MOLECULAR ECOLOGY

Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L.

Journal:	*Molecular Ecology*
Manuscript ID	MEC-15-1040.R2
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Jeffries, Daniel; University of Hull, School of Biological, Biomedical and Environmental Science Copp, Gordon; Centre for Environment, Fisheries and Aquaculture Science, ; Bournemouth University, Department of Life and Environmental Sciences Lawson Handley, Lori; University of Hull, School of Biological, Biomedical and Environmental Science Olsen, Hakan; Södertörn University, School of Natural Science, Technology and Environmental Studies Sayer, Carl; University College London, Department of Geography Haenfling, Bernd; University of Hull, School of Biological, Biomedical and Environmental Science
Keywords:	Approximate Bayesian Computation, Postglacial recolonisation, Landscape Genetics, Conservation Biology, Study design, Population structure

SCHOLARONE™
Manuscripts

"This is the peer reviewed version of the following article: Jeffries, D. L., Copp, G. H., Lawson Handley, L., Olsén, K. H., Sayer, C. D. and Hänfling, B. (2016), Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, Carassius carassius, L.. Mol Ecol. doi:10.1111/mec.13613, which has been published in final form at http://onlinelibrary.wiley.com/doi/10.1111/mec.13613/abstract. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving."

Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius,* L.

Authors: [1]Daniel L Jeffries, [2]Gordon H Copp, [1]Lori Lawson Handley, [3]K. Håkan Olsén, [4]Carl D Sayer, [1]Bernd Hänfling

[1] *Evolutionary Biology Group, School of Biological, Biomedical and Environmental Sciences, Hardy Building, University of Hull, Hull, HU6 7RX, UK*
[2]*Salmon & Freshwater Team, Cefas, Pakefield Road, Lowestoft, Suffolk NR33 0HT, UK, and Department of Life and Environmental Sciences, Faculty of Science and Technology, Bournemouth University, Poole, UK*
[3]*School of Natural Science, Technology and Environmental Studies, Södertörn University, Alfred Nobels allé 7, Flemingsberg, 141 89 Huddinge, Sweden*
[4] *Environmental Change Research Centre, Department of Geography, University College London, Pearson Building, Gower Street, London, WC1E 6BT, UK*

Keywords

Approximate Bayesian Computation, Postglacial recolonisation, Landscape Genetics, Conservation Biology, Study design, Population structure

Corresponding author: Daniel Jeffries (dljeffries86@gmail.com) *Evolutionary Biology Group, School of Biological, Biomedical and Environmental Sciences, Hardy Building, University of Hull, Hull, HU6 7RX, UK*

Running title: The complex phylogeography of the crucian carp

For Review Only

Abstract

The conservation of threatened species must be underpinned by phylogeographic knowledge. This need is epitomised by the freshwater fish *Carassius carassius*, which is in decline across much of its European range. Restriction site associated DNA sequencing (RADseq) is increasingly used for such applications, however RADseq is expensive, and limitations on sample number must be weighed against the benefit of large numbers of markers. This trade-off has previously been examined using simulation studies, however, empirical comparisons between these markers, especially in a phylogeographic context, are lacking. Here, we compare the results from microsatellites and RADseq for the phylogeography of *C. carassius* to test whether it is more advantageous to genotype fewer markers (microsatellites) in many samples, or many markers (SNPs) in fewer samples. These datasets, along with data from the mitochondrial cytochrome b gene, agree on broad phylogeographic patterns; showing the existence of two previously unidentified *C. carassius* lineages in Europe; one found throughout northern and central-eastern European drainages, and a second almost exclusively confined to the Danubian catchment. These lineages have been isolated for approximately 2.15 M years, and should be considered separate conservation units. RADseq recovered finer population structure and stronger patterns of IBD than microsatellites, despite including only 17.6% of samples (38% of populations and 52% of samples per population). RADseq was also used along with Approximate Bayesian Computation to show that the postglacial colonisation routes of *C. carassius* differ from the general patterns of freshwater fish in Europe, likely as a result of their distinctive ecology.

For Review Only

Introduction

Phylogeographic studies have revealed that the contemporary distributions of European taxa and their genetic diversity have been largely shaped by the glacial cycles of the Pleistocene epoch, and in particular by range shifts during recolonisation from glacial refugia (Hewitt 1999). In freshwater fishes, the dynamics of recolonisation are tightly linked to the history of river drainage systems (Bianco 1990; Bănărescu 1990, 1992; Bernatchez & Wilson 1998; Reyjol *et al.* 2006). For example, watersheds pose a significant barrier to fish dispersal, often resulting in strong genetic structuring across separate drainage systems (Durand *et al.* 1999; Hänfling *et al.* 2002). However, during glacial melt periods, ephemeral rivers and periglacial lakes can arise, providing opportunities for colonisation (Gibbard *et al.* 1988) of otherwise isolated drain basins (Grosswald 1980; Arkhipov *et al.* 1995). These processes have resulted in complicated recolonisation scenarios in Europe, which, in contrast to North America (Bernatchez & Wilson 1998), appear to possess few general patterns of population structure. Furthermore, previous phylogeographic studies have predominantly focused on highly mobile, obligatory or facultatively lotic species, with more sedentary, lentic species being largely overlooked.

The crucian carp, *Carassius carassius* (Linnaeus 1758), is native to parts of central, eastern and northern Europe and almost exclusively restricted to lentic ecosystems, including lakes, ponds and river floodplains (Copp 1991; Copp *et al.* 2008). *C. carassius*, has recently experienced sharp declines in the number and sizes of populations throughout its native range, leading to some local population extinctions. The reasons for these declines include habitat loss through drought and terrestrialisation in England (Copp 1991; Wheeler 2000; Sayer *et al.* 2011), acidification (Holopainen & Oikari 1992), poor water quality in the Danube river catchment (Navodaru *et al.* 2002), and hybridisation with several non-native species (Copp *et al.* 2010; Savini *et al.* 2010; Mezhzherin *et al.* 2012; Wouters *et al.* 2012; Rylková *et al.* 2013). The susceptibility of *C.*

For Review Only

carrassius to genetic isolation and bottlenecks is compounded by small population sizes (Hänfling *et al.* 2005) and low dispersal (Holopainen *et al.* 1997). Strong geographic structure is therefore likely in this species. Although the threats to *C. carassius* populations are recognised on a regional level (Lusk *et al.* 2004; Mrakovčić *et al.* 2007; Wolfram & Mikschi 2007; Simic, V *et al.* 2009; Copp & Sayer 2010), a global conservation strategy is missing. Broad scale phylogeographic data and definition of evolutionary significant units are essential for informing unified conservation efforts for this species (Frankham *et al*. 2002).

For Review Only

Phylogeographic data have traditionally been collected using mitochondrial gene regions and/or nuclear markers such as AFLPs and microsatellites. However, cost and time often limits the number of these nuclear markers used, which can result in low power for addressing phylogeographic questions (Cornuet & Luikart 1996; Luikart & Cornuet 2008; Landguth *et al.* 2012; Peery *et al.* 2012; Hoban *et al.* 2013). Single nucleotide polymorphisms (SNPs) are increasingly used in phylogeography for assessments of population structure (for example see Morin *et al.* 2010; Emerson *et al.* 2010; Hess *et al.* 2011; Hauser *et al.* 2011). However, being bi-allelic, SNP loci contain less information than highly polymorphic microsatellites (Coates *et al.* 2009) and therefore large numbers of SNPs are needed to provide adequate statistical power. SNP discovery and assay development, which has been costly and slow in the past, has recently been greatly facilitated by the invention of restriction site associated DNA sequencing (RADseq, (Miller *et al.* 2006)), which enables the fast identification of thousands of orthologous SNP markers in non-model organisms. Nevertheless, although next generation sequencing costs are falling, RADseq remains a relatively expensive approach, which often constrains the number of biological samples that can be included in a given study. Researchers are, therefore, faced with a trade-off between the number of samples and loci during study design. The optimal balance between the two is likely to be based on several important but often unknown properties of the study system in question, for example the strength of population structure (i.e. F_{ST}). Identifying these properties and comparing the relative strengths and weaknesses of different molecular markers have recently been highlighted as priority topics in landscape genetics and phylogeography (Epperson *et al.* 2010; Balkenhol & Landguth 2011). Recent simulation studies have provided some important insights into this trade-off, for example, Schwartz & McKelvey (2009) find that patchy geographic sampling along an IBD gradient could result in falsely identified distinct lineages, whereas Landguth *et al.* (2012) find that increasing the number of loci can strengthen the correlation between genetic and geographic distance for a given sample set. To date, comprehensive empirical comparisons between microsatellite and SNP markers in a phylogeographic context are lacking (but see Bradbury *et al.* 2015).

For Review Only

In the present study, we use a combination of mitochondrial DNA (mtDNA), microsatellites and genome-wide SNPs obtained from RADseq in order to: 1) produce a comprehensive phylogeography for *C. carassius* as a basis for Europe-wide conservation strategies, 2) test competing scenarios of postglacial recolonisation that have potentially contributed to the contemporary distribution of the species, and 3) compare the power of microsatellites and RADseq based population structure analyses, in the context of the first two objectives. In this third aim, we specifically ask, whether the benefits gained by the high numbers of markers obtained from RADseq outweigh the potential loss of power associated by the reduction in the number of samples in our system.

Materials and Methods

Sample collection and DNA extraction

C. carassius is a Cyprinid native to much of continental Europe; latitudinally from the North Sea and Baltic Sea basins, through central Europe north of the Alps down to the Ponto-Caspian region and longitudinally from Belgium and perhaps northern France into Siberia (Lelek 1980). However, the true extent of this native range is unknown, largely due to difficulties in morphologically distinguishing it from three closely related, introduced and widespread species: *Carassius auratus*, *Carassius gibelio*, and *Cyprinus carpio* (Wheeler 2000; Hickley & Chare 2004). We initially collected 1354 samples from 72 populations across 13 European countries, but due to frequent hybridisation between the *C. carassius* and the three species mentioned above, it was necessary to identify and remove hybrids from this sample set. To this end, all samples were first genotyped at 6 species diagnostic microsatellite loci. We removed all samples identified as hybrids from the dataset and, to safeguard against cryptic hybridisation, we also removed all *C. carassius* that were sympatric with hybrids (see SI text for full details of species identification and hybrid detection). This left 867 *C. carassius* samples from 57 populations across the species' distribution in central and northern Europe (Table 1, Fig. 1). Sample sizes ranged from n=4 to n=37, with a mean of n=17 (Table 1). Fish were anaesthetised by a UK Home Office (UKHO) personal license holder (GHC) in a 1 mL L^{-1} bath of 2-phenoxyethanol prior to collection of a 1 cm^{2} tissue sample from the lower-caudal fin, and wounds were treated with a mixture of adhesive powder (Orahesive) and antibiotic (Cicatrin) (Moore *et al.* 1990). Tissue samples were immediately placed in ≥95% ethanol, and stored at -20^{o}C. DNA was extracted from 2–4 mm^{2} of each tissue sample using either the Gentra Puregene DNA isolation kit or the DNeasy DNA purification kit (both Qiagen, Hilden, Germany). For the RADseq library, DNA was quantified using the Quant-iT™ PicoGreen® dsDNA Assay kit (Invitrogen) and normalised to concentrations ≥50 ng ml^{-1}. Gel electrophoresis was then used to check that DNA extractions contained high molecular weight DNA.

Molecular markers and methods

Three types of molecular markers were used in this study. Mitochondrial DNA sequencing was used to identify highly distinct lineages and to date the divergence between them through phylogenetic analysis. Two sets of nuclear markers; microsatellites and RADseq-derived SNPs, were used to investigate more recent and complex structure in a population genetics framework and to compare the relative power of each marker to do so.

For Review Only

Mitochondrial DNA amplification

A total of 83 *C. carassius* individuals, randomly chosen from a subset of 30 populations, which were chosen to represent all major catchment areas and the widest possible geographic range (min. n = 1, max. n = 4, mean n = 2.7), were sequenced at the cytochrome b (*cyt*b) gene (Table 1). PCR reactions were carried out following the protocol in Takada *et al.* (2010) using the forward and reverse primers L14736-Glu and H15923-Thru on an Applied Biosciences® Veriti Thermal Cycler. PCR products were sequenced in both directions on an ABI3700 by Macrogen Europe. The forward and reverse *cyt*b sequence reads were aligned using a GenBank sequence from the UK (accession no. JN412539, Table 1) as a reference and ambiguous nucleotides were manually edited using CodonCode aligner v.2.0.6 (CodonCode Corporation).

Microsatellite amplification

Of the 867 samples identified as pure *C. carassius,* 19 samples were in populations with sample numbers which were too low to be useful for population genetics analyses (< 4). The remaining 848 samples, from 49 populations, were genotyped at 13 microsatellite loci, including the six species diagnostic loci used for hybrid identification (Supporting Information (SI) Table 1). Microsatellites were amplified in three multiplex PCR reactions, using the Qiagen multiplex PCR mix with manufacturer's recommended reagent concentrations, including Q solution and 1 µl of template DNA. Primer concentrations for each locus are provided in SI Table 1 and PCRs were performed on an Applied Biosciences® Veriti Thermal Cycler. The annealing temperature used was 54°C for all reactions, and all other PCR cycling parameters were set to Qiagen multiplex kit recommended values. PCR products were run on a Beckman Coulter CEQ 8000 genome analyser

For Review Only

using a 400 bp size standard and microsatellite alleles scored using the Beckman Coulter CEQ8000 software.

RADseq

A total of 160 individuals (18 populations, min. *n* = 8, max. *n* = 10, mean *n* = 8.9), identified as pure *C. carassius* with the diagnostic microsatellites, were used in the RADseq (Table 1). These samples were chosen to represent a wide geographic range and all major phylogeographic clusters identified using the microsatellite data. These samples were split across 13 libraries prepared at Edinburgh Genomics (University of Edinburgh, UK) according to the protocol in Davey *et al.* (2012) using the enzyme *Sbf*1. Libraries were then sequenced using paired-end sequencing across five lanes of two Illumina HiSeq 2000 flowcells (Edinburgh Genomics).

Data analyses

Phylogenetic analysis of mtDNA

In addition to the 83 sequenced samples (SI Table 2), we retrieved 19 published *C. carassius* and three *C. carpio cyt*b sequences from GenBank to be used as an outgroup. The *C. carpio* samples were chosen to include samples from multiple, distant lineages of *C. carpio* located in Japan, Greece and India. All sequences used were validated through cross checking with their original publications (Table 1). Sequence alignment was performed in MEGA6 (Tamura *et al.* 2013) using default settings, and DNAsp v.5.0 (Librado & Rozas 2009) was used to calculate sequence divergence and to identify haplotypes.

Haplotypes of all *C. carassius* samples and the three *C. carpio* outgroup individuals were exported to BEAST v.1.7.5 (Drummond *et al.* 2012) for phylogenetic analyses in order to identify the major

For Review Only

phylogenetic lineages within European *C. carassius*. Phylogenetic model testing with jModeltest2 v.2.1.7 (Guindon *et al.* 2003; Darriba *et al.* 2012) using Akaike information criterion (AIC), Bayesian Information Criteria (BIC) and the decision-theoretic performance-based (DT) approach showed that HKY (Hasegawa *et al.* 1985) was the most appropriate substitution model for our dataset. Using this model, the splits between the major phylogenetic clades were then dated using a relaxed molecular clock method in BEAST. The widely-used Dowling *et al.* (2002) cyprinid *cyt*b divergence rate of 1.05% pairwise sequence divergence / MY was used after converting to a per lineage value of 0.0053 mutations/site/MY for use in BEAST. We used a 'coalescent: constant size' tree prior, which assumes an unknown but constant population size backwards in time, as recommended for intraspecific phylogenies (Drummond *et al.* 2012) . MCMC chain lengths were 1 x 10^7 with samples taken every 1000 iterations. A gamma site heterogeneity model was used, with the default of four categories. Substitution rates, rate heterogeneity and base frequencies were unlinked between each codon position to allow substitution rate to vary between them. Default values were used for all other parameters and priors.

For Review Only

Population structure and diversity analyses using microsatellites

Allele dropout and null alleles in the microsatellite data were tested using Microchecker (Van Oosterhout *et al.* 2004). FSTAT v. 2.9.3.2 (Goudet 1995) was then used to check for linkage disequilibrium (LD) between loci (using 10,000 permutations), deviations from Hardy-Weinberg equilibrium (HWE) within populations (126500 permutations) and for all population genetic summary statistics. Genetic diversity within populations was estimated using Nei's estimator of gene diversity (H_o) (Nei 1987) and Allelic richness (A_r), which was standardised to the smallest sample size (n = 4) using rarefaction (Petit *et al.* 1998). Pairwise F_{ST} values were calculated according to (Weir & Cockerham 1984) and 23520 permutations and sequential Bonferroni correction were used to test for significance of F_{ST}.

IBD was investigated using a Mantel test in the adegenet v1.6 (Jombart & Ahmed 2011) package in R v3.0.1 (R Core Team 2013). We then tested for an association between A_r and longitude and latitude, which is predicted under a stepping-stone colonisation model (Ramachandran *et al.* 2005; Simon *et al.* 2014), using linear regression analysis in R.

Population structure was then further examined using Discriminant Analyses of Principal Components (DAPC) also in adegenet (DAPC, see SI text and Jombart *et al.* 2010 for more details). DAPC has been shown to perform as well or better than the commonly used program, STRUCTURE (Pritchard *et al.* 2000) for both simple and complex models of population structure (Jombart *et al.* 2010). Furthermore, unlike STRUCTURE, DAPC is free of underlying assumptions of Hardy-Weinberg equilibrium, which are likely to be violated when effective population sizes are small, as is often the case in *C. carassius* (Hänfling *et al.* 2005).

For Review Only

In preliminary DAPC analysis using all 49 *C. carassius* populations, Sweden (SWE9) was found to be so genetically distinct from the rest of the data set that it masked the variation between the other populations. This population was therefore omitted from further DAPC analyses. To infer the appropriate number of genetic clusters in the data, we used BIC scores (SI Fig. 5a), in all cases choosing lowest number of genetic clusters from the range suggested. Spline interpolation (Hazewinkel 1994) was then used to identify the appropriate number of principal components to use in the subsequent discriminant analysis (SI Fig. 5a).

RADseq data filtering and population structure analysis

The quality of the RADseq raw read data was examined using FastQC (Andrews 2010), the dataset was then cleaned, processed and SNPs were called using the Stacks pipeline v 1.19 (Catchen *et al.* 2011). Preliminary tests were carried out in order to identify optimal Stacks parameters (See SI text). Final parameter values for the respective Stacks module were as follows; ustacks: M=2, m=8, removal (-r) and deleveraging (-d) algorithms were also used; cstacks: N=2 (n populations = 18, n individuals = 160); populations module: one SNP per RAD locus was used (--write_single_snp) and SNPs were only retained if they were present in 70% of individuals (r=0.7) in at least 17 out of the 18 populations in the study (p=17), which allows for mutations in restriction sites that may cause loci to dropout in certain lineages. All other parameters were kept at default values. Finally, we filtered out loci which had a heterozygosity of > 0.5 and $F_{\mathrm{IS}} < 0.0$ in one or more populations in order to control for the possibility of erroneously merging ohnologs resulting from the multiple genome duplications that have occurred the in *Cyprinus* and *Carassius* genera (Henkel *et al.* 2012; Xu *et al.* 2014). The resulting refined SNP set was then used in subsequent phylogeographic analyses. The R package Adegenet v. 1.42 was used to calculate H_{o} and pairwise F_{ST}, test for IBD and genetic clusters were inferred using DAPC.

Reconstructing postglacial colonisation routes in Europe

DIYABC v. 2.0 (Windows, Cornuet *et al.* 2014) was used to reconstruct the most likely *C. carassius* recolonisation routes through Europe after the last glacial maximum. We used the RADseq data set for this analysis as it showed a much clearer pattern of population structure than the microsatellite data in DAPC analyses (see Results). Furthermore, preliminary DIYABC analyses using microsatellites failed to identify a scenario which was significantly more likely than its counterparts, suggesting low power in this dataset for reconstructing complex phylogeographic patterns over long timescales.

As DIYABC is a computationally intensive method, it was necessary to perform analyses on a subset of 1000 randomly-selected SNP loci from the full RADseq dataset to reduce computation time. This SNP subset was first analysed with DAPC to confirm that it produced the same population structure as the full dataset and was then used to compare the likelihood of a number of user defined colonisation scenarios (i.e. a specific population tree topology, together with the

For Review Only

parameter prior distributions that are associated with it). First, 1 million datasets were simulated for each scenario. These simulated summary statistic datasets represented the theoretical expectation under each scenario, and were compared to the same summary statistics calculated from the observed data, in order to identify the most likely of the tested scenarios. In DIYABC, two methods of comparison between simulated and observed datasets are used; logistic regression and "direct approach", the latter method identifies the scenario that produces the largest proportion of the *n* number of closest scenarios to the observed, where *n* is specified by the user. The goodness-of-fit of scenarios was also assessed using the model checking function implemented in DIYABC (Cornuet *et al.* 2014). In all analyses, the single-sample summary statistics used were the mean and variance of gene diversity across all polymorphic loci and the mean gene diversity across all loci. The two-sample summary statistics used were the mean and variance of F_{ST} and Nei's distance for loci with F_{ST} greater than zero between two samples and the mean F_{ST} and Nei's distance for all loci. Finally, for scenarios including admixture events, the maximum likelihood estimates of admixture proportions were also used. See Cornuet *et al.* (2014) for the exact equations used and their implementation in DIYABC.

For Review Only

To reduce the number and complexity of possible scenarios, we split DIYABC analysis into three stages (Table 2). In stage 1, we tested 11 broad scale scenarios (Scenarios 1 -11, SI Fig. 1). Populations were grouped into three pools in order to reduce the number and complexity of possible scenarios (Table 2); Pool 1 – all northern European populations (npops = 17, *n* = 155), Pool 2 – Don population (npops = 1, *n* = 9), Pool 3 – Danubian population (npops = 1, *n* = 6). In six scenarios (1, 2, 8-11), northern European and the Don population diverged from each other more recently than from Danubian populations. These scenarios differ in the patterns of effective population size change and the presence or absence of a bottleneck. In scenarios 3 and 4, northern European and Danubian populations are more closely related to each other than to the Don population. And in the remaining three scenarios, one pool of populations is the product of an admixture event between the other two. Population poolings and scenarios were both chosen on the basis of the broad phylogeographic structure identified in the mtDNA and RADseq population structure analysis (see Results).

In the second and third stages, we performed a finer scale analysis, focussing on the 17 northern European populations alone. Populations were again pooled, this time into six groups, on the basis of both population structure and geography (Table 2). In stage 2 we tested five scenarios (Scenarios 12-16, see SI Fig. 2a for graphical description of each scenario), with no bottlenecks included, which represented the major topological variants that were most likely, given population structure results from DAPC. We then identified the most likely of these scenarios in DIYABC and took this forward into the final stage of the analysis where we tested 6 multiple bottleneck combinations (SI Fig. 2b) around this scenario. This three stage approach allowed us to systematically build a complex scenario for the European colonisation of *C. carassius*. Finally, we used the posterior distributions of the time parameters, simulated using the scenario identified as most likely in stages one and three, to estimate the times of the major lineage splits in European *C. carassius*. These parameters, calculated by DIYABC in generations, were converted to years using an average generation time of 2 years (Tarkan *et al.* 2010).

For Review Only

Comparison of microsatellite and RADseq data

Finally, we compared the results derived from population structure analyses on microsatellite and RADseq data to assess their suitability for addressing our phylogeographic question. It is important to note that differences between the full microsatellite and RADseq datasets could be attributable to one or a combination of the following; the number of populations, the geographic distribution of populations, the number of samples per population, the number of markers, or the information content of the marker type. To disentangle these sources of variation, we created two microsatellite data subsets; M2, which included only individuals used in RADseq, (excluding three individuals for which microsatellite data was incomplete, *n* = 146, npops = 19), and M3, which contained all individuals for which microsatellite data was available in populations that were used in RADseq (*n* = 313, npops = 19;

For Review Only

Table 3). This gave us three pairs of datasets for comparison: 1) RADseq Vs. M2: same individuals but different marker types, 2) M1 vs M2: full microsatellite dataset versus a subset of the populations, and 3) M2 vs M3: same populations but different number of individuals per population. This strategy enabled us to test for the influence of marker, sampling of populations and individuals per population respectively. Comparisons were performed between datasets on heterozygosities and pairwise F_{ST}s using both Pearson's product-moment correlation coefficient and paired Student's t-tests in R. IBD results were compared using Mantel tests (Jombart & Ahmed 2011), and DAPC results were compared on the basis of similarity of number of inferred clusters and cluster sharing between populations.

