THE UNIVERSITY OF HULL

Genetic and genomic approaches to the conservation of the threatened crucian carp *Carassius carassius* (L.); phylogeography, hybridisation and introgression

being a Thesis submitted for the Degree of Doctor of Philosophy
School of Biological, Biomedical and Environmental Sciences
University of Hull

by

Daniel Lee Jeffries, BSc, MSc

2015

Acknowledgments

Firstly, I must thank the Fisheries Society of the British Isles (FSBI) for the scholarship funding for this project and the Centre for Environment, Fisheries and Aquaculture Science for consumables funding. Secondly, a thank you to all of the fishermen, farmers, enthusiasts, academics and fellow students who helped me collect the biological material for this PhD, they are:

> 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).

The success of the project was heavily dependent on obtaining samples from all over Europe, which I could not have accomplished without their help. In fact, the majority of samples in the study were collected not only by existing sampling contacts, but by their friends, and friends of friends, and I consider it a great testament to the field that so many people were willing to give their time to help an anonymous PhD student the other side of the continent! Of course, I then had to do something with these carefully collected samples, and for that I relied heavily on the superb guidance of my three supervisors. Gordon, I thank you not only for your scientific expertise, but also for imparting you wise words upon me at many a conference dinner. And thank you Bernd, if you hadn't listened to me ramble on about my undergraduate project all those years ago and encouraged me to do my Masters degree, I wouldn't be here now. So you only have yourself to blame! But seriously, thanks for teaching me the rigour that every scientist should employ, and how it can (and should) go hand in hand with a passion for the work. Lori, thank you firstly for taking me on as a Masters student, I would never have been prepared for this PhD if it wasn't for the scientific grounding that you gave me. Throughout my Masters and PhD you have been a constant source of encouragement, and have given me the confidence to pursue an academic career – I'll be forever grateful. I must also mention all of the other members of the UoH Evolutionary Biology Group, Dave, Domino, Africa, many other past and present, and especially my Bioinformatics Buddies Christoph and Amir. I hope I can find a group as

cohesive and supportive wherever I go! Thank you also to Dave and Kanchon for agreeing to examine my PhD.

On a more personal note, after 9 years in Hull I have many friends that have been instrumental in keeping me sane and it would be impossible to name them all and quantify their value. But two deserve special mention, firstly Adam; possibly the most ridiculous person to have lived. You have been my brother in Hull, and, although sometimes I feel like our friendship and the related shenanigans were not exactly conducive to completing a PhD, indirectly they have been invaluable distractions! Finally, Helen, you are awesome. You have kept me just the right amount of insane during these last stages of the PhD, and I can't wait for our next adventure! Here's to not needing to call beers, "our" special brand of humour, replenishing our whisky collection and, of course, mischief.

So to all of the above, thank you for making this PhD a truly enjoyable endeavour!

Candidates declaration

I declare that the work submitted in this thesis is my own unless otherwise stated.

The work in this thesis represents collaboration with my three main supervisors Dr. Bernd Hänfling (BH), Prof. Gordon Copp (GC), and Dr Lori Lawson-Handly (LLH), and also Dr Veronique Creach (VC), Dr Carl Sayer (CS), and Prof. Håkan Olsén (HO).

BH, GC and LLH were involved in all parts of the project. VC was involved in a supervisory capacity during the early stages of project design and data collection.

CS and HO were instrumental in obtaining samples from Norfolk and northern Europe respectively.

Chapter 3 of this thesis has been submitted to the journal Molecular Ecology. This chapter has therefore been included in this thesis in its submitted form. Incidentally there is some overlap between the Supplementary materials of Chapter 3 and the work carried out in Chapter 2.

Chapter 4 has been submitted to the journal of Conservation Biology however has been slightly reformatted here to be consistent with the rest of the thesis.

I further declare that no part of this thesis has been submitted as part of any other degree.

How to use this thesis

For ease of navigation, all supplementary materials are included at the end of their respective chapter. Citations to figures in the text are also hyperlinks, and for supplementary materials, a “Back to text” hyperlink is included at the end of each caption which will navigate back to the first in-text citation of that table or figure.

All in-text literature citations are hyperlinks to the reference in my paperpile library, the vast majority of which will have associated pdf’s.

Finally, all scripts, used in this thesis are in my thesis Github repository: https://github.com/DanJeffries/DLJeffries_Thesis_data_and_scripts, all raw RADseq data is in the Short Read Archive under the accession number SRP063043, and all remaining raw data and intermediate analyses files are in the dryad repository here: https://figshare.com/articles/Thesis_data_and_scripts/2064345

Table of Contents

List of Tables

List of Figures

General Abstract

Biological invasions can have dramatic detrimental impacts on ecosystems, however they also represent rich opportunities to study the evolutionary processes associated with them. Hybridisation and subsequent introgression are two such processes and are common among native and non-native species. The crucian carp, *Carassius carassius* (L.), is a European freshwater fish that is threatened throughout much of its native range by several factors including hybridisation and introgression with three non-native species, the goldfish, *Carassius auratus* (L.), the gibel carp, *Carassius gibelio* (Bloch), and the common carp, *Cyprinus carpio* (L.). The conservation of *C. carassius* is hampered by a lack of phylogeographic knowledge for the species and no knowledge of the extent or impact of hybridisation and introgression. Contemporary genomic approaches such as Restriction Site Associated DNA sequencing (RADseq) can offer unprecedented insights into such research areas, however RADseq comes with several sources of potential bias. Exploratory analyses in Chapter 2 show that two sources of bias in particular, null alleles and over merged ohnolog loci, are highly important in this dataset, but can be filtered using population genetics statistics. The filtered dataset is used in phylogeographic analyses in Chapter 3, along with microsatellite and mitochondrial DNA and show that *C. carassius* exists as two major lineages in Europe, which diverged approximately 2.26 million years ago, and should be treated as separate units for conservation. These lineages result from the *C. carassius* postglacial recolonisation routes thtough Europe, which are highly distinct from the general patterns seen in other freshwater fish species. These phylogeographic results showed high similarity between *C. carassius* in England and those in continental Europe, calling into question the presently assumed native status of *C. carassius* in England, which has been contentious in the past. Empirical tests of this status using microsatellites showed that, in fact, *C. carassius* is most likely introduced in England around the 15th century, raising interesting discussions pertaining to their conservation in the England. Lastly, in Chapter 5, microsatellite and RADseq approaches show that hybridisation between *C. carassius* and non-native species is prevalent where they are sympatric, however backcrosses are rare, and there is no evidence of further introgression between the species studied. Taken together, these results suggest that postzygotic mechanisms of isolation limit interspecific gene flow, and conservationists should focus further research on the direct impacts of non-native species and F1 hybrids.

Chapter 1 General introduction

The introduction of non-native species into novel regions is a process of great importance in conservation biology (Gurevitch & Padilla 2004; Didham *et al.* 2005; Vitule *et al.* 2009). Although species colonisations are a natural process and are considered an important driver of evolution (Reznick & Ghalambor 2001; Petit 2004), the human mediated transport of organisms beyond their native ranges, be it accidental or intentional, has increased their occurrence far beyond natural rates (Cohen & Carlton 1998). Over time, the huge numbers of species introductions have resulted in countless invasive species around the world (Mooney & Cleland 2001), i.e. species which become self-sustaining and detrimental to native ecosystems and economies (Colautti & Mac Isaac 2004). Such species can cause decline and even extinction of native species (Worthington & Lowe-McConnell 1994; Ricciardi *et al.* 1998; Clavero & García-Berthou 2005), thus knowledge of the mechanisms behind these impacts are of utmost importance for the conservation of native species. However, species invasions also provide opportunities to study the many evolutionary processes that are associated with them (Fitzpatrick *et al.* 2010).

With recent technological advances in High Throughput Sequencing (HTS) technology, genomic techniques now promise unprecedented insights into the evolutionary impact of species invasions (Twyford & Ennos 2011). Furthermore, the availability of genome wide molecular markers is now driving the transition from conservation genetics to conservation genomics, allowing researchers to make management decisions using not only neutral but also selectively important genetic information (Ouborg *et al.* 2010; Avise 2010).

The present thesis spans both conservation and evolutionary biology, in a system containing a threatened native species and several introduced species. The focal native species of this project is the crucian carp, *Carassius carassius*, a freshwater fish native to much of Europe and threatened by multiple factors including three introduced species: goldfish, *Carassius auratus;* gibel carp, *Carassius gibelio;* and common carp, *Cyprinus carpio* (Hänfling *et al.* 2005). These species are introduced or invasive in much of the *C. carassius* range and have been implicated in its decline in several regions (Hänfling *et al.* 2005; Papoušek *et al.* 2008; Copp & Sayer 2010; Sayer *et al.* 2011; Mezhzherin *et al.* 2012; Wouters *et al.* 2012). In light of these declines *C.*

carassius is a species of great conservation concern (Copp & Sayer 2010), however, very little is currently known about the phylogeographic patterns within *C. carassius* and the impacts imposed by the three non-native species. Such knowledge is imperative for the identification of imperilled populations, and the prioritisation of conservation units (Frankham *et al.* 2002). This thesis aims to fill these gaps in the current knowledge for *C. carassius* with the ultimate goal of facilitating the conservation of this species whilst shedding light on the evolutionary processes that occur during the invasion process.

This chapter first summarises the current literature in the field of invasion biology and the conservation and evolutionary implications of invasions, with specific emphasis on non-native fish species. Secondly, the system, comprised of the four species mentioned above, is described in detail and the areas where information is lacking are highlighted. Thirdly, the recent advances in the field of conservation and invasion biology owing to next generation sequencing technology are described with particular reference to the methods used in this thesis. Finally, the specific questions of this thesis are introduced.

Biological invasions

The vast majority of non-native species introductions around the globe are the result of human mediated translocations (Hulme *et al.* 2008), usually for the purposes of food, sport (e.g. fishing), pets or for biocontrol (Hulme *et al.* 2008). In many cases intentionally introduced species escape confinement and subsequently become invasive, which is particularly common in plants and also in fish which have been introduced for aquaculture or the ornamental fish trade (Hulme *et al.* 2008). However, unintentional introductions of species are also commonplace, for example it is well known that many important invasive aquatic species (both vertebrates and invertebrates) have been introduced as stowaways on the hulls of ships or in ballast water (Lavoie *et al.* 1999; Hulme *et al.* 2008). Regardless of human intention however, there is a very strong association between the amount of international trade (i.e. imports) and the number of non-native species present in a given country (Hulme 2009).

Once a species has been introduced into a novel environment, there are multiple factors, relating to either the introduced species or its new environment, that determine whether or not that species becomes established, spreads and has detrimental impacts on resident

species. For example, Marchetti *et al.* (2004) compared the characteristics of successfully established introduced fish species in North America with those that failed to establish and found that two traits: broad physiological tolerance and previous invasion success, were strong predictors of establishment success. These two traits, along with growth rate, were also found to be important for establishment success by Kolar and Lodge (2002), in introduced fish species in the Great Lakes of North America. For the subsequent spread of an established non-native fish, lifespan, distance from nearest native source populations and trophic status have been found to be important (Marchetti *et al.* 2004). Kolar and Lodge (2002) discriminated between fast and slow spreading non-native species and found that those that spread quickly grew more slowly, and could tolerate a wide temperature range, although had poor survival in high temperature extremes. However in the many studies which have aimed to identify commonalities between species invasions, one factor in particular, propagule pressure (which incorporates the number of propagules per introduction and the number of introduction events), stands out as being a major predictor of invasion success (Lockwood *et al.* 2005; Simberloff 2009; Davis 2011). After establishment and spread, some species may then have detrimental impacts on residents, which, in the Great Lakes fishes, was found to be predicted by their ability to survive in low water temperatures, in a wider range of salinities, and was also associated with smaller egg size (Kolar & Lodge 2002). The above characteristics allow for some prediction of the invasive potential of a species (Kolar & Lodge 2002), but ecological factors in the invaded habitat such as resource availability and the abundance of enemies also play an important role. (Davis 2011). Thus the invasive potential of a given species will rely on an interaction between the traits of that species, and the conditions of its introduced range.

Each step of the invasion process described above can be thought of as a filter through which a species must pass in order to become invasive. For example, if a species survives transportation but fails to become established then little or no threat is likely to be posed to the native ecosystem. This pattern is often said to conform, generally, to the tens rule, whereby only 10% of species make it through each of three main filters, introduction, establishment and having negative impacts on residents (Williamson & Fitter 1996). Hence, for every 1000 species introduced, only one is likely to become invasive. Ultimately, invasion biologists aim to use the information from studies such as those described above to predict which species are likely to make it through the filters

of the invasion process to become invasive, and in which environments this is likely to happen. This endeavour is referred to as horizon scanning, and was recently applied by Roy *et al.* (2014) to 591 species which have a possibility of introduction to the UK. The study identified 93 species which have at least a medium risk of becoming invasive in the future and, of these, 30 species were deemed to pose a high risk, which included two brackish water fish: the round goby, *Neogobius melanostomus* (Pallas) and the tubenose goby, *Proterorhinus marmoratus* (Pallas) which are thought to be unintentionally transported as stowaways in the ballast water of ships.

There are many well-known examples of species declines and extinctions that are attributed to invasive species. For example, in the extreme case of the Island of Guam in Micronesia the introduction of several non-native species including the brown tree snake, *Boiga irregularis* (Merrem) is thought to have caused a catastrophic collapse of the native ecosystem, resulting in many species extinctions (Fritts & Rodda 1998). Indeed, invasive species are thought to be among the leading causes of decline and extinctions in birds (Clavero *et al.* 2009), mammals (Clavero & García-Berthou 2005), and amphibians (Gibbons *et al.* 2000). As a general pattern, the most dramatic impacts caused by introduced species appear to occur in Island ecosystems, which is thought to be due to their isolation leading to a lack of evolved defence against colonising species (D'Antonio & Dudley 1995; Fritts & Rodda 1998). Freshwater ecosystems, such as ponds, lakes and river systems can be equally as isolated, and so can, themselves, be viewed as island ecosystems. There are countless examples of invasive freshwater species, which are thought to represent the most significant threat to fish biodiversity after habitat loss (Miller 1989). The rate of non-native fish introductions has more than doubled in the last 30 years, largely as a result of increased international trade and human movements (Gozlan 2008). These introductions have been linked to detrimental impacts on native ecosystems via predation, competition, hybridisation and the introduction of novel pathogens (see Gozlan *et al.* 2010 and references therein). However, in many cases invasive species are only implicated by anecdotal evidence or correlative data, and in only very few cases do mechanistic data exist showing direct causation between non-native species introductions and native species declines (Davis 2003; Gurevitch & Padilla 2004; Didham *et al.* 2005). Obtaining such mechanistic data represents one of the major challenges of contemporary invasion biology, due, in large part, to the fact that human mediated species invasions often go hand in hand with human mediated environmental change (Gurevitch & Padilla 2004; Didham *et al.*

2005). These processes, though often treated as independent, likely interact, and it is has been extensively discussed that invasive species take advantage of opportunities created by ecosystem and habitat change, as opposed to drive it (Vitousek *et al.* 1996; Mack *et al.* 2000; Byers 2002; Gurevitch & Padilla 2004; Didham *et al.* 2005). As such, there is debate as to the true extent of the threat posed by invasive species independent of other drivers of global ecosystem change (Gurevitch & Padilla 2004; Clavero & García-Berthou 2005; Gozlan 2008; Davis *et al.* 2011).

Evolutionary impacts of biological invasions

Biological invasions are implicated in many evolutionary impacts on both the native species, in response to an invasive species, or on the invasive species itself in response to its new environment or via genetic drift (Mooney & Cleland 2001; Lindström *et al.* 2013; Peischl *et al.* 2015). For example, there is evidence that the native North American soapberry bug *Jadera haemotoloma* (Herrich-Schäffer) has adapted its feeding apparatus in order to feed from the introduced goldenrain trees (*Koelreuteria* spp.) (Carroll & Dingle 1996; Yu & Andrés 2014). And in the invasive cane toad, *Rhinella marina* (L.) individuals at the leading invasion edge have higher dispersal ability than those in the core of the population (Lindström *et al.* 2013).

One evolutionary process which often accompanies a species introduction is hybridisation between the native and non-native species (Mooney & Cleland 2001). It is thought that the prevalence of hybridisation during invasions is due to the lack of reinforcement of reproductive barriers between species that would otherwise exist in allopatry (Howard 1993; Arnold 1996). As a result, there are numerous examples of hybridisation between native and non-native species (see examples in Utter 2000; Hänfling 2007). Hybridisation is potentially an important driver of evolution, rapidly creating novel gene combinations (Martinsen *et al.* 2001; Twyford & Ennos 2011) which can, in some cases, lead to rapid adaptive diversification (Seehausen 2004). Indeed it is estimated that 50-70% of all angiosperm plants have hybrid origins (Martinsen *et al.* 2001) and 10% of animal species hybridise (Mallet 2005). Hybridisation has also been implicated in speciation via whole genome duplications, which cause reproductively isolating ploidy differences between pure and hybrid lineages. In many cases, hybridisation can have significant detrimental effects on parental species, for example through the production of vigorous hybrids (Howard

1993; Arnold 1996; Rhymer & Simberloff 1996) which can, in turn, result in highly invasive hybrid lineages. In the context of biological introductions, such vigorous hybrids exacerbate the pressures imposed on the native species. In extreme instances, highly invasive hybrid lineages can outcompete and extirpate parental species, as was the case in the freshwater snail *Melanoides tuberculata* (Facon *et al.* 2005) and in sheepshead minnows *Cyprinodon pecosensis* (Wilde & Echelle 1997). Hybrid lineages may arise via a number of mechanisms. Firstly, recombination may create gene combinations that confer a fitness advantage (Hänfling 2007). Secondly, interactions between the parental genotypes, namely through dominance, overdominance and psuedo-overdominance (reviewed in Birchler *et al.* 2006) may confer fitness advantages in hybrids. Thirdly, novel combinations of parental genotypes may result in transgressive segregation, whereby a hybrid individual possesses phenotypic traits outside of the range of either parental species (Rieseberg *et al.* 1999). Additionally in hybrid lineages which reproduce clonally, advantageous gene combinations are not lost through recombination in subsequent hybrid or backcross generations (Facon *et al.* 2005).

Where hybridisation occurs, there is the potential for introgression; the movement of genetic material from the genepool of one species into that of another. Introgression is mediated by hybridisation and the subsequent backcrossing of F1 individuals with one of the parental species (Barton 2001). This process therefore requires that F1 hybrids and backcrosses are fertile. However, this is not always the case (Arnold & Hodges 1995) as postzygotic isolation mechanisms, such as negative epistatic interactions (Dobzhansky 1937; Muller 1942; Turner *et al.* 2014)), often cause hybrids to be sterile or have low fertility (Orr & Presgraves 2000). These barriers to gene flow are not always absolute, and can be viewed as semi-permeable filters through which gene flow can occur (Mallet 2005; Twyford & Ennos 2011). Indeed F1 hybrids are generally the most problematic to produce and if fertile, subsequent backcrossing is normally much more straightforward (Mallet 2005). Even with only occasional backcrossing events, beneficial alleles are expected to introgress readily (Barton 2001), as has been found between the native red deer *Cervus elaphus* (L.) and the introduced sika deer *Cervus nippon* (Temminck) (Goodman *et al.* 1999). In the context of conservation biology, such introgression may have important implications. Introgression of beneficial locally adapted alleles from the native species might confer a fitness advantage and increasing invasive potential of the introduced species (Hänfling 2007). Evidence of such a

phenomenon has been found between rainbow trout, *Onchorhynchus mykiss* (Walbaum) and westslope cutthroat trout *Onchorhynchus clarkii* (Richardson) whereby several 'super invasive' alleles show much higher rates of introgression than other loci and are therefore thought to be under positive selection (Hohenlohe *et al.* 2013). Alternatively, the movement of genes into the native species can have detrimental effects, as they may disrupt epistatic interactions between co-adapted gene complexes in their new genetic background (Dobzhansky 1937; Muller 1942; Lynch 1997).

Hybridisation is particularly prevalent in fishes due to their external fertilization, and weak behavioural isolating mechanisms, for example shared spawning times and habitats, and high introduction rates owing to stocking practices (Scribner *et al.* 2000; Madeira *et al.* 2005). Despite the fact that this prevalent hybridisation often produces sterile F1 offspring (Hubbs 1955), introgression has been documented in dozens freshwater fish systems (reviewed in Verspoor & Hammart 1991; Smith 1992; Madeira *et al.* 2005). The conservation implications in these systems are considerable, as, in many cases, introgression occurring between anthropogenically introduced species threatens native species (Scribner *et al.* 2000). For example, stocking of hatchery-bred brown trout, *Salmo trutta* (L.) in the Iberian Peninsula has resulted in large amounts of introgression with resident lineages, which is homogenising population structure within this region (Almodóvar *et al.* 2006).

The study system

The crucian carp, *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: *C. auratus*, *C. gibelio*, and *C. carpio* (Wheeler 2000; Hickley & Chare 2004). Hybridisation between *C. carassius* and these introduced species is common (Hänfling *et al.* 2005, discussed in detail below), which further complicates their morphological distinction (Wheeler 2000).

One particularly contentious part of the *C. carassius* range is England, UK. The current consensus is that *C. carassius* is native to the eastern counties of England (Kent, Suffolk, Norfolk and London) and has been anthropogenically introduced to other parts of the UK (Wheeler 2000; Copp *et al.* 2010). This conclusion is based primarily on the similarity of the *C. carassius* distribution to other native UK species such as burbot, *Lota lota* (L.), bream, *Blicca bjoerkna* (L.), spined loach, *Corbitis taenia* (L.) and ruffe, *Gymnocephalus cernuus* (L.) (Wheeler 2000). Further support for its native status in England comes from the identification of pharyngeal bones at a Roman archeological dig site in London (Jones 1978). However this record is ambiguous, as 'crucian carp' is mentioned as having been identified in the text of the archeological report, but is not present in the associated data table (Jones 1978). The opposing argument has been put forward by Maitland (1972), who suggested that *C. carassius* was, in fact, introduced along with common carp in the 15th century. Therefore, as it stands, the body of evidence for the status of *C. carassius* in England does not allow for a concrete conclusion one way or another. Such information is imperative for the conservation of the species, as knowledge of a species native range is required in order to prioritise conservation plans (Frankham *et al.* 2002; Reed & Frankham 2003; Copp & Sayer 2010).

C. carassius is a benthic feeding species, often found in small isolated ponds and lakes and sometimes in the slow moving backwaters of low lying river systems (Holopainen *et al.* 1997; Wheeler 2000). Such water bodies are prone to hypoxia caused by eutrophication or long term ice cover and therefore present intolerable environments for many fish species. However *C. carassius* possesses a number of adaptations which allow it to thrive in such environments. *C. carassius* is able to survive temperatures of between 0 – 38^{O}C (with an optimum of 27^{O}C), pH as low as 4 and has been shown to tolerate anoxia at low temperatures for up to 5-6 months (Holopainen *et al.* 1997). The behavioural adaptations that allow for this remarkable hardiness include inactivity and fasting during the low temperature winter months, slowing the metabolism and avoiding the use of energy (Holopainen & Hyvärinen 1985). *C. carassius* also possess physiological adaptations to anoxic conditions in the form of an alternative metabolic pathway called the 'ethanol cycle' in which glycogen stored in the late summer months is metabolised anaerobically to produce lactate, which is then converted to ethanol and excreted through the gills (Holopainen *et al.* 1997; Vornanen *et al.* 2009).

Despite its hardiness, however, *C. carassius* is threatened throughout much of its native range by a number of factors. For example increased mortality was observed with the acidification of lakes and ponds in Fennoscandia (Holopainen & Oikari 1992), drought and terrestrialisation have contributed to numerous population-level extinctions in Norfolk, UK , with a loss of 72% of populations known to contain *C carassius* in the 1970's and 1980's (Sayer *et al.* 2011) and bad water quality is thought to have contributed to *C. carassius* decline in the Danube river basin (Navodaru *et al.* 2002). The most commonly cited driver of *C. carassius* decline is the presence of the three introduced species previously mentioned (Hänfling *et al.* 2005; Papoušek *et al.* 2008; Copp & Sayer 2010; Sayer *et al.* 2011; Mezhzherin *et al.* 2012; Wouters *et al.* 2012). Morphological similarity and hybridisation between lineages within the *Carassius* genus has led to debate over the number of species or subspecies that it contains. Two of these species in particular, *C. auratus* and *C. gibelio* have a notoriously contentious taxonomy (Rylková *et al.* 2013). *C. gibelio* has previously been suggested to be sub-species or a morphotype of *C. auratus* (reviewed in Mezhzherin & Lisetskii 2004). Additionally, the presence of triploidy in some *C. gibelio* populations further complicated matters, raising the theory that *C. gibelio* in fact resulted from a hybridisation between *C. auratus* and another (unknown) species (Hänfling et al. 2005). However, it has recently been shown that triploid lineages exist in *C. auratus*, *C. gibelio* and *Carassius langsdofii* (Takada *et al.* 2010; Xiao *et al.* 2011) and so this trait cannot be used to separate species.

Recently, Rylková *et al.* (2013) provided the most comprehensive phylogeny of the *Carassius* genus to date, using 404 cytochrome b mitochondrial sequences. In doing so, they found that *C. auratus* and *C. gibelio* represent monophyletic lineages and therefore refer to them as separate species. On the basis Rylková *et al.* (2013), there are now five accepted species in the *Carassius* genus, *C. carassius*, *C. langsdorfii, Carassius cuvieri, C. auratus* and *C. gibelio*. However, as the latter four are morphologically and genetically more similar to one another they are said to belong to the *Carassius auratus* species complex. In this thesis I adhere to this naming system, whereby *C. auratus* and *C. gibelio* are species belonging to the *C. auratus* species complex.

It should also be noted that it is common in Chinese and Japanese literature (when written in English) to refer to *C. auratus* and *C. gibelio* species as the "golden crucian

carp" or the "sliver crucian carp" respectively, which has caused further confusion in the past (e.g. Iguchi *et al.* 2003).

Throughout much of its range, *C. carassius* is found in sympatry with one or a combination of the three non-native species described above (Hänfling *et al.* 2005; Papoušek *et al.* 2008; Mezhzherin *et al.* 2012; Wouters *et al.* 2012; pers. comms. Müller Tamás, András Weiperth, Prof. Sergey Mezhzherin). The pathways of introduction of these species are often human mediated; *C. auratus* arrived in Europe in the 1600's, likely for the purposes of aquaculture, and they are now often imported by the ornamental fish trade and by anglers, who sometimes mistake them for *C. carassius*. *C. auratus* were recently identified as the most widespread introduced freshwater aquaculture fish in Europe (Savini *et al.* 2010). *C. carpio* and *C. gibelio* were introduced to Europe in the 1200's and the 1600's respectively, most likely due to their importance as a source of food and for angling, and, along with *C. auratus*, are now among the top 25 most imported fish for aquaculture in Europe (Savini *et al.* 2010).

Evidence for the impact these non-native species on *C. carassius* is lacking. Although Navodaru *et al.* (2002) found a decrease in *C. carassius* population size with the introduction of *C. carpio*, this evidence is correlative and not causative, and it is possible that *C. carassius* declines were due to human mediated impacts. Copp *et al.* (2010) performed the only known empirical assessment of the ecological impacts of *C. auratus* on *C. carassius*, however they found no significant detrimental impacts in the ponds studied. Where *C.carassius* is sympatric with a non-native species, hybridisation is common, and this has been implicated by many studies as a source of threat to *C. carassius* (Hänfling *et al.* 2005; Papoušek *et al.* 2008; Sayer *et al.* 2011; Mezhzherin *et al.* 2012; Wouters *et al.* 2012). However, to date, the most convincing evidence of negative impacts of hybrids comes from Hanfling *et al.* (2005), who found that 38% of sampled *C. carassius* populations contained hybrids with *C. auratus* spp. or *C. carpio* and, in many populations, hybrids were present but no *C. carassius* individuals were found, indicating their competitive exclusion by vigorous hybrids in these populations. Hanfling *et al.* (2005) also found evidence of backcrossing between *C. carassius* x *C. auratus* F1 hybrids and pure *C. carassius*, suggesting that there is a window for introgression to occur between these species. However, the above studies used only a small number of molecular markers, thus, their power to detect introgression past the first backcross stage was limited. It is therefore not known if subsequent hybrid

generations exist in this system, and to what extent introgression occurs between *C. carassius* and the introduced species.

The conservation of crucian carp in Europe

To date the conservation initiatives of *C. carassius* are few and far between. Despite their being identified as under threat in many countries throughout Europe (Lusk *et al.* 2004; Mrakovčić *et al.* 2007; Wolfram & Mikschi 2007; Simic, V *et al.* 2009; Copp & Sayer 2010) and recognised by the International Union for Conservation of Nature (IUCN) as in decline, *C. carassius* is listed as a species of least concern (Freyhof & Kottelat 2008). As such, only very few local-scale conservation programs exist, for example in Norfolk, UK (Copp & Sayer 2010). Broad scale conservation initiatives are therefore needed. Several lines of information are required to inform such initiatives including, 1) knowledge of the impact of non-native species on *C. carassius*, including the extent of introgression between them, and 2) phylogeographic data describing the amount and distribution of genetic diversity within the species, which is essential for the identification of imperiled populations and distinct conservation units (Ouborg 2009; Ouborg *et al.* 2010). As mentioned, *C. carassius* inhabits still or slow moving water bodies, which are often small and isolated (Holopainen *et al.* 1997). Such habitats predispose species to three important genetic risks; firstly, the gradual loss of alleles through the exaggerated effect of genetic drift in small populations, resulting in the permanent loss of genetic diversity; secondly the gradual accumulation of deleterious alleles due to the reduced efficiency of selection in small populations and thirdly, inbreeding depression resulting from mating with closely related individuals (Keller & Waller 2002). To date, no studies exist which address these potential risks in *C. carassius*.

Conservation and Invasion biology in the genomic age

Several decades ago conservation genetics was a largely theoretical field, however, with the development of genetic markers such as microsatellites, allozymes and amplified fragment length polymorphisms (AFLPs), researchers have been able to test many fundamental theories pertaining to demographic processes like genetic drift, inbreeding and admixture (Ouborg *et al.* 2010). These methods have proven invaluable for the conservation of species, identifying conservation units (Palsbøll *et al.* 2007),

populations of low genetic diversity and shedding light on the general phylogeographic patterns in species distributions (Hewitt 1999). However, despite their invaluable contribution to the fields of conservation and evolutionary genetics, these approaches have their limitations, which can be broadly summarised into two points; firstly, the marker types described above are most often only neutral (except for rare cases where they are linked to selected loci), and thus, do not easily allow for the study of selectively important regions of the genome. And secondly, cost and time limitations allow for only small numbers of these markers (usually 10-20) to be genotyped per study, thus precluding the examination of processes acting differentially throughout the genome (Ouborg *et al.* 2010).

With the advent of HTS, conservation genetics is now transitioning into conservation genomics, which uses data from genome wide markers such as single nucleotide polymorphisms (SNPs) to assess the effects of demographic and selective forces throughout the genome. HTS also allows for the large scale sequencing of transcriptome sequences, which enables researchers to examine gene expression patterns (Ouborg *et al.* 2010). However, perhaps the most important attribute of HTS is its ability to generate genome wide data, at thousands, if not hundreds of thousands of markers, for hundreds of samples, in non-model organisms for which no genomic resources exist (Ouborg *et al.* 2010; Avise 2010). One approach, which is specifically designed for this purpose and is used throughout this study, is Restriction Site Associated DNA sequencing (RADseq (Baird *et al.* 2008; Davey *et al.* 2010)). RADseq is a reduced representation approach to genome sequencing, in which a sample of a genome (typically 2-4%) is sequenced, making its application to large sample sets considerably more practical for conservation and population genetics than whole genome approaches. Importantly, the way RADseq samples a genome is to cut it at specific sites using a restriction enzyme (or pair of enzymes in double digest RADseq (Peterson *et al.* 2012)), resulting in thousands of homologous sequenced genome segments across samples (McCormack *et al.* 2013) (Figure 1.1).

RADseq overcomes many of the challenges associated with applying HTS to conservation biology, phylogenetics and phylogeography (McCormack *et al.* 2013) and as such it is increasingly used in these fields. An important attribute of RADseq, owing to the huge number of SNP markers it identifies, is its ability to resolve phylogeographic structure to a much finer scale than previously achieved with traditional markers such as microsatellites and mitochondrial genes (Emerson *et al.* (2010). The high marker density produced by RADseq also more accurately captures the phylogeographic signal throughout the genome and the sheer quantity of data reduces the confounding effects that loci influenced by selection have on phylogeographic results (Emerson *et al.* 2010). Furthermore, these selected loci can be isolated and used to discriminate between demographic and adaptive forces in the evolution of populations and species (McCormack *et al.* 2012; Catchen *et al.* 2013a).

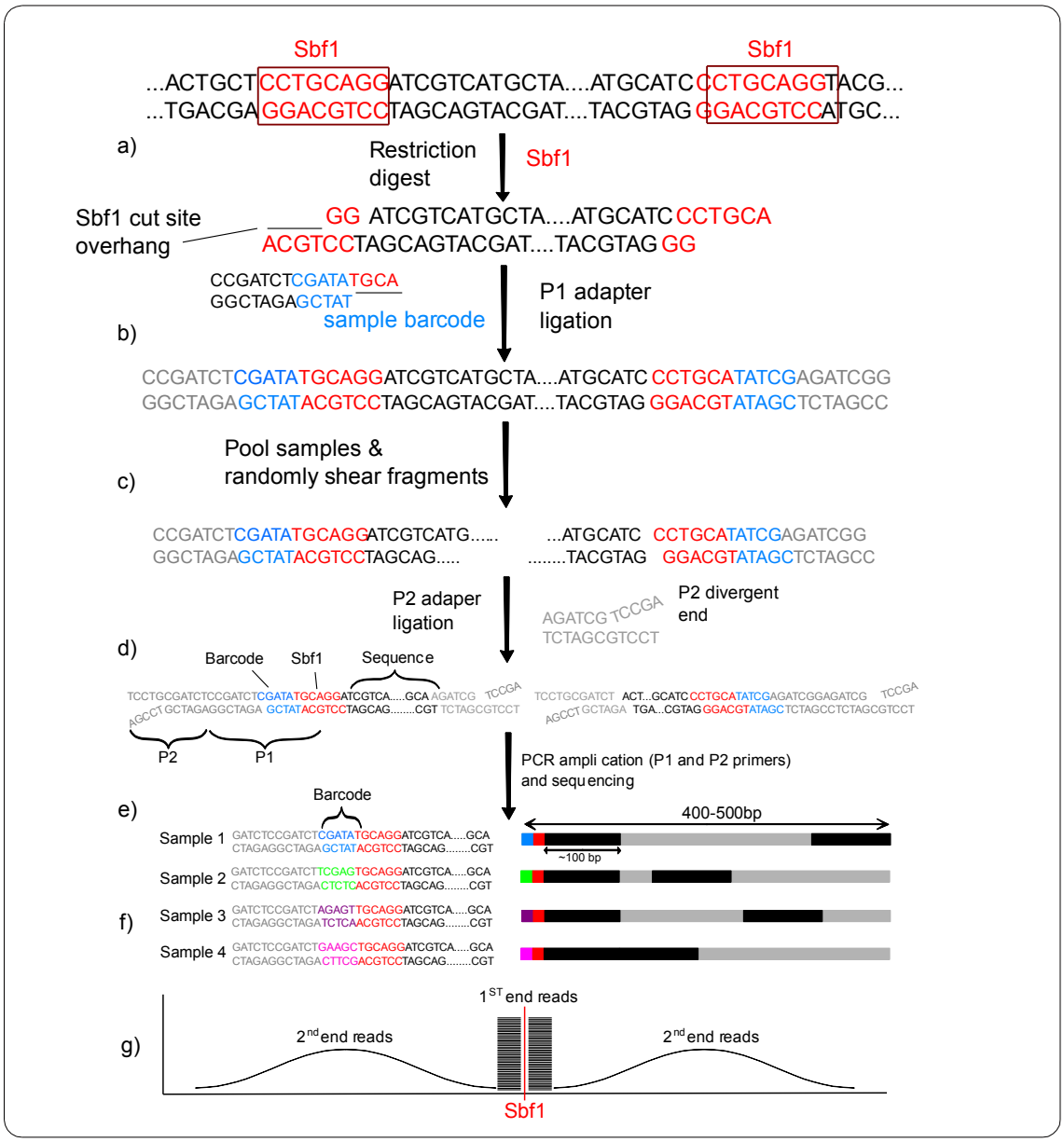


Figure 1.1. Adapted from Davey *et al.* (2010). The steps of paired end RADseq. First, genomic DNA is cut into fragments using a restriction enzyme, in this case Sbf1 (a). The sticky-ends of these fragments are then ligated to the P1 adapter (b), which contains an illumina binding sequence, allowing the fragment to bind to an illumina flow cell, a barcode (blue), allowing for identification of sequences belonging to individual samples and the P1 PCR primer site. Sequences from all samples are then pooled and sheared using sonication, which results in fragments of random length, all with a P1 ligated adapter (c). The P2 adapter is then ligated to both ends of each fragment (d) which contains a second PCR primer site, and has a divergent section of sequence which will not bind to the P2 primer used in PCR unless it has first been completed by amplification using the P1 primer. In this way, only fragments that contain both P1 and P2 adapters will be amplified. The resulting fragment library is then size selected using electrophoresis, to ensure that all fragments are between 200-500 bp in length, and the library is then sequenced as usual on an illumina platform, in the present study, the Illumina Hi-Seq 2000. The resulting data contains paired-end sequences, each approximately 100 bp each, which span a stretch of sequence of between 200-500 bp, with the first end read containing the sample barcode and the remainder of the Sbf1 cut-site (f). The process occurs for sequence either side of a restriction site, and results in the within sample read-depth distribution shown in (g).

RADseq has also driven advances in the field of biological invasions, and in particular, in systems where hybridisation and introgression occur (Hand *et al.* 2015). Although traditional markers such as microsatellites allow for the confident identification of hybridisation (e.g. Hänfling *et al.* 2005), the number of markers required to identify backcrosses and subsequent hybrid generations quickly outstrips the number that can practically be genotyped using these traditional methods (Boecklen & Howard 1997). Therefore the ability of RADseq to quickly genotype thousands of loci across many samples from non-model species makes it perfectly suited to the task of identifying introgression. To date RADseq has been used to identify introgression in several invasive species study systems, for example Lamer *et al.* (2014) identified introgression between the bighead carp, *Hypophthalmichthys nobilis* (Richardson) and the silver carp, *Hypophthalmichthys molitrix* (Valenciennes) in the Mississippi river basin, and much work has been done to characterise the patterns of neutral and adaptive introgression between native westslope cutthroat trout, *O. c. lewisi* and the introduced rainbow *O. mykiss* (Hohenlohe *et al.* 2013; Hand *et al.* 2015).

Although RADseq represents an invaluable tool for conservation and invasion genetics, the data it provides must be treated with caution in order to avoid the incorporation of bias during bioinformatic analysis. For example it has been that, for datasets containing high levels of divergence, allele dropout can be considerable (Gautier *et al.* 2012; McCormack *et al.* 2013; Pante *et al.* 2015). Another bioinformatic challenge arises when analysing polyploid species (Ogden *et al.* 2013), or those diploid species which have undergone whole genome duplications (Hohenlohe *et al.* 2013). It is therefore advocated that extensive tests of bioinformatic parameter values be undertaken for each RADseq dataset (Catchen *et al.* 2013b).

This thesis has three major goals pertaining to the conservation of *C. carassius* and the evolutionary consequences of the non-native species studied here. Firstly, in order to lay down a solid foundation for conservation of *C. carassius* at the European scale, a comprehensive phylogeography is produced, using genetic and genomic approaches to elucidate the patterns that have led to their current distribution. Secondly, again to inform the conservation for *C. carassius* and to address debate in the past literature, Chapter 4 tests the presumed native status of *C. carassius* in the UK using molecular techniques and an approximate bayesian computation approach to phylogeographic hypothesis testing. And thirdly, Chapter 5 assesses the prevalence of hybridisation

between *C. carassius* and the three non-native species examined here, and tests for evidence of introgression beyond the initial stages of hybridisation. The questions of Chapters 3 and 5 are addressed using a RADseq approach, however before these data could be used in the final analyses, it was necessary to perform parameter tests for the bioinformatic analyses carried out in these chapters. Therefore in Chapter 2, the bioinformatics methods employed throughout this thesis are described and tested through a range of parameter values in order to account for any systematic biases in the data.

Chapter 2 . Methods in RADseq analyses

Abstract

In the present thesis, a large multispecies RADseq dataset is used to answer several evolutionary questions pertaining to the impact of invasive species. RADseq data, however, potentially contains several distinct biases, which can significantly influence the results of downstream analyses. Therefore care must be taken to remove all confounding factors, which is often possible at the bioinformatics stages of RADseq data preparation. In this chapter, a python module, Incremental, is developed and used to systematically test a range of core parameter values for the commonly used RADseq analysis pipeline, Stacks. Particular attention is given to two important sources of bias; allele dropout caused by mutations in restriction sites, and the merging of ohnolog loci resulting from whole genome duplications. The results of these tests suggest that a large proportion of RADseq loci in the present study contain null alleles, and these can cause large biases in the data if they are not filtered correctly in the Populations module of Stacks. Furthermore, ohnolog loci appear to be prevalent throughout the dataset, however, using population genetics filters, these loci can be removed from the final SNP datasets. Finally, on the basis of Stacks parameter tests, optimal values are chosen and used, together with population genetics filters, to produce the refined and unbiased SNP datasets used in Chapters 3 and 5 of this thesis.

Introduction

Molecular ecology has been revolutionised by the advent of high-throughput DNA sequencing (HTS, Ekblom and Galindo 2011), which is defined by its ability to sequence template DNA in a massively parallelised way in order to quickly and cost-effectively generate gigabases of genetic information (Hudson 2008). The ability to generate such data has opened up a multitude of opportunities, allowing researchers to address complex ecological and evolutionary questions in species for which no prior genomic resources exist (Stapley et al. 2010; Twyford and Ennos 2011; McCormack et al. 2013).

One HTS based approach which has been widely used by molecular ecologists and evolutionary biologists is Restriction Site Associated DNA sequencing (RADseq),

which is described in detail in Chapter 1 of this thesis. Briefly, through the digestion of whole genomes and the sequencing of the terminal ends of the resulting DNA fragments using HTS, RADseq can be used to identify thousands of genome-wide markers. Importantly, as these RADseq loci are associated with enzyme-specific restriction sites, they are homologous between individuals and, thus, highly applicable to population genetics studies (Davey et al. 2011). RADseq has been applied to a plethora of tasks, from fine-scale linkage mapping (Amores et al. 2011; Baxter et al. 2011; Chutimanitsakun et al. 2011), species identification (Hohenlohe et al. 2013) and genome scaffolding (Davey et al. 2011; The Heliconius Genome Consortium 2012), to phylogeography, phylogenetics (Emerson et al. 2010; Rubin et al. 2012), and population genomics (Hohenlohe et al. 2010).

RADseq data can contain several sources of bias, which can potentially confound biological conclusions (Davey et al. 2012). Many of these biases can be controlled for during RADseq library preparation, for example, the relationship between restriction fragment length and read depth and the GC bias in the library preparation PCR step. Whereas the presence of PCR duplicates (Davey et al. 2012) and allele dropout (Gautier et al. 2012) are biases which must be filtered out in the bioinformatics processing of data.

A number of bioinformatics pipelines are now available for the analyses of RADseq data, including pyRAD (Eaton 2014), RADmapper (github.com/tcezard/RADmapper), RADtools (Baxter et al. 2011), Rainbow (Chong et al. 2012), AftrRAD (Sovic et al. 2015), dDocent (Puritz et al. 2014) and Stacks (Catchen et al. 2011, 2013b). The latter of these is, perhaps, the most widely used, and comprises several modules and supplementary scripts, which are each responsible for a portion of RADseq data processing. These modules fall into two major pipelines within Stacks, the first builds loci from raw read data, based on their sequence identity to one another (hence forth referred to as the *de novo* pipeline), and the second builds loci from the results of the alignment of raw reads to a reference genome (henceforth referred to as the reference guided pipeline). In the *de novo* analyses, the first module, Ustacks, constructs loci by first grouping identical reads into, so-called, stacks, which are in turn grouped into loci based on their similarity to one another (‘-M) and the read depth per stack (-m), among several other parameters. Alternatively, in the reference guided pipeline, reads can be grouped into loci in the Pstacks module (in place of Ustacks), on the basis of their

mapping location to a reference genome. In both Ustacks and Pstacks, SNP calling is also performed. The Cstacks module is then used to create a catalog of loci present in all individuals. Again this is performed on the basis of either sequence similarity or genome mapping location for the *de novo* and reference guided analyses respectively. For both pipelines the Sstacks module then searches each individual sample against this catalog, in order to infer genotypes for each individual at each catalog locus. Finally these genotypes can then either be converted to mappable genotypes using the Genotypes module (for use in the construction of genetic maps), or used in populations genetics analyses performed in the Populations module (Catchen et al. 2013b).

The Stacks analysis steps described above, and especially those in the *de novo* pipeline are heavily parameterised, which allows for a large amount of flexibility in the Stacks analyses, and enables each user to customise the analyses to the properties of their dataset and each parameter. Catchen et al. (2013b) have examined the effects of three core parameters in Ustacks; -M, which specifies the number of mismatches between stacks of unique reads at a locus within an individual; -m, which sets a minimum threshold for the number of reads required at a locus and '--max_locus_stacks', which determines the number of stacks allowed to be merged into a single locus. The authors concluded that the optimal value for each parameter depends largely on the level of polymorphism the genome of study, the amount of sequencing error, and the depth of coverage achieved during sequencing. However, in non-model study systems, these properties are almost always unknown *a priori*, and the misspecification of a parameter can have drastic effects on the SNP datasets outputted from Stacks. For example, if the specified mismatch threshold (-M) is too low, there is an increased risk that the two alleles of a heterozygous locus will be called as separate, homozygous loci. In contrast, if this mismatch threshold is too high, then two loci that have high sequence similarity (e.g. paralogs) can be erroneously merged. In both cases the allele frequencies in the resulting SNP datasets would be incorrect, and potentially bias the results of downstream population genetics analyses (Mastretta-Yanes et al. 2014). Catchen et al. (2013b) therefore "strongly encourage researchers to test a range of values for each parameter when approaching a data set for the first time".

In the present thesis, a large, multi-species RADseq dataset is used to address several evolutionary questions pertaining to the crucian carp, *Carassius carassius* (L.) including the genetic impacts imposed on this species through hybridisation with three non-native

taxa; the goldfish *Carassius auratus* (L.), the gibel carp, *Carassius gibelio* (Bloch), and the common carp, *Cyprinus carpio* (L.). However, this dataset comes with several specific challenges, which must be overcome in order to produce SNP datasets which are suitable for evolutionary analyses. The first of these challenges lies in the possibility of allele dropout, which is caused by mutations in the restriction site of a RAD tag (Gautier et al. 2013). Allele dropout is likely to be present in all RADseq datasets to some extent, for example, Luca et al. (2011) found that, within a human RADseq dataset, null alleles in RAD sequencing data resulted in 9.4% of heterozygous loci being erroneously genotyped as homozygous, which in turn caused an underestimation of sequence diversity of 3%. However the problem is likely to become more pervasive in high-polymorphism, multi-species datasets like the one used in the present study (McCormack et al. 2013). Gautier et al. (2012) showed, with simulated data, that these genotyping errors can cause, somewhat counter intuitively, an inflation of heterozygosity within populations. This is due to null alleles being more likely to occur in ancestral alleles because these are likely to be the more abundant allele in the population. The dropout of these ancestral alleles would therefore increase the minor allele frequency at these loci and thus increase heterozygosity estimates.

The second major challenge presented by the RADseq dataset in this study is the presence of ohnolog loci, that is, paralogous loci that are the result of the multiple genome duplication events (GDE) that have occurred throughout the evolution of Cyprinids (Ohno 1973; Glasauer and Neuhauss 2014). As these loci potentially have very high sequence similarity, they might be erroneously merged into one locus during *de novo* locus construction in Ustacks, resulting in an inflated proportion of heterozygous loci. The probability of this over-merging is dependent on the amount of sequence divergence between the true alleles at a locus, and between ohnolog loci; if the former is greater than the latter at a given diploid locus, or if ohnolog pairs are monomorphic, then there is a strong probability that ohnologs will be merged during the Ustacks phase of Stacks processing. The possibility of merging ohnolog loci is present once again at the catalog building stage (Cstacks), where, as well as true orthologous alleles, there is also the possibility of merging one or both alleles from any of the ohnologs present in other individuals.

Overcoming the problem of ohnologs in individuals from the pure species in this study is theoretically straight forward. Although the divergence of *C. carpio* and the

Carassius genus has not, to our knowledge, been dated, it is assumed that the split between these species occurred more recently than the genome duplication (dated at 8.2 million years ago Xu et al. *(2014)*) on the basis that they share the same number of chromosomes (50). Therefore, the number of mutations between ohnologs should generally exceed the number that exist between these species and so, the use of conservative mismatch thresholds when combining alleles at the same locus in Ustacks (e.g. M=2) should limit their erroneous separation. However, in the case of the F1 hybrids between our study species, the allelic divergence at a given locus is likely to be much higher, as it represents the divergence between the parental species. There will therefore be a larger overlap between the number of mismatches present between homologs and those present between ohnologs in hybrids compared to pure species samples (Figure 2.1). Although this situation again calls for conservative mismatch thresholds in Ustacks and Cstacks, using such cut-offs in hybrids is likely to cause problems of its own, resulting truly heterozygous loci with highly diverged alleles being erroneously identified as two monomorphic loci. Therefore, although it is an important factor in all samples in this dataset, the trade-off between the incorporation of over-merged ohnolog loci and the under-merging of true loci, is likely to be most pronounced in hybrids.

