Effect of metal pollution on genetic variation in natural populations of selected soil invertebrate species with different dispersal potential

Giska, I.

2016

document version
Publisher's PDF, also known as Version of record

Chapter 2

Deeply divergent sympatric mitochondrial lineages of the earthworm *Lumbricus rubellus* are not reproductively isolated

Iwona Giska, Pierfrancesco Sechi, Wiesław Babik

Published: BMC Evolutionary Biology 2015, 15:217

Abstract

The accurate delimitation of species is essential to numerous areas of biological research. An unbiased assessment of the diversity, including the cryptic diversity, is of particular importance for the below ground fauna, a major component of global biodiversity. On the British Isles, the epigeic earthworm *Lumbricus rubellus*, which is a sentinel species in soil ecotoxicology, consists of two cryptic taxa that are differentiated in both the nuclear and the mitochondrial (mtDNA) genomes. Recently, several deeply divergent mtDNA lineages were detected in mainland Europe, but whether these earthworms also constitute cryptic species remains unclear. This information is important from an evolutionary perspective, but it is also essential for the interpretation and the design of ecotoxicological projects. In this study, we used genome-wide RADseq data to assess the reproductive isolation of the divergent mitochondrial lineages of *L. rubellus* that occur in sympatry in multiple localities in Central Europe.

We identified five divergent (up to 16% net p-distance) mitochondrial lineages of *L. rubellus* in sympatry. Because the clustering of the RADseq data was according to the population of origin and not the mtDNA lineage, reproductive isolation among the mtDNA lineages was not likely. Although each population contained multiple mtDNA lineages, subdivisions within the populations were not observed for the nuclear genome. The lack of fixed differences and sharing of the overwhelming majority of nuclear polymorphisms between localities, indicated that the populations did not constitute allopatric species. The nucleotide diversity within the populations was high, 0.7-0.8%. The deeply divergent mtDNA sympatric lineages of *L. rubellus* in Central Europe were not reproductively isolated groups. The earthworm *L. rubellus*, which is represented by several mtDNA lineages in continental Europe, apparently is a single highly polymorphic species rather than a complex of several cryptic species. This study demonstrated the critical importance of the use of multilocus nuclear data for the unbiased assessment of cryptic diversity and for the delimitation of species in soil invertebrates.

Keywords

species delimitation, cryptic species, RADseq, mtDNA, *Lumbricus rubellus*, soil diversity

Introduction

Species delimitation aims to identify species-level biological diversity while delineating interspecific boundaries and estimating the number of species (Fujita et al. 2012; Carstens et al. 2013). The accurate delimitation of species is of paramount importance in numerous fields, including evolutionary biology, systematics, biogeography, conservation biology and many areas of experimental biology (Barley et al. 2013). Traditionally, species have been identified based on morphological traits. However, a large portion of the biological diversity may be impossible to detect by relying only on morphological characters (Bickford et al. 2007). These difficulties are most apparent in taxonomic groups that include closely related, recently diverged species, which form complexes of cryptic species. The morphology-based species delimitation may also severely underestimate the overall diversity in taxonomic groups with morphological uniformity or with a paucity of taxonomically useful morphological characters. Today, many biologists agree that species may be separately evolving metapopulation lineages (de Queiroz 1998 and 2007), which is a deliberately loose definition to proceed beyond the unresolvable debate about the species concepts. This lineage-based interpretation of species shifts the focus to genetic data and the other nonmorphological characters. The DNA data may be a source of valuable additional information to develop new and more accurate species delimitation methods that should be used by alpha taxonomists (Bickford et al. 2007). Distinguishing between the two groups of criteria that are used for species delimitation, the pattern-oriented and the process-oriented criteria, is useful (Reeves and Richards 2011). The pattern-oriented criteria reflect the effect of a lineage existence, e.g., monophyly, diagnosability or formation of distinct genotypic clusters, whereas the process-oriented criteria identify the evolutionary cause of the lineage existence, e.g., the reproductive isolation or occupation of a distinct niche. Species treated as separate evolutionary lineages can be delimited based on these criteria even when the species definition or concept is debated (de Queiroz 2007). The use of multiple criteria is recommended to increase the chance to detect recently separated lineages and to obtain clear evidence of the lineages as separate entities (Reeves and Richards 2011; Bacon et al. 2012).

In the past, the molecular identification of species involved primarily mitochondrial DNA (mtDNA) sequences. The most commonly used mitochondrial marker has been Cytochrome Oxidase I

(*COI*), which is a standard in the DNA barcoding of animal species, under the sometimes questioned (Meier et al. 2006) assumptions of low variation within species and high differentiation between species (Hebert et al. 2003). Within species, high mtDNA differentiation is often observed between allopatric populations; it may or may not be accompanied by a differentiation in the nuclear markers. Sympatric mtDNA divergence is less common, and divergent sympatric lineages often show reproductive isolation and divergence in the nuclear genome (e.g., Haine et al. 2006). However, mtDNA differentiation is often reported to be discordant with the differentiation based on the nuclear genetic markers, and multiple explanations have been proposed for this pattern (Hogner et al. 2012; Toews and Brelsford 2012; Carstens et al. 2013). Thus, the joint analysis of mitochondrial and nuclear markers in sympatry is more likely to provide a robust test to identify cryptic species and assess species boundaries.

The delimitation of species is of primary importance in the study of belowground fauna. Global biodiversity is determined to a large extent by the belowground communities, and soil is one of the most species-rich terrestrial habitats (Wolters 2001). A high percentage of species is estimated to remain undescribed for most soil taxa, and this lack of information is likely due to a lack of taxonomic knowledge and expertise, particularly in the case of small body-sized animals (Decaëns et al. 2006). High cryptic diversity has been detected with DNA based methods in soil invertebrates, including springtails (Timmermans et al. 2005; Porco et al. 2012; Cicconardi et al. 2013), earthworms (King et al. 2008; Klarica et al. 2012), mites (Schäffer et al. 2010) and centipedes (Spelda et al. 2011). Some of these soil invertebrates are considered sentinel species in ecotoxicology. The knowledge of their taxonomy, including the cryptic diversity, is critical for the proper design and for the interpretation of ecotoxicological experiments. The divergent evolution of cryptic species may lead to physiological differences, e.g., differential sensitivity to environmental stressors, including pollution. Individuals belonging to separate evolutionary lineages may display significant differences in the sensitivity to toxicants, and such a phenomenon was reported for the evolutionary lineages of the aquatic oligochaete *Tubifex tubifex*, which consisted of five cryptic species. The sensitivity to Cd, which was assessed based on the mortality and time to death, differed among these lineages (Sturmbauer et al. 1999). The distribution of genetic lineages in nature can be expected to be shaped by pollution when

the different levels of resistance of the mtDNA lineages to toxicants are considered. Thus, more sensitive lineages would be lost at more polluted sites, as predicted by the genetic erosion hypothesis, which posits the loss of genetic diversity because of pollution (Van Straalen and Timmermans 2002).

The species of lumbricid earthworms often consist of highly divergent mitochondrial lineages (King et al. 2008; Klarica et al. 2012; Dupont et al. 2011; Römbke et al. 2015). The epigeic earthworm *Lumbricus rubellus* Hoffmeister, 1843, a sentinel species in ecotoxicology, is found in the UK as two distinct mtDNA lineages, A and B. The mtDNA sequence divergence between these lineages exceeded 13% at both *COI* (King et al. 2008) and *COII* (Donnelly et al. 2013). Recently, differentiation in the nuclear markers was also found between individuals from the two mtDNA lineages in sympatry, which implied reproductive isolation and supported the cryptic species hypothesis (Donnelly et al. 2013). Several other deeply divergent mitochondrial lineages have been found within mainland Europe (Sechi 2013); however, whether these continental lineages are reproductively isolated and represent cryptic species is not known. In Poland, we have identified highly divergent mtDNA lineages in earthworms in sympatry, which represent several of the lineages observed across Europe. Thus, the conditions were favorable to test for reproductive isolation between the mtDNA lineages of *L. rubellus* from continental Europe.

In this work, we tested whether the sympatric mitochondrial lineages of *L. rubellus* that were found in Poland represented cryptic species. We expected the nuclear clustering to be concordant with the sympatric mtDNA lineages if the mtDNA lineages corresponded to reproductively isolated groups. Additionally, this study aimed to estimate the genetic diversity of the *L. rubellus* populations and to compare the haplotypes found in Poland with the mtDNA lineages observed across Europe. The mitochondrial lineages were characterized based on *COI* and *ATP6* sequences, whereas the multilocus genotype data that were generated by the Restriction site Associated DNA Sequencing approach; RADseq (Baird et al. 2008) were used to estimate the differentiation in nuclear DNA. The distribution of the divergent mitochondrial lineages was related to environmental data, including data on pollution.

Materials and methods

Sampling

The earthworms were sampled from the smelting and mining area in southern Poland in the vicinity of the zinc and lead smelter 'Bolesław' that is close to Olkusz along a well-studied metal pollution gradient (Azarbad et al. 2013; Giska et al. 2014). No specific permissions were required because *L. rubellus* is not a protected species and sampling was not done in a protected area. Sufficient numbers of *L. rubellus* were found at three sites with different levels of pollution: OL2 (50°17′44″ N, 19°29′27″ E), OL4 (50°19′05″ N, 19°30′32″ E), and OL5 (50°19′46″ N, 19°32′44″ E). The soil at these sites was primarily contaminated with Cd, Pb and Zn (Table 1). As a reference, we used the TR site (49°49′14″ N, 20°01′22″ E) in Trzemeśnia, which is also in southern Poland and approximately 65 km from the Olkusz area. We sampled only adult earthworms with a fully developed clitellum. The earthworms were collected alive with the use of horse dung traps installed on 15 x 15 m plots. The specimens were washed with distilled water, starved for 48 h and then preserved in 96% ethanol. Additional earthworms were collected to analyze the metal concentration in the tissues (Table 1). The genomic DNA was extracted from the anterior body segment tissues using the Wizard® Genomic DNA Purification Kit (Promega, Madison, USA).