For Review Only

Results

Phylogenetic analyses of mitochondrial data

The combined 1090 bp alignment of 100 *cyt*b *C. carassius* mtDNA sequences yielded 22 haplotypes, which were split across two well supported and highly differentiated phylogenetic lineages (Fig. 2, SI Table 3). Lineage 1 was found in all northern European river catchments sampled, as well as eastern European (Dnieper) and southeastern European (Don and Volga) catchments, whereas Lineage 2 was almost exclusively confined to the River Danube catchment. There were, however, a few exceptions to this clear geographical split; two individuals, one from the Elbe and one from the Rhine in northern Germany, belonged to mtDNA Lineage 2, as did one individual from the River Lahn river catchment in western Germany. Also one population in the Czech Republic, located on the border between the Danube and Rhine river catchments, was found to contain individuals belonging to lineages 1 and 2.

The mean number of nucleotide differences within lineages 1 and 2 was 2.25 and 2.00, respectively, which equated to a sequence divergence 0.2% and 0.18%, respectively. Between the two lineages there was an average of 22.5 nucleotide differences (2.06% mean sequence divergence), with 19 of these being fixed. BEAST molecular clock analysis dated the split between lineages 1 and 2 to be 1.30–3.22 million years ago (MYA), with a median estimate of 2.15 MYA (Fig. 2).

Nuclear marker datasets and quality checking

Microchecker showed no consistent signs of null alleles or allele dropout in microsatellite loci and no significant LD was found between any pairs of loci. No populations showed significant deviation from Hardy-Weinberg proportions (adjusted nominal level 0.0009).

After filtering raw RADseq data, *de novo* construction of loci across the 19 populations produced 35 709 RADseq loci that were present in at least 70% of individuals in at least 17 populations. These loci contained a total of 29 927 polymorphic SNPs (approx. 0.84 SNPs per locus). Only the first SNP in each RADseq locus was retained, to avoid confounding signals of LD. This yielded a total of 18 908 loci with a mean coverage of 29.07 reads (SI Fig. 3b). Finally 5719 of these SNP loci were filtered out due to high (> 0.5) heterozygosity and/or F_{IS} of < 0.0 in at least one population. In doing so, we removed many high coverage tags (SI Fig. 3a), which was consistent with over-merged ohnologs having higher coverage (*i.e.* reads from more than two alleles) than correctly assembled loci. The final dataset therefore contained 13189 SNP loci, with a mean coverage of 27.72 reads.

Within population diversity at nuclear loci

Observed heterozygosity (H_o), averaged across all microsatellite loci within a population, ranged from 0.06 (SWE9) to 0.44 (BLS), with a mean of 0.25 across all populations (SD = 0.105), and was highly correlated with A_r (t= 19.67, P < 0.001, df = 40), which ranged from 1.26 (FIN1) to 2.96

For Review Only

(POL3) with a mean of 1.92 (SD = 0.51). Mean H_o averaged across all RADseq loci for all populations was 0.013 (SD = 0.013), ranged from 0.001 to 0.057 and was significantly correlated with H_o from microsatellite loci at populations shared between both datasets (r = 0.69, t= 3.74, P = 0.002, df = 15). Microsatellite A_r significantly decreased along an east to west longitudinal gradient (adj. R^2 = 0.289, P < 0.001, SI Fig. 4b) consistent with decreasing diversity along colonisation routes. However, A_r did not decrease with increasing latitude (adj R^2=-0.007, P = 0.414, SI Fig. 4a). We also repeated this analysis after removing samples from mtDNA Lineage 2 in the Danube catchment. Again there was no relationship between A_r and latitude (R^2=-0.023, P = 0.254, SI Fig. 4c), but the relationship between A_r and longitude was strengthened (adj. R^2 = 0.316, P < 0.001, SI Fig. 4d).

Population Structure in Europe based on nuclear markers

Population structure was strong, as predicted. Using the full (M1) microsatellite dataset, mean pairwise F_{ST} was 0.413 (min = 0.0; BEL2 and BEL3), max = 0.864 (NOR2 vs GBR2), with 861 of the 1128 pairwise population comparisons being significant F_{ST} (P < 0.05, SI Table 4). Pairwise F_{ST} calculated from the RADseq dataset also showed strong structure (SI Table 5), ranging from 0.067 (DEN1, DEN2) to 0.699 (NOR2, GBR4), and these values were highly correlated with the same population comparisons in the M3 microsatellite dataset (r = 0.66, t = 9.01, P < 0.01, df = 104).

BIC scores obtained from initial DAPC analyses of the microsatellite dataset, using all 49 populations, indicated that between 11 and 19 genetic clusters (SI Fig. 5a) would be an appropriate model of the variation in the data. As a conservative estimate of population structure, we chose 11 clusters for use in the discriminant analysis, retaining eight principal components as recommended by the spline interpolation a-scores (SI Fig. 5a). This initial analysis showed that populations belonging to Cluster 10 (RUS1, Don river catchment) and Cluster 11 (GER3, GER4, CZE1,

For Review Only

Danubian catchment) were highly distinct from clusters found in northern Europe (Fig. 1b). Since the marked genetic differentiation between these three main clusters masked the more subtle population structure among northern European populations (see Fig. 1b), we repeated the DAPC analysis without the populations from the Danube and Don (RUS1, GER3, GER4, CZE1, Fig. 1b). The results of this second DAPC analysis revealed an IBD pattern of population structure, across Europe (Fig. 1). Mantel tests excluding the Danubian and Don populations corroborated these results; showing significant correlation with geographic distance in northern Europe (adjusted $R^2 =$ 0.287, $P < 0.001$, SI Fig. 6a), with Danubian populations shown to be more diverged than their geography would predict (data not shown).

For Review Only

In the RADseq DAPC analysis, BIC scores suggested between four and ten genetic clusters, a lower number than that inferred from the microsatellite data set. Again we chose the lowest number of suggested clusters (four) clusters to take forward in the analysis (SI Fig. 5b). Following spline interpolation, we retained six principal components and kept two of the linear discriminants from the subsequent discriminant analysis (SI Fig. 5b). The inferred population structure showed that the Danubian population (HUN2) and the Don population (RUS1) were highly diverged from the northern European clusters. Unfortunately, HUN2 is not present in the microsatellite dataset for direct comparison, however both datasets, and the mtDNA data show the same pattern of high divergence between northern Europe and Danubian populations. DAPC analyses of RADseq data again showed an IBD pattern in northern European populations, which was confirmed with Mantel tests when the Danubian population HUN2 was excluded (adjusted $R^2 = 0.722$, $P < 0.001$; SI Fig. 6b).

Postglacial recolonisation of C. carassius in Europe

DAPC results of the 1000 SNP RADseq dataset used in DIYABC showed that it produced the same population structure as the full RADseq dataset (SI Fig. 7). For the broad-scale scenario tests in stage one of the DIYABC analysis, both logistic regression and direct approach identified Scenario 9 as being most likely to describe the true broad-scale demographic history (SI Fig. 8). Model checking showed that the observed summary statistics for our data fell well within those of the posterior parameter distributions for scenario 9 (SI Fig. 8c). Scenario 9 agrees with the mtDNA results, suggesting that the Danubian populations have made no major contribution to the colonisation of northern Europe. The median posterior distribution estimate of the divergence time between Danubian and northern European populations is 2.18 MYA (95% CI = 1.03 – 5.12 MYA), assuming a two-year generation time (Tarkan *et al.* 2010)), which is strikingly similar to that of mtDNA dating analysis. Scenario 9 also suggests that the northern European populations experienced a population size decline after the split of Pool 1 from the population in the Don river catchment, which lasted approximately 8920 years (95% CI = 616 – 13700 years) and reduced N_e by 32%.

For Review Only

In stage two of the DIYABC analysis, we tested the major variant scenarios for the colonisation of northern Europe. In assessing the relative probabilities of scenarios, there was some discrepancy between the direct approach, which revealed Scenario 14 to be most likely, and the logistic regression, which favoured Scenario 13 (with Scenario 14 being the second most likely). However, the goodness-of-fit model checking showed that the observed dataset fell well within the posterior parameter distributions for Scenario 14 (SI Fig. 9a), but not for Scenario 13 (not shown). Therefore, Scenario 14 was carried forward into stage three in which we tested six more scenarios (SI Fig. 2b) to compare combinations of bottlenecks using the same population tree topology as in Scenario 14. Direct approach, logistic regression and model checking all found scenario 14d to be the most likely (SI Fig. 9b), we therefore accepted this as the scenario for the colonisation of *C. carassius* in northern Europe (SI Fig. 9b). This scenario infers an initial split between two sub-lineages in northern Europe approximately 33 600 YBP (Fig. 4), one of which re-colonised northwest Europe and one that re-colonised Finland through the Ukraine and Belarus. Scenario 14d also inferred a secondary contact between these sub-lineages approximately 15 940 YBP, resulting in the populations currently present in Poland; these admixed populations provided the source of one colonisation across the Baltic into Sweden, and a second route was inferred into southern Sweden from Denmark (Table 3, SI Fig. 9b).

For Review Only

Comparing microsatellite datasets and RAD sequencing data

The results from the RADseq (n = 149, npops = 16) dataset and the full microsatellite dataset (M1, n = 848, npops = 49) largely agreed on the inferred structure and cluster identity of populations. However, there were some important differences between them. Firstly, the IBD pattern of population structure in northern Europe was much stronger in the RADseq data ($R^2 = 0.722$, $P < 0.001$, SI Fig. 6) compared to the M1 dataset ($R^2 = 0.287$, $P < 0.001$, excluding Danubian populations and SWE9 from both datasets, SI Fig. 6). Secondly, clusters inferred by the RADseq DAPC analysis are much more distinct, *i.e.* there is much lower within-cluster, and higher between-cluster variation in the RADseq results than in the M1 dataset results (Fig. 3).

As the properties of the RADseq and M1 datasets differ in four respects, namely marker type, number of populations, number of samples per population (Table 3) and uniformity of sampling locations, (SI Fig. 10) it was not possible to identify the cause of discrepancies in their results. Therefore, below we report the results from the pair-wise dataset comparisons, which isolate the effects of these parameter differences.

1) *M1 Vs. M3:* the effect that the number of populations and the uniformity of sampling locations might have on inferred population structure. The geographic distribution of sampling locations was more clustered in M1 (full microsatellite dataset) than in M3 (containing microsatellite for samples in populations used in RADseq (SI Fig. 10), and IBD patterns were considerably stronger in the M3 subset (adj. R^2= 0.447, $P < 0.001$) than in the full M1 dataset (adj. R^2= 0.287, $P < 0.001$). In contrast DAPC results were very similar between datasets, with inferred cluster number, structure and population identity of clusters generally agreeing well (Fig. 1, Fig. 3c).

2) M2 Vs. M3: the effect of reducing the number of samples per population on the inferred population structure. The number of samples per population in the M2 subset (microsatellite data only for the samples used in RADseq, mean = 9.125 ± 0.8) was significantly lower than that of the M3 subset (mean, 19.6 ± 9.0, t = -4.66, df = 15, $P < 0.001$), as was the number of alleles per population (M2 mean = 24.4 ± 7.3, M3 mean = 27.4 ± 8.1, t = -5.72, df = 15*,* $P < 0.001$). Population heterozygosities were significantly different between M2 and M3 (M2 mean = 0.21, M3 mean = 0.23, t = -2.4, df = 15, P = 0.012), but highly correlated (r = 0.94, t = -11.13, $P < 0.001$, df = 15). Pairwise F_{ST}s were very strongly correlated (r = 0.97, t = 46.26, $P < 0.001$, df = 105), but again, still significantly different between the two datasets (M2 mean = 0.46, M3 mean = 0.49 , t = -6.21, $P <$ 0.001, df = 15, Table 4). The patterns of IBD were almost identical for M2 (R^2 = 0.455, $P < 0.001$) and M3 (R^2 = 0.447, $P < 0.001$, SI Fig. 6) and population structure inferred by DAPC was again similar. BIC scores suggested a similar range of cluster number for M2 and M3, the smallest of which was nine in both cases.

3) RADseq Vs. M3: The effect of the number and the type of markers used on the phylogeographic results. We compared the results from the RADseq and M2 datasets, which contain exactly the same samples (with the exception of three individuals missing in M2). Significant correlations were again found between heterozygosities estimated for the two datasets (r = 0.69, t = 3.73, P = 0.002,

For Review Only

df = 15) and pair-wise F_{ST}s (r = 0.70, *t* = 10.09, *P* < 0.001, df = 105), but RADseq data yielded much lower pairwise F_{ST}s (mean RAD = 0.29, mean M2 = 0.46, *t* = 13.74, *P* < 0.001, df = 15). DAPC analysis of RADseq data resolved populations into much more distinct clusters (Figs. 3a, 3b), and the IBD pattern found was considerably stronger in the RADseq (R^2 = 0.722, *P* < 0.001) dataset compared to M2 (R^2 = 0.455, *P* < 0.001, SI Fig. 6).

Discussion

For Review Only

In this study, we aimed to simultaneously produce a phylogeographic framework on which to base conservation strategies for *C. carassius* in Europe, and compare the relative suitability of genome-wide SNP markers and microsatellite markers for such an undertaking. Through comparison of the inferred population structure from microsatellite and genome-wide SNP data, we show that there are important differences in the results from each data type, attributable predominantly to marker type, rather than within population sampling or spatial distribution of samples. However, despite these differences, all three data types used (mitochondrial, microsatellite and SNP data) agree that, unlike many other European freshwater fish for which phylogeographic data is available, *C. carassius* has not been able to cross the Danubian catchment boundary into northern Europe. This has resulted in two, previously unknown, major lineages of *C. carassius* in Europe, which we argue should be considered as separate conservation units.

Phylogeography and postglacial recolonisation of C. carassius in Europe

The most consistent result across all three marker types (mtDNA sequences, microsatellites and RADseq) was the identification of two highly-divergent lineages of *C. carassius* in Europe. The distinct geographic distribution of these lineages; Lineage 1 being widely distributed across north and eastern Europe and Lineage 2 generally only in the River Danube catchment, indicates a longstanding barrier to gene flow between these geographic regions. Bayesian inference based on mtDNA phylogeny and ABC analysis of RADseq data showed remarkable agreement, estimating that these lineages have been isolated for 2.15 MYA (95% CI = 1.30–3.22) and 2.18 (95% CI = 2 – 6.12) MYA respectively, which firmly places the event at the beginning of the Pleistocene (2.6 MYA; (Gibbard & Head 2009). This pattern differs substantially from the general phylogeographic patterns observed in other European freshwater fish. Indeed, previous studies have shown that the Danube catchment has been an important source for the postglacial recolonisation of freshwater fish into northern Europe or during earlier interglacials in the last 0.5 MYA. For example, bullhead *Cottus gobio* (Hänfling & Brandl 1998; Hänfling *et al.* 2002), chub *Leuciscus cephalus* (Durand *et al.* 1999), Eurasian perch *Perca fluviatilis* (Nesbø *et al.* 1999), riffle minnow *Leuciscus souffia* (Salzburger *et al.* 2003), grayling *Thymallus thymallus* (Gum *et al.* 2009), European barbel *Barbus barbus* (Kotlík & Berrebi 2001), and roach *Rutilus rutilus* (Larmuseau *et al.* 2009) all crossed the Danube catchment boundary into northern drainages such as those of the rivers Rhine, Rhône and Elbe during the mid-to-late Pleistocene. The above species occur in lotic habitats, and most are capable of relatively high dispersal. In contrast *C. carassius* has a very low propensity for dispersal, and a strict preference for the lentic backwaters, isolated ponds and small lakes (Holopainen *et al.* 1997; Culling *et al.* 2006; Copp 1991). We therefore hypothesise that these ecological characteristics of *C. carassius* have reduced its ability to traverse the upper Danubian watershed, which lies in a region characterised by the Carpathian Mountains and the Central European Highlands. This region may have acted as a barrier to the colonisation of *C. carassius* into northern European drainages during the Pleistocene. It should be noted, however, that the phylogeography of two species, the spined loach *Cobitis taenia* and European weatherfish *Misgurnus fossilus*, does not support this hypothesis as a general pattern for floodplain species (Janko *et al.* 2005; Culling *et al.* 2006). The former is the only species that we know of other than *C. carassius* showing long-term isolation between the Danube and northern European catchments, but has lotic habitat preferences and good dispersal abilities (Janko *et al.* 2005; Culling *et al.* 2006), whereas the latter inhabits

For Review Only

similar ecosystems as *C. carassius*, with low dispersal potential, but has colonised northern Europe from the Danube catchment (Bohlen *et al.* 2006, 2007).

There is one notable exception to the strict separation between Danubian and northern European *C. carassius* populations. The population CZE1, located in the River Lužnice catchment (Czech Republic), which drains into the River Elbe, clusters with Danubian populations in both the microsatellite and mtDNA data. This sample site, from the River Lužnice, is very close to the Danubian catchment boundary and is situated in a relatively low lying area. Therefore, some recent natural movements across the watershed between these river catchments, either through river capture events or ephemeral connections, could have been possible. A similar pattern has been shown in some European bullhead *Cottus gobio* populations along the catchment Danube/Rhine catchment border (Riffel & Schreiber 1995). We also observed the presence of two mtDNA haplotypes from Lineage 2 in some individuals from northern German populations (GER1, GER2, GER8), however, one of these haplotypes was shared with Danubian individuals and the results were not confirmed by nuclear markers. Overall this is most likely to be the result of occasional human mediated long-distance dispersal for the purposes of intentional stocking.

Population structure within Lineage 1 is characterised by a pattern of IBD and a loss of allelic richness from eastern to western Europe. This is consistent with the most likely colonisation scenario identified by the DIYABC analysis, indicating a general southeast to northwest expansion from the Ponto-Caspian region towards central and northern Europe (Fig. 4). The Ponto-Caspian region, and in particular the Black Sea basin, was an important refugium for freshwater fishes during the Pleistocene glacial cycles, and a similar colonisation route has been inferred for many other freshwater species in northern Europe (Nesbø *et al.* 1999; Durand *et al.* 1999; Culling *et al.* 2006; Costedoat & Gilles 2009). The DIYABC analysis also suggests that there was an interval of > 200 000 years between the split of the Don population (≈ 270 000 years ago) and the next split in

For Review Only

the scenario (approx. 33 600 years ago), which marks the main expansion across central and northern Europe. It appears that no further population divergence can be dated back to the time interval between the Riss/Saalian and the Würm/Weichelian glacial periods. This may be because the range of *C. carassius* has not undergone a major change during that time interval, but it is more likely that the signal of expansion during the Riss-Würm interglacial has been eradicated through a subsequent range contraction during the Würm/Weichelian glacial period. The model also suggests that the Würm/Weichelian period was accompanied by a sustained but moderate reduction in population size over almost 9000 years (Bottleneck A, Fig. 4), which may reflect general population size reductions during the Riss glaciations or a series of shorter bottlenecks during subsequent range expansion (Ramachandran et al. 2005, Simon et al 2015, Hewitt 2000).

DIYABC analyses inferred the colonisation of northern Europe by two sub-lineages within the mtDNA Lineage 1, which were isolated from each other approximately 33 600 years ago. These sub-lineages may reflect two glacial refugia resulting from the expansion of the Weichselian ice cap to its maximum extent roughly 22 000 years ago (see hypothetical refugia II and III in Fig. 4). The western sub-lineage underwent a second long period of population decline (Bottleneck B, Fig. 4), which may again represent successive founder effects during range expansion. There is then evidence of secondary contact between these sub-lineages (node b, approximately ≈ 15 940 years ago), contributing to the genetic variation now found in Poland. This inferred admixture event may represent one of the numerous inundation and drainage capture events, which resulted from the melting of the Weichselian ice cap, that are known to have occurred around this time (Grosswald 1980; Gibbard *et al.* 1988; Arkhipov *et al.* 1995). However, as the colonisation of Europe was likely to have occurred via the expansion of colonisation fronts (*i.e.* dashed contour lines in Fig. 4), rather than along linear paths, it could also be indicative of the known IBD gradient between the inferred western and eastern sub-lineages. Such a gradient (eg. between northwestern and

For Review Only

northeastern Europe) may give false signals of admixture between intermediate populations, such as those in Poland.

The colonisation of the Baltic sea basin also seems to have been complex, with three independent routes inferred by DIYABC scenario 14d; one recent route through Denmark into southern Sweden, one to the east of the Baltic Sea, through Finland, and one across the Baltic Sea, from populations related to those in Poland (Pool 4). The first of these agrees well with the findings of Janson *et al.* (2014), whereby populations, including SWE8 from our study (SK3P in Janson *et al.* 2014), in this region were found to be distinct from those in central Sweden. The eastern route shows similarities to the colonisation patterns of *P. fluvilatilis*, which is hypothesised to have had a refugium east of Finland (Nesbø *et al.* 1999) during the most recent glacial period. This is certainly also plausible in *C. carassius* and may account for the distinctiveness of Finnish populations seen in microsatellites and RADseq DAPC analysis. The last colonisation route, across the Baltic Sea from mainland Europe, may have coincided with the freshwater Lake Ancylus stage of the Baltic Sea's evolution, which existed from ≈ 10 600 to 7 500 years ago (Björck 1995; Kostecki 2014). The Lake Ancylus stage likely provided a window for the colonisation of many of the species now resident in the Baltic, and has been proposed as a possible window for the colonisation of *T. thymallus* (Koskinen *et al*. 2000), *C. taenia*, (Culling *et al.* 2006), *C. gobio* (Kontula & Väinölä 2001) and four *Coregonus* species (Svärdson 1998). Consistent with this, we found strong similarity between populations from Fasta Åland, southern Finland and central Sweden, suggesting that shallow regions in the central part of Lake Ancylus (what is now the Åland Archipelago), may have provided one route across Lake Ancylus.

It is also likely that the contemporary distribution of *C. carassius* in the Baltic has been influenced by human translocations. *C. carassius* were often used as a food source in monasteries in many parts of Sweden (Janson *et al.* 2014), and the Baltic island of Gotland (Rasmussen 1959; Svanberg *et al.* 2013) was an important trading port of the Hanseatic League – a commercial confederation

For Review Only

that dominated trade in northern Europe from the 13th to 17th centuries. Previous data suggest that *C. carassius* was transported from the Scania Province, southern Sweden, where *C. carassius* aquaculture was common at least during the 17th century, to parts further north (Svanberg *et al.* 2013; Janson *et al.* 2014).

Implications for the conservation of C. carassius *in Europe*

The two *C. carassius* lineages exhibit highly-restricted gene flow between them and are the highest known organisational level within the species. They therefore meet the genetic criteria for Evolutionarily Significant Units (ESUs) as described in (Fraser & Bernatchez 2001). This is especially important in light of the current *C. carassius* decline in the Danubian catchment (Bănărescu 1990; Navodaru *et al.* 2002; Lusk *et al.* 2010; Savini *et al.* 2010). The conservation of *C. carassius* in central Europe must therefore take these catchment boundaries into consideration, as opposed to political boundaries. A first step would be to include *C. carassius* in Red Lists, not only for individual countries, but at the regional (e.g. European Red List of Freshwater Fishes; (Freyhof & Brooks 2011) and global (IUCN 2015) scales, and we hope that the evidence presented here will facilitate this process. Within the northern European lineage, the Baltic Sea basin shows high levels of population diversity, likely owing to its complex colonisation history. As such, the Baltic represents an important part of the *C. carassius* native range. Although *C. carassius* is not currently thought to be threatened in the Baltic region, *C. gibelio* is invading this region and is considered a threat (Urho & Lehtonen; Deinhardt 2013).

For Review Only

Microsatellites vs RADseq for phylogeography

Broad conclusions drawn from each of our RADseq-derived SNPs, full or partial microsatellite datasets are consistent, demonstrating deep divergence between northern and southern European populations and an IBD pattern of population structure in northern Europe. This similarity in spatial signal between marker types was also observed by (Bradbury *et al.* 2015). However, two striking differences exist in the phylogeographic results produced by RADseq compared to those of the microsatellite datasets. Firstly, the IBD pattern inferred from RADseq data was considerably stronger than for any of the microsatellite datasets. This effect was also found by Coates *et al.* (2009) when comparing SNPs and microsatellites, who postulated that it was driven by the differences in mutational processes of the markers. The second major difference between RADseq and microsatellite results was that clusters inferred by DAPC from the RADseq data were considerably more distinct compared to the full microsatellite dataset, emphasising the fine scale structure in the data (which is particularly apparent in the northern Finnish populations). We ruled out the possibility of these differences being caused by the reduction in number of populations, their spatial uniformity or number of individuals per population used in RADseq by creating two partial microsatellite datasets and comparing these to results from the RADseq-SNPs. Differences between marker types were consistently reproducible whether full or partial microsatellite datasets were used in the analyses.

