

# Supplementary Material

## Part 2

### Identification of species by multiplex analysis of variable-length sequences

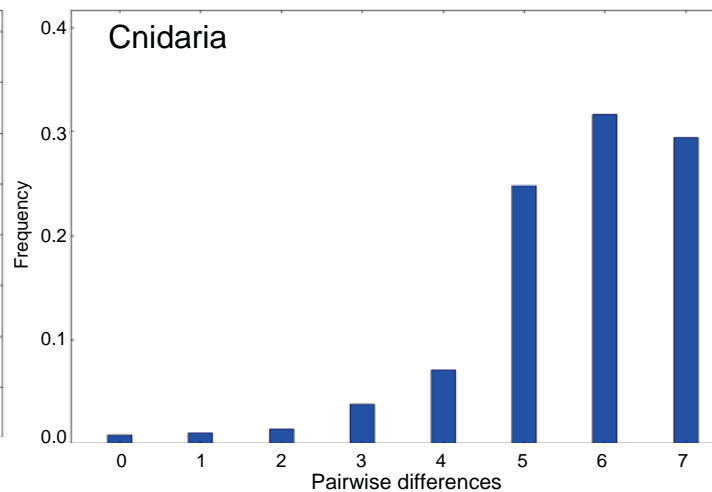
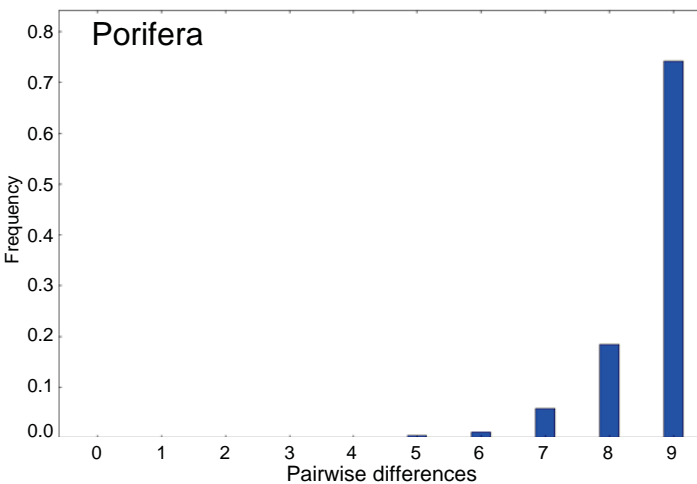
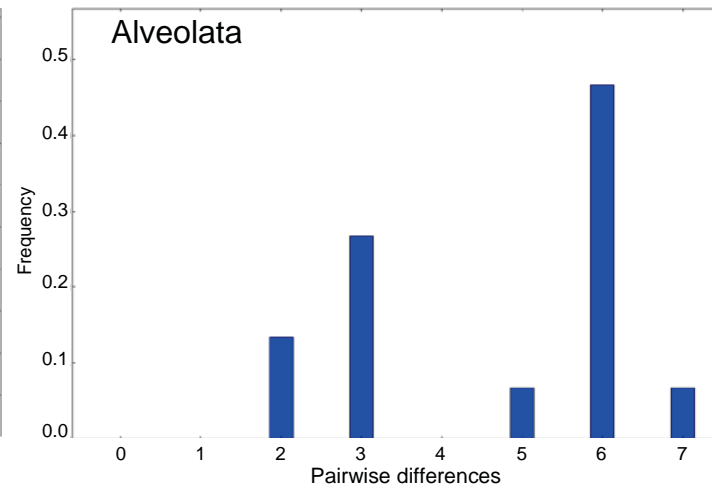
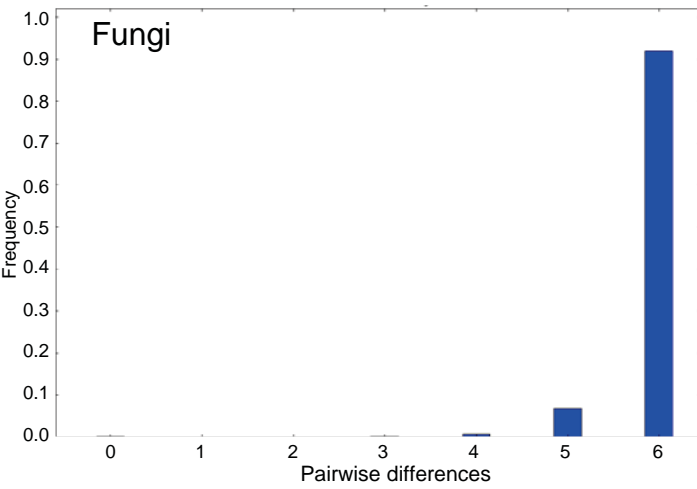
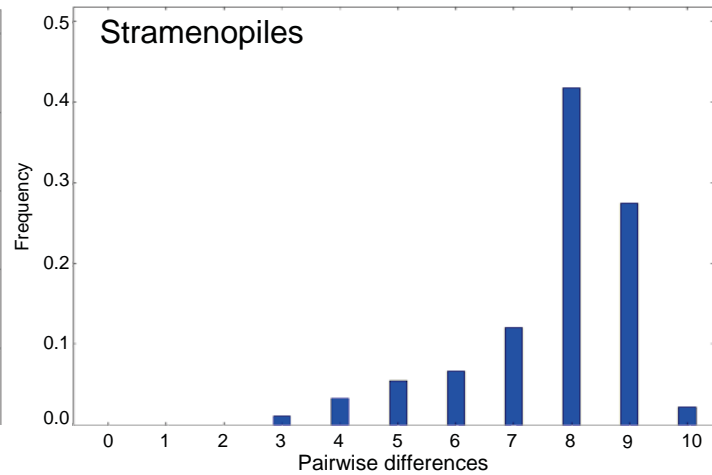
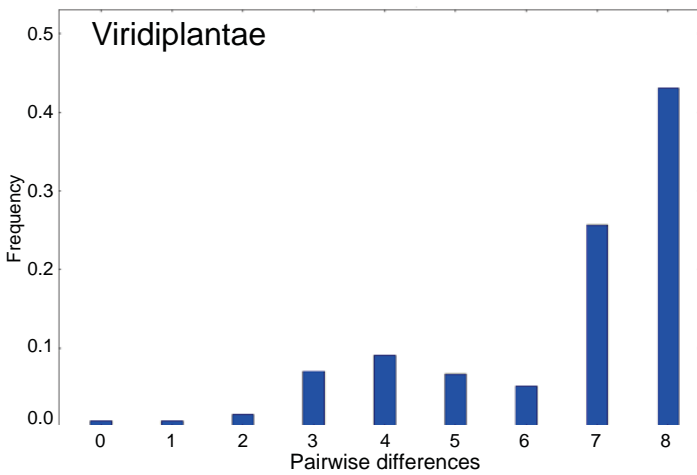
Filipe Pereira<sup>1</sup>, João Carneiro<sup>1</sup>, Rune Matthiesen<sup>1</sup>, Barbara van Asch<sup>1,2</sup>, Nádía Pinto<sup>1,2,3</sup>, Leonor Gusmão<sup>1</sup> and António Amorim<sup>1,2</sup>

<sup>1</sup> Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

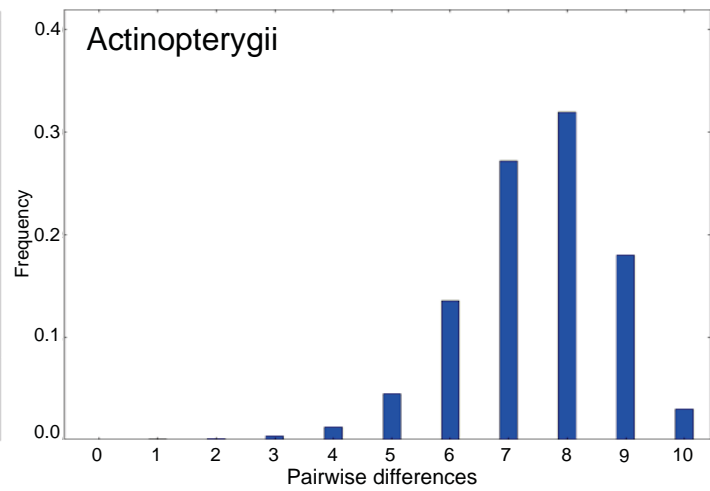
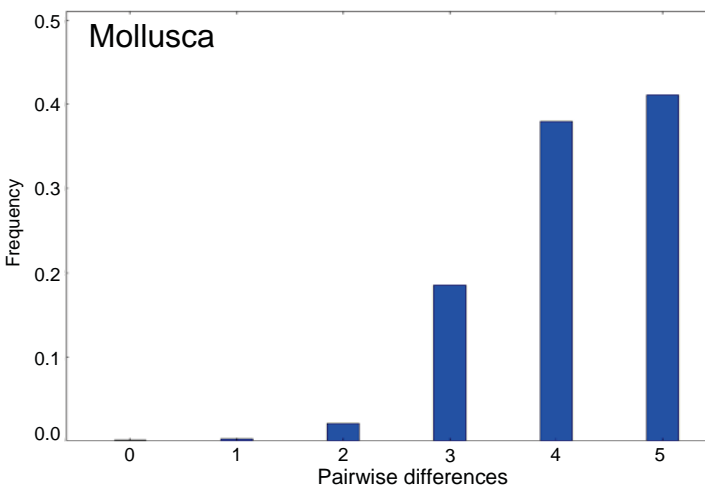
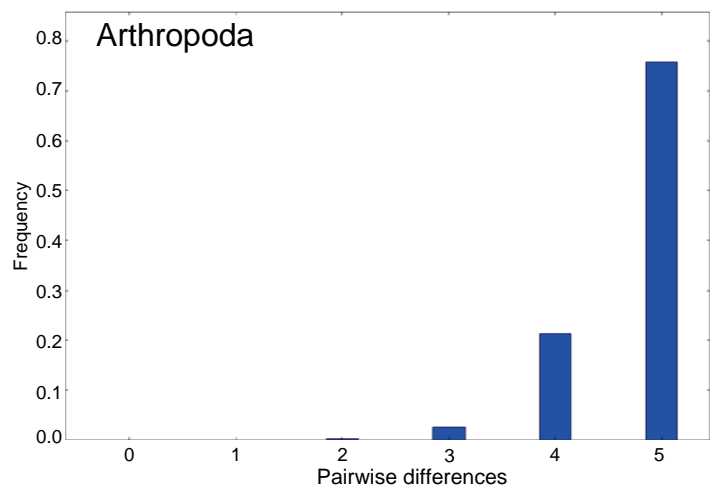
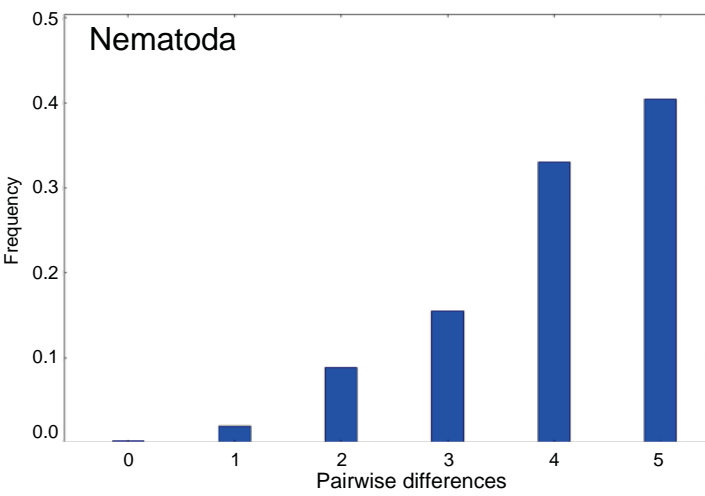
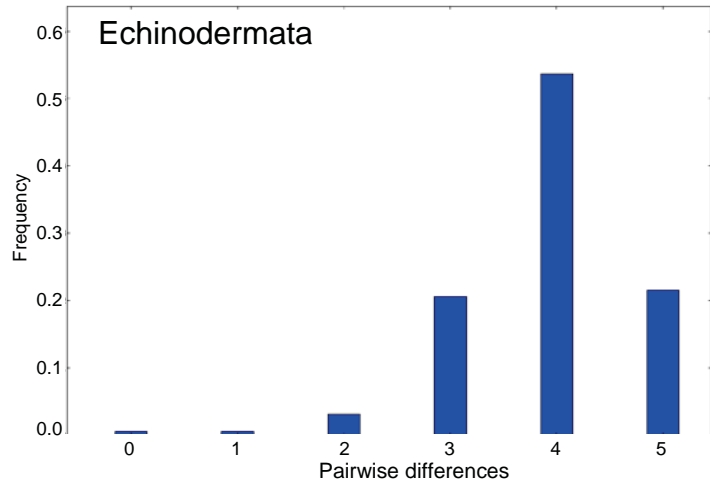
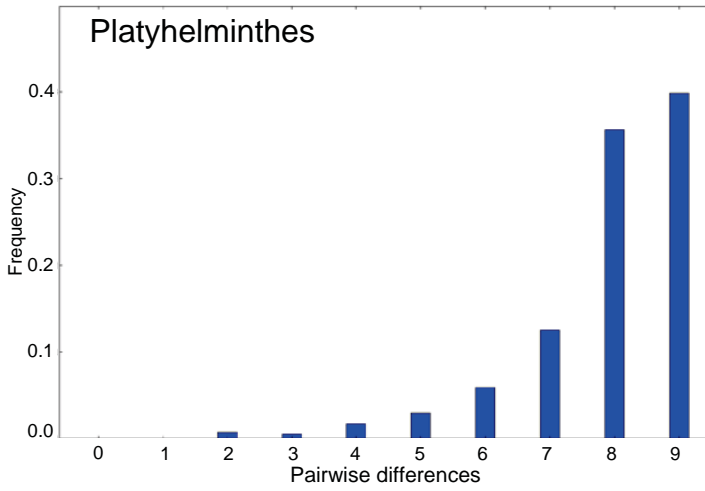
<sup>2</sup> Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

<sup>3</sup> Centro de Matemática da Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

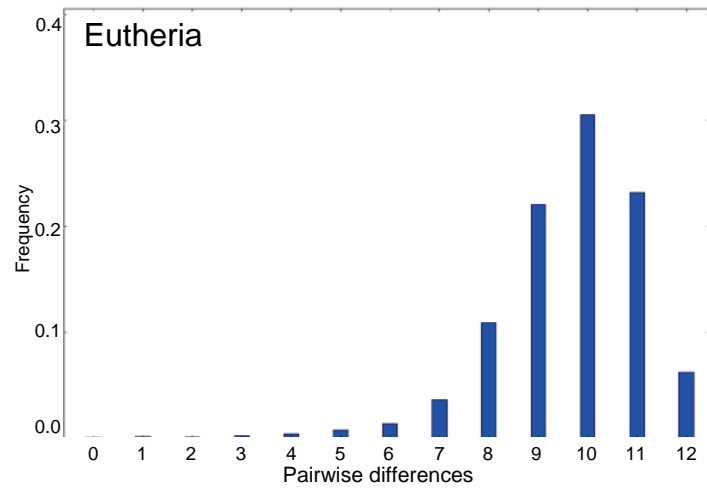
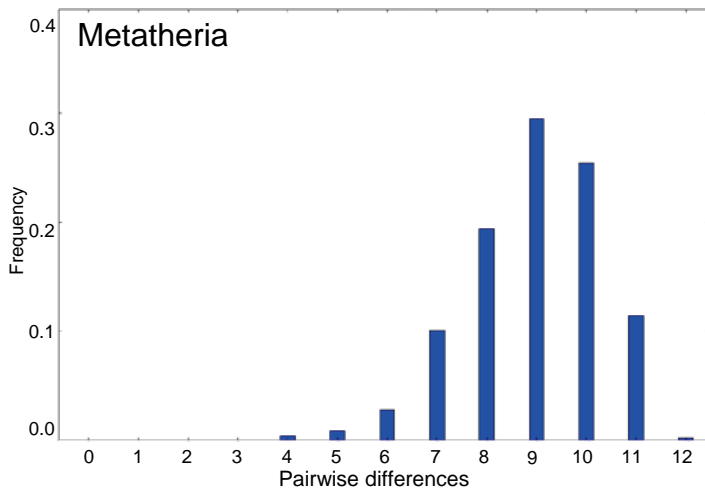
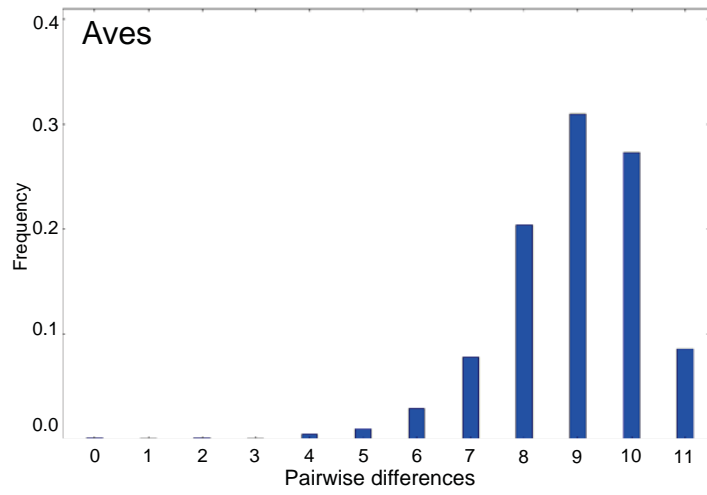
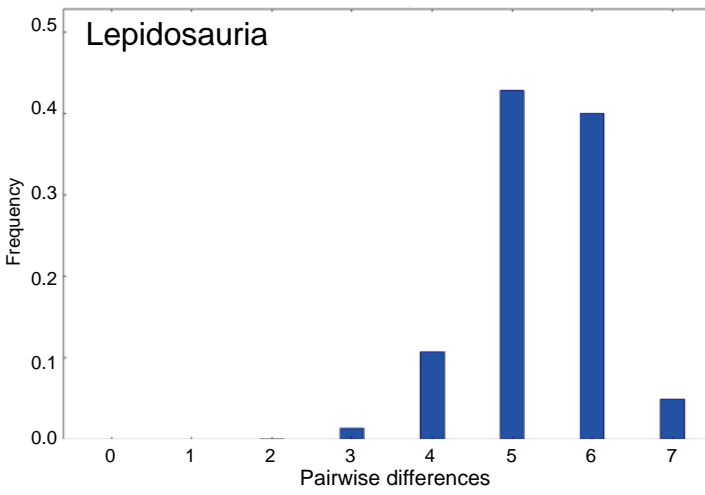
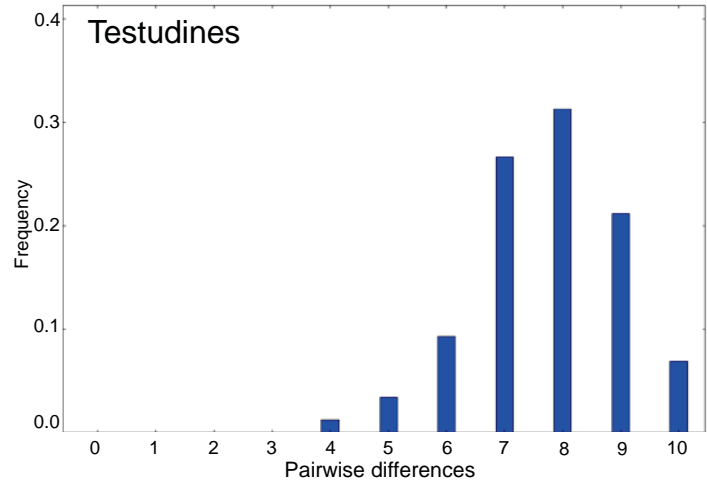
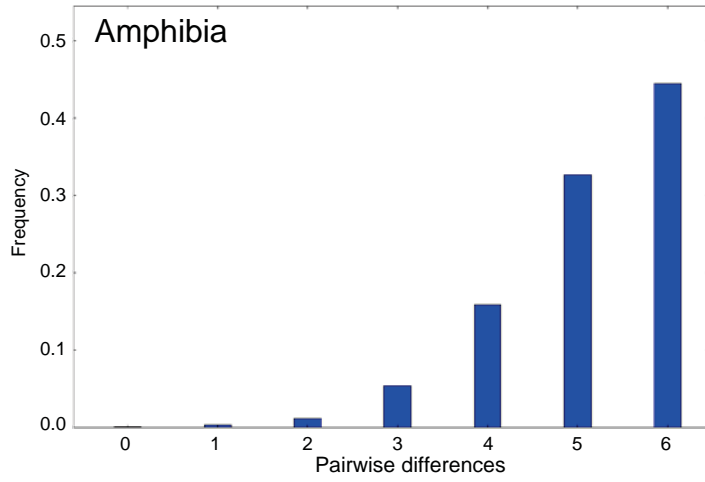
Supplementary Figure S8. Mismatch distribution of SPInDel profiles. The frequency distribution of the number of SPInDel hypervariable regions that differ between all pairs of SPInDel profiles is represented for 18 eukaryotic groups.



