An evolutionary history of the peregrine epigeic earthworm *Lumbricus rubellus*

Pierfrancesco Sechi

Thesis submitted to Cardiff University in candidature for the degree of Doctor of Philosophy

September 2013

Cardiff School of Biosciences

Cardiff University

DECLARATION

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ..... (candidate) Date .....

STATEMENT 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed ..... (candidate) Date .....

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed ..... (candidate) Date .....

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter- library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate) Date .....

STATEMENT 4 – BAR ON ACCESS APPROVED

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter- library loans after expiry of a bar on access previously approved by the Graduate Development Committee.

Signed ..... (candidate) Date .....

An evolutionary history of the peregrine epigeic earthworm *Lumbricus rubellus*

Pierfrancesco Sechi*

Supervisors:

Prof. Michael W. Bruford
Organisms and the Environment research group, Cardiff School of Biosciences, Cardiff University, UK
Address: The Sir Martin Evans Building, Museum Avenue, Cardiff, CF10 3AX

Prof. Peter Kille
Organisms and the Environment research group, Cardiff School of Biosciences, Cardiff University, UK
Address: Main building, Museum Avenue, Cardiff, Cf10 3AT

This work was financed by *Agenzia regionale per il lavoro, Regione Sardegna* (Italy) with a PhD grant relative to the Master and Back program – Alta Formazione (High Education), ref. n. 9847.

**Email for correspondence: pierfrancesco@gmail.com*

"Still round the corner there may wait
A new road or a secret gate,
And though we pass them by today,
Tomorrow we may come this way
And take the hidden paths that run
Towards the Moon or to the Sun."

John Ronald Reuel Tolkien

Summary

Recent studies have indicated the presence of a high degree of cryptic genetic diversity in some clitellate sentinel species. One of these species, the earthworm *Lumbricus rubellus,* has been recently found to comprise two divergent clades in the UK, and are possibly cryptic species. *L. rubellus* is commonly used in ecotoxicological assays, where undetected differences in contaminant responses between cryptic lineages may lead to confusing or misleading results. Furthermore, given the key role that earthworm species play in the soil ecosystem, a better understanding of cryptic diversity is necessary to investigate whether divergent lineages play different roles within their ecosystems.

In this study, the phylogenomics of the acid-tolerant, cosmopolitan, epigeic species *Lumbricus rubellus* was investigated, with regard to demography during the glacial stages of the Pleistocene and the recent post-glacial colonization of North Europe using mitochondrial DNA markers, next-generation sequencing and environmental niche modelling tools.

The niche suitability of *L.rubellus* during the last 120.000 years was inferred, allowing hypotheses on survival and recolonisation to be constructed. Phylogenetic, population structure and coalescent-based analyses resulted in the discovery of 11 deep divergent lineages (with levels of divergence up to 18% for mitochondrial markers), which most likely survived in refugia during Pleistocene glaciations. Signatures of expansions point to a possible recolonisation of central Europe during the last Glaciation, survival of one of the clades in a northern cryptic glacial refugium and a consequent recolonisation of northern Europe during the last 10,000 years. Genetic evidence and divergence time ultimately suggest that *L. rubellus* is a cryptic species complex, which clades diverged as far as ~5MY ago.

The entire mitochondrial genome of the species complex is described here for the first time, and a survey of the deep phylogenetic signal over the mitochondrial genomes of eight selected individuals was carried out, supporting and deepening the phylogeny constructed using only two mitochondrial genes. Finally, whole genome analysis of genetic divergence supported the hypothesis of cryptic divergence for the two most divergent lineages selected.

To my Father Vittorio,
my Mother Speranza
and my Brother Antonello
with all my love and gratitude

ACKNOWLEDGEMENTS

I am not afraid to overstate when I describe this experience as one of the most terrifying, thrilling, painful, exciting, instructive (in terms of life as much as in terms of science) and important experiences of my life. Words cannot express how much these four years changed me and, I think, made me a better person. It was a fantastic trip, and I wish to thank with all my heart all the companions I met along the road.

First of all, I'd like to thank my supervisors. Mike, thank you for your relentless support and the endless trust in me, even in those moments when I was short of it. Thanks for teaching me the real meaning of being a professional, of not giving up in difficult times, and for being an example of what it means to be a great scientist without forgetting to be a human being. It was a privilege to have the opportunity to work with you! Pete, thank you for your great help and support and for inspiring me with your inexhaustible enthusiasm and your precious suggestions. Every chat with you was like a roller coaster in your incredible scientific mind! It's been great to have the opportunity to share your ideas and work together at this project. I hope I'll have many opportunities to collaborate with both of you in the future.

To all my colleagues in the lab, for the many adventures we shared: thank you!!! Leila, Rui, Joana, Tania, Renata, Silke, Neeza, Dave, Niall, Jen, Hannah, Xiang Jiang; all the people at PK's lab: Ceri, Craig, Marta, Dan, Yaser, Gonçalo and Luis; thanks to John, for all the enlightening conversations, for being there at the right time. Thanks to the Portuguese "children", with whom I had great fun during my third year: Jorge and Renatinha. Thanks to my dear ladies, Mafalda and Isa, for the many dinners, drinks, jokes, discussions and laughs, for your constant help with life and work, and for being always close to me. Thanks to Rob, the best companion for worm hunting ever! A special thanks to two persons which arrival literally turned the tide of my Ph. D.: thanks to Pablo, for supporting me, directing me, teaching me, making me believe that I could do it. I cannot imagine how I could have completed this without you. Oh and also thanks for all the attempts to make me drunk (yeah, you succeeded a couple of times…)! Thanks to Mario, for deciding to come here, for all the fun, laughs and jokes, and for the many challenges faced together. A good friend on your side always makes things easier.

Thank you to all my friends outside the lab. First of all, thanks to Lara. I would not be here if it wasn't for you! Eleni, Osian and Axileas: you're a constant in my Cardiff life! My beloved Doctor Arianna: thanks for your endless and unconditioned love and friendship. Thanks to Morena, for all the chats, and cakes, and laughs in our house! And thanks to my former supervisor and colleague Laura… all my gratitude for the cheerleading. It really helped! Thanks to all the friends whose lives are now unfolding in other countries. Even though you went away and you're not close anymore, I will never forget you. Danilo, Argyro, and their beautiful son Alexandros; my little Malaysian sister Faezah; Leonidas, the bloody genius; Juan Carlos and his beautiful family. I look forward to see you all again.

Thanks to all my friends of a life, back in the beautiful Sardinia: Daniele Sani, Daniele Campus, Ignazio, Luigi. Thanks to my former rock band Colt: Luca, Antonio, Bruno and my "surrogate", Francesco! And thanks to my "acquired brothers", Gianni and Joe, and to their ladies Maria Giovanna and Mary. Every time I went back, you made me feel like I'd never left.

Thanks to my family, with all my heart. Thank you for patiently listening my whining in bad moments and sharing my joy in good moments, for the constant encouragement and support and for all the love.

Finally, thank you, Margherita. It was a long and hard path, but you were always on my side on the way long. With you, I feel completed.

TABLE OF CONTENTS

DECLARATION……..I
SUMMARY……..……..IV
DEDICATION……V
ACKNOWLEDGEMENTS……...……...VI
TABLE OF CONTENTS…….……..VIII
LIST OF FIGURES……...XI
LIST OF TABLES……XV

CHAPTER 1 GENERAL INTRODUCTION 1
1.1 Foreword 2
1.2 Soil ecosystem engineers 3
1.3 *Lumbricus rubellus* biology and ecology 4
1.4 Cryptic speciation 8
1.5 Sentinel species in ecotoxicology 9
1.5.1 Earthworms as sentinel species 11
1.6 Phylogeography and the impact of paleoclimatic changes on species distributions 13
1.7 Phylogenomics 16
1.8 Species distribution modelling 17
1.8.1 Maxent 19
1.9 Molecular markers 22
1.9.1 Mitochondrial markers 22
1.9.2 Mitochondrial genomes 24
1.9.3 Genomic SNPs 24
1.9.4 The L. rubellus genome project 28
1.10 Aims and hypotheses 29
1.11 Bibliography 33

CHAPTER 2 PALEOCLIMATIC IMPACT ON CRYPTIC DIVERSITY OF THE EARTHWORM *LUMBRICUS RUBELLUS* IN EUROPE 51
2.1 Introduction 52
2.2 Materials and methods 54
2.2.1 Sampling and DNA extraction 54

2.2.2 PCR and sequencing 54
2.2.3 Population analyses and demographic inference 55
2.2.4 Phylogenetic analyses and divergence time 56
2.2.5 Species and paleodistribution modelling 57
2.2.6 Phylogeographic modelling 60
2.3 Results 61
2.3.1 Phylogenetic and population structure analysis 62
2.3.2 Divergence time estimates 64
2.3.3 Environmental niche modeling: current distribution 67
2.3.4 Paleodistribution modelling 67
2.3.5 Variable contribution 70
2.3.6 Phylogeographic modelling 72
2.4 Discussion 75
2.5 Acknowledgments 80
2.6 Bibliography 82
CHAPTER 3 MITOGENOMICS OF CRYPTIC DIVERSITY: EXPLORING THE DEEP PHYLOGENETIC SIGNAL OF *LUMBRICUS RUBELLUS* 88
3.1 Introduction 89
3.2 Materials and methods 91
3.2.1 Assembly and annotation 91
3.2.2 Phylogenetic and statistical analyses 93
3.3 Results 95
3.3.1 Genome structure and organisation 95
3.3.2 tRNAs and rRNAs 99
3.3.3 Non-coding regions 101
3.3.4 Phylogenetic and statistical analyses 103
3.4 Discussion 112
3.5 Bibliography 118
CHAPTER 4 PHYLOGENOMIC ANALYSIS OF *LUMBRICUS RUBELLUS* CRYPTIC CLADES 125
4.1 Introduction 126
4.2 Materials and methods 128
4.2.1 Illumina reads mapping 128
4.2.2 Variant detection, demographic and phylogenetic analyses 129
4.3 Results 130
4.3.1 Genomes mapping 130

4.3.2 Phylogenetic analyses 134
4.4 Discussion 137
4.5 Bibliography 143
CHAPTER 5 GENERAL DISCUSSION 148
5.1 Overview of main results 149
5.1.1 Overview of chapter 2: Paleoclimatic impact on cryptic diversity of the earthworm Lumbricus rubellus in Europe 149
5.1.2 Overview of Chapter 3: Mitogenomics of cryptic diversity: exploring the deep phylogenetic signal of Lumbricus rubellus 152
5.1.3 Overview of chapter 4: Phylogenomic analysis of Lumbricus rubellus cryptic clades 152
5.2 Is *L. rubellus* really a cryptic species complex? 154
5.3 Implications for *L. rubellus* as a sentinel species 155
5.4 Future perspectives 156
5.5 Bibliography 158
CHAPTER 6 SUPPORTING INFORMATION FOR CHAPTER 2 164
6.1 Demography supporting figures 165
6.2 Species distribution modeling evaluation 170
6.2.1 Standard deviation maps 170
6.2.2 Model performance evaluation: AUC, MESS and MoD analyses 171
6.3 Samples information 175
CHAPTER 7 SUPPORTING INFORMATION FOR CHAPTER 3 182

LIST OF FIGURES

FIGURE 1.1 (A) ILLUMINA LIBRARY PREPARATION. (B) GENERATION OF CLUSTERS WITH BRIDGE-AMPLIFICATION. (C) DESCRIPTION OF SEQUENCING BY REVERSIBLE-DYE TERMINATORS. FROM MARDIS (2013). 26
FIGURE 2.1. PLOT OF THE PRESENCE DATA USED FOR SPECIES DISTRIBUTION MODELLING IN MAXENT. EACH BLUE DOT REPRESENTS A SINGLE OCCURRENCE POINT. THE DATASET COMBINES THE GBIF OCCURRENCE POINTS, CLEANED FOR DUPLICATES AND RE-SAMPLED TO AVOID BIAS, AND THE COORDINATES OF THE SAMPLES COLLECTED FOR THIS STUDY, IN A TOTAL OF 179 OCCURRENCE POINTS. 58
FIGURE 2.2. LEFT: MAXIMUM-LIKELIHOOD TREE BASED ON THE TWO GENE FRAGMENTS AMPLIFIED FOR *L. RUBELLUS* (COI AND COII); RIGHT: BAPS CLUSTERING SOLUTIONS. BAYESIAN-LIKE TRANSFORMATION OF APPROXIMATE LIKELIHOOD RATIO TEST VALUES (ANISIMOVA ET AL. 2011) FOLLOWED BY POSTERIOR PROBABILITIES OVER 0.5 ARE SHOWN ON EACH BRANCH. BAPS CLUSTER SOLUTIONS ON THE RIGHT REFLECT THE TREE TOPOLOGY WITH A FEW EXCEPTIONS DENOTED BY ARROWS, POINTING OUT MISMATCHES BETWEEN PHYLOGENETIC AND BAPS STRUCTURE. 62
FIGURE 2.3. GEOGRAPHIC DISTRIBUTION OF SAMPLES ACCORDING TO THEIR MITOCHONDRIAL LINEAGES. THE AREAS OF THE PIE CHARTS REPRESENT THE NUMBER OF SAMPLES COLLECTED IN EACH LOCATION. 64
FIGURE 2.4. ULTRAMETRIC TREE REPRESENTING DIVERGENCE TIMES IN THE MAIN CLADES UNDER A YULE SPECIATION PROCESS. RESULTS ARE SHOWN IN MILLIONS OF YEARS (95% CONFIDENCE INTERVAL). 65
FIGURE 2.5. MAXENT MODELS OF *LUMBRICUS RUBELLUS* HABITAT SUITABILITY ACROSS EUROPE. WARMER COLOURS REPRESENT HIGHER HABITAT SUITABILITY. VALUES FROM 0 TO 1 ARE AN ESTIMATION OF THE PROBABILITY OF SPECIES PRESENCE. A) SPECIES DISTRIBUTION MODEL OF PRESENT TIME; B) PALEODISTRIBUTION MODEL OF THE LAST GLACIAL MAXIMUM (~21,000 YEARS AGO), ACCORDING TO THE CCSM PALEOCLIMATIC MODEL; C) PALEODISTRIBUTION MODEL OF LAST GLACIAL MAXIMUM, ACCORDING TO THE MIROC PALEOCLIMATIC MODEL; D) PALEODISTRIBUTION MODEL OF THE LAST INTERGLACIAL (EEMIAN STAGE, ~120,000-140,000 YEARS AGO), ACCORDING TO THE MIROC MODEL. 69
FIGURE 2.6. PLOTS REPORTING THE EFFECT OF THE MOST IMPORTANT VARIABLES ON NICHE PREDICTION. EACH CURVE REPRESENTS A MAXENT MODEL CREATED USING ONLY THE CORRESPONDING VARIABLE. THE RED LINE REPRESENTS THE MEAN, AND THE BLUE AREAS THE STANDARD DEVIATION BETWEEN THE RUNS. THE PLOT EVIDENCES HOW THE VARIABLE CONTRIBUTES TO THE NICHE MODEL ACROSS ITS RANGE. THE VARIABLES SHOWN ARE ESTIMATES RELATED TO TEMPERATURE (BIO7: TEMPERATURE ANNUAL RANGE; BIO9: MEAN TEMPERATURE OF THE DRIEST QUARTER; BIO4: TEMPERATURE SEASONALITY) AND PRECIPITATION (BIO17: PRECIPITATION OF THE DRIEST QUARTER; BIO14: PRECIPITATION OF THE DRIEST MONTH; BIO12: ANNUAL PRECIPITATION). WHERE THE TEMPERATURE VARIABLES SHOW HIGHER SUITABILITY ON THE CENTER-LEFT PART OF THE DISTRIBUTIONS, THE PRECIPITATION VARIABLES SHOW THE OPPOSITE, WHERE AREAS WITH HIGHEST PRECIPITATION VALUES ACROSS THE YEAR ACHIEVE THE HIGHEST SUITABILITY PROBABILITY. 71
FIGURE 2.7. JACKKNIFE TEST OF VARIABLE IMPORTANCE. BLUE BARS CORRESPOND TO CONTRIBUTION OF THE VARIABLE TO THE GAIN WHEN USED IN ISOLATION; GREEN BARS SHOW THE DECREASE IN GAIN WHEN THE VARIABLE IS MISSING. A: JACKKNIFE RESULTS ON REGULARIZED TRAINING GAIN; B: JACKKNIFE RESULTS ON THE GAIN OBTAINED CONSIDERING ONLY TEST SAMPLES. C: JACKKNIFE TEST OF A COMBINATION OF BIOCLIMATIC AND SOIL VARIABLES. SOIL VARIABLES SHOW VERY LITTLE CONTRIBUTION TO THE MODEL CONSTRUCTION; THE MOST IMPORTANT SOIL VARIABLE, CATION EXCHANGE CAPACITY, INCREASES THE AUC OF ONLY 1.5 FROM THE RANDOM VALUE. 72

FIGURE 2.8. SCHEMATIC REPRESENTING ALTERNATIVE BIOGEOGRAPHICAL HYPOTHESES FOR FIVE OF THE 11 *L. RUBELLUS* LINEAGES. HYPOTHESIS 1: 3 ANCESTRAL POPULATIONS DERIVED FROM ANCIENT POPULATION B (T3), AND FOLLOWING SPLIT OF THE ANCIENT POPULATIONS A AND Γ INTO THE HAPLOGROUPS A2 – C (T1) AND F – G (T2) BEFORE THE BEGINNING OF THE PLEISTOCENE; LINEAGE B IS THE DIRECT DESCENDANT OF THE ANCESTRAL POPULATION - HYPOTHESIS 2: SPLIT OF ALL THE LINEAGES FROM THE B ANCESTRAL POPULATION DURING THE EEMIAN STAGE (LIG, 140,000 YEARS AGO) – HYPOTHESIS 3: ANCIENT SPLIT OF ALL THE LINEAGES FROM THE B ANCESTRAL POPULATION DURING THE PLIOCENE (~5 MILLION YEARS AGO). 74
FIGURE 3.1. GENE MAP OF THE MITOCHONDRIAL GENOME OF *LUMBRICUS RUBELLUS* (LINEAGE B). ALL THE GENES ARE ENCODED ON THE SAME STRAND. THE TRNAS ARE REPRESENTED BY THE IUPAC SINGLE LETTER CODES. THE INNER CIRCLE REPRESENTS THE GC CONTENT. 96
FIGURE 3.2. TRANSFER RNA (TRNA) STRUCTURES OF *LUMBRICUS RUBELLUS* LINEAGE B. 100
FIGURE 3.3. POSSIBLE SECONDARY STRUCTURES OF THE TWO MAJOR NON-CODING REGIONS IN LINEAGE B. A: AT-RICH REGION (DG = -190.34); B: UNK REGION FOLLOWING THE ND6 GENE (DG = -32.84). 102
FIGURE 3.4. REPRESENTATIVE GENE TREE TOPOLOGIES. EACH TOPOLOGY IS LABELLED BY GENE. THE FIRST TOPOLOGY (TOP-LEFT OF THE FIGURE) REPRESENTS THAT OBTAINED BY CONSTRUCTING A TREE USING ALL THE CONCATENATED GENES. 105
FIGURE 3.5. META-TREE OF BAYESIAN PHYLOGENIES OF 15 DIFFERENT MITOCHONDRIAL GENES FROM THE MITOCHONDRIAL GENOMES OF *LUMBRICUS RUBELLUS* LINEAGES. EACH NODE REPRESENTS INTERMEDIATE TOPOLOGIES BETWEEN THE LEAVES; EACH LEAF REPRESENTS THE PHYLOGENETIC TREE TOPOLOGY OF THE GENES THAT LABEL THEM. 106
FIGURE 3.6. META-TREE FOR THE BOOTSTRAP ANALYSIS OF 15 MITOCHONDRIAL GENES OF *LUMBRICUS RUBELLUS*. EIGHT BOOTSTRAP REPLICATES FOR EACH GENE ARE SHOWN. EACH LABELLED VERTEX CORRESPONDS TO THE TOPOLOGY OF THE GENE OR MULTIPLE GENE REPLICATES THAT LABELS THEM. 107
FIGURE 3.7. PAIRWISE PLOTS OF THE RELATIONSHIPS BETWEEN THE COMPARED PHYLOGENETIC VARIABLES. R12.3: MATRIX OF PARTIAL MANTEL TEST CORRELATION VALUES; LENGTH: MATRIX OF GENE LENGTH DIFFERENCES; POL: MATRIX OF POLYMORPHISM DIFFERENCES BETWEEN GENES; RF: SYMMETRIC DISTANCE MATRIX; D: DIFFERENCES IN TAJIMA'S D BETWEEN GENES; PI: DIFFERENCES IN Π BETWEEN GENES. ALL THE VARIABLES, EXCEPT LENGTH AND POL, AND D AND PI, APPEAR TO BE UNCORRELATED. 109
FIGURE 4.1. SEQUENCE LENGTH DISTRIBUTION. Y AXIS REPRESENTS THE NUMBER OF READS, THE X AXIS REPRESENTS THE SEQUENCE LENGTH. THE GREY AREA REPRESENTS THE 90% OF THE READS. 132
FIGURE 4.2. FASTQC BOX PLOTS OF QUALITY SCORES PER READ POSITION OF ILLUMINA DATA). Y-AXIS: PHRED QUALITY SCORES. HIGHER SCORES MEAN BETTER QUALITY. THE BACKGROUND IS DIVIDED INTO THREE AREAS. THE GREEN AREA COMPRISES GOOD QUALITY SCORES; THE ORANGE AREA = AVERAGE AND THE RED AREA = POOR QUALITY. THE YELLOW AREA IN THE BOX REPRESENTS THE INTER-QUARTILE RANGE FROM THE 25^{TH} TO THE 75^{TH} PERCENTILE. THE ERROR BARS INCLUDE THE 10^{TH} AND THE 90^{TH} PERCENTILE. THE RED LINES INSIDE THE BOX PLOT ARE THE MEDIAN VALUE OF PHRED SCORES FOR THE NUCLEOTIDE RANGE, AND THE BLUE CURVE REPRESENTS THE MEAN VALUE. 133
FIGURE 4.3. PHYLOGENETIC TREES OBTAINED WITH THE TRANSCRIPTOME SNPS. A: MAXIMUM LIKELIHOOD TREE OBTAINED WITH THE PHYML SOFTWARE; SUPPORT VALUES ARE BAYESIAN-LIKE TRANSFORMATIONS OF THE APPROXIMATE LIKELIHOOD RATIO TEST (ABAYES) IMPLEMENTED IN THE PHYML SERVER (GUINDON ET AL. 2010). B: BAYESIAN PHYLOGENETIC TREE OBTAINED WITH MRBAYES. SUPPORT VALUES AT THE NODES ARE POSTERIOR PROBABILITIES. . 135
FIGURE 4.4. BAYESIAN MAJORITY RULE CONSENSUS TREE OBTAINED WITH A DATASET CONSTRUCTED WITH GENOMIC SNPS. SUPPORT VALUES SHOWN ARE

RELATIVE TO POSTERIOR PROBABILITY/BAYESIAN-LIKE TRANSFORMATION OF THE APPROXIMATE LIKELIHOOD RATIO TEST (ABAYES) IMPLEMENTED IN THE PHYML SERVER (GUINDON ET AL. 2010). 136
FIGURE 6.1. BAYESIAN SKYLINE PLOTS (BSP) AND MISMATCH DISTRIBUTIONS (MD) OF THE BAPS CLUSTERS A1 (LEFT) AND A2 (RIGHT). THE BLACK LINE OF THE BSP REPRESENTS THE MEDIAN ESTIMATE OF THE POPULATION SIZE *NE*M OVER COALESCENT INTERVALS, WITH THE BLUE LINES REPRESENTING THE CONFIDENCE INTERVALS. THE MD SHOWS THE OBSERVED (BAR PLOTS) AND THE EXPECTED (LINE PLOTS) VALUES OF THE DISTRIBUTION OF PAIRWISE DIFFERENCES. THE TWO CLUSTERS CLEARLY SHOW A SIGNATURE OF PAST EXPANSION. 165
FIGURE 6.2. BAYESIAN SKYLINE PLOTS (BSP) AND MISMATCH DISTRIBUTIONS (MD) OF THE BAPS CLUSTERS A3 (LEFT) AND C (RIGHT). THE BLACK LINE OF THE BSP REPRESENTS THE MEDIAN ESTIMATE OF THE POPULATION SIZE *NE*M OVER COALESCENT INTERVALS, WITH THE BLUE LINES REPRESENTING THE CONFIDENCE INTERVALS. THE MD SHOWS THE OBSERVED (BAR PLOTS) AND THE EXPECTED (LINE PLOTS) VALUES OF THE DISTRIBUTION OF PAIRWISE DIFFERENCES. A3 GRAPHS SHOW A STABLE POPULATION, WHEREAS THE BALKAN LINEAGE C SHOWS A CLEAR SIGNATURE OF EXPANSION. 166
FIGURE 6.3. BAYESIAN SKYLINE PLOTS (BSP) AND MISMATCH DISTRIBUTIONS (MD) OF THE BAPS CLUSTERS D (LEFT) AND E (RIGHT). THE BLACK LINE OF THE BSP REPRESENTS THE MEDIAN ESTIMATE OF THE POPULATION SIZE *NE*M OVER COALESCENT INTERVALS, WITH THE BLUE LINES REPRESENTING THE CONFIDENCE INTERVALS. THE MD SHOWS THE OBSERVED (BAR PLOTS) AND THE EXPECTED (LINE PLOTS) VALUES OF THE DISTRIBUTION OF PAIRWISE DIFFERENCES. THESE TWO CLUSTERS SHOW SIGNATURES OF STABLE DEMOGRAPHICS THROUGH TIME. 167
FIGURE 6.4. BAYESIAN SKYLINE PLOTS (BSP) AND MISMATCH DISTRIBUTIONS (MD) OF THE BAPS CLUSTERS B (LEFT) AND F (RIGHT). THE BLACK LINE OF THE BSP REPRESENTS THE MEDIAN ESTIMATE OF THE POPULATION SIZE *NE*M OVER COALESCENT INTERVALS, WITH THE BLUE LINES REPRESENTING THE CONFIDENCE INTERVALS. THE MD SHOWS THE OBSERVED (BAR PLOTS) AND THE EXPECTED (LINE PLOTS) VALUES OF THE DISTRIBUTION OF PAIRWISE DIFFERENCES. LINEAGE B SHOWS A SIGNATURE OF EXPANSION STARTING 25,000 YBP, JUST BEFORE THE START OF THE ICE RETREAT PHASE. ACCORDING TO THE BSP, THE SPANISH LINEAGE F SHOWS A STABLE POPULATION DEMOGRAPHY FOR THE LAST MILLION OF YEARS. 168
FIGURE 6.5. BAYESIAN SKYLINE PLOTS (BSP) AND MISMATCH DISTRIBUTIONS (MD) OF THE BAPS CLUSTERS G (LEFT) AND H (RIGHT). THE BLACK LINE OF THE BSP REPRESENTS THE MEDIAN ESTIMATE OF THE POPULATION SIZE *NE*M OVER COALESCENT INTERVALS, WITH THE BLUE LINES REPRESENTING THE CONFIDENCE INTERVALS. THE MD SHOWS THE OBSERVED (BAR PLOTS) AND THE EXPECTED (LINE PLOTS) VALUES OF THE DISTRIBUTION OF PAIRWISE DIFFERENCES. THESE TWO CLUSTERS CLEARLY SHOW SIGNATURES OF STABLE DEMOGRAPHY THROUGH TIME..... 169
FIGURE 6.6. STANDARD DEVIATION (SD) MAPS OF THE SPECIES DISTRIBUTION MODEL RUN IN MAXENT. THE COLOR SCALE REPRESENTS SD POINT VALUES ON THE MAPS, WITH HIGHER VALUES REPRESENTED BY WARMER COLORS. A: SD MAP OF THE 10 REPLICATED MODELS OF THE PRESENT TIME; B,C: MAPS OF THE STANDARD DEVIATION OF THE 10 REPLICATED MODELS APPLIED TO THE ENVIRONMENTAL LAYERS OF CCSM (B) AND MIROC (C). THE STANDARD DEVIATION OF BOTH MODELS SHOWED THERE WAS LITTLE DIFFERENCE AMONG THE 10 MODEL REPLICATES WHEN PROJECTED TO THE CCSM AND MIROC ENVIRONMENTAL VARIABLES, WITH SLIGHTLY HIGHER VARIABILITY RELATIVE TO THE MIROC REPLICATES; D: MAP OF THE STANDARD DEVIATION OF THE 10 REPLICATED MODELS APPLIED TO THE ENVIRONMENTAL LAYERS OF THE LIG. ALSO IN THIS CASE, THE STANDARD DEVIATION OF THE REPLICATE RUNS WAS LOW. 170
FIGURE 6.7. AUC ESTIMATION FOR THE *LUMBRICUS RUBELLUS* CURRENT DISTRIBUTION, CALCULATED ON THE TEST DATA AVERAGED FOR 10

REPLICATES. THE RED LINE REPRESENTS THE MEAN AUC OF THE TEST DATA ACROSS THE REPLICATES, AND THE BLUE AREA REPRESENTS THE STANDARD DEVIATION. THE BLACK LINE REPRESENTS A RANDOM PREDICTION SCENARIO. THE TEST AUC HAS A VALUE OF 0.875, SUGGESTING A GOOD PREDICTIVE CAPACITY OF THE MODEL. 171
FIGURE 6.8. MESS MAPS OF THE PALEODISTRIBUTION MODEL PROJECTIONS. A,B: LGM CCSM MODEL (A) AND THE MIROC MODEL (B). C: LIG MODEL. RED AREAS DEPICT PALEOCLIMATIC VARIABLE VALUES OUT OF RANGE. 173
FIGURE 6.9. A, B: MOD MAPS OF THE LGM CCSM MODEL (A) AND THE MIROC MODEL (B). THE MAIN VARIABLES OUTSIDE THE RANGE ARE MEAN DIURNAL RANGE (BIO2) AND ISOTHERMALITY (BIO3) FOR CCSM, AND MEAN TEMPERATURE OF THE WETTEST QUARTER (BIO8) FOR MIROC; C: THE LIG MOD MAP SHOWS THAT THE MODEL'S EXTRAPOLATED AREA IS MAINLY TOWARDS THE EAST OF THE RANGE, AND THE TEMPERATURE ANNUAL RANGE (BIO7) IS THE VARIABLE RESPONSIBLE OF THIS. 174
FIGURE 7.1**.** BAYESIAN PHYLOGENETIC TREES. EACH TREE REPRESENT A SINGLE GENE ALIGNMENT, EXCEPT THE FIRST ONE (FIRST PAGE, UPPER LEFT), WHICH WAS OBTAINED FROM A DATASET BUILT CONCATENATING ALL THE GENES. 191

LIST OF TABLES

TABLE 2.1. LIST OF THE BIOCLIMATIC VARIABLES AVAILABLE ON THE WWW.WORLDCLIM.COM DATABASE FOR THE PRESENT, THE LAST GLACIAL MAXIMUM (LGM) AND THE LAST INTERGLACIAL (LIG). A QUARTER REPRESENTS ¼ OF THE YEAR. 59
TABLE 2.2. THE NUMBER OF INDIVIDUALS (N), HAPLOTYPE (*H*) AND NUCLEOTIDE DIVERSITY (*Π)* AVERAGED PER LOCUS FOLLOWED BY STANDARD DEVIATION VALUES FOR EACH CLUSTER. 61
TABLE 2.3. ESTIMATES OF EVOLUTIONARY DIVERGENCE OVER SEQUENCE PAIRS BETWEEN (LOW DIAGONAL) AND WITHIN (ON DIAGONAL, IN BOLD) GROUPS. THE NUMBERS OF BASE SUBSTITUTIONS PER SITE FROM AVERAGING OVER ALL SEQUENCE PAIRS BETWEEN GROUPS ARE SHOWN. THE MODEL USED FOR THE ANALYSES WAS KIMURA -2 PARAMETER..... 63
TABLE 2.4. TAJIMA'S D, FU'S FS, SSD AND HARPENDING'S R, WITH THEIR RELATIVE P VALUES, ARE REPORTED. SIGNIFICANT VALUES ARE HIGHLIGHTED IN ORANGE. 67
TABLE 2.5. COALESCENT SIMULATIONS. THE VALUES OF THE SUMMARY STATISTICS OF THE RESTRICTED DATASET ARE SHOWN IN COLUMN 2. COLUMNS 3 TO 5 REPORT THE CONFIDENCE INTERVAL FOR EACH STATISTIC IN THE 3 DIFFERENT COALESCENT HYPOTHESES. HYPOTHESES 2 AND 3 ARE REJECTED BECAUSE THE EMPIRIC VALUES FALL OUTSIDE THE CI. Θ_S AND THE NUMBER OF POLYMORPHIC SITES WERE USED AS AN INPUT PARAMETER FOR ALL THE SIMULATIONS. 73
TABLE 3.1. SAMPLES USED FOR THIS STUDY. GENETIC LABEL, COUNTRY OF ORIGIN, HAPLOTYPE AND SAMPLE COORDINATES ARE SHOWN. THE LAST INDIVIDUAL (S20) COMES FROM A POPULATION STUDIED IN ANDRE ET AL. (2010). 92
TABLE 3.2. MITOCHONDRIAL GENOME PROFILE OF *LUMBRICUS RUBELLUS* LINEAGE B. START AND END POSITION, LENGTH, START CODON, STOP CODON, INTERGENIC NUCLEOTIDES AND AT% ARE SHOWN. THE NUMBER OF INTERGENIC NUCLEOTIDES IS NEGATIVE WHEN THERE IS AN OVERLAP BETWEEN LOCI..... 97
TABLE 3.3. THE CODON USAGE FOR THE LINEAGE B MITOCHONDRIAL GENOME. FOR EACH AMINO ACID (AA), THE RELATIVE CODON, THE NUMBER OF THE CODONS IN THE GENOME, THE RELATIVE PERCENTAGE OF THE CODON TRANSLATING THAT AMINO ACID, AND THE /1000 VALUE ARE SHOWN. THE TOTAL NUMBER OF CODONS IS 3707. 98
TABLE 3.4. SUMMARY STATISTICS FOR EACH GENE ALIGNMENT. EVOLUTIONARY MODELS WERE OBTAINED USING THE AKAIKE INFORMATION CRITERION SELECTION OF THE BEST-FITTING MODEL OF MOLECULAR EVOLUTION USING MRMODELTEST 2.3 (DARRIBA ET AL. 2012). LENGTH, VARIABLE SITES, TAJIMA'S D AND Π STATISTICS WERE OBTAINED WITH THE PROGRAM DNASP (ROZAS AND ROZAS 1999)..... 104
TABLE 3.5. SYMMETRIC DISTANCES CALCULATED WITH THE ROBINSON & FOULDS METHOD (ROBINSON AND FOULDS 1981). EACH NUMBER REPRESENTS THE SYMMETRIC DISTANCE BETWEEN EACH GENE TREE. THE NUMBERS DESCRIBE THE STEPS NECESSARY TO CONVERT ONE GENE TREE INTO THE OTHER..... 110
TABLE 3.6. PARTIAL MANTEL TEST R VALUES (LOWER DIAGONAL) AND P VALUES (UPPER DIAGONAL) OBTAINED FROM THE COMPARISON BETWEEN P-DISTANCE MATRICES CALCULATED FOR EACH GENE, COMPARING BETWEEN THE 8 GENOMES, AND THE P-DISTANCE MATRIX CALCULATED ON THE CONCATENATED SEQUENCE OF ALL THE GENES. SIGNIFICANT VALUES ARE HIGHLIGHTED IN GREY..... 111
TABLE 4.1. COMPARISON BETWEEN PERFORMANCES OF MAPPING ALGORITHMS. THE FIRST COLUMN REPORTS THE LABEL OF THE SEQUENCED INDIVIDUALS (HAPLOGROUP_LOCATION). THE SECOND COLUMN REPRESENTS THE TOTAL NUMBER OF ILLUMINA READS PER INDIVIDUAL. COLUMNS 3, 4, AND 5 REPORT THE PERCENTAGE OF READS MAPPED TO THE REFERENCE USING THREE DIFFERENT MAPPING ALGORITHMS: CLC GENOMIC WORKBENCH, BWA-ALN AND BWA-SW. 131

TABLE 7.1. MITOCHONDRIAL GENOME PROFILES OF *LUMBRICUS RUBELLUS* LINEAGES. START AND END POSITION, LENGTH, START CODON, STOP CODON, INTERGENIC NUCLEOTIDES AND AT% ARE SHOWN. THE NUMBER OF INTERGENIC NUCLEOTIDES IS NEGATIVE WHEN THERE IS AN OVERLAP BETWEEN LOCI..... 183

CHAPTER 1 GENERAL INTRODUCTION

1.1 Foreword

Earthworms, common creatures that we encounter from the day we start to explore nature with childhood curiosity, may appear insignificant creatures to an inexperienced eye. Nevertheless, their importance has been recognised for millennia. In Ancient Egypt, earthworms were considered indispensable to the agricultural economy. It has been told that Cleopatra declared these animals sacred, and any export of earthworms was punished by the death penalty (Minnich 1977). In more recent times, they received the passionate attention of Charles Darwin. Since then, the interest of scientists in these organisms has been constant, increasing consciousness of their importance for ecosystem health.

Charles Darwin's first scientific paper, written in 1837, described the formation of mould as a result of worm activity. His interest for these animals continued all his life and found a symmetrical conclusion with his last book, "*The formation of vegetable mould, through the action of worms, with observations on their habits*". This book, written in 1881, was his most successful book at the time of his death, one year later. For the following 120 years, earthworms have been studied to address important questions in the environmental and ecological sciences. These studies have emphasized the benefits that earthworms bring to soil ecosystems. Processing soil organic matter, they indirectly promote plant growth by increasing the turnover rate of nutrients, preventing run-off and promoting soil microflora (Syers and Springett 1983). According to Lavelle (1997) earthworms have a key role in physically structuring the soil and modifying its environment. Consequently, they can be considered as 'ecosystem engineers' capable of improving soil quality by increasing the availability of nutrients through direct and indirect action (Jouquet et al. 2006). Furthermore, as soil organisms, earthworms are in intimate contact with bioavailable chemicals, and are thus sensitive indicators of soil health. Hence, they can be important biomarkers of how soil organisms adapt, in terms of both physiology and evolution, to changing soil chemistries.

It is commonly believed that the last glaciations eradicated much of the soil fauna in northern Europe and that modern communities are the result of recolonisation after the retreat of the ice (Hewitt 2000). It has been speculated that the

modern earthworm fauna in regions such as UK and Scandinavia are the result of such recolonisations, and that Pleistocene glaciations played an important role in the cryptic differentiation process of many annelid taxa (King et al. 2008). Another important factor that could have shaped the patterns of recolonisation and distribution of earthworms is either the natural or artificial composition of soils, including heavy metal burden (Edwards 2004).

Understanding the genetic structure and adaptations of earthworms potentially aids modelling of present and future ecosystem functions and changes. Despite this, a large-scale genetic analysis of the history and demography of earthworm species, with regards to survival in glacial refugia and the subsequent re-colonization after the ice retreat, is conspicuously lacking. Due to their keystone and ecological engineering status, this is essential to achieve a better knowledge and understanding of their population structures, evolutionary histories, and genetic diversity. The present project therefore aims to investigate the phylogenomics of the acid-tolerant, cosmopolitan, epigeic species *Lumbricus rubellus* with regard to demography during the glacial stages of the Pleistocene and the recent post-glacial colonization of North Europe using mitochondrial DNA markers and next-generation sequencing tools.

1.2 Soil ecosystem engineers

Soil is defined as the upper layer of the planet's surface. It is a complex substrate composed by a mix of mineral particles, organic matter, water, air and living organisms. Soil is considered a non-renewable source (Turbé et al. 2010), performing many functions of vital importance: food production, biomass production, transformation and storage of many substances of primary ecological importance, such as water, nitrogen and carbon. Soil serves as base of human activities and provides raw materials; most importantly, soil is habitat and serves as gene pool for an incredible portion of biodiversity of the planet (Heywood 1995; Decaëns et al. 2006). Indeed, soils form the habitat and resource base for a large element of global biodiversity: over one quarter of all living species are strict soil or litter dwellers (Decaëns et al. 2006). Soils are home to prodigious biodiversity, which can often be several orders of magnitude greater than that present aboveground or in the canopy of rainforests (Heywood 1995; Decaëns et al. 2006). One square metre of land surface

may contain some ten thousand species of soil organisms, whereas aboveground biodiversity may be orders of magnitude less rich (Schaefer and Schauermann 1990; Wardle et al. 2004).

This astounding richness includes organisms that are essential in moulding the soil ecosystem and for this reason they are sometimes called ecosystem engineers. Ecosystem engineers modify the environmental conditions for other organisms through their mechanical activities (Jones et al. 1997). Earthworms, termites, ants and roots have been identified as the most important biota in the soil (Lavelle et al. 1997). Earthworms, in particular, are the single largest contributors to the soil invertebrate biomass in many soil environments. As soil ecosystem engineers, they have the ability to build resistant organo-mineral structures and pores by moving through and mixing the soil, in a process known as bioturbation. They also have important roles in soil aeration and organic matter fragmentation, thus contributing directly to nutrient cycling and structural development of the soil (Darwin 1892; Heemsbergen et al. 2004). It was even discovered that they play a direct role in determining the greenhouse-gas balance of soils worldwide (Lubbers et al. 2013).

1.3 *Lumbricus rubellus* biology and ecology

Earthworms are systematically grouped in the phylum Annelida, class Clitellata, subclass Oligochaeta, order Haplotaxida and suborder Lumbricina. There are 6000 or more described species of Oligochaeta, and half of them are earthworms (Sims and Gerard 1999). The suborder Lumbricina consists of five families, including Lumbricidae, the family of *L. rubellus* (ITIS 2012).

L. rubellus presents the typical earthworm anatomy. The earthworm's body is divided externally in a series of segments separated by furrows. They have a thinly pigmented cuticle, which bears chitinous setae on almost every segment, except the first two. The tubular body wall consists of an external layer of circular muscles and an internal layer of longitudinal muscles. The furrows correspond internally to transverse walls, or septa, filled with coelomic fluid, which divide the internal coelomic space into segmental divisions. Pores on the septa allow the passage of the coelomic fluid within segments (Edwards 1996). They do not possess a skeleton, the function of which has been substituted by a hydraulic system formed by the

incompressible fluid in the coelom. The alimentary canal, the vessels of the vascular system and the nerve chord run over the length of the coelomic cavity. The anatomically and functionally differentiated alimentary canal contains brush-bordered absorptive epithelia. The closed vascular system, with at least a dorsal and a ventral trunk, carries circulating haemoglobin in free suspension. The organized nervous system has cephalic ganglia and neurosecretory activities, and there is evidence of a multifunctional tissue (the chloragog) for which carbohydrate metabolism and storage properties are reminiscent of mammalian hepatocytes (Edwards 1996). A series of paired tubules (nephridia) in each segment has renal urine-forming functions, and a systemic immune system comprises leukocyte-like cells (coelomocytes; Stürzenbaum et al. 2009).

According to Bouché (1971, 1977), earthworm communities can be divided in three major ecological groups: litter dwellers or epigeic species, deep burrowers (anecic), and horizontal burrowers (endogeics). *L. rubellus* belongs to the first category: normally, it lives in the superficial strata of the soil, above the mineral layer, feeding on decaying matter, decomposing logs, manure, compost etc. It is heavily pigmented, as an adaptation to life in the litter. It is characterised by relatively high reproductive rates, quick growth and high mobility. It has high mobility to avoid predation (Edwards 1996; Sims and Gerard 1999). *L. rubellus* presents anatomical features common to the epigeic category: heavy pigmentation on the dorsal side, a small-medium size (60-130 cm) and the anterior part of the body musculature and the anterior septa is reduced, as an adaptive response to life in the upper layers of the soil, where no deep burrows are necessary (Edwards 1996; Sims and Gerard 1999).

L. rubellus is a hermaphrodite lumbricid that reproduces with cross-fertilisation. Each individual has both male and female reproductive systems, mostly restricted to a few anterior segments, that function simultaneously (Edwards 1996). Usually a glandular swelling, the clitellum, develops over several segments in the anterior half of the body of the adults. This is where the cocoons or egg capsules are formed (Sims and Gerard, 1999). As is common in earthworms, *L. rubellus* is a semi-continuous or continuous breeder; it can reproduce periodically through the year, except when it is in diapause (a state in which the worm becomes inactive in response to unfavourable environmental conditions, (Edwards 1996). *L. rubellus* produces a relatively large number of cocoons per year (42-106) as opposed to the anecic species *Aporrectodea longa* (3-13 cocoons/y, Satchell 1967). According to this author, there

is a correlation between number of cocoons produced and exposition to adverse environmental factors such as temperature and predation in earthworm species. Epigeic species, more exposed to environmental hazards, tend to produce more cocoons to increase chances of survival. The time for maturity for *L. rubellus* has been reported to be 37 weeks (Evans and Guild 1948). This gives a generation time of 1/year.

L. rubellus is mainly found in the holartic zone. It is native of Europe, where is widely distributed, mainly in the temperate areas. The species has been reported to actively disperse ~ 14 m/y (Marinissen and Van den Bosch 1992), suggesting that passive dispersal, such as human mediated dispersal could also have played an important part of its distribution: indeed, anthropogenic dispersal have brought populations of the species in North America (Tiunov et al. 2006), peninsular India, South Africa, New Zealand and several other temperate oceanic Islands (Sims and Gerard 1999). *L. rubellus* is part of an ensemble of earthworms species defined 'peregrine' because of their wide distribution over geographically remote localities (Michaelsen 1903). The impact of *L. rubellus* and other peregrine species on formerly earthworm-free environments can be so substantial that it has been recognised as threatening the equilibrium of these ecosystems (Tiunov et al. 2006). The characteristics of peregrine species include many characteristics that make them efficient colonisers, such as ecological plasticity, opportunism in food choice, habitat specificity, tolerance to environmental variables and resistance to chemical stress (Lee 1987).

L. rubellus shows wide adaptability to a broad panel of environmental conditions. It has low levels of desiccation tolerance, thus preferring wetter soils, from 20 to 25% of moisture content (Edwards 1996), but in case of droughts, which can happen over some part of *L. rubellus* range, it has been suggested that it can survive in a cocoon stage, as testified by the finding of immature specimens in soils immediately after a drought (Parmelee and Crossley Jr 1988). Although it has been reported that the optimal temperature range for its development is between 15º and 18º C (Graff 1953), the species has a broad temperature tolerance. As a general rule, high temperature and dry soils are much more limiting for earthworms than low temperature and water saturated or flooded soils (Nordström and Rundgren 1974).

Although it is experimentally known that earthworms can be killed by freezing, in pastures and woodlands, soil does not freeze deeply enough to affect most

species. In general, earthworms have a lower tolerance threshold for higher temperatures than many other invertebrates, whereas tolerance for lower temperatures for many earthworm species has been recorded to be close to freezing point (Edwards 1996). This seems to be true for *L. rubellus*, scarcely distributed in the drought-prone regions of its Southern European range, but thriving at latitudes where soils usually freeze in winter (Tiunov et al. 2006). An additional characteristic that could help a peregrine species such as *L. rubellus* to survive to extreme climate conditions, is their documented capacity to acclimate to different temperatures, allowing for adjustments to seasonal temperature variations (Edwards 1996).

An important challenge for an earthworm can be posed by soil chemistry; it has been recorded that earthworms possess a broad variation in tolerance between species, some being acid tolerant, others being acid intolerant and other being ubiquitous. *L. rubellus*, as other peregrine species, is ubiquitous, tolerating a pH range from 3.5 to 8.4 (Sims and Gerard 1999).

As a result of this adaptability, the species can be found in a wide range of habitats, the usual environment being moist and rich in organic content and decaying vegetable matter. Its presence has been recorded in coniferous forests over the European and introduced North American range (Addison 2009). It has also been found in riparian zones, characterised by high soil moisture and conditions thought to be challenging for earthworm species (Costello and Lamberti 2008). Common environments are pastures (in particular with herds of grazing animals, where dung pats attract them in great quantities; Sims and Gerard 1999), gardens, parks and river banks. It has been proposed that the aggregation behaviour under dung pats may help reproduction in low dispersal species (Whalen 2004).

It has been proposed that the present day *L. rubellus* population of northern Europe countries is a result of recolonisation processes from southern European glacial refugia, occurring after the retreat of the glaciers characterising the last Ice Age (King et al. 2008). Alternatively, the possibility of survival of *L. rubellus* in cryptic northern refugia (Stewart and Lister 2001; Stewart et al. 2010) has been proposed to explain the presence of two deep divergent mitochondrial lineages of the species in UK (Donnelly et al. In Press).

1.4 Cryptic speciation

Historically, zoological taxonomy has focused on patterns of morphological variation in animals, but a classical problem is that in closely related organisms, intraspecific variation may be difficult to distinguish from interspecific differences. In 1948, Mayr argued for the first time against the morphological species concept, pointing out the difficulties in distinguishing clear morphological patterns of differentiation for some species.

During the 19th century and the beginning of the 20th century, a great number of oligochaeta were formally named, but they were often described briefly and/or separated from each other on the basis of rather subtle morphological variation (Brinkhurst et al. 1971). Numerous studies in the last 30 years (Holmquist 1983; Erséus 1988; Coates 1990; Erséus 1997; Wang and Erséus 2004; Rota et al. 2007) indicate that oligochaete diversity is considerably higher than once thought. Taxonomy of earthworms, in particular, suffers from a lack of diagnostic structural characters, a common feature observed in soil organisms (Lee and Frost 2002; Pop et al. 2003) and many characters of taxonomic importance overlap among taxa (Pérez-Losada et al. 2005).

The debate around the morphospecies concept has been recently invigorated by the "revolution" caused by the use of barcoding and molecular analysis to support taxonomy. In particular, invertebrate studies have been characterised, in the last years, by the continuous discovery of hidden deep genetic diversity within morphologically similar taxa. Cryptic speciation has been detected in molluscs (Baker et al. 2003), crustaceans (Adamowicz et al. 2007; Wares et al. 2007) insects (Chenon et al. 2000; Schlick-Steiner et al. 2006) and annelids (Sturmbauer et al. 1999; Erséus and Gustafsson 2009), including sentinel species used in ecotoxicological studies, such as the freshwater oligochaete *Tubifex tubifex* (Sturmbauer et al. 1999), the terrestrial collembolan *Folsomia candida* (Chenon et al. 2000), and the marine isopod *Idotea baltica*, living in intertidal ecosystems (Wares et al. 2007). In particular for earthworms, numerous cases of cryptic divergence in clitellate organisms have been discovered (Heethoff et al. 2004; King et al. 2008; Novo et al. 2009; Pérez-Losada et al. 2009; James et al. 2010; Novo et al. 2010; Buckley et al. 2011; Dupont et al. 2011). All these studies support a scenario where soil conditions and sexual selection independent from visual signals limit morphological changes (Lee and Frost 2002).

The importance of earthworms in soil ecology and ecosystem functioning means the consideration of knowledge of cryptic genetic diversity of primary importance. This knowledge is necessary to further develop research on taxonomy, ecology, systematics and conservation of this variety (King et al. 2008; Fernández et al. 2012).

L. rubellus has been recently found to include two cryptic lineages in Britain (King et al. 2008), with a possible differential response to contaminants (Andre et al. 2009). A recent investigation suggests that this divergence encompasses both mitochondrial and nuclear genomes, hinting at effective reproductive isolation (Donnelly et al. In Press). Therefore, a deeper knowledge of the extent and the patterns of distribution of cryptic variation in peregrine species such as *L. rubellus,* is of primary importance in order to correctly identify species in ecological studies.

1.5 Sentinel species in ecotoxicology

The soil's dynamic system is strongly influenced by environmental variables, such as climate change, human activities and pollution. These factors have a major impact on soil biodiversity (Turbé et al. 2010). Soils reflect the mineral composition of their sources, and are often subject to change in composition due to human activities. Thus, soil concentrations of toxic metals and metalloids will reflect the natural local soil geochemistry as well as the anthropic exploitation of mineral resources.

Soil organisms may respond to toxic levels of soil metals in a number of ways. Toxicity may include mortality, reduced lifespan, reduced fecundity or changes in lifecycle dynamics such as time to maturity (Brown et al. 2004; Spurgeon et al. 2004; Spurgeon et al. 2005; Hodson 2013). Local populations may simply become extinct, as the contaminated soils becomes sterilised. Soil animals may actively migrate out of or avoid the contaminated area (Lukkari and Haimi 2005). Alternatively, the organisms may express tolerance to metal, through phenotypic plasticity or physiological adaptations such as sequestration.

Some organisms are used as models to monitor the health of soil ecosystems. These organisms are called 'sentinel species' and they allow the assessment of the effect of contaminants on the ecosystems without the need to undertake whole-ecosystem surveys (LeBlanc and Bain 1997; Mhatre et al. 1997; Tabor and Aguirre

2004; Fränzle 2006). This approach has been criticised as being an oversimplification, but it has nonetheless been favoured as an affordable way to assess ecosystem health (Caro and O'Doherty 1999; Carignan and Villard 2002). Sentinel species can be used both in the laboratory, in modelling experiments to assess the effect of ecosystem exposure to contaminants, and in the field, to evaluate the effect of real contamination events (Francioni et al. 2007; Quirós et al. 2007).

The selection of sentinel species must be done so that the use of the species in the surveys yields meaningful results in the assessment of the pollution effect on the ecosystem. Firstly, the species must be ecologically relevant in the considered ecosystem, such that any detrimental effect on it can have an impact on the whole ecosystem at different trophic levels (Chapman 2002). Key ecological roles include prey species at different trophic levels (Hamers et al. 2006), species with an active role in nutrient cycling and organic matter decomposition (van Straalen et al. 2001) and ecosystem engineers (Jouquet et al. 2006) which activity affect the physical characteristics of their habitat. Secondly, the species must show sensitivity to the detrimental effects of contaminants, but such effects must be counterbalanced with a certain degree of tolerance, such that a lethal response does not occur at low levels of environment contamination. This would hinder the possibility to assess the effect of chronic levels of exposure to contaminants (Donnelly 2011).

Effects can be characterised using dose-response measurements on the species, and this allows the development of biomarkers for ecotoxicological studies (Vasseur and Cossu-Leguille 2003). Ecotoxicological biomarkers are defined as any biological response to an environmental chemical at individual level that can cause departure from the normal status (Eason and O'Halloran 2002). Biological responses caused by exposure to pollutants can be identified by analyses of morphology, behaviour and can be detected at the cellular and molecular level (Depledge et al. 1995; LeBlanc and Bain 1997). Modern ecotoxicogenomic techniques, including genomics, transcriptomics and metabolomics are capable of identifying gene expression of proteins involved the metabolic pathways of stress response (Spurgeon et al. 2008), allowing the development of biomarkers indicating cellular damage or detoxification (Bundy et al. 2008; Bundy et al. 2009).

It has been proposed that an ideal sentinel species should be sedentary or sessile, with a limited capacity to avoid exposure (Hilty and Merenlender 2000). Limited mobility in a sentinel species would also enable high resolution studies on the

effects of contamination within a given environment (Fränzle 2006). Ecological niche has also been recognised as important since exposure in specialised organisms could be higher than in generalists, able to change their diet to avoid contaminated sources (Hilty and Merenlender 2000). Sentinel species can be selected according to the simplicity with which they can be collected in the field. Abundance within the study area, and ease of identification are also important features (Caro and O'Doherty 1999). Ease in identifying species is also an issue, as it has been shown that even in closely related species, sensitivity to pollutants can differ markedly (Bach et al. 2005).