Table 1. Characteristics of the sites at which *Lumbricus rubellus* was sampled. The distance from the smelter, soil pH, organic matter content at ~10 cm depth (OM%), and metal concentrations [mg kg^{-1} dwt.]: total concentrations in soil (normal font) and concentrations in earthworm tissue (italics), are shown; mean ± SD (soil: n = 3; earthworms: n = 3-6). Some of the data in the table were obtained from (Giska et al. 2014).

Site	Distance[km]	pH_{CaCl2}	OM [%]	Cd [mg kg^{-1}]	Pb [mg kg^{-1}]	Zn [mg kg^{-1}]
OL2	2.5	4.12 ± 0.03	53.5 ± 0.4	49.1 ± 1.1	2 060 ± 37	3 960 ± 54
				244 ± 120	*743 ± 234*	*3568 ± 1158*
OL4	5.3	3.46 ± 0.02	54.2 ± 2.0	14.8 ± 0.2	847 ± 38	966 ± 22
				80.7 ± 33.0	*209 ± 211*	*1672 ± 1602*
OL5	7.7	4.29 ± 0.01	36.3 ± 0.7	12.1 ± 0.7	708 ± 12	756 ± 11
				70.2 ± 35.7	*71.9 ± 75.4*	*1125 ± 653*
TR	~65	5.33 ± 0.04	13.0 ± 0.1	1.77 ± 0.295	65.4 ± 1.10	170 ± 17
				35.0 ± 17.6	*1.35 ± 1.50*	*300 ± 72.3*

Mitochondrial DNA

The fragments of the *COI* and the *ATP6* mitochondrial genes were sequenced. The primers were designed with the Primer3 software (Koressaar and Remm 2007; Untergrasser et al. 2012) based on the conservative fragments of the *L. rubellus* and *L. terrestris* mitochondrial genomes (Table A1). The PCR reactions contained ~50-150 ng of DNA template, 0.5 µM of each primer, 1X *Taq* buffer with $(NH_4)_2SO_4$, 1.5 mM of $MgCl_2$, 0.2 mM of each dNTP, and 0.75 U of *Taq* polymerase (Thermo Fisher Scientific, Waltham, USA) in a total volume of 15 µl; the reactions were performed under the conditions shown in Table A1. After the agarose gel visualization and the Exo-AP cleaning (Exonuclease I and Thermosensitive Alkaline Phosphatase; Thermo Fisher Scientific), the PCR products were sequenced using the BigDye® Terminator v3.1 Cycle Sequencing Kit, cleaned with Ethanol/EDTA precipitation and then analyzed on an ABI 3130xl Genetic Analyzer (Applied Biosystems). The raw sequences were aligned with the SeqScape® software (Applied Biosystems).

The *COI* sequences of *L. rubellus* that were sampled from across Europe by members of the Organisms and the Environment research group from the Cardiff School of Biosciences at the University of Cardiff were downloaded from GenBank [GenBank: KP642090-KP612109]. We selected unique sequences that represented the primary haplotype groups, and we used 16 sequences that originated from individuals sampled in 11 European countries (Austria, France, Germany, Holland, Hungary, Italy, Poland, Serbia, Spain, Sweden, and the UK). We also obtained the *COI* sequences from a few Polish individuals for which no RADseq data were generated, and these individuals originated from the OL3 site (located between OL2 and OL4; individuals PL1-PL4; 50°18’29” N, 19°29’45” E) and from central Poland (individual PL5; 51°32′44″ N, 21°11′22″ E).

RADseq

The genomic DNA was extracted from 25 individuals that were randomly selected from each population and was normalized to a concentration of 50 ng/µl using a Qubit® fluorometer. The RAD sequencing libraries were prepared according to the double-digest RADseq method described by Peterson et al. (2012). For each individual, 500 ng of genomic DNA were digested with SphI-HF and MseI restriction enzymes (NEB). After adapter ligation, the individual samples were pooled into four

libraries, purified and size selected with the LabChip XT (*LabChip XT DNA 300 Assay Kit; PerkinElmer, Waltham, USA*). We selected the 346–406 bp fraction to not exceed ~120,000 RAD tags per earthworm. The libraries were amplified in PCR reactions (20 µl) that contained 1X Phusion buffer, 200 µM of each dNTP, 1.0 µM each of PCR1 and PCR2 primers, 0.5 U of Phusion HF polymerase (Thermo Fisher Scientific) and 2.5 µl of the size-selected library under the following conditions: 98^{o}C for 30 s, followed by10 cycles at 98^{o}C for 10 s, at 62^{o}C for 30 s, and at 72^{o}C for 30 s, and with a final extension at 72^{o}C for 5 min. The size distribution of the amplified libraries was checked on a Bioanalyzer (HS DNA chips; Agilent Technologies). The libraries were sequenced on an Illumina HiSeq 2000 sequencer (single end, 100 bp) at the Center for Genome Research and Biocomputing of Oregon State University, USA (see Supplementary materials for details; Note A1).

The raw Illumina reads were analyzed with the *Stacks* software (Catchen et al. 2011; 2013). To avoid incorrect barcode-individual matches, we first removed all reads that had at least one barcode base with quality < 10 Phred. Subsequently, the reads were demultiplexed and cleaned with *process_radtags.pl* (Table A2), and the SphI recognition site sequence (CATGC) was removed (Note A1). The loci were reconstructed with *denovo_map.pl* with the following parameters: *-m 4, -M 4, -n 4* (see Table A3 for the *Stacks* commands). The *MySQL* database was used for graphical visualization and data filtering. For further analyses, we used the loci found in all four populations that were genotyped in at least 75% of the individuals of each population, had at least 5x coverage in each individual, and contained no more than 10 SNPs. The sequencing resulted in ~100,000 RAD tags per individual, with mean coverage 28 reads per RAD tag (Fig. A1). Numerous RAD tags were discarded because they were not present in the required % of individuals (Fig. A2).

Statistical analyses

The earthworms that were sampled from the different sites were assumed to represent local populations. To assess the population genetic diversity in the mtDNA, we estimated the haplotype diversity, nucleotide diversity and the number of polymorphic sites using DnaSP (Rozas 2009). The measures of population differentiation (mtDNA F_{ST}) were calculated based on haplotype frequencies with *Arlequin 3.5* (statistical significance was assessed with 10,100 permutations; (Excoffier and

Lischer 2010)). The relationships among the haplotypes were illustrated with a Median-Joining haplotype network that was constructed with *Network 4.6* (Bandelt et al. 1999). To show genetic differences between the identified haplotypes, we calculated the pairwise distances (p-distance and K2P distance) in *MEGA 6* (Tamura et al. 2013). To relate the haplotypes found in Poland (including haplotypes PL1-PL4 from the additional sites) to the mtDNA lineages observed across Europe, we reconstructed a phylogenetic tree of the Polish and the primary European haplotypes with the Maximum Likelihood (ML) method in *MEGA 6* (HKY + G model; 1000 bootstraps). The model was selected using the Bayesian Information criterion. Additionally, a Bayesian tree was constructed in *MrBayes* (5 mln generations; GTR + G model).

The RADseq data were analyzed with the *populations* program of *Stacks*. The population genetic statistics, such as the number of haplotypes, haplotype diversity and nucleotide diversity, were estimated. The pairwise differentiation between populations (RADseq F_{ST}) was estimated with Arlequin based on the SNP allele frequency (statistical significance was assessed with 10,100 permutations of individuals between the populations). The number of polymorphic sites that were shared between the populations (Ss) and unique to the individual populations (Sx) as well as the number of fixed differences between the populations (Sf) were calculated in *mstatspop* (Ramos-Onsins 2015). A Bayesian clustering method, implemented in the *Structure* software (Pritchard et al. 2000; Falush et al. 2003 and 2007; Hubisz et al. 2009), was used to examine the population structure with the estimation of the most likely number of genetically differentiated clusters and the fractions of the individual genotypes that were attributable to each cluster. For the *Structure* analysis, we used one randomly selected SNP per RAD tag (*--write_single_snp*), which resulted in 1101 loci. We tested K-values in the range of 1 to 8, with 20 replicates per value. The *Structure* analysis was run with 100,000 burn-in steps and 1,000,000 post-burn-in iterations per run. The admixture model and the correlated allele frequencies model were used. The lambda parameter was set to one (analysis with the estimated value of lambda, λ=0.4, resulted in the same number of clusters). The optimal K was selected based on the inspection of the change in the probability value of the data for a given K (L(K)), analyzed with Structure Harvester (Earl and Von Holdt 2012), assuming the largest value for a correct K. For an additional examination of the population structuring, all SNPs were used to calculate

a pairwise distance matrix between individuals (uncorrected p-distance) to construct a Neighbor-Joining phylogenetic tree in *MEGA 6*.