# 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](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.

**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.

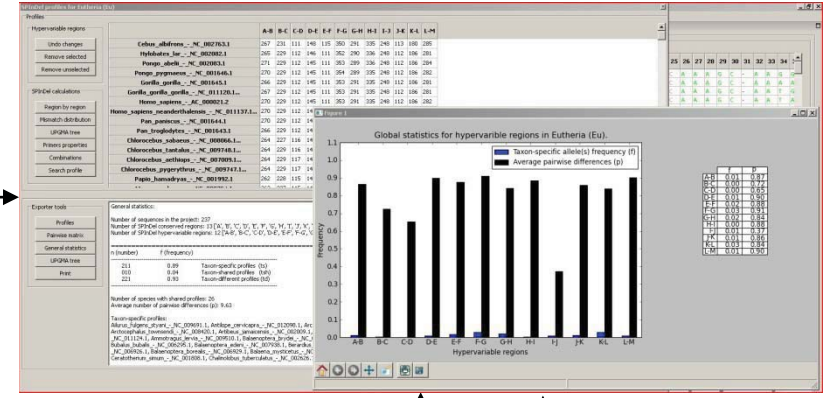
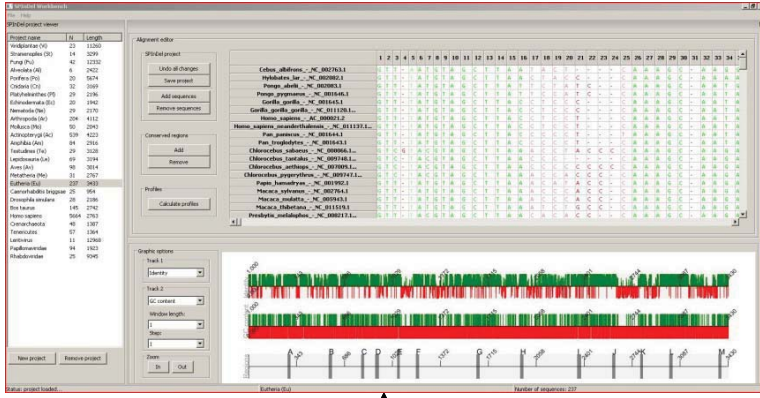
### References

- 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**

**Project Database and Alignment Editor**

**Profile Editor and global calculations**

**b**



Applications

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

SPIInDel profiles

Identification engine

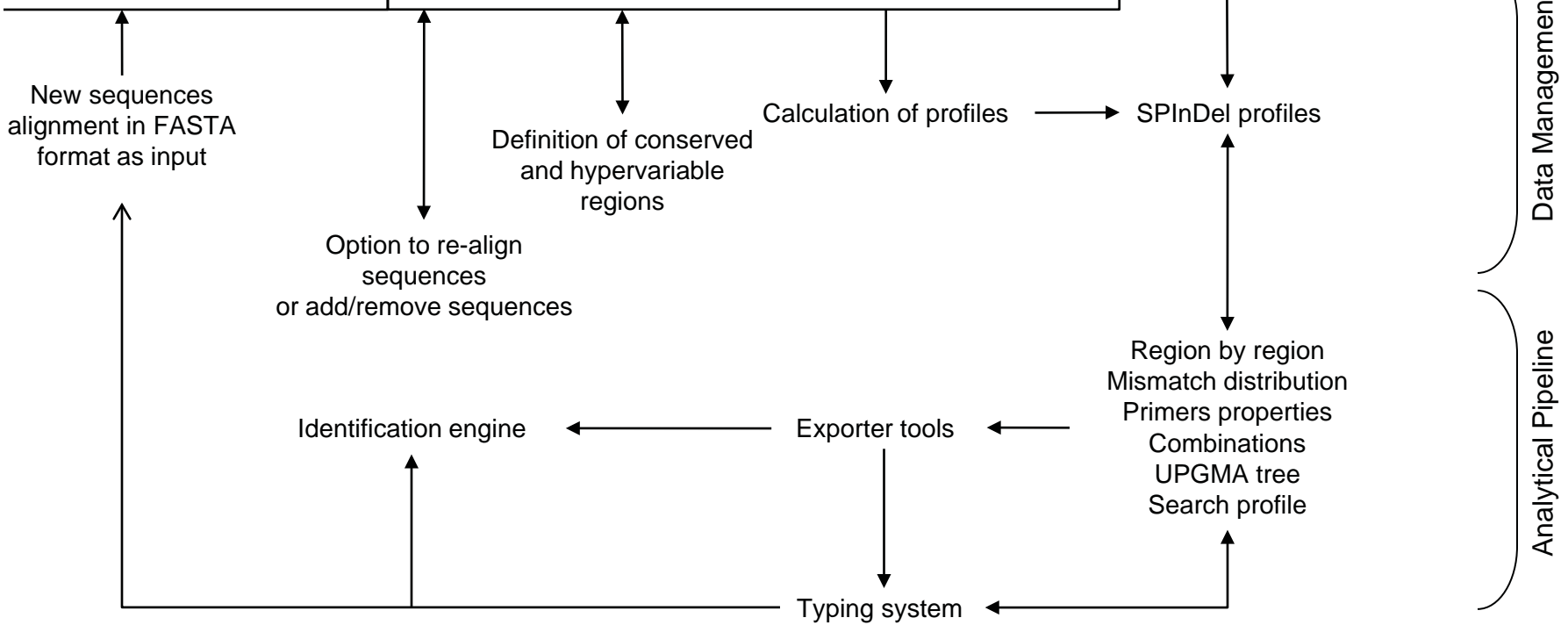
Exporter tools

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

Typing system

Data Management

Analytical Pipeline





# Supplementary Figure S9 (cont.)

## C PROJECTS VIEWER AND ALIGNMENT EDITOR

SPInDel Workbench

File Help

SPInDel project viewer

Project name	N	Length
Viridiplantae (Vi)	23	11260
Stramenopiles (St)	14	3299
Fungi (Fu)	42	12332
Alveolata (Al)	6	2422
Porifera (Po)	20	5674
Cnidaria (Cn)	32	3169
Platyhelminthes (Pl)	29	2196
Echinodermata (Ec)	20	1942
Nematoda (Ne)	29	2170
Arthropoda (Ar)	204	4112
Mollusca (Mo)	50	2043
Actinopterygii (Ac)	539	4223
Amphibia (Am)	84	2916
Testudines (Te)	29	3128
Lepidosauria (Le)	69	3194
Aves (Av)	98	3014
Metatheria (Me)	31	2767
Eutheria (Eu)	237	3433
Caenorhabditis briggsae	25	954
Drosophila simulans	28	2186
Bos taurus	145	2742
Homo sapiens	5664	2763
Crenarchaeota	48	1387
Tenericutes	57	1364
Lentivirus	11	12968
Papillomaviridae	94	1923
Rhabdoviridae	25	9345

Alignment editor

SPInDel project

Undo all changes

Save project

Add sequences

Remove sequences

Conserved regions

Add

Remove

Profiles

Calculate profiles

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34		
Cebus albifrons -_NC_002763.1	G	T	T	-	A	A	T	G	T	A	G	C	T	T	A	A	T	A	C	T	-	-	-	-	C	A	A	A	G	C	-	A	A	G	G	
Hylobates lar -_NC_002082.1	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	A	C	T	A	C	C	-	-	-	C	A	A	A	G	C	-	A	A	A	A	
Pongo abelii -_NC_002083.1	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	T	T	C	T	C	A	T	C	-	-	C	A	A	A	G	C	-	A	A	T	G
Pongo pygmaeus -_NC_001646.1	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	T	T	C	C	A	T	C	-	-	C	A	A	A	G	C	-	A	A	T	A	
Gorilla gorilla -_NC_001645.1	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	C	C	T	C	C	C	-	-	-	C	A	A	A	G	C	-	A	A	T	A	
Gorilla gorilla_gorilla -_NC_011120.1...	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	C	C	T	C	C	C	-	-	-	C	A	A	A	G	C	-	A	A	T	A	
Homo sapiens -_AC_000021.2	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	C	C	T	C	C	T	-	-	-	C	A	A	A	G	C	-	A	A	T	A	
Homo sapiens_neanderthalensis -_NC_011137.1...	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	C	C	T	C	C	T	-	-	-	C	A	A	A	G	C	-	A	A	T	A	
Pan paniscus -_NC_001644.1	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	C	C	C	C	C	T	-	-	-	T	A	A	A	G	C	-	A	A	T	A	
Pan troglodytes -_NC_001643.1	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	C	C	C	C	C	T	-	-	-	C	A	A	A	G	C	-	A	A	T	A	
Chlorocebus sabaeus -_NC_008066.1...	G	T	C	G	T	A	C	G	T	A	G	C	T	T	A	A	A	C	C	T	A	C	C	C	C	A	A	A	G	C	-	A	A	G	A	
Chlorocebus tantalus -_NC_009748.1...	G	T	C	-	T	A	C	G	T	A	G	C	T	T	A	A	C	C	C	-	-	-	-	C	A	A	A	G	C	-	A	A	G	A		
Chlorocebus aethiops -_NC_007009.1...	G	T	C	-	T	A	C	G	T	A	G	C	T	T	A	A	C	C	C	C	C	C	C	C	C	A	A	A	G	C	-	A	A	G	A	
Chlorocebus pygerythrus -_NC_009747.1...	G	T	C	-	T	A	C	G	T	A	G	C	T	T	A	A	A	C	C	A	C	C	C	-	C	A	A	A	G	C	-	A	A	G	A	
Papio hamadryas -_NC_001992.1	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	A	A	C	A	T	A	C	C	-	C	A	A	A	G	C	-	A	A	G	A	
Macaca sylvanus -_NC_002764.1	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	A	A	C	C	C	A	C	C	-	C	A	A	A	G	C	-	A	A	G	A	
Macaca mulatta -_NC_005943.1	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	A	A	C	C	C	A	C	C	-	C	A	A	A	G	C	-	A	A	G	A	
Macaca thibetana -_NC_011519.1	G	T	T	-	T	A	T	G	T	A	G	C	T	T	A	A	A	T	C	T	G	C	C	-	C	A	A	A	G	C	-	A	A	G	A	
Presbytis melalophos -_NC_008217.1...	G	T	T	-	T	A	C	G	T	A	G	C	T	T	A	A	C	A	C	A	C	C	-	-	C	A	A	A	G	C	-	A	A	G	A	

Graphic options

Track 1: Identity

Track 2: GC content

Window length: 25

Step: 1

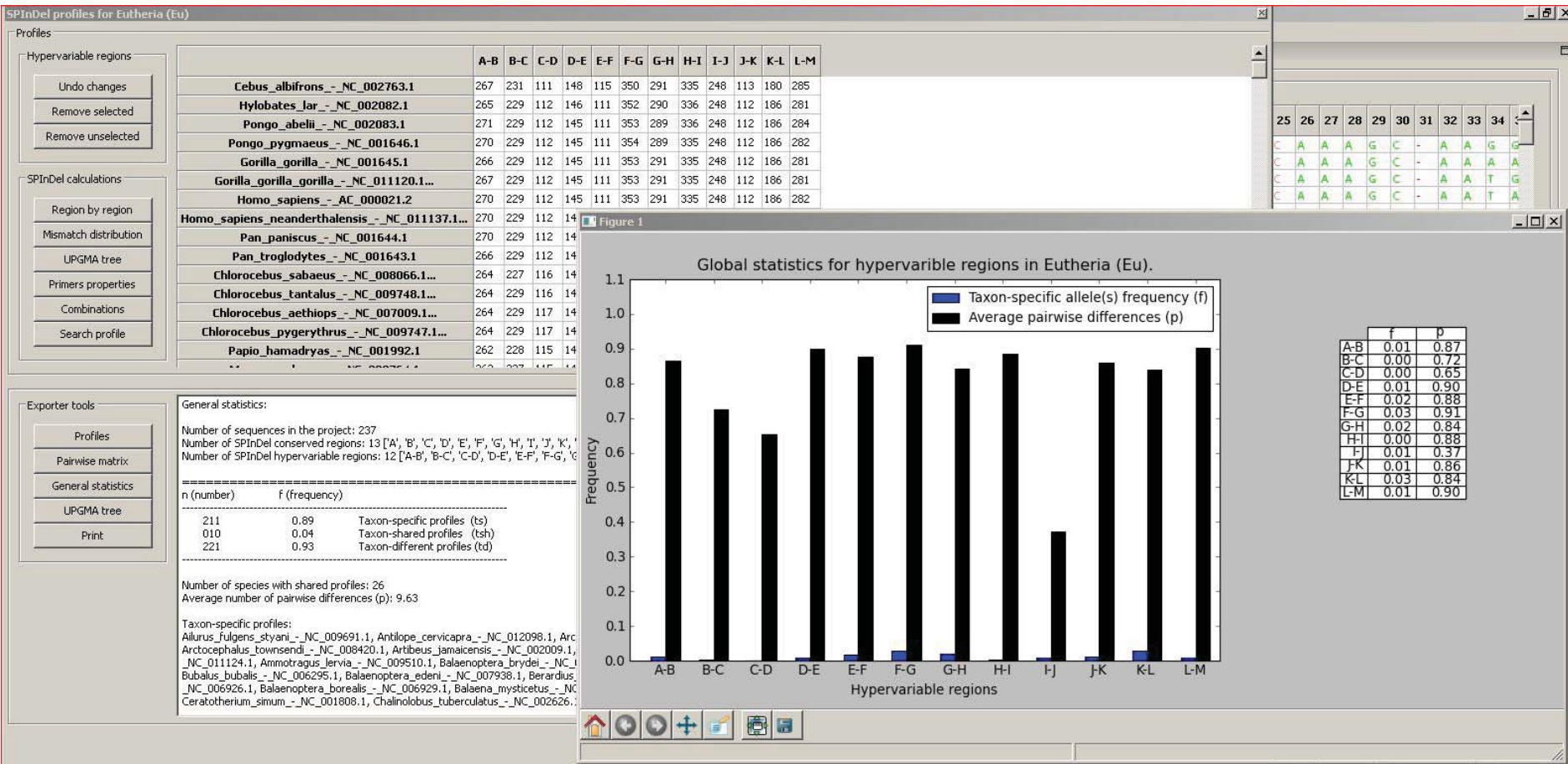
Zoom: In Out

Regions: A 343, B 886, C, D, E 1029, F, G 1372, H 1715, I, J, K 2088, L, M 2401, N 2744, O 3087, P 3430

Status: project loaded... Eutheria (Eu) Number of sequences: 237

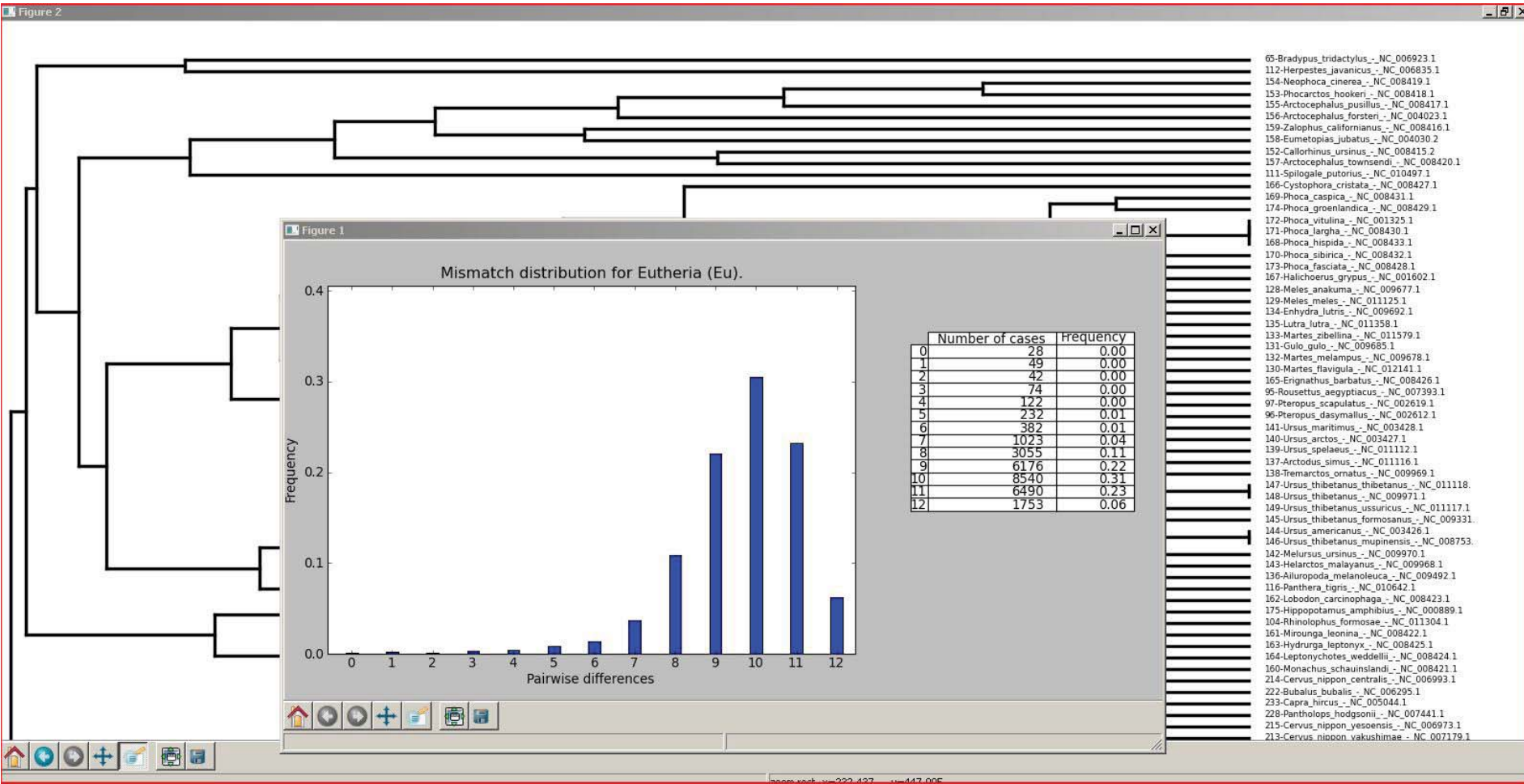
# Supplementary Figure S9 (cont.)

