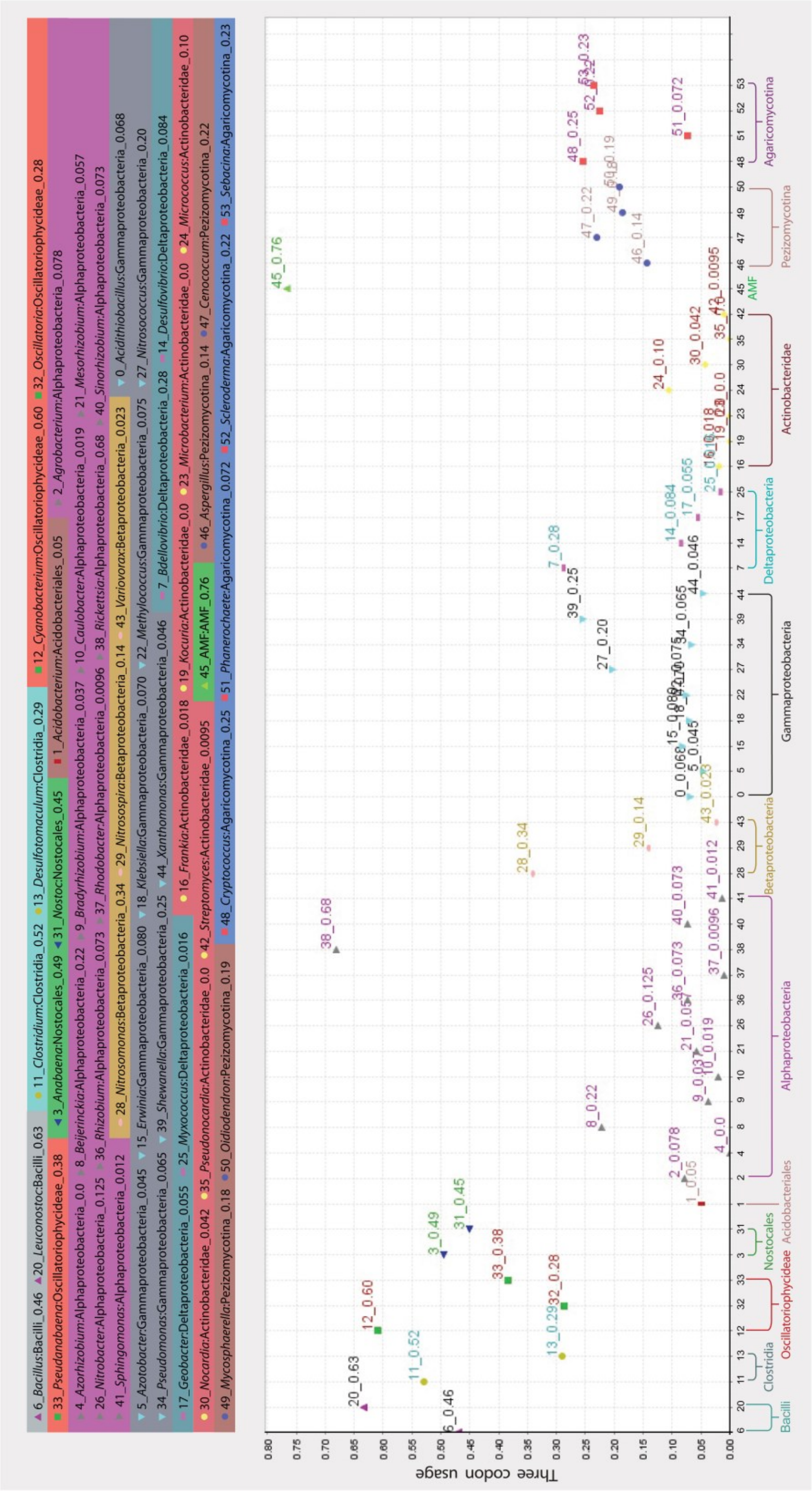
## 3.9 Supplementary Materials



**Supplementary Figure 1 Outliers in the predicted mRNA secondary structures**

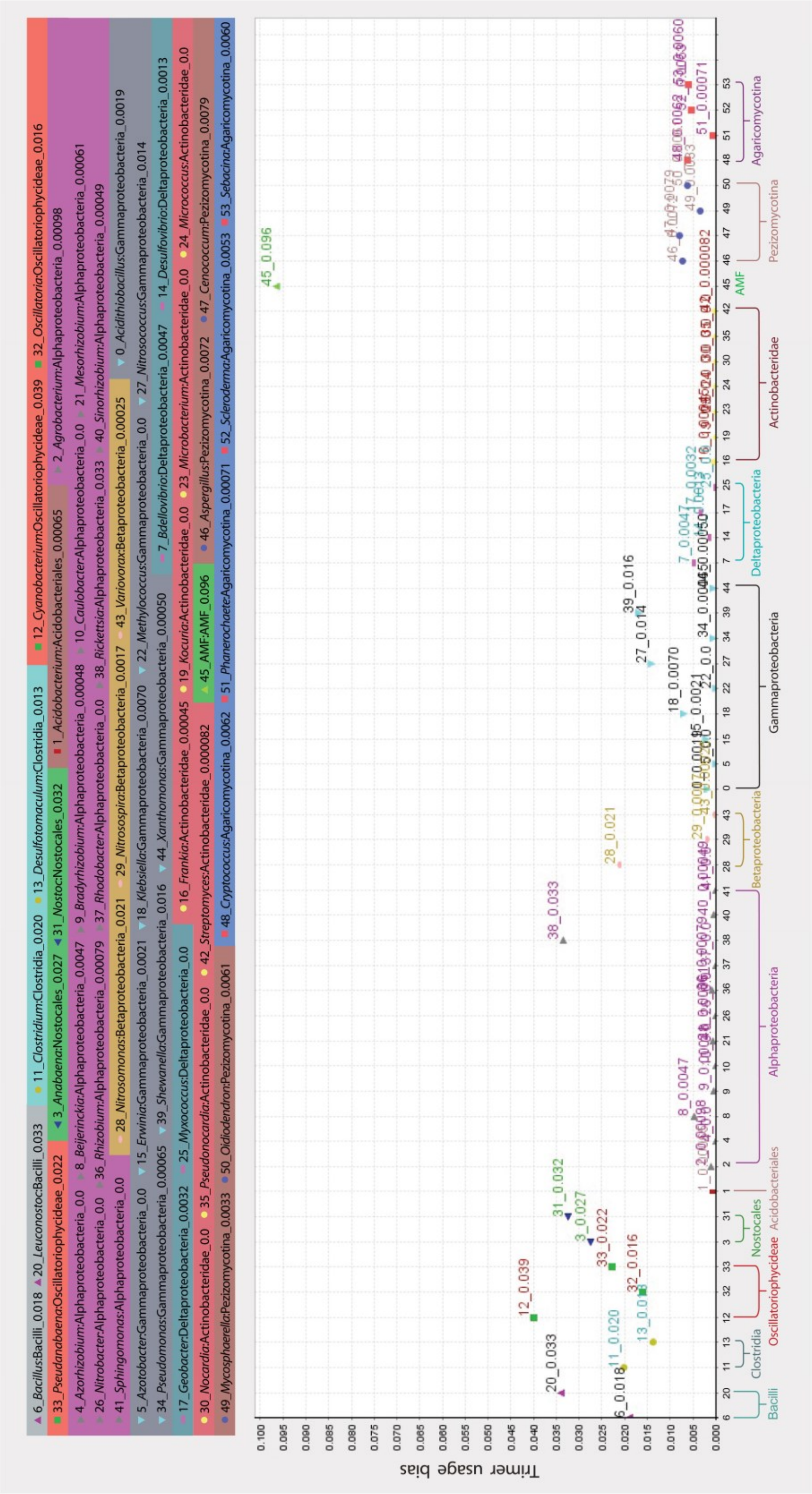
The structures were generated by the bioinformatics program- RNAstructure (https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html).

**A.** The example DNA sequence with T replaced with U was submitted as RNA. The secondary structure was predicted based on the algorithm called Fold. **B.** The example DNA sequence was submitted as DNA. The secondary structure was predicted based on the algorithm called MaxExpect.



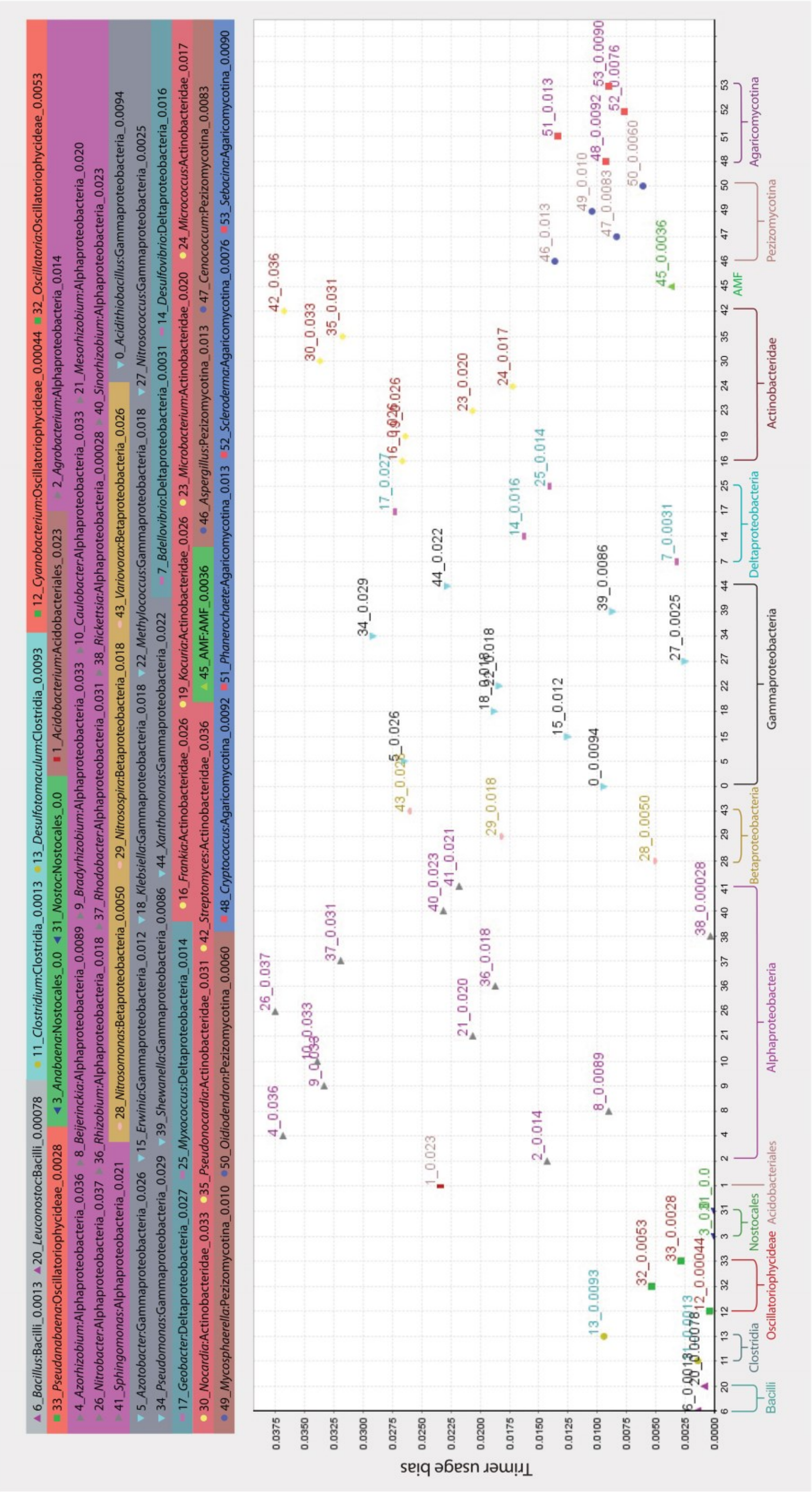
**Supplementary Figure 2 Landscape pattern of the *three codon usage* of 17-CIE-EMQ-GAAATGCAA**

54 genera are arranged into 13 taxonomic groups.



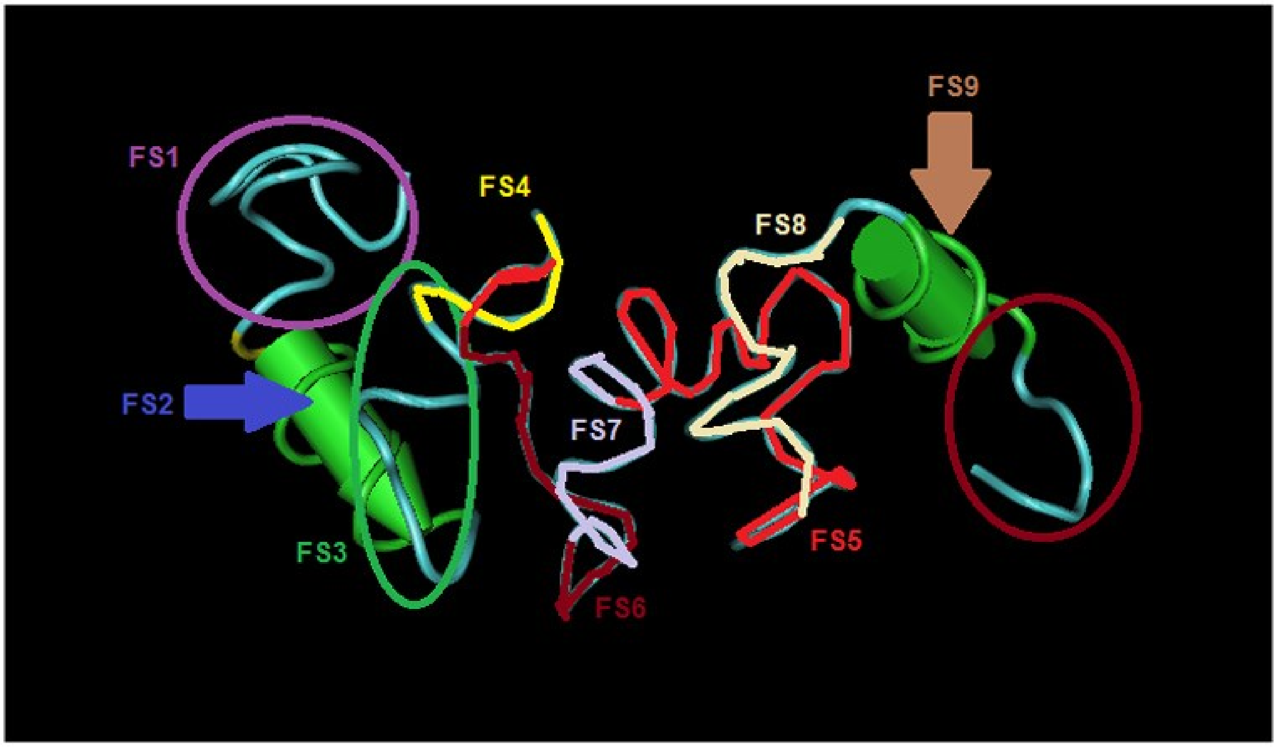
**Supplementary Figure 3 Landscape pattern of the *trimer usage bias* of 18-IEJ-MQL-ATGCAATTA**

54 genera are arranged into 13 taxonomic groups.



**Supplementary Figure 4 Landscape pattern of the *trimer usage bias* of 3-EKD-NWT-AACTGGACC**

54 genera are arranged into 13 taxonomic groups.



**Supplementary Figure 5 FSs of the predicted protein tertiary structure**

The structure was predicted by the bioinformatics program- SCRATCH (http://scratch.proteomics.ics.uci.edu/). The PDB file format was converted to Cn3D format by another bioinformatics program- Vast (https://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html). The 3-dimensional structure was viewed by Cn3D (https://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml).