1.5.1 Earthworms as sentinel species

In the field of soil ecology, many earthworm species possess the required sentinel characteristics, and some of them are commonly used as bioindicators (Spurgeon et al. 2003; Stürzenbaum et al. 2009), including L. *rubellus*. Because of their size, they are easy to collect and recognise as opposed to other invertebrate bioindicator species (Mhatre et al. 1997; Paoletti 1999). The ecology of some species has been thoroughly studied (Bouché 1972; Sims and Gerard 1999) and the key roles they play in the soil ecosystem are well known (nutrient cycling, food webs, physical action via burrowing and bioturbation; Lavelle et al. 1997). As soil organisms, earthworms are in constant and intimate contact with bioavailable toxic substances and earthworm sentinel species show sensitivity towards a wide array of soil pollutants (Owen et al. 2008a; Brulle et al. 2010). It is not clear whether earthworms become exposed to contaminants through water uptake from their body walls, or with contaminated soil ingestion (Vijver et al. 2003; Morgan et al. 2004) but their ability to accumulate very high concentrations of trace heavy metals is well documented (e.g. Ireland 1983; Morgan and Morgan 1988; 1990; Morgan and Morgan 1998). In particular, different studies show that *L. rubellus* has a high metal accumulation capacity when compared to other common earthworm species (Edwards 1996). Laboratory based toxicological experiments showed evidence of precise stress response to pollutants, allowing the development of an array of biomarkers suitable for ecotoxicological studies (Spurgeon et al. 2003; Ricketts et al. 2004). Thus, earthworms possess many characteristics that make them ideal indicators of soil ecosystem health, and markers of how organisms adapt, in terms of both physiology and evolution, to changing soil chemistries (Edwards 2004).

A very important factor for a bioindicator species is genetic homogeneity, which should lead to homogeneity of stress response, and thus repeatability and predictability of the toxicological essays (Erséus and Gustafsson 2009). However, during the last few decades, genetic studies have unravelled a surprising scenario concerning the cryptic diversity of many annelid species (Beauchamp et al. 2001; Gustafsson et al. 2009), including earthworm sentinel species (Pérez-Losada et al. 2005; King et al. 2008).

For earthworms generally, only a few studies, concerning the genetic structure and variability of species and populations in an ecotoxicological context, have been published (Peles et al. 2003; Heethoff et al. 2004; Simonsen and Scott-Fordsmand 2004; King et al. 2008; Andre et al. 2009; Simonsen and Klok 2010; Anderson et al. 2013) but it is interesting to note that other invertebrates, such as collembolans (Simonsen et al. 2004), snails (Arnaud et al. 2001) and slugs (Pinceel et al. 2005a; Pinceel et al. 2005b), have been shown to possess genetic divergence among populations, possible cryptic species and genetic differentiation along Cu gradients. It is therefore possible that these different genotypes display differential responses or susceptibilities to environmental variation, including response to contaminants exposure (Erséus and Gustafsson 2009).

King et al. (2008) found that *L. rubellus* displays a very high degree of genetic lineage diversity, with mitochondrial cytochrome oxidase II sequence divergence of 13-15%, a level considered to be representative of species-level differentiation. There is also evidence that the distribution of different genotypes may not be uniform across a heterogeneous metal-contaminated landscape (Andre et al. 2009; Simonsen and Klok 2010). Early studies on cryptic variation in *L. rubellus* were mainly based on mitochondrial sequences. Nevertheless, a recent study found evidence of lineage separation also at the nuclear level, suggesting reproductive isolation (Donnelly et al. In Press). Therefore, there could be functional genetic differentiation between exposed and non-exposed populations, and cryptic lineages may have different response to pollutants. Because of that, the suitability of *L. rubellus* as a good sentinel species has been discussed. Anderson et al. (2013) found no evidence of differential susceptibility to arsenic toxicity in cryptic lineages of *L. rubellus*, but did find differential maturation times between juveniles, which could lead to differential survival between lineages in extreme conditions.

As an important species for biomonitoring, the extent of *L. rubellus* genetic

diversity and the degree of variation of the diverse metabolic and genetic adaptations to pollution of *L. rubellus* are worthy of investigation, in order to improve our knowledge of this species responses to contamination and thus gain a better evaluation of this species as a tool for ecotoxicological studies.

1.6 Phylogeography and the impact of paleoclimatic changes on species distributions

The climatic history of the European continent from the Miocene has been characterised by numerous climatic shifts with an overall change towards the colder climate of the Quaternary. These shifts have shaped evolutionary processes, divergence and speciation in all European biota. Of all these periods, the genetic and distributional impact of the Pleistocene ice ages on European fauna and flora is the best known (Hewitt 1996; Hewitt 2000; Hewitt 2004). Although the present genetic structure of populations, species and communities was mainly formed during the Quaternary ice ages (Hewitt 2000), previous climatic processes, particularly bound to oscillations in precipitations, must also have had a major impact on sedentary and moisture-dependent species, such as epigeic earthworms. The signature of this impact should still be detectable in their genes.

During the middle Miocene, a long dry period occurred (13 – 11 Million years ago), followed by a "washhouse climate" interval (10.2 – 9.8 Mya), characterised by globally warm conditions and precipitations several orders of magnitude higher than present. A dryer climatic period, between 9.7 and 9.5 Mya was the likely cause of the Vallesian crisis, a transition period which may have led to the extinction of hominoids in Western Eurasia. This period was followed by a second "washhouse" episode (9.0 – 8.5 Mya) at the same time as the onset of warmer global conditions; the precipitation level increased dramatically in Southwest Europe but was less pronounced in Central Europe (Böhme et al. 2008). These long periods of warmer, wetter conditions could have been favourable for soil organisms such as earthworms, and may have formed the distribution we observe today of some widespread peregrine species.

From this time until the beginning of the Pliocene period (5.3 Mya), precipitation recorded for these two areas started to diverge, with Southwest Europe being wetter and Central Europe becoming dryer than present (Böhme et al. 2008).

Progressive cooling of the climate during the Pliocene culminated in the "ice ages" of the Pleistocene, when climatic patterns were dominated by an approximate cycle of 100,000 years interleaved by shorter, warm interglacial periods, such as the period we are experiencing now (Pisias and Moore Jr 1981). The Quaternary ice ages had a marked effect on species distribution, formation and the establishment of major intraspecific lineages of European flora and fauna (Hewitt 1996, 2000, 2004). The most recent glacial period (the Late Devensian for the British Isles, or late Weichselian in Northern Europe) peaked between 25,000 and 13,000 years BP. During this period, glaciers extended as far as south as Northern France. South of the ice line, the climate was of permafrost and tundra. While a few taxa have tolerance to extreme low temperatures and may have survived in high mountains or ice-free coastal refuges within the area of permafrost coverage, most soil species are sensitive to freezing at temperatures only slightly below zero (Holmstrup 2003). Thus, these climate changes are likely to have caused the almost complete eradication of the temperate terrestrial flora and fauna in most of northern Europe, and it is probable that the communities of organisms that today inhabit this area naturally have recolonised since the last ice age.

This phenomenon may also explain the low earthworm species diversity in Britain, compared to mainland Europe. In France, the earthworm fauna comprises 180 species (Bouché 1972), whereas only around 26 species currently exist in Britain, either native or introduced. This scarcity of species in Britain could reflect the slow dispersal rates of earthworms and the comparative short period of persistence of land connections between the British Isles and the continent, when mainland Europe species had the opportunity to colonize the UK (Sims and Gerard, 1999).

At the beginning of the Holocene period (10,000 BP), the early stages of soil reformation would have started in all the previously frozen areas. Rapidly rising temperatures, within a 50-year period, would have triggered secondary ecological successions, with pioneering species inhabiting the nascent soils. After ~3.000 years, large areas of fertile brown earth typical of temperate deciduous forest were widely distributed (Williams et al. 1998, Roberts 1998). Most colonization was achieved in a very rapid time frame, with oaks recorded in north Scotland within 2,000 years of ice retreat. Entire communities may have migrated North with the retreat of the ice, and recolonised suitable areas, or communities may have reassembled gradually from a mixture of first arrivals from southern refugia. In Europe, these refugia were Iberia,

Italy, the Balkans, Turkey and southwestern central Asia. A particularly common finding is that in northern populations is usual to observe a subsample of the genetic variation found in neighboring southern populations, and refugial areas have been found to possess a prevalence of private haplotypes as a consequence (Hewitt 2000, Hewitt 2004). However, in other taxa an increase in intraspecific diversity could in principle occur, due to multiple invasions from glacial refugia. This is probably the case of *L. rubellus* and other earthworm species. *L. rubellus* has been found to be divided into two highly distinct genetic lineages in Britain (King et al. 2008) that may even warrant the status of cryptic species.

Alternatively, this degree of cryptic diversity could be due to survival and recolonisation from a northern cryptic refugium or refugia (Stewart and Lister 2001), with consequent isolation and genetic drift between demes. There is evidence that some widely distributed earthworm species are moderately frost tolerant (Tiunov et al. 2006). These areas, probably characterised by buffered local microclimates, could have been suitable environments (Stewart and Lister 2001) for some *L. rubellus* demes. Earthworms could also have found suitable conditions in the vast emerged landscapes that surrounded the British Isles from west to east, and connected them with mainland Europe. These areas, referred as "Doggerland" in their eastern part (Coles 1999), were colonised by humans and Pleistocene fauna, and could have offered patches of suitable environment for some moderate cold-tolerant species. A suggestion of survival in cryptic refugia may come from a phenomenon called the "Lusitanian pattern" (Corbet 1961; Moore 1987; McDevitt et al. 2009), in which haplotypes shared between Iberia and Ireland suggest the possibility of ample areas of suitability and permeability to gene flow between these zones during the Pleistocene glaciations (Vega et al. 2010). Alternative explanations regarding this pattern point out to possible anthropogenic dispersal (McDevitt et al. 2011).

This study, the first for a Western Paleartic earthworm in the wider context of north-western postglacial history, will analyze and compare the genetic diversity of *L. rubellus* European populations over a wide spatial and temporal scale, thus trying to understand how and when soil invertebrate communities may have reassembled in the UK after the last Ice Age, and from where the recolonising populations putatively originated.

1.7 Phylogenomics

Phylogenetics – the character-based reconstruction of evolutionary relationships among taxa – follows directly from Darwin's evolutionary theory, perfectly exemplified by the sketch of a phylogenetic tree he drew in one of his notebooks (Darwin 1837). Today, this science relies on mathematical reconstruction methods to infer past relationships within and among contemporary species. The study of phylogenetics in conjunction with principles and processes shaping the geographic distribution of genetic lineages at the intraspecific level constitutes a branch of biological sciences called phylogeography (Avise 1998). Since the word itself has appeared in the literature (Avise et al. 1987), the discipline knew an exponential growth and, to date, the word gives back 44,600 hits in Google Scholar.

This growth would not have been possible without the invention of polymerase chain reaction (Saiki et al. 1985; Mullis and Faloona 1987) and the increased availability in commerce of thermal cycler machines which allowed efficient data generation of genetic sequences. However, Because of the limitation of Sanger sequencing methods, which involve sequencing of one gene at a time, classical phylogenetic and phylogeographic studies involved the use of a relatively small number of loci. Limits of this kind of approach have long been recognised (Edwards and Beerli 2000; Hudson and Turelli 2003) but the labour and costs to overcome these problems were a limiting factor for many years, until the advent of the new, modern sequencing technologies. Rather than a single template amplified by PCR, next generation sequencing (NGS), allows sequencing of libraries of template DNA, regardless of platform. NGS allows the parallel sequencing of million or billion of reads. The use of new sequencing technologies in phylogeography and phylogenetics is increasing (Carstens et al. 2012). This wealth of data is giving birth to a new research field, called phylogenomics. The term today describes the intersection between evolution and genomics (Eisen and Fraser 2003). A branch of this field is concerned with the use of genomic data to reconstruct the evolutionary relationships between organisms. Use of genomic data can solve many of the issues related to sample size in previous phylogenetic studies. A useful approach now rendered possible by phylogenomics, is the use of independent sequence features for phylogenetic inferences, such as "morphological-like" characters based on genomic structures (i.e. oligonucleotide fragments; Philippe et al. 2005).

In this study, a subset of individuals was sequenced with Illumina technology, to obtain whole genome data. The aim was to explore the relationship between the major *L. rubellus* cryptic lineages on a geographic and genetic distance gradient, examining the most similar lineages as well as more distant ones. The sequences and SNP data generated potentially allows the support of phylogenetic results obtained with the general mitochondrial monitoring of the sampled European populations, and also to examine demographic and evolutionary signals of divergence on an unprecedented scale for a clitellate model species.

1.8 Species distribution modelling

Species distribution models (SDM) combine concepts from ecology and natural history with recent advances in statistic and informatics (Elith and Leathwick 2009). The purpose of SDMs is to estimate the relationship between species distributions over a defined geographic range, and the environmental or spatial variables in that range (Franklin 2009).

Species distribution models can be also defined as habitat suitability models, as they describe the capacity of a particular habitat to support species survival. In this sense, SDMs are close to the concept of resource selection function (RSF), that include any model which output consists of values proportional to the probability of use of a resource unit, with the resource being a particular habitat for a particular species (Boyce et al. 2002; Franklin 2009). SDMs predict the likelihood of an event happening in a particular location, that is, the probability of species occurrence, conditional on the habitat characteristics of that location (Franklin 2009).

A good SDM framework must include the following elements (Torres-Villaça 2012):

- A theoretical model, controlling for spatial-temporal patterns of the species distribution, that takes into account different scales and the expected form of the prediction functions;
- Geo-referenced data about species occurrence, in the form of presence-only or presence-absence records, and GIS maps of environmental variables, representing factors relevant for the species;

- A modelling framework, defined by a set of rules and thresholds, with the function of linking habitat suitability with the information provided by environmental variables;
- A set of criteria for the validation of the models, as well as to deal with the errors and uncertainties produced by the analysis.

The resulting map of occurrence therefore aims to predict the niche suitability of the species. However, there are different kinds of niche concepts, and the debate on which kind of niche that SDM infers is still open (Phillips et al. 2006; Franklin 2009; Symonds and Moussalli 2011).

There is uncertainty whether SDM outputs reflect the realised niche, that is, the effective niche used by the species taking into account biotic interactions (Hutchinson 1957) or the fundamental niche, or potential distribution, which consists in the potential niche without taking into account other effects on species distribution and niche extent, such as biotic interactions, the effect of dispersal barriers or local extinctions due to human activities (Hutchinson 1957). Most studies consider that SDMs built on occurrence-only data reflect the potential niche, as they are built extrapolating from the environmental conditions recorded at points of species occurrence; that is, occurrence records are used to infer a statistical model of the realised niche of the species, an environmental "hypervolume" space that describes the realised niche, treating variables as dimensions in multi-dimensional space, independently from the spatial dimension (Elith and Leathwick 2009). When this model is extrapolated over geographical space, this realised niche would represent the potential niche of the species (Araujo and Guisan 2006). However, Phillips (2006) suggested that models based on occurrence-only data are a sample of the realised niche in the examined environmental dimensions; thus, they should be regarded as an approximation of the realised niche. This assumption is more realistic when taking into consideration slow moving taxa, such as earthworms.

SDM can be informed using variables (particularly, climatic variables) related to different time frames, to make predictions of the past or future niche of a species. The main assumption behind this method is that the species climatic niche is stable over time, and possible adaptive shifts of habitat preference/suitability in response to environmental changes are non significant over the time frames considered for the models (Nogués-Bravo 2009). However, studies indicate that for some taxa, niche

shifts have occurred over time, mainly because of competitions with different species and climate change (Pearman et al. 2008). However, in some cases, niche stability can be a consequence of constraints imposed by particular environments. In particular, the morphological stability observed in some deeply divergent earthworm lineages such as *L. rubellus*, leads us to assume that the evolutionary changes between lineages may not have had a significant impact on their environmental niche over time.

The development of SDM frameworks has been continuous during the last 20 years, and multiple methods are now available. The main division among these methods is related to what kind of data they use. Some frameworks perform better with presence/absence data, whereas others have been refined for the use of presence data only. Presence/absence data surveys consist of records where both presence and absence of the species are recorded and geo-referenced. Usually, regression methods such as Generalised Linear Models (GLM) and Generalised Additive Models (GAM) are used with this kind of data. Presence/absence datasets require systematic surveys, which usually tend to be sparse and/or limited in coverage for most taxa. However, for many species, a wealth of data is present in the form of presence-only data, recorded mainly by museums and herbariums. These records can sometimes represent over a century of monitoring. This bountiful source of data stimulated the development of many SDM methods, which require presence-data only: the most used are BIOCLIM (Eckert et al. 2008), currently implemented in the DIVA-GIS software (Hijmans et al. 2001) and still commonly used (Budge et al. 2013; Salamone et al. 2013), the "genetic algorithm framework" of GARP (Genetic Algorithm for Rule-set Production, Stockwell 1999) and the maximum entropy model implemented in Maxent (Phillips et al. 2006; Elith et al. 2011). Recently, a method for the use of GLM and GAM models with presence-only data has been developed. This method converts random points on the map, so-called "background points" into "pseudo-absence" values, to use in place of true absences in modelling construction (Phillips et al. 2006; Hijmans and Elith 2013).

1.8.1 Maxent

So far, the most common approach that uses presence-only data to build species distribution models is the Maximum Entropy modelling, implemented in the Maxent software (Phillips et al. 2006). Since its release, the program has been widely

used in SDM. Maxent calculates the probability distribution of maximum entropy, taking into account constraints that represent the incomplete information about the target distribution. The outcome is the probability distribution of species occurrence.

When applying Maxent to presence-only SDM:

- The pixels of the study area constitute the space on which the maxent probability distribution is calculated;
- The pixels with known species occurrence records constitute the sample points;
- The features are transformations of environmental variables (climatic variables, soil features, vegetation type or other environmental variables) usually provided in vector or raster form to the program (Phillips et al., 2006).

A very useful work describes Maxent principles and functioning using statistical terminology (Elith et al. 2011), rather than the machine learning approach appeared in the original description (Phillips et al., 2006). In this statistical approach, the independent environmental variables relevant to habitat suitability can be defined as covariates, predictors or inputs. As the answer to these variables for the modelled species can be complex, fitting non-linear functions is the best option. This is obtained by applying transformations to the covariates in a similar fashion to regression. In the original machine-learning framework in which Maxent has been described, such transformations (or set of transformations) of the original covariates are termed features. The approach described by Elith et al. (2011) aims to model the same quantity that is modelled with presence-absence data using presence-only data. This quantity is the probability of presence of a species.

When compared to other SDM frameworks, the Maxent method presents some advantages and some drawbacks (Phillips et al. 2006). The only requirements are environmental variables and presence-only information. The environmental variables can be continuous (e.g. temperature) or categorical (e. g. vegetation types). There is a great wealth of online resources to obtain both presence and environmental data. The Global Biodiversity Information Facility (GBIF, www.gbif.org) contains now more than 400 millions of indexed presence records: the Food and Agriculture Organization (FAO, http://www.fao.org/nr/land/soils/en/), the Global Climate Data facility (www.worldclim.org) and many more resources provide raster maps of environmental variables to be used for SDM. The algorithms are guaranteed to converge to the

optimal maximum entropy distribution, which should give the best estimate possible of habitat suitability even with small sample sizes (Phillips et al. 2006). Maxent also gives estimates of the variables effect on the species distribution in the form of variable response functions. The output is also continuous, enabling a fine-scale estimation of the suitability landscape. It also has a user-friendly Graphic User Interface (GUI) and recently, a package with Maxent functions has been implemented in the R software (the package dismo, Hijmans et al. 2011). The disadvantages of this approach are mainly related to the fact that is not as mature as GLM or GAM and it has issues of inflated predictability when environmental variables' values fall outside the range of the studied area (Phillips et al. 2006).

Maxent is nowadays one of the most widely used tools in ecological, evolutionary and biogeographic studies and the reason for this success probably lies in the fact that the program has been found to outperform nearly all the other SDM methods based on species presence (Elith et al. 2006). One of the main uses of Maxent in the literature is the estimation of species distribution in the past (Nogués-Bravo 2009; Vega et al. 2010; Svenning et al. 2011; Rebelo et al. 2012), in particular during the last glacial maximum (LGM). Some other studies focused on the time span between the last interglacial (LIG) and today (Nogués-Bravo 2009). Maxent has been widely used with genetic approaches and it has been recognised as a valuable, user-friendly tool for the implementation of SDM in statistical phylogeography (Richards et al. 2007).

This study applied Maxent SDM to infer *L. rubellus* niche and distribution in the present and in the last 120,000 years. The aim is to ascertain possible Mediterranean and cryptic glacial refugia, and infer the contribution of a set of soil and climatic variables on the current distribution of the species.

1.9 Molecular markers

1.9.1 Mitochondrial markers

For this research, a variety of molecular markers are potentially suitable, in order to answer questions on evolutionary history and cryptic variation of the species. The chosen markers include neutral mitochondrial (mt) DNA loci, such as cytochrome oxidase I (COI) and II (COII), which are used to assess the phylogenetic and phylogeographic diversity of the species. In addition, deep intraspecific phylogenetic signal was assessed, taking advantage of recent advances in sequencing technology, with the aim to explore relationships at the genomic level from eight individuals representative of the putative cryptic species complex, examining in first instance the structure and diversity of their whole mitochondrial genomes, and a wide array of Single Nucleotide Polymorphsms (SNPs) from their nuclear genomes.

MtDNA markers were chosen because of their putative neutrality and because they are well-characterized for the studied species (King et al. 2008). These markers can be applied to evaluate evolutionary rates, processes and constraints on molecular change through time or to make inferences about population processes, systematics and phylogeny. The mitochondrial genome of multicellular animals consists of a closed circular DNA molecule, and its usual size range is from 14,000 to 17,000 bp, with some exceptions (Moritz et al. 1987; Snyder et al. 1987; Wolstenholme 1992). In fact, although some rearrangements of mitochondrial genes have been found in different species, the overall structure, size and arrangement of genes is relatively conserved. Animal mtDNA contains 13 protein coding genes, 22 transfer RNAs and two ribosomal RNAs. A non-coding control region (D-Loop) contains sites for replication and transcription initiation.

There are several reasons why mtDNA markers are so popular. Firstly, isolating high-copy number mitochondrial loci is relatively simple, due to the availability of a number of "universal" primers (Folmer et al. 1994; Simons et al. 1994). Secondly, despite its conserved structure, the overall evolutionary rate of mtDNA is higher than that of the nuclear genome (Richter et al. 1988). The rate of synonymous substitutions in mtDNA has been estimated at 5.7×10^{-8} substitutions per site per year (Brown et al. 1982), which is up to ten times higher than the average rate of synonymous substitutions in nuclear protein-coding genes. The high mtDNA

evolutionary rate may be due partly to the oxidative stress of the mitochondrial environment, in combination with inefficient repair mechanisms, in comparison to those acting on nuclear DNA (Wilson et al. 1985). Furthermore, mtDNA is transmitted to the offspring by maternal uniparental inheritance, with few exceptions (e.g. (Schwartz and Vissing 2002; Kvist et al. 2003). This implies a general lack of recombination, as offspring inherit, in the absence of mutations, the same mitochondrial genome of their mother. In out-crossing hermaphrodites such as earthworms, there is simply an exchange of sperm between the breeders, but the assumption is still valid. This is the reason why mtDNA can be seen as a single haplotype that can be identified much more easily than nuclear lineages, which, in sexually reproducing species, are continuously pooling genes by recombination. The effectively clonal inheritance of mtDNA means that individual lineages can be tracked over time and space (Freeland 2005).

However, mitochondrial markers are limited by the fact that mtDNA effectively represents a single locus. Reconstructing population histories from a single locus can be problematic, if that locus has been subjected to evolutionary processes that may have given it an unusual history. For example, mitochondrial DNA can provide a distorted view of inter- and intraspecific hybridization events. Thus, in order to increase the level of accuracy, it is better to look for concordance with genealogies also obtained from the nuclear genome. Analyzing nuclear data is, however, more complicated, because recombination is common in the nuclear genome of sexually reproducing taxa. Nevertheless, recombination can often be detected using statistical methods (e.g. Holmes et al. 1999; Husmeier and Wright 2001; Li and Durbin 2011).

COI and COII markers are widely used to construct phylogenies. The genes encode for the protein subunits I and II of the cytochrome *c* oxidase complex, a protein complex involved in the respiration chain (Tsukihara et al. 1995). This important function implies a certain degree of conservation across phyla. For this reason, COI was proposed to be used as a genetic barcode for species identification (Hebert et al. 2003). The use of COI in earthworm identification has been recognised as capable of resolving the issues related to intraspecific cryptic variation, and an ideal tool to investigate the largely unknown biodiversity of soil animal communities (Klarica et al. 2012; Decaëns et al. 2013). COII often used in conjunction with COI (Roe and Sperling 2007; Jiang et al. 2013; Lee et al. 2013) also has a strong capacity

to discriminate species and cryptic lineages in earthworm studies (Andre et al. 2009; Fernández et al. 2012; Klarica et al. 2012). In conjunction, the two markers provide optimal tools for a large-scale study of *L. rubellus* genetic diversity.

1.9.2 Mitochondrial genomes

One of the aims of this study is to investigate the phylogenetic signal of cryptic divergence at the whole mitochondrial level. Whole mitochondrial genome sequences are a powerful tool in phylogenetic reconstruction. With advances in sequencing technologies and the parallel decrease in cost-per-base, whole mitochondrial sequences can be obtained, often as a by-product of whole genome sequencing projects (Nabholz et al. 2010). Whole mitochondrial phylogenies can be used to detect incongruences and improve the phylogenetic inferences obtained with single mitochondrial genes (Duchene et al. 2011). These authors point out that although different genes can carry out different signals, the use of less than a quarter of the mitochondrial genome can recover the correct phylogeny, suggesting that a proportion of the mitochondrial genes can be used to reconstruct adequate phylogenies even in absence of the complete sequence. Complete mitochondrial genomes are only available for a few annelids (Boore and Brown 1995; Boore and Brown 2000; Boore 2001; Jennings and Halanych 2005; Zhong et al. 2008; Shen et al. 2011; Won et al. 2013). Only one of these sequences belongs to a lumbricid oligochaete, *Lumbricus terrestris* (Boore and Brown 1995). All previous studies have attempted to address interspecific relationships with phylogenetic reconstructions but no mitogenomic work, to date, has tried to investigate relationships within an annelid cryptic species complex.

1.9.3 Genomic SNPs

The mitochondrial genomes obtained in this research area basically by-products of the whole-genome next-generation sequencing (NGS) of individuals representative of *L. rubellus* genetic diversity over its distribution range. Whole genome data from these individuals were used in order to collect a set of Single Nucleotide Polymorphisms (SNPs) representative of *L. rubellus* cryptic diversity at a deep scale. SNPs refer to single base pair positions along a DNA sequence that vary

between individuals. Diploid SNP profiles can be used to genetically characterize both individuals and populations. In the human genome, SNPs account for approximately the 90% of genetic variation (Collins et al. 1998). The extensive use of SNPs as molecular markers for evolutionary and phylogeographic studies was limited for a long time because of the resources required to characterize (Lai et al. 2007; Chen et al. 2008) and genotype them (van Orsouw et al. 2007; Van Tassell et al. 2008). However, advances in next generation sequencing technology allow today the rapid discovery of thousands of SNPs (Li et al. 2009).

The general principles behind NGS are the following. DNA is fragmented and libraries of fragments with a specific size are labelled with synthetic DNA adapters. The adapters are specific for each platform, and contain primer regions to amplify the library fragments during the process. The fragments are amplified *in situ* on a solid surface or a bead containing DNA fragments covalently attached to it, with sequences complementary to those of the adapter fragments. The amplification step allows the replication of DNA fragments around the originally attached fragment, so that cluster points containing a set of replicated sequences of the same fragment are formed. This amplification allows a sufficiently strong signal for the sequencing step. Millions of clusters, each one with a different fragment derived from the original template library at the centre, are distributed on the device surface; in case of beads, each bead is covered by the replicates of the original fragment. The successive sequencing step consists of 1) a nucleotide addition step and 2) a detection step, where the identity of the attached nucleotide is recorded at each cluster, and 3) a wash step, where fluorescent labels and blocking groups are removed. Rather than Sanger sequencing, where sequencing and detection are carried out as separate steps, in NGS the processes are contemporary. Furthermore, the process allows billions of clusters to be sequenced during runs. The denomination “massive parallel sequencing” is therefore appropriate (Mardis 2013). Although the technologies share the same principle, they do differ in some aspect. The technology used to generate the NGS data for this study was Illumina® (Figure 2).

Another main difference between Sanger data and NGS data is the length of the output sequences, also defined as read length. Sanger sequencing methods can amplify fragments up to 900-1000 bp, whilst NGS read length usually ranges from 50 - to 250 bp for Illumina, to 700 bp for some 454 platforms (Liu et al. 2012). The shorter length of NGS reads can be compensated by the fact that they can overlap,

based on the shared portions of their sequences, to obtain longer fragments. However, the low extent of shared sequence makes assemblies with short reads problematic. For this reason, assembly of longer contiguous sequences (contigs) can be problematic when the short-read approach is the only one taken.

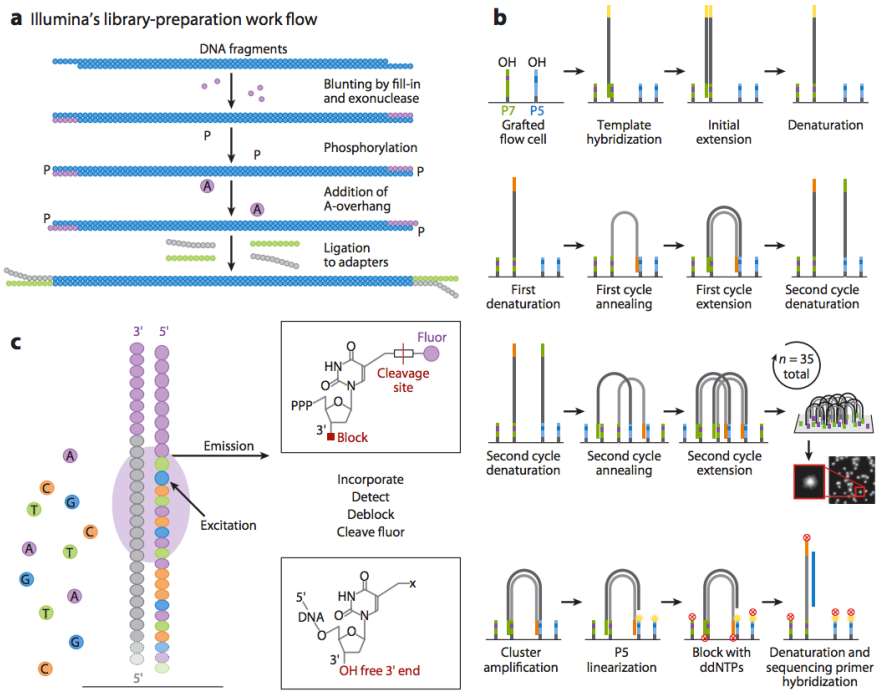


Figure 1.1 (a) Illumina library preparation. (b) Generation of clusters with bridge-amplification. (c) Description of sequencing by reversible-dye terminators. From Mardis (2013).

A way to overcome this limit has been the development of paired-end and mate-pair sequencing. In paired-end sequencing, both ends of a DNA fragment of less than 1kb are ligated to two different adapters and both are sequenced, and the two resulting reads are paired during the subsequent bioinformatic analyses, giving clear mapping information. In mate-pair sequencing, the fragments of DNA are longer than 1kb. Instead of ligating the fragment ends to two separate adapters, both ends of the fragment are ligated to the same adapter, circularising them. The circularised complex DNA-adapter is then fragmented, resulting in a series of linear fragments, which include the fragments containing the extremities of the original DNA complex, ligated in the middle with the adapter sequence. The adapters are then biotinylated, separated from the remaining DNA remnants and sequenced. The resulting reads are paired in the analyses and used in assemblies in the same way as paired-end reads above, but,

as they are representative of much longer fragments, they carry mapping information useful for aligning much longer contigs. The issue of read length, however, will probably be solved soon, as new technologies are constantly appearing, and the latest advances (e.g. Pac-Bio®) are capable of sequence fragments up to 22,000 bp in length (Quail et al. 2012).

After the read dataset is generated, it can be aligned to a reference genome, if available, carrying out whole genome mapping, using bioinformatic tools such as BWA (Li and Durbin 2009), Bowtie (Langmead et al. 2009), SAMtools (Li et al. 2009), or CLC Genomic Workbench (CLC Bio). If no previous genomic data is available, a *de-novo* sequencing pipeline is necessary. Also in this case, a wide array of softwares is available (Miller et al. 2010). The usual criterion is to carry out multiple test runs and evaluate *de novo* performance, usually using the N50 statistic (that is, a weighted median statistic such that 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value; Miller et al. 2010).

The main issues regarding NGS technologies are related to the quality of the data produced. In Sanger sequencing, the sequence obtained is usually a true "snapshot" of the target sequence, with coloured peaks individually evaluated to assess quality of base called. In NGS, with millions of reads generated, it is obviously impossible. The issue is addressed by increasing the coverage (the number of reads overlapping a portion of the sequenced genome) to decrease the possibility of identifying false variants. Another issue is represented by highly repetitive AT-rich portions, which can cause assembly failures in some genome portions, particularly for some technologies (e.g. Ion Torrent; Quail et al. 2012). With the exponential increase in data quantity generated, another limiting factor is related to the computational and human resources needed for data analysis (Chan and Ragan 2013). The assembly of a single large genome can be computationally expensive. Nevertheless, these technologies are revolutionising all the study fields involving genetics, including evolutionary biology, which is now experiencing the transition from studies based on Sanger-sequenced single or multiple loci to the use of genomic information to study evolutionary relationships, in the emerging disciplines of phylogenomics and population genomics.

1.9.4 The L. rubellus *genome project*

The year 2009 was the 200th anniversary of Darwin's Birthday, and 150th anniversary of its most famous book. Fittingly, in 2009 an earthworm, Darwin's favourite creature, became a genetic model organism for soil sciences. The *Lumbricus rubellus* genome project (Elsworth 2012) undertook the task to sequence and annotate the first earthworm genome. The project built on a previous, successful Expressed Sequence Tag (EST) sequencing project on the species transcriptome (Owen et al. 2008b), which data is available as an accessible database website (LumbriBASE) that can be queried for similarity and annotation studies. The specimen used came from a lineage B individual (S17) sampled from an abandoned lead mine site at Cwmystwyth (Wales), as a part of the population examined by Andre et al. (2010) in a study of the molecular differentiation of *L. rubellus* in a lead-contaminated site. The genome of *L. rubellus* is estimated to be 420 Mega bases in size, and distributed over 18 chromosome pairs (Gregory and Hebert 2002). It is still unpublished, but it is available for specific queries via request to the genome website (http://badger.bio.ed.ac.uk/earthworm/). Once published, the genome will be a remarkable resource for evolutionary and functional studies on oligochaete sentinel species and invertebrates. At the moment, the genome is in a 'high quality' draft form – that is, 90% of the genome is represented, it is clean from contaminating sequences and it is in appropriate form for general assessment of gene content (Chain et al. 2009), but it is still highly fragmented. To be published, the genome still needs data to scaffold the 315,000 contigs that compose its current version. This genome will be used in this study as a reference to map genomic data from *L. rubellus* individuals representative of its cryptic divergence, with the aim to assess the genomic extent of a cryptic speciation process and, if possible, to assess past signals of demographic change at the genomic level.

1.10 Aims and hypotheses

The general aim of this study was to investigate the distribution and extent of cryptic genetic and genomic variation within the *L. rubellus* over its native range, assess the effect of environmental variables on the species' distribution and investigate the depth of evolutionary diversity in this cryptic species complex. This PhD had the following specific goals:

i) To describe the mitochondrial genetic diversity of *L. rubellus* and its distribution across its native European range to identify lineages within the species complex;

ii) To localise these lineages, identify possible Pleistocene refugia and pathways of recolonisation;

iii) To assess the environmental variables that have had a major impact on its distribution;

iv) To analyse the evolutionary relationships within and among lineages using well-characterised single mitochondrial genes and compare these results to those using data from the whole mitochondrial genome;

v) To further investigate cryptic diversity within *L. rubellus* at the genomic level.

The following hypotheses were tested:

1) Genetic diversity in northern European populations of L. rubellus *is a subsample of the genetic diversity of southern European populations*

Previous studies on patterns of genetic differentiation caused by the Pleistocene Ice Ages showed that, as an effect of vicariance, genetic drift and post-glacial recolonisation from southern European glacial refugia, the genetic diversity of northern populations is commonly a subsample of the genetic diversity observed in southern refugial populations ('southern richness versus northern purity'; Hewitt 2000; Hewitt 2004). According to this common pattern, I expected to find a simpler genetic structure in populations sampled north of the southern Pleistocene refugia, as opposed to populations sampled in Iberia, Italy and the Balkans. I also expected to find representatives of all lineages in southern refugia, where they are supposed to

have survived during the Pleistocene and from where certain lineages recolonized northern latitudes.

To address this hypothesis, I undertook an extensive sampling programme over continental and insular Europe, involving collaborators from different countries. I 1) used mitochondrial genetic markers commonly used to investigate crypticity in earthworm species (King et al. 2008; Andre et al. 2009; Novo et al. 2010; Fernández et al. 2012), in a large survey to assess the distribution and extent of mitochondrial genetic diversity of *L. rubellus* over the studied range, and 2) selected specimens representative of mitochondrial clades over the considered geographic range, to determine the extent of this diversity at the genomic level.

2) *Genetic divergence of lineage A and B and possibly other major* L. rubellus *lineages occurred in the Pleistocene*

In previous studies regarding *L. rubellus* cryptic diversity, it has been speculated that the emergence of cryptic lineages was mainly related to vicariance and diversification processes occurring during the Pleistocene (King et al. 2008). Demographic signatures of population fluctuations showed that both lineages probably experienced the effect of Quaternary glaciations (Andre et al. 2009). I expected to find evidence of the period of this split in mitochondrial data.

To test this hypothesis, a multidisciplinary analysis was taken, using 1) the mitochondrial mutation rate consistent with previous studies with lumbricids (Chang et al. 2008) to estimate times of divergence and 2) MaxEnt environmental niche models were constructed regarding climatic conditions over the past 135,000 years and 3) coalescent simulations were carried out to test alternative hypothesis regarding survival and divergence in different refugia.

3) Climatic variables are the most important limiting factors in L. rubellus *distribution*

Decades of investigations on earthworm peregrine species widely distributed in the Holarctic zone have led researchers to conclude that climatic variables are the main limiting factors for earthworm species (Edwards 1996). In particular, dry and hot conditions seem to be much less tolerated than opposite extremes, with worms found thriving at latitudes where summer precipitation levels remain high and soils usually freeze in winter (Edwards 1996; Tiunov et al. 2006). Contrastingly, many

peregrine earthworm species are characterised by elevated tolerance to a broad range of soil factors usually capable to limit local life, including pH and heavy metal pollution (Sims and Gerard 1999; (Anderson et al. 2013). I tested whether climate is the primary limiting factor to explain past and present distributions of the epigeic *L. rubellus*, and a key to understand its past patterns of isolation and survival in refugia. To analyse this, a panel of climatic and soil variables were analysed with jackknife and variable importance tests within the MaxEnt framework.

4) Separate mitochondrial gene trees carry the same phylogenetic signal of whole-mitogenome trees

Mitochondrial DNA is widely accepted as a selectively neutral. The argument sustaining this is connected to the assumptions that the translational apparatus of a small genome is subject to less constraints than the nuclear system, which translates many thousands mRNAs (Cann et al. 1984). Additionally, support to this neutrality comes from evidence that the mean rate of divergence of mtDNA is 5 to 10 times greater than nuclear DNA in many taxa, including mammals and invertebrates (Brown et al. 1982). Therefore I expected to observe a constant phylogenetic signal across all the mitochondrial genes, and that signal must be consistent with a phylogeny obtained from the whole mitochondrial genome. To address this hypothesis, I analysed and annotated eight mitochondrial genomes obtained from a NGS experiment, with the objective to isolate genes and build different phylogenies from each mitochondrial gene alignment using a Bayesian phylogenetic framework, assessing distances and correlations between them using graphical (Nye 2008) numerical (Robinson and Foulds 1981) and statistical approaches.

5) The nuclear phylogenetic and demographic signals present similar patterns to those observed in mitochondrial data

The high degree of divergence at the mitochondrial level observed between lineage A and lineage B (King et al. 2008), in addition to the common observed pattern of cryptic divergence in annelids in general (Erséus and Gustafsson 2009), have led researchers to assume *L. rubellus* is a cryptic species complex. Such an hypothesis was worth further testing to explore divergence signals at the nuclear genome level, to assess the extent of genome segregation, or possible hybridisation events that may imply incomplete speciation (King et al. 2008). In addition, recent

techniques developed to exploit recombination information at the genomic level can be used to infer past changes in population size (Li and Durbin 2011). I argue that the cryptic speciation process regarding *L. rubellus* follows the same patterns at the nuclear level as the ones observed at the mitochondrial level, and that genomic signatures of demographic fluctuations in the past match with the ones observed in the mitochondrial genes. To test this consistency hypothesis, eight complete genomes, in the form of NGS Illumina paired-end reads, were mapped against the reference *L. rubellus* genome (Elsworth 2012). I extracted variant SNPs from the genome mappings and built phylogenetic trees to compare phylogenetic signals between nuclear and mitochondrial data. In addition, pairwise sequential Markov coalescent (PSMC) models (Li and Durbin 2011) were applied to study demographic fluctuations in past population size at the genomic level.

1.11 Bibliography

Adamowicz SJ, Menu-Marque S, Hebert PDN, Purvis A. 2007. Molecular systematics and patterns of morphological evolution in the Centropagidae (Copepoda: Calanoida) of Argentina. *Biological Journal of the Linnean Society* **90**: 279-292.

Addison JA. 2009. Distribution and impacts of invasive earthworms in Canadian forest ecosystems. In: Langor D, Sweeney J. *Ecological Impacts of Non-Native Invertebrates and Fungi on Terrestrial Ecosystems*, pp. 59-79. Springer.

Anderson CJ, Kille P, Lawlor AJ, Spurgeon DJ. 2013. Life-history effects of arsenic toxicity in clades of the earthworm *Lumbricus rubellus*. *Environmental Pollution* **172**: 200-207.

Andre J, King RA, Stürzenbaum SR, Kille P, Hodson M, Morgan AJ. 2009. Molecular genetic differentiation in earthworms inhabiting a heterogeneous Pb-polluted landscape. *Environmental Pollution* **158**: 883-890.

Araujo MB, Guisan A. 2006. Five (or so) challenges for species distribution modelling. *Journal of Biogeography* **33**: 1677-1688.

Arnaud JÄ, Madec L, Guiller A, Bellido A. 2001. Spatial analysis of allozyme and microsatellite DNA polymorphisms in the land snail *Helix aspersa* (Gastropoda: Helicidae). *Molecular Ecology* **10**: 1563-1576.

Avise JC. 1998. The history and purview of phylogeography: a personal reflection. *Molecular Ecology* **7**: 371-379.

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics* **18**: 489-522.

Bach L, Palmqvist A, Rasmussen LJ, Forbes VE. 2005. Differences in PAH tolerance between *Capitella* species: Underlying biochemical mechanisms. *Aquatic Toxicology* **74**: 307-319.

Baker AM, Bartlett C, Bunn SE, Goudkamp K, Sheldon F, Hughes JM. 2003. Cryptic species and morphological plasticity in long-lived bivalves (Unionoida: Hyriidae) from inland Australia. *Molecular Ecology* **12**: 2707-2717.

Beauchamp KA, Kathman RD, McDowell TS, Hedrick RP. 2001. Molecular phylogeny of tubificid oligochaetes with special emphasis on *Tubifex tubifex* (Tubificidae). *Molecular Phylogenetics and Evolution* **19**: 216-224.

Böhme M, Ilg A, Winklhofer M. 2008. Late Miocene "washhouse" climate in Europe. *Earth and Planetary Science Letters* **275**: 393-401.

Boore JL. 2001. Complete mitochondrial genome sequence of the polychaete annelid *Platynereis dumerilii*. *Molecular Biology and Evolution* **2**: 1413-1416.

Boore JL, Brown WM. 1995. Complete sequence of the mitochondrial DNA of the annelid worm *Lumbricus terrestris*. *Genetics* **141**: 305-317.

Boore JL, Brown WM. 2000. Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: sequence and gene arrangement comparisons indicate that Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Molecular Biology and Evolution* **17**: 87-106.

Bouché MB. 1972. Lombriciens de France. Écologie et Systématique (n hors-série), Institut National de la Recherche Agronomique. in *Annales de Zoologie-Écologie Animale*.

Boyce MS, Vernier PR, Nielsen SE, Schmiegelow FKA. 2002. Evaluating resource selection functions. *Ecological Modelling* **157**: 281-300.

Brinkhurst RO, Jamieson BGM, Cook DG, Anderson DV, van der Land J. 1971. *Aquatic Oligochaeta of the world*. University of Toronto Press Toronto.

Brown PJ, Long SM, Spurgeon DJ, Svendsen C, Hankard PK. 2004. Toxicological and biochemical responses of the earthworm *Lumbricus rubellus* to pyrene, a non-carcinogenic polycyclic aromatic hydrocarbon. *Chemosphere* **57**: 1675-1681.

Brown WM, Prager EM, Wang A, Wilson AC. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution* **18**: 225-239.

Brulle F, Morgan AJ, Cocquerelle C, Vandenbulcke F. 2010. Transcriptomic underpinning of toxicant-mediated physiological function alterations in three terrestrial invertebrate taxa: A review. *Environmental Pollution* **158**: 2793-2808.

Buckley TR, James S, Allwood J, Bartlam S, Howitt R, Prada D. 2011. Phylogenetic analysis of New Zealand earthworms (Oligochaeta: Megascolecidae) reveals

ancient clades and cryptic taxonomic diversity. *Molecular Phylogenetics and Evolution* **58**: 85-96.

Budge PJ, Little KM, Mues KE, Kennedy ED, Prakash A, Rout J, Fox LM. 2013. Impact of community-based lymphedema management on perceived disability among patients with lymphatic filariasis in Orissa State, India. *PLoS Neglected Tropical Diseases* **7**: e2100.

Bundy JG, Davey MP, Viant MR. 2009. Environmental metabolomics: a critical review and future perspectives. *Metabolomics* **5**: 3-21.

Bundy JG, Sidhu JK, Rana F, Spurgeon DJ, Svendsen C, Wren JF, Stürzenbaum SR, Morgan AJ, Kille P. 2008. 'Systems toxicology' approach identifies coordinated metabolic responses to copper in a terrestrial non-model invertebrate, the earthworm *Lumbricus rubellus*. *BMC Biology* **6**: 25.

Cann RL, Brown WM, Wilson AC. 1984. Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. *Genetics* **106**: 479-499.

Carignan V, Villard M-A. 2002. Selecting indicator species to monitor ecological integrity: a review. *Environmental Monitoring and Assessment* **78**: 45-61.

Caro TM, O'Doherty G. 1999. On the use of surrogate species in conservation biology. *Conservation Biology* **13**: 805-814.

Carstens B, Lemmon AR, Lemmon EM. 2012. The promises and pitfalls of next-generation sequencing data in phylogeography. *Systematic Biology* **61**: 713-715.

Chain PSG, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C. 2009. Genome project standards in a new era of sequencing. *Science* **326**: 236-237.

Chan CX, Ragan MA. 2013. Next-generation phylogenomics. *Biology Direct* **8**: 1-6.

Chang CH, Lin SM, Chen JH. 2008. Molecular systematics and phylogeography of the gigantic earthworms of the *Metaphire formosae* species group (Clitellata, Megascolecidae). *Molecular Phylogenetics and Evolution* **49**: 958-968.

Chapman PM. 2002. Integrating toxicology and ecology: putting the "eco" into ecotoxicology. *Marine Pollution Bulletin* **44**: 7-15.

Chen D, Ahlford A, Schnorrer F, Kalchhauser I, Fellner M, Viràgh E, Kiss I, Syvänen A-C, Dickson BJ. 2008. High-resolution, high-throughput SNP mapping in *Drosophila melanogaster*. *Nature Methods* **5**: 323-329.

Chenon P, Rousset A, Crouau Y. 2000. Genetic polymorphism in nine clones of a parthenogenetic collembolan used in ecotoxicological testing. *Applied Soil Ecology* **14**: 103-110.

Coates KA. 1990. Marine Enchytraeidae (Oligochaeta, Annelida) of the Albany area, Western Australia. *The marine flora and fauna of Albany, Western Australia* **1**: 13-41.

Coles BJ. 1999. *Doggerland's loss and the Neolithic*. WARP Occasional Paper Exeter.

Collins FS, Brooks LD, Chakravarti A. 1998. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research* **8**: 1229.

Corbet GB. 1961. Origin of the british insular races of small mammals and of the Lusitanian Fauna. *Nature* **191**: 1037-1040.

Costello DM, Lamberti GA. 2008. Non-native earthworms in riparian soils increase nitrogen flux into adjacent aquatic ecosystems. *Oecologia* **158**: 499-510.

Darwin C. 1837. Notebook B. *University Library, Cambridge University*: 139.

Darwin C. 1892. *The Formation of Vegetable Mould, Through the Action of Worms, with Observations on their Habits*. J. Murray.

Decaëns T, Jiménez J, Gioia C, Measey G, Lavelle P. 2006. The values of soil animals for conservation biology. *Journal of Soil Biology* **42**: S23-S38.

Decaëns T, Porco D, Rougerie R, Brown GG, James SW. 2013. Potential of DNA barcoding for earthworm research in taxonomy and ecology. *Applied Soil Ecology* **65**: 35-42.

Depledge MH, Aagaard A, Györkös P. 1995. Assessment of trace metal toxicity using molecular, physiological and behavioural biomarkers. *Marine Pollution Bulletin* **31**: 19-27.

Donnelly RK. 2011. An investigation of genetic heterogeneity in a biological sentinel species (*Lumbricus rubellus*). University of Glamorgan/Cardiff University, Cardiff.

Donnelly RK, Harper GL, Morgan AJ, Orozco-terWengel P, Juma GAP, Bruford MW. In Press. Recapitulation of cryptic lineages of *Lumbricus rubellus*. *Biological Journal of the Linnean Society*.

Duchene S, Archer FI, Vilstrup J, Caballero S, Morin PA. 2011. Mitogenome phylogenetics: the impact of using single regions and partitioning schemes on

topology, substitution rate and divergence time estimation. *PloS One* **6**: e27138.

Dupont L, Lazrek F, Porco D, King RA, Rougerie R, Symondson WOC, Livet A, Richard B, Decaëns T, Butt KR. 2011. New insight into the genetic structure of the *Allolobophora chlorotica* aggregate in Europe using microsatellite and mitochondrial data. *Pedobiologia* **54**: 217-224.

Eason C, O'Halloran K. 2002. Biomarkers in toxicology versus ecological risk assessment. *Toxicology* **181**: 517-521.

Eckert CG, Samis KE, Lougheed SC. 2008. Genetic variation across species' geographical ranges: the central-marginal hypothesis and beyond. *Molecular Ecology* **17**: 1170-1188.

Edwards C. 2004. The importance of earthworms as key representatives of the soil fauna. In: Edwards, CA. *Earthworm Ecology*, pp. 3-11. CRC Press.

Edwards CA. 1996. *Biology and Ecology of Earthworms*. Springer.

Edwards S, Beerli P. 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* **54**: 1839-1854.

Eisen JA, Fraser CM. 2003. Phylogenomics: intersection of evolution and genomics. *Science* **300**: 1706-1707.

Elith J, H. Graham C, P. Anderson R, Dudík M, Ferrier S, Guisan A, J. Hijmans R, Huettmann F, R. Leathwick J, Lehmann A, Li J, G. Lohmann L, A. Loiselle B, Manion G, Moritz C, Nakamura M, Nakazawa Y, McC. M. Overton J, Townsend Peterson A, J. Phillips S, Richardson K, Scachetti-Pereira R, E. Schapire R, Soberón J, Williams S, S. Wisz M, E. Zimmermann N. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**: 129-151.

Elith J, Leathwick JR. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* **40**: 677-697.

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* **17**: 43-57.

Elsworth B. 2012. Unearthing the genome of the earthworm *Lumbricus rubellus*. The University of Edinburgh, Edinburgh.

Erséus C. 1988. Taxonomic revision of the *Phallodrilus rectisetosus* complex (Oligochaeta: Tubificidae). *Proceedings of the biological Society of Washington* **101**: 784-793.

Erséus C. 1997. Marine Tubificidae (Oligochaeta) from the Montebello and Houtman Abrolhos Islands, Western Australia, with descriptions of twenty-three new species. *The marine flora and fauna of the Houtman Abrolhos Islands, Western Australia Western Australian Museum, Perth*: 389-459.

Erséus C, Gustafsson D. 2009. Cryptic speciation in clitellate model organisms. In: Shain DA. *Annelids in Modern Biology*: 31-46. Wiley-Blackwell.

Evans AC, Guild WJ. 1948. Studies on the relationships between earthworms and soil fertility. IV. On the life cycles of some British Lumbricidae. *Annals of Applied Biology* **35**: 471-484.

Fernández R, Almodóvar A, Novo M, Simancas B, J. Díaz Cosín D. 2012. Adding complexity to the complex: New insights into the phylogeny, diversification and origin of parthenogenesis in the *Aporrectodea caliginosa* species complex (Oligochaeta, Lumbricidae). *Molecular Phylogenetics and Evolution* **64**: 368-379.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* **3**: 294.

Francioni E, Wagener A, Scofield AL, Depledge MH, Cavalier B. 2007. Evaluation of the mussel *Perna perna* as a biomonitor of polycyclic aromatic hydrocarbon (PAH) exposure and effects. *Marine Pollution Bulletin* **54**: 329-338.

Franklin J. 2009. *Mapping Species Distributions: Spatial Inference and Prediction*. Cambridge University Press, Cambridge, UK.

Fränzle O. 2006. Complex bioindication and environmental stress assessment. *Ecological Indicators* **6**: 114-136.

Freeland J. 2005. *Molecular Ecology*. Wiley and Sons, Chichester.

Graff O. 1953. Investigations in soil zoology with special reference to the terricole Oligochaeta. *Zeitschrift für Pflanzenernährung, Düngung und Bodenkunde* **61**: 12-22.

Gregory TR, Hebert PDN. 2002. Genome size estimates for some oligochaete annelids. *Canadian journal of zoology* **80**: 1485-1489.

Gustafsson DR, Price DA, Erséus C. 2009. Genetic variation in the popular lab worm *Lumbriculus variegatus* (Annelida: Clitellata: Lumbriculidae) reveals cryptic speciation. *Molecular Phylogenetics and Evolution* **51**: 182-189.

Hamers T, Van den Berg JHJ, Van Gestel CAM, Van Schooten F-J, Murk AJ. 2006. Risk assessment of metals and organic pollutants for herbivorous and carnivorous small mammal food chains in a polluted floodplain (Biesbosch, The Netherlands). *Environmental Pollution* **144**: 581-595.