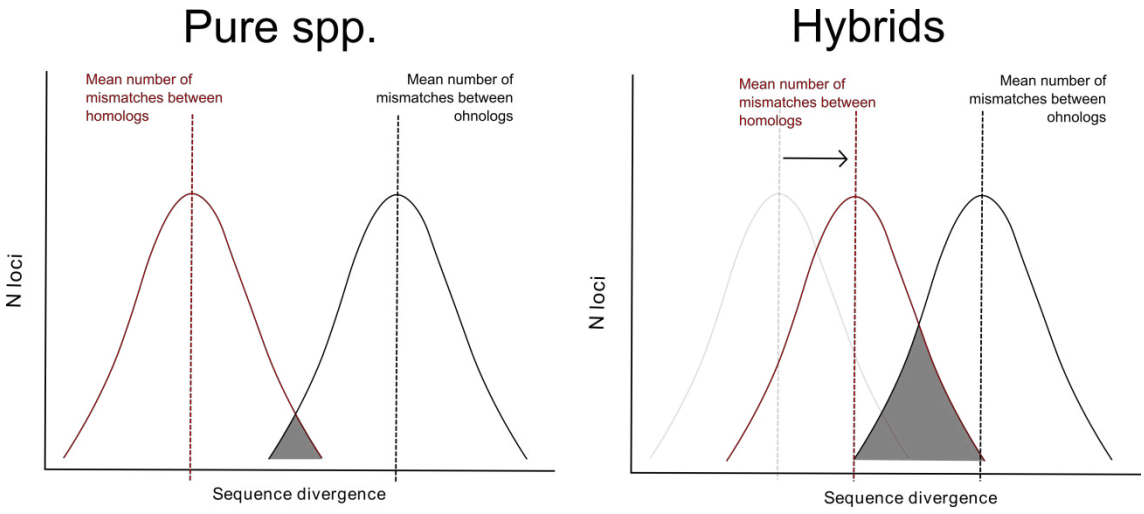


Figure 2.1. Schematic illustrating the hypothetical distribution of sequence diversity between alleles at the same locus (red) and alleles across ohnologous loci pairs (black), showing higher interalleleic divergence in hybrids due to their interspecific parentage. The areas shaded grey represent loci in which alleles from both ohnologs are likely to be wrongly merged into a single locus during the Ustacks stage of Stacks analyses.

A potential way around this problem may be the incorporation of a reference genome into the Stacks analysis pipeline, which confers several advantages over the *de novo* construction of loci. Firstly, reads mapping to more than one location on the reference genome (i.e. ohnologs) can be removed before locus construction, allowing for higher mismatch threshold to be used in the mapping without the danger of merging ohnologs. Secondly, the aligners (e.g. BWA (Li and Durbin 2009), BOWTIE (Langmead and Salzberg 2012)) used can incorporate loci with insertions and deletions (indels) between samples, which Ustacks cannot. This may be particularly important in the present system, as indels are thought to occur more often following whole genome duplications (Guo et al. 2012). And lastly, the use of a reference genome allows for linkage information to be drawn between RAD tags, allowing for more sophisticated analyses downstream (Catchen et al. 2013b).

A second potential way of accounting for ohnologs is to perform post-hoc population genetic filters on SNP datasets resulting from Stacks (Hohenlohe et al. 2011). The erroneous merging of two ohnolog loci is likely to create a false heterozygous locus. Such loci are expected to display elevated observed population heterozygosity or low F_{IS} compared to Hardy-Weinberg expectations, hence, these population genetics indices can be used to filter out putative merged ohnolog loci.

In light of the above factors, there is a need for comprehensive testing of the impacts that various core parameter values have on the present RADseq dataset before it can be used to test the evolutionary questions of this thesis. The aims of this chapter are to: 1) Test multiple values for the core parameters within the Stacks pipeline, in order to characterise their potential effects on the present dataset. To this end, a custom python module, Incremental, is developed, which automates the systematic incrementation of parameter values for Ustacks, Pstacks, Cstacks and the Populations modules of Stacks, and produces useful graphical outputs to enable the user to make informed decisions on optimal parameter values based on their own data. 2) Use the reference guided Stacks pipeline to construct RAD tags using the Xu et al (2014) *C. carpio* genome draft and compare the resulting SNP dataset to that of the *de novo* locus construction pipeline. 3) Use population genetics approaches to identify problem loci, with particular emphasis on those containing null alleles and ohnolog loci. 4) Apply the optimal values, chosen on the basis of Stacks parameter tests and populations genetics filters, in order to produce the final refined SNP datasets for Chapters 3 and 5 of this thesis.

Methods

Sample collection and RAD library preparation

In total, RADseq data were obtained for 247 fish samples from 32 populations, spread across 10 countries. Samples were initially identified as belonging to four different species, the crucian carp, *Carassius carassius* (L.), the goldfish, *Carassius a. auratus* (L.), the gibel carp, *Carassius a. gibelio* (Bloch), the common carp, *Cyprinus carpio* (L.), and hybrids between these species (Table 2.1) on the basis of morphological identification carried out in the field and microsatellite analyses using 6 species diagnostic loci described in Chapter 4.

Tissue sampling and DNA extraction protocols are described in detail in Chapter 3 of this thesis. The resulting DNA was quantified using the Quant-iT™ PicoGreen® dsDNA Assay kit (Invitrogen) and normalised to concentrations greater than 50 ng ml^{-1}. Gel electrophoresis was used to check that DNA extractions contained high molecular weight DNA and the high quality DNA samples were then used for RAD seq library preparation at Genepool (now Edinburgh Genomics) at Edinburgh University, UK, according to the protocol described in Davey et al. (2012). Paired-end sequencing was then performed, again at Genepool on the eight resulting libraries using two lanes of an Illumina HiSeq 2000 sequencer.

Raw data quality checking, PCR-clone filtering and trimming

The analyses were performed using only the first-end reads from the paired-end sequencing, as coverage across 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. Secondly, using the "process_radtags" script distributed with Stacks, data were demultiplexed, using the barcodes which denote individual samples, reads containing low Phred (quality) scores were filtered from the dataset, and partial restriction sites were trimmed. This resulted in refined raw-read files, which each contained only high quality reads of a single sample for either the first

or second read-pair. Thirdly, the demultiplexed raw reads were filtered for PCR duplicates using the "clone_filter" program (also distributed with Stacks), which can arise when two or more PCR copies of the same template DNA molecule are sequenced (Davey et al. 2013; Tin et al. 2014). And finally, all first-end reads were trimmed using a custom python script, from the 5' end, to a length of 92bp for use in the Stacks analysis pipeline.

Stacks parameter tests

Incremental

To explore the effects of the different Stacks parameters on the dataset in this study, it was necessary to perform test-runs of the pipeline using multiple values of each parameter. In order to do this in a time-efficient and systematic way, the python module, Incremental, was written and implemented. Incremental is a set of wrapper functions which run the various modules of Stacks for a user-defined set of values for a specified parameter. The outputs of these Stacks runs are then parsed and used to create informative plots (Figure 2.2.) which can, in turn, be used by the user to make informed decisions about the final analyses parameters. Specifically, in the Ustacks tests, tag number per individual, and average tag coverage within an individual are plotted. For Cstacks, the number of tags present in each catalog are plotted, along with the number of tags shared between individuals in the catalog. And for Populations, the number of SNPs, SNP coverage and number of SNP dropouts in each population are plotted. These plots were used to identify the optimal parameter for at Stacks parameter for this dataset. However, the choice of which parameter value is deemed optimal is inherently subjective (Catchen et al. 2013b; Mastretta-Yanes et al. 2014). In the present study, optimal parameter values were taken to be those where the rate of change in the properties of Stacks outputs, i.e. the number of RAD tags produced or the read depth across RAD tags, slowed or became negligible.

Table 2.1. Locations and sample numbers for each population for which RADseq was obtained, and their use in the chapters of this thesis

				Coordinates		Sample Number		
Code	**Location**	**Country**	**Drainage**	**Lat**	**Long**	**Total**	**Chapter 4**	**Chapter 5**
BEL5	Dendermonde	Belgium	Scheldt River	51	4.09	5	5	5
DEN1	Copenhagan	Denmark	Baltic Sea	17.8	60.2	10	10	10
DEN2	Pederstrup	Denmark	Baltic Sea	12.6	55.8	8	8	8
DEN3	Bornholm Island	Denmark	Baltic Sea	14.9	55.2	5	5	5
FIN2	Helsinki	Finland	Baltic Sea	60.4	25.3	6	-	6
FIN3	Jyväskylä	Finland	Baltic Sea	62.3	25.8	10	10	10
FIN4	Oulu	Finland	Baltic Sea	65	25.5	8	8	8
GBR10	Norfolk	U.K.	U.K	52.9	1.1	15	-	15
GBR14	Epping Forest	U.K.	Thames	51.7	0.04	2	-	2
GBR15*	Buntingford	U.K.	Thames	51.9	-0.01	5	-	5
GBR16	Epping Forest	U.K.	Thames	51.7	0.03	6	-	6
GBR17	Ings Lane Garden Centre, Hull	Ornamental of unknown origin		53.8	-0.36	5	-	5
GBR4	Norfolk	U.K.	U.K	52.8	0.75	9	9	9
GBR6	Norfolk	U.K.	U.K	52.5	0.93	13	10	13
GBR7	Norfolk	U.K.	U.K	52.9	1.15	10	10	10
GBR8	Hertfordshire	U.K.	U.K	52.9	1.1	9	9	9
HUN2	Vörösmocsár	Hungary	Danube River	19.2	46.49	6	6	6
HUN4	Lake Kolon	Hungary	Danube River	47.5	19.1	8	8	8
NOR2	Lake Prestvattnet, Tromsø	Norway	North Sea	19	69.7	9	9	9
POL3	Tupadly	Poland	Vistula River	52.7	19.3	10	10	10
POL4	Orzysz	Poland	Vistula River	53.8	22	10	10	10
RUS1	Proran Lake	Russia	Don River	47.5	40.5	9	9	9
SWE2	Stordammen	Sweden	Baltic Sea	59.8	17.7	10	10	10
SWE10	Norrköping	Sweden	Baltic Sea	58.6	16.3	9	9	9
SWE12	Osterbybruk Mansion	Sweden	Baltic Sea	12.3	55.7	9	9	9
SWE14	Wenngarn Castle, Stockholm	Sweden	Baltic Sea	19	59.7	9	9	9
SWE8	Skabersjo	Sweden	Baltic Sea	55.6	13.2	10	10	10
SWE9	Märsta	Sweden	Baltic Sea	59.6	17.8	10	10	10
SWE20*						4	-	4
UKR1	Vil Laskivtcy, Ternopil	Ukraine	Dniester	49.2	25.6	3	-	3
UKR2	Reut River, Floresti	Ukraine	Dniester	47.8	28.4	5	-	5
					Totals	**247**	**193**	**247**

De novo Stacks pipeline parameter tests

Ustacks

Locus assembly within and between individuals, and subsequent SNP calling was performed using the Ustacks module of Stacks v.1 (Catchen et al. 2013). Two main parameters of Ustacks were tested, using Incremental, for their effect on the resulting SNP datasets, -M, between 0 and 8 maximum mismatches allowed between reads in a stack, -m, between a minimum of 1 and 8 reads per stack (Figure 2.2a, Table 2.2). Previous studies have also examined the –max_locus_stacks parameter (Catchen et al. 2013; Mastretta et al. 2014), which specifies the maximum number of stacks of unique sequence reads allowed to be incorporated into a single locus. Each of these stacks is expected to represent the reads from one allele, therefore, a single diploid locus should contain two stacks of unique reads. However, it is often the case that short stacks of reads containing sequencing errors can be produced, and their incorporation (by specifying –max_locus_stacks = 3) can increase the coverage at a locus (Catchen et al. 2013). As the potential presence of ohnologs in the current dataset precludes the use of a high value for this parameter the --max_locus_stacks parameter was not tested in this study and was kept at a value of 3 (default) for all –M and –m tests in Ustacks. All other non-test parameters were also kept as the Stacks defaults and, in all tests the removal and deleveraging flags (-r, -d) were used to deal with highly repetitive loci. To reduce computation time these parameter ranges were tested in a subset of 33 samples containing 17 randomly chosen *C. carassius*, 2 *C. carpio*, 2 *C. carpio* x *C. carassius* hybrids, 3 samples from both *C. auratus* and *C. gibelio*, and 3 samples for both *C. carassius* x *C. auratus* and *C. carassius* x *C. gibelio* hybrids (all hybrids were confidently identified using microsatellites, see Chapter 5). These samples were chosen so that each species was represented in the Stacks parameter tests and so that in Cstacks tests, each catalog contained a unique combination of samples from each species (see below). On the basis of the results of these tests, final Ustacks parameters were chosen and Ustacks was run a final time on each subsample, in order to produce the final Ustacks outputs for use in the Cstacks and Populations module tests described below. For Ustacks and the other Stacks modules, the general criterion used for choosing the "optimal" parameter values was the rate of change in the properties of the resulting dataset, e.g. the number or RAD tag loci, the coverage of the loci (Ustacks and Cstacks) and the number of loci that dropped out in each population (populations). In cases where these choices were arbitrary due to no clear "optimal" parameter value, the most

conservative values were chosen from the range of those suitable in order to limit biases from allele dropout and ohnolog merging.

Cstacks

For the Cstacks parameter tests, the 33 test samples were split into eight species-pair subsets, which were used to create eight separate loci catalogs in Cstacks. Two of these catalogs contained one *C. carassius,* one *C. carpio* and one hybrid, three contained one *C. carassius*, one *C. auratus* and one hybrids, and the remaining three catalogs contained one *C. carassius,* one *C. auratus* and one hybrid between them. In these tests, the –N parameter was tested for a range of 0-8, this parameter dictates the maximum number of mismatches allowed at a locus between individuals for them to be deemed homologous and incorporated into the locus catalog (Figure 2.2b, Table 2.2). For these tests all other Cstacks parameters were left at their default values.

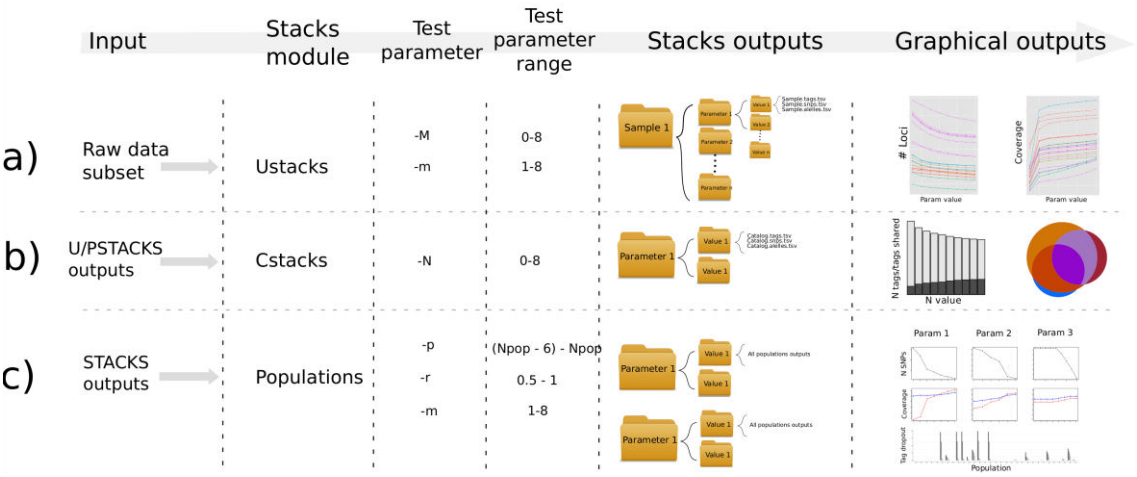


Figure 2.2. Schematic showing the functionality of the Incremental python module for a) Ustacks, b) Cstacks and c) Populations modules of Stacks. Test parameter ranges are those used in the present Chapter, however can be any user specified range within Stacks program limits.

Populations module

Populations module tests were performed separately on the full RADseq data set (including all species and hybrids; n = 247), which was analysed in Chapter 5 (n = 247), and the subset containing only pure Crucian carp (n = 188, Table 2.1), which was analysed in Chapters 4 and 6. Based on the results of the Ustacks and Cstacks tests in the *de novo* Stacks pipeline, optimal parameter values were chosen and used to make separate *de novo* Stacks catalogs for the two data sets. Sstacks was then run on these two catalogs and the resulting outputs were used for the Populations module parameter tests. The parameters tested were the -m (values 1-8) parameter, which specifies the minimum read depth at a locus for the locus to be retained in an individual; -r (0.5 – 1),

which specifies the minimum proportion of individuals in each population that must contain data at a locus for that locus to be retained; and –p ((npops -6) – npops), which specifies the minimum number of populations which must meet the -r cut-off at a locus for that locus to be retained in the dataset (Catchen et al. 2013b) (Figure 2.2, Table 2.2).

Table 2.2. Parameters tested in Stacks tests and the chosen values used in Chapters 3 and 5.

Parameter		Description	Tested value range	Optimal value chosen in *de novo* *C. carassius* only dataset / full dataset
Ustacks	-M	number of mismatches between stacks of unique reads at a locus within an individual	0-8	2 /reference*
	-m	minimum threshold for the number of reads required at a locus	1-8	8 /reference
Cstacks	-N	maximum number of mismatches allowed at a locus between individuals	0-8	2 /reference
Populations	-r	minimum proportion of individuals in each population that must contain data at a locus	0.5-1	0.7/0.8
	-p	minimum number of populations which must meet the -r cut-off at a locus	npops-6 - npops*	17/28
	-m	minimum threshold for the number of reads required at a locus	1-8	8/8

* NOTE: “reference” refers to the use of the reference guided assembly for these steps.

Reference guided Stacks analyses

To allow for comparison with the results of the *de novo* Stacks pipeline, we ran the entire dataset through the reference guided Stacks pipeline. Raw reads were first mapped to the *C. carpio* reference genome (Xu et al. 2014) using the Burrows-Wheeler aligner (BWA, Li and Durbin 2009), allowing for a maximum of 6 mismatches and retaining only reads which mapped uniquely to the reference genome. The aligned raw reads for the remaining 31 populations were then passed to the Pstacks module of stacks, specifying a minimum of four reads required to build a stack. The resulting constructed loci were then passed to the Cstacks module and again loci were merged across individuals using genomic location (i.e. the -N parameter was not used here). Sstacks was then run on this catalog and these outputs were used in the Populations module parameter tests, in which the same parameters and value ranges as in the *de novo* tests were tested (Figure 2.2c, Table 2.2). As this reference guided pipeline uses genomic location to construct loci (Pstacks) and merge loci across individuals (Cstacks), no parameters were tested in these modules.

Identifying and accounting for allele dropout in RAD tags

In order to identify RAD tags which putatively contained null alleles between species, two approaches were taken. The first was the examination of the RAD tag sharing patterns between hybrids and parental species in the Cstacks test catalogs. It was expected that, if a RAD tag has a null allele between two species, the hybrid would contain only one allele for this locus, which would be shared with only the donor parental species in the catalog. The second approach was performed on the final SNP dataset resulting from the full dataset analyses in both *de novo* and reference guided Stacks pipelines. Loci containing a null allele in one of the two species would be homozygous in hybrid samples and should exhibit half the coverage of a correctly assembled locus (Gautier et al. 2013). SNP loci were therefore compared for these two properties in hybrids for all values of -p, in both *de novo* and reference guided Populations module tests.

Identifying and accounting for ohnologs

Putative over-merged ohnologs were filtered from the pure *C. carassius*-only and the full reference aligned datasets only, as the *de novo* full dataset was not used in the final analyses in this thesis (see Results & Discussion). This was achieved using the approach implemented in Hohenlohe et al. (2013), whereby all loci which were heterozygous in more than 50% of individuals in a population or loci for which $F_{IS} < 0$ were removed. However, in the full reference aligned dataset, some populations are almost entirely comprised of hybrids, which, due to their interspecific parents, were expected to truly possess very high heterozygosity. For this reason, hybrids were not included in the assessment of heterozygosity in a population for this dataset, however any loci found to be putative over-merged ohnologs in the *C. carassius* only populations were, indeed, removed from hybrids as well. In order to assess the effectiveness of these filters, the average coverage distributions for all tags across all individuals in the datasets were assessed before and after locus filtering.

Results & Discussion

Raw data quality checking and cleaning

All 247 samples were successfully sequenced using RADseq, which resulted in between approximately 690 000 and 4 500 000 raw sequence reads per sample, with a mean of approximately 2 200 000 reads. However two populations (SWE2, SWE14), which had low DNA concentration in library preparations, showed low read numbers and subsequently showed high locus dropout rates due to low coverage. PCR clone filtering removed an average of 323 015 reads per sample and subsequent FastQC analyses did not identify any individuals that had low overall sequence quality, therefore all samples were retained for further analyses.

De novo pipeline parameter tests

Ustacks

Parameter tests for the Ustacks module showed that both -M and -m have large impacts on the number of tags, tag coverage and the sharing of tags between individuals, and agreed well with those observed in previous studies (Catchen et al. 2013b; Mastretta-Yanes et al. 2014). For all samples the number of RAD tags initially dropped quickly when increasing the mismatch allowance between reads in a stack (-M) and this reduction in tags then slowed as the value for -M increased (Figure 2.3a). This pattern is indicative of alleles at heterozygous loci that are separated at low mismatch thresholds being merged into single loci at higher thresholds. In line with this, the average coverage across loci in all individuals initially increases quickly, likely due to singleton reads resulting from sequencing error being incorporated into loci as suggested in (Catchen et al. 2013a). This increase in coverage then slows as –M is increased further.

Increasing the number of minimum reads required to build a stack (-m) from 1-2 resulted in a drastic decrease in RAD tags constructed, as single reads representing sequence error were discarded (Figure 2.3b). Further increases of –m above two resulted in only minor decreases in the number of loci constructed. Again, the average coverage at each locus mirrored this pattern, increasing dramatically from 0-2 reads required per stack as the number of loci decreased (Figure 2.3b).

Interestingly, a notable difference was observed between hybrid and pure species samples in all Ustacks tests, whereby hybrids possessed considerably more loci and much lower average locus coverage than pure species samples (Figure 2.3). These patterns are consistent with a failure to merge two alleles at a heterozygous locus together in the Ustacks analysis stage. Indeed, hybrids would be expected to contain higher inter-allelic variation than pure species samples, due to the inter-specific origins of the alleles at each locus. Therefore the mismatch threshold at which alleles at a heterozygous locus merge in hybrids would be higher than that in pure species. However, even when allowing for 9 mismatches between alleles at a locus (which would constitute approximately 10% sequence divergence between species), the number of loci in hybrids did not decrease to the same numbers as in the pure species (Figure 2.3). One potential explanation for this may be the presence of indels, which the *de novo* Stacks pipeline cannot process and thus, any loci with indel mutations between alleles would not merge. However, this possibility is refuted by the fact that the pattern of high RAD tag number and lower coverage is also seen in the reference guided locus assembly (see below) in Pstacks, which can account for indel mutations. A more likely explanation, therefore, is the presence of null alleles at these loci, which result from a mutation in the restriction site of one parental species and not another (Gautier et al. 2012). In such a case, the hybrid would possess a RAD tag which represents only one allele for that locus, which has approximately half the coverage of an intact locus and will drop out in one of the parental species.

Cstacks

Based on Ustacks tests, the final Ustacks parameters chosen to create the Cstacks tests inputs for all eight three-sample subsets were -M=4, -m=2 (see "*Final parameter values and SNP dataset production*" section for justifications for using these values). For all values of -N (maximum number of mismatches allowed between individuals at a catalog locus, 0-10), the total number of RAD tags in the catalogs containing samples from more than one species was considerably higher than for those containing only *C. carassius* (Supplementary Figure 2.1). In all catalogs, as -N was increased, the total number of tags decreased and the number of tags shared between all individuals in the catalog increased. Interestingly, for all multi-species catalogs, there were a large proportion of RAD tags which were differentially shared between the hybrid and one parental species, but not the other (areas of Venn diagrams shaded dark orange or lilac Figure 2.4). Due to the observation that, even when using an allowance of ten

mismatches between individuals, this differential sharing pattern remained (Figure 2.4), it is likely that these loci contain null alleles between species and were the same loci that were responsible for the inflated RAD tag numbers in hybrids. Finally, there were a moderate proportion of loci in multispecies catalogs which were present only in the hybrid individual, and not the parents. We hypothesise this is caused by the fact that the pure species individuals in each catalog are not the true parents of each hybrid. As there are multiple lineages within *C. auratus* and *C. gibelio* (Rylokova et al. 2013), it is likely that allele dropout is occurring within these species, for example between ornamental and feral strains of goldfish.

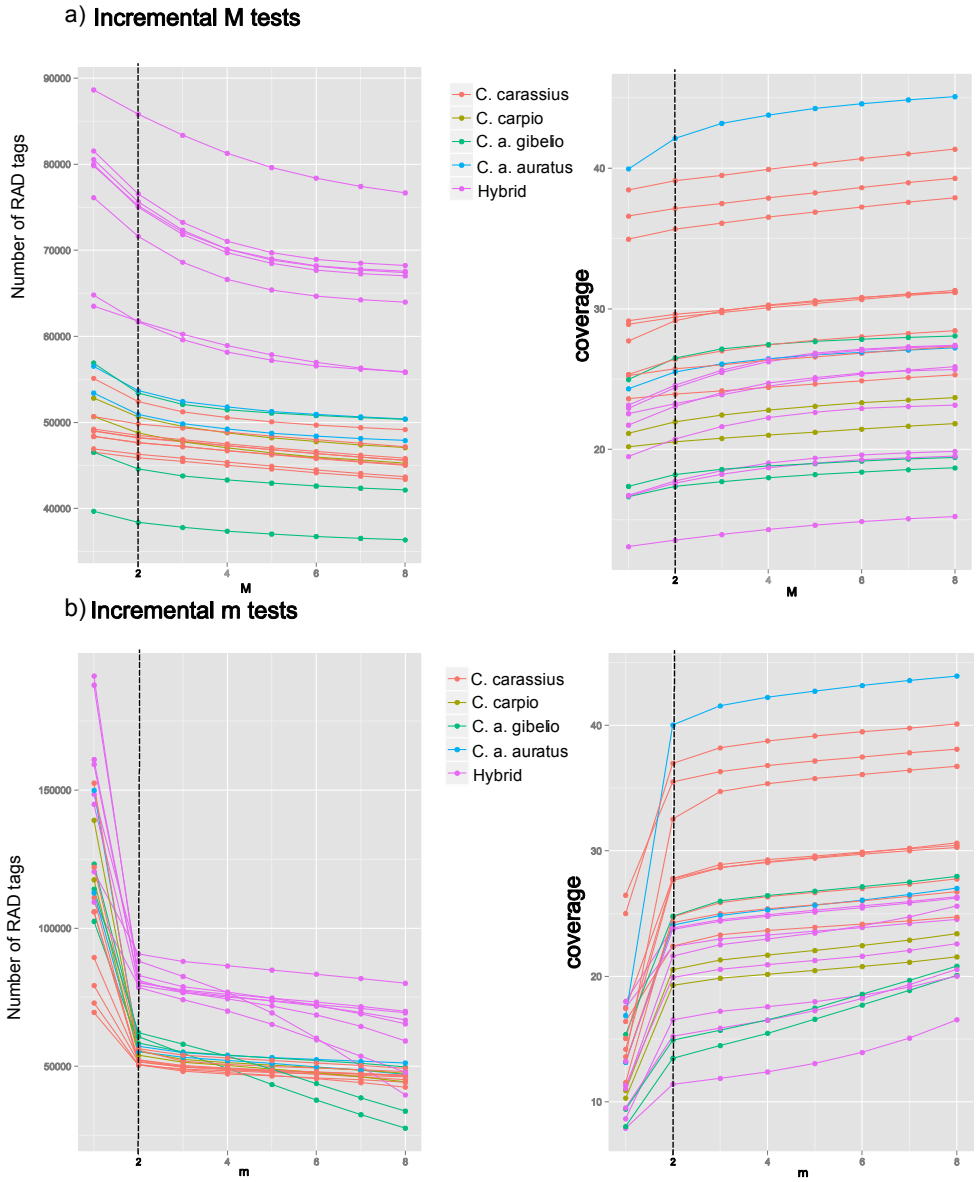


Figure 2.3. Tag number and coverage within each of the 33 stacks parameter test samples for a) the –M parameter and b) the –m parameter in the Ustacks module of Stacks. The chosen values for each parameter are denoted by the dashed black lines.

Populations module

In all *de novo* datasets, tests performed on the -p parameter in Populations module tests showed that, as the minimum number of populations required to be present at a locus (-p) was decreased, the number of SNP loci in the final outputs increased, as loci that drop out in one or a few populations were added to the dataset (Supplementary Figure 2.2a, 3a, 4a). In the full dataset, for both *de novo* and reference guided pipelines, the populations that were most likely to drop out in these extra loci were those containing only, or predominantly *C. auratus* spp., *C. carpio* or hybrid samples (Supplementary Figure 2.2d, 4d), reflecting the number of loci that drop out between these species and *C. carassius*. In the data set containing only pure *C. carassius,* RAD tag dropout was most prevalent in the populations which had low DNA quality (SWE2, SWE14) or contain highly divergent populations from the Danube or the Don river catchments (HUN2, RUS1, Supplementary Figure 2.3d). In all cases the average coverage across tags within individuals decreased as the number of populations required at each locus increased. However, this reduction was much more pronounced in hybrids than in pure species at -p values of 26 and lower (Supplementary Figure 2.2), which is likely attributable to the inclusion, in the hybrids, of loci containing a null allele for the non-*C. carassius* parent. The comparison of coverage and heterozygosity at each SNP locus identified many loci in both the *de novo* and the reference aligned full datasets that fit the predictions of containing null alleles (see circled loci in Figure 2.5). As RAD tags containing a null allele for the non-*C. carassius* spp. dropped out in populations containing only or predominantly non-*C. carassius* spp., these loci were gradually excluded from the dataset as the value of –p was increased above –p=26.

For tests of the –r option (the proportion of individuals in a population required to be genotyped at a locus), the number of RAD tags decreased dramatically as –r increased (Supplementary Figure 2.2b, 3b, 4b). The patterns of this decrease in RAD tag number was almost identical between the *de novo* and reference aligned full datasets and was most pronounced for the change in -r from 0.8 to 0.9. This pattern is likely driven by the presence of two *C. carpio* individuals in populations GBR6. At -r=0.8 these individuals are not required to be present at a locus which therefore allows the inclusion of loci which dropout in *C. carpio*. In contrast, the absence of more than one species in the pure *C. carassius* only dataset, lead to a more steady decrease in the number of RAD tags as -r was increased. In all cases, as the number of tags decreased with increasing -r,

the average tag coverage increased, however, again this was most noticeable in the hybrids in the *de novo* analyses, which showed a dramatic increase in tag coverage, coinciding with a loss of approximately 3000 RAD tags as the -r value increased from 0.8 to 0.9.

Reference guided Stacks analyses

In general, only between 42.8% and 45.7% of reads mapped to the *C. carpio* reference genome in *Carassius* spp. samples and their hybrids, suggesting very high levels of polymorphism at these loci (i.e. higher than 6 bp / 92bp sequence) between the *Cyprinus* and *Carassius* genera. Of these aligned reads, between 18.4% and 21.1% mapped with high confidence to more than one location, likely due to the occurrence of an ohnolog pair in which both loci are less than six mismatches different from the read. These multi-hit reads were therefore discarded, leaving between 24.4% and 25.3% of reads which uniquely mapped in these samples. In contrast, 90% and 66.4% of reads mapped to the genome in *C. carpio* and *C. carpio* x *Carassius* spp. hybrid individuals respectively (Supplementary table 2.1), with 20.1% and 20.3% of these reads being discarded for mapping to more than one location. The remaining 10% of reads that did not map in *C. carpio* individuals likely represents a combination of contamination (at the library preparation stage) and gaps in the *C. carpio* reference genome assembly

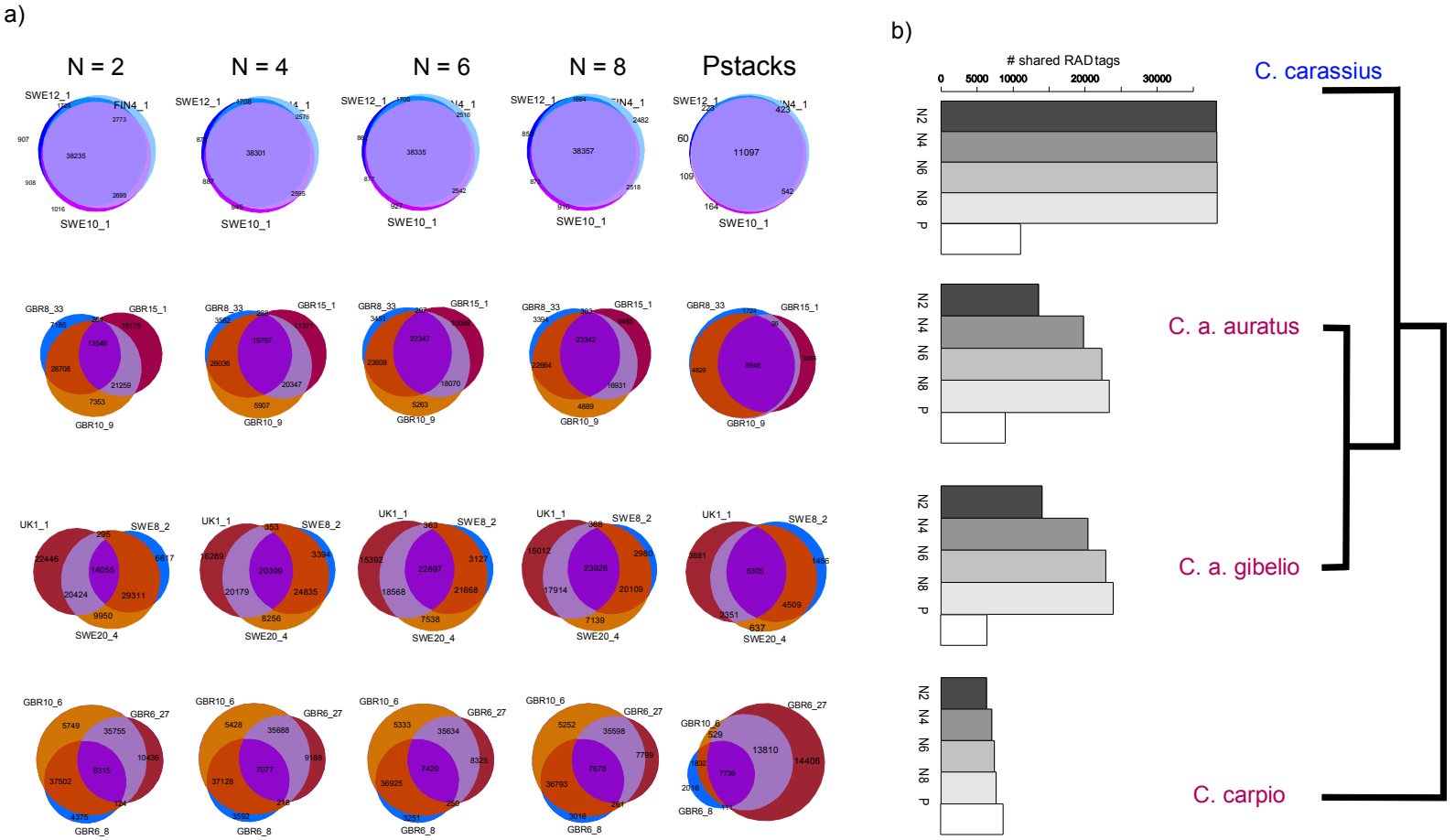


Figure 2.4. a) Venn diagrams showing the differential patterns of tag sharing between species and hybrids within each test catalog produced in the Cstacks module parameter tests. For each combination of species, only one test catalog is shown, however sharing patterns are indicative of those in all other catalogs. Note that Venn area sizes are not proportional. Bar plots in b) show the number of tags shared between all three individuals in each catalog, which decreases with increasing taxonomic distance between species. Chosen value of N (2) denoted by the red box.

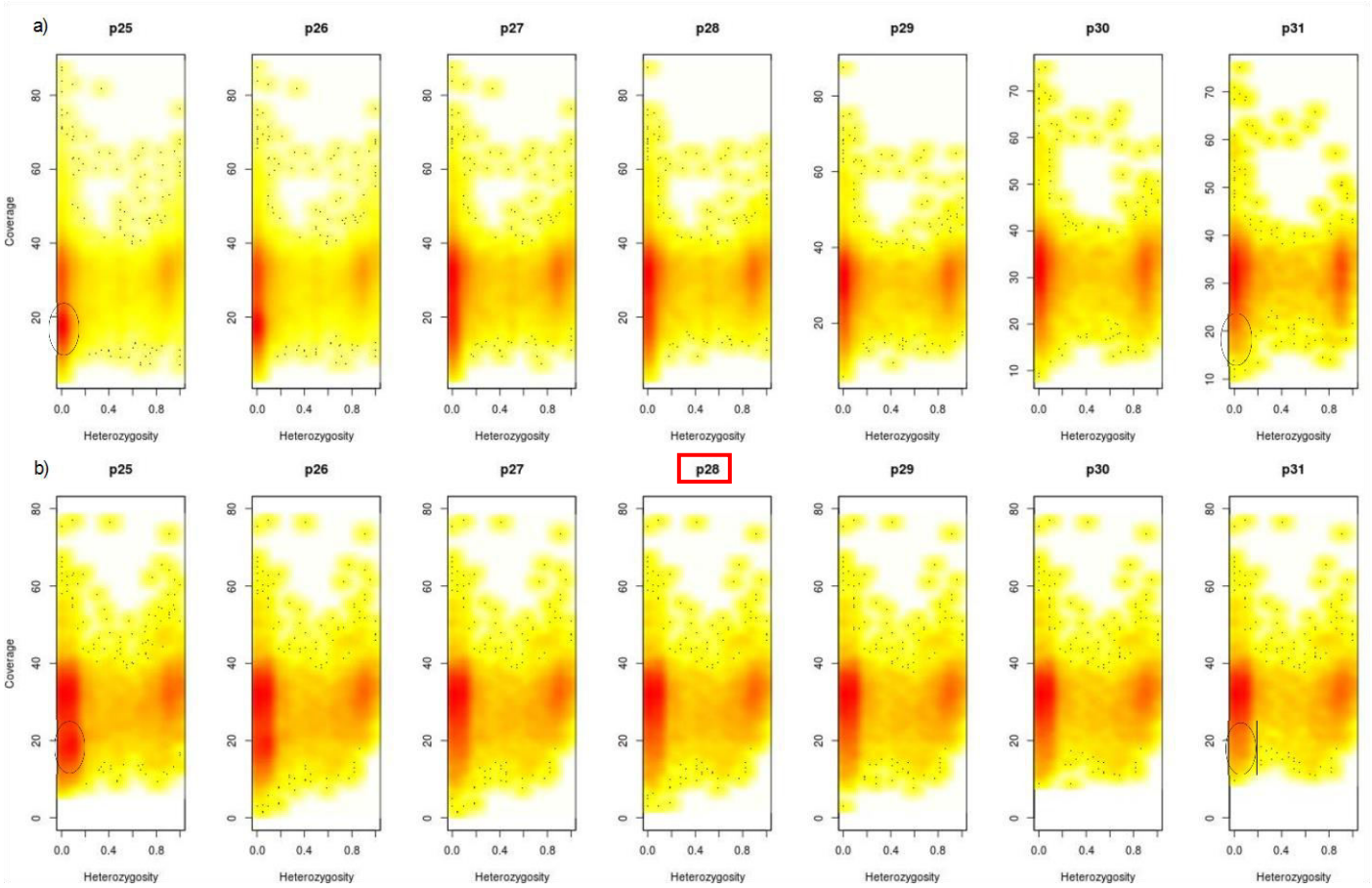


Figure 2.5. Smoothed scatter plots of heterozygosity and coverage at each SNP locus for each value of the –p option tested in the Populations module of stacks for a) the *de novo* and b) the reference guided datasets. Red heat colours represent high numbers of loci, yellows represent low numbers of loci. Circled regions highlight loci which fit the assumptions of possessing null alleles in the non-*C. carassius* samples in the dataset. The value used in the final analyses of this dataset in the reference guided pipeline was –p = 28.

The number of loci constructed for *C. carassius* and *C. auratus* spp. in the reference guided analysis was considerably smaller than for the same catalogs constructed with the *de novo* pipeline (Supplementary Figure 2.1), reflecting the small percentage of reads that were not discarded during mapping for *Carassius* spp. However, the catalogs containing *C. carassius*, *C. carpio* and hybrids contained more RAD tag loci which were shared between all 3 individuals in the catalogs, than the same catalogs constructed with any -N values in the *de novo* Stacks pipeline (Figure 2.4). Nevertheless, the patterns of tag and coverage change in all reference guided catalogs were generally very similar to those of the *de novo* analyses, with the number of tags decreasing and the number of shared tags increasing as the value for -N increased. As in the *de novo* catalogs, the reference guided catalogs also showed the same differential tag sharing between hybrids and one parental species only (Figure 2.4), however in the *C. carassius* x *C. carpio* catalogs, this was considerably biased towards the *C. carpio* individuals (for example see Figure 2.4a). This bias is likely caused by the combination of two types of null alleles, firstly, those that result from restriction site polymorphisms in *C. carassius*, and secondly those caused by only the *C. carpio* allele in the hybrid, and not the *C. carassius* allele, mapping to the reference genome. Finally, in the Populations module, the patterns of RAD tag number, coverage and locus dropout among populations were almost identical to the *de* novo analyses (Supplementary Figure 2.4). However, one notable difference was the increased reduction in coverage as the –p value was increased in *de novo* analyses compared to the reference aligned analyses.

Null alleles in RAD tag loci

As described briefly above, evidence of null alleles in the present dataset is prevalent at each stage of the Stacks analysis pipeline most notably when analysing multi species data sets. These loci have the potential to introduce considerable bias in the final SNP dataset. In particular, the Populations module of Stacks is very important for the inclusion or exclusion of this bias in the present dataset, due to the uneven number of samples representing each species in this dataset. Out of the total 32 populations, five contain only non-*C. carassius* samples and the remaining 27 populations contain only or predominantly *C. carassius* or hybrid individuals, therefore, allowing five populations to drop out at a given locus in the Populations module filters (i.e. -p=27) would allow for the inclusion of loci with a null allele for the non-*C. carassius* species, as these loci

would be present in all *C. carassius* and hybrid populations. Indeed, there is strong evidence that this happens in the *de novo* dataset, as coverage in hybrids decreases rapidly as the -p value decreases to 27 and lower (Supplementary Figure 2.2). Similarly, as the -r parameter is decreased to 0.8 or lower a considerable number of additional loci are added to the final SNP dataset. This effect is likely driven by the presence of the two *C. carpio* individuals in population GBR6. When –r = 0.8 or lower, these two individuals would be allowed to drop out at a given locus, therefore, loci containing a *C. carpio* null allele would be incorporated into the resulting SNP dataset. Importantly, in these cases, the hybrid would contain large numbers of SNP loci which are wrongly genotyped as homozygous due to their null alleles. These effects become particularly relevant in the context of identifying hybridisation and introgression. In Chapter 4 of this thesis, a program called Newhybrids (Anderson and Thompson 2002) is used, which uses the frequency of parental alleles in hybrids to assign them to one of several hybrid classes (e.g. F1, F2, backcross). The use of Populations filters such as those described above would result in many loci which appear to be homozygous for the *C. carassius* allele in hybrids, which, in Newhybrids analysis, would produce false signal for backcrossing when in fact only F1 hybrids exist. Furthermore, in the context of identifying introgression, it is envisaged that such a bias in the data would result in reduced evidence of gene flow between species, as signal in heterozygous loci showing allele dropout between species would go unobserved.

One suggested approach to filtering for loci showing null alleles in RADseq data is the filtering for loci with reduced coverage (Gautier et al. 2012). In the present study, this approach has been used in conjunction with the assumption that loci which show null alleles between species will exist in hybrids, but only in homozygous state. It is also shown here that the use of stringent sample and population filters is efficient at removing these loci from the final SNP datasets.

In the present chapter, only allele dropout *between* species is considered, however, it should be noted that this phenomenon is also likely to occur *within* species. Although such loci are likely to be highly challenging to identify (McCormack et al 2013), within-species allele dropout is likely to be much less prevalent than between species, due to lower polymorphism levels.

Filtering for ohnologs

Filtering loci on the basis of elevated observed heterozygosity and negative F_{IS} removed 15.4%, and 13.9% of RAD tags from the full reference aligned and pure *C. carassius* datasets respectively. In support of the hypothesis that these loci are indeed over-merged ohnologs, examination of the average coverage across all samples at each tag showed a bi-modal distribution consistent with a considerable number of tags showing much higher coverage (approx. 40 - 70 reads) than the mean for each dataset (approx. 30 reads, Figure 2.6). The assessment of average tag coverage after filtering showed that the majority of loci showing high coverage had been removed (Figure 2.6) in both datasets. Given that, in the reference guided analyses, reads which mapped to more than one genomic location were removed, it is perhaps surprising that such a large number of ohnolog loci still existed in this dataset. However, this can be explained by the presence of loci for which only one ohnolog pair is present in the *C. carpio* genome assembly. In such cases reads from both loci would align to the only ohnolog present in the genome draft, and not be filtered due to mapping to more than one location.

The removal of putative ohnolog loci is very important for the present study as their presence would inflate the levels of heterozygosity at a locus, which in itself could drastically affect downstream population genetics analysis. However, it should be noted that such data may prove invaluable for questions relating to the behaviour of ohnolog loci after genome duplication events. Thus, the population genetics filters used here could also be seen as a way of isolating informative loci for such applications.

Final parameter values and SNP dataset production

Based on the results of the tests and filtering steps described above, the following parameter values were chosen for use in the final RADseq data analyses in the subsequent chapters of this thesis (Table 2.2). For the *de novo* pure *C. carassius* dataset used in Chapters 3 and 6, a mismatch value of -M=2 was chosen, as the drop in RAD tag number for increased mismatch thresholds began to level off after this value (Figure 2.3). However, the choice of –M value here was rather arbitrary, as the reduction in tag number is still considerable up to –M = 4 and in some samples up to –M = 6. The more conservative value (-M=2) was decided upon in order to minimise the number of paralogs that would be incorrectly merged, which was known to be a strong possibility

in this dataset. The coverage threshold -m=8 was chosen so that SNP calling could be achieved with high confidence in all retained loci and on the basis that, at a required read depth of eight, the number of RAD tags lost was still low (Figure 2.3). A mismatch threshold of -N=2 was chosen for catalog building, as past this value, there was little increase in the proportion of shared tags in each pure *C. carassius* catalog (Supplementary Figure 2.1). For Populations module analyses, as one population in particular (SWE2) was known to have low DNA quality and indeed showed a lot of locus dropout (Supplementary Figure 2.2d) and the population from the Danube river catchment (HUN2) also showed some locus dropouts, -p=17 was used, allowing at most two of the 19 populations to dropout at each locus. For the -r parameter a final threshold value of 0.7 was used as, for the average population size of 9 samples in this study, this allowed 2 samples to dropout a single locus in most cases. Finally, for the -m parameter in the Populations module, a minimum read depth of -m=8 was used, however at this parameter is largely redundant as no loci will exist in the catalog that have a coverage lower than the values used in Ustacks and Cstacks.

Due to the stronger observed biases caused by allele dropout which were observed in the *de novo* analyses of the full dataset (Supplementary Figure 2.2, Supplementary Figure 2.4), the decision was made to use the reference guided analysis pipeline for the multispecies analyses in Chapter 5. Thus, in Pstacks and Cstacks the –g flag was used to construct loci based on the results of mapping location only. However, as null alleles were still a problem in this dataset, strict filtering options in the Populations module were used (-p = 28, -r 0.8 and –m = 8), which were shown to remove these loci in the tests of the present chapter (Figure 2.5). Putative ohnologs were then filtered on the basis of high heterozygosity and negative F_{IS}.