### Supplementary Table S1 Loading clusters according to the First/Second components

| cluster | type | rank | three_genus | max | char_aa | three_codon |
|---|---|---|---|---|---|---|
| 0 | ASD | 0 | Cyanobacterium | 0.28 | 15_JJC:ILE | 15_ATTTTAGAA |
| 0 | ASD | 1 | Rickettsia | 0.07 | 80_EKJ:NYL | 80_AATTACTTA |
| 0 | Both | 0 | Cyanobacterium | 0.2 | 50_EDJ:NTA | 50_AATACTGCT |
| 0 | None | 1 | Klebsiella | 0.29 | 1_KAE:FKN | 1_TTTAAAAAC |
| 0 | None | 0 | Cyanobacterium | 0.36 | 16_JCI:LEM | 16_TTAGAAATG |
| 0 | None | 2 | Pseudanabaena | 0.27 | 36_EEK:NQF | 36_AATCAGTTT |
| 0 | None | 4 | Bacillus | 0.06 | 81_KJD:YLT | 81_TACTTAACA |
| 0 | None | 3 | Pseudanabaena | 0.16 | 85_CJA:DIK | 85_GATATTAAG |
| 1 | Both | 0 | AMF | 0.14 | 34_HKE:PYN | 34_CCATACAAT |
| 1 | CSA | 0 | AMF | 0.16 | 53_JKD:VYS | 53_GTATATTCA |
| 1 | None | 1 | AMF | 0.26 | 54_KDJ:YSA | 54_TATTCAGCA |
| 1 | None | 2 | AMF | 0.11 | 55_DJJ:SAI | 55_TCAGCAATA |
| 1 | None | 0 | Clostridium | 0.38 | 57_JKA:IWK | 57_ATATGGAAA |
| 2 | ASD | 1 | Leuconostoc | 0.25 | 46_AJG:KVG | 46_AAAGTTGGT |
| 2 | ASD | 2 | Clostridium | 0.07 | 51_DJJ:TAV | 51_ACTGCTGTA |
| 2 | ASD | 0 | AMF | 0.34 | 75_CAJ:EKV | 75_GAAAAAGTT |
| 2 | Both | 1 | Rickettsia | 0.12 | 12_CEJ:DNL | 12_GATAATCTT |
| 2 | Both | 0 | Clostridium | 0.16 | 66_AAA:RKK | 66_AGAAAGAAA |
| 2 | CSA | 0 | Leuconostoc | 0.45 | 11_JCE:IDN | 11_ATTGATAAT |
| 2 | CSA | 5 | Cyanobacterium | 0.22 | 20_JAJ:LKI | 20_TTAAAAATT |
| 2 | CSA | 2 | Rickettsia | 0.38 | 27_CAJ:DKI | 27_GATAAAATA |
| 2 | CSA | 3 | Leuconostoc | 0.27 | 47_JGC:VGD | 47_GTTGGTGAT |
| 2 | CSA | 1 | Rickettsia | 0.43 | 48_GCE:GDN | 48_GGTGATAAT |
| 2 | CSA | 4 | Clostridium | 0.26 | 6_DCE:SEN | 6_AGTGAAAAT |
| 2 | CSA | 6 | Bacillus | 0.09 | 77_JJJ:VVL | 77_GTTGTATTA |
| 2 | None | 9 | Bacillus | 0.09 | 14_JJJ:LIL | 14_CTTATTTTA |
| 2 | None | 2 | Cyanobacterium | 0.35 | 21_AJE:KIN | 21_AAAATTAAT |
| 2 | None | 3 | Rickettsia | 0.34 | 49_CED:DNT | 49_GATAATACT |
| 2 | None | 8 | Clostridium | 0.11 | 52_JJK:AVY | 52_GCTGTATAT |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2 | None | 7 | Rickettsia | 0.13 | 63_JKK:LYY | 63_CTATATTAT |
| 2 | None | 0 | Clostridium | 0.61 | 7_CEC:ENE | 7_GAAAATGAA |
| 2 | None | 1 | Clostridium | 0.53 | 74_ECA:NEK | 74_AATGAAAAA |
| 2 | None | 6 | AMF | 0.15 | 76_AJJ:KVV | 76_AAAGTTGTA |
| 2 | None | 4 | Clostridium | 0.18 | 78_JJE:VLN | 78_GTATTAAAT |
| 2 | None | 5 | Clostridium | 0.15 | 83_DJC:TLD | 83_ACATTAGAT |
| 3 | ASD | 0 | AMF | 0.09 | 13_EJJ:NLI | 13_AATCTTATT |
| 3 | Both | 0 | AMF | 0.27 | 25_DKC:TYD | 25_ACATATGAT |
| 3 | CSA | 0 | AMF | 0.2 | 45_GAJ:GKV | 45_GGAAAAGTT |
| 3 | None | 4 | AMF | 0.18 | 10_JJC:IID | 10_ATAATTGAT |
| 3 | None | 0 | AMF | 0.59 | 18_IEJ:MQL | 18_ATGCAATTA |
| 3 | None | 3 | Clostridium | 0.21 | 44_JGA:IGK | 44_ATAGGAAAA |
| 3 | None | 2 | AMF | 0.21 | 60_EGH:NGP | 60_AATGGTCCA |
| 3 | None | 1 | Clostridium | 0.34 | 64_KKA:YYR | 64_TATTATAGA |
| 3 | None | 5 | Clostridium | 0.05 | 70_JAA:IRK | 70_ATAAGGAAA |
| 4 | ASD | 0 | Clostridium | 0.12 | 69_KJA:WIR | 69_TGGATAAGG |
| 4 | None | 0 | Rhodobacter | 0.63 | 67_AAK:KKW | 67_AAGAAATGG |
| 5 | None | 0 | AMF | 0.5 | 31_CKJ:EWI | 31_GAATGGATA |
| 5 | None | 2 | AMF | 0.4 | 32_KJH:WIP | 32_TGGATACCA |
| 5 | None | 3 | AMF | 0.1 | 33_JHK:IPY | 33_ATACCATAC |
| 5 | None | 1 | AMF | 0.49 | 56_JJK:AIW | 56_GCAATATGG |
| 6 | ASD | 0 | Erwinia | 0.14 | 5_DDC:TSE | 5_ACCAGTGAA |
| 6 | None | 2 | Klebsiella | 0.23 | 0_CKA:EFK | 0_GAGTTTAAA |
| 6 | None | 0 | Klebsiella | 0.71 | 2_AEK:KNW | 2_AAAAACTGG |
| 6 | None | 1 | Klebsiella | 0.28 | 37_EKJ:QFI | 37_CAGTTTATT |
| 6 | None | 3 | Pseudanabaena | 0.17 | 38_KJE:FIN | 38_TTTATTAAC |
| 6 | None | 4 | Klebsiella | 0.17 | 41_JEC:INE | 41_ATTAACGAA |
| 7 | CSA | 0 | Geobacter | 0.02 | 71_AAD:RKS | 71_AGGAAATCC |
| 7 | None | 0 | Bradyrhizobium | 0.32 | 35_KEE:YNQ | 35_TACAATCAG |
| 8 | None | 0 | AMF | 0.14 | 4_KDD:WTS | 4_TGGACCAGT |
| 9 | ASD | 0 | Kocuria | 0.69 | 3_EKD:NWT | 3_AACTGGACC |

| | | | | | | |
|---|---|---|---|---|---|---|
| 10 | ASD | 2 | Clostridium | 0.08 | 24_DDK:STY | 24_AGTACATAT |
| 10 | ASD | 0 | Clostridium | 0.19 | 28_AJJ:KIV | 28_AAAATAGTT |
| 10 | ASD | 1 | Leuconostoc | 0.11 | 39_JEJ:INI | 39_ATTAACATT |
| 10 | Both | 0 | Cyanobacterium | 0.16 | 22_JED:INS | 22_ATTAATAGT |
| 10 | CSA | 2 | Rickettsia | 0.09 | 79_JEK:LNY | 79_TTAAATTAC |
| 10 | CSA | 0 | Cyanobacterium | 0.28 | 84_JCJ:LDI | 84_TTAGATATT |
| 10 | CSA | 1 | Pseudanabaena | 0.13 | 86_JAC:IKE | 86_ATTAAGGAA |
| 10 | None | 0 | AMF | 0.77 | 17_CIE:EMQ | 17_GAAATGCAA |
| 10 | None | 2 | AMF | 0.36 | 19_EJA:QLK | 19_CAATTAAAA |
| 10 | None | 3 | Rickettsia | 0.12 | 23_EDD:NST | 23_AATAGTACA |
| 10 | None | 4 | Sebacina | 0.09 | 40_EJE:NIN | 40_AACATTAAC |
| 10 | None | 1 | Leuconostoc | 0.37 | 59_AEG:KNG | 59_AAAAATGGT |
| 11 | ASD | 0 | Bdellovibrio | 0.15 | 73_DEC:SNE | 73_TCCAATGAA |
| 11 | None | 0 | Nitrosomonas | 0.18 | 72_ADE:KSN | 72_AAATCCAAT |
| 12 | ASD | 1 | Clostridium | 0.17 | 29_JJC:IVE | 29_ATAGTTGAA |
| 12 | ASD | 0 | Cyanobacterium | 0.8 | 58_KAE:WKN | 58_TGGAAAAAT |
| 12 | ASD | 2 | Rickettsia | 0.11 | 65_KAA:YRK | 65_TATAGAAAG |
| 12 | CSA | 0 | AMF | 0.55 | 26_KCA:YDK | 26_TATGATAAA |
| 12 | CSA | 3 | Nitrosospira | 0.09 | 42_ECJ:NEI | 42_AACGAAATA |
| 12 | CSA | 1 | Clostridium | 0.4 | 8_ECJ:NEI | 8_AATGAAATA |
| 12 | CSA | 2 | AMF | 0.1 | 82_JDJ:LTL | 82_TTAACATTA |
| 12 | None | 0 | AMF | 0.56 | 30_JCK:VEW | 30_GTTGAATGG |
| 12 | None | 2 | Clostridium | 0.22 | 43_CJG:EIG | 43_GAAATAGGA |
| 12 | None | 5 | Rickettsia | 0.03 | 61_GHJ:GPL | 61_GGTCCACTA |
| 12 | None | 4 | Leuconostoc | 0.05 | 62_HJK:PLY | 62_CCACTATAT |
| 12 | None | 1 | Clostridium | 0.35 | 68_AKJ:KWI | 68_AAATGGATA |
| 12 | None | 3 | AMF | 0.16 | 9_CJJ:EII | 9_GAAATAATT |

Note: A maximum of *three codon usages* among 54 genera and a genus with the maximum for each *Three codon DNA 9-mer* is indicated in the column- three_max and three_genus, respectively.

**Supplementary Table 2 Correlations of a pair of genera based on *trimer usage bias***

To download the supplementary table 2 (xls file), please click here (at the same time, press Ctrl key)

If the link does not work, please copy and paste the following web address in your browser.

www.codon.kr/thesis/chapter3/supple_table_2.xls

## Supplementary Table 3 The mean of the correlations in 9 FSs

To download the supplementary table 3 (docx file), please click here (at the same time, press Ctrl key)

If the link does not work, please copy and paste the following web address in your browser.

www.codon.kr/thesis/chapter3/supple_table_3.docx

**Supplementary Table 4 The mean and the standard deviation of the correlations of 9 FSs**

To download the supplementary table 4 (xls file), please click here (at the same time, press Ctrl key)

If the link does not work, please copy and paste the following web address in your browser.

www.codon.kr/thesis/chapter3/supple_table_4.xls

**Supplementary Table 5 Genus clusters of five additional sequences**

| | seq252 | seq284 | seq337 | seq475 | seq528 |
|---|---|---|---|---|---|
| **Cluster 0** | 0,1,2,4,5,9,10,16,17,19,21,22,23,24,25,26,29,30,34,35,36,37,40,41,42,43,44,49,51 | 4,5,17,22,34,37,41,43,44 | 1,4,5,9,10,16,19,21,22,23,24,25,26,30,34,35,36,37,40,41,42,43,44,51 | 7,13,28,33 | 1,2,4,5,8,9,10,16,19,21,22,23,24,25,26,30,34,35,36,37,40,41,42,43,44,51 |
| **Cluster 1** | 32,39 | 3,12,20,31,38 | 12,20 | 11,45 | 11,45 |
| **Cluster 2** | 38 | 7,13,27,32,39,47,50,52,53 | 13,31 | 4,5,9,10,16,19,22,23,24,25,30,34,35,37,41,42,43,44 | 6,12 |
| **Cluster 3** | 6,20 | 11,45 | 38,45 | 27,29,39,47,48,50,52,53 | 3,20 |
| **Cluster 4** | 7,8,14,15,18,46,47,48,52,53 | 6,33 | 27,33 | 6,12 | 27,33 |
| **Cluster 5** | 11,45 | 9 | 7,8,14,15,29,39,47,50,52,53 | 0,1,2,18,21,26,36,40 | 28,32,39 |
| **Cluster 6** | 33,50 | 1,2,21,26,36,40 | 0,2,17,18,46,48,49 | 3,20 | 38 |
| **Cluster 7** | 12 | 28,29 | 3,6 | 38 | 7,13,31 |
| **Cluster 8** | 3,13 | 0,8,14,15,18,46,48,49,51 | 11 | 8,14,15,17,46,49,51 | 14,15,18,29,47,48,50,52,53 |
| **Cluster 9** | 27,28,31 | 10,16,19,23,24,25,30,35,42 | 28,32 | 31,32 | 0,17,46,49 |

*Note*: The numbers in the table stand for the genera: 0:*Acidithiobacillus*,1:*Acidobacterium*,2:*Agrobacterium*,3:*Anabaena*,4:*Azorhizobiu*m,5:*Azotobacter*,6:*Bacillus*,7:*Bdellovibrio*,8:*Beijerinckia*,9:*Bradyrhizobium*,10:*Caulobacter*,11:*Clostridium*,12:*Cyanobacterium*,13:*Desulfotomaculu*m,14:*Desulfovibrio*,15:*Erwini*a,16:*Frankia*,17:*Geobacter*,18:*Klebsiella*,19:*Kocuria*,20:*Leuconostoc*,21:*Mesorhizobium*,22:*Methylococcus*,23:*Microbacterium*,24:*Micrococcus*,25:*Myxococcus*,26:*Nitrobacter*,27:*Nitrosococcus*,28:*Nitrosomonas*,29:*Nitrosospira*,30:*Nocardia*,31:*Nostoc*,32:*Oscillatoria*,33:*Pseudanabaena*,34:*Pseudomonas*,35:*Pseudonocardia*,36:*Rhizobium*,37:*Rhodobacter*,38:*Rickettsia*,39:*Shewanella*,40:*Sinorhizobium*,41:*Sphingomonas*,42:*Streptomyces*,43:*Variovorax*,44:*Xanthomonas*,45:AMF,46:*Aspergillus*,47:*Cenococcum*,48:*Cryptococcus*,49:*Mycosphaerella*,50:*Oidiodendron*,51:*Phanerochaete*,52:*Scleroderma*,53:*Sebacina*

## Supplementary Table 6 Loading clusters of five additional sequences

| seq252 | seq284 | seq337 | seq475 | seq528 |
|---|---|---|---|---|
| Cluster 0<br>69_JEJ_VNL_GTGAATTTG | Cluster 0<br>74_HGA_PGK_CCAGGAAAG | Cluster 0<br>12_KKJ_YFL_TATTTTCTC | Cluster 0<br>51_DDD_SSS_AGTTCGAGT | Cluster 0<br>7_AGJ_KGA_AAAGGAGCA |
| Cluster 1<br>59_EKA_NWK_AATTGGAAA | Cluster 1<br>36_JJD_IVT_ATAGTGACT | Cluster 1<br>Major cluster | Cluster 1<br>45_DGD_SGT_TCAGGTACT | Cluster 1<br>Major cluster |
| Cluster 2<br>34_AJE_RAN_AGAGCAAAT | Cluster 2<br>27_JHJ_LPI_TTGCCGATT | Cluster 2<br>8_DDK_TSF_ACATCTTTT | Cluster 2<br>43_GDD_GTS_GGTACTTCA | Cluster 2<br>24_AEJ_KNA_AAAAATGCG |
| Cluster 3<br>30_GAJ_GKV_GGTAAAGTG | Cluster 3<br>52_CHJ_EPI_GAACCAATC,<br>64_JJD_ILS_ATTTTGTCA | Cluster 3<br>46_ICJ_MDA_ATGGATGCA | Cluster 3<br>24_EAD_NKS_AATAAATCA,<br>56_EAK_QKY_CAAAAATAT | Cluster 3<br>31_JDA_ASK_GCCTCGAAA |
| Cluster 4<br>Major cluster | Cluster 4<br>1_GEH_GNP_GGTAACCCA | Cluster 4<br>40_AKC_KFE_AAATTTGAG | Cluster 4<br>11_CAJ_DKL_GACAAATTA,<br>26_DJC_SAE_TCAGCTGAA | Cluster 4<br>17_AEK_RNF_AGAAATTTT |
| Cluster 5<br>37_KKC_WFD_TGGTTTGAT | Cluster 5<br>Major cluster | Cluster 5<br>52_AJE_KLQ_AAACTTCAA | Cluster 5<br>Major cluster | Cluster 5<br>25_EJA_NAK_AATGCGAAA |
| Cluster 6<br>1_BDK_HTF_CATACATTC | Cluster 6<br>43_AJA_KLR_AAACTACGC | Cluster 6<br>27_AEG_RNG_CGTAATGGG | Cluster 6<br>40_AJA_RLR_AGATTACGA | Cluster 6<br>20_EKK_NYY_AATTATTAT |
| Cluster 7<br>70_EJA_NLK_AATTTGAAA,<br>71_JAD_LKS_TTGAAAAGT | Cluster 7<br>26_CJH_ELP_GAATTGCCG | Cluster 7<br>2_JJD_LVS_TTGGTATCA | Cluster 7<br>27_JCE_AEQ_GCTGAACAG | Cluster 7<br>36_AJH_KAP_AAAGCTCCA |
| Cluster 8<br>26_ACC_KDE_AAAGATGAG | Cluster 8<br>17_AAE_RRN_CGTAGAAAC | Cluster 8<br>5_DAK_TKY_ACTAAGTAT | Cluster 8<br>28_CEG_EQG_GAACAGGGA | Cluster 8<br>30_CJD_DAS_GATGCCTCG |
| Cluster 9<br>18_JKC_LYD_TTGTATGAC | Cluster 9<br>37_JDC_VTE_GTGACTGAA | Cluster 9<br>22_EKJ_QFI_CAATTCATA | Cluster 9<br>14_CCD_DDT_GACGACACA | Cluster 9<br>1_CDJ_ESL_GAATCTCTT |
| Cluster 10<br>31_AJK_KVY_AAAGTGTAT | Cluster 10<br>45_AKA_RYR_CGCTATAGA | Cluster 10<br>4_DDA_STK_TCAACTAAG | Cluster 10<br>33_CJA_EAR_GAAGCTAGG | Cluster 10<br>11_JJJ_LLI_TTATTAATC,<br>34_JDA_ITK_ATTACAAAA |
| Cluster 11<br>40_GEJ_GNI_GGGAACATA | Cluster 11<br>57_KIE_YMQ_TATATGCAA | Cluster 11<br>53_JEJ_LQI_CTTCAAATA | Cluster 11<br>29_EGE_QGN_CAGGGAAAT | Cluster 11<br>5_JDA_ATK_GCTACAAAA,<br>18_EKE_NFN_AATTTTAAT |
| Cluster 12<br>36_EKK_NWF_AATTGGTTT | Cluster 12<br>78_KJI_YLM_TATCTCATG | Cluster 12<br>7_KDD_YTS_TATACATCT,<br>35_KCH_YEP_TATGAACCA,<br>36_CHJ_EPI_GAACCAATT,<br>37_HJA_PIK_CCAATTAAA | Cluster 12<br>37_DKK_SFY_AGTTTTTAT | Cluster 12<br>13_JED_INT_ATCAACACT |

*Note*: The major cluster contains the rest of the *Three Codon DNA 9-mers* that were omitted from the table.

