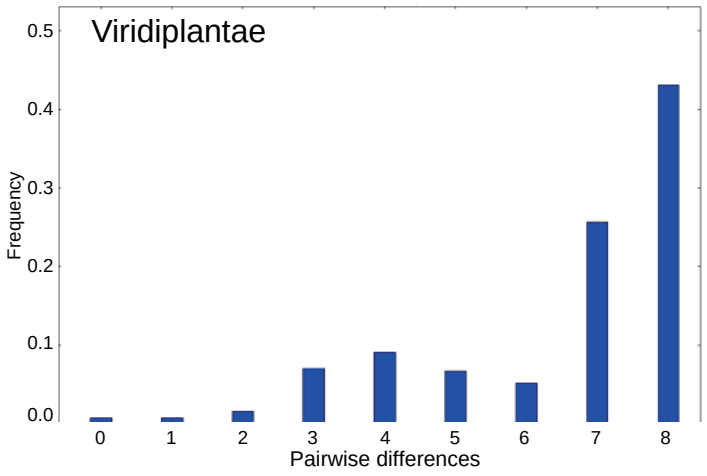
Supplementary Material
Part 2

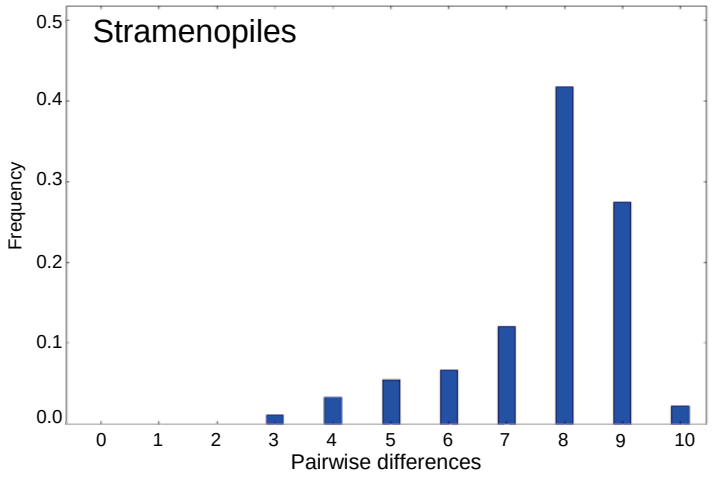
Identification of species by multiplex analysis of variable-length sequences

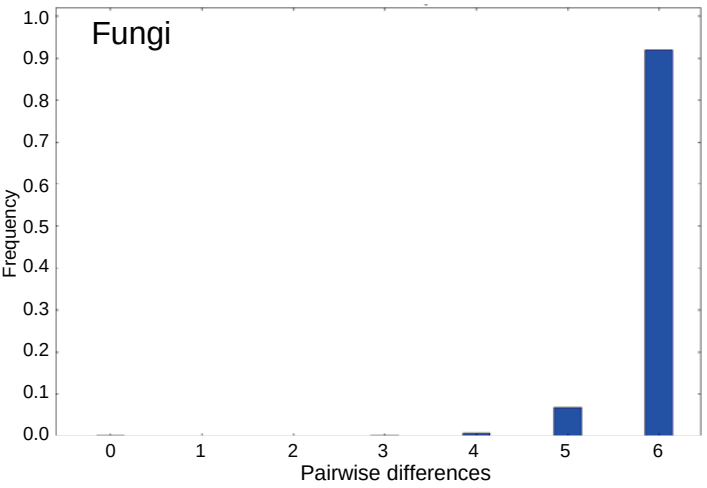
Filipe Pereira[1], João Carneiro[1], Rune Matthiesen[1], Barbara van Asch[1,2], Nádia Pinto[1,2,3], Leonor Gusmão[1] and António Amorim[1,2]

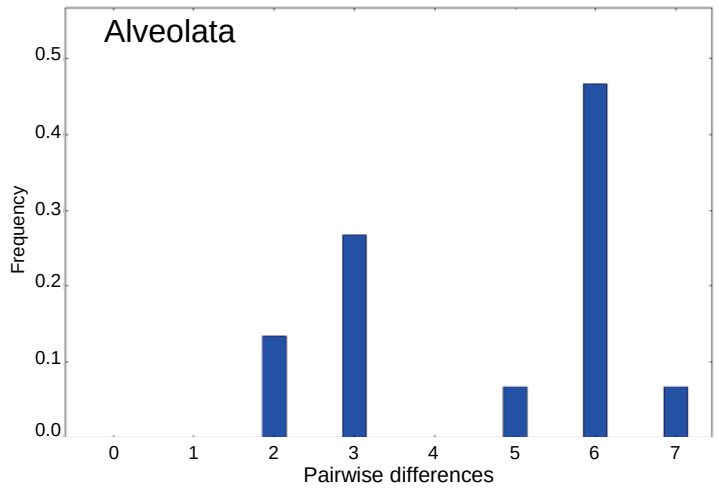
[1] Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
[2] Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal
[3] Centro de Matemática da Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

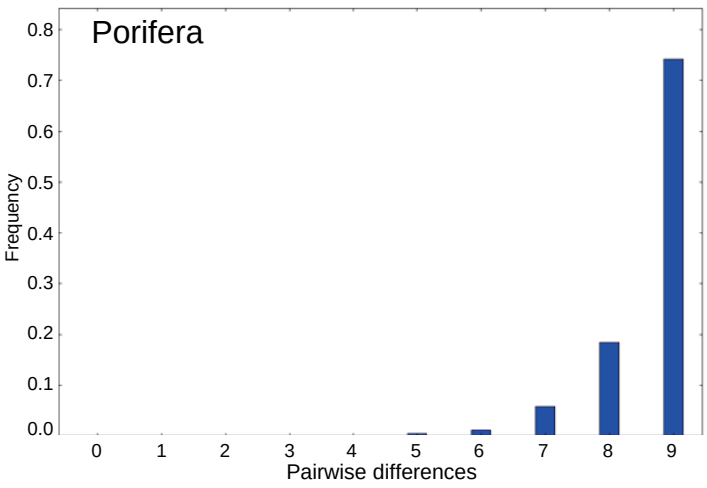
Supplementary Figure S8. Mismatch distribution of SPInDel profiles. The frequency distribution of the number of SPInDel hypervariable regions that differ between all pairs of SPInDel profiles is represented for 18 eukaryotic groups.

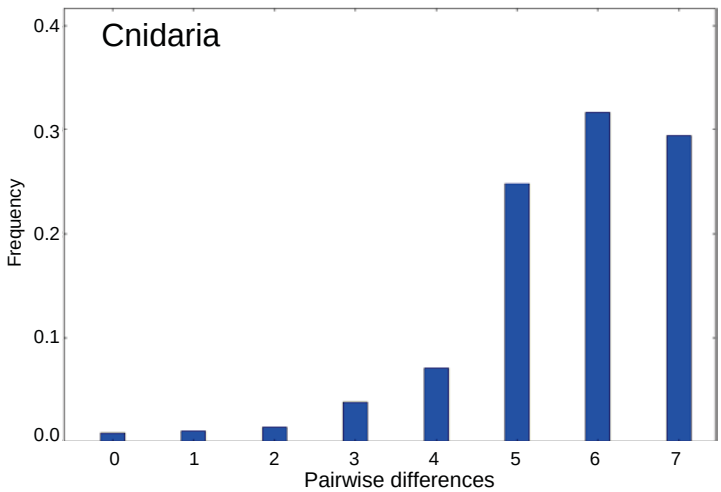




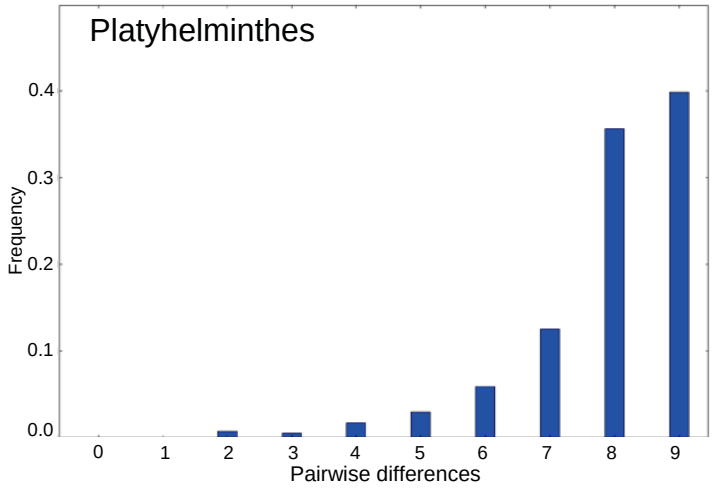


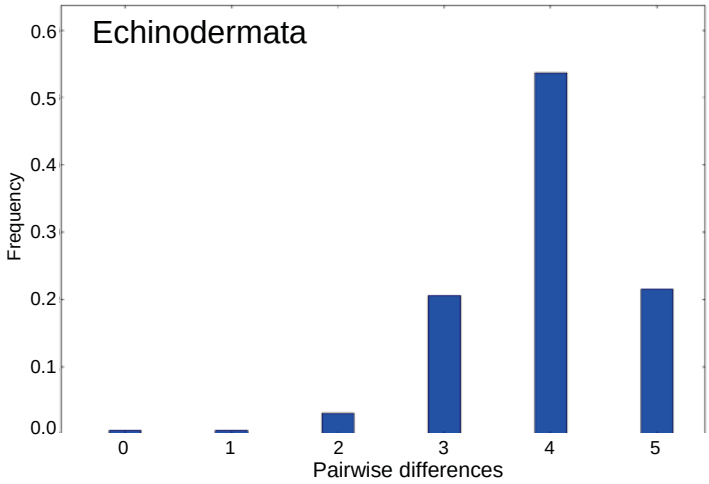


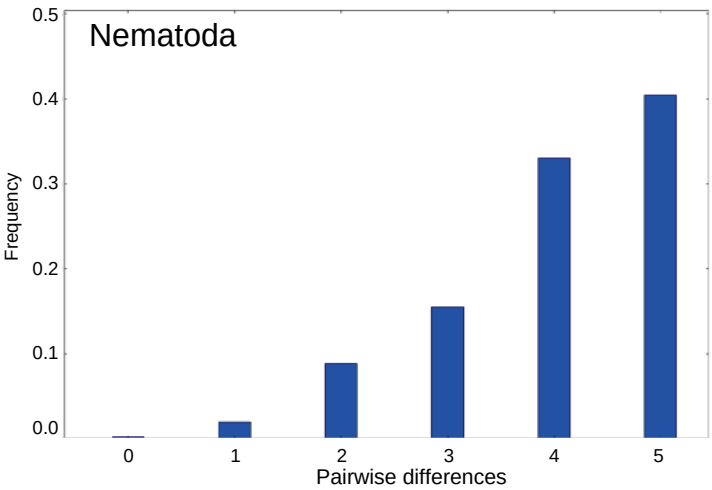


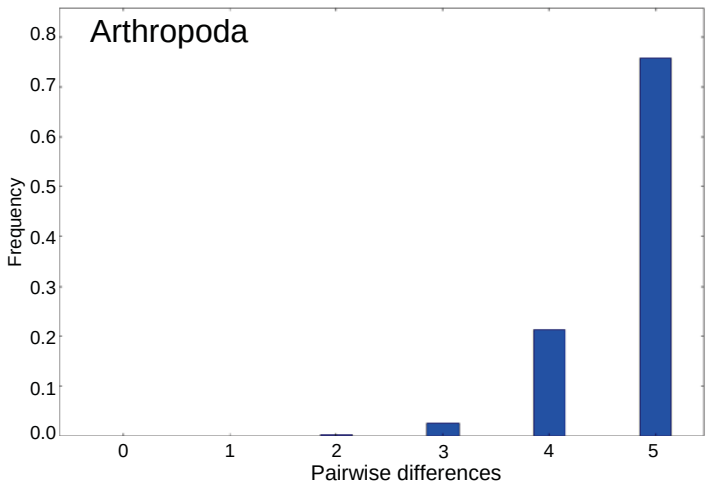


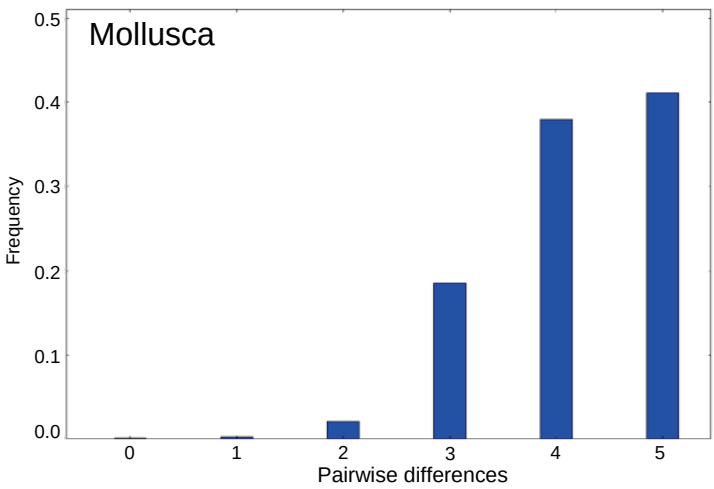
Supplementary Figure S8 (cont.)

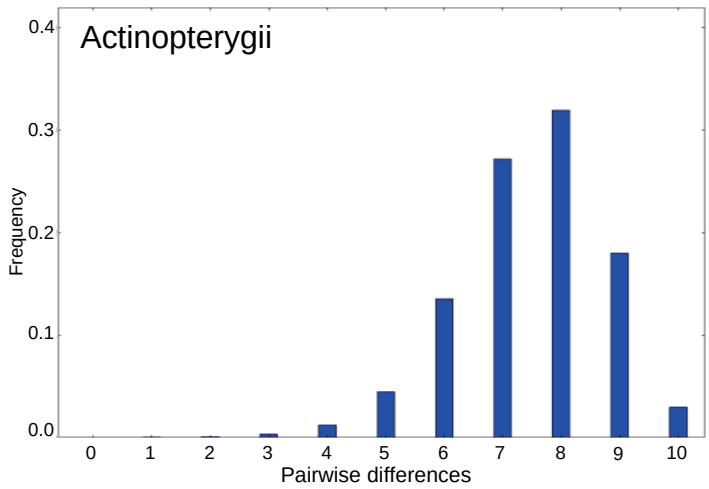




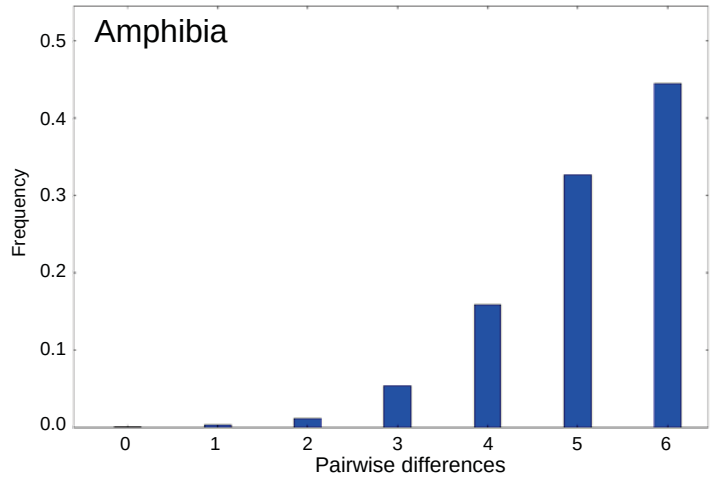


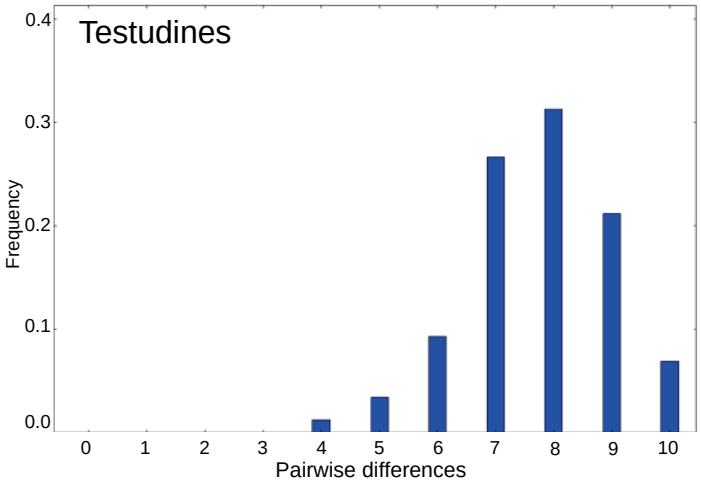


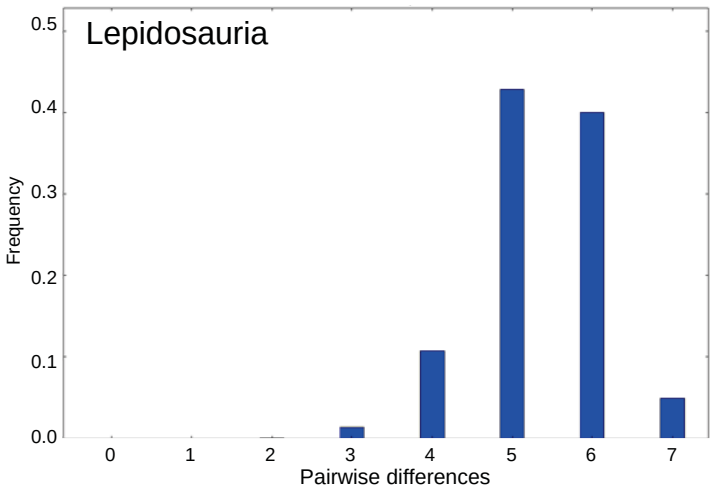


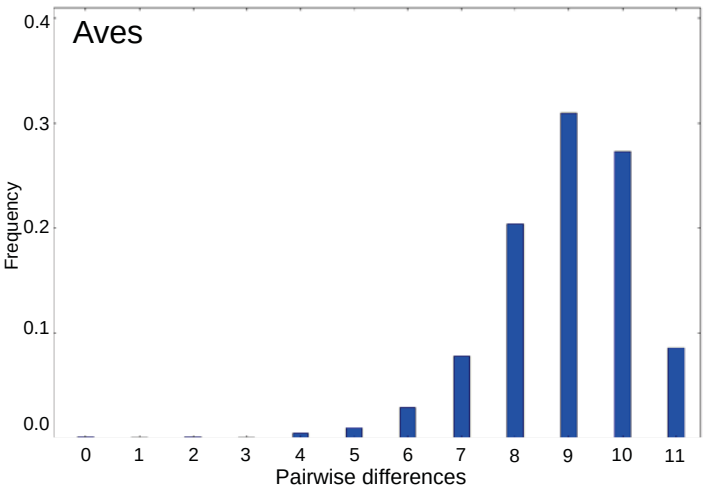


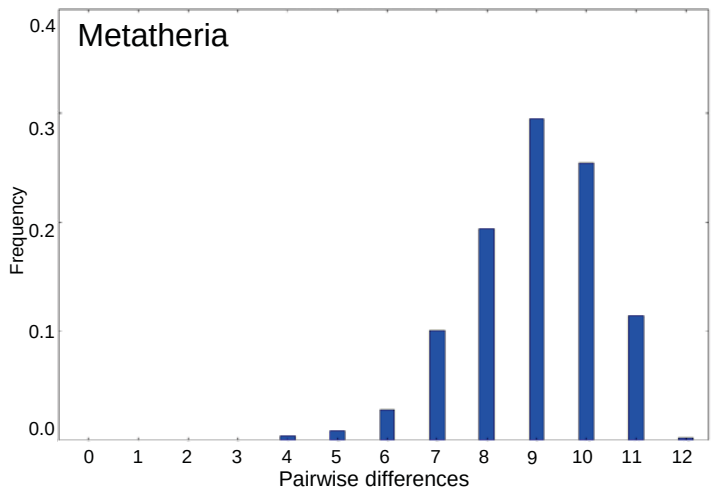
Supplementary Figure S8 (cont.)

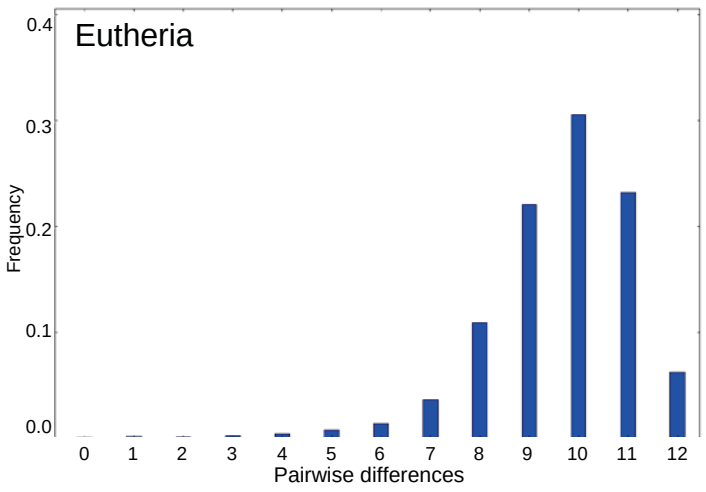












Supplementary Figure S9. SPInDel workbench. (**a**) Brief description of the SPInDel workbench. Further information can be found at http://www.portugene.com/SPInDel/SPInDel_webworkbench.html. (**b**) Flowchart depicting the different steps required for species identification with the SPInDel procedure. (**c**) Screenshots of the different workbench sections: project database, alignment and profile editors and global calculations.

