Supplementary Data

Amplification of EF1α

For EF1α, total RNA was isolated using RNAwiz™ (Ambion) and reverse transcribed using SuperScript™ II (Invitrogen). Amplification of EF1α was carried out using nested degenerate primers [40.6F, 41.2RC, 40.71F designed by Regier and Shultz[1]; and JH16R: 5'-KNRAANKNYTCNACRCACA-3'] with the external and internal pairs being 40.6F /41.2RC and 40.71F/JH16R, respectively. An initial touchdown PCR was performed (from 60°C to 55°C decreasing 1°C per 3 cycles, then from 55°C to 45°C decreasing 0.5°C per cycle for 21 cycles followed by 20 cycles at 45°C), and the products were reamplified at 50°C for 25 cycles.

Alignment and assessing substitutional saturation

Procedures to preliminary analyze our alignments prior to any phylogenetic analysis were slightly modified from the ones described and discussed by Jördens *et al*[2]. To examine the degree of variation more precisely for the structural genes, nuclear and mitochondrial versions of SSU and LSU genes, a sliding window analysis was performed. The percentage of variation within a window of nine bases was determined. The genetic variation is expressed as a percentage of the 27 possible base substitutions in a window of 9 bases. Effectively, this procedure assumes that each substitution is phylogenetically informative (Hennig's auxiliary principle), because the minimally possible number of substitutions to explain the character pattern in the window expresses the genetic variation. For example, if a position exhibits all four nucleotides three substitutions are at least necessary to explain the pattern (e.g., A→C, A→T, A→G). The window is then moved on one position and the percentage of variation is determined anew. For each position the mean is calculated out of the nine windows, which comprise this particular position. Classes of genetic variation were defined in 10% increments ranging from 0-100% (Suppl. Table 3). To obtain more or less the same number of positions per class of variation within the analysis the 10% increments were merged as shown in Suppl. Table 3. This procedure also reflects to some degree a discrete gamma distribution. In contrast to their nuclear counterparts, the mitochondrial SSU and LSU exhibit more variation, and thus they have greater proportional representation in higher classes of variation. Furthermore, the 1st, 2nd and 3rd positions of the protein-coding genes

(EF1α, ATP8, COX 1-3, CYTB, NAD6) were treated as different classes of genetic variation.

Supplementary Table 3: *Classes of genetic variation.* Number of positions in the 10% increments of the sliding window analyses of nucSSU, nucLSU, mtSSU and mtLSU and in the merged increments. Ppi = Positions per increment; Ppmi = Positions per merged increments.

Increments	nucSSU		nucLSU		mtSSU		mtLSU	
	Ppi	Ppmi	Ppi	Ppmi	Ppi	Ppmi	Ppi	Ppmi
0-10%	328	328	351	351	17		34	
10-20%	323	323	431	431	60	184	86	243
20-30%	199	199	453	453	107		123	
30-40%	207	207	419	419	82		116	220
40-50%	206	206	305	305	68	222	104	
50-60%	112	112	313	313	72		69	
60-70%	99		232	232	65		14	
70-80%	31	144	106		5	75	3	86
80-90%	14		34	153	5		0	
90-100%	0		13		0		0	

Homogeneity of base frequencies, uncorrected substitution rates as well as transition (= TI) to transversion (TV) ratio were estimated for each of the classes of variation to determine saturated positions. Using the options “pairdiff”, “include”, and “exclude” of PAUP4.0b[3], a pair wise comparison of the aligned sequences is performed for each class of variation. To detect saturation the TI/TV ratios have been plotted against the uncorrected distances p. Due to both graphical convenience and better comparability between the different genes the plot only displays the mean values of the TI/TV ratios and genetic distances p as well as their corresponding standard deviation (Suppl. Fig. 5). Suppl. Table 4 summarizes the homogeneity of base frequencies for the different genes and their classes of genetic variation.

Overall the results for the structural genes nucSSU, nucLSU, mtSSU and mtLSU are similar independent of the genomic origin. With increasing genetic variation the TI/TV ratios converge on an equilibrium indicating saturation of transitions. The standard deviation for TI/TV ratios decreases whereas the standard deviation of genetic distance increases (Suppl. Fig. 5). Although homogeneity of base frequencies is given for nearly all classes of variations in the different genes (except 60-100% of mtSSU; Suppl. Table 4) we excluded the highest

class of genetic variation (50-100%, 60-100%, 70-100%, respectively) in each of the four genes due to the indication of saturation (Suppl. Fig. 5).

