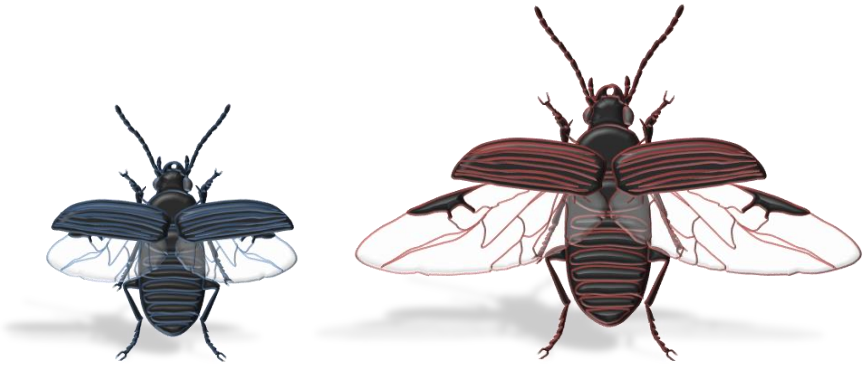
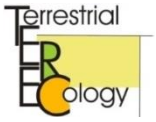
Genomic and behavioral aspects of recurrent ecotypic divergence in the wing polymorphic beetle *Pogonus chalceus*

Steven M. Van Belleghem
Ghent University | Faculty of Sciences



Thesis submitted in fulfillment of the requirements for the degree of Doctor (PhD) in Sciences, Biology
Proefschrift voorgelegd tot het behalen van de graad van Doctor in de Wetenschappen, Biologie



Date of the public defense: 31 October 2014

ISBN: 978-94-6197-235-4

The research presented in this thesis was conducted at the Terrestrial Ecology Unit (TEREC), Ghent University and the Royal Belgian Institute of Natural Sciences (RBINS), Brussels, Belgium and financially supported by:

- Research Foundation Flanders (FWO), Brussels, Belgium
- Belgian Science Policy (BELSPO), Brussels, Belgium

Please refer to this work as:
Van Belleghem SM (2014) Genomic and behavioral aspects of recurrent ecotypic divergence in the wing polymorphic beetle Pogonus chalceus. PhD thesis, Ghent University.

PROMOTERS

Prof. Dr. Frederik Hendrickx	(90%, RBINS and Ghent University, Belgium)
Prof. Dr. Luc Lens	(10%, Ghent University, Belgium)

READING COMMITTEE

Prof. Dr. Thierry Backeljau	(KBIN-IRSNB and University of Antwerp, Belgium)
Prof. Dr. Joost Raeymaekers	(Ghent University, KU Leuven, Belgium and Universität Basel, Switzerland)
Prof. Dr. Dick Roelofs	(VU University Amsterdam, The Netherlands)

EXAMINATION COMMITTEE

Prof. Dr. Thierry Backeljau	(RBINS and University of Antwerp Belgium)
Prof. Dr. Dries Bonte	(Ghent University, Belgium)
Prof. Dr. Frederik Hendrickx	(RBINS and Ghent University, Belgium)
Prof. Dr. Luc Lens	(Ghent University and Belgium)
Prof. Dr. Joost Raeymaekers	(Ghent University, KU Leuven, Belgium and Universität Basel, Switzerland)
Prof. Dr. Dick Roelofs	(VU University Amsterdam, The Netherlands)
Prof. Dr. Carl Vangestel	(RBINS and Ghent University, Belgium)

DANKWOORD

"In closing, I'd like to thank you. "What's that?" you say. "Me, thanking you?" No, it's not a misprint. I enjoyed writing this book as much as you enjoyed reading it. The End." (Burns 1991)

Het lijkt wel een oogwenk die laatste jaren. Echter waren er dankzij vrienden en collega's zoveel amusante en leerrijke momenten dat ik niet weet waar te beginnen met bedanken. Wat ik wel weet is dat ik nagenoeg niets anders zou doen indien ik het nog eens mocht overdoen. Het is overbodig om te zeggen dat het werk in dit doctoraat verre van alleen mijn verdienste is en ik zou dan ook graag een paar personen expliciet willen bedanken.

In de eerste plaats zou ik Frederik willen bedanken. Frederik, mijn dankbaarheid voor u beknopt uitdrukken is op zijn zachtst gezegd een heuse uitdaging. Ondertussen heb ik al sinds begin 2008 het genoegen om onder uw begeleiding te werken. Het is verwonderlijk hoe uw enthousiasme en geduld ten opzichte van mij sindsdien enkel is toegenomen. Ik zou durven zeggen dat dit komt doordat we een gelijkaardige interesse en verwondering hebben wat betreft het bestuderen van evolutie. Ik had hierbij vaak het gevoel dat we samen een pad bewandelden en ons onderweg soms vragen stelden die nu al dom lijken. Het gevoel van appreciatie en samenwerking dat ik hierin kreeg, heeft altijd zeer motiverend gewerkt. Echter moet ik zeer zeker erkennen dat jij mij op elke stap van het doctoraat in de juiste richting hebt gestuurd en ik nog niet de helft zo ver zou staan of zou begrepen hebben zonder uw hulp. De kansen die ik heb gekregen, van Galápagos tot dure experimenten, zijn te kostbaar om met woorden te omschrijven en hebben ervoor gezorgd dat ik (hopelijk) nog even in de wetenschap kan meedraaien. Ik hoop dat de toekomst niets dan meer samenwerking zal brengen. En ook dit nog eens: "You won't believe it, that's my supervisor!".

Zonder het voorgaande uitstekende werk van Konjev Desender en Hilde Dhuyvetter had dit proefschrift niet bestaan. Spijtig genoeg heb ik Konjev nooit kunnen ontmoeten. Naar verluidt had hij de werkijver om vroeg of laat de laatste *Pogonus* weg te vangen. Ik hoop dat hij deze resultaten graag zou gelezen hebben. Jean-Pierre Maelfait stond oorspronkelijk als co-promotor op de aanvraag van mijn doctoraat. Meer dan een paar korte ontmoetingen en vergaderingen waren niet nodig om een onvergetelijke indruk na te laten. Ik denk dat ik en anderen nog veel van hem hadden kunnen leren, maar ongetwijfeld heb ik via collega's en vrienden onrechtstreeks zijn motiverende en amicale invloed op werkethiek ondervonden.

Ik wil ook alle TEREC collega's van de afgelopen 5 jaar ontzettend bedanken. Bedankt: Luc, Dries, Hans, Viki, Charlie, Eduardo, Debbie, Laurence, Carl, Bdog, Brambo, Boeye,

Adriana, Fons, Celine, Davy, Cátia, Bram, Johan, Linda, Pieter, Irene, Helena, Valerie, Lies, Liesbeth, Alex, Alejandro, Maurice, Rein, Boris, Martijn, Christoph, Hannele, Lander, Tom, Greet, Annelies en Jasmijn. Luc en Dries, bedankt voor het creëren van een creatieve, motiverende en leuke werksfeer. Charlie, jij was eigenlijk gewoon de eerste om mij op de TEREC te ontvangen en de juiste richting aan te wijzen in de wondere wereld van het speciatie onderzoek. Van *Hogna's* tot Galápagos en veel te gekke feestje, ik heb veel geleerd en mij altijd ontzettend kunnen amuseren met u. Ook nogmaals sorry van die fles Jack Daniel's ("Is dat trouwens geen soort hond?") in je handtas. Viki, wat zou de TEREC zijn zonder u… Altijd paraat voor hulp in het labo, vervoer naar de Guérande en natuurlijk nachtelijke uitstappen. Thanks! Hans, bedankt om al die problemen in de keuken op te lossen. Boeye, bedankt om altijd alles onder controle te houden op de stages. Eerlijk toegegeven, door uw werk resultaten te zien kreeg ik vaak een stimulans om zelf een tandje bij te steken. Bedankt daarvoor en veel succes met toekomstige ondernemingen! Katrien, gelukkig was jij er altijd om de sfeer op te krikken wanneer al de rest thuis werkte. Is er iets wat jij niet kan breken…? Brambo, ik heb het greiten gemist de laatste twee jaren. De legende die wij kennen als Brambo is echter nog lang niet vergaan. Bdog, als ervaren bureau collega heb jij zeker ook een grote invloed gehad. Eduardo, succes bij de terugkeer naar Spanje. Het Belgisch klimaat zal net iets grauwer zijn zonder uw aanwezigheid. Angelica, bedankt voor raad en hulp op tijd en stond. Wanneer gaan we nog eens dansen in Club Wezon? Linda en Johan, bedankt voor de interesse en gezellige babbels al van toen ik nog aan mijn bachelor proef werkte. Lucien, bedankt voor de hulp bij het verzorgen van de kevers. Léa, good luck in Canada! Adriana, if you ever travel back in time, don't step on anything. Because even the slightest change can alter the future in ways you can't imagine. Alex, "whaaat?" you arriving at TEREC even resulted in capturing beetles in Aveiro. Thanks for all the nice dinners, drinks and (crazy) parties!

Daarnaast wil ik ook een reeks collega's vanop het KBIN bedanken. Carl, Viki en Charlie (opnieuw), Léon, Lore, Patrick, Karin, Arantza, Isa, Koen De Gelas, Séverine, Thierry, Valentina, Vanya, Wouter, Thomas, Stefan, en Jeremy, bedankt voor hulp in het labo en/of lachen tijdens de middagpauze. Gontran, Zoltan and Zohra, thanks for all the help in the lab and the fun environment! Filip, bedankt voor de tips en tricks voor het kweken van kevers. Carl, dankzij uw vervoeging op het KBIN kwam plots het genetische werk in een stroomversnelling. Ik denk niet dat hoofdstuk 5 met misschien wel de interessantste resultaten had bestaan als jij niet de weg had geplaveid.

Limno… bedankt voor de vele feestjes? Bedankt: Dirk, Thijs, Yoeri, Els (dankzij uw rollende ogen wist ik altijd of mijn moppen grappig waren of niet) en Jorunn (succes met je tandjes). Crazy Leen, nooit eerder was ik zo triest als toen jij de Limno verliet. Gelukkig liep je niet te ver weg en ging je uiteindelijk zelfs mee op reis naar Portugal waar ik echter hardhandig moest ingrijpen om de nachtrust te garanderen (en ook een beetje omdat je mijn broer een jurk had aangetrokken).

Elena en Katrien, bedankt voor de uitstekende hulp bij een groot deel van het werk. Katrien, ik denk dat uw werkijver en bewonderenswaardige zelfstandigheid u nog ver zal brengen en, wie weet, het *Pogonus* verhaal tot een volgend niveau tillen.

Lut, bedankt voor de geestige trips en vervoer naar Nieuwpoort en Dudzele. De vangsten daar zijn van cruciaal belang gebleken in het verhaal op de volgende bladzijden. Marc, bedankt voor de hulp bij het vangen van kevers. Francine Ronsse, bedankt voor de steun en interesse al van sinds het indienen van mijn master thesis.