seq252:GCACATACATTCTATGAAGTAAATAATGCATTAGAATGGATACCTTATGATAAATTGTATGACATTAAATAT
ATTACGAAAGATGAGTTAGGTAAAGTGTATAGAGCAAATTGGTTTGATGGGAACATAATTGATAAATATTATAGTT
ATAATTATTGGGGTGATGTATTAAAACATAATTGGAAAAGAAACTATCCTAATATGTTTGTGAATTTGAAAAGTTT
AAATTCTCCAAATGATCTTAC ;

seq284:CCAATGGTAACCCAAATGGAAATGATAATGGTAATGGCAATGGTACAGAACGACGTAGAAACGTAGAA
GATCTTTATTCTGAATTGCCGATTGATAGTAAAACTAAGGAAATAGTGACTGAAGTTAATGCAAAACTACGCTAT
AGATATGTAAATATGGAACCAATCAAGCTTTATATGCAAGTTTGCCAATTTATTTTGTCATTATTTCCTGATGTAC
CGGATCCAGGAAAGTTATATCTCATGTTTCCGGATGGTAAAA ;

seq337:TATTTATTTGGTATCAACTAAGTATACATCTTTTTTATATTTTCTCTTTCCAAAATTAACAAATTTACAATTC

ATAAGAATACGTAATGGGGATAATATTAATAATTATGAACCAATTAAAAAATTTGAGGAATACGCAATGGATGCAAGTTATTATAAACTTCAAATACTTGAGT ;

seq475:AGCTCAATACAATCTTGGAGTTATTTATGAAACTGACAAATTAGACGACACAATTGCAGCACTGTATTGGTATAATAAATCAGCTGAACAGGGAAATCATGAAGCTAGGGAAAGTTTTTATAGATTACGAGGTACTTCAGGTACTAAGACTGTTAGTTCGAGTAGTATACAAAAATATGGTTCTATGGGTAT ;

seq528:CTTGAATCTCTTCTTGCTACAAAAGGAGCAGAGTTATTAATCAACACTTTAAGAAATTTTAATTATTATAAAAAAAATGCGAAAGAACAAGATGCCTCGAAAATTACAAAAGCTCCAAAAATTAAAAAAGAAATGAGTAAAATTAAGTGGTCACAAATT

# Post (POsition Specific genetic code Table): Novel Method for Studying Codon Assignment

Jee Eun Kang[1,*], Antonio Ciampi[2], Mohamed Hijri[1,*]

[1] *Institute de recherche en biologie végétale, Département de Sciences Biologiques, Université de Montréal, QC, H1X 2B2, Canada*

[2] *Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, QC, H3A 1A2, Canada*

Jee Eun Kang orcid.org/0000-0003-2475-0474

Antonio Ciampi orcid.org/0000-0003-4838-8297

Mohamed Hijri orcid.org/0000-0001-6112-8372

E-mail: jennifer.kang@umontreal.ca (Kang JE), mohamed.hijri@umontreal.ca (Hijri M)

Running Title: *Kang JE et al/ Post: Position specific genetic code tables*

| | Words | Figures | Tables | Sup. Figure | Sup. Tables |
|---|---|---|---|---|---|
| Total Counts | Approx. 4950 | 1 | 5 | 0 | 2 |

## 4.1 Abstract

Although most synonymous codons in the standard genetic code table differ only in the third position, synonymous codons encoding for Leucine, Arginine, and Serine differ in the first and/or the second position. Furthermore, a large number of irregular codons have been identified in nuclear and mitochondrial genomes of various organisms. Mainstream hypotheses addressing the origin of the genetic code claim that codon position has played important roles in its evolution; the current genetic codes show an association between the property of the nucleotide either in the first or the second position and the cognate amino acid. In this article, we introduce a novel bioinformatics program- Post (POsition Specific genetic code Table) that provides new perspectives on studying codon assignment. We have developed three different codon tables called Posts according to the codon position. As a case study, we investigated frequencies of irregular codons in 187 mitochondrial genomes of plants and fungi. While there are two different possible types of irregular codons, those that vary within-column and those that vary trans-column type, we observed that the majority of the identified irregular codons belonged to the within-column type. The result suggests that the Post provides a useful means for studying codon assignment.

**Keywords:** Position specific genetic code tables; Post; Arbuscular mycorrhizal fungi; Codon assignment;

## 4.2 Introduction

An amino acid except Methionine and Tryptophan can be encoded by multiple synonymous codons according to the standard genetic code table. Most synonymous codons have a different nucleotide in the third position, an observation explained by the wobble hypothesis. According to the hypothesis, the wobble base pairing between the first RNA in the anticodon of a tRNA and the third RNA in the codon of a mRNA is less specific (1). Therefore, one tRNA is responsible for translating more than one codon; the minimum number of tRNAs required for translating 20 amino acids is much less than 61 (1).

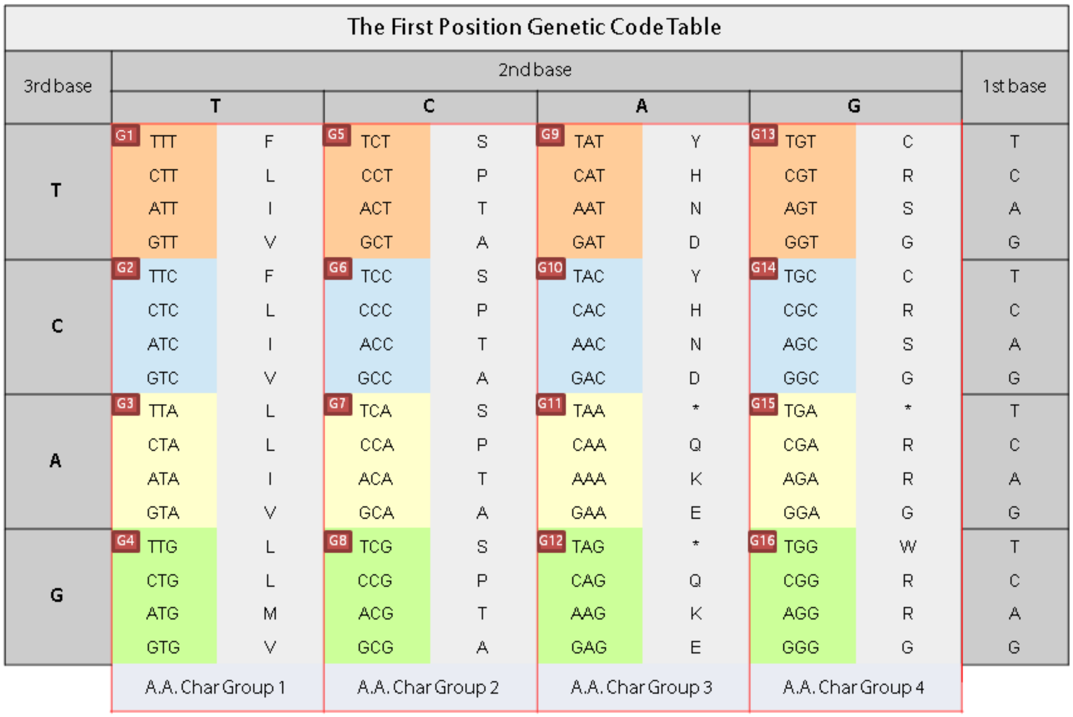
Evolutionary forces acting on codons are known to vary widely for the nucleotides at the first, second, and third position. The second position is known to be under the highest selection pressure while the third position is under the lowest selection pressure, because many of the changes in the third position result in synonymous mutations (2). However, synonymous codons for leucine and arginine exhibit less specificity toward nucleotides not only in the third position but also in the first position; leucine can be encoded by CTN or TTR where N represents adenine (A), thymine (T), G, and cytosine (C) and R represents purines. Similarly, arginine can be encoded by CGN or AGR. Furthermore, Serine showed less specificity in all three positions, as it can be encoded by TCN and AGY where Y represents pyrimidines. In addition to these synonymous changes, some organisms also contain what are termed irregular codons. For example, irregular stop codons that have been found to encode tryptophan, serine, glycine, and glutamine were identified in a human microbiome study (3). Selenocysteine and Pyrrolysine can be also encoded by stop codons. CTG$^{\text{Leu}}$ encodes for Serine in *Candida albicans*, CTG-clade fungi, and some mitochondrial genomes (4). A number of irregular codons have been found in mitochondrial genomes of various organisms. In addition, in plants' mitochondrial genomes, atypical start codons such GTG$^{\text{Val}}$ and ACG$^{\text{Thr}}$ and atypical stop codons such as CAA$^{\text{Gln}}$ and CGA$^{\text{Arg}}$ have been found (5). Interestingly, GTG$^{\text{Val}}$ serving as a start codon has been observed in algae, bryophytes and pteridophyte, while ACG$^{\text{Thr}}$ serving as a start codon has been observed in most land plants (5).

Mainstream hypotheses addressing the origin of the genetic code include the stereochemical hypothesis and the coevolution hypothesis. The stereochemical hypothesis suggested that the origin of the genetic code resulted from the interactions between pre-tRNA molecules and amino acids. The interactions were driven by affinity until they were replaced by aminoacyl-tRNA synthetases (6). The codon-correspondence hypothesis, claiming that the genetic codes arose from direct interactions between nucleotides and their cognate amino acids, is consistent with the claims that the genetic codes emerged before or during the RNA world (7). It has been suggested that the last universal common ancestor already had nearly all components of the translational apparatus, including aminoacyl-tRNA synthetases (8). Ribosomes, found in all living organisms and known to be highly conserved, may have brought some consistency into codon assignments across various taxonomic groups. However, detailed comparison of ribosome structures across diverse organisms remains to be assessed, because ribosomal structures have been solved only for intensively studied model organisms. Evidence supporting the stereochemical hypothesis came from chromatography and experiments involving fitting a cavity in a B-DNA construct with an amino acid (9). Both types of experiment supported the idea that the polarity of an amino acid correlates with the property of the nucleotide in the second position of its cognate tRNA anticodon (9,10). Coevolution hypothesis suggested that mechanisms involving codon assignment and amino acid biosynthesis coevolved through interactions between peptides and RNA-like molecules. It claimed that closely related amino acids should be defined in terms of their biosynthetic pathways and differ by no more than a single base (11). They assigned amino acids into 6 groups: pyruvate (GCN/Ala/A, GTN/Val/V, CTN-TTR/Leu/L); aspartate (GAY/Asp/D, ATY-ATA/Ile/I, ATG/Met/M, ACN/Thr/T, AAY/Asn/N, AAR/Lys/K); glutamate (GAR/Glu/E, CAR/Gln/Q, CCN/Pro/P, CGN-AGR/Arg/R); aromatic (TTN/Phe/F, TAY/Tyr/Y, TGG/Trp/W); serine (GGN/Gly/G, TCN-AGY/Ser/S, TGY/Cys/C); histidine (CAY/His/H) (10). They suggested that codons belonging to the same amino acid group often have the same nucleotide in the first position.

The origin and evolution of codons is a fundamental question that has an enormous impact on a number of areas of biological sciences. However, for decades, the research has been limited to a small number of model organisms. With a vast amount of omics data, variation of codon (re)assignment and the diversity of the translation apparatus can be more fully assessed (12). In order to study association between codon position and genetic code using omics data, we developed a novel bioinformatics method- POsition Specific genetic code Table (Post) that assigns a codon with respect to nucleotide position in the codon. It is widely accepted that position in the codon is an important factor reflecting the evolutionary dynamics of the genetic code. The Post assigns a codon into an amino acid group solely based on codon position in order to analyze evolutionary forces acting on genetic code with respect to codon position. Based on the Post, an irregular codon can be classified either into within-column type or trans-column type. An irregular codon of within-column type encodes for abnormal amino acid that is located in the same column as its cognate standard amino acid in the Post, while that of trans-column type does in a different column. A great majority of irregular codons identified in this study were classified into within-column type. This new method provides a useful tool for studying codon assignment.

We conducted a case study on arbuscular mycorrhizal fungi (AMF), which are plant-root inhabiting fungi that form symbiosis with more than 80% of vascular plants worldwide (13). They provide plants with nutrients, protect them against soil-borne pathogens, and increase tolerance to environmental stresses (14-17). AMF associated microbial community contains beneficial microorganisms such as nitrogen-fixing, phosphate solubilizing, and plant growth promoting microorganisms. A number of researchers have developed AMF inoculants as biofertilizer in order to reduce the use of pesticides and chemical fertilizers that deteriorate agroecosystem (18). In addition to its contribution in establishing sustainable agricultural practice, AMF inoculation has been utilized to clean up soil contaminated with petroleum hydrocarbons, metals, uranium, and iron-cyanide complexes (19-21).

**Table 1 The first position specific genetic code table**

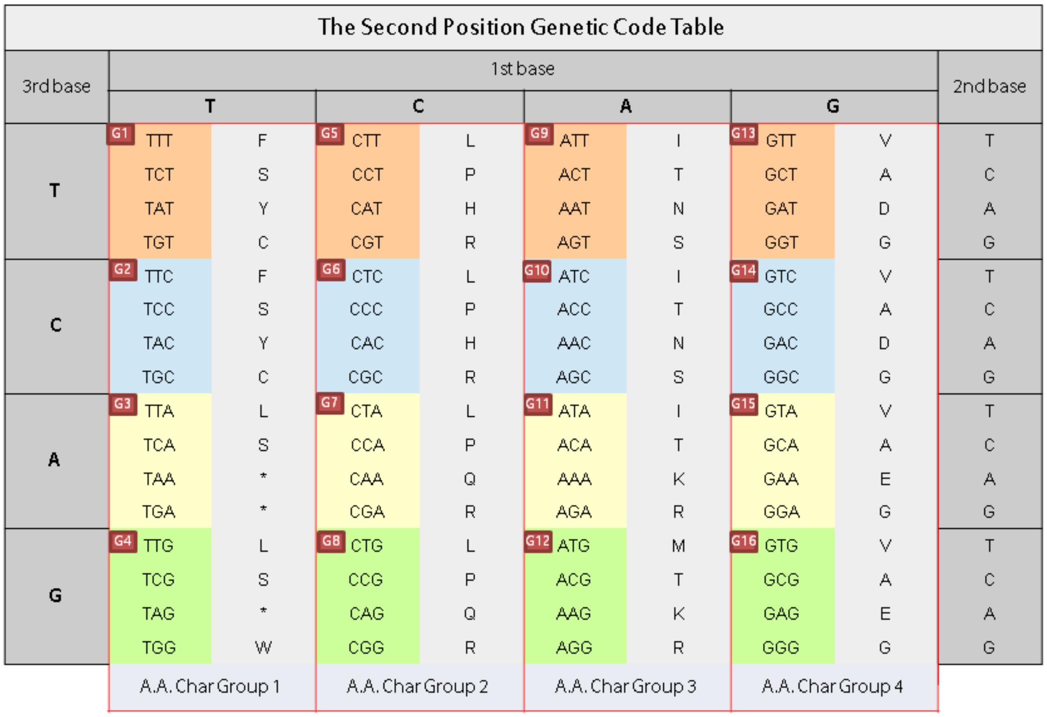


The First Position Genetic Code Table

| 3rd base | 2nd base | | | | | | | | 1st base |
|---|---|---|---|---|---|---|---|---|---|
| | T | | C | | A | | G | | |
| T | G1 TTT | F | G5 TCT | S | G9 TAT | Y | G13 TGT | C | T |
| | CTT | L | CCT | P | CAT | H | CGT | R | C |
| | ATT | I | ACT | T | AAT | N | AGT | S | A |
| | GTT | V | GCT | A | GAT | D | GGT | G | G |
| C | G2 TTC | F | G6 TCC | S | G10 TAC | Y | G14 TGC | C | T |
| | CTC | L | CCC | P | CAC | H | CGC | R | C |
| | ATC | I | ACC | T | AAC | N | AGC | S | A |
| | GTC | V | GCC | A | GAC | D | GGC | G | G |
| A | G3 TTA | L | G7 TCA | S | G11 TAA | * | G15 TGA | * | T |
| | CTA | L | CCA | P | CAA | Q | CGA | R | C |
| | ATA | I | ACA | T | AAA | K | AGA | R | A |
| | GTA | V | GCA | A | GAA | E | GGA | G | G |
| G | G4 TTG | L | G8 TCG | S | G12 TAG | * | G16 TGG | W | T |
| | CTG | L | CCG | P | CAG | Q | CGG | R | C |
| | ATG | M | ACG | T | AAG | K | AGG | R | A |
| | GTG | V | GCG | A | GAG | E | GGG | G | G |
| | A.A. Char Group 1 | | A.A. Char Group 2 | | A.A. Char Group 3 | | A.A. Char Group 4 | | |

Note: G in the red box stands for *A.A. Group*. For example, G1 stands for A.A. Group 1.