The isolation by distance was tested with a simple Mantel test. The effect of pollution on the degree of genetic differentiation between the populations was tested with a partial Mantel test while accounting for geographic distance (Smouse et al. 1986). The tests were performed with the use of the *IBDWS* ((Jensen et al. 2005); http://ibdws.sdsu.edu/; 10,000 randomizations). The following distance matrices were used in the study: pairwise F_{ST} values, log-transformed geographic distance (straight-line distance), and difference in Cd total soil concentration. To test the differences in the genome-wide genetic diversity between populations, we used t-tests with a strict Bonferroni correction for multiple comparisons.

Results

mtDNA

The two analyzed mtDNA fragments totaled 1016 bp (*COI*: 453 bp and *ATP6*: 563 bp). Among the 123 sequenced *L. rubellus* individuals originating from four populations (OL2, OL4, OL5 and TR), 276 polymorphic sites defined 27 unique haplotypes, 10 of which were observed only for one individual (Table 2). *ATP6* showed higher sequence diversity than *COI* (Supplementary materials: Tables A5, A6). The haplotypes formed five deeply divergent lineages (Fig. 1); four of the lineages had been previously described (A1, A2, A3, and E), and one lineage was new and named C2 (Fig. 2, Fig. A3). The sequence divergence between the C2 haplotypes and a haplotype from Serbia assigned to the C lineage was 7.5-8.5%, which led us to distinguish C2 as a separate lineage. An additional mtDNA lineage, which we named D2, was detected in five individuals from another geographic region, and this lineage had not been genotyped with nuclear markers (Fig. 2). The net sequence divergence between the observed lineages was substantial and ranged from 1.3% between the A2 and A3 lineages to 16% between the C2 and E lineages (Table 3).

More than one lineage was detected in each studied population, with four lineages co-occurring in the TR population, which also included the most divergent haplotypes (17.5% uncorrected sequence divergence; Table A7). The A lineages predominated in the Olkusz area but

were found in all studied populations. In contrast to the mtDNA lineages, most mtDNA haplotypes were unique, with only four haplotypes shared by at least two populations and no haplotypes shared by all populations (Fig. 1). Consequently, the mtDNA differentiation (F_{ST}) among all population pairs was substantial and highly significant (Table 4). The highest within-population variation, both for the haplotype (H_d = 0.855 ± 0.056; mean ± SD) and for the nucleotide (π = 0.096 ± 0.008; mean ± SD) diversity, was found in the TR population. Among the Olkusz populations, the mtDNA diversity increased with the level of pollution, and the OL2 population was the most diverse (Table 2).

Table 2. The mtDNA variation in *Lumbricus rubellus* populations. Shown results were based on concatenated mtDNA data (*COI*: 453 bp and *ATP6*:563 bp*).* Site - sampling site, N - number of analyzed individuals, S - number of polymorphic nucleotide positions, H - number of haplotypes, H_d - haplotype (gene) diversity (mean ± SD), π - nucleotide diversity (mean ± SD).

Site	N	S	H	H_d	π
OL2	31	174	8	0.815 ± 0.045	0.0277 ± 0.0072
OL4	31	71	7	0.811 ± 0.044	0.0152 ± 0.0029
OL5	31	62	4	0.578 ± 0.081	0.0167 ± 0.0033
TR	30	254	13	0.855 ± 0.056	0.0960 ± 0.0085
ALL	*123*	*276*	*27*	*0.923 ± 0.012*	*0.0580 ± 0.0065*

Table 3. Evolutionary divergence between mitochondrial lineages of *Lumbricus rubellus*. Mean pairwise sequence divergence - below diagonal; Net sequence divergence - above diagonal; p-distance.

	A1	A2	A3	C2	E
A1		0.032	0.028	0.118	0.157
A2	0.041		0.013	0.114	0.148
A3	0.039	0.023		0.113	0.150
C2	0.133	0.128	0.130		0.160
E	0.162	0.152	0.156	0.170	

Table 4. Pairwise genetic differentiation between *Lumbricus rubellus* populations. mtDNA F_{ST} based on haplotype frequency - above diagonal; RADseq F_{ST} based on SNP allele frequency - below diagonal. All values were significant (10,100 permutations; $p<0.05$, after strict Bonferroni corr.).

	OL2	OL4	OL5	TR
OL2	-	0.1518	0.2892	0.1595
OL4	0.1146	-	0.2457	0.1635
OL5	0.1436	0.1828	-	0.2839
TR	0.1278	0.1608	0.1808	-

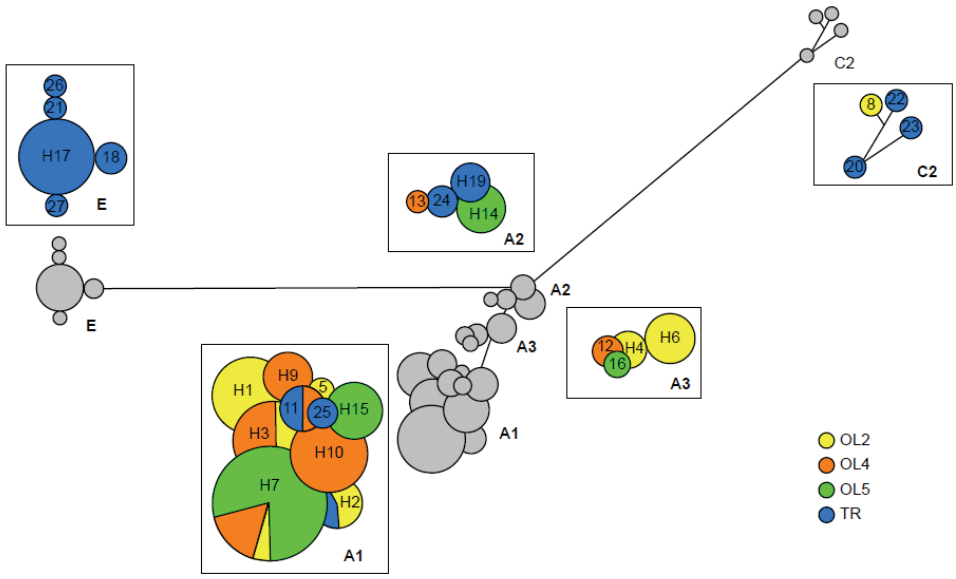

Fig. 1. Haplotype network of mtDNA (*COI* + *ATP6*) sequences of *Lumbricus rubellus*. The network shows the divergent mtDNA lineages (A1, A2, A3, C2, and E). Circles represent distinct haplotypes, which are marked with the labels H1-H27 in the enlarged insertions. The size of each circle is proportional to the total number of individuals that showed that haplotype, and the haplotype distributions within the populations are indicated as pie charts. The smallest circle corresponds to n = 1.

RADseq

The RADseq data were obtained for 100 individuals, 25 individuals per population. After stringent quality control in *Stacks*, our data set consisted of 1101 RADseq loci (~96,800 bp) that contained 5712 biallelic SNPs. The genetic diversity was highest in the TR population ($H = 4.28 \pm 0.07$; $H_d = 0.436 \pm 0.008$; $\pi = 0.0081 \pm 0.0002$; mean ± SE), which also showed the highest number of private polymorphisms ($Sx_{TR} = 1379$). Similar to the mtDNA results, the genetic diversity among the Olkusz populations increased with the level of pollution, and the genetic diversity was highest in population OL2 (Table 5). The RADseq-based differentiation between the earthworm populations was significant, and the pairwise F_{ST} ranged from 0.1146 to 0.1828 (Table 4). Although the populations were significantly differentiated in allele frequencies, the overwhelming majority of polymorphic positions were shared among localities, and fixed differences were not observed between the localities (Table 5).