Supplementary Figure S9

a

SPInDel workbench

The SPInDel software was written in PYTHON 2.6 using Biopython (http://biopython.org), SciPy (http://www.scipy.org), GenomeDiagram (http://bioinf.scri.ac.uk/lp/programs.php), matplotlib (http://matplotlib.sourceforge.net/), NumPy (http://numpy.scipy.org/), Pycogent (http://pycogent.sourceforge.net/) and Pythia (http://pythia.sourceforge.net). The graphical interface was created using the VisualWX Rapid Application Development (RAD) environment (http://visualwx.altervista.org) and the Eclipse platform to debug and test the software. A single EXE file was created using the Inno Setup software (http://www.jrsoftware.org/isinfo.php) for installation purposes.

As input, the software requires a FASTA formatted DNA sequence file. Projects with aligned sequences can be uploaded, although alignments can also be done with the Pycogent progressive alignment implemented on the workbench. The user can select among different nucleotide substitution models to perform the alignment. Sequences can be added or removed at any point. An identity value is plotted for each nucleotide position by estimating the frequency of the most common nucleotide in that position (indels are ignored). Conserved regions can be easily identified by observing the graphic output for identity values (highest conservation represented in green and lowest represented in red) and can be defined directly in the alignment window using column selection.

Supplementary Figure S9 (cont.)

The discriminatory power of selected hypervariable regions can be evaluated by several options. General statistics are calculated for each region and for complete profiles. The UPGMA method is used to build a guide tree to discriminate species in each project. This algorithm allows the clustering of profiles based on a dissimilarity matrix obtained from the number of differences between the profiles from different species. Properties of PCR primers (length, T_m, GC content and folding energies) are estimated as described in the Oligo Calc webserver (Kibbe 2007) and Pythia (Mann et al. 2009).

A function that calculates the discriminatory power of all combinations of hypervariable regions can be used to identify a minimum number of hypervariable regions for accurate discrimination. The algorithm generates *n*-combinations without repetition, which are subsets of *n* distinct elements of the set of all possible regions. For each *n*-combination, all N_{sp} and N_{dp} values are displayed on tables and graphs. The algorithm also included a 'multiplex PCR option' to retrieve only *n*-combinations not sharing conserved regions.

SPInDel profiles of unknown origin can be predicted by a *k*-nearest neighbor method using a database of known profiles. We implemented the algorithm using Biopython and added the discrete distance metric. Classification accuracy can be estimated within the SPInDel workbench by testing the performance of the *k*-nearest neighbors by cross validation using profiles from known species profiles.

All data available in the SPInDel workbench can be readily exported in common file formats for use in other data analysis programs: projects (.sql), sequences (.fasta), PCR primers (.csv), pairwise matrixes (.csv), UPGMA trees (.newick) and graphs (.pdf, .png), among others.

References

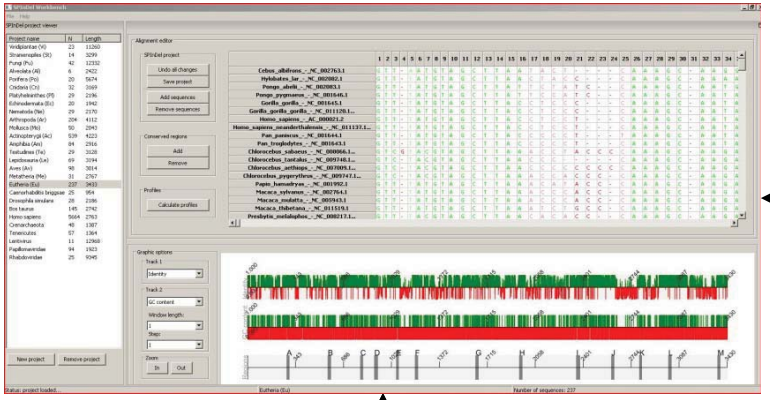
Kibbe,W.A. (2007) OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Res*, **35**, W43-W46.

Mann,T., Humbert,R., Dorschner,M., Stamatoyannopoulos,J. and Noble,W.S. (2009) A thermodynamic approach to PCR primer design. *Nucleic Acids Res*, **37**

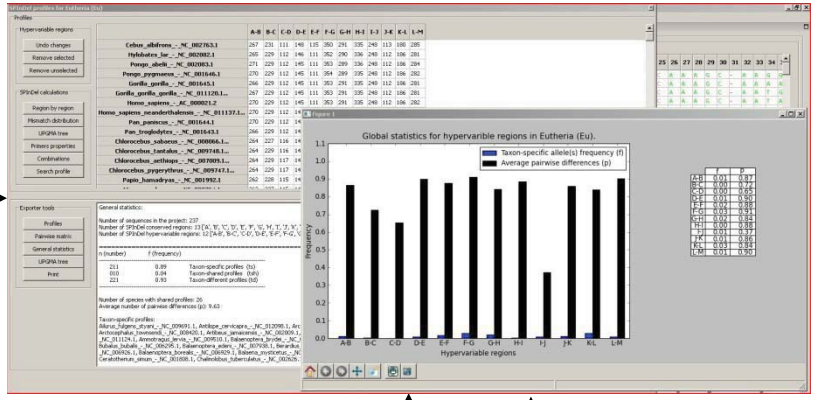
Supplementary Figure S9 (cont.)

b

Project Database and Alignment Editor



Profile Editor and global calculations



Applications

Data Management

Analytical Pipeline

New sequences alignment in FASTA format as input

Option to re-align sequences or add/remove sequences

Definition of conserved and hypervariable regions

Calculation of profiles → SPInDel profiles

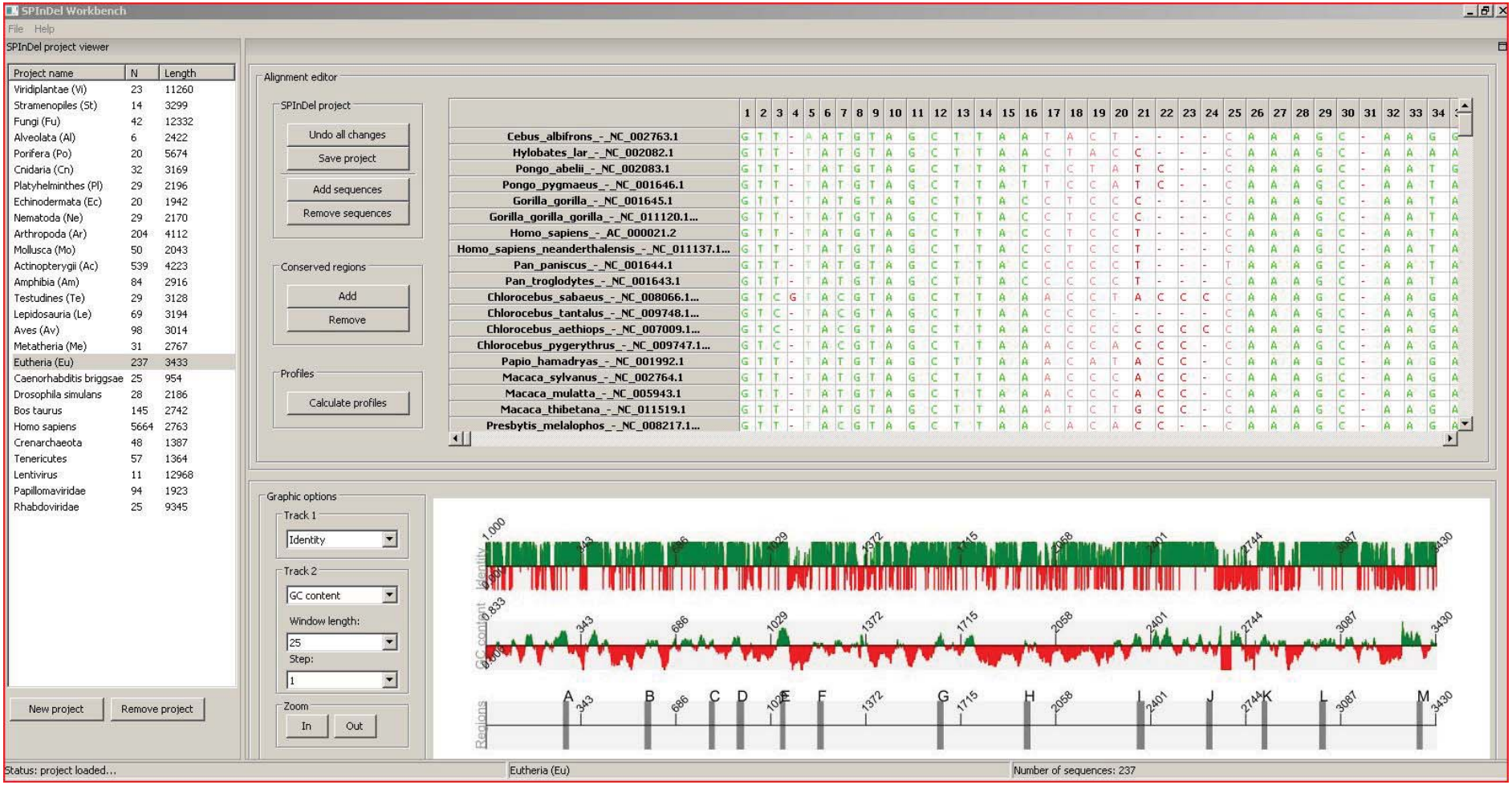
Region by region
Mismatch distribution
Primers properties
Combinations
UPGMA tree
Search profile

Exporter tools → Identification engine

Exporter tools → Typing system

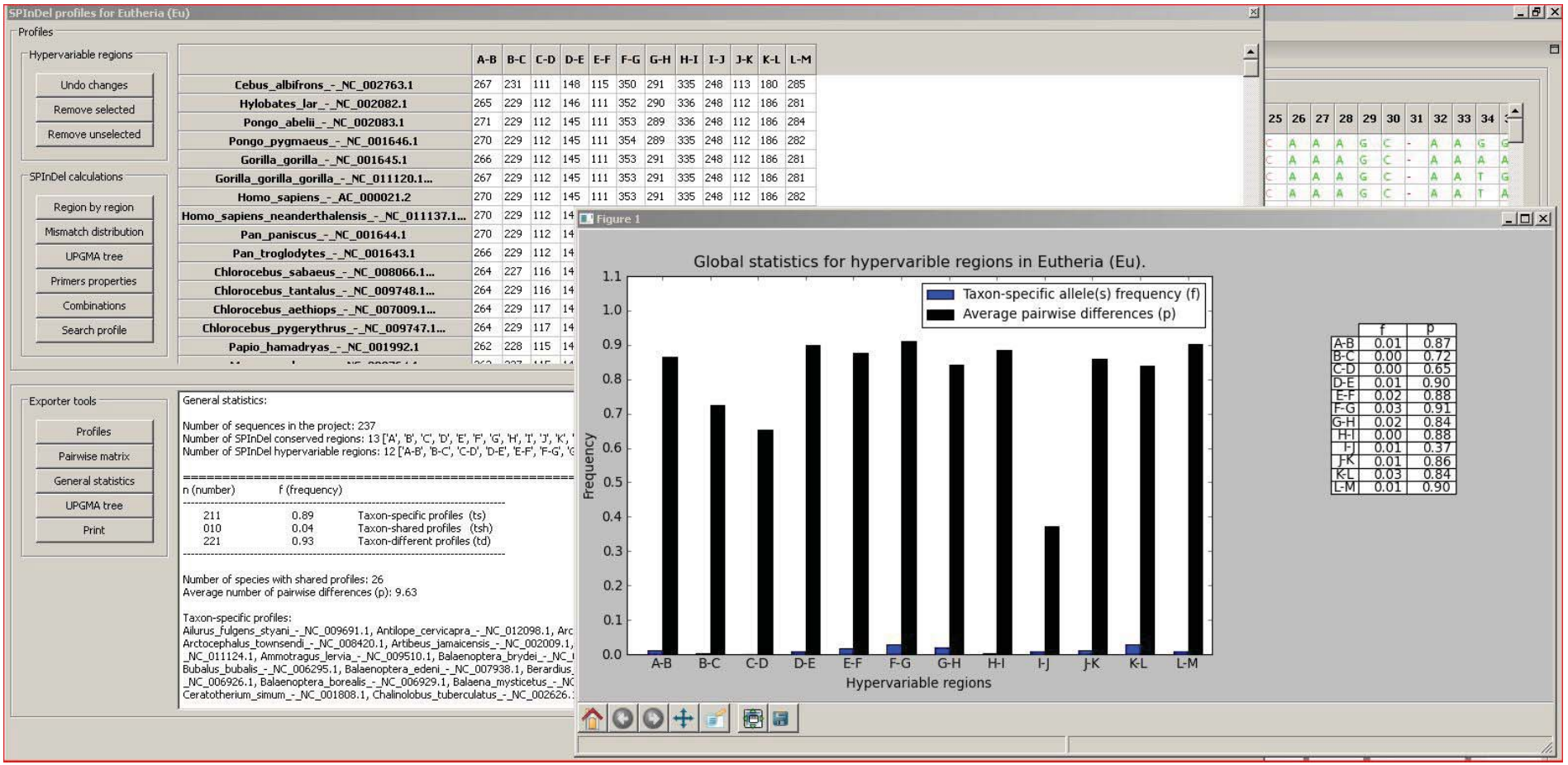
Supplementary Figure S9 (cont.)

c PROJECTS VIEWER AND ALIGNMENT EDITOR



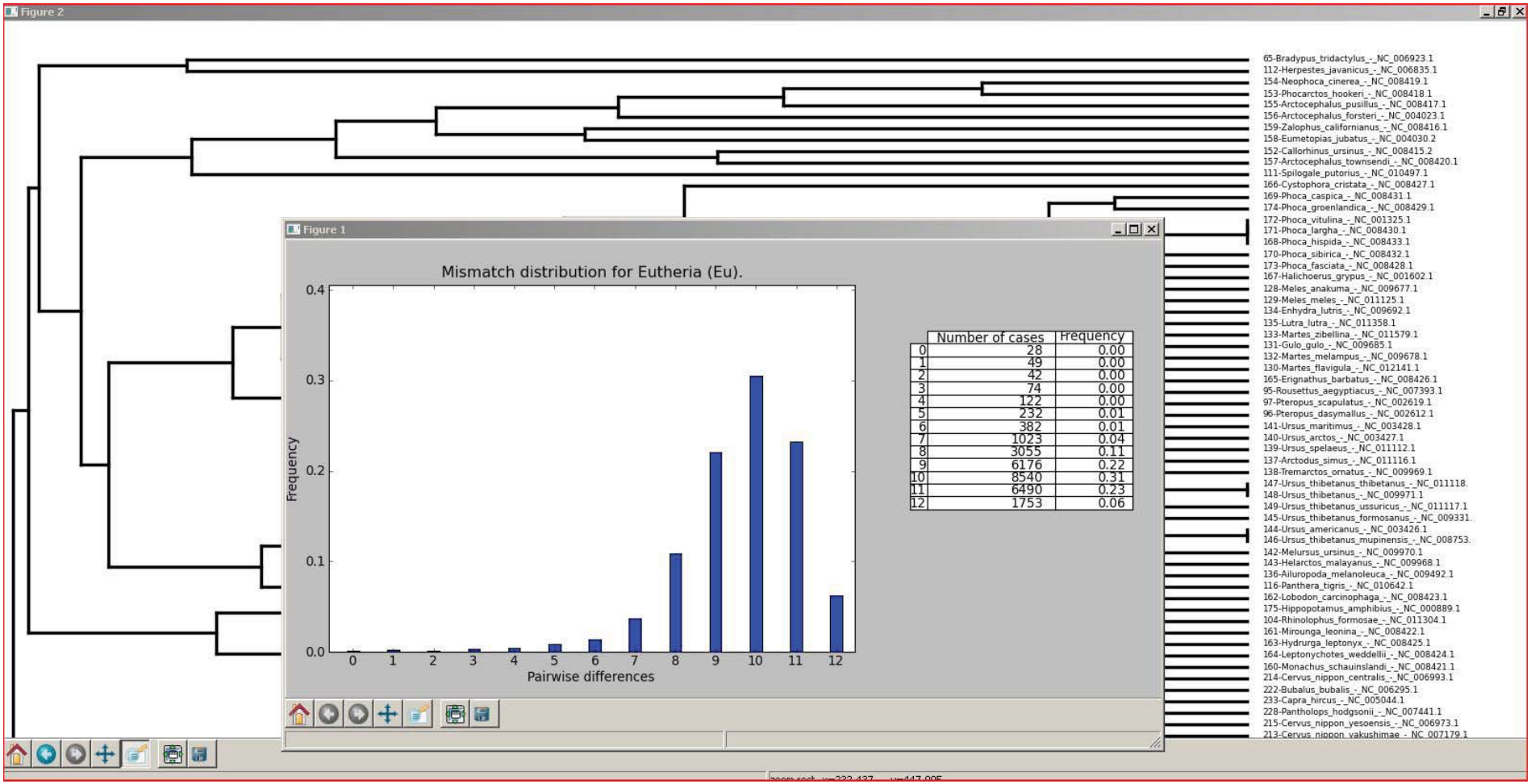
Supplementary Figure S9 (cont.)

PROFILES FRAME AND GLOBAL STATISTICS GRAPH



Supplementary Figure S9 (cont.)

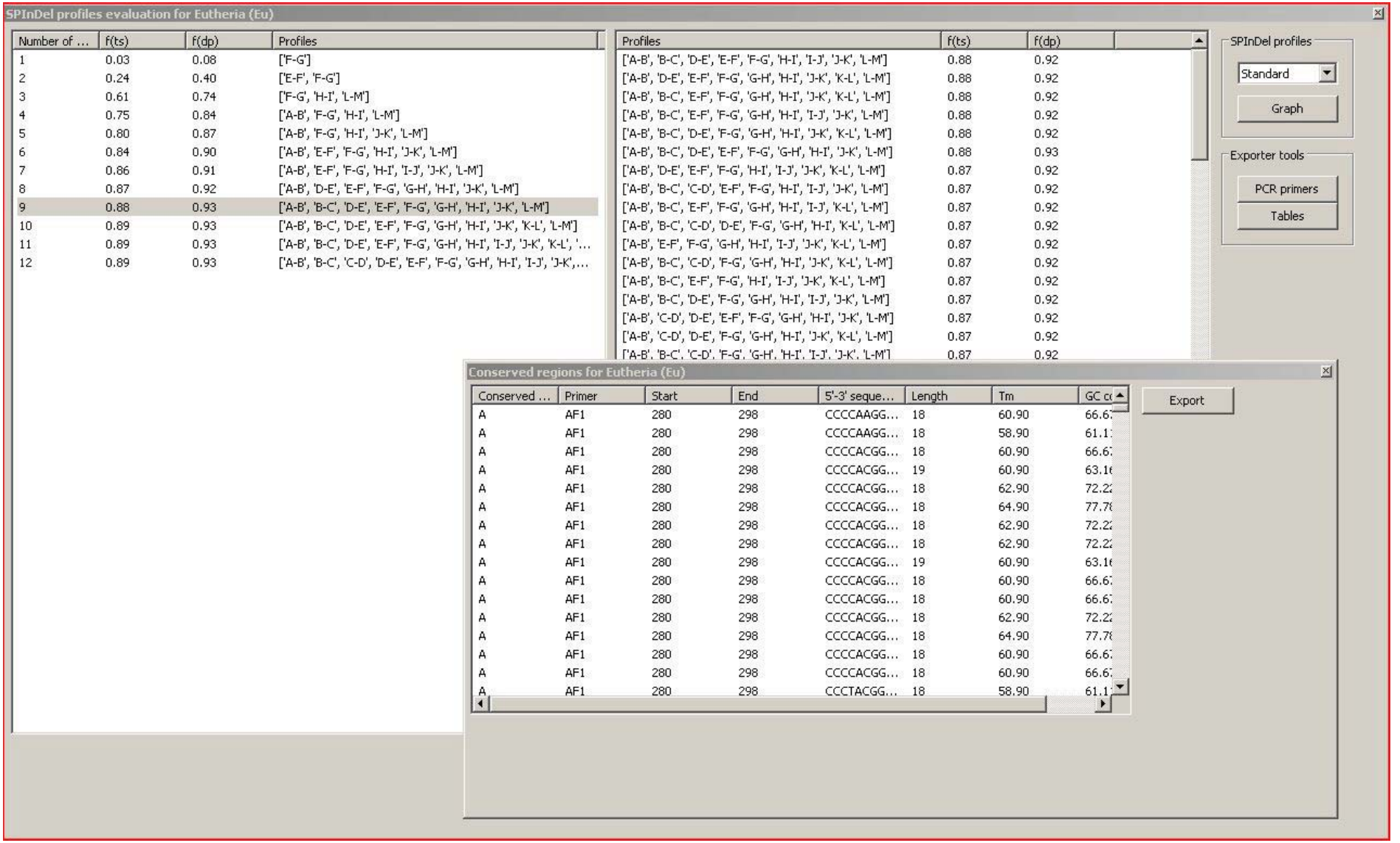
MISMATCH DISTRIBUTION GRAPH AND UPGMA TREE