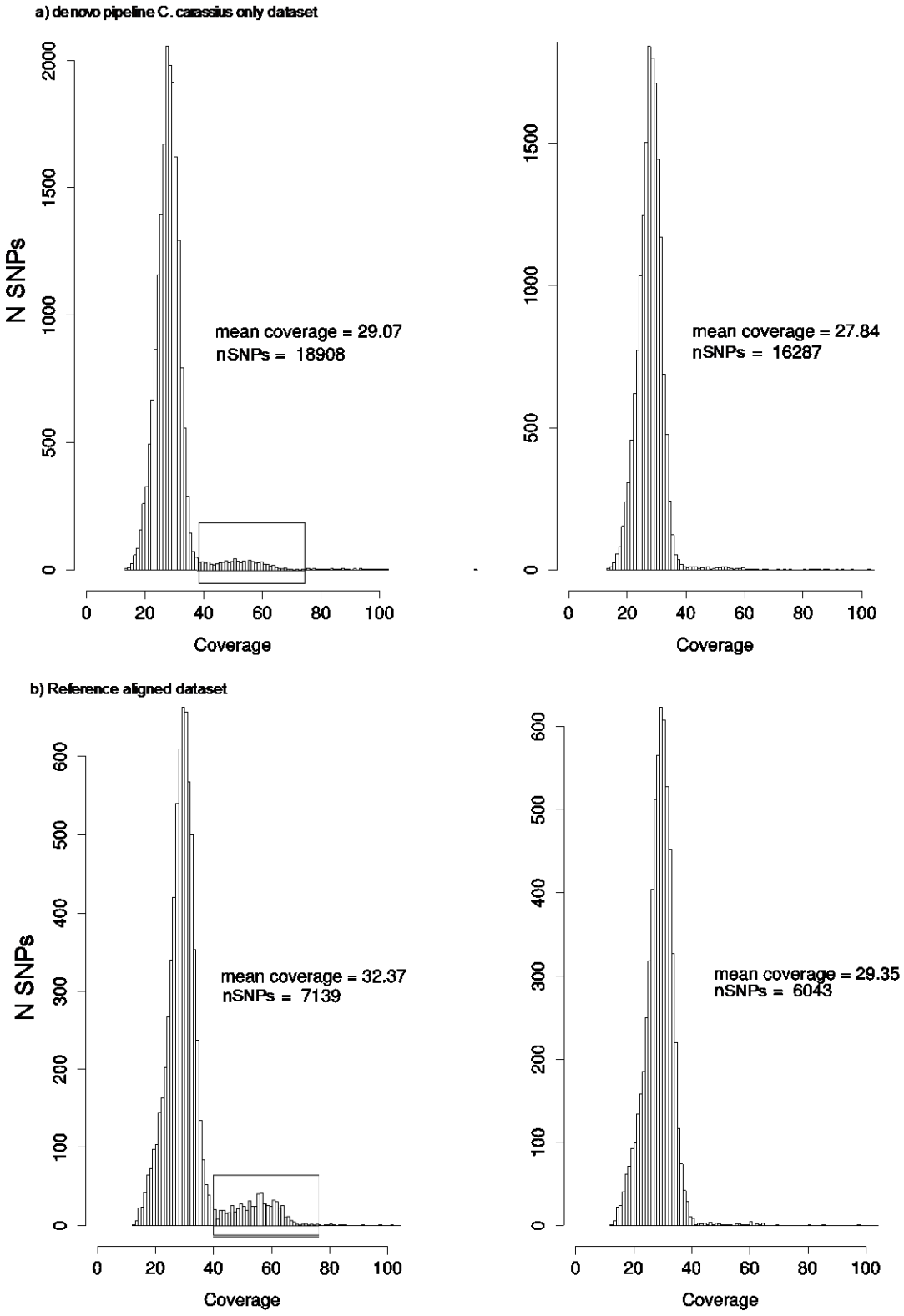


Figure 2.6. Coverage histogram for SNP loci in a) *de novo* and b) reference aligned full datasets before and after filtering for putative ohnolog loci. Boxed regions show large numbers of loci with approximately twice the average coverage, consistent with expectations for over merged ohnolog loci.

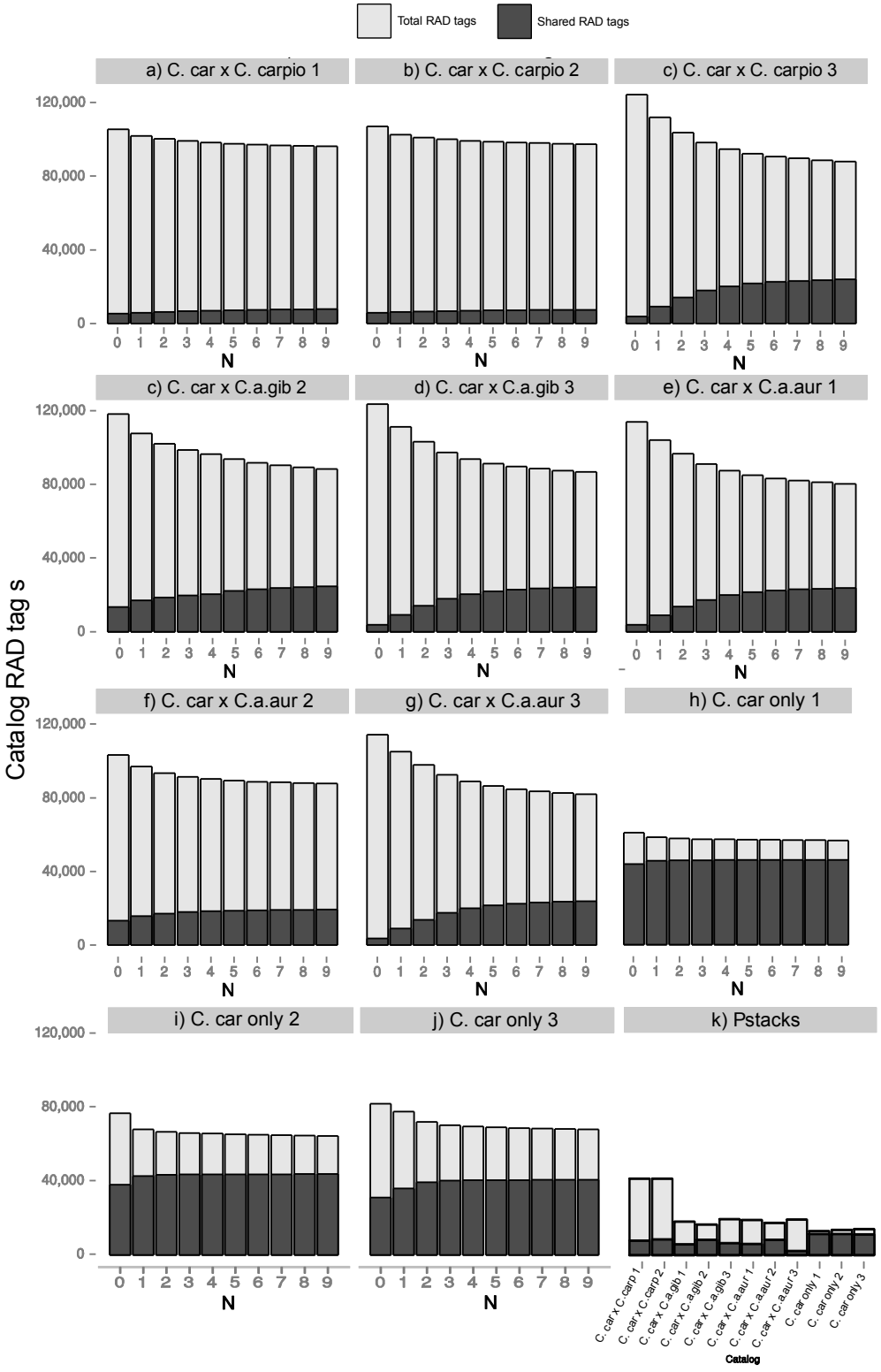
Conclusions

The parameter tests performed in the present chapter have allowed for the identification of two main potential sources of bias in the RADseq dataset used in this study. Firstly, from null alleles in hybrids, which appear to be prevalent throughout this dataset. Incidentally, many interspecific studies report large numbers of locus dropout between species (Hohenlohe et al. 2011; McCormack et al. 2012; Wagner et al. 2013; Wang et al. 2013), which could very well be attributable to null alleles. It is suggested that any study that uses interspecific datasets, or even those that contain high within species diversity, should carry out rigorous tests and filters for allele dropout using these approaches, in order to guard against the significant biases that it can introduce. The second major source of bias identified in the present chapter is the over merging of ohnolog loci, however in contrast to allele dropout, this is a problem that is specific to organisms with recent genome duplications. Reassuringly, however, the results of this chapter suggest that using rigorous parameter testing and population genetics filters, it is possible to identify and filter these loci. Importantly, these results and have allowed for optimal parameter values to be chosen, which reduce systematic biases in the data whilst maximising the number of SNPs for use in the evolutionary analyses in Chapters 3 and 5 of this thesis.

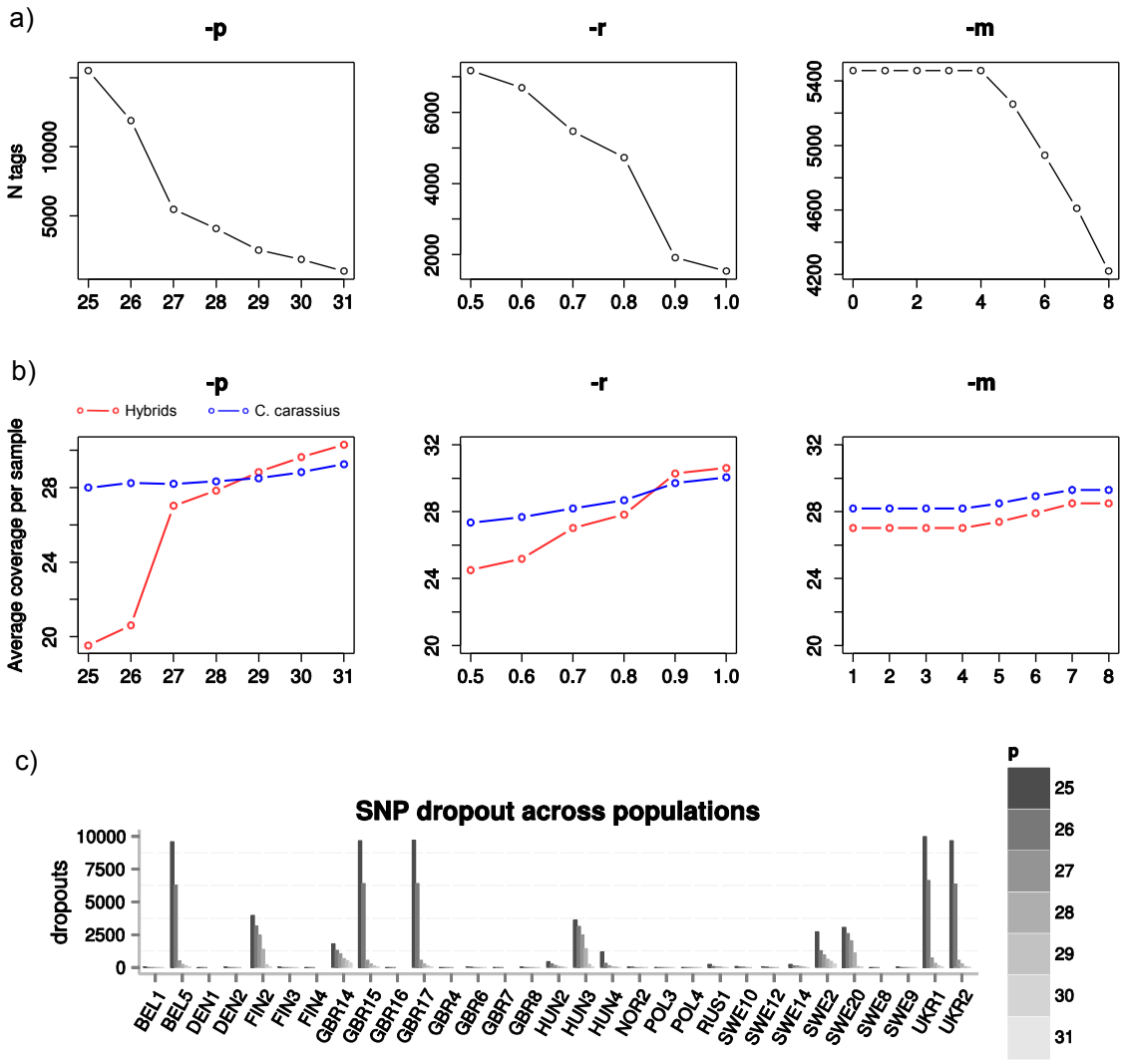
Chapter 2. Supplementary materials

Supplementary table 2.1. Details of raw RADseq read mapping to the Xu et al. (2014) *C. gibelio* draft genome. Back to text.

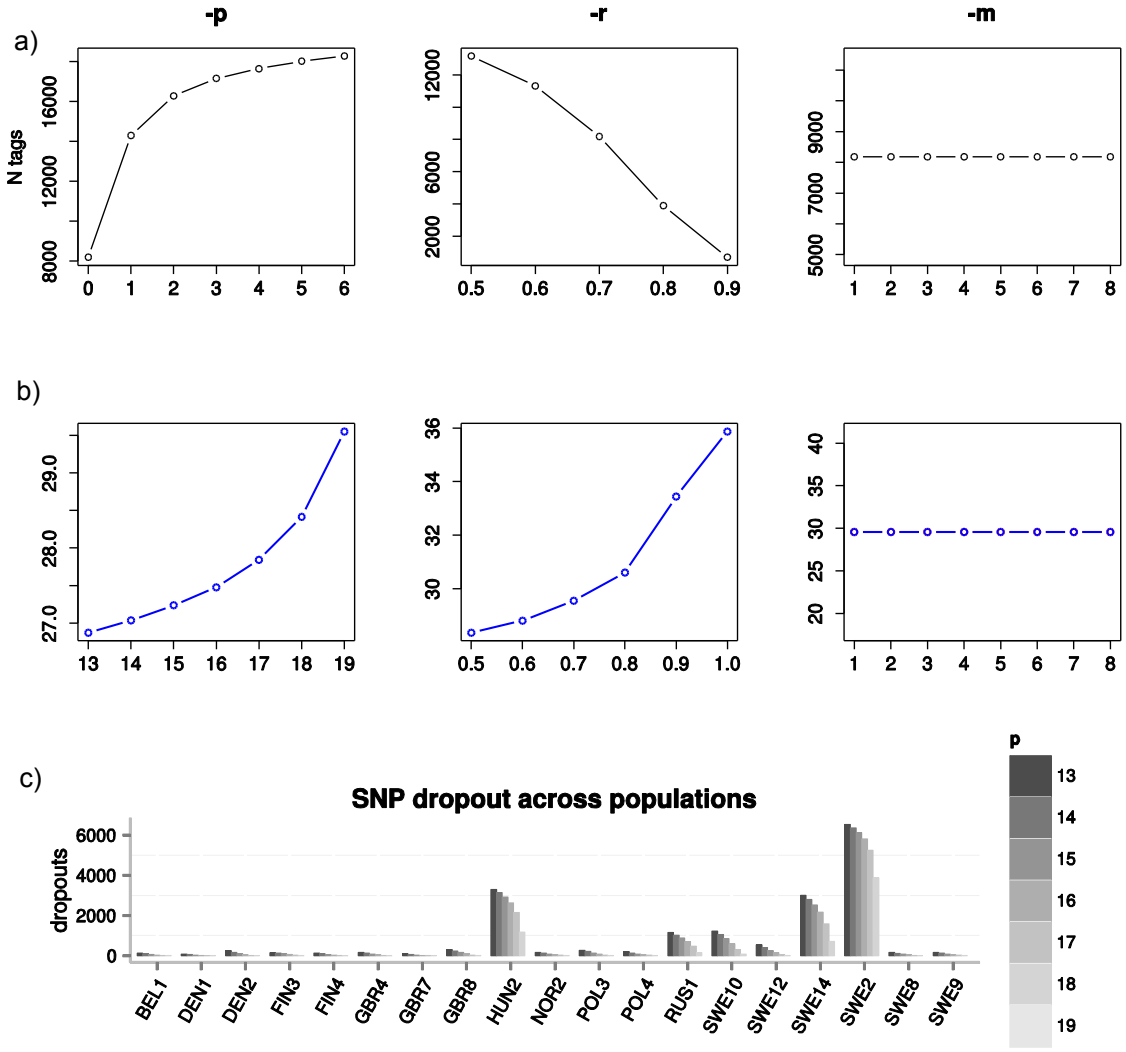
	Total % Mapped	**% Unique**	**% Multi-hits**	**% No-hits**
C. carassius	45.7	24.6	21.1	54.3
C. auratus	42.8	24.4	18.4	57.2
C. gibelio	43.9	25.3	18.6	56.1
C. carassius x C. auratus spp	42.9	24.5	18.4	57.1
C. gibelio	90.1	70	20.1	9.9
C. gibelio x Carassius hybrid	66.4	46.1	20.3	33.6



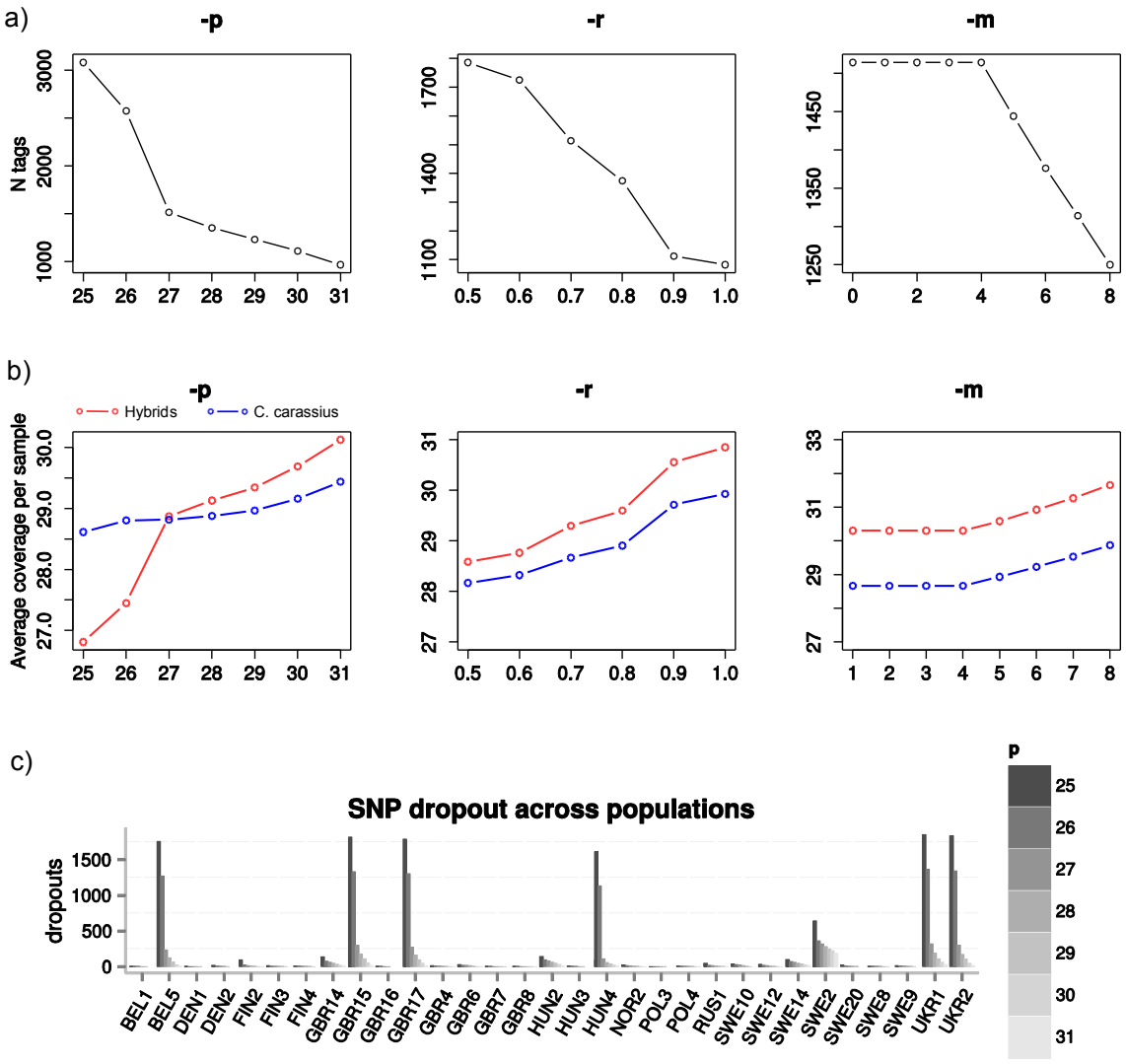
Supplementary Figure 2.1. a-j) Number of loci in each test catalog produced in *de novo* Cstacks parameter tests. k) Number of loci in each test catalog when constructed with the reference guided Stacks pipeline. Total numbers of loci in each Catalog are shown in light grey and loci shared between all three individuals in each catalog are shown in dark grey. Back to text.



Supplementary Figure 2.2. Results of parameter tests for the stacks module Populations in the *de novo* full dataset. 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. Back to text



Supplementary Figure 2.3. Results of the Populations module tests for the –p, -r and –m parameters the *de novo C. carassius* only dataset showing a) Number of SNP loci, b) average coverage per snp per sample and c) the number of loci which drop out in each populations for each test value of the –p parameter. Back to text



Supplementary Figure 2.4. Results of parameter tests for the stacks module Populations in the reference aligned full dataset. 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. Back to text

Chapter 3 Comparing RADseq and microsatellites to infer complex phylogeographic patterns, a real data informed perspective in the Crucian carp, *Carassius carassius,* L.

Authors: Daniel L Jeffries, Gordon H Copp, Lori Lawson Handley, K. Håkan Olsén, Carl D Sayer, Bernd Hänfling

Abstract

The conservation of threatened species must be underpinned by phylogeographic knowledge in order to be effective. This need is epitomised by the freshwater fish *Carassius carassius*, which has recently undergone drastic declines across much of its European range. Restriction Site Associated DNA sequencing (RADseq) is being increasingly used for such phylogeographic questions, however RADseq is expensive, and limitations on sample number must be weighed against the benefit of large numbers of markers. Such tradeoffs have predominantly been addressed using simulated data. Here we compare the results generated from microsatellites and RADseq to the phylogeography of *C. carassius*, to add real-data-informed perspectives to this important debate. 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. These lineages have been isolated for approximately 2.2-2.3 M years, and should arguably be considered as 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.

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 contiguous drainage systems but 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 in European fishes(Costedoat & Gilles 2009). The lack of obvious European pattern could be explained, at least in part, by the focus of phylogeographic studies on highly mobile, obligatory or facultative 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. carrassius* to genetic isolation and bottlenecks is compounded by small population sizes (Hänfling *et al.* 2005) and low dispersal (Holopainen *et al.*). 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.

Phylogeographic data have traditionally been collected using mitochondrial gene regions and/or nuclear markers such as 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) which provide several advantages (Morin *et al.* 2004). One disadvantage of this approach, however, is that bi-allelic SNP loci contain less information than the highly polymorphic microsatellites (Coates *et al.* 2009). Large numbers of SNPs are consequently 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 Restriction Site-Associated DNA sequencing (RADseq, (Miller *et al.* 2006)), which enables thousands of orthologous SNP markers to be quickly isolated from non-model organisms. Despite this new opportunity, microsatellites may still be still more informative and/or cost effective in many cases, allowing for wider geographic coverage and sampling of more individuals per population. A comparison of the utility of RADseq-derived SNPs and microsatellites for phylogeographic studies is needed and will contribute to the important debate on whether it is more advantageous to genotype small numbers of highly polymorphic markers in a large number of samples, or tens of thousands of SNP markers in fewer samples. This trade-off has recently been highlighted as among the most important questions in landscape genetics (Epperson *et al.* 2010; Balkenhol & Landguth 2011).

The optimal phylogeographic study design depends heavily on the properties of the study system; in particular the strength of population structure (i.e. F_{ST}). In *C. carassius* we expect population structure to be strong and driven by isolation by distance (IBD). If this is so, then patchy geographic sampling along the IBD gradient could result in falsely identified distinct lineages (Schwartz & McKelvey 2009). We would, therefore, expect the number of populations sampled and their geographic uniformity to be more important than number of loci, or number of samples per population in this study.

In the present study, we use a combination of mitochondrial DNA (mtDNA), microsatellites and genome-wide single nucleotide polymorphisms (SNPs) obtained from RADseq in order to: 1) produce a range wide phylogeography for *C. carassius* as a basis for Europe-wide conservation strategies, 2) test competing scenarios 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 add perspectives from real biological data to a topic that has almost exclusively been addressed using simulated datasets (but see (Coates *et al.* 2009; Hess *et al.* 2011). Specifically we 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.

Methods

Sample collection and DNA extraction

We collected 848 *C. carassius* tissue samples from 49 populations across the species' distribution in central and northern Europe (Table 3.1, Figure 3.1). Sample sizes ranged from n=4 to n=37, with a mean of n=17 (Table 3.1). 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 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°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.

Table 3.1. Location, number, genetic marker sampled, and accession numbers of all samples and sequences used in the present study for microsatellite and mitochondrial DNA and RADseq analyses.

Code	Location	Country	Drainage	Coordinates		Microsatellites (n)	mtDNA (n)	RAD-seq (n)
				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		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*	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	
SWE1	Grānbrydammen	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		
SWE6	Åland Island	Sweden	Baltic Sea	60.36	19.85	8	3	
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	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	Copenhagan	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	Sweden	Baltic Sea	55.73	12.34	14		9
SWE14	Stockholm	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	
HUN1	Gödöllő	Hungary	Danube River	47.61	19.36		2	
HUN2	Vörösmocsár	Hungary	Danube River	46.49	19.17			6
Total						**848**	**79**	**154**

(Table 3.1 continued)

Genbank mtDNA sequences

Code	Reference	Country	Drainage	Accession
GER6	Kalous et al. (2007)	Germany	Baltic sea	DQ399917
GER6	Kalous et al. (2007)	Germany	Baltic sea	DQ399918
GER6	Kalous et al. (2007)	Germany	Baltic sea	DQ399919
GER7	Rylková et al. (2013)	Germany	Hunte River	JN412540
GER7	Rylková et al. (2013)	Germany	Hunte River	JN412541
GER7	Rylková et al. (2013)	Germany	Hunte River	JN412542
GER7	Rylková et al. (2013)	Germany	Hunte River	JN412543
GER8*	Rylková et al. (2013)	Germany	Lahn River	JN412537
GER8*	Rylková et al. (2013)	Germany	Lahn River	JN412538
CZE2	Rylková et al. (2013)	Czech Republic	Elbe drainage	GU991399
AUS1	Rylková et al. (2013)	Austria	Danube river	JN412533
AUS1	Rylková et al. (2013)	Austria	Danube river	JN412534
AUS2	Rylková et al. (2013)	Austria	Danube river	JN412535
AUS3	Rylková et al. (2013)	Austria	Danube river	JN412536
GBR12	Rylková et al. (2013)	U.K.	U.K	JN412539
GBR12	Kalous et al. (2012)	U.K.	U.K	GU991400
SWE15	Rylková et al. (2013)	Sweden	Baltic sea	JN412545
SWE16	Rylková et al. (2013)	Sweden	Baltic sea	JN412544

† Also present
* Location on Map (Figure 3.1.a) is approximate

Molecular markers and methods

Three types of molecular markers were used in the 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.

Mitochondrial DNA amplification

A total of 82 *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 3.1). 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 3.1) as a reference and ambiguous nucleotides were manually edited using CodonCode aligner v.2.0.6 (CodonCode Corporation).

Microsatellite amplification

All 848 *C. carassius* samples were genotyped at 13 microsatellite loci, which were originally designed for use in *Carassius auratus*, or *Cyprinus carpio* and cross amplify in *C. carassius* (Supplementary table 3.1). Six of these loci were chosen for their species diagnostic properties, allowing us to ensure that all samples used in the present study were *C. carassius* and not one of the closely-related introduced species (*C. carpio*, *C. auratus*, or *Carassius gibelio*) or their hybrids (see Supplementary text for full details of species identification and hybrid detection). 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 Supplementary table 3.1 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 using a 400 bp size standard and microsatellite alleles scored using the Beckman Coulter CEQ8000 software.

RADseq

A total of 149 individuals (16 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 3.1). 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 to a read length of 100bp using paired end sequencing across five lanes of two Illumina HiSeq 2000 flowcells (Edinburgh Genomics).

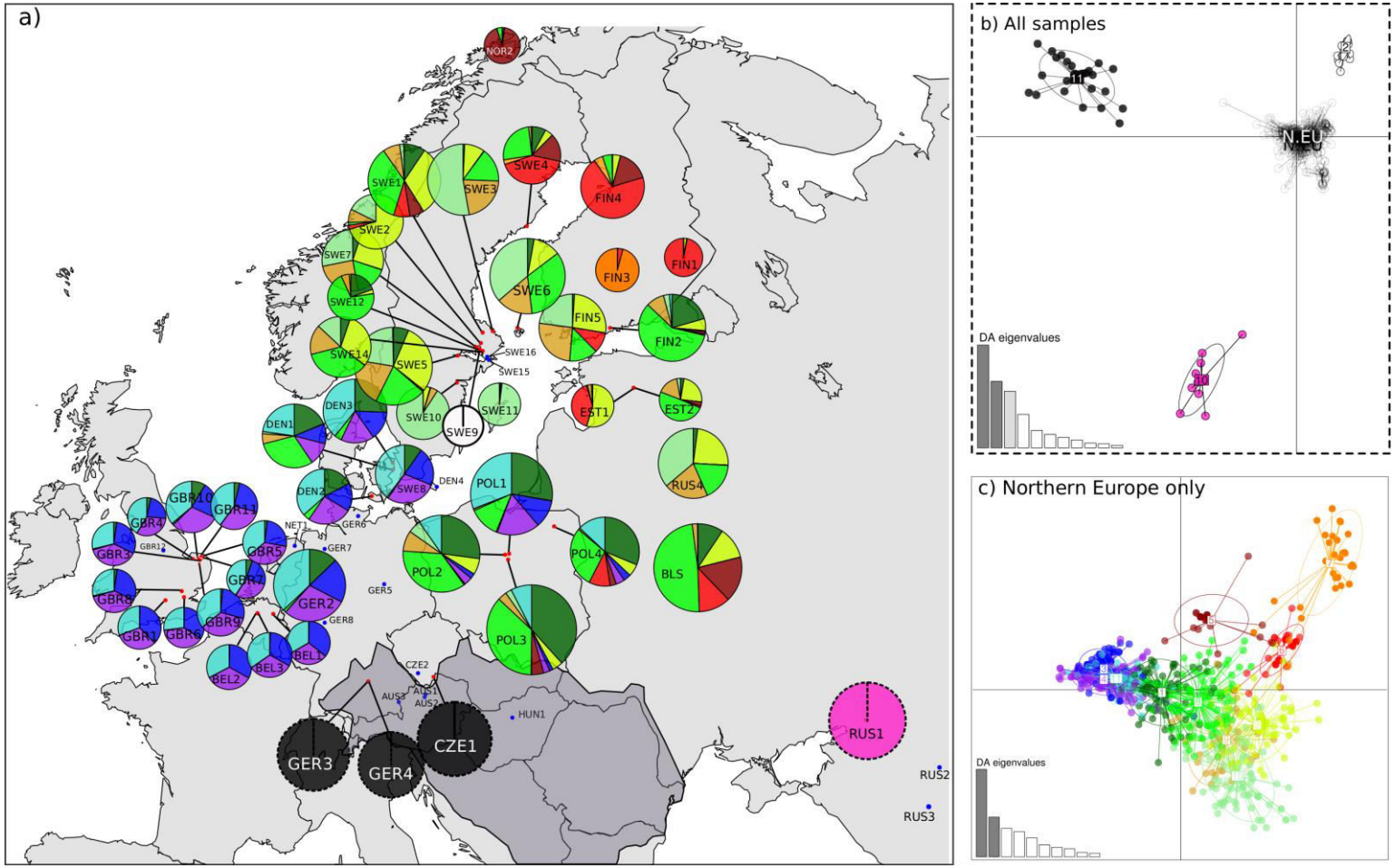


Figure 3.1. Population structure of *C. carassius* in Europe. a) Sampling locations (sites sampled with nuclear and mtDNA markers = red dots, mtDNA only = blue dots) and population cluster memberships from microsatellite DAPC analysis. Pie chart size corresponds to microsatellite allelic richness. Pie 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.

Data analyses

Phylogenetic analysis of mtDNA

In addition to the 82 sequenced samples, we retrieved 18 published *C. carassius cyt*b sequences from GenBank which were validated through cross checking with their original publications (Table 3.1). 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 were exported to BEAST v.1.7.5 (Drummond *et al.* 2012) for phylogenetic analyses in order to identify the major phylogenetic lineages within European *C. carassius*. 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. Initial analyses using the GTR (Tavaré 1986) substitution model yielded multiple parameters with low estimated summary statistic (ESS) values (<100), therefore the less complex HKY (Hasegawa *et al.* 1985) substitution model, which had ESS values >200 for all parameters, was used. We used a 'coalescent: constant size' tree prior, which assumes an unknown but constant population size backwards in time, as recommended for intraspecific phylogenies (*BEAST Tutorial - Tree priors and dating*). 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.

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 2001) 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 =5) 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 Supplementary text and Jombart *et al.* 2010 for more details). 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 was, we used Bayesian Information Criteria (BIC) scores, 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.

RADseq data filtering and population structure analysis

The quality of the RADseq raw read data were first examined using FastQC (Andrews 2010). The data were then demultiplexed using the "process_radtags" script distributed with Stacks (Catchen *et al.* 2013). Raw reads were then clustered into loci within and between individuals and SNPs were called using Stacks. However, this process is heavily parameterised, therefore it was necessary to perform extensive parameter tests before final data processing and SNP calling was carried out. Full details of these tests and the justification of the final parameter values chosen for this dataset can be found in Chapter 2 of this thesis. Briefly, the final parameter values for the respective Stacks

module were as follows; Ustacks: M=2, m=8, removal and deleveraging algorithms were also used; Cstacks: N=2; Populations: 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. Finally, we filtered out loci which had a heterozygosity of > 0.5 and $F_{IS} < 0.0$ 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 *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 adegenet R package 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 (Cornuet *et al.* 2014) was used to reconstruct the most likely *C. carassius* recolonisation routes through Europe after the last glacial maximum. Analyses were performed on1000 randomly-selected SNP loci from the full RAD-seq dataset were used, as microsatellite loci are likely to be affected by homoplasy over the time scales used here (Morin *et al.* 2004). The reduced dataset was first analysed with DAPC to confirm that it produced the same structure as the full dataset. Then datasets of expected summary statistics were simulated for a number of scenarios (i.e. a specific population tree topology, together with the parameter prior distributions that are associated with it). These simulated datasets represent the theoretical expectation under each scenario, and are compared to the same summary statistics calculated from the observed data to identify the most likely of the specified 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).

To reduce the number and complexity of possible scenarios, we split DIYABC analysis into three stages (Table 3.2). In stage 1, we tested 11 broad scale scenarios (Scenarios 1

-11, Supplementary Figure 3.1), in which populations were grouped into three pools; 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). Both population pooling and scenarios were chosen on the basis of the broad phylogeographic structure identified in the mtDNA and RAD-seq population structure analysis (see Results). We tested the likelihood of these 11 scenarios, simulating one million summary-statistic datasets per scenario, for comparison to the real dataset.

Table 3.2. Population pools, parameter priors used and 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** – RUS1 **Pool 3** – HUN2	1 – 11	N1 = 10E+03 - 500E+03 Nb1 = 10 - 100E+03 N2 = 100 - 100E+03 N3 = 100 - 200E+03 t1 = 1E+03 - 1E+06 gens t2 = 1E+03 - 3E+06 gens ra = 0.001-0.999 rb = 0.001-0.999 rc = 0.001-0.999 db = 10- 10E+03 gens	9	N1 = Nb1 = N2 = N3 = t1 = db = t2 =	3.47E+04 2.37E+04 7.49E+04 1.40E+05 1.35E+05 4.46E+03 1.09E+06
2	**Pool 1** – GBR4, GBR7, GBR8 **Pool 2** – DEN1, DEN2, DEN3 **Pool 3** – FIN3, FIN4	12 – 16	N1 = 10-4E+03 N2 = 10 - 10E+03 N3 = 10 - 20E+03 N4 = 10 - 50E+03 N5 = 10 - 20E+03 N6 =10 - 400 t1 = 100- 10E+03 gens t1a = 100- 10E+03 gens t2 =100- 10E+03 t2a =100- 5E+03 gens t2b = 500-20E+03 gens t2c = 100 - 10E+03 gens t2d = 100 - 10E+03 gens t3 = 500 - 20E+03 gens t3c =100 - 10E+03 gens t3d =100 - 10E+03 gens t4 =500 - 20E+03 gens ra = 0.001-0.999 rb = 0.001-0.999	14	N1 = N2 = N3 = N4 = N5 = N6 = t1 = t1a = t2d = t3d = t4 = rb =	3.67E+03 7.52E+03 1.74E+04 1.94E+04 1.18E+04 2.10E+02 6.79E+03 2.51E+03 6.78E+03 8.91E+03 1.20E+04 6.68E-01
3	**Pool 4** – POL3, POL4 **Pool 5** – SWE2, SWE8, SWE9, SWE10, SWE12, SWE14 **Pool 6** – NOR2	14a- 14f	N1 = 10-4E+03 Nb1 = 10-10E+03 N2 = 10 - 10E+03 N3 = 10 - 20E+03 Nb3 = 10-10E+03 N4 = 10 - 50E+03 N5 = 10 - 20E+03 N6 =10 - 400 Nb6 =10-10E+03 t1 = 100- 10E+03 gens t1a = 100- 10E+03 gens t2d = 100 - 10E+03 gens t3d = 100 - 10E+03 gens t4 = 500 - 20E+03 gens rb = 0.001-0.999 da = 10 - 10E+03 gens db = 10 - 10E+03 gens dc = 10 - 10E+03 gens dd = 10 - 10E+03 gens de = 10 - 10E+03 gens	14d	N1 = Nb1 = N2 = N3 = N4 = N5 = N6 = t1 = t1a = t2d = t3d = t4 = rb = dc =	2.39E+03 9.35E+02 8.14E+03 9.36E+03 1.70E+04 1.10E+04 1.38E+02 3.75E+03 2.46E+03 5.90E+03 7.97E+03 1.68E+04 6.19E-01 9.07E+03

In the second and third stages, we performed a finer scale analysis, focussing on the 17 northern European populations alone. Populations were again pooled on the basis of both population structure and geography, in order to reduce scenario complexity (Table 3.2). In stage 2 we tested five scenarios (Scenarios 12-16. Supplementary Figure 3.2a), 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 (Supplementary Figure 3.2b) 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 from the scenario identified as most likely in stages one and three to estimate times of the major lineage splits in European *C. carassius*.

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; Table 3.3). 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.

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 (Figure 3.2, Supplementary table 3.2). 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 a few exceptions to this clear geographical split however; 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 2.26 MYA (Figure 3.2).

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).

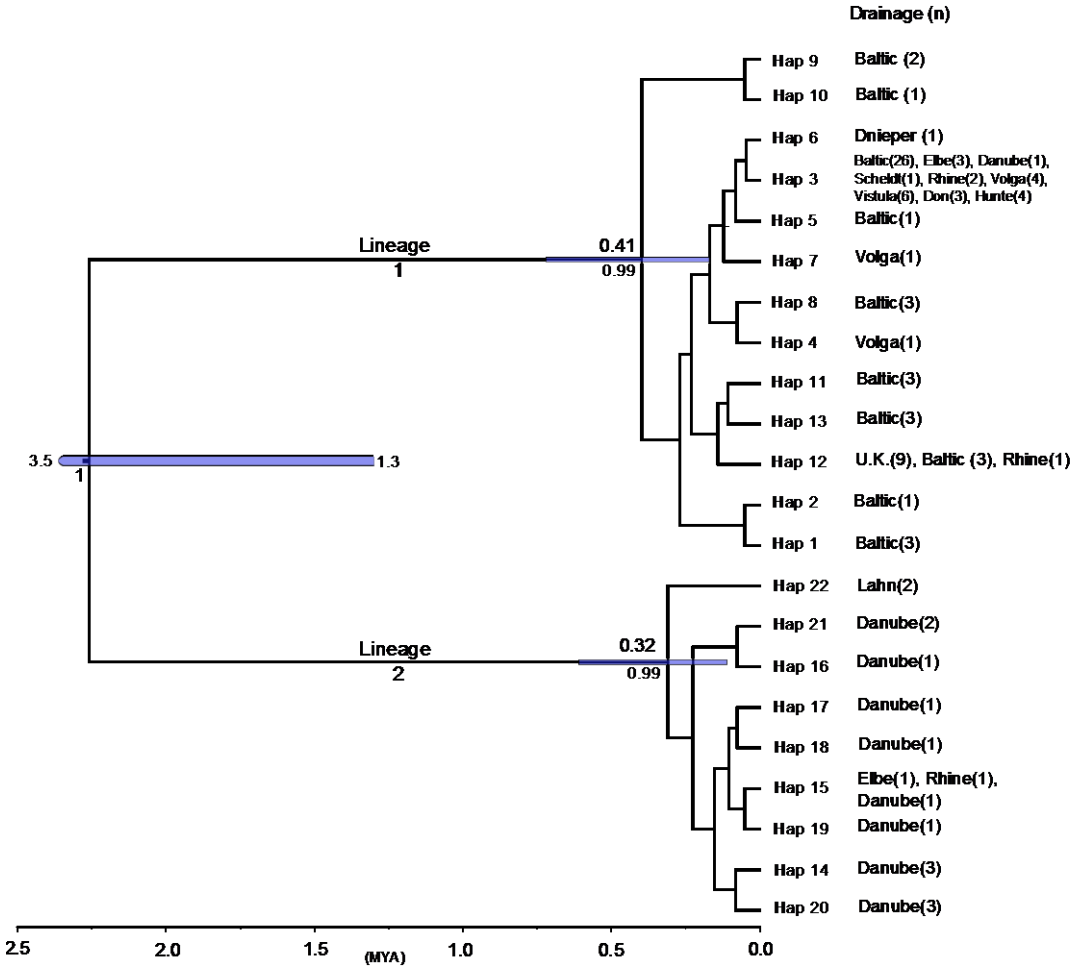


Figure 3.2. Maximum credibility tree calculated in BEAST for 100 *C. carassius* cytb sequences. For the three maximally supported nodes, age is given above and the posterior probability distribution is given below, with 95% CI's represented by blue bars.

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. Finally 5719 of these SNP loci were filtered out due to high (> 0.5) heterozygosity in at least one population. In doing so, we removed many high coverage tags (Supplementary Figure 2.3), 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 (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, Supplementary Figure 3.4b) consistent with decreasing diversity along colonisation routes. However, A_r did not decrease with increasing latitude (Adj R^2 =-0.007, P = 0.414, Supplementary Figure 3.4a). 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, Supplementary Figure 3.4c), but the relationship between A_r and longitude was strengthened (adj. R^2 = 0.316, P < 0.001, Supplementary Figure 3.4d).

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, Supplementary table 3.3). Pairwise F_{ST} calculated from the RADseq dataset also showed strong structure (Supplementary table 3.4), 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, using all 49 populations, indicated that between 11 and 19 genetic clusters (Supplementary Figure 3.5a) 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 (Supplementary Figure 3.5a). This initial analysis showed that populations belonging to

Cluster 10 (RUS1, Don river catchment) and Cluster 11 (GER3, GER4, CZE1, Danubian catchment) were highly distinct from clusters found in northern Europe (Figure 3.1b). Since the marked genetic differentiation between these three main clusters masked the more subtle population structure among northern European populations (see Figure 3.1b), we repeated the DAPC analysis without the populations from the Danube and Don (RUS1, GER3, GER4, CZE1, Figure 3.1b). The results of this second DAPC analysis revealed an IBD pattern of population structure, across Europe (Figure 3.1). Mantel tests excluding the Danubian and Don populations corroborated these results; showing significant correlation with geographic distance in northern Europe (adjusted $R^2 = 0.287$, $P < 0.001$, Supplementary Figure 3.6a), with Danubian populations shown to be more diverged than their geography would predict (data not shown).

In the RADseq DAPC analysis, BIC scores suggested between four and ten genetic clusters, similar to the range inferred in the microsatellite data, and we therefore chose four clusters to take forward in the analysis (Supplementary Figure 3.5b). Following spline interpolation, we retained six principal components and kept two of the linear discriminants from the subsequent discriminant analysis (Supplementary Figure 3.5b). 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 $R^2 = 0.722$, $P < 0.001$; Supplementary Figure 3.6b).

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 (Supplementary Figure 3.7). 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 (Supplementary Figure 3.8). Model checking showed that the observed summary statistics for our data fell well within those

of the posterior parameter distributions for scenario 9 (Supplementary Figure 3.8c). 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 (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 and reduced N_e by 32%.

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 (Supplementary Figure 3.9a), but not for Scenario 13 (not shown). Therefore, Scenario 14 was carried forward into stage three in which we tested six more scenarios (Supplementary Figure 3.2b) 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 (Supplementary Figure 3.9b), we therefore accepted this as the scenario for the colonisation of *C. carassius* in northern Europe (Supplementary Figure 3.9b). This scenario infers an initial split between two sub-lineages in northern Europe approximately 33 600 YBP (Figure 3.4), 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, Supplementary Figure 3.9b).

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 ($R^2 = 0.722$, $P < 0.001$) (Supplementary Figure 3.6) compared to the M1 dataset ($R^2 = 0.287$, $P < 0.001$) (excluding Danubian populations and SWE9 from both datasets, Supplementary Figure 3.6). 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 (Figure 3.3).

Table 3.3. Summary statistics for M1, M2, M3 and RADseq datasets. RAD contains all RAD-seq data, M1 contains all microsatellite data, M2 contains only microsatellite for the individuals used in the RAD-seq, and M3 contains all microsatellite data for all individuals that were available in populations that were used in RAD-seq.

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	9.125 ± 0.8	13	24.4 ± 7.3	7.8.4 ± 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

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, (Supplementary Figure 3.10), 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, Supplementary Figure 3.10), 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 cluster number, structure and population identity of clusters generally agreeing well (Figure 3.1, Figure 3.3c).

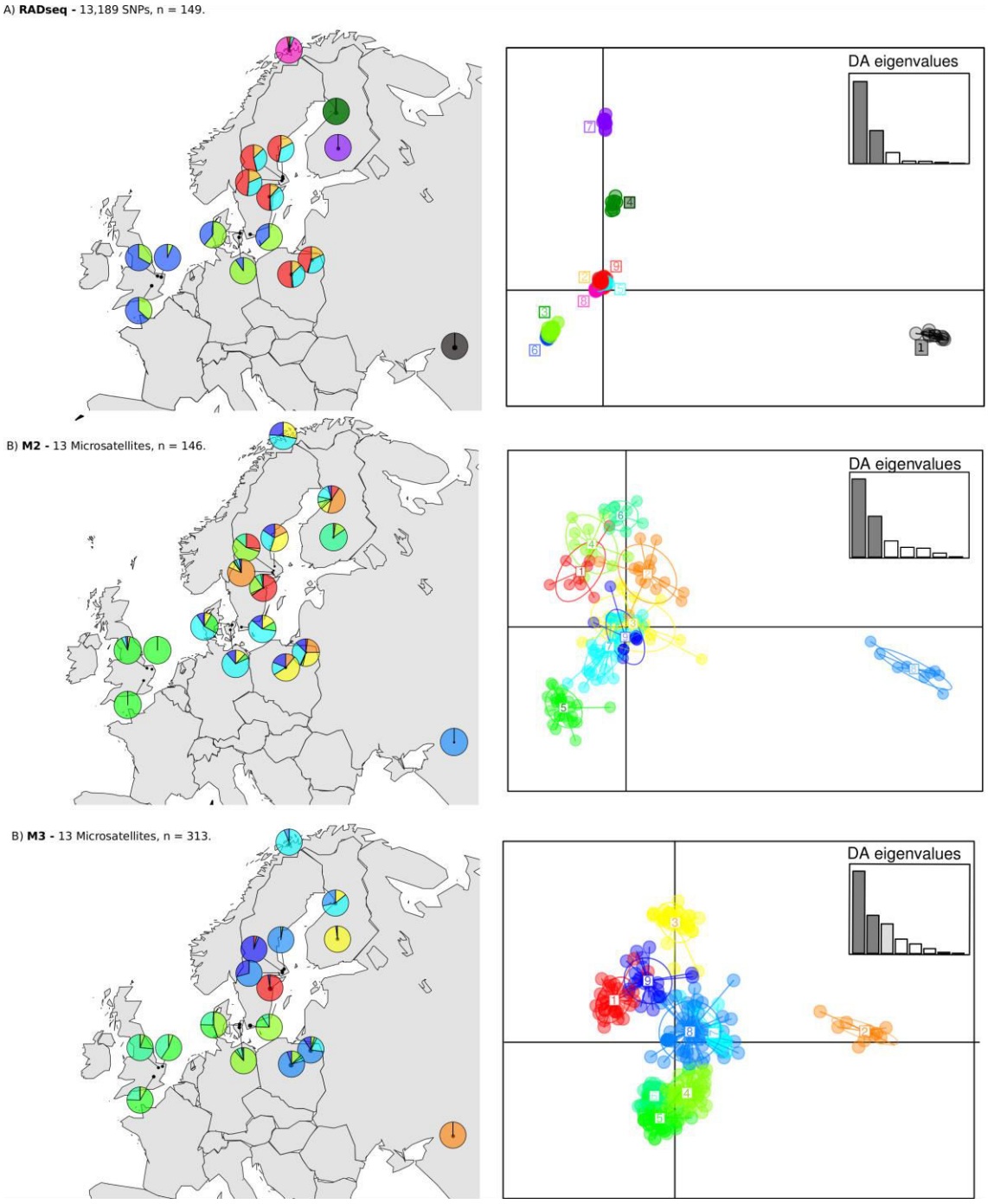


Figure 3.3. Comparison of DAPC results using RADseq dataset a), M2 dataset b) and M2 dataset c). Colours correspond between DAPC scatter plots and maps within but not between panels.