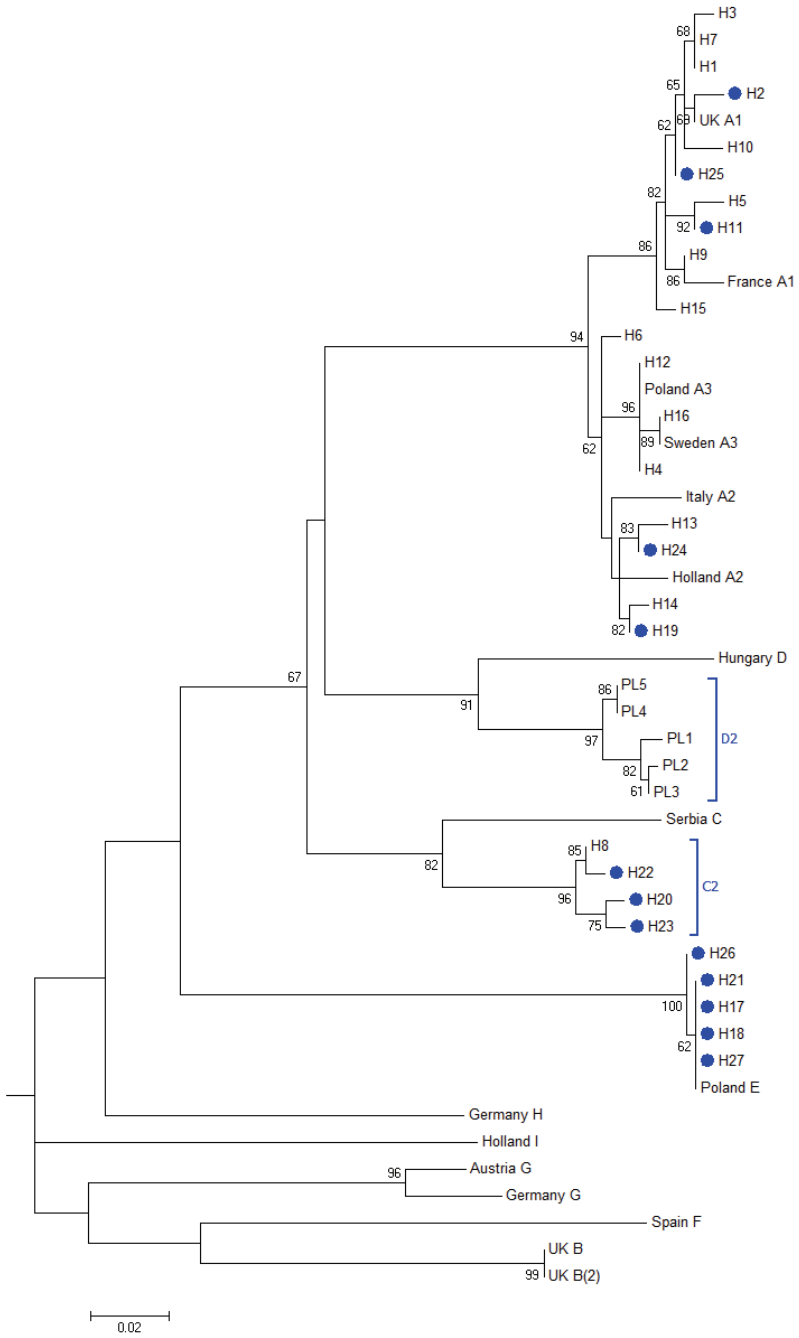

Fig. 2. Maximum-likelihood tree based on the *COI* sequences of *Lumbricus rubellus* collected across Europe. Haplotypes observed in the studied Polish populations are marked with the labels H1-H27. Blue circle labels represent the haplotypes found in the TR population. Blue brackets represent the new mtDNA lineages (C2 and D2). The *COI* sequence of *Hirudo medicinalis* [GenBank: EF446709.1] was used as the outgroup to root the tree. Bootstrap percentages ≥ 50 are shown at the branch points.

The Bayesian clustering identified a clear population structure, the individuals were grouped into four clusters according to their populations of origin with a low level of admixture observed mainly between the neighboring OL2 and OL4 sites (Fig. 3, Fig. A4). The four clusters revealed by *Structure* were recovered also in the Neighbor-Joining tree that was based on the genetic distances between individuals calculated from all 5712 RADseq SNPs (Fig. 4). We did not identify a clear pattern of isolation by distance or a correlation between the level of pollution and the genetic differentiation between the populations (Fig. A5, Table A8).

The clustering of the RADseq data according to the population of origin and not the mtDNA lineage indicated the lack of reproductive isolation between the mtDNA lineages. Although each population contained multiple mtDNA lineages, subdivisions within the populations were not observed in the nuclear genome, not even in the separate *Structure* analysis of the most diverse TR population.

Table 5. Polymorphism of *Lumbricus rubellus* populations estimated from the RADseq data (1101 RAD tags that contained 5712 SNPs). Site – sampling site, H - number of haplotypes (mean ± SE), H_d - haplotype (gene) diversity (mean ± SE), π - nucleotide diversity (mean ± SE); S – number of polymorphic sites, Sf – number of fixed differences, Sx – number of polymorphic sites unique for a population, Ss – number of polymorphic sites shared with other populations. Means with different letters are significantly different (t-test; $p < 0.05$, after strict Bonferroni correction).

Site	H	H_d	π	S	Sf	Sx	Ss
OL2	3.68 ± 0.06^{a}	0.427 ± 0.008^{a}	0.0081 ± 0.0002^{a}	3239	0	541	2696
OL4	3.23 ± 0.05^{b}	0.395 ± 0.008^{b}	0.0074 ± 0.0002^{ab}	2775	0	372	2403
OL5	2.69 ± 0.04^{c}	0.360 ± 0.008^{c}	0.0068 ± 0.0002^{b}	2284	0	310	1974
TR	4.28 ± 0.07^{d}	0.436 ± 0.008^{a}	0.0081 ± 0.0002^{a}	3816	0	1379	2437

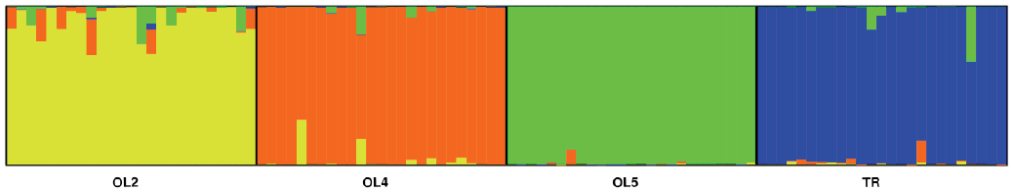

Fig. 3. Population genetic structure of *Lumbricus rubellus.* The graph shows the results of the *Structure* analysis of the RAD tags (single SNP selected from each of 1101 RAD tags). Each vertical bar represents a different individual from one of the four populations.

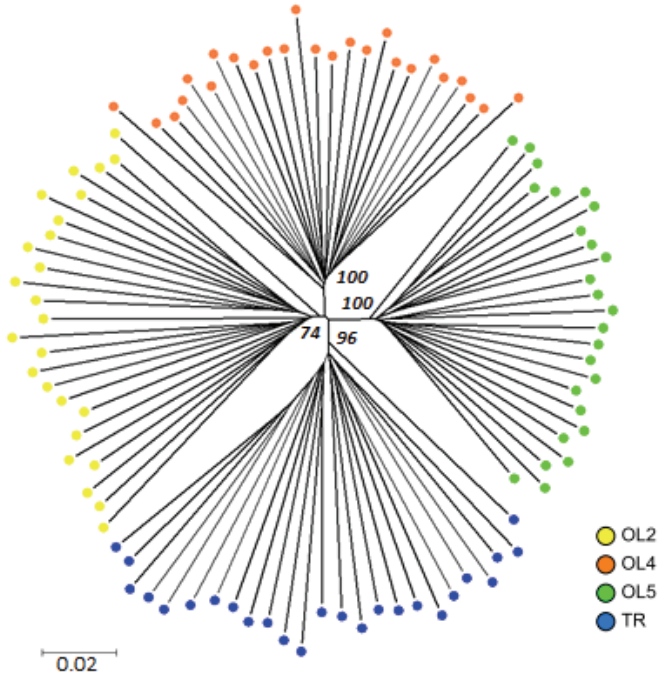

Fig. 4. Neighbor-Joining tree generated from the between-individual distance matrix (uncorrected p-distance) based on all 5712 SNPs from the RADseq data. Each dot represents an individual *Lumbricus rubellus* earthworm, and the color represents the population of origin. Bootstrap support values indicate grouping of individuals according to the population of origin.

Discussion

Our analyses of the *L. rubellus* individuals sampled from multiple sites in Poland revealed deeply divergent mtDNA lineages that occurred in sympatry. However, these divergent lineages were not reproductively isolated as evidenced by patterns of clustering in the nuclear data and therefore did not represent cryptic species. The situation we observed in Poland contrasts with that in the UK because the two main mtDNA lineages of *L. rubellus* in the UK, A and B, whose divergence was comparable to that observed in our study, were also differentiated at nuclear microsatellite markers (Donnelly et al. 2013). The morphological data further supported the hypothesis that the two British lineages represent cryptic species (Donnelly et al. 2014). Thus, the level of mtDNA divergence in *L. rubellus* within and between reproductively isolated lineages may be similar. This similarity is not surprising because the divergence at which reproductive isolation evolves varies extensively between and within taxonomic groups (Johns and Avise 1998; Meier et al. 2008). Additionally, Torres-Leguizamon et al. (2014) found different patterns of mitochondrial and nuclear structuring in another earthworm species that consist of two highly divergent (8.7%) mtDNA lineages, *Apporectodea icterica*. This finding suggested the random interbreeding of the mtDNA lineages. Because the nuclear markers of earthworms examined in the present study clustered according to the sample location, we theoretically could have sampled several microallopatric species that share mtDNA lineages. However, the genetic data did not support such a hypothesis. First, fixed differences were not detected among the localities in the nuclear genome; second, the overwhelming majority of polymorphisms were shared among the localities; and third, the signatures of genetic admixture between populations, particularly those separated by small distances in the Olkusz area, were detected. Because our data indicate no reproductive isolation between the lineages, a question arises which processes and mechanisms might explain the presence of highly divergent mtDNA lineages of *L. rubellus* in sympatry? In the following sections, we discuss several plausible, although not mutually exclusive, hypotheses.

The highly divergent mtDNA lineages in Poland might be the result of admixture that followed the postglacial recolonization among previously geographically but not reproductively isolated lineages derived from separate glacial refugia. Such scenario was suggested for the British *L. rubellus* (Donnelly et al. 2013) as well as other earthworm species (King et al. 2008; Torres-

Leguizamon et al. 2014). Recent studies (Sechi 2013; Vega et al. 2010) note that cryptic refugium may have occurred on the northwestern coasts of Europe, where one or more of the *L. rubellus* mtDNA lineages could have survived the periods of unfavorable climate during the Pleistocene. The area of present-day Poland may have been colonized from refugia located in central, southern and eastern Europe (Schmitt 2007; Schmitt and Varga 2012), and such a pattern appears common because zones of contact between divergent evolutionary lineages have been described for multiple taxa in Poland (Babik et al. 2005; Durka et al. 2005; Gratton et al. 2008; Wójcik et al. 2010). However, the large genetic distances among the mtDNA lineages of *L. rubellus* indicated that their origin predated the Last Glacial Maximum. Nevertheless, multiple cycles of changes in the ranges of the earthworms during the Pleistocene may have caused major changes in the distribution of the ancient mtDNA lineages, which resulted in their sorting into separate refugia. The changes in range may also have prevented the accumulation of reproductive isolation mechanisms between geographically separated populations because of the opportunities for multiple contacts and genetic exchange (Hofreiter et al. 2004; Hewitt 2011). The sampling of earthworms in potential refugial areas would provide a direct test of the multiple refugia hypothesis, and a single lineage in a particular area would suggest that the situation observed in Poland resulted from a postglacial admixture. Moreover, admixture might also have contributed to the high nuclear polymorphism detected in our study.