## PROFILES FRAME AND GLOBAL STATISTICS GRAPH



# Supplementary Figure S9 (cont.)

## MISMATCH DISTRIBUTION GRAPH AND UPGMA TREE



# Supplementary Figure S9 (cont.)

## PROFILES EVALUATION FRAME AND PRIMERS PROPERTIES

SPInDel profiles evaluation for Eutheria (Eu)

Number of ...	f(ts)	f(dp)	Profiles
1	0.03	0.08	[F-G]
2	0.24	0.40	[E-F, F-G]
3	0.61	0.74	[F-G, H-I, L-M]
4	0.75	0.84	[A-B, F-G, H-I, L-M]
5	0.80	0.87	[A-B, F-G, H-I, J-K, L-M]
6	0.84	0.90	[A-B, E-F, F-G, H-I, J-K, L-M]
7	0.86	0.91	[A-B, E-F, F-G, H-I, I-J, J-K, L-M]
8	0.87	0.92	[A-B, D-E, E-F, F-G, G-H, H-I, J-K, L-M]
9	0.88	0.93	[A-B, B-C, D-E, E-F, F-G, G-H, H-I, J-K, L-M]
10	0.89	0.93	[A-B, B-C, D-E, E-F, F-G, G-H, H-I, J-K, K-L, L-M]
11	0.89	0.93	[A-B, B-C, D-E, E-F, F-G, G-H, H-I, I-J, J-K, K-L, L-M]
12	0.89	0.93	[A-B, B-C, C-D, D-E, E-F, F-G, G-H, H-I, I-J, J-K, L-M]

Profiles	f(ts)	f(dp)
[A-B, B-C, D-E, E-F, F-G, H-I, I-J, J-K, L-M]	0.88	0.92
[A-B, D-E, E-F, F-G, G-H, H-I, J-K, K-L, L-M]	0.88	0.92
[A-B, B-C, E-F, F-G, G-H, H-I, J-K, K-L, L-M]	0.88	0.92
[A-B, B-C, E-F, F-G, G-H, H-I, I-J, J-K, L-M]	0.88	0.92
[A-B, B-C, D-E, E-F, F-G, G-H, H-I, J-K, L-M]	0.88	0.92
[A-B, B-C, D-E, E-F, F-G, G-H, H-I, J-K, L-M]	0.88	0.93
[A-B, D-E, E-F, F-G, H-I, I-J, J-K, K-L, L-M]	0.87	0.92
[A-B, B-C, C-D, E-F, F-G, H-I, I-J, J-K, L-M]	0.87	0.92
[A-B, B-C, E-F, F-G, G-H, H-I, I-J, J-K, L-M]	0.87	0.92
[A-B, B-C, C-D, D-E, F-G, G-H, H-I, K-L, L-M]	0.87	0.92
[A-B, E-F, F-G, G-H, H-I, I-J, J-K, K-L, L-M]	0.87	0.92
[A-B, B-C, C-D, F-G, G-H, H-I, J-K, K-L, L-M]	0.87	0.92
[A-B, B-C, E-F, F-G, H-I, I-J, J-K, K-L, L-M]	0.87	0.92
[A-B, B-C, D-E, E-F, F-G, G-H, H-I, I-J, J-K, L-M]	0.87	0.92
[A-B, C-D, D-E, E-F, F-G, G-H, H-I, J-K, L-M]	0.87	0.92
[A-B, C-D, D-E, E-F, F-G, G-H, H-I, J-K, L-M]	0.87	0.92
[A-B, B-C, C-D, F-G, G-H, H-I, I-J, J-K, L-M]	0.87	0.92

Conserved regions for Eutheria (Eu)

Conserved ...	Primer	Start	End	5'-3' seque...	Length	Tm	GC c...
A	AF1	280	298	CCCCAAGG...	18	60.90	66.6%
A	AF1	280	298	CCCCAAGG...	18	58.90	61.1%
A	AF1	280	298	CCCCACGG...	18	60.90	66.6%
A	AF1	280	298	CCCCACGG...	19	60.90	63.1%
A	AF1	280	298	CCCCACGG...	18	62.90	72.2%
A	AF1	280	298	CCCCACGG...	18	64.90	77.7%
A	AF1	280	298	CCCCACGG...	18	62.90	72.2%
A	AF1	280	298	CCCCACGG...	18	62.90	72.2%
A	AF1	280	298	CCCCACGG...	19	60.90	63.1%
A	AF1	280	298	CCCCACGG...	18	60.90	66.6%
A	AF1	280	298	CCCCACGG...	18	60.90	66.6%
A	AF1	280	298	CCCCACGG...	18	62.90	72.2%
A	AF1	280	298	CCCCACGG...	18	64.90	77.7%
A	AF1	280	298	CCCCACGG...	18	60.90	66.6%
A	AF1	280	298	CCCCACGG...	18	60.90	66.6%
A	AF1	280	298	CCCTACGG...	18	58.90	61.1%

SPInDel profiles

Standard

Graph

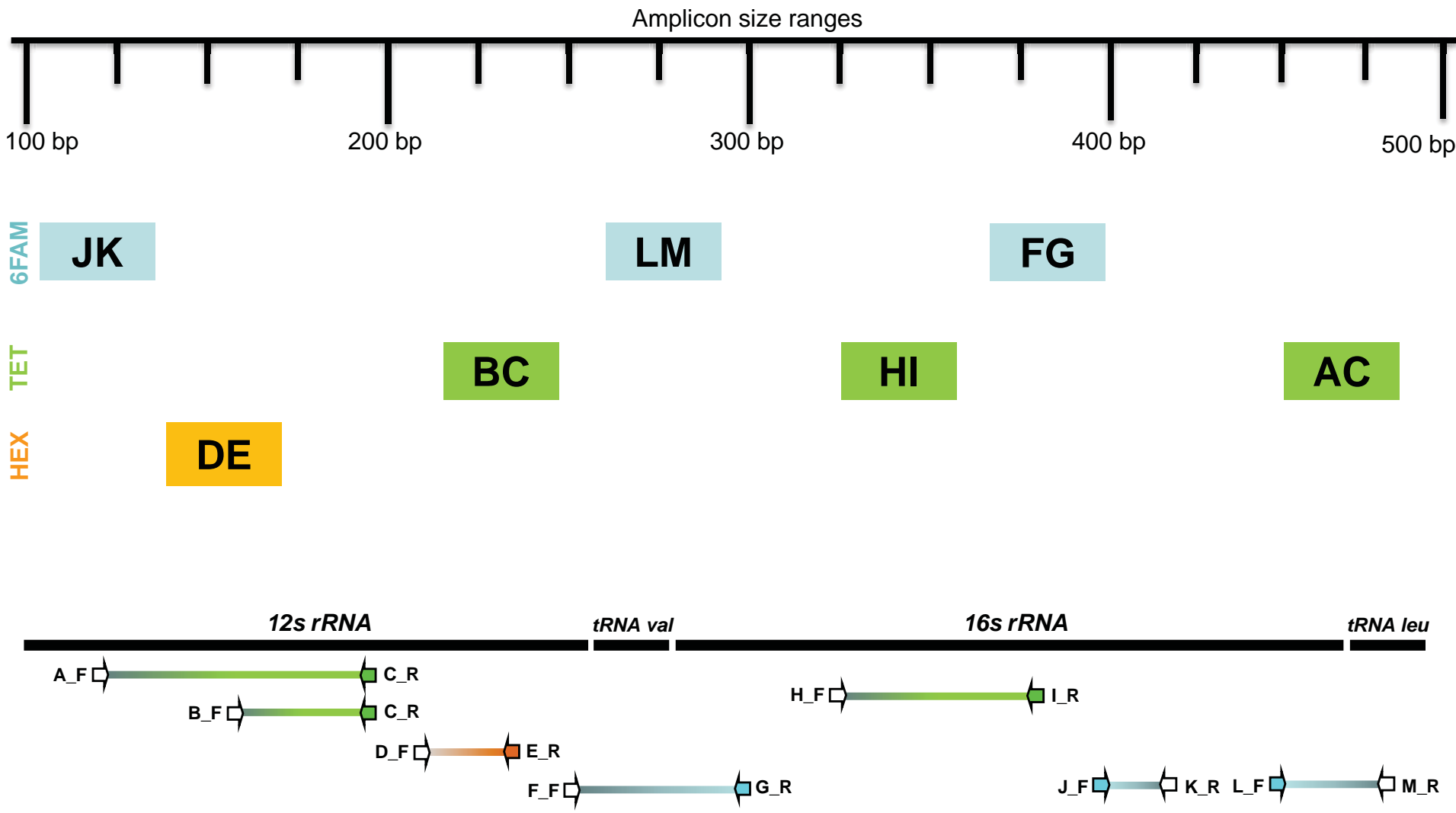
Exporter tools

PCR primers

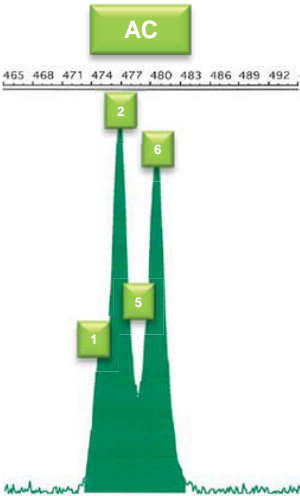
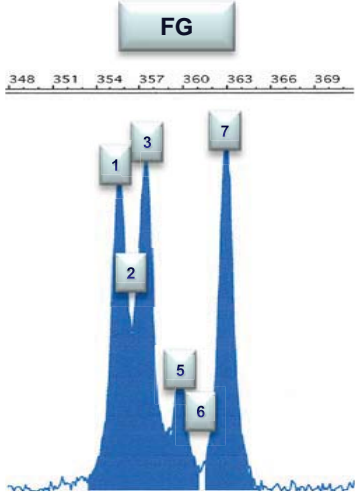
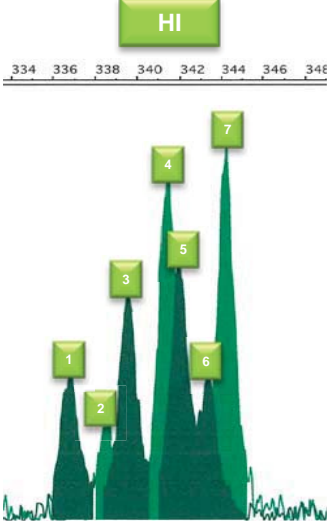
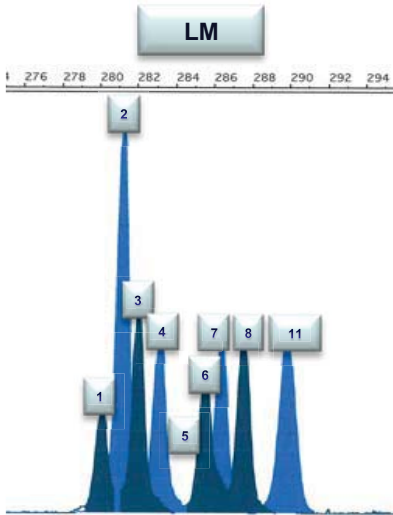
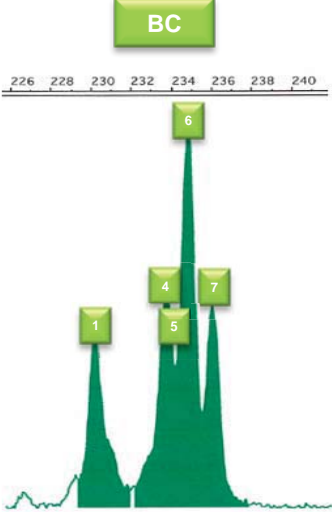
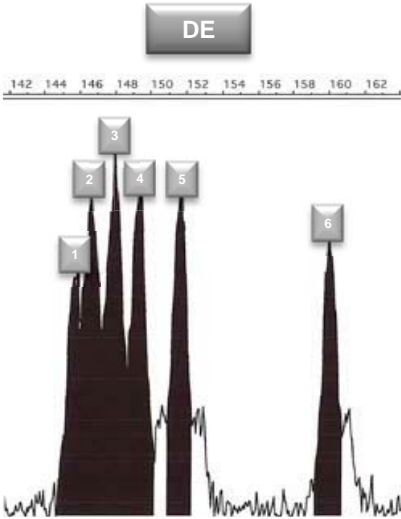
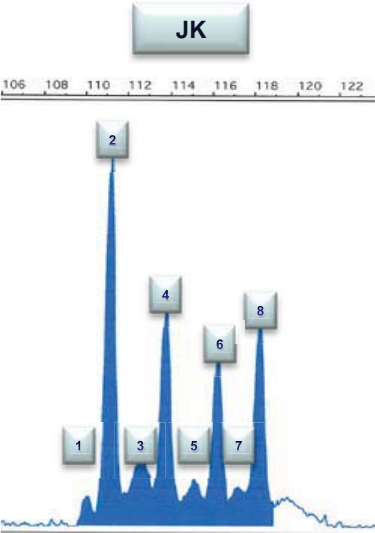
Tables

Export

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.

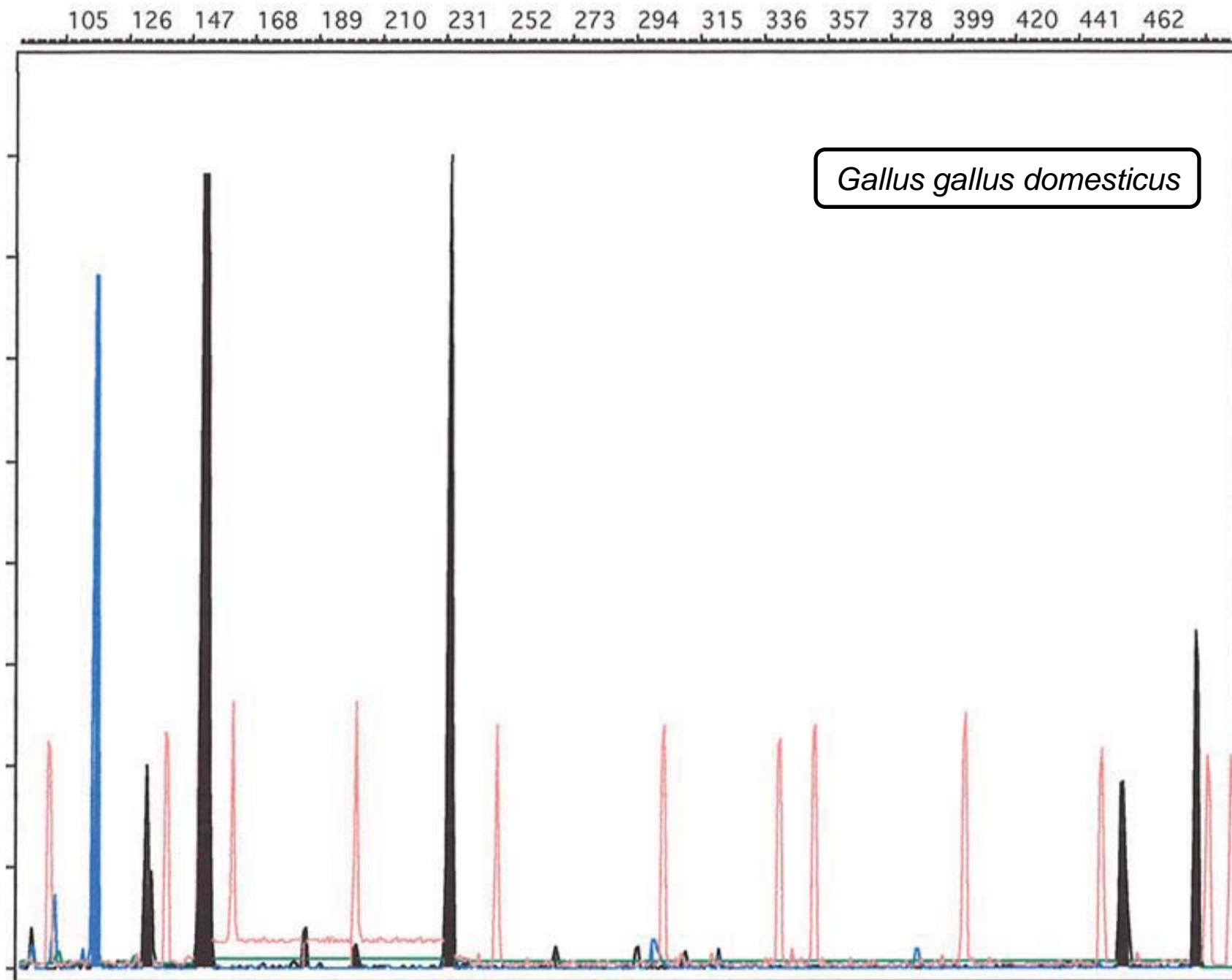


Supplementary Figure S11. Allelic ladders used in the SPInDel profiling kit. The number of each allele is indicated above peaks.



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

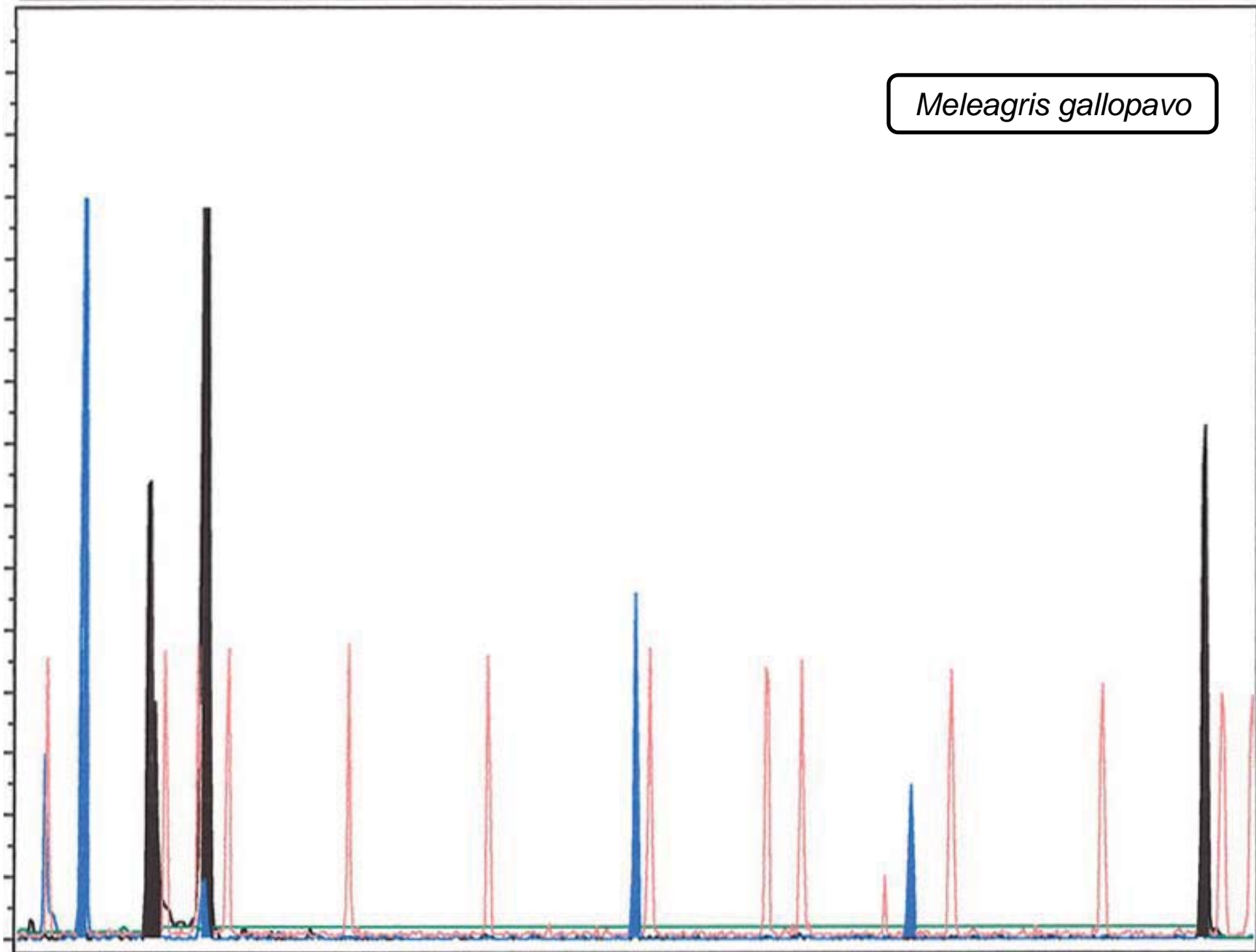




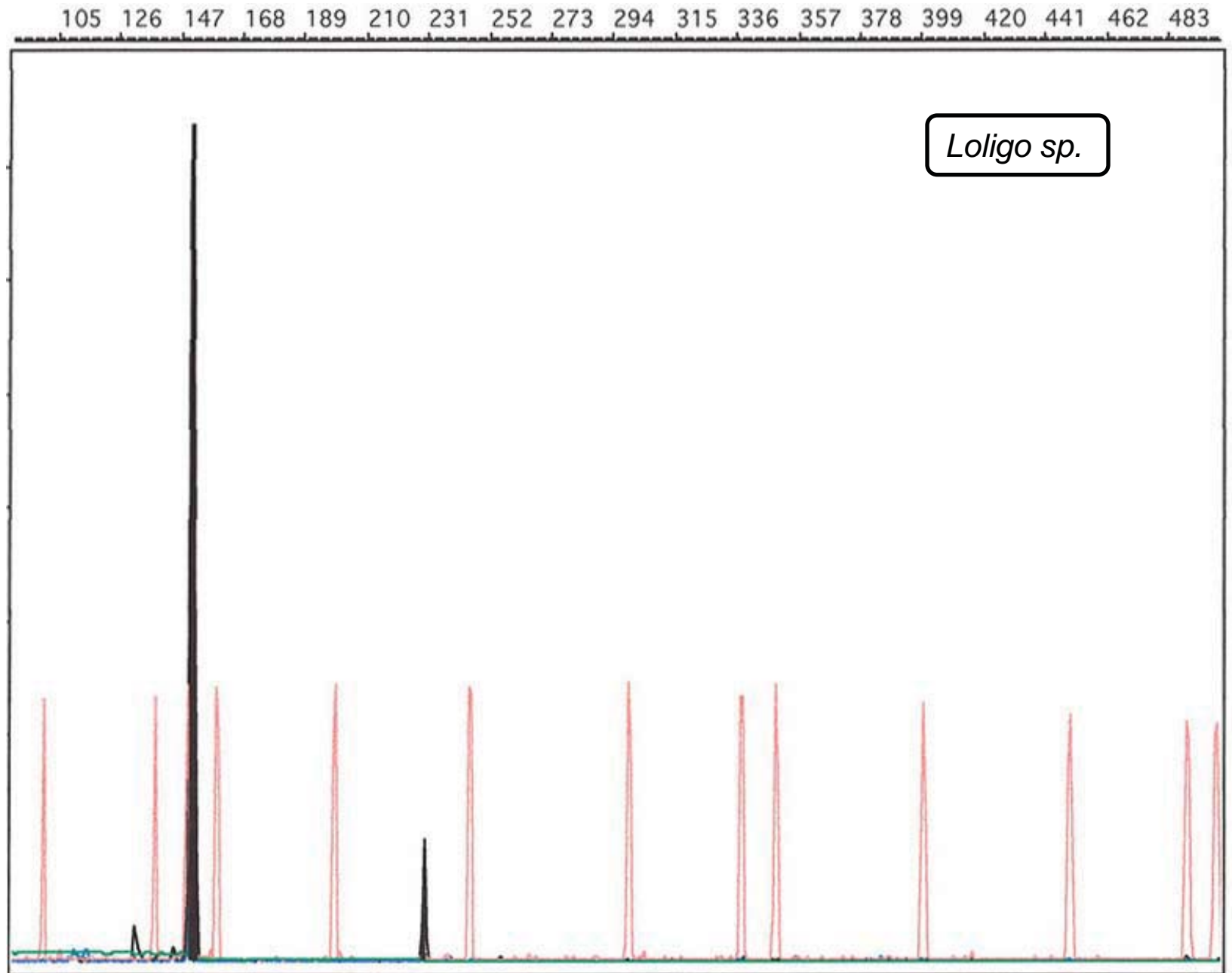
Supplementary Figure S12 (cont.)