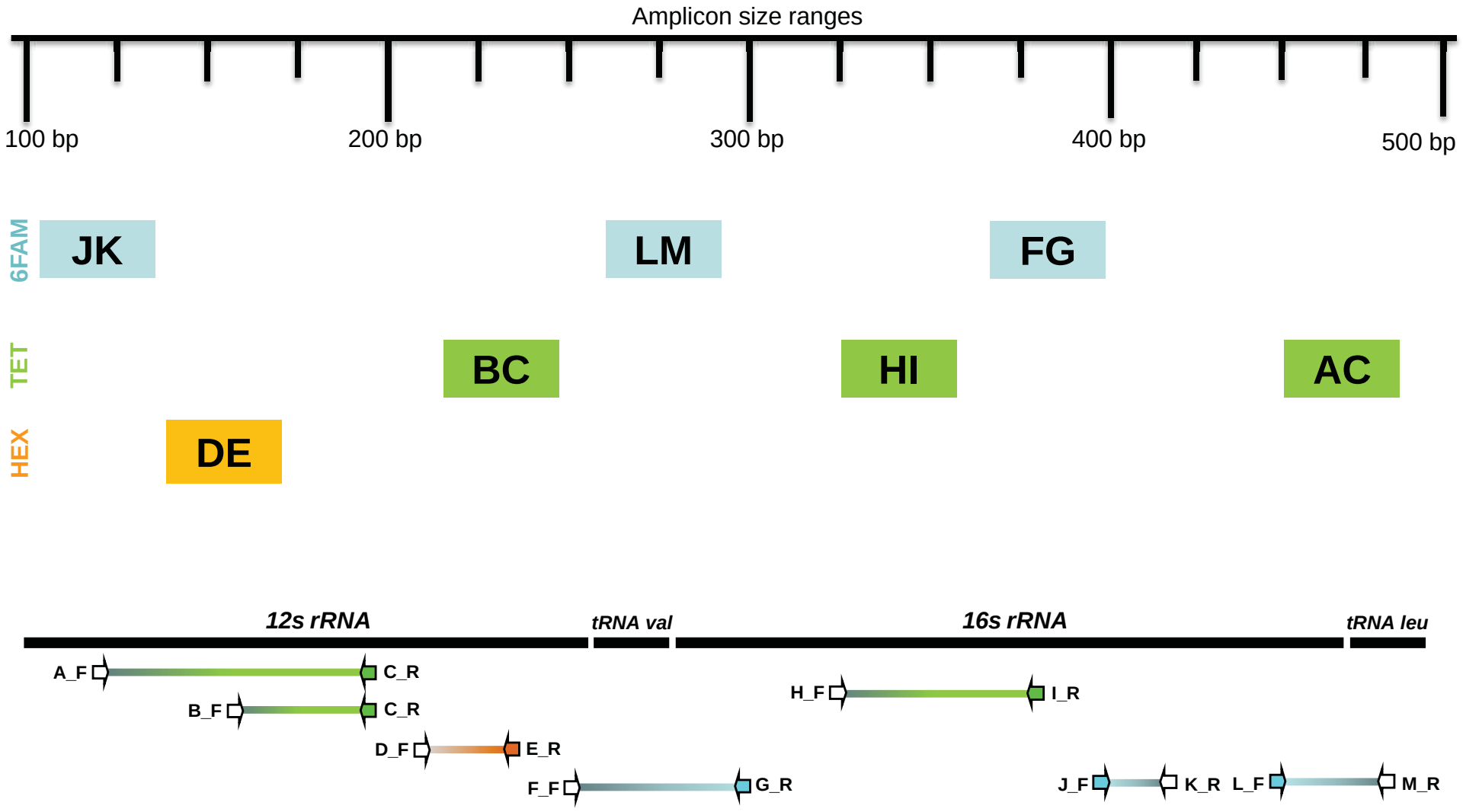
	Number of cases	Frequency
0	28	0.00
1	49	0.00
2	42	0.00
3	74	0.00
4	122	0.00
5	232	0.01
6	382	0.01
7	1023	0.04
8	3055	0.11
9	6176	0.22
10	8540	0.31
11	6490	0.23
12	1753	0.06

Supplementary Figure S9 (cont.)

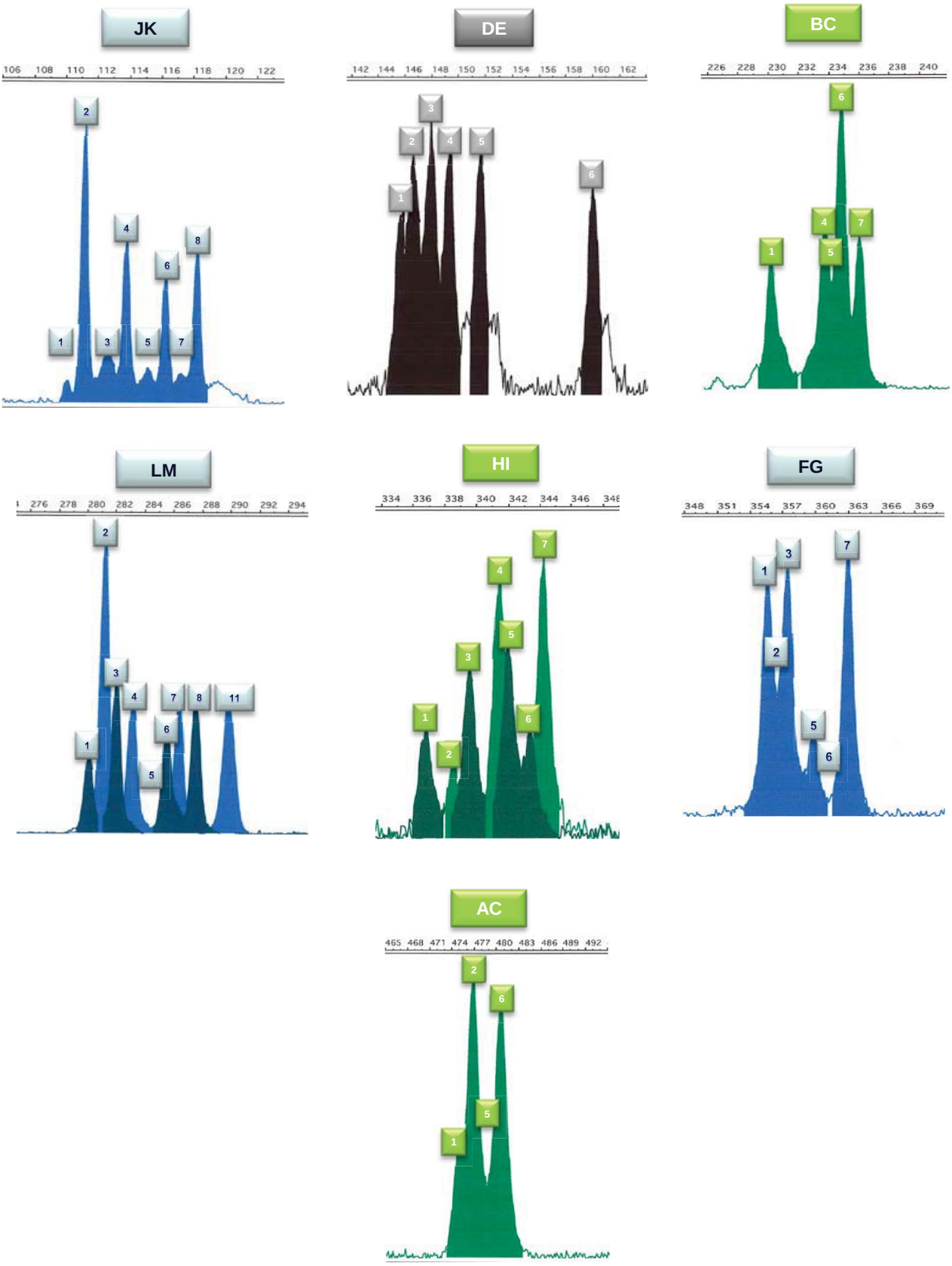
PROFILES EVALUATION FRAME AND PRIMERS PROPERTIES



Supplementary Figure S10. Schematic representation of size ranges and dye labels used in designing the SPInDel assay. The profiling kit uses three spectrally distinguishable fluorescent dyes in the filter set C: 6–FAM (blue), TET (green), and HEX (yellow). We devised a simple way of retrieving all information enclosed in two contiguous hypervariable regions by multiplex PCR, as exemplified for hypervariable regions AB and BC (bottom image). Instead of using a reverse primer for conserved region B to amplify amplicon AB (B_R would be complementary to B_F used to amplify BC), we used the same reverse primer (C_R) for both regions, which meant that region AC was used instead of AB. In this way, we avoided the problem of indels that eliminate size differences in large regions.

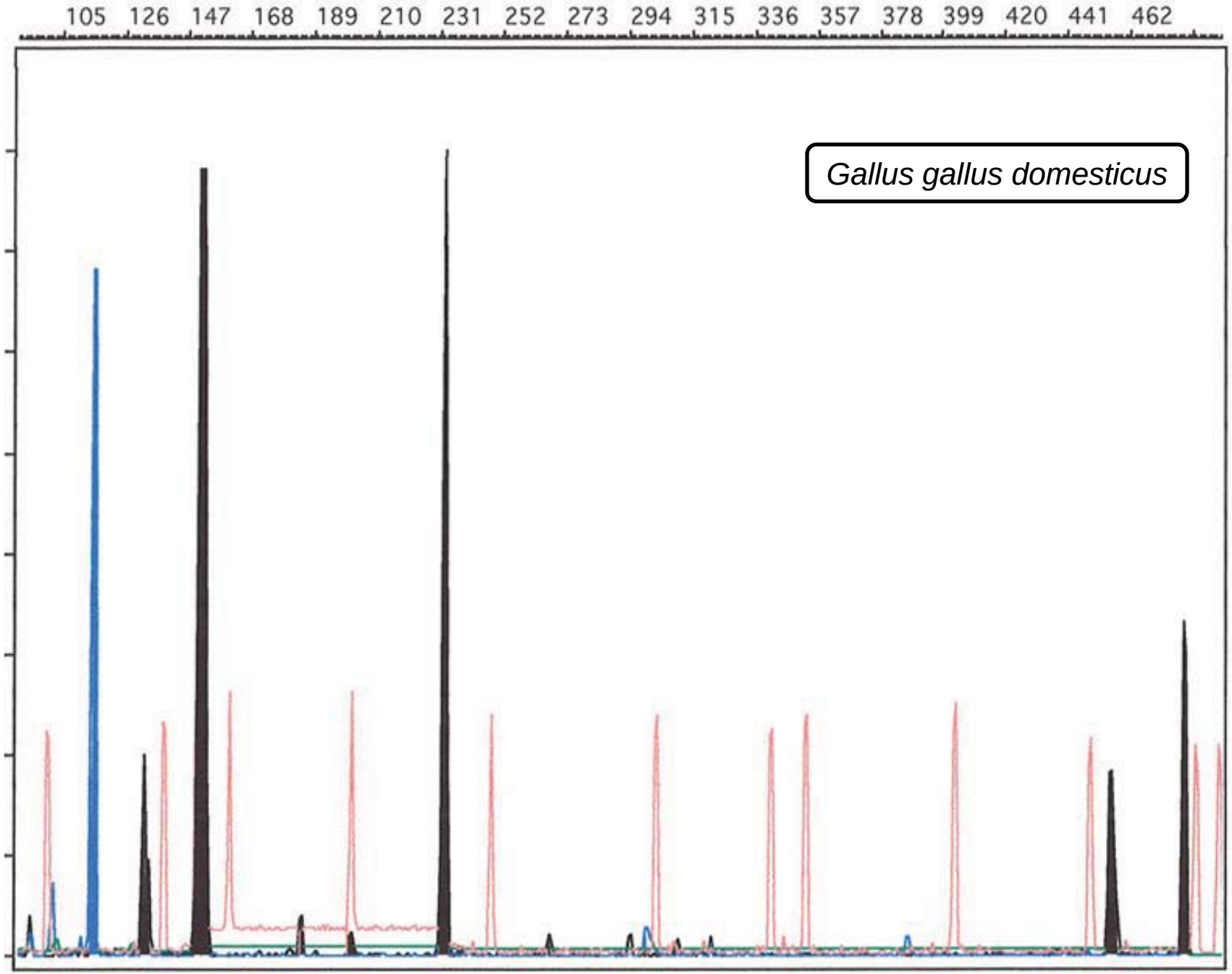


Supplementary Figure S11. Allelic ladders used in the SPInDel profiling kit. The number of each allele is indicated above peaks.

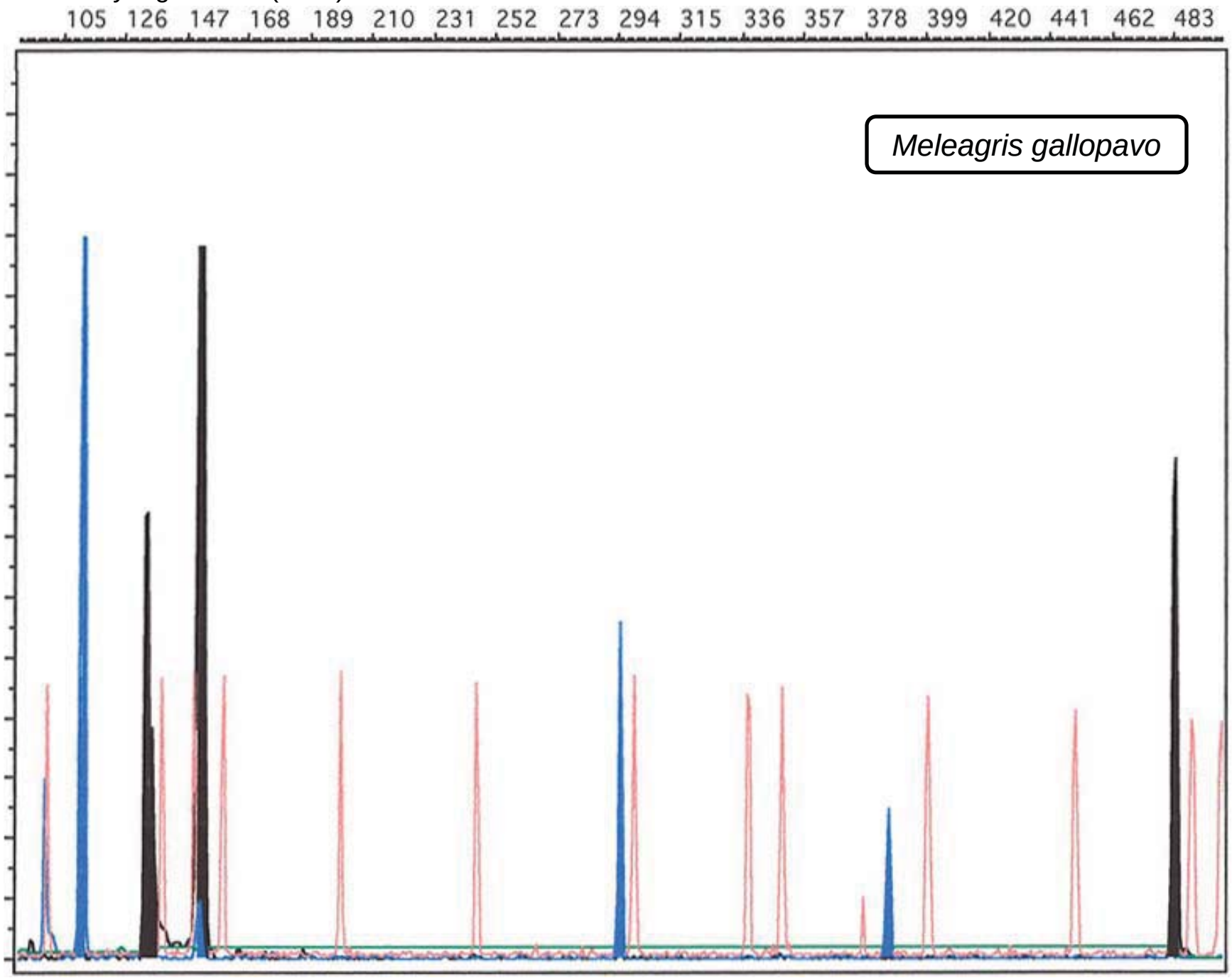


Supplementary Figure S12. Non-eutherian profiles obtained with the SPInDel profiling kit designed for identification of eutherian species. The images represent examples of electropherograms obtained by capillary electrophoresis with multicolor fluorescence detection in representatives of eight species from Arthropoda, Mollusca, Actinopterygii and Aves. The profiles are displayed in a four-dye fluorescent system, with the green, blue and yellow channels used for detection of amplified products and the red channel used as a size marker.

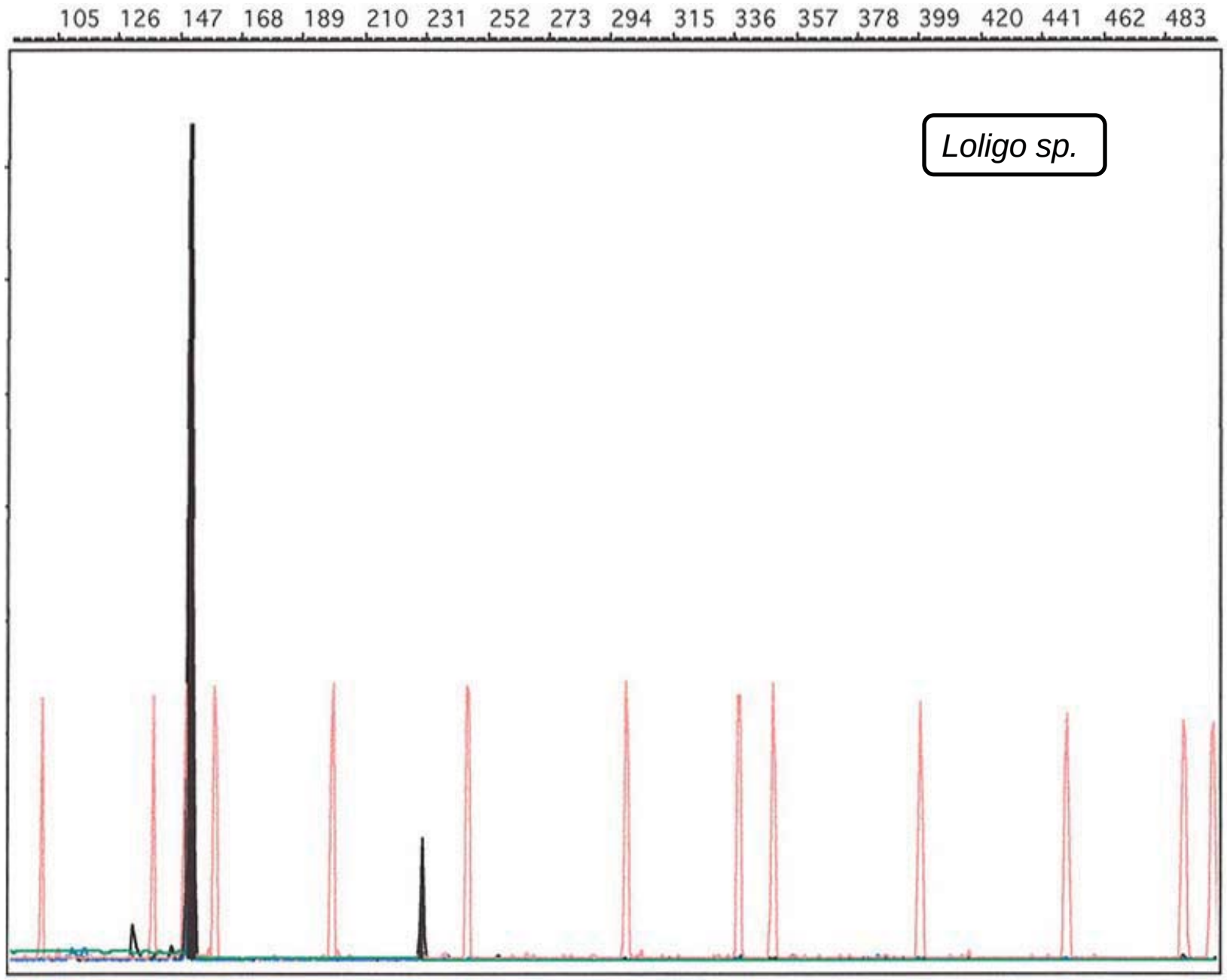
Supplementary Figure S12



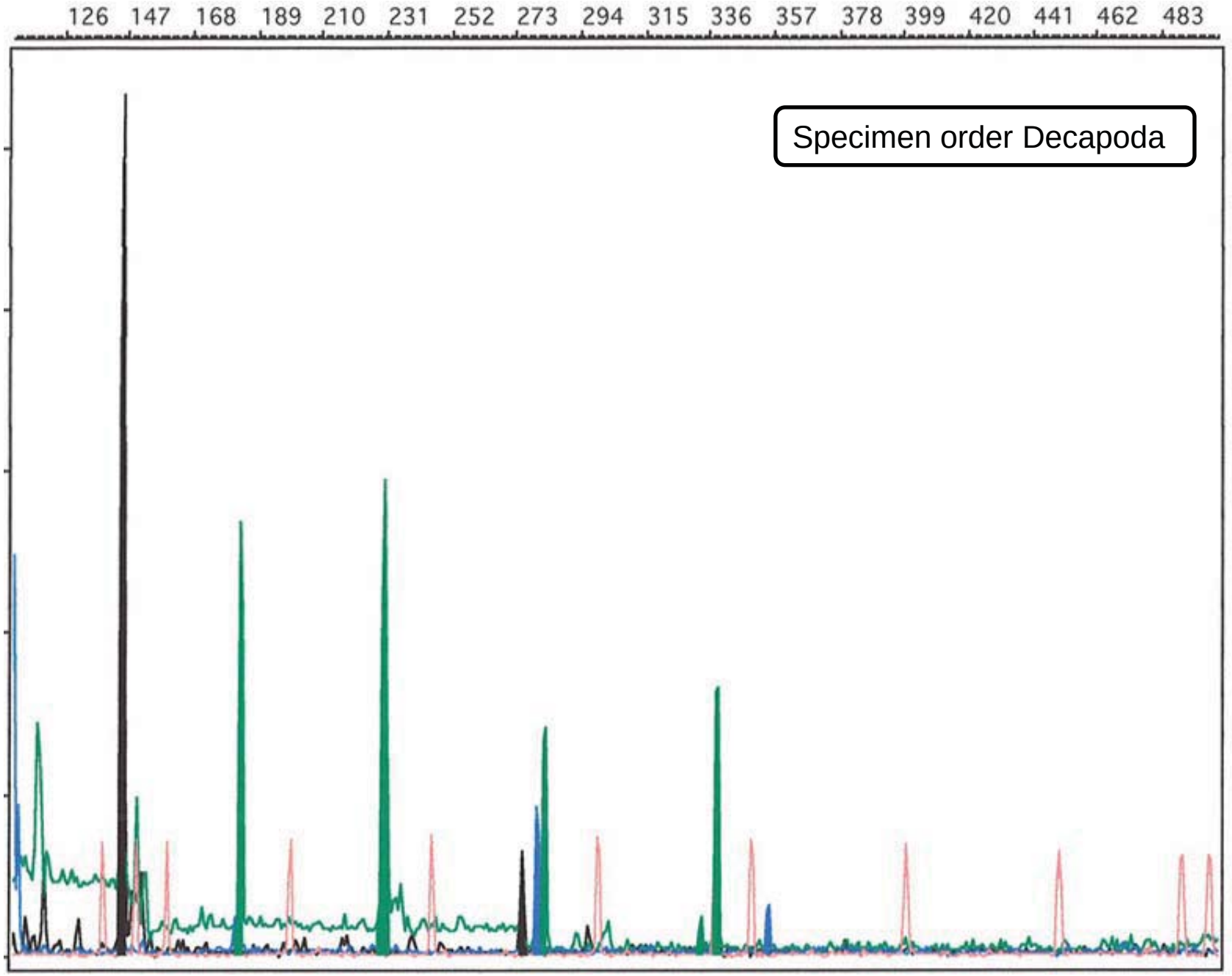
Supplementary Figure S12 (cont.)