The highly divergent sympatric mtDNA lineages might also be a simple consequence of a large effective population size (*N*e); the estimates of the *N*e may be inflated when coupled with population subdivision and low migration rates (Charlesworth et al. 2003). Earthworms are generally considered highly polymorphic organisms that are characterized by a large *N*e and very low migration rates (Marinissen and Van den Bosch 1992). High intraspecific mtDNA divergence, which often exceeds 5% in *COI*, is not limited to *L. rubellus* and has been reported in numerous earthworm species, including *Allolobophora chlorotica*, *Aporrectodea rosea*, *Octolasion lacteum*, *Dendrobaena octaedra*, *L. castaneus*, and *L. terrestris*, e.g., (King et al. 2008; James et al. 2010; Klarica et al. 2012). Earthworms appear to also be highly polymorphic in the nuclear genome, although most available estimates are based on microsatellites (Velavan et al. 2009; Novo et al. 2010; Dupont et al. 2011; Donnelly et al. 2013) and are therefore difficult to compare among taxa. However, the genome-wide

synonymous nucleotide diversity in *Allolobophora chlorotica* exceeds 1% (Romiguier et al. 2014), which is comparable with the values obtained in our study (0.7-0.8%). These values might underestimate the true diversity because many polymorphic RADseq loci were filtered out due to the high incidence of missing data. This could result from mutations in the restriction recognition sites but could be also simply due to the random loss of RAD tags during the preparation of RADseq library and random variation in coverage depth. The multiple divergent mtDNA lineages caused by long genealogies in a large population and a high mtDNA mutation rate might be particularly plausible in *L. rubellus*. This species is an obligate cross-fertilizing hermaphrodite: each individual passes its mtDNA to its progeny, which increases the ratio of nuclear to mitochondrial *N*e. For a large *N*e, even colonization from a single refugium could explain our results. Additionally, this hypothesis can be tested by directly sampling populations in multiple putative refugial areas. The comparison of the nuclear diversity between the recently recolonized areas and the refugial areas should indicate whether the colonization was accompanied by a reduction in genetic diversity, as postulated by many models of range expansion (Excoffier et al. 2009).

The evaluation of the two hypotheses presented above would require additional sampling and data on genome wide variation and differentiation among populations. However, the RADseq markers are less than ideal for such purposes because a large fraction of the loci are not usable due to the high frequency of missing data. This problem has been previously recognized as a serious issue in highly polymorphic species (Arnold et al. 2013; Gautier et al. 2013). Therefore, alternative approaches to estimate nucleotide variation could focus either on the protein-coding genes that harbor extensive synonymous variation and can be analyzed using various targeted resequencing methods (Mamanova et al. 2010) or on the ultraconserved (Faircloth et al. 2012) or conserved elements (Lemmon et al. 2012), which also capture more polymorphic flanking regions.

The high intraspecific mtDNA divergence may be due to an introgression from a related species, as commonly observed in animals (Toews and Brelsford 2012). However, the mtDNA sequences of *L. rubellus* found in our study were highly divergent from the sequences of the related species, *L. castaneus* and *L. terrestris*, that were available in GenBank, which eliminated the

possibility of introgression from currently known *Lumbricus* lineages. Nevertheless, introgression from an undescribed or extinct lineage remains a viable option.

The multiple divergent mtDNA lineages might also be maintained by natural selection, particularly selection acting in a highly heterogeneous environment like soil. A recent study of Kozancioğlu and Arnqvist (2014) suggested that negative frequency-dependent selection (NFDS), a form of balancing selection that favors rare variants, could maintain mtDNA polymorphism (Fijarczyk and Babik 2015). Kozancioğlu and Arnqvist (2014) showed an increase in the rare mtDNA haplotype frequency and a decrease in the common haplotype frequency in experimental populations of the seed beetle (*Callosobruchus*) over the course of 10 generations. NFDS is expected under conditions with environmental heterogeneity, genotype-by-environment interactions and competition for resources, which are conditions likely to be common in earthworm populations.

Soil contamination may also affect the distribution of genetic lineages in nature. If the degrees of sensitivity to soil pollution differ among mtDNA lineages, some lineages will be lost in polluted areas, which reduces variation and is consistent with the genetic erosion hypothesis (Van Straalen and Timmermans 2002). For example, Andre et al. (2010) investigated the highly differentiated populations of *L. rubellus* from a Pb-polluted habitat near Cwmystwyth in Wales, UK. The predominant linage differed by study site depending on the level of contamination, and this pattern supported the loss of distinct mtDNA lineages due to pollution. In our research, four mtDNA lineages occurred in the least polluted TR site. In contrast, the E lineage was not found at any of the polluted Olkusz sites. However, either experimental manipulations or field data from multiple pollution gradients are necessary to demonstrate that individuals carrying the mtDNA of the E lineage are more sensitive to metal pollution. On the other hand, among the three contaminated OL sites, the most polluted site OL2 was characterized by relatively high haplotype and nucleotide diversity and the largest number of the polymorphic sites and private SNPs; therefore, this result did not support the genetic erosion hypothesis and was consistent with a pattern we identified also for the rove beetle *Staphylinus erythropterus*, inhabiting the same gradient (Giska et al. 2015).

In ecotoxicology, earthworms are used for standard toxicity tests. The recommended and most commonly used species are *Eisenia fetida* and *Eisenia andrei* (e.g., (OECD, ISO)). However, the

taxonomy of these species is not clear because of cryptic diversity: The earthworm *E. fetida* has been suggested to be a species complex. Römbke et al. (2015) reported two distinct mtDNA *COI* clusters of *E. fetida* that were separated by a p-distance of 11.2%. Based on the assumption that an uncorrected p-distance > 10% indicates species level differentiation, these authors hypothesized that *E. fetida* consisted of cryptic species; this result calls the quality and the comparability of ecotoxicological tests into question because cultures of *Eisenia* earthworms are rarely barcoded (Römbke et al. 2015). Nuclear markers were not applied to confirm the mtDNA clustering of the *E. fetida* reported by Römbke et al. (2015), although previous analysis of nuclear *28S* gene indicated possibility that *E. fetida* from Ireland might be a cryptic species (Pérez-Losada et al. 2009). Therefore, the findings of our study are particularly relevant because we showed that high mtDNA divergence, even values exceeding 15%, did not necessarily indicate the presence of cryptic earthworm species. Thus, in addition to crossbreeding experiments, we recommend the use of multilocus nuclear data to test for cryptic species in *E. fetida*.

The results of our study have consequences for the estimation of the diversity of soil fauna and the delimitation of species. As emphasized by Emerson et al. (2011), the soil mesofauna is more diverse than previously thought, and describing the cryptic diversity of soil remains a challenge for ecologists. Thus, the choice of the proper molecular techniques is of crucial importance. The identification of cryptic diversity in the previously mentioned soil invertebrates was often based on a small number of loci (Schäffer et al. 2010; Spelda et al. 2011; Porco et al. 2012; Cicconardi et al. 2013), although the accuracy of species delimitation is known to depend on the number of loci sampled (Knowles and Carstens 2007). Although the cryptic diversity of earthworms has long been a matter of debate, the conclusions about cryptic species remain primarily based on mitochondrial data, which is sometimes complemented with a single nuclear gene or morphological traits (e.g., Pérez-Losada et al. 2005; Pérez-Losada et al. 2009; James et al. 2010). In this study, the NGS methods were used to our advantage and showed that genome-wide data supplied valuable knowledge for the study of cryptic diversity.

Conclusions

The highly divergent mtDNA lineages of the earthworm *Lumbricus rubellus* that sympatrically co-occurred in multiple localities in Poland did not constitute reproductively isolated groups. We concluded that *L. rubellus*, which is represented by several mtDNA lineages in continental Europe, is a single highly polymorphic species rather than a complex of several cryptic species. This study demonstrated the critical importance of multilocus nuclear data for the unbiased assessment of cryptic diversity and species delimitation in soil invertebrates.

Availability of supporting data

The mtDNA sequences of *L. rubellus* generated in this study are available on GenBank under accession numbers: KT731474 – KT731500 (*COI*), and KT731501 – KT731525 (*ATP6*). The RADseq data (*denovo_map.pl* output tsv files) are available in the Dryad Digital Repository at the http://dx.doi.org/10.5061/dryad.8070m. Raw Illumina reads are available at NCBI BioProject number PRJNA296755 (http://www.ncbi.nlm.nih.gov/bioproject/296755) or upon request to the corresponding author.