Hebert PDN, Cywinska A, Ball SL. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Series B: Biological Sciences* **270**: 313-321.

Heemsbergen DA, Berg MP, Loreau M, Van Hal JR, Faber JH, Verhoef HA. 2004. Biodiversity effects on soil processes explained by interspecific functional dissimilarity. *Science* **306**: 1019-1020.

Heethoff M, Etzold K, Scheu S. 2004. Mitochondrial COII sequences indicate that the parthenogenetic earthworm *Octolasion tyrtaeum* (Savigny 1826) constitutes of two lineages differing in body size and genotype. *Pedobiologia* **48**: 9-13.

Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. *Nature* **405**: 907-913.

Hewitt GM. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society* **58**: 247-276.

Hewitt GM. 2004. Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* **359**: 183-195.

Heywood VH. 1995. *Global Biodiversity Assessment*. Cambridge University Press.

Hijmans R, Cruz M, Rojas E, Guarino L. 2001. DIVA-GIS version 1.4: A geographic information system for the analysis of biodiversity data, manual.

Hijmans RJ, Elith J. 2013. Species distribution modeling with R. R Project.

Hijmans RJ, Phillips S, Leathwick J, Elith J. 2011. Package 'Dismo'. Available online at: http://cran.r-project.org/web/packages/dismo/index.html.

Hilty J, Merenlender A. 2000. Faunal indicator taxa selection for monitoring ecosystem health. *Biological Conservation* **92**: 185-197.

Hodson ME. 2013. Effects of heavy metals and metalloids on soil organisms. in *Heavy Metals in Soils*, pp. 141-160. Springer.

Holmes E, Worobey M, Rambaut A. 1999. Phylogenetic evidence for recombination in dengue virus. *Molecular Biology and Evolution* **16**: 405-409.

Holmquist C. 1983. What is *Tubifex tubifex* (OF Müller)(Oligochaeta, Tubificidae)? *Zoologica Scripta* **12**: 187-201.

Hudson RR, Turelli M. 2003. Stochasticity overrules the "three-times rule": genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. *Evolution* **57**: 182-190.

Husmeier D, Wright F. 2001. Probabilistic divergence measures for detecting interspecies recombination. *Bioinformatics* **17**: S123-131.

Hutchinson GE. 1957. Concluding remarks. in *Symposium on Quantitative Biology*, pp. 415-427, Cold Spring Harbor

Ireland MP. 1983. Heavy metal uptake and tissue distribution in earthworms. In: Edwards CA. *Earthworm Ecology*, pp. 247-265. Springer.

ITIS. 2012. ITIS report for Lumbricina, taxonomic serial No.: 69069. Retrieved May 14, 2012.

James SW, Porco D, Decaëns T, Richard B, Rougerie R, Erséus C. 2010. DNA barcoding reveals cryptic diversity in *Lumbricus terrestris* L., 1758 (Clitellata): resurrection of *L. herculeus* (Savigny, 1826). *PloS One* **5**: e15629.

Jennings RM, Halanych KM. 2005. Mitochondrial genomes of *Clymenella torquata* (Maldanidae) and *Riftia pachyptila* (Siboglinidae): evidence for conserved gene order in annelida. *Molecular Biology and Evolution* **22**: 210-222.

Jiang F, Li ZH, Deng YL, Wu JJ, Liu RS, Buahom N. 2013. Rapid diagnosis of the economically important fruit fly, *Bactrocera correcta* (Diptera: Tephritidae) based on a species-specific barcoding cytochrome oxidase I marker. *Bulletin of Entomological Research* **103**: 363-371.

Jones CG, Lawton JH, Shachak M. 1997. Positive and negative effects of organisms as physical ecosystem engineers. *Ecology* **78**: 1946-1957.

Jouquet P, Dauber J, Lagerlöf J, Lavelle P, Lepage M. 2006. Soil invertebrates as ecosystem engineers: intended and accidental effects on soil and feedback loops. *Applied Soil Ecology* **32**: 153-164.

King RA, Tibble AL, Symondson WOC. 2008. Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Molecular Ecology* **17**: 4684-4698.

Klarica J, Kloss-Brandstätter A, Traugott M, Juen A. 2012. Comparing four mitochondrial genes in earthworms – Implications for identification, phylogenetics, and discovery of cryptic species. *Soil Biology and Biochemistry* **45**: 23-30.

Kvist L, Martens J, Nazarenko AA, Orell M. 2003. Paternal leakage of mitochondrial DNA in the great tit (*Parus major*). *Molecular Biology and Evolution* **20**: 243-247.

Lai C-Q, Leips J, Zou W, Roberts JF, Wollenberg KR, Parnell LD, Zeng Z-B, Ordovas JM, Mackay TFC. 2007. Speed-mapping quantitative trait loci using microarrays. *Nature Methods* **4**: 839-841.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: R25.

Lavelle P, Bignell D, Lepage M, Wolters W, Roger P, Ineson P, Heal OW, Dhillion S. 1997. Soil function in a changing world: the role of invertebrate ecosystem engineers. *European Journal Of Soil Biology* **33**: 159-193.

LeBlanc GA, Bain LJ. 1997. Chronic toxicity of environmental contaminants: sentinels and biomarkers. *Environmental Health Perspectives* **105**: 65.

Lee CE, Frost BW. 2002. Morphological stasis in the *Eurytemora affinis* species complex (Copepoda: Temoridae). *Hydrobiologia* **480**: 111-128.

Lee KE. 1987. Peregrine species of earthworms. *On Earthworms*: 315-327.

Lee W, Park J, Lee G-S, Lee S, Akimoto S-i. 2013. Taxonomic status of the *Bemisia tabaci* complex (Hemiptera: Aleyrodidae) and reassessment of the number of its constituent species. *PloS One* **8**: e63817.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493-496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. 2012. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* **2012**: 11.

Lubbers IM, van Groenigen KJ, Fonte SJ, Six J, Brussaard L, van Groenigen JW. 2013. Greenhouse-gas emissions from soils increased by earthworms. *Nature Climate Change*.

Lukkari T, Haimi J. 2005. Avoidance of Cu- and Zn-contaminated soil by three ecologically different earthworm species. *Ecotoxicology and Environmental Safety* **62**: 35-41.

Mardis ER. 2013. Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*.

Marinissen JCY, Van den Bosch F. 1992. Colonization of new habitats by earthworms. *Oecologia* **91**: 371-376.

McDevitt AD, Rambau RV, O'Brien J, McDevitt CD, Hayden TJ, Searle JB. 2009. Genetic variation in Irish pygmy shrews *Sorex minutus* (Soricomorpha: Soricidae): implications for colonization history. *Biological Journal of the Linnean Society* **97**: 918-927.

McDevitt AD, Vega R, Rambau RV, Yannic G, Herman JS, Hayden TJ, Searle JB. 2011. Colonization of Ireland: revisiting 'the pygmy shrew syndrome' using mitochondrial, Y chromosomal and microsatellite markers. *Heredity* **107**: 548-557.

Mhatre GN, Pankhurst CE, Pankhurst C, Doube BM, Gupta V. 1997. Bioindicators to detect contamination of soils with special reference to heavy metals. *Biological Indicators of Soil Health*: 349-369.

Michaelsen W. 1903. *Die geographishe Verbreitung der Oligochaeten*, Berlin.

Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95**: 315-327.

Minnich J. 1977. *The earthworm book: How to raise and use earthworms for your farm and garden*. Rodale Press.

Moore PD. 1987. Snails and the Irish question. *Nature* **328**: 381-382.

Morgan AJ, Sturzenbaum SR, Winters C, Grime GW, Aziz NAA, Kille P. 2004. Differential metallothionein expression in earthworm (*Lumbricus rubellus*) tissues. *Ecotoxicology and Environmental Safety* **57**: 11-19.

Morgan JE, Morgan AJ. 1988. Calcium-lead interactions involving earthworms. Part 2: The effect of accumulated lead on endogenous calcium in *Lumbricus rubellus*. *Environmental Pollution* **55**: 41-54.

Morgan JE, Morgan AJ. 1990. The distribution of cadmium, copper, lead, zinc and calcium in the tissues of the earthworm *Lumbricus rubellus* sampled from one uncontaminated and four polluted soils. *Oecologia* **84**: 559-566.

Morgan JE, Morgan AJ. 1998. The distribution and intracellular compartmentation of metals in the endogeic earthworm *Aporrectodea caliginosa* sampled from an unpolluted and a metal-contaminated site. *Environmental Pollution* **99**: 167-175.

Moritz C, Dowling T, Brown W. 1987. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology and Systematics* **18**: 269-292.

Mullis KB, Faloona FA. 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in enzymology* **155**: 335.

Nabholz B, Jarvis ED, Ellegren H. 2010. Obtaining mtDNA genomes from next-generation transcriptome sequencing: a case study on the basal Passerida (Aves: Passeriformes) phylogeny. *Molecular Phylogenetics and Evolution* **57**: 466-470.

Nogués-Bravo D. 2009. Predicting the past distribution of species climatic niches. *Global Ecology and Biogeography* **18**: 521-531.

Nordström S, Rundgren S. 1974. Environmental factors and lumbricid associations in southern Sweden. *Pedobiologia* **14**: 1-27.

Novo M, Almodóvar A, Díaz-Cosín DJ. 2009. High genetic divergence of hormogastrid earthworms (Annelida, Oligochaeta) in the central Iberian Peninsula: evolutionary and demographic implications. *Zoologica scripta* **38**: 537-552.

Novo M, Almodóvar A, Fernández R, Trigo D, Díaz Cosín DJ. 2010. Cryptic speciation of hormogastrid earthworms revealed by mitochondrial and nuclear data. *Molecular Phylogenetics and Evolution* **56**: 507-512.

Nye TMW. 2008. Trees of trees: an approach to comparing multiple alternative phylogenies. *Systematic Biology* **57**: 785-794.

Owen J, Ann BA, Svendsen C, Wren J, Jonker MJ, Hankard PK, Lister LJ, Stürzenbaum SR, Morgan AJ, Spurgeon DJ, Blaxter ML, Kille P. 2008a. Transcriptome profiling of developmental and xenobiotic responses in a keystone soil animal, the oligochaete annelid *Lumbricus rubellus*. *BMC Genomics* **9**.

Owen J, Hedley BA, Svendsen C, Wren J, Jonker M, Hankard P, Lister L, Sturzenbaum S, Morgan AJ, Spurgeon D, Blaxter M, Kille P. 2008b. Transcriptome profiling of developmental and xenobiotic responses in a keystone soil animal, the oligochaete annelid *Lumbricus rubellus*. *BMC genomics* **9**: 266.

Paoletti MG. 1999. Using bioindicators based on biodiversity to assess landscape sustainability. *Agriculture, Ecosystems & Environment* **74**: 1-18.

Parmelee RW, Crossley Jr DA. 1988. Earthworm production and role in the nitrogen cycle of a no-tillage agroecosystem on the Georgia Piedmont. *Pedobiologia* **32**: 353-361.

Pearman PB, Guisan A, Broennimann O, Randin CF. 2008. Niche dynamics in space and time. *Trends in Ecology & Evolution* **23**: 149-158.

Peles JD, Towler WI, Guttman SI. 2003. Population genetic structure of earthworms (*Lumbricus rubellus*) in soils contaminated by heavy metals. *Ecotoxicology* **12**: 379-386.

Pérez-Losada M, Eiroa J, Mato S, Domínguez J. 2005. Phylogenetic species delimitation of the earthworms *Eisenia fetida* (Savigny, 1826) and *Eisenia andrei* Bouché, 1972 (Oligochaeta, Lumbricidae) based on mitochondrial and nuclear DNA sequences. *Pedobiologia* **49**: 317-324.

Pérez-Losada M, Ricoy M, Marshall JC, Domínguez J. 2009. Phylogenetic assessment of the earthworm *Aporrectodea caliginosa* species complex (Oligochaeta: Lumbricidae) based on mitochondrial and nuclear DNA sequences. *Molecular Phylogenetics and Evolution* **52**: 293-302.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*: 541-562.

Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**: 231-259.

Pinceel J, Jordaens K, Backeljau T. 2005a. Extreme mtDNA divergences in a terrestrial slug (Gastropoda, Pulmonata, Arionidae): accelerated evolution, allopatric divergence and secondary contact. *Journal of Evolutionary Biology* **18**: 1264-1280.

Pinceel J, Jordaens K, Pfenninger M, Backeljau T. 2005b. Rangewide phylogeography of a terrestrial slug in Europe: evidence for Alpine refugia

and rapid colonization after the Pleistocene glaciations. *Molecular Ecology* **14**: 1133-1150.

Pisias NG, Moore Jr TC. 1981. The evolution of Pleistocene climate: a time series approach. *Earth and Planetary Science Letters* **52**: 450-458.

Pop AA, Wink M, Pop VV. 2003. Use of 18S, 16S rDNA and cytochrome *c* oxidase sequences in earthworm taxonomy (Oligochaeta, Lumbricidae): The 7th international symposium on earthworm ecology · Cardiff · Wales · 2002. *Pedobiologia* **47**: 428-433.

Quail M, Smith M, Coupland P, Otto T, Harris S, Connor T, Bertoni A, Swerdlow H, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics* **13**: 341.

Quirós L, Piña B, Solé M, Blasco J, López MA, Riva MC, Barceló D, Raldúa D. 2007. Environmental monitoring by gene expression biomarkers in *Barbus graellsii*: laboratory and field studies. *Chemosphere* **67**: 1144.

Rebelo H, Froufe E, Brito JC, Russo D, Cistrone L, Ferrand N, Jones G. 2012. Postglacial colonization of Europe by the barbastelle bat: agreement between molecular data and past predictive modelling. *Molecular Ecology* **21**: 2761-2774.

Richards CL, Carstens BC, Lacey Knowles L. 2007. Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography* **34**: 1833-1845.

Richter C, Park J-W, Ames BN. 1988. Normal oxidative damage to mitochondrial and nuclear DNA is extensive. *Proceedings of the National Academy of Sciences* **85**: 6465-6467.

Ricketts HJ, Morgan AJ, Spurgeon DJ, Kille P. 2004. Measurement of annetocin gene expression: a new reproductive biomarker in earthworm ecotoxicology. *Ecotoxicology and Environmental Safety* **57**: 4-10.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* **53**: 131-147.

Roe AD, Sperling FAH. 2007. Patterns of evolution of mitochondrial cytochrome *c* oxidase I and II DNA and implications for DNA barcoding. *Molecular Phylogenetics and Evolution* **44**: 325-345.

Rota E, Wang H, Erséus C. 2007. The diverse *Grania* fauna (Clitellata: Enchytraeidae) of the Esperance area, Western Australia, with descriptions of two new species. *Journal of Natural History* **41**: 999-1023.

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**: 1350-1354.

Salamone I, Govindarajulu R, Falk S, Parks M, Liston A, Ashman T-L. 2013. Bioclimatic, ecological, and phenotypic intermediacy and high genetic admixture in a natural hybrid of octoploid strawberries. *American Journal of Botany* **100**: 939-950.

Schaefer M, Schauermann J. 1990. The soil fauna of beech forests: comparison between a mull and a moder soil. *Pedobiologia* **34**: 299-314.

Schlick-Steiner BC, Steiner FM, Moder K, Seifert B, Sanetra M, Dyreson E, Stauffer C, Christian E. 2006. A multidisciplinary approach reveals cryptic diversity in Western Palearctic *Tetramorium* ants (Hymenoptera: Formicidae). *Molecular Phylogenetics and Evolution* **40**: 259-273.

Schwartz M, Vissing J. 2002. Paternal Inheritance of Mitochondrial DNA. *New England Journal of Medicine* **347**: 576-580.

Shen X, Wu Z, Sun Ma, Ren J, Liu B. 2011. The complete mitochondrial genome sequence of *Whitmania pigra* (Annelida, Hirudinea): The first representative from the class Hirudinea. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics* **6**: 133-138.

Simons C, Frati F, Beckenbach A, Crespi B, Liu H, Floors P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the entomological Society of America* **87**: 651-701.

Simonsen V, Holmstrup M, Niklasson M. 2004. Genetic differentiation of the parthenogenetic soil collembolan *Isotoma notabilis* along a copper gradient based on random amplified polymorphic DNA. *Pedobiologia* **48**: 297-303.

Simonsen V, Klok C. 2010. Genetic and ecological impacts of heavy metal and flooding stress on the earthworm *Lumbricus rubellus* in floodplains of the Rhine river. *Soil Biology and Biochemistry* **42**: 270-275.

Simonsen V, Scott-Fordsmand JJ. 2004. Genetic variation in the enzyme esterase, bioaccumulation and life history traits in the earthworm *Lumbricus rubellus*

from a metal contaminated area, Avonmouth, England. *Ecotoxicology* **13**: 773-786.

Sims R, Gerard B. 1999. *Earthworms*. Linnean Society, London.

Snyder M, Fraser AR, LaRoche J, Gartner-Kepkay KE, Zouros E. 1987. Atypical mitochondrial DNA from the deep-sea scallop *Placopecten magellanicus*. *Proceedings of the National Academy of Sciences* **84**: 7595-7599.

Spurgeon DJ, Morgan AJ, Kille P. 2008. Current research in soil invertebrate ecotoxicogenomics. *Advances in Experimental Biology* **2**: 133-326.

Spurgeon DJ, Ricketts H, Svendsen C, Morgan AJ, Kille P. 2005. Hierarchical responses of soil invertebrates (earthworms) to toxic metal stress. *Environmental Science & Technology* **39**: 5327-5334.

Spurgeon DJ, Stürzenbaum SR, Svendsen C, Hankard PK, Morgan AJ, Weeks JM, Kille P. 2004. Toxicological, cellular and gene expression responses in earthworms exposed to copper and cadmium. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* **138**: 11-21.

Spurgeon DJ, Weeks JM, Van Gestel CAM. 2003. A summary of eleven years progress in earthworm ecotoxicology: The 7th international symposium on earthworm ecology · Cardiff · Wales · 2002. *Pedobiologia* **47**: 588-606.

Stewart JR, Lister AM. 2001. Cryptic northern refugia and the origins of the modern biota. *Trends in Ecology & Evolution* **16**: 608-613.

Stewart JR, Lister AM, Barnes I, Dalén L. 2010. Refugia revisited: Individualistic responses of species in space and time. *Proceedings of the Royal Society B: Biological Sciences* **277**: 661-671.

Stockwell D. 1999. The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science* **13**: 143-158.

Sturmbauer C, Opadiya GB, Niederstatter H, Riedmann A, Dallinger R. 1999. Mitochondrial DNA reveals cryptic oligochaete species differing in cadmium resistance. *Molecular Biology and Evolution* **16**: 967-974.

Stürzenbaum SR, Andre J, Kille P, Morgan AJ. 2009. Earthworm genomes, genes and proteins: the (re) discovery of Darwin's worms. *Proceedings of the Royal Society B: Biological Sciences* **276**: 789-797.

Svenning J-C, Fløjgaard C, Marske KA, Nógues-Bravo D, Normand S. 2011. Applications of species distribution modeling to paleobiology. *Quaternary Science Reviews* **30**: 2930-2947.

Syers JK, Springett JA. 1983. Earthworm ecology in grassland soils. in *Earthworm Ecology*, pp. 67-83. Springer.

Symonds MRE, Moussalli A. 2011. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology* **65**: 13-21.

Tabor GM, Aguirre AA. 2004. Ecosystem health and sentinel species: adding an ecological element to the proverbial "canary in the mineshaft". *EcoHealth* **1**: 226-228.

Tiunov AV, Hale CM, Holdsworth AR, Vsevolodova-Perel TS. 2006. Invasion patterns of Lumbricidae into the previously earthworm-free areas of northeastern Europe and the western Great Lakes region of North America. in *Biological Invasions Belowground: Earthworms as Invasive Species*, pp. 23-34. Springer.

Torres-Villaça S. 2012. Spatial and temporal distribution of mitochondrial lineages in the European wild boar. Università degli Studi di Ferrara, Ferrara (Italy).

Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S. 1995. Structures of metal sites of oxidized bovine heart cytochrome *c* oxidase at 2.8 A. *Science* **269**: 1069-1074.

Turbé A, De Toni A, Benito P, Lavelle P, Lavelle P, Ruiz N, Van Der Putten H, Labouze E, Mudgal S. 2010. Soil biodiversity: functions, threats and tools for policy makers. Report for European Commission (DG environment).

van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, van der Poel H, van Oeveren J, Verstegen H. 2007. Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PloS One* **2**: e1172.

van Straalen NM, Butovsky RO, Pokarzhevskii AD, Zaitsev AS, Verhoef SC. 2001. Metal concentrations in soil and invertebrates in the vicinity of a metallurgical factory near Tula (Russia). *Pedobiologia* **45**: 451-466.

Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS. 2008. SNP

discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* **5**: 247-252.

Vasseur P, Cossu-Leguille C. 2003. Biomarkers and community indices as complementary tools for environmental safety. *Environment International* **28**: 711-717.

Vega R, Fløjgaard C, Lira-Noriega A, Nakazawa Y, Svenning J-C, Searle JB. 2010. Northern glacial refugia for the pygmy shrew *Sorex minutus* in Europe revealed by phylogeographic analyses and species distribution modelling. *Ecography* **33**: 260-271.

Vijver MG, Vink JPM, Miermans CJH, van Gestel CAM. 2003. Oral sealing using glue: a new method to distinguish between intestinal and dermal uptake of metals in earthworms. *Soil Biology and Biochemistry* **35**: 125-132.

Wang H, Erséus C. 2004. New species of *Doliodrilus* and other Limnodriloidinae (Oligochaeta, Tubificidae) from Hainan and other parts of the north-west Pacific Ocean. *Journal of Natural History* **38**: 269-299.

Wardle DA, Bardgett RD, Klironomos JN, Setälä H, Van Der Putten WH, Wall DH. 2004. Ecological linkages between aboveground and belowground biota. *science* **304**: 1629-1633.

Wares JP, Daley S, Wetzer R, Toonen RJ. 2007. An evaluation of cryptic lineages of *Idotea balthica* (Isopoda: Idoteidae): Morphology and microsatellites. *Journal of Crustacean Biology* **27**: 643-648.

Whalen JK. 2004. Spatial and temporal distribution of earthworm patches in corn field, hayfield and forest systems of southwestern Quebec, Canada. *Applied Soil Ecology* **27**: 143-151.

Wilson AC, Cann RL, Carr SM, George M, Gyllensten UB, Helm-Bychowski KM, Higuchi RG, Palumbi SR, Prager EM, Sage RD, Stoneking M. 1985. Mitochondrial DNA and two perspectives on evolutionary genetics. *Biological Journal of the Linnean Society* **26**: 375-400.

Wolstenholme DR. 1992. Animal mitochondrial DNA: structure and evolution. In: Jeon KW, Wolstenholme DR. *Mitochondrial Genomes*. Academic Press.

Won E-J, Rhee J-S, Shin K-H, Lee J-S. 2013. Complete mitochondrial genome of the marine polychaete, *Perinereis nuntia* (Polychaeta, Nereididae). *Mitochondrial DNA*.

Zhong M, Struck TH, Halanych KM. 2008. Phylogenetic information from three mitochondrial genomes of Terebelliformia (Annelida) worms and duplication of the methionine tRNA. *Gene* **416**: 11-21.

CHAPTER 2 PALEOCLIMATIC IMPACT ON CRYPTIC DIVERSITY OF THE EARTHWORM *LUMBRICUS RUBELLUS* IN EUROPE

Pierfrancesco Sechi[1], Pablo Orozco-terWengel[1], Isa-Rita M. Russo[1], Peter Kille[1] and Michael W. Bruford[1]*

[1] Cardiff University, School of Biosciences, Museum Avenue, Cardiff CF10 3AX, UK

Keywords: *Lumbricus rubellus*, earthworm, cryptic species, phylogeography, species distribution modelling

2.1 Introduction

The climatic shifts that characterized the European continent from the middle to late Miocene, ~13 Million years ago (Mya), may have caused major changes in species distribution, playing an important role in the divergence and speciation of many taxa (Hewitt 2000). These climate changes were characterized by oscillations between dry and wet phases, with a general tendency towards a colder climate (Böhme et al. 2008). In particular, two different "washhouse" climatic intervals (10.2 to 9.8 Mya, and 9.0 to 8.5 Mya), characterized by globally warm conditions and precipitation several orders of magnitude higher than at present (Böhme et al. 2008), may have had a major impact on the range of species dependent on mesic environments. This progressive cooling of the world's climate culminated in the "Ice Ages" of the Pleistocene, separated by warm interglacial periods with a rough cyclity of 100,000 years (Pisias and Moore Jr 1981).

This epoch has to date been studied in more detail, as genetic signatures of the Pleistocene glaciations have been used to assess the role of climate in the process of speciation, wherein species disappeared over broad parts of their range, migrated to new locations or survived in refugia (Hewitt 1996; Hewitt 2000; Hewitt 2004). Following ice retreats, some taxa expanded their range as habitable regions became available. These demographic changes occurred repeatedly during the Pleistocene, and are likely to have had profound consequences for European biota (Hewitt 1996; Hewitt 2000; Hewitt 2004).

The effects of these historical climatic shifts on soil ecosystems are likely to have had dramatic consequences for epigeic earthworms inhabiting the surface soil layer. Earthworms are a dominant component of the animal biomass in soils for many habitats (Lavelle et al. 1997). They act as ecosystem engineers, playing a key role in physically shaping soil structure, modifying and refining soil particulates and improving aeration and drainage with their burrows (bioturbation; Lavelle et al. 1997). They are also involved in the nutrient cycle, enriching the soil by casting stable organo-mineral structures with a low C/N ratio, thus contributing to making nitrogen available for plant growth (Lavelle et al. 1997; Sims and Gerard 1999). Given their continuous and intimate contact with soil, bio-geochemical soil conditions are another limiting factor for earthworm distributions (Edwards 1996).

Lumbricus rubellus (Hoffmeister, 1843) is a widely distributed earthworm found in Continental and Insular Europe (Sims and Gerard 1999) and is recognizable because of its dark-red pigmentation. It is an epigeic, out-crossing hermaphrodite (Bouché 1972; Sims and Gerard 1999), widely used in the field of soil ecotoxicology as a sentinel species (Spurgeon et al. 2003; Fränzle 2006; Bundy et al. 2007). Ecological studies on this species have led researchers to hypothesize that temperature and precipitation, more than soil characteristics such as pH and chemistry limit habitat suitability for the species (Edwards 1996). *L. rubellus* is usually restricted to moist pastures, but can survive in a broad range of soil environments, tolerating pH values from 3.5 to 8.4 (Sims and Gerard 1999) and soils contaminated by heavy metals (Morgan and Morgan 1988a; b; Langdon et al. 1999; Spurgeon and Hopkin 1999). An assessment of the importance of environmental variables on the current and historical geographic range of *L. rubellus* is a key to understanding the spatial distribution of this species and its genetic variation.

It has been proposed that extant northern European populations of *L. rubellus* are derived from re-colonizations from southern European glacial refugia after the retreat of the ice sheets ~10,000 years ago (King et al. 2008). The recent discovery of a subdivision of the species into two deeply divergent mitochondrial lineages inhabiting the UK in sympatry (probably cryptic species; King et al. 2008; Andre et al. 2009) could be the legacy of such re-colonization events. Alternatively, cryptic refugia during the Last Glacial Maximum (LGM) on the coasts of Northwestern Europe, may have served as the source of the extant but largely divergent British *L. rubellus* genetic diversity. There is increasing genetic evidence that many European thermophilic species not only survived in southern glacial refugia, separated by great distances from the range of the glacial front, but also in cryptic northern refugia (Stewart and Lister 2001; Stewart et al. 2010; Vega et al. 2010; Finnegan et al. 2013).

In this study, we explored the phylogeographic patterns and population history of the *L. rubellus* cryptic species complex in Europe. For this purpose the mitochondrial DNA (mtDNA) Cytochrome c oxidase subunits I and II (COI and COII) sequences, integrated with environmental niche modeling, were used in order to assess the spatial and temporal dimension of *L. rubellus* cryptic variation, and the effect of environmental variables on the current and historic species distribution. In particular, we investigated whether *L. rubellus* comprises a complex of cryptic species across Europe, and if the two divergent lineages found in UK re-colonized from

southern European locations. On the basis of the results of paleoclimatic modeling, we developed alternative scenarios that could explain the current distribution of *L. rubellus*. In particular, we addressed the following questions: 1) is *L. rubellus* composed of more than two cryptic lineages over its range? 2) do the northern European genetic lineages derive from southern populations that survived in glacial refugia? 3) did the current differentiation pattern arise during the Quaternary?

2.2 Materials and methods

2.2.1 Sampling and DNA extraction

We collected 299 *L. rubellus* samples from across western and central Europe during the spring and autumn of 2010-2011. Samples (227) were received from collaborators across Europe and stored in liquid nitrogen for DNA extraction (whole individuals for juveniles, and the posterior segments after the clitellum in adults, with adult anterior segments including the clitellum being stored in 100% EtOH for species identification). The remainder of the samples were stored in 70% EtOH. Precise sampling coordinates, location and number of samples per site are provided in Chapter 6, section 6.3. Genomic DNA was extracted from tissue using the DNEasy blood and tissue Kit following the manufacture's instructions (QIAGEN® Hilden, Germany).

2.2.2 PCR and sequencing

A 710 bp fragment of the COI gene was amplified using the general invertebrate primers LCO1490 and HCO2198 (Folmer et al. 1994). When PCR was unsuccessful, an alternative 356 bp fragment of the Folmer region was amplified with the primers LRBCOIF and LRBCOIR (Donnelly et al., in press). Specific primers designed to amplify a 469 bp fragment of the Cytochrome Oxidase II gene (Andre et al. 2009) were also used. All PCR reactions included ~ 100 ng DNA template, 0,25 U of GoTaq® Flexi DNA Polymerase (Promega), 1x Green GoTaq® Buffer, 2,5 mM of $MgCl_2$, 0,5 µM for both the forward and reverse primers and a 0.5 mM dNTPs mix (dATP, dCTP, dGTP and dTTP, Invitrogen, UK) in a 20 µl final reaction volume. PCR conditions had an initial denaturing step at 95 °C for 5 min, followed by 35 cycles of 95 °C for 30 s, 55 °C for 30 s and 72 °C for 30 s, and a final elongation step at 72 °C

for 10 minutes. PCR products where cleaned with Biolabs Exonuclease I (20,000 U/ml) and Promega Shrimp Alkaline Phosphatase (1 U/µl) following the manufacturer's protocol. PCR products were sequenced using the ABI PRISM BigDye v3.1 Terminator kit (Applied Biosystems, USA - 4337458) on an ABI PRISM 3100 automated DNA analyser, Molecular Biology Support Unit. Raw sequences were confirmed using Sequencher 4.9 (Gene Codes Corporation) and aligned using Mega v5.0 (Tamura et al. 2011).

2.2.3 Population analyses and demographic inference

Summary statistics (Haplotype diversity - h and nucleotide diversity averaged over all loci - π) on the concatenated dataset were calculated using ARLEQUIN v3.5 (Excoffier et al. 2005). Pairwise estimates of the nucleotide sequence divergence within and among clusters were calculated in MEGA 5 (Tamura et al. 2011) using the Kimura-2 parameter model (Kimura 1980; King et al. 2008). The minimum number of mutational steps between *L. rubellus* haplotypes was estimated based on the concatenated dataset using TCS v.1.21 (Clement et al. 2000). A Bayesian analysis of population structure using BAPS v.6.0 (Corander et al. 2008) was carried out to infer the most probable number of clusters in the data. Three independent analyses were performed to assess convergence of the results, using as prior upper bound vector between 6 and 20.

Inference of demographic history was assessed with mismatch distributions (MD) (Harpending 1994) in ARLEQUIN v3.5 (Excoffier et al. 2005). The deviation of the observed data from the exponential null model was determined with Harpending's Raggedness Index (RI) and the Sum of Squared Deviations (SSD). A total of 1,000 simulations of the model were carried out to define the 95% confidence intervals (CI) of RI, SSD and the model parameters. Tajima's D (Tajima 1989b) and Fu's F (Fu 1997) were also calculated as these statistics are normally more sensitive to demographic changes. Additionally, Bayesian skyline plot (BSP) analyses were carried out using BEAST v1.7.5 (Drummond and Rambaut 2007) based on each cluster as identified in BAPS. These analyses were run for 10^7 generations, but for some clusters more iterations were necessary to reach an Effective Sample Size (ESS) over 200 for all parameters. All the runs implemented a HKY model of evolution with four rate categories. Convergence of the runs was evaluated using Tracer v1.5.0

(Rambaut and Drummond 2007). The molecular clock was calibrated using the per-lineage mitochondrial mutation rate estimated for the megascolecid *Metaphire* (0.024 substitutions/My^{-1}) (Chang et al. 2008), as this rate is consistent with other lumbricid levels of genetic divergence for COI and COII (Perez-Losada et al. 2011).

2.2.4 Phylogenetic analyses and divergence time

The model of sequence evolution for both genes was calculated using MrModeltest v.2.3 (Nylander 2008). A Maximum Likelihood (ML) phylogeny was estimated using PhyML (Guindon et al 2010), with a Bayesian-like transformation to estimate branch support (Anisimova et al 2011). Additionally, a Bayesian phylogeny was estimated using MrBayes v3.2.1 (Ronquist et al. 2012). This analysis consisted of two independent runs of four chains each for 5 x 10^6 generations, discarding the initial 25% as burn-in and sampling trees and parameters every 100^{th} generation. The convergence of the runs (SD of split frequencies <0.01) was checked using the MrBayes diagnostic. Sequences of *L. terrestris* (JN869943.1 – COI and JN869610.1 - COII) were used as an outgroup.

Estimating divergence times between different lineages of *L. rubellus*, and for earthworms in general, is problematic, as no fossil records are available. Novo (2011) calibrated the divergence between cryptic species of the *Hormogastridae* complex using the split between the Sardinian microplate and the Iberian Peninsula (~ 33 Mya). However, such vicariant events could not be linked to this study as Continental Europe and the British isles have been a geologically stable environment for the last 10 My. Consequently, we dated our lineages according to the substitution rate described above by implementing the software BEAST v1.7.5 (Drummond et al. 2012). An analysis using a relaxed lognormal clock model was carried out for 2 x 10^8 iterations, consisting of 20 runs of 10^7 steps each, sampling every 10^3 iterations and with a burn-in phase of 20%. The resulting log files were combined with LogCombiner (Drummond and Rambaut 2007) and checked for ESS values above 200 and convergence of all parameters. The combined tree files were thinned with LogCombiner, to obtain a final dataset of 10,000 trees on which maximum clade credibility intervals and a consensus tree were estimated using TreeAnnotator.

2.2.5 Species and paleodistribution modelling

The extent of potential distribution for the species during present and past environmental conditions was assessed with the maximum entropy statistical method implemented in the software MaxEnt v.3.3.3 (Phillips et al. 2006; Richards et al. 2007). The aim was to 1) model the environmental niche of *L. rubellus* using data from weather measurements from a recent time interval (1950 to 2000), during the Last Glacial Maximum (LGM; ~21,000 Ybp) and during the last interglacial period (Eemian stage or LIG, ~120,000 – 140,000 Ybp), and 2) to build a framework to identify potential past population refugia. These models were used as alternative genealogical hypotheses to explain the current distribution of *L. rubellus* lineages in Europe.

The MaxEnt approach allows the modeling of species distribution by estimating the density of environmental covariates (precipitation, temperature etc.) depending on presence data (Franklin 2009). The data necessary for the MaxEnt modeling was compiled by integrating *L. rubellus* sample occurrences used in this study along with 2152 geo-referenced occurrences for *L. rubellus* downloaded from the Global Biodiversity Information Facility (GBIF, http://www.gbif.org). Duplicates were removed from the data, and the occurrence points were subsampled in order to limit the possibility of sample bias. For this purpose the study area was divided in a raster grid of 1 degree (approximately 90 x 110 km) and it was re-sampled by allowing a maximum of one sample per degree with R (R Developement Core Team 2013). The reduced dataset consisted of 146 points across the continent. After the addition of the 33 sampled points from this study, the final dataset consisted of 179 points across the study area (Figure 2.1).

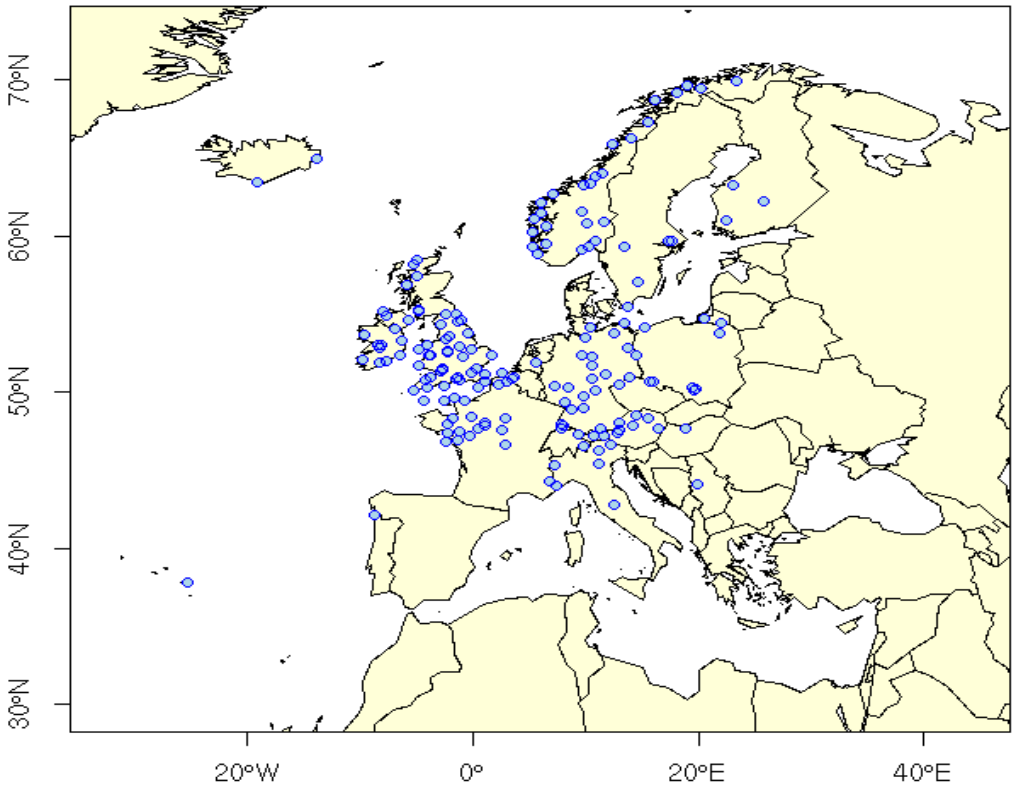


Figure 2.1. Plot of the presence data used for species distribution modelling in MaxEnt. Each blue dot represents a single occurrence point. The dataset combines the GBIF occurrence points, cleaned for duplicates and re-sampled to avoid bias, and the coordinates of the samples collected for this study, in a total of 179 occurrence points.

A total of 19 bioclimatic variables (Table 2.1) relative to the present day conditions (Hijmans et al. 2005), the LGM (from the CCSM and MIROC climate models; Collins et al. 2006; Otto-Bliesner et al. 2006) and the LIG (from the MIROC climate models; Otto-Bliesner et al. 2006) were downloaded from the www.worldclim.org website to construct present and paleodistribution models of habitat suitability for *L. rubellus*. These rasters comprised variables obtained by different combinations of temperature and precipitation over different time frames (year or quarter of year), with the aim of obtaining biologically meaningful variables for niche modeling. A second present day model was constructed to evaluate the effect of soil variables on the environmental niche modeling. For this model, five additional global topsoil variables were downloaded from the FAO Geonetwork database (http://www.fao.org/geonetwork/srv/en/main.home), namely: 1) topsoil organic carbon pool, 2) topsoil nitrogen percentage, 3) topsoil cation exchange capacity, 4) topsoil carbon/nitrogen ratio and 5) topsoil pH. Maps were clipped and re-projected to the

desired spatial extent and resolution using Quantum GIS v.1.8 (Developement Team 2009) and command line tools within the Geospatial Data Abstraction Library (GDAL; Warmerdam 2008). The contribution of each variable to explain the species distribution was evaluated with a jackknife test implemented in MaxEnt.

BIO1 = Annual Mean Temperature
BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3 = Isothermality (BIO2/BIO7) (* 100)
BIO4 = Temperature Seasonality (standard deviation *100)
BIO5 = Max Temperature of Warmest Month
BIO6 = Min Temperature of Coldest Month
BIO7 = Temperature Annual Range (BIO5-BIO6)
BIO8 = Mean Temperature of Wettest Quarter
BIO9 = Mean Temperature of Driest Quarter
BIO10 = Mean Temperature of Warmest Quarter
BIO11 = Mean Temperature of Coldest Quarter
BIO12 = Annual Precipitation
BIO13 = Precipitation of Wettest Month
BIO14 = Precipitation of Driest Month
BIO15 = Precipitation Seasonality (Coefficient of Variation)
BIO16 = Precipitation of Wettest Quarter
BIO17 = Precipitation of Driest Quarter
BIO18 = Precipitation of Warmest Quarter
BIO19 = Precipitation of Coldest Quarter

Table 2.1. List of the bioclimatic variables available on the www.worldclim.com database for the present, the Last Glacial Maximum (LGM) and the Last InterGlacial (LIG). A quarter represents ¼ of the year.

Three different model runs were carried out: the present distribution model of the species and the related LGM and LIG projections were analyzed with a "subsample" replicate run type in MaxEnt. Ten replicates of the analyses were carried out where 75% of occurrence data was used as training data and the remaining 25% as test data with default parameters. The inferred niche models were visualized as maps and habitat suitability was expressed as a color log scale, where a higher probability was represented by a warmer color (i.e. closer to red). Model performance was evaluated by an estimate of the Area Under the Curve (AUC) of the Receiving

Operating Characteristics (ROC). ROC measures the ability of the prediction to discriminate between the species presence and absence (Elith et al. 2010). AUC values range between 0.5 (the model performs as a random variable) and 1 (the model can reliably discriminate between suitable and unsuitable points in the landscape). A Multivariate Similarity Surface (MESS) analysis and the Most Dissimilar Variable (MoD) analysis were used to assess the reliability of the projection of the present day model to the past and to evaluate the sustainability of the model, respectively. Descriptions of the evaluation methods of model performance (standard deviation, AUC, MESS and MoD maps) are in Chapter 6.

2.2.6 Phylogeographic modelling

Phylogeographic hypotheses explaining the European distribution of genetic variation in *L. rubellus* were tested. For this purpose a set of samples for the concatenated COI and COII genes was selected, and the following summary statistics were calculated using Arlequin v3.5 (Excoffier et al 2005): θ_H (Chakraborty and Weiss 1991), Tajima's D and π (Tajima 1989a; b). Three alternative genealogies reflecting different phylogeographic origins of the extant genetic diversity (Figure 2.7) were simulated with the software ms (Hudson 2002). Times of divergence were expressed as generations/θ_s/μ, assuming a generation length of 1 year as expected from field observations for *L. rubellus* (Edwards 1996) and the *Metaphire* μ rate of 2.4% substitutions/year (Chang et al. 2008). A total of 1000 simulations were carried out for each phylogeographic model parameterized by θ_s (the estimation of the $4Ne\mu$ parameter using the proportion of segregating sites; Watterson 1975), the number of segregating sites and the times of divergence between lineages. Each simulation was summarized as observed data, and the quantile at which the observed data occurs in the distribution of simulated summary statistics was calculated in R. Models were rejected if the observed summary statistics were found outside the 95% CI of the simulated statistics.

2.3 Results

Both sequences (COI and COII) were amplified in 297 individuals combined sequence length was 828bp (428 from COI and 402 from COII). The concatenated dataset was used for all analyses. Among the 297 individuals, 125 haplotypes were found from 305 segregating sites. The average haplotype diversity (h) and nucleotide diversity in the whole dataset was 0.97 (SD = 0.005) and 0.08 (SD = 0.003), respectively. The Bayesian analysis of population structure using BAPS inferred 11 clusters in the data with a posterior probability of 0.99 (Figure 2.2). However, as one of the clusters was characterized by only two samples and thus was not suitable for population-based inferences (Lineage I), we did not include it in most analyses. The distribution of genetic variation in these clusters is shown in Table 2.2.

	N	*h*	π		
A1	76	20	0.0113	+/-	0.006
A2	42	27	0.0191	+/-	0.01
A3	67	11	0.01	+/-	0.005
C	17	17	0.0122	+/-	0.007
D	9	5	0.0063	+/-	0.004
E	7	7	0.0865	+/-	0.049
B	31	18	0.0068	+/-	0.004
F	26	6	0.0084	+/-	0.005
G	11	9	0.0332	+/-	0.018
H	9	3	0.015	+/-	0.009
I	2	2	0.0061	+/-	0.007
Total	297	125			

Table 2.2. The number of individuals (N), haplotype (*h*) and nucleotide diversity (π) averaged per locus followed by standard deviation values for each cluster.

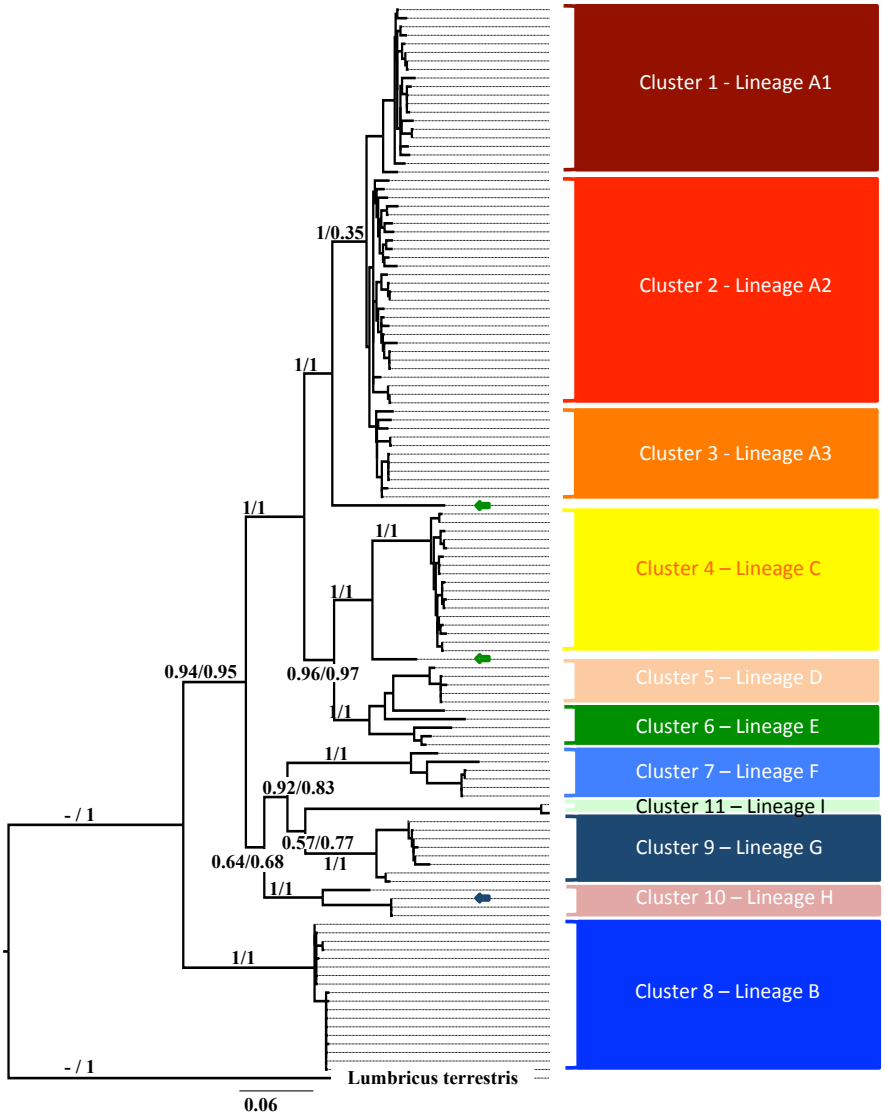


Figure 2.2. Left: Maximum-likelihood tree based on the two gene fragments amplified for *L. rubellus* (COI and COII); right: BAPS clustering solutions. Bayesian-like transformation of approximate Likelihood Ratio Test values (Anisimova et al. 2011) followed by posterior probabilities over 0.5 are shown on each branch. BAPS cluster solutions on the right reflect the tree topology with a few exceptions denoted by arrows, pointing out mismatches between phylogenetic and BAPS structure.

2.3.1 Phylogenetic and population structure analysis

The model of sequence evolution inferred for each gene was the General Time Reversible model with a linked Gamma function for the rates (GTR+G), and for COI additionally a parameter of invariant sites (I) was inferred. As both genes presented a similar model of evolution we concatenated these genes for all phylogenetic analyses and the combined dataset was analyzed under a GTR+G+I model of substitution. The

Bayesian and ML phylogenetic analyses resulted in eleven lineages that recapitulated the BAPS results (Figure 2.2). However, in the phylogenetic analysis, Lineage E was polyphyletic, and one haplotype of Lineage G grouped with Lineage H. The individuals within Lineage E comprised mostly examples of various divergent lineages only represented by a few individuals that were clustered in BAPS as one group. Lineage B was monophyletic and basal to all other lineages.

Sequence divergence values (based on the Kimura-2-parameter model) between lineages ranged between 2.8% (between Lineage A2 and Lineage A3) and 17.05% (between the Lineage I and Lineage D; Table 2.3). Within group divergences ranged from 0.8% in lineage D to 9% in lineage E, the latter value supporting the hypothesis that this is a polyphyletic lineage consisting of multiple independent lineages rather than a monophyletic clade. Lineages A1-A2-A3 were monophyletic with a maximum degree of divergence of 4.2%; C-D-E and F-G-H formed reciprocally monophyletic clades, but with much higher degrees of inter-lineage divergences (~ 10% and 14% respectively for the two clades).

	A1	A2	A3	C	D	E	B	F	G	H	I
A1	**0.017**										
A2	0.041	**0.021**									
A3	0.042	0.028	**0.016**								
C	0.105	0.113	0.114	**0.012**							
D	0.119	0.109	0.106	0.105	**0.008**						
E	0.112	0.108	0.104	0.107	0.089	**0.095**					
B	0.153	0.146	0.139	0.161	0.158	0.156	**0.008**				
F	0.152	0.144	0.142	0.147	0.144	0.143	0.145	**0.035**			
G	0.148	0.142	0.137	0.159	0.141	0.147	0.142	0.135	**0.04**		
H	0.145	0.135	0.129	0.157	0.147	0.138	0.143	0.136	0.102	**0.051**	
I	0.167	0.167	0.158	0.164	0.175	0.173	0.158	0.171	0.154	0.155	**0**

Table 2.3. Estimates of Evolutionary Divergence over Sequence Pairs between (low diagonal) and within (on diagonal, in bold) Groups. The numbers of base substitutions per site from averaging over all sequence pairs between groups are shown. The model used for the analyses was Kimura -2 parameter.

The geographic distribution of the mitochondrial lineages is depicted in Figure 2.3. Lineages A1 to A3 and Lineage E were found to be widely distributed across Europe, while Lineages C and D were restricted to the Balkans. Lineages G and H were restricted to Germany and Austria, while Lineage F was only present in Spain. Lineage B was only found in the UK.

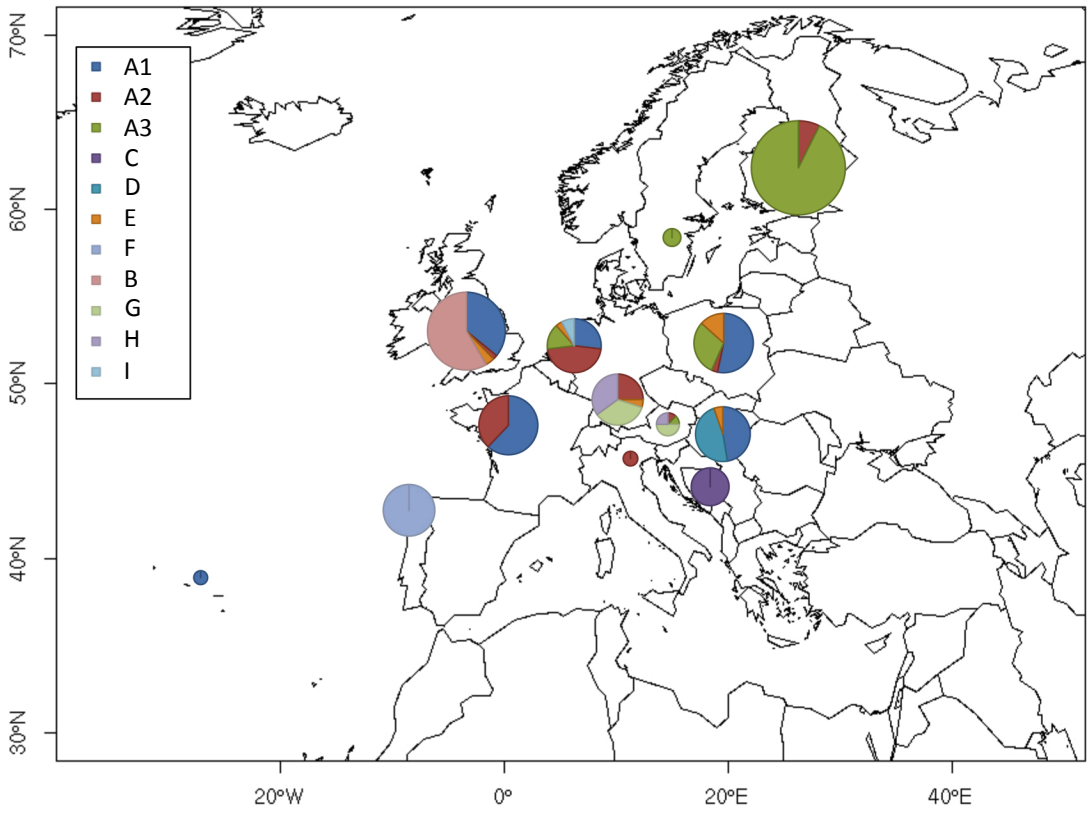


Figure 2.3. Geographic distribution of samples according to their mitochondrial lineages. The areas of the pie charts represent the number of samples collected in each location.

2.3.2 Divergence time estimates

The divergence between the *L. rubellus* species complex was estimated as starting between 5.68 and 2.99 Mya, with Lineage B being the first to diverge, the most basal and the direct descendant of the MRCA of all the lineages. The split between the monophyletic group of Lineages A1-A3 and the other group of Lineages C, D and E was estimated at 3.83 to 2.1 Mya, diversification within Lineages F-H at 4.21 to 1.78 Mya. The divergence between *L. rubellus* and *L. terrestris* was estimated to have occurred 6.27 to 2.99 Mya, as ancient as the split between Lineage B and the other lineages. Estimates of divergence times are found in Figure 2.4.