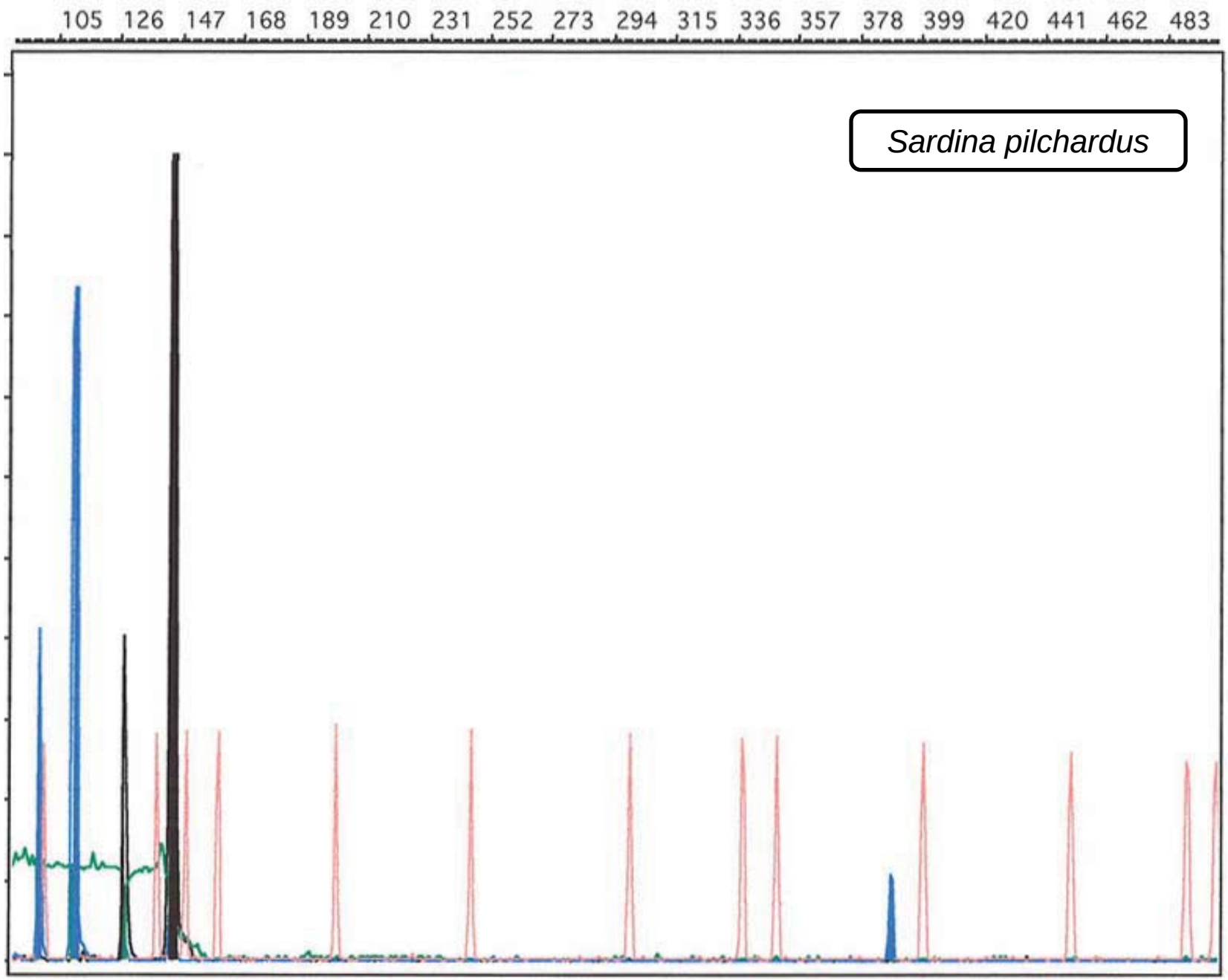
Supplementary Figure S12 (cont.)



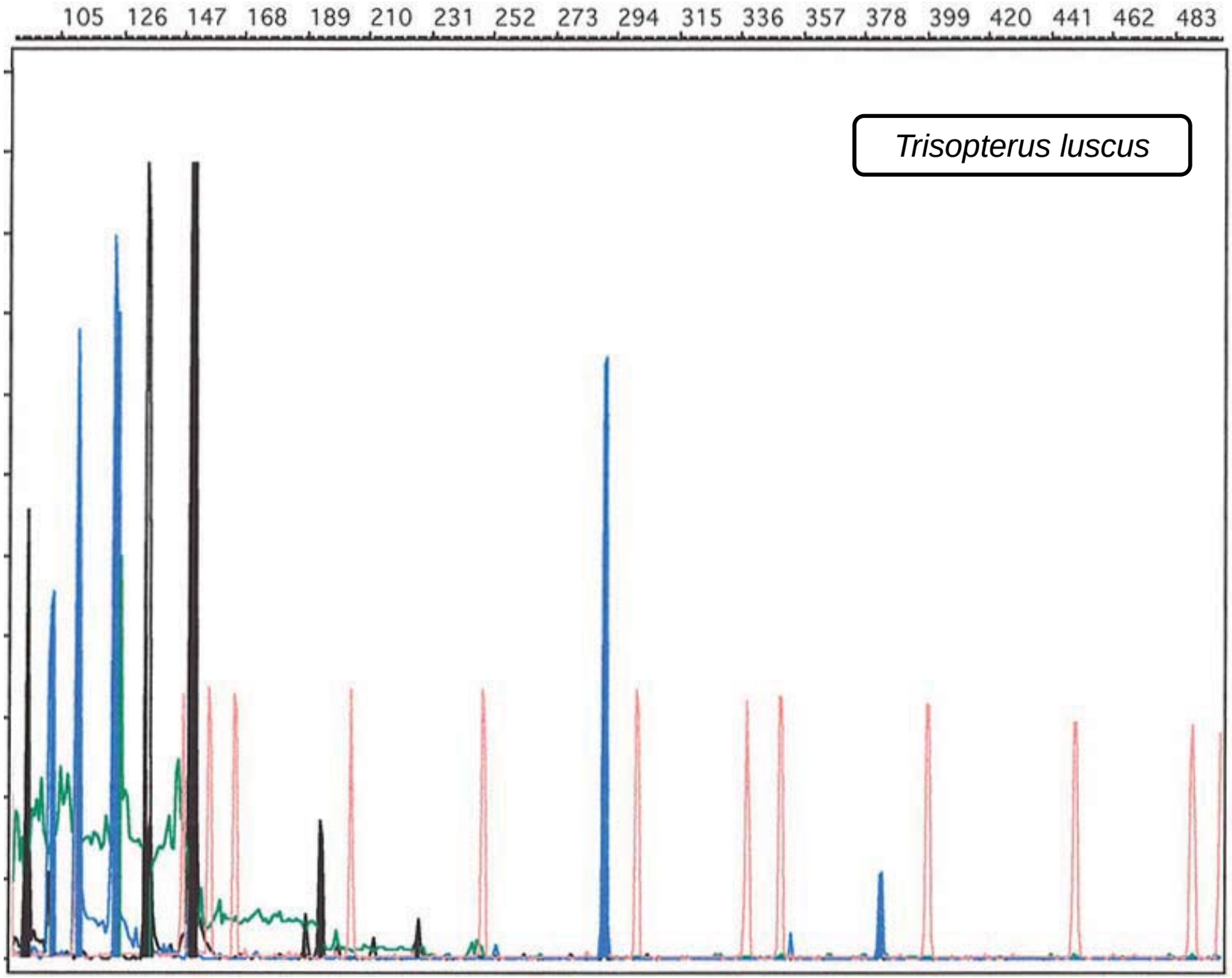
Supplementary Figure S12 (cont.)



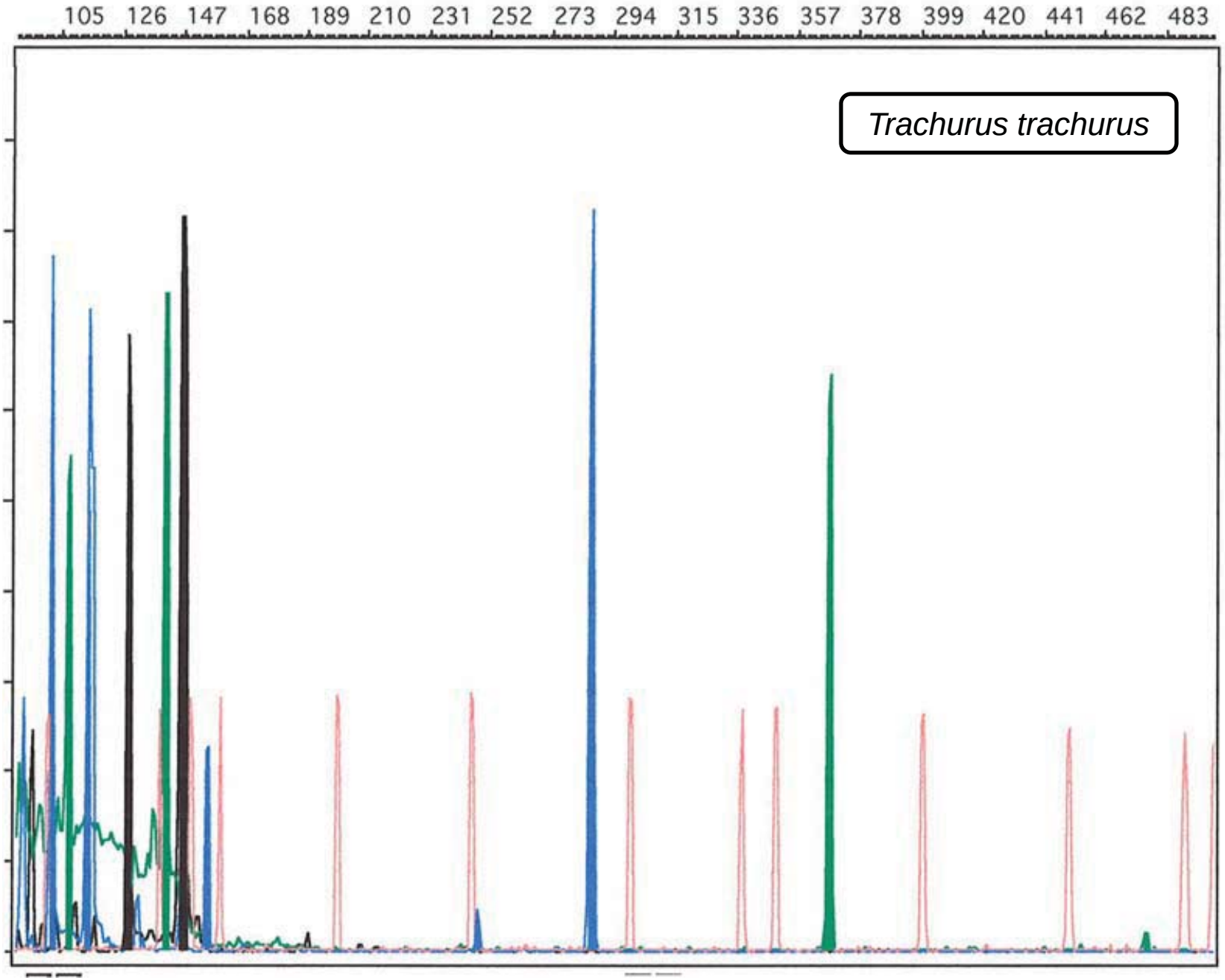
Supplementary Figure S12 (cont.)



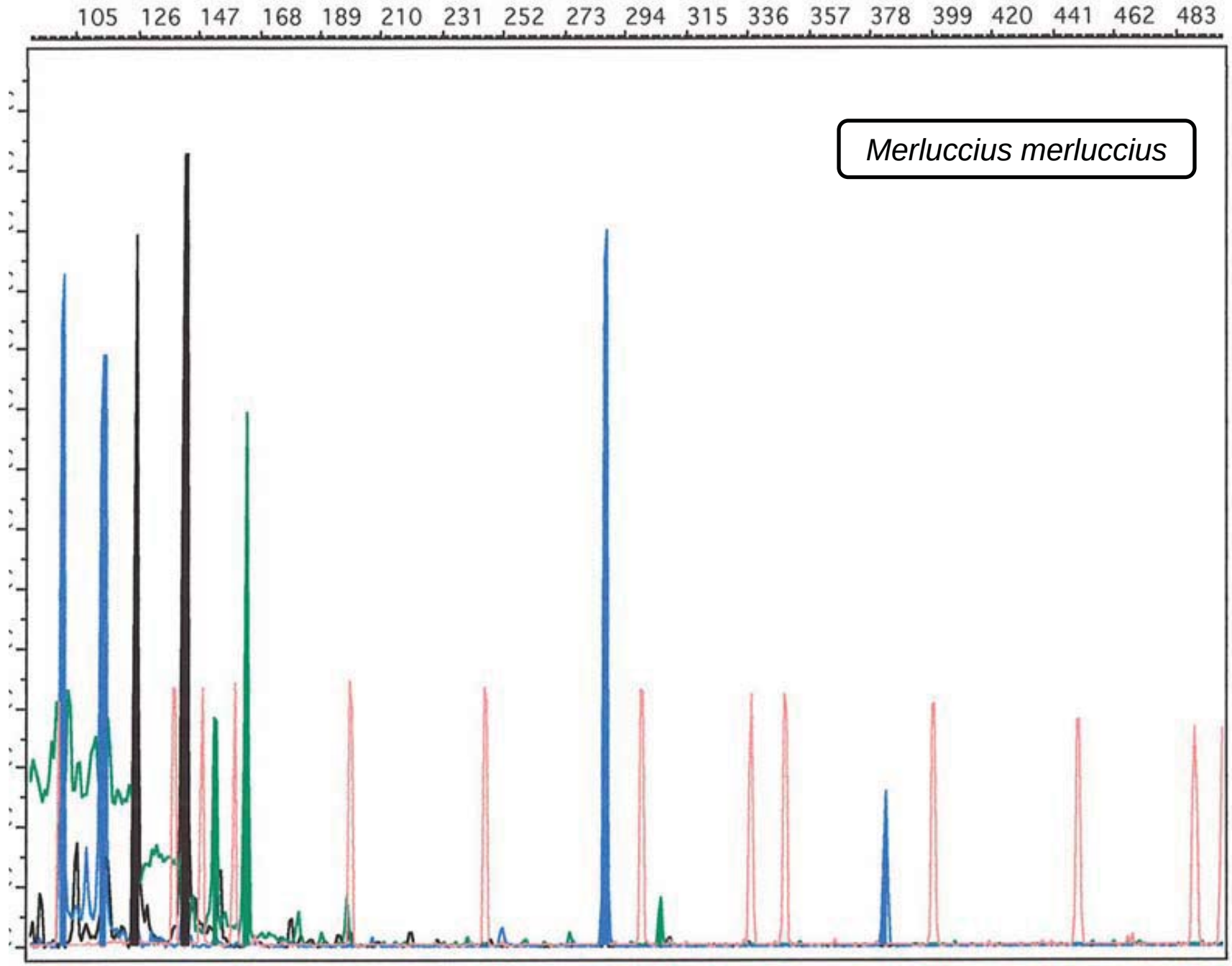
Supplementary Figure S12 (cont.)



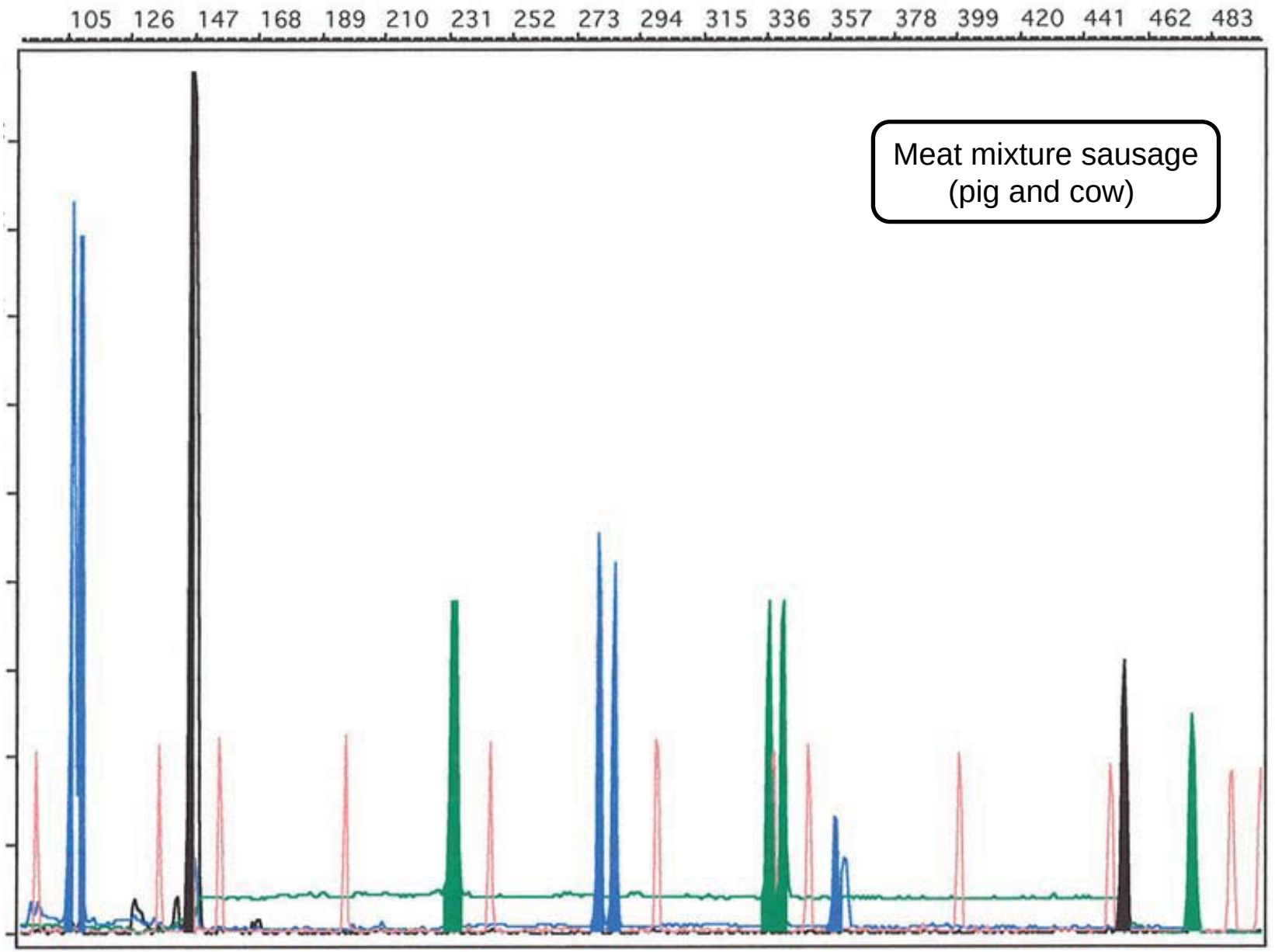
Supplementary Figure S12 (cont.)