Dick Roelofs, bedankt voor de gouden raad voornamelijk bij het begin van mijn doctoraat. Janine Mariën, bedankt voor hulp bij het qPCRen en RNA extracties. Jeroen en de Genomics Core (Leuven), bedankt bij de hulp van het opzetten van de eerste sequenerings projecten. Ook veel dank aan de mensen van het labo verouderingsfysiologie en moleculaire evolutie. Andy Vierstraete, bedankt voor de hulp bij het sequeneren (en een hele reeks andere labo technieken) en server problemen.

Ook niet te vergeten: Team Galápagos! Charlie, Carl, Wouter, Léon en Frederik, bedankt voor misschien wel de plezantste en meest leerrijke tijden ooit! Ik was altijd vol verwondering van het enthousiasme waarmee arthropoda kunnen gevangen worden na het trotseren van eindeloze lavavelden en duizenden hoogtemeters onder een loodrechte meedogenloze zon. En "Oh zo spijtig dat ze den haaientand hebben vervangen door een souvenirwinkel!". Henri Herrera, thanks for the hospitality and helping us survive the 'deadly islands'! Thanks guides and crue of el Pirata and CDRS!

Bij dezen zou ik ook graag alle leden van de lees- en examencommissie bedanken voor hun constructieve commentaar en suggesties.

(H)A(r|n)ne, Babera, Maes, Jasper, Kim, Uwe, voetbalmakkers en pintendrinkers, wat zou ik zijn zonder jullie? Bedankt voor alle uitspattingen en ontspanningen. Woorden schieten mij tekort…

Ma en Pa, bedankt voor alle steun sinds het begin van mijn bestaan, mij te laten uitslapen in het weekend en mijn egoïsme om te doen wat ik graag doe te steunen. Marnix, had ik ooit hard gestudeerd had jij niet zo een extreme bolleboos geweest die met alle pluimen ging lopen? Jonas, niemand heeft zoveel over mijn werk moeten aanhoren als jij en dankzij uw studies kon ik zelfs op les komen als het over moleculaire technieken ging. Ondertussen ben je misschien wel de meest ambitieuze weg ingeslaan. Marijke, Nand en <…in progress…>, bedankt!

Bedankt merci gracias obrigado danke thanks!

Gent, oktober 2014
Steven Van Belleghem

CONTENTS

GENERAL INTRODUCTION

Steven M. Van Belleghem

Speciation – the origin of new species – is a complex and challenging topic in which subtle arguments sometimes make large differences. First, speciation is difficult to observe in natural populations and is mostly studied indirectly. Moreover, inferring past processes from present patterns necessitates caution as different evolutionary scenarios are often difficult to disentangle. Second, speciation is subject to the interaction of both ecological and genetic processes. These processes are diverse and individual study systems allow addressing only subsets of the ecological and genetic factors involved. Extracting empirical evidence from these study systems is pivotal to our understanding of the ecological and genetic mechanisms underlying adaptive divergence and speciation and their relative importance. Third, difficulty in the distinction between the geographical mode and the mechanism of speciation impedes a simple understanding of the roles of selection, genetic drift and external factors on speciation as well as their interactions. Finally, identifying the genes underlying ecologically important traits and traits that are involved in reproductive isolation seems a simple goal, but it has proven to be a challenging task. Identifying these traits usually involves measurable phenotypes, large numbers of genome wide markers and specific breeding designs. However, recent advances in sequencing technologies provide novel opportunities to study adaptation and speciation from a genomic perspective. In the following introduction I elaborate on several topics from the vast literature on speciation which are relevant to the work presented in this thesis.

SPECIATION

"That mystery of mysteries" (Darwin 1959, p. 1)

It is well known that the diversity of life at every level of biological organization, including species, individual organisms and molecules, arose by common descent through a branching pattern of evolution (Darwin 1859). Over successive generations, changes in inherited characteristics become either more or less frequent in a population through the differential effect of the inherited traits on the reproductive success of organisms interacting with their environment. Charles Darwin considered this so called process of natural selection as the direct cause for the origin of new species and made little distinction between speciation and adaptation (Coyne & Orr 2004). However, this view had two main unsolved problems. First, it was interpreted and argued to be inadequate that Darwin emphasized on the evolution of species within single interbreeding populations (i.e. sympatric speciation) (Sulloway 1979). Second, the idea that speciation was gradual and driven by natural selection left it difficult to understand how a continuous process could create discontinuous entities such as species (Bateson 1922).
Reconciling Mendelism, biogeography and natural selection into the Modern Evolutionary Synthesis accounted for these issues. Mainly, within the advent of population genetics, Theodosius Dobzhansky described how gradual changes in allele frequencies could produce genetically and morphologically discrete entities (Dobzhansky 1935, 1937). Moreover, in his work, Dobzhansky recognized the importance of barriers to gene exchange for the coexistence of ecologically distinct forms. The study of the origin of these reproductive isolating mechanisms became essential for understanding the origin of species. Ernst Mayr elaborated on this work by defining the 'biological species concept' in which species are considered groups of interbreeding populations that are reproductively isolated from other groups (Mayr 1942). Studying the biogeography of speciation, Mayr was a strong advocate of the importance of genetic drift and allopatry in speciation, in which species arise from populations that are geographically isolated. Although both Dobzhansky and Mayr recognized the importance of ecology and natural selection in evolution, little attention was given to studying the connection between adaptation and speciation (Coyne & Orr 2004). Therefore, in the Modern Synthesis, a major emphasis was given to measuring and explaining genetic variation within species. Around the 1980s, protein gel electrophoresis and DNA sequencing further lead to advances in this field of empirical as well as theoretical population genetics. Molecular-genetic and phylogenetic advances have been essential in understanding the biogeography of speciation and the factors that

may promote the evolution of reproductive isolation. Eventually, these molecular techniques allowed tackling unresolved questions concerning the origin of species. The question how natural selection shapes reproductive isolation revived as geneticists attempted to identify the genes causing isolation, locate their positions on chromosomes, measure their relative effects and ultimately, identify their functions and the evolutionary forces that drove their divergence (Coyne & Orr 2004).

Understanding the processes – mutation, natural selection, sexual selection, genetic drift, gene flow and isolation – that drive the evolution of new species is one of the major themes in evolutionary biology. The brief history of evolutionary thinking emphasizes that the concept of speciation, how species originate and evolve, is a complex and vigorously discussed topic. Despite decades of empirical and theoretical research, consensus on many of these factors has not been reached, especially when considering the likelihood of speciation in the face of gene flow. The dominant view of allopatric speciation has become debated when empirical and theoretical insights suggested that species could evolve without geographic isolation and, hence, in the face of limited or even strong gene flow (e.g. Rice and Hostert 1993; Rundle et al. 2000; Dhuyvetter et al. 2007; Butlin et al. 2013). This debate has strongly stressed the interrelatedness of adaptation and speciation in recent years (i.e. ecological speciation; Nosil 2012).

REPRODUCTIVE ISOLATION

> *"We cannot study how species form until we determine what they are." (Coyne and Orr 2004, p. 25)*

To understand the nature of speciation, it is important to consider a useful concept of species. Coyne and Orr (2004) define species according to the biological species concept (Mayr 1942), but argue that "distinct species are characterized by substantial but not necessarily complete reproductive isolation". Hence, this definition allows limited gene flow. Resulting from this definition of species, the process of "the origin of species" requires the evolution of reproductive barriers which should maintain distinct groups even in the same area and if they occasionally hybridize. Therefore, speciation research largely focuses on the evolution of reproductive barriers. The point at which sympatric taxa should be called "species" is, however, arbitrary (Coyne & Orr 2004) and to some extent irrelevant as the goal of speciation research is to understand how coexisting populations evolve (McPhail 1994).

A wide range of reproductive isolating barriers have been discovered. These mechanisms include all "biological features of organisms that impede the exchange of genes with members of other populations" (Coyne & Orr 2004). Isolating mechanisms can take place pre- or postmating and can be classified as either extrinsic or intrinsic. Extrinsic barriers result from divergent selection and fitness reduction in hybrids is dependent on the environment. In case of intrinsic barriers, fitness reduction of hybrids is independent of the environment. In animals, important extrinsic postmating isolating barriers include immigrant inviability. Intrinsic postmating mechanisms often result from epistatic incompatibilities. Premating isolating mechanisms may evolve as a consequence of divergent sexual as well as natural selection. Premating mechanisms include mechanical (e.g. lack of mechanical fit), behavioral (e.g. mate choice) and ecological isolation. Ecological isolating barriers result from species' ecology and are thus considered direct byproducts of ecological divergence. These include habitat isolation (e.g. matching habitat choice; Edelaar et al. 2008) and temporal isolation (e.g. different breeding times; Friesen et al. 2007).
Divergence of a trait is only relevant to speciation if it contributes to reproductive isolation. The effects of genes that evolve differently among groups may be transmitted, as a by-product, pleiotropically to affect reproductive isolation (Coyne & Orr 2004, Rundle & Nosil 2005). For instance, adaptation to a new habitat may directly guarantee spatial isolation. Alternatively, diverged genes may be linked to genes involved in reproductive isolation (via linkage disequilibrium) yielding reproductive isolation as a by-product. Epistatic effects between genes may cause reproductive isolation when genes that evolved in one genetic background result in less fit hybrids in another genetic background.

Determining which reproductive barriers were involved in the initial reduction of gene flow and which evolutionary forces produced these barriers is a formidable task (Coyne & Orr 2004). First, current isolating barriers may not have been the most import in the initial stages of restricting gene flow. Second, several isolating mechanisms may act successively at multiple stages in the life history and the proportional effect strongly depends on the life history stage in which the mechanism takes place. Even though later barriers may have strong absolute effects, earlier-acting reproductive barriers will reduce gene flow proportionally more than later-acting barriers (Ramsey *et al.* 2003). Therefore, prezygotic barriers are often considered the most important factor for restricting gene flow (Kirkpatrick & Ravigné 2002). However, in the initial speciation process, postzygotic barriers may still have been significant (Coyne & Orr 2004, Seehausen *et al.* 2014).

SYMPATRIC SPECIATION

> *"One would think that it should no longer be necessary to devote much time to this topic, but past experience permits one to predict that the issue will be raised again at regular intervals." (Mayr 1963)*

DEFINITION

The splitting of a homogeneous population in two or more adaptive lines in the absence of a physical barrier, remains a contentious issue in evolutionary biology (Via 2001, Coyne & Orr 2004, Bolnick & Fitzpatrick 2007, Mallet *et al.* 2009, Pinho & Hey 2010). 'Sympatry' originally meant ''in the same geographical area'' (Poulton 1904). However, 'pure sympatric speciation' may be precisely defined as "the origin of an isolating mechanism (i.e. the evolution of a barrier to gene flow) among the members of an interbreeding population" (Futuyma & Mayer 1980). This is interpreted as speciation between two populations that show free migration (m) or complete panmixia (m = 0.5). This definition allows focusing on the actual mechanisms such as gene flow and selection parameters, rather than being concerned about the geography of speciation (Kirkpatrick & Ravigné 2002, Dieckmann *et al.* 2004). However, it has been argued that for natural populations this demic definition is unsuitable, because it is a theoretical end point of a continuum (Fitzpatrick *et al.* 2008) and it omits consideration of space (Mallet et al. 2009). For instance, consider situations where populations may begin to specialize on different resources or habitat characteristics. If these resources are not perfectly mixed, such populations will not fall within the scope of pure sympatry. In contrast, resources and habitats are often distributed patchy, leading to situation with reduced migration and gene flow ($m < 0.5$). These situations are more feasible in nature and are often called sympatric mosaics (Fitzpatrick *et al.* 2008, Mallet *et al.* 2009). Alternatively, 'divergence-with-gene-flow' may be considered as a spectrum of models with at its left extreme a single population with simultaneous selection for two opposing phenotypes (sympatry) and at its right extreme divergently selected geographically isolated subpopulations (Rice & Hostert 1993).