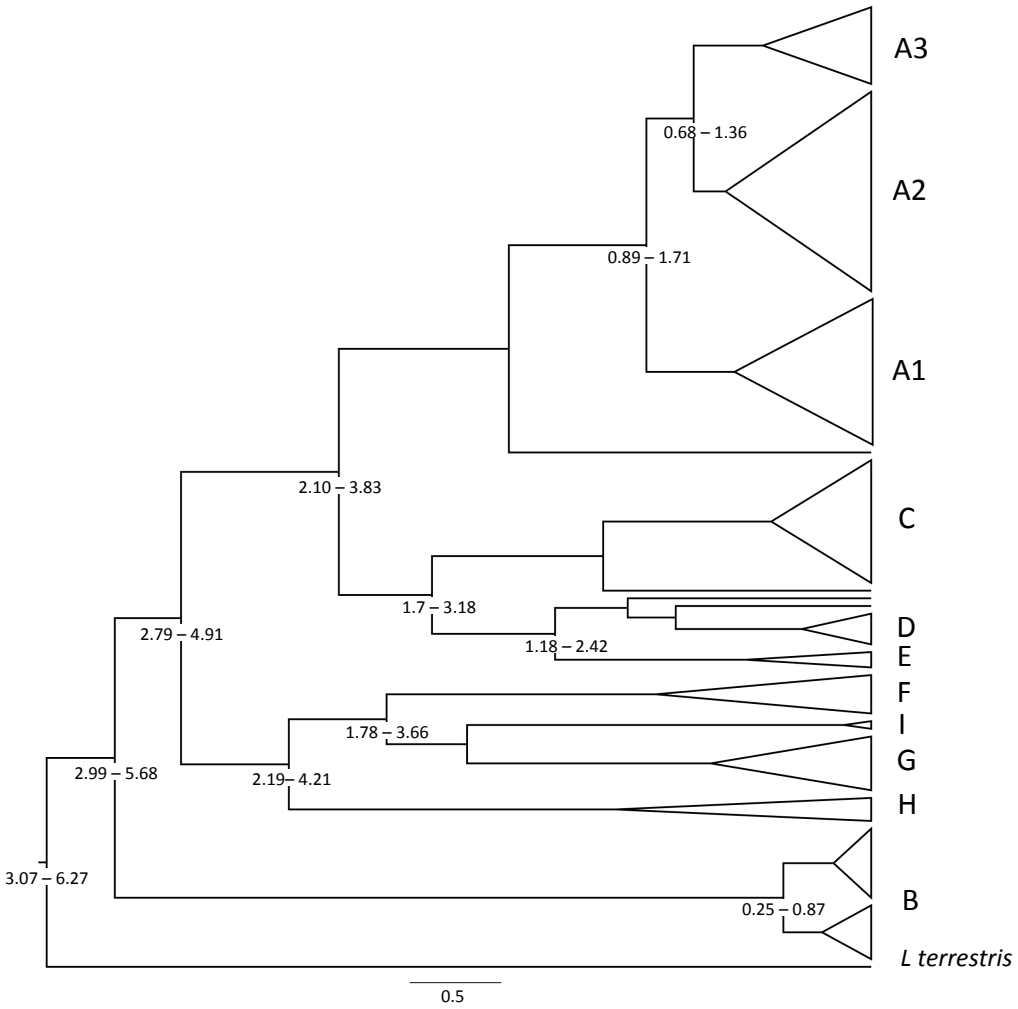


Figure 2.4. Ultrametric tree representing divergence times in the main clades under a Yule speciation process. Results are shown in millions of years (95% confidence interval).

Demographic history

Lineages A3, D, E, F, G and H all showed multimodal frequency distributions of pairwise nucleotide differences characteristic of demographically stable populations. Lineage A3 showed a ragged multimodal MD (Chapter 6, Figure 6.2), a stable BSP over the last 300,000 years (except for the last 25,000 years, when a decline in population size has been observed), and a positive Fu's F (Table 2.4). Similarly, the BSP of Lineage D showed a similar pattern (Chapter 6, Figure 6.3). Contrastingly, Lineage E showed evidence of a slow increase in population size starting 250,000 years ago until reaching a stable demography for the last 125,000 years (Chapter 6, Figure 6.3). With the exception of a bottleneck during the last 10,000 years, Lineage F showed evidence of a stable population demography during the last 900,000 years (Figure 6.4). The BSP of Lineages G and H suggested stable demographies over the last 400,000 and one Million years, respectively (Chapter 6, Figure 6.5).

A consistent signature of demographic expansion (MD, BSP) was, however, found in four lineages (A1, A2, C and B), two of which are distributed throughout Europe. For A1 and A2 the MD (RI 0.01 and SSD *p-value* $\geq$ 0.05 for both lineages; Table 2.4) and BSP analyses supported an expansion 300-350,000 years ago during the Holstenian (northern Europe) second interglacial period (Richmond and Fullerton 1986). In addition the MD and BSP graphs also suggested a bottleneck signature at around 10,000 years ago (Chapter 6, Figure 6.1). The BSP for Lineage C suggested a demographic increase approximately 200-300,000 years ago, followed by an increase in population size towards 60,000 years ago (Chapter 6, Figure 6.2). The bimodal MD of Lineage B could be an indication that this lineage is subdivided (Chapter 6, Figure 6.4); the demographic parameters R and SSD indicated an expansion event. Similarly, the BSP showed that a population expansion event started ~25,000 years ag, during the LGM.

Lineage	Tajima's D	P	Fu's Fs	P	SSD	P	R	P
A1	-1.199	0.095	0.686	0.622	0.01	0.13	0.019	0.05
A2	-0.925	0.177	-3.848	0.105	0.007	0.08	0.008	0.12
A3	-0.532	0.331	5.478	0.934	0.168	0.17	0.112	0.12
C	-0.967	0.168	-9.326	0	0.004	0.86	0.01	0.94
D	-0.547	0.307	1.31	0.749	0.08	0.26	0.138	0.36
E	0.474	0.707	0.544	0.351	0.031	0.3	0.062	0.33
F	-2.331	0.002	7.133	0.989	0.069	0.13	0.366	0.52
B	-0.728	0.238	-5.565	0.021	0.03	0.022	0.029	0.54
G	-0.936	0.157	1.361	0.704	0.03	0.7	0.03	0.76
H	-2.024	0	8.41	0.999	0.051	0.24	0.255	0.62

Table 2.4. Tajima's D, Fu's Fs, SSD and Harpending's R, with their relative P values, are reported. Significant values are highlighted in orange.

2.3.3 Environmental niche modeling: current distribution

The MaxEnt model for the current *L. rubellus* distribution was consistent with the current distribution of the species, including its sparsely populated distribution in the South-Central parts of the Iberian Peninsula and France (M. Novo, pers. comm.) (Figure 2.6a). A discrepancy between the observed and expected distribution was in the model's estimated null habitat suitability for extreme North-eastern part of Europe where the species has recently been described as invasive (Tiunov et al. 2006). The average AUC for the current distribution across 10 replicates performed better than the random model, with an average of 0.93 for the training data and 0.87 for the test data.

2.3.4 Paleodistribution modelling

The two LGM climate models gave different results in terms of habitat suitability for the species. The model applied to the CCSM environmental variables (Figure 2.5b) showed habitat suitability mainly across a landscape including the Iberian Peninsula's North-western Atlantic coastline, extending eastwards towards the Pyrenees and northwards towards Ireland's western seafloor, a landscape covered by tundra during the LGM (Ray and Adams 2001). The other European glacial refugia (i.e. the Italian Peninsula and the Balkans) showed low habitat suitability under this model. Remarkably, a suitable area in the South-western part of the black Sea (modern

day Georgia and the Russian Federation) was found where the species has previously been deemed invasive (Global Invasive Species Database 2013).

The MIROC LGM environmental variables model rendered a larger habitat suitability area (Figure 2.5c) spanning across Central Europe. This distribution reflects the taiga-tundra that once covered this area, south of the ice sheet threshold and the frozen desert (Ray and Adams 2001). When compared to the CCSM, the MIROC model showed higher habitat suitability also with the southern glacial refugia, were the habitat in South-western Spain and the Pyrenees appears extended and the niche suitability range over the Italian Peninsula and the Balkans was significantly higher than in the CCSM model.

The habitat suitability MIROC model at the time of the LIG depicts a scenario of a northward shift of the species niche (Figure 2.5d). During this period patches of suitability are inferred for northern Spain, while its southern range featured a great reduction in niche probability compared to both the LGM and the current models. Contrastingly, the most suitable areas during the LIG were Brittany, the British Isles and northern Scandinavia. The MESS and MoD analyses, described in detail in Chapter 6, section 6.2, showed that for the CCSM model, some variables were out of range over great part of the studied area.

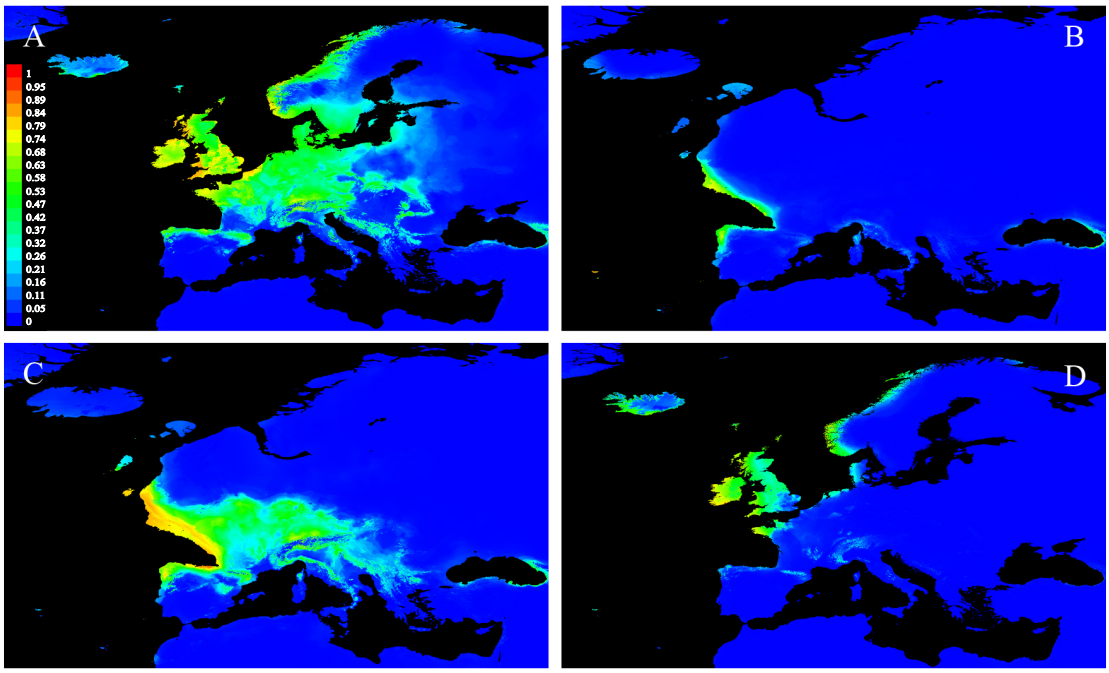


Figure 2.5. MaxEnt models of *Lumbricus rubellus* habitat suitability across Europe. Warmer colours represent higher habitat suitability. Values from 0 to 1 are an estimation of the probability of species presence. A) Species distribution model of present time; B) Paleodistribution model of the last glacial maximum (~21,000 years ago), according to the CCSM paleoclimatic model; C) Paleodistribution model of last glacial maximum, according to the MIROC paleoclimatic model; D) Paleodistribution model of the last interglacial (Eemian stage, ~120,000-140,000 years ago), according to the MIROC model.

2.3.5 Variable contribution

The most significant variables in the MaxEnt model were annual temperature range (BIO7; 26.6%), maximum temperature of the warmest month (BIO5; 14.6%), precipitation in the driest quarter (BIO17; 13.7%), precipitation in the driest month (BIO14; 13.5%) and mean temperature in the coldest quarter (BIO11; 6.8%). All the other variables contributed less than 5% to the model. The MaxEnt jackknife test showed that annual temperature range was the most important variable when used in isolation, while mean temperature in the driest quarter, was the most important variable for the test when the exclusion of a single variable at a time was employed (Figure 2.6 and 2.7 a,b). Hence, these two climatic variables seem to carry the most important information explaining the species' past and present distributions. The model evaluation of the importance of soil factors (Edwards 1996) on the species' distribution and the jackknife test (Figure 2.7c) showed that the five soil variables contributed very little to the model (less than 1.2% and a maximum AUC of 0.67, respectively). Thus, the effect of precipitation and temperature was inferred to have a greater impact on *L. rubellus* distribution than soil conditions.

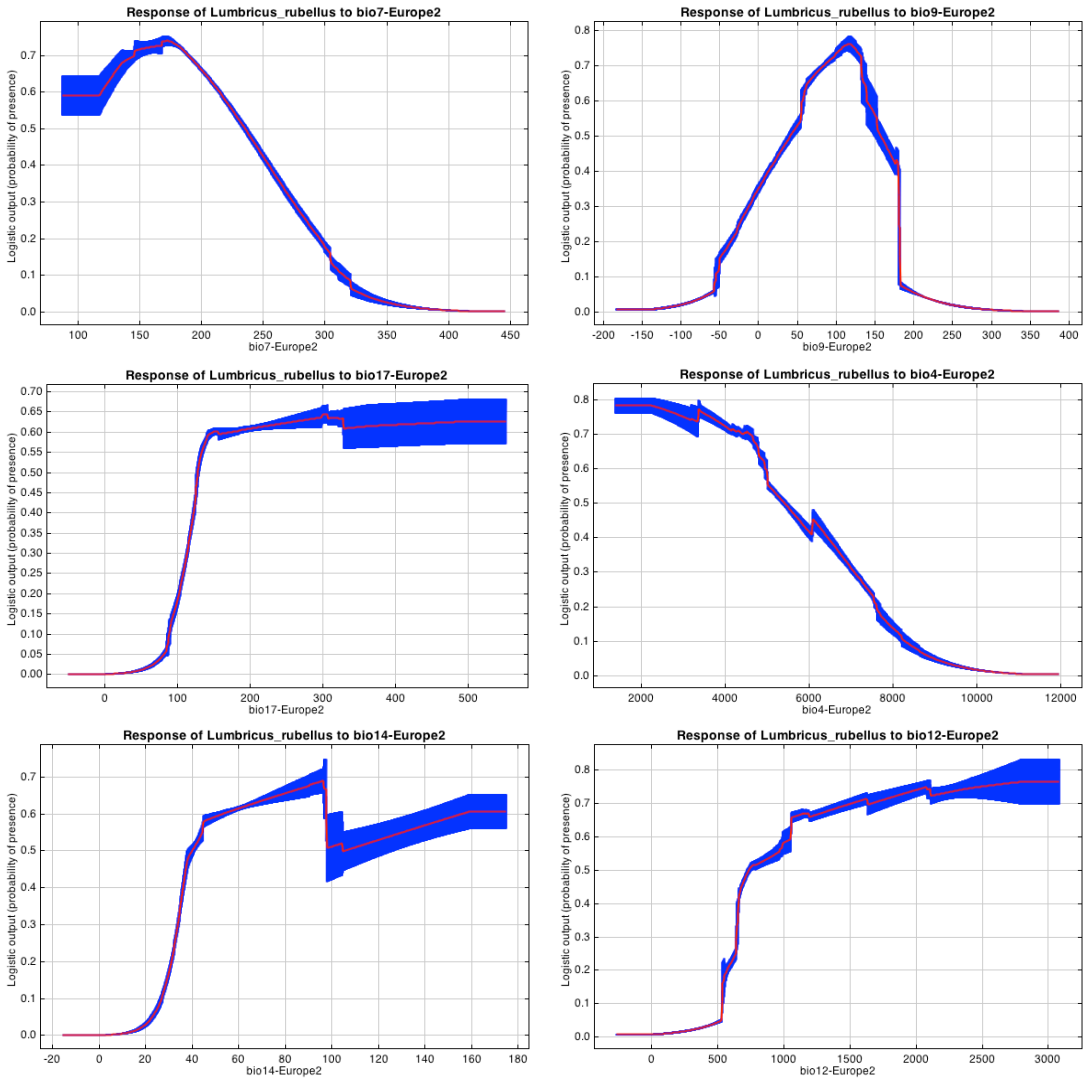


Figure 2.6. Plots reporting the effect of the most important variables on niche prediction. Each curve represents a Maxent model created using only the corresponding variable. The red line represents the mean, and the blue areas the standard deviation between the runs. The plot evidences how the variable contributes to the niche model across its range. The variables shown are estimates related to temperature (BIO7: temperature annual range; BIO9: mean temperature of the driest quarter; BIO4: temperature seasonality) and precipitation (BIO17: precipitation of the driest quarter; BIO14: Precipitation of the driest month; BIO12: annual precipitation). Where the temperature variables show higher suitability on the center-left part of the distributions, the precipitation variables show the opposite, where areas with highest precipitation values across the year achieve the highest suitability probability. See table 2.1 for a complete list of bioclimatic variable labels.

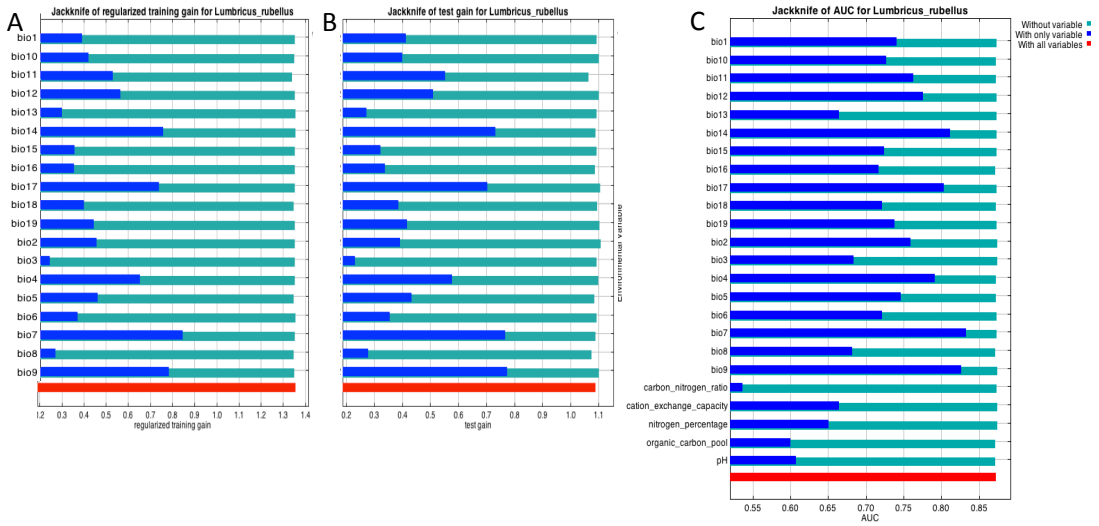


Figure 2.7. Jackknife test of variable importance. Blue bars correspond to contribution of the variable to the gain when used in isolation; green bars show the decrease in gain when the variable is missing. A: jackknife results on regularized training gain; B: jackknife results on the gain obtained considering only test samples. C: Jackknife test of a combination of bioclimatic and soil variables. Soil variables show very little contribution to the model construction; the most important soil variable, cation exchange capacity, increases the AUC of only 1.5 from the random value. See table 2.1 for a complete list of variable labels.

2.3.6 Phylogeographic modelling

Three different phylogeographic hypotheses were tested using coalescent simulations. The coalescent simulations were built on a subset of the individuals (n=127) representing only five of the original 11 clusters: Lineage A2 (n=42; representing the monophyletic Lineages A1-A2-A3), Lineage C (n=17; representing the Balkan lineages), Lineage G (n=11; representing a potential German refugium), Lineage F (n=26; representing the Iberian refuge), and the British Lineage B (n=31). The hypotheses tested differed from each other in the timing at which lineage divergence took place. Under Hypothesis 1) the ancestral lineages to B, the F and G clade, and the A clade started diverging form their ancestral population (β) 4.5-5 Mya. At the beginning of the Quaternary, the lineages F and G diverged from their common ancestor (γ: 3.2 Mya), followed not much later by the divergence between lineages A2 and C from their common ancestor (α: 2.9 Mya). Contrastingly, the other two hypotheses consisted of a simultaneous radiation of the five lineages at either (Hypothesis 2) the beginning of the Last interglacial or (Hypothesis 3) 4.5-5 Mya. The summary statistics (θ_H, D and π) calculated using the observed data and their 95% CI for the simulated data under the three alternative hypotheses are shown in Table 2.5.

The second and third hypotheses were rejected as the observed summary statistics were significantly smaller than the 95% CI of the simulations under the corresponding scenarios (Figure 2.8). In contrast, the observed data statistically supports the first phylogeographic hypothesis as all summary statistics of the observed data occurred within the 95% CI of the simulated data. Consequently, an initial divergence of 3-5 Mya of ancestral lineages followed by a final diversification close to the 2.58 Mya best explained the observed data.

Statistic	Empirical Value	Hyp1 5% < CI < 95%	Hyp2 5% < CI < 95%	Hyp3 5% < CI < 95%
θ_s	45.6	-	-	-
π	81.6	78.39 - 83.81	70.93 - 76.58	73.88 - 78.48
θ_H	27.56	27.31 - 32.92	20.96 - 25.45	21.77 - 25.31
D	2.6	2.37 - 2.76	1.83 - 2.24	2.04 - 2.37
N. Pol. sites	247	-	-	-

Table 2.5. Coalescent simulations. The values of the summary statistics of the restricted dataset are shown in column 2. Columns 3 to 5 report the confidence interval for each statistic in the 3 different coalescent hypotheses. Hypotheses 2 and 3 are rejected because the empiric values fall outside the CI. θ_s and the number of polymorphic sites were used as an input parameter for all the simulations.

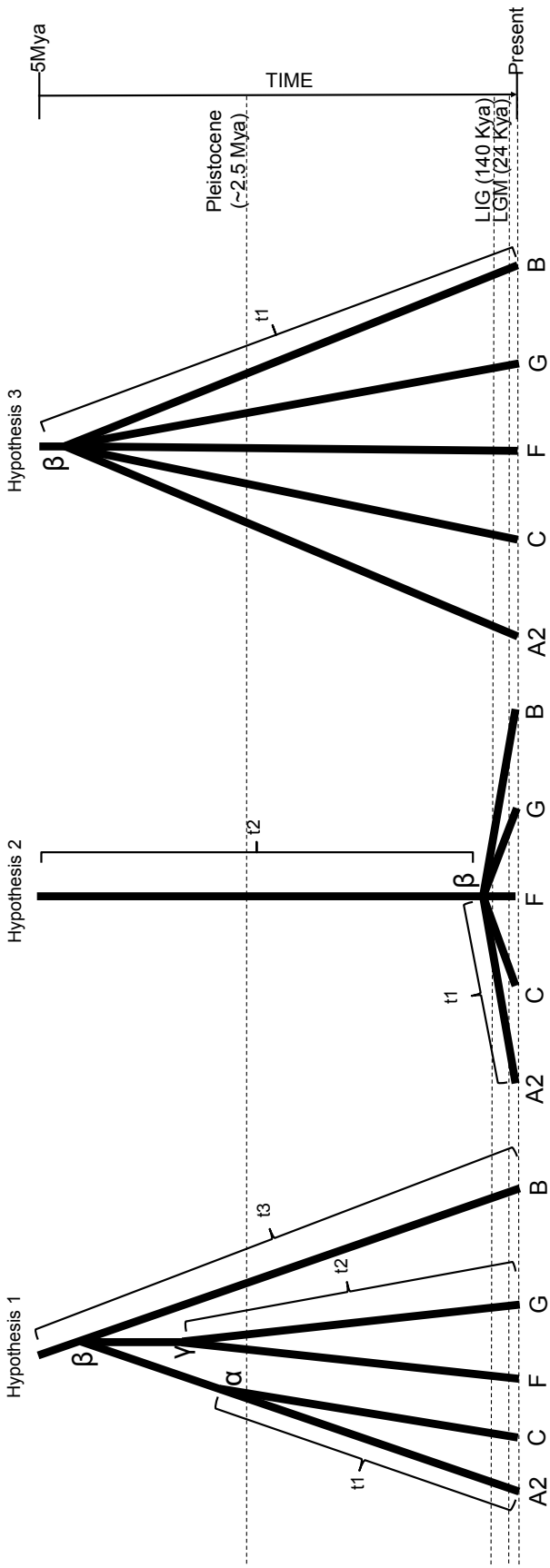


Figure 2.8. Schematic representing alternative biogeographical hypotheses for five of the 11 *L. rubellus* lineages. Hypothesis 1: 3 ancestral populations derived from ancient population β (t3), and following split of the ancient populations α and γ into the haplogroups A2 – C (t1) and F – G (t2) before the beginning of the Pleistocene; lineage B is the direct descendant of the ancestral population - Hypothesis 2: Split of all the lineages from the β ancestral population during the Eemian stage (LIG, 140,000 years ago) – Hypothesis 3: Ancient split of all the lineages from the β ancestral population during the Pliocene (~5 million years ago).

2.4 Discussion

In this study, we investigated the phylogeography and population structure of the *L. rubellus* cryptic species complex across Europe. Eleven well-supported lineages/clusters were identified according to both phylogenetic and population structure analyses. Some of these lineages showed clear geographical patterns that were widely distributed across the continent (A1-A2-A3 and E), while others represent deeply divergent localized lineages and probably descend from refugial populations (C, D, B, F, G, H). These localized lineages are the most divergent ones, separated from the A1-A2-A3 cluster with a minimum of 10.5%. Divergence time estimates point out to a contrasting history between these lineages with the oldest divergence occurring at ~6 Mya and the later divergences during the Pleistocene. Some lineages associated with refugia showed signatures of stable demographics, as well as expansion events possibly linked to the shifting climatic conditions of the Pleistocene. Environmental niche modelling confirmed the current species distribution pattern and inferred likely suitable areas for the species during the last glacial and interglacial phases. An assessment of the effect of environmental variables on species distribution in the SDM framework suggested that climate, as opposed to soil conditions, is the limiting factor for the species distribution, confirming past studies regarding the general preference of earthworm peregrine taxa for cold and wet climates (Nordström and Rundgren 1974; Edwards 1996; Tiunov et al. 2006).

The distribution of lineages in combination with divergence time estimates, suggested three main radiations, i.e. the radiation that resulted in Lineages A1-A3 and the C-D-E lineages (3.83-2.1 Mya, during the Pleistocene), the radiation that originated in the F-G-H refugial populations (4.2-1.78 Mya), and the one that gave rise to the contemporary lineage B. Lineage B appears to be the direct descendant of the common ancestor of all current lineages, with a divergence age as old as the split between *L. terrestris* and *L. rubellus*, i.e. 6.27 to 3.07 Mya. Thus, the data suggest an ancient split and consequent isolation between what must now be regarded as separate species (Donnelly et al. In Press). Once the lineages were disconnected from each other, they became characterized by independent species demographic histories and deep divergences. Consistent with this hypothesis, we found a mixture of signatures demographic history in the various lineages for the recent past. Most lineages were

probably already well defined at the time of the LIG (as reflected by sequence divergences between 1.7% and 17.5%) and survived in the same areas where they are found today. The stable population sizes detected by BSP analyses for Lineages F (Spain), G, H (Germany and Austria), and D (Balkans) confirmed this expectation and suggested that stable conditions were a constant feature for more than just the last glacial stage (e.g. a stable population size is inferred for Lineage F for the last 900 Ky). The only refugial population, other than Lineage B, that showed a signature of expansion was Lineage C (Serbia). This expansion dates back to the last glacial stage and is consistent with survival in a suitable area such as a southern European refugium, where expansion signatures could be due to the opening of new environments. The paleo-distribution models support such a scenario, although they can only infer habitat suitability for the last 135 Ky. Nevertheless, our results supported the likely survival of most of the reciprocally divergent lineages in refugial areas during the LIG and LGM. While some lineages clearly showed a survival pattern in southern European peninsulas, confirming the classical pattern of survival in Mediterranean refugia (Hewitt 2000; Schmitt 2007), others lineages presented distributions and demographies pointing out to survival in cryptic refugia (Stewart and Lister 2001).

The LIG niche distribution model predicted areas of suitability close to the postulated refugia, even though in some areas, the suitability was greatly reduced (i.e. the eastern Alpine range - Lineages G and H; the northern Iberia - Lineage F; Brittany and the British Isles - Lineage B). The model only fails to predict suitability for the Balkans, but for this region the MESS and MoD maps indicated that the range of the most important variable in the analysis (annual temperature) was out of range for that area (Chapter 6, 6.9) and thus predictions of the Balkans for the LIG could not be resolved. Nevertheless, with its patterns of suitability shifting towards the north, the LIG SDM is concordant with what is known about the late Eemian stage, most remarkably the occurrence, in its last 400 years, of a dry period that affected central Europe just before the beginning of the last Ice Age (Sirocko et al. 2005).

The two LGM models differed greatly in terms of output concerning central Europe. Nevertheless, both allowed inference that some of the lineages could have survived in the southern European glacial refugia. In addition, the LGM MIROC model showed a vast area of suitability corresponding to the tundra that occupied the centre of Europe during the LGM. Currently Lineages A1-A3 are found at northern

latitudes and are therefore moderately cold tolerant. Thus, it is likely that the Lineages A1-A3 survived in this geographic area, as testified by the signatures of expansion over the last 300 Ky for A1 and A2. Lineage A3 seems to have been demographically stable over this period of time. This is consistent with previous evidence of central European refugia, where temperate species could have survived, sometimes in co-occurrence with cold adapted Pleistocene Fauna. Using mitochondrial markers and species distribution models, Vega et al. (2010) found evidence of areas of suitability consistent with the survival of mitochondrial haplotypes private of northern and central Europe in the pigmy shrew (*Sorex minutus*). Further expansions after the retreat of the ice probably led to the re-colonization of other parts of the northern range, such as Poland and Scandinavia. The land bridge that connected Britain and continental Europe until ~8,000 years ago was most likely the path through which Lineages A1-A3 re-colonized Great Britain as previously suggested for other lumbricid species (Sims and Gerard 1999). An alternative explanation for the widespread distribution of Lineage A in Europe and in the British Isles could be due to anthropogenic crop and soil movement (Edwards 2004). The MIROC model also point out to a possible alternative cryptic refugium for the C and D lineages, considered Balkan according to classical signatures of recolonisation from southern refugia. The Carpathian basin is indicated as a cryptic glacial refugium according to recent studies on the bank vole (*Clethrionomys glareolus*, Kotík et al. 2006) and temperate deciduous plants (Willis and van Andel 2004; Magri et al. 2006).

The environmental variables tested in MaxEnt identified precipitation and temperature as the most important limiting factors for *L. rubellus* distribution as opposed to topsoil conditions. These results can be strongly linked to the species' ecology. Epigeic worms live in the upper 30 cm of the soil litter and consequently are heavily exposed to climatic elements (Bouché 1972). Therefore, past climate conditions are the key to underpin this species' past geographic distribution, as well as the range history of similar lumbricid species. According to previous studies, many lumbricid species are likely share to various degrees the same sensitivity of *L. rubellus* to climatic variables (Edwards 1996). Therefore, SDM may be an essential tool in future large-scale studies concerning earthworms phylogeography.

In order to evaluate the hypotheses generated by the SDM results and the genetic data, three phylogeographic hypotheses that could explain the extant distribution of genetic variation in European *L. rubellus* were tested in a statistical phylogeographic

framework (Figure 2.8; Table 2.5). The scenarios where five lineages (A1, C, F, G and B) simultaneously radiated from a common ancestor at either an early time point (5 Mya) or a later one (LIG) were not supported by our data. However, the simulations where divergence events occurred in two separate phases (a Pliocene event ~6 Mya and a Pleistocene event ~3 Mya, with Lineage B as the direct descendant population of the MRCA), resulted in patterns of diversity which did not significantly differ from the empirical data. These results suggest that the extant genetic diversity is the result of the disruption of a once "continuous" distribution. The "washhouse" climatic periods (Böhme et al. 2008) of the late Miocene caused an intensification of humid conditions on a continental scale. According to our results of the divergence time estimates for *L. rubellus* and its affinity to humid environments, it is likely that the species' dispersal over the continent was boosted by one of the "washhouse" periods. However, the populations became subsequently structured due to limited dispersal and vicariant events caused by successive climatic shifts such as the ice ages. Estimates of the rates of dispersal obtained from the Netherlands postulate that 10,000 years is a necessary timespan for worms to move over a distance of 100 km (Marinissen and Van den Bosch 1992) and the two "washhouse" periods of the Miocene described in the introduction spanned 500 Ky each. This "washhouse" effect could have provided favorable conditions for the spread of *L. rubellus* across the continental landmass. At the same time, such low levels of dispersal accounted for isolation and absence of gene flow during the Quaternary glaciations. Therefore, it is probable that some refugial populations of *L. rubellus* are descendants of the first demes that colonised the same geographical areas in the late Miocene. Successive climatic shifts, coinciding with the onset of the Ice Ages during the Pleistocene led to successive vicariant events (i.e. population isolation due to ice sheet movements and sea level fluctuations) that resulted in the divergence between demes, which are in some cases geographically proximate (i.e. G-H split; C-D split; A1-A2-A3 differentiation). In this scenario, Lineage B seems to be the direct descendant of the Miocene ancestral population. However, this hypothesis should be tested with the addition of nuclear DNA loci, which are likely to present deeper genealogies than mtDNA (Li and Durbin 2011).

The most remarkable result of our modelling is that both LGM models show a strip of continuous suitable landscape starting from the northwestern Iberian Atlantic shore, extending to the Pyrenees and northwards over the French Atlantic coast reaching Southwest Ireland and Cornwall (UK), a vast extension of land currently

under water (Figure 2.5). This corridor, inferred by our modelling, could have harboured a cryptic refugium for lineage B and possibly other lineages described in this study. The results support the previously observed cryptic species complex hypothesis for *L. rubellus* (King et al. 2008; Andre et al. 2009), and contribute to the evidence of cryptic divergence, a tendency observed in many annelid species since the application of genetic markers for taxonomic inferences (Erséus and Gustafsson 2009). There is increasing evidence in favour of the hypothesis that the presence of haplotypes private to northern populations may be a result of their survival in refugia in areas previously thought unsuitable for thermophilic species (Stewart and Lister 2001). For example, a population of South-West Scottish Caledonian Scot pines, appears from both genetic and paleo-botanical evidence to have spread from a cryptic South-Western refugium in Ireland (Kinloch et al. 1986). Finnegan *et al.* (2013) also found genetic evidence of a cryptic refugium in North-western France for populations of the Atlantic salmon (*Salmo salar* L.).

The data presented in this work potentially links the evidence for cryptic refugia in Ireland to a long-known biogeographic enigma known as the "Lusitanian element" (Corbet 1961; Moore 1987). Irish plant and animal species might be expected to be most similar to those of Britain. However, biodiversity is lower in Ireland and some of its species are absent or scarce in Britain. Contrastingly, there is sometimes a higher affinity between species present in Ireland and those present in the northern Iberian Peninsula and western France (Searle 2008), e.g. mammals such as the pigmy shrew (*Sorex minutus*) and invertebrates such as C*epaea nemoralis* (Grindon and Davison 2013). A possible link between the Lusitanian element and a similar SDM outcome as the one found in this study has been already suggested in a phylogeographic study on the pigmy shrew (Vega et al. 2010). However, while an effect of human mediated migration between these two regions has been confirmed as the most likely for some taxa (e.g. McDevitt et al. 2009), it is plausible that the observed North-western private haplotypes and the "Lusitanian element" are explained by the existence of a large coastal area of warmer climate and wet conditions that offered suitable habitat for many cold tolerant species, allowing gene flow and migration between the Iberian Peninsula and Ireland.

Lineage B peculiarly shows genetic evidence of survival in isolation from all the other southern European populations for millions of years, and no evidence has been found of its survival within Iberia, Italy or the Balkans. We hypothesize that Lineage

B survived glacial periods at the edge of the most northern periglacial part of the cryptic Irish refugium, and expanded into the British Isles as soon as the ice sheet retreated and suitable habitat became available. Consistent with such a hypothesis, we observed a signature of demographic expansions for this lineage ~25,000 years ago (Chapter 6, Figure 6.4). Very recently, a single lineage B *L. rubellus* from Ireland was found in a sample of earthworms recently genotyped from Ireland, adding support to this hypothesis.

Our results contribute to the understanding of cryptic diversity in annelid species, and add evidence to the cryptic refugia concept. Divergence values between lineages were high, with evidence of vicariant events occurring during the last 6 million years. This study shows that vicariance over great distances and isolation over long periods of time in combination with low dispersal rates are probably the main causes of cryptic speciation in *L. rubellus*. It is likely that these results are indicative of cryptic speciation for other similar earthworm species, but further studies need to confirm this hypothesis. Further studies are also needed to address the phenomenon known as the "Lusitanian syndrome". The SDM carried out in this study offers a possible avenue for further work in other species, and evidence from climatic, fossil and genetic data from other taxa is needed to confirm the existence of the Atlantic cryptic refugium we infer in this study.

2.5 Acknowledgments

This study was funded with the 2009 High Education Master & Back fellowship by the Sardinian government to Pierfrancesco Sechi, and Cardiff University. The authors thank all the collaborators who contributed samples, sequences or gave assistance in the sample collection process: Csuzdi Csaba (Hungarian Natural History Museum), Robert Donnelly (Royal Botanic Gardens, London, UK), Barbara Płytycz, Iwona Giska (Jagiellonian University, Poland), Jari Haimi (University of Jyväskylä, Finland), Mari Ivask (Tallin University, Estonia), Anita Juen (University of Innsbruck, Austria), Jan Lagerlöf (Swedish University of Agricultural Sciences, Sweden), Fuencisla Mariño (University of Vigo, Spain), Luis Cunha, John Morgan, Marta Novo (Cardiff University), Visa Nuutinen (MTT Agrifood Research, Finland), Maurizio Paoletti, Tommaso Zanetti (Università degli

Studi di Padova), Emilia Rota (Università degli Studi di Siena), Mirjana Stojanovic (University of Kragujevac, Serbia) and Frank Vandenbulcke (Université Lille 1, France).

2.6 Bibliography

Andre J, King RA, Stürzenbaum SR, Kille P, Hodson M, Morgan AJ. 2009. Molecular genetic differentiation in earthworms inhabiting a heterogeneous Pb-polluted landscape. *Environmental Pollution* **158**: 883-890.

Böhme M, Ilg A, Winklhofer M. 2008. Late Miocene "washhouse" climate in Europe. *Earth and Planetary Science Letters* **275**: 393-401.

Bouché MB. 1972. Lombriciens de France. Écologie et Systématique (n hors-série), Institut National de la Recherche Agronomique. in *Annales de Zoologie-Écologie Animale*.

Bundy JG, Keun HC, Sidhu JK, Spurgeon DJ, Svendsen C, Kille P, Morgan AJ. 2007. Metabolic profile biomarkers of metal contamination in a sentinel terrestrial species are applicable across multiple sites. *Environmental Science and Technology* **41**: 4458-4464.

Chakraborty R, Weiss KM. 1991. Genetic variation of the mitochondrial DNA genome in American Indians is at mutation-drift equilibrium. *American Journal of Physical Anthropology* **86**: 497-506.

Chang CH, Lin SM, Chen JH. 2008. Molecular systematics and phylogeography of the gigantic earthworms of the *Metaphire formosae* species group (Clitellata, Megascolecidae). *Molecular Phylogenetics and Evolution* **49**: 958-968.

Clement M, Posada D, Crandall KA. 2000. TCS: a computer program to estimate gene genealogies. *Molecular Ecology* **9**: 1657-1659.

Collins WD, Bitz CM, Blackmon ML, Bonan GB, Bretherton CS, Carton JA, Chang P, Doney SC, Hack JJ, Henderson TB. 2006. The community climate system model version 3 (CCSM3). *Journal of Climate* **19**: 2122-2143.

Corander J, Marttinen P, Sirén J, Tang J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* **9**: 539.

Corbet GB. 1961. Origin of the british insular races of small mammals and of the 'lusitanian' fauna. *Nature* **191**: 1037-1040.

Developement Team. 2009. Quantum GIS Geographic Information System, Open Source. Geospatial Foundation Project.

Donnelly RK, Harper GL, Morgan AJ, Orozco-terWengel P, Juma GAP, Bruford MW. 2013. Nuclear DNA recapitulates the cryptic mitochondrial lineages

of *Lumbricus rubellus* and suggests the existence of cryptic species in an ecotoxological soil sentinel. *Biological Journal of the Linnean Society* **4**: 780-795.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**: 214.

Edwards CA. 1996. *Biology and Ecology of Earthworms*. Springer.

Edwards CA. 2004. *Earthworm Ecology*. CRC Press.

Elith J, Kearney M, Phillips S. 2010. The art of modelling range-shifting species. *Methods in Ecology and Evolution* **1**: 330-342.

Erséus C, Gustafsson D. 2009. Cryptic speciation in clitellate model organisms. *Annelids in Modern Biology*: 31-46.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1**: 47.

Finnegan AK, Griffiths AM, King RA, Machado-Schiaffino G, Porcher JP, Garcia-Vazquez E, Bright D, Stevens JR. 2013. Use of multiple markers demonstrates a cryptic western refugium and postglacial colonisation routes of Atlantic salmon (*Salmo salar* L.) in northwest Europe. *Heredity* **111**: 34-43.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* **3**: 294.

Franklin J. 2009. *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, Cambridge, UK.

Fränzle O. 2006. Complex bioindication and environmental stress assessment. *Ecological Indicators* **6**: 114-136.

Fu Y-X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**: 915-925.

Global Invasive Species Database. 2013. *Lumbricus rubellus*. Available from: http://www.issg.org/database/species/ecology.asp?si=1711&fr=1&sts=&lang=EN [Acccessed 30 May 2013].

Grindon AJ, Davison A. 2013. Irish *Cepaea nemoralis* Land snails have a cryptic franco-iberian origin that is most easily explained by the movements of mesolithic humans. *PloS One* **8**: e65792.

Harpending HC. 1994. Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Human Biology*: 591-600.

Hewitt GM. 2000. The genetic legacy of the Quaternary ice ages. *Nature* **405**: 907-913.

Hewitt GM. 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. *Biological Journal of the Linnean Society* **58**: 247-276.

Hewitt GM. 2004. Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* **359**: 183-195.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**: 1965-1978.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337-338.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* **16**: 111-120.

King RA, Tibble AL, Symondson WOC. 2008. Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Molecular Ecology* **17**: 4684-4698.

Kinloch BB, Westfall RD, Forrest GI. 1986. Caledonian Scots pine: origins and genetic structure. *New Phytologist*: 703-729.

Kotík P, Deffontaine V, Mascheretti S, Zima J, Michaux JR, Searle JB. 2006. A northern glacial refugium for bank voles (*Clethrionomys glareolus*). *Proceedings of the National Academy of Sciences* **103**: 14860-14864.

Langdon CJ, Piearce TG, Black S, Semple KT. 1999. Resistance to arsenic-toxicity in a population of the earthworm *Lumbricus rubellus*. *Soil Biology and Biochemistry* **31**: 1963-1967.

Lavelle P, Bignell D, Lepage M, Wolters W, Roger P, Ineson P, Heal OW, Dhillion S. 1997. Soil function in a changing world: the role of invertebrate ecosystem engineers. *European Journal of Soil Biology* **33**: 159-193.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493-496.

Magri D, Vendramin GG, Comps B, Dupanloup I, Geburek T, Gömöry D, Latałowa M, Litt T, Paule L, Roure JM, Tantau I, Van Der Knaap WO, Petit RJ, De Beaulieu J-L. 2006. A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytologist* **171**: 199-221.

Marinissen JCY, Van den Bosch F. 1992. Colonization of new habitats by earthworms. *Oecologia* **91**: 371-376.

McDevitt AD, Rambau RV, O'Brien J, McDevitt CD, Hayden TJ, Searle JB. 2009. Genetic variation in Irish pygmy shrews *Sorex minutus* (Soricomorpha: Soricidae): implications for colonization history. *Biological Journal of the Linnean Society* **97**: 918-927.

Moore PD. 1987. Snails and the Irish question. *Nature* **328**: 381-382.

Morgan JE, Morgan AJ. 1988a. Calcium-lead interactions involving earthworms. Part 2: The effect of accumulated lead on endogenous calcium in *Lumbricus rubellus*. *Environmental Pollution* **55**: 41-54.

Morgan JE, Morgan AJ. 1988b. Earthworms as biological monitors of cadmium, copper, lead and zinc in metalliferous soils. *Environmental Pollution* **54**: 123-138.

Nordström S, Rundgren S. 1974. Environmental factors and lumbricid associations in southern Sweden. *Pedobiologia* **14**: 1-27.

Nylander JAA. 2008. MrModeltest v2.3. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Otto-Bliesner BL, Marshall SJ, Overpeck JT, Miller GH, Hu A, members CLIP. 2006. Simulating arctic climate warmth and icefield retreat in the last interglaciation. *Science* **311**: 1751-1753.

Perez-Losada M, Breinholt JW, Porto PG, Aira M, Dominguez J. 2011. An earthworm riddle: systematics and phylogeography of the spanish lumbricid *Postandrilus*. *PloS One* **6**: e28153.

Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**: 231-259.

Pisias NG, Moore Jr TC. 1981. The evolution of Pleistocene climate: a time series approach. *Earth and Planetary Science Letters* **52**: 450-458.

R Developement Core Team. 2013. R: A language and environment for statistical computing. *R Core Team*. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut A, Drummond AJ. 2007. Tracer v1.5, Available from http://beast.bio.ed.ac.uk/Tracer

Ray N, Adams JM. 2001. A GIS-based vegetation map of the world at the last glacial maximum (25,000-15,000 BP). *Internet Archaeology* **11**.

Richards CL, Carstens BC, Lacey Knowles L. 2007. Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography* **34**: 1833-1845.

Richmond GM, Fullerton DS. 1986. Summation of Quaternary glaciations in the United States of America. *Quaternary Science Reviews* **5**: 183-196.

Schmitt T. 2007. Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology* **4**: 1-13.

Searle JB. 2008. The colonization of Ireland by mammals. *The Irish Naturalists' Journal* **29**: 109-115.

Sims R, Gerard B. 1999. *Earthworms*. Linnean Society, London.

Sirocko F, Seelos K, Schaber K, Rein B, Dreher F, Diehl M, Lehne R, Jager K, Krbetschek M, Degering D. 2005. A late Eemian aridity pulse in central Europe during the last glacial inception. *Nature* **436**: 833-836.

Spurgeon DJ, Hopkin SP. 1999. Tolerance to zinc in populations of the earthworm *Lumbricus rubellus* from uncontaminated and metal-contaminated ecosystems. *Archives of Environmental Contamination and Toxicology* **37**: 332-337.

Spurgeon DJ, Weeks JM, Van Gestel CAM. 2003. A summary of eleven years progress in earthworm ecotoxicology: The 7th international symposium on earthworm ecology · Cardiff · Wales · 2002. *Pedobiologia* **47**: 588-606.

Stewart JR, Lister AM. 2001. Cryptic northern refugia and the origins of the modern biota. *Trends in Ecology & Evolution* **16**: 608-613.

Stewart JR, Lister AM, Barnes I, Dalén L. 2010. Refugia revisited: Individualistic responses of species in space and time. *Proceedings of the Royal Society B: Biological Sciences* **277**: 661-671.

Tajima F. 1989a. The effect of change in population size on DNA polymorphism. *Genetics* **123**: 597-601.

Tajima F. 1989b. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585-595.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**: 2731-2739.

Tiunov AV, Hale CM, Holdsworth AR, Vsevolodova-Perel TS. 2006. Invasion patterns of Lumbricidae into the previously earthworm-free areas of northeastern Europe and the western Great Lakes region of North America. in *Biological Invasions Belowground: Earthworms as Invasive Species*, pp. 23-34. Springer.

Vega R, Fløjgaard C, Lira-Noriega A, Nakazawa Y, Svenning J-C, Searle JB. 2010. Northern glacial refugia for the pygmy shrew *Sorex minutus* in Europe revealed by phylogeographic analyses and species distribution modelling. *Ecography* **33**: 260-271.

Warmerdam F. 2008. The geospatial data abstraction library. in *Open Source Approaches in Spatial Data Handling*, pp. 87-104. Springer.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical population biology* **7**: 256-276.

Willis KJ, van Andel TH. 2004. Trees or no trees? The environments of central and eastern Europe during the Last Glaciation. *Quaternary Science Reviews* **23**: 2369-2387.

CHAPTER 3 MITOGENOMICS OF CRYPTIC DIVERSITY: EXPLORING THE DEEP PHYLOGENETIC SIGNAL OF *LUMBRICUS RUBELLUS*

3.1 Introduction

Earthworms are the largest contributors to invertebrate biomass in many soil environments. As soil ecosystem engineers, they play key roles in soil ecology end ecosystem functioning (Lavelle et al. 1997). Interest in the applied ecology and biology of earthworms has resulted in an ever-increasing body of scientific literature (Edwards, 2004). Furthermore, earthworm species are widely used as ecotoxicological bioindicators (Spurgeon et al. 2003; Stürzenbaum et al. 2009). An increasing number of studies show evidence of the separation of earthworm taxa into distinct divergent genetic lineages which in many cases probably represent cryptic species (Chang et al. 2008; King et al. 2008; Erséus and Gustafsson 2009; Novo et al. 2010; Perez-Losada et al. 2011). It has also been shown that annelid intraspecific lineages can differ markedly in their tolerance to contaminants (Sturmbauer et al. 1999; Linke-Gamenick et al. 2000). Without an appropriate understanding of species boundaries, scientific results pertaining to earthworm biology can potentially be confusing or misleading.

Lumbricus rubellus is a lumbricid earthworm widely distributed in Europe, easily recognised by its dark red pigmentation. It is an out-crossing, hermaphrodite epigeic earthworm, with a lifestyle consisting of burrowing in the upper soil layer, feeding on decaying organic matter (Sims and Gerard 1999). *L. rubellus* was recently found to be divided into two highly distinct genetic lineages in Britain, thought to be the result of re-colonisation from different periglacial refugia (King et al. 2008; Andre et al. 2009; Donnelly et al. In Press). The previous chapter of this thesis revealed a broad pattern of cryptic diversity of the species over continental Europe.

Animal mitochondrial DNA comprises a small and simply structured extrachromosomal genome, usually including 13 protein coding genes, 22 tRNAs and 2 rRNAs and a few noncoding segments, in a circular double strand of ~ 15 Kb in size (Boore 1999). Whole mitochondrial genome sequences provide a useful tool for phylogenetic reconstruction in animal taxa and such studies range from tunicates (Singh et al. 2009) to mammoths (Gilbert et al. 2008). Nevertheless, the phylogenetic use of mitogenomic data in annelid studies is limited to a few studies, usually focusing on inter-species relationships (Boore and Brown 2000; Bleidorn et al. 2006; Zhong et al. 2008; Mwinyi et al. 2009; Shen et al. 2011). No mitogenomic study to

date has focused on questions related to phylogenetic relationships within an annelid cryptic species complex.

So far, studies of the genetic diversity of *L. rubellus* have included mtDNA markers (COI and COII), AFLPs and microsatellites (Harper et al. 2006; King et al. 2008; Andre et al. 2009; Donnelly 2011; Donnelly et al. In Press). In this chapter, the whole mitochondrial genome sequence from eight individuals from within each mitochondrial haplogroup identified in the previous dataset and the mitochondrial genome sequenced in the *L. rubellus* Genome Project are examined, with the aim to investigate the phylogenetic signal of the *L. rubellus* putative cryptic species complex at the whole mitochondrial level.

Using next-generation sequencing, we describe the whole mitochondrial genome of *L. rubellus*, its gene composition and arrangement, tRNA folding, nucleotide patterns, investigating whether there are differences with its closest mitochondrial genome in the literature, *Lumbricus terrestris* (Boore and Brown 1995). We investigated the phylogenetic signal contained in each gene, assessing whether different phylogenies constructed for each ortholog differed from each other and from a tree topology constructed on all concatenated genes. The aim was to assess whether some genes carried a stronger phylogenetic signal at the mitochondrial level than others, and if common phylogenetic statistical parameters were correlated to these differences. This is the first study where mitochondrial diversity of a cryptic species complex is examined at the genome level.

3.2 Materials and methods

3.2.1 Assembly and annotation

We selected eight individuals for whole genome next-generation sequencing (Table 3.1) to represent mitochondrial COI/COII haplogroups identified by Bayesian Analysis of Population Structure and initial phylogenetic analyses of *L. rubellus* (Chapter 2) and taking the geographic source for each sample into account. In order to extend the amount of phylogenetic information and explore divergence in this cryptic species complex, and to compare the evolutionary information contained at the different mitochondrial loci, a few representatives of the Lineages in chapter 2 were chosen for whole genome sequencing. Two A1 individuals coming from distant locations were selected (one individual is from UK, the other from Hungary), in conjunction with an A2 individual from France and an A3 individual from Finland, with the aim to gather deeper information about the A1-A2-A3 cluster diversity over a broad geographic range. Single individuals from haplotypes C, D (representing the Balkan lineages), F (representing the Spanish lineage) and B were included in the analysis as representatives of the other deep divergent lineages. Nuclear SNP data produced within this study is presented in chapter 4. Individuals from the other deep divergent clusters D and E were unfortunately not included in this analysis because they were discovered after Illumina sequencing was carried out.

Genomic DNA (gDNA) was extracted using the DNeasy blood and Tissue kit (Qiagen) with a modified protocol, assuring high purity and yield. Quantity and quality of the template were assessed using a NanoDrop® spectrophotometer (Nanodrop Technologies, Oxfordshire, UK) and a Qubit® 2.0 Fluorometer (Invitrogen, UK). Genomic DNA was sent to the GenePool (Edinburgh) for short-read (100 bp) paired-end library production and sequencing in an Illumina Hi-Seq 2000 system. The eight short read paired-end libraries yielded on average of approximately 80 million reads per individual. The paired-end output files of the analysis were imported and *de novo* assembled using CLC genomic workbench 5.5 (CLC bio), and contigs corresponding to the mitochondrial genomes were extracted as consensus sequences. The high average coverage of each mtDNA genome (~10^3 per individual) allowed us to extract mitochondrial genome sequences without ambiguous base calling, except for within the putative control region.

Genetic code	Country	Haplotype	Coordinates	
ECO.15	UK	A1	52°21'35"N	3°44'59"W
HUN.B.03	Hungary	A1	47°38'24"N	18°51'18"E
FRA.A.11	France	A2	50°41'07"N	3°02'37"E
FIN.A.11	Finland	A3	63°17'6.3"N	23°7'26"E
SER.A.16	Serbia	C	44°05'00"N	19°55'00"E
HUN.A.05	Hungary	D	47°39'50"N	16°26'10"E
SPA.A.03	Spain	F	42°09'59"N	8°41'00"W
S20	UK	B	52°24'57"N	3°51'47"W

Table 3.1. Samples used for this study. Genetic label, country of origin, haplotype and sample coordinates are shown. The last individual (S20) comes from a population studied in Andre et al. (2010).

All the individuals were chosen from the samples used in the study described in Chapter 2, except one, a lineage B individual (S20, Table 3.1) that originates from the same population previously studied by Andre et al. 2009, and where the specimen used for the *Lumbricus rubellus* Genome Project originated (Elsworth 2012). The aim of including this individual in the dataset was to integrate the dataset to the *L. rubellus* Genome Project as a re-sequencing project. In addition, another Lineage B mitochondrial sequence was downloaded from the *L. rubellus* Genome Project website and used to compare the outcomes of different assemblies.