105 126 147 168 189 210 231 252 273 294 315 336 357 378 399 420 441 462 483

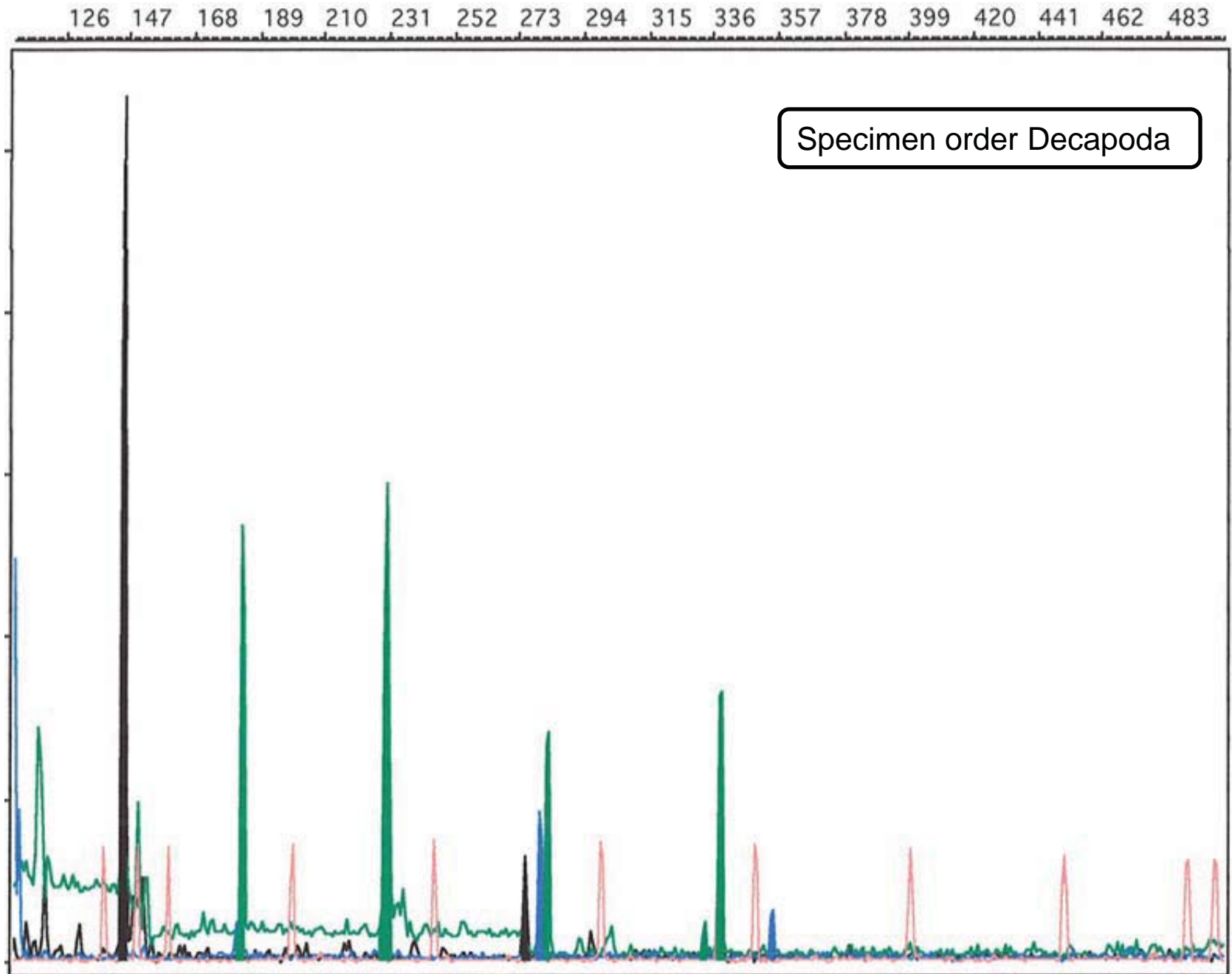
*Meleagris gallopavo*



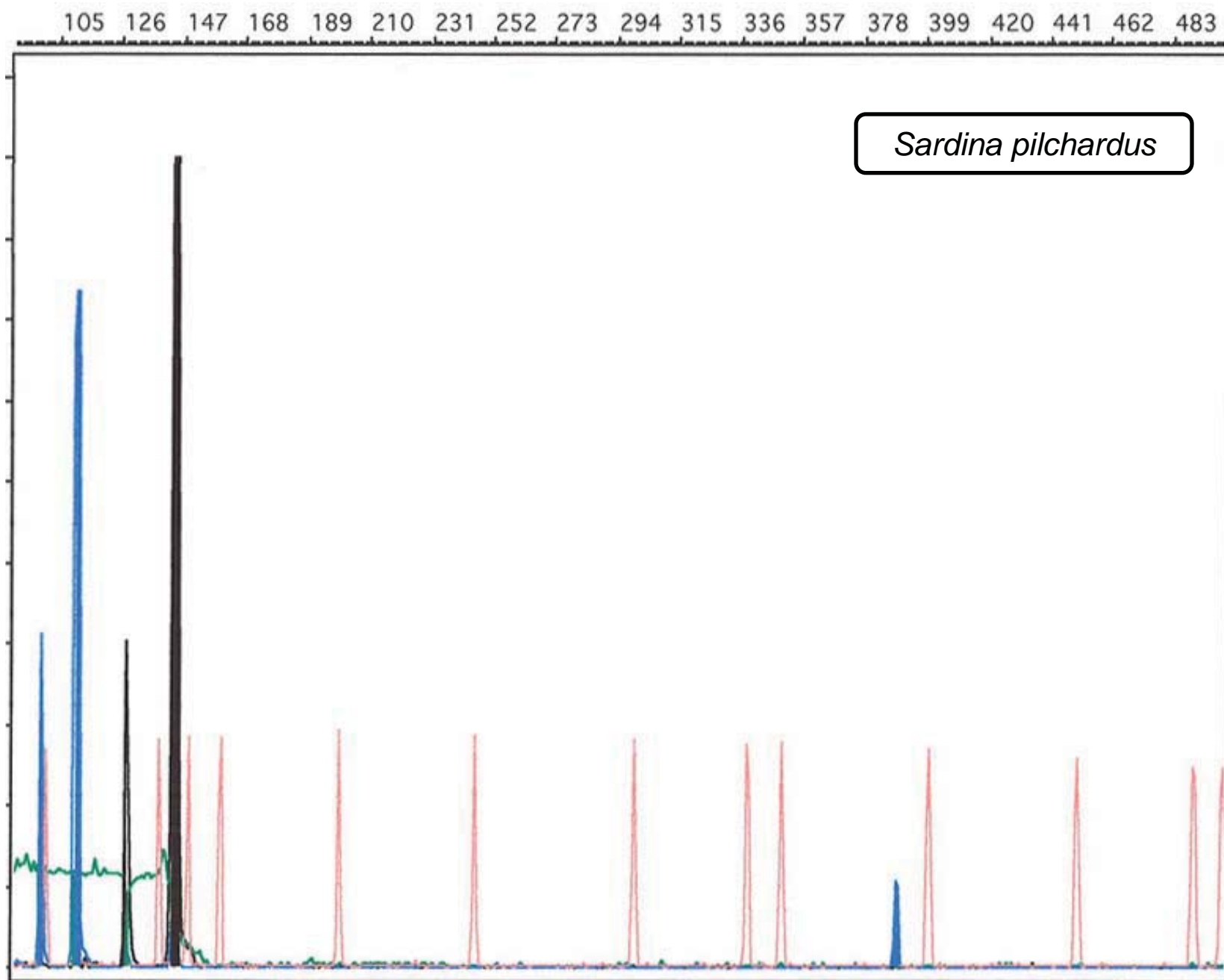
Supplementary Figure S12 (cont.)



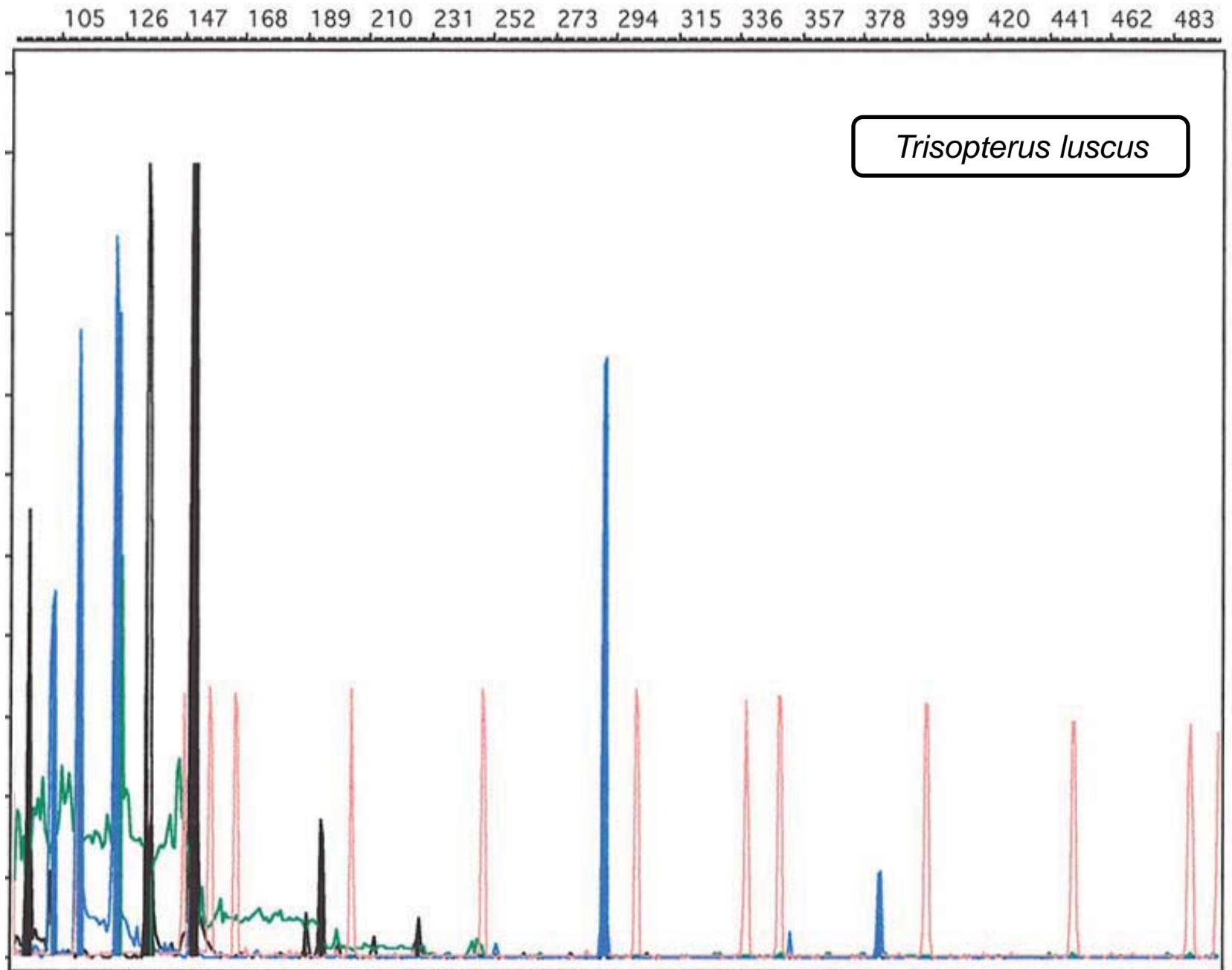
Supplementary Figure S12 (cont.)



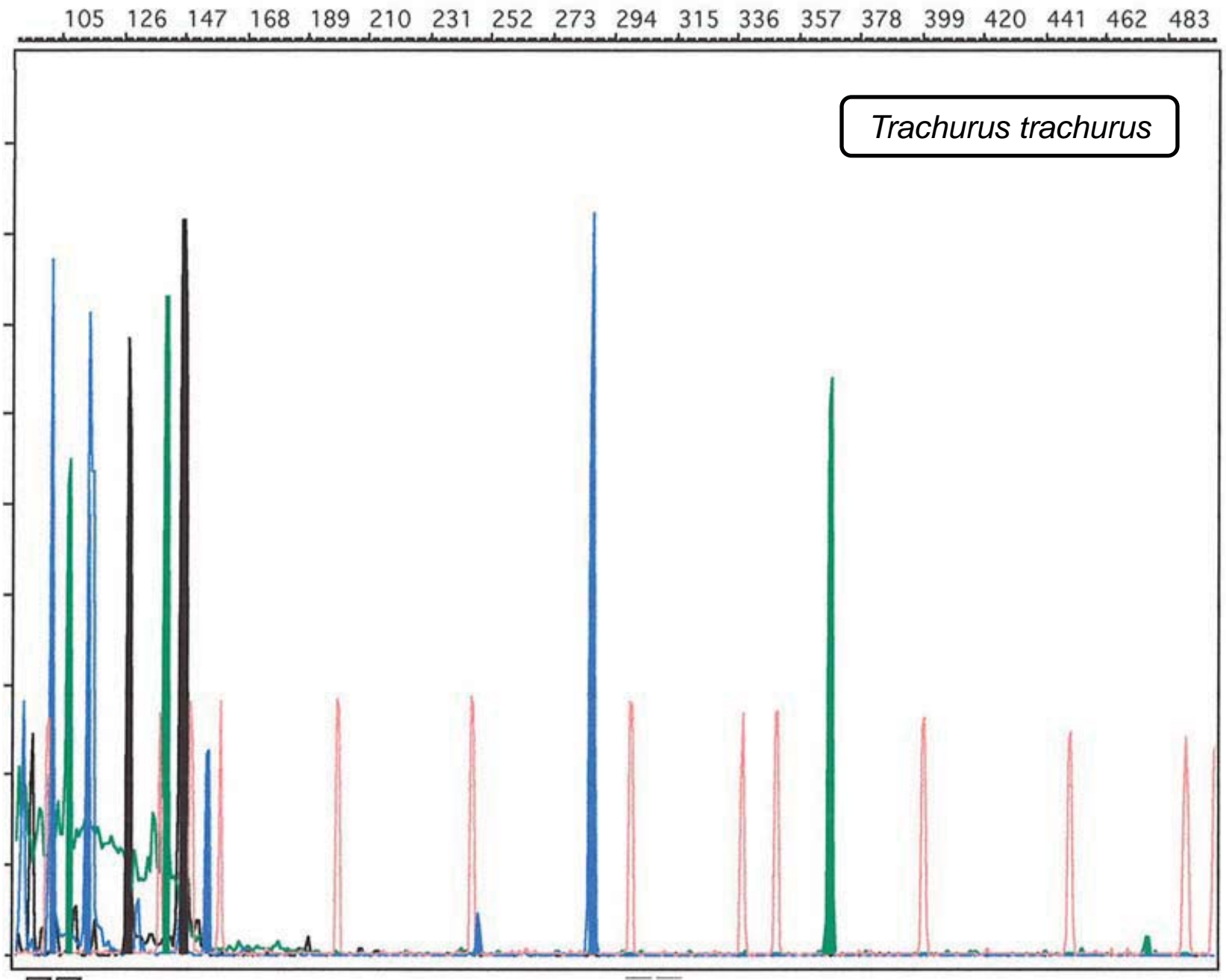
Supplementary Figure S12 (cont.)



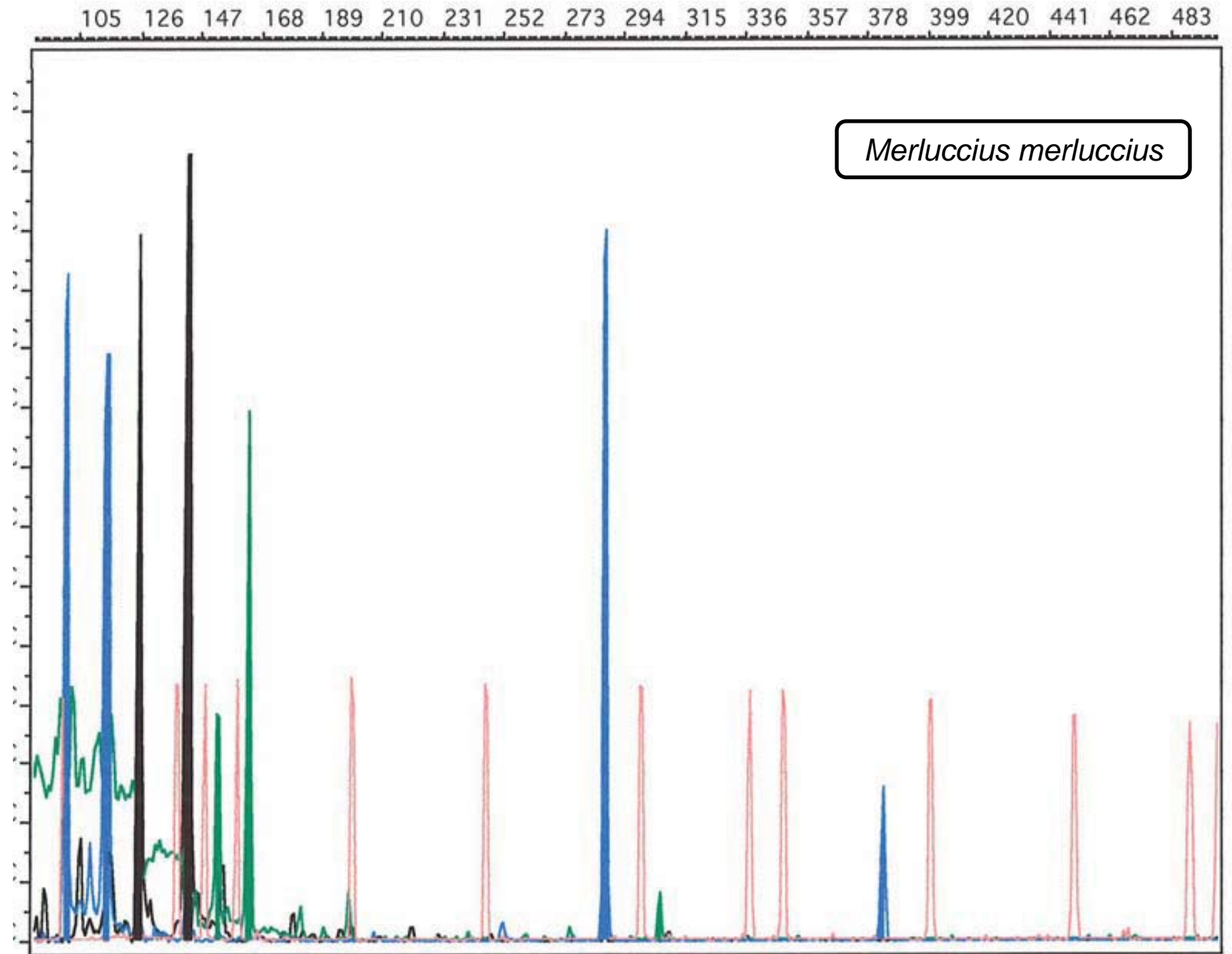
Supplementary Figure S12 (cont.)