THE PROBLEM

A major stumbling block in the plausibility of sympatric speciation is the antagonism between natural selection and recombination (Felsenstein 1981). More precisely, in the absence of geographical isolation, interbreeding and recombination between differently selected populations will hamper the evolution of gene complexes that result in

reproductive isolation. For instance, when reproductive isolation results from habitat *preference* and habitat preference has a different genetic basis than the traits involved in *performance* (fitness) in those habitats, recombination will break down the association between alleles that increase performance in each habitat and alleles involved in the preference of those habitats. This would break down associations between alleles that contribute to the same trait (e.g. multiple loci involved in performance or habitat preference) as well as lead to allele combinations preferring habitats in which they have low performance. In contrast, in allopatric speciation scenarios, geographical isolation prevents recombination.

A second problem that is often put forward is coexistence (Coyne & Orr 2004). Populations have to sufficiently diverge in their resource use to coexist during and after the speciation process. However, when speciation results from disruptive natural selection this issue is readily solved as disruptive selection will result in ecological divergence. In case of disruptive sexual selection, the problem is more complex because of the absence of ecological divergence (Coyne & Orr 2004), but solutions have been proposed (M'Gonigle *et al.* 2012).

Solutions

> *"When migration rates and gene exchange is high, the initial restriction to gene exchange has to be caused not by geography or distance, but by biological features of the organisms." (Futuyma and Mayer 1980 in Coyne and Orr 2004)*

One often used example of sympatric speciation includes host race shifts in phytophagous insects in which genetic linkage between ecological specialization and reproductive isolation has been proposed (Hawthorne & Via 2001, Berlocher & Feder 2002). However, most work on sympatric speciation involves theoretical models incorporating assumptions about selection and the genetic architecture of genes involved in performance (niche adaptation) and assortative mating (e.g. Rice 1987; Fry 2003; Bolnick and Fitzpatrick 2007). Kirkpatrick & Ravigné (2002) dissected a set of important (and interacting) elements of speciation from these models (Figure 1).

SOURCE OF DISRUPTIVE SELECTION First, sympatric speciation requires a form of disruptive selection that creates a force that causes the evolution of reproductive isolation. This selection can result from spatial variation in fitness, frequency dependent selection or sexual disruptive selection. In case of disruptive natural selection, this

should result in races that differ in their ability to survive in different niches (Coyne & Orr 2004).

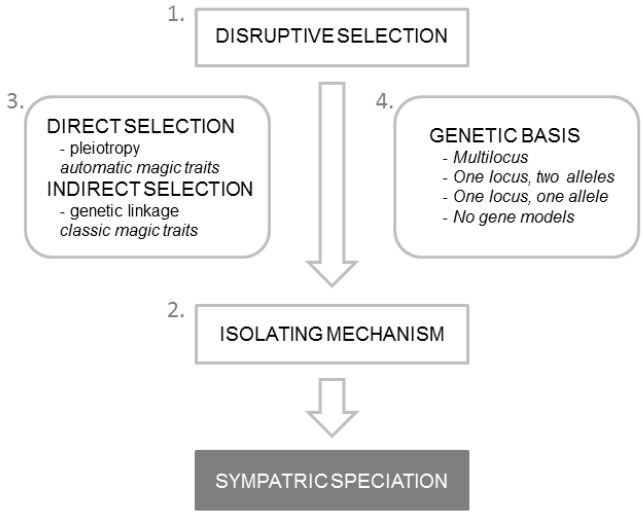


Figure 1. Elements involved in the evolution of sympatric speciation. Numbers refer to the discussion in the text.

ISOLATING MECHANISM Second, a mechanism is needed that reduces gene exchange among individuals with different traits. In sympatric speciation, prezygotic isolation mechanisms are argued to be the most important factor keeping populations separate (Kirkpatrick & Ravigné 2002). These mechanisms can depend on mating preference (e.g. female preference for display traits in males) or assortative mating resulting from niche preference (Edelaar *et al.* 2008, Edelaar & Bolnick 2012) or result from an automatic by-product of phenotypic variation in traits (e.g. differences in developmental timing or breeding time imposed by the environment; Friesen *et al.* 2007). In this sense, geography can be simply seen as another form of assortative mating (Kirkpatrick & Ravigné 2002).

LINK BETWEEN DISRUPTIVE SELECTION AND ISOLATING MECHANISM Third, there must exist a link between disruptive selection and the isolating mechanism for it to cause speciation. Using lab experiments, for instance, disruptive selection on habitat preference resulted in the evolution of reproductive isolation in *Drosophila melanogaster* (Rice 1985). Alternatively, body size is clearly under divergent selection between limnetic and benthic stickleback, and female stickleback prefer male stickleback closer to their own size (Nagel & Schluter 1998). In Darwin's finches beak size is under divergent

selection and affects reproductive isolation by means of song differences (Huber *et al.* 2007). Further, pleiotropy and linkage have been argued to provide a link between disruptive selection on color pattern and mate choice in mimetic *Heliconius* butterflies (Jiggins 2008) and between performance and habitat choice in phytophagous insects (Hawthorne & Via 2001). The effect of disruptive selection can be transmitted either directly or indirectly to an isolating mechanism. Direct selection may result in isolation if genes that affect fitness also directly influence mate preference or assortative mating (e.g. divergent selection on phenology or traits that produce isolation pleiotropically). As reproductive isolation evolves automatically as a result of selection from divergent environments, such traits are referred to as *automatic magic traits* (Servedio *et al.* 2011). Indirect selection includes scenarios in which genes that are affected by selection are in linkage disequilibrium with genes that cause isolation. Hence, in this latter case, selection is not directly on the genes that influence mate preference or assortative mating, but indirectly selects closely linked genes or against maladaptive gene combinations that lower the fitness of hybrids. These latter traits are generally referred to as *classic magic traits* (Servedio *et al.* 2011). Direct selection is more efficient than indirect selection in the evolution of isolating mechanisms as imperfect correlations between genes lowers the efficiency of linking selection and isolation.

GENETIC BASIS Fourth, the genetic basis of the isolation mechanism has to be taken into consideration (Felsenstein 1981). In '*multilocus*' models, changes in preference and assortative mating traits involves multiple genes (Johnson *et al.* 1996a). Models have shown that under strong selection, preference genes can become associated with unlinked performance genes resulting into reproductively isolated specialist populations (Dieckmann & Doebeli 1999, Fry 2003). However, in most multilocus models, close genetic linkage between the multiple loci is often opted for the evolution of reproductive barriers (Kondrashov & Mina 1986). In '*one-locus, two-alleles*' models, different alleles at the same locus promote mate preference or assortative mating among populations (Felsenstein 1981). For instance, individuals with allele A at locus X prefer habitat A, whereas individuals with allele B at locus X prefer habitat B. In this latter model, the alleles, however, must become associated with performance genes (i.e. build up genetic differences between populations) to promote sympatric speciation. Hence, two-allele mechanisms may suffer from the selection-recombination antagonism. Therefore, when the isolating mechanism depends on genetic variation in habitat preference, a close genetic linkage might be expected between genes involved in habitat preference and genes involved in ecological adaptation to that habitat which would facilitate the coevolution of preference and performance traits. This linkage has, for instance, been found in pea aphids specializing on different host plants (Hawthorne & Via 2001).

Alternatively, the traits subject to divergent selection may also contribute to non-random mating, resulting in an 'automatic' link between selection and assortative mating (~automatic magic traits; Servedio *et al.* 2011). In '*one-locus, one-allele*' models, the same allele increases isolation in both populations (e.g. an allele that causes animals of similar size to mate with each other, or philopatric behavior) solving the selection-recombination antagonism (Felsenstein 1981). However, this model requires that there is already differentiation between the populations (e.g. genetic variation for size). Direct genetic evidence for a one-allele assortative mating locus has been provided for sympatric fruit fly species, *Drosophila pseudoobscura* and *D. persimilis*. For these sympatric species, serial backcrosses allowed the reciprocal introgression of the so called *Coy-2* chromosomal region between the two fruit fly species and demonstrated that the same allele at the *Coy-2* chromosome region confers assortative mating in both a *D. pseudoobscura* and *D. persimilis* genetic background (Ortíz-Barrientos & Noor 2005). Another speciation mechanism that can be interpreted as one-allele mechanism includes learned habitat preferences (e.g. natal habitat experience), as alleles that strengthen learning necessarily have the same effect in both habitats (Davis & Stamps 2004, Beltman & Metz 2005). Finally, in '*no-gene models*', there are no genes that directly affect mate preference or assortative mating, but isolation pleiotropically results from genes affected by selection (Rice 1987, Servedio *et al.* 2011). Imagine for instance disruptive selection for different flowering time imposed by a different environment.

The distinction between the genetic models is orthogonal to the distinction between direct and indirect selection; all combinations of factors are possible and will have different effectiveness in producing sympatric speciation (see Kirkpatrick and Ravigné 2002; Servedio *et al.* 2011) and, moreover, multiple mechanisms of non-random mating may interact (Thibert-Plante & Gavrilets 2013). Using these elements, the models indeed show the possibility of sympatric speciation and allow exploring the set of assumptions for which it is most likely. Moreover, modeling the selection-recombination antagonism results in some explicit predictions of the genomic architecture of adaptation, such as few and linked loci with large effects (Yeaman & Whitlock 2011) and even selection for chromosomal rearrangements during the process of local adaptation as genomic architectures that eliminate or decrease recombination are expected to facilitate coupling and hence adaptation in multiple loci and speciation (Yeaman 2013).