In the case of the nuclear protein-coding gene EF1α saturation and heterogeneity of base frequencies can be detected in the 3rd positions and thus they were excluded.

For the 1st positions of the mitochondrial protein-coding genes (ATP8, COX 1-3, CYTB, NAD6) convergence of TI/TV ratios can be shown and thus an indication of saturation of transitions. Furthermore, their 3rd positions exhibit a cloud-like relationship of TI/TV ratios and genetic distance p. In combination with extremely high genetic distances p of about 0.6 this is indicative for additionally saturated transversions. Furthermore, homogeneity of base frequencies is not given for the 1st and 3rd positions of these genes. Thus, we excluded both 1st and 3rd positions of ATP8, COX 1-3, CYTB and NAD6.

Supplementary Table 4: *Homogeneity of base frequencies.* P values of the classes of genetic variation of the different genes used. Significant differences ($p < 0.05$) in bold.

<table>
<tr><th>Increments</th><th>nucSSU</th><th>nucLSU</th><th>mtSSU</th><th>mtLSU</th></tr>
<tr><td>0-10%</td><td>1.00</td><td>1.00</td><td rowspan="3">0.99</td><td rowspan="3">0.99</td></tr>
<tr><td>10-20%</td><td>1.00</td><td>1.00</td></tr>
<tr><td>20-30%</td><td>1.00</td><td>1.00</td></tr>
<tr><td>30-40%</td><td>1.00</td><td>1.00</td><td rowspan="2">0.30</td><td rowspan="2">0.21</td></tr>
<tr><td>40-50%</td><td>1.00</td><td>1.00</td></tr>
<tr><td>50-60%</td><td>1.00</td><td>1.00</td><td rowspan="5"><0.01</td><td rowspan="5">0.10</td></tr>
<tr><td>60-70%</td><td rowspan="3">1.00</td><td>1.00</td></tr>
<tr><td>70-80%</td><td rowspan="2">0.99</td></tr>
<tr><td>80-90%</td></tr>
<tr><td>90-100%</td><td></td><td></td></tr>
</table>

<table>
<tr><th>Position</th><th>EF1α</th><th>ATP8</th><th>COX1</th><th>COX2</th></tr>
<tr><td>1st</td><td>1.00</td><td><0.01</td><td><0.01</td><td><0.01</td></tr>
<tr><td>2nd</td><td>1.00</td><td>0.98</td><td>0.99</td><td>0.99</td></tr>
<tr><td>3rd</td><td><0.01</td><td><0.01</td><td><0.01</td><td><0.01</td></tr>
</table>

<table>
<tr><th>Position</th><th>COX3</th><th>CYTB</th><th>NAD6</th></tr>
<tr><td>1st</td><td>0.02</td><td>0.03</td><td><0.01</td></tr>
<tr><td>2nd</td><td>0.99</td><td>0.99</td><td>0.03</td></tr>
<tr><td>3rd</td><td><0.01</td><td><0.01</td><td><0.01</td></tr>
</table>