Visualisation and annotation of the protein coding genes, transfer RNAs (tRNAs), short and long ribosomal RNAs (s- and lrRNAs) and non-coding regions was carried out by aligning the nine genomes to the *Lumbricus terrestris* annotated mitochondrial genome sequence (Boore and Brown 1995) using CLC Genomic Workbench 5.5 (CLC bio, Aarhus, Denmark). The secondary structure of tRNAs was inferred using ARWEN online (Laslett and Canbäck 2008). Mitochondrial maps were generated with the OGdraw suite (Lohse et al. 2007). The putative secondary structure of the non-coding genes were examined using the online service “mfold” (Zuker 2003).

3.2.2 Phylogenetic and statistical analyses

Each DNA alignment was analysed separately in order to obtain a phylogenetic tree from each protein coding and rRNA gene. Tests for the best fit models of molecular evolution were carried out using MrModeltest 2.3 (Nylander 2008). Phylogenetic analyses of each protein-coding and rRNA loci were carried out with mrBayes 3.2.1 (Ronquist et al. 2012). The Bayesian analyses were carried out over four independent MCMC chains running for 10^6 generations, with a tree sampled every 1,000 generations and a relative burn in of 25%. Convergence of runs and parameters was checked with the diagnostic implemented in MrBayes. Another phylogenetic tree was obtained from the concatenated dataset containing each coding and rRNA sequences in mrBayes. The Bayesian analysis was carried out implementing the proper model for each partition, over four independent MCMC chains running for 10^6 generations, with a tree sampled every 1,000 generations and a relative burn in of 25%.

Comparison of the alternative phylogenies obtained was carried out using a meta-tree approach (Nye 2008) using the online META-TREE java applet (http://www.mas.ncl.ac.uk/~ntmwn/phylo_comparison/multiple.html). This tool, based on minimisation of the total Robinson & Foulds distance (Robinson and Foulds 1981), provides a means to summarise, represent and compare sets of multiple alternative phylogenies via "meta-trees". This method does not provide a way to see "real" patterns of evolution, but is a useful way to visualise similarities and differences between phylogenies. In conjunction with bootstrapping, meta-trees can offer an approach to visualise and interpret whether different topologies produced by orthologous genes effectively carry different evolutionary patterns, or if there is a lack of significant difference due to poor phylogenetic signal (Nye 2008). Bootstrap analyses and construction of Maximum-Likelihood trees from the bootstrapped sequences for the meta-tree comparison of the single genes, as described by Nye (2008), were carried out using SEQBOOT and DNAML in the PHYLIP package (Felsenstein 1989).

The Symmetric Difference (Robinson and Foulds 1981) estimates how many partitions are present in one tree but not on the other. The metric was calculated to estimate the distance between the phylogenetic trees (Robinson and Foulds 1981) using the software TREEDIST in the PHYLIP package (Felsenstein 1989). Given two

unrooted trees and a set of taxa represented by different labels, the metric finds the number of required operations to convert one tree into another; the number of operations needed is the symmetrical distance definition. The authors defined two trees to be the same if they are isomorphic both in terms of branch structure and in terms of labelling.

A correlation between the phylogenetic information carried out by each gene and other variables from its alignment was hypothesised. To investigate this hypothesis, different variables were measured for each gene: p-distance between individuals, nucleotide diversity (π), gene length, number of polymorphic sites and Tajima's D (Tajima 1989) using DnaSP (Rozas and Rozas 1999). Secondly, matrices of residuals for gene length, polymorphic sites and π between each gene were built, and the fit to a normal distribution for each variable matrix was tested with Shapiro-Wilk tests (Shapiro and Wilk 1965). Thirdly, we checked how much of the variability between genes is explained by the variability between whole mtDNA sequences. A partial mantel test (Mantel 1967; Legendre and Legendre 1998) was carried out, comparing the p-distance matrices between each pairwise combination of genes, controlling for a third p-distance matrix calculated on the sequence obtained from the concatenation of all the genes. The results obtained express the level of partial correlation between the p-distance of two genes, while controlling for the p-distances of the whole coding part of the mitogenome. Lastly, correlations between the r values obtained from the partial mantel test, the symmetric distance of trees and the aforementioned variables, were calculated using mantel tests and visualised using pairwise plots in R (R Developement Core Team 2013).

3.3 Results

3.3.1 Genome structure and organisation

The mtDNA genomes of the oligochaete *L. rubellus* analysed in this work varied in length between 15640 and 15945. The nucleotide composition varied slightly between lineages, but was generallyAT-rich, with AT content between 62.4% (lineage B) to 63.4% (A1 from UK), close to the *L. terrestris* value of 61.6% (Boore and Brown 1995). *L. rubellus* mtDNA genomes contain the typical metazoan mitochondrial gene composition, comprising 13 protein-coding genes (COI-COIII, ND1-ND6, ND4L, ATPase6/8 and Cytochrome b), two ribosomal RNAs (srRNA and lrRNA) and 22 tRNAs, all encoding on the same strand. The gene order is identical to that observed in *Lumbricus terrestris* (Boore and Brown 1995). There are a few short intergenic regions, the largest between Cytochrome b and $tRNA^{Trp}$, and two main non-coding sites, one located between $tRNA^{Arg}$ and $tRNA^{His}$, corresponding to the *Lumbricus terrestris* AT-rich non-coding region, and a second one, variable in length between all the genomes, between ND6 and Cytochrome b. Most of the genes adjoin, with a few overlaps. Gene content and organization can be visualised in Figure 3.1 and are summarised in Table 3.2. Each of the 13 protein-coding genes features ATG as a start codon, a common feature in annelid mitochondrial genomes (Boore and Brown 2000; Boore 2001; Jennings and Halanych 2005; Bleidorn et al. 2006). Codon usage bias, represented in Table 3.3, appeared to be affected by the AT-richness of the genome, with 66% of codons ending with T or A, including the most used codons (CTA - Leu, 4.9%; ATA - Met, 4.9%: ATT - Ile, 4,6%; TTT - Phe, 4.1%; TTA - Leu, 3.9%).

Except for ND6 and Cytochrome b, the other genes have stop codons either overlapping with downstream genes or are truncated (T- or TA-) for completion with post-transcriptional polyadenylation, another common characteristic in annelid mtDNAs (Ojala et al. 1981; Wolstenholme 1992). In all lineages, the ND6 gene was 474 bp long and had a TAA or TAG termination codon preceding a large noncoding region (see Chapter 7 - Table 7.1, Table 3.1 and Figure 3.1). This region is not present in the *L. terrestris* genome, or in any other known annelid mitochondrial genome.

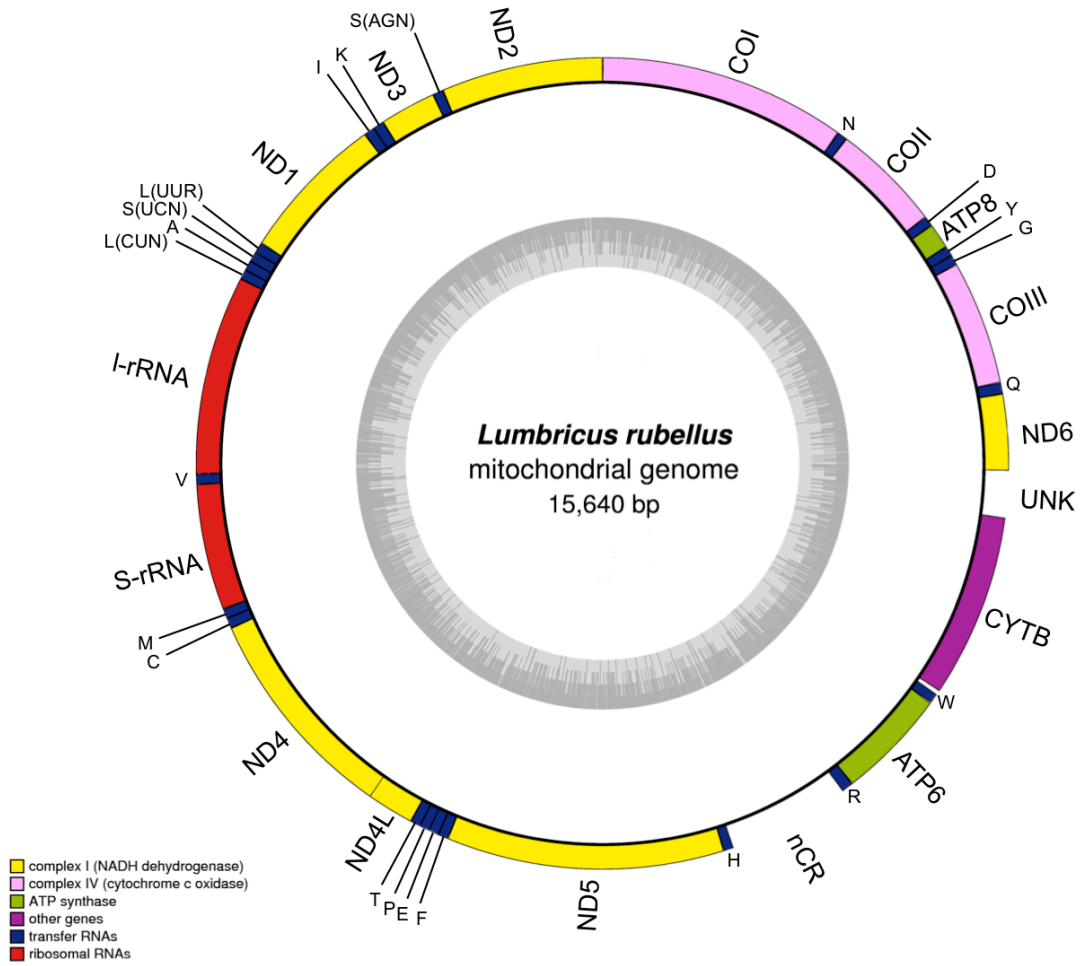


Figure 3.1. Gene map of the mitochondrial genome of *Lumbricus rubellus* (Lineage B). All the genes are encoded on the same strand. The tRNAs are represented by the IUPAC single letter codes. The inner circle represents the GC content.

Gene name	position from-to	Size	Start Codon	Stop Codon	Intergenic Nucleotides	AT%
COI	1-1540	1540	ATG	T[1]		58.57
$tRNA^{Asn}$	1541-1601	61				
COII	1602-2286	685	ATG	T[1]	1	59.42
$tRNA^{Asp}$	2288-2348	61				
ATP8	2349-2508	160	ATG	T[1]		67.50
$tRNA^{Tyr}$	2509-2571	63			-2	
$tRNA^{Gly}$	2570-2631	62			1	
COIII	2633-3410	778	ATG	T[1]	1	55.78
$tRNA^{Gln}$	3412-3480	69				
ND6	3481-3954	474	ATG	TAA		68.86
UNK	3955-4243	289				
CYTB	4244-5392	1149	ATG	TAA	24	61.51
$tRNA^{Trp}$	5417-5478	62			2	
ATP6	5481-6174	694	ATG	T[1]		59.65
$tRNA^{Arg}$	6175-6247	73				
nCR	6248-7002	755				
$tRNA^{His}$	7003-7065	63				
ND5	7066-8793	1728	ATG	T[1]	1	61.05
$tRNA^{Phe}$	8795-8856	62				
$tRNA^{Glu}$	8857-8920	64			-2	
$tRNA^{Pro}$	8919-8980	62				
$tRNA^{Thr}$	8981-9042	62				
ND4L	9043-9339	297	ATG	TAA	-7	63.64
ND4	9333-10689	1357	ATG	T[1]		62.71
$tRNA^{Cys}$	10690-10753	64				
$tRNA^{Met}$	10754-10816	63				
s-rRNA	10817-11601	785				60.59
$tRNA^{Val}$	11602-11664	63			1	
l-rRNA	11666-12906	1241				66.34
$tRNA^{Leu}$	12907-12967	61				
$tRNA^{Ala}$	12968-13029	62				
$tRNA^{Ser}$	13030-13092	63			1	
$tRNA^{Leu}$	13094-13157	64			1	
ND1	13159-14085	927	ATG	T[1]		62.57
$tRNA^{Ile}$	14086-14150	65				
$tRNA^{Lys}$	14151-14217	67				
ND3	14218-14569	352	ATG	T[1]		59.38
$tRNA^{Ser}$	14570-14633	64				
ND2	14634-15640	1007	ATG	T[1]		64.45

Table 3.2. Mitochondrial genome profile of *Lumbricus rubellus* Lineage B. Start and end position, length, start codon, stop codon, intergenic nucleotides and AT% are shown. The number of intergenic nucleotides is negative when there is an overlap between loci.

[1] Truncated stop codons, terminated by post-transcriptional modification (polyadenilation).

AA	Codon	N	%	/1000	AA	Codon	N	%	/1000
Ala	GCG	16	0.4	4.3	Pro	CCG	10	0.3	2.7
Ala	GCA	87	2.3	23.5	Pro	CCA	53	1.4	14.3
Ala	GCT	93	2.5	25.1	Pro	CCT	53	1.4	14.3
Ala	GCC	79	2.1	21.3	Pro	CCC	63	1.7	17
Cys	TGT	16	0.4	4.3	Gln	CAG	21	0.6	5.7
Cys	TGC	18	0.5	4.9	Gln	CAA	48	1.3	12.9
Asp	GAT	34	0.9	9.2	Arg	CGG	8	0.2	2.2
Asp	GAC	34	0.9	9.2	Arg	CGA	35	0.9	9.4
					Arg	CGT	7	0.2	1.9
Glu	GAG	23	0.6	6.2	Arg	CGC	11	0.3	3
Glu	GAA	46	1.2	12.4					
					Ser	AGG	15	0.4	4
Phe	TTT	152	4.1	41	Ser	AGA	69	1.9	18.6
Phe	TTC	104	2.8	28.1	Ser	AGT	12	0.3	3.2
					Ser	AGC	14	0.4	3.8
Gly	GGG	45	1.2	12.1	Ser	TCG	13	0.4	3.5
Gly	GGA	86	2.3	23.2	Ser	TCA	100	2.7	27
Gly	GGT	28	0.8	7.6	Ser	TCT	71	1.9	19.2
Gly	GGC	29	0.8	7.8	Ser	TCC	79	2.1	21.3
His	CAT	36	1.0	9.7	Thr	ACG	9	0.2	2.4
His	CAC	58	1.6	15.6	Thr	ACA	99	2.7	26.7
					Thr	ACT	78	2.1	21
Ile	ATT	172	4.6	46.4	Thr	ACC	55	1.5	14.8
Ile	ATC	123	3.3	33.2					
					Val	GTG	28	0.8	7.6
Lys	AAG	15	0.4	4	Val	GTA	108	2.9	29.1
Lys	AAA	73	2.0	19.7	Val	GTT	58	1.6	15.6
					Val	GTC	32	0.9	8.6
Leu	TTG	32	0.9	8.6					
Leu	TTA	146	3.9	39.4	Trp	TGG	11	0.3	3
Leu	CTG	36	1.0	9.7	Trp	TGA	92	2.5	24.8
Leu	CTA	182	4.9	49.1					
Leu	CTT	99	2.7	26.7	Tyr	TAT	65	1.8	17.5
Leu	CTC	58	1.6	15.6	Tyr	TAC	67	1.8	18.1
Met	ATG	73	2.0	19.7	End	TAG	0		0
Met	ATA	182	4.9	49.1	End	TAA	2	0.1	0.5
Asn	AAT	68	1.8	18.3					
Asn	AAC	78	2.1	21					

Table 3.3. The codon usage for the lineage B mitochondrial genome. For each amino acid (AA), the relative codon, the number of the codons in the genome, the relative percentage of the codon translating that amino acid, and the /1000 value are shown. The total number of codons is 3707.

The terminal part of the Cytochrome b gene was different from *L. terrestris* in length, and the stop codon position varied among *L. rubellus* genomes (Chapter 7, Table 7.1). As opposed to *L. terrestris*, the stop codon TAA was complete and a small non-coding region followed it. In order to validate the inference related to the termination of translation of Cytochrome b in all the lineages, the gene was aligned with the Cytochrome b sequences of *Urechis caupo* (NC_006379.1), *Orbinia latreillii* (NC_007933.1), *Perionyx excavatus* (NC_009631.1), *Nephytis* sp. (NC_010559.1), *Pista cristata* (NC_011011.1), *Terebellides stroemii* (NC_011014.1), *Sipunculus nudus* (NC_011826.1), *Phascolosoma esculenta* (NC_012618.1), *Urechis unicinctus* (NC_012768.1), *Whitmania pigra* (NC_013569.1). All Cytochrome b protein-coding regions of these species have a stop codon near the end point in the *L. rubellus* genomes, being shorter by 1-4 amino acids. Therefore, the inferred stop codon is probably correct.

ND4L peculiar termination is a common feature found in annelid genomes: an overlap of 5-7 nt over the ND4 gene (Boore and Brown 1995). ND3, ND5, ATP6 and COII genes either ended with a truncated stop codon or a full stop codon, but in the second case the gene overlapped with the following tRNA (Table 3.2) for one or two base pairs.

3.3.2 tRNAs and rRNAs

L. rubellus has the typical 22 tRNA genes contained in metazoan mitochondrial genomes (Boore 1999). The average length is 63.6 bp: the shortest tRNAs are 61 bp in length and the longest ($tRNA^{Arg}$) is 73 bp. In terms of structure and organisation, these genes were found to be conserved across lineages, and were organised in the same way as in *L. terrestris*. Lineage B tRNA structures were visualised in Figure 3.2 and the tRNA structure for the other genomes were reported in tRNA.pdf (CD-ROM).

The small and large ribosomal RNA (srRNA and lrRNA) genes were conserved among the lineages and with *L. terrestris*, with similar sizes, GC content (Table 3.2) and low levels of polymorphisms when compared to the rest of the genome (Table 3.4).

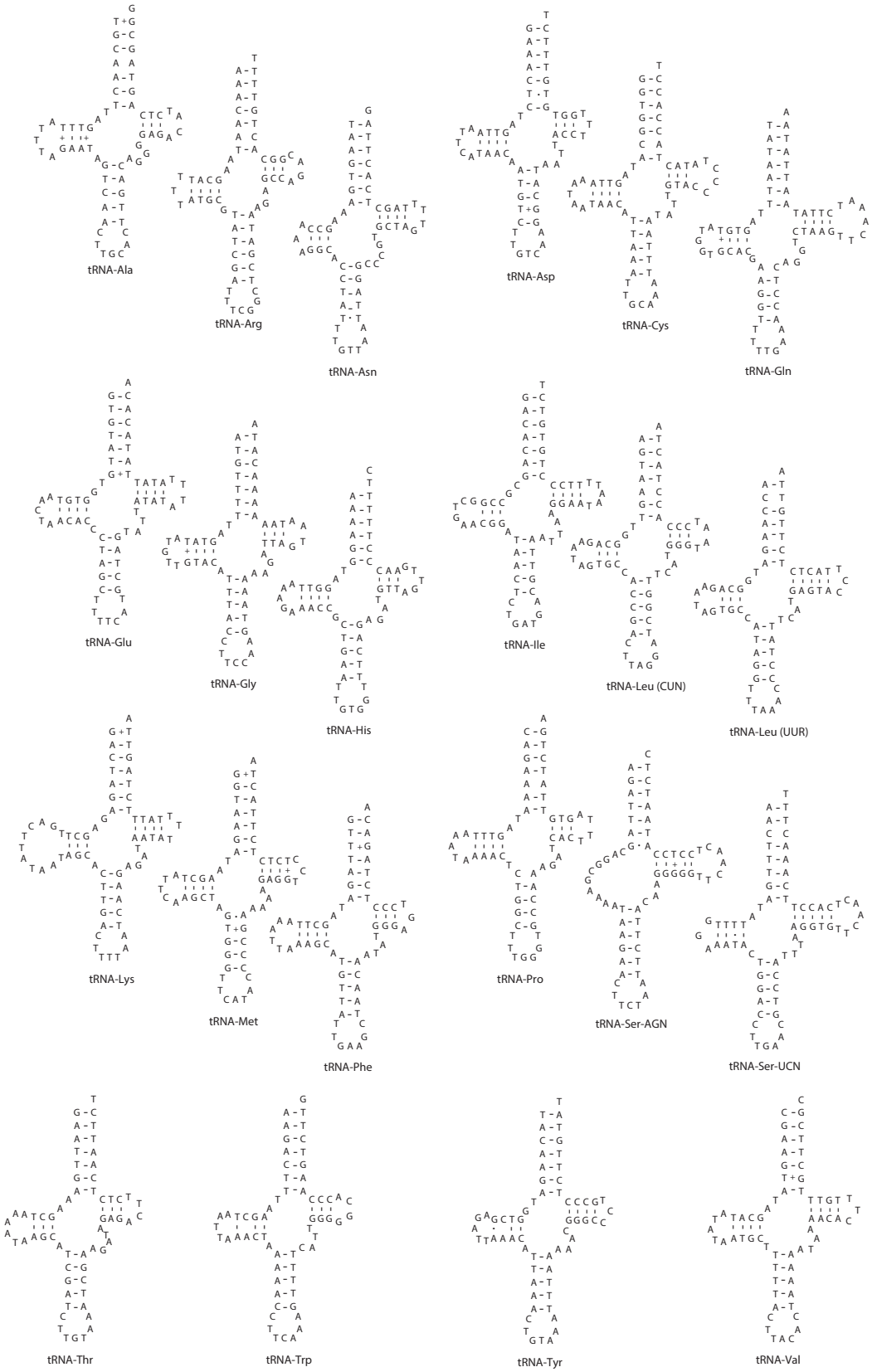


Figure 3.2. Transfer RNA (tRNA) structures of *Lumbricus rubellus* lineage B.

3.3.3 Non-coding regions

The region between $tRNA^{Arg}$ and $tRNA^{His}$ is the putative control region (Figure 3.1). However, it can be inferred to be complete only in ORI_B; the high AT content, resulting in many repetitive regions with microsatellite-like structures, is likely to have hindered the assembly software to completely recover the consensus sequence of this gene for the assemblies operated with the Illumina data. The ORI_B putative control region is more complete, based on the alignment with *L. terrestris*. The possible secondary structures in this region were examined using the "mfold" online server (Zuker 2003) (Figure 3.3) and revealed a high capacity of folding in a high variety of hairpins and complex structures within the temperature range at which *L. rubellus* survives.

The region between the ND6 gene and the Cytochrome B gene showed the most unusual divergence from the *L. terrestris* mitochondrial genome. In *L. terrestris* ND6 adjoins directly with CYTB, however, in all the *L. rubellus* lineages, after the stop codon of ND6, a non-coding sequence was highly divergent between lineages. This sequence was divergent both in sequence (average evolutionary divergence 69%, Kimura 2 parameter model), and in length, as this region ranged from 437 bp in A3 to 289 bp in B.

The consensus sequence of this region in B was identical to ORI_B. The two individuals originated from the same population and have substantially identical mitochondrial genomes; the fact that the same result has been obtained independently via different assemblies, leads us to assume in this case that the recovered consensus sequence of this gene is reliable. A possible secondary structure is shown in Figure 3.3.

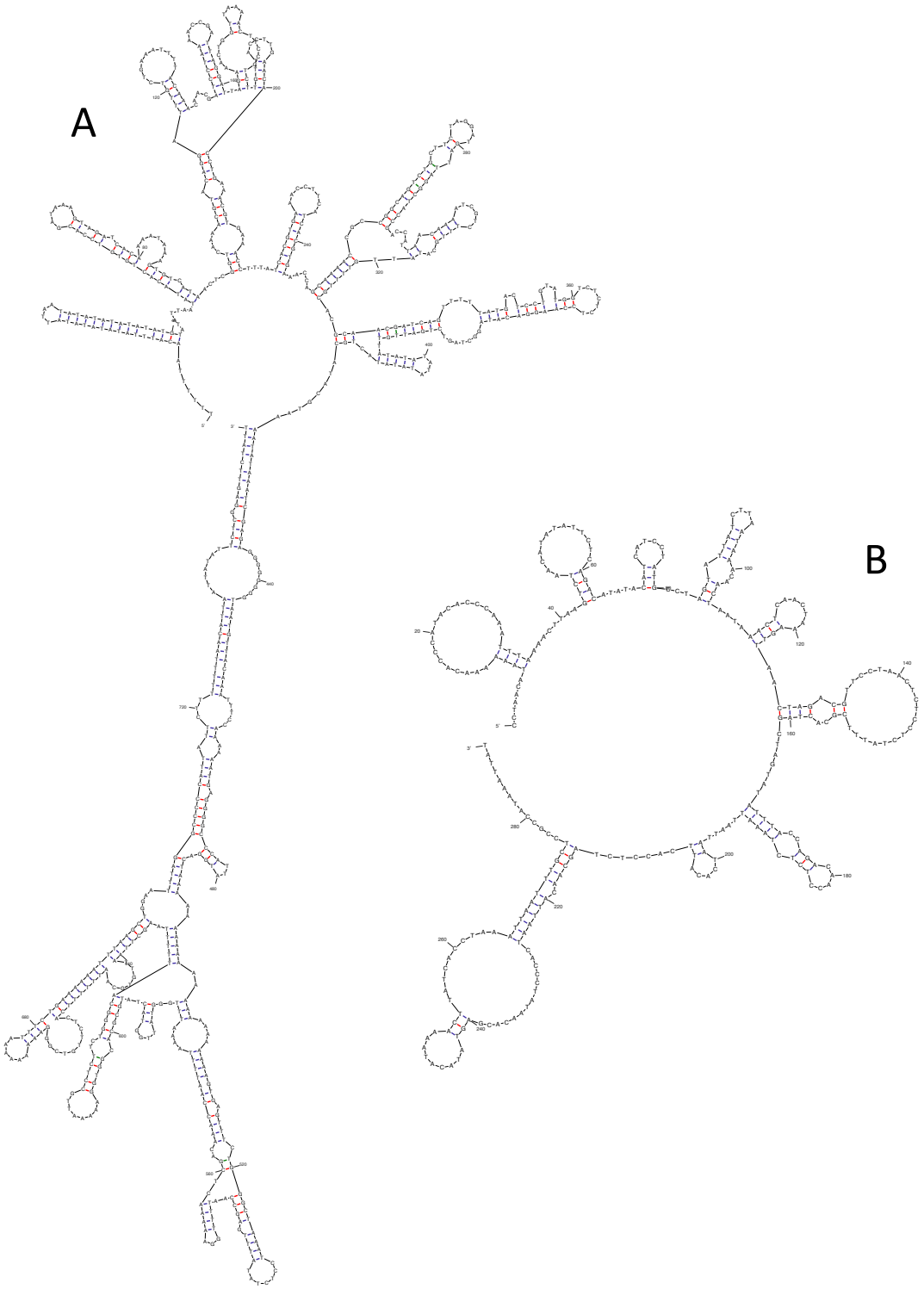


Figure 3.3. Possible secondary structures of the two major non-coding regions in Lineage B. A: AT-Rich region (dG = -190.34); B: UNK region following the ND6 gene (dG = -32.84).

3.3.4 Phylogenetic and statistical analyses

Assuming that a phylogenetic tree constructed using the concatenated sequence would represent the optimal phylogenetic hypothesis for the mitochondrial genealogy of the species, the main aim of this study was to assess how single mitochondrial genes would conform to that phylogeny, thus investigating which genes carry the strongest phylogenetic signal, and how differences could be related with common variables calculated using aligned sequences.

The evolutionary models obtained for each gene-coding region are represented in Table 3.4. Bayesian analyses of the different orthologous mitochondrial genes produced different topologies. Trees with posterior probability support values are reported in Chapter 7 – Figure 7.1. The 15 phylogenetic trees obtained resulted in ten different topologies. The topologies of COI, lrRNA, ND1 and ND5 matched the main topology of the tree obtained with all genes concatenated (Figure 3.4). In general, good support was evident, with posterior probability values close to 1 for almost all nodes (Chapter 7 – Figure 7.1). Other topologies deviated somewhat, usually in the resolution of the internal branching of A1-A2-A3 or the position of the C and D (Balkan) lineages (ND3, COII-CYTB, COIII-srRNA, ND4, ATP8). Others featured a different position for the (Spanish) lineage F (ND2, ATP6) but the most unusual topologies were represented by ND6 and ND4L (Figure 3.4). In some cases (COII, ATP8, ND3, ND6, ND4L), support for branches conflicting with the main topology was low (Chapter 7 – Figure 7.1).

Gene	Length	Evolutionary model	Variable sites	Tajima's D	π
COI	1540	GTR+I+G	355	0.833	0.112
COII	685	HKY+G	144	0.634	0.099
ATP8	160	HKY+I	45	0.238	0.137
COIII	778	GTR+I+G	187	0.478	0.127
ND6	474	GTR+I	147	0.393	0.151
CytB	1138	HKY+I+G	309	0.877	0.13
Atp6	694	GTR+G	194	0.842	0.139
ND5	1726	GTR+I+G	486	0.475	0.145
ND4L	297	GTR+I	84	0.676	0.141
ND4	1357	GTR+I+G	373	0.647	0.132
s-rRNA	793	GTR+I+G	74	-0.666	0.053
l-rRNA	1257	GTR+I+G	196	0.086	0.078
ND1	925	GTR+I+G	236	0.646	0.126
ND3	352	HKY+I	109	0.664	0.158
ND2	1006	GTR+I+G	291	0.326	0.145

Table 3.4. Summary statistics for each gene alignment. Evolutionary models were obtained using the Akaike Information Criterion selection of the best-fitting model of molecular evolution using mrmodeltest 2.3 (Darriba et al. 2012). Length, variable sites, Tajima's D and π statistics were obtained with the program DnaSP (Rozas and Rozas 1999).

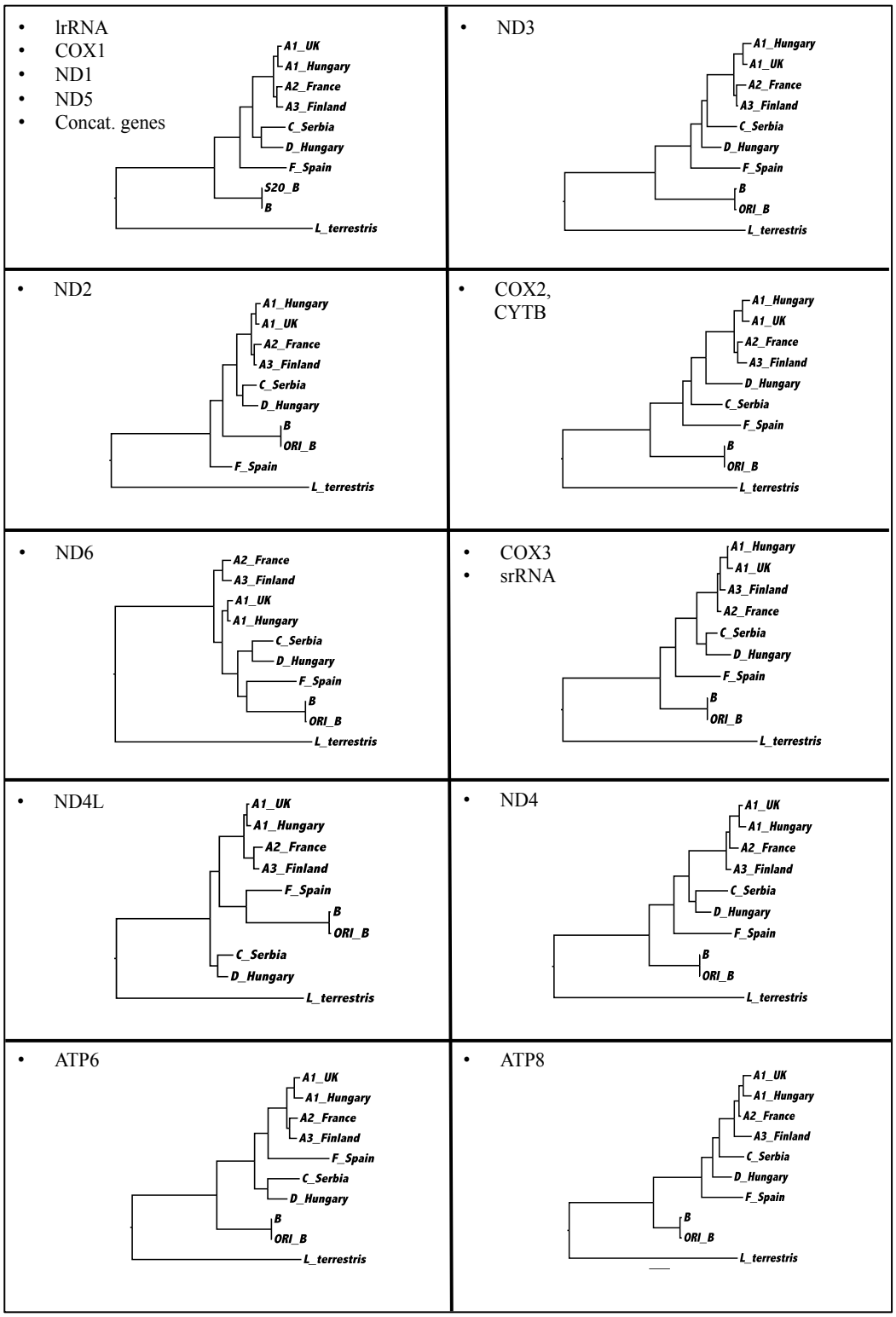


Figure 3.4. Representative gene tree topologies. Each topology is labelled by gene. The first topology (top-left of the figure) represents that obtained by constructing a tree using all the concatenated genes.

The difference in topology between the trees was summarised using a meta-tree approach (Figure 3.5). The meta-tree featured ten different topologies, represented by the small circles at the branching nodes, labelled according to the genes, showing a star-like appearance, with a central topology shared by 4 genes, lrRNA, ND5, ND1 and COI, concordant with the whole-genome topology as seen before. The four other branches summarised topologies differing by a maximum of three steps. The most divergent topologies are represented by ATP8 on the left branch and ND4L and ND6 on the right branch, divergent for 3 partitions.

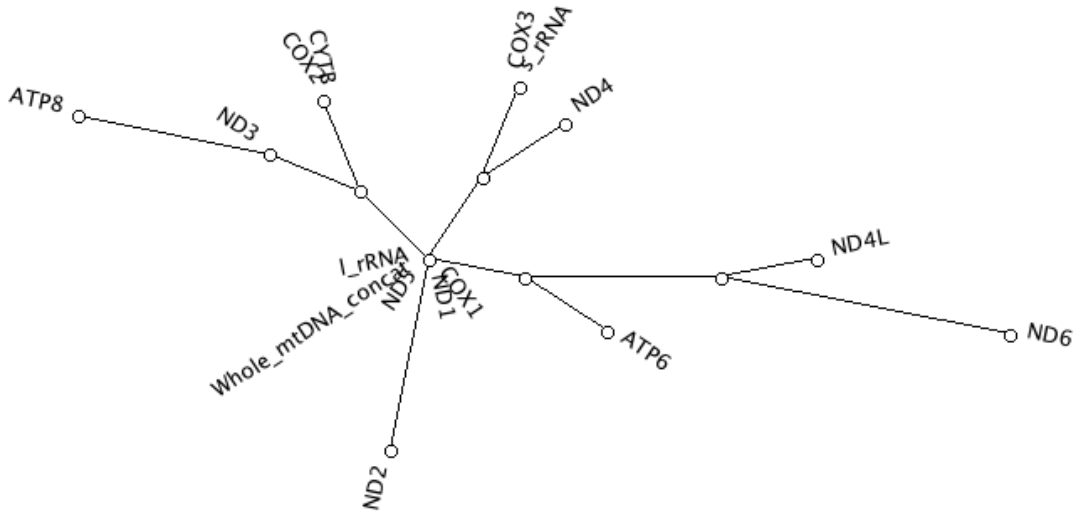


Figure 3.5. Meta-tree of Bayesian phylogenies of 15 different mitochondrial genes from the mitochondrial genomes of *Lumbricus rubellus* lineages. Each node represents intermediate topologies between the leaves; each leaf represents the phylogenetic tree topology of the genes that label them.

In order to assess whether these discrepancies were caused by statistically significant differences in the trees, or whether the differences were due to a lack of phylogenetic signal, we used a bootstrap approach as suggested by Nye (2008). We constructed a meta-tree in Figure 3.6 with eight bootstrap replicates obtained from each of the 15 genes. The tree again produced a star like pattern. Some genes, such as CYTB, ND6, ATP8 and srRNA were scattered, rather than forming separate clusters. A cluster occurred at the centre of the star-shaped tree, where the most of COI, lrRNA, ND1 and ND2 bootstrap topologies were found together.

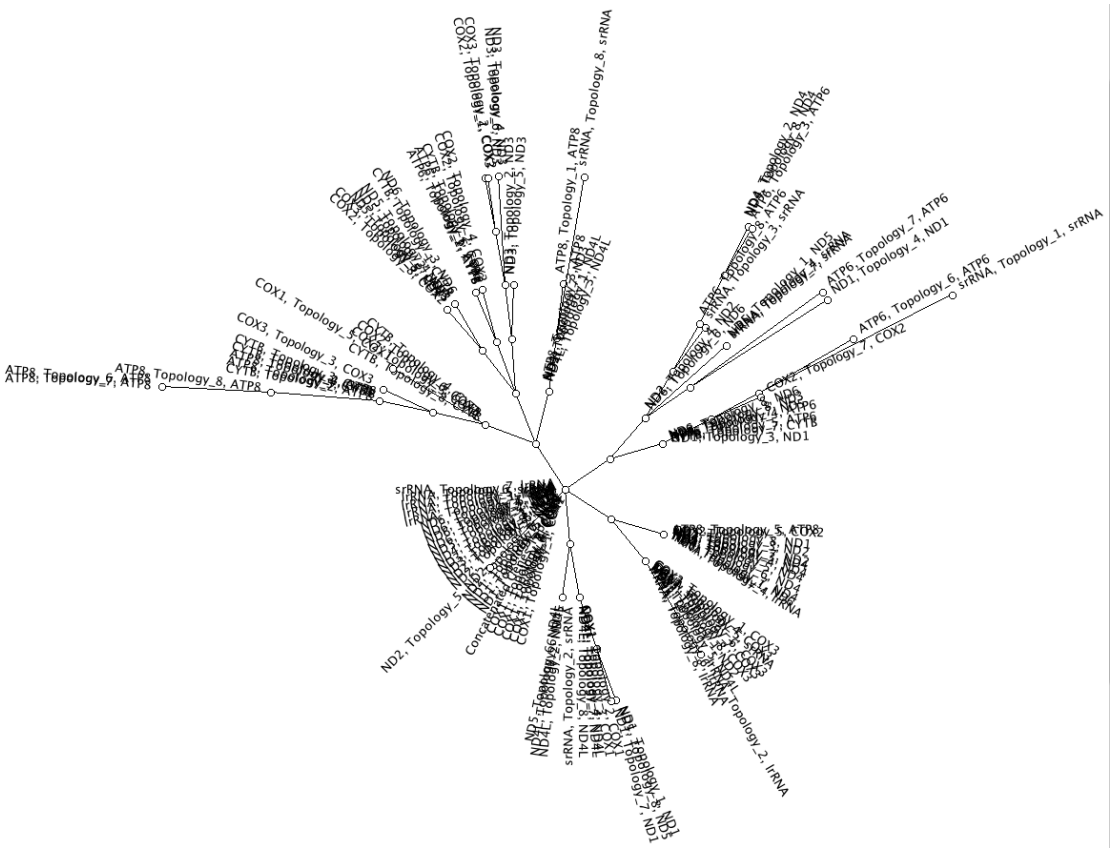


Figure 3.6. Meta-tree for the bootstrap analysis of 15 mitochondrial genes of *Lumbricus rubellus*. Eight bootstrap replicates for each gene are shown. Each labelled vertex corresponds to the topology of the gene or multiple gene replicates that labels them.

The Robinson & Foulds matrix in Table 3.5 show that the most divergent genes were very distinct from the reference topology (represented by COI) by 4 to 6 partitions. The symmetric distances between most genes ranged from 0 to 4, except for ATP8, ND6 and ND4L, for which distances ranged from 4 to 10.

The summary statistics calculated for each gene are shown in Matrices.docx (CD-ROM), and in Table 3.4 (length, number of variable sites, Tajima's D and π). The matrices constructed for each of the last four variables, showing the differences between each gene's variable values, are shown in Matrices.docx (CD-ROM). The Shapiro-Wilk test, calculated on all matrices, led us to reject the null hypothesis of normality for all the considered variables. Therefore, non-parametric statistics were used for correlation tests.

The PMT values calculated comparing p-distance matrices of each gene pair, using the Spearman's rank correlation method, controlling for the p-distance matrix

calculated on the concatenated gene sequence, were tabulated (Table 3.6). Out of 98 comparisons, only 33 comparisons were significantly correlated.

Correlation between summary variables, PMT and RF distance were calculated with Mantel tests using Spearman's rank correlation. Pairwise comparisons between variables were significantly correlated only between length and polymorphism, and between Tajima's D and π. These two quantities are linearly correlated as Tajima's $D = \theta_\pi - \theta_s$, where θ_π is the estimation of the $4Ne$ parameter using π, and θ_s is the estimation of $4Ne$ according to the proportion of segregating sites. Relationships between the variables, fitted with a polynomial curve are shown in Figure 3.7. The plot clearly shows that most of the variables were uncorrelated.

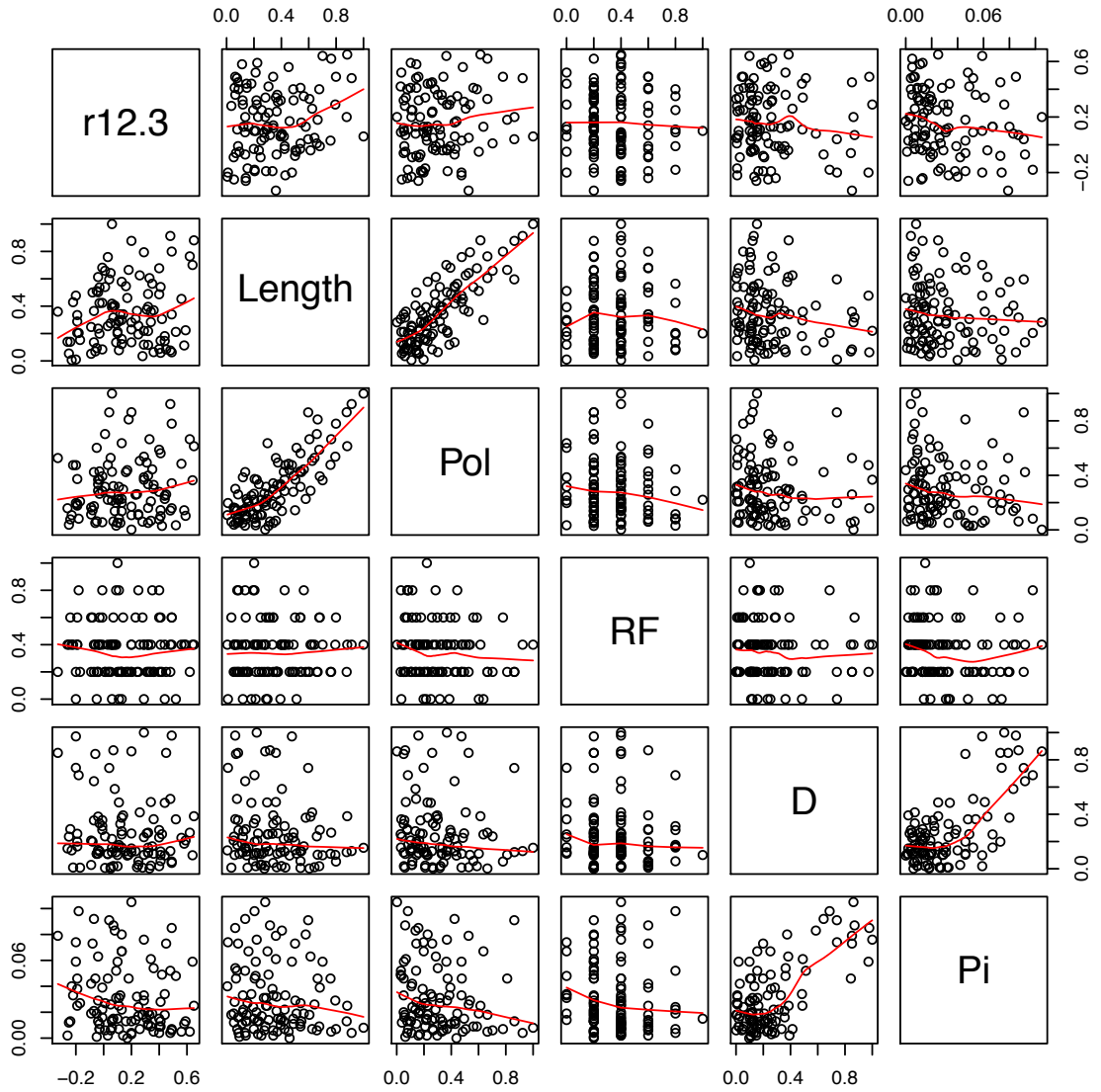


Figure 3.7. Pairwise plots of the relationships between the compared phylogenetic variables. R12.3: matrix of partial mantel test correlation values; Length: matrix of gene length differences; Pol: matrix of polymorphism differences between genes; RF: Symmetric distance matrix; D: differences in Tajima's D between genes; Pi: differences in π between genes. All the variables, except Length and Pol, and D and Pi, appear to be uncorrelated.

	COI	COII	Atp8	COIII	ND6	CYTB	ATP6	ND5	ND4L	ND4	srRNA	lrRNA	ND1	ND3	ND2
COI	0	2	4	2	6	2	2	0	4	2	2	0	0	2	2
COII	2	0	4	4	8	0	4	2	6	4	4	2	2	2	4
ATP8	4	4	0	4	10	4	6	4	8	2	4	4	4	2	6
COIII	2	4	4	0	8	4	4	2	6	2	0	2	2	4	4
ND6	6	8	10	8	0	8	6	6	4	8	8	6	6	8	6
CYTB	2	0	4	4	8	0	4	2	6	4	4	2	2	2	4
ATP6	2	4	6	4	6	4	0	2	4	4	4	2	2	4	4
ND5	0	2	4	2	6	2	2	0	4	2	2	0	0	2	2
ND4L	4	6	8	6	4	6	4	4	0	6	6	4	4	6	4
ND4	2	4	2	2	8	4	4	2	6	0	2	2	2	4	4
srRNA	2	4	4	0	8	4	4	2	6	2	0	2	2	4	4
lrRNA	0	2	4	2	6	2	2	0	4	2	2	0	0	2	2
ND1	0	2	4	2	6	2	2	0	4	2	2	0	0	2	2
ND3	2	2	2	4	8	2	4	2	6	4	4	2	2	0	4
ND2	2	4	6	4	6	4	4	2	4	4	4	2	2	4	0

Table 3.5. Symmetric distances calculated with the Robinson & Foulds method (Robinson and Foulds 1981). Each number represents the symmetric distance between each gene tree. The numbers describe the steps necessary to convert one gene tree into the other.

	COI	COII	ATP8	COIII	ND6	CYTB	ATP6	ND5	ND4L	ND4	srRNA	lrRNA	ND1	ND3	ND2
COI	1	0.23	0	0.01	0.53	0.1	0.51	0.07	0.06	0.03	0.84	0	0.42	0.45	0.29
COII	0.16	1	0.81	0.03	0.3	0.29	0.87	0.21	0.66	0.86	0.57	0.59	0.77	0.3	0.8
ATP8	0.65	-0.19	1	0.02	0.29	0.05	0.21	0.35	0.01	0	0.32	0	0.02	0.22	0
COIII	0.4	0.41	0.32	1	0.11	0.16	0.85	0.05	0.14	0.03	0.81	0.01	0.61	0.37	0.01
ND6	-0.01	0.09	0.1	0.25	1	0.54	0.89	0	0	0.04	0.76	0.64	0.82	0.31	0
CYTB	0.25	0.11	0.26	0.2	-0.01	1	0.01	0.02	0.56	0.91	0.02	0	0.09	0.66	0.57
ATP6	0.01	-0.23	0.14	-0.25	-0.24	0.34	1	0.46	0.43	0.58	0	0.01	0	0.01	0.57
ND5	0.29	0.18	0.06	0.33	0.49	0.31	0.01	1	0	0.59	0.42	0.31	0.6	0.19	0.25
ND4L	0.29	-0.09	0.4	0.23	0.58	-0.03	0.03	0.48	1	0.01	0.38	0.56	0.36	0.08	0
ND4	0.32	-0.22	0.62	0.36	0.35	-0.26	-0.03	-0.04	0.4	1	0.95	0.02	0.41	0.41	0
srRNA	-0.2	-0.06	0.08	-0.2	-0.18	0.29	0.49	0.04	0.07	-0.33	1	0.32	0.03	0.21	0.59
lrRNA	0.44	-0.07	0.64	0.42	-0.08	0.48	0.34	0.13	-0.04	0.36	0.11	1	0	0.21	0.18
ND1	0.06	-0.17	0.32	-0.05	-0.19	0.2	0.41	-0.05	0.07	0.05	0.45	0.52	1	0.13	0.69
ND3	0.03	0.1	0.13	0.09	0.11	-0.04	0.32	0.2	0.28	0.05	0.2	0.13	0.24	1	0.28
ND2	0.12	-0.19	0.41	0.44	0.49	-0.01	-0.03	0.17	0.56	0.59	-0.07	0.13	-0.09	0.06	1

Table 3.6. Partial Mantel test r values (lower diagonal) and p values (upper diagonal) obtained from the comparison between p-distance matrices calculated for each gene, comparing between the 8 genomes, and the p-distance matrix calculated on the concatenated sequence of all the genes. Significant values are highlighted in grey.

3.4 Discussion

The mitochondrial genome structure of *L. rubellus* was elucidated and reflects the structure and organisation of its closest studied mitochondrial genome (*L. terrestris*). All genes initiated with a unique start codon, ATG and although this represents a deviation from other metazoan genomes, where two or more codons are normally employed for the initiation of translation, this feature seems to be common in annelid mitochondrial genomes (Boore and Brown 1995; Boore and Brown 2000; Boore 2001; Jennings and Halanych 2005; Bleidorn et al. 2006).

The high frequency of truncated stop codons is also a common feature in annelids. Stop codons can either be truncated or overlap downstream genes after initiation of translation. It has been speculated that overlapping stop codons are not normally used and the transcripts are preferentially completed by cleavage and polyadenylation of the transcript, and the encoded overlapping TAA codons work only as a backup measure in case of cleavage errors (Boore and Brown 2000).

The AT-rich non-coding region was variable in length between the sequenced genomes (Chapter 7, Table 7.1), but this variation in length is likely to be a result of problems in the assembly of the putative control region. Given the fact that the control region of the ORI_B individual seems to be the most complete according to the alignment with *L. terrestris*, the estimated length of this individual can be considered the closest to the true length of the genome, notwithstanding that the B lineage could be unusual among all *L. rubellus* lineages. In Sanger sequencing – based mitogenomic studies, this region has previously been recognised to be difficult to amplify, particularly in long-targeted PCRs, probably because of regulatory secondary structures and microsatellite regions (Boore and Brown 2000; Boore 2001; Jennings and Halanych 2005; Bleidorn et al. 2006; Zhong et al. 2008). Repetitive regions have proven to be difficult to assemble with NGS data, particularly when repeats concur with the length of the reads (Li et al. 2010; Peng et al. 2013; Zhan et al. 2013). A possible way to overcome the problem, now that the flanking regions are known in all lineages, would require PCRs and Sanger sequencing targeted to amplify only the AT rich region.

The most remarkable difference between *L. terrestris* and *L. rubellus* genomes can be found in the non-coding strand at the end of the ND6 gene, a large and variable region without any detectable function. In all lineages, this gene is AT rich (average

72.7%) and showed a high capacity for forming folding structures and hairpins (Figure 3.3), suggesting that this region may have a regulatory role. The other Annelida mitochondrial genomes published so far do display only one large non-coding sequence, with minor interspersed non-coding sequences of unknown function, like in the polychaete *Orbinia latreillii* (Bleidorn et al. 2006). In invertebrates, a second large non coding region of 664 bp has been found recently in the mitochondrial genome of the clam *Mactra veneriformis* (Meng et al. 2013). They are common in cnidarians, where they have been proposed to be used as phylogenetic characters (Concepcion et al. 2006). These features have been recognised in some cnidarians, some sponges and a Placozoan, to be type II introns (Beagley et al. 1998; Dellaporta et al. 2006; Rot et al. 2006). Type II introns are self-catalytic ribozymes with different functions, common in plant and fungal mitochondria (Michel et al. 1989) but now found in all three life domains (Vallès et al. 2008). Remarkably, a polychaete annelid was recently found to be the only bilateran animal to have a type II intron in its mitochondrial genome (Vallès et al. 2008). It could be speculated that the described ND6-associated non-coding region in *L. rubellus* lineages could have some type II - intronic function, but this hypothesis must be tested further, as type II introns must correspond to determinate, measurable characteristics. A possible origin of this locus would account for horizontal gene transfer from endosymbiotic bacteria or viruses (Vallès et al. 2008). Horizontal gene transfer from endosymbionts to an insect host has been documented (Kondo et al. 2002), as well as the symbiotic relationship between bacteria of the *Verminephrobacter* genus and various earthworm species, including *L. rubellus* (Lund et al. 2010). The relationship between *Verminephrobacter* and earthworms was inferred to be ancient, suggesting that early symbionts were already present in the most recent common ancestor of lumbricids (Lund et al. 2010); however, no similarity between the genome and the plasmid of *Verminephrobacter* and the examined non-coding region in the earthworm mitogenome was found. An alternative explanation for interspersed non-coding regions could be described by the duplication/random loss model (Moritz et al. 1987), in which non-coding fragments could be remnants of duplicated and degenerated tRNAs. The tendency to form secondary loop-like structures would support this notion.

Phylogenetic analyses on single genes revealed a diverse result in terms of phylogenetic information carried out by each locus studied. Over the ten different topologies obtained from the Bayesian phylogenetic analysis of the 15 analysed loci

(all the coding genes plus the two rRNAs), four where identical with the tree topology obtained from the concatenated genes, but the others deviated by different degrees, the most different topologies being ND4L and ND6. In addition, some topologies were uncertain, with poor PP support for some branches (Chapter 7). These discrepancies may be due to different reasons. The ND4L and ATP8 genes featured the shortest sequence lengths (Tables 3.2 and 3.4); therefore, sampling bias, in terms of insufficient polymorphism to infer phylogenetic relationships, may be the reason why these gene trees are either incongruent or poorly supported (Felsenstein 1988). The ribosomal srRNA gene tree was the most conserved gene of all, having a maximum p-distance of 5% between individuals (Matrices.docx, CD-ROM). However, the gene topology showed concordance with the general topology, although with reduced branch length due to the small proportion of polymorphisms. In this case, balancing selection, which probably caused the retention of ancestral polymorphism, may be the reason for poor support (Maddison 1997). The conservation of rRNAs and tRNAs genes is probably due to their structural importance, thus accounting for their higher sequence similarity among taxa, despite the high mitochondrial mutation rate (Zhuang et al. 2013).