ECOLOGICAL SPECIATION

A related concept that focuses explicitly on the mechanism of speciation concerns 'ecological speciation', defined as "the process by which barriers to gene flow evolve between populations as a result of ecologically based divergent selection between habitats" (Nosil 2012). Ecological speciation differs from sympatric speciation in that it can occur under any mode of geographical isolation between populations as long as divergent selection drives the divergence process (Schluter 2001, Rundle & Nosil 2005). Hence, ecological speciation mainly excludes allopatric divergence scenarios in which the evolution of reproductive isolation largely results from genetic drift (Schluter 2001), mutation-order speciation (Schluter 2009) and cases in which selection is not necessarily ecologically based such as sexual conflict or runaway sexual selection (e.g. Sauer and Hausdorf 2009). However, in this concept the geography is still important to consider as it can affect the source of divergent selection and rates of gene flow (Nosil 2012). Furthermore, different geographical contexts might alternate and influence the course of ecological speciation (e.g. Feder *et al*. 2005). Like most models of sympatric speciation by disruptive selection, ecological speciation presumes the existence of a direct link between ecologically based divergent selection and reproductive isolation (Hendry 2009, Faria *et al.* 2014). Moreover, focusing on approaches that ask how these reproductive barriers evolve and are selected is of major interest to advance the understanding of the field of speciation (Faria *et al.* 2014). An interesting implication is that when ecological speciation occurs, habitat and temporal isolation as forms of reproductive isolation are expected because adaptation to different environments will itself generate selection favoring individuals with appropriate habitat preferences and developmental timing (Nosil 2012).

THE SPECIATION CONTINUUM

Although speciation can be fast and sudden such as in polyploidy speciation in plants (Rieseberg & Willis 2007), the speciation process is generally argued to be a continuous process with different stages (Darwin 1859, Wu 2001a, Hendry *et al.* 2009, Nosil 2012) and the frequency of hybridization gradually decreasing with genetic distance (Mallet *et al.* 2007). A central task of speciation studies is to reconstruct the sequence in which different barriers evolved in order to distinguish between causes and consequences of speciation. However, in natural populations the continuum is difficult to reconstruct from single study systems (Nosil 2012, Seehausen *et al.* 2014). At best, comparing different populations within a single species pair that vary in their degree of reproductive isolation can provide valuable insights about transitions along the

continuum (e.g. Peccoud et al. 2009; Seehausen 2009). In the light of ecological speciation, one might recognize several states along this continuum (Hendry 2009). These states range from (i) continuous adaptive variation without reproductive isolation, to (ii) discontinuous adaptive divergence with minor reproductive isolation, up to (iii) adaptive differences with strong but reversible reproductive isolation and, finally, (iv) irreversible reproductive isolation. When gene flow is ample, several 'cluster-generating' factors are thought to be involved in the transition between the initial stages of the evolution of reproductive isolation, such as strong disruptive selection and assortative mating.

Further, if speciation progresses along a continuum, how far does it proceed and what factors determine the extent of progression? Partial reproductive isolation may represent a stable outcome maintained at a balance between selection and gene flow (Matessi *et al.* 2001, Gavrilets 2003, Bolnick & Fitzpatrick 2007, Nosil 2012). Alternatively, speciation might be driven by ongoing feedback loops between selection and gene flow in which adaptive divergence reduces gene flow and lowered gene flow allows adaptive divergence and vice versa (Räsänen & Hendry 2008, Gourbière & Mallet 2010). This would lead to an increasingly higher degree of reproductive isolation. Moreover, Nosil (2012) argues that in essence partial reproductive isolation will never be completely stable, but results from the timescale examined. More precisely, reproductive isolation will appear partial if (i) speciation involves only a few genes of large effect and the waiting time for mutations causing increased reproductive isolation is long (Bolnick & Near 2005) and (ii) if the rate of increasing reproductive isolation slows down in the latter stages of speciation (Gourbière & Mallet 2010). The latter point may occur when initially reduced gene flow lowers the selection strength for premating isolation (i.e. reinforcement) at later stages. Finally, the temporal stability of divergent natural selection is an important factor to consider (Siepielski *et al.* 2009), because over long time scales, the direction and strength of selection may fluctuate and influence the evolution of adaptations and reproductive isolation. One example is selection for reduced dispersal and increased reproduction over short time scales, but selection for higher dispersal ability at longer time scales when the probability of habitat changes and extinction increases (Olivieri *et al.* 1995, Mathias *et al.* 2001, Roff & Fairbairn 2007).

The strength of selection and the nature of ecological shifts affect how far speciation can proceed (Nosil *et al.* 2009b). However, how a finite amount of selection is distributed across a few versus many traits or genes may also be important. The effect of selection on few or multiple genes on speciation is subject to discussion, but Nosil (2012) summarizes that (i) strong selection on a few genes better allows adaptive divergence in the face of gene flow, but is expected to cause little correlated response (e.g. in linked genes or traits causing genetic incompatibilities) and will, therefore, result more often in

single trait polymorphisms rather than speciation. Alternatively, (ii) selection on multiple genes may be too weak to overcome gene flow, but the combined effect of reproductive mechanisms might be strong and correlated responses causing reproductive isolation as by-product may increase under widespread genomic divergence. Furthermore, it can be argued that initial strong selection on a single or few traits may provide the onset for divergence in multiple genes and traits, converting single trait polymorphisms to speciation (McKinnon & Pierotti 2010, Nosil 2012).

The former issue relates to the major question about the genomic architecture of incipient speciation and how this architecture either facilitates or impedes further divergence (Feder *et al.* 2012b). First, as discussed above, several genetic factors such as pleiotropy and one-allele assortative mating mechanisms may promote speciation. Alternatively, indications of widespread genomic divergence in incipient diverging species in the face of gene flow (e.g. Lawniczak et al. 2010; Michel et al. 2010; Parchman et al. 2013) have triggered the hypothesis that describes a process by which physical linkage of gene regions and strong divergent selection can reduce gene exchange for large genomic regions (Feder & Nosil 2010, Via 2012). This so called 'Divergence Hitchhiking' process is suggested to allow new mutations, with even weak effect, to differentiate owing to locally reduced gene exchange at the few tightly linked already diverged genes. In later stages of the divergence process, genetic divergence across the entire genome, even for loci unlinked to those under selection, may become facilitated by global reductions in gene flow caused by genome-wide selection (i.e. Genomic Hitchhiking; Feder, Gejji, et al. 2012; Feder, Egan, et al. 2012). However, the detection of widespread genomic divergence may also be pronounced by other factors that lead to a high occurrence of false positives in genomic scans such as neutral mutations that arise in the front of a wave of expansion (Excoffier & Ray 2008) and correlated coancestry in highly structured populations (Bierne *et al.* 2013, Fourcade *et al.* 2013). Furthermore differences in recombination rates may strongly effect the heterogeneity of divergence along the genome (Roesti *et al.* 2012).

GENETICS OF ADAPTATION

> *"Any locus under divergent natural selection between parental environments contributes to immigrant inviability and therefore may contribute to speciation." (Schluter & Conte 2009)*

Satisfying the criteria to identify those genes whose divergence made a significant contribution to the evolution of reproductive isolation between populations (i.e. speciation genes) is a daunting task (Orr *et al.* 2004, Nosil & Schluter 2011). First, adaptation and ecological divergence are not a strict prerequisite for speciation (e.g. genetic drift and mutation-order speciation). Second, the persistence of species differences in geographical proximity typically requires the evolution of prezygotic barriers as divergent adaptation rarely causes sufficient reproductive isolation on its own (Seehausen *et al.* 2014). However, in many cases there will be considerable association between factors that prevent gene flow between sympatric species and traits that are involved in ecological divergence that allow coexistence between species (Coyne & Orr 2004). Therefore, as natural selection is widely considered a major force in speciation (see Wu 2001b; Nosil 2012), understanding the genetic architecture and ecological mechanisms involved in the evolution of adaptations is a prerequisite in the study of speciation.

ECOLOGICALLY IMPORTANT GENES

In contrast to the complex discussions of speciation, textbook examples of adaptation, such as melanism in the peppered moth (Saccheri *et al.* 2008, Van't Hof *et al.* 2011), are often presented in a straightforward manner. Genes mutate and advantageous mutations spread in a population. As time since its origin progresses, nucleotide and protein sequence divergence will increase in agreement with neutral theory and hypothesis about the molecular clock (Kimura 1983). The expression of the dark and light color morphs in the peppered moth is controlled by a polymorphism at a single locus, evolved only once and carries a signature of recent strong selection (Grant 2004, Van't Hof *et al.* 2011). Unraveling this genetic architecture of adaptations is crucial for understanding adaptation dynamics, but becomes far more complicated when adaptation involves complex evolutionary histories, multiple loci and epistatic interactions (e.g. Steiner *et al*. 2007; Linnen *et al*. 2013).

Identifying genes underlying ecologically important traits allows addressing a host of questions relevant to genetics, ecology and speciation that have long intrigued evolutionary biologists. These questions include: What ecological and evolutionary forces maintain variation at these loci (Mitchell-olds *et al.* 2007)? How many genes are involved in adaptation and speciation, what are their effect sizes and how does this affect the availability of suitable genetic variation for adaptation and speciation (Orr & Coyne 1992, Allen Orr 2001, Pritchard & Di Rienzo 2010, Yeaman & Whitlock 2011, Seehausen *et al.* 2014)? What is the importance and impact of linkage and the genetic architecture of traits on adaptation (Yeaman & Whitlock 2011, Feder *et al.* 2012b, Flaxman *et al.* 2013)? Are the same genes involved repeatedly in independent adaptation and speciation events (Colosimo *et al.* 2005, Conte *et al.* 2012, Stern 2013)? Did mutations involved in adaptation and speciation arise *de novo* or from older preexisting variation (Barrett & Schluter 2008, Sousa & Hey 2013)? Are changes at regulatory sites or coding regions more likely to underlie adaptation (Hoekstra & Coyne 2007, Stern & Orgogozo 2008)? Are structural variants such as inversions, duplications or chromosomal rearrangements important for speciation (Kirkpatrick & Barton 2006, Hoffmann & Rieseberg 2008, Yeaman 2013)? Several of these questions have been thoroughly investigated in a few unique ecological model systems. Some well-known examples include the genetic basis of repeated armor plate evolution in threespine sticklebacks (Colosimo *et al.* 2005, Jones *et al.* 2012), coat color variation in beach mice (Hoekstra *et al.* 2006, Linnen *et al.* 2009, 2013), involvement of the *optix* gene in repeated convergent evolution of butterfly wing pattern mimicry (Reed *et al.* 2011), chromosomal rearrangements that result in tightly linked genetic loci that are inherited as a single unit or 'supergenes' and that provide integrated control of complex adaptive phenotypes in *Heliconius* butterflies (Joron *et al.* 2011), increased divergence in inverted regions in *Drosophila* genomes (Machado *et al.* 2002, Noor *et al.* 2007, Stevison *et al.* 2011), allopatric origins of inversions in sympatric races of the apple maggot fly *Rhagoletis pomonella* (Feder *et al.* 2003b, a) and parallel evolution of local adaptation and reproductive isolation in the face of gene flow in the rocky-shore gastropod *Littorina saxatilis* (Johannesson *et al.* 2010, Butlin *et al.* 2013). These study systems have proven to be pivotal in our understanding of adaptive evolution. However, to obtain a clear picture of the ecology and genetics of adaptation, these questions need to be addressed in a large number of taxa (Stinchcombe & Hoekstra 2008).