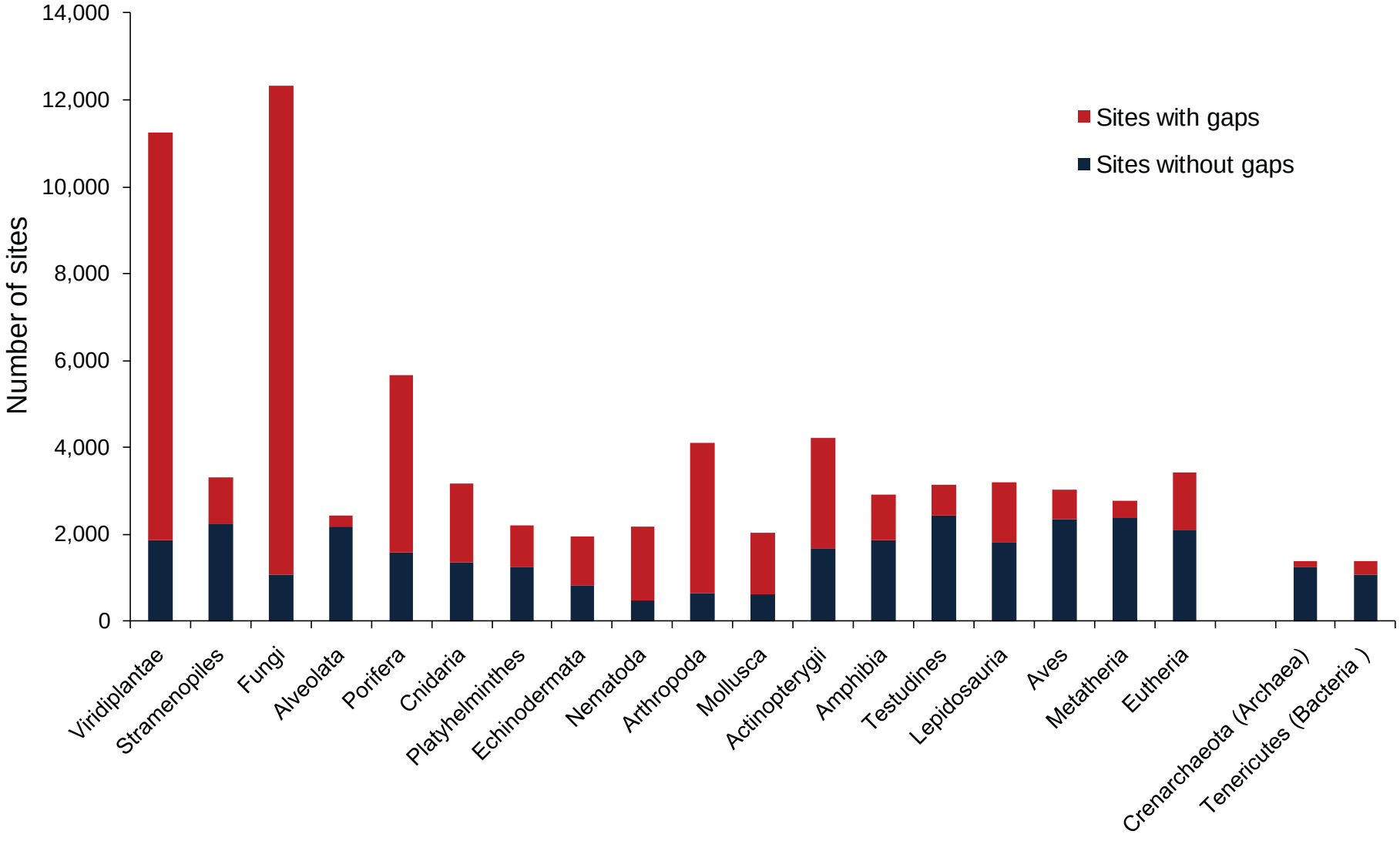
Supplementary Figure S12 (cont.)



Supplementary Figure S13. Detection of mixtures with the SPInDel assay. The image represents an electropherogram obtained by capillary electrophoresis with multicolor fluorescence detection in a food product with mixture of porcine and bovine biological material. The profile is displayed in a four-dye fluorescent system, with the green, blue and yellow channels used for detection of amplified products and the red channel used as a size marker.

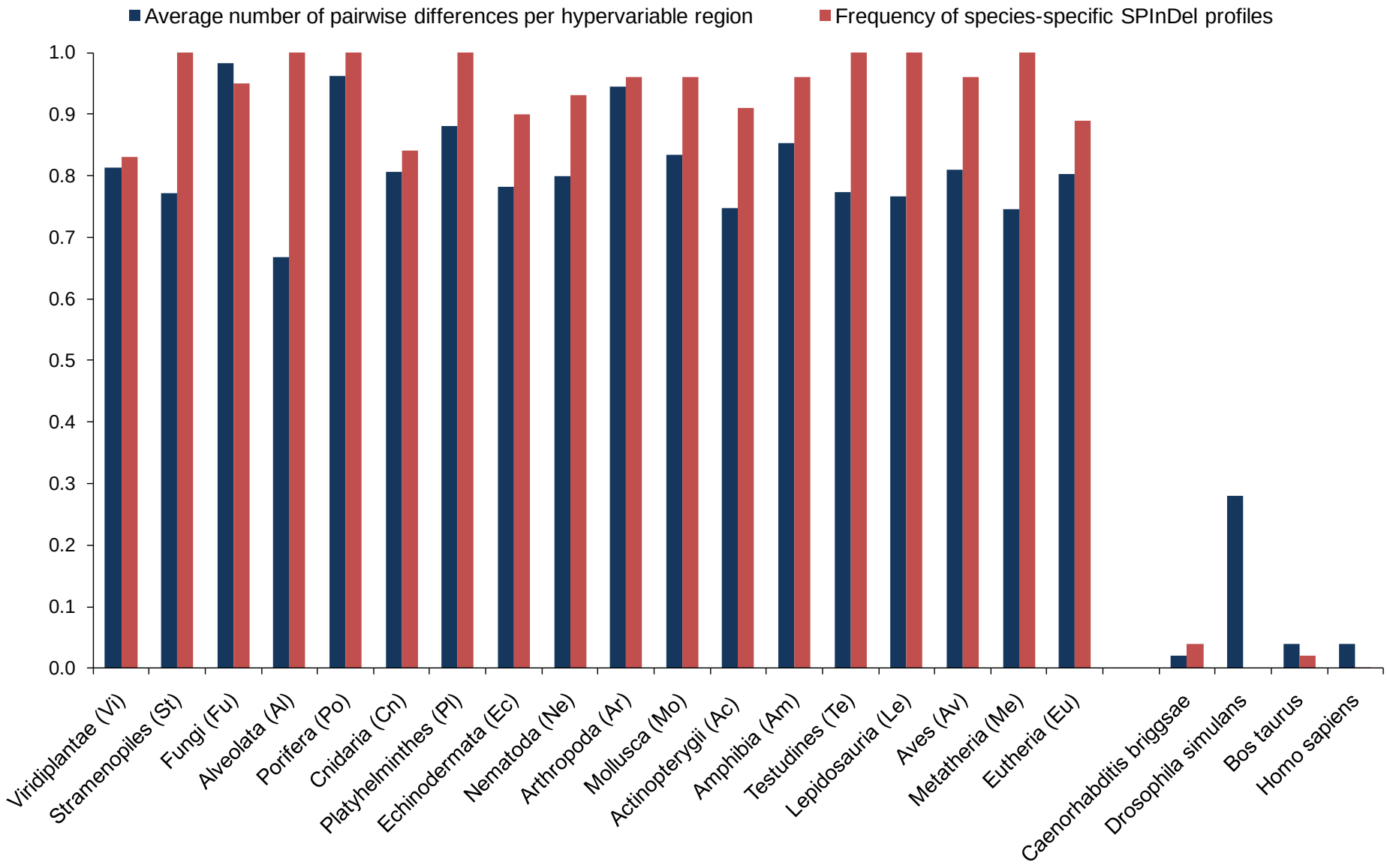


Supplementary Figure S14. Alignment gaps in ribosomal RNA genes. Distribution of sites with and without gaps in the sequence alignment of ribosomal RNA genes from eukaryotic and prokaryotic groups.

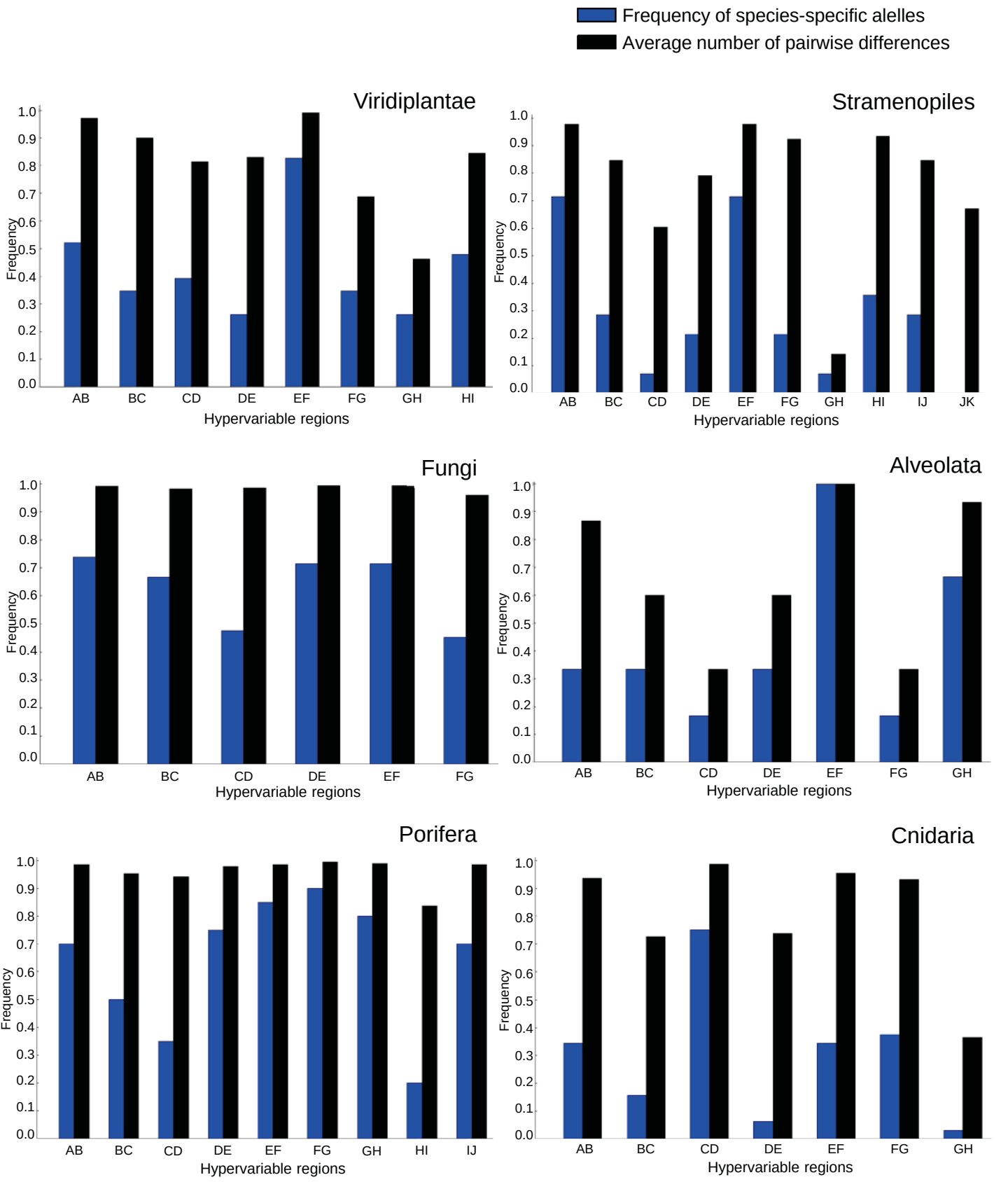


Supplementary Figure S15. Sequence alignment of duplicated ribosomal RNA (rRNA) genes. Nucleotide differences are indicated by white bars on identity plots (obtained in the Geneious software) for the alignment of rRNA genes on 10 species belonging to four eukaryotic groups.

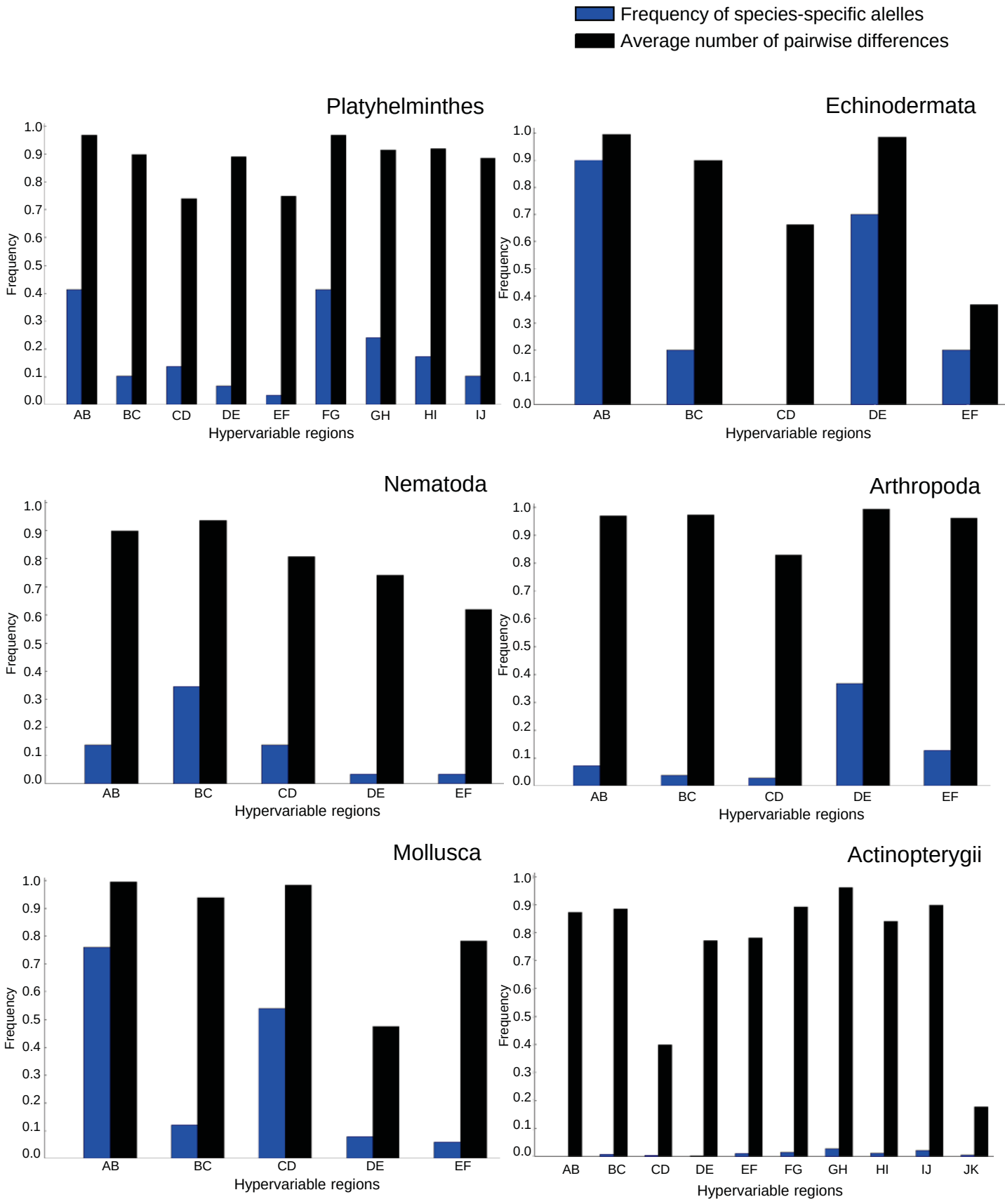
Supplementary Figure S16. Effectiveness of SPInDel as an identification tool. The frequency of species-specific SPInDel profiles (f_n^G) and the average number of pairwise differences per hypervariable region is represented for 18 eukaryotic groups and 4 intra-species datasets.



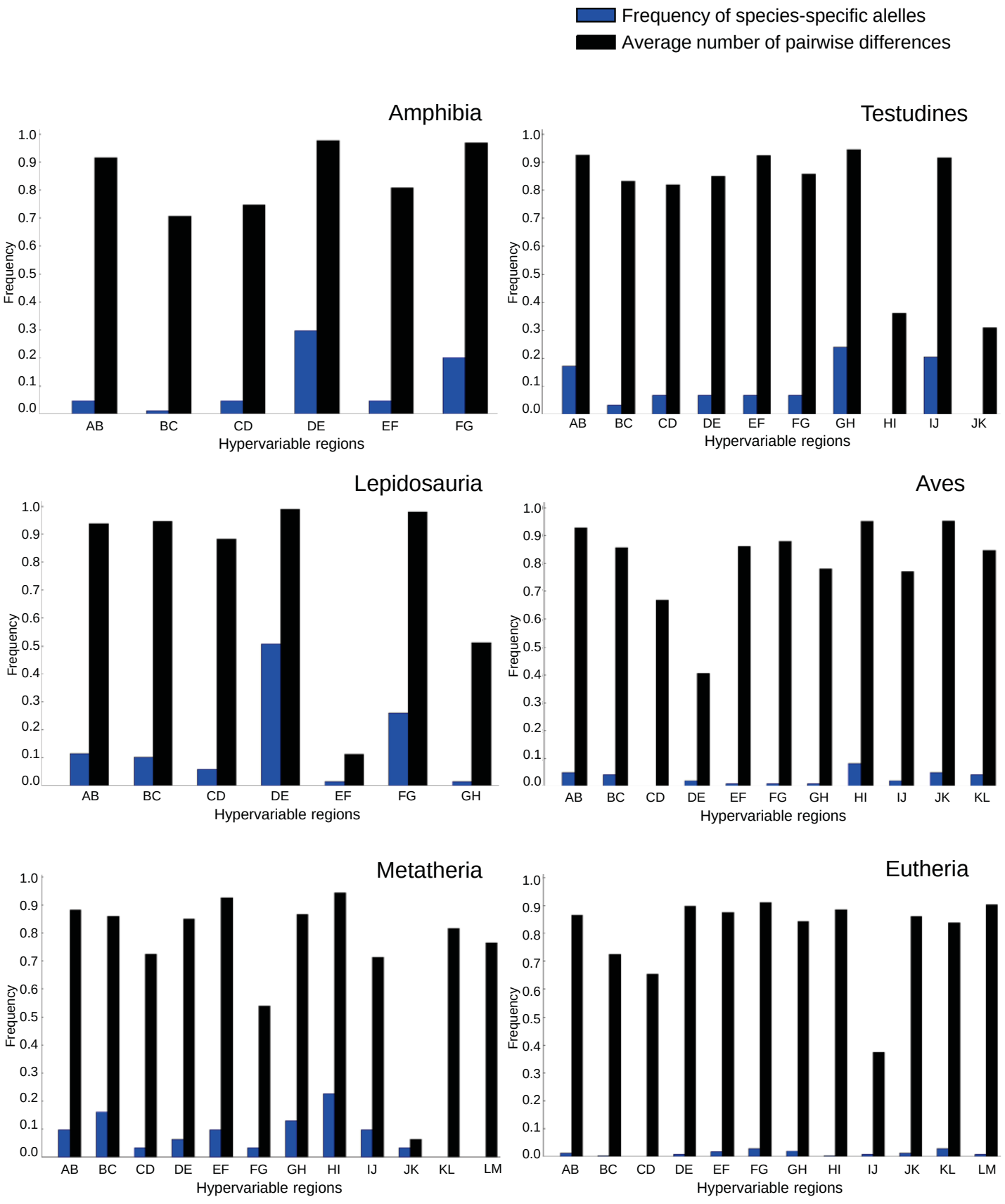
Supplementary Figure S17. Power of discrimination in each SPInDel hypervariable region. The frequency of species-specific alleles (f_1^G) and the average number of pairwise differences ($\bar{p}_1^G$) were estimated for 18 eukaryotic groups using the SPInDel workbench.