It is also worth noting that the number of populations or the number of samples per population had no apparent impact on IBD and DAPC results between the microsatellite datasets. This is in contrast to predictions of patchy sampling of IBD made by Schwartz and McKelvey (2009), perhaps because of the strong population structure in *C. carassius,* and likelihood that a sufficiently informative number of populations was included even in the reduced datasets.

For Review Only

SNP loci provide several advantages over microsatellites additional to those highlighted here. SNPs are more densely and evenly distributed across the genome (Xing *et al.* 2005) and have been shown to display lower error rates during genotyping (Montgomery *et al.* 2005). For example, Morin *et al.* (2009a) showed that HW proportions are very sensitive to microsatellite genotyping errors. SNPs also lend themselves to a plethora of evolutionary applications, including the identification of outlier loci (Hohenlohe *et al.* 2012) or small regions of introgression in the genome (Hohenlohe *et al.* 2013). Lastly, SNPs are also much less susceptible to homoplasy than microsatellites (Morin *et al.* 2004). Van Oppen *et al*. (2000) found evidence of homoplasy in 10 out of 13 microsatellite loci, which had accumulated in approximately 700,000 years and Cornuet *et al.* (2010) show that such homoplasy makes microsatellites unreliable and error prone when used in DIYABC for inference over long time scales. For these reasons, SNPs have a clear advantage over microsatellites for the purposes of characterising population divergence over long time scales. This may explain why preliminary microsatellite analyses in DIYABC showed insufficient power to identify a most likely colonisation scenario.

For Review Only

Conclusions

We have identified the most likely routes of post-glacial colonisation in *C. carassius*, which deviate from the general patterns observed in other European freshwater fishes. This has resulted in two, previously-unidentified major lineages in Europe, which future broad-scale monitoring and conservation strategies should take into account.

Although our RADseq sampling design included only 17.6% of samples included in the full microsatellite dataset this was sufficient to produce a robust phylogeography in agreement with the microsatellite dataset, and emphasised the fine scale structure among populations. We therefore conclude that, if made to choose between the comprehensively sampled microsatellite approach or the RADseq approach with fewer samples but many more loci, the RADseq approach presents the better option for the phylogeography of *C. carassius,* with the huge number of SNP loci overcoming the limitations imposed by reduced sample number. We also predict that this will hold true for systems with similar genetic characteristics to ours, *i.e.* strong population structure characterised by IBD.

For Review Only

Acknowledgements

The authors thank the FSBI (www.fsbi.org.uk) and Cefas (Lowestoft, UK) for funding the PhD of DLJ in which this research was performed. Additionally KHO received financial support from The Foundation for Baltic and Eastern European Studies (Östersjöstiftelsen). We thank the following landowners, and contributors of fish tissue; Keith Wesley, Ian Patmore and Dave Emson (England), L. Urho (Helsinki, Finland), M. Himberg (Salo, Finland), J. Krekula (Steninge Castle, Uppland, Sweden), B.-M. Josephson and G. Josephson (Styrstad Vicarage, Sweden), G. Hellström (Umeå University, Sweden), K.Ø. Gjelland (NINA, Tromsø, Norway), N. Hellenberg (Gotland Island, Sweden), A. Tuvikene (Center for Limnology, Tartu, Estonia), S.V. Mezhzherin (Kiev, Ukraine), K. Lindström (Kvicksund, Sweden), K.-J. Dahlbom and G. Sundberg (Åland Island, Finland), O. Sandström and M. Andersson (Skutab, Öregrund, Sweden), B. Tengelin (Structor Miljöteknik AB), A. Olsén-Wannefjord (Uppsala, Sweden), Müller Tamás (Godollo, Hungary), András Weiperth (Hungary), Peter D Rask Møller and Henrik Carl (Copenhagen, Denmark), Oksana Stoliar, (Ternopil, Ukraine), Manuel Deinhardt (Jyväskylä, Finland), Filip Volckaert and Greg Maes (Leuven, Belgium).

References

Andrews S (2010) *FastQC: a quality control tool for high throughput sequence data*. Babraham Bioinformatics. Available at www.bioinformatics.babraham.ac.uk/projects/fastqc.

Arkhipov SA, Ehlers J, Johnson RG, Wright HE Jr (1995) Glacial drainage towards the Mediterranean during the Middle and Late Pleistocene. *Boreas*, **24**, 196–206.

Balkenhol N, Landguth EL (2011) Simulation modelling in landscape genetics: on the need to go further. *Molecular ecology*, **20**, 667–670.

Bernatchez L, Wilson CC (1998) Comparative phylogeography of Nearctic and Palearctic fishes. *Molecular ecology*, **7**, 431–452.

Bianco P (1990) Potential role of the palaeohistory of the Mediterranean and Paratethys basins on the early dispersal of Euro-Mediterranean freshwater fishes. *Ichthyological exploration of freshwaters*, **1**, 167 - 184

Björck S (1995) A review of the history of the Baltic Sea, 13.0-8.0 ka BP. *Quaternary international: the journal of the International Union for Quaternary Research*, **27**, 19–40.

Bohlen J, Šlechtová V, Bogutskaya N, Freyhof J (2006) Across Siberia and over Europe: Phylogenetic relationships of the freshwater fish genus Rhodeus in Europe and the phylogenetic position of *R. sericeus* from the River Amur. *Molecular phylogenetics and evolution*, **40**, 856–865.

Bohlen J, ŠLechtová V, Doadrio I, Ráb P (2007) Low mitochondrial divergence indicates a rapid expansion across Europe in the weather loach, *Misgurnus fossilis* (L.). *Journal of fish biology*, **71**, 186–194.

Bănărescu P (1990) *Zoogeography of Fresh Waters. Vol. 1. General Distribution and Dispersal of Freshwater Animals*. Aula-Verlag, Wiesbaden.

For Review Only

Bănărescu P (1992) *Zoogeography of fresh waters. Vol. 2. Distribution and dispersal of freshwater animals in North America and Eurasia*. Aula-Verlag, Wiesbaden.

Bradbury IR, Hamilton LC, Dempson B *et al.* (2015) Transatlantic secondary contact in Atlantic Salmon, comparing microsatellites, a single nucleotide polymorphism array and restriction-site associated DNA sequencing for the resolution of complex spatial structure. *Molecular Ecology* **24**, 5130–5144.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular ecology*, **22**, 3124–3140.

Coates BS, Sumerford DV, Miller NJ *et al.* (2009) Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis. *The Journal of heredity*, **100**, 556–564.

Copp GH (1991) Typology of aquatic habitats in the great ouse, a small regulated lowland river. *Regulated Rivers: Research & Management*, **6**, 125–134.

Copp G, Sayer C (2010) *Norfolk Biodiversity Action Plan–Local Species Action Plan for Crucian Carp (Carassius carassius). Norfolk Biodiversity Partnership Reference: LS/ 3*. Fisheries & Aquaculture Science, Lowestoft.

Copp G, Tarkan S, Godard M, Edmonds N, Wesley K (2010) Preliminary assessment of feral goldfish impacts on ponds, with particular reference to native crucian carp. *Aquatic invasions*, **5**, 413–422.

Copp GH, Černý J, Kováč V (2008) Growth and morphology of an endangered native freshwater fish, crucian carp *Carassius carassius*, in an English ornamental pond. *Aquatic conservation: marine and freshwater ecosystems*, **18**, 32–43.

For Review Only

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, **144**, 2001–2014.

Cornuet JM, Ravigne V, Estoup A (2010) Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics* **11**, 401.

Cornuet J-M, Pudlo P, Veyssier J *et al.* (2014) DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, **30,** 1187-1189.

Costedoat C, Gilles A (2009) Quaternary pattern of freshwater fishes in Europe: comparative phylogeography and conservation perspective. *The Open Conservation Biology Journal*, **3**, 36-48.

Culling MA, Janko K, Boron A *et al.* (2006) European colonization by the spined loach (*Cobitis taenia*) from Ponto-Caspian refugia based on mitochondrial DNA variation. *Molecular ecology*, **15,** 173–190.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nature Methods **9**, 772.

Davey JW, Cezard T, Fuentes-Utrilla P *et al.* (2012) Special features of RAD Sequencing data: implications for genotyping. *Molecular ecology*, **22**, 3151–3164.

Deinhardt M (2013) The invasive potential of Prussian carp in Finland under the light of a novel semi-clonal reproductive mechanism, Masters thesis, University of Jyväskylä.

Dowling TE, Tibbets CA, Minckley WL, Smith GR, McEachran JD (2002) Evolutionary Relationships of the Plagopterins (Teleostei: Cyprinidae) from Cytochrome b Sequences. *Copeia*, **2002**, 665–678.

For Review Only

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution*, **29**, 1969–1973.

Durand JD, Persat H, Bouvet Y (1999) Phylogeography and postglacial dispersion of the chub (*Leuciscus cephalus*) in Europe. *Molecular ecology*, **8**, 989–997.

Eaton DAR (2014) PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences*, **107**, 16196–16200.

Epperson BK, Mcrae BH, Scribner K *et al.* (2010) Utility of computer simulations in landscape genetics. *Molecular ecology*, **19**, 3549–3564.

Frankham R, Briscoe D, Ballou J (2002) *Introduction to Conservation Genetics*. Cambridge University Press, Cambridge, UK.

Fraser DJ, Bernatchez L (2001) Adaptive evolutionary conservation: towards a unified concept for defining conservation units. *Molecular ecology*, **10**, 2741–2752.

Freyhof J, Brooks E (2011) European red list of freshwater fishes. Luxembourg: Publications Office of the European Union.

Gibbard P, Head MJ (2009) The Definition of the Quaternary System/Era and the Pleistocene Series/Epoch. *Quaternaire*, **20**, 125–133.

Gibbard PL, Rose J, Bridgland DR (1988) The history of the great northwest European rivers during the past three million years [and discussion]. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **318**, 559–602.

Goudet J (1995) FSTAT, a computer program to calculate F-Statistics. *Heredity*, **86**, 485-486

For Review Only

Grosswald MG (1980) Late Weichselian ice sheet of Northern Eurasia. *Quaternary Research*, **13**, 1–32.

Gum B, Gross R, Geist J (2009) Conservation genetics and management implications for European grayling, *Thymallus thymallus*: synthesis of phylogeography and population genetics. *Fisheries management and ecology*, **16**, 37–51.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology, **52**, 696–704.

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, **22**, 160–174.

Hauser L, Baird M, Hilborn R, Seeb LW, Seeb JE (2011) An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. *Molecular ecology resources*, **11 Suppl 1**, 150–161.

Hazewinkel M (Ed.) (1994) *Encyclopeidia of Mathematics (set)*. Kluwer, Dordrecht, Netherlands.

Henkel CV, Dirks RP, Jansen HJ *et al.* (2012) Comparison of the exomes of common carp (*Cyprinus carpio*) and zebrafish (*Danio rerio*). *Zebrafish*, **9**, 59–67.

Hess JE, Matala AP, Narum SR (2011) Comparison of SNPs and microsatellites for fine-scale application of genetic stock identification of Chinook salmon in the Columbia River Basin. *Molecular ecology resources*, **11**, 137–149.

Hewitt GM (1999) Post-glacial re-colonization of European biota. *Biological journal of the Linnean Society. Linnean Society of London*, **68**, 87–112.

Hoban SM, Gaggiotti OE, Bertorelle G (2013) The number of markers and samples needed for detecting bottlenecks under realistic scenarios, with and without recovery: a simulation-based study. *Molecular ecology*, **22**, 3444–3450.

For Review Only

Holopainen IJ, Aho J, Vornanen M, Huuskonen H (1997) Phenotypic plasticity and predator effects on morphology and physiology of crucian carp in nature and in the. *Journal of fish biology*, **50**, 781–798.

Holopainen IJ, Oikari A (1992) Ecophysiological effects of temporary acidification. *Annales Zoologici Fennici*, **29**, 29–38.

Holopainen IJ, Tonn WM, Paszkowski CA (1997) Tales of two fish: the dichotomous biology of crucian carp (*Carassius carassius* (L.)) in northern Europe.

Hänfling B, Hellemans B, Volckaert F, Carvalho GR (2002) Late glacial history of the cold-adapted freshwater fish *Cottus gobio*, revealed by microsatellites. *Molecular ecology,* **11,** 1717–1729.

Hänfling B, Bolton P, Harley M, Carvalho GR (2005) A molecular approach to detect hybridisation between crucian carp (*Carassius carassius*) and non-indigenous carp species (*Carassius* spp. and *Cyprinus carpio*). *Freshwater biology*, **50**, 403–417.

Hickley P, Chare S (2004) Fisheries for non-native species in England and Wales: angling or the environment? *Fisheries management and ecology*, **11**, 203-212.

IUCN 2015. The IUCN Red List of Threatened Species. Version 2015-4, http://www.iucnredlist.org . Downloaded on 19 August 2015.

Janko K, Culling MA, Ráb P, Kotlík P (2005) Ice age cloning--comparison of the Quaternary evolutionary histories of sexual and clonal forms of spiny loaches (*Cobitis*; Teleostei) using the analysis of mitochondrial DNA variation. *Molecular ecology*, **14**, 2991–3004.

Janson S, Wouters J, Bonow M, Svanberg I, Olsén KH (2014) Population genetic structure of crucian carp (*Carassius carassius*) in man-made ponds and wild populations in Sweden. *Aquaculture international: journal of the European Aquaculture Society*, **23**, 359-368.

For Review Only

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* , **27**, 3070–3071.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics*, **11**, 94.

Kontula T, Väinölä R (2001) Postglacial colonization of Northern Europe by distinct phylogeographic lineages of the bullhead, *Cottus gobio*. *Molecular ecology*, **10**, 1983–2002.

Koskinen MT, Ranta E, Piironen J *et al.* (2000) Genetic lineages and postglacial colonization of grayling (*Thymallus thymallus*, Salmonidae) in Europe, as revealed by mitochondrial DNA analyses. *Molecular ecology*, **9**, 1609–1624.

Kostecki R (2014) Stages of the Baltic Sea evolution in the geochemical record and radiocarbon dating of sediment cores from the Arkona Basin. *Oceanological and Hydrobiological Studies*, **43**, 237–246.

Kotlík P, Berrebi P (2001) Phylogeography of the barbel (*Barbus barbus*) assessed by mitochondrial DNA variation. *Molecular ecology*, **10**, 2177–2185.

Landguth EL, Fedy BC, Oyler-McCance SJ *et al.* (2012) Effects of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. *Molecular ecology resources*, **12**, 276–284.

Larmuseau MHD, Freyhof J, Volckaert FAM, Van Houdt JKJ (2009) Matrilinear phylogeography and demographical patterns of *Rutilus rutilus*: implications for taxonomy and conservation. *Journal of fish biology*, **75**, 332–353.

Lelek A (1980) Threatened freshwater fishes of Europe. Council of Europe, Strasbourg, France.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* , **25**, 1451–1452.

For Review Only

For Review Only

Luikart G, Cornuet J-M (2008) Empirical Evaluation of a Test for Identifying Recently Bottlenecked Populations from Allele Frequency Data. *Conservation biology*, **12**, 228-237.

Lusk S, Hanel L, Luskova V (2004) Red List of the ichthyofauna of the Czech Republic: Development and present status. *Folia Zoologica*, **53**, 215–226.

Lusk S, Lusková, V, Hanel L (2010) Alien fish species in the Czech Republic and their impact on the native fish fauna. *Folia Zoology*, **59**, 57–72.

Mabuchi K, Senou H, Suzuki T, Nishida M (2005) Discovery of an ancient lineage of *Cyprinus carpio* from Lake Biwa, central Japan, based on mtDNA sequence data, with reference to possible multiple origins of koi. *Journal of Fish Biology,* **66**, 1516–1528.

Mezhzherin SV, Kokodii SV, Kulish AV, Verlatii DB, Fedorenko LV (2012) Hybridization of crucian carp *Carassius carassius* (Linnaeus, 1758) in Ukrainian reservoirs and the genetic structure of hybrids. *Cytology and genetics*, **46**, 28–35.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2006) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome research*, **17**, 240–248.

Moore A, Russell IC, Potter ECE (1990) The effects of intraperitoneally implanted dummy acoustic transmitters on the behaviour and physiology of juvenile Atlantic salmon, *Salmo salar* L. *Journal of fish biology*, **37**, 713–721.

Morin PA, Luikart G, Wayne RK, the SNP workshop group (2004) SNPs in ecology, evolution and conservation. *Trends in ecology & evolution*, **19**, 208–216.

Morin PA, Pease VL, Hancock BL *et al.* (2010) Characterization of 42 single nucleotide polymorphism (SNP) markers for the bowhead whale (*Balaena mysticetus*) for use in discriminating populations. *Marine Mammal Science*, **26**, 716–732.

Mrakovčić M, Buj I, Mustafić P, Ćaleta M, Zanella D (2007) *Croatian Red List: Freshwater fish*. Department of Zoology, Faculty of Science, Zagreb.

Navodaru I, Buijse AD, Staras M (2002) Effects of Hydrology and Water Quality on the Fish Community in Danube Delta Lakes. *International review of hydrobiology*, **87**, 329–348.

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York.

Nesbø CL, Fossheim T, Vøllestad LA, Jakobsen KS (1999) Genetic divergence and phylogeographic relationships among European perch (*Perca fluviatilis*) populations reflect glacial refugia and postglacial colonization. *Molecular ecology*, **8**, 1387–1404.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) micro-checker: software for identifying and correcting genotyping errors in microsatellite data. *Molecular ecology notes*, **4**, 535–538.

Peery MZ, Kirby R, Reid BN *et al.* (2012) Reliability of genetic bottleneck tests for detecting recent population declines. *Molecular ecology*, **21**, 3403–3418.

Petit RJ, El Mousadik A, Pons O (1998) Identifying Populations for Conservation on the Basis of Genetic Markers. *Conservation biology: the journal of the Society for Conservation Biology*, **12**, 844–855.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Core Team (2013) *R: a language and environment for statistical computing*, Vienna, Austria.

Ramachandran S, Deshpande O, Roseman CC *et al.* (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15942–15947.

For Review Only

Rasmussen H (1959) Fish ponds and fish rearing. In: *Kulturhistoriskt lexikon för nordisk medeltid* (eds: Andersson I, Granlund J)*,* pp. 307–309, Malmö, Allhem.

Reyjol Y, Hugueny B, Pont D *et al.* (2006) Patterns in species richness and endemism of European freshwater fish. *Global ecology and biogeography*, **16**, 65–75.

Rylková K, Kalous L, Bohlen J, Lamatsch DK, Petrtýl M (2013) Phylogeny and biogeographic history of the cyprinid fish genus Carassius (Teleostei: Cyprinidae) with focus on natural and anthropogenic arrivals in Europe. *Aquaculture* , **380,** 13–20.

Salzburger W, Brandstätter A, Gilles A *et al.* (2003) Phylogeography of the vairone (Leuciscus souffia, Risso 1826) in Central Europe. *Molecular ecology*, **12**, 2371–2386.

Savini D, Occhipinti-Ambrogi A, Marchini A *et al.* (2010) The top 27 animal alien species introduced into Europe for aquaculture and related activities. *Journal of applied ichthyology*, **26**, 1–7.

Sayer CD, Copp GH, Emson D *et al.* (2011) Towards the conservation of crucian carp Carassius carassius: understanding the extent and causes of decline within part of its native English range. *Journal of fish biology*, **79**, 1608–1624.

Schwartz MK, McKelvey KS (2009) Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. *Conservation genetics*, **10**, 441–452.

Simic, V, Simic S, Cirkovic M, Pantovic N (2009) *Preliminary red list of the fishes of Serbia*. COMBAFF-First Conference on Conservation and Management of Balkan Freshwater Fishes.

Simon A, Gozlan RE, Robert Britton J, van Oosterhout C, Hänfling B (2014) Human induced stepping-stone colonisation of an admixed founder population: the spread of topmouth gudgeon (*Pseudorasbora parva*) in Europe. *Aquatic sciences*, **77**, 17–25.

For Review Only

For Review Only

Svanberg I, Bonow M, Olsén H (2013) Fish ponds in Scania, and Linnaeus's attempt promote aquaculture in Sweden. In: *Svenska Linnésällskapets årsskrift* (eds David B, Gunnar D), pp. 85–100. Svenska Linnésällskapet, Uppsala.

Svärdson G (1998) Plostglacial dispersal and reticulate evolution of Nordic Coregonids. *Nordic journal of freshwater research*, **74**, 3–32.

Takada M, Tachihara K, Kon T *et al.* (2010) Biogeography and evolution of the *Carassius auratus*-complex in East Asia. *BMC evolutionary biology*, **10**, 7.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution*, **30**, 2725–2729.

Tarkan AS, Cucherousset J, Zięba G, Godard MJ, Copp GH (2010) Growth and reproduction of introduced goldfish *Carassius auratus* in small ponds of southeast England with and without native crucian carp *Carassius carassius*. *Journal of applied ichthyology*, **26**, 102–108.

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. In: *Lectures in mathematics in the life sciences* (ed Miura RM), pp. 57–86. American Mathematical Society, Providence, RI.

Tsipas G, Tsiamis G, Vidalis K, Bourtzis K (2009) Genetic differentiation among Greek lake populations of *Carassius gibelio* and *Cyprinus carpio carpio*. *Genetica* **136**, 491–500.

Urho L, Lehtonen H (2008) Fish species in Finland. Finnish Game and Fisheries Research Institute, Helsinki.

Van Oppen MJH, Rico C, Turner GF, Hewitt GM (2000) Extensive Homoplasy, Nonstepwise Mutations, and Shared Ancestral Polymorphism at a Complex Microsatellite Locus in Lake Malawi Cichlids. *Moleculary Biology and Evolution*, **17**, 489–498.

Weir B, Cockerham C (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38***, 1358-1370.*

Wheeler A (2000) Status of the crucian carp, Carassius carassius (L.), in the UK. *Fisheries management and ecology*, **7**, 315–322.

Wolfram G, Mikschi E (2007) Rote Liste der Fische (Pisces) Österreichs. In: *Rote Liste gefährdeter Tiere Österreichs, Teil 2.* Grüne Reihe des Lebensministeriums Band 14/2. (ed Zulka K), pp. 61–198. Böhlau-Verlag, Wien.

Wouters J, Janson S, Lusková V, Olsén KH (2012) Molecular identification of hybrids of the invasive gibel carp *Carassius auratus gibelio* and crucian carp *Carassius carassius* in Swedish waters. *Journal of fish biology*, **80**, 2595–2604.

Xu P, Zhang X, Wang X *et al.* (2014) Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nature genetics*, **46**, 1212-1219.

For Review Only

Author contributions

DLJ collected, samples, performed lab work, analysed data and wrote the manuscript. BH was involved with conception of the project, advised on all steps of analyses and commented on the manuscript. GHC was involved in the conception of the project, contributed samples and commented on the manuscript. LJLH advised on the analysis and commented on the manuscript. CDS and KHO provided samples and commented on manuscript.

Data accessibility

Genbank accession numbers for mtDNA sequences are provided in Table 1 of this manuscript. The microsatellite data files for all DAPC analyses, mantel test matrices, mtDNA raw sequences, sequence alignments, model testing outputs, tree files, RADseq loci catalog and VCF files used for analyses and the DIYABC project files (containing all inputs, scenarios and parameter priors for each analysis stage) have now been uploaded to Dryad (http://dx.doi.org/10.5061/dryad.t2j45). All scripts used for DAPC, mantel tests and comparisons of RADseq and microsatellite datasets can be found on GitHub (https://github.com/DanJeffries/Jeffries-et-al-2016-crucian-phylogeography). All demultiplexed RADseq reads have been uploaded to the short read archive (Project accession: SRP063043).