ENZYME POLYMORPHISMS

Many population studies have used allozyme polymorphisms that involve soluble enzymes separated by size and charge on electrophoresis gels by methods that use the cofactor NAD^+ or $NADP^+$ either directly or in conjunction with enzymes that are coupled to them (Eanes 1999). Therefore, much information exists about polymorphisms in metabolic enzymes in the glycolytic pathway, the Krebs cycle and their branches (Marden 2013). In some organisms, associations between markers and ecologically relevant variation have been found when investigating a small number of these molecular markers in population studies. Interesting examples include polymorphism in the alcohol dehydrogenase (ADH) enzyme in *Drosophila melanogaster* (Johnson & Schaffer 1973, Kreitman 1983), balanced polymorphisms in the phosphoglucose isomerase (*Pgi*) and succinate dehydrogenase d (*Sdhd*) loci associated to flight distance in Glanville fritillary butterflies (Watt *et al.* 2003, Wheat *et al.* 2006, Marden *et al.* 2012) and polymorphisms associated with local adaptation in the $NADP^+$-dependent isocitrate dehydrogenase (IDH) enzyme in the ground beetle *Pogonus chalceus* (Dhuyvetter *et al.* 2004, 2007) and the cricket species *Allonemobius socius* (Huestis & Marshall 2006, Huestis *et al.* 2009) and *Gryllus firmus* (Zhao & Zera 2006). Although it is difficult to infer whether these markers are the target of selection or closely linked to genes involved in adaptation, studying genetic variation associated with these genes has provided valuable insights into the evolutionary history of adaptive evolution (Kreitman 1983, McDonald & Kreitman 1991, Wheat *et al.* 2009). One of the reasons is that analyses of population structure and phylogenetic relationships based on these markers enable to reveal patterns of adaptive divergence that could be obscured by ongoing gene exchange at genomic regions unaffected by divergent selection (i.e. neutral markers).

QUANTITATIVE GENETICS

In most cases, unraveling the genetic architecture of ecologically important traits necessitates evaluating large amounts of variable genetic markers in multiple individuals and populations. Quantitative genetics, using techniques such as linkage disequilibrium (LD) mapping and mapping Quantitative Trait Loci (QTL mapping), has a rich history in its ability to identify functionally relevant genes in model organisms, such as for instance bristle number in *Drosophila melanogaster* (Mackay & Langley 1990, Mackay & Lyman 2005, Mackay *et al.* 2009). Classical approaches using variation in restriction sites (e.g. amplified fragment length polymorphisms (AFLP's)) for LD and QTL mapping have also allowed mapping genes involved in ecologically relevant variation, such as armor plate variation in threespine sticklebacks (Colosimo *et al.* 2004) and growth associated

loci in lake whitefish (Rogers & Bernatchez 2005, 2007). Alternatively, by focusing on markers in candidate genes identified in related model species associations can be identified, such as for melanism in pocket and beach mice (Nachman *et al.* 2003, Hoekstra *et al.* 2006). Unfortunately, thoroughly examining the association of markers and phenotypes by QTL mapping necessitates large amounts of variable markers and needs laborious crossbreeding, which is, however, unfeasible for many ecological model systems. Moreover, techniques such as QTL mapping often only allow identifying large genomic regions as the amount of variable genetic markers is often limited. Subsequently, these genomic regions have to be investigated and mapped on a finer scale to find the exact loci associated with adaptation (Colosimo *et al.* 2005, Mackay *et al.* 2009).

POPULATION GENOMICS

More recently, technological advances allow scoring a massive number of molecular markers in multiple individuals from different environments. These advances largely involve high throughput sequencing techniques (Metzker 2010) which allow sequencing and comparing up to complete genomes for any kind of organism (Lawniczak *et al.* 2010, Jones *et al.* 2012, Dasmahapatra *et al.* 2012, Ellegren *et al.* 2012, Nadeau *et al.* 2013) or comparing many individuals from different populations by focusing on randomly distributed but consistent genomic regions (Hohenlohe *et al.* 2010a, Davey *et al.* 2010, 2011, Nadeau *et al.* 2013, Keller *et al.* 2013). Applying these techniques for scoring massive amounts of genetic markers in ecological model systems allows unprecedented opportunities for detecting genomic regions and genes under selection and, hence, involved in adaptation and phenotypic differentiation (Feder & Micthell-Olds 2003, Luikart *et al.* 2003, Storz 2005, Stinchcombe & Hoekstra 2008, Stapley *et al.* 2010, Butlin 2010, Hohenlohe *et al.* 2010b, Rice *et al.* 2011, Savolainen *et al.* 2013).

In population genomics, population genetic analyses of a large number of markers distributed throughout the genome aims at identifying loci showing unusual patterns of variation, potentially due to selection at linked sites (Schlötterer 2003, Hohenlohe *et al.* 2010b). When mutations are positively selected to high frequencies or fixation, closely linked neutral sites will show similar patterns of genetic variations because of genetic hitchhiking (Maynard Smith & Haigh 1974). According to this principle, studying genome wide patterns of variation allows separating locus-specific effects that affect one or a few loci at a time (e.g. recombination, selection and mutation) from genome-wide demographic effects (e.g. population size increase, genetic bottlenecks, founder events and inbreeding) (Luikart *et al.* 2003, Stinchcombe & Hoekstra 2008).

This approach allows identifying ecologically relevant loci and studying the genetics of adaptation from a genome wide perspective; it allows studying genome wide patterns of hard and soft selective sweeps (Messer & Petrov 2013) or balancing selection (Charlesworth 2006) and like quantitative genetic approaches, allows studying the genetic architecture of adaptive traits (Savolainen *et al.* 2013). Moreover, complementary to artificial crossings, this approach can be used to study the genetic basis of (ecological) speciation when species divergence is recent and gene flow is ongoing so that outliers stand out from relatively low levels of background divergence (Gompert & Buerkle 2009, Malek *et al.* 2012). Finally, loci that contribute to phenotypic differences between ancestral populations can be identified by investigating genotype-phenotype correlations in a population of mixed ancestry (i.e. admixture mapping; Buerkle and Lexer 2008; Malek et al. 2012).

However, population genomics also has its limitations (Stinchcombe & Hoekstra 2008). First, the loci showing unusual patterns of variation (i.e. outlier loci) are most likely not the causal loci but linked with the selected site(s). However, this can also be problematic in QTL mapping. Second, many factors may affect the extent of linkage disequilibrium and their effect may vary across the genome (e.g. recombination rate, strength of selection, population history and long term balancing selection). Third, without a linkage map or reference genome, the size and the position of the differentiated genomic regions will be unknown. Fourth, the phenotypic effect of differentiating loci remains largely unknown, limiting ecological investigation. Finally, loci involved in adaptation may be missed or, given the large amount of markers, falsely associated with adaptation by chance. If feasible, population genetic structure can be controlled using controlled crosses (i.e. QTL mapping). Furthermore, QTL mapping can be used for fine scale mapping of the large chromosomal regions identified by the population genomics approach (Stinchcombe & Hoekstra 2008).

In sum, identifying and studying ecologically relevant genes is cumbersome, but an increasingly powerful set of approaches exists. The relative ease by which these techniques can be used on so called non-model organisms makes this a strongly developing research field answering long-standing evolutionary questions.

STUDY SYSTEM: *POGONUS CHALCEUS*

BIOLOGY

Pogonus chalceus (Marsham 1802) is a halobiontic ground beetle (Coleoptera, Carabidae) about 6-8 mm in length (Figure 2). The species mainly occurs in marine marshes with a salinity exceeding 0.1% and prefers partly vegetated zones on heavy sea clay(Desender & Maelfait 1999). These habitats include tidal marshes along the coast as well as inland salt marshes that are separated from the tidal influence of the sea. These latter inland habitats generally become flooded on an irregular base for several months during winter. *P. chalceus* individuals are mostly found hidden under washed up debris, vegetation, crevices and loose sand clods. Typically, individuals are found buried in the soil. In tidal habitats, specimens most likely reside in crevices during high tides. In the inland habitats, beetles are most easily found between shrinkage cracks and in heaps of sand which protrude above the more humid surroundings (personal observation). Rough density estimations of *P. chalceus* beetles among a wide set of populations ranges from 25 up to 57 individuals per 10 m^2 (Desender *et al.* 1998). Habitat sizes along the Atlantic European coasts range from 0.1 ha (e.g. Oostende) up to 4,000 ha (e.g. Mont Saint Michel).

P. chalceus most likely feeds on amphipods and other small invertebrates that are abundant in their habitat. Except for *P. chalceus,* which is wing-polymorphic, all species of the genus *Pogonus* are long-winged with functional flight musculature during the entire year.

Figure 2. *Pogonus chalceus* (Marsham 1802). Photo adopted from www.eurocarabidae.de.

GEOGRAPHIC DISTRIBUTION

The geographical distribution of *P. chalceus* extends along the Atlantic coasts from Denmark down to the major part of the Mediterranean coasts, including North-Africa (Turin 2000). In Belgium, *P. chalceus* is classified as vulnerable in the Red data book (Desender *et al.* 1995).

REPRODUCTIVE BIOLOGY

The development of beetles is holometabolous (indirect development or complete morphosis) with a dramatic pupal reorganization between juvenile and adult phases. *P. chalceus* has three larval stages (instar I-III) before developing into an adult through a pupal stage. Most adult *P. chalceus* beetles live for one year, however, some can survive and reproduce for more than one year (Desender 1985). *P. chalceus* is a univoltine (one generation per year) spring-reproducing ground beetle (Desender 1985). They hibernate as adult and have their reproductive activity mainly during spring. As the larvae of *P. chalceus* need high temperatures for their development, they develop during summer and the new beetle generation emerges during late summer and autumn. It is argued that in contrast to true spring breeders the maturation of the gonads only depends on the conditions of temperature and not on the photoperiod in *P. chalceus* (Paarmann 1976). No significant differences have been found in egg production between long -and short-winged populations (Desender 1989a). Phenological differences have not been studied between different populations. However, Dhuyvetter *et al.* (2007) reported marked temperature and salinity differences between closely located sets of populations. As temperature has been found to have strong effects on larval development as well as on propagation rhythm by affecting dormancy of the gonads (Paarmann 1976), these environmental differences may possibly result in a slight offset of the emergence of the new generation of adult beetles as well as the reproductive period.

DISPERSAL POLYMORPHISM

Wing polymorphism and flight muscle dimorphism occurs in *P. chalceus*. Wing size is highly polymorphic in this species with a percentage ranging from approximately 15 % to 100 % of the maximum realizable wing size (MRWS; Desender and Serrano 1999). No differences are found in relative wing development for different months or years within one site (Desender *et al.* 2000). On the other hand, flight muscle development shows seasonal variation, with a higher percentage of individuals with functional flight muscles during seasons with high temperatures and long days (Desender 1985). Variation in wing and flight muscle development gives indirect measures for variation in dispersal by flight and give an idea of the maximal proportion of individuals in a population that are able to fly and disperse. It is noteworthy that, especially for Atlantic populations, flight observations are rare for *P. chalceus* and a main annual flight period is unknown (Desender 2000). When wing development values decrease below values of about 70 %, individuals with functional flight muscles become increasingly rare or completely absent and beetles lose their capability of flying completely (Desender, 1985).