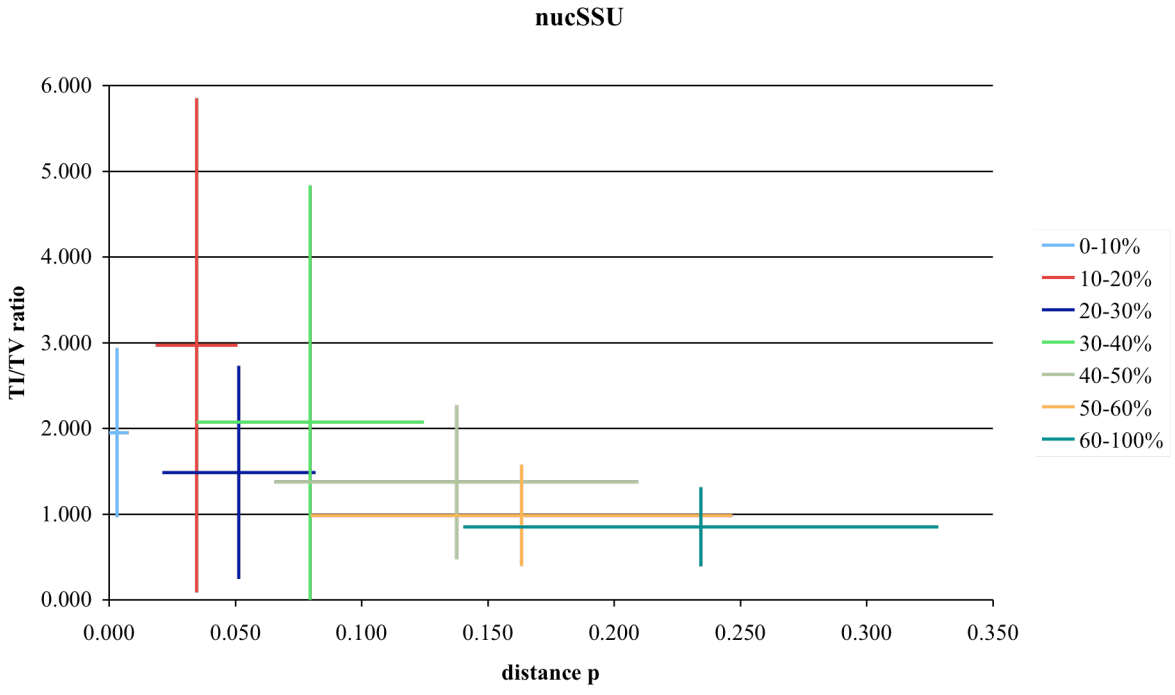
Phylogenetic results of partitions

Topologies produced by ML and BI are similar or identical within data sets; however, the best trees differed among the individual partitions and combined data sets. Individual partitions are congruent with previous studies based on single genes in that they are inconclusive for higher-level relationships concerning Annelida, Clitellata, and Sipuncula[4,5]. Therefore, we focus mainly on results of combined data sets. Even though trees obtained by the two data sets (Nuc and NucMt) are not congruent with each other, most conclusions regarding the affiliation of major groups like Sipuncula and Clitellata to polychaetes are the same, especially considering explicit hypotheses tested. Due to taxon sampling, the NucMt

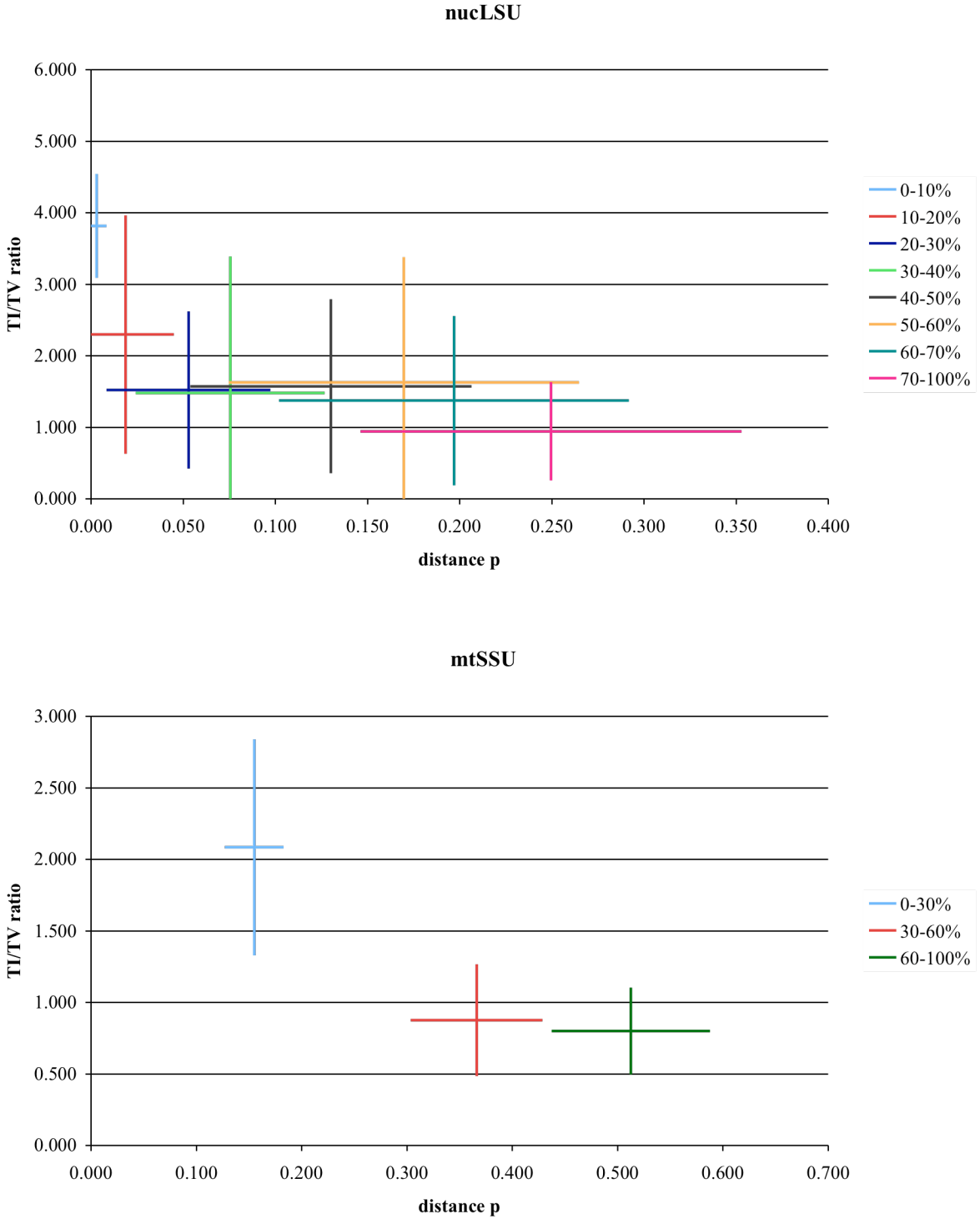
data set is limited concerning inference of intra-polychaete relationships and conclusions herein are based on the more inclusive Nuc data set with 45 polychaete “families”.