For Review Only

Figure 1. Population structure of *C. carassius* in Europe. a) Sampling locations (sites sampled with nuclear and mtDNA markers = red dots, mtDNA only = blue dots) and population cluster memberships from DAPC analysis. Pie chart size corresponds to microsatellite allelic richness. Pie chart colours for Danubian populations and RUS1 correspond to clusters in the broad scale DAPC analysis b) and for all northern European populations colours correspond to clusters in the northern European DAPC analysis (mtDNA lineage 1 only) c). The Danube river catchment is shaded dark grey.

Figure 2. Maximum credibility tree calculated in BEAST for 100 *C. carassius cyt*b sequences. For the three maximally supported nodes, age is given above and the posterior probability distribution is given below, with 95% CI's represented by blue bars.

Figure 3. Comparison of DAPC results using a) RADseq dataset, b) M2 dataset and c) M3 dataset. Colours correspond between DAPC scatter plots and maps within but not between panels.

Figure 4. The postglacial recolonisation of *C. carassius* in Europe. Arrows represent the relationships between population pools used in DIYABC (grey circles) as inferred from Stage 1, scenario 9 (arrows outlined in black) and Stage 3, scenario 14d (arrows with no outline) analyses on RADseq data. Bottlenecks are represented by white-striped sections of arrows. Posterior time estimates in years for each demographic event are given in black, and estimates of Ne are given in blue. Blue diamonds represent ancestral populations inferred by DIYABC and the labels (a-f) correspond to their mention in the text. Hypothetical expansion fronts are represented by dashed contour lines and the Danube river catchment is shaded red. Hypothetical glacial refugia are represented by dashed blue circles (I - III). The blue dashed box (?) represents our inference that *C. carassius* expanded into central and perhaps northern Europe during the Riss-Würm interglacial, however we cannot estimate this range.

SI Figure 1. DIYABC scenarios used in broad-scale analysis (Stage 1). See text for population poolings. See Table 3 for population poolings and prior parameter values.

SI Figure 2. All scenarios tested in stage 2 a) and stage 3 b) of DIYABC analysis. See Table 3 for population poolings and prior parameter values.

SI Figure 3. Filtering out merged ohnologs. a) Distribution of SNP locus coverage prior to removing loci that had observed heterozygosity higher than 0.5 in one or more population. b) Distribution of locus coverage after filtering, showing a loss of many high coverage loci and a reduction in mean SNP coverage. Note the loss of loci with high coverage.

SI Figure 4. Linear regressions for all samples a) *A*r against latitude; b) *A*r against longitude and for only samples in mtDNA lineage 1 c) *A*r against latitude; d) *A*r against longitude.

SI Figure 5. DAPC analysis of a) full microsatellite dataset (Excluding NOR2); for results used in Fig. 1) and b) Full RADseq dataset.

SI Figure 6. Isolation by distance a) in M1 dataset for mtDNA lineage 1 only (excluding NOR2), b) full RADseq dataset, c) M2 dataset and d) M3 dataset.

SI Figure 7. DAPC scatter plot for the 1000 SNP RADseq dataset used in the DIYABC analysis, showing the same population structure as inferred from the full RADseq dataset.

SI Figure 8. Broad scale DIYABC analysis (Stage 1) results. a) Direct approach (left) and Logistic regression (right) showing support for scenario 9. b) Model checking for scenario 9, showing that the observed data fall well within the cloud of datasets simulated from the posterior parameter distribution. c) Scenario 9 schematic.

SI Figure 9. Fine scale DIYABC analysis in northern Europe. a) Stage 2 - major topological variants of scenarios. Direct approach (top left) and Logistic regression (top right) showing support for scenario 14 and 13 respectively. Model checking (Middle) for scenario 14 (bottom), showing that the observed data fall well within the cloud of datasets simulated from the posterior parameter distribution. Note the model checking placed the observed data outside of the cloud of posterior datasets for scenario 13. b) Stage 3 - Minor scenario variants of scenario 14 from stage 2. Direct

For Review Only

approach (top left), logistic regression (top right) and model checking (middle) all support scenario 14d (bottom).

SI Figure 10. Comparison of spatial patterns of uniformity in geographic sampling regimes of the full M1 dataset locations (a, c) and the sampling location subset used in M2, M3, and RAD datasets (b,d). Estimates of G and L from true sampling locations are plotted using the black solid lines. Estimates of G and L from simulated locations based on random Poisson distribution is represented by the red dashed line. Grey shaded areas are the 95% confidence intervals around the random estimates. Both the G and L function estimates show that there is more clustering of sampling locations in the M1 dataset than in the M2, M3 and RAD subsets.

SI Figure 11. Change in a) number of RAD tags and b) average tag coverage for three individuals used in the preliminary Stacks tag mismatch parameter (M) tests.

SI Figure 12. Results of parameter tests for the Stacks module Populations. a) Number of SNP loci in final dataset for incrementing values of parameters –p, -r and –m; b) average coverage per SNP and per sample for the same parameter values; c) the number of loci which drop out in each population for each test value of the –p parameter

For Review Only

Table 1. Location, number, genetic marker sampled, and accession numbers of samples and sequences used in the present study for microsatellite and mitochondrial DNA analyses. mtDNA sequence accession numbers can be found in SI table 2.

Code	Accession	Location	Country	Drainage	Coordinates		Microsatellites	mtDNA	RADseq
					lat	long			
GBR1		London	U.K.	U.K	51.5	0.13	9		
GBR2		Reading	U.K.	U.K	51.45	-0.97	4		
GBR3		Norfolk	U.K.	U.K	52.86	1.16	7		
GBR4		Norfolk	U.K.	U.K	52.77	0.75	27		9
GBR5		Norfolk	U.K.	U.K	52.77	0.76	14		
GBR6		Norfolk	U.K.	U.K	52.54	0.93	29	3	
GBR7		Norfolk	U.K.	U.K	52.9	1.15	24	1	10
GBR8		Hertfordshire	U.K.	U.K	52.89	1.1	37	3	9
GBR9		Norfolk	U.K.	U.K	52.8	1.1	27		
GBR10		Norfolk	U.K.	U.K	52.89	1.1	14		
GBR11		Norfolk	U.K.	U.K	52.92	1.16	20		
BEL1		Bokrijk	Belgium	Scheldt River	50.95	5.41	13	1	
BEL2		Meer van Weerde	Belgium	Scheldt River	50.97	4.48	12		
BEL3		Meer van Weerde	Belgium	Scheldt River	50.97	4.48	8		
GER1*		Kruegersee	Germany	Elbe River	52.03	11.97		3	
GER2		Münster	Germany	Rhine River	51.89	7.56	21	3	
GER3		Bergheim	Germany	Danube River	48.73	11.03	9	3	
GER4		Bergheim	Germany	Danube River	48.73	11.03	8	3	
CZE1		Lužnice	Czech Republic	Danube River	48.88	14.89	9	3	
POL1		Sarnowo	Poland	Vistula River	52.93	19.36	33		
POL2		Kikót-Wies	Poland	Vistula River	52.9	19.12	34		
POL3		Tupadly	Poland	Vistula River	52.74	19.3	17	3	10
POL4		Orzysz	Poland	Vistula River	53.83	22.02	13	3	10
EST1		Tartu	Estonia	Baltic Sea	58.39	26.72	5	3	
EST2		Vehendi	Estonia	Baltic Sea	58.39	26.72	5		
RUS4*		Small lake, Velikaya river	Russia	Baltic Sea	55.9	30.25	29	3	
FIN1		Joensuu	Finland	Baltic Sea	62.68	29.68	32	3	
FIN2		Helsinki	Finland	Baltic Sea	60.36	25.33	32		
FIN3		Jyväskylä	Finland	Baltic Sea	62.26	25.76	37	3	10
FIN4		Oulu	Finland	Baltic Sea	65.01	25.47	7	3	8
FIN5		Salo	Finalnd	Baltic Sea	60.37	23.1	10	3	
FIN6		Åland Island	Sweden	Baltic Sea	60.36	19.85	8	3	
SWE1		Gränbrydammen	Sweden	Baltic Sea	59.87	17.67	25		
SWE2		Stordammen	Sweden	Baltic Sea	59.8	17.71	21	3	10
SWE3		Östhammar	Sweden	Baltic Sea	60.26	18.38	27	3	
SWE4		Umeå	Sweden	Baltic Sea	63.71	20.41	9	3	
SWE5		Kvicksund	Sweden	Baltic Sea	59.45	16.32	9		
SWE7		Grillby	Sweden	Baltic Sea	59.64	17.37	10		
SWE8		Skabersjo	Sweden	Baltic Sea	55.55	13.15	19	3	10
SWE9		Märsta	Sweden	Baltic Sea	59.6	17.8	31	3	
SWE10		Norrköping	Sweden	Baltic Sea	58.56	16.27	29		9
SWE11		Gotland Island	Sweden	Baltic Sea	57.85	18.79	11	3	
NOR1		Oslo	Norway	North Sea	60.05	9.94		2	
NOR2		Lake Prestvattnet, Tromsø	Norway	North Sea	69.65	18.95	16		9
BLS			Belarus	Dnieper	52.47	30.52	7	1	
RUS1		Proran Lake	Russia	Don River	47.46	40.47	10	3	9
DEN1		Copenhagen	Denmark	Baltic Sea	60.21	17.79	12		10
DEN2		Pederstrup	Denmark	Baltic Sea	55.77	12.55	14		8
DEN3		Gammel Holte	Denmark	Baltic Sea	56	12.5	14		
DEN4		Bornholm Island	Denmark	Baltic Sea	55.17	14.86			5
SWE12		Osterbybruk Mansion	Sweden	Baltic Sea	55.73	12.34	14		9
SWE14		Wenngarn Castle	Sweden	Baltic Sea	59.66	18.95	16		9
RUS2*		Karma	Russia	Volga River	52.9	58.4		2	
RUS3*		Saygach'yedake	Russia	Volga River	47.5	48.5		4	
TNO			Netherlands	North Sea	-	-		1	
HUN1		Gödöllő	Hungary	Danube River	47.61	19.36		2	6
HUN2		Vörösmocsár	Hungary	Danube River	46.49	19.17			
							848	**83**	**160**
								Total number of fish =	**867**

Genbank mtDNA Sequences

Code	Accession	Reference	Country	Drainage
GER6	DQ399917	Kalous et al. (2007)	Germany	Baltic sea
GER6	DQ399918	Kalous et al. (2007)	Germany	Baltic sea
GER6	DQ399919	Kalous et al. (2007)	Germany	Baltic sea
GER7	JN412540	Rylková et al. (2013)	Germany	Hunte River
GER7	JN412541	Rylková et al. (2013)	Germany	Hunte River
GER7	JN412542	Rylková et al. (2013)	Germany	Hunte River
GER7	JN412543	Rylková et al. (2013)	Germany	Hunte River
GER8*	JN412537	Rylková et al. (2013)	Germany	Lahn River
GER8*	JN412538	Rylková et al. (2013)	Germany	Lahn River
CZE2	GU991399	Rylková et al. (2013)	Czech Republic	Elbe drainage
Milevsko	DQ399938	Kalous et al. (2012)	Czech Republic	Elbe drainage
AUS1	JN412534	Rylková et al. (2013)	Austria	Danube river
AUS1	JN412533	Rylková et al. (2013)	Austria	Danube river
AUS2	JN412535	Rylková et al. (2013)	Austria	Danube river
AUS3	JN412536	Rylková et al. (2013)	Austria	Danube river
GBR12	JN412539	Rylková et al. (2013)	U.K.	U.K

GBR12	GU991400	Kalous et al. (2012)	U.K.	U.K
SWE15	JN412545	Rylková et al. (2013)	Sweden	Baltic sea
SWE16	JN412544	Rylková et al. (2013)	Sweden	Baltic sea
Ccarp1	AB158807	Mabuchi et al (2005)	Japan	-
Ccarp2	DQ868875	Tsipas et al. (2009)	Greece	-
Ccarp3	KF574490	Unpublished	India	-

† Also present

* Location on Map (Fig. 1.a) is approximate

For Review Only

Table 2. Population pools, parameter priors used and median posterior parameter values inferred in the three stages of DIYABC analysis.

Analysis stage	Population Pools	Scenarios tested	Parameter priors	Most likely Scenario	Median of posterior distributions of most likely scenario
1	**Pool 1** – GBR4, GBR7, GBR8, DEN1, DEN2, DEN3, FIN3, FIN4, POL3, POL4, SWE2, SWE8, SWE9, SWE10, SWE12, SWE14, NOR2 **Pool 2** – DEN1, DEN2, DEN3 **Pool 3** – FIN3, FIN4	1 - 11	N1 = 10E+03 - 500E+03	9	N1 =34700
			Nb1 = 10 - 100E+03		Nb1 =23700
			N2 = 100 - 100E+03		N2 =74900
			N3 = 100 - 200E+03		N3 =140000
			t1 = 1E+03 - 1E+06 gens		t1 =135000
			t2 = 1E+03 - 3E+06 gens		db =4460
			ra = 0.001-0.999		
			rb = 0.001-0.999		t2 =1090000
			rc = 0.001-0.999		
			db = 10- 10E+03 gens		
2	**Pool 1** – GBR4, GBR7, GBR8 **Pool 2** – DEN1, DEN2, DEN3 **Pool 3** – FIN3, FIN4	12 - 16	N1 = 10-4E+03	14	N1 =3670
			N2 = 10 - 10E+03		N2 =7520
			N3 = 10 - 20E+03		N3 =17400
			N4 = 10 - 50E+03		N4 =19400
			N5 = 10 - 20E+03		N5 =11800
			N6 =10 - 400		N6 =210
			t1 = 100- 10E+03 gens		t1 =6790
			t1a = 100- 10E+03 gens		t1a =2510
			t2 =100- 10E+03		
			t2a =100- 5E+03 gens		
			t2b = 500-20E+03 gens		
			t2c = 100 - 10E+03 gens		
			t2d = 100 - 10E+03 gens		t2d =6780
			t3 = 500 - 20E+03 gens		
			t3c =100 - 10E+03 gens		
			t3d =100 - 10E+03 gens		t3d =8910
			t4 =500 - 20E+03 gens		t4 =12000
			ra = 0.001-0.999		
			rb = 0.001-0.999		rb =0.668
3	**Pool 4** – POL3, POL4 **Pool 5** – SWE2, SWE8, SWE9, SWE10, SWE12, SWE14 **Pool 6** – NOR2	14a - 14f	N1 = 10-4E+03	14d	N1 =2390
			Nb1 = 10-10E+03		Nb1 =935
			N2 = 10 - 10E+03		N2 =8140
			N3 = 10 - 20E+03		N3 =9360
			Nb3 = 10-10E+03		
			N4 = 10 - 50E+03		N4 =17000
			N5 = 10 - 20E+03		N5 =11000
			N6 =10 - 400		N6 =138
			Nb6 =10-10E+03		
			t1 = 100- 10E+03 gens		t1 =3750
			t1a = 100- 10E+03 gens		t1a =2460
			t2d = 100 - 10E+03 gens		t2d =5900
			t3d = 100 - 10E+03 gens		t3d =7970
			t4 = 500 - 20E+03 gens		t4 =16800
			rb = 0.001-0.999		rb =0.619
			da = 10 - 10E+03 gens		
			db = 10 - 10E+03 gens		
			dc = 10 - 10E+03 gens		dc =9070
			dd = 10 - 10E+03 gens		
			de = 10 - 10E+03 gens		

Table 3. Summary statistics for M1, M2, M3 and RADseq datasets. RAD contains all RADseq data, M1 contains all microsatellite data, M2 contains only microsatellite for the individuals used in the RADseq, and M3 contains all microsatellite data for all individuals that were available in populations that were used in RADseq.

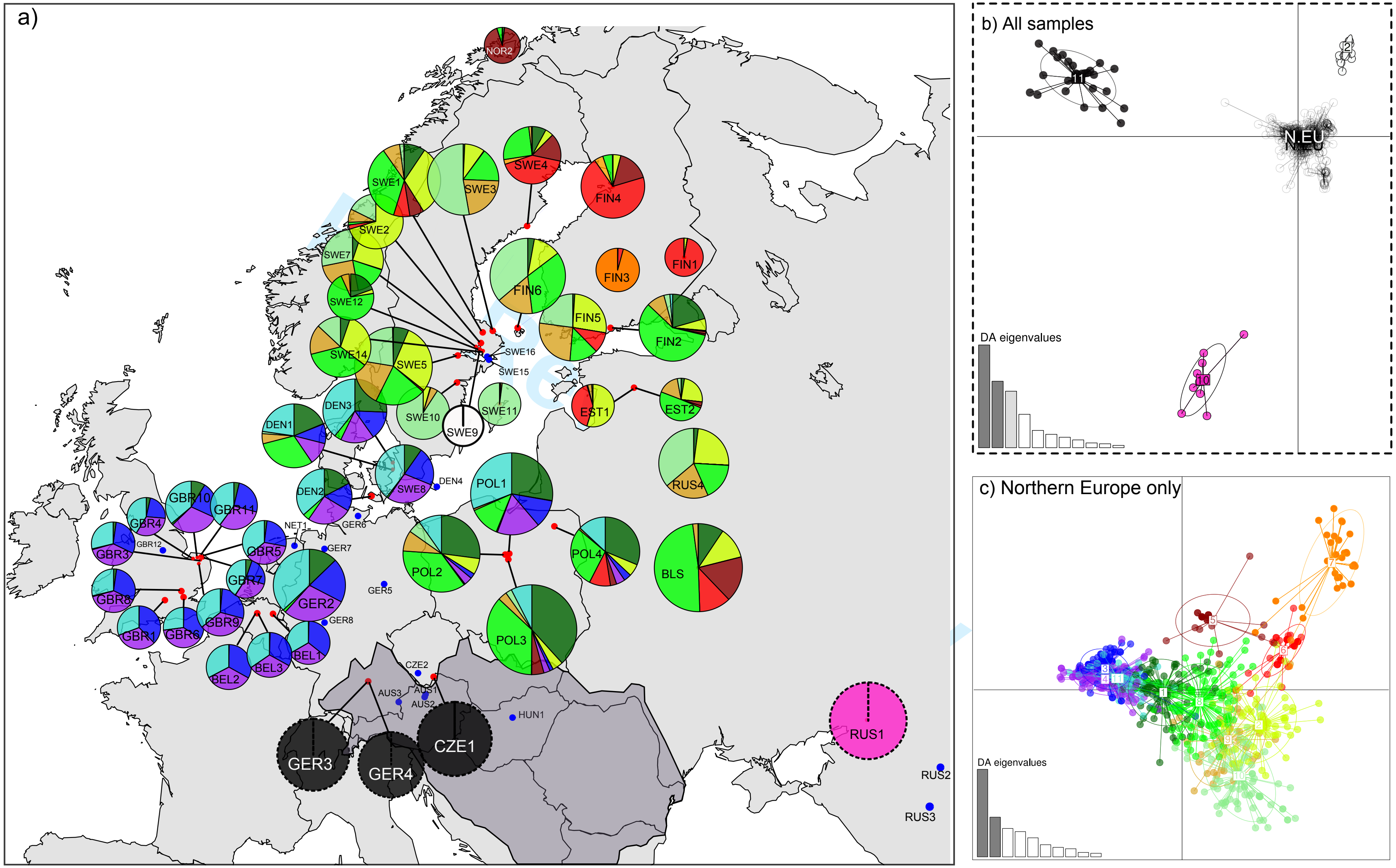
Dataset	Description	N samples	Mean N samples/pop	N. loci	Mean N.alleles/pop	Mean N.alleles/locus
RAD	RADseq data only	149	8.95 ± 1.4	13189	6723	2
M1	Full Microsatellite dataset	848	17.2 ± 9.5	13	27 ± 8.8	7.6
M2	Microsatellites for RADseq samples only	146	9.13 ± 0.8	13	24.4 ± 7.3	7.84 ± 5.1
M3	Microsatellites for all samples in populations used in RADseq	313	19.6 ± 9.0	13	27.4 ± 8.1	11.23 ± 7.6

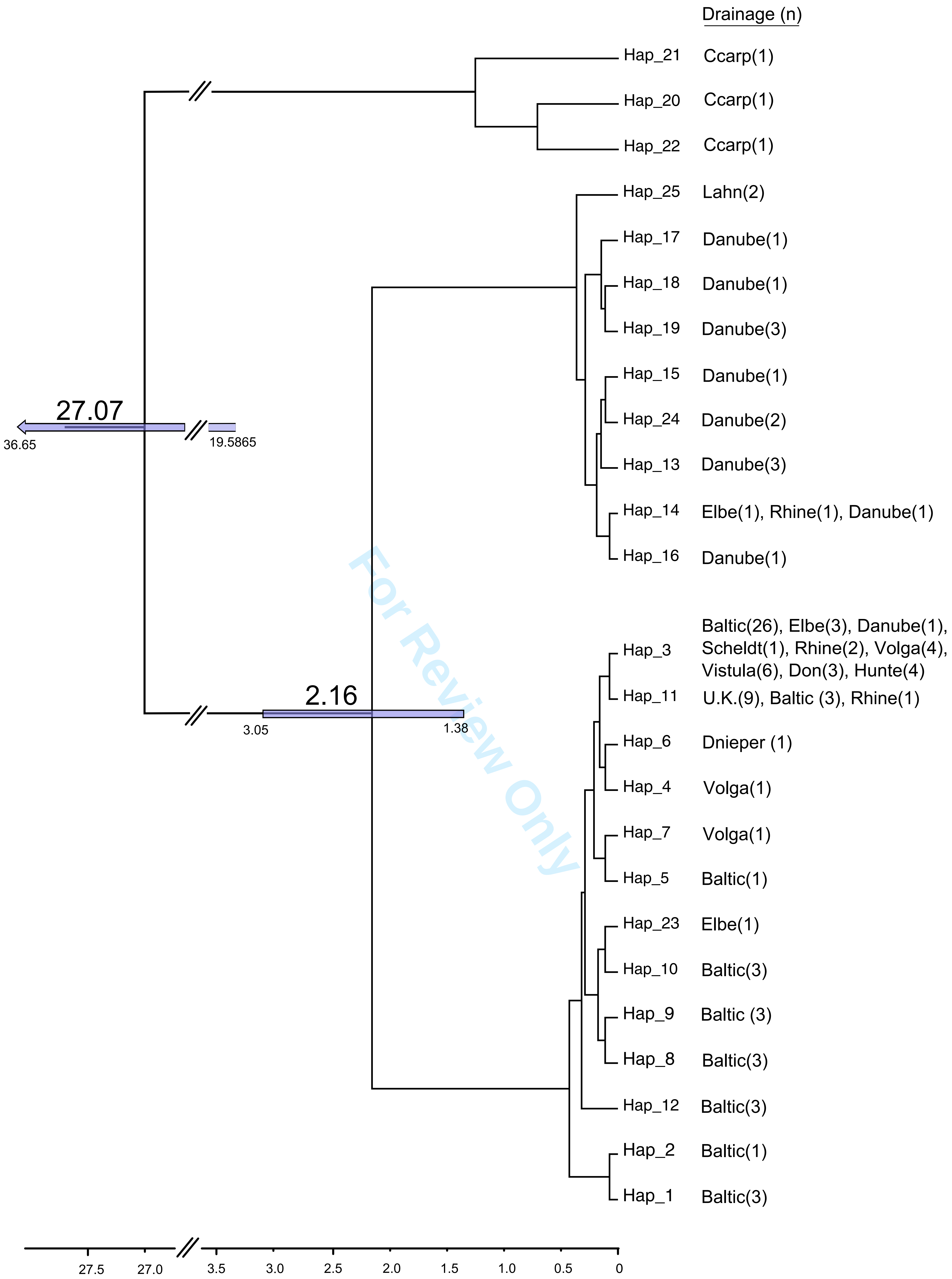
Table 4. Pearson's product-moment correlation coefficients and paired t-tests comparing heterozygosities and F_{ST}s between M2, M3 and RADseq datasets. *** P = <0.001, ** P = < 0.005, * P = < 0.05.