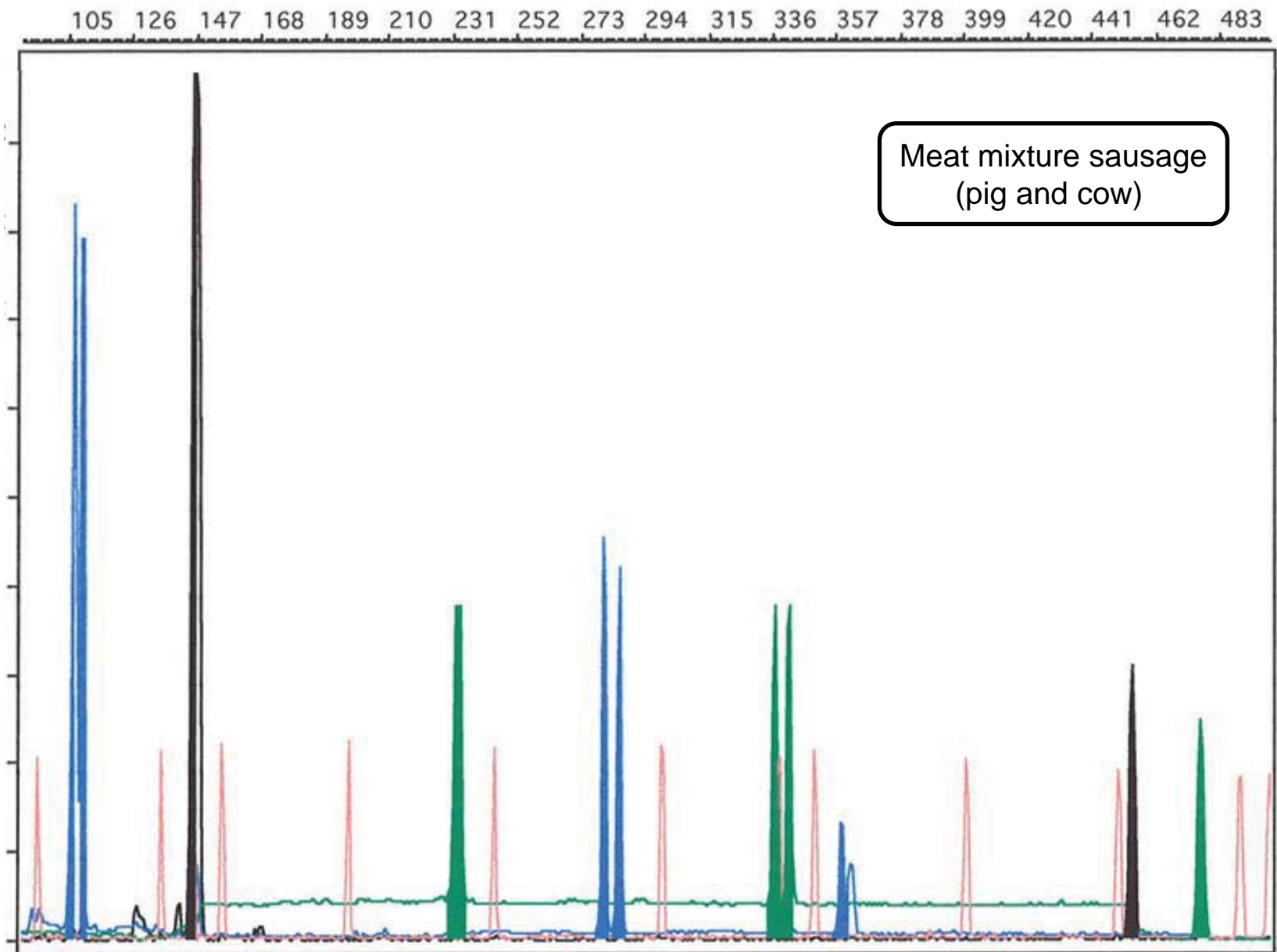
Supplementary Figure S12 (cont.)



Supplementary Figure S12 (cont.)

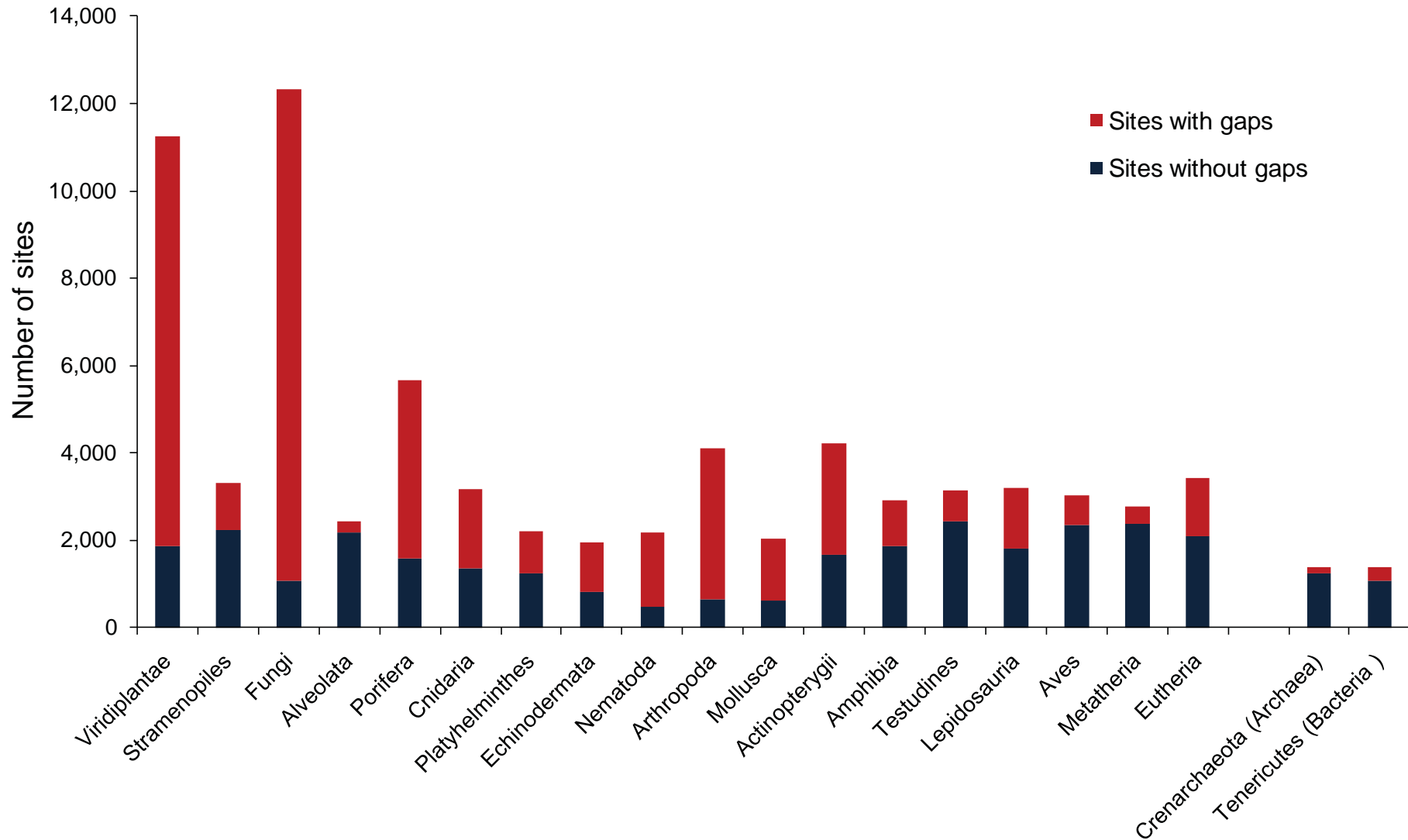


Supplementary Figure S13. Detection of mixtures with the SPInDel assay. The image represents an electropherogram obtained by capillary electrophoresis with multicolor fluorescence detection in a food product with mixture of porcine and bovine biological material. The profile is displayed in a four-dye 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.

Viridiplantae

Beta vulgaris subsp. vulgaris - NC\_002511.2 BevupMr001  
 Beta vulgaris subsp. vulgaris - NC\_002511.2 BevupMr002  
 Beta vulgaris subsp. vulgaris - NC\_002511.2 BevupMr005

Ostreococcus tauri - NC\_008290.1 OstapMr03  
 Ostreococcus tauri - NC\_008290.1 OstapMr04

Oryza sativa Indica Group - NC\_007886.1 OrsaiPr03  
 Oryza sativa Indica Group - NC\_007886.1 OrsaiPr06

Stramenopiles

Saprolegnia ferax - NC\_005984.1\_rnl1  
 Saprolegnia ferax - NC\_005984.1\_rnl2

Alveolata

Tetrahymena malaccensis - NC\_008337.1\_rnlb1  
 Tetrahymena malaccensis - NC\_008337.1\_rnlb2

Tetrahymena paravorax - NC\_008338.1\_rnlb1  
 Tetrahymena paravorax - NC\_008338.1\_rnlb2

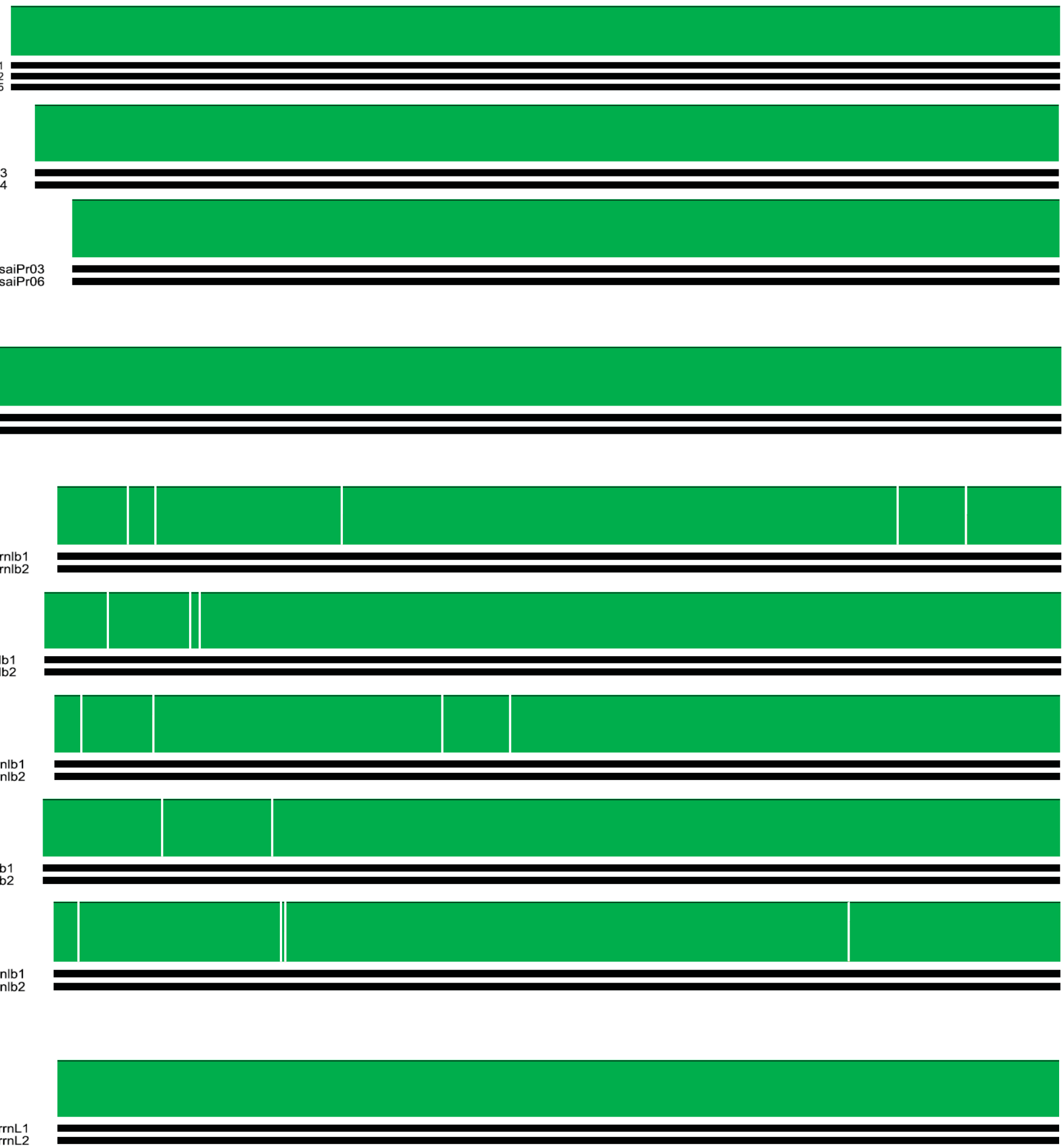
Tetrahymena pigmentosa - NC\_008339.1\_rnlb1  
 Tetrahymena pigmentosa - NC\_008339.1\_rnlb2

Tetrahymena pyriformis - NC\_000862.1\_rnlb1  
 Tetrahymena pyriformis - NC\_000862.1\_rnlb2

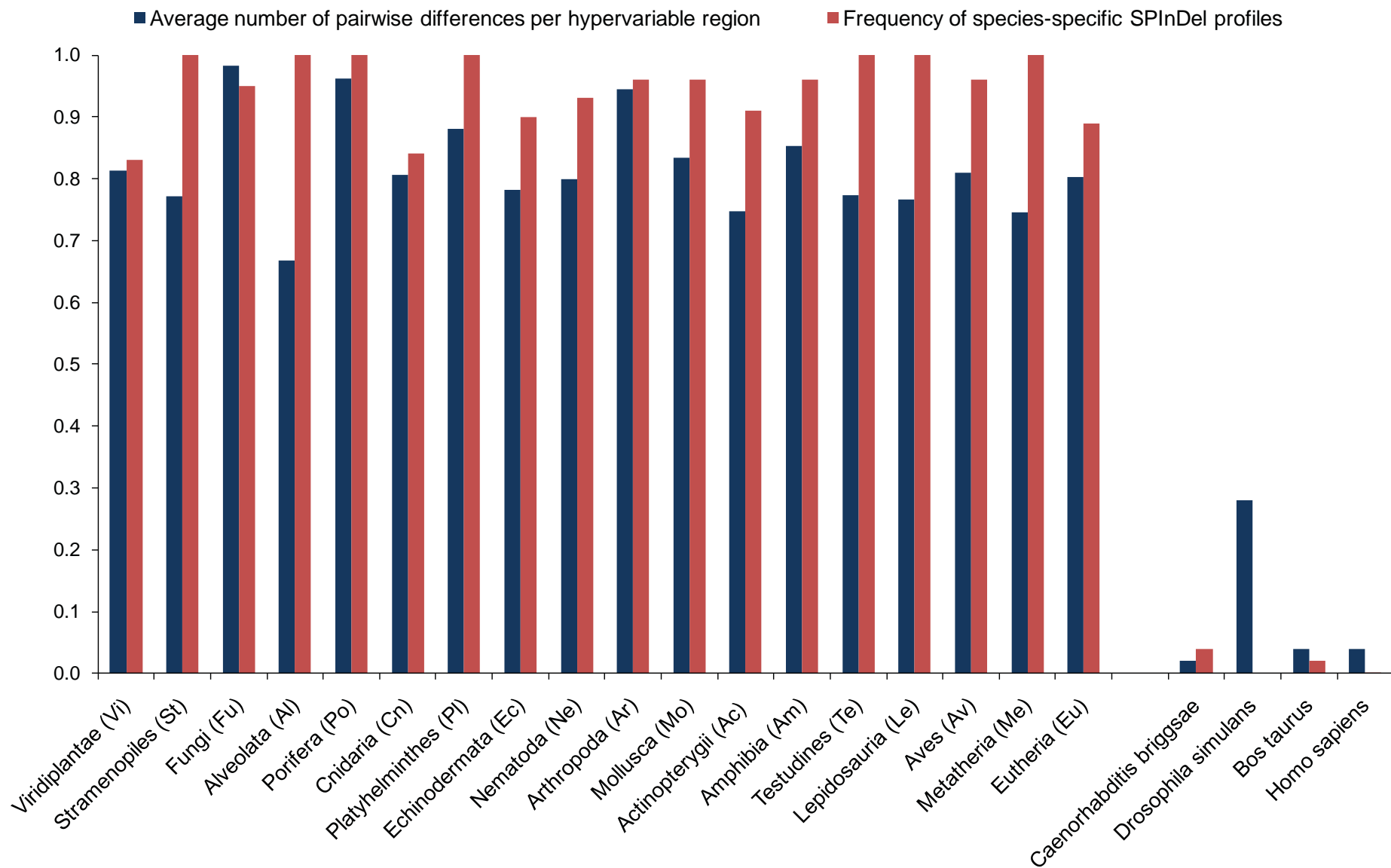
Tetrahymena thermophila - NC\_003029.1\_rnlb1  
 Tetrahymena thermophila - NC\_003029.1\_rnlb2

Nematoda

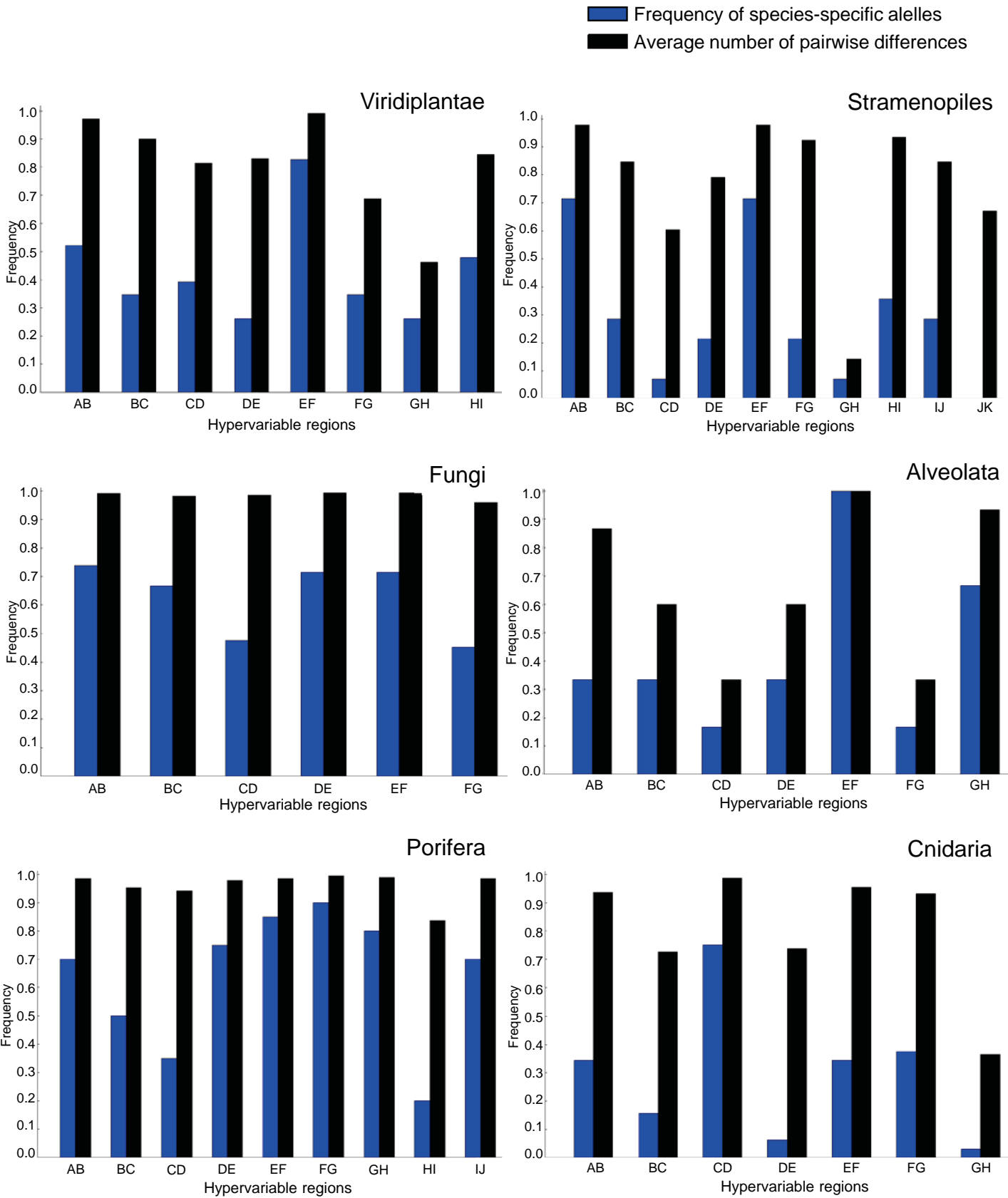
Strelkovimermis spiculatus - NC\_008047.1\_rrnL1  
 Strelkovimermis spiculatus - NC\_008047.1\_rrnL2



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.