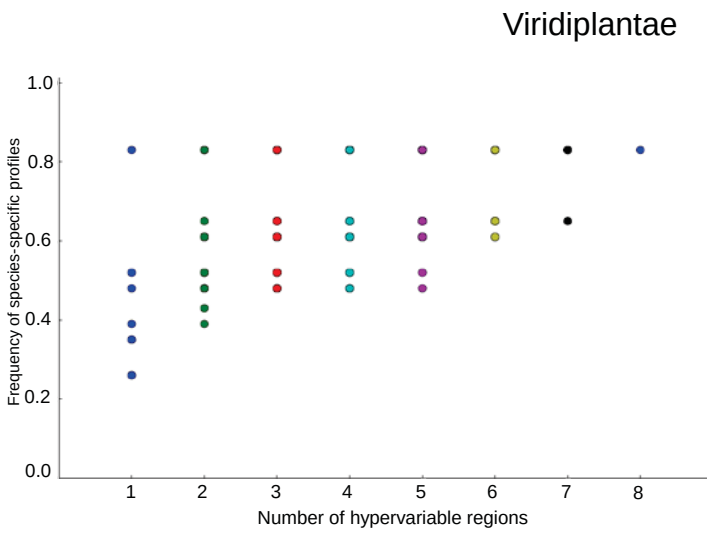
Supplementary Figure S17 (cont.)

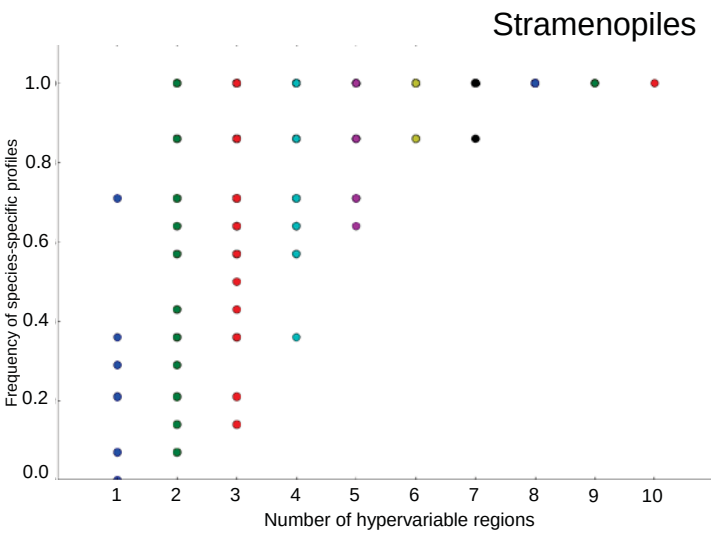


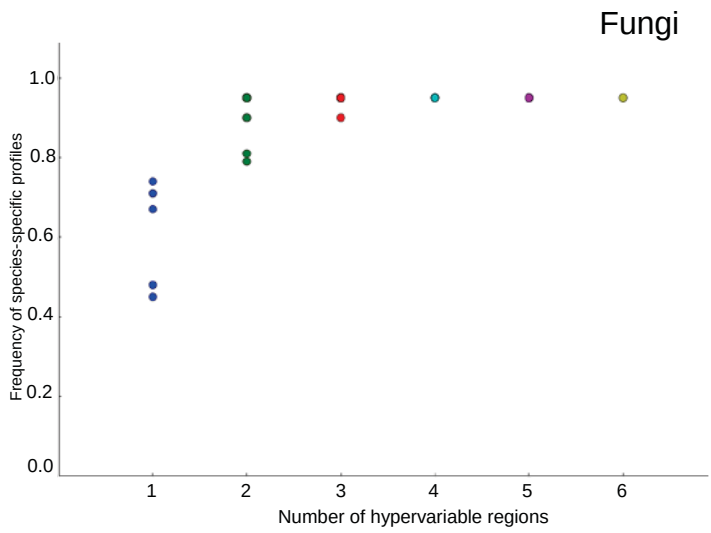
Supplementary Figure S17 (cont.)

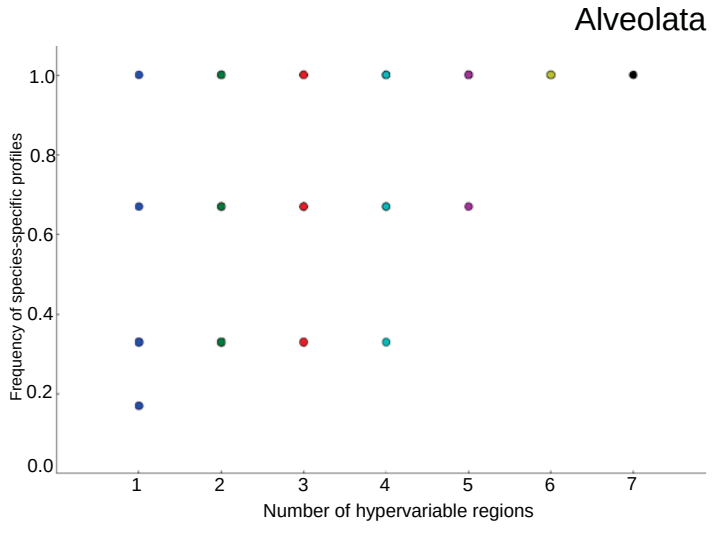


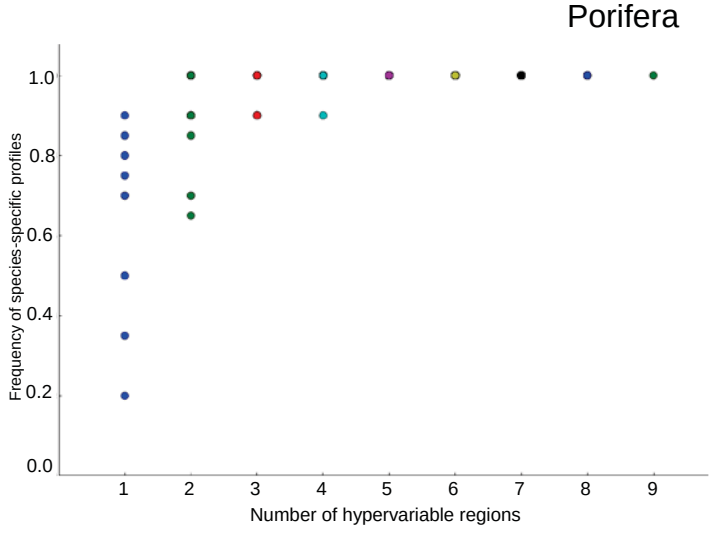
Supplementary Figure S18. Species discrimination achieved with different numbers of SPInDel hypervariable regions. The frequency of species-specific SPInDel profiles (y-axis) is plotted for all *m*-combinations from a set with *n* hypervariable regions (x-axis), for *m* from 1 to *n*.

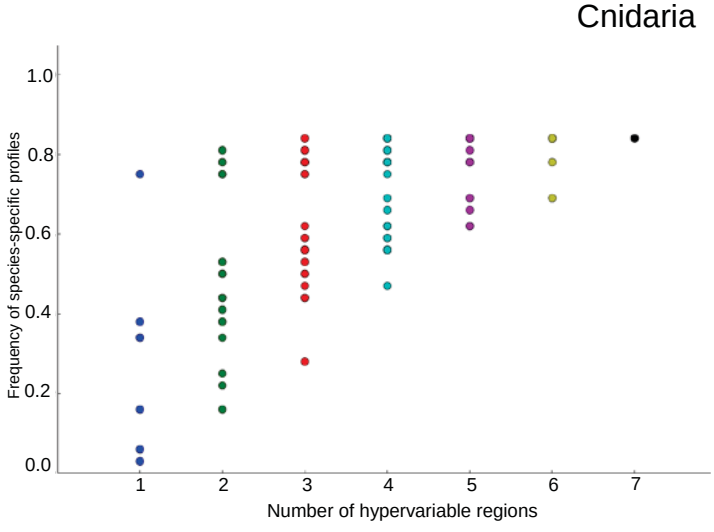




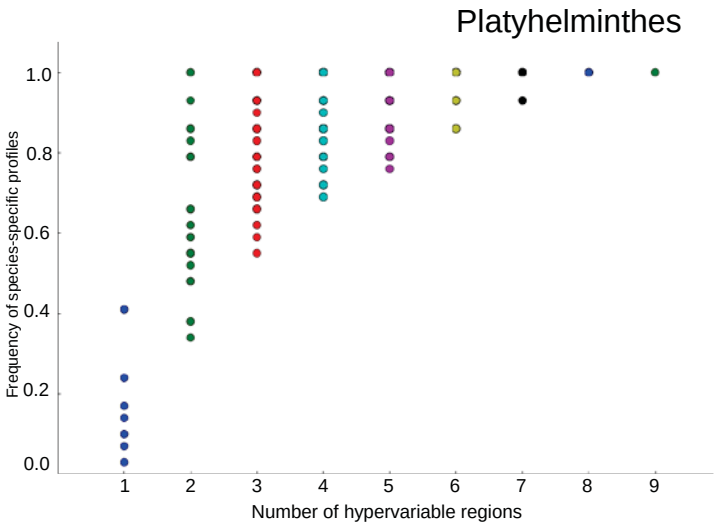


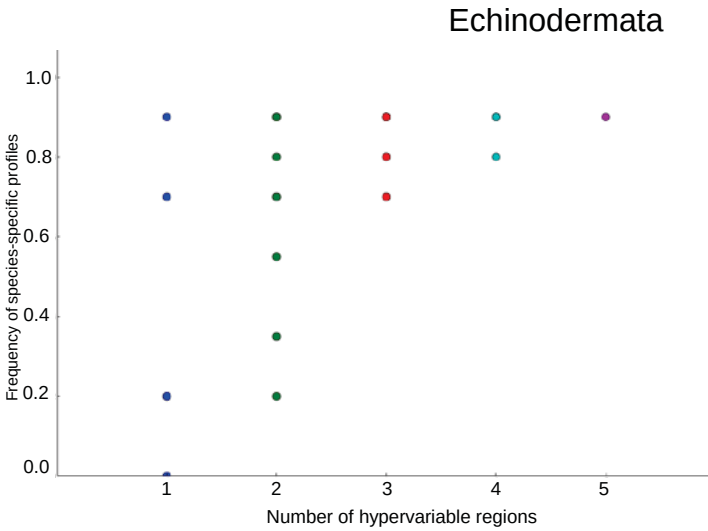


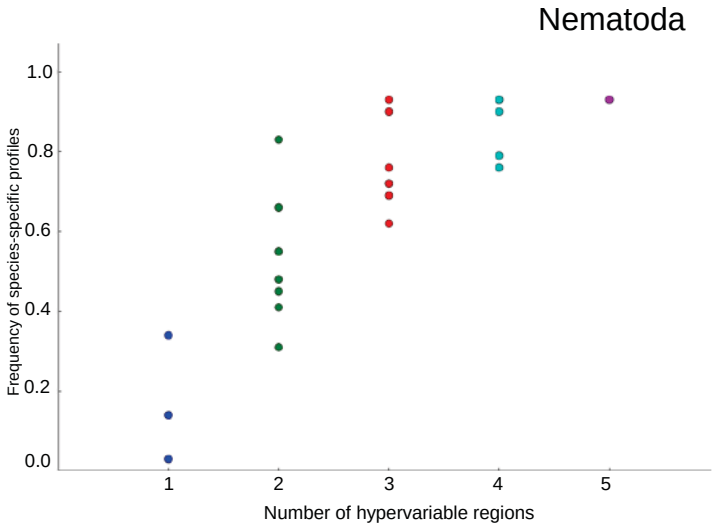


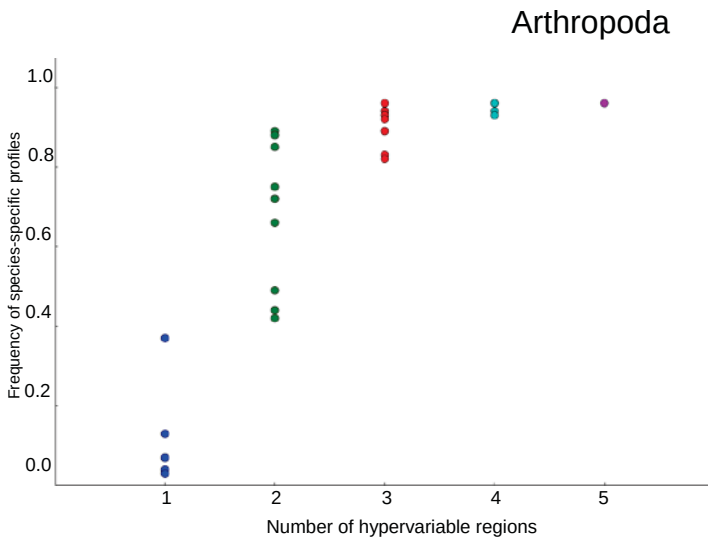


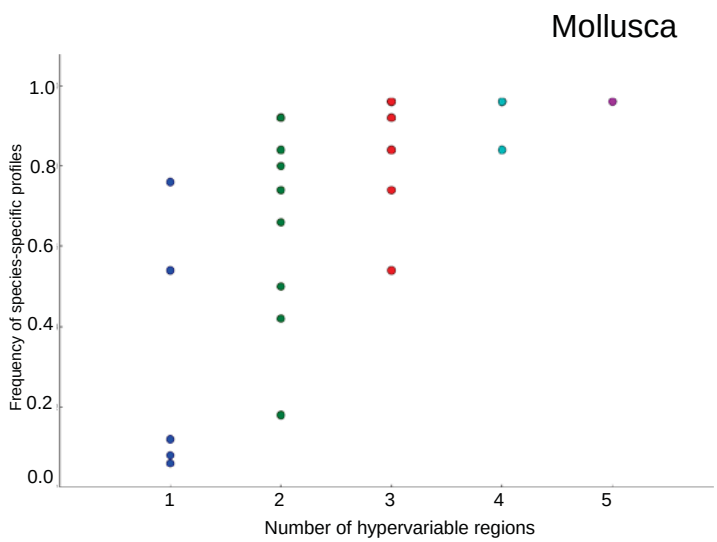
Supplementary Figure S18 (cont.)

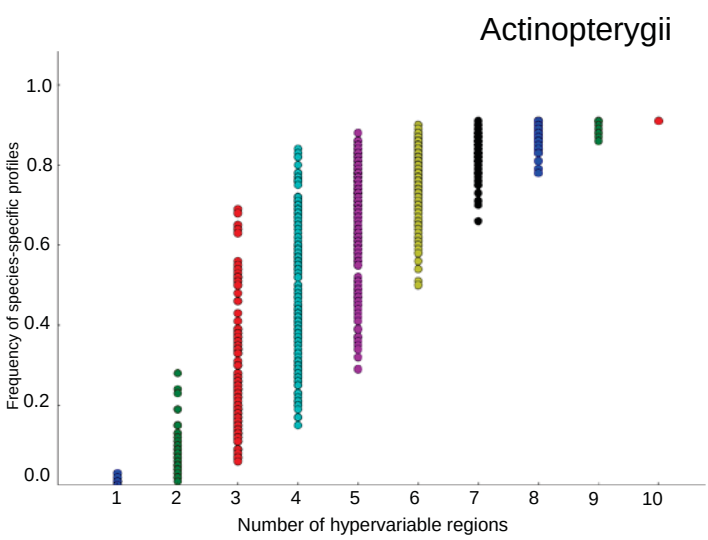




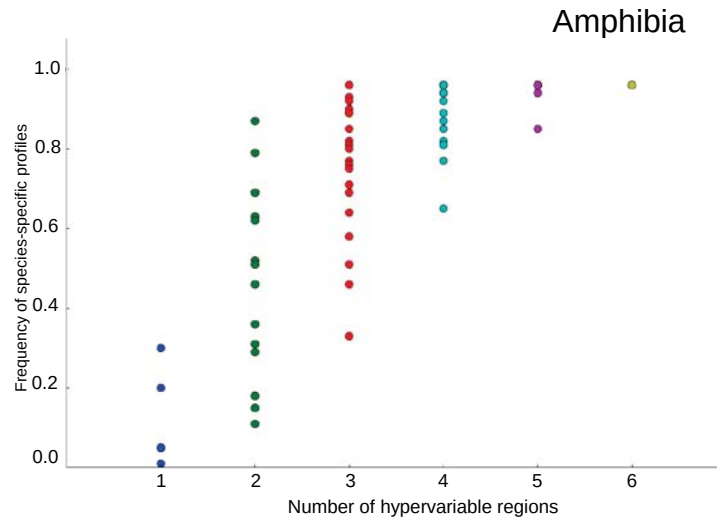


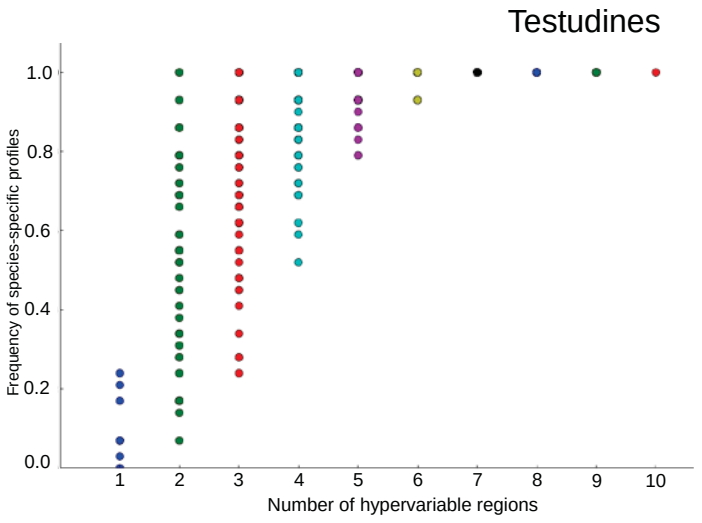


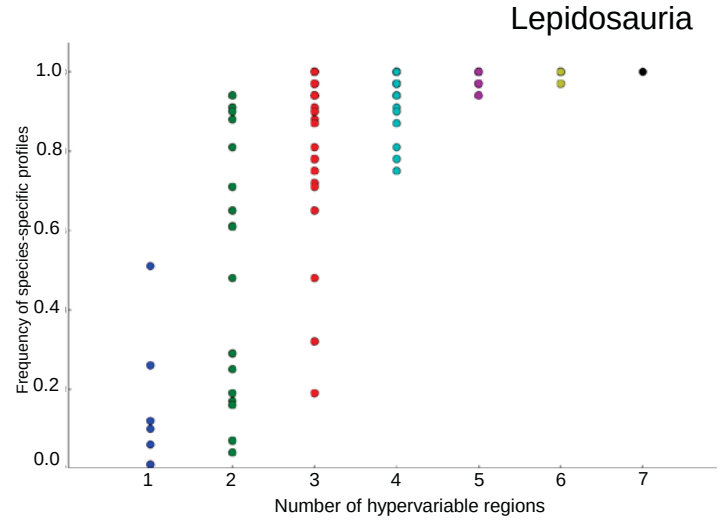


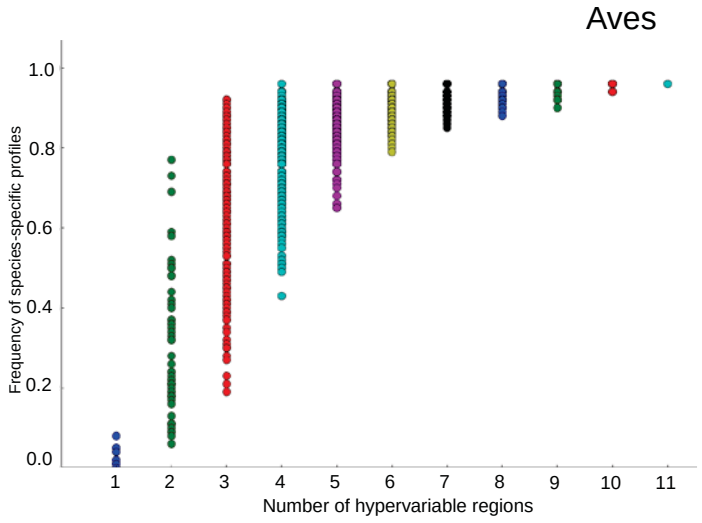


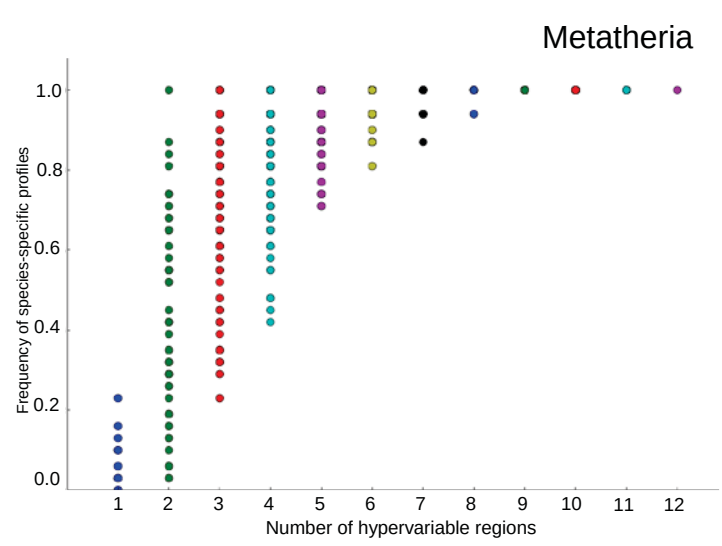
Supplementary Figure S18 (cont.)