Mediterranean populations all possess high dispersal ability with fully developed wing sizes as well as high frequencies of individuals with functional flight muscles (Desender *et al.* 2000). Atlantic populations, on the other hand, show a more varying degree of wing polymorphism and dispersal ability. Crossbreeding experiments between beetles from allopatric populations (i.e. Nieuwpoort, Zwin and Oostende) showed that variation in relative wing size has a high heritability of 0.819 (SE= 0.07) (Desender 1989a), meaning that a high fraction of the phenotypic variability between these populations in this trait can be attributed to genetic variation. Populations with large mean wing sizes also tend to have larger body sizes, however, this trait has lower heritability estimates (0.68, SE= 0.21; Desender 1989). It has been suggested that body size might be related to the need to accommodate functional flight musculature (Desender 1989a, Dhuyvetter *et al.* 2004).

HABITAT STABILITY AND WING SIZE

Desender and colleagues identified two different habitat types that select differently for dispersal ability; *stable* and *temporary* habitats (Desender *et al.* 1998, Dhuyvetter *et al.* 2004). They argued that small and unstable or temporary populations select for retention of a high dispersal morph despite the associated reproductive costs. In habitats with unpredictable dynamics it is advantages to invest in dispersal to be able to escape these changing environments when they become unsuitable and to colonize new locations. In stable habitats, on the other hand, individuals with reduced dispersal ability may invest more resources in reproduction. Hence, when there is a cost to dispersal, individuals with reduced dispersal ability are expected to increase in frequency in populations inhabiting stable and permanent habitats (Roff 1986, Roff & Fairbairn 2007). Moreover, higher emigration rates of the long-winged morph would lead to an increased frequency of the short-winged morph in this situation (Roff 1986). Indeed, in *P. chalceus* a clear association exists between habitat persistence (i.e. age of the salt marshes) and wing-size among populations along the Atlantic coast (Desender *et al.* 1998). Furthermore, Mediterranean populations all possess high dispersal ability in accordance with low permanence due to prolonged inundation of these habitats (Paarmann 1976, Desender *et al.* 2000, Dhuyvetter *et al.* 2004).

In this dissertation we recognize two main habitats that strongly differ in hydrological regime and select differently for dispersal capacity; *tidal salt marshes* and *seasonally inundated inland salt marshes*. Tidal salt marshes are flooded year-round on a regular basis (i.e. tides), but for short periods of at maximum a few hours only. Seasonally flooded marshes are disconnected from the sea and are permanently inundated for long periods, forcing the beetles to escape these inundations. In accordance with these habitat dynamics, populations inhabiting tidal and seasonal salt marshes have a low and high

average wing size, respectively. These differences in dispersal ability most likely result from divergent selection forcing beetles to disperse to drier patches or habitats during long term flooding in the seasonal habitats. Conversely, in the tidal habitats it is most likely advantageous to stay submerged during quickly rising, but relatively short flooding as this decreases the risk of predation if beetles would repeatedly attempt to escape these frequent inundations. Hence, dispersal behavior is expected to be costly in the tidal habitats, whereas staying submerged for extensive periods in the seasonal habitats would result in mortality as beetles are expected to tolerate submergence for only short periods.

The definition of the different habitats as tidal or seasonal is similar to the distinction of the stable and temporary habitats, but makes an important distinction in that it emphasis on adaptation to the different hydrological dynamics in the habitats (not only habitat persistence). Moreover, this has important implication for the understanding of the evolution and preservation of the dispersal ecotypes (see Chapter 6).

MITOCHONDRIAL NADP$^+$-DEPENDENT ISOCITRATE DEHYDROGENASE

Population genetic studies of *P. chalceus* revealed that variation in allozyme frequencies of the mitochondrial NADP$^+$-dependent isocitrate dehydrogenase (mtIDH) protein shows a similar pattern as variation in wing size differences among populations (Dhuyvetter *et al.* 2004). Populations with high frequencies of long-winged individuals, which mostly occupy seasonally flooded inland habitats, have high frequencies of the mtIDH-B allozyme. In contrast, populations from tidal habitats with high frequencies of short-winged individuals have high frequencies of the mtIDH-D allozyme. The consistent association of wing length and mtIDH allozyme frequencies with habitat stability in independent populations strongly implies natural selection as the cause of differentiation, as random genetic drift is unlikely to produce such a pattern (Dhuyvetter *et al.* 2004). However, heritability of variation in wing size has not been determined among sympatric populations. Furthermore, how mtIDH allozyme variation is translated at the genomic level and whether the association between wing size and mtIDH allozymes results from close genetic linkage or similar selection pressures has not been investigated.

Whether selection is acting on the *mtIdh* gene itself or on a tightly linked locus is unclear. NADP$^+$-dependent isocitrate dehydrogenases (i.e. mitochondrial and cytoplasmic NADP$^+$-IDH) catalyze the oxidative decarboxylation of isocitrate to α-oxogluterate with the concomitant reduction of NADP$^+$ to NADPH. NADP$^+$-specific IDH isozymes are not

directly involved in the Krebs cycle (i.e. NAD^+-dependent isocitrate dehydrogenase) and the precise metabolic functions of the $NADP^+$-specific IDH isozymes are unclear. Two $NADP^+$-dependent isocitrate dehydrogenases have been reported in eukaryotes, one of which is located in the mitochondria (mtIDH) and the other predominantly in the cytoplasm (cytIDH) (Jenningss *et al.* 1994, Zhao & Mcalister-henns 1996). Both $NADP^+$-dependent enzymes are homodimers that are encoded in the nuclear genome (Ceccarelli *et al.* 2002, Xu *et al.* 2004). According to the allozyme protocol of Hebert and Beaton (1993), the IDH protein associated with wing size in *P. chalceus* is most likely the mitochondrial IDH isozyme.

It has been demonstrated that flight muscles of beetles contain high activities of $NADP^+$-IDH, indicating a possible importance of $NADP^+$-IDH for flight metabolism (Alp *et al.* 1976). Further, it has been suggested that mtIDH provides NADPH for maintenance of proper oxidation-reduction balance and protection against oxidative damage (Jo *et al.* 2001, Lee & Koh 2002, Kim *et al.* 2005). However, these functional associations are only suggestive for possible adaptive differences between the mtIDH allozymes. Examples of other allozyme polymorphism that are shown to be involved in flight metabolism are phosphoglucose isomerase (PGI) and succinate dehydrogenase (SDH) in *Colias eurytheme* and *Melitaea cinxia* butterflies (Wheat *et al.* 2006, 2009, Marden *et al.* 2012). Activity variation in cytoplasmic $NADP^+$-dependent isocitrate dehydrogenase has been inferred to differ between dispersal morphs of the crickets *Gryllus firmus* (Zhao & Zera 2006). However, it has been demonstrated that this variation in enzyme activity is exclusively attributable to variation in enzyme concentration, which in turn stems from allelic differences in transcription rates (Schilder *et al.* 2011). Former enzymes are involved in energy metabolism pathways, which have been found to be frequent targets of selection (Marden 2013), as might be the case for the mtIDH enzyme.

THE GUÉRANDE SALTERNS

The Guérande salterns in France spread over more than 2,000 hectares and have been constructed and cultivated for over a millennium. A mosaic of two contrasting habitats can be found here at distances of only a few meters and in hundreds of replicates; *canals* and *ponds* (Figure 3). The ponds are used to evaporate water and concentrate salts. The canals bring Atlantic seawater into the ponds. These canals show similar hydrological dynamics as tidal marshes. The artificially constructed ponds are periodically completely flooded every few years when silt deposits prevent further cultivation. The hydrological dynamics of ponds resembles that of seasonally flooded inland marshes.

In these salterns, short-winged populations with on average 2.2 times smaller wing size than long-winged populations and a high frequency of the mtIDH-D allozyme (0.96 ±

0.04) are found along the edge of the tidally flooded (stable) canals, while long-winged populations with a high frequency of the mtIDH-B allele (0.58 ± 0.02) are found in dry ponds or in dry edges of seasonally flooded (temporary) ponds (Dhuyvetter *et al.* 2007). Despite the strong divergence in both wing size and mtIDH alleles, microsatellite and allozyme data confirmed that genetic differentiation among these ecotypes in neutral markers is very low and smaller compared to allopatric populations from the same ecotype (Dhuyvetter *et al.* 2007). Hence, also at micro geographical scales with ample opportunity of gene flow, the mtIDH allozyme frequencies and wing size distribution show a strong correlation with habitat dynamics.

Figure 3. The Guérande salterns. Above: satellite view of the Guérande salterns (Google earth). Down: Panoramic view showing a canal (left) and a flooded pond (right) separate by only few meters (June 2013).

DISPERSAL POLYMORPHISMS

A large body of theoretical and empirical literature exists discussing the evolution of dispersal polymorphisms (e.g. Roff 1994a; Roff and Fairbairn 2007). These studies often consider the evolution of dispersal polymorphism as resulting from fitness trade-offs (Mole & Zera 1993, Roff & Fairbairn 2007, Stevens *et al.* 2012) or metapopulation dynamics with dispersal between patches that have different rates of extinction or fluctuations in carrying capacity (McPeek & Holt 1992, Holt & McPeek 1996, Mathias *et al.* 2001, Hendrickx *et al.* 2013). Theoretical models mostly discuss how these factors result in stable dimorphisms within populations. *P. chalceus* populations inhabit environments that can strongly differ in their dynamics (i.e. tidal and seasonal habitats). However, in contrast to most theoretical models discussing the evolution of dispersal polymorphisms, in *P. chalceus* differentiation in dispersal ability has resulted in spatial separation and local adaptation to these differing habitats. Nevertheless, to obtain a good understanding of the evolutionary processes leading to ecological divergence in the ground beetle *P. chalceus* it is interesting to discuss the ultimate causes (selective forces) that result in the evolution of dispersal polymorphisms. Furthermore, how ultimate causes affect the evolution of dispersal polymorphisms requires knowledge about the genetic architecture and physiology of the dispersal strategy and its developmental pathways (proximate causes).

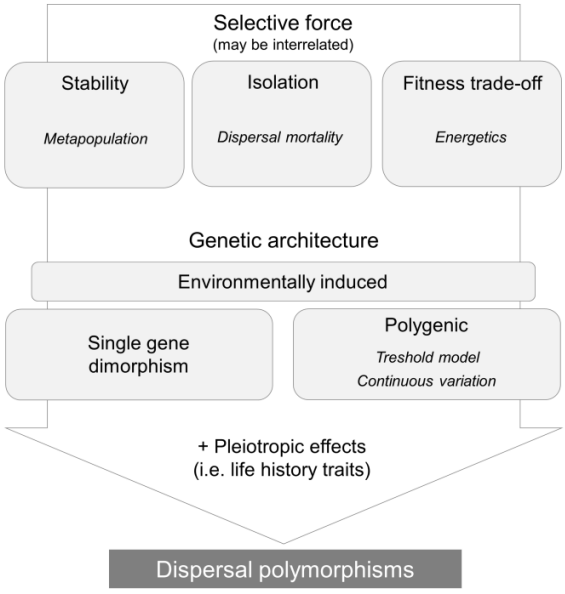


Figure 4. Selective forces that may result in dispersal polymorphisms and their genetic architecture.