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 (mean = 0.21) and 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, Supplementary Figure 3.6) 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.

Table 3.4. Pearson's product-moment correlation coefficients and Pared t-tests comparing Heterozygosities and FSTs 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**

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

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, 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, Supplementary Figure 3.6).

Discussion

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 long-standing 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 2.3 MYA (95% CI = 1.30–3.22) and 2.2 (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, 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 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 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. The sample site from the River Lužnice is located in very close proximity 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 (Figure 3.4). The Ponto-Caspian region, and in particular the Black Sea basin, was an important refugium for freshwater fishes during the Pleistocene glacials, 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 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, Figure 3.4), 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 Figure 3.4). The western sub-lineage underwent a second long period of population decline (Bottleneck B, Figure 3.4), 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 Figure 3.4), 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 northeastern Europe) may give false signals of admixture between intermediate populations, such as those in Poland.

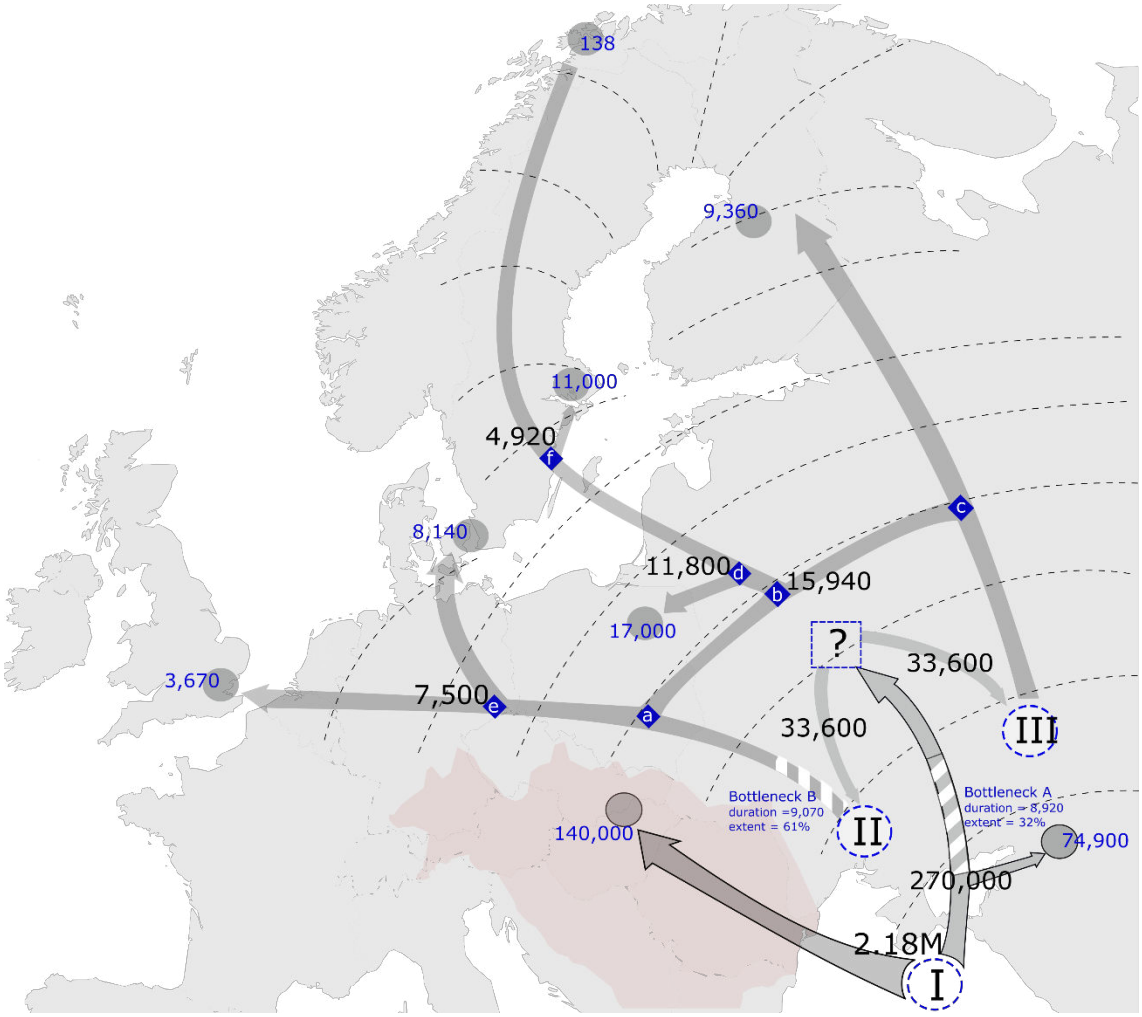


Figure 3.4. The postglacial recolonisation of *C. carassius* in Europe. Arrows represent the relationships between population pools used in DIYABC (grey circles) as inferred from Stage 1, scenario 9 (arrows outlined in black) and Stage 3, scenario 14d (arrows with no outline) analyses on RADseq data. 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.

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 *Thymallus thymallus(Koskinen et al. 2000)*, *Cobitis taenia*,(Culling *et al.* 2006), *Cottus 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 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 2014) 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).

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. 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.

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 RADseq would present the better option for the phylogeography of *C. carassius,* with the huge number of SNP loci overcoming the limitations imposed by reduced sample number.

Chapter 3. Supplementary materials

Detecting hybrids

Methods

In total we acquired tissue samples of 1078 Fish during sampling for this study. All of which were first genotyped using multiplex 1 (Supplementary table 3.1) 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 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, Supplementary table 3.1).

Results

Of the 1087 genotyped fish from 58 populations, 942 individuals across 55 populations (86.7%) were identified as pure crucian 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*, 848 in 49 populations existed in sites where hybrids or non-native species were not detected by microsatellite genotyping. To safeguard against cryptic introgression which may produce erroneous results only these 848 pure *C. carassius*

were used for the main phylogeographic analyses and tests of the status of *C. carassius* in England.

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 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 (Supplementary Figure 3.5) 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 (Supplementary Figure 3.5c) and we kept 2 linear discriminants for plotting (Figure 3.1b).

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 diveregence 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, Figure 3.1b) 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

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 (Supplementary Figure 3.10).

Additional discussion

Population structure in northwest Europe

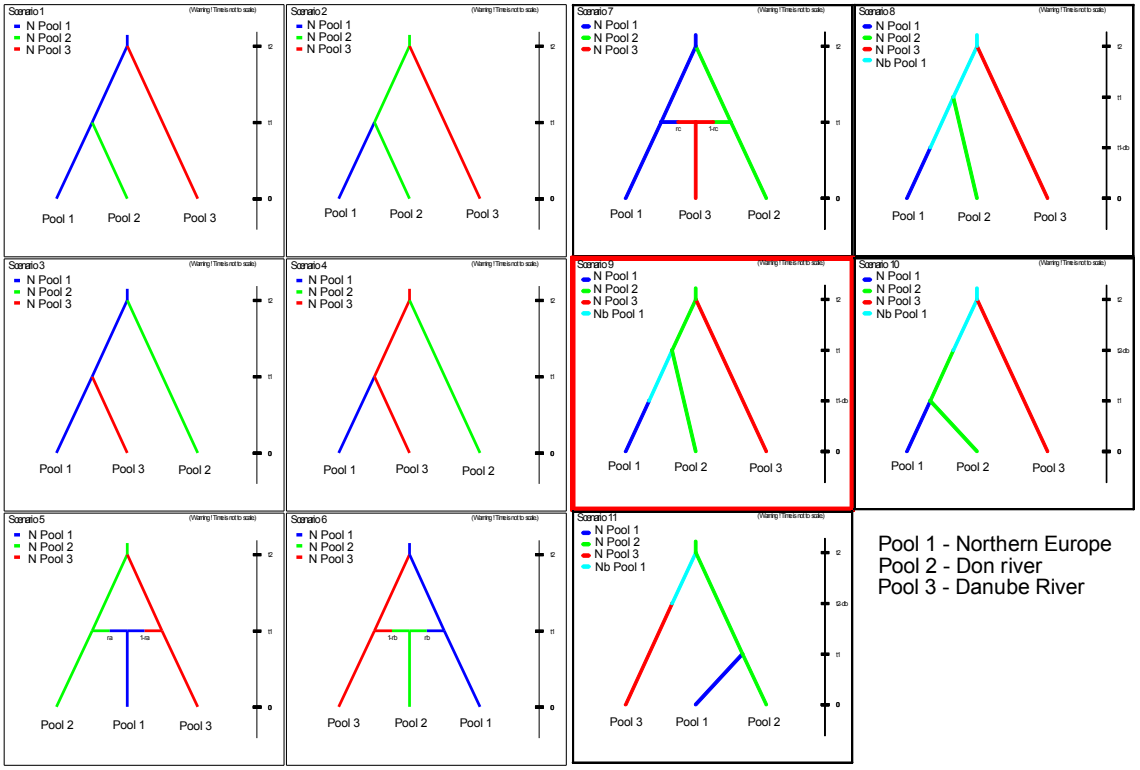
An intriguing result lies in the genetic similarity between populations in England with those in Belgium and Germany. *C. carassius* has been designated as native to England, however this status has been contentious in the past (Maitland 1972). Under the

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.

Supplementary table 3.1. 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



Supplementary Figure 3.1. DIYABC scenarios used in broad-scale analysis (Stage 1). See text for population poolings. See Table 3.3 for population poolings and prior parameter values. Back to text.

Supplementary table 3.2. Haplotype memberships for 101 Cytochrome B sequences used in Figure 3.2. Back to text.

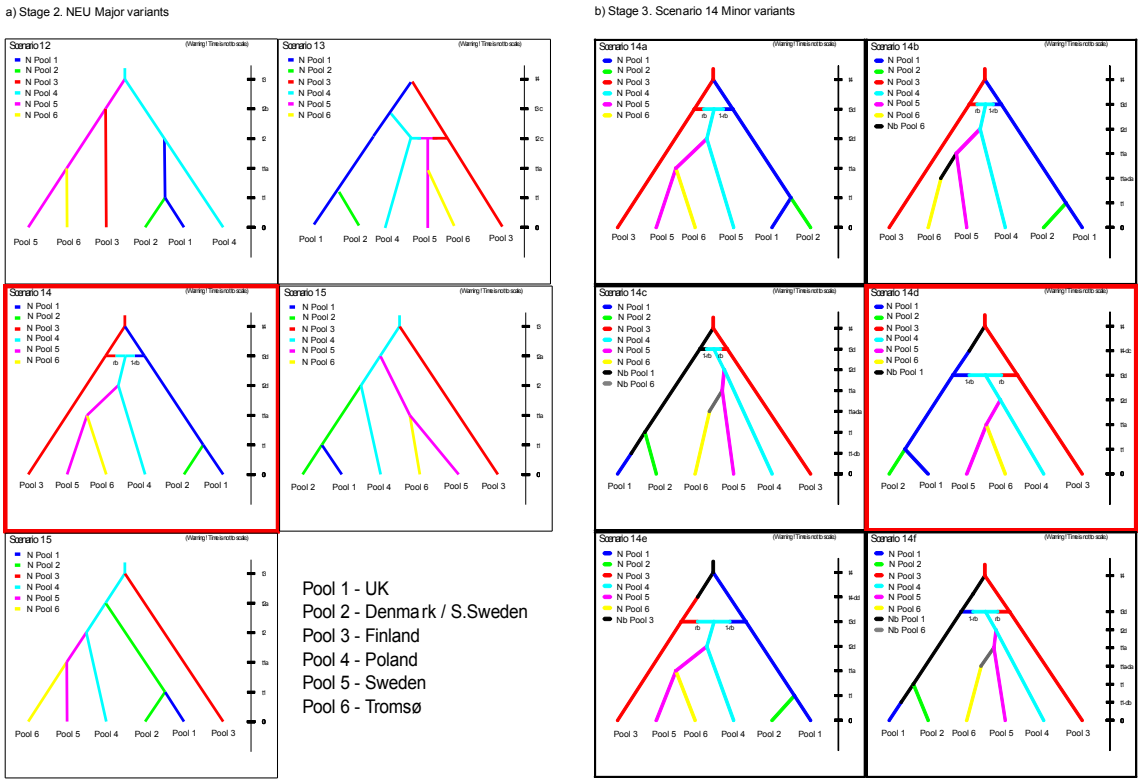
	Haplotype	N	Drainage (n populations)	Sample sequence
Lineage 1	Hap 1	3	Baltic	FIN5 1-3
	Hap 2	1	Baltic	EST1 2
	Hap 3	49	Elbe(2), Baltic(24), Scheldt(1), Rhine(2), North sea(2), Vistula(6), Volga(4), Don(3), Danube(1), Hunte(4)	GER1 1,3, EST1 1, 3, SWE6 1 -3, BEL1 3 , GER2 2, 3, GER4 2, NOR 1, 2, SWE11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

Supplementary table 3.3. Pairwise FST values calculated using the M1 dataset. Back to text.

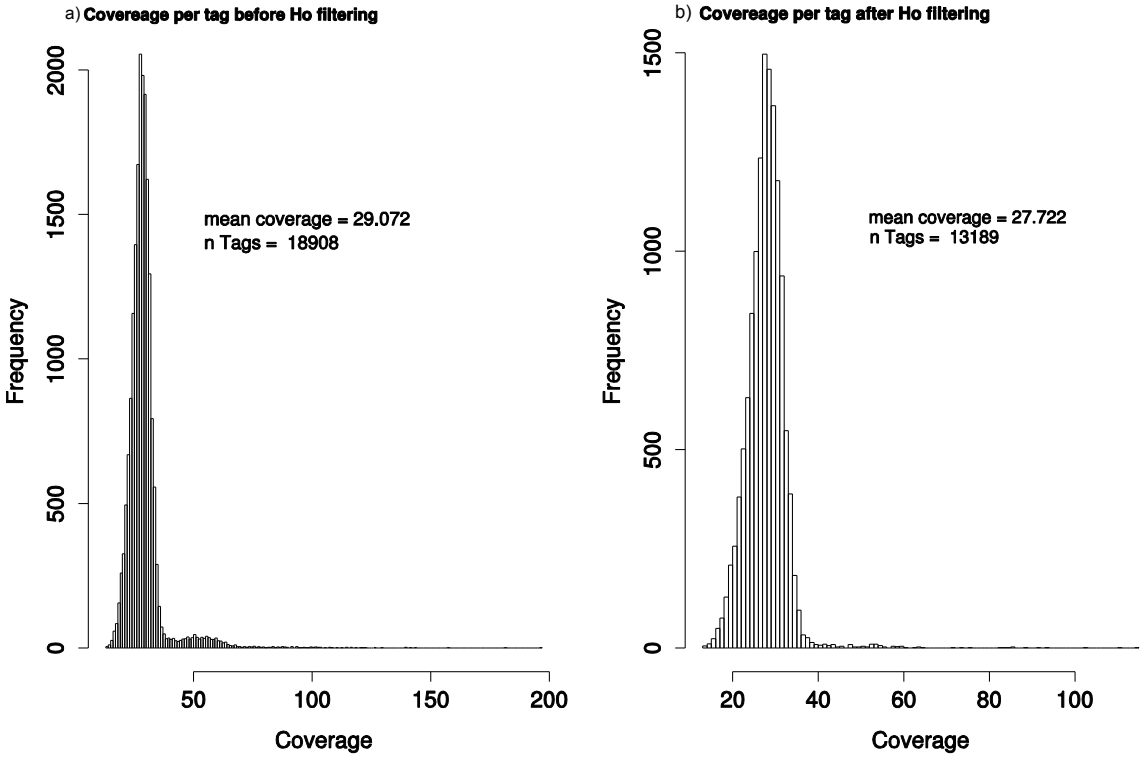
	GBR1	GBR2	GBR4	BEL1	BEL2	BEL3	FIN1	RUS4*	FIN2	CZE1	GER2	GER3	GER4	POL1	POL2	POL3	POL4	GBR7	GBR3	GBR8	GBR9	GBR11	GBR5	GBR6	GBR10	SWE4	SWE3	SWE5	SWE6	SWE7	SWE2	SWE1	SWE9	SWE10	SWE11	SWE8	FIN5	FIN3	FIN4	EST1	EST2	BLS	RUS1	DEN1	SWE12	DEN2	NOR2	SWE14	DEN3
GBR1		0.307	0.531	0.312	0.198	0.346	0.785	0.472	0.407	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	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		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	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		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
SWE6	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		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	NA	*	*	*	NS	*	*	*	

Supplementary table 3.4. Pairwise FST values calculated using the RADseq dataset. Back to text.

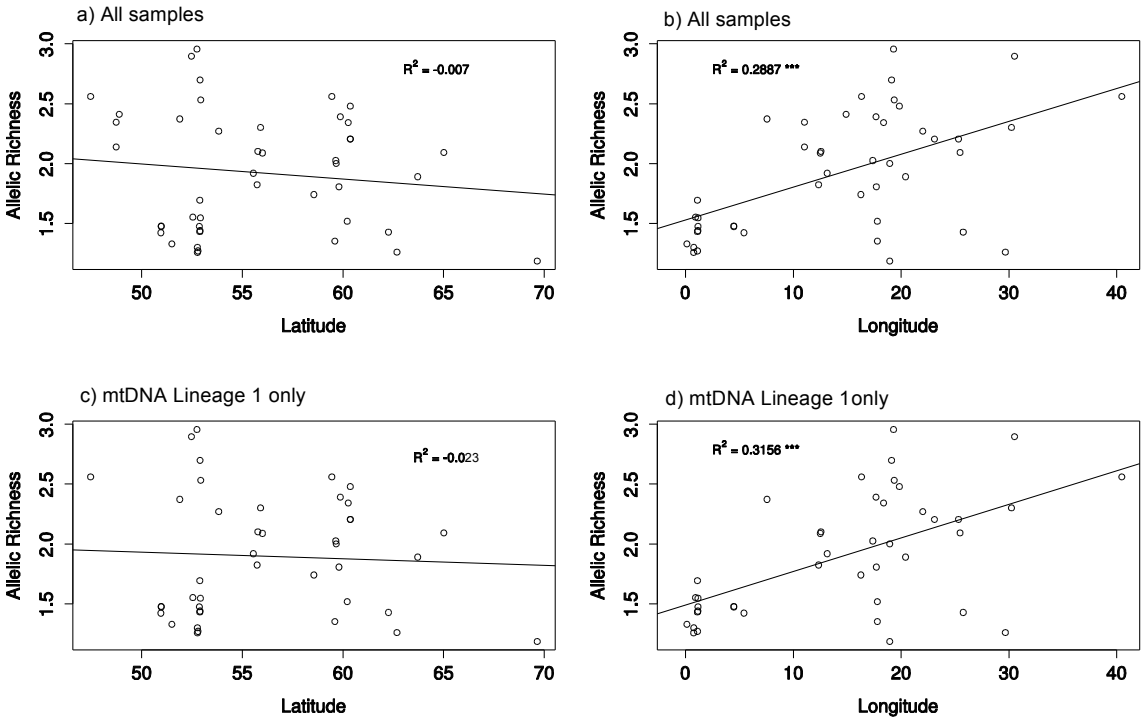
	GBR8	BEL1	GBR4	FIN3	DEN1	GBR7	SWE12	FIN4	DEN2	POL4	RUS1	SWE2	SWE8	SWE9	SWE10	NOR2	POL3	HUN2	WEN
GBR8		0.34971	0.35695	0.49475	0.223897	0.35613	0.406544	0.295019	0.293628	0.211876	0.38775	0.308973	0.273693	0.412263	0.321365	0.650207	0.146766	0.61801	0.397239
BEL1			0.370425	0.390916	0.098308	0.381154	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



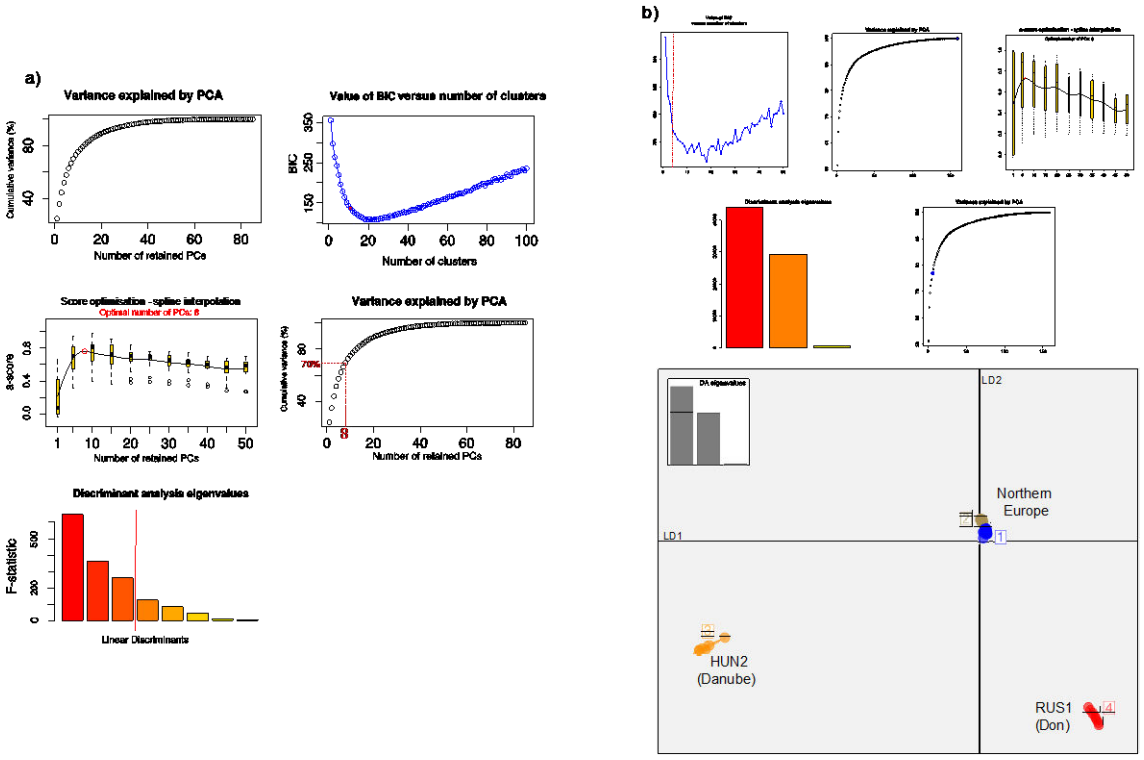
Supplementary Figure 3.2. All scenarios tested in stage 2 **a)** and stage 3 **b)** of DIYABC analysis. See Table 3 for population poolings and prior parameter values. Back to text.



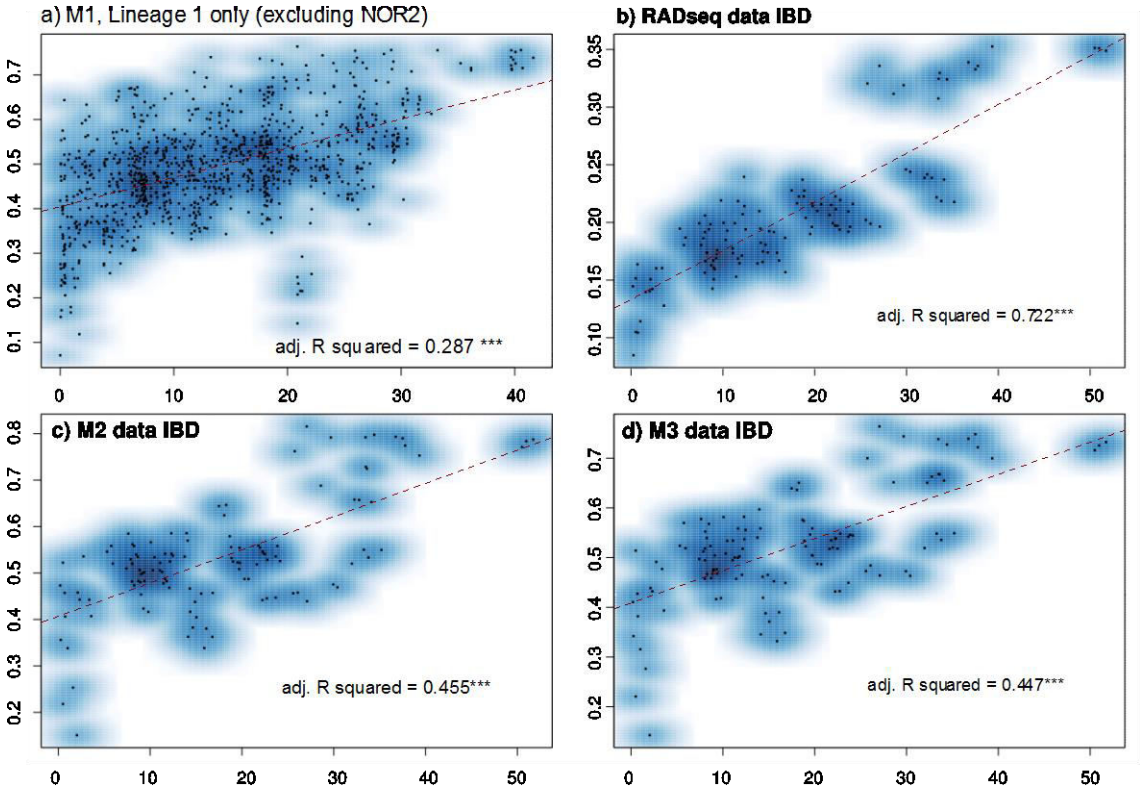
Supplementary Figure 3.3. Filtering out merged ohnologs. **a)** Distribution of SNP locus coverage prior to removing loci that had observed heterozygosity higher than 0.5 in one or more population. **b)** Distribution of locus coverage after filtering, showing a loss of many high coverage loci and a reduction in mean SNP coverage. Note the loss of loci with high coverage.



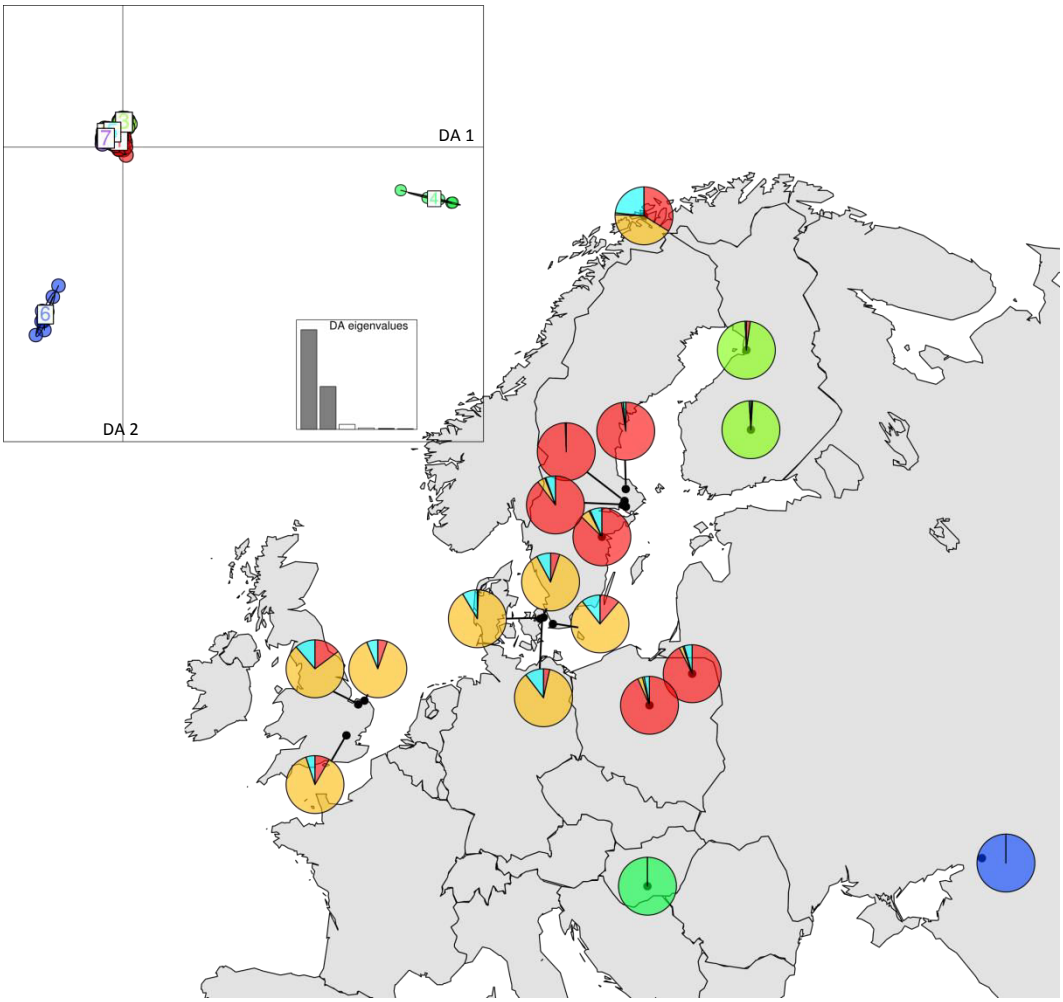
Supplementary Figure 3.4. Linear regressions for all samples **a)** Ar against latitude; **b)** Ar against longitude and for only samples in mtDNA lineage 1 **c)** Ar against latitude; d) Ar against longitude. Back to text.



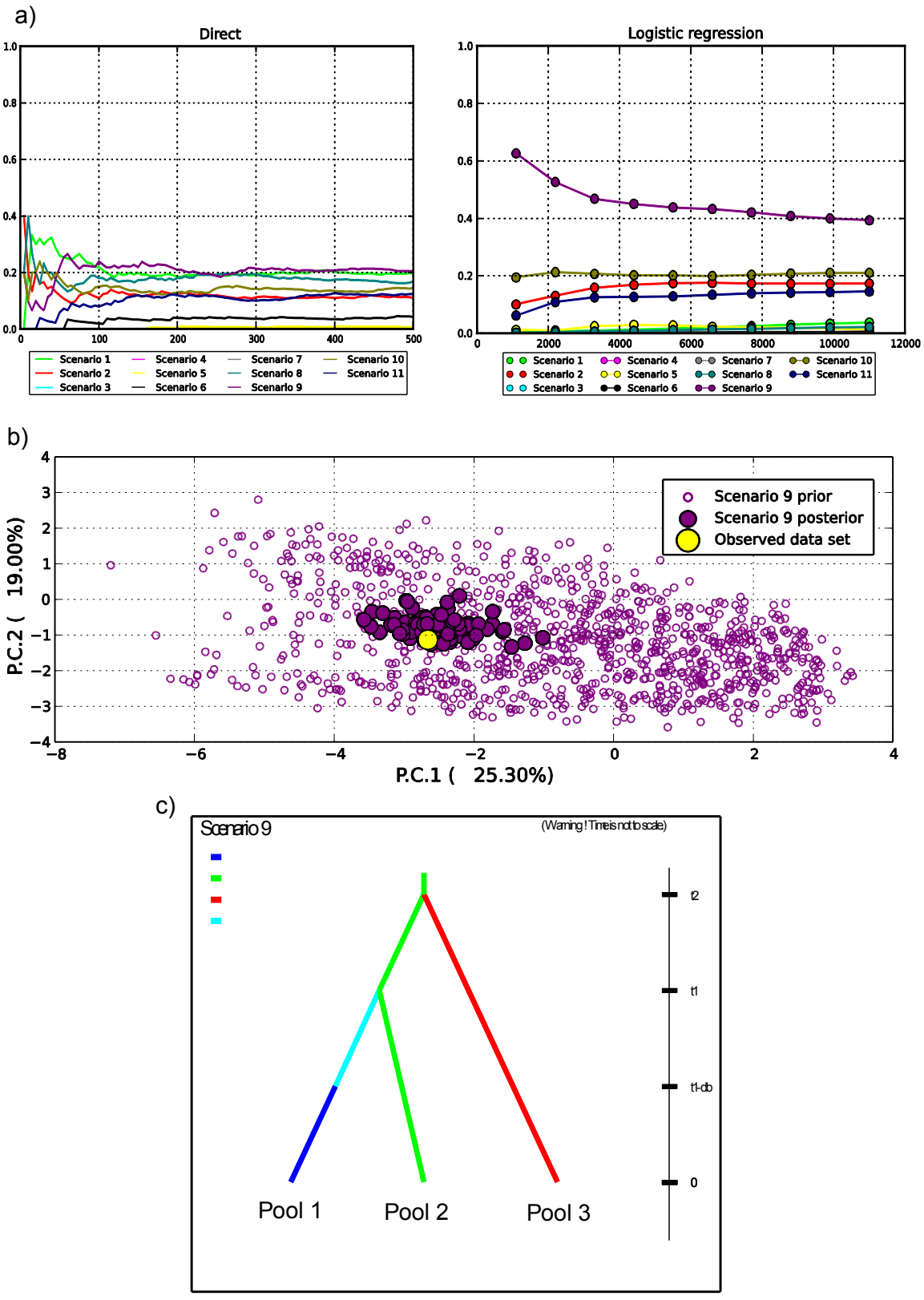
Supplementary Figure 3.5. DAPC analysis of **a)** full microsatellite dataset (Excluding NOR2); for results used in Figure 3.1) and **b)** Full RADseq dataset. Back to text.



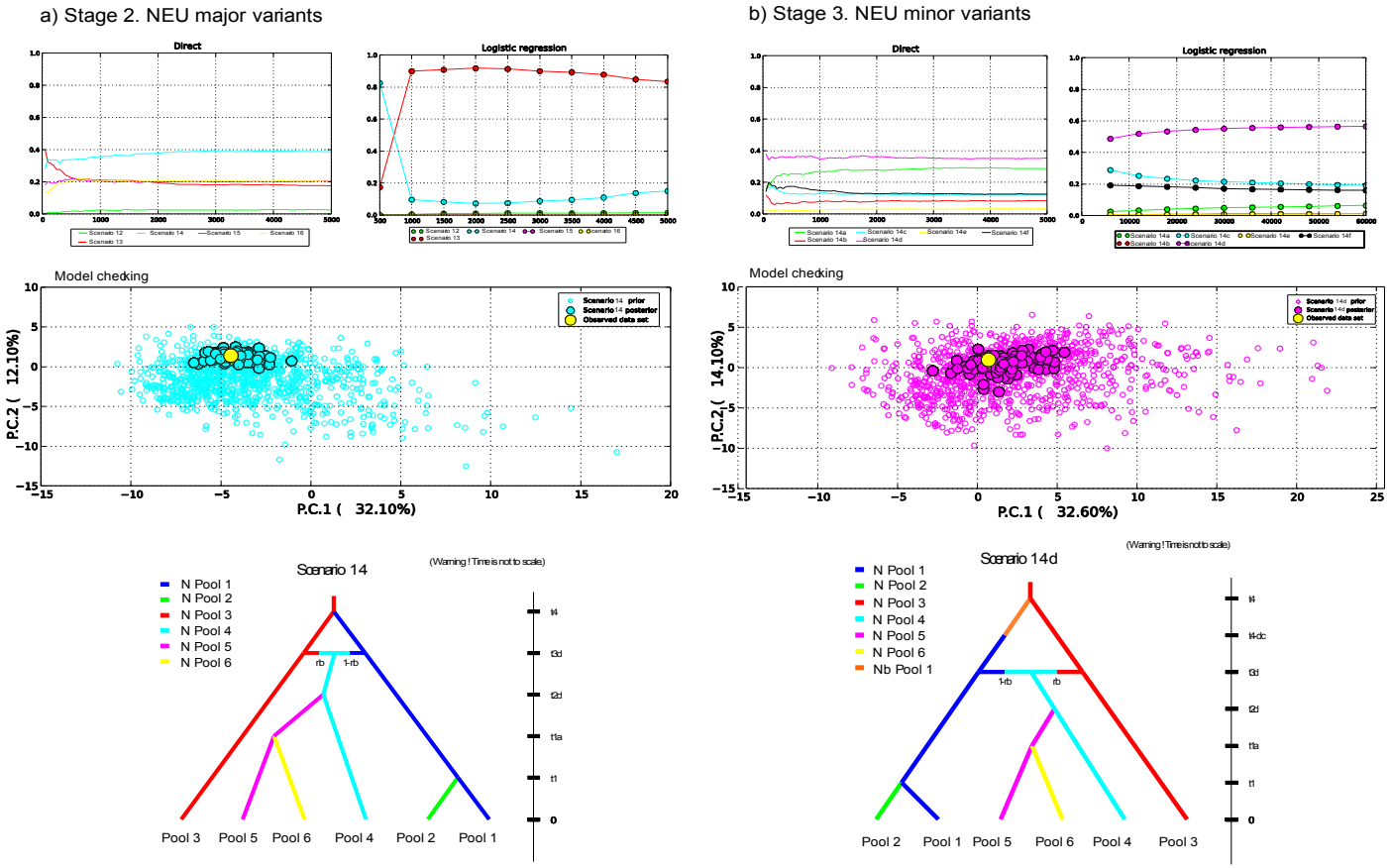
Supplementary Figure 3.6. Isolation by distance **a)** in M1 dataset for mtDNA lineage 1 only (excluding NOR2), **b)** Full RADseq dataset, **c)** M2 dataset and **d)** M3 dataset. Back to text.



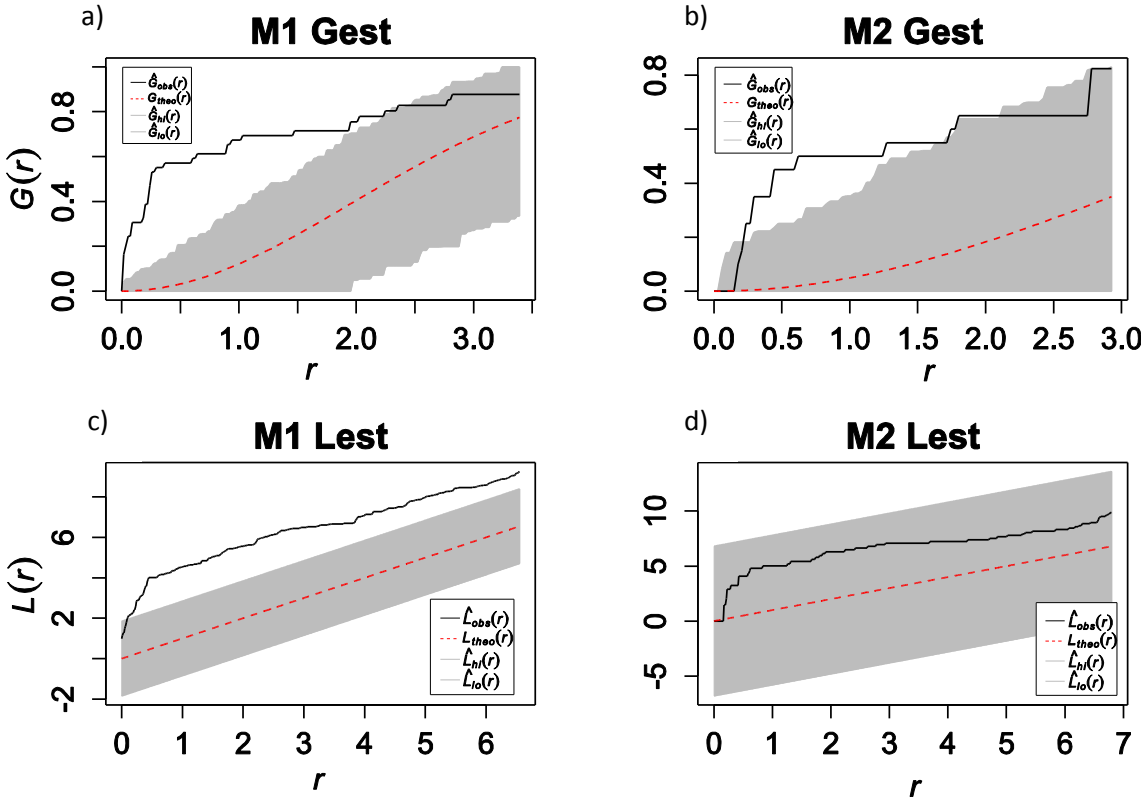
Supplementary Figure 3.7. DAPC scatter plot for the 1000 SNP RADseq dataset used in the DIYABC analysis, showing the same population structure as inferred from the full RADseq dataset. Back to text.



Supplementary Figure 3.8. 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. Back to text.



Supplementary Figure 3.9. Fine scale DIYABC analysis in northern Europe.. a) Stage 2 - major topological variants of scenarios. Direct approact (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 approach (top left), logistic regression (top right) and model checking (middle) all support scenario 14d (bottom). Back to text.



Supplementary Figure 3.10. 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. Back to text.

Chapter 4 Genetic evidence challenges the native status of a threatened freshwater fish (*Carassius carassius*) in England

Authors: Daniel L Jeffries, Gordon H Copp, Lori Lawson Handley, Carl D Sayer, Bernd Hänfling

Abstract

A fundamental consideration for the conservation of a species is the extent of its native range, however defining a native range is often challenging as changing environments drive shifts in species distributions over time. The crucian carp, *Carassius carassius* (L.) is a threatened freshwater fish native to much of Europe, however the extent of this range is ambiguous. One particularly contentious region is England, in which *C. carassius* is currently considered native on the basis of anecdotal evidence. Here, we use 13 microsatellite loci, population structure analyses and approximate bayesian computation (ABC), to empirically test the native status of *C. carassius* in England. Contrary to the current consensus, ABC yields strong support for introduced origins of *C. carassius* in England, with posterior distribution estimates placing their introduction in the 15th century, well after the loss of the Doggerland landbridge. This result brings to light an interesting and timely debate surrounding our motivations for the conservation of species. We discuss this topic, and make arguments for the continued conservation of *C. carassius* in England, despite its non-native origins.

Introduction

Obtaining a detailed understanding of a species' native range and the distribution of its diversity within that range is fundamental for species conservation (Frankham *et al.* 2002; Reed & Frankham 2003; Scoble & Lowe 2010; IUCN 2012). However, this is complicated by the fact that species' ranges are not static but often change dramatically over time in response to changing environments and newly arising dispersal corridors. A species is usually considered native if it has colonised an area naturally. Thus, it follows that areas which have been colonised with human intervention are not included as part of the native range. This has profound implications for the areas in which a threatened species may be conserved (e.g. Copp et all 2005). During the last 2.5 MY, the ranges of European biota have been impacted most strongly by the glacial cycles (Hewitt 1999). These processes have been extensively studied in particularly in freshwater fish, whose postglacial recolonisation dynamics have been determined by the history of river drainage systems (Bianco 1990; Bănărescu 1990, 1992; Bernatchez & Wilson 1998; Reyjol *et al.* 2006). For example, ephemeral rivers and periglacial lakes that result from glacial meltwater have provided opportunities for fish colonisations (Gibbard *et al.* 1988) of otherwise isolated drainages (Grosswald 1980; Arkhipov *et al.* 1995). However, human-mediated translocations also had a significant impact on the current distributions of European freshwater fish have also been determined, which have enabled some species to overcome natural dispersal barriers like watersheds (Copp *et al.* 2005; Gozlan *et al.* 2010). Knowing whether natural or human mediated dispersal, is responsible for an organism's contemporary distribution, is fundamental in determining its native range.

However, this distinction is particularly difficult to make in the UK. With very few exceptions such as groundwater invertebrates (McInerney *et al.* 2014), it is thought that the vast majority of terrestrial and freshwater animals were forced South, into continental refugia, by the expansion of the Weichselian ice sheet during the last glaciation. At its maximum extent, approximately 25000 years before present (YBP), this ice sheet covered almost the entirety of the UK, with frozen tundra covering the remaining unglaciated land area (Coles 2000). Native UK species have therefore recolonised this region over the last 18,000 years, when the Weichselian ice sheet began to recede. In the case of primary freshwater fish, this was made possible by connections between English and Continental river systems that existed in Doggerland, the land

bridge connection between southeast England and continental Europe. However, this window of opportunity was relatively short, as Doggerland was inundated at around 7800 YBP with rising sea levels resulting from the continued melting of the Weichselian ice sheet (Coles 2000).

After the loss of the Doggerland land bridge, the only means by which freshwater species could colonise the UK, precluding the very unlikely possibility of fertilised eggs being transported by migrating waterfowl (for which no empirical evidence exists, to our knowledge), would have been via human mediated introductions. The earliest known record of live fish translocations into the UK was the movement of common carp, *Cyprinus carpio*, into the southeast of England by monks in the 15th century (Lever 1977). Although, it cannot be ruled out that they were introduced by earlier civilisations, e.g. the Romans, in the 1st century A.D or in the following few centuries by Viking invaders.

The dates described above therefore allow us to make a clear distinction between the possible arrival times of a primary freshwater fish in the UK under two hypotheses; if native, then it must have colonised naturally before 7800 YBP, if introduced, then realistically it could not have arrived earlier than approximately 2000 YBP.

One species, which, in the past, has had a particularly contentious status in the UK is the crucian carp (*Carassius carassius*, Linneaus 1758); a primary freshwater fish, native to much of central and Eastern Europe. The crucian carp is of conservation concern in much of its range due to sharp declines in the number and sizes of populations in recent times, which has led to local population extinctions (Copp *et al.* 2010; Savini *et al.* 2010; Sayer *et al.* 2011; Mezhzherin *et al.* 2012; Rylková *et al.* 2013). Awareness of the threats to *C. carassius* is building, and it often appears on red-lists at the national level e.g , Czech Republic (Lusk *et al.* 2004), Ukraine (Andrievskiy 2009), Austria (Wolfram & Mikschi 2007), Croatia (Mrakovčić *et al.* 2007) and Serbia (Simic, V *et al.* 2009). Despite this, however, there are still very few active conservation initiatives for *C. carassius* in Europe and, to our knowledge, one of the most comprehensive of these exists in Norfolk, in eastern England (Copp & Sayer 2010; Sayer *et al.* 2011).

Characterising the native range of *C. carassius* has been hampered in the past, largely due to morphological confusion with closely related species (Wheeler 2000; Hickley &

Chare 2004). *C. carassius* is presently assumed to be native in southeast England on the basis of two pieces of evidence. Firstly, he identification of *C. carassius* pharyngeal bones found at a Roman archaeological dig site Southwark, London (Lever 1977; Jones 1978), and secondly the similarity of its distribution, in southeast England, to those of other native freshwater fish species, such as silver bream, *Blicca bjoerka* (L.), Ruffe, *Gymnocephalus cernuus* (L.), burbot *Lota lota* (L.) and spined loach, *Corbitis taenia* (L.) (Wheeler 1977, 2000). However, in contrast, Maitland (1972) suggested that *C. carassius* was introduced to south east England along with common carp in the 15th century. More recently, Chapter 3 inferred substantial shared ancestry between UK and several Belgian and German populations from microsatellite and genome wide SNP markers supporting the hypothesis of a more recent origin.

Recently, Approximate Bayesian Computation (ABC) methods have been developed (Cornuet *et al.* 2008), that allow such questions to be addressed more explicitly in a population genetic framework, which is suitable for investigating events on a post-Pleistocene timescale. In the present study, we employ ABC to empirically test the status of *C. carassius* in southeast England, using highly polymorphic microsatellite markers. Specifically we test three possible alternative hypotheses for the *C. carassius* colonisation of England; i) all English populations originate from natural colonisation from Continental Europe more than 7800 YBP, ii) all English populations were introduced by humans from Continental Europe sometime in the last 2000 years or iii) some English populations are native and some have been more recently introduced. Our ultimate aim is to increase the knowledge available for the assessment of status and conservation of *C. carassius* in England and Continental Europe.

Methods

Samples, DNA extraction and microsatellite amplification

The samples used in this study include 257 *C. carassius*, from 11 English populations, three Belgian populations and one German population (Table 4.1, Figure 4.1). These represent a subset of samples from a Europe-wide phylogeographic study, which used the same 13 microsatellite loci as used here, as well as mitochondrial DNA sequences and genome wide SNP data (see Jeffries et al 2015 for Methods). In Chapter 3, population structure analyses of the Europe-wide dataset showed that these fall into a

single genetic cluster, which was distinct from the other genetic clusters found in Europe. The Belgian and German samples used in the present study therefore represent the closest known relatives of English *C. carassius* populations in Europe (Jeffries et al 2015) and are the most likely of our sampled populations to have been the source of their colonisation.

Table 4.1. Location, number and summary statistics of samples used in the present study for microsatellite analyses.