Acknowledgements

We thank Edyta Podmokła and Sebastian Żmudzki for their help with sampling earthworms and Barbara Kowalczyk for her contribution to *COI* sequencing. This study was supported by the Polish National Science Center Grant No. 2011/03/N/NZ8/00013 and the Foundation for Polish Science International PhD Projects Programme co-financed by the EU European Regional Development Fund in the frame of the "Environmental Stress, Population Viability and Adaptation" project; MPD/2009-3/5. Support from Jagiellonian University in Kraków, DS/MND/WBiNoZ/INoŚ/11/2014, is also acknowledged.

Appendix 2

Supplementary materials to Chapter 2

Table A1. Information on mtDNA sequence markers of *Lumbricus rubellus* and conditions of the PCR amplification; the gene, length of the sequenced product after alignment trimming [bp], and PCR primer sequences are shown. The PCR cycle scheme used for sequencing the mitochondrial genes of *Lumbricus rubellus* is presented.

gene	length [bp]	primer sequence	PCR profile
COI	453	F1: CCGAATCGAACTAAGrCAAC	95 ^{0}C – 3 min
		F2: GGTCAACAAATCATAAAGATATTGG*	30 cycles:
		R: TCAGAAGAGGTGTTGGTAkAGGA	95 ^{0}C – 30 s
			55 ^{0}C – 30 s
ATP6	563	F: GAGTATCCAAGTCTTGCCATGAT	72 ^{0}C – 1 min
		R: TGkGCGTGrTCrTCTGAGTAT	72 ^{0}C – 10 min

* Folmer universal primer used for some TR individuals for which the F1 primer did not work.

Note A1. Detailed description of the Illumina sequencing and *Stack*s analysis of RAD tags

A single library run on one HiSeq 2000 lane included 25 individuals that originated from two populations (13 individuals from one population and 12 individuals from another). The populations were distinguished by two indices (6 bp); whereas the individuals were distinguished by 13 different barcodes (5 bp). Because a RAD library is a low-diversity library, the sequencing was performed at a relatively low cluster density (~700 K/mm^2), with a dedicated PhiX lane and a sample PhiX spike in (~15%). In total, the sequencing yielded in 764.4 million (M) reads of *L. rubellus* (Table A2).

The raw Illumina reads were analyzed with the *Stacks* software (Table A3). We discarded all reads that had at least one barcode base with quality < 10 (in the case of one sample, the quality level was increased to 15). This filtering step was performed to ensure the removal of low-quality barcodes because the quality scores were not checked during the *Stacks* analysis. With this filtering, ~23% of the raw reads were discarded. The remaining reads were cleaned and demultiplexed with the *process_radtags.pl* program. Only the reads with the correct barcode and a high sequence quality were used. We applied the following filters: *-w 0.15, -s 10* (default sliding window), *--filter_illumina* (to discard reads failing the Illumina chastity filter), *-c* (to discard reads that contained uncalled bases), *-t 93* (to trim the last three nucleotides), and *-r* (to rescue RAD tags with one sequencing error in the

restriction enzyme overhang). After filtering with *process_radtags.pl* we removed the SphI recognition site sequence (CATGC) from all reads, which resulted in a final read length of 88 bp. For each individual, the loci were reconstructed *de novo* with the following parameters of the *denovo_map.pl* program: *-m 4* (required at least four identical reads to form a stack), *-M 4* (allowed a four nucleotide distance between stacks), *-N M+0* (no secondary reads), *-t* (removed highly repetitive RAD tags), *--max_locus_stacks 3* (the maximum number of stacks allowed at a single locus was set to three), *-d* (enabled deleveraging algorithm), and *-n 4* (allowed four mismatches between catalog loci when constructing the catalog). The bounded-error implementation of the maximum-likelihood SNP calling model was used with the upper bound of the error set to a value of 0.05. The genotype likelihood ratio test critical value in the SNP calling model was set to 0.1 (results of different α values were compared; Table A4). This implementation resulted in a set of RAD tags with a mean coverage equal to ~28 reads per RAD tag (Figure A1). From this set, we used only the RAD tags that contained no more than 10 SNPs, which were filtered from the *MySQL* database. These parameters were selected after testing various values and accounting for the high polymorphism within *L. rubellus.* For further analyses, we used loci that had at least 5x coverage for an individual, were found in all four populations and genotyped in at least 75% of the individuals of each population (Table A3). We observed a substantial decrease in the usable RAD tags when the *r* parameter was increased (Fig. A2).

Table A2. Quality control of Illumina reads from the HiSeq 2000. Raw reads obtained from the Illumina platform, reads retained after filtering sequences with at least one barcode base with a QV < 10 (in the case of samples OL2/II and OL5/II, the quality was increased to 15), reads filtered by *process_radtags.pl* (ambiguous barcodes, failed chastity filter, ambiguous RAD tag, and low QV reads) and reads used for the final analyses (retained reads) are included. Samples marked with the same letter (A, B, C, and D) were pooled and sequenced on one HiSeq lane.

Parameter\Sample	OL2/I [A]	OL2/II [B]	OL4/I [C]	OL4/II [D]	OL5/I [C]	OL5/II [B]	TR/I [A]	TR/II [D]
index	GCCAAT	CTTGTA	CTTGTA	GCCAAT	GCCAAT	GCCAAT	CTTGTA	CTTGTA
raw reads	76,507,384	154,737,876	69,365,166	75,925,083	84,589,867	171,315,380	65,816,468	66,159,114
barcode QV	67,862,301	87,914,574	65,109,910	72,942,177	79,282,359	96,230,896	58,595,180	63,643,136
ambiguous barcodes	26,469,975 (39.0%)	28,530,783 (32.5%)	17,082,714 (26.2%)	17,187,901 (23.6%)	20,770,618 (26.2%)	31,385,971 (32.6%)	22,489,101 (38.4%)	14,658,013 (23.0%)
failed chastity filter	3,998,262	8,318,041	3,928,411	4,024,905	4,967,865	9,340,745	3,430,429	3,467,755
ambiguous RAD tag	209,655	154,546	214,674	190,451	212,073	216,087	148,705	202,056
low QV reads	2,081,921	1,990,208	2,436,512	2,737,386	2,970,886	2,229,686	1,796,958	2,363,596
retained reads	35,102,488 (45.9%)	48,920,996 (31.6%)	41,447,599 (59.8%)	48,801,534 (64.3%)	50,360,917 (59.5%)	53,058,407 (31.0%)	30,729,987 (46.7%)	42,951,716 (64.9%)

Table A3. The *Stacks* commands used to process the RADseq data.

program	command
process_radtags.pl	`process_radtags -p … -b … -o … -c -q -r -t 93 \` `-e sphI -i fastq --barcode_dist 1 --filter_illumina`

denovo_map.pl	`denovo_map.pl -T … -m 4 -M 4 -n 4 -N M+0 -t -H -B … \`
	`-b … -X "ustacks:-d" \`
	`-X "ustacks:--max_locus_stacks 3" \`
	`-X "ustacks:--model_type bounded" \`
	`-X "ustacks:--bound_high 0.05" \`
	`-X "ustacks:--alpha 0.1"`

export_sql.pl	`export_sql.pl -D … -b … -f … -o tsv -F snps_u=10`
populations	`populations -b … -P … -M … -r 0.75 -p 4 -m 5 -W …`

Table A4. Comparison of *Stacks* results for different analysis parameters. alpha = the genotype likelihood ratio test critical P value of the SNP calling model (in *denovo_map.*pl), m = minimum stack depth required for individuals at a locus (in *populations*), S = polymorphic sites, π = nucleotide polymorphism (SE).

alpha	m	# loci	S				π			
			OL2	OL4	OL5	TR	OL2	OL4	OL5	TR
0.1	5	1101	3239	2775	2284	3816	0.0081 (0.0002)	0.0074 (0.0002)	0.0068 (0.0002)	0.0081 (0.0002)
	10	723	2110	1803	1493	2524	0.0082 (0.0002)	0.0074 (0.0002)	0.0070 (0.0002)	0.0082 (0.0002)
0.05	5	1103	3238	2774	2287	3815	0.0081 (0.0002)	0.0074 (0.0002)	0.0068 (0.0002)	0.0081 (0.0002)
	10	723	2106	1801	1494	2521	0.0082 (0.0002)	0.0074 (0.0002)	0.0070 (0.0002)	0.0082 (0.0002)
0.01	5	1103	3229	2762	2279	3802	0.0081 (0.0002)	0.0074 (0.0002)	0.0068 (0.0002)	0.0081 (0.0002)
	10	724	2102	1796	1492	2518	0.0081 (0.0002)	0.0074 (0.0002)	0.0070 (0.0002)	0.0081 (0.0002)

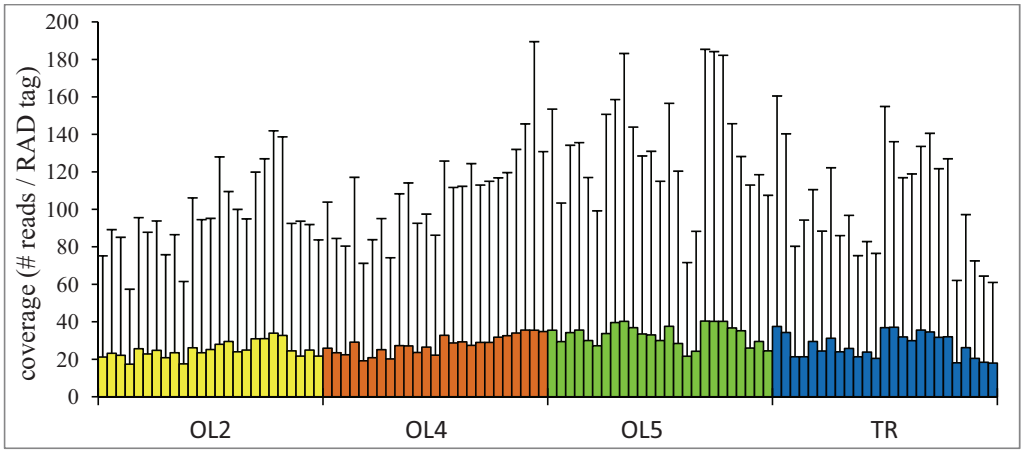

Fig. A1. Final coverage per RAD tag (mean ± SD) for individual earthworms of *Lumbricus rubellus*. The coverage was calculated after merging stacks in *denovo_map.pl*. The individuals are represented by single bars, and the colors represent the populations.