Heterozygosities (df = 18)	Pearsons correlation coefficient (t)		
	M2	11.13***	3.85**
Paired T-tests	-2.4*	**M3**	3.86**
	-9.71***	-9.29***	**RAD**

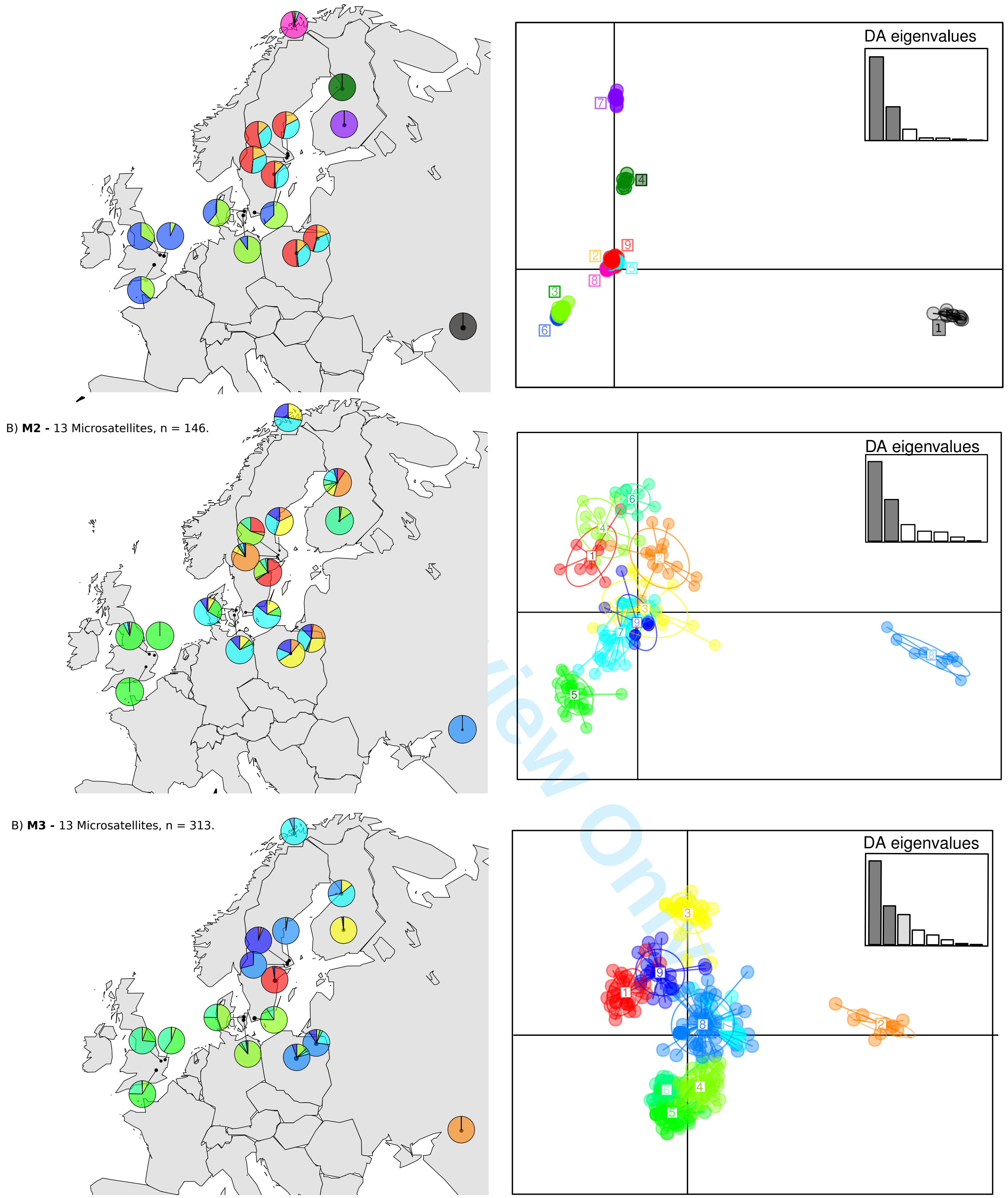
F_{ST} (df = 105)	Pearsons correlation coefficient (t)		
	M2	46.26***	10.09***
Paired T-tests	-6.21***	**M3**	9.05***
	13.74***	15.12***	**RAD**

For Review Only


a)
NOR2
SWE1
SWE3
SWE4
FIN4
SWE2
SWE7
SWE12
FIN6
FIN3
FIN1
FIN5
FIN2
SWE14
SWE5
SWE16
SWE15
SWE10
SWE11
SWE9
DEN3
DEN1
EST1
EST2
RUS4
SWE8
DEN4
DEN2
POL1
GBR10
GBR11
GBR4
GBR3
GBR12
NET1
GER6
GBR5
GER7
POL2
POL4
BLS
GBR7
GER2
GER5
GBR8
GBR9
GER8
POL3
GBR1
GBR6
BEL1
BEL3
BEL2
CZE2
AUS3
AUS1
AUS2
HUN1
RUS1
GER3
GER4
CZE1
RUS2
RUS3
b) All samples
11
2
N.EU
10
DA eigenvalues
c) Northern Europe only
7
5
6
3
4
1
8
9
10
DA eigenvalues



For Review Only

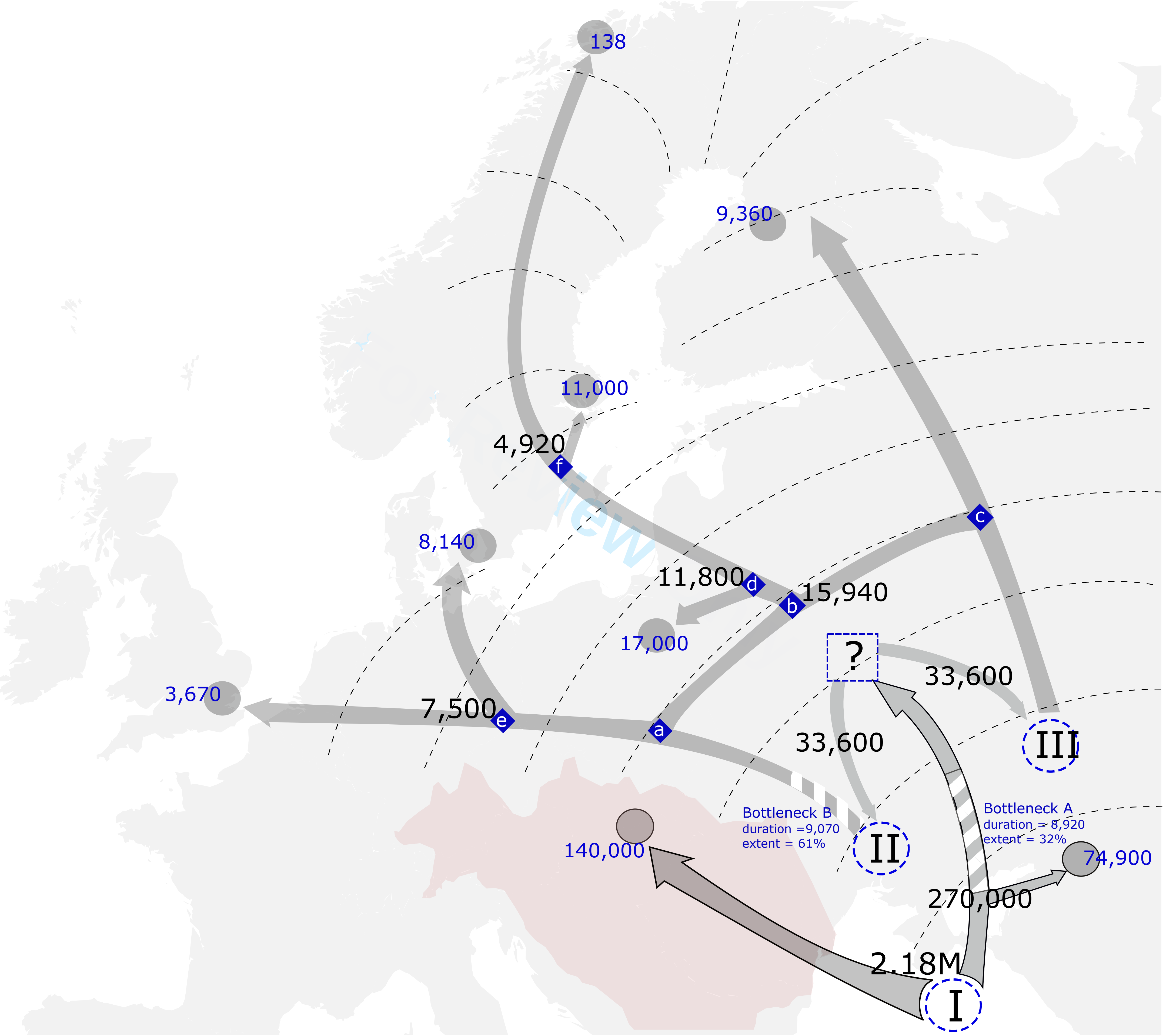


A) **RADseq -** 13,189 SNPs, n = 149.

B) **M2 -** 13 Microsatellites, n = 146.

B) **M3 -** 13 Microsatellites, n = 313.

For Review Only


138
9,360
11,000
4,920
f
8,140
11,800
d
15,940
b
c
17,000
?
33,600
3,670
7,500
e
a
33,600
III
Bottleneck B
duration =9,070
extent = 61%
II
Bottleneck A
duration = 8,920
extent = 32%
140,000
74,900
270,000
2.18M
I

Comparing RADseq and microsatellites to infer complex phylogeographic patterns, a real data informed perspective in the Crucian carp, *Carassius carassius,* L.

Authors: [1]Daniel L Jeffries, [2]Gordon H Copp, [1]Lori Lawson Handley, [3]K. Håkan Olsén, [4]Carl D Sayer, [1]Bernd Hänfling

Supporting Information

For Review Only

Detecting hybrids

Methods

In total we acquired tissue samples of 1354 fish from 72 populations. All samples were first genotyped using multiplex 1 (SI table 1) which contained the 6 species diagnostic microsatellite loci. These data were then analysed using the NewHybrids v. 1.1 (Anderson & Thompson 2002) software package in order to determine whether each fish was *C. carassius, C. auratus, C. gibelio* or a hybrid between any of these species.

NewHybrids uses allele frequencies to give a likelihood probability that an individual belongs to one species or another, or if the individual belonged to one of several hybrid classes (F1, F2 or backcross). Data from 20 *C. carassius* samples, which were confidently identified as pure from both morphology and genotypes, and were not sympatric with non-native species, were included in each analysis as baseline data. Priors were then added to the analyses specifying that these individuals were indeed pure in order to give the software more power with which to assess allele frequencies associated with *C. carassius*. To be sure to account for allele frequency differences between different geographic regions, only pure individuals from regions neighbouring the hybrid population were used. Individuals which had more than a 25% chance of being an F1 hybrid, F2 hybrid, or a backcross were removed from population structure analyses and were not genotyped at the additional 7 microsatellite loci (Multiplexes 2.1 and 2.2, SI table 1).

Results

Of the 1354 fish which were genotyped with microsatellites, 942 individuals across 55 populations (86.7%) were identified as pure *C. carassius* using the first set of 6 species diagnostic loci in NewHybrids analyses. 19 (1.8%) from 2 different populations were identified as *C. auratus*, 15 fish (1.4%) from 4 populations were identified as. *C. gibelio* and 10 fish (0.93%) from two populations were identified as *C. carpio*. NewHybrids identified 60 (5.5%) *C. carassius* x *C. auratus* hybrids, 25 (2.2%) *C. carassius* x *C. gibelio* hybrids, and 16 (1.5%) *C. carassius* x *C. carpio* hybrids. Of the 942 fish identified as pure *C. carassius*, 867 in existed in locations (49 populations) where hybrids or non-native species were not detected by microsatellite genotyping. To safeguard against cryptic introgression which may produce erroneous results only these 867 pure *C. carassius* were used for the main phylogeographic analyses and tests using either microsatellites, mtDNA or RADseq.

RADseq data filtering and Stacks analysis parameter testing

RADseq analyses were performed using only the first-end reads from the paired-end sequencing, as coverage across the length of the second-end contigs was not consistent enough to call SNPs in all individuals. For these first-end reads, raw data was first quality checked using FastQC (Andrews 2010), which assesses the per-base sequence quality and content of reads, and provides comprehensive graphical outputs with which to assess the overall quality of raw sequencing data. These analyses did not identify any individuals that had low overall sequence quality, therefore all samples were retained for further analyses.

For Review Only

Preliminary analyses were also carried out using PyRAD (Eaton 2014), which allows for the incorporation of allelic variants resulting from insertions and deletions. However, no significant difference in the number of usable loci was shown. As Stacks provides more downstream populations genetics facilities, this program was used for the final analyses.

Raw RADseq reads were first, demultiplexed using the “process_radtags” module distributed with Stacks and our inline barcodes. Second, reads were filtered for any sequences containing Illumina adapters or primers and trimmed to a length of 92 bp. Third, PCR duplicates introduced during library preparation were removed using the “clone_filter” program (also distributed with Stacks). Finally, preliminary tests of parameter values for each module of the de novo stacks pipeline were performed in order to identify “optimal” parameter values (i.e. where loci number and read depth were stable) for use in the final Stacks analysis. These tests were carried out for 5 sets of 3 randomly chosen individuals from the RADseq dataset and, for each test, all non-test parameters were kept as default. In the ustacks module, which groups identical reads into stacks and then stacks into loci, Parameters M and m were tested (See Catchen et al. 2013 for detailed description of parameters). M values were increased in increments of 2 from 0 to 10. The efficiency of ustacks in finding real loci was then examined with simple counts of the number of constructed loci at each M parameter value and the read coverage of these loci. The expectation was that, at low parameter values, divergent alleles (percentage divergence > M) at a locus will not merge (under-merging), thus increasing the number of loci overall and decreasing the average coverage. In contrast high parameter values could cause over-merging of paralogous loci and have the opposite effects on the number of loci and coverage (Catchen et al. 2013). SI Fig. 11 shows the outputs for a single subset of *C. carassius* samples, which was typical of all 5 subsets tried. In ustacks, an ‘m’ parameter value of zero (minimum of 0 reads required to form a stack) resulted in a very large number of tags (49000-54000) as expected. Likely due to many single reads containing sequencing error being called as loci. The number of loci decreased by approximately 3000 – 4000 tags in the samples

For Review Only

tested at a required read depth of 2 (approx. 50,000), after which further increases in ‘m’ resulted in small decreases in the number of tags. This likely reflects merging of paralogous loci, or low coverage loci. Mean coverage across all loci within an individual of course reflected the ‘m’ parameter increase, jumping initially from approx. 16 reads per locus with zero read depth required, to 20-35 at a minimum required depth of two reads. On the basis of these results we chose an m = 8, to ensure high power for SNP calling.

Incrementing over values of ‘M’ again met our expectations, with the number of loci dropping significantly as the ‘M’ parameter was increased from zero to 2 mismatches allowed, and then dropping more slowly with higher mismatch allowance. These further drops may again be allowing for paralog merging between loci. The mean coverage of loci behaved as expected, with higher mismatch allowance, more divergent reads can be added to existing stacks, inflating coverage for those loci. On the basis of these results M=2 was chosen for final analyses.

Parameter tests were also performed for the cstacks parameter N, which is responsible for setting the maximum mismatch threshold allowed between homologous loci among individuals in the locus catalog. First, ustacks was run using chosen “optimal” parameters to obtain the inputs necessary for cstacks. Cstacks was then run separately on each of the 5 sample subsets with values of N between 0 – 10, with increments of 2.

For Review Only

Finally, we tested three core parameters in the Populations module of Stacks, -m which is analogous to the parameter of the same name in the ustacks module, -r, which specifies the number of individuals within a give population that a locus must be present in, and –p which specifies the number of populations that a locus must be present in (above the –r threshold) for it to be retained in the final dataset (SI Fig. 12). –p was tested for values of between 13 – 19 populations, -r was tested for values between 0.5 – 1.0 and –m was tested for values between 1-8 however, a the dataset had previously been filtered at previous stages for loci present with a depth of 8 reads or higher, the tests of –m in the populations stage were redundant.

Final running parameters used

For all parameter tests, the optimal values were taken to be those where the rate of change in either RAD tag number, or coverage began to decrease. In ustacks, a maximum of two mismatches were allowed between alleles at a given locus (M=2) and at least eight identical reads per stack (m=8) were required. Default values were used for all other parameters. ustacks also called SNPs within individuals at each locus. The cstacks module was then used to merge loci across individuals into a catalog, where N=2 mismatches were allowed between individuals at a given locus. Individuals were then searched against this catalog using Sstacks to determine their genotype at each catalog locus. For the Populations module, optimal values were chosen so that loci that were shared between at least 70% of individuals in each population (-r = 0.7), allowing loci to drop out in one or two individuals in a population for reasons of low DNA sample quality or low coverage. Loci must have also been present in 17 of the 19 populations (-p = 17), and have read depth of at least 8 (-m 8) in each individual.

DAPC & Running parameters

Methods

Population structure was examined using Discriminant Analyses of Principal Components (DAPC, (Jombart *et al.* 2010)) in adegenet. Similar to the more commonly used program, STRUCTURE (Pritchard et al. 2000), DAPC is an individual-based approach that uses Principal Components Analysis (PCA) to transform population genetic data and Discriminant Analysis (DA) to identify clusters. The number of clusters is assessed using the K-means method, which is also used in STRUCTURE (Pritchard *et al.* 2000). Unlike STRUCTURE, DAPC does not assume underlying population genetics models such as Hardy-Weinberg Equilibrium (Jombart *et al.* 2010) and is

For Review Only

therefore more suitable for analysing C. carassius since populations are often bottlenecked (Hänfling *et al.* 2005). An additional benefit of DAPC is that it maximizes between-group variation, while minimizing variation within groups, allowing for optimal discrimination of between-population structure (Jombart *et al.* 2010).

Results

For the full microsatellite dataset (M1), BIC scores indicated that between 11 and 19 genetic clusters (**Error! Reference source not found.**) would be an appropriate model of the variation in the data. We therefore chose 11 clusters to use in the discriminant analysis, retaining 8 principal components as recommended by the spline interpolation a-scores (**Error! Reference source not found.**c) and we kept 2 linear discriminants for plotting (**Error! Reference source not found.**b).

Three major lineages were found, one located in the Danube, one in the Don, and one spread across northern Europe. However the large amount of divergence between them masked the population structure present in northern Europe. We therefore subsetted the data, separating NEU populations from RUS1, GER3, GER4, CZE1 (and SWE9, which was an outlier within NEU, **Error! Reference source not found.**b) and reanalysed them with DAPC in order to better infer fine population structure between them.

For the RADseq dataset, BIC scores suggested between 9 and 14 genetic clusters, similar to the range inferred in the microsatellite data, we therefore chose 9 clusters to take forward in the analysis. As recommended by spline interpolation, we retained 7 principal components and we kept 2 of the linear discriminants from the subsequent discriminant analysis

For Review Only

Assessment of spatial uniformity of sampling locations

Methods

In order to assess the geographic uniformity of the sampling regimes in each data subset, we used two measures of spatial patterns. The nearest neighbour distance distribution function (G), measures the distance of each sampling location to its nearest neighbour (Ripley 1991). The L-function is a transformation (for ease of interpretation) of Ripley's K-function (Ripley 1991), which measures the number of sampling locations within a given radius from each point. K has the advantage of assessing the uniformity of the sampling regime over multiple scales, as opposed to only measuring distances between closest neighbours as with G. In both cases, the estimates of G or K from our sampling locations were compared against random Poisson distributions, which would represent uniformly spaced sampling locations. 5% and 95% confidence thresholds for these Poisson distributions were also calculated to allow us to determine whether our sampling regimes significantly deviated from random ($p < 0.05$). These calculations were performed using the Gest and Lest functions (for G and L respectively) in the package "spatstats" in R (Baddeley & Turner 2005).

Results

Both methods used for the assessment of geographic uniformity of sampling locations shows that the M1 dataset locations are more patchily distributed than those of the M2, M3 and RAD datasets (**Error! Reference source not found.**).

For Review Only

Additional discussion

Population structure in northwest Europe

An intriguing result lies in the genetic similarity between populations in England with those in Belgium and Germany. *C. carassius* has been designated as native to England, however this status has been contentious in the past (Maitland 1972). Under the assumption that it is native, and considering the observed diversity and divergence times between populations across mainland Europe, we would expect to see stronger population structure between English and continental Europe, which have been separated for approximately 7800 years (Coles 2000). Given the observed diversity between populations across mainland Europe, which, according to DIYABC analysis, has arisen relatively recently. Clearly further examination of this issue is warranted and molecular data would be a value addition to the current evidence, which is predominantly anecdotal.

For Review O

SI table 1. Microsatellite loci used, grouped by their combinations in multiplex reactions. Multiplex primer mix ratios for PCR were chosen so as to give even peak strengths when analysing PCR products. Allele size ranges are those present in C. carassius for all 43 putatively pure crucian populations.

Locus	Multiplex #	Primer mix Ratios*	# Alleles	Allele size range	Ho	GenBank Accession no.	Reference
GF1	1	0.1	1	299	0	U35614	Zheng et al. 1995
GF17	1	0.1	2	182-186	0.024	U35616	Zheng et al. 1995
GF29	1	0.2	8	191-226	0.348	U35618	Zheng et al. 1995
J7	1	0.07	10	202-228	0.109	AY115095	Yue & Orban 2002
MFW2	1	0.1	1	161	0	-	Croojimans et al. 1997
Ca07	1	0.2	9	122-140	0.286	D85428	Yue & Orban 2004
TE Buffer	1	0.23					
J69	2.1	0.4	14	213-241	0.404	AY115106	Yue & Orban 2002
HJLY17	2.1	0.1	9	152-168	0.223	DQ378986	Zhi-Ying et al. 2006
HJLY35	2.1	0.1	18	261-307	0.377	DQ403242	Zhi-Ying et al. 2006
TE Buffer	2.1	0.4					
J20	2.2	0.2	9	171-218	0.149	AY115099	Yue & Orban 2002
J58	2.2	0.1	14	119-147	0.398	-	Yue & Orban 2002
MFW7	2.2	0.35	25	160-206	0.464	-	Croojimans et al. 1997
MFW17	2.2	0.35	26	185-262	0.41	-	Croojimans et al. 1997

* All primers used at 10mM per ul concentration, diluted in ddH20 from 100mM per ul stock

SI table 2. Genbank accession numbers for the mtDNA sequences used in this study.

Sample code	Accession number
FIN5_01	KT630314
FIN5_02	KT630315
FIN5_03	KT630316
EST1_02	KT630317
GER1_01	KT630318
EST1_01	KT630319
GER1_03	KT630320
FIN6_01	KT630321
FIN6_02	KT630322
FIN6_03	KT630323
BEL1_03	KT630324
EST1_03	KT630325
GER2_02	KT630326
GER2_03	KT630327
GER4_02	KT630328
NOR1_01	KT630329
NOR1_02	KT630330
SWE11_01	KT630331
SWE11_02	KT630332
SWE11_03	KT630333
RUS2_02	KT630334
RUS4_01	KT630335
RUS4_03	KT630336
FIN1_01	KT630337
FIN1_02	KT630338
FIN1_03	KT630339
FIN4_01	KT630340
FIN4_02	KT630341
FIN4_03	KT630342
POL4_01	KT630343
POL4_02	KT630344
POL4_03	KT630345
RUS1_01	KT630346
RUS1_02	KT630347
RUS1_03	KT630348
SWE8_01	KT630349
SWE8_02	KT630350
SWE8_03	KT630351
POL3_01	KT630352
POL3_02	KT630353
POL3_03	KT630354
SWE4_01	KT630355
SWE4_02	KT630356
SWE4_03	KT630357
RUS3_01	KT630358
RUS3_03	KT630359
RUS3_04	KT630360
RUS2_01	KT630361
RUS4_02	KT630362
BLS_03	KT630363
RUS3_02	KT630364
SWE3_01	KT630365
SWE3_02	KT630366
SWE3_03	KT630367
SWE2_01	KT630368
SWE2_02	KT630369
SWE2_03	KT630370
SWE9_01	KT630371
SWE9_02	KT630372
SWE9_03	KT630373
GBR7_01	KT630374
GBR6_01	KT630375
GBR8_01	KT630376
GBR8_02	KT630377
GBR8_03	KT630378
GBR6_02	KT630379
GBR6_03	KT630380
CZE1_01	KT630381
CZE1_02	KT630382
CZE1_03	KT630383
GER4_01	KT630384
GER4_03	KT630385
GER1_02	KT630386
GER2_01	KT630387
FIN3_01	KT630388
FIN3_02	KT630389
FIN3_03	KT630390
HUN1_02	KT630391
GER3_01	KT630392
GER3_02	KT630393
GER3_03	KT630394

For Review Only

SI table 3. Haplotype memberships for 101 Cytochrome B sequences used in Fig. 2.