Code	Location	Country	Drainage	Coordinates		N	H_{obs}	A_r
				lat	long			
GBR1	London	U.K.	U.K	51.5	0.13	9	0.11	1.33
GBR2	Reading	U.K.	U.K	51.45	-0.97	4	0.03	NA
GBR3	Norfolk	U.K.	U.K	52.86	1.16	7	0.16	1.48
GBR4	Norfolk	U.K.	U.K	52.77	0.75	27	0.12	1.26
GBR5	Norfolk	U.K.	U.K	52.77	0.76	14	0.13	1.30
GBR6	Norfolk	U.K.	U.K	52.54	0.93	20	0.22	1.55
GBR7	Norfolk	U.K.	U.K	52.9	1.15	24	0.15	1.44
GBR8	Hertfordshire	U.K.	U.K	52.89	1.1	37	0.16	1.43
GBR9	Norfolk	U.K.	U.K	52.8	1.1	27	0.09	1.27
GBR10	Norfolk	U.K.	U.K	52.89	1.1	14	0.21	1.69
GBR11	Norfolk	U.K.	U.K	52.92	1.16	20	0.18	1.55
BEL1	Bokrijk	Belgium	Scheldt River	50.95	5.41	13	0.15	1.42
BEL2	Meer van Weerde	Belgium	Scheldt River	50.97	4.48	12	0.19	1.48
BEL3	Meer van Weerde	Belgium	Scheldt River	50.97	4.48	8	0.16	1.47
GER2	Münster	Germany	Rhine River	51.89	7.56	21	0.4	2.37
						257		

DNA was extracted from tissue samples using either the Puregene DNA isolation kit or the DNeasy DNA purification kit (Qiagen, Hilden, Germany). Samples were then genotyped at 13 microsatellite loci, which were amplified in three multiplex reactions using the Qiagen multiplex PCR mix with manufacturer's recommended reagent concentrations, including Q solution and 1 µl of template DNA. The annealing temperature was 54°C for all reactions and individual primer pair concentrations within each multiplex reaction were optimised depending on the relative PCR product yield for each locus (see Chapter 3). PCR reactions were run on an Applied Biosciences® Veriti Thermal Cycler and microsatellite fragment lengths were analysed on a Beckman Coulter CEQ 8000 genome analyser using a 400 bp size standard.

Standard Population statistics

First, allele dropout and null alleles in the data were tested for using Microchecker (Van Oosterhout *et al.* 2004). FSTAT v. 2.9.3.2 (Goudet 2001a) was then used to check for

linkage disequilibrium (LD) between loci, deviations from Hardy-Weinberg equilibrium (HWE) within populations and for all population genetic summary statistics. Genetic diversity within populations was estimated using Nei's estimator of gene diversity (H_e) (Nei 1987) and Allelic richness (A_r), which was standardised to the smallest sample size (n =7) using the rarefaction method (Petit *et al.* 1998). In order to quantify differentiation among populations, pairwise F_{ST} values were calculated in FSTAT (Goudet 2001b) using the multilocus (Weir & Cockerham 1984) F_{ST} estimator. Sequential Bonferroni correction and permutation tests (2100 permutations) were used to test for significance of F_{ST}. We also used the Hierfstat package (Goudet 2005) in R (R Core Team 2013), to quantify the genetic variation (F_{ST}) at 4 hierarchical levels of population isolation, the population-level (separate ponds within countries), the country-level (between Belgium and Germany) the landmass-level (between England and continental Europe) and also at the level of the DIYABC pools used (described below). In the latter case, hierarchical F_{ST}s were used to validate the population poolings used for the DIYABC as in Pedrischi et al. (2013)

Testing the native status of C. carassius in England

In order to test our three alternative hypotheses for the colonisation of *C. carassius* in England, an Approximate Bayesian Computation (ABC) approach was taken, implemented in the program DIYABC (Cornuet *et al.* 2014). DIYABC simulates datasets of expected summary statistics (ESS) for user-defined demographic scenarios ('scenario' is used herein to describe a specific population tree topology together with the parameter distribution priors that are associated with it). These scenarios were then statistically compared to the actual observed data, allowing us to identify those that are most likely to represent the true history of populations (Cornuet *et al.* 2008).We then estimated the divergence time between populations based on posterior parameter distributions to provide a likely date for the arrival of *C. carassius* in the UK.

.

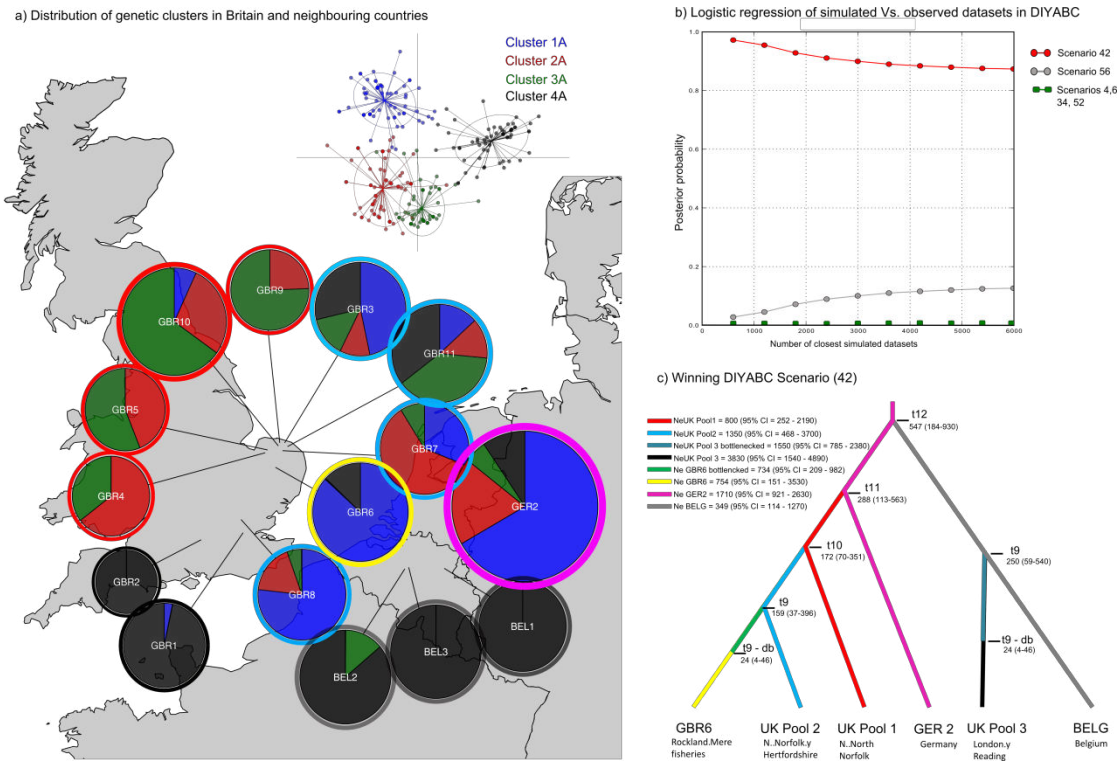


Figure 4.1. a) DAPC analysis of *C. carassius* in northwest Europe, showing similar genetic composition of English and Continental populations. Pie charts represent the a single population, the size of the pie chart is relative to A_r and the colours within them correspond to clusters inferred from DAPC analyses. Coloured rings around the pies show the pooling of populations for DIYABC analyses and correspond the the colours in c). b) Posterior probabilities that each of the of the 6 likely DIYABC scenarios explains the distribution of diversity in the northwest European *C. carassius*, calculated using linear regression between the observed dataset and the closest 6000 simulated datasets. c) Scenario 42 - the winning DIYABC, in which *C. carassius* were brought to the UK approx. 288 generations ago (t11). Numbers next to nodes represent the estimated time in generations for that event, with error margins in parentheses. "db" stands for duration of bottleneck.

In order to reduce the number of scenarios to be tested from the huge number possible, we grouped populations in DIYABC analyses into pools of populations with shared history, a method also employed by Pedreshi et al. (2013). To inform these poolings it was first necessary to perform a fine scale population structure analysis of the 15 populations used. This was done using Discriminant Analyses of Principal Components (DAPC), implemented in the Adegenet R package (Jombart *et al.* 2010). Bayesian Information Criteria (BIC) scores were used to choose the appropriate number of genetic clusters in the dataset. Spline interpolation (Hazewinkel 1994) was then used to

identify the appropriate number of principal components for use in the subsequent discriminant analysis.

Based on the results of the DAPC analysis, populations were grouped into six pools. Those of similar genetic composition (and therefore very likely to have a shared history) were pooled together (see results section). However, if populations from either side of the English Channel shared similar genetic composition, they were separated across pools, to allow for hypothesis testing.

In total, 56 scenarios were tested: six, 39 and 11 representing hypothesis i), ii) and iii) respectively (Supplementary Figure 4.1). The number of scenarios for each hypothesis reflects the number and plausibility of the possible population histories for the different hypotheses given the results of the populations structure analysis. The discriminating factors between scenarios representing different hypotheses were tree topology and, most importantly, the parameter priors for the divergence times between populations (Supplementary table 4.1). These divergence time priors were set in order to represent the possible time windows of *C. carassius* introduction under our three hypotheses. To test hypothesis i) – the natural colonisation of *C. carassius* more than 7800 YBP - the time prior for the oldest split between English and Continental European populations was set to 4000-10000 generations (equivalent to 8000 – 20000 YBP, assuming an average generation time of two years (Tarkan et al. 2010), Supplementary Figure 4.1: scenarios 1-6). To test hypothesis ii) – that English *C. carassius* were introduced after the 15th century - the same prior was set to 10-1000 generations (2 – 2000 YBP, scenarios 25 - 44), which very conservatively encompasses all dates of possible live fish translocations to the UK by humans. Finally, to test hypothesis iii) - that some populations were native and some introduced we used multiple combinations of both native and introduced prior dates (as used in hypothesis i and ii) scenarios respectively) for different population splitting events (scenarios 45 – 56). In the interests of completeness, we also tested an intermediate time window of 10 – 2500 generations (20 – 5000 YBP, scenarios 7-24). Analyses were performed in a sequential manner, whereby a million datasets per scenario were first simulated in DIYABC. Then, for the computationally intensive part of the analysis, simulated datasets were grouped according to the hypothesis they represented (i.e. (i), (ii) or (iii)) and these groups were separately compared to the observed data using both approaches offered in DIYABC, logistic regression and “direct estimate”. The latter of which is a count of the number of

times that a given scenario simulates one of the closest datasets to the real data set (Cornuet *et al.* 2008). The resulting posterior probabilities were used to identify the top two most likely scenarios for each hypothesis (six in total). These were then used in a final test, again using logistic regression and direct estimate, to identify the single most likely scenario of the final 6. Model checking analyses, which measures the discrepancy between the model parameter posterior combination and the actual data (Cornuet *et al.* 2010), were then carried out to test the robustness of scenario choice. Finally, posterior parameter distributions for effective population size, divergence times and bottleneck parameters were estimated on the basis of the most likely scenario.

Results

Microsatellite data analyses

Microchecker showed no consistent signs of null alleles or allele dropout in populations of pure *C. carassius* and no LD was found between loci pairs. Tests of Hardy-Wienberg proportions did not identify any populations that significantly deviated from HWE.

Population Structure in England, Belgium and Germany

Population structure was weakest (F_{ST} = 0.0) between the two Belgian populations, strongest (F_{ST} = 0.736) between GBR2 and GBR4 (Supplementary table 4.2) and followed a weak IBD pattern, being significantly associated with geographic distance (adjusted R^2 = 0.248, $P < 0.001$, Supplementary Figure 4.2). Hierarchical assessment of population structure showed that variation between individuals was significantly explained by population assignment and country (F_{pop} = 0.36, P =0.001; $F_{country}$ = 0.154, P =0.001). However the landmass (continental Europe or Britain) had no significant effect on variation between individuals ($F_{landmass}$= -0.04, P =0.482). Importantly, the pools used in DIYABC analysis explained a large of the genetic between individuals in total (F_{pools} = 0.244, P =0.001) and within the poolings the remaining variation between individuals was considerably lower than at the landmass level or the country level, though still highly significant ($F_{Ind/pools}$ = 0.142, P = 0.001), confirming that these population groupings were appropriate groupings for the populations in DIYABC analyses.

Observed heterozygosity (averaged across all loci within a population) ranged from 0.03 (GBR2) to 0.4 (GER2). A_r ranged from 1.26 (GBR4) to 2.37 (GER2), and correlated with H_o (adjusted $R^2 = 0.543$, $P = 0.001$).

In the DAPC analysis of population structure, ten genetic clusters were indicated by BIC scores (Supplementary Figure 4.3c). The resulting population-cluster identities were complex (Supplementary Figure 4.3b), with most populations containing many closely related clusters (Supplementary Figure 4.3a) making it difficult to identify sets of closely related population for pooling. Therefore in order to reliably inform our DIYABC poolings, we incrementally dropped the number of clusters to four which seem to reflect the large scale patterns of genetic differentiation better. Seven principal components and two linear discriminants were retained in this final, four-cluster DAPC analysis (Figure 4.1a). The resulting inferred population structure showed that many of the English populations showed higher similarities to Continental populations than to neighbouring English populations. For example, GBR1 and GBR2 were extremely similar to Belgian populations, and GBR3, 6, 7, 8 and 11 were more similar to populations in northern Germany (Figure 4.1a). However, GBR4, 5, 9, 10, all in north Norfolk (eastern England), showed some distinctiveness from continental populations.

Testing the native status of C. carassius in England

For the DIYABC analyses, populations were grouped into six pools on the basis of the above DAPC results (pools are denoted by coloured rings around pie charts in Figure 4.1a). Within-hypothesis logistic regressions of simulated vs. observed data, performed in DIYABC, showed that the two most likely scenarios for each hypothesis were scenarios 4 and 6 for hypothesis i); 42 and 34 for hypothesis ii) and 52 and 56 for hypothesis iii). These final six scenarios were then tested against each other, again using logistic regression to find the single most likely scenario of all 56 tested. Scenario 42, representing hypothesis ii), produced data sets that were, by far, the closest to the real data, with a posterior probability of 0.91 (Figure 4.1b).

Scenario 42 (Figure 4.1c) had prior constraints on the split between English and Continental populations (t11) of 10 – 1000 generations and thus supports a human introduction of *C. carassius* into southeast England <2000 YBP. Under this scenario, the oldest demographic event was the split between German and Belgian populations

approximately 547 generations ago (1094 YBP). However, the most important demographic event for the purposes of testing our hypotheses is the split between English populations (UK pools 1, 2 and RM) and continental populations (pools GER2 and BELG), at time "t11" in Scenario 42 (Figure 4.1c). Furthermore, this scenario suggests that the ancestral source population of the initial English introduction was more closely related to the German than the Belgian populations sampled here. The date of this English/Continental population split is estimated at 288 (95% CI = 113-563, Supplementary table 4.3) generations ago, which corresponds to 576 (95% CI = 226 - 1126) YBP, approximately 7400 years after the loss of the Doggerland land bridge. DIYABC also outputs posterior estimates of population split times scaled by mutation rate and effective population size. The estimated time for the English/Continental population split, scaled by mutation rate estimated by the model was $t11(u+_{SNI}) = 9.83 \times 10^{-2}$ (where u+SNI is the median estimate of the microsatellite mutation rate using the generalised stepwise mutation model, (1.11×10^{-4} mutations/locus/generation) and SNI is the single nucleotide insertion rate (6.18×10^{-8}/mutations/locus/generation) Supplementary table 4.3). The median estimate of this mutation rate ($u = 1.11 \times 10^{-4}$ /locus/generation), although slow, is still within the realms of that observed in the closely related *C. carpio* (mean = 5.56×10^{-4} mutations/locus/generation, 95% CI = 1.52×10^{-4} - 1.63×10^{-3}, (Yue *et al.* 2007)) and indeed in humans (Ellegren 2004).

To validate this result we first tested the "goodness-of-fit" of Scenario 42 using statistical model checking as implemented in DIYABC, which showed that the observed data fell well within the predictive posterior parameter distribution of the simulated data (Supplementary Figure 4.4). Secondly, we calculated the oldest possible date of the English/Continental population split using its upper 95% confidence value under Scenario 42 (563 generations), and assumed the unrealistic, but sometimes possible generation time of 5 years (Tarkan *et al.* 2010). Despite these extremely conservative values, the split between English and Continental populations was still estimated at 2815 YBP, approximately 5000 years after the flooding of Doggerland. Finally, we inferred t11 (the English/Continental population split) of scenario 42 using the scaled parameter estimate, t11(u+SNI). This gave an estimate of 885 generations, or 1770 years (with a two year generation time), which, although older than the un-scaled estimate, is still over 6000 years later than the possible natural colonisation window. In fact, in order for the scaled estimate to fit the hypothesis of natural colonisation (more than 8000 years ago), assuming a two year generation time, the mutation rate would

have to be approximately 1.0 x 10^{-5} mutations/locus/generation, at least one order of magnitude lower than reported for microsatellite loci (reference).

Further population splits have occurred more recently from this initial introduction, and there is also support for a second independent introduction of *C. carassius* into the UK (t9) approximately 250 (95% CI = 59-540) generations or 500 (95% CI = 118-1080) years ago (UK pool 3), from a source population closely related to the Belgian populations sampled here.

Discussion

The primary aim of the present study was to test the contentious assumption that *C. carassius* arrived in southeast England naturally. Owing to its hydrogeological history during the last glaciation, the UK presents a rare opportunity to test such a question amongst its inhabitants. Our analyses suggest that *C. carassius* was anthropogenically introduced into England and on this basis we therefore discuss the potential implications for *C. carassius* conservation.

Non-native origins of C. carassius in England

Analyses of the population structure within southeast England and closely neighbouring countries revealed that many English populations are more similar genetically to continental populations than to their English counterparts, implying multiple independent colonisation events or introductions into England. DIYABC analyses supported this, suggesting that populations GBR1 and GBR2 split from Belgian populations more recently than they did from other English populations (Figure 4.1c). Indeed these populations are known to be managed and therefore have likely been stocked in the recent past; GBR1 being a conservation pond, and GBR2 a fish farm. Therefore, our results indicate that these fish came from recently imported stocks closely related to the sampled Belgian populations.

In contrast to GBR1 and GBR2, DIYABC analyses suggest that all north Norfolk and Hertfordshire populations share a most recent common ancestor with the sampled German population; indicative of a separate introduction. The central question of this

analysis however, was; how long ago was the first colonisation or introduction of *C. carassius* into England? DIYABC analyses predicted that the oldest possible date for the arrival of *C. carassius* in England was approximately 1126 YBP but most likely 576 YBP; over 7000 years after the loss of the Doggerland land bridge, and that there were in fact two independent introductions around this time.

As this result could have important implications for the conservation of *C. carassius* in the UK (see below), we performed rigorous results checking. Tests for the goodness-of-fit of the winning scenario (42) confirmed that this was the most likely out of all scenarios tested, and even when using the 95% confidence interval limits of the posterior time parameter distribution or using the unrealistically long generation time of 5 years (to convert DIYABC results from generations to years), it still was not possible to achieve estimates of the split between English and continental populations older than 2815 YBP. Only with a mutation rate an order of magnitude slower than that estimated here (and elsewhere, e.g in *C. carpio* (Yue *et al.* 2007), mice (Dallas 1992), sheep (Crawford & Cuthbertson 1996) and humans (Ellegren 2004)) would the time for this split support a natural introduction of *C. carassius* into England.

Although our sampling is not exhaustive, it comprehensively covers the areas of England previously thought to contain native *C. carassius* populations, in particular Norfolk, which is thought to have been a stronghold for *C. carassius* in the past (Patterson 1905; Ellis 1965; Sayer *et al.* 2011). It is therefore unlikely that there are unsampled populations of *C. carassius* in England that show further divergence from those of continental Europe. Furthermore, broad scale phylogeographic results in Jeffries et al (2015) show that Belgian and German populations are the closest relatives of English *C. carassius* in Europe. In fact, adding currently unsampled populations from continental Europe could only result in a lower estimate of divergence between English and continental European samples. We are, therefore, confident that our estimate represents the earliest possible timeframe for the first *C. carassius* introductions into England. It should also be noted that the estimate for this split does not directly predict when populations were introduced to England, only when they were separated from the sampled continental European populations, which must have been at the same time as, or prior to, their introduction. Thus, it is entirely possible that the arrival time of *C. carassius* in the UK was even more recent than the DIYABC estimate of population divergence time.

However, we cannot rule out the possibility that *C. carassius* colonised naturally, but either then went extinct, or were extirpated by the current English *C. carassius* strains when they were introduced. If these scenarios were true, only dated fossil evidence, and perhaps ancient molecular studies would allow for a definitive answer.

The results of this study therefore strongly point to the anthropogenic introduction of English *C. carassius* and, in fact, fall perfectly in line with the first known record of *C. carpio* introductions into England by monks for food in the 15th Century (Lever 1977). However, we can only speculate as to the motivations behind these introductions. To our knowledge, *C. carassius* are not mentioned in the literature until 1766 (Pennant 1766), however it is possible that *C. carassius* was intentionally introduced as a source of food, as with *C. carpio*. Indeed there are mentions of *C. carassius* used as food in 1778 in Norfolk (Woodforde *et al.* 2008), and although *C. carassius* does not grow to the size of other carp species, its ability to survive in small, isolated and often anoxic ponds may have made it an attractive species for use in medieval aquaculture. It is possible, however, that the introduction of *C. carassius* in England was unintentional. For example, it can be very difficult to tell *C. carassius* and *C. carpio* apart, especially if they are found in sympatry and if hybrids are present (Wheeler 2000), as is often the case (Hänfling *et al.* 2005; Sayer *et al.* 2011). Irrespective of the initial motivations however, intentional movements of *C. carassius* have since been common, predominantly for angling purposes (Sayer *et al.* 2011).

Conclusions and implications for the conservation of C. carassius

A fundamental consideration in the conservation of a species is its native range, and, contrary to current belief, the results of this study support the human-mediated introduction of *C. carassius* into England. But what does this mean for the conservation of *C. carassius* in England, a country which has one of the few active projects in place for its conservation (Copp & Sayer 2010)? In light of these results, should England cease efforts conservation of *C. carassius*? There has been a call recently, for a change in the conservation paradigm, moving away from the unfounded assumption that all non-native species have detrimental impacts on native ecosystems (Davis *et al.* 2011). Instead the authors advocate embracing the idea of constantly changing communities, and moving towards impact-driven conservation, whereby only those species that have

been empirically shown to be invasive and detrimental to native ecosystems and economies are actively managed. Indeed only a small proportion of freshwater fish introductions have been shown to have detrimental impacts on the native ecosystem, whereas many provide significant ecological and economical benefits (Gozlan 2008; Schlaepfer *et al.* 2011), and sometimes replace ecosystem services lost in extinct species (Schlaepfer *et al.* 2011). Currently, *C. carassius* could not be labelled as invasive in England, as they are not expanding, in fact, they are declining in numbers in England (Sayer *et al.* 2011). To date, there has been no attempt to assess the impact of *C. carassius* on ecosystems due to the assumption that they were native, however, available studies show that *C. carassius* are widely associated with species-rich, macrophyte-dominated ponds (Sayer *et al.* 2011), which are extremely important ecosystems for conservation (Oertli *et al.* 2002). There is no evidence that *C. carassius* negatively impact these habitats, unlike *C. carpio* (Miller & Crowl 2006), and despite concerns that *C. carassius* may impact the threatened great crested newt (*Triturus cristatus,* Laurenti 1768), this does not seem to be the case in UK ponds, with *C. carassius* often co-existing with recruiting *T. cristatus* populations (Chan 2010).

A further important consideration in the case of *C. carassius* is its threatened status in much of its native European range. Copp *et al.* (2005) pose the question; should we treat all introduced species in the same way, even if one such species is endangered in its native range? Indeed, if the goal of conservation science is to protect and enhance biodiversity, it would seem counterproductive to abandon the conservation of *C. carassius* populations in one region when they are threatened in another. Our Europe-wide population structure results show that English populations, along with those in Belgium and Germany, comprise a distinct part of the overall diversity of *C. carassius* in Europe. And this is made all the more important by the expansion of *C. gibelio* through Europe, especially into the Baltic Sea basin from the south (Wouters *et al.* 2012; Deinhardt 2013); Lauri Urho. Pers. comms). Although the invasive *C. auratus* is present and poses a threat to *C. carassius* in England (as it does in continental Europe), *C. gibelio* is not yet present and therefore England may represent an important refuge from this threat.

A final consideration for the continued conservation of *C. carassius* is their status as an English heritage species. *C. carassius* is affectionately regarded by the zoological and angling communities of England and as such, has regularly featured in the writings of

both groups over the past three centuries (see the many examples in (Rolfe 2010), pp. 50-64). Therefore, although our results indicate that *C. carassius* can probably not be regarded as a native species in the true sense, the species been has been an important part of the cultural landscape in England for around 500 years.

As outlined above, despite the evidence that *C. carassius* is non-native in England, strong arguments can be made for its continued conservation in that important part of its range. However, our results bring to light much broader and timely questions in invasion and conservation biology; how many assumptions about the native status of other freshwater species in the UK would stand up to the same tests as performed here, if the data were available to perform it? And what do we do about it if they don't?

Chapter 4. Supplementary materials

Supplementary table 4.1. Prior parameters for all scenarios used in DIYABC analyses. Back to text.

Hypothesis tested	Parameter	Defined Prior Distribution	Times in Years*	Conditions
	N1 – N6	Uniform[10 - 5000]		
All	*ra*	Uniform[0.001 – 0.999]		
	db	Uniform[1 - 100]	2 - 500	
i)	t1	Uniform[10 - 10000]	20 – 20000	< t2, t3, t4
	t2	Uniform[10 - 10000]	20 – 20000	< t3, t4
	t3	Uniform[10 - 10000]	20 – 20000	< t4
	t4	Uniform[4000 - 10000]	8000 – 20000	
ii)	t5	Uniform[10- 2500]	20-5000	<t6, t7, t8
	t6	Uniform[10 - 2500]	20-5000	< t7, t8
	t7	Uniform[10 - 2500]	20-5000	< t8
	t8	Uniform[10 - 2500]	20-5000	
	t9	Uniform[10 - 1000]	20 - 2000	<t10, t11, t12
	t9b	Uniform[10 - 1000]	20-2000	<t10, t11, t12
	t10	Uniform[10 - 1000]	20 - 2000	< t11, t12
	t11	Uniform[10 - 1000]	20 - 2000	< t12
	t12	Uniform[10 - 1000]	20 - 2000	
iii)	t12a	Uniform[10 - 2500]	20-5000	<t13, t14, t15, t16
	t13	Uniform[10 - 2500]	20-5000	< t14, t15, t16
	t14	Uniform[10 - 2500]	20-5000	< t15, t16
	t15	Uniform[10 - 2500]	20-5000	< t16
	t16	Uniform[4000 - 10000]	8000-20000	
	t17	Uniform[10 - 1000]	20-2000	<t18, t19, t20
	t18	Uniform[10 - 1000]	20-2000	< t19, t20
	t19	Uniform[10 - 1000]	20-2000	< t20
	t20	Uniform[4000 - 10000]	8000-20000	

Supplementary table 4.2. Pairwise FST values for 15 *C. carassius* populations in northwest Europe. Back to text.

	GBR1	GBR2	GBR4	BEL1	BEL2	BEL3	GER2	GBR7	GBR3	GBR8	GBR9	GBR11	GBR5	GBR6	HUN3
GBR1	0	0.3148	**0.6323**	**0.3711**	**0.207**	0.366	**0.2842**	**0.5017**	0.3676	**0.4369**	**0.6114**	**0.2975**	**0.5956**	**0.3518**	**0.4373**
GBR2		0	**0.7369**	0.353	**0.2644**	0.3795	**0.3334**	**0.6109**	0.535	**0.5616**	**0.7164**	**0.3809**	**0.7**	0.4851	0.5358
GBR4			0	0.6616	**0.494**	0.5857	**0.3076**	**0.1943**	0.4229	**0.3497**	**0.2982**	**0.2058**	**0.2351**	**0.4976**	**0.1883**
BEL1				0	0.0755	0.0204	**0.269**	**0.5601**	**0.3835**	**0.4882**	**0.5886**	**0.2985**	**0.5648**	**0.3322**	**0.4858**
BEL2					0	-0.0149	**0.1878**	**0.4225**	**0.2504**	**0.3869**	**0.4159**	**0.157**	**0.3934**	**0.2715**	**0.3002**
BEL3						0	**0.1909**	**0.4879**	**0.3133**	**0.4379**	**0.5058**	**0.2165**	**0.4763**	**0.299**	**0.3754**
GER2							0	**0.1669**	0.0682	**0.1352**	**0.3144**	**0.134**	**0.2451**	**0.1502**	**0.1845**
GBR7								0	0.1526	**0.0717**	**0.3643**	**0.164**	**0.2876**	**0.3282**	**0.161**
GBR3									0	0.0208	**0.422**	0.0896	**0.3539**	0.0795	0.2133
GBR8										0	**0.4205**	**0.1837**	**0.3569**	**0.2124**	**0.2581**
GBR9											0	**0.2054**	0.0287	**0.4394**	**0.2007**
GBR11												0	**0.1633**	**0.2274**	**0.1071**
GBR5													0	**0.3894**	**0.1972**
GBR6														0	**0.318**
HUN3															0

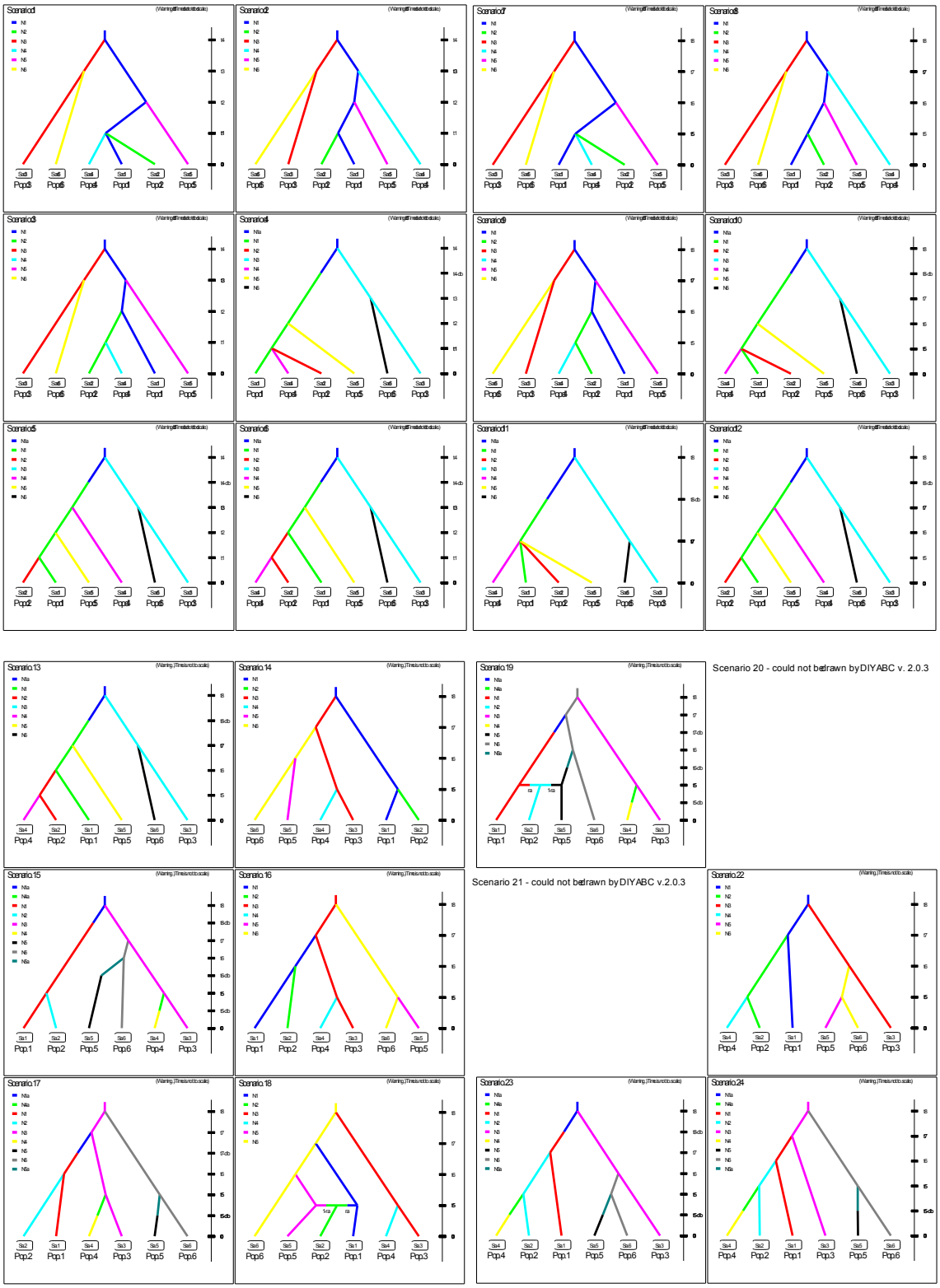
P-values obtained after:2100 permutations

Indicative adjusted nominal level (5%) for multiple comparisons is: 0.000476

Supplementary table 4.3. All posterior parameter distributions for all scenario 42 - identified as the most likely scenario for the colonisation of *C. carassius* into England by DIYABC analyses.

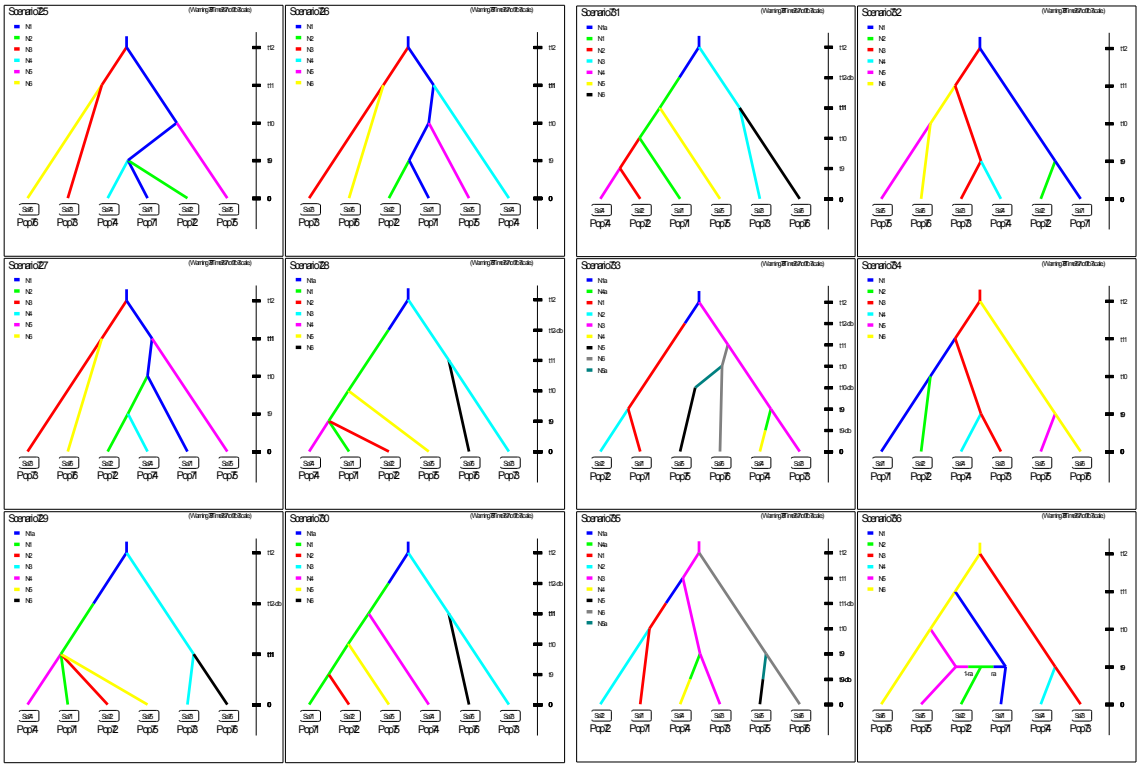
Back to text.

Parameter	mean	median	mode	q025	q050	q250	q750	q950	q975
Original									
N1	9.45E+02	8.00E+02	5.60E+02	2.02E+02	2.52E+02	5.42E+02	1.12E+03	2.19E+03	2.99E+03
N2	1.58E+03	1.35E+03	1.19E+03	3.45E+02	4.68E+02	8.84E+02	2.04E+03	3.70E+03	4.15E+03
N3	1.75E+03	1.71E+03	1.58E+03	8.00E+02	9.21E+02	1.38E+03	2.11E+03	2.63E+03	2.80E+03
N4	1.14E+03	7.54E+02	3.31E+02	1.14E+02	1.51E+02	4.00E+02	1.54E+03	3.53E+03	4.15E+03
N5	3.63E+03	3.83E+03	4.79E+03	1.13E+03	1.54E+03	2.98E+03	4.49E+03	4.89E+03	4.95E+03
N6	4.70E+02	3.49E+02	3.20E+02	9.89E+01	1.14E+02	2.36E+02	5.18E+02	1.27E+03	1.81E+03
t9	2.72E+02	2.50E+02	2.34E+02	4.00E+01	5.92E+01	1.54E+02	3.67E+02	5.42E+02	6.23E+02
db	2.43E+01	2.43E+01	8.67E+00	2.04E+00	3.60E+00	1.22E+01	3.62E+01	4.61E+01	4.73E+01
N5a	1.67E+02	7.89E+01	1.88E+01	1.19E+01	1.41E+01	3.54E+01	2.03E+02	7.03E+02	7.97E+02
t9b	1.79E+02	1.59E+02	1.05E+02	2.54E+01	3.73E+01	9.58E+01	2.39E+02	3.96E+02	4.51E+02
N4a	7.81E+02	8.49E+02	9.69E+02	2.47E+02	3.48E+02	6.77E+02	9.44E+02	9.89E+02	9.96E+02
t10	1.88E+02	1.72E+02	1.43E+02	6.01E+01	7.08E+01	1.25E+02	2.32E+02	3.51E+02	3.94E+02
t11	3.06E+02	2.88E+02	2.68E+02	8.08E+01	1.13E+02	1.99E+02	3.83E+02	5.63E+02	6.76E+02
N1a	6.52E+02	7.17E+02	9.82E+02	8.10E+01	1.42E+02	4.61E+02	8.84E+02	9.78E+02	9.86E+02
t12	5.52E+02	5.47E+02	4.37E+02	1.46E+02	1.84E+02	3.72E+02	7.38E+02	9.30E+02	9.53E+02
µmic_1	1.17E-04	1.11E-04	1.00E-04	1.00E-04	1.00E-04	1.04E-04	1.22E-04	1.57E-04	1.72E-04
pmic_1	2.86E-01	2.95E-01	3.00E-01	2.21E-01	2.42E-01	2.80E-01	3.00E-01	3.00E-01	3.00E-01
snimic_1	4.00E-07	6.18E-08	1.00E-08	1.02E-08	1.08E-08	2.16E-08	2.59E-07	1.95E-06	2.99E-06
Composite									
N1(u+sni)_1	5.69E-02	9.04E-03	9.04E-03	9.04E-03	9.04E-03	9.04E-03	9.04E-03	9.04E-03	1.05E-01
N2(u+sni)_1	3.58E-01	6.49E-03	6.49E-03	6.49E-03	6.49E-03	6.49E-03	8.36E-02	1.74E+00	1.74E+00
N3(u+sni)_1	1.30E-01	1.99E-03	1.99E-03	1.99E-03	1.99E-03	1.99E-03	1.38E-02	7.29E-01	7.29E-01
N4(u+sni)_1	1.56E+00	2.03E+00	2.03E+00	3.97E-03	3.97E-03	1.55E+00	2.03E+00	2.03E+00	2.03E+00
N5(u+sni)_1	2.09E+00	2.20E+00	2.20E+00	3.75E-03	9.07E-01	2.20E+00	2.20E+00	2.20E+00	2.20E+00
N6(u+sni)_1	1.98E-02	1.53E-03	1.53E-03	1.53E-03	1.53E-03	1.53E-03	1.53E-03	1.53E-03	1.53E-03
t9(u+sni)_1	8.58E-02	6.26E-02	6.35E-04	6.35E-04	6.35E-04	6.35E-04	1.74E-01	1.74E-01	1.74E-01
db(u+sni)_1	1.13E-03	1.04E-04	1.04E-04	1.04E-04	1.04E-04	1.04E-04	1.04E-04	9.59E-03	2.05E-02
N5a(u+sni)_1	1.55E-01	6.24E-02	1.15E-03	1.15E-03	1.15E-03	3.59E-03	3.58E-01	4.03E-01	4.03E-01
t9b(u+sni)_1	2.83E-02	1.42E-03	6.20E-04	6.20E-04	6.20E-04	6.27E-04	1.73E-02	1.97E-01	2.11E-01
N4a(u+sni)_1	2.85E-02	1.55E-03	1.55E-03	1.55E-03	1.55E-03	1.55E-03	2.05E-03	3.35E-01	3.82E-01
t10(u+sni)_1	3.22E-02	1.10E-02	1.10E-02	1.10E-02	1.10E-02	1.10E-02	1.30E-02	2.10E-01	2.51E-01
t11(u+sni)_1	2.62E-01	2.92E-01	2.92E-01	2.85E-02	2.85E-02	2.92E-01	2.92E-01	2.92E-01	2.92E-01
N1a(u+sni)_1	8.40E-02	1.06E-03	1.06E-03	1.06E-03	1.06E-03	1.06E-03	8.16E-03	4.18E-01	4.18E-01
t12(u+sni)_1	3.11E-01	3.75E-01	3.75E-01	4.41E-02	4.66E-02	3.13E-01	3.75E-01	3.75E-01	3.75E-01
Scaled									
N1/Mean(N)	1.16E+00	4.30E-01	3.11E-02	3.11E-02	3.11E-02	5.64E-02	2.46E+00	3.43E+00	3.43E+00
N2/Mean(N)	4.70E-01	3.07E-02	3.07E-02	3.07E-02	3.07E-02	3.07E-02	3.07E-02	3.27E+00	3.27E+00
N3/Mean(N)	2.35E-01	4.39E-03	4.39E-03	4.39E-03	4.39E-03	4.39E-03	4.39E-03	2.88E+00	2.88E+00
N4/Mean(N)	1.49E+00	1.61E-01	2.25E-02	2.25E-02	2.25E-02	2.25E-02	3.37E+00	3.37E+00	3.37E+00
N5/Mean(N)	1.27E-01	8.40E-03	8.40E-03	8.40E-03	8.40E-03	8.40E-03	8.82E-03	2.95E-01	1.83E+00
N6/Mean(N)	5.35E-02	6.82E-03	6.82E-03	6.82E-03	6.82E-03	6.82E-03	6.82E-03	6.82E-03	6.82E-03
t9/Mean(N)	2.58E-01	2.27E-03	2.24E-03	2.24E-03	2.24E-03	2.24E-03	7.18E-01	7.18E-01	7.18E-01
db/Mean(N)	6.09E-03	5.00E-04	5.00E-04	5.00E-04	5.00E-04	5.00E-04	1.05E-03	4.72E-02	7.57E-02
N5a/Mean(N)	1.11E+00	1.56E+00	1.56E+00	3.96E-03	3.96E-03	1.38E-01	1.56E+00	1.56E+00	1.56E+00
t9b/Mean(N)	4.25E-01	6.97E-01	6.97E-01	2.27E-03	2.27E-03	2.27E-03	6.97E-01	6.97E-01	6.97E-01
N4a/Mean(N)	1.54E-01	3.73E-03	3.73E-03	3.73E-03	3.73E-03	3.73E-03	3.73E-03	1.57E+00	1.57E+00
t10/Mean(N)	2.06E-01	3.38E-02	3.20E-02	3.20E-02	3.20E-02	3.20E-02	2.78E-01	7.87E-01	7.87E-01
t11/Mean(N)	8.10E-01	1.08E+00	1.08E+00	7.54E-02	7.54E-02	5.39E-01	1.08E+00	1.08E+00	1.08E+00
N1a/Mean(N)	8.43E-01	1.28E+00	1.28E+00	4.68E-03	4.68E-03	1.13E-02	1.28E+00	1.28E+00	1.28E+00
t12/Mean(N)	8.18E-01	1.07E+00	1.27E+00	9.18E-02	9.18E-02	2.37E-01	1.27E+00	1.27E+00	1.27E+00



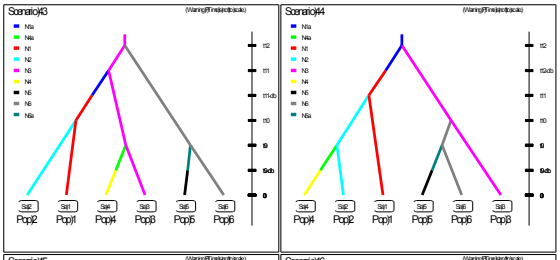
Scenario 20 - could not be drawn by DIYABC v. 2.0.3

Scenario 21 - could not be drawn by DIYABC v.2.0.3

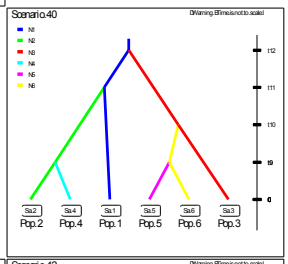


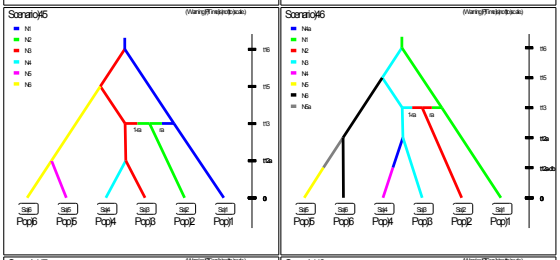
Scenario.37

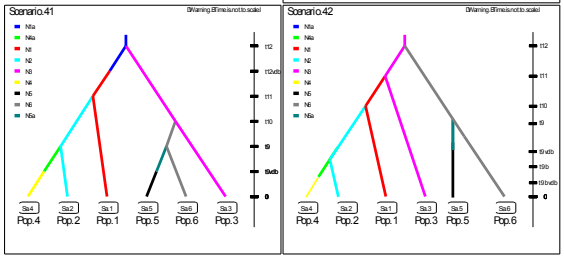
Scenario.38.v.Cannot.be.drawn.by.DIYABC.v..2.0.3

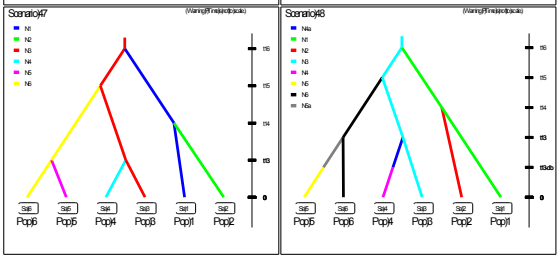


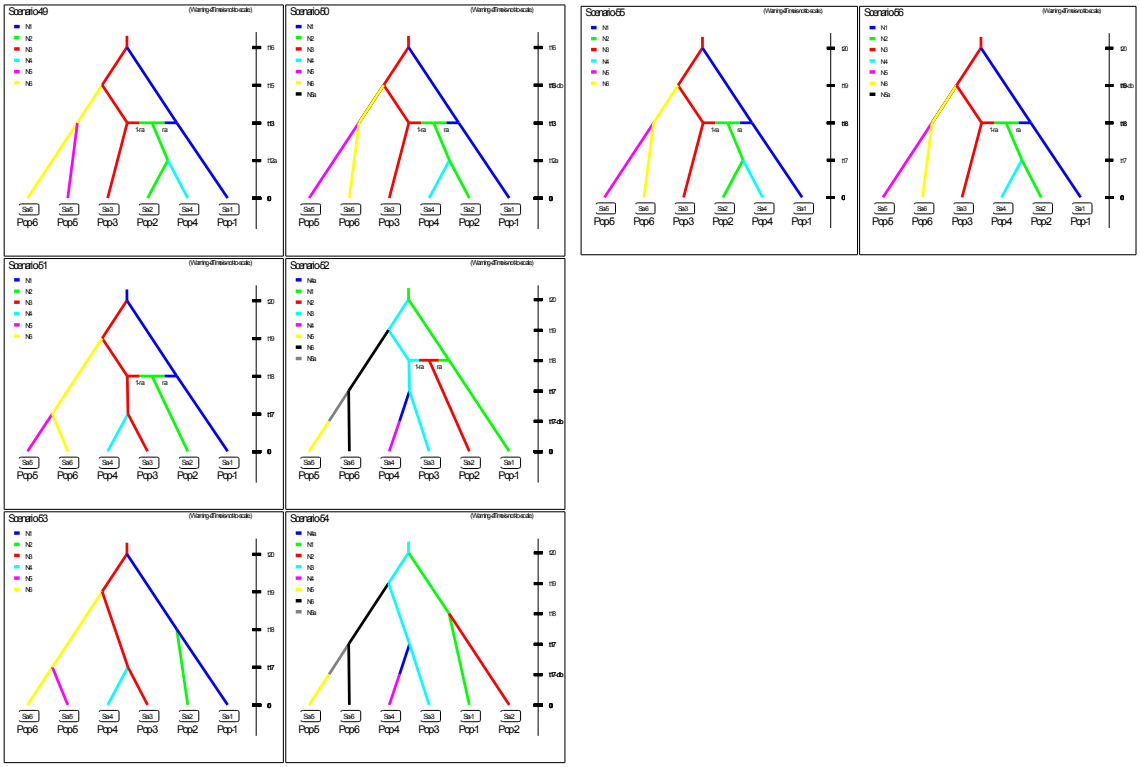
Scenario.39.v.Cannot.be.drawn.by.DIYABC.v..2.0.3