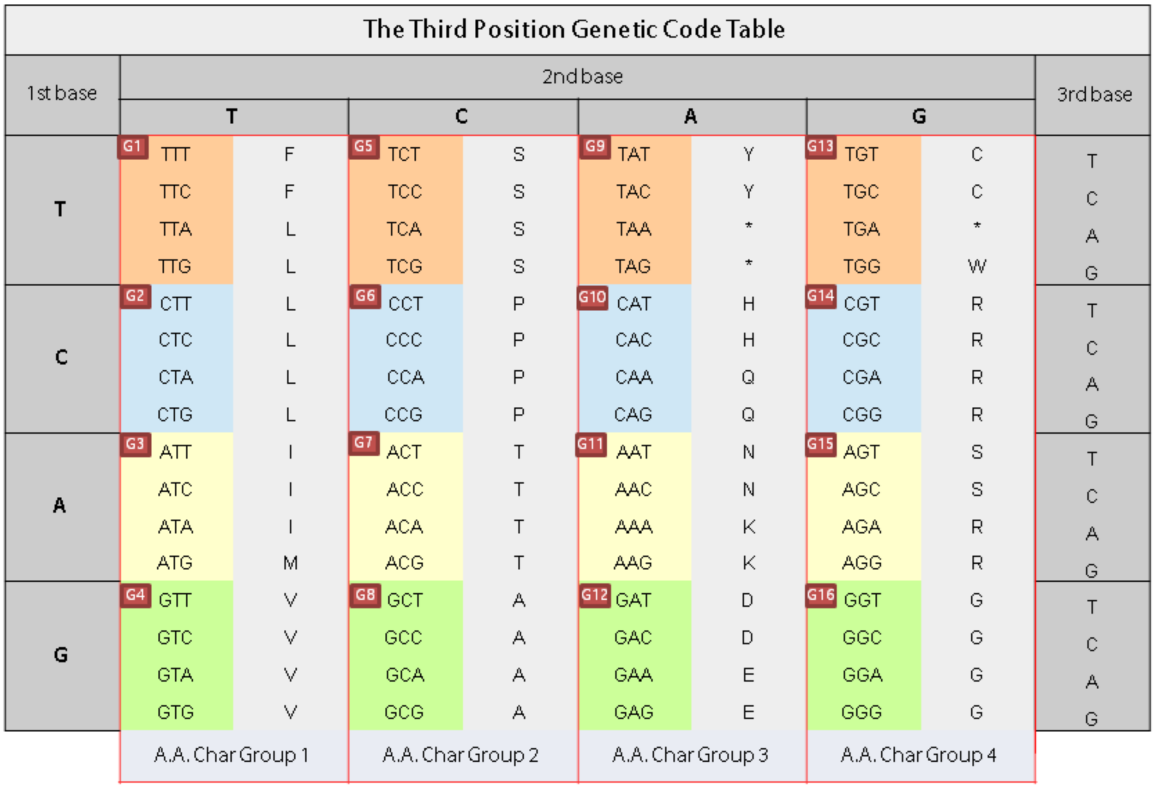
### Table 2 The second position specific genetic code table



| The Second Position Genetic Code Table | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 3rd base | 1st base | | | | | | | | 2nd base |
| | T | | C | | A | | G | | |
| T | G1 TTT | F | G5 CTT | L | G9 ATT | I | G13 GTT | V | T |
| | TCT | S | CCT | P | ACT | T | GCT | A | C |
| | TAT | Y | CAT | H | AAT | N | GAT | D | A |
| | TGT | C | CGT | R | AGT | S | GGT | G | G |
| C | G2 TTC | F | G6 CTC | L | G10 ATC | I | G14 GTC | V | T |
| | TCC | S | CCC | P | ACC | T | GCC | A | C |
| | TAC | Y | CAC | H | AAC | N | GAC | D | A |
| | TGC | C | CGC | R | AGC | S | GGC | G | G |
| A | G3 TTA | L | G7 CTA | L | G11 ATA | I | G15 GTA | V | T |
| | TCA | S | CCA | P | ACA | T | GCA | A | C |
| | TAA | * | CAA | Q | AAA | K | GAA | E | A |
| | TGA | * | CGA | R | AGA | R | GGA | G | G |
| G | G4 TTG | L | G8 CTG | L | G12 ATG | M | G16 GTG | V | T |
| | TCG | S | CCG | P | ACG | T | GCG | A | C |
| | TAG | * | CAG | Q | AAG | K | GAG | E | A |
| | TGG | W | CGG | R | AGG | R | GGG | G | G |
| | A.A. Char Group 1 | | A.A. Char Group 2 | | A.A. Char Group 3 | | A.A. Char Group 4 | | |

Note: G in the red box stands for *A.A. Group*. For example, G1 stands for A.A. Group 1.

### Table 3 The third position specific genetic code table



| The Third Position Genetic Code Table | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1st base | 2nd base | | | | | | | | 3rd base |
| | T | | C | | A | | G | | |
| T | G1 TTT | F | G5 TCT | S | G9 TAT | Y | G13 TGT | C | T |
| | TTC | F | TCC | S | TAC | Y | TGC | C | C |
| | TTA | L | TCA | S | TAA | * | TGA | * | A |
| | TTG | L | TCG | S | TAG | * | TGG | W | G |
| C | G2 CTT | L | G6 CCT | P | G10 CAT | H | G14 CGT | R | T |
| | CTC | L | CCC | P | CAC | H | CGC | R | C |
| | CTA | L | CCA | P | CAA | Q | CGA | R | A |
| | CTG | L | CCG | P | CAG | Q | CGG | R | G |
| A | G3 ATT | I | G7 ACT | T | G11 AAT | N | G15 AGT | S | T |
| | ATC | I | ACC | T | AAC | N | AGC | S | C |
| | ATA | I | ACA | T | AAA | K | AGA | R | A |
| | ATG | M | ACG | T | AAG | K | AGG | R | G |
| G | G4 GTT | V | G8 GCT | A | G12 GAT | D | G16 GGT | G | T |
| | GTC | V | GCC | A | GAC | D | GGC | G | C |
| | GTA | V | GCA | A | GAA | E | GGA | G | A |
| | GTG | V | GCG | A | GAG | E | GGG | G | G |
| | A.A. Char Group 1 | | A.A. Char Group 2 | | A.A. Char Group 3 | | A.A. Char Group 4 | | |

Note: G in the red box stands for *A.A. Group*. For example, G1 stands for A.A. Group 1.

We conducted comparative mitochondrial genomics using a total of 187 mitochondrial genomes of taxonomically diverse organisms covering plants and fungi including lichen-forming fungi and AMF. We restricted our analyses to irregular codons, those codons whose nucleotide sequence and its cognate protein sequence provided by NCBI do not conform with the standard genetic code table (although it is unclear whether it was a result of (pre/post) transcription or (pre/post) translation related processes). The goal of this study was to assess the degree of variations in the occurrences of irregular codons, to classify an irregular codon into within-column type or trans-column type, and to assess the differences of codon frequencies in mitochondrial genomes according to the Post.

## 4.3 Methods

### 4.3.1 The Post: new perspectives for studying codon property and codon assignment

The Post was created to facilitate analysis of the position of a nucleotide in a codon. Each Post contains 16 amino acid groups contained in a 4 x 4 matrix. Each amino acid group (*A.A. Group*) consists of 4 codons where one codon position can be A, T (U), G, or C, while the other two positions are fixed (Tables 1-3). For example, TTC, CTC, ATC, and GTC belong to the same *A.A. Group* in the first Post (Post 1) (Table 1). A.A. Groups located in the same column of the table were assigned into what we will hereafter refer to as an amino acid characteristics group (*A.A. Char group*). For example, A.A. Group1 (TTT, TCT, TAT, TGT), A.A. Group2 (TTC, TCC, TAC, TGC), A.A. Group3 (TTA, TCA, TAA, TGA), and A.A. Group4 (TTG, TCG, TAG, TGG) belong to A.A. Char Group 1 in the second Post (Post 2) (Table 2). Codons belonging to the same *A.A. Char Group* in the Post 2 have the same nucleotide in the first position, while those in the Post 1 have the same nucleotide in the second position.

In this study, a codon frequency is defined as an observed frequency ratio of a codon within an *A.A. Char Group*. The program calculates a *codon rank* based on codon frequency in increasing order; ranks range from 0 to 15 with rank 0 the lowest codon frequency. Codons belonging to *an A.A. Char Group* of Post 1 are identical to those of the Post3. Therefore both of the Post 1 and the Post3 will be hereafter referred as to the Post 1st/3rd. The Post program compares DNA sequence with its provided amino acid sequence, and identifies those codons whose translated amino acid, indicated in the amino acid sequence, does not conform with the standard genetic code table. Although DNA sequences and their amino acid sequences are not given an annotation to identify what might have caused an irregular codon, e.g., nucleotide substitution, RNA editing, amino acid substitution, or codon substitution, we used the term, irregular codon, to describe all of these throughout the article. The program identifies an irregular codon in a query sequence and indicates whether the normally encoded amino acid and the actually encoded amino acid are from the same *A.A. Char Group* either in the Post 1st/3rd or in the Post 2. An irregular codon is classified into one of four types: WW, WT, TW, and TT. WW and WT stand for Within-column type in Post1st/3rd and Within-column type in Post2, and for Within-column type in Post1st/3rd and Trans-column type in Post2, respectively. TW and TT stand for Trans-column type in Post1st/3rd and Within-column type in Post2, and for Trans-column type in Post1st/3rd and Trans-column type in Post2, respectively.

### 4.3.2 Implementation and program availability

The Post has been implemented using the Java programming language (www.java.net, www.oracle.com (Java8)). The programs have been tested and confirmed to work on Linux system-CentOS Linux 7 (www.centos.org), and are currently being used at the Biodiversity Center, Institut de Recherche en Biologie Végétale, Département de Sciences Biologiques, Université de Montréal. They are freely available for download at www.codon.kr. There are no restrictions to use the programs by academic or non-academic organizations as long as they comply with the terms and conditions of the license agreement.

### 4.3.3 Input, output, and options

The Post utilizes a command-line interface. There are several Post programs available. The Post program introduced in this paper requires users to give two mandatory command line arguments (e.g., /home/input-file each). The first argument is an input file path that should contain DNA sequence(s) in fasta format. The second argument should be either “each” or “all” to indicate whether a user wants the program to calculate codon frequencies in “each” query sequence or in “all” the sequences in an input file. The Post produces output files containing two different types of information per reading frame. One is codon frequency according to the Post and the other is identification of irregular codons such as shown in the case study in the following section. Another Post programs identify irregular codons based on the protein sequence that a user provides. Each program comes with the instruction file that explains how to use the program.

### 4.3.4 A case study: an example application for the program

#### 4.3.4.1 Sequence databases

We created two sequence databases based on a batch file called mitochondrion.1.genomic provided by the NCBI GenBank; one contains DNA coding sequences and the other contains their cognate protein sequences from 179 plant and 8 fungal mitochondrial genomes. We assigned sequences from 187 genomes into 23 taxonomic groups: 20 plants, 2 lichen-forming fungi (*Xylariaceae* and *Peltigera*), and 1 AMF (*Gigaspora* and *Rhizoglomus*). The databases were highly skewed in terms of the number of genomes available. While a large number of mitochondrial genomes have been completely sequenced for the seed plants belonging to Spermatophyta, only a few mitochondrial genomes have been sequenced for the early diverged plants and fungi. The databases were created without manual curation. If an exon contained a large mismatching gap that does not match to its provided protein sequence, it was excluded. Per taxonomic group, we created sub-databases according to the following gene groups: atp6, atp9, cox1, cox2, cox3, cob, cytb, nad1, nad2, nad3, nad4, nad4L, nad5, nad6, and orf. Only the listed genes were included in this study.

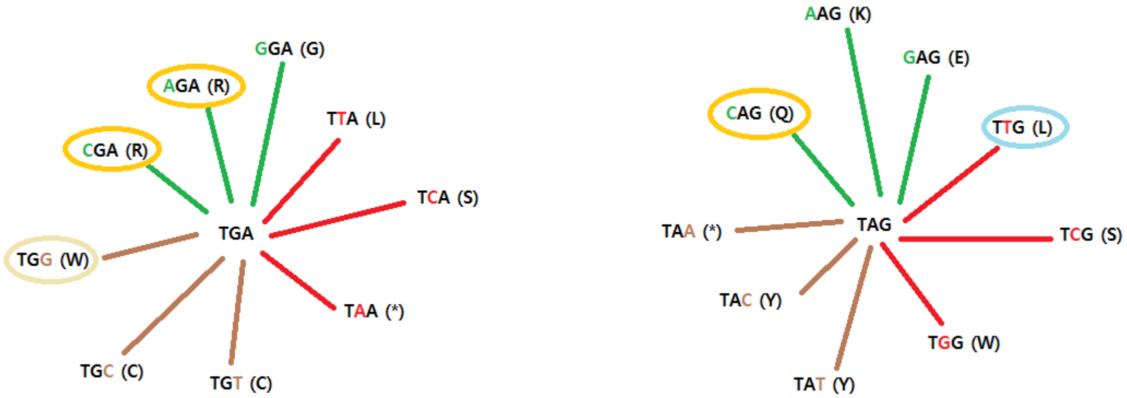


**Figure 1. Position specific nucleotide substitution of stop codons- TGA and TAG**

Note: Circles indicate the irregular codons belonging to the list of a pair of amino acids.

### 4.3.4.2 Rank database- codon rank per gene group and per taxonomic group

Per taxonomic group and per gene group in the sequence databases, we calculated *codon rank* based on codon frequency in a gene according to the Post 1st/3rd and the Post 2. In most of the cases, there were multiple genes per gene group because more than one genome belonged to a taxonomic group. We calculated the mean of the ranks of a codon from multiple genes belonging to the same gene group according to the Post 1st/3rd and the Post 2, excluding those having 0 frequency per gene group and per taxonomic group. The rank data will be called hereafter gene-taxon rank data. We created two different sets of gene-taxon rank data, one with all the genes and the other only with genes containing at least one irregular codon.

### 4.3.4.3 Mann Whitney U Tests applied to the gene-taxon rank data

For the identified irregular codons of within-column type, we made a list of amino acid pairs, where the first one is what an irregular codon is supposed to encode according to the standard genetic code table and the second one is what an irregular codon actually encoded as indicated in its protein sequence: *_R, R_W, *_L, R_L, T_M, *_W, *_Q, R_C, P_S, H_Y, L_F, L_I, A_V, S_L, S_F, P_L, and T_I where * stands for stop codon. For Mann Whitney U Tests, we analyzed only the within-column type in the list. For example, TAG is supposed to be a stop codon according to the standard genetic code table. However, its cognate amino acid written in its protein sequence was leucine. We included the irregular codon TAG in this study, because the stop codon and leucine belonged to the same A.A. Group 4 and consequently to the same A.A. Char Group 1 in the Post 2 (Figure 1). We applied the Mann Whitney U Test to each codon of the list in two different sets of gene-taxon rank data in order to assess the equality of the ranks of the codon between two different sets of genes where one set contains all the genes regardless of the occurrence of the irregular codons and the other contains only the genes containing the irregular codons.

**Table 4 Frequencies of irregular codons with the type- WW/WT/TW/TT**

| A.A. | Codon | WW | WT | TW | TT |
|---|---|---|---|---|---|
| * | TAG | 0 | *_Q:7 | *_L:250 | 0 |
| * | TGA | *_W:2317 | *_R:21 | 0 | *_Q:1 |
| * | TAA | 0 | *_Q:51 | 0 | 0 |
| A | GCG | 0 | A_P:1 | A_D:1:A_V:4 | A_Y:1 |
| A | GCC | 0 | A_P:1:A_S:1 | A_E:1:A_V:6 | A_F:2:A_I:2:A_K:1:A_M:2 |
| A | GCT | 0 | A_S:3:A_T:1 | A_V:2:A_G:1 | A_C:2:A_F:2:A_I:2:A_Y:2:A_L:4:A_M:3 |
| A | GCA | 0 | A_T:1 | A_D:1:A_V:2:A_G:2 | A_Q:1:A_R:1:A_L:1 |
| C | TGC | C_S:1 | C_G:1 | 0 | C_N:1 |
| C | TGT | C_S:1 | 0 | C_F:1 | C_N:1:C_A:1:C_I:1:C_H:1 |
| D | GAC | 0 | D_K:1 | 0 | D_F:2:D_I:1 |
| D | GAT | D_E:1 | D_K:1:D_N:1:D_Q:2 | D_A:2:D_V:1 | D_L:1:D_M:1:D_F:1 |
| E | GAG | 0 | E_Q:1 | E_G:2 | E_L:1:E_R:1:E_I:1 |
| E | GAA | 0 | E_Y:1 | E_A:1:E_V:1 | 0 |
| F | TTC | F_L:4 | F_M:1 | F_W:1 | F_K:1:F_N:1:F_P:1:F_R:1:F_Q:1:F_T:1:F_D:1:F_E:1:F_G:3 |
| F | TTT | F_L:3 | F_I:2:F_V:1 | F_Y:1:F_C:2:F_S:2:F_W:1 | F_P:1:F_R:1:F_A:4:F_Q:2:F_T:3:F_E:1:F_H:1:F_G:1 |
| G | GGG | 0 | G_S:1:G_R:1 | 0 | G_I:1:G_Y:1:G_L:1:G_N:1 |
| G | GGA | 0 | G_S:2:G_W:1 | G_A:1:G_D:1:G_V:1 | G_I:1:G_Y:1:G_K:1:G_L:1:G_F:3 |
| G | GGC | 0 | G_S:1 | G_V:1 | G_I:1:G_Y:1:G_P:1 |
| G | GGT | 0 | G_C:1:G_S:1 | G_A:3:G_D:1 | G_I:1:G_L:1:G_P:1 |
| H | CAC | 0 | H_Y:21 | 0 | H_A:1 |
| H | CAT | 0 | H_Y:54:H_D:1 | H_L:1 | H_A:2:H_T:1:H_F:1 |
| I | ATA | 0 | I_V:1:I_L:3 | I_K:1:I_S:1:I_T:1 | I_G:1:I_A:2:I_P:1 |
| I | ATC | 0 | I_F:1:I_L:2 | I_S:2 | 0 |
| I | ATT | 0 | I_F:2:I_V:1:I_L:2 | I_S:2:I_T:1 | I_G:2:I_H:3:I_P:2:I_C:1:I_E:1 |