Peculiar evolutionary patterns were also observed for ND4L, ATP6 and ND6, which have already been reported in Boore and Brown (1995) for *L. terrestris*. Different amino-acid patterns between taxa for these genes were also observed in the brachiopod *Terebratulina retusa* (Stechmann and Schlegel 1999). These genes seem to evolve relatively fast: the low level of sequence conservation gave problems of identification of these proteins during annotations in the cited studies. In a study involving the annelids *Platynereis dumerilii*, *Helobdella robusta* and the pogonophoran *Galathealinum brachiosum*, the high degree of divergence of ND6 between taxa caused the decision to exclude it from whole mitochondrial genome phylogenetic analyses (Boore and Brown 2000). Hydrophilicity profiles were necessary in order to compare the protein product function. These patterns were very similar among these proteins in different taxa, suggesting that this factor is a functional constraint on their sequence evolution (Boore and Brown 1995). Comparisons between Annelida genomes revealed gene rearrangement events (e.g. Zhong et al. 2008); a certain degree of correlation between high mutation rates in mitochondrial genes and high degree of gene rearrangements has been observed in Hymenoptera and some gastropods (Yamazaki et al. 1997; Dowton et al. 2002). It is

possible that certain diverging evolutionary patterns may be caused by ancient rearrangement events.

The symmetric distance among trees, and the meta-tree analyses underlined the observed differences in phylogenetic signal both quantitatively and graphically (Figure 3.5 and Table 3.5). According to Nye (2008), a high degree of discordance between the tree topologies results in star-shaped meta-trees where genes are scattered on the leaves (Nye 2008, elaborated using the dataset in Li et al. 2007); on the other hand, a high frequency of genes clustered in one or more central nodes may imply high concordance between gene topologies, giving also an idea (but not an estimation) of the frequency of a topology in a gene set, which could relate to the probability of a particular phylogeny for the species tree (Nye 2008). The scenario depicted in this study points out that mitochondrial genes carry, as expected, a common pattern of phylogenetic information, with few cases of deviation from the general evolutionary signal. The meta-tree central clustering of topologies reflect the mitochondrial nature of the data, inherited effectively as a single locus (Moritz et al. 1987).

The bootstrap meta-tree (Figure 3.6) confirmed that a coherent phylogenetic signal was dominant among the gene trees. Bootstrap tree topologies clustered coherently overall with the meta-tree in Figure 3.5, with the genes closer to the main topology at the centre and the others clustering coherently in the other four branches, for the most part, for a maximum of three steps. Gene tree topologies that coherently cluster together in the bootstrapped meta-tree carry consistent phylogenetic information (Nye 2008). This method was used to test the effect of different data partitioning in phylogenies on tree topologies over a concatenated dataset of ten nuclear genes in ray-finned fishes (Actinopterigii), showing in general that high congruence between topologies build a star-like meta-tree, where topologies with the better phylogenetic signal cluster in internal nodes (Li et al. 2008).

The partial mantel test run comparing p-distance matrices of each pair of genes, controlling for the p-distance matrix calculated over the concatenated gene sequence, resulting in the matrix in Table 3.6, showed that only 33 comparisons resulted significant. The expectation was that significant correlation would relate genes which phylogenetic information was closer to the genomic phylogenetic information content, even when removing the effect of whole genome diversity from the correlation calculation. The results were partially discordant with other analyses. This may be because p-distance could be too simplistic to infer relationships in this

case. Some changes between pairs of taxa, such as back mutations, are undetectable for distance methods (Felsenstein 2004). The mantel tests between matrices of different statistical parameters (partial mantel test correlation values, p-distance, π, symmetric distance of trees, gene length, number of polymorphic sites and Tajima's D) gave significant results only in two expected cases, the comparison between length and polymorphism matrices and the comparison between Tajima's D and π matrices. The failure to detect correlation between variables in this dataset is probably due to reasons not easily detectable by the approaches taken. Perhaps, a data standardisation would help to detect underlying relationships within variables, as shown in a study that used a mantel test approach to detect correlation patterns between multiple distance matrices (Legendre and Lapointe 2004). In addition, this dataset may be affected by systematic errors that would hinder the considered variable to be good estimators of phylogenetic signals. Systematic errors are also called non-phylogenetic signals, and may consist of heterogeneity of nucleotide composition among taxa, rate variation across lineages and possibly within-site rate variation; such errors can be problematic also over large datasets, where sampling bias has been ruled out (Philippe et al. 2005).

These results support the hypothesis that the mitochondrial genome is not merely a selectively neutral locus; clearly, some genes either evolve faster or are more conserved than others, pointing out that purifying selection may exert differential constraints on different parts of the genome. There is increasing evidence pointing out that the mitochondrial selective neutrality assumption is invalid (Ballard and Kreitman 1995; Popadin et al. 2013). Researchers hypothesise that selective pressure in mtDNA may be maintained within populations via the joint mitochondrial-nuclear genotype, which can be considered as a unit of selection (Dowling et al. 2008).

Clearly, there are single genes which manage to identify the general phylogeny in a satisfactory way, with good support values, and confirm the status of COI as a suitable marker to recover cryptic differentiation among closely related earthworm taxa (Klarica et al. 2012), validating the results of the second chapter. However, the incongruences between phylogenetic information and bootstrap support between phylogenies point out that the use of multiple loci is more likely to recover correct phylogenies, overcoming sample bias (Jeffroy et al. 2006). However, non-phylogenetic signals, such as nucleotide composition bias among genes and within-rate variation (Philippe et al. 2005) may still cause the recovery of an incorrect

phylogeny from concatenated data, even with good support. Nucleotide compositional bias has been recognised as causing most problems in terms of congruence, but this occurs mainly in GC-biased datasets (Jeffroy et al. 2006) and all the genes considered are AT-rich. AT-rich genes in phylogenomic studies have been shown to minimise conflicts caused by compositional bias (Romiguier et al. 2013); therefore, we may assume that this dataset is free from this kind of error; on the other hand, within-rate variation could be the non-phylogenetic signal affecting ND4L, ND6 and ATP6, but the fact that the majority of the trees show topologies identical or close to the whole genome tree topology allows the inference that this kind of bias does not affect the general topology.

These results may have implications for future works on the biology and ecology of he *L. rubellus* species complex, particularly in the light of the discovery of its broad cryptic diversity. Single genes reported in this study and capable of retrieving the main topology with good statistical support, can be used singularly or in combination to reliably assess the lineage, when needed, in a quicker and cost effective manner.

3.5 Bibliography

Andre J, King RA, Stürzenbaum SR, Kille P, Hodson M, Morgan AJ. 2009. Molecular genetic differentiation in earthworms inhabiting a heterogeneous Pb-polluted landscape. *Environmental Pollution* **158**: 883-890.

Ballard JWO, Kreitman M. 1995. Is mitochondrial DNA a strictly neutral marker? *Trends in Ecology & Evolution* **10**: 485-488.

Beagley CT, Okimoto R, Wolstenholme DR. 1998. The mitochondrial genome of the sea anemone *Metridium senile* (Cnidaria): introns, a paucity of tRNA genes, and a near-standard genetic code. *Genetics* **148**: 1091-1108.

Bleidorn C, Podsiadlowski L, Bartolomaeus T. 2006. The complete mitochondrial genome of the orbiniid polychaete *Orbinia latreillii* (Annelida, Orbiniidae) - A novel gene order for Annelida and implications for annelid phylogeny. *Gene* **370**: 96-103.

Boore JL. 1999. Animal mitochondrial genomes. *Nucleic acids research* **27**: 1767-1780.

Boore JL. 2001. Complete mitochondrial genome sequence of the polychaete annelid *Platynereis dumerilii*. *Molecular Biology and Evolution* **2**: 1413-1416.

Boore JL, Brown WM. 1995. Complete sequence of the mitochondrial DNA of the annelid worm *Lumbricus terrestris*. *Genetics* **141**: 305-317.

Boore JL, Brown WM. 2000. Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: sequence and gene arrangement comparisons indicate that Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Molecular Biology and Evolution* **17**: 87-106.

Chang CH, Lin SM, Chen JH. 2008. Molecular systematics and phylogeography of the gigantic earthworms of the *Metaphire formosae* species group (Clitellata, Megascolecidae). *Molecular Phylogenetics and Evolution* **49**: 958-968.

Concepcion GT, Medina M, Toonen RJ. 2006. Noncoding mitochondrial loci for corals. *Molecular Ecology Notes* **6**: 1208-1211.

Dellaporta SL, Xu A, Sagasser S, Jakob W, Moreno MA, Buss LW, Schierwater B. 2006. Mitochondrial genome of *Trichoplax adhaerens* supports Placozoa as the basal lower metazoan phylum. *Proceedings of the National Academy of Sciences* **103**: 8751-8756.

Donnelly RK. 2011. An investigation of genetic heterogeneity in a biological sentinel species (*Lumbricus rubellus*). Ph. D. Thesis. University of Glamorgan/Cardiff University, Cardiff.

Donnelly RK, Harper GL, Morgan AJ, Orozco-terWengel P, Juma GAP, Bruford MW. 2013. Nuclear DNA recapitulates the cryptic mitochondrial lineages of *Lumbricus rubellus* and suggests the existence of cryptic species in an ecotoxological soil sentinel. *Biological Journal of the Linnean Society* **4**: 780-795.

Dowling DK, Friberg U, Lindell J. 2008. Evolutionary implications of non-neutral mitochondrial genetic variation. *Trends in Ecology & Evolution* **23**: 546-554.

Dowton M, Castro L, Austin A. 2002. Mitochondrial gene rearrangements as phylogenetic characters in the invertebrates: the examination of genome 'morphology'. *Invertebrate Systematics* **16**: 345-356.

Elsworth B. 2012. Unearthing the genome of the earthworm *Lumbricus rubellus*. Ph. D. Thesis. The University of Edinburgh, Edinburgh.

Erséus C, Gustafsson D. 2009. Cryptic speciation in clitellate model organisms. *Annelids in Modern Biology*: 31-46.

Felsenstein J. 1988. Phylogenies from Molecular Sequences: Inference and Reliability. *Annual Review of Genetics* **22**: 521-565.

Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5**: 164-166.

Felsenstein J. 2004. *Inferring phylogenies*. Sinauer Associates Sunderland.

Gilbert MTP, Drautz DI, Lesk AM, Ho SYW, Qi J, Ratan A, Hsu C-H, Sher A, Dalén L, Götherström A, Tomsho LP, Rendulic S, Packard M, Campos PF, Kuznetsova TV, Shidlovskiy F, Tikhonov A, Willerslev E, Iacumin P, Buigues B, Ericson PGP, Germonpré M, Kosintsev P, Nikolaev V, Nowak-Kemp M, Knight JR, Irzyk GP, Perbost CS, Fredrikson KM, Harkins TT, Sheridan S, Miller W, Schuster SC. 2008. Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. *Proceedings of the National Academy of Sciences* **105**: 8327-8332.

Harper GL, Cesarini S, Casey SP, Morgan AJ, Kille P, Bruford MW. 2006. Microsatellite markers for the earthworm *Lumbricus rubellus*. *Molecular Ecology Notes* **6**: 325-327.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *TRENDS in Genetics* **22**: 225-231.

Jennings RM, Halanych KM. 2005. Mitochondrial genomes of *Clymenella torquata* (Maldanidae) and *Riftia pachyptila* (Siboglinidae): evidence for conserved gene order in Annelida. *Molecular Biology and Evolution* **22**: 210-222.

King RA, Tibble AL, Symondson WOC. 2008. Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Molecular Ecology* **17**: 4684-4698.

Klarica J, Kloss-Brandstätter A, Traugott M, Juen A. 2012. Comparing four mitochondrial genes in earthworms – Implications for identification, phylogenetics, and discovery of cryptic species. *Soil Biology and Biochemistry* **45**: 23-30.

Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T. 2002. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. *Proceedings of the National Academy of Sciences* **99**: 14280-14285.

Laslett D, Canbäck B. 2008. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics (Oxford, England)* **24**: 172-175.

Lavelle P, Bignell D, Lepage M, Wolters W, Roger P, Ineson P, Heal OW, Dhillion S. 1997. Soil function in a changing world: the role of invertebrate ecosystem engineers. *European Journal of Soil Biology* **33**: 159-193.

Legendre P, Lapointe F-J. 2004. Assessing congruence among distance matrices: single-malt scotch whiskies revisited. *Australian & New Zealand Journal of Statistics* **46**: 615-629.

Legendre P, Legendre L. 1998. *Numerical Ecology*. Elsevier.

Li C, Lu G, Ortí G. 2008. Optimal data partitioning and a test case for ray-finned fishes (actinopterygii) based on ten nuclear loci. *Systematic Biology* **57**: 519-539.

Li C, Ortí G, Zhang G, Lu G. 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evolutionary Biology* **7**: 44.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC-C, Zhou Y, Cao J, Sun X, Fu

Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z ,Zheng H, Li Y, Steiner CC, Lam TT-Y, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q ,Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam T-W, Yiu S-M, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK-S, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J. 2010. The sequence and de novo assembly of the giant panda genome. *Nature* **463**: 311-317.

Linke-Gamenick I, Vismann B, Forbes VE. 2000. Effects of fluoranthene and ambient oxygen levels on survival and metabolism in three sibling species of *Capitella* (Polychaeta). *Marine Ecology Progress Series* **194**: 169-177.

Lohse M, Drechsel O, Bock R. 2007. OrganellarGenomeDraw (OGDRAW): a tool for the easy generation of high quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* **52**: 267-274.

Lund MB, Davidson SK, Holmstrup M, James S, Kjeldsen KU, Stahl DA, Schramm A. 2010. Diversity and host specificity of the *Verminephrobacter*-earthworm symbiosis. *Environmental Microbiology* **12**: 2142-2151.

Maddison WP. 1997. Gene trees in species trees. *Systematic biology* **46**: 523-536.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**: 209-220.

Meng X, Shen X, Zhao N, Tian M, Liang M, Hao J, Cheng H, Yan B, Dong Z, Zhu X. 2013. The complete mitochondrial genome of the clam *Mactra veneriformis* (Bivalvia: Mactridae): Has a unique non-coding region, missing atp8 and typical $tRNA^{Ser}$. *Mitochondrial DNA*.

Michel F, Kazuhiko U, Haruo O. 1989. Comparative and functional anatomy of group II catalytic introns - a review. *Gene* **82**: 5-30.

Moritz C, Dowling T, Brown W. 1987. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology and Systematics* **18**: 269-292.

Mwinyi A, Meyer A, Bleidorn C, Lieb B, Bartolomaeus T, Podsiadlowski L. 2009. Mitochondrial genome sequence and gene order of *Sipunculus nudus* give additional support for an inclusion of Sipuncula into Annelida. *BMC Genomics* **10**: 27.

Novo M, Almodóvar A, Fernández R, Trigo D, Díaz Cosín DJ. 2010. Cryptic speciation of hormogastrid earthworms revealed by mitochondrial and nuclear data. *Molecular Phylogenetics and Evolution* **56**: 507-512.

Nye TMW. 2008. Trees of trees: an approach to comparing multiple alternative phylogenies. *Systematic Biology* **57**: 785-794.

Nylander JAA. 2008. MrModeltest v2.3. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Ojala D, Montoya J, Attardi G. 1981. tRNA punctuation model of RNA processing in human mitochondria. *Nature* **290**: 470-474.

Peng Z, Lu Y, Li L, Zhao Q, Feng Q, Gao Z, Lu H, Hu T, Yao N, Liu K, Li Y, Fan D, Guo Y, Li W, Lu Y, Weng Q, Zhou C, Zhang L, Huang T, Zhao Y, Zhu C, Liu X, Yang X, Wang T, Miao K, Zhuang C, Cao X, Tang W, Liu G, Liu Y, Chen J, Liu Z, Yuan L, Liu Z, Huang X, Lu T, Fei B, Ning Z, Han B, Jiang Z. 2013. The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). *Nature Genet*ics **45**: 456-461.

Perez-Losada M, Breinholt JW, Porto PG, Aira M, Dominguez J. 2011. An earthworm riddle: systematics and phylogeography of the spanish lumbricid *Postandrilus*. *PloS One* **6**: e28153.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*: 541-562.

Popadin KY, Nikolaev SI, Junier T, Baranova M, Antonarakis SE. 2013. Purifying selection in mammalian mitochondrial protein-coding genes is highly effective and congruent with evolution of nuclear genes. *Molecular Biology and Evolution* **30**: 347-355.

R Developement Core Team. 2013. R: A language and environment for statistical computing. *R Core Team*. R Foundation for Statistical Computing, Vienna, Austria.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* **53**: 131-147.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Molecular Biology and Evolution* **30**: 2134-2144.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard Ma, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**: 539-542.

Rot C, Goldfarb I, Ilan M, Huchon D. 2006. Putative cross-kingdom horizontal gene transfer in sponge (Porifera) mitochondria. *BMC Evolutionary Biology* **6**: 71.

Rozas J, Rozas R. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**: 174-175.

Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples). *Biometrika* **52**: 591-611.

Shen X, Wu Z, Sun Ma, Ren J, Liu B. 2011. The complete mitochondrial genome sequence of *Whitmania pigra* (Annelida, Hirudinea): The first representative from the class Hirudinea. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics* **6**: 133-138.

Sims R, Gerard B. 1999. *Earthworms*. Linnean Society, London.

Singh TR, Tsagkogeorga G, Delsuc Fdr, Blanquart S, Shenkar N, Loya Y, Douzery EJP, Huchon De. 2009. Tunicate mitogenomics and phylogenetics: peculiarities of the *Herdmania momus* mitochondrial genome and support for the new chordate phylogeny. *BMC Genomics* **10**: 534.

Stechmann A, Schlegel M. 1999. Analysis of the complete mitochondrial DNA sequence of the brachiopod *Terebratulina retusa* places Brachiopoda within the protostomes. *Proceedings of the Royal Society of London Series B: Biological Sciences* **266**: 2043-2052.

Sturmbauer C, Opadiya GB, Niederstatter H, Riedmann A, Dallinger R. 1999. Mitochondrial DNA reveals cryptic oligochaete species differing in cadmium resistance. *Molecular Biology and Evolution* **16**: 967-974.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585-595.

Vallès Y, Halanych KM, Boore JL. 2008. Group II introns break new boundaries: presence in a bilaterian's genome. *PloS One* **3**: e1488.

Wolstenholme DR. 1992. Animal mitochondrial DNA: structure and evolution. *International Review of Cytology* **141**: 173-216.

Yamazaki N, Ueshima R, Terrett JA, Yokobori S-i, Kaifu M, Segawa R, Kobayashi T, Numachi K-i, Ueda T, Nishikawa K. 1997. Evolution of pulmonate gastropod mitochondrial genomes: comparisons of gene organizations of *Euhadra*, *Cepaea* and *Albinaria* and implications of unusual tRNA secondary structures. *Genetics* **145**: 749-758.

Zhan X, Pan S, Wang J, Dixon A, He J, Muller MG, Ni P, Hu L, Liu Y, Hou H. 2013. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nature Genetics* **45**: 563-566.

Zhong M, Struck TH, Halanych KM. 2008. Phylogenetic information from three mitochondrial genomes of Terebelliformia (Annelida) worms and duplication of the methionine tRNA. *Gene* **416**: 11-21.

Zhuang X, Qu M, Zhang X, Ding S. 2013. A comprehensive description and evolutionary analysis of 22 grouper (Perciformes, epinephelidae) mitochondrial genomes with emphasis on two novel genome organizations. *PloS One* **8**: e73561.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research* **31**: 3406-3415.

CHAPTER 4 PHYLOGENOMIC ANALYSIS OF *LUMBRICUS RUBELLUS* CRYPTIC CLADES

4.1 Introduction

The life-long interest of Charles Darwin in earthworms expressed in his first article (Darwin 1838) and in his last, best-selling book during his lifetime (Darwin 1892), had the effect to change prior common views on these organisms. From being considered soil pests, their useful and beneficial effects for soil and agriculture were finally appreciated, which subsequently stimulated research on them. Today, the primary functional roles that these animals play in their ecosystems have been thoroughly investigated and recognised (Jones et al. 1997; Lavelle et al. 1997).

Fittingly, the 200th anniversary of Darwin's Birthday in 2009 was coincident with the start of the first efforts to sequence and annotate an earthworm genome, the *L. rubellus* Genome Project (Elsworth 2012). *Lumbricus rubellus* is a common Holarctic worm used as a model and a sentinel species by ecotoxicologists, evolutionary biologists and soil ecologists. At the moment, the genome is in draft form: 90% of the genome is represented, it is free of contaminating sequences and it is in an appropriate form for general assessment of gene content (Chain et al. 2009), but it is still highly fragmented, comprising 315,000 scaffolds with an average length of 1,380 bp. Despite this high fragmentation level, the draft genome can be used as a reference to find homologous regions between genomes of individuals representative of cryptic lineages. The use of these genomic data could be also beneficial to infer the past demographic histories of these individuals. Recently, a pairwise sequentially Markovian coalescent (PSMC) model was applied to a complete human genome to infer the history of its population size changes in the last million years (Li and Durbin 2011). This technical breakthrough has enabled the possibility for researchers to infer population size changes using a single diploid genome.

During recent years, the development of sequencing technologies, and the consequent decrease in the cost-per base of sequencing experiments, have been primary factors for the constant shift from phylogenetics and phylogeography towards an 'omics' dimension, and have found an answer to the need of more data in these new technologies. Early phylogenetic and phylogeographic studies would compare one or a few genes, with a strong focus on cost-effective and rapid sequencing of mitochondrial DNA (mtDNA), to reconstruct evolutionary relationships among taxa and infer the spatial-temporal dimension of genetic variation (McCormack et al. 2013). However, data limitation was recognised to be an issue, and the field

consequently shifted to multilocus approaches, supported by theoretical developments in statistical phylogeography (Knowles 2009) and the species-tree paradigm argument in phylogenetics, which emphasises the inclusion of multiple loci to infer population and species histories to account for random variation in gene inheritance patterns, or coalescent stochasticity (Edwards 2009). However, the time and resources necessary to screen and develop multiple molecular markers have been a limiting factor for a long time. Hence, the introduction of Next Generation Sequencing (NGS) a few years ago, was enthusiastically welcomed by researchers in phylogenetics and phylogeography alike (McCormack et al. 2013). These disciplines are now at the interface of the field of phylogenomics.

Phylogenomics occupies the intersection between evolutionary biology and genomics (Eisen and Fraser 2003). A branch of this field is concerned with the use of genomic data to reconstruct the evolutionary relationship between organisms. Access to genomic data can solve issues related to sample size in previous phylogenetic studies. Even though phylogenomics studies have shown their power in reconstructing deep phylogenetic relationships in the tree of life (Delsuc et al. 2005), it is also important to use genomic information to investigate the degree of divergence between closely related taxa (eg peregrine and saker falcon, Zhan et al 2013), with the aim to broaden the knowledge of divergence processes in *L. rubellus* and discover if the observed mitochondrial (Chapter 2) and mitogenomic (Chapter 3) patterns can be observed also at the nuclear level.

In this chapter, single nucleotide polymorphisms from eight whole genomes of *L. rubellus* mapped against the *L. rubellus* draft genome (Elsworth 2012), and a Lineage B transcriptome, were used to a) investigate phylogenetic relationships between lineages at the genomic level, with the preliminary hypothesis that genomic variation patterns would reflect mitochondrial variation patterns, and b) to assess whether demographic signals could be retrieved from the genome data, to test the hypothesis that they match with demographic trajectories previously elaborated with the mitochondrial data. We wanted to answer the following questions: do we find the same pattern of observed mitochondrial diversity and inferred population history in the *L. rubellus* cryptic species complex at the nuclear genome level as with mtDNA? And can we recover the demographic history of each lineage, given the limitation of the current assembled genome?

4.2 Materials and methods

4.2.1 Illumina reads mapping

The data used for this chapter were produced in the context of the same NGS experiment used to gather the mitogenomic data analysed in the Chapter 3, from the same eight *L. rubellus* individuals selected from different lineages in different geographic locations across Europe. One of the individuals was a lineage B specimen from the same population of the animal originally used to sequence the genome, and it was chosen because of its AFLP and mitochondrial similarity with it (from the same population of Andre et al., 2009). The genomic DNA was purified and sent to the Gene Pool (University of Edinburgh). Library preparation and sequencing were carried out by the Gene Pool technicians according to the details described in Chapter 1 and in Chapter 3.

Bioinformatic analysis of the NGS data was carried out in a Bio-Linux 7 operative system (Field et al. 2006), using the informatics resources of the Organisms and the Environment (OnE) division, School of Biosciences, Cardiff University. The produced Illumina raw reads where quality-checked with the fastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and then mapped to the reference genome. For the genome alignment, a mapping performance test was carried out by comparing how the algorithms BWA-ALN (Li and Durbin 2009), BWA-SW (Li and Durbin 2010) and the CLC genomic workbench 6.5 suite (CLC bio) performed in terms of percentage of mapped reads produced. The difference between BWA-ALN and BWA-SW algorithms, both contained in the BWA package, is in the way they carry out the alignment. BWA-ALN performs a global alignment, which attempts to align every residue in every sequence; BWA-SW performs a local alignment, which looks for similarity patterns within the sequences to be aligned (Li and Durbin 2009; 2010). The CLC mapping algorithm is a local alignment by default, but the program also offers the option to perform a global alignment. Local alignments are more useful for more dissimilar sequences with expected patterns of similarity.

Two different mappings were carried out. For the first mapping, the transcriptome data of *L. rubellus* lineage B was used as a reference sequence in CLC. The lineage B transcriptome data were provided by Craig Anderson (Commonwealth

Scientific and Industrial Research Organisation, Australia) for analysis. The second mapping was carried out using the draft genome of *L. rubellus* as a reference (Elsworth 2012). The purpose of this strategy was to compare phylogenies obtained where the polymorphic information was restricted to exonic regions with those obtained with the whole genome SNP information. The aim was to check whether a subset of only transcriptome SNPs, supposedly coming from more conserved regions, could retrieve the phylogenetic signal when compared with whole genome SNP data.

4.2.2 Variant detection, demographic and phylogenetic analyses

A PSMC (Li and Durbin 2011) analysis was carried out on the Lineage B genome mapping obtained with BWA-SW. Firstly, a diploid genome was called, using the reference lineage B genome and the Illumina data from the S20 lineage B individual, using the mpileup function in SAMtools (Li et al. 2009). Secondly, the resulting file was used in the PSMC pipeline to retrieve the demographic signal.

Identification of variants was carried out for both datasets using the CLC suite. Variant calls were exported as tab-delimited text files. A variant filtering software (PedCreator, by Mario Barbato, Cardiff University) was used to identify and select polymorphic SNPs with allele count ≥5 and coverage values between 10 (inclusive) and 50. As the expected sequencing coverage was 20x, the filtering for SNPs with coverage lower than 10 was carried out to set a minimum threshold for reliable SNPs calls, whereas the filtering for a coverage higher than 50 was set to prevent the inclusion of high-coverage mitochondrial or bacterial genetic signals in the genotypes. The same software was used produced an output file in PLINK format (Purcell et al. 2007) only including SNPs shared by the eight populations. Further filtering was carried out to include only biallelic SNPs, assuming that polyallelic SNPS derived from sequencing errors, and only polymorphic SNPs were included in the final dataset.

Both SNP files were converted into Phylip format with the program TASSEL 3 (Bradbury et al. 2007) and a maximum likelihood phylogenetic analysis was carried out using the PHYML server (Guindon et al. 2010), with a General Time Reversible (GTR) substitution model, leaving the program to estimate the empirical equilibrium frequencies between sites and the gamma shape parameter. The branch support method implemented was a Bayesian-like transformation of the approximate

Likelihood Ratio Test (aBayes, Anisimova et al. 2011) implemented in the PhyML server. For the transcriptome mapped SNPs, a Bayesian phylogenetic analysis was carried out with MrBayes v3.2 (Ronquist et al. 2012). The analysis ran for 10^6 iterations, sampling every 1,000 iterations, over 4 chains in two independent runs. A General Time Reversible model with a gamma-shape parameter (GTR+G) was used, as it is the most general and inclusive model of molecular evolution, ideal for a selection of non-contiguous loci. The I parameter (proportion of invariable sites) was not used as no invariable sites were included in the analysis. Even though discarding invariant sites could cause errors in the estimation of the rate of variation among sites and bias the phylogenetic inference (Steel et al. 2000), the invariable sites were discarded from the analysis because A) it was necessary to reduce the amount of data to a more manageable size B) the calling of an unknown proportion of these sites was biased, as the reference genome was in a consensus haploid form and had a single random allele call chosen at each heterozygous position and C) the called invariant SNPs were still an underestimation of the real proportion of invariable sites across the genome. Convergence of the runs was checked with the MrBayes implemented diagnostic (standard deviation of split frequencies <0.05). A Bayesian phylogenetic analysis was also carried out with the genome-mapped SNPs. The dataset was run for 10^5 iterations, with 2 independent runs of 4 chains each, sampling every ten iterations. A GTR+G parameter was used also in this case and the convergence of the runs was checked with the diagnostic implemented in MrBayes.

4.3 Results

4.3.1 Genomes mapping

For each of the eight genomes, four files, corresponding to two paired-end libraries (two forward read files and two reverse read files, total size ~20 Gigabytes per individual), were received in compressed fastq format (see Table 4.1), yielding on average ~88 million reads per individual, with an estimated 20-fold coverage over the *L. rubellus* genome. The read length distribution is reported in Figure 4.1; as expected, the length distribution was uniform, with all the reads being between 100 and 102 in length. A significant reduction in the reads quality from the 85th nucleotide

towards the end could be observed (Figure 4.2). Illumina outputs were consistent in all the files, and the reported plots are representative of the overall features of each raw read library.

Individual	Reads n.	% CLC	% BWA-ALN	% BWA-SW
A2_France	89,803,646	72.18	20.39	57.92
A3_Finland	102,040,060	73.10		
A1_UK	81,111,033	71.72		
A1_Hungary	77,938,900	74.82		
C_Serbia	96,505,662	72.18		
D_Hungary	84,599,256	73.70		
B_UK	86,237,138	82.38	54.66	77.87
F_Spain	86,580,302	73.50		

Table 4.1. Comparison between performances of mapping algorithms. The first column reports the label of the sequenced individuals (Haplogroup_Location). The second column represents the total number of Illumina reads per individual. Columns 3, 4, and 5 report the percentage of reads mapped to the reference using three different mapping algorithms: CLC genomic workbench, BWA-ALN and BWA-SW.

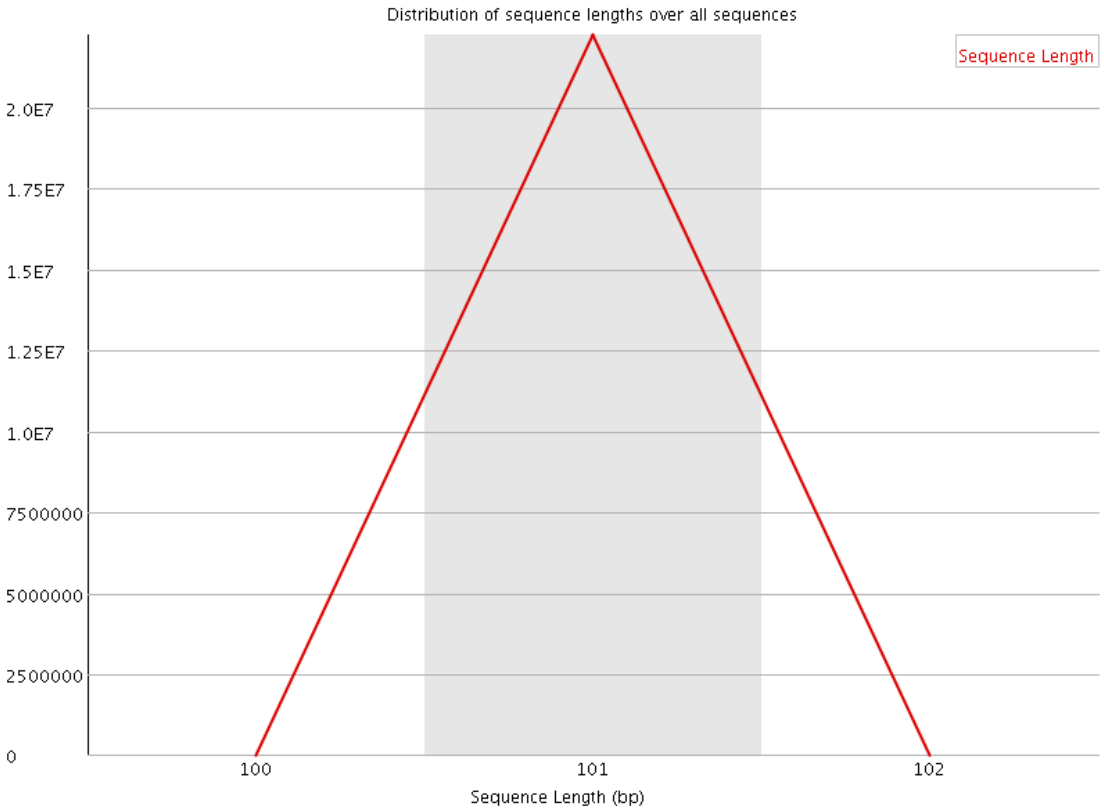


Figure 4.1. Sequence length distribution. Y axis represents the number of reads, the X axis represents the sequence length. The grey area represents the 90% of the reads.

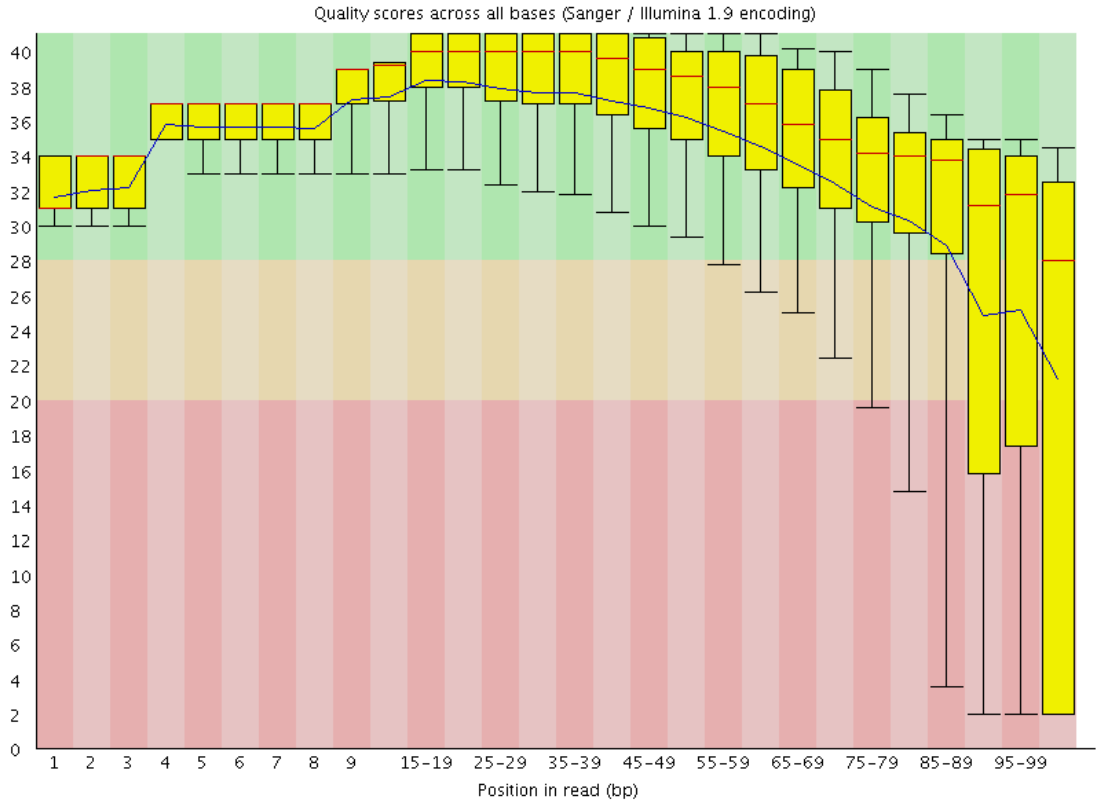


Figure 4.2. FastQC box plots of quality scores per read position of Illumina data). Y-axis: Phred quality scores. Higher scores mean better quality. The background is divided into three areas. The green area comprises good quality scores; the orange area = average and the red area = poor quality. The yellow area in the box represents the inter-quartile range from the 25th to the 75th percentile. The error bars include the 10th and the 90th percentile. The red lines inside the box plot are the median value of phred scores for the nucleotide range, and the blue curve represents the mean value.

The test carried out between the default settings of BWA-ALN, BWA-SW and the CLC genomic workbench mapping algorithm showed that, as expected, a local alignment approach was most efficient in order to obtain the best mapping efficiency. The test using two of the genomes (A2_France, from a distant lineage, and B, from the same population as the reference) was carried out using the three methods. As is shown in Table 4.1, the local alignment algorithm implemented in CLC was the most efficient in mapping the reads. PSMC tests conducted on the lineage B mapped genome were unsuccessful, probably because of the highly fragmented nature of the reference genome.

4.3.2 Phylogenetic analyses

After retrieval of the variant calls for each genome, the calling of common variants, the selection of only biallelic positions and the filtering of invariable polymorphisms, the final PLINK-format file obtained from the variant comparison of the eight *L. rubellus* genomes comprised 94,327 SNPs. After the same procedure, the genome-to-transcriptome variant comparison resulted in, as expected, a significantly smaller SNPs dataset of 4,993 markers.

The ML and Bayesian trees obtained with the exome SNPs are shown in Figure 4.3. The trees display remarkably different topologies, both showing good statistical support; in both cases, lineage B roots the tree, and the Hungarian A1 individual clusters together with the representative of the Balkan lineage C, but the other lineages differ in terms of position between the trees. In the ML tree, the Balkan lineage D follows lineage B as the most ancestral lineage, whereas in the Bayesian tree the British A1 individual takes this position. The Spanish lineage appears close to the A lineages in the ML tree, but additionally clusters with the Balkan D lineage in the Bayesian tree. The trees are quite discordant, and show very different configurations in comparison to the mitochondrial phylogenetic inference.

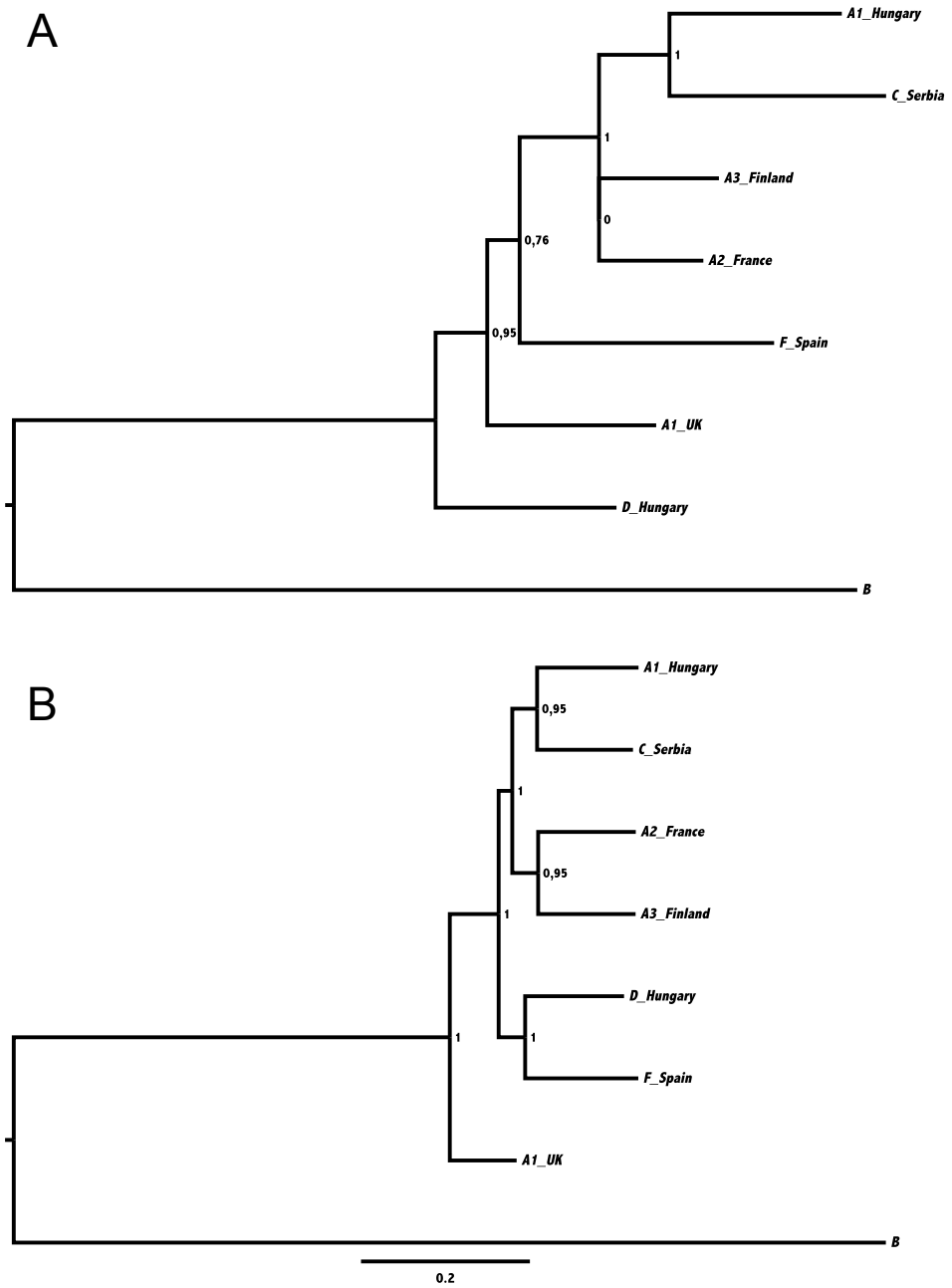


Figure 4.3. Phylogenetic trees obtained with the transcriptome SNPs. A: Maximum Likelihood tree obtained with the PHYML software; support values are Bayesian-like transformations of the approximate Likelihood Ratio Test (aBayes) implemented in the PhyML server (Guindon et al. 2010). B: Bayesian phylogenetic tree obtained with Mrbayes. Support values at the nodes are posterior probabilities.

As the topologies obtained in the ML and Bayesian analysis with the genomic data were identical, the tree shown in Figure 4.4 shows just one common topology. There is maximum branch support, both with posterior probabilities and Bayesian-like transformations of the approximate likelihood ratio test (aBayes) over all the nodes. The tree shows Lineage B as the most divergent lineage, followed by the Spanish Lineage F, the Balkan specimen represented by lineage D, and the monophyletic group between the specimens A1 from UK, A2 from France and A3 from Finland. This configuration is mostly in common with the mitochondrial phylogenies observed in previous chapters; the discrepancies occur in the topology relative to the British A1 specimen, which clusters with one of the Balkan specimens, and the fact that the two Balkan lineages cluster separately.

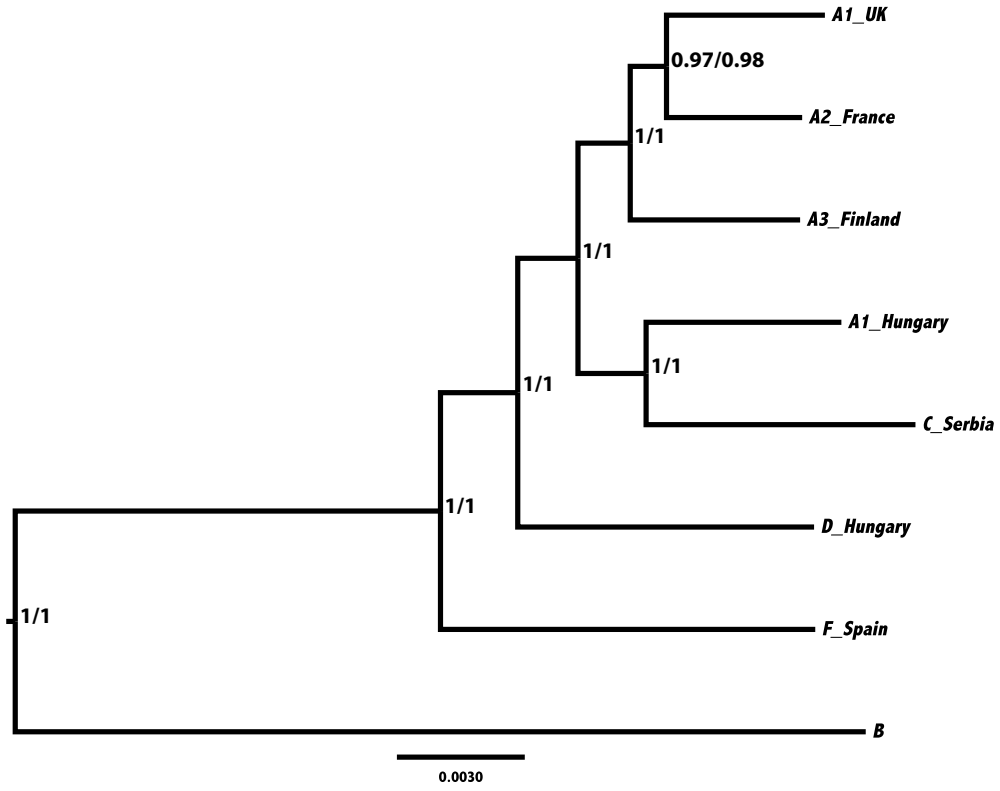


Figure 4.4. Bayesian majority rule consensus tree obtained with a dataset constructed with genomic SNPs. Support values shown are relative to posterior probability/Bayesian-like transformation of the approximate Likelihood Ratio Test (aBayes) implemented in the PhyML server (Guindon et al. 2010).

4.4 Discussion

In this study, eight individuals, representative of *L. rubellus* mitochondrial cryptic clades across Europe, were selected for whole-genome NGS sequencing. The plot in Figure 4.3 shows that the data quality is good overall, with the most of the reads having an average phred score between 30 and 40, with significant dropping in quality only over a minor fraction of the distribution of the read nucleotides (Figure 4.2). As expected, mappings using a local alignment algorithm were more efficient than global alignment approaches. Global alignments perform better with similar sequences, with roughly equal size (Needleman and Wunsch 1970). However, the divergence between the genomes analysed in this study was probably too high to obtain good quality mappings using a global alignment. Nonetheless, the Lineage B individual chosen for this study comes from the same population as the reference genome individual. In Chapter 3 it was shown that these individuals are almost identical at the mitochondrial level. A Principal Component Analysis (PCA) carried out with AFLP data in Andre et al (2009) clusters these individuals together also at the nuclear level. The low global alignment mapping success of this individual (only 54%, Table 4.1) could perhaps be explained by previously undetected patterns of genetic divergence between these individuals. Early mapping tests using less stringent parameters and increased thresholds for mismatch penalties did not improve the assemblies and were more computationally intensive than local alignment approaches. Hence, the reason behind the average mapping quality when mapping Lineage B against the reference B genome with global alignment requires further parameter testing to be fully understood. A possible reason could be the fragmented nature of the reference genome. Local alignment algorithms attempt to map the entire sequence of the read to the reference: if the genome is highly fragmented, a significant part of the reads could map at the extremities of the scaffolds, resulting in the reads that do not overlap with enough bases being discarded, according to the threshold rules of the global aligner. Local alignment mapping problems when assembling illumina short reads were in agreement with other studies (Kofler et al. 2011), but inconsistencies of global alignment were already noted in earlier studies about genome mapping (Brudno et al. 2003), where it was observed that early local alignment algorithms, such as Needleman and Wunsch (1970), Dialign (Morgenstern 1999), MUMer

(Delcher et al. 1999, 2002) Avid (Bray et al. 2003) could not handle efficiently rearrangements, gaps and mismatches when mapping sequences to a reference.

The two local alignment strategies, carried out with the BWA-SW and the CLC genomic workbench, performed better than the global BWA-ALN algorithm. The CLC genomic workbench mapping performance proved to be the best, as is clearly shown in Table 4.1. The reason why CLC worked better than others is perhaps due to a less stringent/more flexible mapping algorithm, with lower mismatch and/or gap penalties. Unfortunately, a deeper understanding of the mechanisms of the CLC mapping algorithm is not possible, as this is commercial software for which the source code is not available. Although the local alignment algorithms mapped the genomes most efficiently, this increase in mapping success could hide a bias: it has been observed that local alignment algorithms can operate soft masking: if a mismatch between the reference and a locally aligned read is observed at the end of the reads, this mismatch can be ignored, resulting in an increased likelihood of incorrectly mapped reads (Degner et al. 2009). This kind of bias has resulted in an erroneous increase of reference allele frequencies in previous studies that have compared mapping performances (Degner et al. 2009; Kofler et al. 2011). A way to limit the problem, described in Kofler et al. (2011), is to carry out a mixed approach (Paired End Smith-Waterman remap, PE-SW), using local alignment to map paired end reads separately, and in case of only one read of the pair mapping, a local alignment approach can be used to map the other read. The success of assemblies carried out in this way was assessed comparing numbers of pairs correctly mapped in different runs optimising for different gap and mismatch parameters. The opportunity to use "glocal" (global+local) alignment approaches was already considered and tested with success in early studies about genome mapping (Brudno et al. 2003). This method was not implemented here for two reasons: A) a limit in informatics resources and time B) the fact that the comparison of properly paired reads requires a defined structure of *L. rubellus* chromosomes; thus, it was impossible to assess the proportion of incorrectly mapped paired reads in local alignment, which could have been consistent considering the fragmentation of the reference (Elsworth 2012).

The PSMC analysis attempted on lineage B genome did not succeed, probably because of the nature of the PSMC algorithm. The fragmented nature of the draft genome probably did not allow a reliable inference of past recombination events. The program infers the time to the most recent common ancestor (TMRCA) on the basis

of the density of heterozygotes in a diploid sequence, and estimates past changes in population size from a model that approximates the coalescent from ancestral recombination events (Li and Durbin 2011). All the studies that have used this technique to date, used complete genomes, or drafts with high N50 scaffold values, allowing them to infer recombination events over great distances (Cho et al. 2013; Freedman et al. 2013; Ibarra-Laclette et al. 2013; Orlando et al. 2013; Prado-Martinez et al. 2013; Wan et al. 2013; Zhan et al. 2013). The *L. rubellus* genome will probably be completed and ready for publication in 2014. The use of these data in conjunction with a completed reference genome may allow interesting insights on the different demographic histories of the lineages.

Despite being derived from the same 4,993 SNP dataset, the ML and Bayesian phylogenetic trees for the transcriptome SNPs dataset (in Figure 4.4) showed very different cluster solutions. Both trees showed that Lineage B is the most divergent lineage and that the Hungarian A1 clusters with the Serbian C, but the other inferred phylogenetic relationships varied widely. It is remarkable that in almost all cases, these incongruent relationships are statistically well supported. An important challenge to phylogenetic inference, particularly in the era of phylogenomics, is to ensure that topologies converge towards the correct answer, as an increasing number of characters are included in the analysis (Felsenstein 1988). The construction of phylogenies from single gene or a few orthologous genes often lead to incongruent topologies, but usually in these cases the lack of statistical support helps to solve the issue. In phylogenomics, inconsistency can be often masked by perfectly supported statistical outcomes, whereas different phylogenetic reconstruction methods lead to mutually discordant topologies (Phillips et al. 2004). Errors in phylogenetic inference are mainly due to A) stochastic error, e.g. sampling bias due to insufficient polymorphism over short DNA sequences; B) violation of the orthology assumption, caused by gene duplications, horizontal gene transfer or lineage sorting (Jeffroy et al. 2006) and C) systematic error, given by non-historical processes which lead to non-phylogenetic signals that violate assumptions of the models of phylogenetic inference (Phillips et al. 2004). These non-phylogenetic biases can be interpreted as rate signal (unaccounted variable rates across lineages; Felsenstein 1978), compositional signal (heterogeneous nucleotide/amino acid compositions, which cause erroneous clustering of taxa that share the same bias; Lockhart et al. 1994) and heterotachy (shift of site-specific evolutionary rates over time; Philippe and Germot 2000;

Kolaczkowski and Thornton 2004). Large datasets certainly overcome sampling bias, and in most of the cases, non-orthologous biases can be overcome by the fact that only a few genes are affected and the erroneous signal is supposedly buffered by the genomic true signal. However, systematic error could appear regardless of the quantity of the data (Felsenstein 1978).

In this case, a non-orthologous bias caused by incomplete lineage sorting at the considered loci between *L. rubellus* populations could be the problem. The incomplete sorting may be due to the fact that the protein coding genes, from which this SNP dataset was drawn, are generally more conserved. Even over millions of years of divergence between lineages inferred from mitochondrial data, a high degree of conservation over a great proportion of protein-coding genes could be expected (Maddison, 1997), considering the selective constraints imposed by the ecological niche these animals occupy. However, the discrepancy could be due to systematic bias. Given that only the transcriptome phylogenetic signal was considered, the outcome could be due to the different evolutionary patterns of protein coding genes and gene families, with different selection and evolutionary rates among the considered sites. The methods used, ML and Bayesian inference, can arguably be preferred over other methods, as they explicitly incorporate the processes of sequence evolution in their models. More complex models should reduce the probability of inconsistency. However, given the complexity of phylogenomic data and the heterogeneity of evolutionary patterns, this might not hold in all cases. Therefore, phylogenomics needs more realistic models of sequence evolution, and research in this area is ongoing (Delsuc et al. 2005). These results also show that transcriptomic data (e.g. Hartmann et al. 2012; Zhao et al. 2013) in phylogenomics should be carefully considered and checked for errors, particularly when trying to assess phylogenetic distance in closely related taxa.

The BS and ML phylogenetic inferences obtained with the full genomic dataset, using 94,327 SNPs, and shown in the tree in Figure 4.5, show complete topological concordance and high statistical support. The trees in this case also confirm the deep divergent split between Lineage B and Lineage F and the other lineages. It also shows relationships between A individuals as already inferred by mitochondrial data. Discrepancies between this topology and the mitochondrial topologies are found in the independent clustering between Balkan lineages C and D, and the clustering of lineage C with the Hungarian A1 individual. Given the

geographical proximity of these individuals, it can be hypothesised that a hybridisation event between mitochondrial A1 individuals and refugial Balkan populations, which came into contact in the Carpathian area, is the cause of this particular configuration. Alternatively, this signature could be due to systematic error. A recent study indicates that GC-rich regions of the genome could lead to systematic bias, because recombination events drive GC-content evolution through biased gene conversion, causing problem in phylogenetic reconstructions, e.g. when there could be incomplete lineage sorting, due to some gene genealogies not being concordant with the species phylogeny (Romiguier et al. 2013). The *L. rubellus* genome AT percentage is estimated at 59.22%. This increased AT content could help in retrieving a correct phylogeny as far as the genome is concerned.

Finally, an unequivocal outcome of the analyses carried out in this chapter, is the confirmation of the divergence of lineage B from all the other lineages. In all the analyses, even the ones affected by systematic biases or incomplete lineage sorting, Lineage B clearly stands out as a genetically isolated entity. This study confirms at the genomic level what has been observed both at the nuclear and mitochondrial level by previous studies (King et al. 2008; Donnelly et al. In Press). Nevertheless, it may be possible that rare hybridisation events can happen between *L. rubellus* A and B. An old study reported hybridisation between *L. rubellus* and a closely related species, *L. friendi* (Evans and Guild 1948), an epigeic European worm, whose habitat overlaps with *L. rubellus* (Sims and Gerard 1999). A hybridisation signal between lineages A and B was also observed in Andre et al. (2009). However, it may be an artefact, as AFLPs can incur fragment size homoplasy, with consequent problems in reliably inferring patterns of genetic diversity (Caballero et al. 2008). The more recent evidence points to genetic segregation between lineages, a possible hint that these lineages are effectively reproductively isolated species.