References

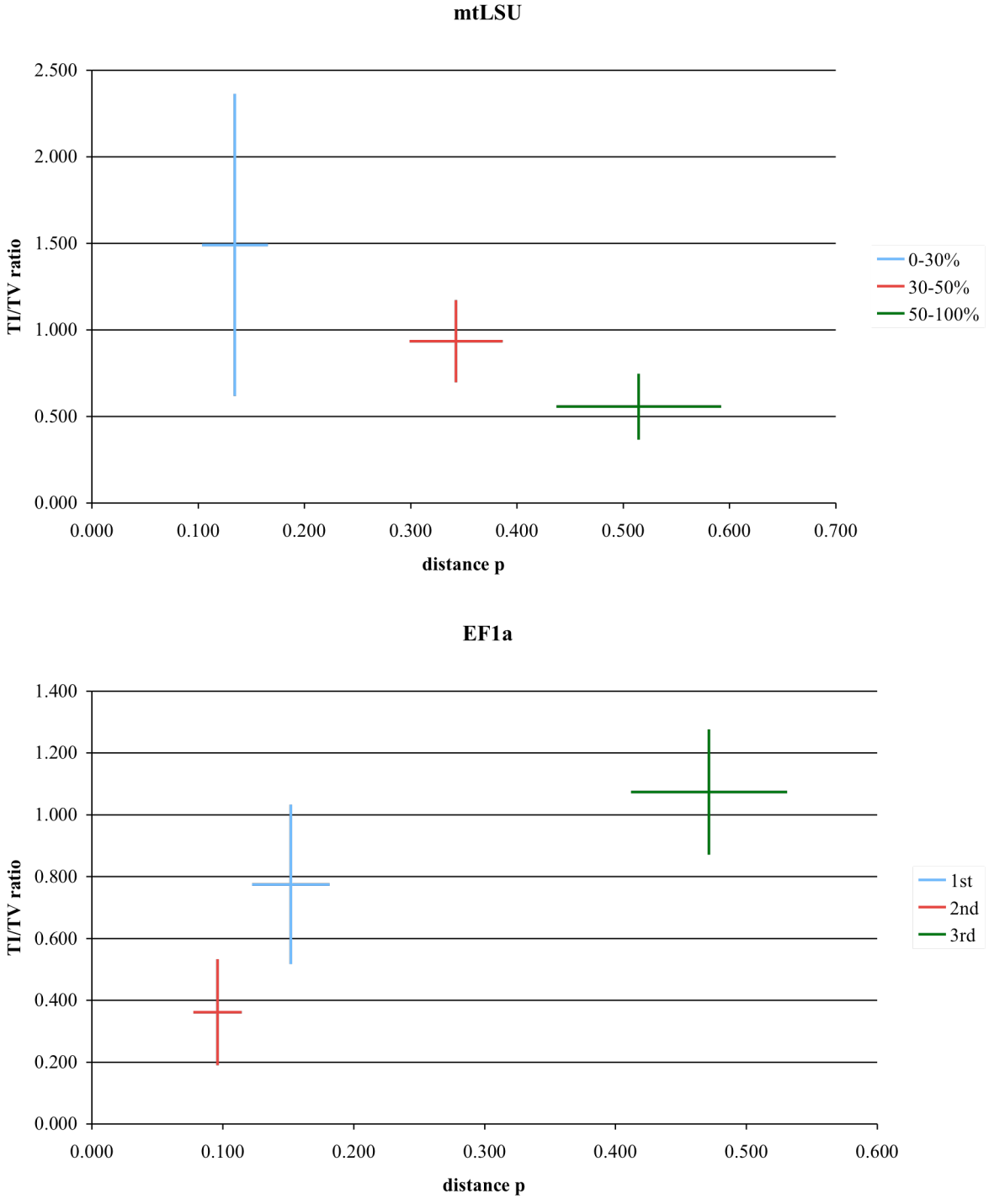
1. Regier, J. & Shultz, J. Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods. *Mol. Biol. Evol.* **14**, 902-913 (1997).
2. Jördens, J., Struck, T. H. & Purschke, G. Phylogenetic inference of Parergodrilidae and *Hrabeiella periglandulata* ("Polychaeta", Annelida) based on sequences of the CO I, 18S- and 28S-rDNA. *J. Zool. Syst. Evol. Res.* **42**, 270-280 (2004).
3. Swofford, D. L. (Sinauer Associates, Sunderland, MA, 2002).
4. Struck, T. H. & Purschke, G. The sistergroup relationship of Aeolosomatidae and Potamodrilidae - a molecular phylogenetic approach based on 18S rDNA and Cytochrome Oxidase I. *Zool. Anz.* **243**, 281-293 (2005).
5. Hall, K. A., Hutchings, P. A. & Colgan, D. J. Further phylogenetic studies of the Polychaeta using 18S rDNA sequence data. *J. Mar. Biol. Ass. UK* **84**, 949-960 (2004).