Ultimately, flight polymorphisms result from different selective advantages and disadvantages of the normal and flightless morph (Harrison 1980). Dispersal may be advantageous for avoiding competition among kin (Bitume *et al.* 2013) and preventing inbreeding (Bengtsson 1978). Further, if populations go extinct, dispersal is favored because it allows the recolonization of empty patches (Olivieri *et al.* 1995). Disadvantages include increased mortality associated with dispersal and costs associated with the allocation of resources to morphological structures that facilitate flight but which decrease the amount of resources available for reproduction (Bonte *et al.* 2012). Here I discuss several aspects that select differently for dispersal ability and which may be involved in the evolution of flight polymorphisms (Figure 4).

TEMPORAL AND SPATIAL HETEROGENEITY "The world is heterogeneous in both time and space, and migration is an evolved response to this heterogeneity" (Roff & Fairbairn 2007). Oppositely, when dispersal ability is under direct genetic control, permanent habitats will have a higher proportion of flightless individuals. A simple explanation is that after a site is initially colonized by long-winged individuals, the fraction of long-winged morphs will decline because a greater fraction of the genes determining long-winged morphs will leave the population each generation which may result in short-winged populations in permanent habitats (Järvinen & Vepsäläinen 1976, Olivieri *et al.* 1995).

ISOLATION Dispersal involves risks. When dispersal becomes increasingly costly, the advantage of being able to find new suitable habitats becomes offset. Therefore, if habitat patches become increasingly isolated, the chances of finding new suitable habitat reduce and the mortality of dispersers will likely increase. Hence, it is predicted that species from isolated locations will have lower levels of dispersal (Harrison 1980). This could for instance explain the observed high percentage of flightless species on oceanic islands (Darwin 1859, Harrison 1980).

FITNESS TRADE-OFF In numerous wing-polymorphic species, flight capability is negatively correlated with key life history traits such as fecundity and age at first reproduction (Mole & Zera 1993, Desender 2000, Oliveira *et al.* 2006, Stevens *et al.* 2012). From these findings it is argued that flight capability and reproduction are energetically expensive and compete for internal resources, resulting in a fitness trade-off between flight capability and reproduction called the 'flight-oogenesis syndrome' (Rankin *et al.* 1986).

GENETIC ARCHITECTURE The evolution of dispersal requires that dispersal ability is under genetic control and, therefore, can be subjected to selection (Mathias *et al.* 2001). The vast majority of insects are descended from winged ancestors, and, hence, wing size reduction is considered the derived state (Roff & Fairbairn 2007). The presence or absence of wings may be controlled by a single locus with two alleles or a polygenic system. In most wing-dimorphic Coleoptera, wing dimorphism is under control of a *single locus*, with reduced wings being dominant (Roff & Fairbairn 2007). It is argued that dominance of the allele that reduces wing size has been repeatedly favored because dominant alleles are expressed in heterozygotes and are, hence, readily available for selection when they evolve and are less likely to be lost from the population by chance. On the other hand, it can be considered that dispersal syndromes are complex traits (Stevens *et al.* 2014) and likely involve many genes (*polygenic*). Indeed, wing size inheritance involving multiple loci has also been found in insects (Desender 1989a, Fairbairn & Roff 1991). This may result in continuous variation in dispersal ability or, alternatively, these polygenic traits may be threshold driven, resulting in distinct dispersal polymorphisms (Roff 1994b). According to this latter view, continuous genetic variation may result in distinct phenotypes; depending on whether individuals lie above a certain threshold they will develop in one or the other morph. This model is useful for understanding the development of different morphologies and can also be applied to other types of dichotomous traits, such as the decision to migrate or not. For instance, *P. chalceus* has a continuous wing size distribution, but effective dispersal may be discrete. Furthermore, within populations, genetic linkage becomes of less importance as selection works on the threshold value. Depending on the allele distributions within the population, a certain frequency of dispersive individuals will occur. Finally, morph determination can be *environmentally induced*. For instance wing dimorphism in both *Calathus cinctus* and *Calathus melanocephalus* is genetically determined by a single locus with short-wings dominant compared to long-wings (Aukema 1990, 1995). In *C. melanocephalus*, however, the expression of the long-winged genotype is modified by environmental factors such as temperature and food supply, whereas in *C. cinctus* wing-length is independent of these factors. Further, in the cricket *Gryllus firmus*, variation in dispersal is under polygenic control and is also influenced by a variety of environmental factors such as density, photoperiod and nutrition (Zera & Larsen 2001, Zera & Zhao 2003, Vellichirammal *et al.* 2014). Morph expression in crickets is best viewed as a polygenic threshold trait, the threshold level of which is determined by both multiple loci and environmental inputs (Roff 1994b). Finally, among aphids, some species alternate between environmentally sensitive and genetic control of wing morph determination in their life cycle or between males and females (Braendle *et al.* 2006, Brisson 2010).

DEVELOPMENT OF THE WING Insect limb and wing development has been intensively studied in *Drosophila melanogaster* by systematic screens of loss-of-function mutations (Nüsslein-volhard & Wieschaus 1980), gain-of -function phenotypes (Rørth *et al.* 1998), expression studies (Calleja *et al.* 1996) and employing reverse genetic approaches such as RNA-mediated interference (Kennerdell & Carthew 1998). Weihe *et al.* (2005) give a comprehensive overview of the pathways involved in wing development. The wing is derived from the wing imaginal disc, which is developed from outgrowths of the body wall (i.e. ectoderm). During development of the imaginal disc, three major axes are established, i.e. anterior-posterior, dorsal-ventral, and proximal-distal axes (Figure 5). The imaginal disc receives these patterning cues by diffusible signals called morphogens which provide cells with positional information. Hence, changes in expression of these genes affect cell identity and will, consequently, affect growth and development. The anterior-posterior axis is structured a.o. by the morphogens *engrailed, invected, decapentaplegic* and *hedgehog*. The dorsal-ventral axis is structured a.o. by the transcription factor *wingless*, *apterous* and *vestigial*. The proximal-distal axis is structured a.o. by the gene *distal-less*. For a complete list of genes involved in limb and wing development we refer to Weihe *et al. (*2005).

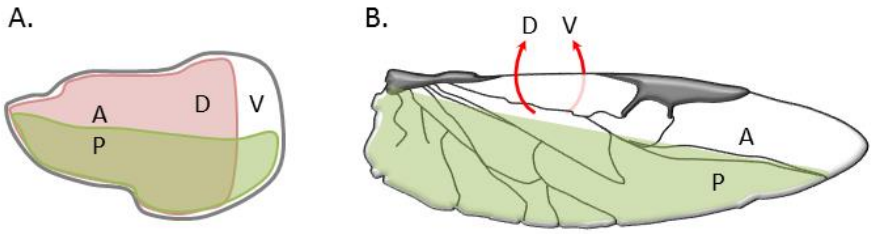


Figure 5. Organization of the dorsal-ventral (DV) and anterior-posterior (AP) axes in the wing imaginal disc (A.) and in the adult wing (B.). Wing patterning genes or morphogens are expressed in different compartments (e.g. *apterous* in D compartment, *engrailed* in P compartments) and direct the development of the wings.

Apart from studying how wings develop, few studies have investigated the genes directly involved in wing dimorphisms or polymorphisms. A large set of 34 candidate genes affecting wing development in *Drosophila* have also been studied in the red flour beetle *Tribolium castaneum* using RNA-mediated interference (Richards *et al.* 2008). Twenty two of these genes have been identified in the genome of the pea aphid *Acyrthosiphon pisum* (Brisson *et al.* 2010). To examine the effect of these wing development genes on wing polymorphisms in pea aphids, Brisson et al. (2010) examined the expression levels of eleven of these wing development genes via quantitative PCR at different developmental stages, for both winged and unwinged

parthenogenetic pea aphid females. One gene, *apterous1*, exhibited significantly different expression levels between winged and unwinged morphs, suggesting this gene is involved in polyphenic development in pea aphids. In ants, studying the expression of several wing development genes showed that expression of these genes is conserved in the winged castes of different ant species, whereas the wingless castes have evolved from different points of interruption in the wing development network (Abouheif & Wray 2002). Hence, wing polyphenism in different ant species results from interrupting or down regulating different genes in the wing development network. Finally, it has to be noted that alternative dispersal morphs most often exhibit systemic differences that go well beyond the presence or absence of wings and (pleiotropically) involve multiple key life history traits (Stevens *et al.* 2012). For instance, in the monarch butterfly, *Danaeus plexippus*, differences in dispersal behavior have been found to be associated with genes that are essential for flight muscle morphogenesis (i.e. collagen IV subunit α-1 and α-2 and *kettin*; Zhan *et al.* 2014).

OBJECTIVES AND OUTLINE

In this thesis we attempt to obtain a better understanding of the ecological mechanisms, the genetics and evolutionary history of adaptive divergence in the wing-polymorphic ground beetle *Pogonus chalceus*. Studying taxa in the process of splitting and with incipient reproductive isolation is of major interest for understanding how adaptation and speciation progresses (Via 2009). Moreover, mechanisms involved in reproductive isolation are best studied in a sympatric setting as geographic distance between *P. chalceus* populations in itself is unlikely to result in reduced gene flow. In *P. chalceus*, several factors allow addressing a wealth of evolutionary relevant questions discussed in the former section. These factors include (i) divergent selection among populations resulting in local adaptation, (ii) the environmental gradient being uncorrelated with geographical distance, (iii) the presence of multiple geographically separated replicates of the divergence process, and (iv) sufficient gene flow among populations subjected to divergent selection to distinguish neutral from non-neutral processes. The following questions are addressed in this thesis:

- What is the evolutionary history of the mtIDH alleles which are selected to high frequencies in different allopatric as well as sympatric populations?

- What are the ecological and genetic factors that maintain divergence in sympatry?

- How might divergent selection result in reproductive isolation in sympatric *P. chalceus* populations?

- How does the ecotypic divergence translate to the genomic level? Can we identify multiple loci associated with divergence, are they physically linked and what is their evolutionary history?

In CHAPTER 1, we investigate heritability of wing size in the sympatric Guérande populations and combine all available population genetic data on wing size and mtIDH allozymes from *P. chalceus* to study the association between these two genetic traits. Although wing size divergence has shown to have a high heritability between the compared geographically isolated populations (Desender 1989a), plastic responses may increase when gene flow levels are increased (Sultan & Spencer 2002). Whether wing size differences in the Guérande result from a plastic response or are constitutively expressed

is of major importance to the interpretation of the ecological and evolutionary factors influencing these differences in sympatry. Further, the across population association between wing size and mtIDH allozymes suggests that the *mtIdh* locus and loci involved in wing development show low levels of recombination, which would strongly facilitate the persistence of both locally adapted ecotypes under high levels of gene flow. Whether these traits constitute physically linked variation or whether the association results from similar selection pressures working on different loci is investigated.