A limitation of this study is related to the fact that, for reasons of time and computational limitations, the full potential of the genomes data produced could not be explored in its entirety. In particular, possible gene-specific analyses were not carried out. Further investigations on these genomes could lead to discover possible different functional pathways and key genes/gene families involved in responses to environmental variables, such as climate and toxicity. Functional studies have identified conserved orthologous genes involved in drug response in species of ecotoxicological importance: a high degree of conservation in these regions may

imply a selected response to a range of pollutants (Gunnarsson et al. 2008; Celander et al. 2011). In the light of these findings, a possible future line of research would be to explore these orthologs and the implications of their differentiation in response to pollutants in the genomes of the *L. rubellus* lineages.

Further studies should focus on the extent of this divergence related to ecological and functional differences, to fully understand the differential response these clades may have to environmental changes, and better assess their role and usefulness in ecotoxicological assays.

4.5 Bibliography

Andre J, King RA, Stürzenbaum SR, Kille P, Hodson M, Morgan AJ. 2009. Molecular genetic differentiation in earthworms inhabiting a heterogeneous Pb-polluted landscape. *Environmental Pollution* **158**: 883-890.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**: 2633-2635.

Bray N, Dubchak I, Pachter L. 2003. AVID: a global alignment program. *Genome Research* **13**: 97–102.

Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S. 2003. Glocal alignment: finding rearrangements during alignment. *Bioinformatics* **19**: 154-162.

Caballero A, Quesada H, Rolán-Alvarez E. 2008. Impact of amplified fragment length polymorphism size homoplasy on the estimation of population genetic diversity and the detection of selective loci. *Genetics* **179**: 539-554.

Celander MC, Goldstone JV, Denslow ND, Iguchi T, Kille P, Meyerhoff RD, Smith BA, Hutchinson TH, Wheeler JR. 2011. Species extrapolation for the 21st century. *Environmental Toxicology and Chemistry* **30**: 52-63.

Chain PSG, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C. 2009. Genome project standards in a new era of sequencing. *Science* **326**: 236-237.

Cho YS, Hu L, Hou H, Lee H, Xu J, Kwon S, Oh S, Kim H-M, Jho S, Kim S, Shin Y-A, Kim BC, Kim H, Kim C-u, Luo S-J, Johnson WE, Koepfli K-P, Schmidt-Küntzel A, Turner JA, Marker L, Harper C, Miller SM, Jacobs W, Bertola LD, Kim TH, Lee S, Zhou Q, Jung H-J, Xu X, Gadhvi P, Xu P, Xiong Y, Luo Y, Pan S, Gou C, Chu X, Zhang J, Liu S, He J, Chen Y, Yang L, Yang Y, He J, Liu S, Wang J, Kim CH, Kwak H, Kim J-S, Hwang S, Ko J, Kim C-B, Kim S, Bayarlkhagva D, Paek WK, Kim S-J, O'Brien SJ, Wang J, Bhak J. 2013. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nature Communications* **4**.

Darwin C. 1838. On the formation of mould. *Proceedings of Geological Society* **2**: 574 - 576.

Darwin C. 1892. *The Formation of Vegetable Mould, Through the Action of Worms, with Observations on Their Habits*. J. Murray.

Delcher AL, Kasif S, Fleischman R, Peterson J, White O, Salzberg SL 1999. Alignment of whole genomes. *Nucleic Acids Research* **27**: 2369–2376.

Delcher AL, Phillippy A, Carlton J, Salzberg SL 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* **30**: 2478–2483.

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**: 3207-3212.

Delsuc Fdr, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* **6**: 361-375.

Donnelly RK, Harper GL, Morgan AJ, Orozco-terWengel P, Juma GAP, Bruford MW. In Press. Recapitulation of cryptic lineages of *Lumbricus rubellus*. *Biological Journal of the Linnean Society*.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* **63**: 1-19.

Eisen JA, Fraser CM. 2003. Phylogenomics: intersection of evolution and genomics. *Science* **300**: 1706-1707.

Elsworth B. 2012. Unearthing the genome of the earthworm *Lumbricus rubellus*. The University of Edinburgh, Edinburgh.

Evans AC, Guild WJ. 1948. Studies on the relationships between earthworms and soil fertility. IV. On the life cycles of some British Lumbricidae. *Annals of Applied Biology* **35**: 471-484.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22**: 521-565.

Field D, Tiwari B, Booth T, Houten S, Swan D, Bertrand N, Thurston M. 2006. Open software for biologists: from famine to feast. *Nature Biotechnology* **24**: 801-804.

Freedman AH, Schweizer RM, Gronau I, Han E, Vecchyo DO-D, Silva PM, Galaverni M, Fan Z, Marx P, Lorente-Galdos B. 2013. Genome sequencing highlights genes under selection and the dynamic early history of dogs. *arXiv preprint arXiv:1305.7390*.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**: 307-321.

Gunnarsson L, Jauhiainen A, Kristiansson E, Nerman O, Larsson DGJ. 2008. Evolutionary conservation of human drug targets in organisms used for environmental risk assessments. *Environmental Science & Technology* **42**: 5807-5813.

Hartmann S, Helm C, Nickel B, Meyer M, Struck TH, Tiedemann R, Selbig J, Bleidorn C. 2012. Exploiting gene families for phylogenomic analysis of myzostomid transcriptome data. *PloS One* **7**: e29843.

Ibarra-Laclette E, Lyons E, Hernandez-Guzman G, Perez-Torres CA, Carretero-Paulet L, Chang T-H, Lan T, Welch AJ, Juarez MJA, Simpson J, Fernandez-Cortes A, Arteaga-Vazquez M, Gongora-Castillo E, Acevedo-Hernandez G, Schuster SC, Himmelbauer H, Minoche AE, Xu S, Lynch M, Oropeza-Aburto A, Cervantes-Perez SA, de Jesus Ortega-Estrada M, Cervantes-Luevano JI, Michael TP, Mockler T, Bryant D, Herrera-Estrella A, Albert VA, Herrera-Estrella L. 2013. Architecture and evolution of a minute plant genome. *Nature* **498**: 94-98.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *TRENDS in Genetics* **22**: 225-231.

Jones CG, Lawton JH, Shachak M. 1997. Positive and negative effects of organisms as physical ecosystem engineers. *Ecology* **78**: 1946-1957.

King RA, Tibble AL, Symondson WOC. 2008. Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Molecular Ecology* **17**: 4684-4698.

Knowles LL. 2009. Statistical phylogeography. *Annual Review of Ecology, Evolution, and Systematics* **40**: 593-612.

Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PloS One* **6**: e15925.

Kolaczkowski B, Thornton JW. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**: 980-984.

Lavelle P, Bignell D, Lepage M, Wolters W, Roger P, Ineson P, Heal OW, Dhillion S. 1997. Soil function in a changing world: the role of invertebrate ecosystem engineers. *European Journal Of Soil Biology* **33**: 159-193.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754 - 1760.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493-496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution* **11**: 605-612.

Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* **46**: 523-536.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* **66**: 526-538.

Morgenstern B. 1999. DIALIGN 2: improvement of the segment–to–segment approach to multiple sequence alignment. *Bioinformatics* **15**: 211–218.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**: 443-453.

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, Johnson PLF, Fumagalli M, Vilstrup JT, Raghavan M, Korneliussen T, Malaspinas A-S, Vogt J, Szklarczyk D, Kelstrup CD, Vinther J, Dolocan A, Stenderup J, Velazquez AMV, Cahill J, Rasmussen M, Wang X, Min J, Zazula GD, Seguin-Orlando A, Mortensen C, Magnussen K, Thompson JF, Weinstock J, Gregersen K, Roed KH, Eisenmann V, Rubin CJ, Miller DC, Antczak DF, Bertelsen MF, Brunak S, Al-Rasheid KAS, Ryder O, Andersson L, Mundy J, Krogh A, Gilbert MTP, Kjaer K, Sicheritz-Ponten T, Jensen LJ, Olsen JV, Hofreiter M, Nielsen R,

Shapiro B, Wang J, Willerslev E. 2013. Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**: 74-78.

Philippe H, Germot A. 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Molecular Biology and Evolution* **17**: 830-834.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution* **21**: 1455-1458.

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G. 2013. Great ape genetic diversity and population history. *Nature* **499**: 471-475.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**: 559-575.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Molecular Biology and Evolution* **30**: 2134-2144.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard Ma, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**: 539-542.

Sims R, Gerard B. 1999. *Earthworms*. Linnean Society, London.

Wan Q-H, Pan S-K, Hu L, Zhu Y, Xu P-W, Xia J-Q, Chen H, He G-Y, He J, Ni X-W, Hou H-L, Liao S-G, Yang H-Q, Chen Y, Gao S-K, Ge Y-F, Cao C-C, Li P-F, Fang L-M, Liao L, Zhang S, Wang M-Z, Dong W, Fang S-G. 2013. Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator. *Cell Research* **23**: 1091-1105.

Zhan X, Pan S, Wang J, Dixon A, He J, Muller MG, Ni P, Hu L, Liu Y, Hou H. 2013. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nature Genetics* **45**: 563-566.

Zhao S, Shalchian-Tabrizi K, Klaveness D. 2013. Sulcozoa revealed as a paraphyletic group in mitochondrial phylogenomics. *Molecular Phylogenetics and Evolution*.

CHAPTER 5 GENERAL DISCUSSION

This dissertation has examined the patterns of genetic cryptic differentiation in a model earthworm species at a variety of genetic scales – from the single gene to the genome. The hypothesis that *L. rubellus* is actually a cryptic species complex, with deep mitochondrial divergent lineages scattered over its distributional range, was examined using genetic and genomic data. I tested whether the *L. rubellus* diversity pattern previously observed was consistent over a larger scale, which environmental variables could be inferred to be driving this pattern and I explored the temporal and geographical dimension of this variation, clearly influenced by the quaternary glaciations. Finally, I selected a subsample of eight specimens for Next Generation Sequencing, to test whether the differentiation pattern observed at the mitochondrial level could be observed at the whole genome level, also investigating how phylogenetic signals changed within different parts of the mitochondrial genome. Combining species distribution modelling and mitochondrial data, I provided the first study investigating an earthworm cryptic species complex at a continental scale using a statistical phylogeography framework. This was also the first time that the genetic diversity of a cryptic sentinel species was investigated at the genomic level.

5.1 Overview of main results

5.1.1 Overview of chapter 2: Paleoclimatic impact on cryptic diversity of the earthworm Lumbricus rubellus *in Europe*

In this chapter, the spatial and temporal patterns of genetic diversity of *L. rubellus* lineages were studied in a statistical phylogeographic framework. A broad survey using mitochondrial markers commonly used in studies to detect cryptic species in earthworms (King et al. 2008; Andre et al. 2009; James et al. 2010; Fernández et al. 2012; Klarica et al. 2012; Decaëns et al. 2013) was used. We tested the hypothesis that genetic diversity in northern refugia is a subsample of the genetic diversity of southern glacial refugia. This hypothesis was rejected, as eleven cryptic lineages were found, with a phylogeographic structure suggesting a much more complex scenario. Some lineages did show signatures of probable survival and differentiation in southern glacial refugia, but other lineages could not be directly linked with these areas and were probably descendants of demes that survived in

cryptic northern or intermediate latitude refugia. The finding that *L. rubellus* is actually a cryptic species complex, comprising different lineages over its geographic range, was in agreement with what was already observed for the species (King et al. 2008; Andre et al. 2009; Donnelly et al. In Press) and another confirmation of the high frequency of cryptic divergence processes in earthworms (Sturmbauer et al. 1999; Pérez-Losada et al. 2005; Erséus and Gustafsson 2009; Gustafsson et al. 2009; Novo et al. 2009; Pérez-Losada et al. 2009; James et al. 2010; Novo et al. 2010; Buckley et al. 2011; Novo et al. 2011; Fernández et al. 2012; Cunha et al. 2013; Shekhovtsov et al. 2013). It must be noted that the observed diversity of *L. rubellus* across Europe could itself be a subsample of the real picture. A wider, more systematic sampling is needed in order to have a better knowledge of this genetic diversity over its natural range, the Holartic Europe, and the newly colonized areas (North-Eastern and Eastern Europe and the Americas, (Sims and Gerard 1999; Tiunov et al. 2006)). In particular, some areas that could not be covered in this study or where the number of samples collected was small, for reasons depending on time, budget or difficulty to find specimens (most remarkably, the Italian peninsula), could harbour interesting signatures of past demographic history and survival both in Mediterranean and in northern refugia.

The hypothesis regarding divergences within and among lineages A and B from a common ancestor in the Pleistocene, was only partially accepted. Signatures of past demographic change and divergence time estimates, integrated with SDM, depicted a scenario of late Miocene colonisation of the continent mainly driven by climatic factors, followed by vicariant events and isolation, most likely caused by the impact of glaciations in the Quaternary period. Hypotheses on divergence were tested in a statistical phylogeographic framework, highlighting that the current genetic diversity scenario is likely the result of two main divergence events, occurring A) during the Quaternary and B) during the late Miocene. While estimates of the time of divergence were in agreement with genetic studies regarding the well-known patterns of differentiation in Quaternary refugia occurring during the Pleistocene glaciations (Hewitt 2000; Hewitt 2004; Stewart et al. 2010), some much older signatures of divergence more likely link to the "washhouse climate" periods at the end of the Miocene (Böhme et al. 2008) and the main divergence event of Lineage B from Lineage A (and the other lineages) happened long before the Pleistocene. In addition, the "Washhouse" phases may have played an important part in the continental

colonisation of *L. rubellus*' common ancestor. Tests on this hypothesis may shed a new light on studies concerning the pre-Quaternary distribution of many taxa, in particular those dependent to mesic environments, such as earthworms. The outcome of SDM pointed out that some lineages could have survived in multiple cryptic glacial refugia, including central Europe, the Carpathian basin and southern Ireland, consistently with an increasing body of evidence of suitable conditions in territories previously considered inhospitable for temperate biota (Kotík et al. 2006; Vega et al. 2010; Parducci et al. 2012; Finnegan et al. 2013).

The hypothesis regarding climatic variables being the most important factors limiting the distribution of *L. rubellus*, rather than soil variables was accepted. The SD modelling showed a strong correlation between certain climatic conditions and the presence of *L. rubellus*, as opposed to soil characteristics, which minimally contributed in the outcome of the model, and this was found both in the test and in the training models. SDM appears to have captured a number of relevant features for the requirements of *L. rubellus*: modelling sedentary species allows a higher probability to identify the realised niche of the species, as opposed to the potential niche. This increases the likelihood to capture the effective distribution of the species (Phillips et al. 2006). Furthermore, the fact that such evolutionary distinct clades have a conserved epigeic niche supports the SDM assumption of niche stability through time (Phillips et al. 2006; Elith et al. 2011). A possible limitation of the models built in this study concerned the soil variables chosen. Although climatic effects on epigeic worms remain the best candidates in shaping their distributions, further testing with an increased number of soil variables (e.g. heavy metal content, organic and inorganic pollutants, land use etc.) which could not be implemented in the present modelling because of data limitations and permission issues may be needed in order to build a more complete assessment of soil variable effects, given their demonstrated sensibility to chemical factors and soil pollutants (e.g. Spurgeon et al. 2003; Spurgeon et al. 2004; Bundy et al. 2008; Svendsen et al. 2008).

5.1.2 *Overview of Chapter 3: Mitogenomics of cryptic diversity: exploring the deep phylogenetic signal of* Lumbricus rubellus

In this chapter, eight mitochondrial genomes isolated from NGS data obtained from individuals representative of *L. rubellus* genetic diversity and distribution, were chosen in order to describe the structure of the genome and investigate the cryptic diversity signal at the mitogenomic level. The study gave interesting insights on the mitochondrial genome structure. Although a high degree of similarity between *L. rubellus* and *L. terrestris* mitochondrial genomes was found, an interesting difference was highlighted, corresponding to a variable non-coding region situated between the ND6 and CYTB genes. Further studies should investigate the intriguing possibility that this region could include a type II intron, as the existence of such a structure in a bilateran genome has already been documented (Vallès et al. 2008).

The hypothesis that single gene trees could reproduce a phylogeny obtained using the concatenated sequence of all the genes was rejected. Although the majority of genes gave phylogenies quite similar to the whole mitochondrial phylogenetic signal, some topologies varied, hinting at different evolutionary pressures acting on different parts of the mitochondrial genome. These results support evidence against selective neutrality of the mitochondrial genome (Popadin et al. 2013). The hypothesis of correlation between phylogenetic and certain statistical parameters was also rejected. No correlation was found between statistical parameters calculated from the dataset and phylogenetic signals. Probably, the technique I used was not refined enough to detect correlations; an alternative approach, accounting for standardisation of parameters, may be better in detecting some relationships (Legendre and Lapointe 2004). There is also the possibility that non-phylogenetic signals masked such correlations, most likely within-site rate variation (Phillips et al. 2004).

5.1.3 *Overview of chapter 4: Phylogenomic analysis of* Lumbricus rubellus *cryptic clades*

In this chapter, NGS data from the eight whole genomes of the same individuals used for Chapter 3 were investigated, in order to explore the phylogenetic relationships between selected individuals at the nuclear level. The aim was to compare phylogenetic signals between mitochondrial and nuclear data, so as to

investigate the real extent of the diversification process between lineages. The analyses consisted, in the first instance, of a methodological comparison between global and local alignment methods to map the reads to two reference datasets, a transcriptome and a genome of *L. rubellus* lineage B. While local alignment was found to be more efficient, there is also a higher likelihood of incorrectly mapped reads using only local alignment approaches (Degner et al. 2009; Kofler et al. 2011). A mixed approach combining local and global alignment mapping algorithms could potentially maximise mapping efficiency at the same time as reducing mapping errors (Brudno et al. 2003; Kofler et al. 2011).

To test the hypothesis that nuclear and demographic signals present similar patterns to those observed in mitochondrial data, SNPs were extracted from both datasets and a comparison of phylogenies was carried out. The hypothesis was partially rejected, as the outcomes of the analysis indicate that the phylogenetic information contained in the genome is only consistent with the mitochondrial phylogeny in certain features. Firstly, the phylogenies constructed with SNP data from the transcriptome led to contrasting results in term of topology, but with high support values. This is likely a limitation due to the fact that SNP data from transcriptomes is not appropriate for building reliable phylogenies, because of likely violations of orthology assumptions, such as incomplete lineage sorting (Jeffroy et al. 2006) or systematic errors due to non-phylogenetic signals (Phillips et al. 2004) which can appear regardless of the quantity of the data (Felsenstein 1978). The phylogenies constructed using genomic SNPs showed congruency both in topology and in branch support, but these phylogenies, while congruent with mitochondrial data regarding the separation between the A, B and F lineages, seem to indicate a possible hybridisation event has occurred between the Serbian lineage C and A1. It is possible that populations from A1 and C came into contact during previous postglacial expansions in the Carpathian area, following the common process of hybrid zone formation after post-glacial recolonisation (Hewitt 2001). Remarkably, lineage A and B live in sympatry, but the outcome of this study, in agreement with previous microsatellite evidence (Donnelly et al. In Press) inferred reproductive isolation between the two, supporting the hypothesis that these clades are effectively cryptic species. The only contrasting result (Andre et al. 2009) inferring a hybridisation event between clades is now likely to have been an artefact due to size homoplasy in AFLPs (Caballero et al. 2008). Unfortunately, it was not possible to test the hypothesis of whether genomic

demography was consistent with mitochondrial data. The genome of *L. rubellus*, although in draft form, was too fragmented to allow the PSMC algorithm (Li and Durbin 2011) to calculate past demographic changes with an approximation of the coalescent from recombination events.

The key limitation of this study is that I could not properly exploit the full potential of the genomes obtained during this study, in terms of identifying genes involved in important adaptive pathways and investigating their evolutionary signals. Although fragmented, the *L. rubellus* draft genome is in appropriate form to assess gene content (Chain et al. 2009; Elsworth 2012) and could be used as a reference to find homologous regions and investigate important functional differences in key genes or gene families involved in adaptation to e.g. epigeic niche, response to pollutants and local adaptation to environmental variables (e.g. Qiu et al. 2012; Zhan et al. 2013).

5.2 Is *L. rubellus* really a cryptic species complex?

This is the first study where a multidisciplinary approach involving species distribution modelling, barcoding markers and genome-wide differentiation has been used to explore cryptic variation and evolution in earthworms. Despite some incongruences, mainly concerning the more closely related lineages, the phylogenetic relationships inferred both at the genomic and mitogenomic level are convergent in the definition of at least three entities (The A-C-D lineages cluster, lineage F from Spain and lineage B from UK) which clearly carry genomic signatures of ancient segregation, genetic drift and divergence. The fact that such a high phylogenetic resolution strongly supports genetic segregation between these entities, indicates that these clades are different species, descending from ancestral demes which became separate because of climatic vicariant events during the last 6 millions years. Low dispersal capacities (Marinissen and Van den Bosch 1992), genetic drift and isolation by distance before and during the Pleistocene (Hewitt 2000) may be the causes of such variation. This evidence is supported by their patterns of geographical distribution, as well as the estimated niche during the last 135,000 years evidenced by species distribution modelling. In addition, the particular patterns of distributions of lineages A, B, the German lineages G and H and perhaps the Balkan lineages C and

D, suggest the existence of different cryptic refugia where they could have survived in isolation and where they started recolonizing once the climatic conditions became warmer (Stewart and Lister 2001). In particular, niche models showed the existence of a vast and continuous area of habitat suitability corresponding to a region of tundra including parts of southern Ireland, Great Britain and the seafloor of the Celtic Sea, extending southwards over the Atlantic coast of France and Iberia until the western Galician coastline. This area was previously identified in niche modelling studies (Vega et al. 2010) and could link evidence of Irish (Stewart and Lister 2001) and north-western French northern refugia (Finnegan et al. 2013) with the observed presence of haplotypes shared between Iberian regions and Ireland, a phenomenon called "Lusitanian pattern" (Corbet 1961; Searle 2008). Further studies should test this hypothesis, as such a region could have an important impact in the understanding of gene flow, migration and diversification processes during the Pleistocene. Nevertheless, a human contribution to their distribution and its implication in the Lusitanian pattern in many species cannot be ruled out, and it is the most probable in some cases (McDevitt et al. 2011). What emerges from our study and from the literature is a very tangled scenario where human mediated dispersal, survival and divergence in refugia and in microclimatic patches of habitat suitability, in addition to postglacial recolonisation patterns, could all contribute to explaining past distribution and divergence processes. Most likely, the signature of diversification processes observed in *L. rubellus* during its Miocene and Pleistocene history had an impact on other peregrine lumbricid species. Cryptic speciation in peregrine lumbricid species is now a recognised phenomenon; I argue, in the light of my results, that future investigations on the origins of cryptic divergence in holartic earthworm taxa cannot disregard SDM frameworks as essential tools to explain their past distributions.

5.3 Implications for *L. rubellus* as a sentinel species

This study showed that, in *L. rubellus*, some commonly used barcoding genes had the capacity to give a good estimation of the deep phylogenetic relationships between lineages. These results confirm the power of barcoding markers in assessing phylogenetic relationships between closely related cryptic clades (Decaëns et al. 2013). In addition, the concordance between the genomic data and mitochondrial

phylogenetic signals for the most distant clades implies that the high degrees of mitochondrial divergence reflect real speciation processes, with a concurrently increased likelihood of differential genetic responses to pollutants in ecotoxicological assays (Simonsen et al. 2004; Erséus and Gustafsson 2009). In the light of the power of these markers in detecting cryptic variation, and the need to have genetically homogeneous animals in ecotoxicological essays (Erséus and Gustafsson 2009), it is of primary importance to implement COI barcoding in the standardisation of every ecotoxicological study using sentinel earthworm species. However, what has been found for *L. rubellus* in terms of correlation between mitochondrial and nuclear divergence degrees might not hold for all sentinel species and, furthermore, analyses could be complicated by the retention of paralogous mitochondrial copies in the nuclear genome (numts; Funk and Omland 2003). Therefore, future assessments may include nuclear markers, such as in some recent phylogeographic studies on earthworm cryptic speciation (Novo et al. 2010; Fernández et al. 2012) or could be conducted in a phylogenomics framework, using NGS data to infer the real depth of phylogenetic relationships and also acquiring data that can give insights on gene function and adaptation (Gunnarsson et al. 2008).

5.4 Future perspectives

The *L. rubellus* genome project is at an advanced stage. Additional data in the form of multiple insert mate-pair libraries or 3rd generation sequencing long reads (e.g. Pacific Biosciences) is likely to solve the issues regarding the assembly of the complete genome from a series of short reads (Elsworth 2012). Once complete, inference of recombination events and past demography using the genomes produced will be possible. This will give an unprecedented insight into the complex scenario of a cryptic species' complex history. Probably, access to these data will lead to more insights regarding the species' divergence and speciation in the light of the different histories obtained from mitochondrial and nuclear genes (Li and Durbin 2011), and the evolutionary pressure exerted on these animals by the climatic shifts of the Pleistocene.

A more immediate task, as it is already feasible with the present draft genome, may concern the assessment of orthologous genes, so as to undertake studies on

functional genes and gene families, and the functional differences these genes may have evolved. An interesting line of research could be related to the presence of conserved gene targets in wild species with high affinity interactions to a wide range of chemical substances (Gunnarsson et al. 2008). The presence of such evolutionary targets in a given species is associated with an increased risk of environmental contamination, and it has been recommended that wildlife taxa in mesic/aquatic environments should be used in risk assessments for environmental pollution, extrapolating effects from a series of test species to all the species of that group in the environment. This process, labelled "species extrapolation" (Celander et al. 2011) could be applied to *L. rubellus* genome data in order to evaluate the differential presence of such target in the different genomes, which could lead to the identification of an "optimal" *L. rubellus* clade to be selected for environmental risk assessment.

5.5 Bibliography

Andre J, King RA, Stürzenbaum SR, Kille P, Hodson M, Morgan AJ. 2009. Molecular genetic differentiation in earthworms inhabiting a heterogeneous Pb-polluted landscape. *Environmental Pollution* **158**: 883-890.

Böhme M, Ilg A, Winklhofer M. 2008. Late Miocene "washhouse" climate in Europe. *Earth and Planetary Science Letters* **275**: 393-401.

Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S. 2003. Glocal alignment: finding rearrangements during alignment. *Bioinformatics* **19**: i54-i62.

Buckley TR, James S, Allwood J, Bartlam S, Howitt R, Prada D. 2011. Phylogenetic analysis of New Zealand earthworms (Oligochaeta: Megascolecidae) reveals ancient clades and cryptic taxonomic diversity. *Molecular Phylogenetics and Evolution* **58**: 85-96.

Bundy JG, Sidhu JK, Rana F, Spurgeon DJ, Svendsen C, Wren JF, Stürzenbaum SR, Morgan AJ, Kille P. 2008. 'Systems toxicology' approach identifies coordinated metabolic responses to copper in a terrestrial non-model invertebrate, the earthworm *Lumbricus rubellus*. *BMC Biology* **6**: 25.

Caballero A, Quesada H, Rolán-Alvarez E. 2008. Impact of amplified fragment length polymorphism size homoplasy on the estimation of population genetic diversity and the detection of selective loci. *Genetics* **179**: 539-554.

Celander MC, Goldstone JV, Denslow ND, Iguchi T, Kille P, Meyerhoff RD, Smith BA, Hutchinson TH, Wheeler JR. 2011. Species extrapolation for the 21st century. *Environmental Toxicology and Chemistry* **30**: 52-63.

Chain PSG, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, Ali J, Birren B, Bruce DC, Buhay C. 2009. Genome project standards in a new era of sequencing. *Science* **326**: 236-237.

Coles BJ. 1999. *Doggerland's loss and the Neolithic*. WARP Occasional Paper Exeter.

Corbet GB. 1961. Origin of the british insular races of small mammals and of the Lusitanian Fauna. *Nature* **191**: 1037-1040.

Cunha L, Montiel R, Novo M, Orozco-terWengel P, Rodrigues A, Morgan AJ, Kille P. 2013. Living on a volcano's edge: genetic isolation of an extremophile terrestrial metazoan. *Heredity*.

Decaëns T, Porco D, Rougerie R, Brown GG, James SW. 2013. Potential of DNA barcoding for earthworm research in taxonomy and ecology. *Applied Soil Ecology* **65**: 35-42.

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**: 3207-3212.

Donnelly RK, Harper GL, Morgan AJ, Orozco-terWengel P, Juma GAP, Bruford MW. 2013. Nuclear DNA recapitulates the cryptic mitochondrial lineages of *Lumbricus rubellus* and suggests the existence of cryptic species in an ecotoxological soil sentinel. *Biological Journal of the Linnean Society* **4**: 780-795.

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* **17**: 43-57.

Elsworth B. 2012. Unearthing the genome of the earthworm *Lumbricus rubellus*. The University of Edinburgh, Edinburgh.

Erséus C, Gustafsson D. 2009. Cryptic speciation in clitellate model organisms. *Annelids in Modern Biology*: 31-46.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

Fernández R, Almodóvar A, Novo M, Simancas B, J. Díaz Cosín D. 2012. Adding complexity to the complex: New insights into the phylogeny, diversification and origin of parthenogenesis in the *Aporrectodea caliginosa* species complex (Oligochaeta, Lumbricidae). *Molecular Phylogenetics and Evolution* **64**: 368-379.

Finnegan AK, Griffiths AM, King RA, Machado-Schiaffino G, Porcher JP, Garcia-Vazquez E, Bright D, Stevens JR. 2013. Use of multiple markers demonstrates a cryptic western refugium and postglacial colonisation routes of Atlantic salmon (*Salmo salar* L.) in northwest Europe. *Heredity* **111**: 34-43.

Funk DJ, Omland KE. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics* **34**: 397-423.

Gunnarsson L, Jauhiainen A, Kristiansson E, Nerman O, Larsson DGJ. 2008. Evolutionary conservation of human drug targets in organisms used for

environmental risk assessments. *Environmental science & technology* **42**: 5807-5813.

Gustafsson DR, Price DA, Erséus C. 2009. Genetic variation in the popular lab worm *Lumbriculus variegatus* (Annelida: Clitellata: Lumbriculidae) reveals cryptic speciation. *Molecular Phylogenetics and Evolution* **51**: 182-189.

Hewitt GM. 2000. The genetic legacy of the Quaternary ice ages. *Nature* **405**: 907-913.

Hewitt GM. 2001. Speciation, hybrid zones and phylogeography — or seeing genes in space and time. *Molecular Ecology* **10**: 537-549.

Hewitt GM. 2004. Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* **359**: 183-195.

James SW, Porco D, Decaëns T, Richard B, Rougerie R, Erséus C. 2010. DNA barcoding reveals cryptic diversity in *Lumbricus terrestris* L., 1758 (Clitellata): resurrection of *L. herculeus* (Savigny, 1826). *PloS One* **5**: e15629.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *TRENDS in Genetics* **22**: 225-231.

King RA, Tibble AL, Symondson WOC. 2008. Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Molecular Ecology* **17**: 4684-4698.

Klarica J, Kloss-Brandstätter A, Traugott M, Juen A. 2012. Comparing four mitochondrial genes in earthworms – Implications for identification, phylogenetics, and discovery of cryptic species. *Soil Biology and Biochemistry* **45**: 23-30.

Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. 2011. PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PloS One* **6**: e15925.

Kotík P, Deffontaine V, Mascheretti S, Zima J, Michaux JR, Searle JB. 2006. A northern glacial refugium for bank voles (*Clethrionomys glareolus*). *Proceedings of the National Academy of Sciences* **103**: 14860-14864.

Legendre P, Lapointe F-J. 2004. Assessing congruence among distance matrices: single-malt scotch whiskies revisited. *Australian & New Zealand Journal of Statistics* **46**: 615-629.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493-496.

Marinissen JCY, Van den Bosch F. 1992. Colonization of new habitats by earthworms. *Oecologia* **91**: 371-376.

McDevitt AD, Vega R, Rambau RV, Yannic G, Herman JS, Hayden TJ, Searle JB. 2011. Colonization of Ireland: revisiting 'the pygmy shrew syndrome' using mitochondrial, Y chromosomal and microsatellite markers. *Heredity* **107**: 548-557.

Novo M, Almodóvar A, Díaz-Cosín DJ. 2009. High genetic divergence of hormogastrid earthworms (Annelida, Oligochaeta) in the central Iberian Peninsula: evolutionary and demographic implications. *Zoologica Scripta* **38**: 537-552.

Novo M, Almodóvar A, Fernández R, Giribet G, Díaz Cosín DJ. 2011. Understanding the biogeography of a group of earthworms in the Mediterranean basin--the phylogenetic puzzle of Hormogastridae (Clitellata: Oligochaeta). *Molecular Phylogenetics and Evolution* **61**: 125-135.

Novo M, Almodóvar A, Fernández R, Trigo D, Díaz Cosín DJ. 2010. Cryptic speciation of hormogastrid earthworms revealed by mitochondrial and nuclear data. *Molecular Phylogenetics and Evolution* **56**: 507-512.

Parducci L, Jørgensen T, Tollefsrud MM, Elverland E, Alm T, Fontana SL, Bennett KD, Haile J, Matetovici I, Suyama Y, Edwards ME, Andersen K, Rasmussen M, Boessenkool S, Coissac E, Brochmann C, Taberlet P, Houmark-Nielsen M, Larsen NK, Orlando L, Gilbert MTP, Kjær KH, Alsos IG, Willerslev E. 2012. Glacial Survival of Boreal Trees in Northern Scandinavia. *Science* **335**: 1083-1086.

Pérez-Losada M, Eiroa J, Mato S, Domínguez J. 2005. Phylogenetic species delimitation of the earthworms *Eisenia fetida* (Savigny, 1826) and *Eisenia andrei* Bouché, 1972 (Oligochaeta, Lumbricidae) based on mitochondrial and nuclear DNA sequences. *Pedobiologia* **49**: 317-324.

Pérez-Losada M, Ricoy M, Marshall JC, Domínguez J. 2009. Phylogenetic assessment of the earthworm *Aporrectodea caliginosa* species complex (Oligochaeta: Lumbricidae) based on mitochondrial and nuclear DNA sequences. *Molecular Phylogenetics and Evolution* **52**: 293-302.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-Scale Phylogeny and the Detection of Systematic Biases. *Molecular Biology and Evolution* **21**: 1455-1458.

Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**: 231-259.

Popadin KY, Nikolaev SI, Junier T, Baranova M, Antonarakis SE. 2013. Purifying selection in mammalian mitochondrial protein-coding genes is highly effective and congruent with evolution of nuclear genes. *Molecular Biology and Evolution* **30**: 347-355.

Qiu Q, Zhang G, Ma T, Qian W, Wang J, Ye Z, Cao C, Hu Q, Kim J, Larkin DM, Auvil L, Capitanu B, Ma J, Lewin HA, Qian X, Lang Y, Zhou R, Wang L, Wang K, Xia J, Liao S, Pan S, Lu X, Hou H, Wang Y, Zang X, Yin Y, Ma H, Zhang J, Wang Z, Zhang Y, Zhang D, Yonezawa T, Hasegawa M, Zhong Y, Liu W, Zhang Y, Huang Z, Zhang S, Long R, Yang H, Wang J, Lenstra JA, Cooper DN, Wu Y, Wang J, Shi P, Liu J. 2012. The yak genome and adaptation to life at high altitude. *Nature Genetics* **44**: 946-949.

Searle JB. 2008. The colonization of Ireland by mammals. *The Irish Naturalists' Journal* **29**: 109-115.

Shekhovtsov SV, Golovanova EV, Peltek SE. 2013. Cryptic diversity within the Nordenskiold's earthworm, *Eisenia nordenskioldi* subsp. *nordenskioldi* (Lumbricidae, Annelida). *European Journal Of Soil Biology* **58**: 13-18.

Simonsen V, Holmstrup M, Niklasson M. 2004. Genetic differentiation of the parthenogenetic soil collembolan *Isotoma notabilis* along a copper gradient based on random amplified polymorphic DNA. *Pedobiologia* **48**: 297-303.

Sims R, Gerard B. 1999. *Earthworms*. Linnean Society, London.

Spurgeon DJ, St¸rzenbaum SR, Svendsen C, Hankard PK, Morgan AJ, Weeks JM, Kille P. 2004. Toxicological, cellular and gene expression responses in earthworms exposed to copper and cadmium. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology* **138**: 11-21.

Spurgeon DJ, Weeks JM, Van Gestel CAM. 2003. A summary of eleven years progress in earthworm ecotoxicology: The 7th international symposium on earthworm ecology · Cardiff · Wales · 2002. *Pedobiologia* **47**: 588-606.

Stewart JR, Lister AM. 2001. Cryptic northern refugia and the origins of the modern biota. *Trends in ecology & evolution (Personal edition)* **16**: 608-613.

Stewart JR, Lister AM, Barnes I, Dalén L. 2010. Refugia revisited: Individualistic responses of species in space and time. *Proceedings of the Royal Society B: Biological Sciences* **277**: 661-671.

Sturmbauer C, Opadiya GB, Niederstatter H, Riedmann A, Dallinger R. 1999. Mitochondrial DNA reveals cryptic oligochaete species differing in cadmium resistance. *Molecular Biology and Evolution* **16**: 967-974.

Svendsen C, Owen J, Kille P, Wren J, Jonker MJ, Headley BA, Morgan AJ, Blaxter M, Stürzenbaum SR, Hankard PK, Lister LJ, Spurgeon DJ. 2008. Comparative transcriptomic responses to chronic cadmium, fluoranthene, and atrazine exposure in *Lumbricus rubellus*. *Environmental Science and Technology* **42**: 4208-4214.

Tiunov AV, Hale CM, Holdsworth AR, Vsevolodova-Perel TS. 2006. Invasion patterns of Lumbricidae into the previously earthworm-free areas of northeastern Europe and the western Great Lakes region of North America. in *Biological Invasions Belowground: Earthworms as Invasive Species*, pp. 23-34. Springer.

Vallès Y, Halanych KM, Boore JL. 2008. Group II introns break new boundaries: presence in a bilaterian's genome. *PloS One* **3**: e1488.

Vega R, Fløjgaard C, Lira-Noriega A, Nakazawa Y, Svenning J-C, Searle JB. 2010. Northern glacial refugia for the pygmy shrew *Sorex minutus* in Europe revealed by phylogeographic analyses and species distribution modelling. *Ecography* **33**: 260-271.

Zhan X, Pan S, Wang J, Dixon A, He J, Muller MG, Ni P, Hu L, Liu Y, Hou H. 2013. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nature Genetics* **45**: 563-566.

CHAPTER 6 SUPPORTING INFORMATION FOR CHAPTER 2

6.1 Demography supporting figures

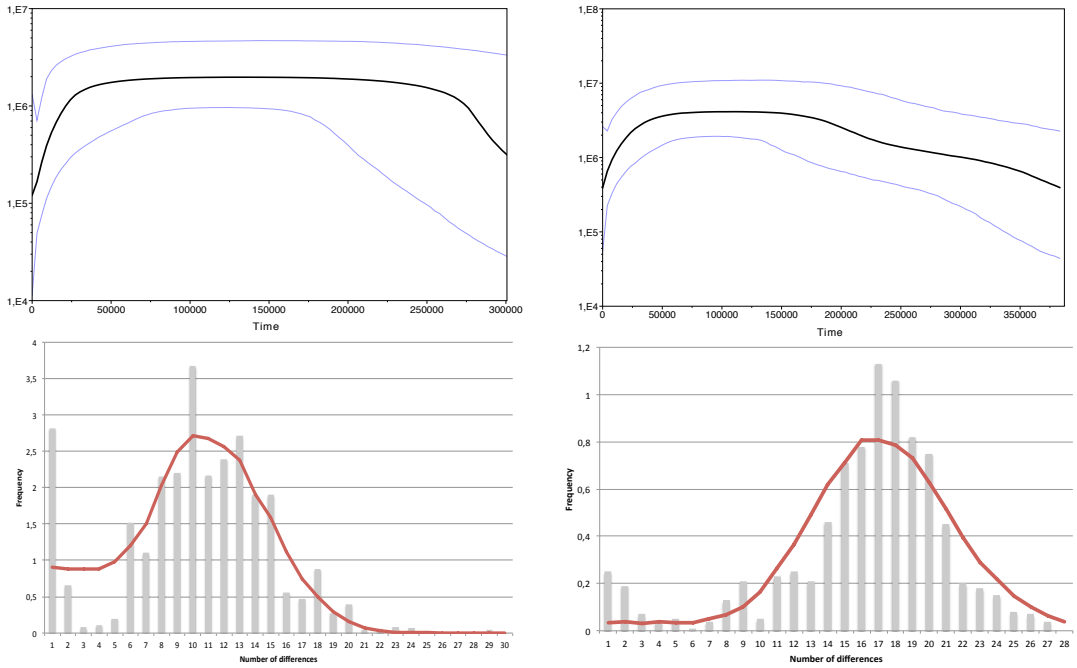


Figure 6.1**.** Bayesian Skyline Plots (BSP) and Mismatch Distributions (MD) of the BAPS clusters A1 (left) and A2 (right). The black line of the BSP represents the median estimate of the population size *Ne*μ over coalescent intervals, with the blue lines representing the confidence intervals. The MD shows the observed (bar plots) and the expected (line plots) values of the distribution of pairwise differences. The two clusters clearly show a signature of past expansion.

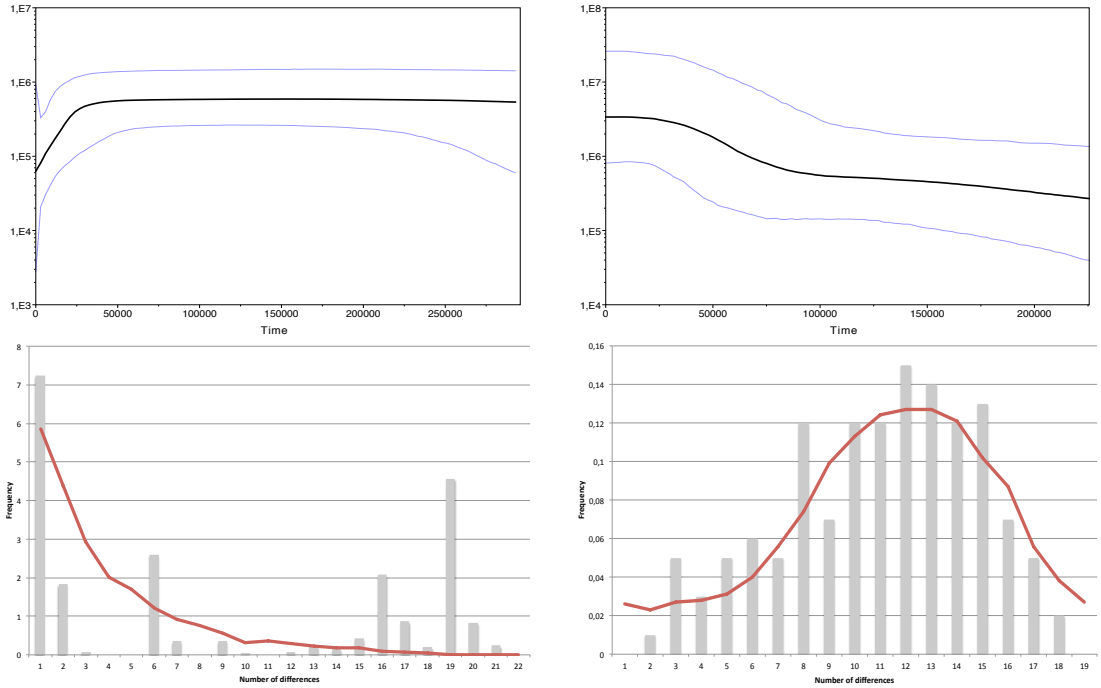


Figure 6.2. Bayesian Skyline Plots (BSP) and Mismatch Distributions (MD) of the BAPS clusters A3 (left) and C (right). The black line of the BSP represents the median estimate of the population size *Ne*μ over coalescent intervals, with the blue lines representing the confidence intervals. The MD shows the observed (bar plots) and the expected (line plots) values of the distribution of pairwise differences. A3 graphs show a stable population, whereas the Balkan lineage C shows a clear signature of expansion.

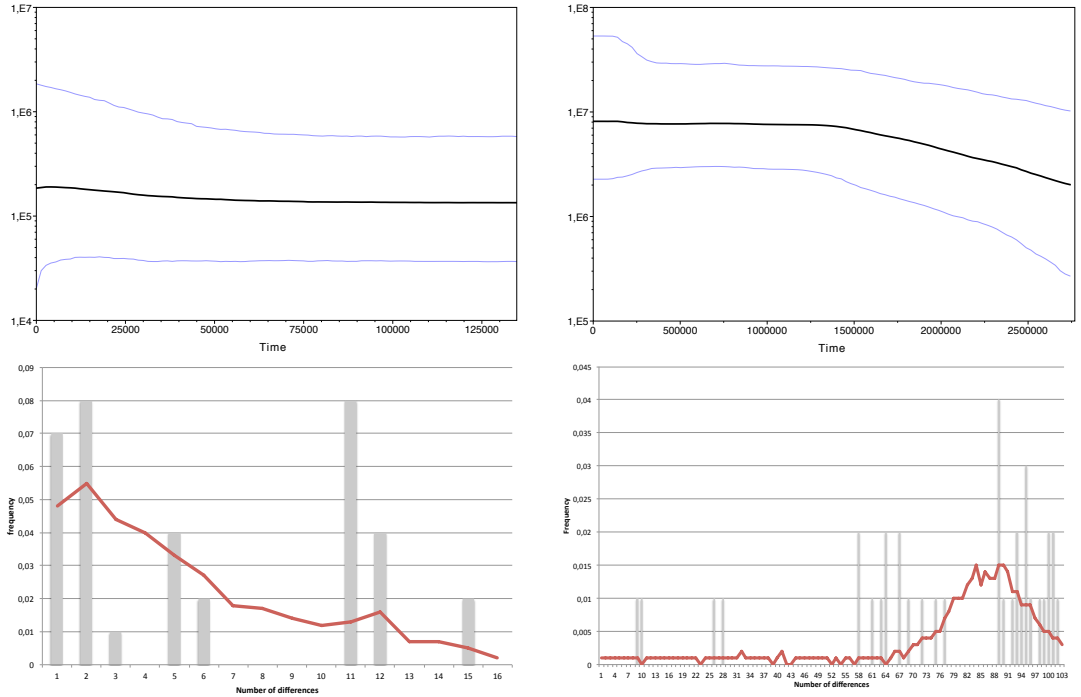


Figure 6.3. Bayesian Skyline Plots (BSP) and Mismatch Distributions (MD) of the BAPS clusters D (left) and E (right). The black line of the BSP represents the median estimate of the population size *Ne*μ over coalescent intervals, with the blue lines representing the confidence intervals. The MD shows the observed (bar plots) and the expected (line plots) values of the distribution of pairwise differences. These two clusters show signatures of stable demographics through time.

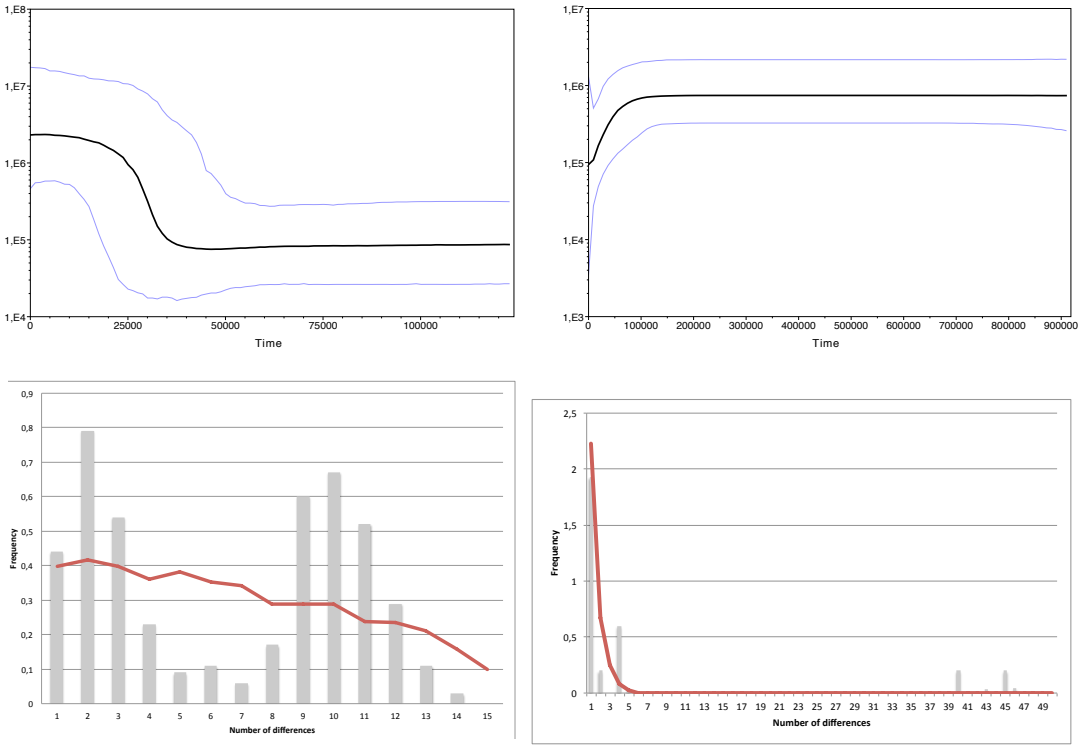


Figure 6.4. Bayesian Skyline Plots (BSP) and Mismatch Distributions (MD) of the BAPS clusters B (left) and F (right). The black line of the BSP represents the median estimate of the population size *Ne*μ over coalescent intervals, with the blue lines representing the confidence intervals. The MD shows the observed (bar plots) and the expected (line plots) values of the distribution of pairwise differences. Lineage B shows a signature of expansion starting 25,000 Ybp, just before the start of the Ice retreat phase. According to the BSP, the Spanish lineage F shows a stable population demography for the last million of years.

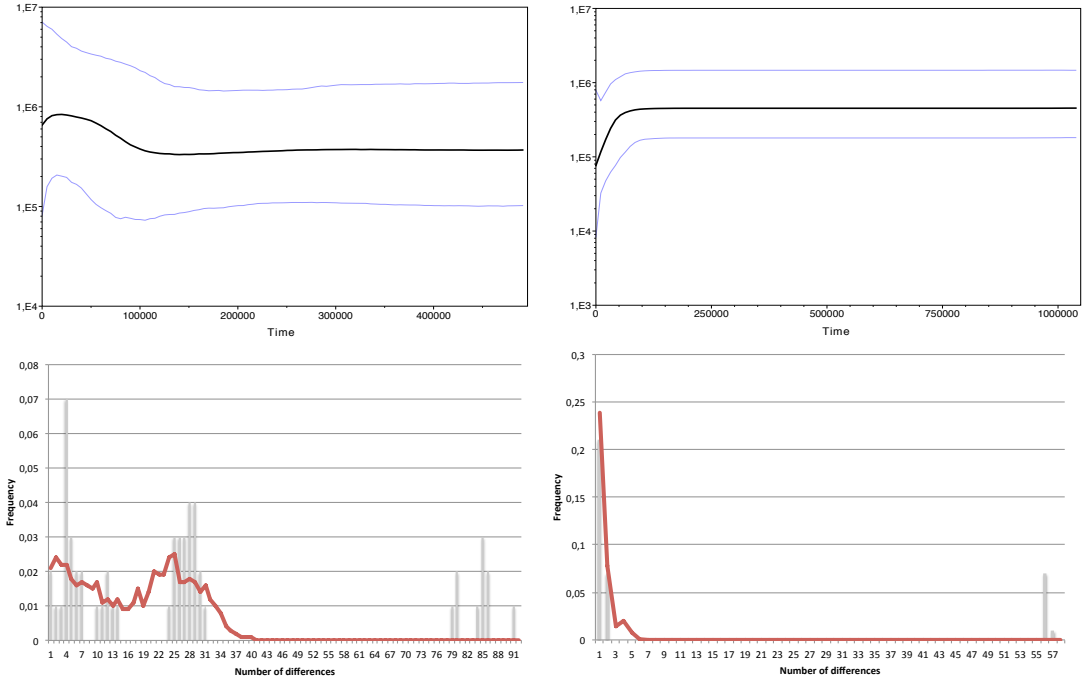


Figure 6.5. Bayesian Skyline Plots (BSP) and Mismatch Distributions (MD) of the BAPS clusters G (left) and H (right). The black line of the BSP represents the median estimate of the population size *Ne*μ over coalescent intervals, with the blue lines representing the confidence intervals. The MD shows the observed (bar plots) and the expected (line plots) values of the distribution of pairwise differences. These two clusters clearly show signatures of stable demography through time.

6.2 Species distribution modeling evaluation

6.2.1 Standard deviation maps

Each map of niche distribution modeling had a standard deviation map (SD) associated, which expressed the variation of habitat suitability obtained across the ten replicates. Small SD values across the predicted landscape indicate high predictive capacity of a variable, while large SD values indicate a variable with little information to discriminate environmental niche suitability. SD maps are reported in Figure 6.6.

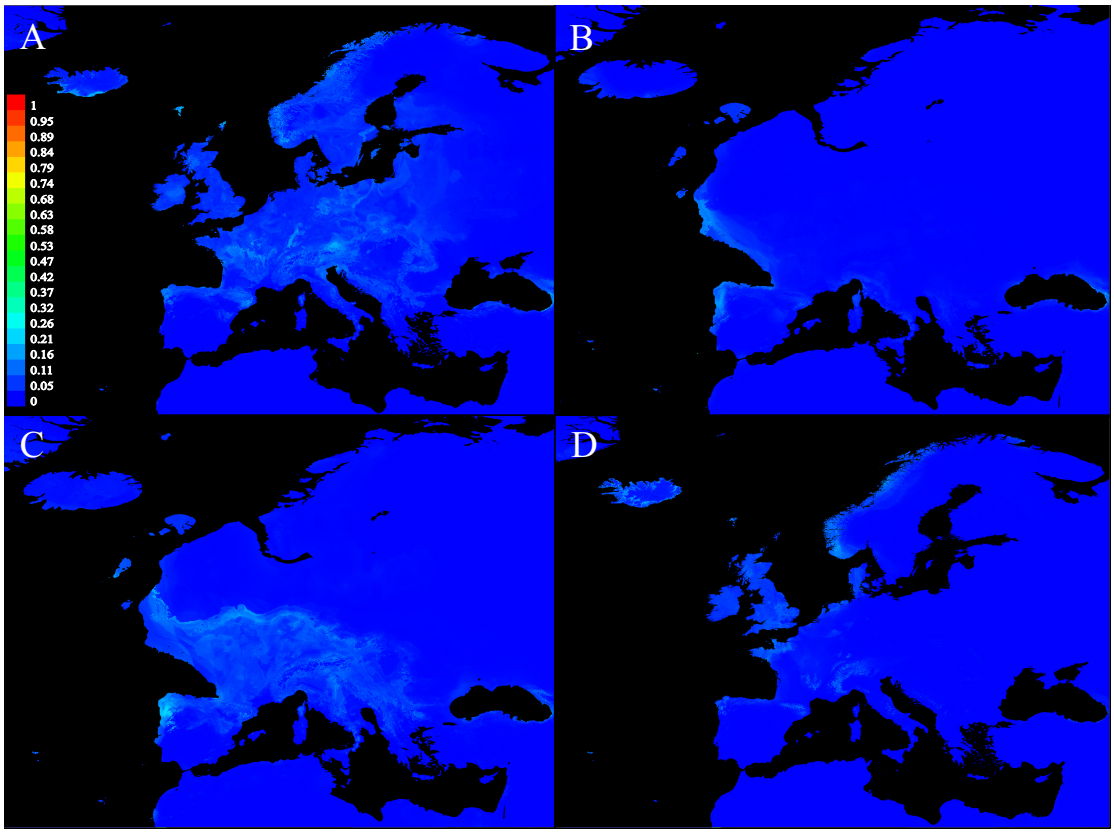


Figure 6.6. Standard deviation (SD) maps of the species distribution model run in Maxent. The color scale represents SD point values on the maps, with higher values represented by warmer colors. A: SD map of the 10 replicated models of the present time; B,C: maps of the standard deviation of the 10 replicated models applied to the environmental layers of CCSM (B) and MIROC (C). The standard deviation of both models showed there was little difference among the 10 model replicates when projected to the CCSM and MIROC environmental variables, with slightly higher variability relative to the MIROC replicates; D: map of the standard deviation of the 10 replicated models applied to the environmental layers of the LIG. Also in this case, the standard deviation of the replicate runs was low.