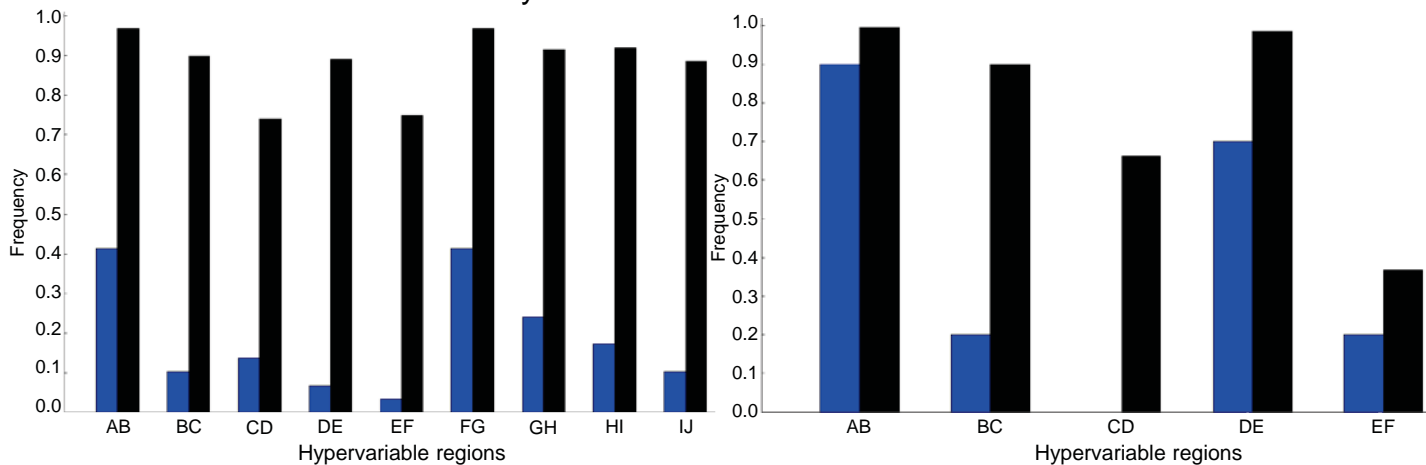


Supplementary Figure S17 (cont.)

■ Frequency of species-specific alleles  
■ Average number of pairwise differences

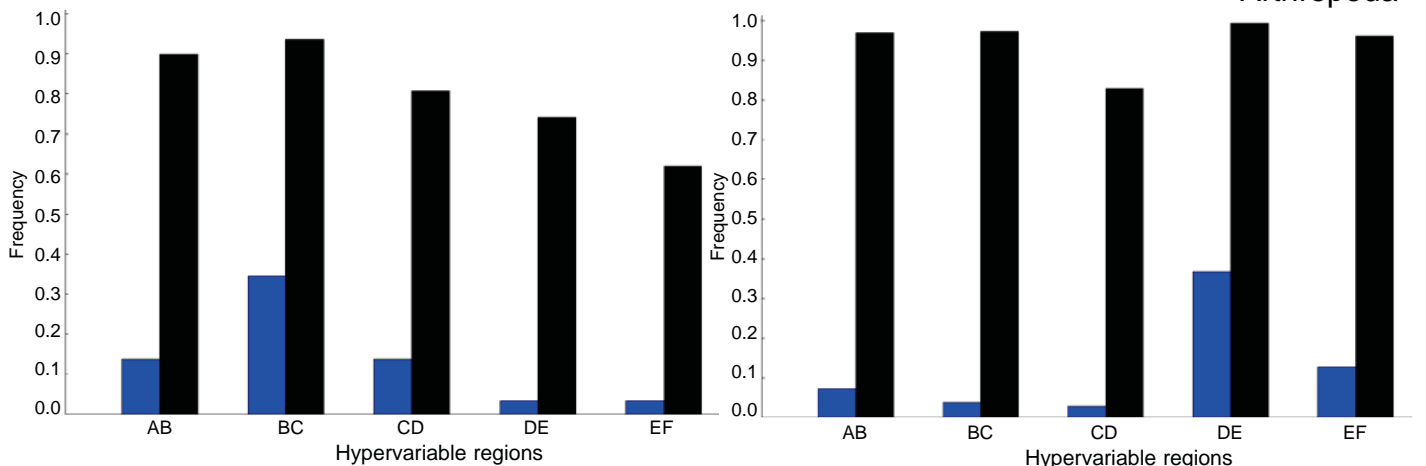
### Platyhelminthes

### Echinodermata



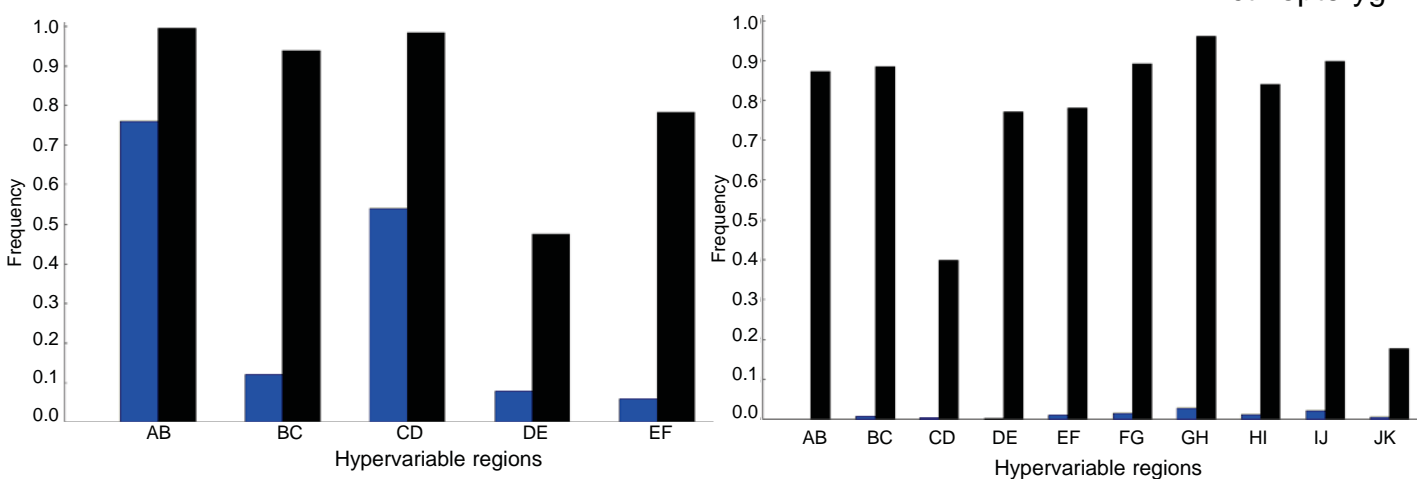
### Nematoda

### Arthropoda



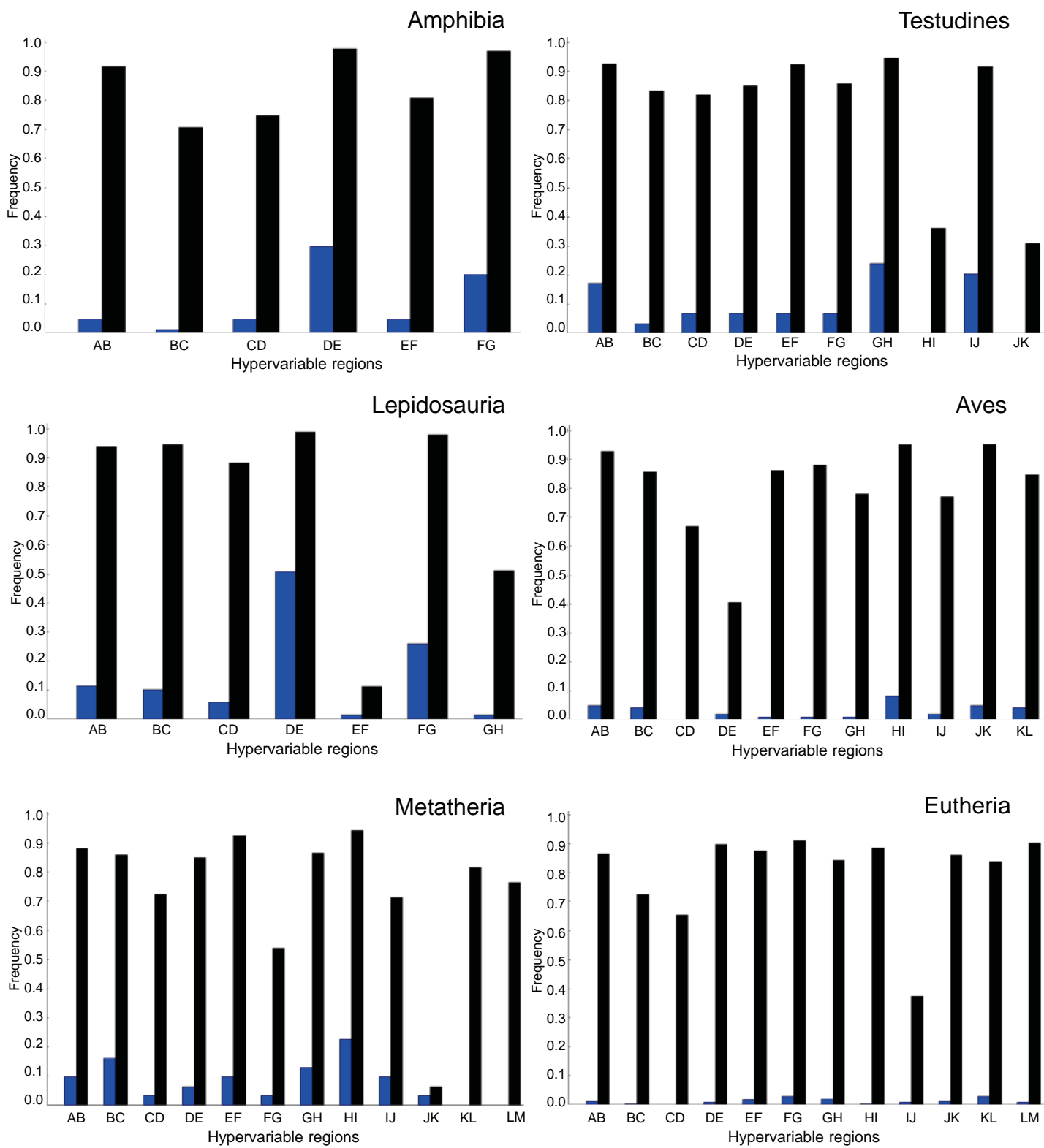
### Mollusca

### Actinopterygii

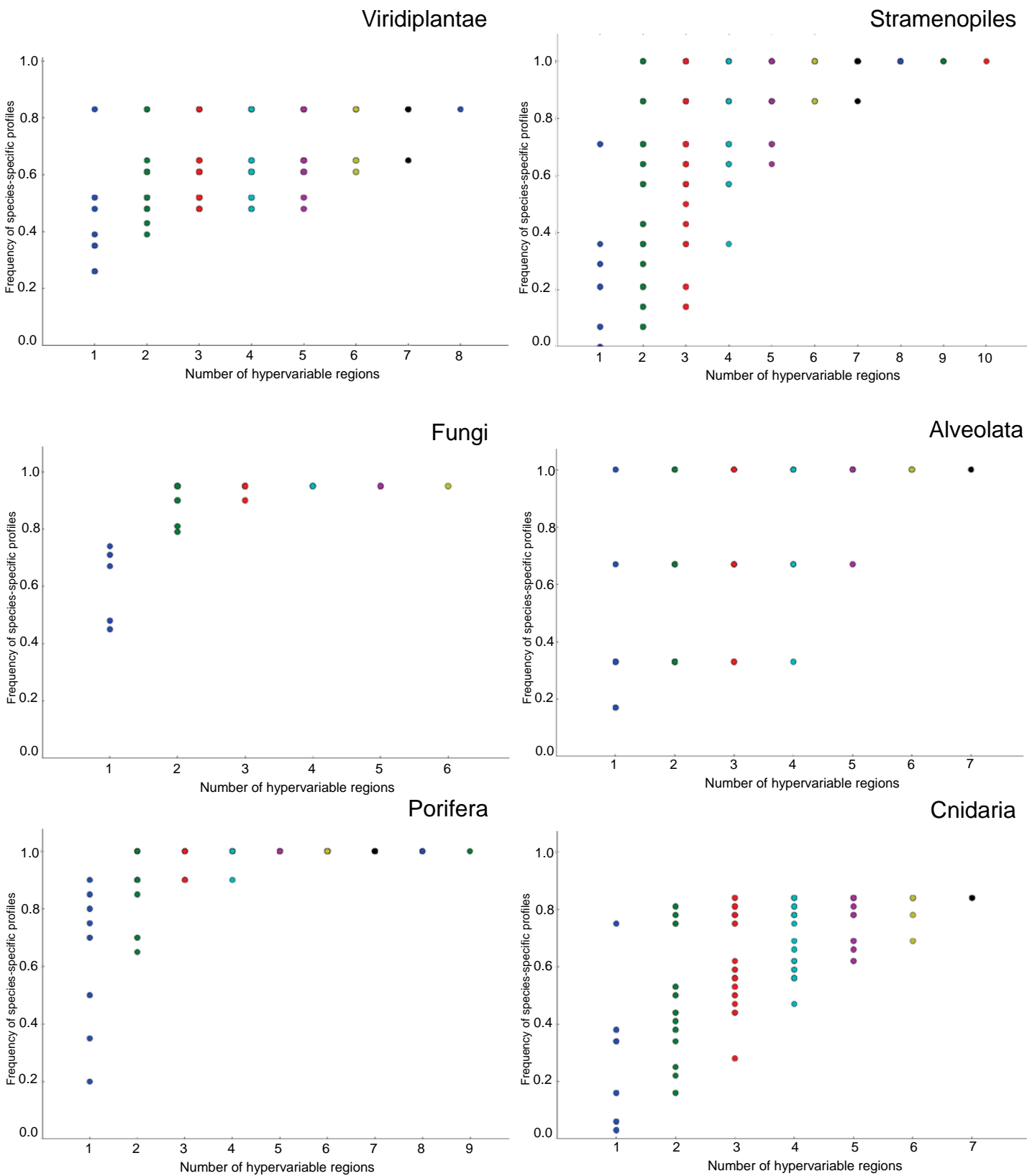


Supplementary Figure S17 (cont.)

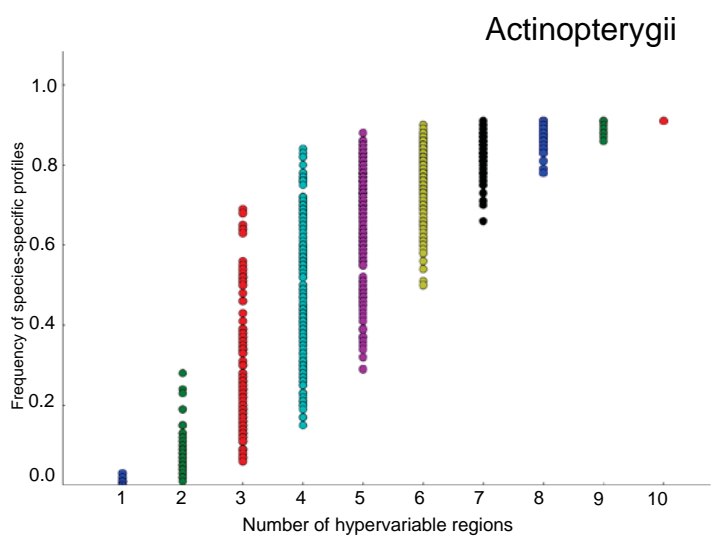
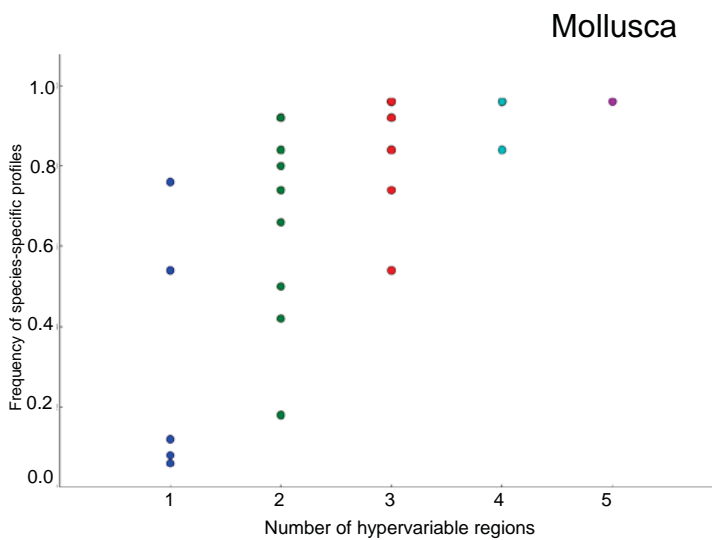
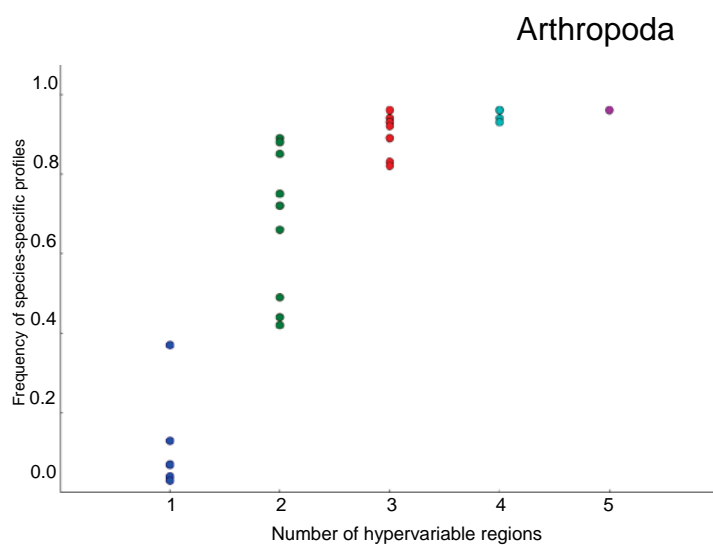
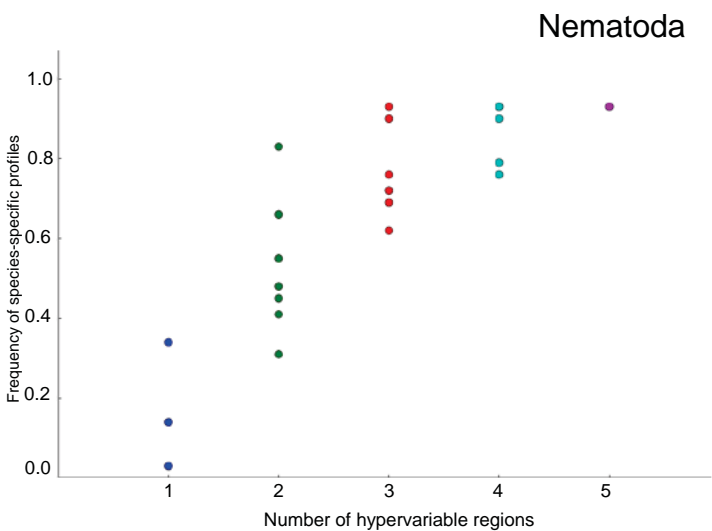
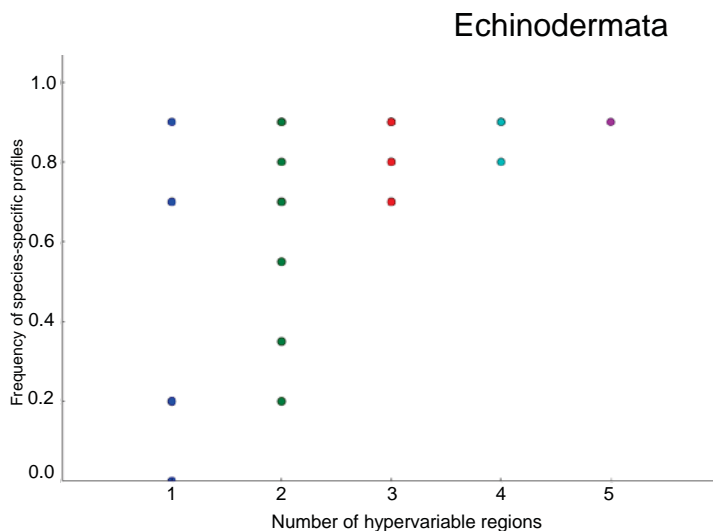
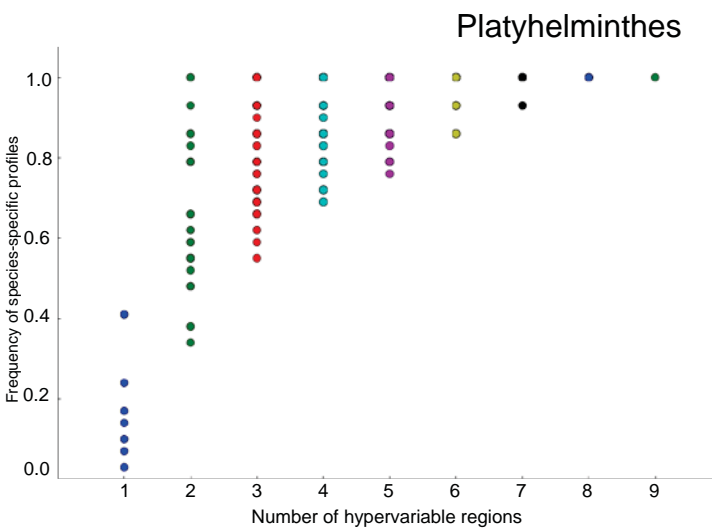
Frequency of species-specific alleles  
Average number of pairwise differences



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  $n$ .