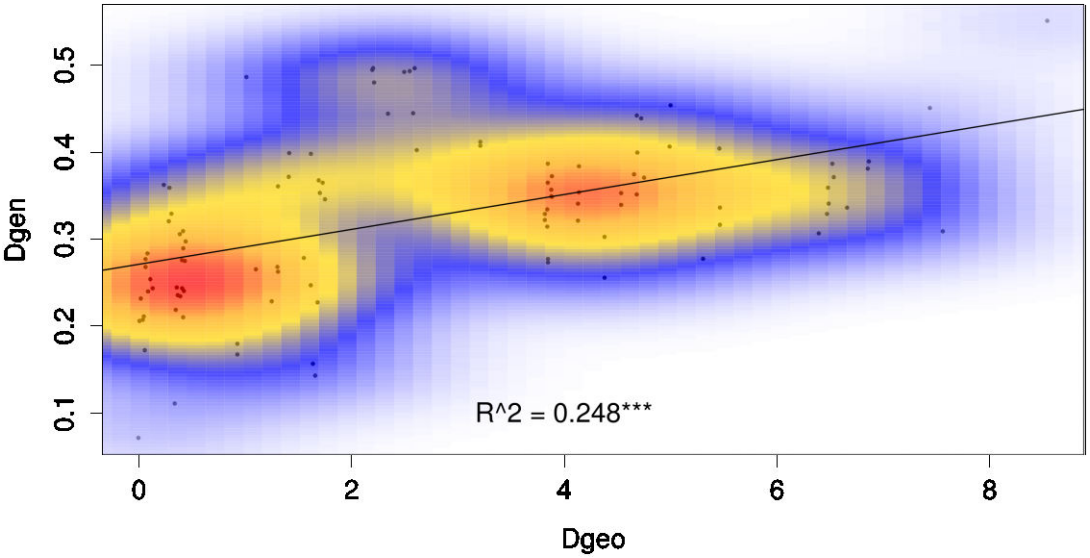




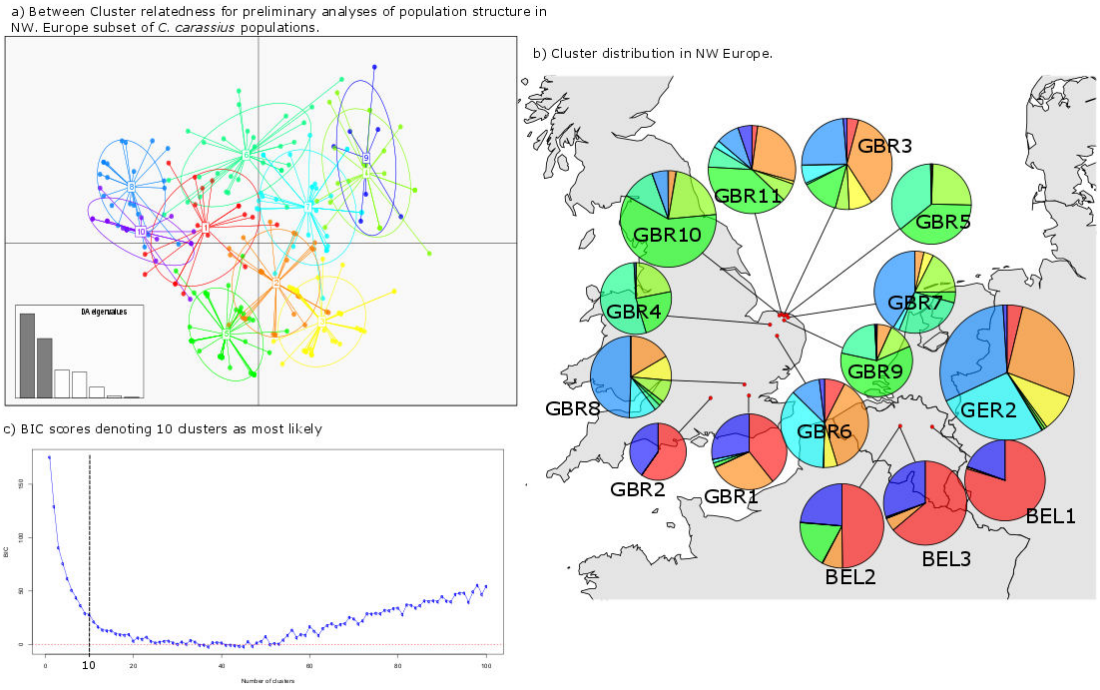




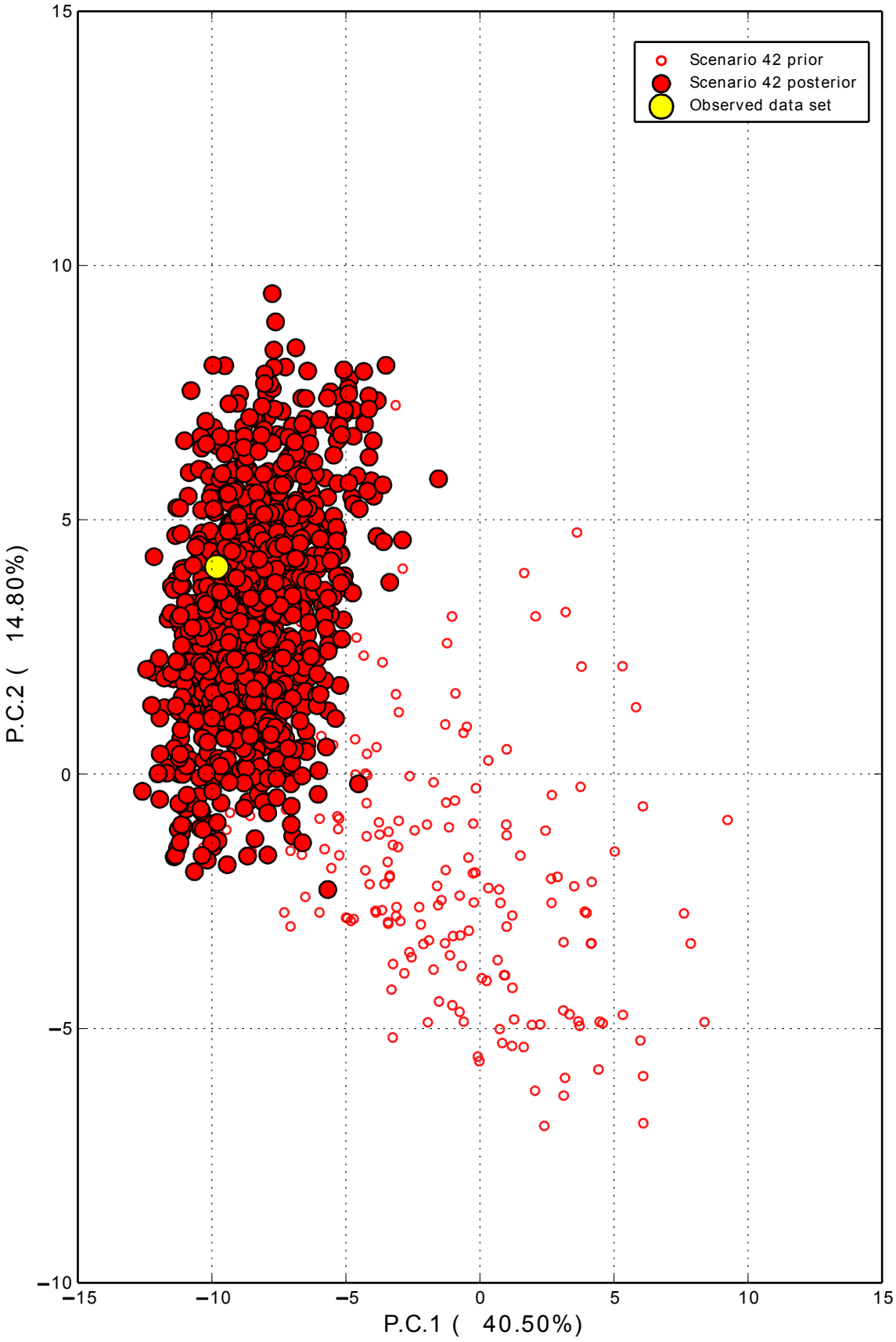
Supplementary Figure 4.1. All scenarios tested in DIYABC analysis. Pop1 = UK Pool 1, Pop2 = UK Pool 2, Pop3 = GER, Pop4 = UK pool 4, Pop5 = UK Pool 3, Pop6 = BELG. For the user-defined prior parameter distributions see S.Table 2. Back to text.



Supplementary Figure 4.2. Heat map showing isolation by distance in the 15 populations sampled. Colours represent the density of points on the plot (blue = low, red = high). Back to text.



Supplementary Figure 4.3. DAPC analysis of English, German and Belgian C, carassius populations. a) Shows relatedness between inferred clusters, b) shows geographic distribution of those clusters within populations and c) gives the BIC scores denoting 10 clusters as the most likely (the number of clusters after which no significant change in BIC score is observed). Back to text.



Supplementary Figure 4.4. The results of Model Checking of the most likely scenario identified in DIYABC. Note that Observed dataset lies well within the cloud of the predictive posterior parameter distribution. Back to text.

Chapter 5 Prolific hybridisation but no evidence for introgression between the native crucian carp, (*Carassius carassius* L.) and three highly invasive non-native species in Europe.

Abstract

Non-native species can impose significant detrimental impacts on native species, and two important mechanisms by which this can occur are hybridisation and subsequent introgression. *C. carassius* is a European freshwater cyprinid species, which is thought to be threatened by hybridisation and introgression from three non-native species, *Carassius auratus auratus* (L.), *Carassius auratus gibelio* (Bloch) or *Cyprinus carpio* (L.). In this study, 72 populations (1352 fish) were screened for hybridisation and introgression between *C. carassius* and one of the three non-native species, using six species diagnostic microsatellites, and thousands of genome-wide single nucleotide polymorphism (SNP) loci, identified using Restriction Site Associated sequencing (RADseq). A subset of fish identified in microsatellite and RADseq data as belonging to each of the four parental species were then used to test for signatures of introgression in between these species. The results reveal that F1 hybrids occur, often in high numbers, in 86% of populations where *C. carassius* is sympatric with diploid sexual forms of the non-native species, with one population containing triploid parthenogenetic *C. gibelio*. F2 generation hybrids were identified in microsatellite analyses in four populations. Despite the prevalence of F1 hybrids, advanced generation hybrids (F2, backcrosses) were rare, with three populations showing evidence of F2 hybrids in microsatellite data, and only one population found to contain backcrossed individuals between *C. carassius* and *C. gibelio*. In line with this result, no evidence was found for introgression having played a significant role in the evolution of the species in this study. Based on these findings, the threat posed to *C. carassius* by introgression is low and it is therefore suggested that conservationists concentrate their research on the direct ecological impacts of these non-native species on *C. carassius*, for which there is presently very little data.

Introduction

It is widely recognised that non-native species can impose dramatic detrimental impacts on native species (Mooney & Cleland 2001; Clavero & García-Berthou 2005), which can, in many cases lead to their decline or even extinction (for example Worthington & Lowe-McConnell 1994; Ricciardi *et al.* 1998). One often-cited mechanism by which this can occur is hybridisation, (Rhymer & Simberloff 1996; Mooney & Cleland 2001; Perry *et al.* 2002; Muhlfeld *et al.* 2009) which is a common occurrence between native and closely-related non-native species (Perry *et al.* 2002), and can lead to the decline or extinction of parental species in a number of ways. Firstly heterospecific reproduction can produce vigorous hybrids, which can subsequently out-compete the native species (Arnold & Hodges 1995; Facon *et al.* 2005). Secondly, the wasted reproductive resources that are committed to the hybrid offspring can reduce the number of pure species offspring produced and, in turn, lead to reduced reproductive the decline of the parental species (Rhymer & Simberloff 1996). And lastly, where hybrids are not sterile, backcrossing between hybrids and a parental species can lead to introgression, which in turn may result in outbreeding depression and loss of locally adapted genotypes or the transfer of beneficial locally adapted genes to the non-native species (Rhymer & Simberloff 1996; Mooney & Cleland 2001; Perry *et al.* 2002; Muhlfeld *et al.* 2009). Introgression is hypothetically likely in the case of anthropogenically introduced species, which would otherwise have been allopatric, as reinforcement of barriers to gene flow has not occurred between them (Arnold 1996).

Approaches to the detection of hybridisation in its initial F1 stage are well established, and this can often be accomplished through the use of meristic characters, or a relatively small number of molecular markers. However, the unambiguous identification of hybrid classes beyond this initial F1 generation and the detection introgression can be problematic (Boecklen and Howard 1997; Currat et al. 2008) and requires the application of a large set of genomic markers. Recent advances in high-throughput sequencing technology and approaches such as Restriction Site Associated DNA sequencing (RADseq, described in detail in Chapter 1 of this thesis) are revolutionising the study of hybridisation and introgression (Twyford & Ennos 2011). The resulting datasets can consist of tens of thousands of homologous, genome wide single nucleotide polymorphisms (SNPs) and provide an unprecedented ability to identify specific regions of the genome that have introgressed between species. For example, Hohenlohe et al. (2011, 2013) identified over 3000 species diagnostic SNPs in *Onchorhynchus mykiss*

and *Onchorhynchus clarkii* and used these to identify extensive admixture between these species, and to identify loci that show particularly high admixture rates, indicative of adaptive introgression. The methodology employed in (Hohenlohe *et al.* 2013) represents one of several methods used to identify introgression, whereby diagnostic markers for species were identified in populations known, *a priori*, to contain only pure parental species. These markers were then used to look for admixture between these species in a hybrid zone. Other approaches to the identification of introgression with genome scale data aim to build models for the demographic history of populations to test the null hypothesis of isolation with no gene flow between them. When the observed data do not fit this model, these approaches allow for the estimation of magnitude and direction of gene flow (Durand *et al.* 2011; Harris & Nielsen 2013; Sousa & Hey 2013). One example of this approach is *Treemix* (Pickrell & Pritchard 2012), which uses a graph-based approach to construct null models of isolation without gene flow in the form of bifurcating phylogenetic trees. Migration events can then be added to this model to test for the presence of gene flow between populations.

The crucian carp, *Carassius carassius* (L.), which is native to most of central and northern Europe is a freshwater cyprinid species often found in spatially restricted waters such as ponds, lakes and slow moving river backwaters (Holopainen *et al.* 1997). *C. carassius* population declines have recently been observed in many regions throughout its range, and have been attributed to several human-mediated factors including habitat loss, drought and acidification (Holopainen & Oikari 1992; Navodaru *et al.* 2002; Sayer *et al.* 2011). However, two often-cited drivers of *C. carassius* declines are hybridisation and subsequent introgression with three non-native species; the goldfish, *Carassius auratus auratus* (L.); the gibel carp, *Carassius auratus gibelio* (Bloch); and the common carp, *Cyprinus carpio* (L.) (Hänfling *et al.* 2005; Papoušek *et al.* 2008; Sayer *et al.* 2011; Mezhzherin *et al.* 2012; Wouters *et al.* 2012), all of which are among the top 25 most important non-native freshwater fish in Europe (Savini *et al.* 2010). These three non-native species are now introduced or invasive across much of the *C. carassius* native range (Savini *et al.* 2010) to the extent that, in Ukraine, Hungary, Czech Republic and likely many other countries in Europe, it is difficult to find *C. carassius* populations without one or more non-native species (*pers. comms.* Müller Tamás, András Weiperth, Prof. Sergey Mezhzherin). Where this is the case, hybridisation between *C. carassius* and a non-native species is commonly observed (Hänfling *et al.* 2005; Papoušek *et al.* 2008; Sayer *et al.* 2011; Mezhzherin *et al.* 2012;

Wouters *et al.* 2012) and has been purported, by a number of studies, to result in the extirpation of pure *C. carassius* in some populations (Hänfling *et al.* 2005; Papoušek *et al.* 2008; Sayer *et al.* 2011). For example, pure *C. carassius* are often in decline or absent in ponds containing *C. carassius* x *C. auratus* hybrids (Hänfling *et al.* 2005; Mezhzherin *et al.* 2012), suggesting that hybrids are vigorous and can outcompete pure *C. carassius*.

Where there is hybridisation, there is also the possibility for introgression, which has been highlighted as a potential threat to the *C. carassius* gene pool (Hänfling *et al.* 2005; Smartt 2007). The process of introgression is primarily mediated by backcrossing, however, in the present system, previous studies have shown that, although F1 hybrids are very common, further hybrid generations only occur at very low frequencies. For example, Hänfling et al. (2005) found that, in 42 sampled populations, 38% contained F1 hybrids between *C. carassius* and either *C. auratus* or *C. carpio*, however only 4 individuals, from one population, were identified as being backcrosses between *C. carassius* and *C. carassius* x *C. auratus* hybrids. In line with this result, Mezhzherin et al. (2012) found only 2 samples that could be putatively assigned as backcrosses between *C. carassius* and *C. auratus*, from a total of 1638 fish from 36 locations. Furthermore, Smartt (2007) found that in controlled crossing experiments, F1s could be readily produced between *C. carassius* and *C. auratus*, but failed to produce F2 generation hybrids or backcrosses. Existing evidence, therefore, predicts that between *C. carassius* and *C. auratus*, backcrossing and subsequent introgression is rare. Similar results have been found in studies assessing hybridisation between *C. carassius* and *C. gibelio* or *C. carpio*; hybrids are common where species are sympatric, but hybridisation past the initial F1 stage is rare (Hänfling *et al.* 2005; Papoušek *et al.* 2008; Wouters *et al.* 2012). In many of these studies, however, the number of samples (Smartt 2007; Papoušek *et al.* 2008; Wouters *et al.* 2012) or species-diagnostic markers (Wouters *et al.* 2012) were low. Although microsatellite markers allow for good discrimination between F1s and simple backcross classes, upwards of 70 markers would be required in order to discriminate between pure parental species and complex backcross generations (i.e. those beyond the first one or two backcrossing events) (Boecklen & Howard 1997). This raises the possibility that advanced generation backcrosses have gone undetected in these previous studies and that backcrossing does occur more frequently than so far observed.

In the current chapter I revisit the question of whether backcrossing results in introgression between these species. Firstly, a traditional microsatellite approach is employed to comprehensively assess the levels of hybridisation and backcrossing among *C. carassius* and the three non-native species studied here. Secondly RADseq is used to test for genomic regions that have introgressed between native and non-native species. Finally, SNP loci that are fixed between species are identified and reported in order to add to the genetic resources available for the discrimination between the species of this study and the conservation of the threatened *C. carassius*.

Methods

Sample collection and DNA extraction

There were three main objectives of our sampling regime. Firstly to collect baseline data for samples belonging to each of the four species studied here (*C. carassius*, *C. auratus*, *C. gibelio* and *C. carpio*) we sampled individuals that had been putatively identified as such on morphological grounds in the field. Secondly, we collected samples from populations which had been morphologically identified as containing hybrids between *C. carassius* and any of the three invasive species. As it was known that hybridisation was occurring in these populations, they were seen to be good candidates for containing backcrossed individuals if they existed. And lastly, we devoted the large majority of the sampling to fish morphologically identified as pure *C. carassius*, in order to screen for any cryptic introgression resulting from hybridisation past the F1 stage.

This sampling regime resulted in 1353 fish from 72 populations across 14 counties (Table 5.1), which contained individuals morphologically identified as *C. carassius*, *C. auratus*, *C. gibelio* or *C. carpio*. However, a number of these samples were collected in the context of other studies and prior to the start of this project with an uncertain *a priori* identification of hybrids. Therefore, the exact number of samples morphologically identified as belonging to each species or hybrid category is, unfortunately, unavailable.

Table 5.1. All samples used in this study. Numbers of samples identified for each species or hybrid class are based on the final assignment of individuals in NewHybrids analyses. Where a population contains samples genotyped at both microsatellite and RADseq loci, final assignment is based on RADseq NewHybrids results and the number of samples assigned to each species and hybrid class are given in the form of microsatellite/RADseq.

Code	Non-native species present	Location	Country	Drainage	Coordinates		Microsa-tellites	RAD	Molecular assignment										
					Lat	Long			*CA*	*AU*	*GI*	*CY*	*CAxAU*	*CAxAU F2*	*CAxGI*	*CAxGI F2*	*CAxCY*	*AUxCY*	*CAx(CAxGI)*
BEL1		Bokrijk	Belgium	Scheldt River	51	5.41	14		14										
BEL2		Meer van Weerde	Belgium	Scheldt River	51	4.48	12		12										
BEL3		Meer van Weerde	Belgium	Scheldt River	51	4.48	9		9										
BEL4		Bocholt	Belgium	Muese River	51.2	5.56	4		4										
BEL5	GI	Dendermonde	Belgium	Scheldt River	51	4.09	na	5			5								
CZE1		Lužnice	Czech Rep.	Danube River	48.9	14.89	9		9										
DEN1		Copenhagan	Denmark	Baltic Sea	17.8	60.21	12	10	12										
DEN2		Pederstrup	Denmark	Baltic Sea	12.6	55.77	14	8	14										
DEN3		Bornholm Island	Denmark	Baltic Sea	14.9	55.17	6	5	6										
DEN3		Gammel Holte	Denmark	Baltic Sea	12.5	56	14		14										
EST1	GI	Tartu	Estonia	Baltic Sea	58.4	26.72	5		5										
EST2	GI	Vehendi	Estonia	Baltic Sea	58.4	26.72	6		6										
FIN1		Joensuu	Finland	Baltic Sea	62.7	29.68	38		38										
FIN2	GI	Helsinki	Finland	Baltic Sea	60.4	25.33	39	6	33		1				5				
FIN3		Jyväskylä	Finland	Baltic Sea	62.3	25.76	39	10	39										
FIN4		Oulu	Finland	Baltic Sea	65	25.47	9	8	9										
FIN5	GI (3n)	Salo	Finalnd	Baltic Sea	60.4	23.1	10		10										
GBR1		London	U.K.	U.K	51.5	0.13	30		30										
GBR10	AU/CY	Norfolk	U.K.	U.K	52.9	1.1	45	15	21				22	1				1	
GBR11		Norfolk	U.K.	U.K	52.9	1.16	20		20										
GBR13	AU/CY	Norfolk	U.K.	North Sea	52.9	1.15	7		4				3						
GBR14	AU	Epping Forest	U.K.	Thames	51.7	0.04	60	2	6	16			35	2			1		
GBR15		Buntingford	U.K.	Thames	51.9	-0.01	na	5	5										
GBR16	AU	Epping Forest	U.K.	Thames	51.7	0.03	37	6	26	2			8	1					
GBR17		Hull	Ornamental-unknown origin		53.8	-0.36	na	5	4									1	
GBR18	AU/CY	Nofolk	U.K.	U.K.	52.8	1.08	5					1					4		
GBR2		Reading	U.K.	U.K	51.5	-0.97	4		4										
GBR3		Norfolk	U.K.	U.K	52.9	1.16	47		47										
GBR4		Norfolk	U.K.	U.K	52.8	0.75	29	9	29										
GBR5	CY	Norfolk	U.K.	U.K	52.8	0.76	60		49								11		
GBR6	CY	Norfolk	U.K.	U.K	52.5	0.93	39	13	29			7					3		
GBR7		Norfolk	U.K.	U.K	52.9	1.15	24	10	24										
GBR8		Hertfordshire	U.K.	U.K	52.9	1.1	39	9	39										

(Table 5.1 continued)

Code	Non-native species present	Location	Country	Drainage	Coordinates		Microsa-tellites	RAD	Molecular assignment										
					Lat	Long			*CA*	*AU*	*GI*	*CY*	*CAxAU*	*CAxAU F2*	*CAxGI*	*CAxGI F2*	*CAxCY*	*AUxCY*	*CAx(CAxGI)*
GBR9		Norfolk	U.K.	U.K	52.8	1.1	27		27										
GER1		Kruegersee	Germany	Elbe River	52.03	11.97	5		5										
GER2		Münster	Germany	Rhine River	51.9	7.56	25		25										
GER3		Bergheim	Germany	Danube River	48.7	11.03	8		8										
GER4		Bergheim	Germany	Danube River	48.7	11.03	9		9										
HUN1	GI	Gödöllő	Hungary	Danube River	19.4	47.61	9		2		1				6				
HUN2	GI	Vörösmocsár	Hungary	Danube River	19.2	46.49	10	6	10										
HUN3	GI	Ócsa:	Hungary	Danube River	47.5	19.05	9				7				2				
HUN4	GI	Lake Kolon	Hungary	Danube River	47.5	19.1	16	8	16										
HUN5		Tisza Canal	Hungary	Danube River	47.3	19.2	6		6										
NOR1		Oslo	Norway	North Sea	60.05	9.94	19		12		1				6				
NOR2		Tromsø	Norway	North Sea	19	69.65	16	9	16										
POL1		Sarnowo	Poland	Vistula River	52.9	19.36	38		38										
POL2	GI	Kikót-Wies	Poland	Vistula River	52.9	19.12	38		38										
POL3		Tupadly	Poland	Vistula River	52.7	19.3	20	10	20										
POL4		Orzysz	Poland	Vistula River	53.8	22.02	14	10	14										
RUS1		Proran Lake	Russia	Don River	47.5	40.47	10	9	10										
RUS4		Velikaya river	Russia	Baltic Sea	55.9	30.25	30		30										
SWE1		Gränbrydammen	Sweden	Baltic Sea	59.9	17.67	29		29										
SWE10		Norrköping	Sweden	Baltic Sea	58.6	16.27	29	9	29										
SWE11		Gotland Island	Sweden	Baltic Sea	57.9	18.79	11		11										
SWE12		Osterbybruk	Sweden	Baltic Sea	12.3	55.73	14	9	14										
SWE14		Stockholm	Sweden	Baltic Sea	19	59.66	16	9	16										
SWE18		Lake Krauke, Skane	Sweden	Baltic Sea	n	n	6		6										
SWE19		Lund	Sweden	Baltic Sea	n	n	31		31										
SWE2		Stordammen	Sweden	Baltic Sea	59.8	17.71	22	-	22										
SWE3		Östhammar	Sweden	Baltic Sea	60.3	18.38	30		30										
SWE4		Umeå	Sweden	Baltic Sea	63.7	20.41	10		10										
SWE5		Kvicksund	Sweden	Baltic Sea	59.5	16.32	10		10										
SWE6		Åland Island	Sweden	Baltic Sea	60.4	19.85	8		8										
SWE7		Grillby	Sweden	Baltic Sea	59.6	17.37	10		10										
SWE8		Skabersjo	Sweden	Baltic Sea	55.6	13.15	20	10	20										
SWE9		Märsta	Sweden	Baltic Sea	59.6	17.8	31	10	31										
SWE17	GI	Gotland Island	Sweden	Baltic Sea	57.5	18.14	11		5		1				5				
SWE20	GI	-	-	-	-	-	15	4			4				7	4*			3
TL		High Wycombe	U.K.	U.K.	51.6	-0.74	20		20										
UKR1		Ternopil	Ukraine	Dniester	49.2	25.56	8	3			8								
UKR2		Reut River, Floresti	Ukraine	Dniester	47.8	28.42	na	5			5								
BLS		Ochkino	Belarus	Dnieper	52.5	30.52	7		7										
							1333	**237**	**1166**	**18**	**33**	**8**	**68**	**4**	**31**	**0**	**19**	**2**	**3**

* conflict between assignment in microsatellite and RADseq data. See text.

For all samples, approximately 1cm^2 of tissue was taken from the lower caudal fin and immediately placed in 95% ethanol for storage at -20°C. Samples collected specifically for this study by the author were anaesthetised using a 1 mL L^{-1} anaesthetic bath containing 2-phenoxyethanol, and tissue samples were taken from the lower caudal fin and immediately placed in 95% ethanol for storage at -20°C. Wounds were then treated with adhesive powder (Orahesive) and antibiotic (Cicatrin) (Moore *et al.* 1990) to prevent infection. DNA extraction was performed from approximately 2-4mm^2 of tissue using the Qiagen DNeasy extraction kit (Qiagen, Hilden, Germany).

Microsatellite amplification and scoring

In order to identify samples as one of the four species or their hybrids 1333 out of the total 1353 samples were genotyped at six species diagnostic microsatellite loci (GF1, GF17, GF29 (Zheng *et al.* 1995), MFW2 (Crooijmans *et al.* 1997), J7 (Yue & Orban 2002) and Ca07 (Yue & Orban 2004)), which were originally developed for use in either *C. auratus* (GF1, GF17, GF29, Ca07), *C. a. gibleio* (J7) or *C. carpio* (MFW2). The diagnostic properties of these loci have been established in previous studies (Maes *et al.* 2003; Hänfling *et al.* 2005). GF1, GF29, J7 and MFW2 are diagnostic for all four species, whereas GF17 and Ca07 are diagnostic between all three *Carassius* species, but do not amplify consistently in *C. carpio*, and were therefore removed from *C. carpio*-specific analyses. Loci were optimised for use in a single multiplex PCR reaction, performed using Qiagen multiplex PCR mix in 10 µl volumes, with manufacturer's recommended reagent concentrations, including Q solution and 1 µl of template DNA (see multiplex 1 in Supplementary table 3.1 in Chapter 3 for PCR specifications). PCR reactions were run on an Applied Biosciences® Veriti Thermal Cycler and microsatellite fragment lengths were analysed on a Beckman Coulter CEQ 8000 genome analyser using a 400 bp size standard. Microsatellite fragment lengths were analysed and alleles scored using the Beckman Coulter CEQ8000 software.

RADseq library preparation and data processing

RADseq was performed for 237 fish samples from 32 populations (Table 5.1). 217 of these samples were included in the 1333 samples genotyped at microsatellite loci. The remaining 20 samples included 5 *C. auratus* samples, identified as such in (Hänfling *et*

al. 2005)(GBP), 5 samples morphologically identified as ornamental *C. auratus* (GF) and 10 samples identified as *C. gibelio* in microsatellite analyses prior to this study (DEND, MY-5, (Maes *et al.* 2003)). To confirm microsatellite identification of hybrids, we devoted 15 and 6 RADseq samples to individuals identified by microsatellite analysis (see results) as being *C. carassius* x *C. auratus* hybrids and *C. carassius* x *C. gibelio* hybrids respectively. Although microsatellite analyses identified several samples which belong to advanced hybrid classes (e.g. F2 hybrid, see results), low DNA quality unfortunately precluded the use of most of these samples for RADseq. However three samples (SWE20_7, SWE20_8, SWE20_11) which were found to have high probability to belong to the F2 hybrid class in microsatellite analyses, had sufficient DNA quality for RADseq and were included. The remaining 183 RADseq samples were comprised of fish identified as pure crucian on morphological grounds and preliminary microsatellite analyses. Of these putatively pure *C. carassius,* 47 were from populations known to currently or historically contain hybrids (Table 5.1). These samples were used, firstly, to confirm that no cryptic hybridisation beyond the initial F1 generation has taken place that was not picked up during initial morphological or microsatellite identification, and secondly, to test for introgression between *C. carassius* (see below) and any of the invasive species in this study.

To ensure high quality RADseq library preparations, DNA was quantified using the Quant-iT™ PicoGreen® dsDNA Assay kit (Invitrogen) and normalised to concentrations greater than 50 ng ml^{-1}. Gel electrophoresis was then used to check that DNA extractions contained high molecular weight (i.e. low-fragmentation) DNA. Samples were then prepared in 13 RADseq libraries at Edinburgh Genomics (University of Edinburgh, UK), using the enzyme *Sbf*1, according to the protocol in Davey et al. (2012). Libraries were sequenced on 5 lanes of 2 Illumina HiSeq 2000 flowcells (Edinburgh Genomics). Libraries 1-8 were sequenced using the V3 Illumina chemistry, and libraries 9-13 were sequenced with the V4 chemistry.

RADseq raw data was processed according to the methods detailed in chapter 2. Briefly, data was quality assessed using FastQC (Andrews 2010), filtered for PCR duplicates and aligned to the *C. carpio* reference genome (Xu *et al.* 2014). Only reads which aligned uniquely with a maximum of six mismatches per read were retained for subsequent processing in the reference guided STACKS pipeline (Catchen *et al.* 2013). This pipeline was used to cluster reads into loci based on their mapping locations to the

reference genome and simultaneously call SNPs at loci with a read depth greater than 8 reads. Initially, this pipeline was run using all species and all populations, where only loci present in 70% of all populations were retained and only one SNP per RADtag was retained, to reduce confounding effects of linkage between SNPs. The resulting dataset was used for preliminary species delimitation using PCA (see below). However, due to the amounts of divergence between species (Rylkova et al. 2013), many loci did not meet the filtering criteria of being present in 70% of all populations, when the data was processed in its entirety. This is likely due to some loci containing more than four SNPs between the *C. carassius* and the *C. carpio* reference genome, insertions or deletions (which are not processed by STACKS) or mutations in the restriction site (Chapter 2). Therefore, once samples had been preliminarily assigned to species on the basis of the initial PCA, the reference guided STACKS pipeline was run separately for each pairwise combination of species (*C. carassius x C. auratus, C. carassius x C. gibelio, C. carassius x C. carpio*), again filtering for loci present in at least 70% of individuals in all populations in each subset and retaining only one SNP per RAD tag. This approached maximised the number of loci that were available between species pairs.

Species delimitation and identification of ongoing hybridisation

In order to identify samples as pure species or hybrid, genotypic class assignment was performed in Newhybrids (Anderson & Thompson 2002) for both microsatellite and RADseq data separately. Newhybrids uses genotype frequency data and Bayesian computation to calculate the posterior probability that an individual belongs to one of two species, or one of several user-specified hybrid classes between them, for example F1, F2, or backcrosses, and requires that pairs of species be analysed together. The first step of this analyses was, therefore, to putatively group samples into species and hybrid classes. To do this we performed Principle Component Analysis (PCA) separately for microsatellites and the full RADseq SNP dataset using the Adegenet package (Jombart & Ahmed 2011) in R v3.2.0 (R Core Team 2015). Based on the clustering of samples in the PCA, individuals were assigned to either a single parental species or as a hybrid between two particular species. Once assigned, samples were grouped into subsets, each containing two species and the hybrids between them. However, in Chapter 3 of this thesis, 2.26 (95% CI = 1.30-3.22) million years of divergence was observed between northern European *C. carassius* populations and those in the Danube and Don river catchments. As this amount of divergence is analogous to species level divergence, it

could potentially confound the Newhybrids analyses, therefore, the Danubian and Don *C. carassius* samples were analysed separately. As no *C. carpio* samples were obtained in the Danube region, *C. carpio* samples from northern European population GBR6 were used as baseline data in the Danube-specific analyses. For samples from the Don river catchment, no samples were obtained for *C. auratus* spp. or *C. carpio* species, therefore northern European samples from populations BEL5, GBR15 and GBR6 were used as baseline data for *C. gibelio*, *C. auratus* and *C. carpio* respectively.

Each pairwise species subset was analysed in NewHybrids for both microsatellites and RADseq data. NewHybrids was used to calculate the probability that individuals belonged to either parental species (Par1, Par2), or one of several hybrid classes; F1, F2, Backcross 1 (Par1xF1), Backcross 2 (Par2xF2). For the RADseq data, we extended the hybrid classes tested to include second generation backcrosses (Par1 x Backcross 1, Par2 x Backcross 2) and F3 generation hybrids (F1 x F2), since the high number of loci provides more power when discriminating between complex hybrid categories. Where a sample was found to have a posterior assignment probability of greater than zero for more than one genotype class, it was assigned to the class for which the posterior probability was highest. For samples that were represented by both microsatellite loci and RADseq data, if the two datasets disagreed on their assignment, then the result from the RADseq data was taken to be correct, as a higher number of loci in was expected to produce the more accurate assignments (Boecklen & Howard 1997).

Testing for introgression between native and invasive species.

A genome wide SNP dataset obtained from RADseq was used to test for introgression between the native *C. carassius* and the invasive species in this study. To acquire the data for this analysis, we used the species assignments from the above NewHybrids analysis to group all samples identified as pure invasive species into three species pools: *C. auratus* (n=10)*, C. gibelio* (n = 10)*, and C. carpio* (n = 2). The 183 *C. carassius* samples were kept in their 18 separate populations to allow for the identification of population-specific introgression events. The raw RADseq data for these samples were then re-processed in the reference guided STACKS pipeline, and SNPs were retained only if they were present in 70% of individuals in all three invasive species pools and at least 70% of individuals in each *C. carassius* population. This approach, again, maximised the number of SNPs available for this analysis. Individuals identified as

belonging to a hybrid class in NewHybrids were not included in this analysis as such high amounts of gene flow (i.e. 50% in F1 hybrids) would confound the Treemix approach used (Pickrell & Pritchard 2012). The resulting genome wide SNP dataset was analysed in TreeMix (Pickrell & Pritchard 2012), which uses a graph based approach to infer the relationships between the populations analysed. Treemix first constructs a bifurcating tree and then infers gene flow events on the basis of the residual covariance matrix of the bifurcating model. Here we followed the approach employed in Pickrell and Pritchard (2012) and Decker et al. (2014) whereby we first constructed a bifurcating tree in TreeMix, specifying no historic migration events, and assessed the proportion of the variation in the data explained by this tree model, using custom R scripts distributed with the Treemix software (Pickrell & Pritchard 2012). We then sequentially added migration events to this model, and at each point, again assessed the variation in the data explained. The significance of each specific migration event itself was calculated using a jackknifing approach implemented in Treemix (Pickrell & Pritchard 2012), to test whether it significantly improved the overall fit of the model. The tree model that explained the most variation in the data was retained as the most likely tree to explain the demographic history of the samples in this study.

Identification of species diagnostic RAD tag loci

Finally the RADseq dataset was filtered for loci that are fixed between the species used in this study in order to provide genomic resources for the identification of species and hybrids in future studies. RADseq samples were pooled into 4 species specific groups based on the PCA and Newhybrids analyses, and hybrids were removed from the dataset. The populations module of STACKS was then run separately for each of the 6 pairwise combinations of species pools to identify SNPs between each species pair and to calculate locus-specific F_{ST} values. SNP loci were retained if they were present in at least 70% of individuals in each species pool and, unlike STACKS analysis for PCA and NewHybrids datasets, all SNPs in a tag were retained. A custom python script was then used to identify SNP loci which were fixed between species. These SNP loci were then outputted in variant call format (VCF) files, containing individual genotypes for the samples in this study and the alignment position for each locus on the Xu et al. (2014) *C. carpio* draft genome assembly.

Results

Species delimitation and identification of hybridisation

In the microsatellite dataset, all loci displayed highly diagnostic alleles or allele ranges between *C. carassius* and at least one of the invasive species (Table 5.2). In preliminary analyses, *C. carassius* samples from the Danube river catchment (V, GEW, GODO1, HK1 and HK2) and Don (PRO) were found to contain alleles (J7-202, J7-204, GF29-213, GF29-215) that were previously thought to be specific to *C. auratus* (Maes *et al.* 2003; Hänfling *et al.* 2005). This resulted in the false assignment of Danubian individuals as *C. gibelio* or *C. gibelio* hybrids in the NewHybrids analysis (described below). These loci were therefore removed in Danubian populations for the final analyses. One population, FIN5, was found to contain triploid *C. gibelio*, and, interestingly, a single triploid hybrid, which possessed two *C. carassius* alleles and one *C. gibelio* allele at each locus, was found in population SWED20. As these triploid individuals could not be analysed in PCA or NewHybrids with diploid individuals, they were removed from further analyses.

Table 5.2. Diagnostic properties of 6 microsatellite loci

Locus	Size range				Diagnostic properties					
	CA	AU	GI	CY	CA x AU	CA x GI	CA x CY	AU x GI	AU x CY	GI x CY
GF1	299	301-314	301-337	297	A	A	A	R	A	A
GF17	182, 186	192-214	190-214	-	R	R	-	F	-	-
GF29	213-223	191-207	191-215	254-282	R(ND)F	R(ND)F	RF	RF	R	R
MFW2	161	157	157	239-267	A	A	A	N	A	A
J7	202-(220-228*)	202, 204	202-212	208	R(ND)	R(ND)	R(ND)	RF	F	F
Ca07	122-140	130-136	126-155	-	RF	RF	-	RF	-	-

CA = C. carassius, AU = *C. auratus, GI* = *C. gibelio*, CY = *C. carpio*
A = Diagnostic allele, R = Diagnostic range, F = Diagnostic Frequency, N= Not diagnostic, ND = Not diagnostic in Danube,
* = Northern European crucian allele range

The initial PCA for this microsatellite dataset was effective at discriminating between northern European *C. carassius* and all three invasive species. However, the Danubian samples overlapped to some extent with *C. carassius* x *C. carpio* hybrids in the PCA analyses (Figure 5.1). PC1 captured the variation between *C. carassius* and *C. auratus spp.* (explaining 9.64% of variation in the data) and PC2 captured the variation (4.90%) between the *Carassius* genus and *C. carpio* (Table 5.3). Principal components 3 and 4 explained the variation between the two lineages of *C. carassius* (3.3%) and between *C. auratus* and *C. gibelio* (2.98%) respectively (Supplementary Figure 5.1). However, this

initial PCA showed that the loci used here had low discriminatory power between *C. auratus* and *C. gibelio* (see PC1, Table 5.3 and PC4, Supplementary Figure 5.1). Samples were, therefore, grouped into only 2 species-pair subsets for microsatellite NewHybrids analyses; the first contained *C. carassius*, *C. carpio* samples and hybrids between them (n = 1116, n loci = 4). The second subset contained *C. carassius*, both *C. auratus* spp. and hybrids between them (n = 1299, n loci = 6). However after the analysis of both microsatellite and RADseq datasets it was possible to confidently assign *C. auratus* spp. samples to individual species based firstly on the fact that *C. gibelio* has, to date, never been observed in the UK, secondly on *a priori* knowledge of sampling locations and morphological identifications in the field, and lastly through the cross checking between microsatellite and RADseq datasets for individuals that were genotyped in both.

Out of the total 1333 fish samples genotyped using microsatellites, NewHybrids identified 1166 as *C. carassius*, 18 as *C. auratus*, 33 as *C. gibelio*, 68 *C. carassius* x *C. auratus*, 31 as *C. carassius* x *C. gibelio*, 8 as *C. carpio*, 19 as *C. carassius* x *C. carpio*, 4 as *C. carassius* x *C. gibelio* F2 hybrids and 4 as *C. carassius* x *C. auratus* F2 hybrids (Figure 5.2, Table 5.1). In eight of the above samples (GBR14_26, GBR14_33, GBR14_37, GBR16_21, GBR18_2, SWE20_7, SWE20_14, SWE20_15), these assignments were ambiguous, whereby more than one genotype class had high assignment probabilities (Figure 5.2). In all of these samples, the ambiguity existed between different hybrid categories, i.e. between F1 and F2 hybrid classes, or F1 and Backcross classes. These NewHybrids results generally agreed very well with the clustering of samples in the PCA (see colours in Table 5.3). However, there were some exceptions, for example, several individuals identified as F2 or backcrosses in NewHybrids analysis, clustered close to samples identified as being F1 hybrids in the PCA (see labels in Table 5.3). Also, individuals SWE17_6 and GBR14_8, which were identified as pure *C. gibelio* and *C. auratus* respectively, clustered close to F1 hybrids, however they were on the periphery of this group (Table 5.3). Importantly, hybrids were found in 82% of populations where *C. carassius* and non-native species were found in sympatry, excluding population FIN5, in which *C. carassius* was found with triploid, asexually reproducing *C. gibelio*.

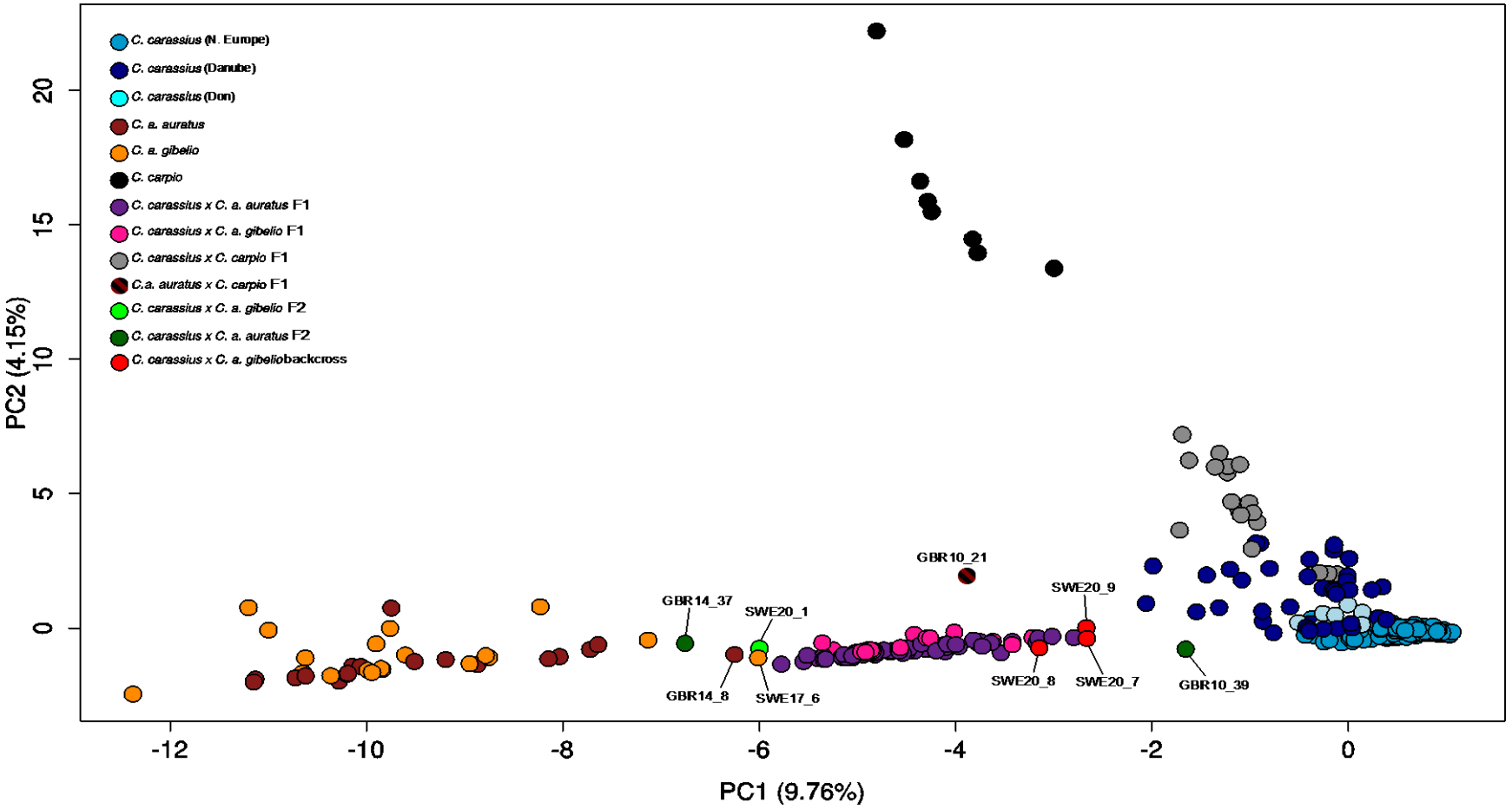


Figure 5.1. Principle components 1 and 2 for 1333 samples genotyped at species diagnostic microsatellite loci. PC1 captures the variation between *C. carassius* and the two *C. auratus* spp, whereas PC2 captures the variation between the Carassius and Cyprinus genera. Colours represent the final consensus assignments of individuals to species or hybrid class based on Newhybrids analyses of both microsatellite and RADseq datasets.

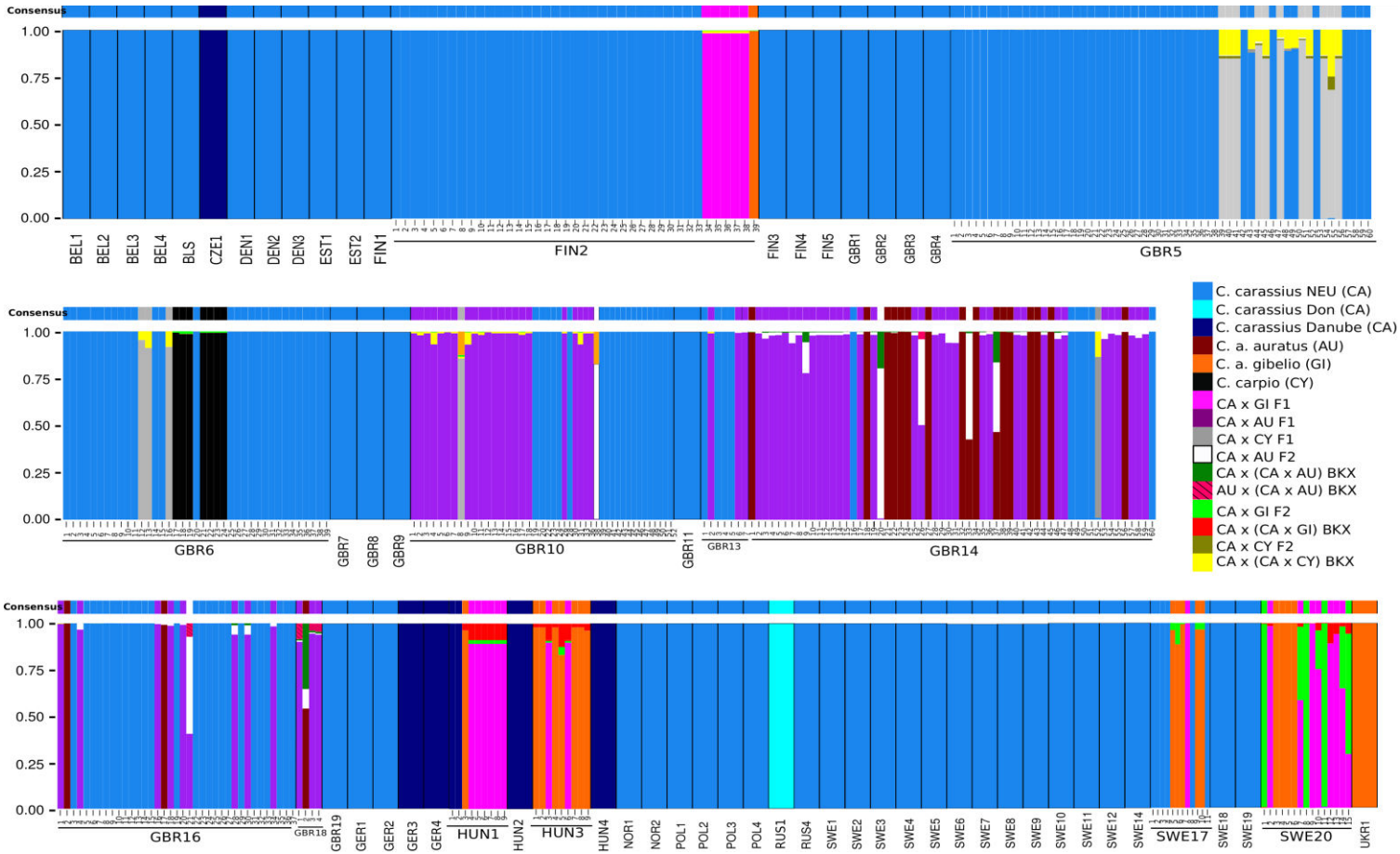


Figure 5.2. Genotypic class assignments for *C. carassius* x *C. auratus* spp. and *C. carassius* x *C. gibelio* species pair Newhybrids analysis using species diagnostic microsatellite loci. For each sample (column), the segments of the stacked bar represent the posterior probability that an individual belongs to the corresponding genotypic class. The consensus assignment based on the microsatellite data alone is shown by the coloured boxes above each panel.