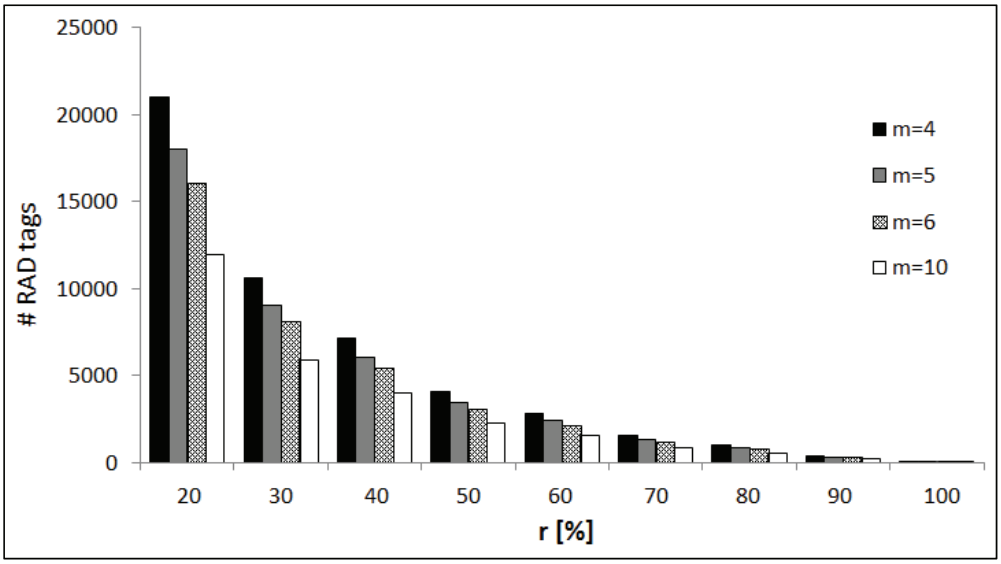

Fig. A2. Effect of the *Stacks* analysis parameters on the final number of usable RAD tags in *Lumbricus rubellus*. r = minimum percentage of individuals in a population required to process a locus for that population, and m = minimum stack depth required for individuals at a locus.

Table A5. Variation at *COI* (453 bp) in *Lumbricus rubellus* populations. Site - sampling site, S - number of polymorphic nucleotide positions, H - number of haplotypes, H_d - haplotype (gene) diversity (mean ± SD), and π - nucleotide diversity (mean ± SD).

Site	S	H	H_d	π
OL2	67	7	0.791 ± 0.049	0.02514 ± 0.00589
OL4	33	7	0.811 ± 0.044	0.01808 ± 0.00270
OL5	25	4	0.578 ± 0.081	0.01477 ± 0.00268
TR	97	10	0.743 ± 0.081	0.08347 ± 0.00715
ALL	*107*	*22*	*0.880 ± 0.019*	*0.05261 ± 0.00566*

Table A6. Variation at *ATP6* (563 bp) in *Lumbricus rubellus* populations. Site - sampling site, N - number of analyzed individuals, S - number of polymorphic nucleotide positions, H - number of haplotypes, H_d - haplotype (gene) diversity (mean ± SD), and π - nucleotide diversity (mean ± SD).

Site	S	H	H_d	π
OL2	107	8	0.815 ± 0.045	0.02976 ± 0.00834
OL4	38	7	0.811 ± 0.044	0.01291 ± 0.00307
OL5	37	4	0.578 ± 0.081	0.01821 ± 0.00386
TR	157	12	0.853 ± 0.056	0.10609 ± 0.00956
ALL	*169*	*25*	*0.923 ± 0.011*	*0.06227 ± 0.00726*

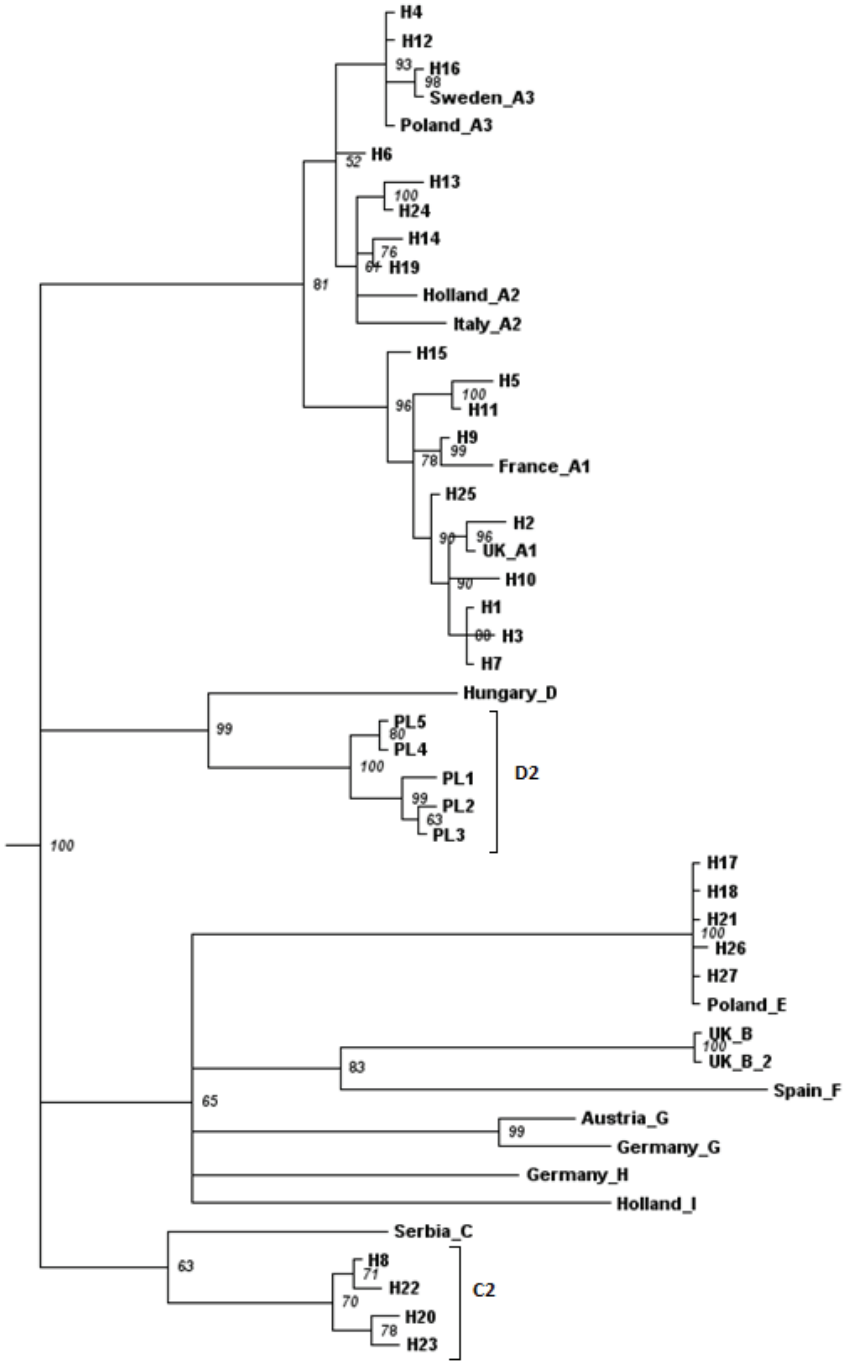

Fig. A3. Bayesian tree based on the *COI* sequences of *Lumbricus rubellus*. The posterior probabilities ≥ 50% are shown for each node. Haplotypes observed in the studied populations are marked with the labels H1-H27. The *COI* sequence of *Hirudo medicinalis* (GenBank EF446709.1) was used to root the tree.

Table A7. Pairwise genetic distances between mtDNA haplotypes H1-H27 found in Poland. Uncorrected p-distance below the diagonal (black), and K2P distance above the diagonal (blue); the distances were computed with *MEGA6* based on concatenated *COI* and *ATP6* sequences of *Lumbricus rubellus*.