| | | | | | |
|---|---|---|---|---|---|
| K | AAG | 0 | K_Y:1 | K_M:1:K_R:1 | 0 |
| K | AAA | 0 | K_D:1 | K_S:1:K_R:1 | K_G:1:K_A:3 |
| L | CTA | 0 | L_F:1:L_V:1:L_I:2 | 0 | L_T:2:L_K:1 |
| L | CTC | 0 | L_F:20:L_I:1:L_M:1 | 0 | L_Y:1:L_A:2 |
| L | CTG | 0 | L_F:1 | 0 | L_S:1:L_A:1 |
| L | CTT | 0 | L_F:44:L_V:1:L_I:1 | L_Q:3 | L_G:1:L_A:2 |
| L | TTG | 0 | L_V:1:L_I:1 | 0 | L_G:3:L_N:2:L_R:1:L_A:2 |
| L | TTA | L_F:2 | L_I:3:L_M:1 | L_S:1 | L_D:1:L_E:1:L_G:1:L_R:1:L_A:1:L_Q:1 |
| M | ATG | M_I:3 | M_V:1:M_L:3 | M_S:1:M_R:2:M_T:1:M_N:1 | M_E:1:M_G:2:M_A:1 |
| N | AAC | 0 | N_D:1:N_Y:1 | N_S:1 | N_F:1 |
| N | AAT | N_K:1 | N_D:1:N_H:1 | N_R:2:N_S:2:N_M:1 | N_A:1:N_V:1:N_P:1 |
| P | CCC | 0 | P_T:1:P_S:18 | P_L:29 | P_D:1:P_F:15:P_V:1:P_I:1:P_Y:1 |
| P | CCG | 0 | P_S:2 | P_L:66 | 0 |
| P | CCT | 0 | P_T:1:P_S:56 | P_R:1:P_L:79 | P_F:7:P_I:1:P_M:1 |
| P | CCA | 0 | P_T:1:P_S:27 | P_L:94 | P_V:1:P_M:1 |
| Q | CAG | 0 | Q_N:1 | 0 | Q_A:1:Q_I:1 |
| Q | CAA | 0 | Q_D:1 | Q_R:1:Q_L:3 | Q_S:1:Q_T:2:Q_F:1:Q_V:1 |
| R | AGG | 0 | R_C:1 | 0 | R_V:1 |
| R | CGG | 0 | R_W:99 | R_L:1 | R_K:1 |
| R | AGA | R_S:1 | 0 | R_N:1 | R_L:1 |
| R | CGA | 0 | R_C:1 | R_L:3 | R_A:1:R_D:1:R_V:1:R_I:1 |
| R | CGC | 0 | R_C:23:R_W:1 | 0 | 0 |
| R | CGT | 0 | R_C:53 | R_L:1 | R_F:1 |
| S | TCC | 0 | S_A:1:S_P:1 | S_F:89 | S_K:1 |
| S | TCG | 0 | 0 | S_L:98:S_F:2:S_Y:1 | S_D:1:S_G:1:S_V:2:S_H:1 |
| S | AGC | 0 | 0 | 0 | S_L:1:S_V:1:S_F:1 |
| S | TCT | 0 | S_A:1:S_T:1 | S_L:2:S_W:1:S_F:153:S_Y:1 | S_N:1:S_V:2 |
| S | AGT | 0 | S_G:2 | S_T:2 | S_L:1:S_A:2:S_Q:2:S_V:1 |
| S | TCA | 0 | S_P:1:S_T:1 | S_L:221:S_F:3:S_Y:1 | S_G:1 |
| T | ACG | T_S:1 | T_A:1 | T_M:22 | 0 |
| T | ACT | T_S:3 | T_P:1 | T_K:1:T_N:1:T_M:1:T_I:11 | T_L:2:T_C:1:T_F:2:T_G:3:T_W:1 |
| T | ACA | 0 | T_A:1 | T_R:1:T_I:22 | T_F:2:T_V:1 |
| T | ACC | 0 | 0 | T_I:8 | T_Y:1 |
| V | GTC | 0 | V_I:1:V_L:1:V_F:1 | 0 | V_N:1:V_R:1:V_S:2 |
| V | GTA | 0 | 0 | V_A:1:V_D:1:V_G:3 | V_N:1:V_P:1:V_T:1:V_S:1 |
| V | GTG | 0 | V_I:2 | 0 | V_T:1 |
| V | GTT | 0 | V_I:1:V_L:1:V_F:1 | V_A:1:V_G:1 | V_S:1 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| W | TGG | W_S:1 | 0 | W_L:1:W_F:2 | W_I:1:W_E:1 |
| Y | TAC | 0 | 0 | Y_L:1:Y_S:1 | Y_P:1:Y_T:2 |
| Y | TAT | 0 | Y_N:1:Y_E:1 | Y_L:1:Y_S:4 | Y_G:1:Y_M:1:Y_A:1:Y_T:1 |
| Total |  | 2,339.00 | 591 | 1258 | 234 |

Note: Pair of amino acids indicates standard amino acid and actual amino acid that an irregular codon encoded for. It is followed by a number that indicates the frequency of the irregular codon.

### 4.3.5 Results of the case study

#### 4.3.5.1 Irregular codons

The most frequently occurring irregular codon was TGA (WW type) in the sequence databases. The second and third most frequent irregular codons were TAG (TW type) and TCA (TW type) (Table 4). Irregular codons of within-column type either or both in Post1st/3rd and Post2 occurred far more frequently than those of TT type. The great majority of the irregular codons belonged to the within-column type either or both in Post1st/3rd and Post2. Irregular codons with a frequency of 20 or higher were only found in WW, WT, or TW types. Most of frequently occurring irregular codons had ring structure in the side chain of either or both of standard amino acid and actually encoded amino acid (Table S1). They also frequently showed either or both uncharged polar and hydrophobic properties.

**Table 5 Irregular codons and *P* values from Mann Whitney U test**

| Pair | Irr. codon | Post | *P* value min. | *P* value max. | Pair | Irr. codon | Post | *P* value min. | *P* value max. |
|---|---|---|---|---|---|---|---|---|---|
| R_C | CGC | 1st/3rd | 0.05 | 0.73 | S_L | TCA | 2nd | 0.04 | 0.95 |
| R_C | CGT | 1st/3rd | 0.04 | 0.68 | S_L | TCG | 2nd | 0.09 | 0.95 |
| L_F | CTC | 1st/3rd | 0.23 | 0.82 | T_M | ACG | 2nd | 0.15 | 0.88 |
| L_F | CTT | 1st/3rd | 0.06 | 0.89 | P_S | CCA | 1st/3rd | 0.03 | 0.88 |
| S_F | TCC | 2nd | 0.11 | 0.96 | P_S | CCC | 1st/3rd | 0.04 | 0.94 |
| S_F | TCT | 2nd | 0.01 | 0.93 | P_S | CCG | 1st/3rd | 0.37 | 0.37 |
| L_I | CTC | 1st/3rd | 0.23 | 0.23 | P_S | CCT | 1st/3rd | 0.12 | 0.96 |
| T_I | ACA | 2nd | 0.04 | 0.49 | A_V | GCA | 2nd | 0.93 | 0.93 |
| T_I | ACC | 2nd | 0.57 | 0.57 | A_V | GCC | 2nd | 0.23 | 0.23 |
| T_I | ACT | 2nd | 0.89 | 0.99 | A_V | GCG | 2nd | 0.52 | 0.89 |
| P_L | CCA | 2nd | 0.02 | 0.84 | A_V | GCT | 2nd | 0.01 | 0.69 |
| P_L | CCC | 2nd | 0.05 | 0.62 | R_W | CGG | 1st/3rd | 0.00 | 0.89 |
| P_L | CCG | 2nd | 0.00 | 0.81 | H_Y | CAC | 1st/3rd | 0.02 | 0.3 |
| P_L | CCT | 2nd | 0.01 | 0.93 | H_Y | CAT | 1st/3rd | 0.04 | 0.77 |
| R_L | CGG | 2nd | 0.02 | 0.02 |  |  |  |  |  |

The most widely spread irregular codon in the mitochondrial genomes was the stop codon TGA (*) actually encoding for Tryptophan. This is a well-documented phenomenon in the mitochondrial genomes of various organisms. In plant taxonomic groups, Bangiophyceae, Florideophyceae, Pedinophyceae, and Prasinophytes had only this irregular codon. In the fungal taxonomic groups (Glomeromycota, Xylariaceae, and Peltigerales) in the database, only this irregular codon was observed as well**.** It is noteworthy that this irregular codon is a within-column type both of Post1st/3rd and of Post2.

In contrast to the widely observed irregular codon, TGA (*) encoding for tryptophan, Chlorophyceae had less common irregular stop codon- TAG (*) encoding leucine. Lycopodiidae included a few irregular stop codons (TAG and TAA) encoding Glutamine. Irregular codons encoding amino acids are shown in the Table 5; irregular stop codons were excluded due to a high number of occurrences. In Anthocerotophyta, irregular codons were dispersed across species. Anthocerotophyta had three stop codons (TAA, TAG, and TGA) encoding Arginine and Glutamine and CGC$^{Arg}$ encoding Cysteine. Spermatophyta had the highest occurrences of irregular codons as well as the most diverse irregular codons (Table S2). Interestingly, the widely spread irregular stop codon was not observed in the Spermatophyta, the most evolved plant division.

#### 4.3.5.2 Mann Whitney U Tests applied to gene-taxon rank data

Mann Whitney U Tests, applied to two different sets of gene-taxon rank data, indicated that the irregular codons- CCG$^{Pro}$ of the pair P_L and CGG$^{Arg}$ of the pair R_W, that occurred in the gene-nad4 in Spermatophyta, had the lowest P values (Table 5, Table S2). While most irregular codons showed a wide range of P values across various gene groups, CAC$^{His}$ of the pair H_Y showed the low P values. Irregular codons that occurred in only one gene group of one taxonomic group had only one P value, the same P value for both the minimum and the maximum: CTC$^{Leu}$ of the pair L_I, ACC$^{Thr}$ of the pair T_I, CGG$^{Arg}$ of the pair R_L, CCG$^{Pro}$ of the pair P_S, GCA$^{Ala}$ of the pair A_V, and GCC$^{Ala}$ of the pair A_V.

## 4.4 Discussions

The case study showed that the irregular stop codon, UGA encoding for Tryptophan, was WW type. This irregular codon was by far the most frequently observed phenomenon in mitochondrial genomes. It was also the only observed irregular codon in the ancient plants, Rhodophyta and Chlorophyta (Bangiophyceae, Florideophyceae, Pedinophyceae, and Prasinophytes), as well as in the fungi (Glomeromycota, Xylariaceae, and Peltigerales) in this study. Another frequent irregular stop codon, TAG encoding for Leucine, was TW type, and observed only in the green algae, Chlorophyceae. The stop codons encoding for Glutamine or Arginine were mostly WT type, and observed in non-vascular plant, Anthocerotophyta, and the oldest lineage of vascular plant, Lycopodiidae. The most evolved plant division, Spermatophyta, had the most diverse irregular codons. However, none of isolates belonging to them have the widespread irregular stop codon, UGA encoding for Tryptophan. These observations lead to a question whether a taxonomic group has preference in type of irregular codons.

If so, do four types of irregular stop codons (WW, WT, TW, and TT) reflect taxonomically specific property? In addition, a category of non-coding RNA may be a key determinant in gene regulation and codon assignment, considering that it is speculated that the genetic code emerged in (pre) RNA world. A large number of studies have documented that non-coding RNAs play crucial roles in many steps of (pre/post) transcriptional and translational processes (21-24). Transcriptional and translational components may vary considerably with respect to a category of non-coding RNA.

AMF harbor a large number of microorganisms inside of their spores and mycelia, some of which are obligate endosymbionts (25-32). If AMF contain different types of transcriptional and translational components tailored to their endosymbionts, their codon assignments may not abide by the standard genetic code table and vary considerably. In future, we may utilize Post in order to conduct comparative study in codon assignments of endosymbionts with respect to category of the non-coding RNA across taxonomic groups as well as across gene types.

## 4.5 Authors' contributions

KJE designed the new method- the Post, created the databases, implemented the new method, and conducted the case study using Java programming language. CA provided insight and advice about statistical methods. HM outlined the goals of the project, provided insightful advice about mitochondrial genomes of AMF, and helped to draft the manuscript.

## 4.6 Acknowledgements

The authors gratefully acknowledge AFE (Aide financière aux études), FESP (Faculté des études supérieures et postdoctorales de l'UdeM), and IRBV (Institute de recherche en biologie végétale - Université de Montréal) for awarding scholarships to Kang Jee Eun. We thank Maryam Nadimi for helpful comments about mitochondrial genomes of AMF and Sylvie Hamel for the insightful selection of the scientific article- Xu et al 2015. We also thank David Morse for his helpful comments on the manuscript.

## 4.7 References

1. Crick, F.H.C. (1966) Codon-anticodon pairing: The wobble hypothesis. Journal of Molecular Biology, 19, 548–555.
2. Bofkin, L. and Goldman, N. (2007) Variation in evolutionary processes at different codon positions. . Molecular Biology and Evolution, 24, 513-521.
3. Ivanova, N.N., Schwientek, P., Tripp, H.J., Rinke, C., Pati, A., Huntemann, M., Visel, A., Woyke, T., Kyrpides, N.C. and Rubin, E., M. (2014) Stop codon reassignments in the wild. Science, 344, 909-913.

4. Rocha, R., Pereira, P.J., M.A., S. and Macedo-Ribeiro, S. (2011) Unveiling the structural basis for translational ambiguity tolerance in a human fungal pathogen. Proc Natl Acad Sci USA, 108, 14091–14096.

5. Xu, W., Xing, T., Zhao, M., Yin, X., Xia, G. and Wang, M. (2015) Synonymous codon usage in plant mitochondrial genes is associated with intron number and mirrors species evolution. . PLoS One, 10, e0131508.

6. Fox, G.E. (2010) Origin and evolution of the ribosome. Cold Spring Harb Perspect Biol., 3, a003483.

7. Knight, R.D., Freeland, S.J. and Landweber, L.F. (1999) Selection, history and chemistry: the three faces of the genetic code. Trends Biochem Sci., 24, 241-247.

8. Fournier, G.P., Andam, C.P., Alm, E.J. and Gogarten, J.P. (2011) Molecular evolution of aminoacyl tRNA synthetase proteins in the early history of life. Orig Life Evol Biosph., 41, 621-632.

9. Hendry, L.B., Bransome, E.D., Hutson, M.S. and Campbell, L.K. (1981) First approximation of a stereochemical rationale for the genetic code based on the topography and physicochemical properties of "cavities" constructed from models of DNA. Proc. Natl Acad. Sci. USA, 78, 7440-7444.

10. Foltan, J.S. (2008) tRNA genes and the genetic code. J Theor Biol., 253, 469-482.

11. Wong, J.T. (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci U S A. 72, 1909-12

12. Hernández, G., Proud, C., Preiss, T. and Parsyan, A. (2012) On the Diversification of the Translation Apparatus across Eukaryotes. Comparative and Functional Genomics, 2012, 256848.

13. Smith, S.E. and Read, D.J. (2008) Mycorrhizal Symbiosis. Third Edition ed. Academic Press, London.

14. Roy-Bolduc, A. and Hijri, M. (2011) The Use of Mycorrhizae to Enhance Phosphorus Uptake: A Way Out the Phosphorus Crisis. Journal of Biofertilizers & Biopesticides, 02.

15. Hijri, M. (2016) Analysis of a large dataset form field mycorrhizal inoculation trials on potato showed highly significant increase in yield. Mycorrhiza, 26, 209-214.

16. Bunn, R., Lekberg, Y. and Zabinski, C. (2009) Arbuscular mycorrhizal fungi ameliorate temperature stress in thermophilic plants. Ecology, 90, 1378-1388.

17. Zarik, L., Meddich, A., Hijri, M., Hafidi, M., Ouhammou, A., Ouahmane, L., Duponnois, R. and Boumezzough, A. (2016) Use of arbuscular mycorrhizal fungi to improve the drought tolerance of Cupressus Atlantica G. Comptes Rendus Biologies 339, 185-196.

18. Barea, J.-M., Azcon, R. and Azcon-Aguilar, C. (2002) Mycorrhizosphere interactions to improve plant fitness and soil quality. Antonie van Leeuwenhoek, 81, 343-351.

19. Iffis, B., St-Arnaud, M. and Hijri, M. (2017) Petroleum Contamination and Plant Identity Influence Soil and Root Microbial Communities While AMF Spores Retrieved from the Same

Plants Possess Markedly Different Communities. Front Plant Sci, 8, 1381.

20. Chen, B.D., Zhu, Y.G., Zhang, X.H. and Jakobsen, I. (2005) The influence of mycorrhiza on uranium and phosphorus uptake by barley plants from a field-contaminated soil. Environ Sci Pollut Res Int, 12, 325—331.

21. Sut, M., Boldt-Burisch, K. and Raab, T. (2016) Possible evidence for contribution of arbuscular mycorrhizal fungi (AMF) in phytoremediation of iron-cyanide (Fe-CN) complexes. Ecotoxicology, 25, 1260-1269.

22. Takemata, N. and Ohta, K. (2017) Role of non-coding RNA transcription around gene regulatory elements in transcription factor recruitment. RNA Biology, 2, 1-5.

23. Ward, M., McEwan, C., Mills, J.D. and Janitz, M. (2015) Conservation and tissue-specific transcription patterns of long noncoding RNAs. . Journal of Human Transcriptome, 1, 2-9.

24. Bazin, J., Baerenfaller, K., Gosai, S.J., Gregory, B.D., Crespi, M. and J., B.-S. (2017) Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. Proc Natl Acad Sci USA, 114, E10018-E10027.

## 4.8 Supplementary tables

**Supplementary Table 1 Relationship between amino acid properties and types of irregular codons: WW/WT/TW/TT**

To download Supplementary Table 1, please click here (at the same time, press Ctrl key)
If the link does not work, kindly copy and paste the following website address in your browser

www.codon.kr/thesis/chapter4/supple_table_1_relationship.xls

**Supplementary Table 2  P values from Mann Whitney U tests**

To download Supplementary Table 2, please click here (at the same time, press Ctrl key)

If the link does not work, kindly copy and paste the following website address in your browser

www.codon.kr/thesis/chapter4/supple_table_2_mann_whitney_test.xls

# Conclusions, discussions, and future work

## 5.1 Conclusions

### 5.1.1 SeSaMe

Among existing approaches for taxonomical classification of metagenome sequencing data, two methods, 16S rRNA based method and composition-based method, have been widely used. The former only classifies rRNA for estimating microbial diversity and the latter focuses on prokaryotic organisms and is inefficient for taxonomical classification of fungal sequences in metagenome data. In contrast to the composition-based method that relies on frequency of k-mers without considering their biological properties, the SeSaMe distinguished between CDS and non-CDS, identifies an open reading frame, and classifies a query sequence into one genus group out of 54 genera used as references. It provides a useful means for taxonomical classification for the WGS of the AMF grown *in-vivo*. In addition, P-value scores may serve as guidelines to help users to judge whether a query sequence comes from a predicted reference genus.