PCA of the 1296 SNP RADseq dataset, representing all species in the present study, was again very effective at discriminating between species and hybrids. PC1 identified the variation between the *Carassius* and *C. carpio*, (47.56%, Figure 5.3), PC2 explained the variation between the two major lineages within *C. carassius* (identified in Chapter 3 of this thesis, 13.23%, Supplementary Figure 5.2), PC3 captured the variation between *C. carassius* and both *C. auratus* spp. (9.12%, Figure 5.3), PC4 explained the variation between *C. carassius* in the Don river catchment and all other *C. carassius* populations (3.13%, Supplementary Figure 5.2), and PC5 identified the variation between *C. gibelio* and *C. auratus* (2.52%, Figure 5.3). On the basis of these results, individuals were grouped into 3 species-pair subsets, *C. carassius* x *C. auratus* (n =197, n loci =17146), *C. carassius* x *C. gibelio* (n = 205, n loci =17383) and *C. carassius* x *C. carpio* (n = 192, n loci = 7783), for use in the NewHybrids analyses. However, as in the microsatellite PCA, clustering of samples on principal components did not allow for unambiguous identification of specific hybrid classes.

Analyses of the RADseq species pair subsets, in NewHybrids, identified 191 *C. carassius*, 9 *C. auratus*, 16 *C. gibelio*, 2 *C. carpio*, 15 *C. carassius* x *C. auratus*, 16 *C. carassius* x *C. gibelio*, 3 *C. carassius* x *C. carpio* and 3 *C. carassius* x *(C. carassius* x *C. gibelio)* backcrosses (Figure 5.4b). Sample assignments in NewHybrids were highly unambiguous, with all samples having posterior assignment probabilities of >0.99 to a single species or hybrid class (Figure 5.4). Interestingly, the three backcrossed individuals, SWE20_7, SWE20_8 and SWE20_11 were found in the same populations as the triploid *C. carassius* x *C.* gibelio hybrid.

In the 217 samples that were genotyped at both microsatellite loci and using RADseq, the assignment of individuals to species or hybrid class was identical in both datasets except for the three backcrossed samples, SWE20_7, SWE20_8 and SWE20_11 (Figure 5.4). In the microsatellite analysis SWE20_8 and SWE20_11 were identified as being F2 generation *C. carassius* x *C. gibelio* hybrids and SWE20_7 was ambiguously identified as an F1 *C. carassius* x *C. gibelio* hybrid, with a strong probability of assignment to the F2 hybrid class. Whereas in the RADseq analysis all three of these samples were unambiguously identified as *C. carassius* (*C. carassius* x *C. a.gibelio*) backcrosses. As the RADseq data was assumed to be more likely to be correct, due to its greater number of loci, and, thus, increased assignment power (Boecklen & Howard

1997), the final assignment for these samples was to the backcross class *C. carassius* x (*C. carassius* x *C. gibelio*). NewHybrids did not identify any individuals belonging to a hybrid class beyond the first generation of backcrossing.

Based on the final assignments of all individuals from both microsatellite and RADseq data, of the 18 populations putatively identified as containing *C. carassius* and non-native species, 14 were found to contain hybrids (Table 5.1). Therefore, discounting FIN5 which contained only triploid *C. gibelio* (which are therefore unable to sexually reproduce with *C. carassius*), hybridisation was observed in 82% of populations where *C. carassius* was found in sympatry with sexual non-native species (Table 5.1).

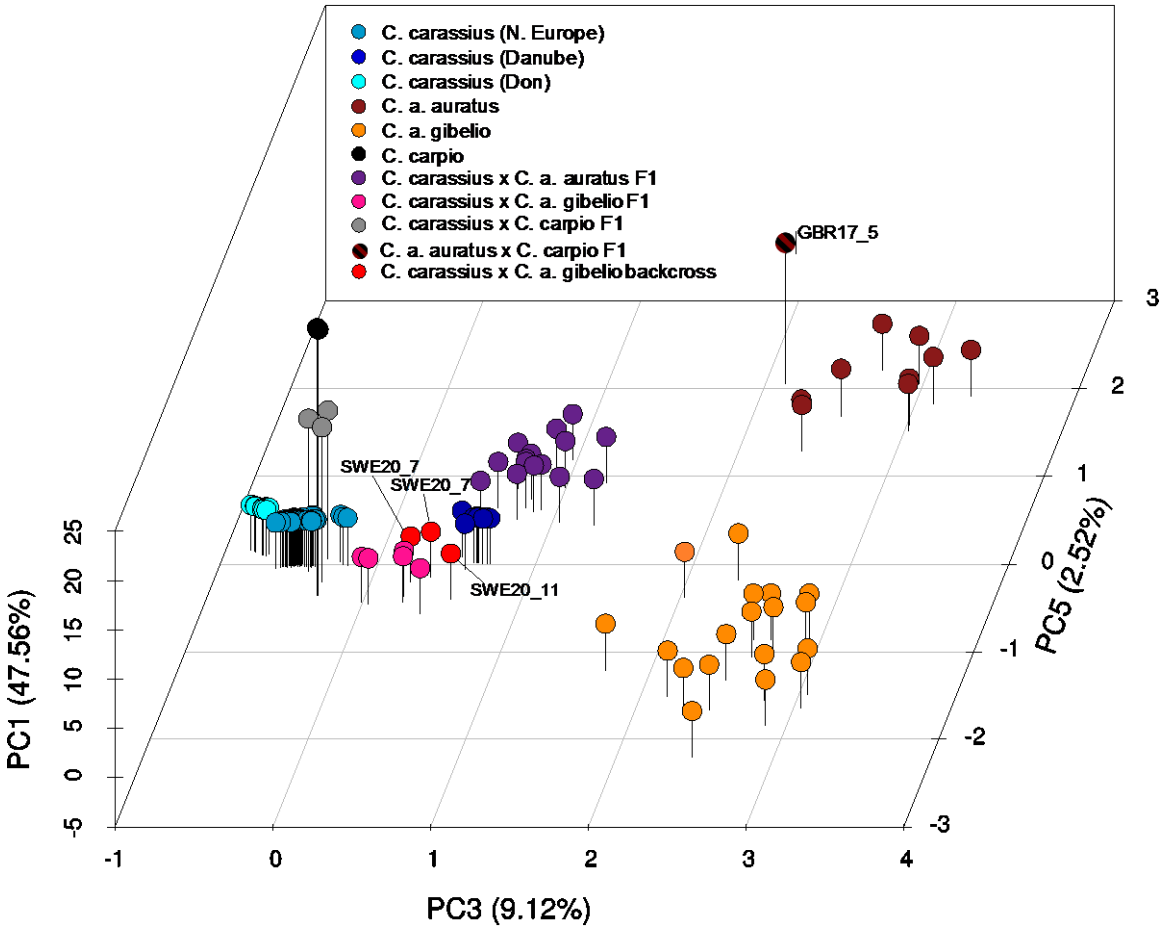


Figure 5.3. Principle components 1, 3 and 5 for the entire RADseq dataset of 247 genotyped individuals. PC1 captures the variation between the *Carassius* and *C. carpio* genera, PC3 captures the variation between *C. carassius* and both *C. auratus* and *C. gibelio*. and PC5 explains the variation between *C. auratus* and *C. gibelio*. Colours represent the final consensus assignments of individuals to species or hybrid class based on Newhybrids analyses of both microsatellite and RADseq datasets.

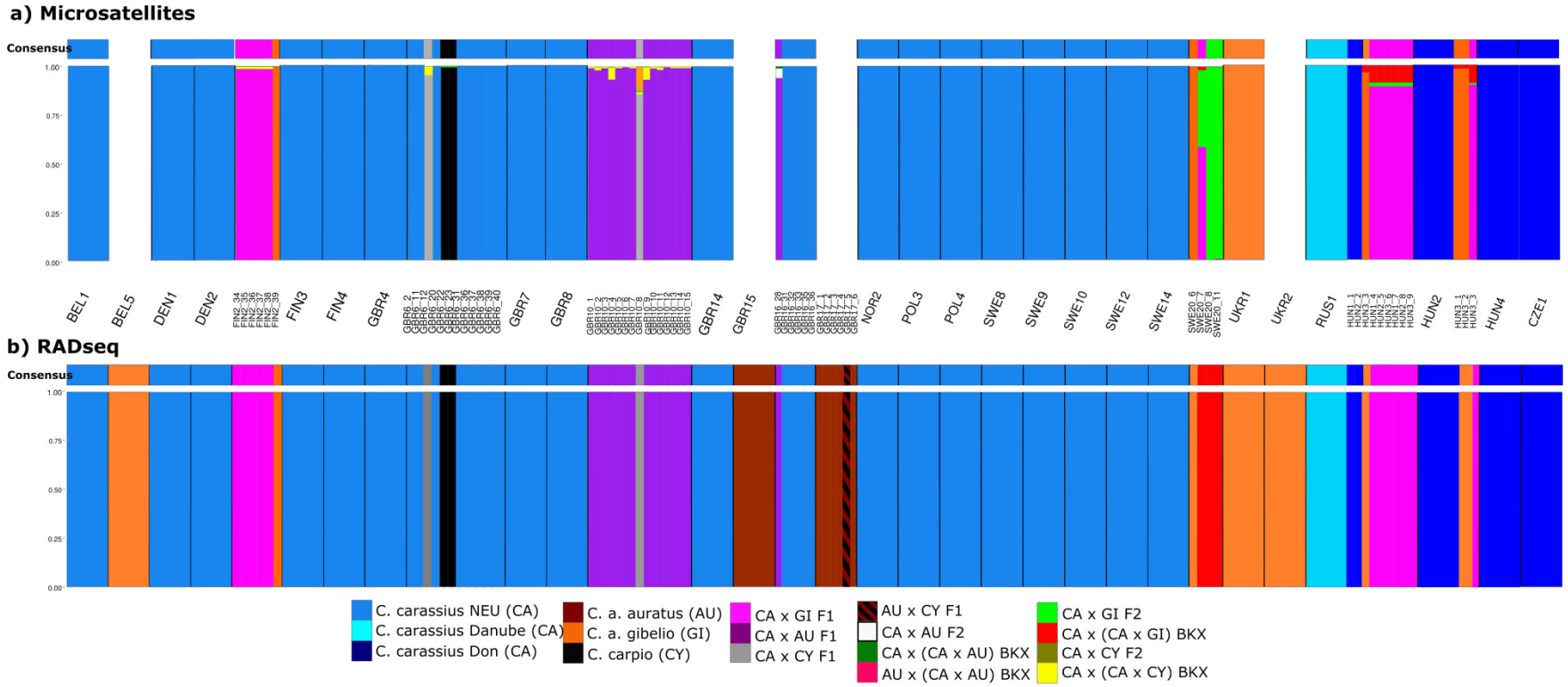


Figure 5.4. Genotypic class assignments for *C. carassius* x *C. auratus* spp. and *C. carassius* x *C. carpio* species pair Newhybrids analysis using RADseq data. For each sample, the segments of the stacked bar represent the posterior probability that an individual belongs to the corresponding genotypic class. The consensus assignment based on the microsatellite data alone is shown by the coloured boxes above each panel.

Testing for signatures of introgression

STACKS analysis of pooled species yielded a dataset of 4494 SNPs present in at least 70% of all invasive species pools and *C. carassius* populations. The analysis of this dataset in TreeMix showed no evidence of introgression between any of the species examined here, with the initial bifurcating tree model explaining a larger proportion of the variance in population relatedness (99.983%, Figure 5.5) than models containing migration events. Sequentially adding one to five migration events to the population tree only resulted in a decrease in the explanatory power of the model (Figure 5.5b) and P-values for specific migration edges were non-significant in all cases.

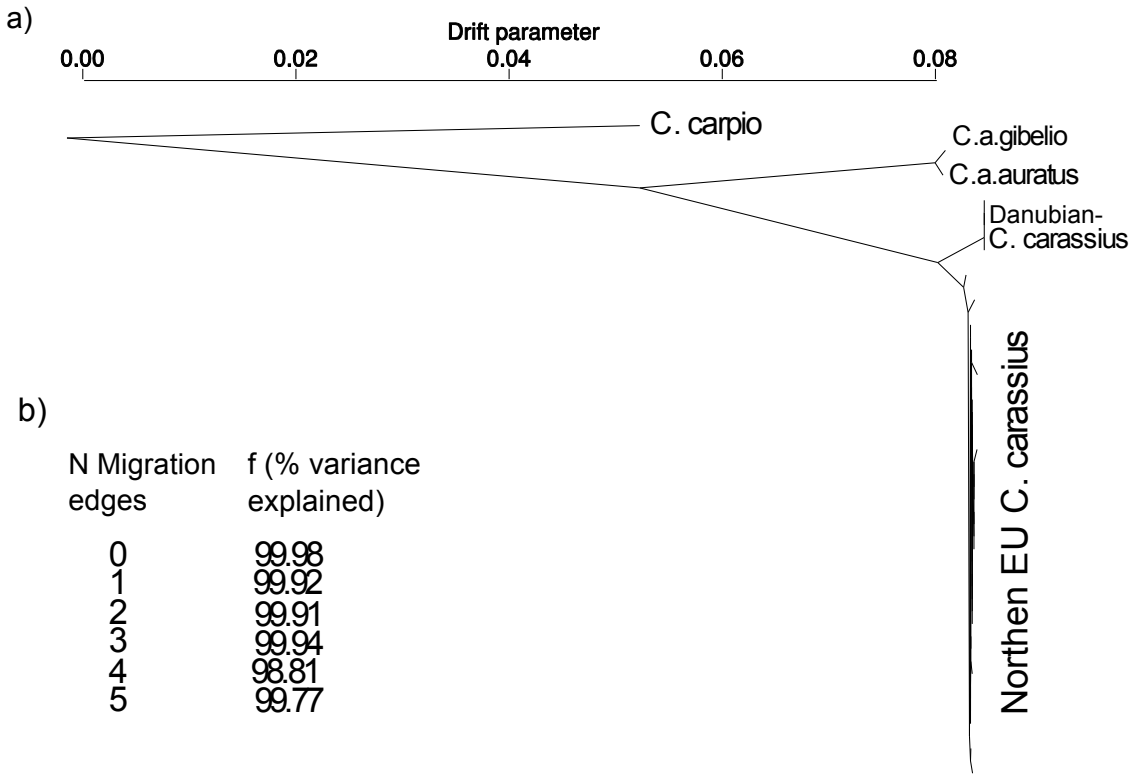


Figure 5.5. a) Bifurcating Maximum likelihood tree with no modelled migration events, which is the most likely Treemix model to explain the demographic history of the populations and species used in this study. b) as migration edges are added to the model, the percentage of variation it explains (f) is reduced.

Identification of species diagnostic loci between European carp species

The number of SNPs in the STACKS analyses of the 6 species-pair datasets ranged from 21607 - 37193 SNPs in 6396 - 7332 RAD tag loci, with an average of 4.4 SNPs per RAD tag (Table 5.3). When filtering these datasets for SNPs fixed between species, one species pair comparison, *C. auratus* and *C. gibelio*, showed a very low number (4)

of diagnostic SNPs. However, filtering the datasets of the remaining species pairs identified between 7813-23551 SNPs that were fixed between them (Table 5.3). VCF files containing these loci and their position on the (2014) *C. carpio* reference genome can be found on GitHub at https://github.com/DanJeffries/DLJeffries_Thesis_data_and_scripts.

Table 5.3. Fixed species diagnostic SNP loci identified between species pairs

	SNPs identified	**RADtags identified**	**N SNPs/ N tags**	**Fixed SNPs**
CA x AU	24753	6396	3.87	7813
CA x GI	30241	6767	4.47	7995
CA x CY	35924	7332	4.90	23551
AU x CY	32654	6625	4.93	19359
GI x CY	37193	7164	5.19	21481
AU x GI	21607	7042	3.07	4

CA = C. carassius, AU = *C. auratus, GI* = *C. gibelio*, CY = *C. carpio*

Discussion

Prevalent hybridisation between C. carassius and non-native species

Microsatellite data and NewHybrids analysis revealed that hybridisation occurred in 82% of populations where *C. carassius* and non-native species were found in sympatry. However, this is a minimum estimate, as our sampling of each population was not exhaustive, and hybrids may have been present but unsampled in the 3 remaining populations. This result corroborates those of previous studies (Hänfling *et al.* 2005; 2007; Mezhzherin *et al.* 2012), adding to the consensus that hybridisation between *C. carassius* and closely related non-natives is almost certain to occur where they are found in sympatry together. If the inferences of (Hänfling *et al.* 2005; 2007; Mezhzherin *et al.* 2012). (2005) are correct, and hybrids can indeed impose strong negative impacts on *C. carassius* populations, then these levels of hybridisation could prove to be a significant threat to this already endangered species. There are several examples of hybrid vigour contributing to the extirpation of native species (Rosenfield *et al.* 2004), however, current evidence for this occurring in *C. carassius* is anecdotal. There is, therefore, a clear need for studies that explicitly test for the ecological impacts of hybrids on *C. carassius* and the mechanisms by which they occur. It should be noted that, for populations which were *a priori* thought to contain hybrids, samples were

chosen for molecular analyses in such a way as to ensure that data were obtained from both pure and hybrid individuals. Therefore, the frequency of hybridisation within populations found in this study is likely not to be indicative of the true frequencies in nature.

Microsatellite and RADseq screening of 54 populations that were *a priori* thought to contain only pure *C. carassius*, did not identify any non-native species, hybrids or cryptic introgression, suggesting that the morphological identification carried out in the field was highly accurate and the probability of mis-characterising *C. carassius* populations as containing no hybridisation is low. Such a result bodes well for the management of *C. carassius* populations where it is not practical to undertake molecular analyses on the scale of this study.

Rare backcrossing and no evidence for further introgression

In the present study we identify only three diploid backcrossed individuals from one population (SWE20) out the 11 in which *C. carassius* were sympatric with a non-native species. A similarly low frequency of backcrossing was found by Hänfling *et al.* (2005), who identified only four individuals as diploid backcrosses between *C. carassius* and *C. auratus* (CA x (CA x AU)) in the UK. This rarity of backcrossed individuals and also of F2 generation hybrids suggests that the fertility of *C. carassius* x *C. auratus* spp. F1 hybrids is low. This is consistent with the findings of (Smartt 2007), who found that, in experimental crosses of *C. carasssius* and *C. auratus*, F1 hybrids in some cases produced no eggs, and when eggs were produced they failed to develop. However, given that some backcrosses do exist, F1 hybrids must be capable of producing viable eggs in rare cases. Although these backcrossed individuals are rare, it has been shown that introgression can occur between species despite very low frequency of hybridisation (Goodman *et al.* 1999; Barton 2001), thus introgression past the first backcrossing stage between *C. carassius* and *C. auratus* spp. may indeed be possible.

Despite the presence of backcrosses, Treemix analysis of 4494 genome wide SNPs identified no evidence of introgression beyond the initial backcross stage between any of the four species in this study. This result may be attributable to one or a combination of mechanisms which are acting to prevent gene flow beyond the backcross generation. Firstly, the rarity of backcrosses may, in itself, be enough to prevent introgression

between species. Secondly, despite being diploid, backcrosses may be sterile or have low fertility. Or thirdly, backcrosses or offspring beyond the initial backcross stage may suffer from outbreeding depression, and therefore have low reproductive success. It is possible, however, that introgression does occur but that the approaches used in this study were unable to detect it. For example, introgression can, in some cases, occur at the scale of a few genes (for example see The Heliconius Genome Consortium 2012). The genome of *C. carassius* is thought to be approximately 1.9GB (Xu *et al.* 2014), therefore, the 4494 SNP loci used in this study constitutes approximately one marker every 420kb (under the assumption that these markers are evenly spread throughout the genome). Thus, it is possible that small genomic regions have introgressed between these species but are not represented in our dataset. If such small introgressed regions existed, it is likely that they resulted from old introgression events, as the size of introgressed linkage groups is eroded over time be recombination (Twford and Ennos 2011). The amount of time required to reduce a linkage group to a size undetectable in the present study would be dependent on the size of the original introgressed region, and any adaptive forces acting upon it, both of which are unknown. If such ancient introgression existed, there will have been more time for these regions to have segregated throughout the species, making them more likely to be detected with a limited sample number. A logical next step would therefore be to use a more frequently cutting restriction enzyme for RADseq, for example Pst1, which will yield SNPs much more densely spaced across the genome than that of the present dataset.

One interesting result was the observation of a triploid hybrid between *C. carassius* and *C. gibelio*, in the same population as the three samples identified as diploid backcrosses. This triploid hybrid possessed 2 *C. carassius* alleles, and one *C. gibelio* allele at each microsatellite locus. Interestingly, Hänfling *et al. (2005)* found two polyploid hybrids which also co-occurred with diploid backcross individuals. It is unknown how such individuals arise; one possibility is that they result from the mating of a pure *C. carassius* individual with a pure *C. auratus* spp. individual, in which the *C. carassius* parent contributes an unreduced gamete. If this was true, the co-occurrence of triploid hybrids and backcrosses observed here and in Hänfling et al. *(2005)* would suggest that these triploid F1 hybrids have increased fertility and, in fact, facilitate backcrossing. However, if this were the case, it is expected that triploid *C. carassius* would occasionally be observed, yet, to date no such fish have been found. Given that meiosis is more likely to be disrupted in hybrids (Choleva *et al.* 2012) than in pure species, a

more probable explanation is that triploid hybrids are the result of a backcrossing event, whereby the unreduced egg was contributed by a *C. carassius* x *C. auratus* spp. F1 hybrid and the paternal contribution came from a pure *C. carassius* male. This process has previously been observed in diploid female F1 hybrids of *Corbitis taenia* and *Corbitis elongatoides*, which produce unreduced oocytes, and, during backcrossing with diploid males of either species, produce triploid backcross offspring (Janko *et al.* 2007; Choleva *et al.* 2012). The same has been seen in *Poeciliid* fish, whereby *Poecilia mexicana limantouri* x *Poecilia latipinna* F1 hybrids have been shown to produce diploid oocytes through automixis, which, when fertilised produce triploid offspring (Lampert *et al.* 2007).

In summary of the information above, it is hypothesised that F1 hybrids between *C. carassius* and *C. auratus* spp. observed here and in Hänfling *et al. (2005)* are often sterile but when they are fertile, they can produce either reduced or unreduced eggs. In both cases these eggs can be fertilised by *C. carassius* males leading to either diploid or triploid backcross offspring. It is expected that, although viable, the triploid backcross offspring identified here and in Hänfling *et al.* (2005) are not fertile, as was the case in *Corbitis* (Janko *et al.* 2007)*, Poeciliids (Lampert et al. 2007)* and in many other triploid fish (Vrijenhoek 1994). If this hypothesis is true, and a proportion of backcrossing events lead to triploid sterile offspring, this phenomenon may constitute a further barrier to introgression between *C. carassius* and *C. auratus* spp.

Identifying highly-diagnostic loci between native and non-native species

One of the major challenges to identifying hybridisation, backcrossing and introgression in the past has been the development of a high number of genome-wide markers. Prior to the advent of high throughput sequencing this was not possible unless the study was focused on a model organism (Twyford & Ennos 2011). In the present study RADseq data filtering has identified thousands of SNPs that are fixed between species for all but one of the species pair combinations. These loci provide a valuable genomic resource which will allow future studies to unequivocally identify hybrids or backcrosses between the species examined in this study, in a cost effective manner. Unlike morphological or microsatellite approaches, these loci will allow for the confident distinction between *C. carassius* hybrids with *C. auratus, C. gibelio* or *C. carpio*. Interestingly however, only 4 fixed SNPs were identified between *C. auratus* and *C.*

gibelio. Although this is surprisingly low, it likely reflects the small amount of divergence between the species within the *C. auratus* complex. This pattern may also be driven by an ascertainment bias in the RADseq data resulting from the mapping of raw reads to the *C. carpio* reference genome during STACKS analysis. Loci that have high sequence similarity between the species of this study are more likely to have a large proportion of their reads map to the reference genome, and thus be correctly assembled, in all species. Thus, RADseq loci that meet the STACKS filtering criteria of being present in the majority of individuals in the study are likely to represent conserved regions of the genome. It is therefore possible that the low divergence rates at these loci are emphasising the low general divergence between *C. auratus* and *C. gibelio*.

RADseq vs Microsatellites for the study of hybridisation and introgression

As expected, the high number of loci produced by RADseq allowed for much higher certainty for the assignment of samples to species or hybrid class than the microsatellite data. In all individuals in the RADseq analysis, posterior probabilities calculated in NewHybrids were > 0.99. In contrast there was much more ambiguity in the microsatellite analyses, and in some cases, multiple hybrid classes had high probability of assignment in the same individual. This was true for the three individuals identified in the SNP data as being *C. carassius* x (*C. carassius* x *C. gibelio*) backcrosses. NewHybrids analyses of the microsatellite data assigned SWE20 11 and 13 to the F2 hybrid class between *C. carassius* and *C. gibelio*, and SWE15 to the F1 hybrid class. However NewHybrids analyses of the RADseq data for these individuals calculated a posterior probability of 1.0 that they were in fact backcrosses. The disparity between these results and those of the microsatellite analyses is suggestive of a lack of power in the 6 microsatellite loci to discriminate between the genotype patterns expected under F2 and backcross scenarios. If this is the case then the samples in microsatellite dataset that were identified as F2 hybrids between *C. carassius* and *C. auratus*, may, instead, be backcrosses. Without genotyping these samples at additional loci this cannot be confirmed, and could, instead, be attributable to allele dropout at one or two loci in these individuals.

As the three individuals identified in the RADseq as backcrosses were identified as F2 hybrids in the microsatellite data set, there is the possibility that the other samples identified as F2 hybrids in microsatellite analyses (between *C. carassius* and *C. auratus*

in populations GBR10 and GBR14) may, in fact, also be backcrosses. Unfortunately, RADseq data was not available for these individuals so it is not possible to verify these identifications.

One challenge faced by any study aiming to identify diagnostic alleles between species, is that the geographic distribution of the genetic diversity can lead to loci being diagnostic between them in one region but not another (Amish *et al.* 2012). In the present study, this was seen at two microsatellite loci in particular (J7 and GF29), in which alleles that were diagnostic between northern European *C. carassius* lineages and non-native species were not diagnostic between them in the *C. carassius* lineage found in the Danubian catchment. These lineages are known to have diverged approximately 2.26 million years ago and it is likely that these alleles have been lost through the successive bottlenecks known to have occurred during the postglacial expansion of *C. carassius* into northern Europe (see Chapter 3). This finding not only highlights the importance of comprehensively sampling the diversity within each species when developing species diagnostic loci, but also it also identifies an advantage that RADseq possesses over microsatellites for studies such as this. As RADseq allows for the identification of thousands of SNP markers in non-model systems, the impact of the small proportion of loci that differ in their diagnostic power between geographic regions is likely to be less important.

Conclusions

The results of the present study confirm the high hybridisation rates between *C. carassius* and the three non-native species studied here, with hybridisation occurring in almost all populations where they are sympatric. In line with previous studies, backcrossing rates are low, which could be attributable to one or a number of possible post zygotic barriers to gene flow. This is further supported by the lack of evidence for any significant introgression between these species, which would require backcrossing to occur. Although the clear identification of backcrossed individuals in the present study suggests that there is the possibility of introgression, at least between *C. carassius* and *C. gibelio*, there is no evidence that introgression beyond the backcross generation to or from the three non-native species studied here is a major threat to *C. carassius*. Instead, it is suggested that conservationists should focus their attention on the ecological impacts imposed by the parental non-native species themselves, for example,

through assessment of the potential increased fitness in F1 hybrids and the reproductive burden of hybridisation on *C. carassius*.

Chapter 5. Supplementary materials

Supplementary table 5.1. Allele frequencies in all species at all 6 microsatellite loci.

	C.carassius	C.a.auratus	C.a.gibelio	C.carpio
GF1.297	0.000	0.000	0.000	1.000
GF1.299	1.000	0.028	0.000	0.000
GF1.301	0.000	0.167	0.542	0.000
GF1.305	0.000	0.000	0.083	0.000
GF1.307	0.000	0.222	0.208	0.000
GF1.311	0.000	0.000	0.063	0.000
GF1.312	0.000	0.389	0.000	0.000
GF1.314	0.000	0.194	0.000	0.000
GF1.321	0.000	0.000	0.021	0.000
GF1.325	0.000	0.000	0.063	0.000
GF1.334	0.000	0.000	0.000	0.000
GF1.337	0.000	0.000	0.021	0.000
GF17.182	0.993	0.000	0.167	-
GF17.186	0.006	0.000	0.000	-
GF17.190	0.000	0.000	0.042	-
GF17.192	0.000	0.222	0.063	-
GF17.194	0.000	0.000	0.271	-
GF17.196	0.000	0.250	0.000	-
GF17.200	0.000	0.000	0.042	-
GF17.202	0.000	0.000	0.104	-
GF17.204	0.000	0.056	0.000	-
GF17.208	0.000	0.000	0.000	-
GF17.212	0.000	0.028	0.250	-
GF17.214	0.000	0.167	0.063	-
GF29.191	0.015	0.250	0.042	0.000
GF29.195	0.007	0.139	0.021	0.000
GF29.199	0.011	0.361	0.229	0.000
GF29.207	0.002	0.083	0.021	0.000
GF29.209	0.000	0.000	0.042	0.000
GF29.213	0.019	0.000	0.021	0.000
GF29.215	0.018	0.000	0.021	0.000
GF29.219	0.015	0.000	0.000	0.000
GF29.221	0.670	0.000	0.000	0.000
GF29.223	0.235	0.000	0.000	0.000
GF29.224	0.000	0.000	0.000	0.000
GF29.226	0.000	0.000	0.000	0.000
GF29.229	0.003	0.000	0.000	0.000
GF29.254	0.001	0.000	0.000	0.188
GF29.272	0.002	0.000	0.000	0.438
GF29.275	0.001	0.000	0.000	0.313
GF29.278	0.000	0.000	0.000	0.000
GF29.282	0.000	0.000	0.000	0.063
MFW2.000	0.000	0.000	0.000	0.000
MFW2.157	0.002	1.000	0.958	0.000
MFW2.161	0.998	0.000	0.042	0.063
MFW2.204	0.000	0.000	0.000	0.000
MFW2.221	0.000	0.000	0.000	0.000
MFW2.233	0.000	0.000	0.000	0.000
MFW2.239	0.000	0.000	0.000	0.063
MFW2.241	0.000	0.000	0.000	0.063
MFW2.249	0.000	0.000	0.000	0.250
MFW2.251	0.000	0.000	0.000	0.313

(Supplementary table 5.1 continued)

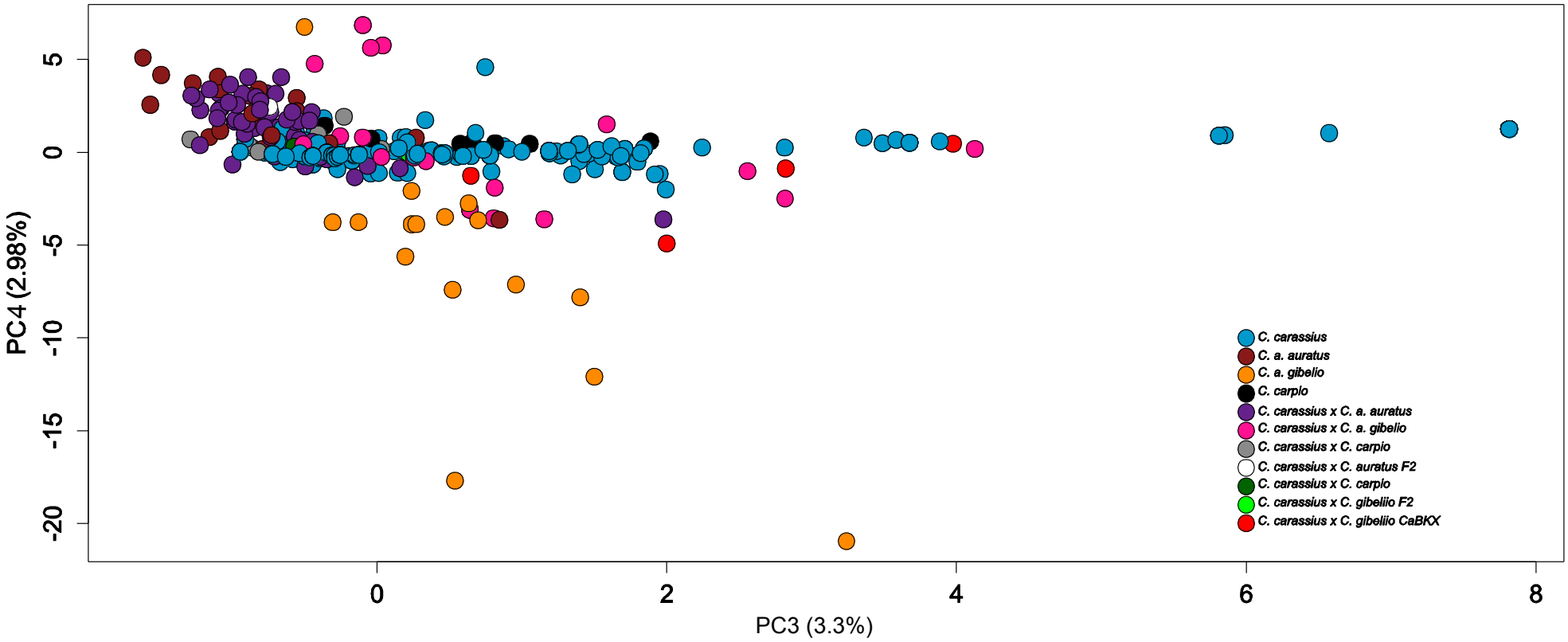
MFW2.253	0.000	0.000	0.000	0.063
MFW2.257	0.000	0.000	0.000	0.063
MFW2.261	0.000	0.000	0.000	0.000
MFW2.265	0.000	0.000	0.000	0.125
MFW2.267	0.000	0.000	0.000	0.000
J7.124	0.003	0.000	0.000	0.000
J7.202	0.020	0.389	0.125	0.000
J7.204	0.002	0.611	0.771	0.000
J7.206	0.000	0.000	0.000	0.000
J7.208	0.015	0.000	0.083	0.813
J7.210	0.015	0.000	0.000	0.000
J7.212	0.002	0.000	0.021	0.000
J7.214	0.003	0.000	0.000	0.000
J7.216	0.010	0.000	0.000	0.000
J7.218	0.002	0.000	0.000	0.000
J7.220	0.822	0.000	0.000	0.188
J7.221	0.000	0.000	0.000	0.000
J7.222	0.045	0.000	0.000	0.000
J7.224	0.026	0.000	0.000	0.000
J7.228	0.035	0.000	0.000	0.000
Ca07.120	0.001	0.000	0.000	-
Ca07.122	0.008	0.000	0.000	-
Ca07.124	0.565	0.000	0.000	-
Ca07.126	0.013	0.000	0.208	-
Ca07.128	0.019	0.000	0.063	-
Ca07.130	0.012	0.500	0.271	-
Ca07.132	0.019	0.333	0.104	-
Ca07.134	0.001	0.028	0.083	-
Ca07.136	0.324	0.083	0.021	-
Ca07.138	0.010	0.000	0.208	-
Ca07.140	0.027	0.000	0.000	-
Ca07.144	0.002	0.000	0.000	-
Ca07.148	0.000	0.000	0.000	-
Ca07.150	0.000	0.000	0.021	-
Ca07.155	0.000	0.000	0.000	-

Supplementary table 5.2. Species and hybrid class assignments based on Newhybrids analysis of samples genotyped with both RADseq and Microsatellites.

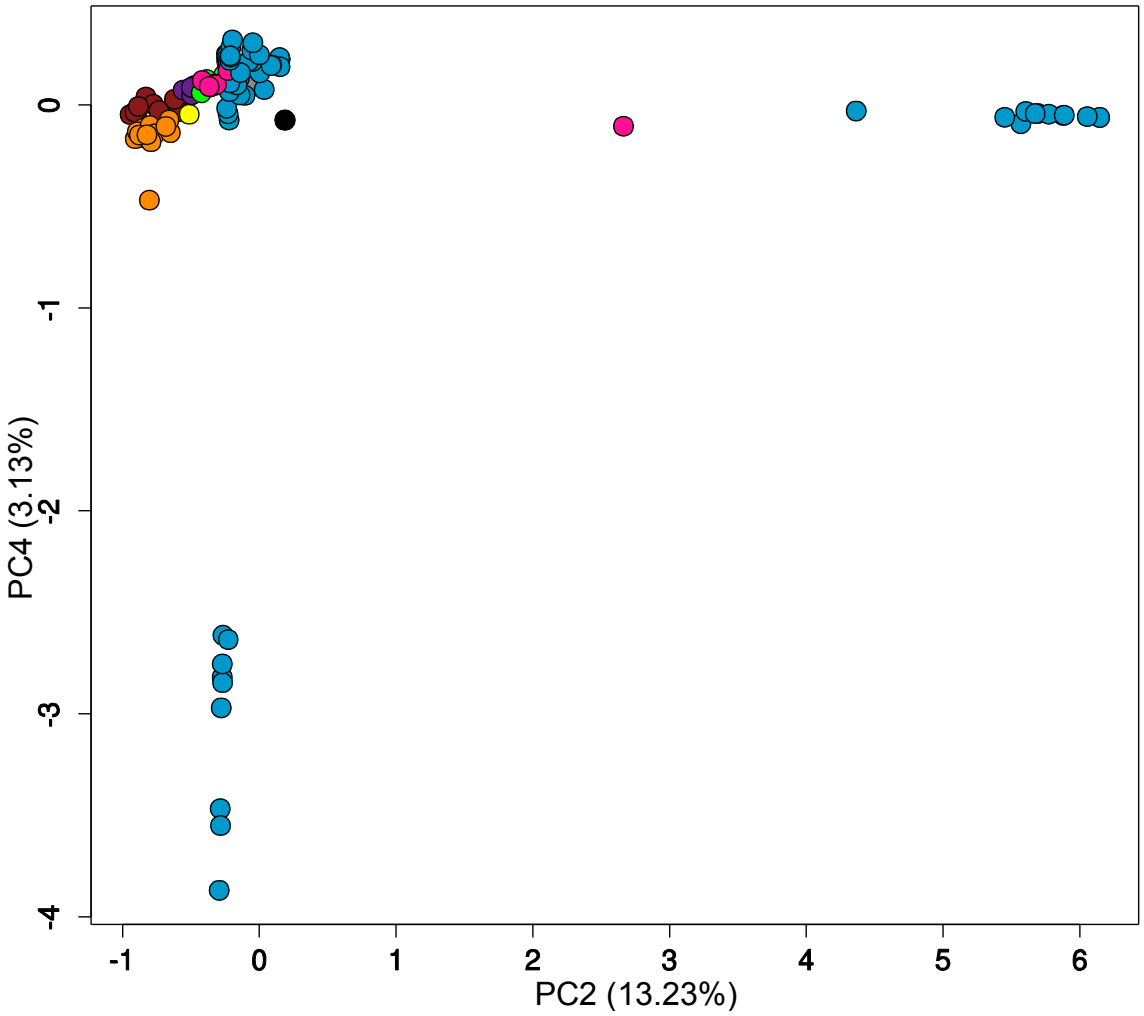
Sample	RAD	Microsatellites
GBR8_04	*C. carassius*	C.carassius
GBR8_05	*C. carassius*	C.carassius
GBR8_06	*C. carassius*	C.carassius
GBR8_08	*C. carassius*	C.carassius
GBR8_09	*C. carassius*	C.carassius
GBR8_10	*C. carassius*	C.carassius
GBR8_11	*C. carassius*	C.carassius
GBR8_12	*C. carassius*	C.carassius
GBR8_13	*C. carassius*	C.carassius
BEL1_01	*C. carassius*	C.carassius
BEL1_02	*C. carassius*	C.carassius
BEL1_03	*C. carassius*	C.carassius
BEL1_04	*C. carassius*	C.carassius
BEL1_06	*C. carassius*	C.carassius
GBR4_10	*C. carassius*	C.carassius
GBR4_2	*C. carassius*	C.carassius
GBR4_3	*C. carassius*	C.carassius
GBR4_4	*C. carassius*	C.carassius
GBR4_5	*C. carassius*	C.carassius
GBR4_6	*C. carassius*	C.carassius
GBR4_7	*C. carassius*	C.carassius
GBR4_8	*C. carassius*	C.carassius
GBR4_9	*C. carassius*	C.carassius
FIN3_01	*C. carassius*	C.carassius
FIN3_02	*C. carassius*	C.carassius
FIN3_03	*C. carassius*	C.carassius
FIN3_04	*C. carassius*	C.carassius
FIN3_05	*C. carassius*	C.carassius
FIN3_06	*C. carassius*	C.carassius
FIN3_07	*C. carassius*	C.carassius
FIN3_08	*C. carassius*	C.carassius
FIN3_09	*C. carassius*	C.carassius
FIN3_10	*C. carassius*	C.carassius
DEN1_06	*C. carassius*	C.carassius
DEN1_07	*C. carassius*	C.carassius
DEN1_08	*C. carassius*	C.carassius
DEN1_09	*C. carassius*	C.carassius
DEN1_10	*C. carassius*	C.carassius
DEN1_11	*C. carassius*	C.carassius
DEN1_12	*C. carassius*	C.carassius
DEN1_13	*C. carassius*	C.carassius
DEN1_14	*C. carassius*	C.carassius
DEN1_15	*C. carassius*	C.carassius
BEL5_2	*C. gibelio*	na
BEL5_3	*C. gibelio*	na
BEL5_4	*C. gibelio*	na
BEL5_5	*C. gibelio*	na
BEL5_6	*C. carassius*	na
GBR14_01	*C. carassius* x *C. gibelio*	na
GBR14_10	*C. carassius*	C.carassius
GBR16_01	*C. carassius* x *C. auratus*	*C. carassius* x *C. auratus*
GBR16_03	*C. carassius*	C.carassius
GBR16_04	*C. carassius*	C.carassius
GBR16_05	*C. carassius*	C.carassius
GBR16_07	*C. carassius*	C.carassius
GBR16_08	*C. carassius*	C.carassius
GBR3_17	*C. auratus*	*C. auratus*
GBR3_19	*C. auratus*	*C. auratus*
GBR3_4	*C. auratus*	*C. auratus*
GBR3_5	*C. auratus*	*C. auratus*
GBR3_7	*C. auratus*	*C. auratus*
GBR17_10	*C. auratus*	*C. auratus* ornamental
GBR17_1	*C. auratus*	*C. auratus* ornamental
GBR17_2	*C. auratus*	*C. auratus* ornamental
GBR17_3	*C. auratus*	*C. auratus* ornamental
GBR17_7	*C. auratus* x *C. gibelio*	*C. auratus* ornamental
HUN4_10	*C. carassius*	C.carassius
HUN4_3	*C. carassius*	C.carassius
HUN4_5	*C. carassius*	C.carassius
HUN4_H1	*C. carassius*	C.carassius
HUN4_H2	*C. carassius*	C.carassius
HUN4_11	*C. carassius* x *C. gibelio*	C.carassiusxC.a.gibelio
HUN4_1	*C. gibelio*	C.a.gibelio
HUN4_2	*C. gibelio*	C.a.gibelio
GBR10_10	*C. carassius* x *C. auratus*	*C. carassius* x *C. auratus*
GBR10_11	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_12	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_13	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_14	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_15	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_16	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_1	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_2	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_3	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus

GBR10_4	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_5	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_6	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_7	*C. carassius* x *C. auratus*	C.carassiusxC.a.auratus
GBR10_9	*C. carassius* x *C. gibelio*	*C. carassius* x *C. gibelio*
UKR1_01	*C. gibelio*	C.a.gibelio
UKR1_08	*C. gibelio*	C.a.gibelio
UKR1_09	*C. gibelio*	C.a.gibelio
GBR7_10	*C. carassius*	C.carassius
GBR7_1	*C. carassius*	C.carassius
GBR7_2	*C. carassius*	C.carassius
GBR7_3	*C. carassius*	C.carassius
GBR7_4	*C. carassius*	C.carassius
GBR7_5	*C. carassius*	C.carassius
GBR7_6	*C. carassius*	na
GBR7_7	*C. carassius*	C.carassius
GBR7_8	*C. carassius*	C.carassius
GBR7_9	*C. carassius*	C.carassius
UKR2_57	*C. gibelio*	na
UKR2_60	*C. gibelio*	na
UKR2_62	*C. gibelio*	na
UKR2_63	*C. gibelio*	na
UKR2_64	*C. gibelio*	na
SWE12_01	*C. carassius*	C.carassius
SWE12_02	*C. carassius*	C.carassius
SWE12_07	*C. carassius*	C.carassius
SWE12_10	*C. carassius*	C.carassius
SWE12_11	*C. carassius*	C.carassius
SWE12_12	*C. carassius*	C.carassius
SWE12_13	*C. carassius*	na
SWE12_14	*C. carassius*	C.carassius
SWE12_15	*C. carassius*	C.carassius
FIN4_01	*C. carassius*	*C. carassius*
FIN4_03	*C. carassius*	*C. carassius*
FIN4_04	*C. carassius*	C.carassius
FIN4_05	*C. carassius*	C.carassius
FIN4_06	*C. carassius*	C.carassius
FIN4_07	*C. carassius*	C.carassius
FIN4_08	*C. carassius*	C.carassius
FIN4_09	*C. carassius*	C.carassius
DEN2_01	*C. carassius*	C.carassius
DEN2_02	*C. carassius*	C.carassius
DEN2_03	*C. carassius*	C.carassius
DEN2_04	*C. carassius*	C.carassius
DEN2_05	*C. carassius*	C.carassius
DEN2_06	*C. carassius*	C.carassius
DEN2_07	*C. carassius*	C.carassius
DEN2_08	*C. carassius*	C.carassius
POL4_10	*C. carassius*	C.carassius
POL4_1	*C. carassius*	C.carassius
POL4_2	*C. carassius*	C.carassius
POL4_3	*C. carassius*	C.carassius
POL4_4	*C. carassius*	C.carassius
POL4_5	*C. carassius*	C.carassius
POL4_6	*C. carassius*	C.carassius
POL4_7	*C. carassius*	C.carassius
POL4_8	*C. carassius*	C.carassius
POL4_9	*C. carassius*	C.carassius
RUS1_01	*C. carassius*	C.carassius
RUS1_02	*C. carassius*	C.carassius
RUS1_03	*C. carassius*	C.carassius
RUS1_04	*C. carassius*	C.carassius
RUS1_05	*C. carassius*	C.carassius
RUS1_06	*C. carassius*	C.carassius
RUS1_07	*C. carassius*	C.carassius
RUS1_08	*C. carassius*	C.carassius
RUS1_9	*C. carassius*	C.carassius
GBR6_10	*C. carassius*	C.carassius
GBR6_1	*C. carassius*	C.carassius
GBR6_20	*C. carassius* x *C. gibelio*	*C. carassius* x *C. gibelio*
GBR6_2	*C. carassius*	C.carassius
GBR6_31	*C. gibelio*	C.carpio
GBR6_32	*C. gibelio*	C.carpio
GBR6_3	*C. carassius*	C.carassius
GBR6_4	*C. carassius*	C.carassius
GBR6_5	*C. carassius*	C.carassius
GBR6_6	*C. carassius*	C.carassius
GBR6_7	*C. carassius*	C.carassius
GBR6_8	*C. carassius*	C.carassius
GBR6_9	*C. carassius*	na
SWE2_10	*C. carassius*	C.carassius
SWE2_1	*C. carassius*	C.carassius
SWE2_2	*C. carassius*	C.carassius
SWE2_3	*C. carassius*	C.carassius
SWE2_4	*C. carassius*	C.carassius
SWE2_5	*C. carassius*	C.carassius
SWE2_6	*C. carassius*	C.carassius
SWE2_8	*C. carassius*	C.carassius
SWE2_9	*C. carassius*	C.carassius
SWE8_10	*C. carassius*	C.carassius
SWE8_1	*C. carassius*	C.carassius
SWE8_2	*C. carassius*	C.carassius
SWE8_3	*C. carassius*	C.carassius

SWE8_4	*C. carassius*	C.carassius
SWE8_5	*C. carassius*	C.carassius
SWE8_6	*C. carassius*	C.carassius
SWE8_7	*C. carassius*	C.carassius
SWE8_8	*C. carassius*	C.carassius
SWE8_9	*C. carassius*	C.carassius
SWE9_10	*C. carassius*	C.carassius
SWE9_1	*C. carassius*	C.carassius
SWE9_2	*C. carassius*	C.carassius
SWE9_3	*C. carassius*	C.carassius
SWE9_4	*C. carassius*	C.carassius
SWE9_5	*C. carassius*	C.carassius
SWE9_6	*C. carassius*	C.carassius
SWE9_7	*C. carassius*	C.carassius
SWE9_8	*C. carassius*	C.carassius
SWE9_9	*C. carassius*	C.carassius
SWE10_01	*C. carassius*	C.carassius
SWE10_02	*C. carassius*	C.carassius
SWE10_03	*C. carassius*	C.carassius
SWE10_04	*C. carassius*	C.carassius
SWE10_05	*C. carassius*	C.carassius
SWE10_06	*C. carassius*	C.carassius
SWE10_07	*C. carassius*	C.carassius
SWE10_08	*C. carassius*	C.carassius
SWE10_09	*C. carassius*	C.carassius
SWE20_06	*C. gibelio*	C.a.gibelio
SWE20_07	C.carassiusx (*C. carassius* x *C. gibelio*)	C.carassiusxC.a.gibelio F2
SWE20_08	C.carassiusx (*C. carassius* x *C. gibelio*)	C.carassiusxC.a.gibelio F2
SWE20_11	C.carassiusx (*C. carassius* x *C. gibelio*)	C.carassiusxC.a.gibelio F2
NOR2_02	*C. carassius*	C.carassius
NOR2_03	*C. carassius*	C.carassius
NOR2_04	*C. carassius*	C.carassius
NOR2_09	*C. carassius*	C.carassius
NOR2_11	*C. carassius*	C.carassius
NOR2_12	*C. carassius*	C.carassius
NOR2_16	*C. carassius*	C.carassius
NOR2_17	*C. carassius*	C.carassius
NOR2_18	*C. carassius*	C.carassius
POL3_10	*C. carassius*	C.carassius
POL3_1	*C. carassius*	C.carassius
POL3_2	*C. carassius*	C.carassius
POL3_3	*C. carassius*	C.carassius
POL3_4	*C. carassius*	C.carassius
POL3_5	*C. carassius*	C.carassius
POL3_6	*C. carassius*	C.carassius
POL3_7	*C. carassius*	C.carassius
POL3_8	*C. carassius*	C.carassius
POL3_9	*C. carassius*	C.carassius
HUN2_11	*C. carassius*	C.carassius
HUN2_1	*C. carassius*	C.carassius
HUN2_13	*C. carassius*	C.carassius
HUN2_3	*C. carassius*	C.carassius
HUN2_5	*C. carassius*	C.carassius
HUN2_7	*C. carassius*	C.carassius
FIN2_1	*C. carassius* x *C. gibelio*	C.carassiusxC.a.gibelio
FIN2_2	*C. carassius* x *C. gibelio*	C.carassiusxC.a.gibelio
FIN2_3	*C. carassius* x *C. gibelio*	C.carassiusxC.a.gibelio
FIN2_4	*C. carassius* x *C. gibelio*	C.carassiusxC.a.gibelio
FIN2_5	*C. carassius* x *C. gibelio*	C.carassiusxC.a.gibelio
FIN2_6	*C. gibelio*	C.a.gibelio
SWE14_02	*C. carassius*	C.carassius
SWE14_03	*C. carassius*	C.carassius
SWE14_04	*C. carassius*	C.carassius
SWE14_06	*C. carassius*	C.carassius
SWE14_08	*C. carassius*	C.carassius
SWE14_10	*C. carassius*	C.carassius
SWE14_11	*C. carassius*	C.carassius
SWE14_13	*C. carassius*	C.carassius
SWE14_16	*C. carassius*	C.carassius



Supplementary Figure 5.1. Principle components 3 and 4 for the whole microsatellite dataset of 1333 genotyped individuals, explaining the variation between the two lineages of *C. carassius* and between the two *C. auratus* spp. respectively. Colours represent the final consensus assignments of individuals to species or hybrid class based on Newhybrids analyses of both microsatellite and RADseq datasets. Back to text.