6.2.2 Model performance evaluation: AUC, MESS and MoD analyses

Model performance was evaluated by an estimate of the Area Under the Curve (AUC) of the Receiving Operating Characteristics (ROC). ROC measures the ability of the prediction to discriminate the presence of the species from the absence (Elith et al. 2010). AUC values range between 0.5 (the model performs like a random variable) and 1 (the model can reliably discriminate between suitable and unsuitable points in the landscape). The average AUC for the current distribution across 10 replicates performed better than the random model, with averaged values of 0.93 for the training data and 0.87 for the test data. Figure 6.7 shows the AUC curve for test data averaged for all the 10 replicate runs; the averaged curve of the test data represents the real test of the model predictive power.

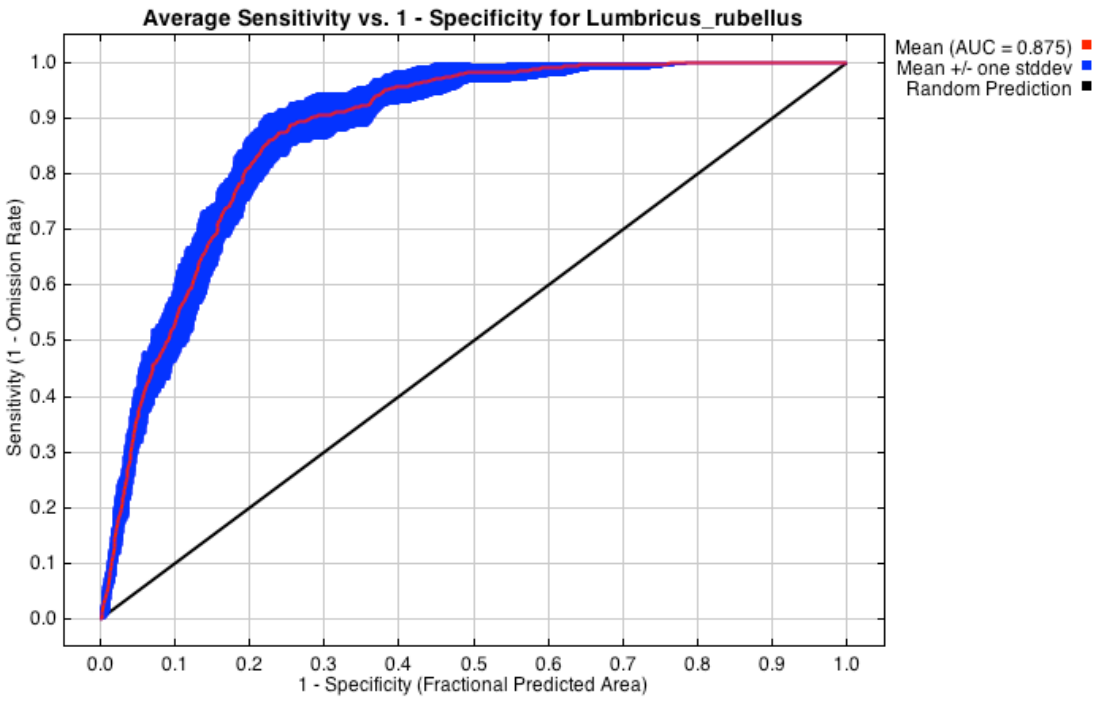


Figure 6.7. AUC estimation for the *Lumbricus rubellus* current distribution, calculated on the test data averaged for 10 replicates. The red line represents the mean AUC of the test data across the replicates, and the blue area represents the standard deviation. The black line represents a random prediction scenario. The test AUC has a value of 0.875, suggesting a good predictive capacity of the model.

A Multivariate Similarity Surface (MESS) analysis and the Most Dissimilar Variable (MoD) analysis were performed to assess the reliability of the projection of the present day model to the past, and to evaluate the sustainability of the model, respectively. MESS measures the similarity of any given point to a reference set of points, relative to the chosen predictor variables (used for model training), it gives negative values for dissimilar points and builds a map of the continuous distribution of these values across the whole region of the projection (Elith et al. 2010). The resulting map enables inference of values outside the predictor value range, where model predictions should be interpreted with caution. The MESS maps help to understand what areas are affected with extrapolation and to what extent, as they are built comparing the environmental variables of the projections with the contemporary ones, used for training. The MoD maps underline what are the variables outside the training range.

When the MESS maps of CCSM and MIROC model of the LGM (Figure 6.8) are commpared, it is evident that the first model is much more affected by variables outside their range, than the MIROC one. The CCSM MoD map points out that the variables outside their training range are BIO2 (Mean monthly diurnal temperature range) and BIO3 (Isothermality, that is, the ratio between BIO2 and temperature annual range) (Figure 6.9). The LIG niche distribution model predicts areas of suitability close with the postulated refugia, even though in some areas, the suitability is greatly reduced; that is, the Alp region, for lineages D and E; the North of Spain, for lineage C; and Brittany and the British Isles, where the B lineage could have survived. The model only fails to predict suitability over the Balkan area, but the MESS and MoD maps of the model evidence that one of the variables is out of range in that area; the variable responsible is BIO7 (temperature annual range), which is the most important variable according to all the estimates of variable importance. Therefore, predictions for the Balkan area for the LIG could not be resolved.

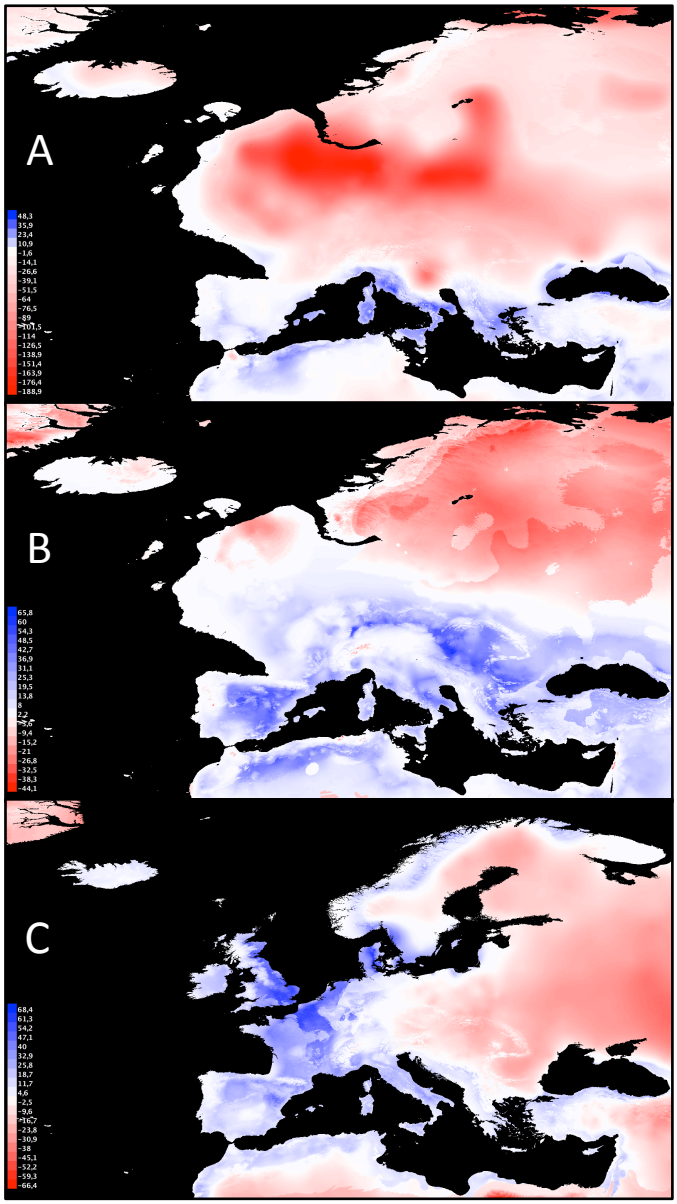


Figure 6.8. MESS maps of the paleodistribution model projections. A,B: LGM CCSM model (A) and the MIROC model (B). C: LIG model. Red areas depict paleoclimatic variable values out of range.

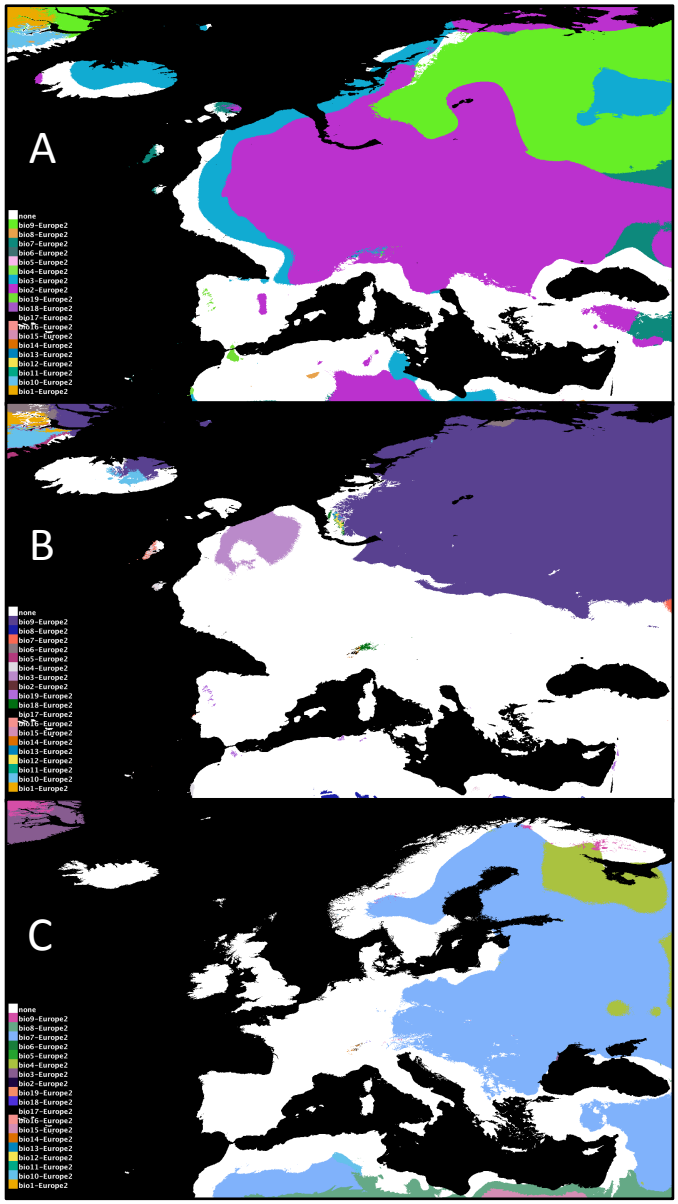


Figure 6.9. A, B: MoD maps of the LGM CCSM model (A) and the MIROC model (B). The main variables outside the range are mean diurnal range (Bio2) and Isothermality (Bio3) for CCSM, and mean temperature of the wettest quarter (Bio8) for MIROC; C: The LIG MoD map shows that the model's extrapolated area is mainly towards the east of the range, and the temperature annual range (Bio7) is the variable responsible of this.

6.3 Samples information

Individual	Haplo group	Haplo type	Locality	lon	lat	Collector	Collection method
AUS.A.02	A3	1	Austria	11.60	47.25	Anita Juen	silico (fasta sequences)
AUS.A.04	A2	2	Austria	11.60	47.25	Anita Juen	silico (fasta sequences)
AUS.A.05	H	3	Austria	11.60	47.25	Anita Juen	silico (fasta sequences)
AUS.A.14	G	4	Austria	11.60	47.25	Anita Juen	silico (fasta sequences)
AUS.A.15	H	3	Austria	11.60	47.25	Anita Juen	silico (fasta sequences)
AUS.A.16	G	5	Austria	11.60	47.25	Anita Juen	silico (fasta sequences)
AUS.A.17	G	6	Austria	11.60	47.25	Anita Juen	silico (fasta sequences)
AUS.A.18	G	6	Austria	11.60	47.25	Anita Juen	silico (fasta sequences)
AZO.A.01	A1	7	Azores	-25.39	37.78	Mike Bruford, Peter Kille, John Morgan, Marta Novo, Luis Cunha	Hand sorting
AZO.A.02	A1	7	Azores	-25.39	37.78	Mike Bruford, Peter Kille, John Morgan, Marta Novo, Luis Cunha	Hand sorting
AZO.A.03	A1	7	Azores	-25.39	37.78	Mike Bruford, Peter Kille, John Morgan, Marta Novo, Luis Cunha	Hand sorting
AZO.A.04	A1	7	Azores	-25.39	37.78	Mike Bruford, Peter Kille, John Morgan, Marta Novo, Luis Cunha	Hand sorting
AZO.A.05	A1	7	Azores	-25.39	37.78	Mike Bruford, Peter Kille, John Morgan, Marta Novo, Luis Cunha	Hand sorting
AZO.A.06	A1	7	Azores	-25.39	37.78	Mike Bruford, Peter Kille, John Morgan, Marta Novo, Luis Cunha	Hand sorting
AZO.A.07	A1	7	Azores	-25.39	37.78	Mike Bruford, Peter Kille, John Morgan, Marta Novo, Luis Cunha	Hand sorting
AZO.A.08	A1	7	Azores	-25.39	37.78	Mike Bruford, Peter Kille, John Morgan, Marta Novo, Luis Cunha	Hand sorting
B10A	I	47	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B10B	A2	48	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B1A	A3	35	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B1B	A1	36	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B2B	A1	37	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B3A	A3	38	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B3B	A2	39	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B4A	A2	40	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B4B	A2	41	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B5A	E	42	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B6A	A3	43	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B7A	A1	44	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B7B	A2	40	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B8A	A2	45	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B8B	A3	43	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
B9B	I	46	Holland	4.77	51.76	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
C1B	A1	49	Holland	5.65	51.9	Pierfrancesco Sechi,	Hand sorting

						Robert Donnelly	
C2A	A1	37	Holland	5.65	51.9	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
C2B	A2	50	Holland	5.65	51.9	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
C4A	A2	51	Holland	5.65	51.9	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
C5A	A1	52	Holland	5.65	51.9	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
C6A	A2	53	Holland	5.65	51.9	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
C9A	A1	54	Holland	5.65	51.9	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
ECO.01	A1	21	UK	-3.74	52.35	John Morgan, Robert Donnelly	Hand sorting
ECO.02	B	104	UK	-3.74	52.35	John Morgan, Robert Donnelly	Hand sorting
ECO.05	B	105	UK	-3.74	52.35	John Morgan, Robert Donnelly	Hand sorting
ECO.06	B	106	UK	-3.74	52.35	John Morgan, Robert Donnelly	Hand sorting
ECO.07	A1	122	UK	-3.74	52.35	John Morgan, Robert Donnelly	Hand sorting
ECO.09	B	104	UK	-3.74	52.35	John Morgan, Robert Donnelly	Hand sorting
ECO.10	B	107	UK	-3.74	52.35	John Morgan, Robert Donnelly	Hand sorting
ECO.13	A2	124	UK	-3.74	52.35	John Morgan, Robert Donnelly	Hand sorting
ECO.14	A1	122	UK	-3.74	52.35	John Morgan, Robert Donnelly	Hand sorting
ECO.15	A1	122	UK	-3.74	52.35	John Morgan, Robert Donnelly	Hand sorting
ECO.16	B	108	UK	-3.74	52.35	John Morgan, Robert Donnelly	Hand sorting
FIN.A.02	A3	8	Finland	23.12	63.28	Jari Haimi	Received in vivo
FIN.A.03	A3	8	Finland	23.12	63.28	Jari Haimi	Received in vivo
FIN.A.04	A3	9	Finland	23.12	63.28	Jari Haimi	Received in vivo
FIN.A.05	A3	10	Finland	23.12	63.28	Jari Haimi	Received in vivo
FIN.A.06	A3	10	Finland	23.12	63.28	Jari Haimi	Received in vivo
FIN.A.07	A3	9	Finland	23.12	63.28	Jari Haimi	Received in vivo
FIN.A.08	A3	8	Finland	23.12	63.28	Jari Haimi	Received in vivo
FIN.A.10	A2	11	Finland	23.12	63.28	Jari Haimi	Received in vivo
FIN.A.11	A3	10	Finland	23.12	63.28	Jari Haimi	Received in vivo
FIN.B.01	A2	12	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.03	A3	9	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.05	A3	9	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.06	A3	9	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.08	A2	12	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.09	A3	9	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.10	A3	9	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.11	A3	9	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.12	A2	13	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.13	A3	9	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.14	A3	9	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.15	A3	9	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.16	A3	9	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.17	A3	10	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.B.18	A3	10	Finland	25.79	62.21	Jari Haimi	Received in vivo
FIN.C.01	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.02	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.03	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.04	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.05	A3	14	Finland	22.42	60.99	Visa Nuutinen	Received in vivo

FIN.C.06	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.07	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.08	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.09	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.10	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.11	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.12	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.13	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.14	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.15	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.16	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.17	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.18	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.19	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.20	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.21	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.22	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.23	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.24	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.25	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.26	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.27	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.28	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.29	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FIN.C.30	A3	10	Finland	22.42	60.99	Visa Nuutinen	Received in vivo
FRA.A.02	A1	15	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.03	A2	16	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.04	A2	17	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.05	A1	7	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.07	A1	18	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.08	A1	7	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.09	A2	19	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.10	A1	7	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.11	A2	16	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.12	A1	18	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.13	A1	15	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.14	A2	16	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.15	A2	20	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.17	A1	18	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.18	A2	16	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.19	A2	19	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.21	A1	15	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.22	A1	21	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.23	A1	15	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.24	A1	7	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.25	A1	7	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.26	A2	19	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.27	A2	22	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.28	A2	20	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
FRA.A.29	A1	7	France	3.04	50.68	Frank Vanderbulcke	Received in vivo
GER.A.01	A2	23	Germany	11.43	47.73	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.A.03	A2	23	Germany	11.43	47.73	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.A.04	H	3	Germany	11.43	47.73	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.A.05	H	3	Germany	11.43	47.73	Pierfrancesco Sechi,	Hand sorting

						Robert Donnelly	
GER.A.06	H	3	Germany	11.43	47.73	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.A.07	H	24	Germany	11.43	47.73	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.A.13	H	3	Germany	11.43	47.73	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.A.14	H	3	Germany	11.43	47.73	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.C.02	G	25	Germany	7.97	47.92	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.C.03	G	26	Germany	7.97	47.92	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.C.05	E	27	Germany	7.97	47.92	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.C.07	G	28	Germany	7.97	47.92	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.C.08	H	29	Germany	7.97	47.92	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.C.11	G	30	Germany	7.97	47.92	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.D.01	G	31	Germany	8.00	47.91	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.D.04	A2	32	Germany	8.00	47.91	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.D.09	A2	32	Germany	8.00	47.91	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.D.10	A2	33	Germany	8.00	47.91	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.D.11	G	34	Germany	8.00	47.91	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GER.D.14	G	31	Germany	8.00	47.91	Pierfrancesco Sechi, Robert Donnelly	Hand sorting
GLF.01	B	109	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.02	B	110	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.04	B	111	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.05	B	112	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.06	B	110	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.07	B	109	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.08	B	113	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.11	B	114	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.14	B	109	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.15	B	109	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.16	B	109	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.17	B	109	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.19	B	115	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
GLF.20	B	116	UK	-2.79	51.31	John Morgan, Robert Donnelly	Hand sorting
HAH.02	A1	122	UK	-2.65	51.51	John Morgan, Robert Donnelly	Hand sorting
HAH.03	A1	122	UK	-2.65	51.51	John Morgan, Robert Donnelly	Hand sorting
HAH.04	A1	125	UK	-2.65	51.51	John Morgan, Robert Donnelly	Hand sorting
HAH.05	A1	122	UK	-2.65	51.51	John Morgan, Robert Donnelly	Hand sorting
HAH.06	A1	126	UK	-2.65	51.51	John Morgan, Robert Donnelly	Hand sorting
HAH.07	A1	21	UK	-2.65	51.51	John Morgan, Robert Donnelly	Hand sorting

HAH.08	A1	7	UK	-2.65	51.51	John Morgan, Robert Donnelly	Hand sorting
HAH.09	A1	122	UK	-2.65	51.51	John Morgan, Robert Donnelly	Hand sorting
HAH.10	A1	122	UK	-2.65	51.51	John Morgan, Robert Donnelly	Hand sorting
HUN.A.01	D	55	Hungary	16.43	47.66	Csuzdi Csaba	Received in vivo
HUN.A.02	D	56	Hungary	16.43	47.66	Csuzdi Csaba	Received in vivo
HUN.A.03	D	55	Hungary	16.43	47.66	Csuzdi Csaba	Received in vivo
HUN.A.04	D	55	Hungary	16.43	47.66	Csuzdi Csaba	Received in vivo
HUN.A.05	D	55	Hungary	16.43	47.66	Csuzdi Csaba	Received in vivo
HUN.A.08	D	57	Hungary	16.43	47.66	Csuzdi Csaba	Received in vivo
HUN.A.09	D	58	Hungary	16.43	47.66	Csuzdi Csaba	Received in vivo
HUN.A.10	D	57	Hungary	16.43	47.66	Csuzdi Csaba	Received in vivo
HUN.A.11	D	59	Hungary	16.43	47.66	Csuzdi Csaba	Received in vivo
HUN.B.01	A1	60	Hungary	18.85	47.64	Csuzdi Csaba	Received in vivo
HUN.B.02	E	61	Hungary	18.85	47.64	Csuzdi Csaba	Received in vivo
HUN.B.03	A1	60	Hungary	18.85	47.64	Csuzdi Csaba	Received in vivo
HUN.B.04	A1	60	Hungary	18.85	47.64	Csuzdi Csaba	Received in vivo
HUN.B.05	A1	62	Hungary	18.85	47.64	Csuzdi Csaba	Received in vivo
HUN.B.07	A1	63	Hungary	18.85	47.64	Csuzdi Csaba	Received in vivo
HUN.B.08	A1	63	Hungary	18.85	47.64	Csuzdi Csaba	Received in vivo
HUN.B.09	A1	64	Hungary	18.85	47.64	Csuzdi Csaba	Received in vivo
HUN.B.10	A1	60	Hungary	18.85	47.64	Csuzdi Csaba	Received in vivo
HUN.B.11	A1	60	Hungary	18.85	47.64	Csuzdi Csaba	Received in vivo
ITA.A.09	A2	65	Italy	11.18	45.44	Tommaso Zanetti, Pierfrancesco Sechi	Hand sorting
ITA.A.14	A2	66	Italy	11.18	45.44	Tommaso Zanetti, Pierfrancesco Sechi	Hand sorting
ITA.A.17	A2	67	Italy	11.18	45.44	Tommaso Zanetti, Pierfrancesco Sechi	Hand sorting
ITA.A.20	A2	68	Italy	11.18	45.44	Tommaso Zanetti, Pierfrancesco Sechi	Hand sorting
ITA.A.30	A2	68	Italy	11.18	45.44	Tommaso Zanetti, Pierfrancesco Sechi	Hand sorting
ITA.A.42	A2	69	Italy	11.18	45.44	Tommaso Zanetti, Pierfrancesco Sechi	Hand sorting
POL.A.01	A3	43	Poland	19.63	50.12	Barbara Ptytycz	Received in EtOH
POL.A.03	A2	39	Poland	19.63	50.12	Barbara Ptytycz	Received in EtOH
POL.A.04	E	70	Poland	19.63	50.12	Barbara Ptytycz	Received in EtOH
POL.A.06	A3	71	Poland	19.63	50.12	Barbara Ptytycz	Received in EtOH
POL.A.07	E	72	Poland	19.63	50.12	Barbara Ptytycz	Received in EtOH
POL.A.08	A2	39	Poland	19.63	50.12	Barbara Ptytycz	Received in EtOH
POL.A.10	E	73	Poland	19.63	50.12	Barbara Ptytycz	Received in EtOH
POL.B.01	E	74	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.02	A1	75	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.03	A1	76	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.05	A1	77	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.06	A3	78	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.08	A1	75	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.09	A1	77	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.10	A2	39	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.11	A1	77	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.12	A3	78	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.13	A1	37	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.14	A1	77	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.15	A1	75	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.16	A1	75	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.17	A3	78	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.18	A3	78	Poland	19.49	50.29	Iwona Giska	Received in vivo

POL.B.20	A2	79	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.21	A1	77	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.22	A3	78	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.23	A1	75	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.24	A3	78	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.25	A2	39	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.26	A3	78	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.27	A1	75	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.28	A1	37	Poland	19.49	50.29	Iwona Giska	Received in vivo
POL.B.30	A1	75	Poland	19.49	50.29	Iwona Giska	Received in vivo
SER.A.02	C	80	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.03	C	81	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.04	C	82	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.05	C	83	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.06	C	84	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.07	C	85	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.08	C	86	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.09	C	87	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.10	C	88	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.11	C	89	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.12	C	90	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.13	C	91	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.14	C	92	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.16	C	93	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.18	C	94	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.19	C	95	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SER.A.23	C	96	Serbia	19.91	44.08	Mira Stojanovic	Received in vivo
SPA.A.01	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.02	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.03	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.04	F	98	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.05	F	99	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.06	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.07	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.08	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.09	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.10	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.11	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.12	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.13	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.14	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.15	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.16	F	100	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.17	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.18	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.19	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.20	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.21	F	101	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.22	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.23	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.24	F	97	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.25	F	100	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SPA.A.26	F	102	Spain	-8.68	42.16	Fuencisla Mariño	Received in vivo
SWE.A.05	A3	10	Sweden	17.71	59.72	Jan Lagerlof	Received in vivo
SWE.A.11	A3	10	Sweden	17.71	59.72	Jan Lagerlof	Received in vivo

SWE.A.14	A3	103	Sweden	17.71	59.72	Jan Lagerlof	Received in vivo
WEM.01	B	117	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
WEM.02	A1	122	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
WEM.03	A1	122	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
WEM.04	B	118	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
WEM.05	A1	122	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
WEM.06	A1	122	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
WEM.08	A1	21	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
WEM.09	A1	122	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
WEM.10	A1	21	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
WEM.11	A1	21	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
WEM.12	B	119	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
WEM.14	A1	122	UK	-3.88	52.34	John Morgan, Robert Donnelly	Hand sorting
YST.01	B	120	UK	-3.69	52.36	John Morgan, Robert Donnelly	Hand sorting
YST.02	B	104	UK	-3.69	52.36	John Morgan, Robert Donnelly	Hand sorting
YST.03	B	104	UK	-3.69	52.36	John Morgan, Robert Donnelly	Hand sorting
YST.04	B	104	UK	-3.69	52.36	John Morgan, Robert Donnelly	Hand sorting
YST.05	B	104	UK	-3.69	52.36	John Morgan, Robert Donnelly	Hand sorting
YST.06	B	121	UK	-3.69	52.36	John Morgan, Robert Donnelly	Hand sorting
YST.07	B	104	UK	-3.69	52.36	John Morgan, Robert Donnelly	Hand sorting
YST.09	A1	122	UK	-3.69	52.36	John Morgan, Robert Donnelly	Hand sorting
YST.10	B	104	UK	-3.69	52.36	John Morgan, Robert Donnelly	Hand sorting

CHAPTER 7 SUPPORTING INFORMATION FOR CHAPTER 3

Table 7.1**.** Mitochondrial genome profiles of *Lumbricus rubellus* lineages. Start and end position, length, start codon, stop codon, intergenic nucleotides and AT% are shown. The number of intergenic nucleotides is negative when there is an overlap between loci.

[1] Truncated stop codons, terminated by post-transcriptional modification (polyadenilation).

	position from-to	Size	Start Codon	Stop Codon	Intergenic Nucleotides	AT%
COI	1-1540	1540	ATG	T[1]		57.86
$tRNA^{Asn}$	1541-1601	61				
COII	1602-2286	685	ATG	T[1]	1	61.02
$tRNA^{Asp}$	2288-2348	61				
ATP8	2349-2508	160	ATG	T[1]		71.25
$tRNA^{Tyr}$	2509-2571	63			-2	
$tRNA^{Gly}$	2570-2632	63			1	
COIII	2634-3411	778	ATG	T[1]	1	56.17
$tRNA^{Gln}$	3413-3481	69				
ND6	3482-3955	474	ATG	TAG		73.38
UNK	3956-4387	432				
CYTB	4388-5536	1149	ATG	TAA	29	60.81
$tRNA^{Trp}$	5566-5627	62			2	
ATP6	5630-6323	694	ATG	T[1]		59.94
$tRNA^{Arg}$	6324-6390	67				
nCR	6391-6675	285				
$tRNA^{His}$	6676-6738	63				
ND5	6739-8464	1726	ATG	T[1]	1	62.40
$tRNA^{Phe}$	8466-8527	62				
$tRNA^{Glu}$	8528-8591	64			-2	
$tRNA^{Pro}$	8590-8652	63				
$tRNA^{Thr}$	8653-8714	62				
ND4L	8715-9011	297	ATG	TAA	-7	64.65
ND4	9005-10361	1357	ATG	T[1]		63.89
$tRNA^{Cys}$	10362-10426	65				
$tRNA^{Met}$	10427-10489	63				
s-rRNA	10490-11273	784				59.92
$tRNA^{Val}$	11274-11337	64			1	
l-rRNA	11339-12580	1242				66.53
$tRNA^{Leu}$	12581-12641	61				
$tRNA^{Ala}$	12642-12703	62				
$tRNA^{Ser}$	12704-12766	63			1	
$tRNA^{Leu}$	12768-12831	64			1	
ND1	12833-13757	925	ATG	T[1]		62.70
$tRNA^{Ile}$	13758-13821	64				
$tRNA^{Lys}$	13822-13887	66				
ND3	13888-14239	352	ATG	T[1]		64.49
$tRNA^{Ser}$	14240-14303	64				
ND2	14304-15309	1006	ATG	T[1]		66.30

A1 (UK)

Gene name	position from-to	Size	Start Codon	Stop Codon	Intergenic Nucleotides	AT%
COI	1-1540	1540	ATG	T[1]		58.25
$tRNA^{Asn}$	1541-1601	61				
COII	1602-2286	685	ATG	T[1]	1	61.31
$tRNA^{Asp}$	2288-2348	61				
ATP8	2349-2508	160	ATG	T[1]		70.00
$tRNA^{Tyr}$	2509-2571	63			-2	
$tRNA^{Gly}$	2570-2632	63			1	
COIII	2634-3411	778	ATG	T[1]	1	56.56
$tRNA^{Gln}$	3413-3481	69				
ND6	3482-3955	474	ATG	TAG		75.23
UNK	3956-4387	432				
CYTB	4388-5536	1149	ATG	TAA	29	61.60
$tRNA^{Trp}$	5566-5627	62			2	
ATP6	5630-6323	694	ATG	T[1]		60.95
$tRNA^{Arg}$	6324-6390	67				
nCR	6391-6717	327				
$tRNA^{His}$	6718-6780	63				
ND5	6781-8506	1726	ATG	T[1]	1	62.80
$tRNA^{Phe}$	8508-8569	62				
$tRNA^{Glu}$	8570-8633	64			-2	
$tRNA^{Pro}$	8632-8694	63				
$tRNA^{Thr}$	8695-8756	62				
ND4L	8757-9053	297	ATG	TAA	-7	63.64
ND4	9047-10403	1357	ATG	T[1]		64.63
$tRNA^{Cys}$	10404-10468	65				
$tRNA^{Met}$	10469-10531	63				
s-rRNA	10532-11315	784				60.97
$tRNA^{Val}$	11316-11379	64			1	
l-rRNA	11381-12622	1242				66.69
$tRNA^{Leu}$	12623-12683	61				
$tRNA^{Ala}$	12684-12745	62				
$tRNA^{Ser}$	12746-12808	63			1	
$tRNA^{Leu}$	12810-12873	64			1	
ND1	12875-13799	925	ATG	T[1]		62.92
$tRNA^{Ile}$	13800-13863	64				
$tRNA^{Lys}$	13864-13929	66				
ND3	13930-14281	352	ATG	T[1]		64.77
$tRNA^{Ser}$	14282-14345	64				
ND2	14346-15351	1006	ATG	T[1]		66.00

A1 (Hungary)

Gene name	position from-to	Size	Start Codon	Stop Codon	Intergenic Nucleotides	AT%
COI	1-1540	1540	ATG	T[1]		58.25
$tRNA^{Asn}$	1541-1601	61				
COII	1602-2286	685	ATG	T[1]	1	60.73
$tRNA^{Asp}$	2288-2348	61				
ATP8	2349-2508	160	ATG	T[1]		70.00
$tRNA^{Tyr}$	2509-2571	63			-2	
$tRNA^{Gly}$	2570-2632	63			1	
COIII	2634-3411	778	ATG	T[1]	1	56.43
$tRNA^{Gln}$	3413-3481	69				
ND6	3482-3954	473	ATG	TAG		75.06
UNK	3955-4391	437				
CYTB	4392-5543	1152	ATG	TAA	26	60.81
$tRNA^{Trp}$	5570-5631	62			2	
ATP6	5634-6327	694	ATG	T[1]		60.81
$tRNA^{Arg}$	6328-6394	67				
nCR	6395-6717	323				
$tRNA^{His}$	6718-6780	63				
ND5	6781-8506	1726	ATG	T[1]	1	72.97
$tRNA^{Phe}$	8508-8569	62				
$tRNA^{Glu}$	8570-8633	64			-2	
$tRNA^{Pro}$	8632-8694	63				
$tRNA^{Thr}$	8695-8756	62				
ND4L	8757-9053	297	ATG	TAA	-7	61.62
ND4	9047-10403	1357	ATG	T[1]		63.96
$tRNA^{Cys}$	10404-10468	65				
$tRNA^{Met}$	10469-10531	63				
s-rRNA	10532-11315	784				61.05
$tRNA^{Val}$	11316-11379	64			1	
l-rRNA	11381-12623	1243				66.88
$tRNA^{Leu}$	12624-12684	61				
$tRNA^{Ala}$	12685-12746	62				
$tRNA^{Ser}$	12747-12809	63			1	
$tRNA^{Leu}$	12811-12874	64			1	
ND1	12876-13800	925	ATG	T[1]		63.46
$tRNA^{Ile}$	13801-13864	64				
$tRNA^{Lys}$	13865-13930	66				
ND3	13931-14282	352	ATG	T[1]		65.91
$tRNA^{Ser}$	14283-14346	64				
ND2	14347-15352	1006	ATG	T[1]		66.80

A2 (France)

Gene name	position from-to	Size	Start Codon	Stop Codon	Intergenic Nucleotides	AT%
COI	1-1540	1540	ATG	T[1]		58.51
$tRNA^{Asn}$	1541-1601	61				
COII	1602-2286	685	ATG	T[1]	1	60.88
$tRNA^{Asp}$	2288-2348	61				
ATP8	2349-2508	160	ATG	T[1]		69.38
$tRNA^{Tyr}$	2509-2571	63			-2	
$tRNA^{Gly}$	2570-2632	63			1	
COIII	2634-3411	778	ATG	T[1]	1	56.68
$tRNA^{Gln}$	3413-3481	69				
ND6	3482-3955	474	ATG	TAG		76.44
UNK	3956-4388	433				
CYTB	4389-5540	1152	ATG	TAA	26	61.42
$tRNA^{Trp}$	5567-5628	62			2	
ATP6	5631-6324	694	ATG	T[1]		61.10
$tRNA^{Arg}$	6325-6391	67				
nCR	6392-7312	921				
$tRNA^{His}$	7313-7375	63				
ND5	7376-9101	1726	ATG	T[1]	1	62.57
$tRNA^{Phe}$	9103-9164	62				
$tRNA^{Glu}$	9165-9228	64			-2	
$tRNA^{Pro}$	9227-9289	63				
$tRNA^{Thr}$	9290-9351	62				
ND4L	9352-9648	297	ATG	TAA	-7	62.29
ND4	9642-10998	1357	ATG	T[1]		65.14
$tRNA^{Cys}$	10999-11063	65				
$tRNA^{Met}$	11064-11126	63				
s-rRNA	11127-11909	783				60.59
$tRNA^{Val}$	11910-11973	64			1	
l-rRNA	11975-13216	1242				66.45
$tRNA^{Leu}$	13217-13277	61				
$tRNA^{Ala}$	13278-13339	62				
$tRNA^{Ser}$	13340-13402	63			1	
$tRNA^{Leu}$	13404-13467	64			1	
ND1	13469-14393	925	ATG	T[1]		63.35
$tRNA^{Ile}$	14394-14457	64				
$tRNA^{Lys}$	14458-14523	66				
ND3	14524-14875	352	ATG	T[1]		66.19
$tRNA^{Ser}$	14876-14939	64				
ND2	14940-15945	1006	ATG	T[1]		66.80

A3 (Finland)

Gene name	position from-to	Size	Start Codon	Stop Codon	Intergenic Nucleotides	AT%
COI	1-1540	1540	ATG	T[1]		58.44
$tRNA^{Asn}$	1541-1601	61				
COII	1602-2286	685	ATG	T[1]	1	62.92
$tRNA^{Asp}$	2288-2348	61				
ATP8	2349-2508	160	ATG	T[1]		64.38
$tRNA^{Tyr}$	2509-2570	62			-2	
$tRNA^{Gly}$	2569-2631	63			1	
COIII	2633-3410	778	ATG	T[1]	1	55.91
$tRNA^{Gln}$	3412-3480	69				
ND6	3481-3954	474	ATG	TAA		72.10
UNK	3955-4377	423				
CYTB	4378-5526	1149	ATG	TAA	33	60.11
$tRNA^{Trp}$	5560-5621	62			2	
ATP6	5624-6317	694	ATG	T[1]		61.53
$tRNA^{Arg}$	6318-6391	74				
nCR	6392-6687	296				
$tRNA^{His}$	6688-6750	63				
ND5	6751-8476	1726	ATG	T[1]	1	62.46
$tRNA^{Phe}$	8478-8539	62				
$tRNA^{Glu}$	8540-8603	64			-2	
$tRNA^{Pro}$	8602-8663	62				
$tRNA^{Thr}$	8664-8725	62				
ND4L	8726-9022	297	ATG	TAA	-7	61.95
ND4	9016-10372	1357	ATG	T[1]		64.19
$tRNA^{Cys}$	10373-10437	65				
$tRNA^{Met}$	10438-10500	63				
s-rRNA	10501-11286	786				60.18
$tRNA^{Val}$	11287-11348	62			1	
l-rRNA	11350-12590	1241				66.75
$tRNA^{Leu}$	12591-12651	61				
$tRNA^{Ala}$	12652-12713	62				
$tRNA^{Ser}$	12714-12776	63			1	
$tRNA^{Leu}$	12778-12841	64			1	
ND1	12843-13767	925	ATG	T[1]		62.16
$tRNA^{Ile}$	13768-13831	64				
$tRNA^{Lys}$	13832-13896	65				
ND3	13897-14248	352	ATG	T[1]		65.06
$tRNA^{Ser}$	14249-14312	64				
ND2	14313-15318	1006	ATG	T[1]		65.21

C (Serbia)

Gene name	position from-to	Size	Start Codon	Stop Codon	Intergenic Nucleotides	AT%
COI	1-1540	1540	ATG	T[1]		59.09
$tRNA^{Asn}$	1541-1601	61				
COII	1602-2286	685	ATG	T[1]	1	60.73
$tRNA^{Asp}$	2288-2348	61				
ATP8	2349-2508	160	ATG	T[1]		66.88
$tRNA^{Tyr}$	2509-2571	63			-2	
$tRNA^{Gly}$	2570-2632	63			1	
COIII	2634-3411	778	ATG	T[1]	1	56.17
$tRNA^{Gln}$	3413-3481	69				
ND6	3482-3955	474	ATG	TAA		73.51
UNK	3956-4359	404				
CYTB	4360-5508	1149	ATG	TAA	28	60.46
$tRNA^{Trp}$	5537-5598	62			2	
ATP6	5601-6294	694	ATG	T[1]		61.96
$tRNA^{Arg}$	6295-6361	67				
nCR	6362-6688	327				
$tRNA^{His}$	6689-6751	63				
ND5	6752-8477	1726	ATG	T[1]	1	62.17
$tRNA^{Phe}$	8479-8540	62				
$tRNA^{Glu}$	8541-8604	64			-2	
$tRNA^{Pro}$	8603-8664	62				
$tRNA^{Thr}$	8665-8726	62				
ND4L	8727-9023	297	ATG	TAA	-7	61.62
ND4	9017-10373	1357	ATG	T[1]		63.45
$tRNA^{Cys}$	10374-10438	65				
$tRNA^{Met}$	10439-10501	63				
s-rRNA	10502-11287	786				60.38
$tRNA^{Val}$	11288-11349	62			1	
l-rRNA	11351-12591	1241				66.91
$tRNA^{Leu}$	12592-12652	61				
$tRNA^{Ala}$	12653-12714	62				
$tRNA^{Ser}$	12715-12777	63			1	
$tRNA^{Leu}$	12779-12842	64			1	
ND1	12844-13768	925	ATG	T[1]		62.27
$tRNA^{Ile}$	13769-13832	64				
$tRNA^{Lys}$	13833-13897	65				
ND3	13898-14249	352	ATG	T[1]		65.34
$tRNA^{Ser}$	14250-14313	64				
ND2	14314-15319	1006	ATG	T[1]		66.60

D (Hungary)

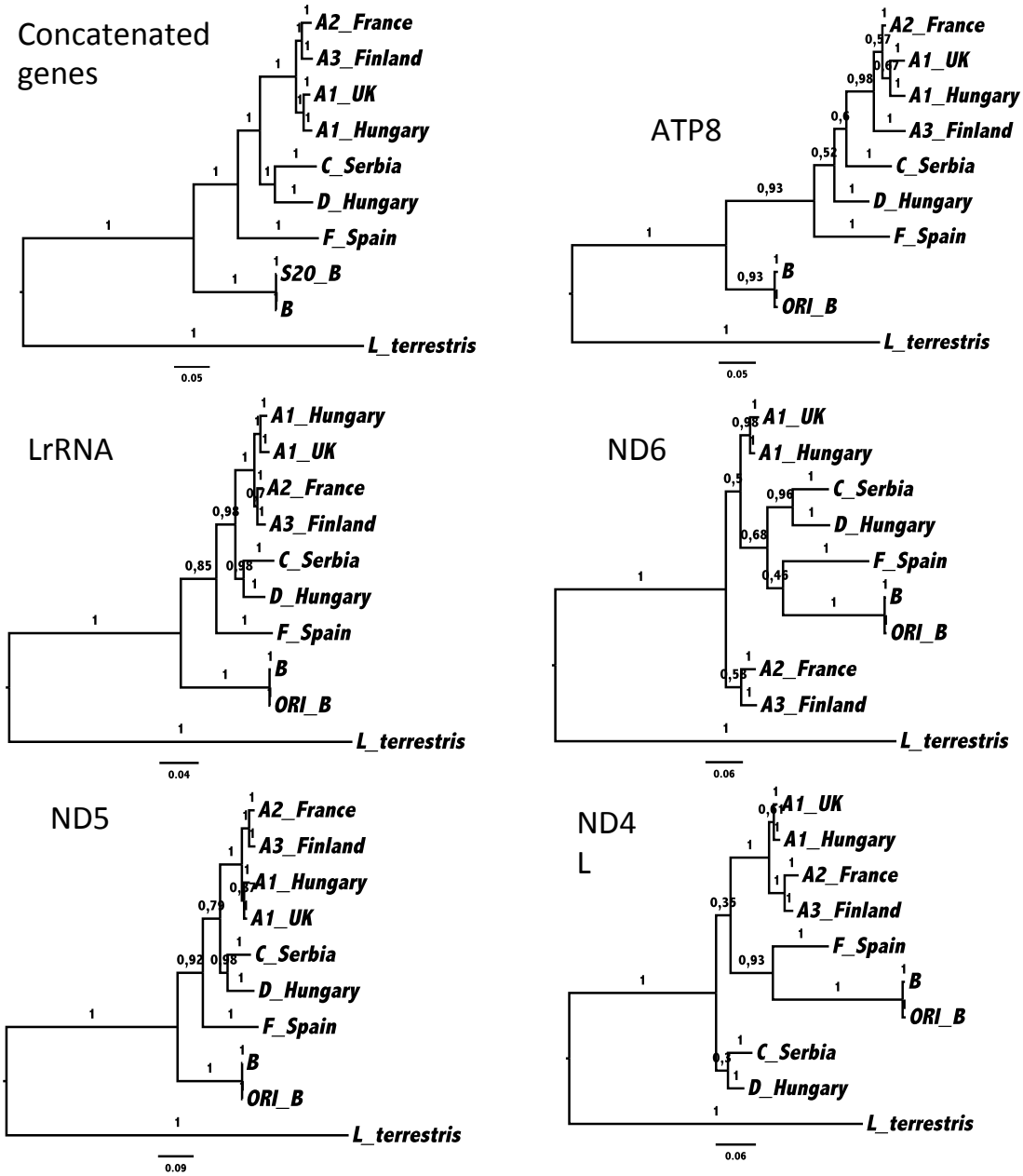
Gene name	position from-to	Size	Start Codon	Stop Codon	Intergenic Nucleotides	AT%
COI	1-1540	1540	ATG	T[1]		58.90
tRNA^{Asn}	1541-1601	61				
COII	1602-2286	685	ATG	T[1]	1	61.02
tRNA^{Asp}	2288-2348	61				
ATP8	2349-2508	160	ATG	T[1]		66.88
tRNA^{Tyr}	2509-2571	63			-2	
tRNA^{Gly}	2570-2632	63			1	
COIII	2634-3411	778	ATG	T[1]	1	55.40
tRNA^{Gln}	3413-3481	69				
ND6	3482-3955	474	ATG	TAG		70.07
UNK	3956-4356	401				
CYTB	4357-5508	1152	ATG	TAA	27	60.90
tRNA^{Trp}	5536-5597	62			2	
ATP6	5600-6293	694	ATG	T[1]		62.39
tRNA^{Arg}	6294-6359	66				
nCR	6360-6633	274				
tRNA^{His}	6634-6696	63				
ND5	6697-8422	1726	ATG	T[1]	1	61.41
tRNA^{Phe}	8424-8485	62				
tRNA^{Glu}	8486-8549	64			-2	
tRNA^{Pro}	8548-8609	62				
tRNA^{Thr}	8610-8672	63				
ND4L	8673-8969	297	ATG	TAA	-7	61.95
ND4	8963-10319	1357	ATG	T[1]		63.38
tRNA^{Cys}	10320-10384	65				
tRNA^{Met}	10385-10447	63				
s-rRNA	10448-11232	785				61.15
tRNA^{Val}	11233-11294	62			1	
l-rRNA	11296-12535	1240				65.43
tRNA^{Leu}	12536-12596	61				
tRNA^{Ala}	12597-12658	62				
tRNA^{Ser}	12659-12721	63			1	
tRNA^{Leu}	12723-12786	64			1	
ND1	12788-13712	925	ATG	T[1]		62.59
tRNA^{Ile}	13713-13776	64				
tRNA^{Lys}	13777-13842	66				
ND3	13843-14194	352	ATG	T[1]		61.93
tRNA^{Ser}	14195-14258	64				
ND2	14259-15264	1006	ATG	T[1]		64.21

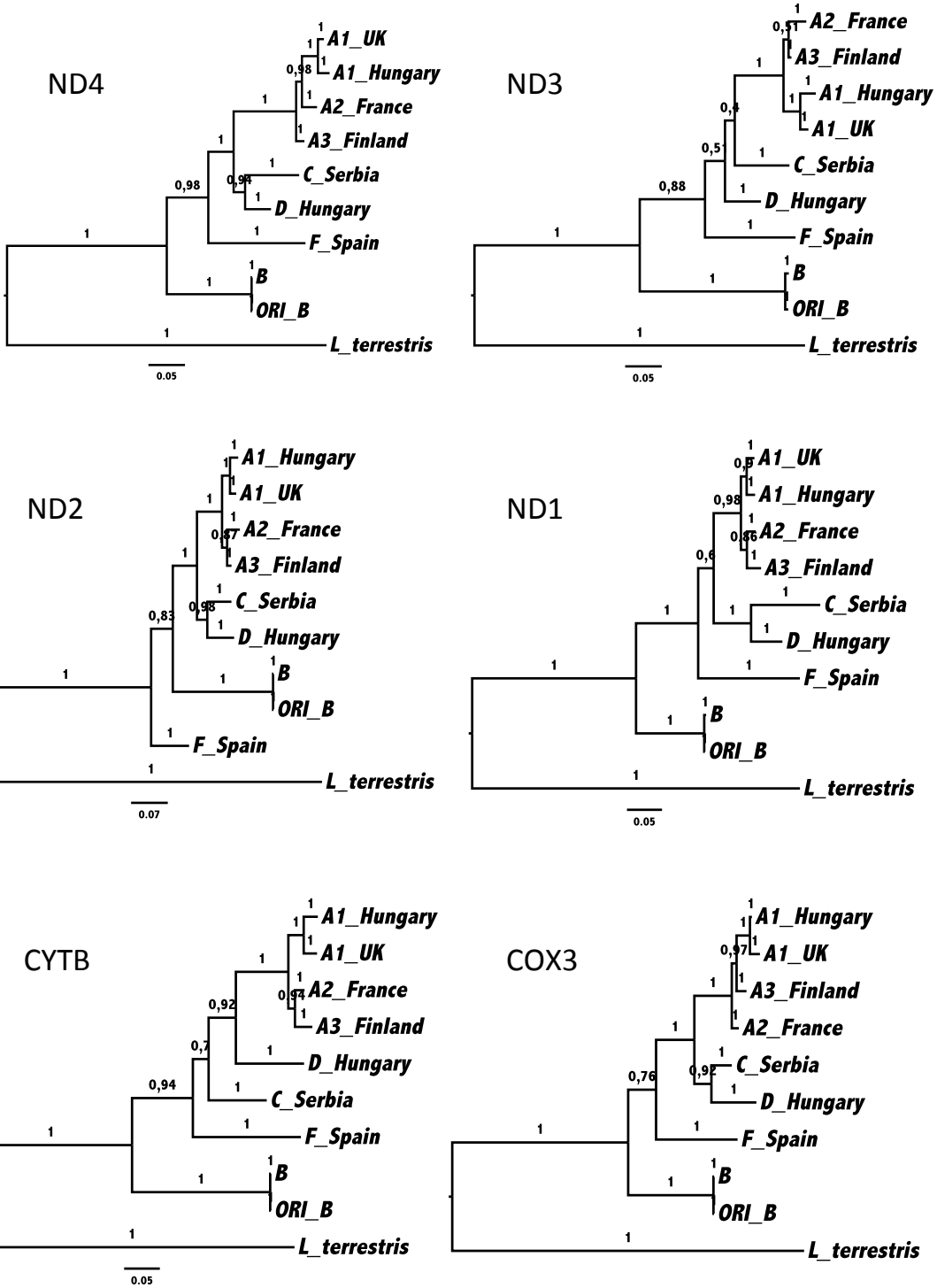
F (Spain)

Gene name	position from-to	Size	Start Codon	Stop Codon	Intergenic Nucleotides	AT%
COI	1-1540	1540	ATG	T[1]		58.57
tRNA^{Asn}	1541-1601	61				
COII	1602-2286	685	ATG	T[1]	1	59.42
tRNA^{Asp}	2288-2348	61				
ATP8	2349-2508	160	ATG	T[1]		67.50
tRNA^{Tyr}	2509-2571	63			-2	
tRNA^{Gly}	2570-2631	62			1	
COIII	2633-3410	778	ATG	T[1]	1	55.78
tRNA^{Gln}	3412-3480	69				
ND6	3481-3954	474	ATG	TAA		68.86
UNK	3955-4243	289				
CYTB	4244-5392	1149	ATG	TAA	24	61.51
tRNA^{Trp}	5417-5478	62			2	
ATP6	5481-6174	694	ATG	T[1]		59.65
tRNA^{Arg}	6175-6247	73				
nCR	6248-7002	755				
tRNA^{His}	7003-7065	63				
ND5	7066-8793	1728	ATG	T[1]	1	61.05
tRNA^{Phe}	8795-8856	62				
tRNA^{Glu}	8857-8920	64			-2	
tRNA^{Pro}	8919-8980	62				
tRNA^{Thr}	8981-9042	62				
ND4L	9043-9339	297	ATG	TAA	-7	63.64
ND4	9333-10689	1357	ATG	T[1]		62.71
tRNA^{Cys}	10690-10753	64				
tRNA^{Met}	10754-10816	63				
s-rRNA	10817-11601	785				60.59
tRNA^{Val}	11602-11664	63			1	
l-rRNA	11666-12906	1241				66.34
tRNA^{Leu}	12907-12967	61				
tRNA^{Ala}	12968-13029	62				
tRNA^{Ser}	13030-13092	63			1	
tRNA^{Leu}	13094-13157	64			1	
ND1	13159-14085	927	ATG	T[1]		62.57
tRNA^{Ile}	14086-14150	65				
tRNA^{Lys}	14151-14217	67				
ND3	14218-14569	352	ATG	T[1]		59.38
tRNA^{Ser}	14570-14633	64				
ND2	14634-15640	1007	ATG	T[1]		64.45

B (UK)

Figure 7.1. Bayesian phylogenetic trees. Each tree represent a single gene alignment, except the first one (first page, upper left), which was obtained from a dataset built concatenating all the genes.




ND4
A1_UK
A1_Hungary
A2_France
A3_Finland
C_Serbia
D_Hungary
F_Spain
B
ORI_B
L_terrestris
0.05
ND3
A2_France
A3_Finland
A1_Hungary
A1_UK
C_Serbia
D_Hungary
F_Spain
B
ORI_B
L_terrestris
0.05
ND2
A1_Hungary
A1_UK
A2_France
A3_Finland
C_Serbia
D_Hungary
B
ORI_B
F_Spain
L_terrestris
0.07
ND1
A1_UK
A1_Hungary
A2_France
A3_Finland
C_Serbia
D_Hungary
F_Spain
B
ORI_B
L_terrestris
0.05
CYTB
A1_Hungary
A1_UK
A2_France
A3_Finland
D_Hungary
C_Serbia
F_Spain
B
ORI_B
L_terrestris
0.05
COX3
A1_Hungary
A1_UK
A3_Finland
A2_France
C_Serbia
D_Hungary
F_Spain
B
ORI_B
L_terrestris