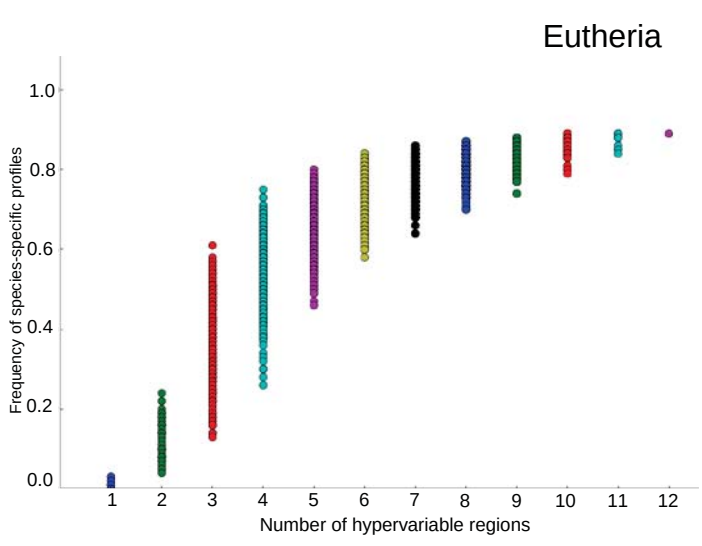




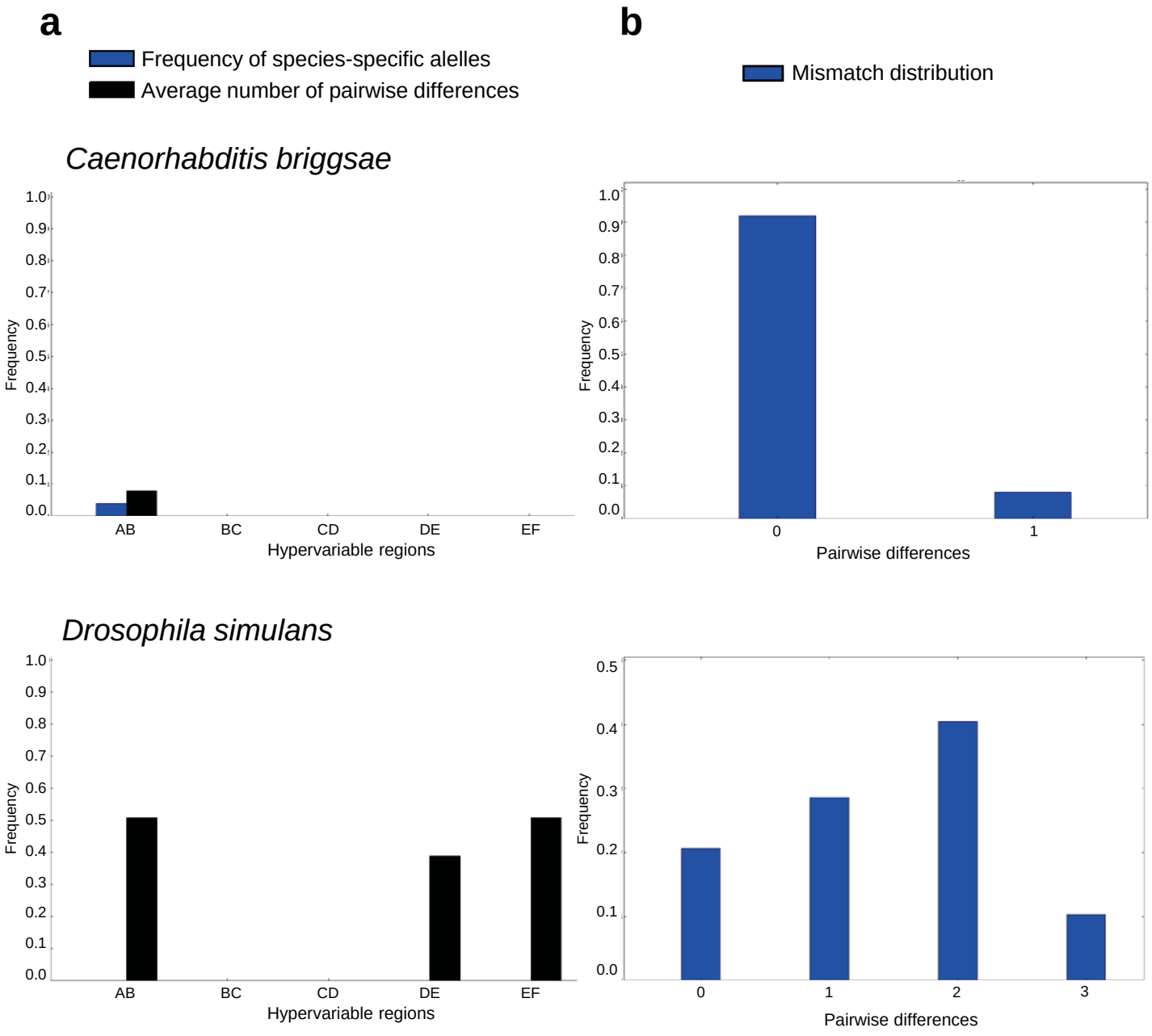




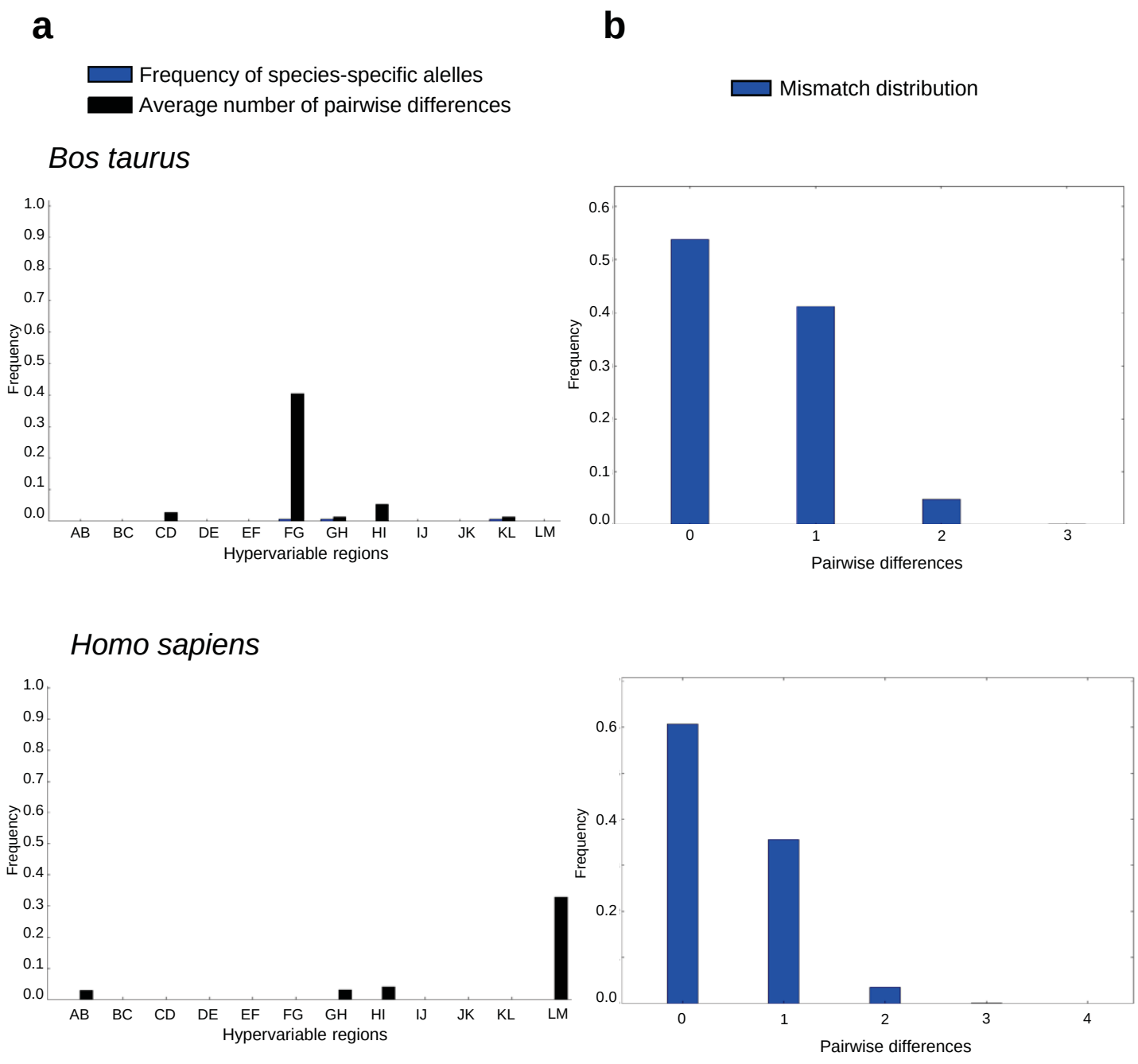




Supplementary Figure S19. Reduced diversity of SPInDel profiles in four intra-species datasets. (**a**) Frequency of species-specific alleles (f_n^G) and the average number of pairwise differences ($\bar{p}_n^G$) per hypervariable region. (**b**) Mismatch distribution.

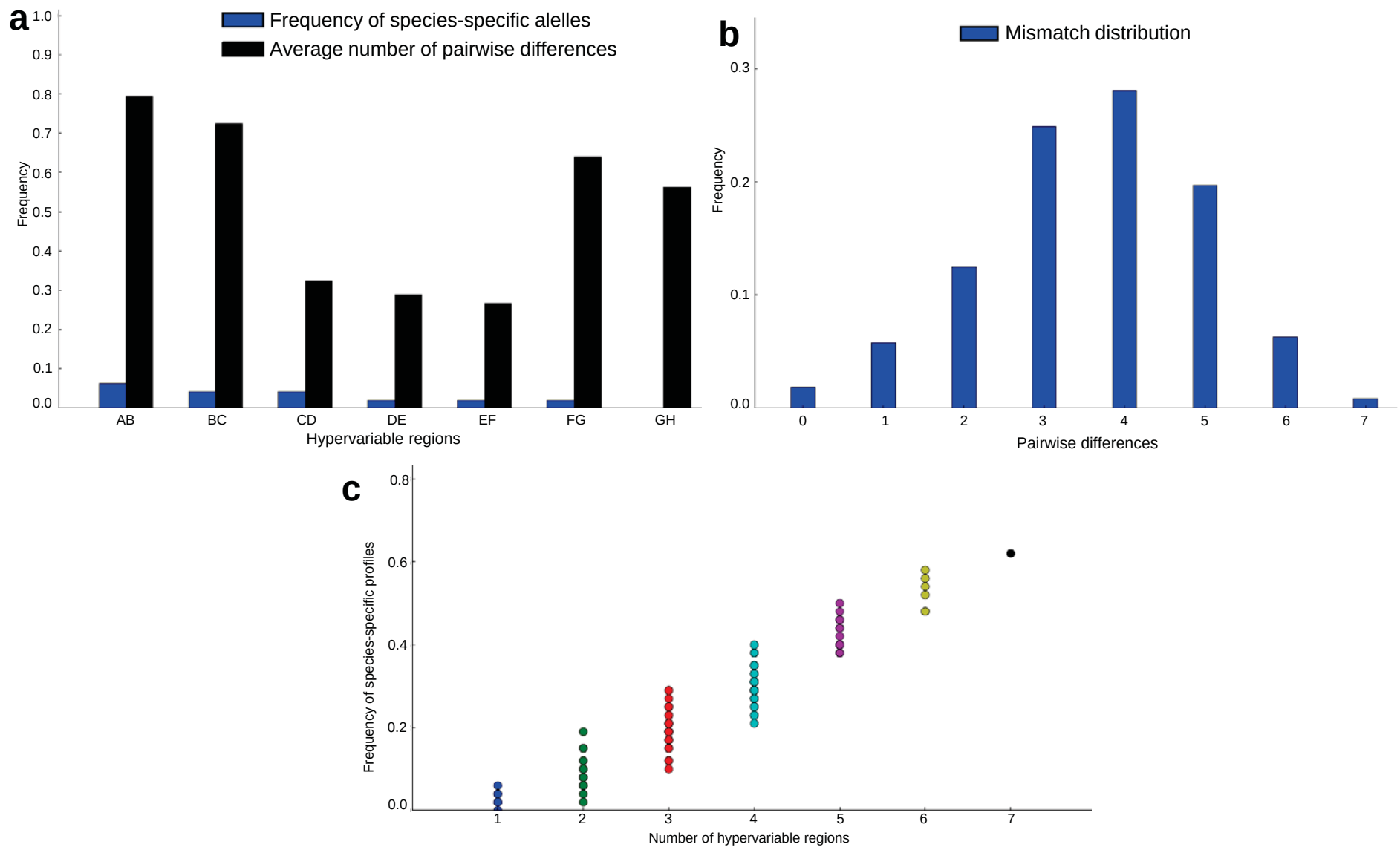


Supplementary Figure S19 (cont.)



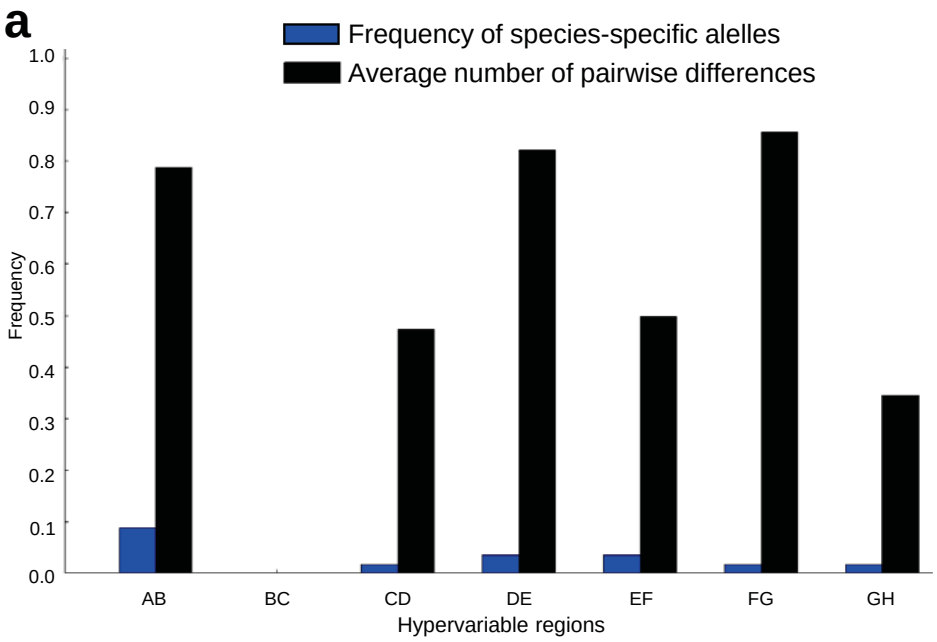
Supplementary Figure S20. Species identification in prokaryotic and viral species using the SPInDel method. (**a**) Frequency of species-specific alleles (f_n^G) and the average number of pairwise differences ($\bar{p}_n^G$) in each hypervariable region. (**b**) Mismatch distribution. (**c**) Frequency of species-specific SPInDel profiles (y-axis) for all *m*-combinations from a set with *n* hypervariable regions (x-axis), for *m* from 1 to *n*.

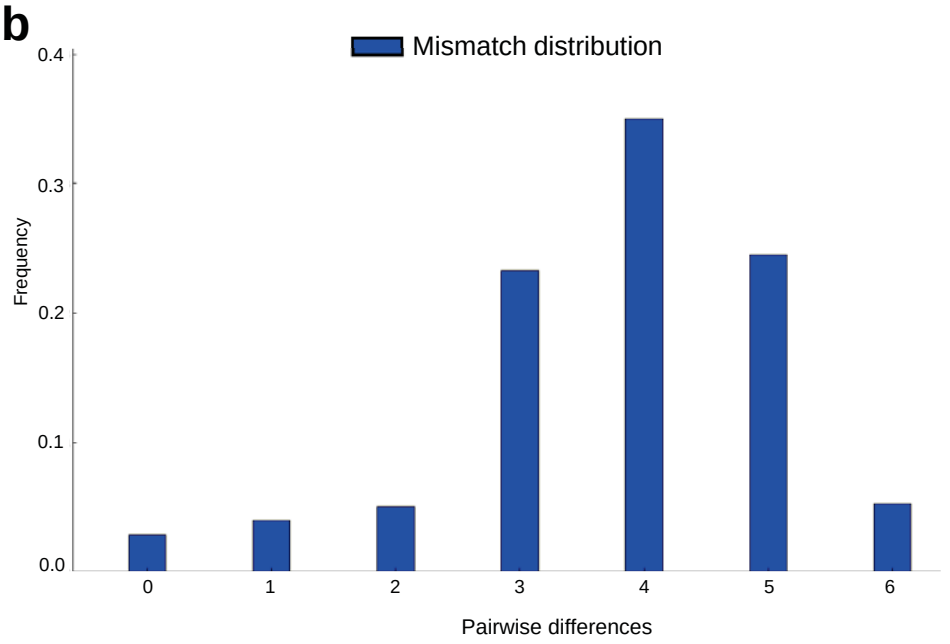
Crenarchaeota (Archaea)

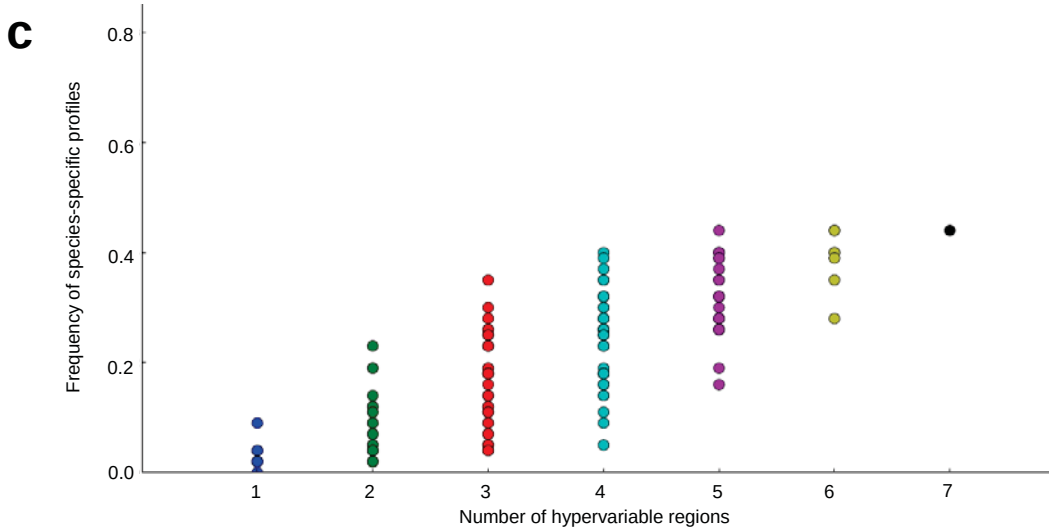


Supplementary Figure S20 (cont.)