Lineage	Haplotype	N	Drainage (n populations)	Sample code
1	1	3	Baltic	FIN5 1-3
	2	1	Baltic	EST1 2
	3	49	Elbe(2), Baltic(9), Scheldt(1), Rhine(2), North sea(2), Vistula(6), Volga(4), Don(3), Danube(1), Hunte(4)	GER1 1,3, EST1 1, 3, SWE6 1 -3, BEL1 3 , GER2 2, 3, GER4 2, NOR 1, 2, SWE11 1-3, RUS2 2, RUS4 1, 3, FIN1 1-3, FIN4 1-3, POL4 1-3, RUS1 1-3, SWE8 1-3, POL5 1-3, SWE4 1-3, RUS3 1, 3, 4, CZE2 1, GER6 1 – 4, SWE14 1, SWE15 1
	4	1	Volga	RUS2 1
	5	1	Baltic	RUS4 2
	6	1	Dnieper	BLS 3
	7	1	Volga	RUS3 2
	8	3	Baltic	SWE3 1-3
	9	2	Baltic	SWE2 1 - 3
	10	3	Baltic	SWE9 1-3
	11	13	UK(4), Rhine(1), Baltic (2)	GBR7 1, GBR6 1-3, GBR8 1-3, NET 1, GER5 1-3, GBR12 1, 2
	12	3	Baltic	FIN3 1-3
2	13	3	Danube	GER4 1, 2, AUS3 1
	14	3	Elbe(1), Rhine(1), Danube(1)	GER1 2, GER2 1, AUS2 1
	15	1	Danube	CZE1 1
	16	1	Danube	CZE1 2
	17	1	Danube	CZE1 3
	18	2	Danube	HUN 1, 2
	19	3	Danube	GER3 1-3
	23	1	Elbe	CZE2 2
	24	2	Danube	AUS1 1, 2
	25	2	Lahn	GER7 1, 2
Outgroup	20	1		Ccarp 1
	21	1		Ccarp 2
	22	1		Ccarp 3

SI table 4. Pairwise FST values calculated using the M1 dataset.

	GBR1	GBR2	GBR4	BEL1	BEL2	BEL3	FIN1	RUS4*	FIN2	CZE1	GER2	GER3	GER4	POL1	POL2	POL3	POL4	GBR7	GBR3	GBR8	GBR9	GBR11	GBR5	GBR6	GBR10	SWE4	SWE3	SWE5	FIN6	SWE7	SWE2	SWE1	SWE9	SWE10	SWE11	SWE8	FIN5	FIN3	FIN4	EST1	EST2	BLS	RUS1	DEN1	SWE12	DEN2	NOR2	SWE14	DEN3
GBR1		0.307	0.531	0.312	0.198	0.346	0.785	0.472	0.407	0.604	0.256	0.613	0.628	0.226	0.291	0.342	0.368	0.436	0.364	0.378	0.518	0.317	0.517	0.302	0.376	0.479	0.444	0.419	0.458	0.542	0.591	0.404	0.839	0.548	0.793	0.39	0.428	0.72	0.596	0.628	0.526	0.491	0.623	0.319	0.626	0.261	0.768	0.457	0.233
GBR2	NS		0.67	0.316	0.247	0.366	0.783	0.482	0.446	0.588	0.332	0.6	0.618	0.266	0.309	0.357	0.378	0.611	0.535	0.562	0.716	0.381	0.651	0.451	0.501	0.476	0.478	0.443	0.518	0.566	0.594	0.396	0.853	0.616	0.826	0.454	0.444	0.725	0.572	0.645	0.522	0.459	0.59	0.346	0.664	0.357	0.864	0.454	0.268
GBR4	*	NS		0.588	0.445	0.532	0.774	0.498	0.327	0.69	0.267	0.708	0.716	0.19	0.325	0.315	0.484	0.15	0.401	0.288	0.223	0.248	0.185	0.432	0.145	0.508	0.41	0.433	0.422	0.543	0.57	0.402	0.817	0.506	0.774	0.501	0.439	0.717	0.601	0.663	0.497	0.488	0.683	0.472	0.648	0.362	0.627	0.525	0.312
BEL1	*	NS	*		0.065	0.023	0.732	0.479	0.427	0.601	0.253	0.609	0.617	0.284	0.293	0.359	0.347	0.512	0.36	0.442	0.523	0.295	0.502	0.291	0.436	0.449	0.447	0.446	0.483	0.524	0.586	0.412	0.8	0.583	0.75	0.47	0.436	0.696	0.569	0.614	0.481	0.467	0.608	0.363	0.569	0.362	0.73	0.462	0.283
BEL2	*	NS	*	NS		0	0.711	0.438	0.363	0.571	0.195	0.582	0.588	0.193	0.24	0.288	0.296	0.396	0.24	0.361	0.38	0.156	0.356	0.249	0.278	0.39	0.395	0.374	0.393	0.465	0.525	0.359	0.779	0.536	0.705	0.425	0.359	0.673	0.508	0.558	0.394	0.398	0.57	0.327	0.523	0.287	0.683	0.422	0.198
BEL3	NS	NS	*	NS	NS		0.724	0.447	0.382	0.563	0.204	0.573	0.581	0.232	0.249	0.303	0.296	0.472	0.306	0.423	0.482	0.215	0.439	0.279	0.353	0.407	0.412	0.381	0.418	0.474	0.54	0.368	0.807	0.561	0.731	0.462	0.369	0.686	0.521	0.577	0.41	0.39	0.559	0.352	0.534	0.34	0.738	0.428	0.233
FIN1	*	NS	*	*	*	*		0.498	0.537	0.742	0.586	0.746	0.745	0.513	0.475	0.508	0.532	0.745	0.761	0.738	0.797	0.695	0.763	0.718	0.737	0.419	0.515	0.532	0.587	0.627	0.55	0.437	0.75	0.642	0.756	0.685	0.56	0.569	0.43	0.521	0.456	0.487	0.717	0.632	0.697	0.666	0.676	0.485	0.591
RUS4*	*	*	*	*	*	*	*		0.309	0.484	0.33	0.506	0.51	0.311	0.3	0.286	0.334	0.462	0.416	0.482	0.5	0.41	0.434	0.442	0.437	0.291	0.301	0.215	0.191	0.354	0.367	0.262	0.555	0.38	0.462	0.433	0.286	0.494	0.304	0.28	0.113	0.231	0.495	0.367	0.455	0.371	0.522	0.27	0.317
FIN2	*	NS	*	*	*	*	*	*		0.488	0.225	0.526	0.521	0.191	0.142	0.125	0.235	0.286	0.302	0.325	0.395	0.286	0.314	0.312	0.302	0.284	0.142	0.166	0.161	0.212	0.295	0.172	0.649	0.271	0.482	0.271	0.182	0.442	0.289	0.28	0.137	0.168	0.484	0.265	0.264	0.206	0.448	0.193	0.159
CZE1	NS	NS	*	*	*	*	*	*	*		0.38	0.342	0.364	0.43	0.421	0.364	0.462	0.573	0.546	0.572	0.672	0.596	0.637	0.555	0.571	0.471	0.444	0.347	0.408	0.445	0.587	0.456	0.791	0.555	0.615	0.448	0.395	0.69	0.535	0.479	0.402	0.388	0.477	0.384	0.484	0.44	0.677	0.418	0.408
GER2	*	NS	*	*	*	*	*	*	*	*		0.379	0.381	0.146	0.189	0.181	0.232	0.142	0.111	0.113	0.269	0.177	0.226	0.139	0.168	0.263	0.256	0.2	0.186	0.275	0.39	0.226	0.654	0.355	0.507	0.207	0.22	0.552	0.351	0.358	0.228	0.237	0.458	0.168	0.337	0.146	0.453	0.299	0.128
GER3	NS	NS	*	*	*	NS	*	*	*	NS	*		0.113	0.445	0.445	0.397	0.48	0.579	0.543	0.567	0.673	0.61	0.649	0.542	0.57	0.502	0.492	0.402	0.454	0.492	0.609	0.489	0.805	0.589	0.642	0.438	0.441	0.708	0.532	0.499	0.435	0.412	0.472	0.411	0.54	0.467	0.691	0.47	0.435
GER4	NS	NS	*	*	*	NS	*	*	*	NS	*	NS		0.442	0.443	0.399	0.465	0.584	0.553	0.569	0.687	0.612	0.661	0.546	0.575	0.488	0.487	0.387	0.45	0.494	0.61	0.486	0.812	0.593	0.657	0.439	0.435	0.697	0.54	0.501	0.435	0.405	0.492	0.415	0.542	0.481	0.703	0.463	0.431
POL1	*	*	*	*	*	*	*	*	*	*	*	*	*		0.105	0.074	0.191	0.182	0.202	0.242	0.21	0.153	0.175	0.237	0.105	0.218	0.195	0.194	0.183	0.246	0.3	0.187	0.587	0.317	0.477	0.235	0.186	0.487	0.259	0.298	0.156	0.161	0.426	0.194	0.314	0.138	0.356	0.246	0.11
POL2	*	*	*	*	*	*	*	*	*	*	*	*	*	*		0.061	0.113	0.292	0.253	0.317	0.358	0.237	0.298	0.243	0.242	0.241	0.148	0.149	0.169	0.111	0.219	0.146	0.598	0.266	0.417	0.244	0.112	0.438	0.228	0.239	0.125	0.114	0.427	0.203	0.184	0.157	0.422	0.17	0.124
POL3	*	NS	*	*	*	NS	*	*	*	*	*	*	*	*	*		0.142	0.31	0.271	0.368	0.392	0.234	0.274	0.294	0.227	0.246	0.16	0.16	0.185	0.214	0.283	0.154	0.642	0.253	0.448	0.26	0.154	0.456	0.197	0.261	0.086	0.057	0.355	0.203	0.268	0.155	0.427	0.194	0.117
POL4	*	NS	*	*	*	*	*	*	*	*	*	NS	*	*	*	*		0.416	0.301	0.418	0.491	0.323	0.413	0.281	0.358	0.263	0.285	0.184	0.246	0.24	0.34	0.22	0.69	0.391	0.547	0.344	0.211	0.446	0.269	0.269	0.204	0.177	0.464	0.266	0.286	0.261	0.53	0.257	0.202
GBR7	*	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		0.153	0.072	0.364	0.164	0.244	0.286	0.134	0.497	0.405	0.388	0.391	0.514	0.529	0.321	0.8	0.452	0.74	0.355	0.426	0.685	0.542	0.63	0.424	0.406	0.608	0.37	0.606	0.277	0.637	0.499	0.279
GBR3	NS	NS	*	NS	NS	NS	*	*	*	NS	NS	NS	NS	*	*	*	*	NS		0.021	0.422	0.09	0.336	0.097	0.22	0.435	0.387	0.322	0.343	0.479	0.525	0.297	0.827	0.516	0.751	0.284	0.364	0.673	0.509	0.573	0.396	0.394	0.592	0.232	0.591	0.182	0.752	0.442	0.175
GBR8	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	NS		0.42	0.184	0.31	0.181	0.22	0.518	0.444	0.426	0.424	0.534	0.561	0.356	0.784	0.479	0.734	0.301	0.464	0.686	0.564	0.636	0.453	0.447	0.631	0.332	0.605	0.254	0.661	0.524	0.287
GBR9	*	NS	*	*	*	*	*	*	*	*	*	*	NS	*	*	*	*	*	*	*		0.205	0.021	0.38	0.159	0.577	0.483	0.528	0.517	0.661	0.621	0.458	0.841	0.528	0.814	0.61	0.529	0.728	0.651	0.723	0.519	0.495	0.652	0.553	0.751	0.504	0.757	0.608	0.395
GBR11	*	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	NS	*	*		0.178	0.235	0.138	0.369	0.346	0.342	0.336	0.438	0.475	0.285	0.746	0.509	0.689	0.368	0.344	0.641	0.46	0.542	0.342	0.384	0.603	0.287	0.52	0.211	0.584	0.418	0.161
GBR5	NS	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	NS	*		0.339	0.161	0.452	0.367	0.387	0.365	0.538	0.555	0.375	0.819	0.489	0.759	0.489	0.398	0.681	0.561	0.604	0.401	0.415	0.619	0.422	0.645	0.366	0.655	0.483	0.27
GBR6	*	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	NS	*	*	*	*		0.278	0.452	0.39	0.358	0.36	0.463	0.533	0.366	0.773	0.474	0.686	0.293	0.387	0.634	0.513	0.519	0.398	0.413	0.599	0.272	0.511	0.235	0.666	0.429	0.228
GBR10	NS	NS	*	*	*	*	*	*	*	*	*	NS	NS	*	*	*	*	*	NS	*	*	*	*	*		0.403	0.352	0.332	0.346	0.447	0.478	0.325	0.787	0.469	0.703	0.376	0.335	0.662	0.481	0.537	0.378	0.365	0.571	0.341	0.54	0.284	0.583	0.427	0.221
SWE4	*	NS	*	*	*	NS	*	*	*	NS	*	*	NS	*	*	*	*	*	NS	*	*	*	*	*	*		0.233	0.235	0.192	0.363	0.329	0.176	0.652	0.357	0.578	0.458	0.224	0.436	0.224	0.294	0.132	0.222	0.47	0.351	0.473	0.378	0.507	0.294	0.25
SWE3	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		0.166	0.115	0.188	0.209	0.134	0.652	0.116	0.378	0.358	0.117	0.465	0.307	0.292	0.083	0.171	0.447	0.32	0.315	0.292	0.519	0.175	0.205
SWE5	NS	NS	*	*	*	NS	*	*	*	NS	*	NS	NS	*	*	*	*	*	NS	*	*	*	*	*	NS	NS	*		0.084	0.168	0.248	0.115	0.625	0.168	0.404	0.337	0.103	0.462	0.235	0.214	0.064	0.113	0.378	0.274	0.26	0.271	0.513	0.137	0.191
FIN6	NS	NS	*	NS	NS	NS	*	*	*	NS	NS	NS	NS	*	*	NS	NS	*	NS	*	*	*	*	NS	NS	NS	NS	NS		0.258	0.295	0.141	0.687	0.135	0.429	0.385	0.127	0.532	0.32	0.29	0.08	0.175	0.426	0.311	0.411	0.294	0.63	0.229	0.187
SWE7	NS	NS	*	*	*	NS	*	*	*	*	*	NS	*	*	*	*	*	*	NS	*	*	*	*	*	*	*	*	NS	NS		0.205	0.141	0.77	0.279	0.501	0.406	0.12	0.555	0.37	0.36	0.253	0.195	0.463	0.362	0.15	0.345	0.641	0.201	0.297
SWE2	*	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		0.129	0.695	0.389	0.515	0.491	0.202	0.495	0.329	0.334	0.266	0.29	0.585	0.433	0.448	0.435	0.567	0.235	0.361
SWE1	*	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		0.589	0.318	0.439	0.281	0.136	0.389	0.193	0.205	0.108	0.157	0.489	0.2	0.29	0.201	0.368	0.168	0.174
SWE9	*	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		0.721	0.838	0.768	0.686	0.753	0.706	0.756	0.65	0.699	0.776	0.734	0.829	0.753	0.828	0.661	0.702
SWE10	*	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		0.444	0.437	0.176	0.578	0.477	0.374	0.274	0.331	0.558	0.419	0.407	0.386	0.61	0.25	0.3
SWE11	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS		0.64	0.36	0.701	0.642	0.636	0.381	0.442	0.62	0.605	0.64	0.642	0.851	0.378	0.545
SWE8	NS	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	NS		0.387	0.604	0.429	0.472	0.342	0.309	0.532	0.181	0.441	0.159	0.638	0.363	0.233
FIN5	*	NS	*	*	*	NS	*	*	*	*	*	NS	NS	*	*	*	*	*	NS	*	*	*	*	*	*	NS	*	NS	NS	*	*	*	*	*	NS	*		0.48	0.288	0.265	0.118	0.155	0.405	0.314	0.271	0.295	0.561	0.178	0.199
FIN3	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	NS	*	*		0.402	0.357	0.43	0.423	0.687	0.566	0.606	0.605	0.657	0.464	0.528
FIN4	NS	NS	*	*	*	NS	*	*	*	NS	*	NS	NS	*	*	*	*	*	NS	*	*	*	*	*	*	NS	*	NS	NS	NS	*	*	*	*	NS	*	NS	*		0.23	0.166	0.165	0.519	0.34	0.458	0.361	0.462	0.267	0.269
EST1	NS	NS	*	NS	NS	NS	*	*	*	NS	*	NS	NS	*	*	*	NS	*	NS	*	NS	*	NS	*	NS	NS	*	NS	NS	NS	*	*	*	*	NS	NS	NS	*	NS		0.158	0.191	0.474	0.39	0.461	0.452	0.674	0.234	0.34
EST2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA		0.042	0.367	0.253	0.356	0.274	0.504	0.119	0.146
BLS	NS	NS	*	*	*	NS	*	*	*	NS	*	NS	NS	*	*	*	*	*	NS	*	*	*	*	*	NS	*	*	NS	NS	*	*	*	*	*	NS	*	*	*	NS	NS	NA		0.364	0.245	0.269	0.265	0.406	0.138	0.181
RUS1	NS	NS	*	*	*	NS	*	*	*	*	*	NS	NS	*	*	*	*	*	NS	*	*	*	*	*	NS	NS	*	NS	NS	*	*	*	*	*	NS	*	*	*	NS	NS	NA	NS		0.483	0.467	0.503	0.646	0.405	0.437
DEN1	NS	NS	*	*	*	*	*	*	*	NS	*	NS	NS	*	*	*	*	*	NS	*	*	*	*	*	NS	NS	*	*	NS	*	*	*	*	*	NS	*	*	*	NS	NS	NA	NS	*		0.428	0.121	0.588	0.316	0.132
SWE12	NS	NS	*	*	NS	NS	*	*	*	*	*	NS	NS	*	*	*	*	*	NS	*	*	*	NS	*	*	NS	*	NS	NS	NS	NS	*	*	*	NS	NS	NS	*	NS	NS	NA	NS	NS	NS		0.431	0.723	0.265	0.359
DEN2	*	NS	*	*	*	*	*	*	*	*	*	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	NS	*	*	*	*	NS	NA	NS	*	NS	*		0.611	0.335	0.099
NOR2	*	NS	*	*	*	*	*	*	*	*	*	*	NS	*	*	*	*	*	NS	*	*	*	*	*	*	*	*	*	NS	*	*	*	*	*	NS	*	*	*	*	NS	NA	*	*	*	NS	*		0.532	0.541
SWE14	*	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	NS	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	NS	*	NS	*	NS	NS	NA	NS	*	*	NS	*	*		0.261
DEN3	*	NS	*	*	*	*	*	*	*	*	*	NS	NS	*	*	*	*	*	NS	*	*	*	*	*	*	*	*	*	NS	*	*	*	*	*	NS	*	*	*	NS	NS	NA	*	*	*	NS	*	*	*	

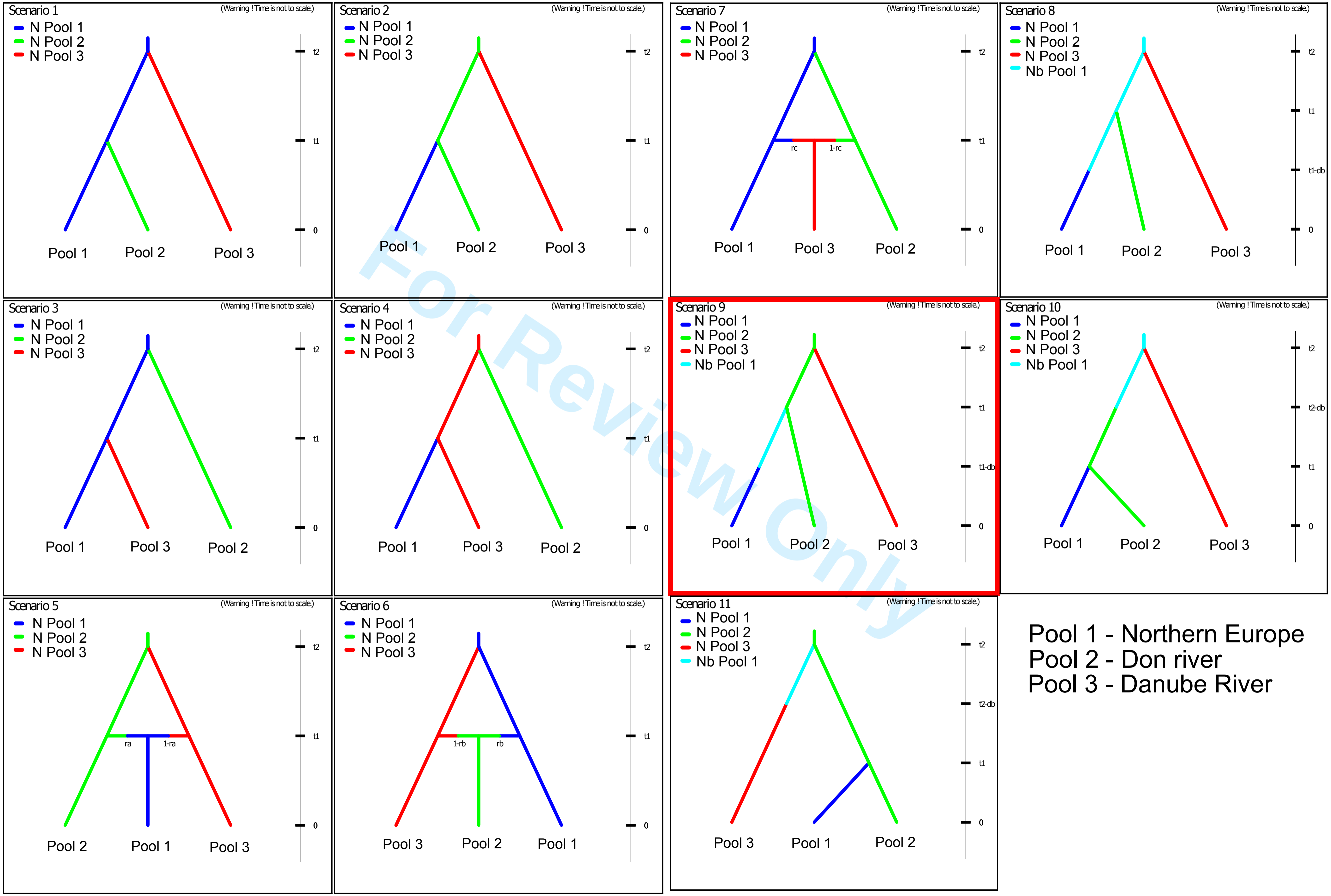
For Review Only

SI table 5. Pairwise FST values calculated using the RADseq dataset.