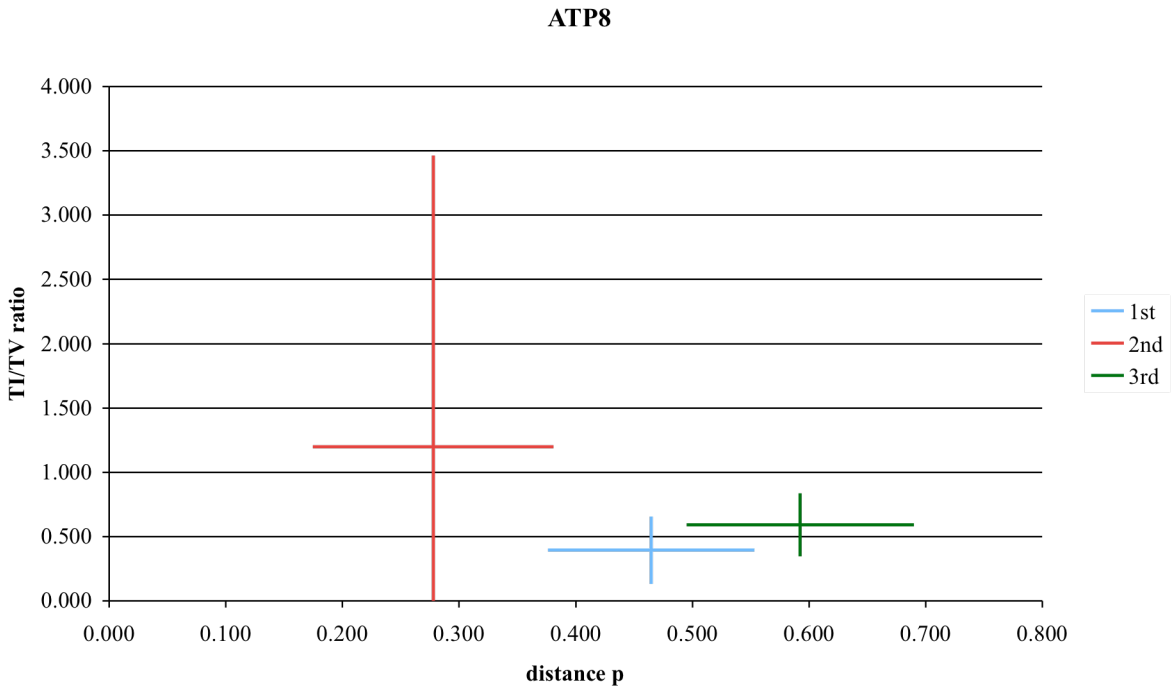
Supplementary Fig. 5. Plots of mean values of the TI/TV ratios and genetic p distances p as well as their corresponding standard deviation.