Supplementary Figure 5.2. Principle components 2 and 4 for the entire RADseq dataset of 247 genotyped individuals. PC2, captures the variation between the two major lineages within *C. carassius* (identified in Chapter 3 of this thesis, and PC4 captures the variation between samples in the Don River catchment and the rest of pure *C. carassius*. Colours represent the final consensus assignments of individuals to species or hybrid class based on Newhybrids analyses of both microsatellite and RADseq datasets. Back to text.

Chapter 6 General Discussion

The overarching goals of this thesis were twofold: firstly, to lay down a solid foundation of phylogeographic knowledge on which to build conservation plans for the threatened crucian carp, *Carassius carassius* (L.); and secondly to explore the evolutionary processes associated with their hybridisation with three of the most widely introduced fish species in Europe, the goldfish, *Carassius auratus* (L.), the gibel carp, *Carassius gibelio* (Bloch) and the common carp, *Cyprinus carpio* (L.). Recent advances in sequencing technology now allow for goals such as these to be addressed using a genomic approach, and one such approach, Restriction Site Associated DNA sequencing (RADseq) has been used throughout this study.

The complex datasets produced by the RADseq approach come with unique bioinformatics challenges, which must be overcome before the data can be used to its full potential (Davey *et al.* 2013). Therefore, in the present thesis the first task, addressed in Chapter 2, was to account for several inherent biases of the RADseq dataset used. Exploratory analyses revealed that two sources of bias were particularly important in this dataset. Firstly, allele dropout between species was prolific, with drastic reductions in the number of homologous RADseq loci as divergence between species increased. Allele dropout has been found by several other studies (Hohenlohe *et al.* 2011; McCormack *et al.* 2012; Wagner *et al.* 2013; Wang *et al.* 2013; Pante *et al.* 2015) and Gautier et al. (2012) show that it can inflate heterozygosity estimates, which in turn could reduce F_{ST} estimates and result in an underestimation of population structure. In the context of species conservation, these erroneous results could lead conservationists to incorrect decisions, for example, not recognising separate management units, which could seriously threaten the success of conservation plans. Furthermore, allele dropout could have significant impact on hybrid analyses, such as tests for introgression carried out in Chapter 5, and lead to false conclusions about the amount of introgression occurring between species. Fortunately, strict population and sample filters for each locus and the examination of several key properties of the data shown in Chapter 2, including heterozygosity and read coverage, can helped to identify and exclude such loci.

The second potentially confounding attribute of the RADseq dataset used here was the presence of ohnologous loci; duplicated genome regions resulting from whole genome duplications (WGD (Ohno 1970)). These loci were treated here, as in many other studies (Hohenlohe *et al.* 2011; Ogden *et al.* 2013) as nuisance data and so were filtered out using population genetics and coverage filters in Chapter 2. However, the role that ohnologs play in evolution is a topic of much interest to evolutionary biologists. Several studies have implicated WGDs as important drivers of adaptive radiations and evolutionary innovation (e.g. Schranz *et al.* 2012; Berthelot *et al.* 2014), although general patterns in the fate of genes after whole genome duplications remain elusive. For example, large lineage-specific differences have been found in the amount of genome rearrangement following WGDs (Sémon & Wolfe 2007a; Kasahara *et al.* 2007; Hufton *et al.* 2008), and it is not known why some ohnolog pairs remain intact whereas others undergo neofunctionalization, subfunctionalization and pseudogenisation (Sémon & Wolfe 2007b). *C. carassius* have undergone several genome duplications, the most recent of which is estimated to be approximately eight million years ago (Li *et al.* 2015). *C. carassius* is therefore potentially an ideal study system in which to study the early processes that follow a genome duplication event. By clustering reads which were identified in Chapter 2 as putatively belonging to ohnolog pairs, for example using Uclust (Edgar 2010), pairs of loci could theoretically be isolated and used to address these questions. Furthermore, as analyses of ohnologous gene fates has been performed in *C. carpio* (Li *et al.* 2015), a comparative analysis between these two very closely related species may yield some highly interesting insights and perhaps allow for the identification of convergent fates of ohnologous gene pairs.

The conservation of C. carassius in Europe

Prior to this study, no knowledge existed for the broad-scale genetic structure of the threatened *C. carassius* in Europe. In Chapter 3, the Europe-wide phylogeographic patterns within *C. carassius* were examined using mitochondrial, microsatellite and genome wide Single Nucleotide Polymorphism (SNP) markers. *C. carassius* were found to exist in two previously unknown major lineages which are geographically separated by the Danubian watershed and have been isolated for approximately 2.26 million years. This pattern is distinct among European freshwater fish, with the Danubian river catchment having been an important source for the postglacial recolonisation of northern Europe in almost all species documented to date (Nesbø *et al.* 1999; Durand *et*

al. 1999; Kotlík & Berrebi 2001; Salzburger *et al.* 2003; Gum *et al.* 2009; Larmuseau *et al.* 2009). Therefore, to elucidate the processes that have lead to this distinct distribution, the postglacial recolonisation routes of *C. carassius* were reconstructed for the Pleistocene epoch using RADseq data and an Approximate Bayesian Computation (ABC) approach. This analysis confirmed the separation of these lineages and showed that northern and northwestern Europe have been colonised by a single lineage and that the Danubian populations made no major contribution in this colonisation. It is hypothesised that this is a result of the distinct ecology of *C. carassius*, which prefer spatially restricted habitats such as ponds and lakes, and thus presumably have low dispersal capacity (Holopainen *et al.* 1997). This information is invaluable for the conservation of crucian carp in central Europe and every effort should be made to avoid stocking these lineages together so as not to homogenise the variation within the species.

The power of a given phylogeographic study comes from a combination of the number and spatial distribution of samples, and the number and type of genetic marker (Morin *et al.* 2009; Schwartz & McKelvey 2009). These elements of study design exist as a trade-off mediated by time and monetary constraints. Molecular ecologists have, given much attention to identifying the best balance of sample and marker number to maximise the power of a phylogeographic study, however the vast majority of these studies have been simulation-based (e.g. Schwartz & McKelvey 2009; Epperson *et al.* 2010; Landguth *et al.* 2012; Oyler-McCance *et al.* 2012). In Chapter 3, the phylogeographic results obtained from microsatellite and RADseq datasets were compared in order to lend perspectives to this discussion from real biological data. These comparisons, revealed, that the hugely increased number of loci in the RADseq dataset overcame the greatly reduced sample size (17.6%) to produce a comparable phylogeography to that of the microsatellite approach, which emphasised the fine scale structure among populations. This result agrees with simulation studies performed by Morin *et al.* (2009), which predicted that the higher number of SNPs reduced the number of individuals required to confidently differentiate between populations. However, these results should be interpreted in the context of this study alone; *C. carassius* has strong population structure compared to many other fish species, likely owing to the isolated nature of its preferred habitats. In a system with lower population structure, the number of samples and their spatial uniformity may play a more important role than the number of loci genotyped. Analyses such as those in Chapter 3, from

freshwater fish species with lower population structure would, therefore be a valuable contribution to this topic.

The phylogeographic analyses in Chapter 3 focus only on the neutral genetic signal in the data, however, RADseq datasets produced here lend themselves to many evolutionary analyses, and, although outside of the scope of this study, it would be interesting to examine phylogeographic processes at loci under selection (or those linked to selected regions). It has been suggested that genomic tools could be used to identify loci for inbreeding depression, local adaptation and those with the potential to cause outbreeding depression (Allendorf *et al.* 2010; Funk *et al.* 2012). In butterflies, the Phosphoglucose isomerase (Pgi) gene, which is important in glycolysis is thought to be highly important for dispersal performance, and, as such, is a candidate locus for informing the conservation of the Glanville fritillary, *Melitaea cinxia* (L.) (Wheat 2010). Incorporating information on such adaptive variation into conservation plans is a major goal of the incipient field of conservation genomics, as it can be used to prioritise populations to be conserved and decide on those to use a source populations for translocations (Funk *et al.* 2012). To this end, further study is already underway, examining the association between the diversity within *C. carassius* with temperature, a known driver of selection in many fish species, and indeed ectotherms in general (Narum et al. 2013).

A crucial aspect of conservation plans is knowledge of a species' native range (Frankham *et al.* 2002; Scoble & Lowe 2010; IUCN 2012). In *C. carassius*, defining this range has proved challenging in the past (Wheeler 2000), and in England, the status of *C. carassius* has been particularly contentious (Maitland 1972; Wheeler 2000). The currently held belief is that *C. carassius* is native in England, however, in opposition to this, Chapter 4 shows strong support for an introduced origin of *C. carassius* in England, with an estimated time of arrival around 576 years ago. This result raises an interesting debate, which has received increasing attention in the last few decades, that is; how to deal with non-native species? In the past, non-native species have been vilified and implicated in the decline of countless native species (Davis *et al.* 2011). Indeed, there are many examples where non-native species have been shown to have detrimental impacts on the native fauna (Simberloff *et al.* 2013). However, there are also many examples of non-native species being implicated where no causal data exists (Davis *et al.* 2011). Some non-native species have even been of conservation benefit,

for example the tamarisk, *Tamarix* spp. (L.), which is introduced in southwestern United States, was initially blamed for the decline in willow flycatcher, *Empidonax traillii extimus* (Audubon) populations. Further studies revealed that, in fact *E. t. extimus* relied heavily on *Tamarix* spp. for nest sites (Schlaepfer *et al.* 2011). There is now a growing consensus that the subjective labelling of non-native species as detrimental to native ecosystems is not a progressive approach (Gozlan 2008; Davis *et al.* 2011; Schlaepfer *et al.* 2011) and the naïve assumption that all non-native species are a threat can dilute conservation efforts that would be better focused on a few problematic species. Instead, an impact driven threat assessment for non-native species is advocated (Davis *et al.* 2011). Such discussions become even more important in the light of the increasing numbers of species being translocated by humans (Madeira *et al.* 2005) and the increasing number of natural species range shifts that are being driven by climate change (Parmesan *et al.* 1999; Muhlfeld *et al.* 2014).

For *C. carassius*, no impact studies have previously been performed due to the assumption that they were native to England. However, the pond ecosystems that they inhabit in England are extremely important for conservation of many aquatic species (Oertli *et al.* 2002). Further work is therefore needed to characterise the interactions between *C. carassius* and the aquatic life present in such environments. This work may be facilitated by the characteristics of the small ponds that *C. carassius* inhabits. For example, in Norfolk *C. carassius* is found in many small ponds, known as “Marl Pits”, which were created by historic farming practices (Sayer *et al.* 2011). These Marl Pits, which are often very closely spaced (adjacent fields), are likely to be highly similar in their environmental properties and therefore provide easy to control arenas for natural experiments to assess the impact of *C. carassius* presence/absence. However, until such impact data becomes available, an argument can be made for the continued conservation of *C. carassius* in England in light of the growing consensus that *C. carassius* is threatened throughout most, if not all, of its range (Holopainen & Oikari 1992; Navodaru *et al.* 2002; Hänfling *et al.* 2005; Papoušek *et al.* 2008; Copp & Sayer 2010; Sayer *et al.* 2011; Mezhzherin *et al.* 2012; Wouters *et al.* 2012). Therefore, to cease conservation in one region would be counter productive in the context of conserving the entire species (Copp *et al.* 2005). This consideration is even more important in light of the invasion of *C. gibelio* in continental Europe (Wouters *et al.* 2012; Deinhardt 2013). As England is one of the only countries in Europe which is actively conserving *C.*

carassius (Copp & Sayer 2010), decisions regarding its continued conservation in this region should not be taken lightly.

The result that *C. carassius* has introduced origins in England calls into question our assumptions about the nativeness of other freshwater fish species in this region. Like *C. carassius*, all primary freshwater fish will have had to naturally disperse northwards from the glacial refugia in continental Europe and cross Doggerland into the UK. It may be that low dispersal rates of *C. carassius* set it apart from other UK species, which were able to naturally colonise this area in the time window between the last glacial maximum, approximately 18 000 years ago, and the inundation of Doggerland approximately 7 800 years ago. However, given that the hydrological history of the UK offers the rare opportunity to test such hypotheses, and that the approaches, such as those used in Chapter 4, now exist to perform these tests, it would be interesting to address the presumed native status of other English fish species.

Though *C. carassius* is threatened by a number of factors throughout its range, the most commonly cited threat is that of hybridisation with the three non-native species: the goldfish, *C. auratus,* the gibel carp, *C. gibelio* and the common carp, *C. carpio,* which has been shown to be prevalent by previous studies (Hänfling *et al.* 2005; Papoušek *et al.* 2008; Copp & Sayer 2010; Sayer *et al.* 2011; Mezhzherin *et al.* 2012; Wouters *et al.* 2012). The results of Chapter 5 confirm these high levels of hybridisation showing that hybrids were present in 86% of populations where *C. carassius* was sympatric with a non-native species. In a system of such high hybridisation rates, there has been understandable concern over the potential for introgression between *C. carassius* and non-native species (Hänfling *et al.* 2005; Mezhzherin *et al.* 2012; Wouters *et al.* 2012). However, previous studies have been unable to comprehensively assess the potential for introgression in this system, due to the limitations of using only small numbers of microsatellite, allozyme loci or morphological characters. In Chapter 5, the use of almost 4 500 genome-wide SNP markers present in all four species allowed for tests of the presence of introgression in this system. Interestingly, despite the occurrence of diploid (and thus presumably fertile) backcross hybrids between *C. carassius* and *C. gibelio* in one population, no evidence of introgression was found. This may be due to one of several postzygotic isolating mechanisms including outcrossing depression, low fertility of F1 hybrids or backcross generations, behavioural isolation, or prezygotic barriers like meiotic dysfunction in F1 hybrids, leading to triploid infertile backcrosses.

However, unfortunately no *C. carassius* tissue samples suitable for RADseq analysis were available from population in which we found the backcrossed individuals. It was, therefore, not possible to screen this population for cryptic introgression past the initial backcross stage. RADseq data from this population and others containing backcrossed individuals would be invaluable for more confidently ruling out the occurrence of introgression here. There is the possibility that introgression has occurred at the scale of only a few genes, as has been documented in *Heliconius* butterflies (The Heliconius Genome Consortium 2012), and as such were missed in this study due to gaps between SNP loci across the genome. If the latter is the case, it would be likely that this introgression is old, and that the length of introgressed linkage blocks have been eroded by recombination over time (Twyford & Ennos 2011). A useful extension of this study would, therefore, be to use an enzyme such as Pst1 or Nsi1, which yields many more loci in RADseq library preparation than the Sbf1 enzyme used here (Davey *et al.* 2010). Importantly, an advantage of SNP data and the RADseq approach is that data can be combined across studies, regardless of the enzyme used, and so upscaling the current dataset to include more markers would be straightforward.

One interesting result of Chapter 5 was the presence of a triploid hybrid between *C. carassius* and *C. gibelio*. It is hypothesised that this fish was produced through the backcrossing between an F1 *C. carassius* x *C. gibelio* hybrid female, which produced an unreduced oocyte, and a pure *C. carassius* male, based on similar findings in *Poeciliid* fish species (Lampert *et al.* 2007). Hänfling et al. (2005) observed triploid hybrids between *C. carassius* and *C. auratus* in the UK, which were also likely to result from a backcrossing event, suggesting that they are not an isolated occurrence. Interestingly, those identified in Hänfling et al. (2005) had identical genotypes, raising the possibility that they belonged to the same clonal lineage. This inference is particularly important in light of the documented importance of hybridisation in the emergence of polyploid, gynogenetic fish lineages (Lampert *et al.* 2007; Lampert & Schartl 2008). Furthermore, the study of the mechanisms behind meiotic dysfunction in such lineages may shed light on the barriers to gene flow that appear to prevent introgression between *C. auratus* spp. and *C. carassius*.

Conclusions

In this thesis, genetic and genomic approaches have provided valuable insights into the phylogeographic structure of the threatened *C. carassius* and allow for the clear identification of the most important conservation units present in Europe. Furthermore, the introduced origins of *C. carassius* in England puts to rest the long standing debate surrounding their status, and highlights the need for impact assessments to ensure that they have no detrimental impact on native English ecosystems. The threat from hybridisation with non-native species requires significant attention. The work carried out here adds to the consensus that hybridisation is prolific between these species, but to date the impact of the potentially vigorous hybrids is unknown. This should be seen as a priority for the conservation of *C. carassius*, especially in light of the continued spread of all three non-native species. Reassuringly, however, there is no evidence that introgression is occurring between these species and, thus, conservationists should prioritise assessments of the direct impact of non-native species and their hybrids. Further genomic studies are still required to confidently rule out the possibility of localised introgression in this system.

Bibliography

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. Nature reviews. Genetics, 11, 697–709.

Almodóvar A, Nicola GG, Elvira B, García-Marín JL (2006) Introgression variability among Iberian brown trout Evolutionary Significant Units: the influence of local management and environmental features. Freshwater biology, 51, 1175–1187.

Amish SJ, Hohenlohe PA, Painter S et al. (2012) RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. Molecular Ecology resources, 12, 653–660.

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics, 188, 799–808.

Anderson EC, Thompson EA (2002) A model-based method for identifying species hybrids using multilocus genetic data. Genetics, 160, 1217–1229.

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. http://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.

Arnold ML (1996) Natural Hybridization and Evolution (ML Arnold, Ed,). Oxford University Press, Oxford, UK.

Arnold ML, Hodges SA (1995) Are natural hybrids fit or unfit relative to their parents? Trends in ecology & evolution, 10, 67–71.

Avise JC (2010) Perspective: conservation genetics enters the genomics era. Conservation genetics , 11, 665–669.

Baddeley A, Turner R (2005) spatstat: An R Package for Analyzing Spatial Point Patterns. Journal of Statistical Software, 12, 1–42.

Baird NA, Etter PD, Atwood TS et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one, 3, e3376.

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

Barton NH (2001) The role of hybridization in evolution. Molecular Ecology, 10, 551–568.

Baxter SW, Davey JW, Johnston JS et al. (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PloS one, 6, e19315.

BEAST Tutorial - Tree priors and dating

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

Berthelot C, Brunet F, Chalopin D et al. (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nature communications, 5, 3657.

Bianco PG (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. Munchen, 1, 167–184.

Birchler JA, Yao H, Chudalayandi S (2006) Unraveling the genetic basis of hybrid vigor. Proceedings of the National Academy of Sciences of the United States of America, 103, 12957–12958.

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.

Boecklen WJ, Howard DJ (1997) Genetic Analysis of Hybrid Zones: Numbers of Markers and Power of Resolution. Ecology, 78, 2611–2616.

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.

Byers JE (2002) Impact of non-indigenous species on natives enhanced by anthropogenic alteration of selection regimes. Oikos , 97, 449–458.

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

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

Carroll SP, Dingle H (1996) The biology of post-invasion events. Biological conservation, 78, 207–214.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: Building and Genotyping Loci *De novo* From Short-Read Sequences. G3: Genes, Genomes, Genetics, 1, 171–182.

Catchen J, Bassham S, Wilson T et al. (2013a) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. Molecular Ecology, 22, 2864–2883.

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

Chan K (2010) Can great crested newts (Triturus cristatus) co-exist with fish. master's thesis Thesis. University College London.

Choleva L, Janko K, De Gelas K et al. (2012) Synthesis of clonality and polyploidy in vertebrate animals by hybridization between two sexual species. Evolution; international journal of organic evolution, 66, 2191–2203.

Chong Z, Ruan J, Wu C-I (2012) Rainbow: an integrated tool for efficient clustering and assembling RAD-seq reads. Bioinformatics , 28, 2732–2737.

Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A et al. (2011) Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. BMC genomics, 12, 4.

Clavero M, Brotons L, Pons P, Sol D (2009) Prominent role of invasive species in avian biodiversity loss. Biological conservation, 142, 2043–2049.

Clavero M, García-Berthou E (2005) Invasive species are a leading cause of animal extinctions. Trends in ecology & evolution, 20, 110.

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.

Cohen AN, Carlton JT (1998) Accelerating invasion rate in a highly invaded estuary. Science, 279, 555–558.

Colautti RI, Mac Isaac HJ (2004) A neutral terminology to define "invasive." Diversity & distributions, 10, 135–141.

Coles BJ (2000) Doggerland: the cultural dynamics of a shifting coastline. Geological Society, London, Special Publications, 175, 393–401.

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

Copp GH, Bianco PG, Bogutskaya NG et al. (2005) To be, or not to be, a non-native freshwater fish? Journal of Applied Ichthyology, 21, 242–262.

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 / European Research Network on Aquatic Invasive Species, 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.

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 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 , 2014.

Cornuet J-M, 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, Santos F, Beaumont MA et al. (2008) Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. Bioinformatics , 24, 2713–2719.

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

Crawford AM, Cuthbertson RP (1996) Mutations in sheep microsatellites. Genome research, 6, 876–879.

Crooijmans RPMA, Poel JJV der, Groenen MAM, Bierbooms VAF, Komen J (1997) Microsatellite markers in common carp (Cyprinus carpio L.). Animal genetics, 28, 129–134.

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.

Dallas JF (1992) Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. Mammalian genome: official journal of the International Mammalian Genome Society, 3, 452–456.

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

Davey JW, Cezard T, Fuentes-Utrilla P et al. (2013) Special features of RAD Sequencing data: implications for genotyping. Molecular Ecology, 22, 3151–3164.

Davey JW, Davey JL, Blaxter ML, Blaxter MW (2010) RADSeq: next-generation population genetics. Briefings in functional genomics, 9, 416–423.

Davey JW, Hohenlohe PA, Etter PD et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature reviews. Genetics, 12, 499–510.

Davis MA (2003) Biotic Globalization: Does Competition from Introduced Species Threaten Biodiversity? Bioscience, 53, 481–489.

Davis M (2011) Invasion Biology. In: Encyclopedia of Biological Invasions (eds Simberloff D, Rejmanek M), pp. 364–369. University of California Press.

Davis MA, Chew MK, Hobbs RJ et al. (2011) Don't judge species on their origins. Nature, 474, 153–154.

Decker JE, McKay SD, Rolf MM et al. (2014) Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. PLoS genetics, 10, e1004254.

Deinhardt M (2013) The invasive potential of Prussian carp in Finland under the light of a novel semi-clonal reproductive mechanism. Bio- ja ympäristötieteiden laitos, Jyväskylän yliopisto.

Didham RK, Tylianakis JM, Hutchison MA, Ewers RM, Gemmell NJ (2005) Are invasive species the drivers of ecological change? Trends in ecology & evolution, 20, 470–474.

Dobzhansky T (1937) Genetics and the origin of species. University Press, New York: Columbia.

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.

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 EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. Molecular biology and evolution, 28, 2239–2252.

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

D'Antonio CM, Dudley TL (1995) Biological Invasions as Agents of Change on Islands Versus Mainlands. In: Islands Ecological Studies., pp. 103–121. Springer Berlin Heidelberg.

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

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics , 26, 2460–2461.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in Molecular Ecology of non-model organisms. Heredity, 107, 1–15.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nature reviews. Genetics, 5, 435–445.

Ellis EA (1965) The Broads. Collins Books, London, UK.

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.

Facon B, Jarne P, Pointier JP, David P (2005) Hybridization and invasiveness in the freshwater snail Melanoides tuberculata: hybrid vigour is more important than increase in genetic variance. Journal of evolutionary biology, 18, 524–535.

Fitzpatrick BM, Johnson JR, Kump DK et al. (2010) Rapid spread of invasive genes into a threatened native species. Proceedings of the National Academy of Sciences of the United States of America, 107, 3606–3610.

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

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.

Freyhof J, Kottelat M (2008) Carassius carassius. In: IUCN Red List of Threatened Species 2013. IUCN.

Fritts TH, Rodda GH (1998) The Role of Introduced Species in the Degradation of Island Ecosystems: A Case History of Guam. Annual review of ecology and systematics, 29, 113–140.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. Trends in ecology & evolution, 27, 489–496.

Gautier M, Gharbi K, Cezard T et al. (2012) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology, 22, 3165–3178.

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.

Gibbons JW, Scott DE, Ryan TJ et al. (2000) The Global Decline of Reptiles, Déjà Vu Amphibians. Bioscience, 50, 653–666.

Glasauer SMK, Neuhauss SCF (2014) Whole-genome duplication in teleost fishes and its evolutionary consequences. Molecular genetics and genomics: MGG, 289, 1045–1060.

Goodman SJ, Barton NH, Swanson G, Abernethy K, Pemberton JM (1999) Introgression through rare hybridization: A genetic study of a hybrid zone between red and sika deer (genus Cervus) in Argyll, Scotland. Genetics, 152, 355–371.

Goudet J (2001) FSTAT, a program to estimate and test gene diversities and fixation indices. Université de Lausanne.

Goudet J (2005) HIERFSTAT, a package for to compute and test hierarchical F-statistics. Molecular Ecology notes, 5, 184–186.

Gozlan RE (2008) Introduction of non-native freshwater fish: is it all bad? Fish and fisheries , 9, 106–115.

Gozlan RE, Britton JR, Cowx I, Copp GH (2010) Current knowledge on non-native freshwater fish introductions. Journal of fish biology, 76, 751–786.

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.

Guo B, Zou M, Wagner A (2012) Pervasive indels and their evolutionary dynamics after the fish-specific genome duplication. Molecular biology and evolution, 29, 3005–3022.

Gurevitch J, Padilla DK (2004) Are invasive species a major cause of extinctions? Trends in ecology & evolution, 19, 470–474.

Hand BK, Hether TD, Kovach RP et al. (2015) Genomics and introgression: Discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. Current zoology, 61, 146–154.

Harris K, Nielsen R (2013) Inferring demographic history from a spectrum of shared haplotype lengths. PLoS genetics, 9, e1003521.

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.

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

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.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Molecular Ecology resources, 11, 117–122.

Hohenlohe PA, Bassham S, Etter PD et al. (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS genetics, 6, e1000862.

Hohenlohe PA, Day MD, Amish SJ et al. (2013) Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. Molecular Ecology, 22, 3002–3013.

Holopainen IJ, Aho J, Vornanen M, Huuskonen H (1997a) 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, Hyvärinen H (1985) Ecology and physiology of crucian carp (Carassius carassius L.) in small Finnish ponds with anoxic conditions in winter. Verhandlungen der Internationalen Vereinigung fur Theoretische und Angewandte Limnologie. International Association of Theoretical and Applied Limnology, 22, 2566–2570.

Holopainen IJ, Oikari A (1992) Ecophysiological effects of temporary acidification on crucian carp, Carassius carassius (L.): a case history of a forest pond in eastern Finland. Annales Zoologici Fennici, 29, 29–38.

Holopainen IJ, Tonn WM, Paszkowski CA (1997b) Tales of two fish: the dichotomous biology of crucian carp (Carassius carassius (L.)) in northern Europe. Annales zoologici Fennici, 34, 1–22.

Howard DJ (1993) Reinforcement: origin, dynamics, and fate of an evolutionary hypothesis. In: Hybrid zones and the evolutionary process (ed Harrison RG), pp. 46–69. Oxford University Press, New York, USA.

Hubbs CL (1955) Hybridization between Fish Species in Nature. Systematic zoology, 4, 1–20.

Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. Molecular Ecology resources, 8, 3–17.

Hufton AL, Groth D, Vingron M et al. (2008) Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. Genome research, 18, 1582–1591.

Hulme PE (2009) Trade, transport and trouble: managing invasive species pathways in an era of globalization. The Journal of applied ecology, 46, 10–18.

Hulme PE, Bacher S, Kenis M et al. (2008) Grasping at the routes of biological invasions: a framework for integrating pathways into policy. The Journal of applied ecology, 45, 403–414.

Hänfling B (2007) Understanding the establishment success of non-indigenous fishes: lessons from population genetics. Journal of fish biology, 71, 115–135.

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.

Iguchi K, Yamamoto G, Matsubara N, Nishida M (2003) Morphological and genetic analysis of fish of a Carassius complex (Cyprinidae) in Lake Kasumigaura with reference to the taxonomic status of two all-female triploid morphs. Biological journal of the Linnean Society. Linnean Society of London, 79, 351–357.

IUCN (2012) IUCN RED LIST CATEGORIES AND CRITERIA. Gland, Switzerland and Cambridge, UK.

Janko K, Bohlen J, Lamatsch D et al. (2007) The gynogenetic reproduction of diploid and triploid hybrid spined loaches (Cobitis: Teleostei), and their ability to establish successful clonal lineages—on the evolution of polyploidy in asexual vertebrates. Genetica, 131, 185–194.

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, 1–10.

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.

Jones A (1978) A note on the fish remains. In: Southwark Excavations 1972–74 Joint Publication No 1. (eds Bird J, Graham AH, Sheldon H, Townend P). The London and Middlesex Archaeological Society with the Surrey Archaeological Society.

Kasahara M, Naruse K, Sasaki S et al. (2007) The medaka draft genome and insights into vertebrate genome evolution. Nature, 447, 714–719.

Keller LF, Waller DM (2002) Inbreeding effects in wild populations. Trends in ecology & evolution, 17, 230–241.

Kolar CS, Lodge DM (2002) Ecological predictions and risk assessment for alien fishes in North America. Science, 298, 1233–1236.

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.

Lamer JT, Sass GG, Boone JQ et al. (2014) Restriction site-associated DNA sequencing generates high-quality single nucleotide polymorphisms for assessing hybridization between bighead and silver carp in the United States and China. Molecular Ecology resources, 14, 79–86.

Lampert KP, Lamatsch DK, Fischer P et al. (2007) Automictic reproduction in interspecific hybrids of poeciliid fish. Current biology: CB, 17, 1948–1953.

Lampert KP, Schartl M (2008) The origin and evolution of a unisexual hybrid: Poecilia formosa. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 363, 2901–2909.

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.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature methods, 9, 357–359.

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.

Lavoie DM, Smith LD, Ruiz GM (1999) The Potential for Intracoastal Transfer of Non-indigenous Species in the Ballast Water of Ships. Estuarine, coastal and shelf science, 48, 551–564.

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

Lever C (1977) Naturalized animals of the British Isles. Hutchinson & Co Limited, London.

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

Li J-T, Hou G-Y, Kong X-F et al. (2015) The fate of recent duplicated genes following a fourth-round whole genome duplication in a tetraploid fish, common carp (Cyprinus carpio). Scientific reports, 5.

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

Lindström T, Brown GP, Sisson SA, Phillips BL, Shine R (2013) Rapid shifts in dispersal behavior on an expanding range edge. Proceedings of the National Academy of Sciences of the United States of America, 110, 13452–13456.

Lockwood JL, Cassey P, Blackburn T (2005) The role of propagule pressure in explaining species invasions. Trends in ecology & evolution, 20, 223–228.

Luca F, Hudson RR, Witonsky DB, Di Rienzo A (2011) A reduced representation approach to population genetic analyses and applications to human evolution. Genome research, 21, 1087–1098.

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.

Lynch M (1997) Inbreeding depression and outbreeding depression. In: Genetic Effects of Straying of Non-native Hatchery Fish into Natural Populations: Proceedings of the Workshop. US Dept. Comm., NOAA Tech Memo. NMFS-NWFSC-30, pp. 59–67.

Mack RN, Simberloff D, Lonsdale WM et al. (2000) Biotic Invasions: Causes, Epidemiology, Global Consequences, and Control. Ecological applications: a publication of the Ecological Society of America, 10, 689–710.

Madeira MJ, Gómez-moliner BJ, Barbé AM (2005) Genetic introgression on freshwater fish populations caused by restocking programmes. Biological invasions, 7, 117–125.

Maes G, Van Houdt J, Volckaert F et al. (2003) Onderzoek van het geslacht Carassius in het Vlaamse Gewest.

Maitland PS (1972a) A key to the freshwater fishes of the British Isles: with notes on their distribution and ecology. Biological Association, Ambleside.

Maitland PS (1972b) A key to the freshwater fishes of the British Isles: with notes on their distribution and ecology. Freshwater Biological Association.

Mallet J (2005) Hybridization as an invasion of the genome. Trends in ecology & evolution, 20, 229–237.

Marchetti MP, Moyle PB, Levine R (2004) Invasive species profiling? Exploring the characteristics of non-native fishes across invasion stages in California. Freshwater biology, 49, 646–661.

Martinsen GD, Whitham TG, Turek RJ, Keim P (2001) Hybrid populations selectively filter gene introgression between species. Evolution; international journal of organic evolution, 55, 1325–1335.

Mastretta-Yanes A, Arrigo N, Alvarez N et al. (2014) Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. Molecular Ecology resources.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular phylogenetics and evolution, 66, 526–538.

McCormack JE, Maley JM, Hird SM et al. (2012) Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. Molecular phylogenetics and evolution, 62, 397–406.

McInerney CE, Maurice L, Robertson AL et al. (2014) The ancient Britons: groundwater fauna survived extreme climate change over tens of millions of years across NW Europe. Molecular Ecology, 23, 1153–1166.

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.

Mezhzherin SV, Lisetskii IL (2004) The genetic structure of European goldfish *Carassius auratus* s. lato (Cyprinidae) in Ukranian water bodies: an analysis of bisexual samples. Biological Bulletin, 6, 689-697.

Miller RR (1989) Extinctions of North American fishes during the past century. Fisheries, 14, 22–38.

Miller SA, Crowl TA (2006) Effects of common carp (Cyprinus carpio) on macrophytes and invertebrate communities in a shallow lake. Freshwater biology, 51, 85–94.

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.

Mooney HA, Cleland EE (2001) The evolutionary impact of invasive species. Proceedings of the National Academy of Sciences of the United States of America, 98, 5446–5451.

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, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. Molecular Ecology resources, 9, 66–73.

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.

Muhlfeld CC, Kalinowski ST, McMahon TE et al. (2009) Hybridization rapidly reduces fitness of a native trout in the wild. Biology letters, 5, 328–331.

Muhlfeld CC, Kovach RP, Jones LA et al. (2014) Invasive hybridization in a threatened species is accelerated by climate change. Nature climate change, 4, 620–624.

Muller HJ (1942) Isolating mechanisms, evolution and temperature. Biol Symp, 6, 71–125.

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, NY 10023.

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.

Oertli B, Joye DA, Castella E et al. (2002) Does size matter? The relationship between pond area and biodiversity. Biological conservation, 104, 59–70.

Ogden R, Gharbi K, Mugue N et al. (2013) Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. Molecular Ecology, 22, 3112–3123.

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York, New York, USA.

Ohno S (1973) Fish and nature's extensive experiments with gene duplication. In: Genetics and Mutagenesis of Fish, pp. 213–219. Springer.

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.

Orr HA, Presgraves DC (2000) Speciation by postzygotic isolation: forces, genes and molecules. BioEssays: news and reviews in molecular, cellular and developmental biology, 22, 1085–1094.

Ouborg NJ (2009) Integrating population genetics and conservation biology in the era of genomics. Biology letters, 6, 3–6.

Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma RK, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. Trends in genetics: TIG, 26, 177–187.

Oyler-McCance SJ, Fedy BC, Landguth EL (2012) Sample design effects in landscape genetics. Conservation genetics , 14, 275–285.

Palsbøll PJ, Bérubé M, Allendorf FW (2007) Identification of management units using population genetic data. Trends in ecology & evolution, 22, 11–16.

Pante E, Abdelkrim J, Viricel A et al. (2015) Use of RAD sequencing for delimiting species. Heredity, 114, 450–459.

Papoušek I, Vetešník L, Halačka K et al. (2008) Identification of natural hybrids of gibel carp Carassius auratus gibelio (Bloch) and crucian carp Carassius carassius (L.) from lower Dyje River floodplain (Czech Republic). Journal of fish biology, 72, 1230–1235.

Parmesan C, Ryrholm N, Stefanescu C et al. (1999) Poleward shifts in geographical ranges of butterfly species associated with regional warming. Nature, 399, 579–583.

Patterson AH (1905) Nature in Eastern Norfolk. Methuen & Co., London, UK.

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

Peischl S, Kirkpatrick M, Excoffier L (2015) Expansion load and the evolutionary dynamics of a species range. The American naturalist, 185, E81–93.

Pennant T (1766) British Zoology. Benjamin White, London.

Perry WL, Lodge DM, Feder JL (2002) Importance of hybridization between indigenous and nonindigenous freshwater species: an overlooked threat to North American biodiversity. Systematic biology, 51, 255–275.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PloS one, 7, e37135.

Petit RJ (2004) Biological invasions at the gene level. Diversity and Distributions, 10, 159–165.

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.

Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS genetics, 8, e1002967.

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

Puritz JB, Hollenbeck CM, Gold JR (2014) dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ, 2, e431.

R Core Team (2013) R: a language and environment for statistical computing. ISBN 3-900051-07-0, Vienna, Austria.

R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation 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.

Rasmussen H (1959) Fish ponds and fish rearing. In: Kulturhistoriskt lexikon för nordisk medeltid, pp. 307–309.

Reed D, Frankham R (2003) Correlation between fitness and genetic diversity. Conservation biology: the journal of the Society for Conservation Biology.

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

Reznick DN, Ghalambor CK (2001) The population ecology of contemporary adaptations: what empirical studies reveal about the conditions that promote adaptive evolution. Genetica, 112-113, 183–198.

Rhymer JM, Simberloff D (1996) Extinction by Hybridization and Introgression. Annual review of ecology and systematics, 27, 83–109.

Ricciardi A, Neves RJ, Rasmussen JB (1998) Impending extinctions of North American freshwater mussels (Unionoida) following the zebra mussel (Dreissena polymorpha) invasion. The Journal of animal ecology, 67, 613–619.

Rieseberg LH, Archer MA, Wayne RK (1999) Transgressive segregation, adaptation and speciation. Heredity, 83, 363–372.

Ripley BD (1991) Statistical inference for spatial processes. Cambridge university press.

Rolfe P (2010) Crock of Gold. M Press (Media) Ltd, Romford, UK.

Rosenfield JA, Nolasco S, Lindauer S, Sandoval C, Kodric-Brown A (2004) The Role of Hybrid Vigor in the Replacement of Pecos Pupfish by Its Hybrids with

Sheepshead Minnow. Conservation biology: the journal of the Society for Conservation Biology, 18, 1589–1598.

Roy HE, Peyton J, Aldridge DC et al. (2014) Horizon scanning for invasive alien species with the potential to threaten biodiversity in Great Britain. Global change biology, 20, 3859–3871.

Rubin BER, Ree RH, Moreau CS (2012) Inferring phylogenies from RAD sequence data. PloS one, 7, e33394.

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 , 380–383, 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.

Schlaepfer MA, Sax DF, Olden JD (2011a) The Potential Conservation Value of Non-Native Species. Conservation biology: the journal of the Society for Conservation Biology, 25, 428–437.

Schlaepfer MA, Sax DF, Olden JD (2011b) The potential conservation value of non-native species. Conservation biology: the journal of the Society for Conservation Biology, 25, 428–437.

Schranz ME, Mohammadin S, Edger PP (2012) Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Current opinion in plant biology, 15, 147–153.

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

Scoble J, Lowe AJ (2010) A case for incorporating phylogeography and landscape genetics into species distribution modelling approaches to improve climate adaptation and conservation planning. Diversity & distributions, 16, 343–353.

Scribner KT, Page KS, Bartron ML (2000) Hybridization in freshwater fishes: a review of case studies and cytonuclear methods of biological inference. Reviews in fish biology and fisheries, 10, 293–323.

Seehausen O (2004) Hybridization and adaptive radiation. Trends in ecology & evolution, 19, 198–207.

Simberloff D (2009) The Role of Propagule Pressure in Biological Invasions. Annual review of ecology, evolution, and systematics, 40, 81–102.

Simberloff D, Martin J-L, Genovesi P et al. (2013) Impacts of biological invasions: what's what and the way forward. Trends in ecology & evolution, 28, 58–66.

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.

Smartt J (2007) A possible genetic basis for species replacement: preliminary results of interspecific hybridisation between native crucian carp Carassius carassius (L.) and introduced goldfish Carassius auratus (L.). Aquatic Invasions, 2, 59–62.

Smith GR (1992) Introgression in Fishes: Significance for Paleontology, Cladistics, and Evolutionary Rates. Systematic biology, 41, 41–57.

Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. Nature reviews. Genetics, 14, 404–414.

Sovic MG, Fries AC, Gibbs HL (2015) AftrRAD: a pipeline for accurate and efficient *de novo* assembly of RADseq data. Molecular Ecology resources, 15, 1163–1171.

Stapley J, Reger J, Feulner PGD et al. (2010) Adaptation genomics: the next generation. Trends in ecology & evolution, 25, 705–712.

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.

Sémon M, Wolfe KH (2007a) Rearrangement rate following the whole-genome duplication in teleosts. Molecular biology and evolution, 24, 860–867.

Sémon M, Wolfe KH (2007b) Consequences of genome duplication. Current opinion in genetics & development, 17, 505–512.

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. Zeitschrift fur angewandte Ichthyologie = 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.

The Heliconius Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature, 487, 94–98.

The IUCN Red List of Threatened Species (2014) IUCN.

Tin MMY, Rheindt FE, Cros E, Mikheyev AS (2014) Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. Molecular Ecology resources.

Turner LM, White MA, Tautz D, Payseur BA (2014) Genomic networks of hybrid sterility. PLoS genetics, 10, e1004162.

Twyford AD, Ennos RA (2011) Next-generation hybridization and introgression. Heredity, 108, 179–189.

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

Utter F (2000) Patterns of subspecific anthropogenic introgression in two salmonid genera. Reviews in fish biology and fisheries, 10, 265–279.

Verspoor E, Hammart J (1991) Introgressive hybridization in fishes: the biochemical evidence. Journal of fish biology, 39, 309–334.

Vitousek PM, D'Antonio CM, Loope LL, Westbrooks R, Others (1996) Biological invasions as global environmental change. American scientist, 84, 468–478.

Vitule JRS, Freire CA, Simberloff D (2009) Introduction of non-native freshwater fish can certainly be bad. Fish and fisheries , 10, 98–108.

Vornanen M, Stecyk JAW, Nilsson GE (2009) Chapter 9 The Anoxia-Tolerant Crucian Carp (Carassius Carassius L.). In: Fish Physiology (ed Jeffrey G. Richards APFACJB), pp. 397–441. Academic Press.

Vrijenhoek RC (1994) Unisexual Fish: Model Systems for Studying Ecology and Evolution. Annual review of ecology and systematics, 25, 71–96.

Wagner CE, Keller I, Wittwer S, Selz OM (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Molecular Ecology, 22, 787–798.

Wang N, Thomson M, Bodles W (2013) Genome sequence of dwarf birch (Betula nana) and cross-species RAD markers. Molecular.

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution; international journal of organic evolution, 38, 1358–1370.

Wheat CW (2010) Phosphoglucose isomerase (Pgi) performance and fitness effects among Arthropods and its potential role as an adaptive marker in conservation genetics. Conservation genetics , 11, 387–397.

Wheeler A (1977) The Origin and Distribution of the Freshwater Fishes of the British Isles. Journal of biogeography, 4, 1–24.

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

Wilde GR, Echelle AA (1997) Morphological variation in intergrade pupfish populations from the Pecos River, Texas, USA. Journal of fish biology, 50, 523–539.

Williamson M, Fitter A (1996) The Varying Success of Invaders. Ecology, 77, 1661–1666.

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, Köln, Weimar.

Woodforde J, Winstanley RL, Jameson P (2008) The Diary of James Woodforde: Norfolk 1778-1779. Parson Woodforde Society, Norfolk, UK.

Worthington EB, Lowe-McConnell R (1994) African Lakes Reviewed: Creation and Destruction of Biodiversity. Environmental conservation, 21, 199–213.

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.

Xiao J, Zou T, Chen Y et al. (2011) Coexistence of diploid, triploid and tetraploid crucian carp (Carassius auratus) in natural waters. Biomedical chromatography: BMC.

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.

Yu Y, Andrés JA (2014) Genetic architecture of contemporary adaptation to biotic invasions: quantitative trait locus mapping of beak reduction in soapberry bugs. G3 , 4, 255–264.

Yue GH, David L, Orban L (2007) Mutation rate and pattern of microsatellites in common carp (Cyprinus carpio L.). Genetica, 129, 329–331.

Yue GH, Orban L (2002) Polymorphic microsatellites from silver crucian carp (Carassius auratus gibelio Bloch) and cross-amplification in common carp (Cyprinus carpio L.). Molecular Ecology notes, 2, 534–536.

Yue GH, Orban L (2004) Characterization of microsatellites located within the genes of goldfish (Carassius auratus auratus). Molecular Ecology notes, 4, 404–405.

Zheng W, Stacey NE, Coffin J, Strobeck C (1995) Isolation and characterization of microsatellite loci in the goldfish Carassius auratus. Molecular Ecology, 4, 791–792.