After applying the SeSaMe to the AMF WGS data, a user may use homology search methods to draw inference about functionality of a query sequence within a frame of reference genus. Therefore, the developed program enables users to identify gene reservoir of each reference genus and to study symbiotic interactions of AMF with their associated microbial community. In addition to the WGS data of the AMF, the program can be applied to environmental data from hyphosphere/ rhizosphere and to the WGS data of other organisms grown in soil.

### 5.1.2 SeSaMe PS Function

While existing bioinformatics tools for position-specific sequence annotation depend on alignment of a query sequence to annotated sequences in motif databases, SeSaMe PS Function identifies outliers of *Three codon DNA 9-mers* by applying PCA to comparative data created based on *trimer usage biases* of *Three codon DNA 9-mers* identified in a query sequence. The outliers with unique landscape patterns are often structurally and functionally distinctive sites. The program offers higher sensitivity toward undiscovered motifs with unknown functions and structures. Considering that existing bioinformatics databases and gene prediction programs are biased toward intensively studied model organisms and those culturable under laboratory conditions, the developed program may provide an efficient means for studying novel properties of genes from the relatively ancient fungi- AMF. The program is versatile for studying gene functions with a wide spectrum of applications, for example identifying a novel gene or studying new mechanisms in a known gene. SeSaMe and SeSaMe PS Function are freely available at www.fungalsesame.org.

### 5.1.3 Post

Recent studies have documented important regulatory roles of codon usage and codon context. Codon appears to play key roles both in transcriptional and translational processes. Considering that the standard genetic code table offers a limited view of codon property, we may be able to understand fundamental properties of the genetic codes by studying their origin. Mainstream hypotheses of genetic code origin support the ideas that codon position is a key determinant of codon assignment. The Post assigns a codon into amino acid group with respect to codon position. Although only a limited number of 3-D structures belonging to transcriptional and translational apparatuses are available, we can use a vast amount of sequencing data for studying the diversity of the transcriptional and the translational apparatuses. By applying the Post to the immense amount of omics data, we may be able to advance our knowledge of the flow of the genetic information. The Post is a versatile program with many applications. For example, it may enable us to conduct comparative analysis of transcriptional/ translational apparatuses and codon assignment of their target genes.

## 5.2 Discussions

### 5.2.1 SeSaMe

Genera belonging to the same taxonomic group demonstrated considerable similarities with each other in terms of *trimer usage bias* of *Three codon DNA 9-mer* derived from protein secondary structure. However, AMF showed a similar range of *trimer usage bias* most frequently with Firmicutes, Cyanobacteria and Rickettsia but much less frequently with Dikarya. With respect to fungal lifestyles, AMF and Dikarya have developed different survival strategies. While most fungi belonging to Dikarya have a variety of digesting enzymes to get carbon sources, AMF have evolved to form a mutualistic symbiosis with plant root from which they acquire carbon, and consequently lost some digesting enzymes (Tisserant et al 2013). Recent discoveries of fossil records and molecular analysis suggest that it appears likely that lichenization and de-lichenization events have independently occurred across different fungal lineages- not exclusively to Dikarya- during evolution, and rates of loss of lichenization were higher than those of gain (Hawksworth 2015). The same research group documented that some fungi might have formed symbioses that were similar to but not strictly lichenization not only with photosynthetic partners but also with non-photosynthetic bacterial and fungal partners. There have been several fossil records implicating AMF forming a lichenization like symbiosis. For example, a fossil evidence from marine deposits documented that Cyanobacteria and fungal filaments formed lichen like structures where the fungi produced spores recalling AMF rather than Ascomycota (Hawksworth 2015, Yuan et al 2005). A recent study has documented that putative horizontal gene transfer of class I ribonuclease III protein coding genes occurred from autotrophic cyanobacteria genomes to the AMF- *R. irregularis*. It is likely that other horizontal gene transfer events could have occurred between Firmicutes and AMF during evolution (Lee et al 2018).

The most preferred *three codon DNA 9-mers* of AMF, Fimicutes especially *Clostridium, Bacillus*, and *Leuconostoc*, Cyanobacteria particularly *Anabaena, Cyanobacterium*, and *Nostoc*, and *Rickettsia* were extremely biased towards A/T- ending codons in all three positions in the CDS Genus Specific DB; the similarity between AMF and the early diverged bacteria were shown in a majority of 7674 most preferred *Three codon DNA 9-mers* computed based on the entire CDS within a genus. John et al. (1975) hypothesized that *Rickettsia* is the mitochondrial progenitor 285 (John et al 1975). Whatley et al. hypothesized that hydrogenosome has multiple origins, evolved from *Clostridium* like bacterium, and is an ancestor of mitochondrion (1979). If the hypotheses hold true, they are the progenitors of endosymbiont organelles- hydrogenosome, mitochondrion, and plastid that AMF showed considerable similarities with. Xia discovered that distributions of A-ending codons are higher in mitochondria due to high ATP concentration (1996). Tisserant et al. documented that AMF contain an exceptionally higher number of High Mobility Group (HMG) box genes compared to Dikarya, *Rhizopus oryzae*, and *Dictyostelium discoideum* (2013), and recent studies have reported that roles of HMG box proteins include mitochondrial quality control (Tang et al 2011). Furthermore, recent studies have discovered the presence of heritable endosymbiont bacteria called MRE inside mycelia of AMF as well as evidences of horizontal gene transfer between MRE and their host- AM fungus (Naito et al 2015, Torres-Cortes 2015). Trachtenberg hypothesized that Mollicutes have been derived from Clostridia (1998). The hypotheses, the documented observations, the program test results, and the bias towards A/T ending codons all raise questions on the same subject- endosymbiont organelles: i) Have AMF harbored a considerable number of hydrogenosome like and/ or *Clostridium* like endosymbiont organelles? Had AMF once harbored and lost a large number of *Clostridium* like organelles during evolution and had their genes been integrated into AMF genomes via horizontal gene transfer?, ii) Given that AMF harbor a considerable number of associated microorganisms inside their mycelia (Bonfante 2003), to what extent intimately associated bacteria and endosymbionts have contributed to acquisitions of genes via horizontal transfer during AMF evolution? Numerous studies have documented that a majority of plants' and animals' mitochondrial genes have migrated to their nuclear genomes (Berg et al 2000, Farrelly et al 1983, Henze et al 2001). However, in the case of AMF, little is known about a degree of heterogeneity of nuclei within an isolate and the extent of the complexity of its genome organization.

It should be noted that the genera in the program are limited to those that are dominant in soil environment. Several studies documented that AMF are capable of performing dark CO2 fixation at a significant level (Bago et al 1999). CO2 fixation is an important means for microorganisms inhabiting in aquatic habits for producing organic matters (Santoro et al 2013). Rates of horizontal gene transfer between microorganisms are extremely higher in aquatic environment (Hermansson et al 1994). Some chytrid fungi have hydrogenosome (Hackstein et al 2007), and those living in aquatic environment may have capacity for CO2 fixation (Amon 1986). Therefore, comparison of AMF with early diverged fungi such as chytrid fungi may provide additional insights into AMF evolution.

It should be taken into consideration that publicly available gene prediction programs are highly biased towards genes that comply with a set of standard coding rules created based on most frequently studied organisms, and microbial sequences in public bioinformatics databases represent a small fraction of existing microorganisms. Therefore, considerable portions of AMF's genes may not be detectable with currently available programs. Consequently, genes detectable with conventional approaches are probably overrepresented in CDS lists of AMF. Due to the small sample size of AMF test set, we could not estimate a proportion of sequences that showed similarities with the early diverged bacteria in the entire repertoire of genes in *R. irregularis*, but made a relative comparison with Dikarya. In addition, given that AMF have high inter/ intra genome variations, we should investigate various species of AMF in order to assess whether the attributes shown in *R. irregularis* are common within Glomeromycotina or limited only to the individual fungus. It is plausible to assume that high inter/ intra genome variations of AMF reflect evolutionary history of genome diversification stimulated by adaptations associated with various types of symbioses (Joy 2013). The program has provided a platform for studying diversification of AMF genomes in the context of symbiosis associated adaptations; Comparative study on program results of WGS data of various AMF species may enable us to compare repertoires of proteins distinctive of the species and of its associated microorganisms, and provides insights into AMF genome diversification influenced by evolution of their symbiotic interactions.

### 5.2.2 SeSaMe PS Function

The Trimer Ref. DB was created based on secondary structures extracted from PDB where a majority of proteins with solved structures involve in binding activities. In PDB, A.A. trimers with polar amino acid are overrepresented. Therefore, Trimer Ref. DB is also biased toward A.A. trimers with polar amino acid.

*Trimer usage biases* of the Genus Specific DB were calculated based on all CDS within a genus irrespective of species, gene family and domain, polycistronic mRNA, regulon, and molecular evolutionary origin. Additionally, overlapping ORFs adds another layer of complexity to interpretation of the loading clusters. Therefore, some *three codon DNA 9-mers* within Genus Specific DB may have large intra variations. They may also have large inter variations due in part to that folding signatures of three codon DNA 9-mer vary widely across different taxonomic groups. For such reasons, loading clusters with 80% components may have higher detection capacity for outliers- functionally/ structurally distinctive sites. Due in part to the same reasons, the First/Second components represent more complicated properties than a single factor such as protein secondary structure or solvent accessibility, and so interpretation of loading clusters is not straight forward. We may be able to add a feature- reference map connecting *Three codon DNA 9-mer* with its regulatory role(s)- in future when more sequences are annotated with roles of codon usage and of codon context.

We created a conversion rule, in other words based on which properties we group similar amino acids into an amino acid characteristics, with respect to general biochemical properties to acquire comparability of amino acids among various taxonomic groups. The program calculates codon usage bias within amino acid characteristics rather than within amino acid (Berman 2000). Both *three codon usage* and *trimer usage bias* showed comparable prediction capacity in taxonomical classification, which may suggest that they may provide coherent explanation to describe the undiscovered regulatory mechanisms.

### 5.2.3 Post

AMF harbor a large number of microorganisms inside of their spores and mycelia, some of which are obligate endosymbionts. Considering that most proteins and RNAs involving in transcriptional and translational processes of mitochondria are different from those of nuclear genome, the heterogeneous transcriptional and translational apparatuses may have, at least partially, contributed to the mitochondrial specific codon assignments. If AMF contain different versions of transcriptional and translational apparatuses specific for their endosymbionts, their codon assignments may not abide by the standard genetic code table and vary considerably. Developing endosymbiont-specific Post and incorporating it into the SeSaMe may enable us to identify novel genes from endosymbionts in the WGS of the AMF and to study their symbiotic roles in AMF adaptation.

## 5.3 Future work

### 5.3.1 Heterogeneous regulators of DNA replication, transcription, and translation

A large number of researchers have documented that AMF inoculation is a promising candidate for phytoremediation that cleans up contaminated soil (Chanda et al 2014, Marchand et al 2017 (b), Iffis et al 2014). AMF and their associated microbial community need to adapt to environmental stresses in contaminated soil. Recent studies reported that biases of codon usage, codon context, and amino acid composition play important roles in microbial adaptation in response to abiotic stresses (Su et al 2016, Ding et al 2012, Paul et al 2008, Sanjukta et al 2012). The Genus Specific DB of the developed programs was created based on bias information of codon usage and codon context obtained from completely sequenced genomes of microorganisms mostly grown in optimal condition. Due to lack of completely sequenced genomes of microorganisms sampled from the phytoremediation field study, we will need to incorporate the new factors (environmental stresses) into the developed program. We will need to understand the dynamics underlying sequence modification during microbial adaptation in response to environmental stresses, especially with respect to transcriptional and translational apparatuses.

#### 5.3.1.1 Different types of DNA polymerase in response to environmental stresses

It is well documented that microorganisms have multiple heterogeneous DNA polymerase[31] subpopulations, some of which have error-prone activities to increase adaptability. For example, the

SOS system responds to DNA damages caused by environmental stresses. SOS-induced DNA polymerases are capable of repairing damaged DNA sequences; they produce DNA sequences with high rates of mutations to increase the chances of survival under environmental stress (Radman 1999, Yeiser et al 2002).

Bacteria increase mutational rates under environmental stress via mutagenesis mechanisms. Rosenberg et al documented that a network with 93 genes, involving in sensing and activating stress response system, promoted mutagenesis under stress (2014). They also suggested that a local cluster of mutations and mutational hotspots implied that some mutations were not random. The assumption of non-random mutation leads to another important question; if it is alternative regulator that produces mutations, what rules they may have other than Watson-Crick base pairs? A number of researchers have employed molecular biology methods to identify gene regulatory networks in response to environmental stresses and to understand what are happening inside of regulators during DNA repair, mRNA synthesis, and protein synthesis. We will need to study the heterogeneous regulators to improve the developed programs tailored to analyze metagenome data from phytoremediation projects.

#### 5.3.1.2 Heterogeneous transcriptional regulator, RNA polymerase, in response to environmental stresses

Bacterial RNA polymerase[32] has subunits called sigma. Sigma factors[33], belonging to sigma subunit, play key roles in recognition of promotors in transcription initiation process (Saecker et al 2011). A microorganism contains a number of heterogeneous sigma factors that are induced in response to environmental stresses. They not only provide functionally different RNA polymerase subpopulations but also involve in regulation of a regulon- a set of a large number of target genes. A major sigma factor σ70, that is also called RpoD, manages most of transcription activities during active growth while alternative sigma factors, such as sigma factor σ54 that is also called RpoN, regulate a variety of adaptive responses according to environmental stresses in E. coli. RNA polymerase with sigma factor σ54 requires enhancer[34] recognition and a specialized transcription activator[35] with ATP hydrolysis in initiation process (Paget 2015, Zhang et al 2015). With microarray analyses, sigma factor 54s were shown to regulate expressions of 100s of target genes in *Borrelia burgdorferi* (Fisher et al 2005). It suggests that prokaryotes fine tune their activities via transcriptional regulator with respect to cellular stress. In addition, error-prone DNA polymerases can be induced by sigma factor RpoS, which promotes mutagenesis (Foster et al 2007). It remains unknown what roles the structural change of RNA polymerase plays in terms of codon assignment. Comparison and contrast of sequence properties of regulons regulated by different heterogeneous regulator subpopulations will promote understanding of the association between codon assignment and a type of stress.

#### 5.3.1.3 Heterogeneous translational regulators in response to environmental stresses

Pseudomonas sp. UW4 had multiple copies of rRNA genes whose operons[36] were differentially activated under varying environmental situations in terms of temperature, nutrient availability, and developmental stage (Duan et al 2014). In order to differentially express rRNA operons under varying environmental conditions, it is possible that transcription factors, activators and repressors, may play important roles. Their operons had unique promotors, some of which were similar to those of heat shock proteins recognized by heat shock sigma factor. It implies a possibility that alternative transcriptional regulator may control overall cellular processes including alternative translational regulator. Another researcher has implied a view with slight differences; while heterogeneous transcriptional regulator subpopulation regulates expressions of regulon under relatively severe and prolonged environmental stresses, heterogeneous translational regulator subpopulation may provide an express lane in managing immediate environmental stresses (Sauert et al 2015).

rrnDB is a public database that provides rRNA operon copy number for bacteria and archaea via web service at rrndb.umms.med.umich.edu (Stoddard et al. 2014). It also provides links to completely sequenced genomes used in their calculation. The tool can be useful for studying the association between rRNA operon and microbial adaptation to environmental stresses. To investigate effects of heterogeneity of translational regulator subpopulations within a single isolate in context of codon assignment, sequence comparison of rRNA genes within a single isolate as well as among closely related taxa will be required. In addition, it seems also important to investigate whether there are associations between hypervariable regions of heterogeneous regulators and environmental stimuli.

### 5.3.2 Incorporation of heterogeneous regulator into analysis of mycorrhizosphere microbiota sampled from stressed environments

Most of completely sequenced genomes in public databases are from microorganisms grown in their optimal growth conditions. Analysis of environmental data sampled from stressful environment proves to be challenging, because we need to make inference about evolved microbiota based on their prior sequences obtained when they were under their optimal growth conditions.

Microorganism's adaptation process varies widely according to its morphological, physiological, and biochemical characteristics. For example, the impact of environmental stress becomes stronger when a microorganism does not have protective components such as cell wall. Fadiel et al experimented on *Mycoplasma genitalium, M. pneumoniae*, and *Ureaplasma urealyticum* to study properties of codon usages using complete transcriptomes of the microorganisms (2005). All of them belong to Mollicutes that lack of cell wall. Comparative study revealed that *M. genitalium and U. urealyticum* exhibited more similarity in transcriptome structure although *M. genitalium* and *M. pneumoniae* were phylogenetically more closely related to each other. Cell wall and cell membranes are one of the most important contributors that protect prokaryotes from environmental stresses. The study suggests that an environmental factor has stronger impact on microorganisms compared to phylogeny, especially if they are not equipped with cellular components that help them to cope with environmental stresses. Therefore, a degree of impact of an environmental stress on microorganism is assumed to vary widely across different taxa. In addition, different types of environmental factors are considered to have varying degrees of effect on microbial adaptation across taxa. We will need to employ computational biology and bioinformatics approach to conduct comparative study on how the heterogeneous regulator subpopulations affect codon assignment in their target genes across taxonomic groups. Then, we may be able to create taxo-specific Posts or one similar to it, which we will need to incorporate into the developed programs, SeSaMe and SeSaMe PS Function.

# References (Definitions, Chapter 1, and Chapter 5)

## Definitions

1. Antonets KS, Nizhnikov AA. A Novel Algorithm to Assess Compositional Biases in Protein Sequences. Evol Bioinform Online. 2013 Jul 11;9:263-73

2. Behura SK, Severson DW. Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes. PLoS One. 2012;7(8):e43111

3. R.R. Burgess, in Encyclopedia of Genetics, https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/sigma-factor

## Chapter 1

1. Ahmad S, Selvapandiyan A, Bhatnagar RK. A protein-based phylogenetic tree for gram-positive bacteria derived from hrcA, a unique heat-shock regulatory gene. Int J Syst Bacteriol. 1999 Oct;49 Pt 4:1387-94.

2. Akashi H. Synonymous Codon Usage in Drosophila Melanogaster: Natural Selection and Translational Accuracy. Genetics 1994; 136(3): 927–35.