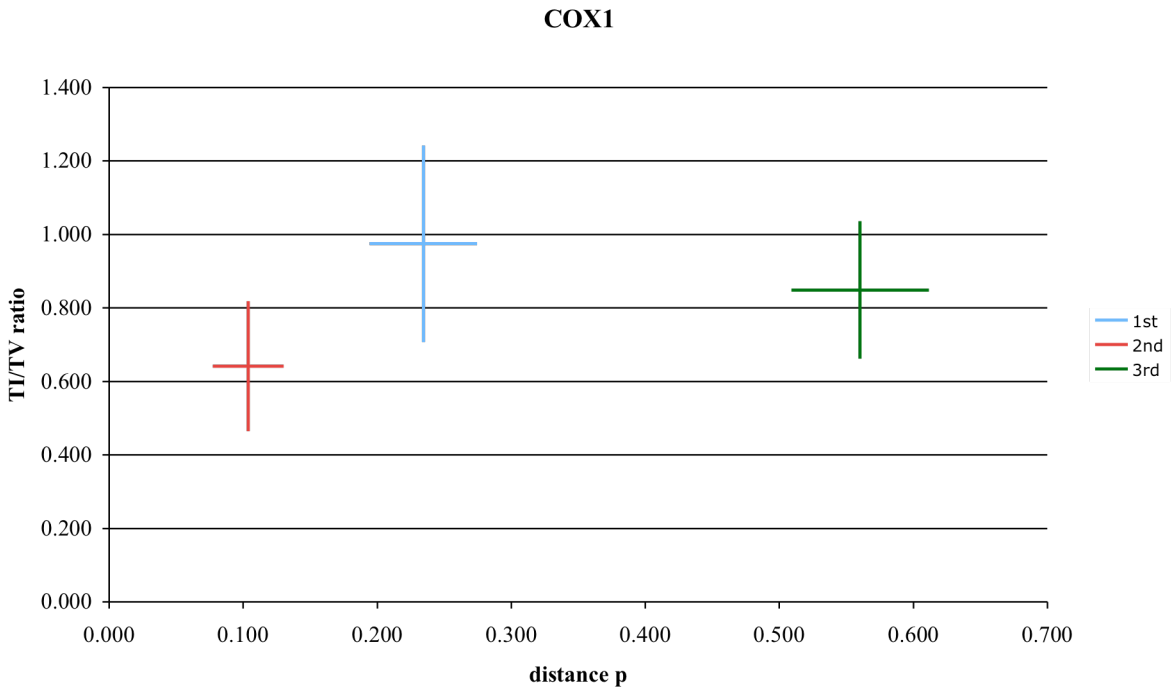


Suppl. Fig. 5. Continued.