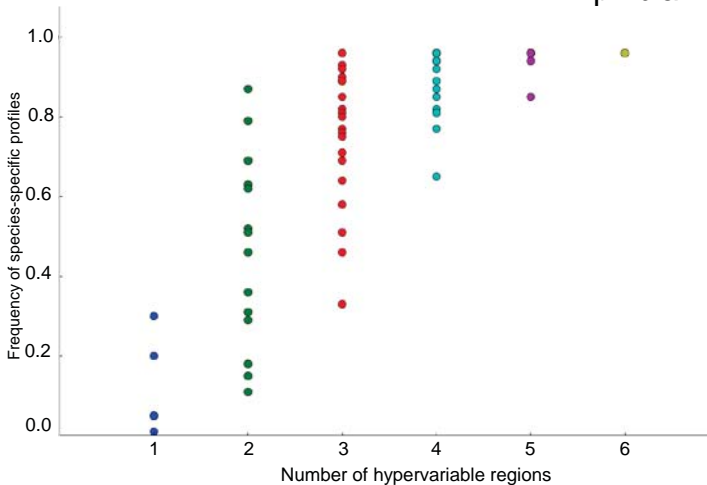
Supplementary Figure S18 (cont.)



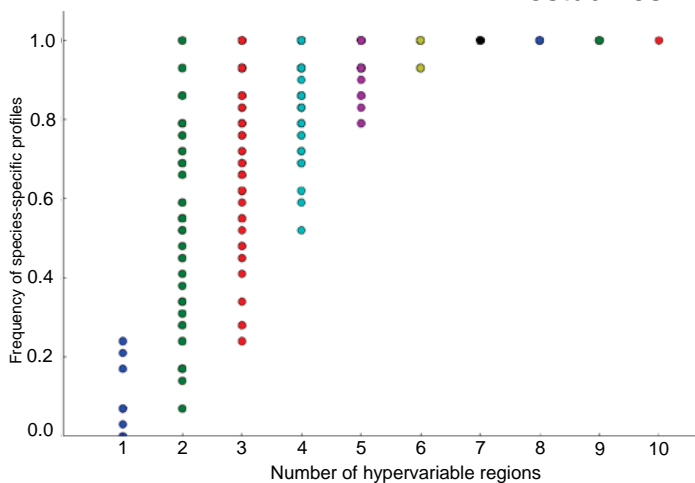


# Supplementary Figure S18 (cont.)

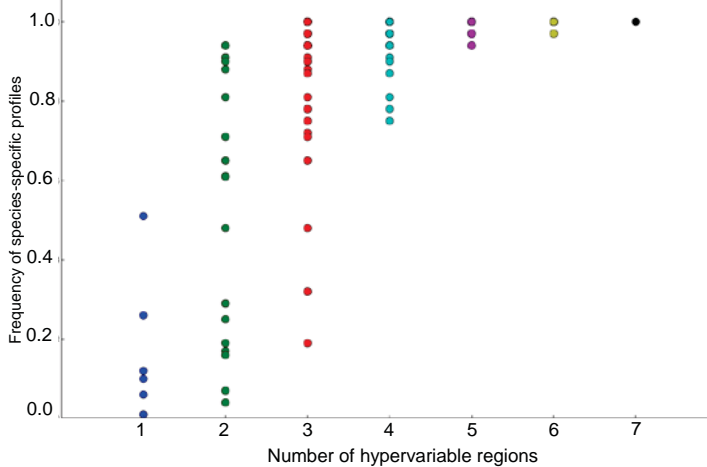
## Amphibia



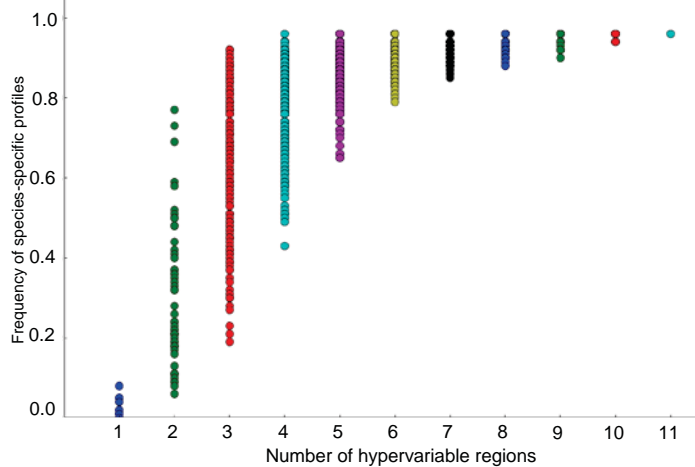
## Testudines



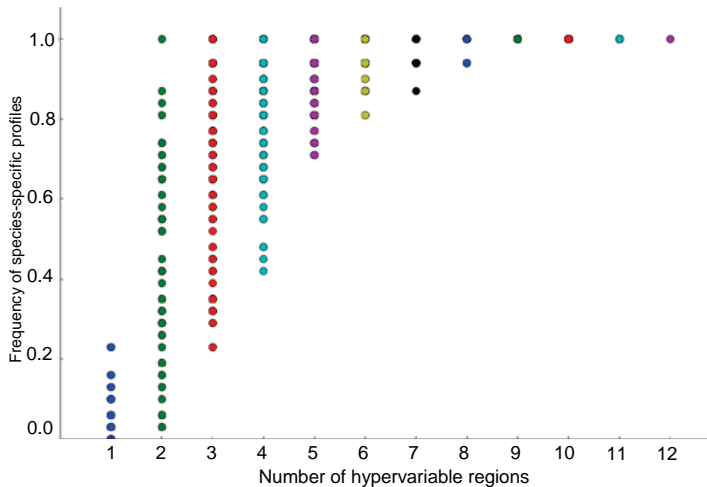
## Lepidosauria



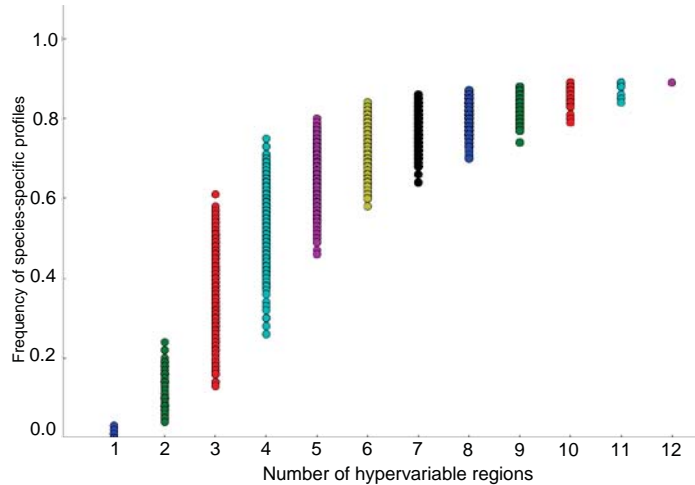
## Aves



## Metatheria



## Eutheria



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.

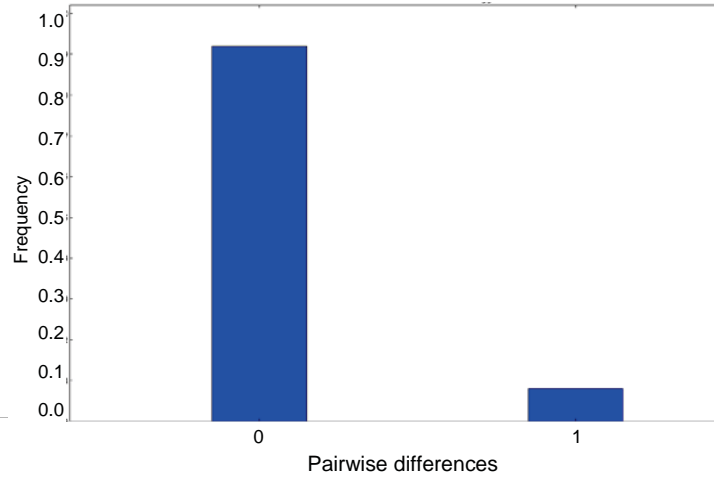
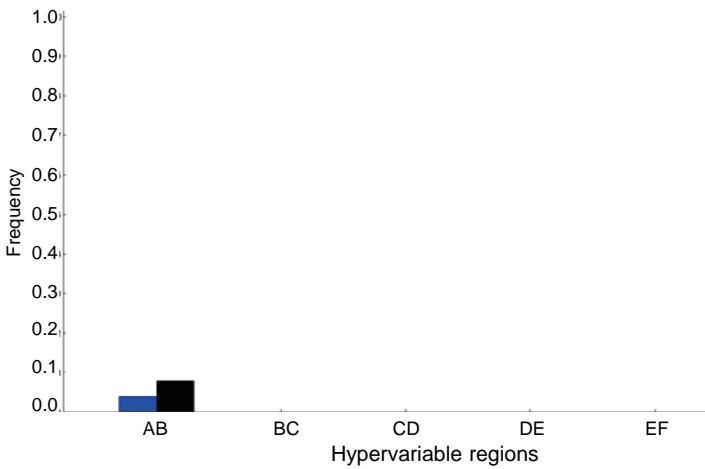
**a**

■ Frequency of species-specific alleles  
 ■ Average number of pairwise differences

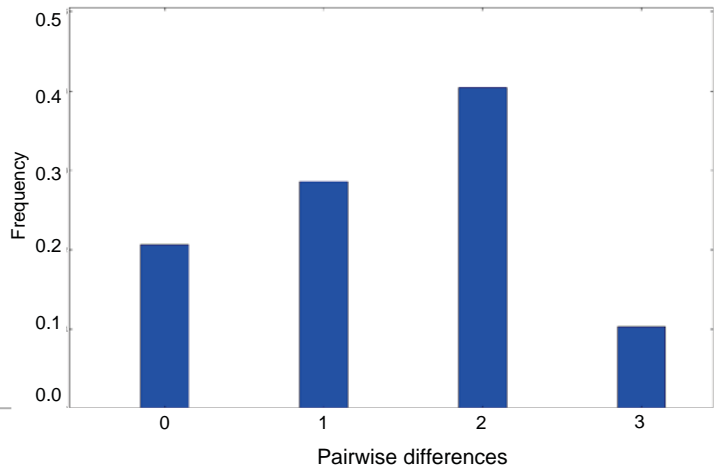
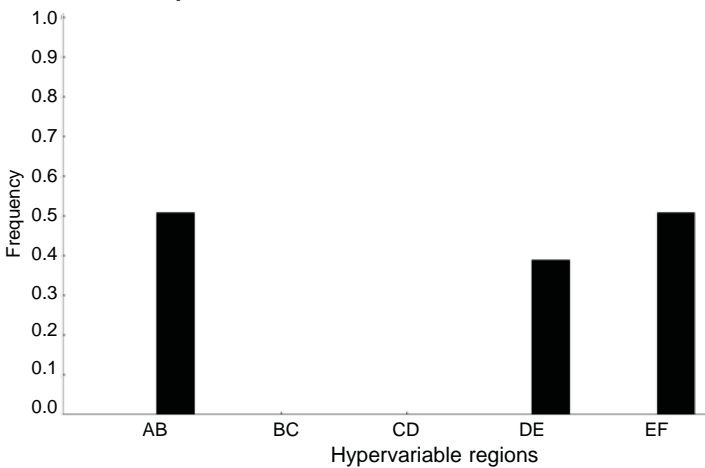
**b**

■ Mismatch distribution

*Caenorhabditis briggsae*



*Drosophila simulans*

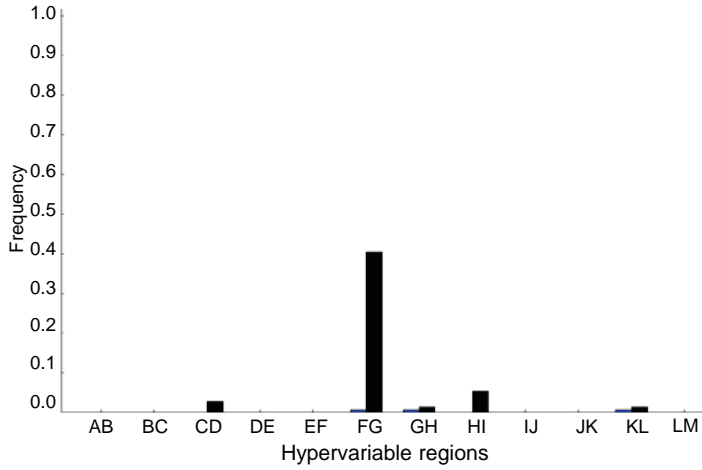


Supplementary Figure S19 (cont.)

**a**

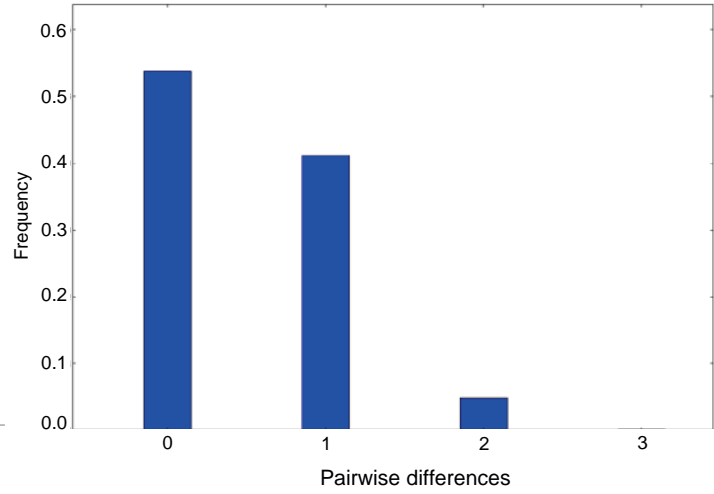
■ Frequency of species-specific alleles  
 ■ Average number of pairwise differences

*Bos taurus*

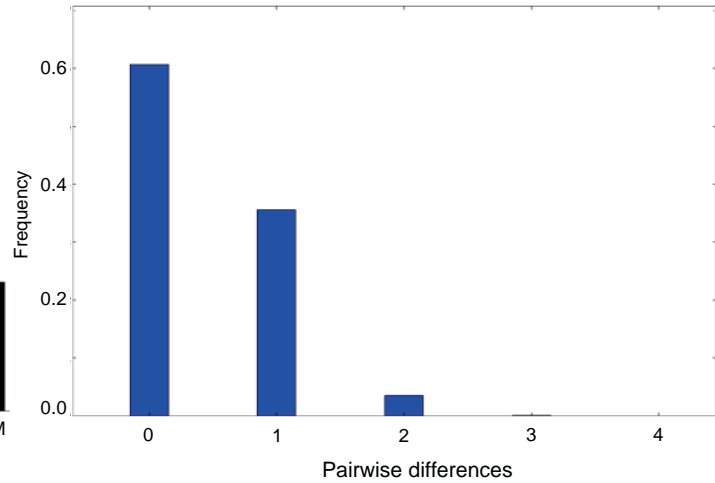
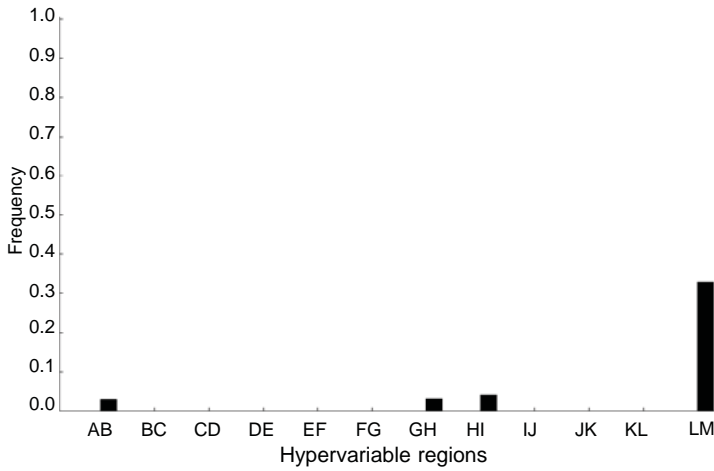