Tenericutes (Bacteria)

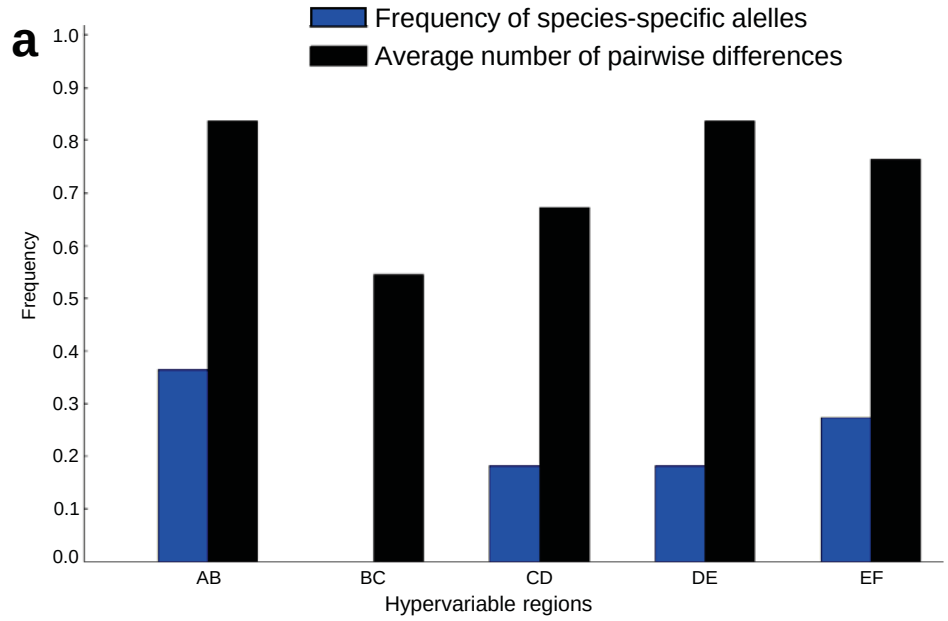


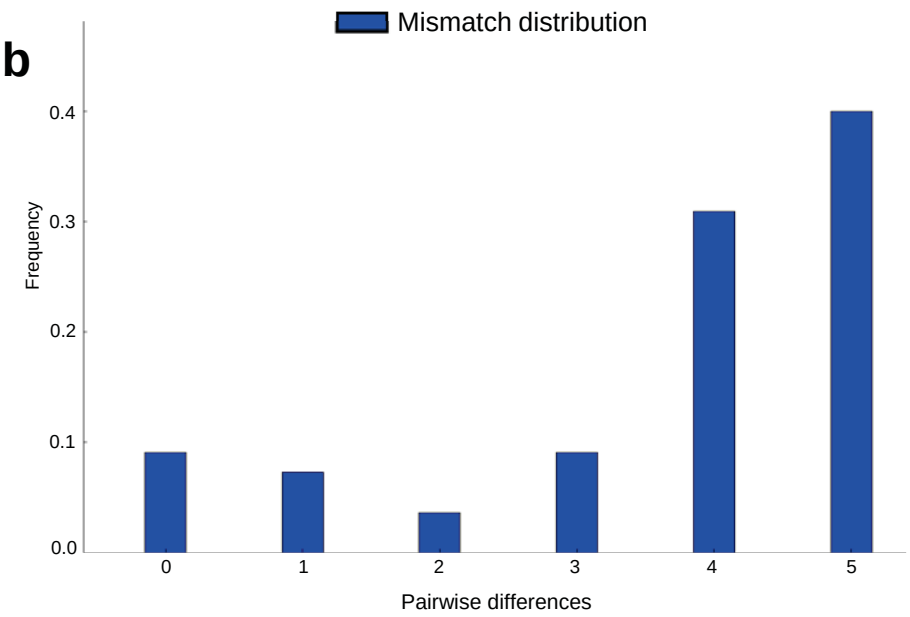


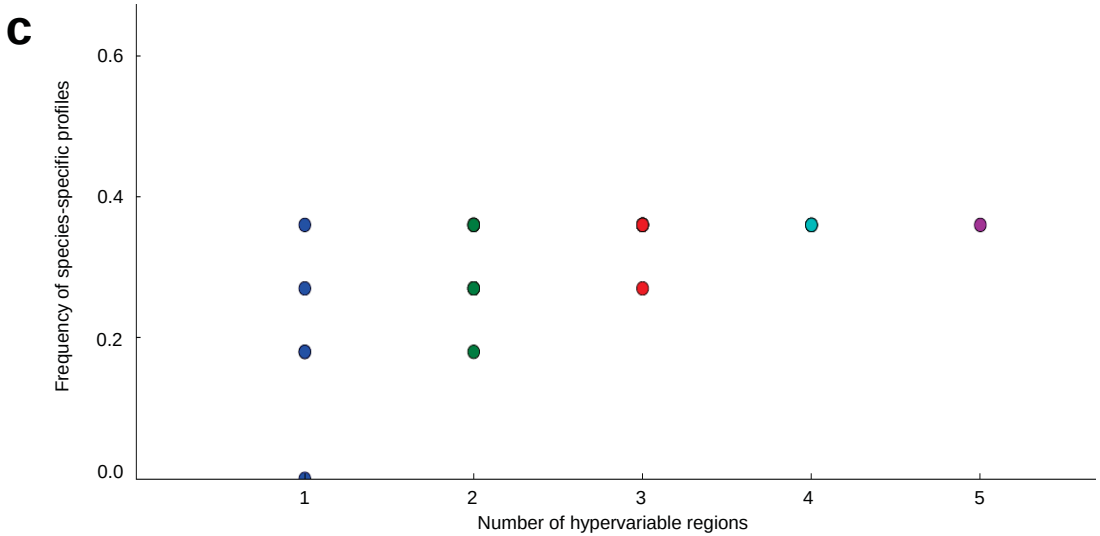


Supplementary Figure S20 (cont.)

Lentivirus (Retroviridae)

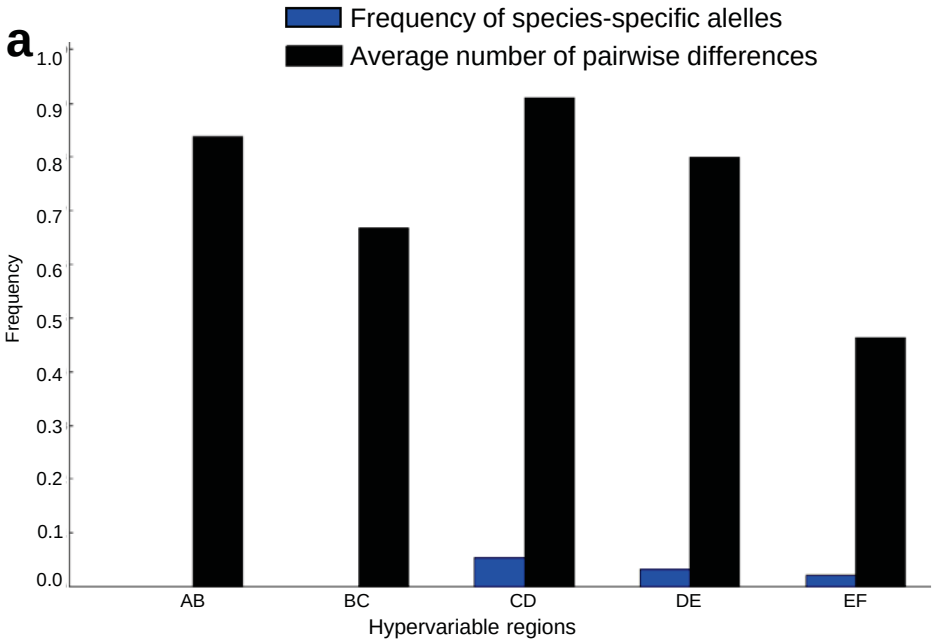


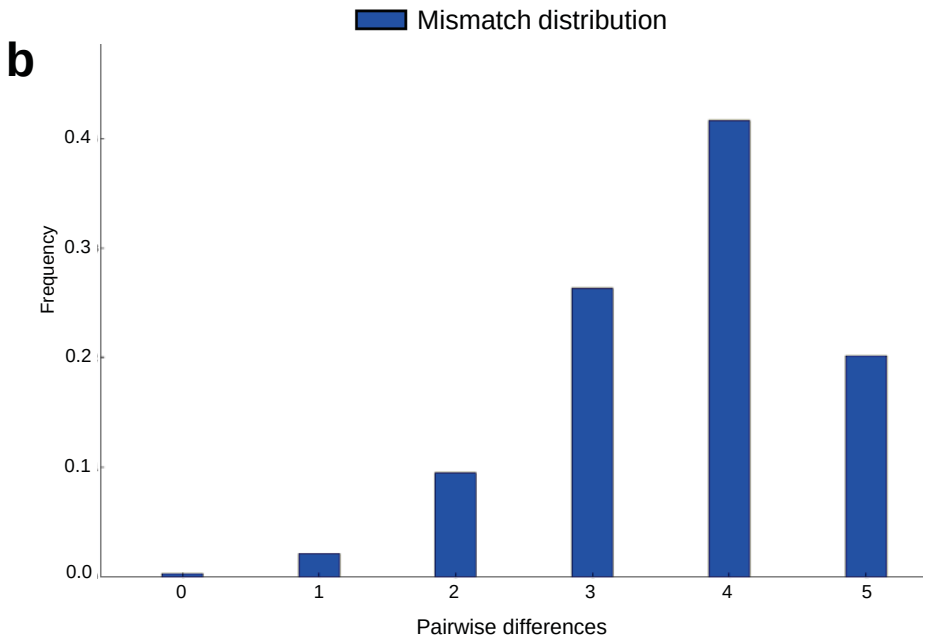


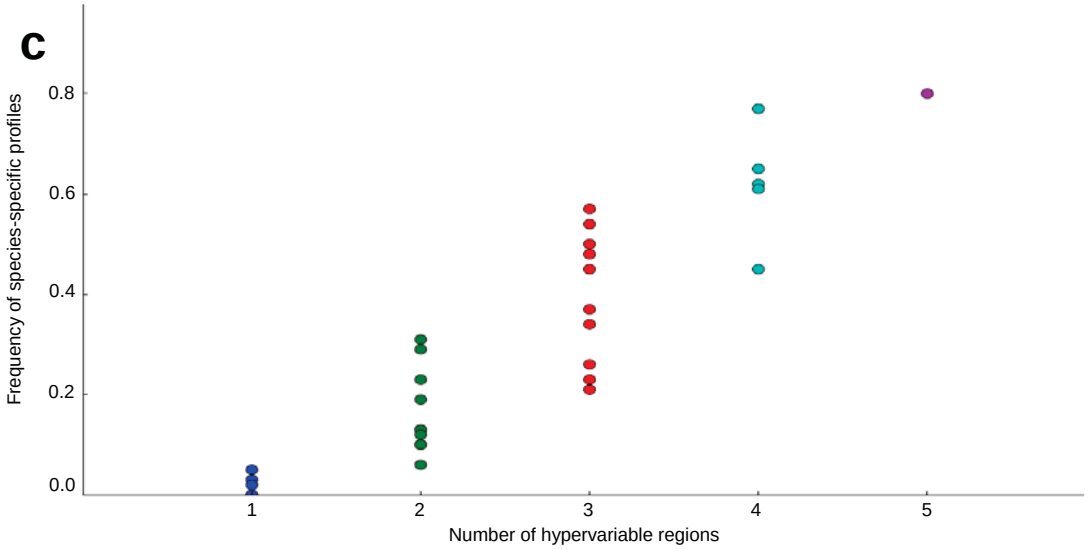


Supplementary Figure S20 (cont.)

Papillomaviridae (dsDNA viruses)

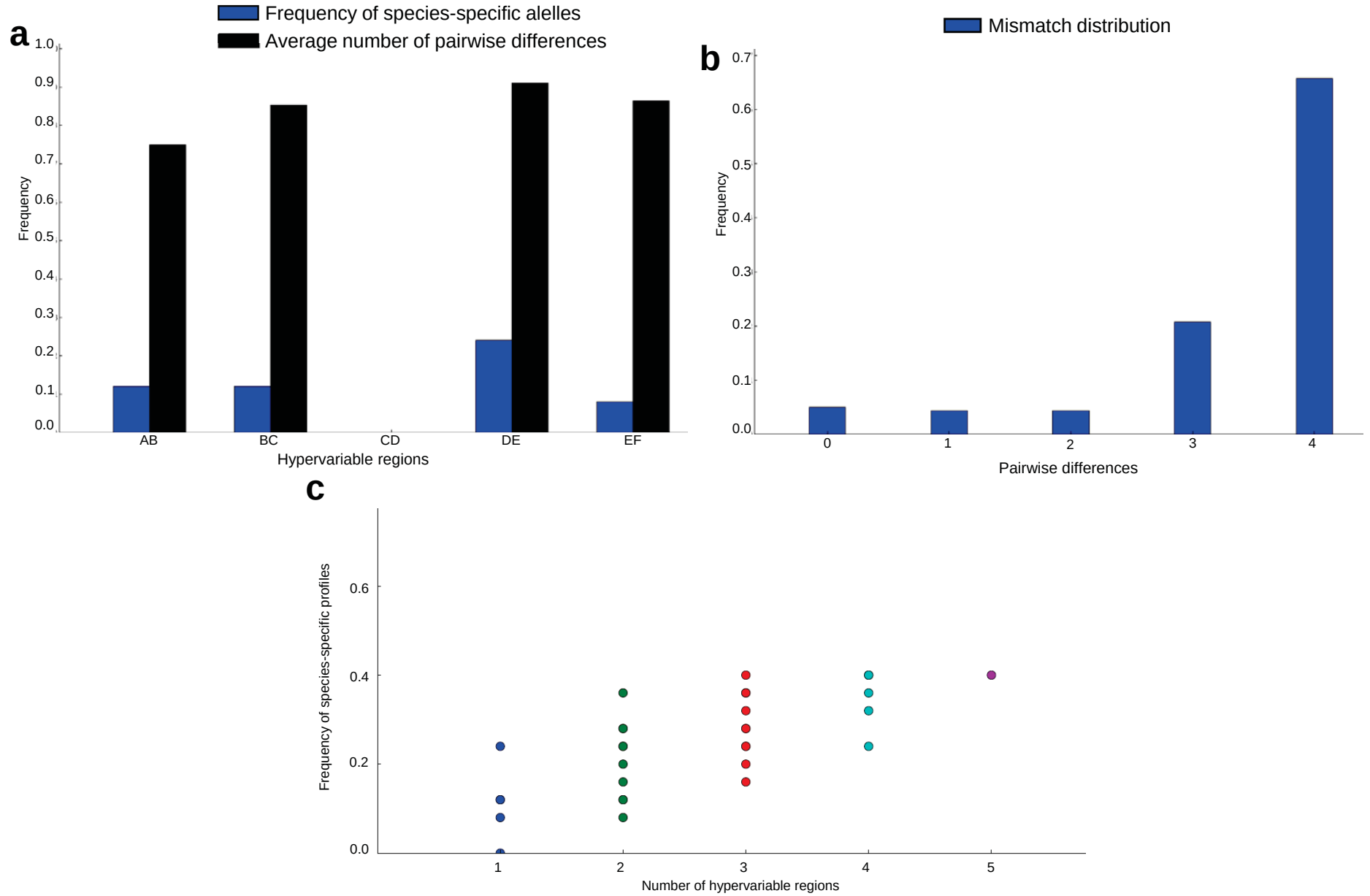




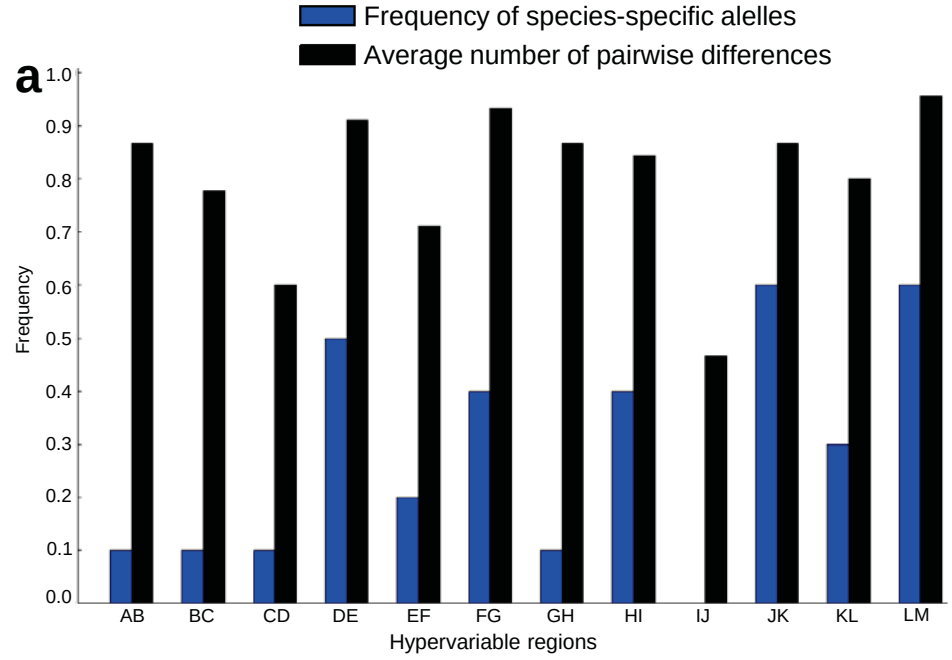


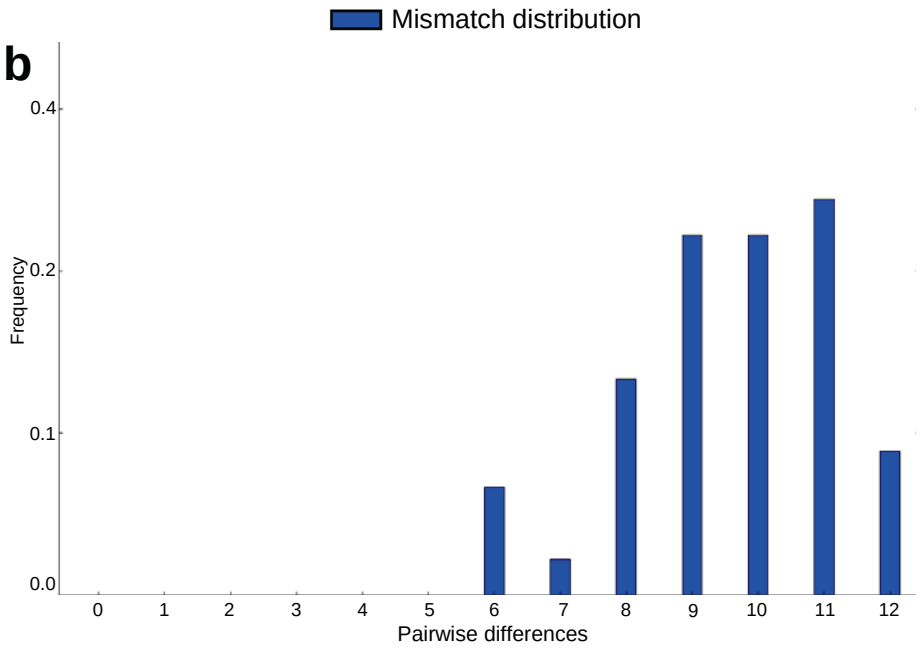
Supplementary Figure S20 (cont.)

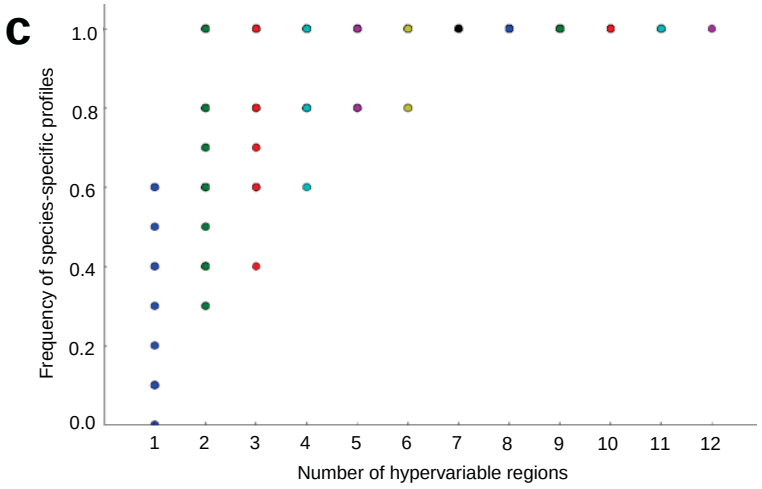
Rhabdoviridae (ssRNA viruses)



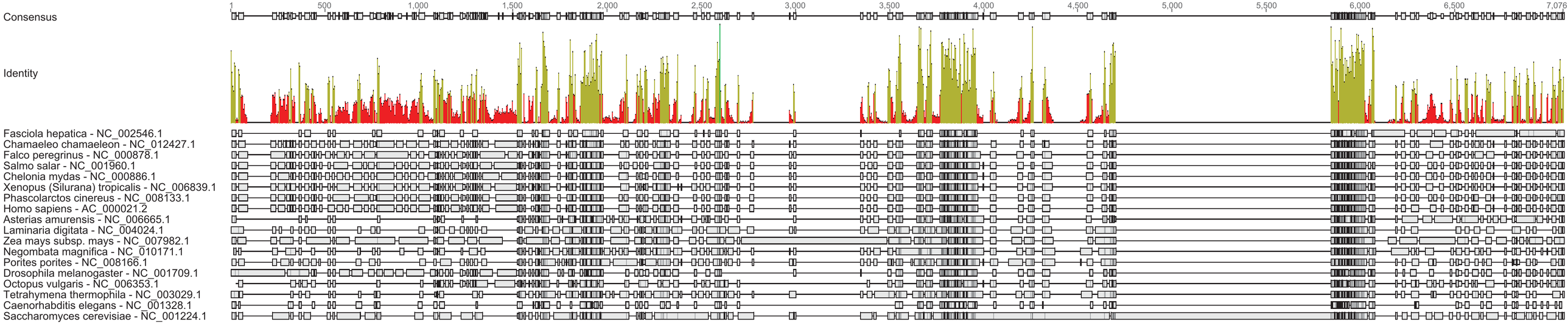
Supplementary Figure S21. Efficacy of the SPInDel method in the identification of 10 eutherian species. (**a**) Frequency of species-specific alleles (f_n^G) and the average number of pairwise differences ($\bar{p}_n^G$) per hypervariable region. (**b**) Mismatch distribution. (**c**) Frequency of species-specific SPInDel profiles (y-axis) for all *m*-combinations from a set with *n* hypervariable regions (x-axis), for *m* from 1 to *n*.







Supplementary Figure S22. The potential use of SPInDel for discrimination of divergent eukaryotic species. The image shows the alignment of mitochondrial large subunit ribosomal RNA genes sequences from a representative sample of 18 eukaryotic taxonomic groups. The identity plot represents the distribution of conserved (green and yellow bars) and variable (red bars) sites across the sequence alignment (obtained in Geneious software).


Consensus
Identity
1
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
4,500
5,000
5,500
6,000
6,500
7,0
Fasciola hepatica - NC_002546.1
Chamaeleo chamaeleon - NC_012427.1
Falco peregrinus - NC_000878.1
Salmo salar - NC_001960.1
Chelonia mydas - NC_000886.1
Xenopus (Silurana) tropicalis - NC_006839.1
Phascolarctos cinereus - NC_008133.1
Homo sapiens - AC_000021.2
Asterias amurensis - NC_006665.1
Laminaria digitata - NC_004024.1
Zea mays subsp. mays - NC_007982.1
Negombata magnifica - NC_010171.1
Porites porites - NC_008166.1
Drosophila melanogaster - NC_001709.1
Octopus vulgaris - NC_006353.1
Tetrahymena thermophila - NC_003029.1
Caenorhabditis elegans - NC_001328.1
Saccharomyces cerevisiae - NC_001224.1