3. Angov E. Codon usage: Nature's roadmap to expression and folding of proteins. Biotechnol. j. 2011; 6: 650-9.

4. Baeza M., Alcaíno J., Barahona S., Sepúlveda D., and Cifuentes V. “Codon Usage and Codon Context Bias in Xanthophyllomyces Dendrorhous.”BMC Genomics 16, no. 1 (April 13, 2015).

5. Barea, J.-M., Azcón, R. and Azcón-Aguilar, C. (2002) Mycorrhizosphere interactions to improve plant fitness and soil quality. Antonie van Leeuwenhoek 81, 343–351.

6. Bartoszewski R, Króliczewski J, Piotrowski A, Jasiecka AJ, Bartoszewska S, Vecchio-Pagan B et al. Codon bias and the folding dynamics of the cystic fibrosis transmembrane conductance regulator. Cellular & Molecular Biology Letters 2016; 21: 23.

7. Bazin J, Baerenfaller K, Gosai SJ, Gregory BD, Crespi M, Bailey-Serres J. (2017) Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. Proc Natl Acad Sci U S A. 114(46):E10018-E10027

8. Beauchemin S. and Simard R.R. (1999) Soil phosphorus saturation degree: review of some indices and their suitability for P management in Quebec, Canada. Canadian Journal of Soil Science 79, 615–625.

9. Beaudet D, Terrat Y, Halary S, de la Providencia IE, Hijri M. Mitochondrial genome rearrangements in Glomus species triggered by homologous recombination between distinct mtDNA haplotypes. Genome Biol Evol. 2013A;5(9):1628–43.

10. Bécard G, Fortin JA. Early events of vesicular-arbuscular mycorrhiza formation on Ri T-DNA transformed roots. New Phytol. 1988; 108: 211-8

11. Behura S. K., and Severson D. W. “Comparative Analysis of Codon Usage Bias and Codon Context Patterns between Dipteran and Hymenopteran Sequenced Genomes.”PLoS ONE 7, no. 8 (August 17, 2012).

12. Bell T, Stefani FOP, Abram K, Champagne J, Yergeau E, Hijri M, & St-Arnaud M (2016). A diverse soil microbiome degrades more crude oil than specialized bacterial assemblages obtained in culture. Applied and Environmental Microbiology 82(18): 5530-5541.

13. Bianciotto V, Lumini E, Orgiazzi A, Borriello R, Bonfante P. Metagenomics applied to arbuscular mycorrhizal fungal communities. In: Metagenomics: Current Innovations and Future Trends. Marco D (ed). Norfolk, UK: Caister Academic Press pp. 161-78.

14. Bonfante P. Plants, mycorrhizal fungi and endobacteria: a dialog among cells and genomes. The Biological bulletin 2003; 204: 215-20.

15. Bonfante P and Anca IA. Plants, mycorrhizal fungi, and bacteria: a network of interactions. Annu Rev Microbiol. 2009; 63: 363-83.

16. Boon E, Zimmerman E, St-Arnaud M & Hijri M (2013). Allelic differences within and among sister spores of the arbuscular mycorrhizal fungus Glomus etunicatum suggest segregation at sporulation. PLoS One 8(12): e83301.

17. Boon E, Halary S, Bapteste E & Hijri M (2015). Studying Genome Heterogeneity within the Arbuscular Mycorrhizal Fungal Cytoplasm. Genome Biology and Evolution, 7(2): 505-521.

18. Brandao MM, Silva-Filho MC. (2011) Evolutionary History of Arabidopsis thaliana Aminoacyl-tRNA Synthetase Dual-Targeted Proteins. Mol. Biol. Evol. 28(1):79–85.

19. Bunn R, Lekberg Y, and Zabinski C. Arbuscular Mycorrhizal Fungi Ameliorate Temperature Stress in Thermophilic Plants. Ecology 2009; 90(5): 1378–88.

20. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK and et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010 May;7(5):335-6.

21. Carbone A, Képès F, Zinovyev A. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. Mol. Biol. Evol. 2005 22(3): 547–561. 98. dos

22. Chanda D, Sharma GD, Jha DK, Hijri M. Associations of arbuscular mycorrhizal (AM) fungi in the phytoremediation of trace metal (TM) contaminated soils. Journal of Research in Biology 2014; 4: 1247−63.

23. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci U S A, 101 (2004), pp. 3480-3485.

24. Chevance F. F. V., Le Guyon S., and Hughes K. T. “The Effects of Codon Context on In Vivo Translation Speed.”PLoS Genetics 10, no. 6 (June 5, 2014). doi:10.1371/journal.pgen.1004392.

25. Coenye, Tom, and Peter Vandamme. “Simple Sequence Repeats and Compositional Bias in the Bipartite Ralstonia Solanacearum GMI1000 Genome.”BMC Genomics 4, no. 1 (March 17, 2003): 10.

doi:10.1186/1471-2164-4-10.

26. Costafreda I. M., Pérez-Rodriguez F. J., D'Andrea L., Guix S., Ribes E., Bosch A., and PintóR. M. "Hepatitis A Virus Adaptation to Cellular Shutoff Is Driven by Dynamic Adjustments of Codon Usage and Results in the Selection of Populations with Altered Capsids."Journal of Virology 88, no. 9 (May 2014): 5029–41.

27. Cruz, A. F., Horii, S., Ochiai, S., Yasuda, A., and Ishii, T. (2008). Isolation and analysis of bacteria associated with spores of Gigaspora margarita. J. Appl. Microbiol. 104, 1711-1717

28. Cruz AF & Ishii T (2012) Arbuscular mycorrhizal fungal spores host bacteria that affect nutrient biodynamics and biocontrol of soil-borne plant pathogens. Biol Open 1: 52–57.

29. Declerck S, Strullu D, Fortin AJ. In vitro culture of mycorrhizas. Berlin; New York: Springer; 2005

30. Del Campo C., Bartholomäus A., Fedyunin I., and Ignatova Z. "Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function."PLoS Genetics 11, no. 10 (October 2015): e1005613.

31. de la Providencia I, Stéfani FOP, Labridy M, St-Arnaud M, & Hijri M (2015). Arbuscular mycorrhizal fungal diversity associated with Eleocharis obtusa and Panicum capillare growing in an extreme petroleum hydrocarbon-polluted sedimentation basin. FEMS Microbiology Letters, 362(12):fnv081.

32. Ding Y, Cai Y, Han Y, Zhao B. Comparison of the structural basis for thermal stability between archaeal and bacterial proteins. Extremophiles. 2012 Jan;16(1):67-78.

33. Ermolaeva MD. Synonymous codon usage in bacteria. Curr Issues Mol Biol. 2001 Oct;3(4):91-7.

34. Gao Feng, and Chun-Ting Zhang. Comparison of Various Algorithms for Recognizing Short Coding Sequences of Human Genes. Bioinformatics (Oxford, England) 20, no. 5 (March 22, 2004): 673–681. doi:10.1093/bioinformatics/btg467.

35. Glaeser SP, Kämpfer P. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Syst Appl Microbiol. 2015 Jun;38(4):237-45.

36. Golding GB, Gupta RS. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol Biol Evol. 1995 Jan;12(1):1-6.

37. Grantham R, Gautier C, Gouy M, Jacobzone M, and Mercier R. Codon Catalog Usage Is a Genome Strategy Modulated for Gene Expressivity. Nucleic Acids Research 1981; 9(1): r43–74.

38. Gulbis, N. and Robinson-Boyer, Louisa and Robinson, Gary K. (2013) Studying the microbiome of AMF cultivated in vitro. Aspects of Applied Biology (Positive Plant Microbial Interactions: Their role in maintaining sustainable and natural ecosystems), 120 . pp. 71-76. ISSN 0265-1491.

39. Gupta A, Sharma VK. Using the taxon-specific genes for the taxonomic classification of bacterial genomes. BMC Genomics. 2015 May 20;16:396.

40. Hafidi M, Qaddoury A, Duponnois R, Wipf D, Hijri M, & Ba A (2015). International Congress on Mycorrhizae: mycorrhizal symbiosis a key factor for improving plant productivity and ecosystems restoration. Mycorrhiza 25(8): 673-674.

41. Handa Y, Nishide H, Takeda N, Suzuki Y, Kawaguchi M, Saito K. (2015) RNA-seq Transcriptional Profiling of an Arbuscular Mycorrhiza Provides Insights into Regulated and Coordinated Gene

Expression in Lotus japonicus and Rhizophagus irregularis. Plant Cell Physiol: 56(8):1490-511.

42. Harigaya Y, Parker R. The link between Adjacent Codon Pairs and mRNA Stability. BMC Genomics 18, no. 1 (May 10, 2017): 364.

43. Hassan SE, Liu A, Bittman S, Forge, TA Hunt DE, Hijri M, & St-Arnaud M (2013). Impact of 12-year field treatments with organic and inorganic fertilizers on crop productivity and mycorrhizal community structure. Biology and Fertility of Soils 49: 1109-1121.

44. Hassan SE, Bell T, Stefani FOP, Denis D, Hijri M, Yergeau E & St-Arnaud M (2014). Contrasting the community structure of arbuscular mycorrhizal fungi from hydrocarbon-contaminated and uncontaminated soils following willow (Salix spp. L.) planting. PLoS ONE, 9(7): e102838.

45. Herriges MJ, Swarr DT, Morley MP, Rathi KS, Peng T, Stewart KM et al (2014). Long noncoding RNAs are spatially correlated with transcription factors and regulate lung development. Genes Dev. 28(12): 1363–1379.

46. Hershberg R, Petrov DA. (2008). Selection on codon bias. Annual Review of Genetics. 42: 287-99.

47. Hijri M, Redecker D, Petetot JAM-C, Voigt K, Wöstemeyer J, and Sanders IR. Identification and isolation of two ascomycete fungi from spores of the arbuscular mycorrhizal fungus. Appl Environ Microbiol 2002; 68: 4567-73.

48. Hijri M, Niculita H, & Sanders IR (2007). Molecular characterization of chromosome termini of the arbuscular mycorrhizal fungus Glomus intraradices (Glomeromycota). Fungal Genetics and Biology 44: 1380-1386.

49. Hijri M. Analysis of a Large Dataset of Mycorrhiza Inoculation Field Trials on Potato Shows Highly Significant Increases in Yield. Mycorrhiza 2016; 263): 209–14.

50. Iffis B, St-Arnaud M & Hijri M (2014). Bacteria associated with arbuscular mycorrhizal fungi within roots of plants growing in a soil highly contaminated with aliphatic and aromatic petroleum hydrocarbons. FEMS Microbiology Letters, 358: 44-54.

51. Iffis B, St-Arnaud M & Hijri M (2016). Petroleum hydrocarbon contamination, plant identity and arbuscular mycorrhizal fungal community determine assemblages of the AMF spore-associated microbes. Environmental Microbiology 18(8): 2689-2704.

52. Iffis B, St-Arnaud M & Hijri M (2017). Petroleum Contamination and Plant Identity Influence Soil and Root Microbial Communities While AMF Spores Retrieved from the Same Plants Possess Markedly Different Communities. Frontiers in Plant Science 8: 1381.

53. Jacobson G. N., and Clark P. L. “Quality over Quantity: Optimizing Co-Translational Protein Folding with Non-’optimal’Synonymous Codons.”Current Opinion in Structural Biology 38 (June 2016): 102–10. doi:10.1016/j.sbi.2016.06.002.

54. Jargeat P, Cosseau C, Ola’h B, Jauneau A, Bonfante P, Batut J & Becard G (2004) Isolation, free-living capacities, and genome structure of “Candidatus Glomeribacter gigasporarum”, the endocellular bacterium of the mycorrhizal fungus Gigaspora margarita. J Bacteriol 186: 6876–6884.

55. Jeffery S, Gardi C, Jones A, Montanarella L, Marmo L, Miko L et al. European Atlas of Soil Biodiversity. Luxembourg: Publications Office of the European Union; 2010.

56. Johansson JF, Paul LR, Finlay RD. Microbial interactions in the mycorrhizosphere and their significance for sustainable agriculture. FEMS Microbiol Ecol. 2004 Apr 1;48(1):1-13.

57. Karlin, S., J. Mrázek, and A. M. Campbell. “Compositional Biases of Bacterial Genomes and Evolutionary Implications.”Journal of Bacteriology 179, no. 12 (June 1, 1997): 3899–3913.

58. Khabou B., Olfa S., Gargouri L., Mkaouar-Rebai E., Keskes L., Hachicha M., and Fakhfakh F. “In Silico Investigation of the Impact of Synonymous Variants in ABCB4 Gene on mRNA Stability/Structure, Splicing Accuracy and Codon Usage: Potential Contribution to PFIC3 Disease.”Computational Biology and Chemistry 65 (2016): 103–9.

59. Kim M., Lee K.H., Yoon S.W., Kim B.S., Chun J., and Yi H.. Analytical Tools and Databases for Metagenomics in the Next-Generation Sequencing Era. Genomics Inform. 2013 Sep; 11(3): 102–113.

60. Klappenbach J.A., Dunbar J.M., and Schmidt T.M. rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria. Appl. Environ. Microbiol. April 2000 vol. 66 no. 4:1328-1333

61. Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, and et al. (2013) Towards a unified paradigm for sequence-based identification of fungi. Mol Ecol 22:5271–5277

62. Komar A. A. “The Yin and Yang of Codon Usage.”Human Molecular Genetics 25, no. R2 (October 1, 2016): R77–85.

63. Kuhn, G. 2003. Organisation of Genetic Variation in Multinucleate Arbuscular Mycorrhizal Fungi. University of Lausanne, Lausanne

64. Lee J, Young JP. 2009. The mitochondrial genome sequence of the arbuscular mycorrhizal fungus Glomus intraradices isolate 494 and implications for the phylogenetic placement of Glomus. New Phytol. 183: 200–211.

65. Lee S, Weon S, Lee S, and Kang C. Relative Codon Adaptation Index, a Sensitive Measure of Codon Usage Bias. Evolutionary Bioinformatics Online 2010; 6: 47.

66. Lecomte J, St-Arnaud M, & Hijri M (2011). Isolation and identification of soil bacteria growing at the expense of arbuscular mycorrhizal fungi. FEMS Microbiology Letters 317: 43-51.

67. Lindahl BD, Nilsson RH, Tedersoo L et al. (2013) Fungal community analysis by high-throughput sequencing of amplified markers – a user’s guide. New Phytologist, 199, 288–299.

68. MacDonald RM, Chandler MR, Mosse B. The occurrence of bacterium-like organelles in vesicular-arbuscular mycorrhizal fungi. New Phytol. 1982; 90, no. 4: 659–63.

69. Marchand M, Hogland W, Kaczala F, Jani Y, Marchand L, Augustsson A, & Hijri M (2016). Effect of Medicago sativa L. and compost on organic and inorganic pollutant removal from a mixed contaminated soil and risk assessment using ecotoxicological tests. International Journal of Phytoremediation 18(11): 1136-1147.

70. Marchand Ma, St-Arnaud M, Hogland W, Bell TH, & Hijri M (2017) Petroleum biodegradation capacity of bacteria and fungi isolated from petroleum-contaminated soil. International Biodeterioration & Biodegradation 116: 48-57.

71. Marchand Mb, Mench M, Jani Y, Kaczala F, Notini P, Hijri M & Hogland M (2017) Field pilot scale aided-phytoremediation of a soil contaminated by petroleum hydrocarbons and metals. Science of the

Total Environment (In Press).

72. Marleau J, Dalpe Y, St-Arnaud M, & Hijri M (2011). Spore development and nuclear inheritance in arbuscular mycorrhizal fungi. BMC Evolutionary Biology 11: 51.

73. Mathy NW, Chen XM. (2017). Long non-coding RNAs (lncRNAs) and their transcriptional control of inflammatory responses. J Biol Chem. 292(30):12375-12382.

74. McCarthy C., Carrea A., and Diambra L. Bicodon Bias Can Determine the Role of Synonymous SNPs in Human Diseases. BMC Genomics 18, no. 1 (March 13, 2017): 227.

75. McDonald JH. Patterns of Temperature Adaptation in Proteins from the Bacteria Deinococcus radiodurans and Thermus thermophilus. Mol. Biol. Evol. 2001 18(5):741–749.

76. Merget B., Koetschan C., Hackl T., Förster F., Dandekar T., Müller T., Schultz J., and Wolf M. The ITS2 Database. J Vis Exp. 2012; (61): 3806.

77. More RP, Purohit HJ. The Identification of Discriminating Patterns from 16S rRNA Gene to Generate Signature for Bacillus Genus. J Comput Biol. 2016 Aug;23(8):651-61.

78. Nadimi M, Beaudet D, Forget L, Hijri M, Lang BF. Group I intron-mediated trans-splicing in mitochondria of Gigaspora rosea and a robust phylogenetic affiliation of arbuscular mycorrhizal fungi with Mortierellales. Mol Biol Evol. 2012;29(9):2199–210.

79. Nadimi M, Stefani FOP, Hijri M. The Mitochondrial genome of the Glomeromycete Rhizophagus sp. DAOM 213198 reveals an unusual organization consisting of two circular chromosomes. Genome Biol Evol. 2015;7(1):96–105.

80. Nadimi M, Daubois L, & Hijri M (2016). Mitochondrial comparative genomics and phylogenetic signal assessment of mtDNA among arbuscular mycorrhizal fungi. Molecular Phylogenetics and Evolution, 98: 74-83.

81. Naito M, Morton JB, and Pawlowska TE. Minimal genomes of mycoplasma-related endobacteria are plastic and contain host-derived genes for sustained life within Glomeromycota. Proc Natl Acad Sci U S A 2015; 112: 7791-6.

82. Nguyen N., Warnow T., Pop M., and White B. "A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity," Npj Biofilms And Microbiomes, v.2, 2016.

83. Paul S, Bag SK, Das S, Harvill ET, Dutta C. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biol. 2008 Apr 9;9(4):R70.