**b**

■ Mismatch distribution

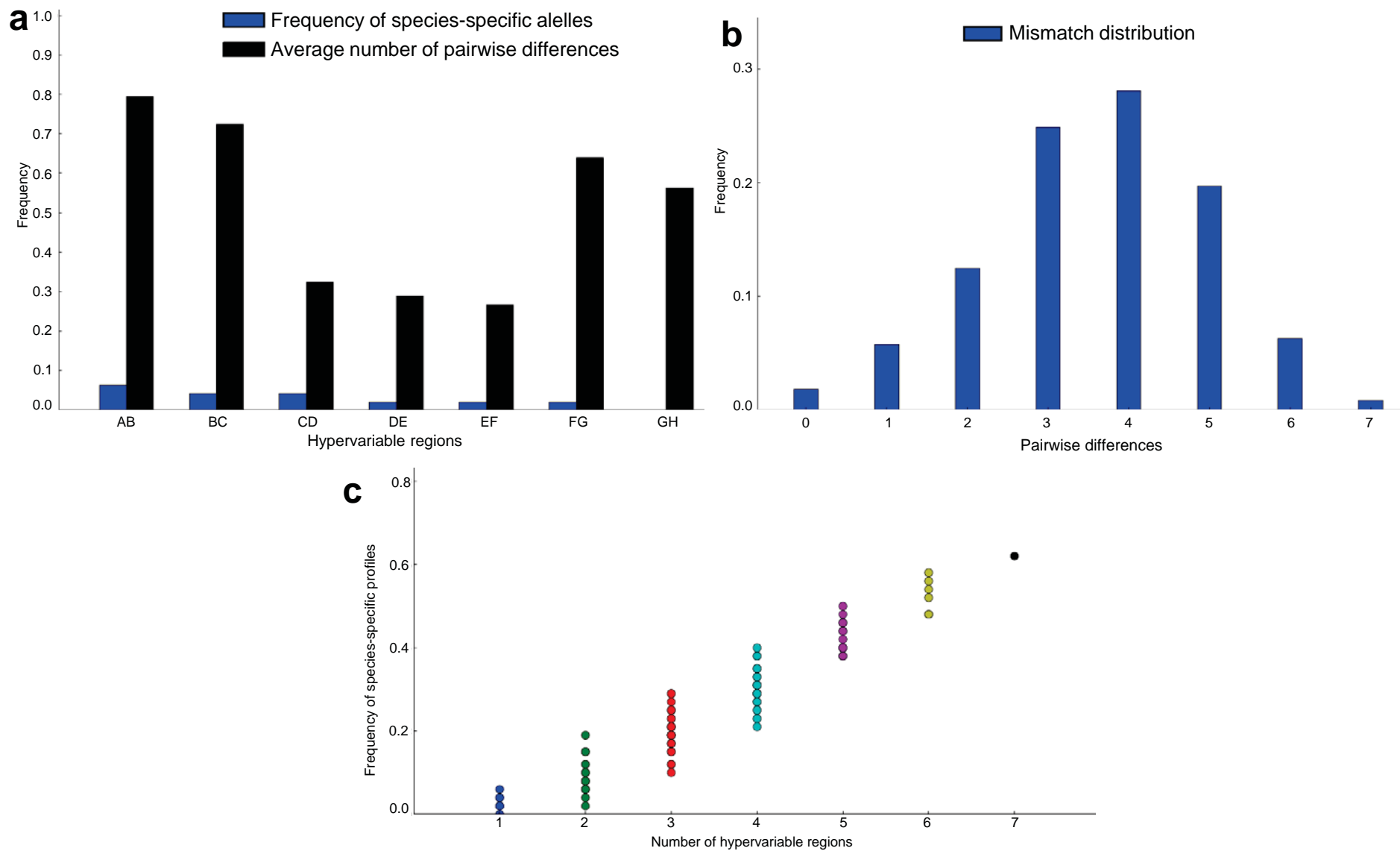


*Homo sapiens*



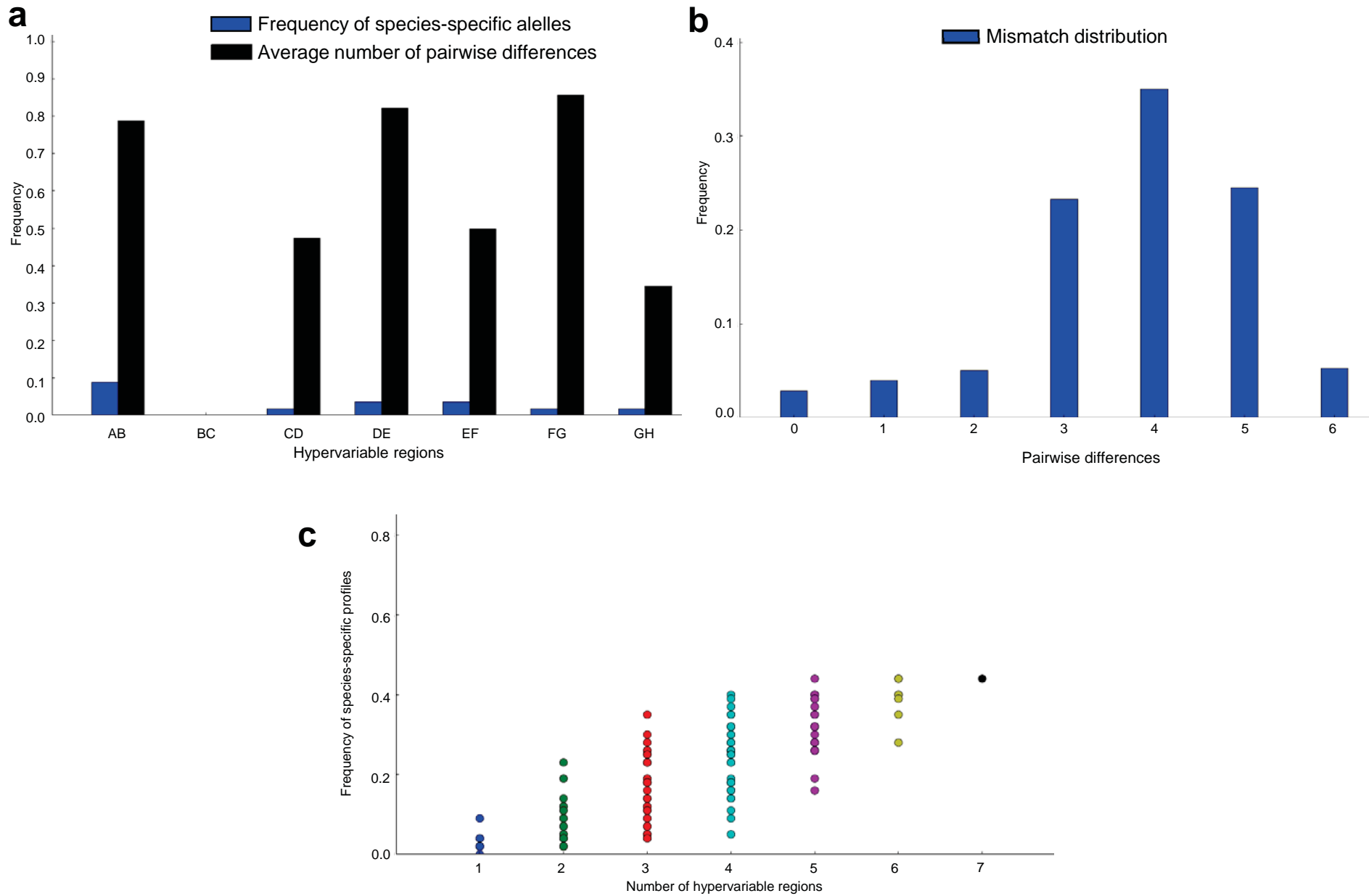
Supplementary Figure S20. Species identification in prokaryotic and viral species using the SPInDel method. **(a)** Frequency of species-specific alleles ( $f_n^G$ ) and the average number of pairwise differences ( $\bar{p}_n^G$ ) in each hypervariable region. **(b)** Mismatch distribution. **(c)** Frequency 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)



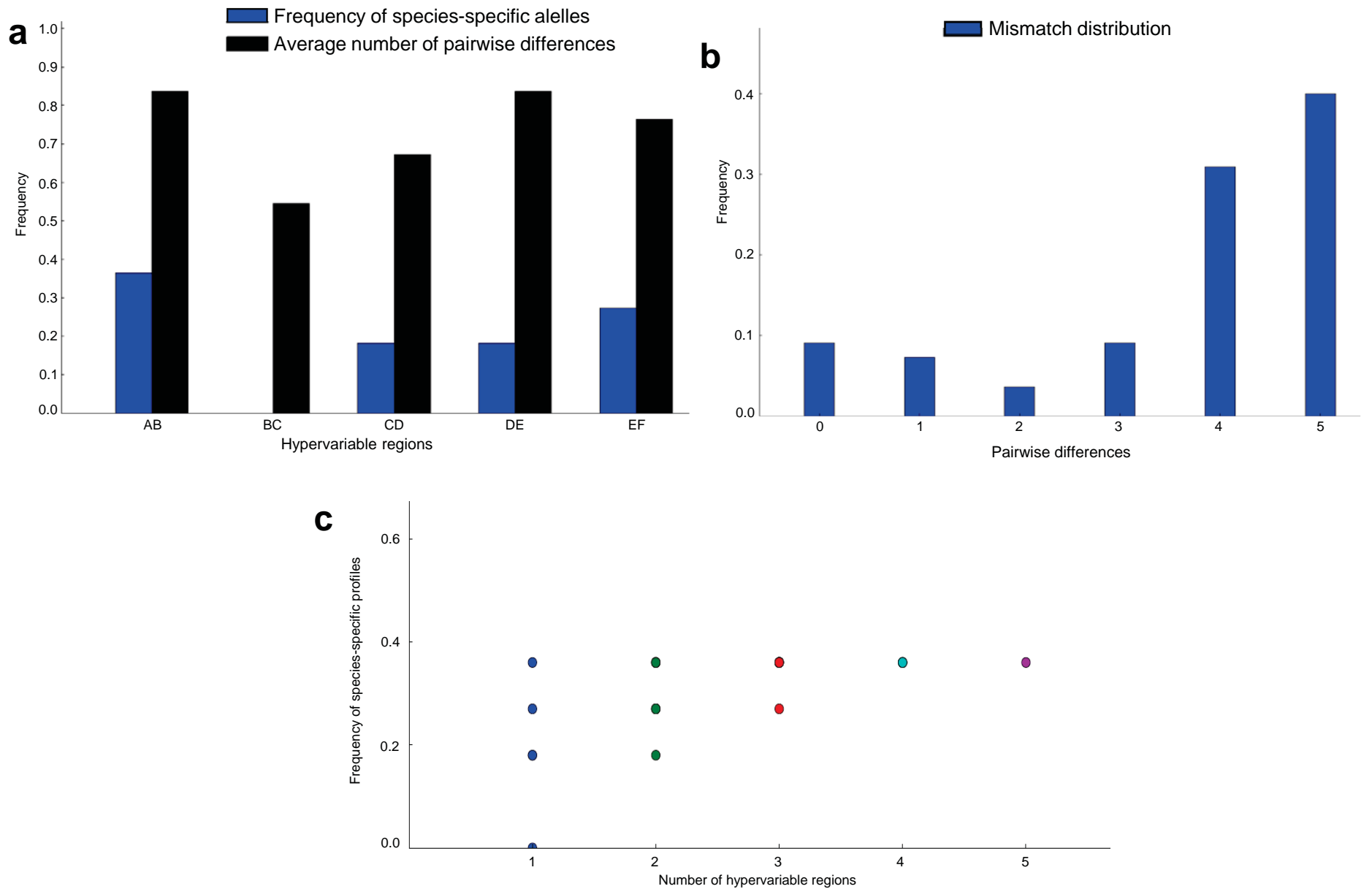
# Supplementary Figure S20 (cont.)

## Tenericutes (Bacteria)



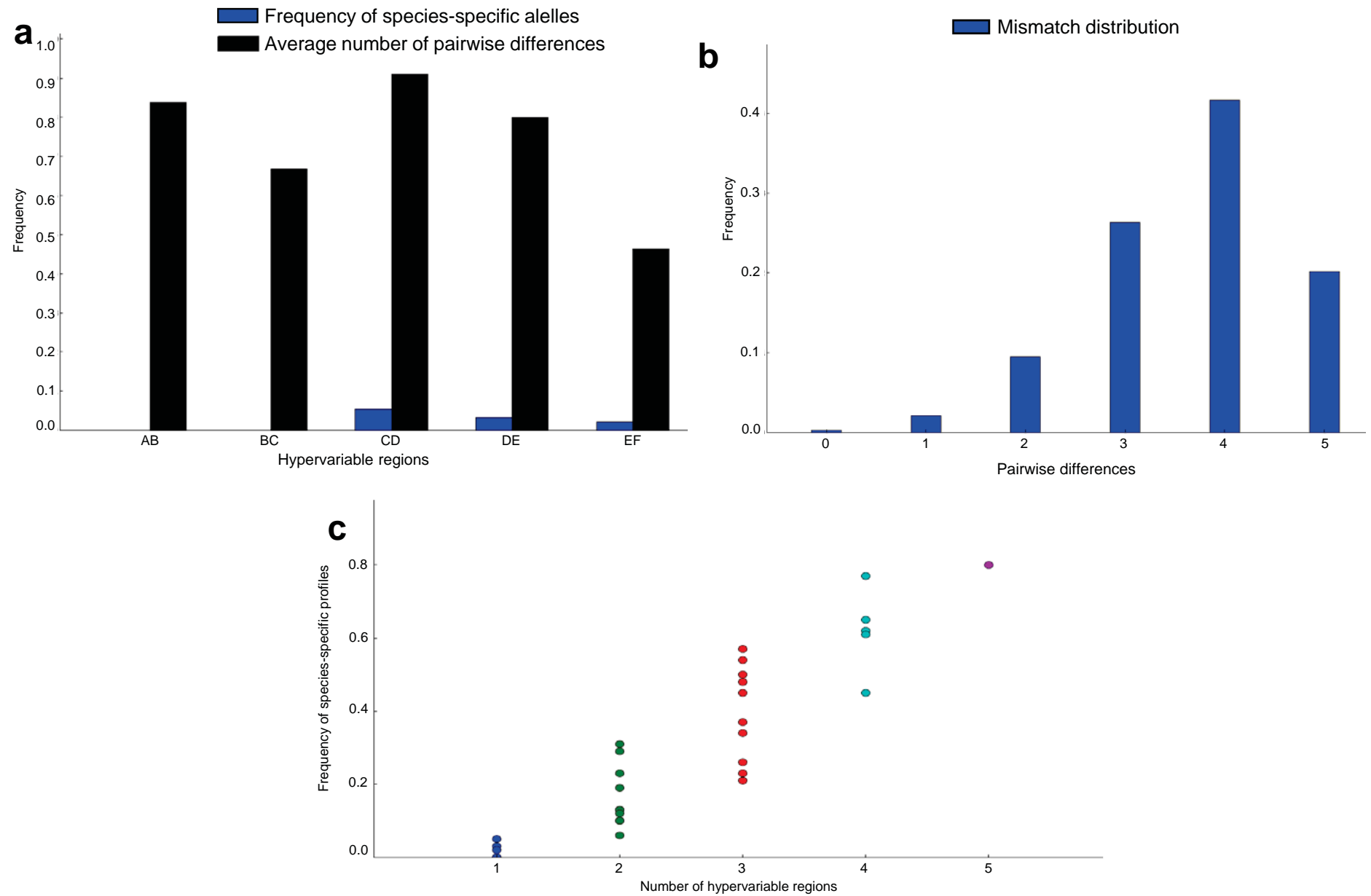
# Supplementary Figure S20 (cont.)

## Lentivirus (Retroviridae)



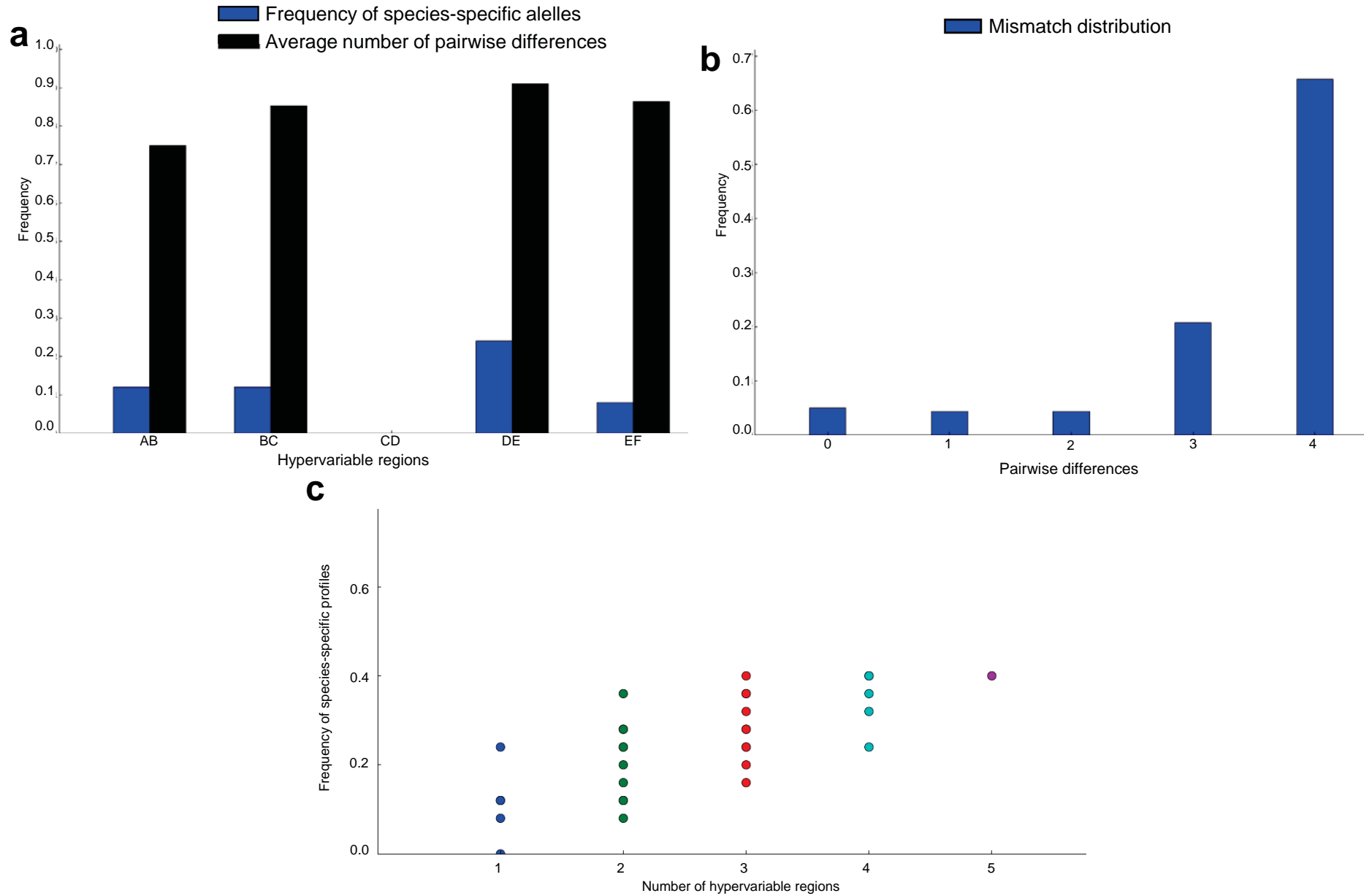
# Supplementary Figure S20 (cont.)

## Papillomaviridae (dsDNA viruses)



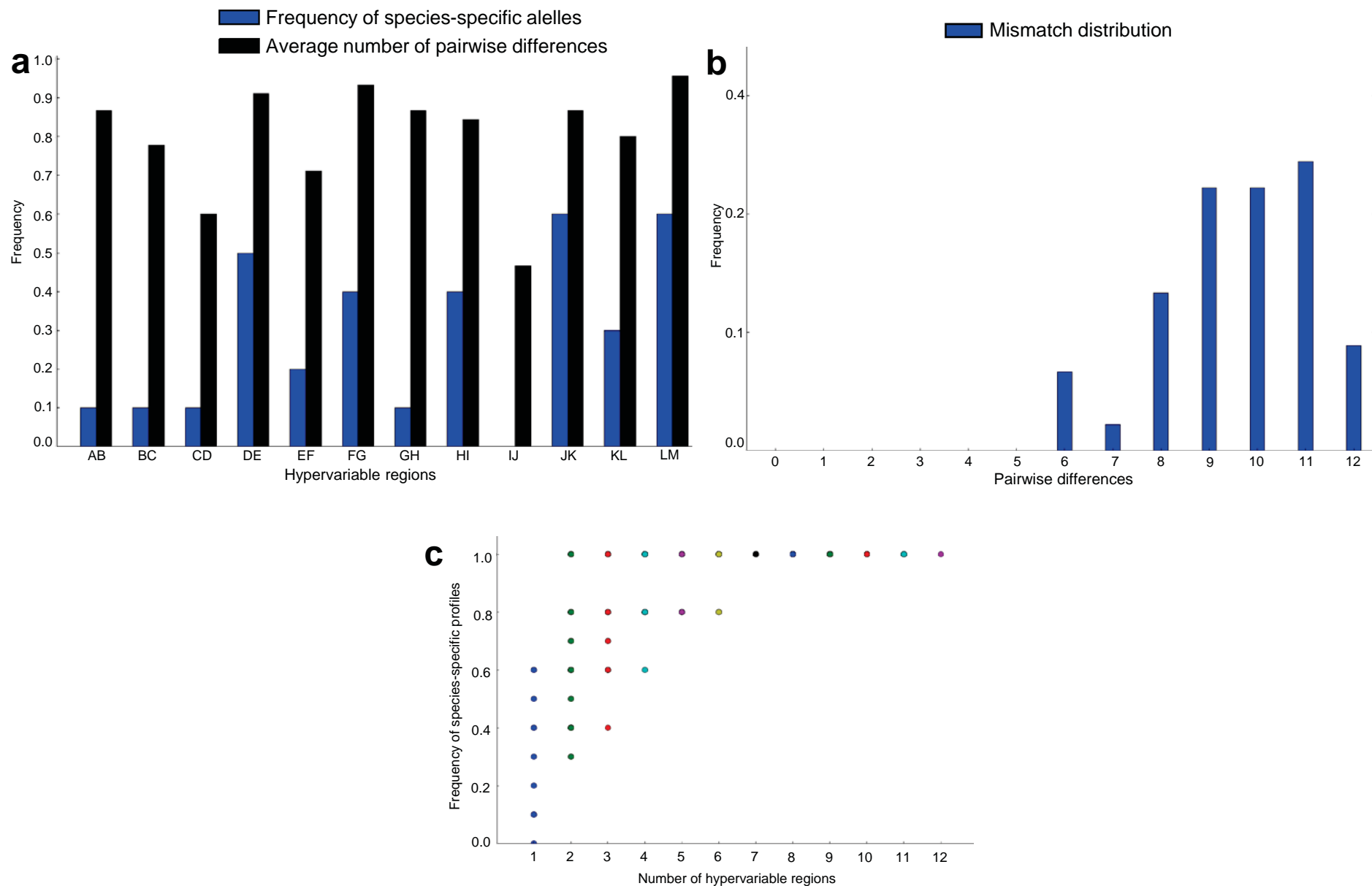
# Supplementary Figure S20 (cont.)

## 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).