	GBR8	BEL1	GBR4	FIN3	DEN1	GBR7	SWE12	FIN4	DEN2	POL4	RUS1	SWE2	SWE8	SWE9	SWE10	NOR2	POL3	HUN2	WEN
GBR8		0.34971	0.35695	0.49475	0.223897	0.35613	0.406544	0.295019	0.293628	0.211876	0.38775	0.308973	0.273693	0.412263	0.321365	0.650207	0.146766	0.61801	0.397239
BEL1			0.370425	0.390916	0.098308	0.381154	0.300836	0.225496	0.130398	0.152617	0.343954	0.22423	0.08032	0.326848	0.235947	0.522507	0.103445	0.597677	0.31111
GBR4				0.513779	0.231153	0.195241	0.423664	0.302246	0.316185	0.218776	0.392539	0.314975	0.284155	0.422534	0.331921	0.698989	0.149208	0.620806	0.412409
FIN3					0.308284	0.517114	0.341754	0.198275	0.364426	0.222674	0.378729	0.27048	0.328488	0.331267	0.286862	0.562015	0.149991	0.614832	0.341565
DEN1						0.244594	0.239562	0.194342	0.06762	0.136982	0.356985	0.182005	0.085513	0.266461	0.190793	0.362014	0.102429	0.602815	0.237037
GBR7							0.430574	0.31162	0.32391	0.229621	0.396753	0.319608	0.295939	0.433712	0.340292	0.692819	0.157339	0.621918	0.422803
SWE12								0.209406	0.282835	0.173199	0.363912	0.198857	0.259513	0.303204	0.211775	0.459576	0.122381	0.606576	0.250115
FIN4									0.218225	0.142389	0.328888	0.154803	0.211809	0.203425	0.174944	0.316929	0.099233	0.586541	0.198636
DEN2										0.153556	0.362177	0.212179	0.101801	0.307702	0.222051	0.459347	0.108029	0.60623	0.284015
POL4											0.321777	0.128672	0.150743	0.192186	0.138734	0.250273	0.073063	0.579543	0.161299
RUS1												0.341129	0.358602	0.368371	0.349288	0.396194	0.278006	0.516158	0.358584
SWE2													0.19768	0.218326	0.145195	0.325228	0.094924	0.591579	0.151258
SWE8														0.289356	0.208013	0.401551	0.110433	0.604134	0.262799
SWE9															0.257245	0.429544	0.136442	0.607715	0.29715
SWE10																0.350951	0.100275	0.598136	0.184722
NOR2																	0.165304	0.625602	0.426179
POL3																		0.547371	0.111018
HUN2																			0.604399

For Review Only

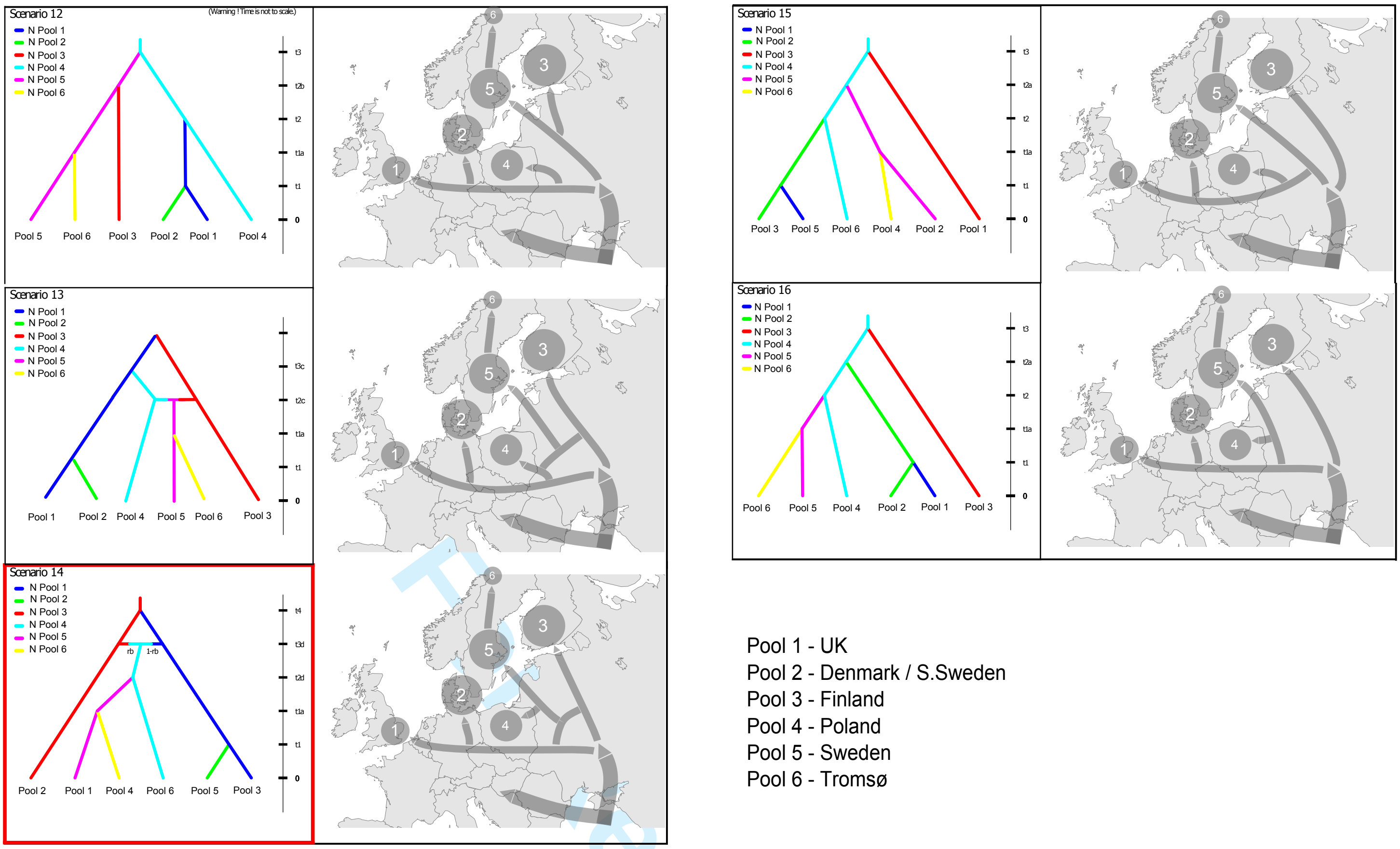
For Review Only



Pool 1 - Northern Europe
Pool 2 - Don river
Pool 3 - Danube River

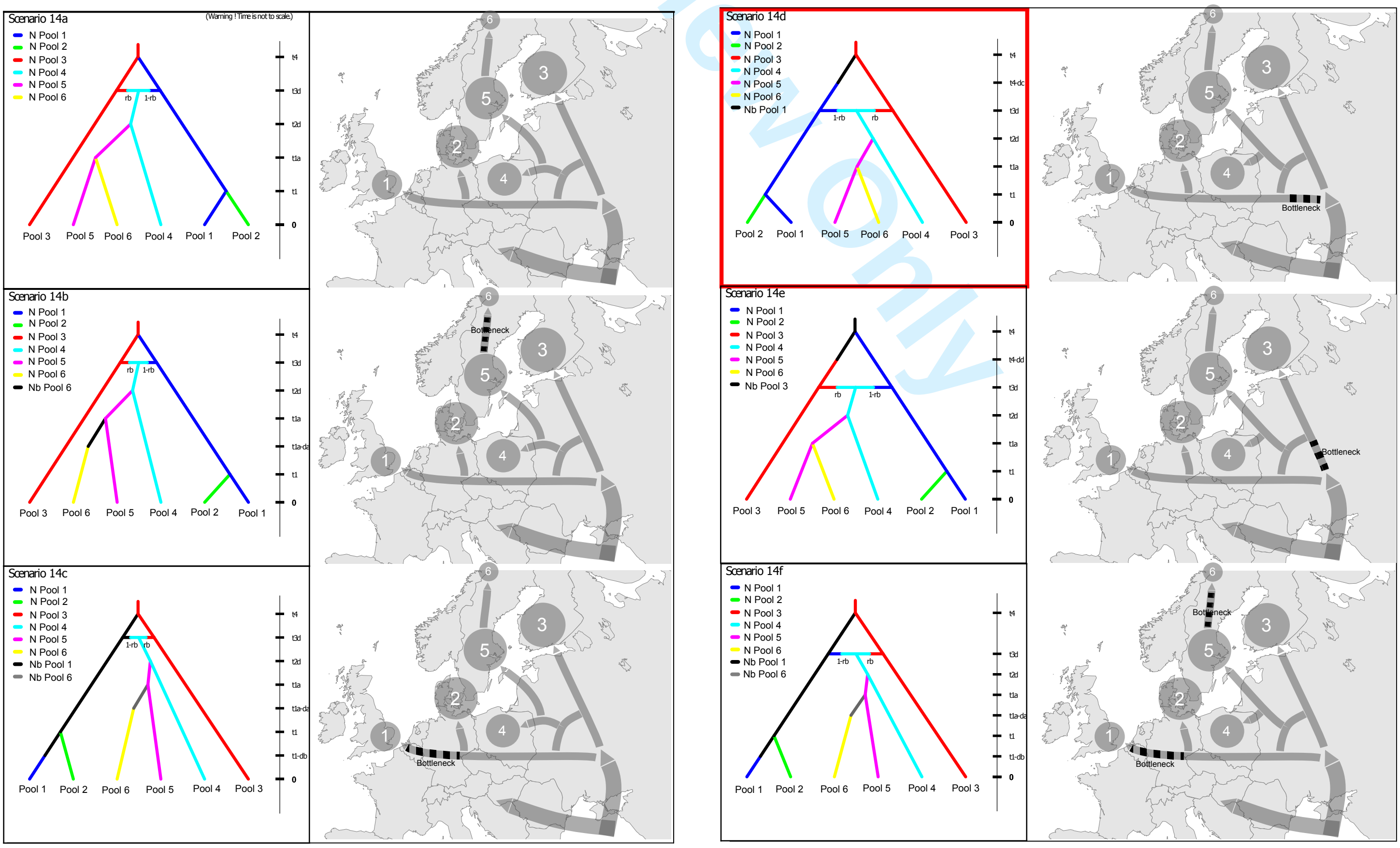
For Review Only

a) Stage 2. NEU Major variants

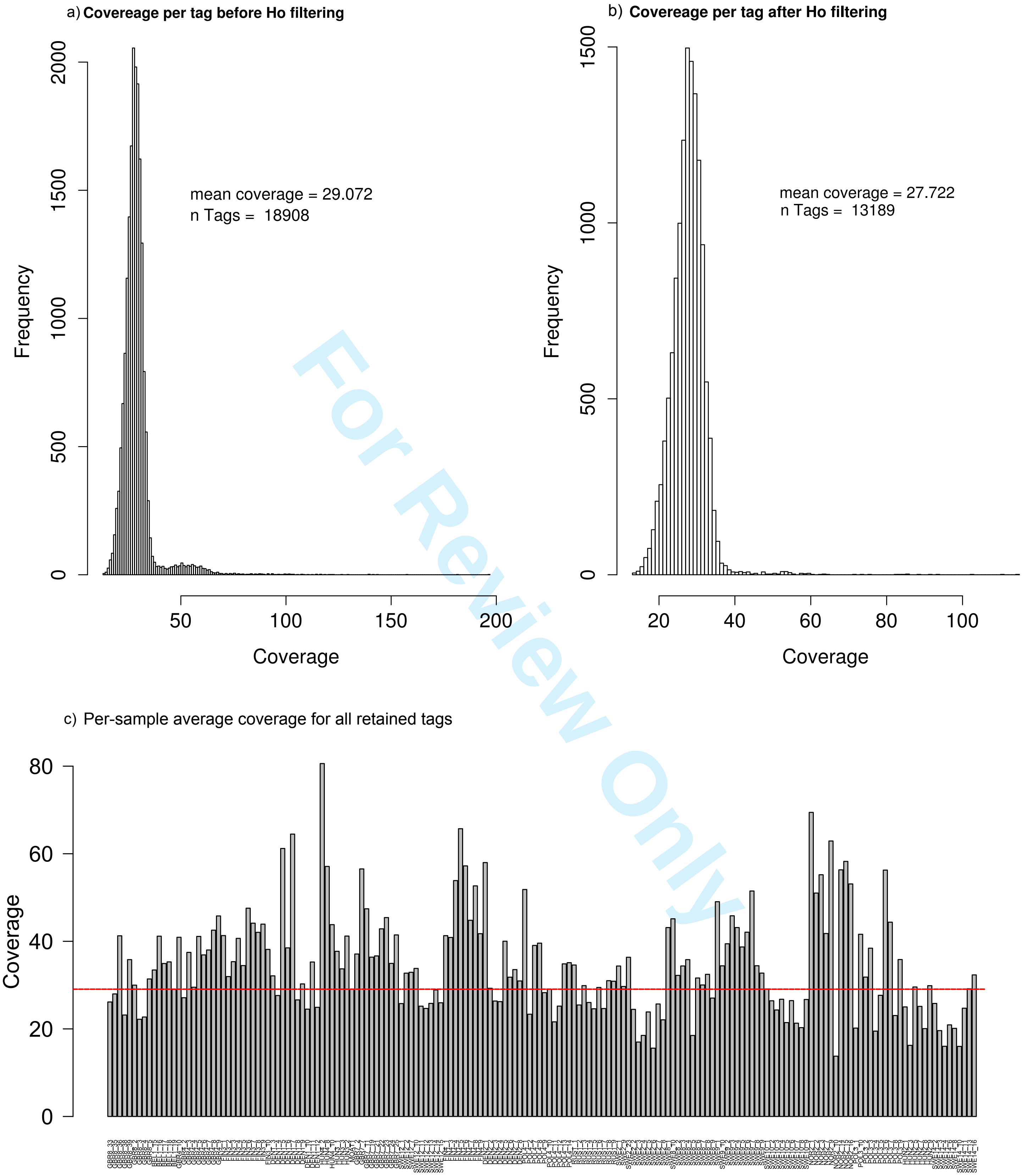

Scenario 12
(Warning ! Time is not to scale.)
N Pool 1
N Pool 2
N Pool 3
N Pool 4
N Pool 5
N Pool 6
Pool 5 Pool 6 Pool 3 Pool 2 Pool 1 Pool 4
t3 t2b t2 t1a t1 0
Scenario 13
Pool 1 Pool 2 Pool 4 Pool 5 Pool 6 Pool 3
t3c t2c t1a t1 0
Scenario 14
rb 1-rb
Pool 2 Pool 1 Pool 4 Pool 6 Pool 5 Pool 3
t4 t3d t2d t1a t1 0
Scenario 15
Pool 3 Pool 5 Pool 6 Pool 4 Pool 2 Pool 1
t3 t2a t2 t1a t1 0
Scenario 16
Pool 6 Pool 5 Pool 4 Pool 2 Pool 1 Pool 3
t3 t2a t2 t1a t1 0
1
2
3
4
5
6

Pool 1 - UK
Pool 2 - Denmark / S.Sweden
Pool 3 - Finland
Pool 4 - Poland
Pool 5 - Sweden
Pool 6 - Tromsø

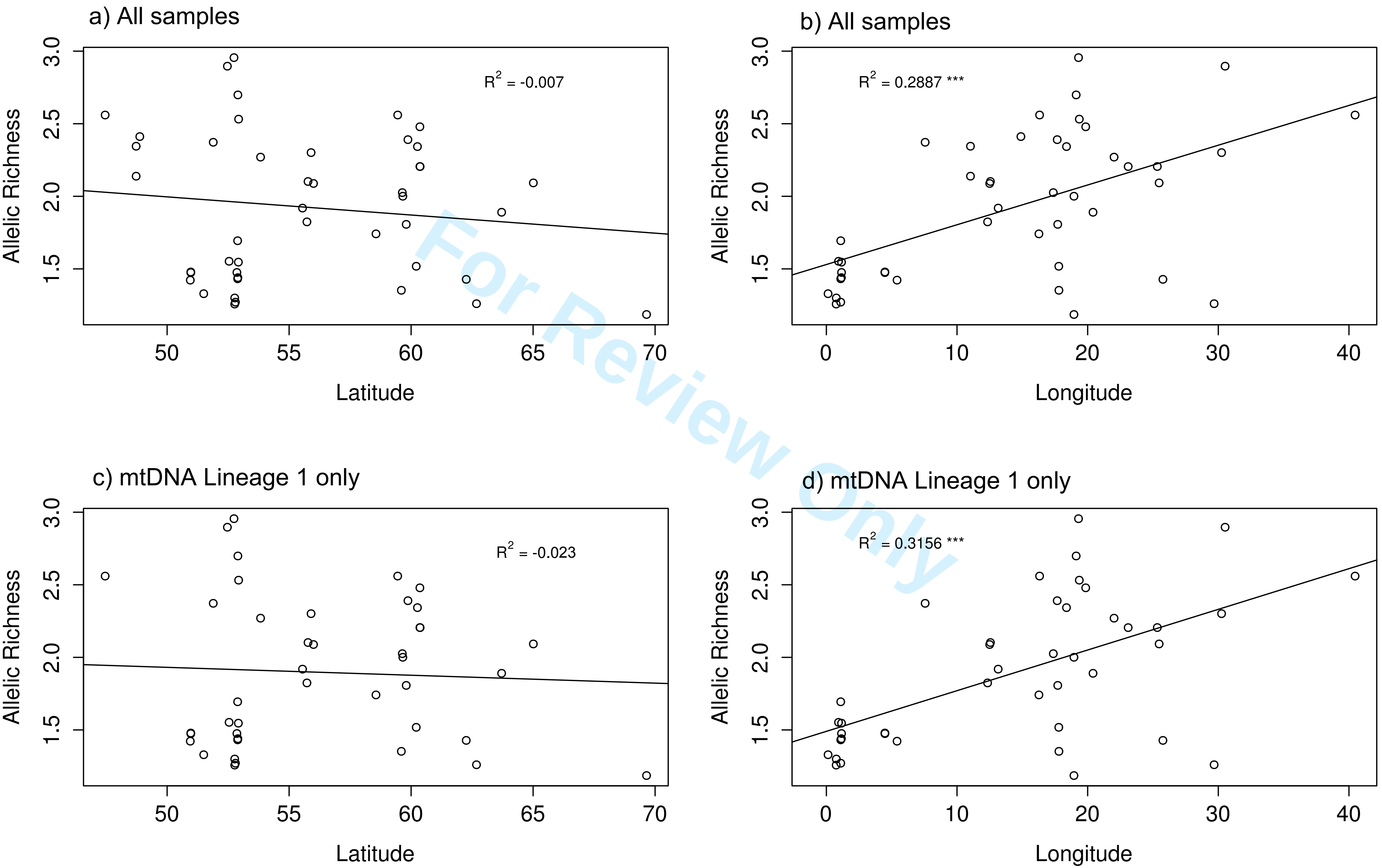
b) Stage 3. Scenario 14 Minor variants


Scenario 14a
(Warning ! Time is not to scale.)
N Pool 1
N Pool 2
N Pool 3
N Pool 4
N Pool 5
N Pool 6
rb 1-rb
Pool 3 Pool 5 Pool 6 Pool 4 Pool 1 Pool 2
t4 t3d t2d t1a t1 0
Scenario 14b
Nb Pool 6
Pool 3 Pool 6 Pool 5 Pool 4 Pool 2 Pool 1
t4 t3d t2d t1a t1a-db t1 0
Bottleneck
Scenario 14c
Nb Pool 1
Nb Pool 6
1-rb rb
Pool 1 Pool 2 Pool 6 Pool 5 Pool 4 Pool 3
t4 t3d t2d t1a t1a-db t1 t1-db 0
Bottleneck
Scenario 14d
Nb Pool 1
Pool 2 Pool 1 Pool 5 Pool 6 Pool 4 Pool 3
t4 t4-dc t3d t2d t1a t1 0
Bottleneck
Scenario 14e
Nb Pool 3
Pool 3 Pool 5 Pool 6 Pool 4 Pool 2 Pool 1
t4 t4-dd t3d t2d t1a t1 0
Bottleneck
Scenario 14f
Nb Pool 1
Nb Pool 6
Pool 1 Pool 2 Pool 6 Pool 5 Pool 4 Pool 3
t4 t3d t2d t1a t1a-db t1 t1-db 0
Bottleneck
1
2
3
4
5
6