84. Rajendhran J1, Gunasekaran P. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res. 2011 Feb 20;166(2):99-110.

85. Raymond-Bouchard I, Goordial J, Zolotarov Y, Ronholm J, Stromvik M, Bakermans C, Whyte LG. Conserved genomic and amino acid traits of cold adaptation in subzero-growing Arctic permafrost bacteria. FEMS Microbiology Ecology. 2018 Vol. 94, No. 4

86. Reddy, Rachamalla Maheedhar, Monzoorul Haque Mohammed, and Sharmila S Mande. “TWARIT: An Extremely Rapid and Efficient Approach for Phylogenetic Classification of Metagenomic Sequences.”Gene 505, no. 2 (September 1, 2012): 259–265.

87. Roy-Bolduc A, & Hijri, M (2011). The use of mycorrhizae to enhance phosphorus uptake: a way out

the phosphorus crisis. Journal of Biofertilizers & Biopesticides, 2: 104.

88. Sanjukta R, Farooqi MS, Sharma N, Rai A, Mishra DC, Singh DP. Trends in the codon usage patterns of Chromohalobacter salexigens genes. Bioinformation. 2012;8(22):1087-95.

89. Sarma R.K., Saikia Ratul, and Talukdar N.C. (2017) Mitochondrial DNA Based Molecular Markers in Arbuscular Mycorrhizal Fungi (AMF) Research. In: Molecular Markers in Mycology: Diagnostics and Marker Developments. Bhim Pratap Singh, Vijai Kumar Gupta (Eds). Cham, Switzerland : Springer.

90. Schieweck R., Popper B., and Kiebler M. A. “Co-Translational Folding: A Novel Modulator of Local Protein Expression in Mammalian Neurons?”Trends in Genetics: TIG 32, no. 12 (December 2016): 788–800.

91. Schloss PD, Westcott SL, Ryabin T et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology, 75, 7537–7541.

92. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Bergeron MJ, Hamelin RC, Vialle A, and Fungal Barcoding Consortium (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Science (PNAS). 109: 6241-6246.

93. Smith SE and Read DJ. (2008) Mycorrhizal symbiosis. Cambridge: Academic Press.

94. Sokolski S., Dalpé Y., and Piché Y. Phosphate Transporter Genes as Reliable Gene Markers for the Identification and Discrimination of Arbuscular Mycorrhizal Fungi in the Genus Glomus. Appl Environ Microbiol. 2011 Mar; 77(5): 1888–1891.

95. Stackebrandt, E., Rainey, F. A. & Ward-Rainey, N. L. (1997). Proposal for a new hierarchic classification system, Actinobacteria classis nov. Int J Syst Bacteriol47, 47949 1.

96. Stanke, Mario, Rasmus Steinkamp, Stephan Waack, and Burkhard Morgenstern. “AUGUSTUS: A Web Server for Gene Finding in Eukaryotes.”Nucleic Acids Research 32, no. suppl 2 (July 1, 2004): W309–W312. doi:10.1093/nar/gkh379.

97. Sun M, Kraus WL. (2015) From discovery to function: the expanding roles of long noncoding RNAs in physiology and disease. Endocr Rev. 36(1):25-64

98. Supek F, Skunca N, Repar J, Vlahovicek K, Smuc T. Translational selection is ubiquitous in prokaryotes. PLoS Genet. 2010 6(6): e1001004.

99. Suzuki H, Saito R, Tomita M. Measure of synonymous codon usage diversity among genes in bacteria. BMC Bioinformatics. 2009 Jun 1;10:167.

100. Su Y, Jiang X, Wu W, Wang M, Hamid MI, Xiang M, Liu X. Genomic, Transcriptomic and Proteomic Analysis Provide Insights into the Cold Adaptation Mechanism of the Obligate Psychrophilic Fungus Mrakia psychrophila. G3 (Bethesda). 2016 Sep 15. pii: g3.116.033308.

101. Ter-Hovhannisyan, Vardges, Alexandre Lomsadze, Yury O. Chernoff, and Mark Borodovsky. “Gene Prediction in Novel Fungal Genomes Using an Ab Initio Algorithm with Unsupervised Training.”Genome Research 18, no. 12 (December 2008): 1979–1990.

102. Tisserant E, Kohler A, Dozolme-Seddas P, Balestrini R, Benabdellah K, Colard A et al. The

Transcriptome of the Arbuscular Mycorrhizal Fungus Glomus Intraradices (DAOM 197198) Reveals Functional Tradeoffs in an Obligate Symbiont. The New Phytologist 2012; 193(3): 755–69.
103. Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R et al. Genome of an Arbuscular Mycorrhizal Fungus Provides Insight into the Oldest Plant Symbiosis. Proc Natl Acad Sci U S A 2013; 110(50): 20117–22.
104. Torres-Cortes G, Ghignone S, Bonfante P, and SchuSsler A. Mosaic genome of endobacteria in arbuscular mycorrhizal fungi: Transkingdom gene transfer in an ancient mycoplasma-fungus association. Proc Natl Acad Sci U S A 2015; 112: 7785-90.
105. Ueno K, Ibarra M, Gojobori T. Structural adaption of extremophile proteins to the environments with special reference to hydrophobic networks. Ecol. Genet. Genomics. 2016 1:1–5, 95.
106. Vangelisti A, Natali L, Bernardi R, Sbrana C, Turrini A, Hassani-Pak K, Hughes D, Cavallini A, Giovannetti M, Giordani T. (2018) Transcriptome changes induced by arbuscular mycorrhizal fungi in sunflower (Helianthus annuus L.) roots. Sci Rep: 8;8(1):4
107. Xu Y., Ma P., Shah P., Rokas A., Liu Y., and Johnson C. H. Non-Optimal Codon Usage Is a Mechanism to Achieve Circadian Clock Conditionality. Nature 495, no. 7439 (March 7, 2013): 116–20. doi:10.1038/nature11942.
108. Yang J.. “Does mRNA Structure Contain Genetic Information for Regulating Co-Translational Protein Folding?”Zoological Research 38, no. 1 (January 18, 2017): 36–43.
109. Zarik L, Meddich A, Hijri M, Hafidi M, Ouhammou A, Ouahmane L, Duponnois R, & Boumezzough A (2016) Use of arbuscular mycorrhizal fungi to improve the drought tolerance of Cupressus atlantica G. C R Biol. pii: S1631-0691(16) 30032-4.
110. Zhao F., Yu C., and Liu Y. Codon Usage Regulates Protein Structure and Function by Affecting Translation Elongation Speed in Drosophila Cells. Nucleic Acids Research, June 5, 2017.
111. Zhou M., Guo J., Cha J., Chae M., Chen S., Barral J. M., Sachs M. S., and Liu Y. “Non-Optimal Codon Usage Affects Expression, Structure and Function of Clock Protein FRQ.”Nature 495, no. 7439 (March 7, 2013): 111–15.
112. Zimmerman E, St-Arnaud M, & Hijri M (2009). Sustainable Agriculture and the Multigenomic Model: How Advances in the Genetics of Arbuscular Mycorrhizal Fungi will Change Soil Management Practices (p. 269-287). In: Molecular Plant-microbe interactions. Bouarab K. Brisson N. and Daayf F. (Eds). CAB International.

## Chapter 5

1. Amon JP. Growth of marine chytrids at ambient nutrient levels. In: Moss ST, editor. The biology of marine fungi. London: Cambridge University Press; 1986, p. 70–80.
2. Bago B, Pfeffer PE, Douds DDJ, Brouillette J, Bécard G, Shachar-Hill Y. Carbon metabolism in spores of the arbuscular mycorrhizal fungus Glomus Intraradices as revealed by nuclear magnetic resonance spectroscopy. Plant Physiol 1999; 121: 263–72.

3. Berg OG, Kurland CG. Why mitochondrial genes are most often found in nuclei. Mol Biol Evol 2000; 17: 951–61.

4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000; 28: 235−42.

5. Bonfante P. Plants, mycorrhizal fungi and endobacteria: a dialog among cells and genomes. Biol Bull 2003; 204: 215–20.

6. Duan J, Reimer L, Heikkila JJ, Glick BR. Differential expression of the seven rRNA operon promoters from the plant growth-promoting bacterium Pseudomonas sp. UW4. FEMS Microbiol Lett. 2014 Dec; 361(2):181-9.

7. Fadiel A, Lithwick S, Naftolin F. The influence of environmental adaptation on bacterial genome structure. Lett Appl Microbiol. 2005;40(1):12-8.

8. Farrelly F, Butow RA. Rearranged mitochondrial genes in the yeast nuclear genome. Nature 1983; 301: 296–301.

9. Fisher MA, Grimm D, Henion AK, Elias AF, Stewart PE, Rosa PA, Gherardini FC. Borrelia burgdorferi sigma54 is required for mammalian infection and vector transmission but not for tick colonization. Proc Natl Acad Sci U S A. 2005 Apr 5;102(14):5162-7.

10. Foster PL. Stress-induced mutagenesis in bacteria. Crit Rev Biochem Mol Biol. 2007 Sep-Oct;42(5):373-97.

11. Hackstein JH, Thaden J, Koopman WJH, Huynen MA. Hydrogenosomes (and related organelles, either) are not the same. In: Martin WF, Müller M, editors. Origin of mitochondria and hydrogenosomes. Berlin: Springer; 2007, p. 135–59.

12. Hawksworth DL. Lichenization: the origins of a fungal life-style. In: Upreti DK, Divakar PK, Shukla V, Bajpai R, editors. Recent advances in lichenology: modern methods and approaches in lichen systematics and culture techniques. New Delhi: Springer India; 2015, p. 1–10

13. Henze K, Martin W. How do mitochondrial genes get into the nucleus? Trends Genet. 2001; 17:383–7.

14. Hermansson M, Linberg C. Gene transfer in the marine environment. FEMS Microbiol Ecol 1994; 15: 47–54

15. John P, Whatley FR. Paracoccus denitrificans and 573 the evolutionary origin of the mitochondrion. Nature 1975; 254: 495–8.

16. Joy JB. Symbiosis catalyses niche expansion and diversification. Proc Biol Sci 2013; 280: 20122820.

17. Lee SJ, Kong M, Harrison P, Hijri M. Conserved proteins of the RNA interference system in the arbuscular mycorrhizal fungus Rhizoglomus irregulare provide new insight into the evolutionary history of Glomeromycota. Genome Biol Evol 2018; 10: 328–343.

18. Naito M, Morton JB, Pawlowska TE. Minimal genomes of mycoplasma-related endobacteria are plastic and contain host-derived genes for sustained life within Glomeromycota. Proc Natl Acad Sci U S A 2015; 112: 7791–6.

19. Paget MS. Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution. Biomolecules. 2015 Jun 26;5(3):1245-65. doi: 10.3390/biom5031245.

20. Radman M. Enzymes of evolutionary change. Nature. 1999. 401:866–868.

21. Rajendhran J, Gunasekaran P. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res 2011; 166: 99–110.

22. Rosenberg S.M. and Queitsch C.. Combating Evolution to Fight Disease. Science. 2014 Mar 7; 343(6175): 1088–1089.

23. Saecker RM, Record MT Jr, Dehaseth PL. Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. J Mol Biol. 2011 Oct 7;412(5):754-71.

24. Santoro AL, Bastviken D, Gudasz C, Tranvik L, Enrich-Prast A. Dark carbon fixation: an important process in lake sediments. PLoS One. 2013; 8: e65813.

25. Sauert M, Temmel H, Moll I. Heterogeneity of the translational machinery: Variations on a common theme. Biochimie. 2015 Jul;114:39-47.

26. Stoddard SF, Smith BJ, Hein R, Roller BR, Schmidt TM. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res. 2015 Jan; 43(Database issue):D593-8.

27. Tang D, Kang R, Livesey KM, Kroemer G, Billiar TR, Houten BV, et al. HMGB1 [High Mobility Group Box 1] is essential for mitochondrial quality control. Cell Metab 2011; 13: 701–11.

28. Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, et al. Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. Proc Natl Acad Sci U S A 2013; 110: 20117–22.

29. Torres-Cortes G, Ghignone S, Bonfante P, SchuSsler A. Mosaic genome of endobacteria in arbuscular mycorrhizal fungi: transkingdom gene transfer in an ancient mycoplasma-fungus association. Proc Natl Acad Sci U S A 2015; 112: 7785–90.

30. Trachtenberg S. Mollicutes-wall-less bacteria with internal cytoskeletons. J Struct Biol 1998; 124: 244–56.

31. Whatley JM, John P, Whatley FR. From extracellular to intracellular: the establishment of mitochondria and chloroplasts. Proc R Soc Lond B Biol Sci 1979; 204: 165–87.

32. Xia X. Maximizing transcription efficiency causes codon usage bias. Genetics 1996; 144: 1309–20.

33. Yeiser B, Pepper ED, Goodman MF, Finkel SE. SOS-induced DNA polymerases enhance long-term survival and evolutionary fitness. Proc Natl Acad Sci U S A. 2002 Jun 25;99(13):8737-41.

34. Yuan X, Xiao S, Taylor TN. Lichen-like symbiosis 600 million years ago. Science 2005; 308: 1017–20.

35. Zhang N, Buck M. (2015) A perspective on the enhancer dependent bacterial RNA polymerase. Biomolecules. 21;5(2):1012-9.

# Appendix I Methods for finding an optimal k- the number of clusters

For SeSaME PS Function, it may be helpful to provide users with an additional method for finding an optimal k- the number of clusters. Here, I outline the methods for finding an optimal k value for k parameter of k-means clustering method. For a range of k values, the following methods will produce scores for evaluating the k values. The first method is elbow method using the sum of squared errors that is also called within cluster sum of squares (SSE) (Eqn. 1). SSE belongs to internal index and measures a degree of cluster cohesion without considering external information. For this method, an optimal k value is found at the point looking like an elbow where the line chart looks like an arm, when we make XY axis scatter chart where SSE is in Y axis and k value is in X axis. The second method is SSB_SSE, a sum of SSB/SSE where SSB is between cluster sum of squares that measures cluster separation (Eqn. 2). An optimal k has the maximum of SSB_SSE. The third method is from a JAVA class called SumOfClusterVariances under the package ml.evaluation in apache math 3.6. It is similar to the first method. The fourth and the fifth methods are silhouette coefficient that incorporates both cluster cohesion and cluster separation for individual points as well as for clusters. An optimal k has the highest silhouette coefficient. Silhouette coefficient 1 results from averaging all silhouette coefficients while the method called silhouette coefficient 2 calculates an average of silhouette coefficients belonging to each cluster and then calculates an average of the averages of all clusters.

Eqn 1. SSE = $\sum_{i}^{k} \sum_{x}^{y} dist^2(x, m_i)$

where there are k clusters. Point x belongs to cluster i. There are y members in cluster i. mi is the centroid of cluster i.

Eqn 2. SSB_SSE = $\sum_{i}^{k} SSB/SSE$

Where SSB = $\sum_{i}^{k} |Ci| \, dist^2(m, m_i)$ |Ci| is the size of cluster i. m is the centroid of the overall data, mi is the centroid of cluster i.

If data include a cluster with a single member (version 1), for the cluster SSB/SSE = 0

Silhouette coefficients s = 1 – a/b if a < b, or s = b/a -1 if a >= b

Where a = average distance of i to the points in its cluster, b = min (average distance of i to points in another cluster). If data include a cluster with a single member (version 1), a = 0 for the cluster.

Each method has two versions depending on whether it includes a cluster with a single member in calculation; version 1 includes it while version 2 excludes it.

I implemented the five methods with Java programming language. I ran the program, SeSaMe PS Function, with 5 newly selected sequences and the example sequence (chapter 3). I provided the option to specify k values (5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29). Then, I applied the implemented methods to the coordinates of component spaces, the result data from SeSaMe PS Function. Because the result of 80% components contained the coordinates in the components spaces that accounted for 80% inertia, the number of dimensions was larger than that of the first/second components, representing the data with higher resolution. Consequently, the data were more dispersed compared to the first/second components. Therefore, XY axis scatter chart, created based on 80% components, where SSE is in Y axis and k value is in X axis, had a shape of linear line with a negative slope, while the first/ second components had a usual shape of converged line that has an elbow.

Determining a single optimal K value was risky because both SSB_SSE and silhouette coefficients had large fluctuations, probably due to the complex properties of the main variable- trimer usages (Table 1). With real world data, it may be more realistic to find an optimal range of k values, rather than a single optimal k. For example, we may find an intersection of optimal ranges of k values from multiple methods. I will provide users with an access to the methods for evaluating k values to find an optimal k value or an optimal range of k values for his/ her query sequence. In near future, they will be available at the website, www.fungalsesame.org

To download the Table 1, please click here (at the same time, press Ctrl key).
If the link does not work, please copy and paste the following website address in your browser.

www.codon.kr/thesis/appendixI/appendixI_table1.xls

# Appendix II Publications and Conference Presentations

## Publications

Two articles (the second and the third chapters) have been accepted to the journal, Genomics Proteomics, and Bioinformatics (GPB) in August 2018.

1) Jee Eun Kang1*, Antonio Ciampi2, Mohamed Hijri1

**SeSaMe PS Function: Functional Analysis of the Whole Metagenome Sequencing Data of the Arbuscular Mycorrhizal Fungi**

2) Jee Eun Kang1*, Antonio Ciampi2, Mohamed Hijri1

**SeSaMe: Metagenome Sequence Classification of Arbuscular Mycorrhizal Fungi Associated Microorganisms**

3rd article (chapter 4) has been in preparation for submission.

## Conference presentations

1) 2017 ISERD – 125th International Conference on Environment and Natural Science (ICENS)
Sponsored by the IIER (International Institute of Engineers and Researchers).
Jan. 2017 Seoul, Republic of Korea
Abstract submission/ Oral presentation - Position Specific Genetic Code Approach for Comparative Study of Mitochondrial Genomes in Plant, Lichen associated Fungi, and Arbuscular Mycorrhizal Fungi

2) 2016 5th International Conference on Environment, Energy, and Biotechnology
May 23-25, 2016 Jeju Island, Republic of Korea
ICEEB 2016
Abstract submission/ Oral presentation - Taxonomical Classification of the Genome Sequencing Data from Arbuscular Mycorrhizal Fungi and their Associated Bacteria