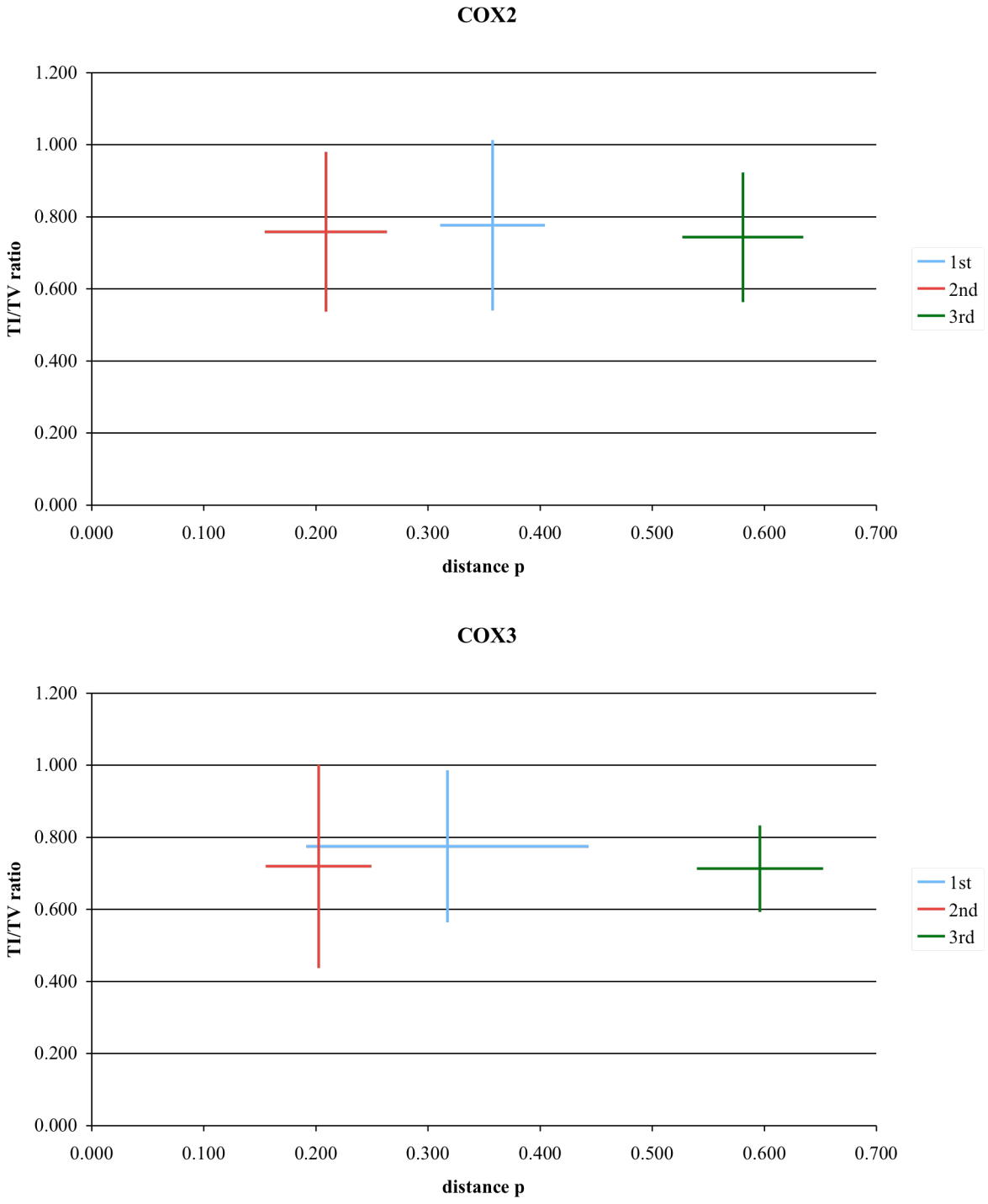


Suppl. Fig. 5. Continued.