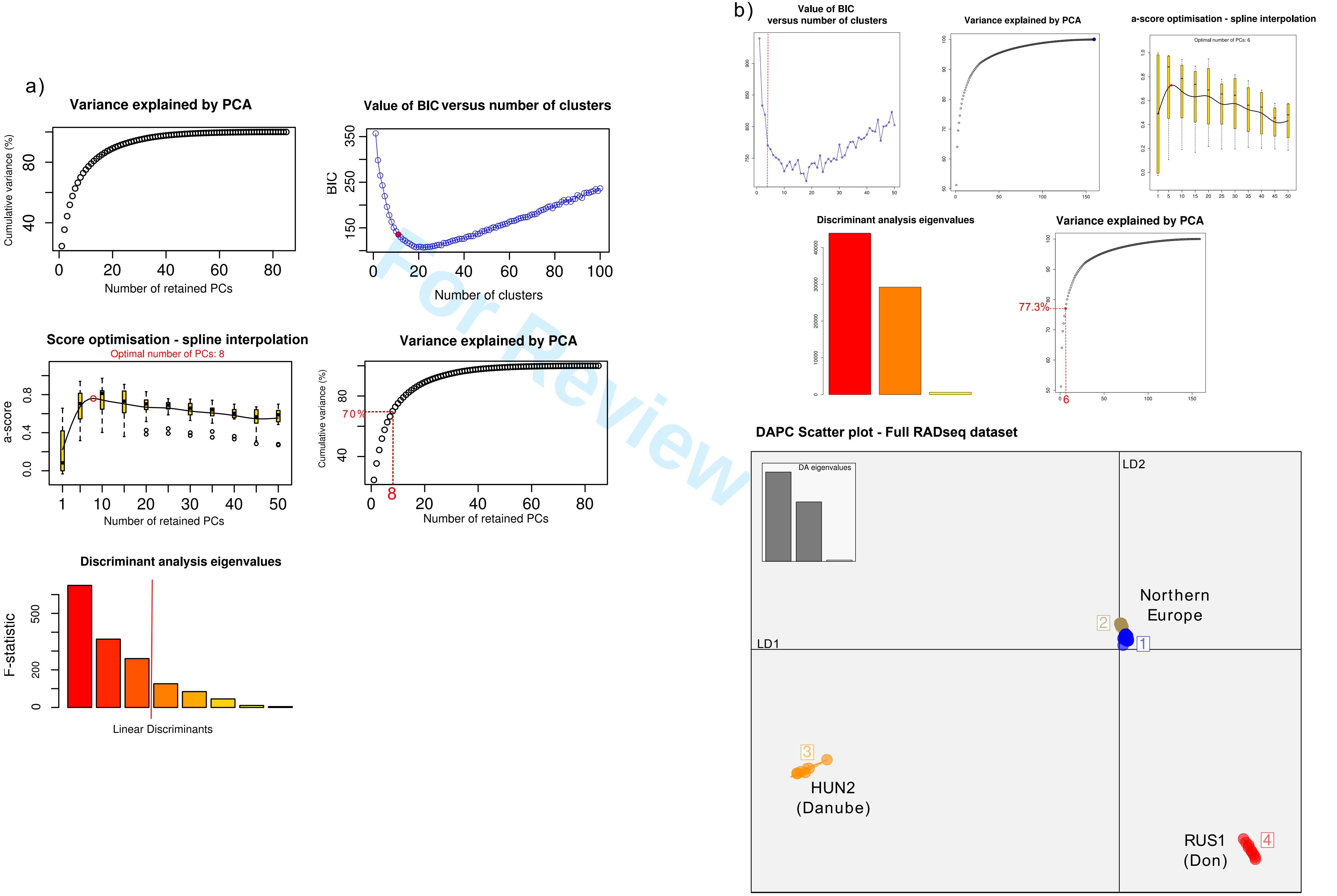
For Review Only



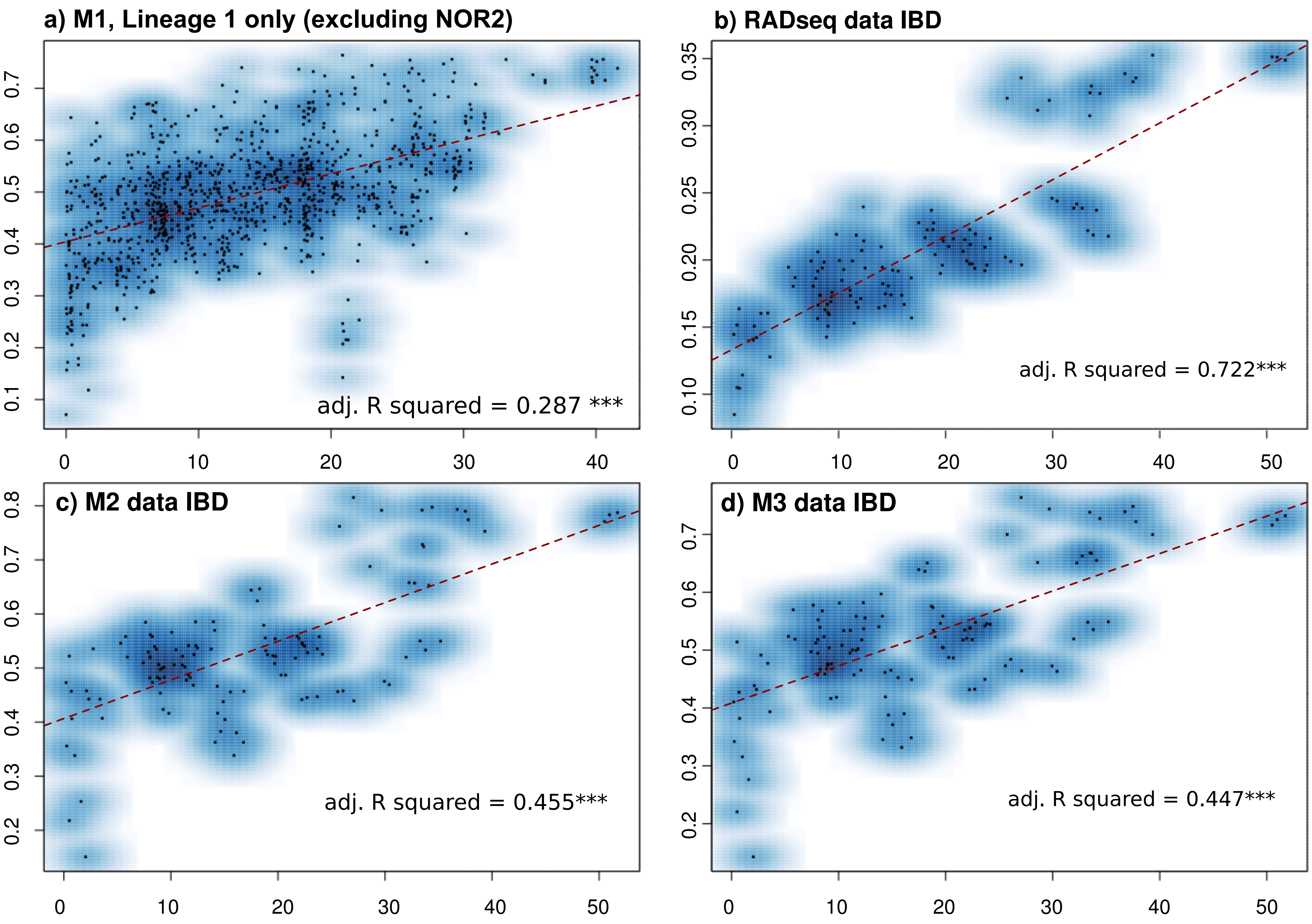
For Review Only



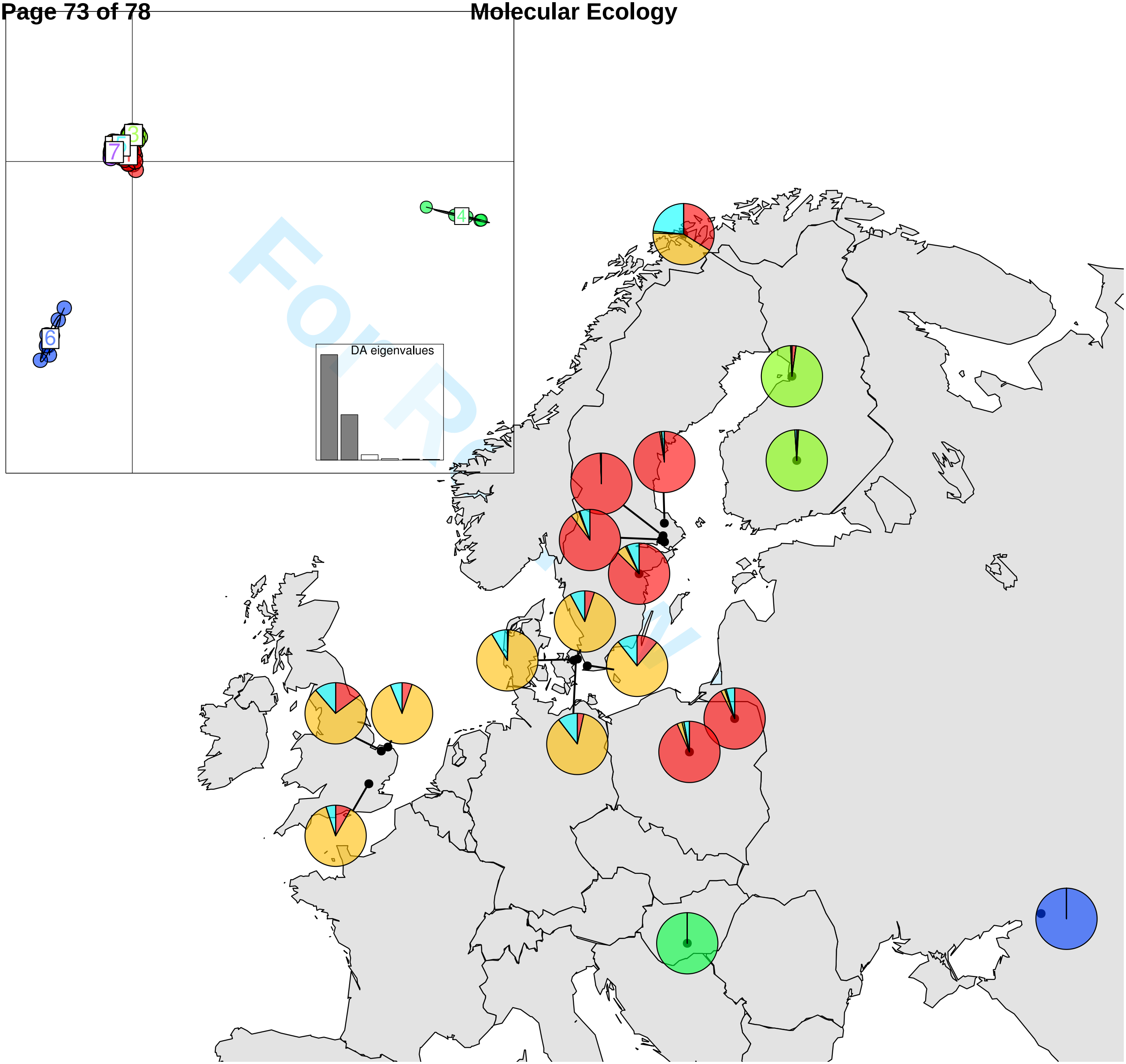
For Review Only

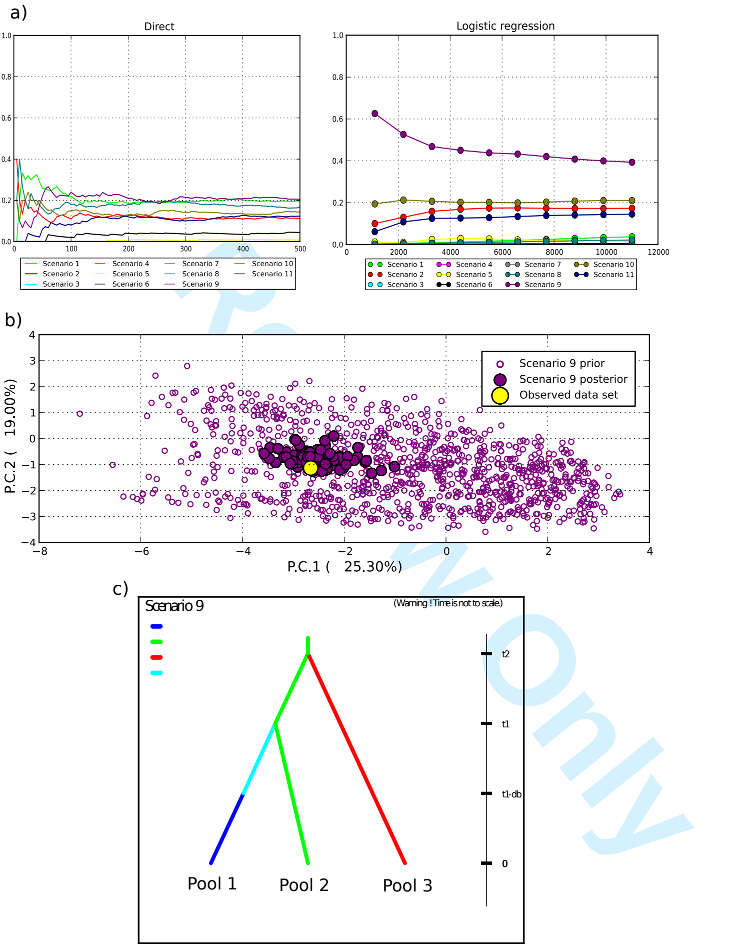

a)
Variance explained by PCA
Cumulative variance (%)
40
80
0
20
40
60
80
Number of retained PCs
Value of BIC versus number of clusters
BIC
150
250
350
0
20
40
60
80
100
Number of clusters
Score optimisation - spline interpolation
Optimal number of PCs: 8
a-score
0.0
0.4
0.8
1
10
20
30
40
50
Number of retained PCs
Variance explained by PCA
Cumulative variance (%)
40
80
70%
8
0
20
40
60
80
Number of retained PCs
Discriminant analysis eigenvalues
F-statistic
0
200
500
Linear Discriminants
b)
Value of BIC versus number of clusters
750
800
850
900
0
10
20
30
40
50
Variance explained by PCA
50
60
70
80
90
100
0
50
100
150
a-score optimisation - spline interpolation
Optimal number of PCs: 6
0.0
0.2
0.4
0.6
0.8
1.0
1
5
10
15
20
25
30
35
40
45
50
Discriminant analysis eigenvalues
0
10000
20000
30000
40000
Variance explained by PCA
77.3%
50
60
70
80
90
100
0
6
50
100
150
DAPC Scatter plot - Full RADseq dataset
DA eigenvalues
LD2
LD1
Northern Europe
2
1
3
HUN2 (Danube)
RUS1 (Don)
4

For Review

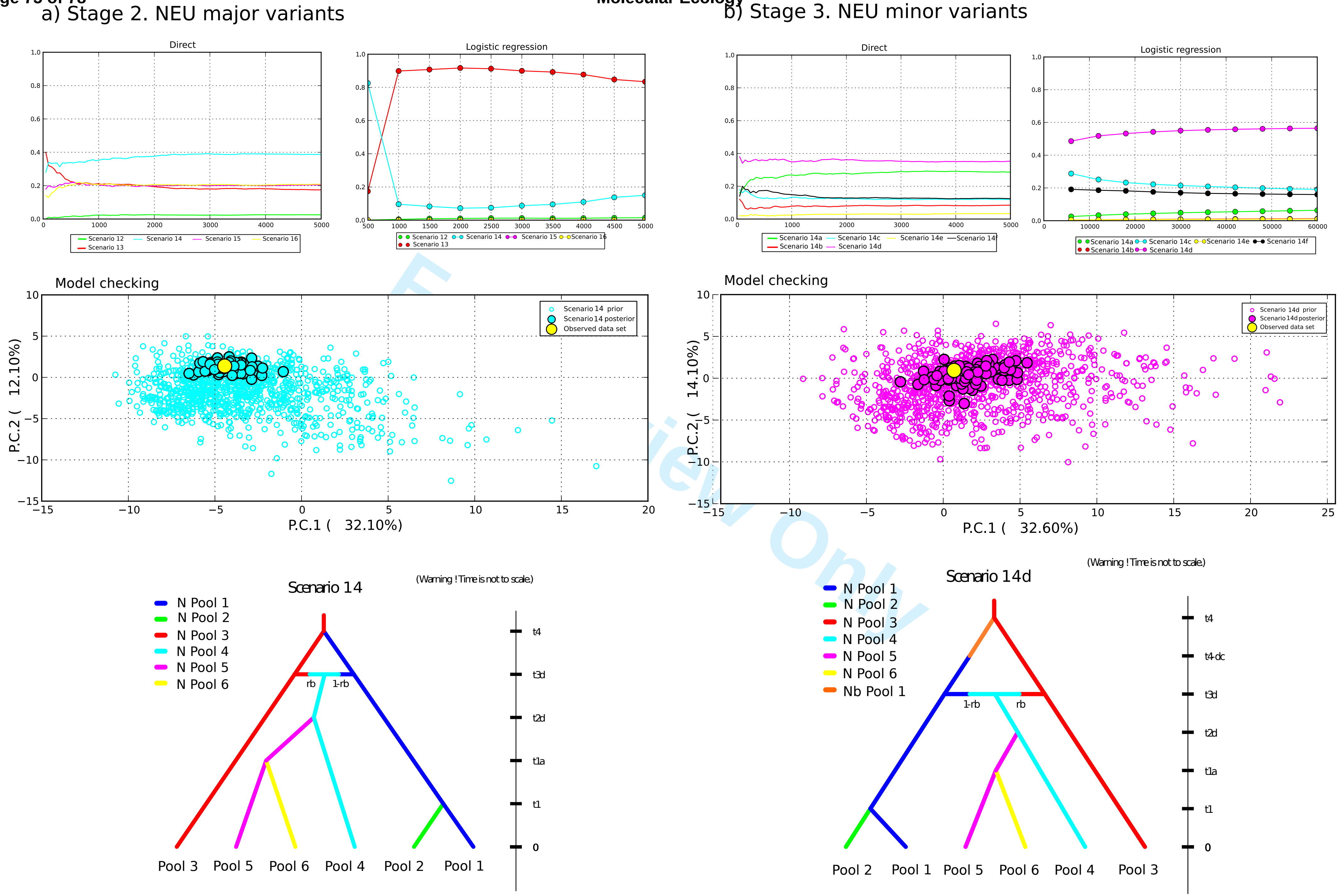

a) M1, Lineage 1 only (excluding NOR2)
adj. R squared = 0.287 ***
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
b) RADseq data IBD
adj. R squared = 0.722***
0.10
0.15
0.20
0.25
0.30
0.35
0
10
20
30
40
50
c) M2 data IBD
adj. R squared = 0.455***
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
10
20
30
40
50
d) M3 data IBD
adj. R squared = 0.447***
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
50

For Review Only





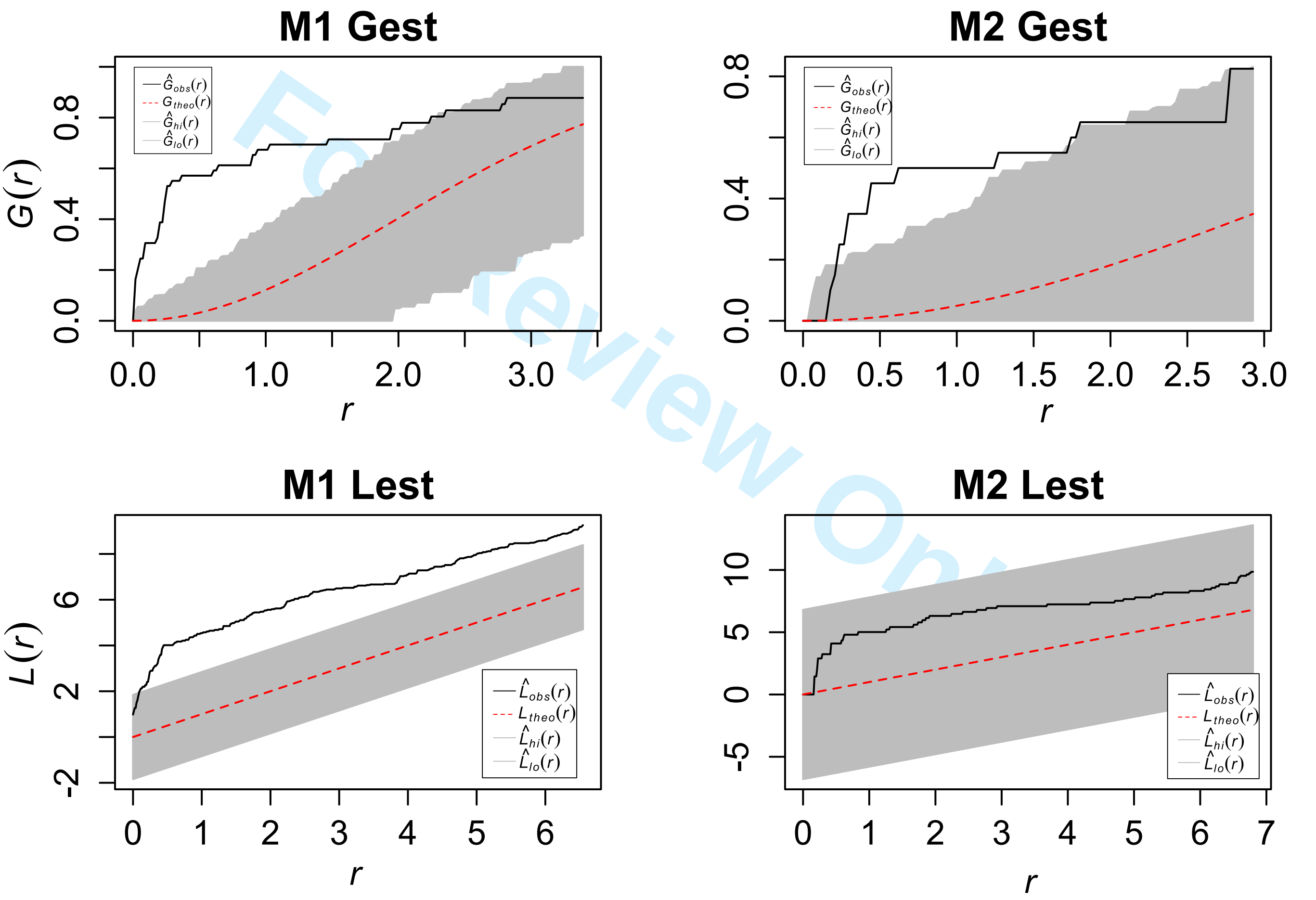
For Review Only

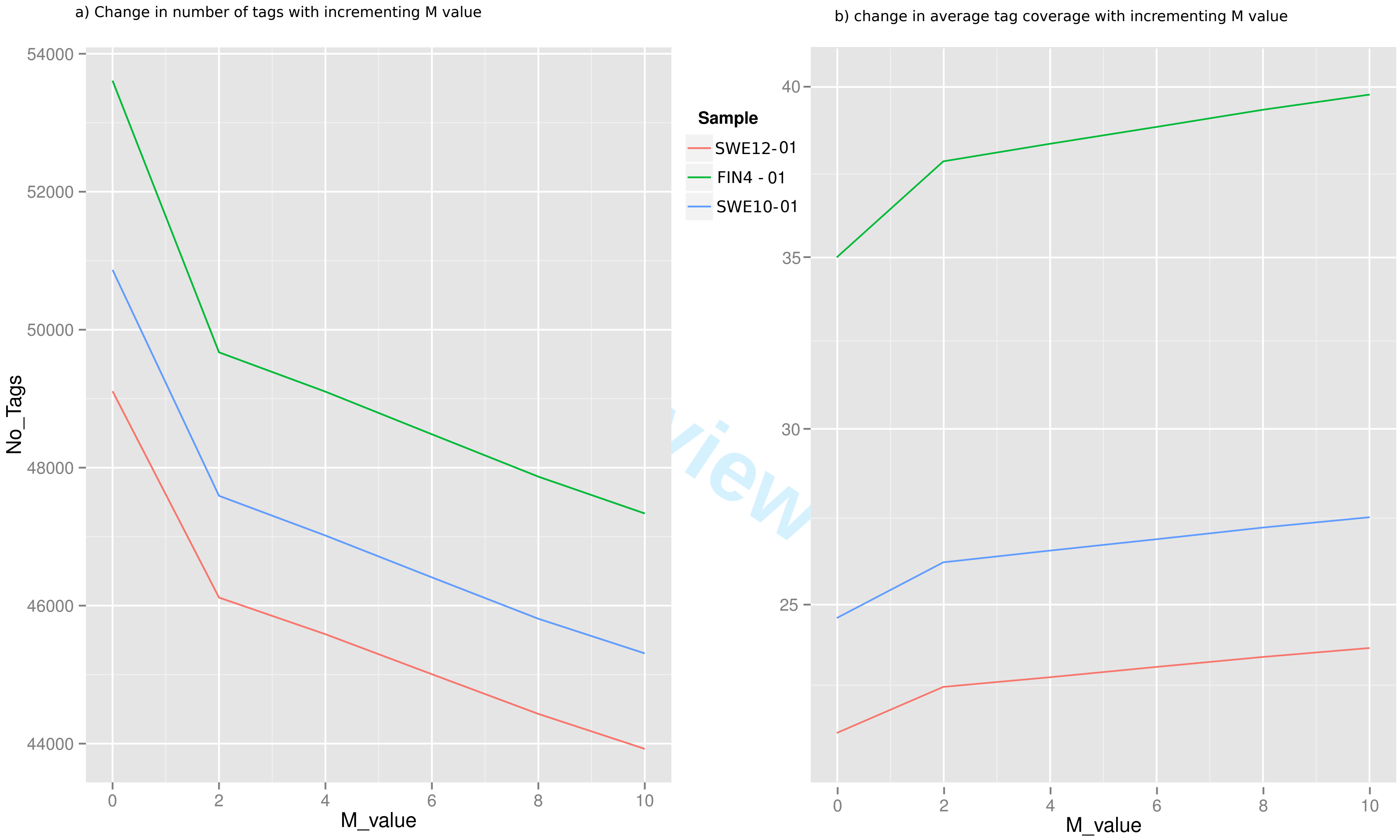


For Review Only

Spatial patterns in M2 vs M1 datasets

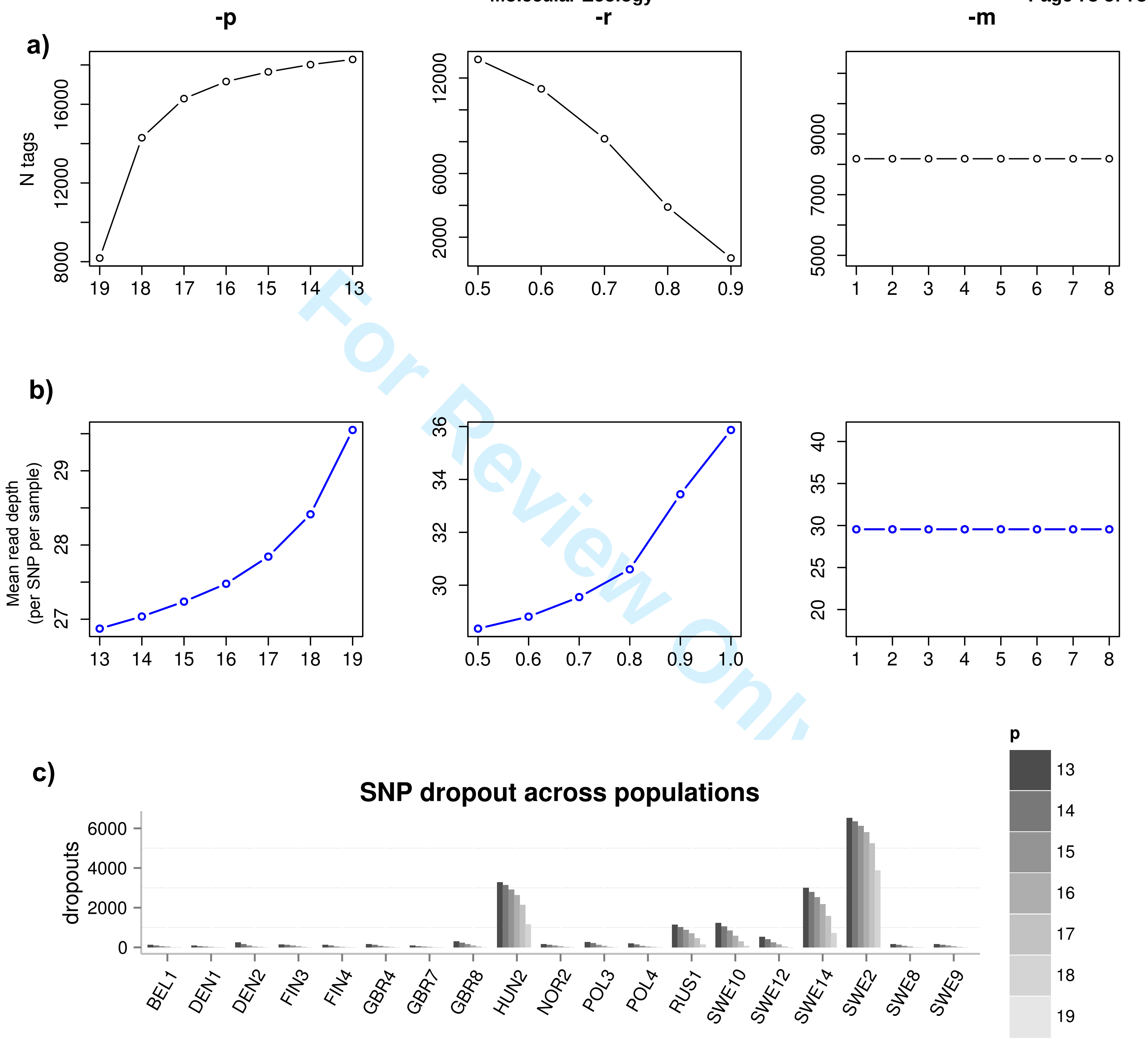
For Review Only





a) Change in number of tags with incrementing M value

b) change in average tag coverage with incrementing M value



For Review Only