	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16	H17	H18	H19	H20	H21	H22	H23	H24	H25	H26	H27
H1		0.012	0.005	0.040	0.022	0.035	0.000	0.128	0.012	0.012	0.015	0.040	0.046	0.038	0.015	0.046	0.183	0.183	0.032	0.132	0.183	0.135	0.142	0.040	0.005	0.187	0.183
H2	0.014		0.017	0.048	0.030	0.043	0.012	0.136	0.020	0.020	0.022	0.048	0.051	0.046	0.022	0.054	0.184	0.184	0.040	0.139	0.184	0.143	0.150	0.048	0.012	0.188	0.184
H3	0.007	0.013		0.046	0.027	0.041	0.005	0.135	0.017	0.017	0.019	0.046	0.051	0.043	0.020	0.051	0.190	0.190	0.038	0.138	0.190	0.142	0.149	0.046	0.010	0.195	0.190
H4	0.040	0.046	0.041		0.040	0.015	0.040	0.128	0.038	0.043	0.037	0.000	0.027	0.022	0.035	0.005	0.166	0.166	0.017	0.124	0.166	0.135	0.135	0.019	0.035	0.169	0.166
H5	0.020	0.024	0.019	0.045		0.043	0.022	0.127	0.019	0.030	0.007	0.040	0.048	0.046	0.022	0.045	0.177	0.177	0.040	0.130	0.177	0.134	0.141	0.043	0.017	0.181	0.177
H6	0.036	0.042	0.037	0.022	0.044		0.035	0.129	0.033	0.043	0.035	0.015	0.022	0.017	0.030	0.019	0.163	0.163	0.012	0.125	0.163	0.136	0.136	0.015	0.030	0.167	0.163
H7	0.001	0.013	0.006	0.039	0.019	0.035		0.128	0.012	0.012	0.015	0.040	0.046	0.038	0.015	0.046	0.183	0.183	0.032	0.132	0.183	0.135	0.142	0.040	0.005	0.187	0.183
H8	0.136	0.132	0.136	0.132	0.137	0.127	0.135		0.125	0.132	0.124	0.128	0.128	0.118	0.115	0.125	0.191	0.191	0.115	0.015	0.191	0.005	0.015	0.125	0.121	0.191	0.191
H9	0.009	0.013	0.008	0.037	0.015	0.033	0.008	0.132		0.020	0.012	0.038	0.043	0.035	0.012	0.043	0.171	0.171	0.030	0.128	0.171	0.132	0.138	0.038	0.007	0.175	0.171
H10	0.015	0.019	0.014	0.043	0.025	0.040	0.014	0.132	0.014		0.022	0.043	0.054	0.046	0.022	0.048	0.195	0.195	0.041	0.135	0.195	0.139	0.146	0.049	0.012	0.200	0.195
H11	0.013	0.017	0.012	0.038	0.009	0.035	0.012	0.133	0.008	0.018		0.037	0.040	0.038	0.015	0.043	0.177	0.177	0.032	0.127	0.177	0.131	0.137	0.035	0.010	0.181	0.177
H12	0.041	0.047	0.042	0.001	0.046	0.023	0.040	0.132	0.038	0.044	0.039		0.027	0.022	0.035	0.005	0.166	0.166	0.017	0.124	0.166	0.135	0.135	0.019	0.035	0.169	0.166
H13	0.040	0.045	0.041	0.029	0.046	0.019	0.039	0.130	0.037	0.044	0.037	0.030		0.020	0.035	0.032	0.167	0.167	0.015	0.124	0.167	0.135	0.135	0.007	0.040	0.171	0.167
H14	0.041	0.047	0.042	0.029	0.049	0.019	0.040	0.127	0.038	0.045	0.040	0.030	0.018		0.027	0.027	0.159	0.159	0.005	0.115	0.159	0.125	0.124	0.012	0.032	0.163	0.159
H15	0.015	0.019	0.014	0.041	0.021	0.037	0.014	0.130	0.010	0.020	0.014	0.042	0.039	0.039		0.040	0.159	0.159	0.022	0.118	0.159	0.121	0.128	0.030	0.010	0.163	0.159
H16	0.041	0.047	0.042	0.005	0.046	0.023	0.040	0.130	0.038	0.044	0.039	0.006	0.030	0.030	0.042		0.166	0.166	0.022	0.121	0.166	0.132	0.131	0.024	0.040	0.169	0.166
H17	0.162	0.164	0.164	0.159	0.164	0.153	0.161	0.166	0.158	0.162	0.161	0.159	0.154	0.153	0.156	0.158		0.000	0.156	0.191	0.000	0.199	0.194	0.163	0.175	0.002	0.000
H18	0.161	0.163	0.163	0.158	0.163	0.152	0.160	0.165	0.157	0.161	0.160	0.158	0.153	0.152	0.155	0.157	0.001		0.156	0.191	0.000	0.199	0.194	0.163	0.175	0.002	0.000
H19	0.039	0.045	0.040	0.027	0.047	0.017	0.038	0.126	0.036	0.043	0.038	0.028	0.016	0.004	0.037	0.028	0.151	0.150		0.111	0.156	0.121	0.121	0.007	0.027	0.159	0.156
H20	0.134	0.130	0.134	0.130	0.135	0.125	0.133	0.022	0.130	0.132	0.131	0.130	0.128	0.123	0.129	0.128	0.170	0.169	0.124		0.191	0.020	0.010	0.121	0.124	0.191	0.191
H21	0.163	0.165	0.165	0.160	0.165	0.154	0.162	0.167	0.159	0.163	0.162	0.160	0.155	0.154	0.156	0.159	0.001	0.002	0.152	0.171		0.199	0.194	0.163	0.175	0.002	0.000
H22	0.134	0.130	0.134	0.134	0.135	0.131	0.133	0.010	0.130	0.132	0.133	0.134	0.134	0.131	0.128	0.132	0.171	0.170	0.130	0.022	0.172		0.019	0.132	0.128	0.199	0.199
H23	0.135	0.131	0.135	0.131	0.134	0.128	0.134	0.023	0.131	0.131	0.134	0.131	0.131	0.128	0.129	0.129	0.173	0.172	0.129	0.021	0.174	0.023		0.131	0.135	0.194	0.194
H24	0.038	0.044	0.039	0.026	0.044	0.016	0.037	0.129	0.035	0.042	0.035	0.027	0.003	0.015	0.037	0.027	0.153	0.152	0.013	0.127	0.154	0.133	0.130		0.035	0.167	0.163
H25	0.010	0.014	0.009	0.038	0.018	0.034	0.009	0.132	0.007	0.012	0.011	0.039	0.038	0.039	0.013	0.039	0.158	0.157	0.037	0.132	0.159	0.132	0.131	0.036		0.179	0.175
H26	0.164	0.166	0.166	0.161	0.166	0.155	0.163	0.168	0.160	0.165	0.163	0.161	0.156	0.155	0.157	0.160	0.003	0.004	0.153	0.172	0.002	0.173	**0.175**	0.155	0.160		0.002
H27	0.163	0.165	0.165	0.160	0.165	0.154	0.162	0.167	0.159	0.163	0.162	0.160	0.155	0.154	0.156	0.159	0.001	0.002	0.152	0.171	0.002	0.172	0.174	0.154	0.159	0.004	

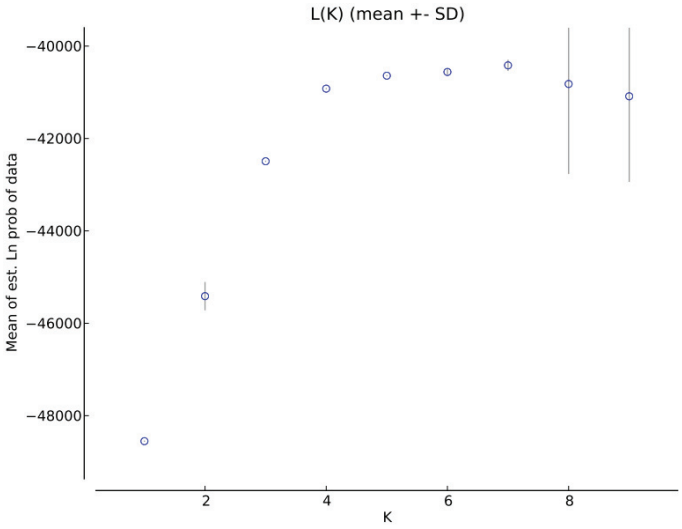

Fig. A4. The Ln P(D) in *Structure* analysis of *Lumbricus rubellus*.

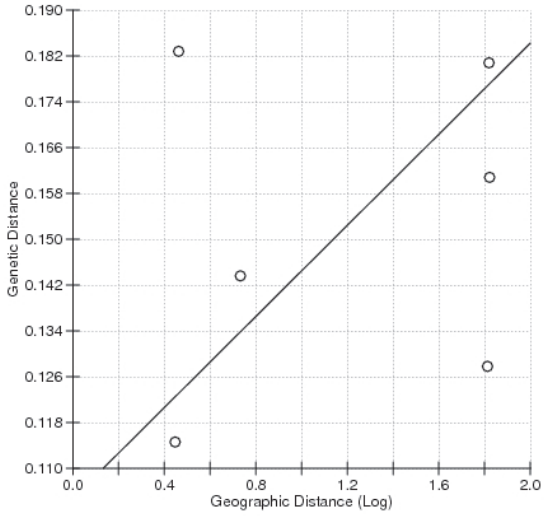

Fig. A5. Relation between the genetic distance (RADseq F_{ST}) and the log(geographic distance) in *Lumbricus rubellus* populations; a reduced major axis regression based on the Mantel test.

Table A8. Mantel test and partial Mantel test statistics for *Lumbricus rubellus* populations sampled at sites with different levels of metal pollution.

	RADseq F_{ST}
Correlation of genetics and log (geographic distance)	Z = 1.09, r = 0.181, p = 0.363
Correlation of genetics and contamination (indicator) matrix	Z = 19.7, r = -0.887, p = 0.886
Partial corr. of genetics and log (geographic distance), controlling for indicator matrix	r = 0.419, p = 0.324
Partial corr. of genetics and indicator matrix, controlling for geography	r = -0.905, p = 0.839