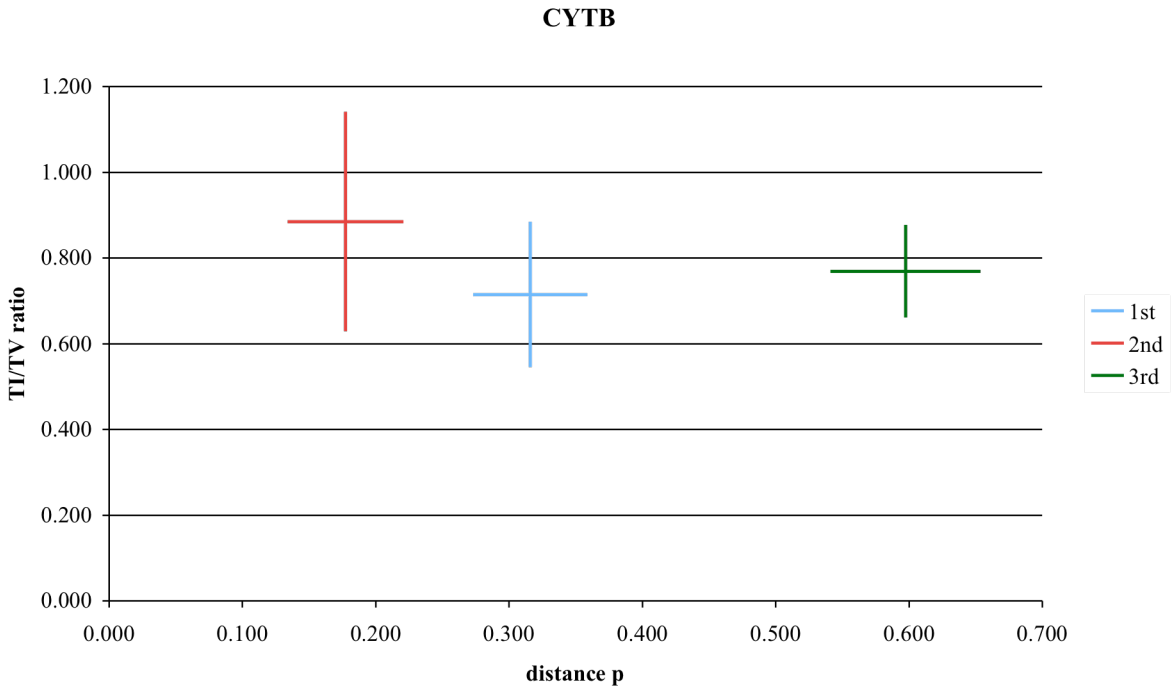


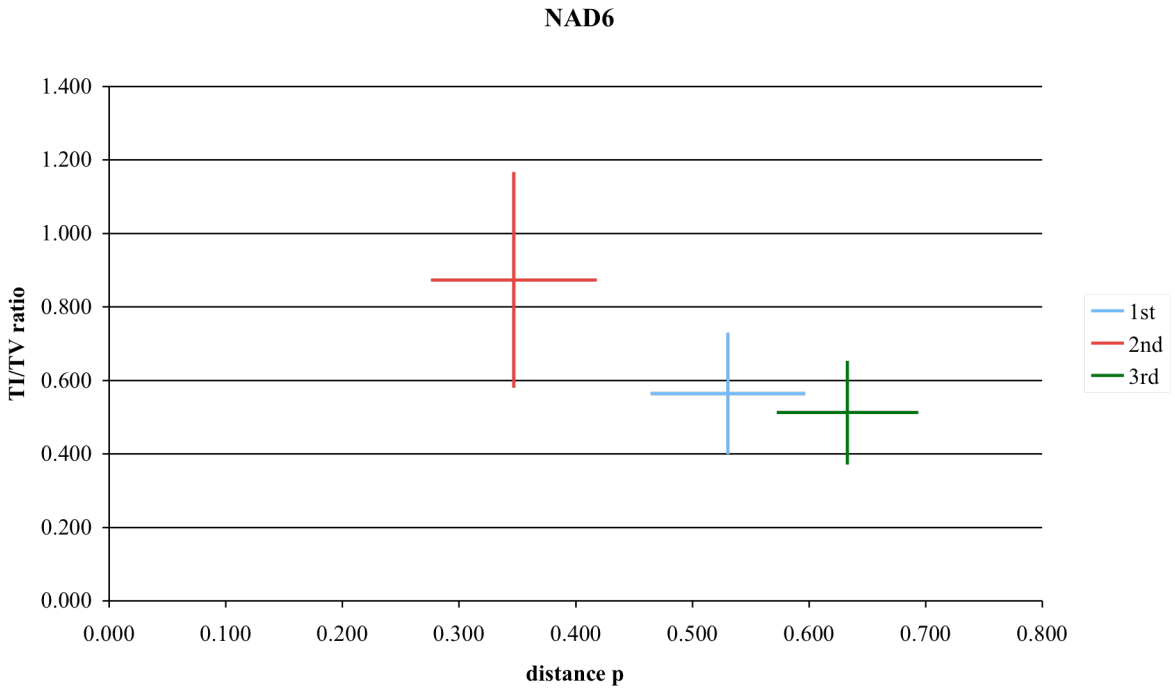


Suppl. Fig. 5. Continued.



Suppl. Fig. 5. Continued.





Suppl. Fig. 5. Continued.