Next, in CHAPTER 2, we identify the DNA sequence of the *mtIdh* gene and study its evolutionary history by comparing sequences from different mtIDH allozymes and populations. This allows us to determine the origin of the adaptation associated with the *mtIdh* locus; whether it evolved *de novo* in several populations separately or whether the observed patterns result from a single origin and subsequent spread into other populations. Further, comparing sequence variation in the *mtIdh* locus with coalescent simulations allows us to evaluate several evolutionary scenarios.

In CHAPTER 3, we expand the available genetic resources for *P. chalceus* by sequencing, assembling and annotating the complete transcriptome. Subsequently, in CHAPTER 4, we present the *de novo* assembly and analysis of a draft genome sequence for *P. chalceus*.

In CHAPTER 5, the genetic basis of population and ecotypic divergence between eight *P. chalceus* populations distributed across Europe is investigated using Restriction site Associated DNA markers (RAD tags). This technique allows comparing many individuals from different populations by focusing on randomly distributed genetic markers but at consistent genomic regions.

In CHAPTER 6, we try to identify a link between disruptive selection and assortative mating that may promote the divergence and preservation of sympatric beetle populations by testing the response and adaptation to inundation of *P. chalceus* beetles from tidal versus seasonal habitats. If populations from tidal and seasonal habitats respond differently to inundation events in terms of dispersal or staying in a habitat, this may lead to spatial sorting and offer an explanation for the coexistence of distinct ecotypes in sympatric settings.

Finally, in CHAPTER 7, we explore the possible effect of natal habitat experience on adult habitat preferences. Existence of natal habitat preferences may provide an easy explanation for the initial colonization of new habitats and may strongly affect the evolution of the distinct *P. chalceus* ecotypes in sympatric settings.

CHAPTER 1

A TIGHT ASSOCIATION IN TWO GENETICALLY UNLINKED DISPERSAL RELATED TRAITS IN SYMPATRIC AND ALLOPATRIC SALT MARSH BEETLE POPULATIONS

Steven M. Van Belleghem [1,2]
Frederik Hendrickx [1,2]

Modified from: *Genetica* (2014) 142:1-9

[1] Terrestrial Ecology Unit, Biology Department, Ghent University, K. L. Ledeganckstraat 35, 9000 Gent, Belgium
[2] Department Entomology, Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000 Brussel, Belgium

ABSTRACT

Local adaptation likely involves selection on multiple, genetically unlinked traits to increase fitness in different habitats. Conversely, recombination is expected to counteract local adaptation under gene flow by breaking down adaptive gene combinations. Western European populations of the salt marsh beetle *Pogonus chalceus* are characterized by large interpopulation variation at various geographical ranges in two traits related to dispersal ability, i.e. wing size and different allozymes of the mitochondrial $NADP^+$-dependent isocitrate dehydrogenase (*mtIdh*) gene. In this study, we tested whether variation in wing length was as strongly genetically determined in locally adapted populations in a sympatric mosaic compared to allopatric populations, and if variation in mtIDH and wing size was genetically unlinked. We demonstrate that variation in wing size is highly genetically determined ($h^2 = 0.90$) in sympatry and of comparable magnitude as geographically separated populations. Second, we show that, although frequencies of mtIDH allozymes are tightly associated with mean population wing size across Western European populations, the correlation is strongly reduced within some of the populations. These findings demonstrate that the divergence involves at least two traits under independent genetic control and that the genetically distinct ecotypes are retained at geographical distances with ample opportunity for gene flow.

INTRODUCTION

When multiple unlinked loci are involved in local adaptation, recombination between environment specific alleles counteracts their joint inheritance, hampering the independent evolution of these lineages (Felsenstein 1981, Lenormand 2002). Consequently, theory predicts that phenotypic differentiation resulting from genetic changes in multiple traits under sympatry is more likely to involve pleiotropy or strong linkage disequilibrium among the involved loci (Felsenstein 1981, Via 2001, Rundle & Nosil 2005). More recently, theoretical models even predict that when levels of migration between locally adapted populations is high, this will play an important role in shaping the genetic architecture of a trait wherein QTL's will be clustered into fewer loci of large effect (Lenormand 2002, Griswold 2006, Yeaman & Whitlock 2011). Indeed, strong genetic linkage has been suggested to play an important role in maintaining contrasting ecotypes in hybridizing populations of, for instance, sticklebacks (Jones *et al.* 2012) and pea aphids (Hawthorne & Via 2001).

Despite a multitude of mathematical and simulation models (Dieckmann & Doebeli 1999, Fry 2003, Gavrilets 2003, Bolnick & Fitzpatrick 2007), only a few well known examples of adaptive divergence under high levels of gene flow have been documented in recent decades (Nosil 2012). These examples include the genetic linkage of ecological specialization and reproductive isolation in pea aphids, *Acyrthosiphon pisum pisum* (Hawthorne & Via 2001), and genome wide divergence between sympatric races of the apple maggots, *Rhagoletis pomonella* (Michel *et al.* 2010). These systems proved to be pivotal to gain a better understanding of the first steps within ecological divergence and speciation processes (Nosil 2012). Unfortunately, few of these studies investigated to what extent multiple traits are involved in adaptation to these divergent habitats, and in particular how they are associated at the genetic level.

The salt marsh beetle *Pogonus chalceus* (Marsham 1802) represents a case of pronounced genetic divergence in multiple traits related to dispersal ability. These differences in dispersal ability are associated with differences in habitat stability within a set of highly interconnected populations. *P. chalceus* is a halobiontic ground beetle (Carabidae), found along the Atlantic Western European coasts down to and including the major parts of the Mediterranean coasts (Turin 2000). Wing size is highly polymorphic in this species with a percentage ranging from approximately 15 % to 100 % of the maximum realizable wing size (Desender & Serrano 1999). Although it has not been directly shown for this species, the high energetic costs associated with flight capability is expected to result in lower fecundity for long-winged morphs favoring the short-winged morph if the habitat is permanent (Roff 1994a). The retention of a high dispersal morph is interpreted as an adaptation to survival in temporary, more unstable environments (Dhuyvetter et al. 2004, but see Hendrickx et al. 2013). Stable habitats comprise coastal tidal marshes in which beetles are submerged during the regular short periods of flooding. The inland, unstable, salt marshes become inundated unpredictably for longer periods, most likely forcing the beetles to escape these unsuitable conditions. However, also age and size of the habitats have been shown to affect wing size distributions, with young and small salt marshes being occupied by individuals with on average high dispersal ability (Desender *et al.* 1998).

Moreover, previous work demonstrated that the average population wing size is strongly associated with the population frequencies of mitochondrial $NADP^+$-dependent isocitrate dehydrogenase allozymes (mtIDH; KEGG orthology (KO): k00031) across populations (Dhuyvetter *et al.* 2004), with long-winged populations having higher frequencies of the mtIDH-B allele. This divergence in both traits is apparently retained under high levels of interpopulation gene flow, as demonstrated for the Guérande salterns in France where both ecotypes coexist at distances of only a few meters in a

sympatric mosaic (Dhuyvetter *et al.* 2007). In these salterns, short-winged populations with on average 2.2 times smaller wing size than long-winged populations and a high frequency of the mtIDH-D allozyme (0.96 ±0.04) are found along the edge of tidally flooded (stable) canals that bring Atlantic seawater into ponds, while long-winged populations with a high frequency of the mtIDH-B allele (0.58 ±0.02) are found in the seasonally flooded (unstable) pond habitat. Despite the strong divergence in both wing size and mtIDH alleles, microsatellite and allozyme data confirmed that genetic differentiation among these ecotypes in these supposedly neutral markers is very low and smaller compared to allopatric populations from the same ecotype (Dhuyvetter *et al.* 2007). Although a biochemical link between the mtIDH allozymes and flight metabolism has not been confirmed and selection may be acting on a tightly linked locus. The mtIDH enzyme is encoded by a nuclear gene and catalyzes the oxidative decarboxylation of isocitrate to α-oxogluterate with the concomitant reduction of $NADP^+$ to NADPH. The enzyme is not directly involved in the Krebs cycle (i.e. NAD^+-dependent isocitrate dehydrogenase) and the precise metabolic function of the mtIDH enzyme is unclear. It has been suggested that mtIDH provides NADPH for maintenance of proper oxidation-reduction balance and protection against oxidative damage (Jo *et al.* 2001, Lee & Koh 2002, Kim *et al.* 2005). Further, it has been demonstrated that flight muscles of beetles have high activities of $NADP^+$-IDH, indicating possible involvement of $NADP^+$-IDH in flight metabolism (Alp *et al.* 1976). However, these functional associations are only suggestive for possible adaptive differences between the mtIDH allozymes.

In this study, we further analyze the previously demonstrated tight association between mtIDH allozyme frequencies and wing size at the population level. This across population association suggested that the *mtIdh* locus and loci involved in wing development show low levels of recombination, which would strongly facilitate the persistence of both locally adapted ecotypes under high levels of gene flow. Whether these traits constitute physically linked variation or whether the association results from similar selection pressures working on different loci has not been previously tested. Further, variation in wing size, which has previously been shown to have strong genetic component in this species (Desender 1989a), could be expected to be less genetically determined in sympatric populations compared to geographically isolated populations that have been studied to date. Therefore, we subjected a long- and short winged sympatric population from the Guérande salterns to crossbreeding in order to estimate the additive genetic contributions of the divergence in wing size and body size in this sympatric mosaic and compare this with estimates from allopatric populations. Next, we analyzed the available population data for wing length and mtIDH allozyme frequencies from previous studies to test whether nearby populations were more similar in mtIDH frequencies (i.e. spatial autocorrelation), which would imply a role of gene flow in the

distribution of mtIDH allele frequencies. Finally, we tested whether variation at the *mtIdh* locus was physically linked with genes involved in wing size by studying the association of mtIDH allozyme frequencies and wing size within populations.

MATERIALS & METHODS

HERITABILITY ESTIMATES

SAMPLING AND CROSSBREEDING

Individuals of the halobiont carabid beetle *P. chalceus* were sampled in the salterns of the Guérande region in France from two contrasting habitats, a tidal canal and an unpredictably flooded pond, in both September 2010 and April 2011. These paired sampling sites were located only 30 m apart (Figure 6; Pond 3 and Canal 3). Beetles captured in September 2010 were not in the reproductive stage and were not used for the crossbreeding experiment. Females captured in April 2011 were in the egg laying stage and fertilized in the field. Eggs of this parental (P) generation were raised in the lab and the emerged adults constitute the F0 generation. Adult F0 beetles of the pond and canal population were kept randomly in pairs and used to produce a F1 generation. Breeding and crossbreeding were performed under identical laboratory conditions; constant long day conditions (16h light, 8h dark) at 20 °C in plastic jars 5 cm in diameter with plaster. The plaster was initially moisturized with salt water from the field and kept saturated with fresh water. Both adults and larvae were fed pieces of mealworm every two days. The beetles tend to burry small holes into the plaster in which they lay their eggs separately and encapsulate the hole with gnawed plaster. Adult beetles were exposed to winter conditions (5 °C for five weeks) after emergence from the pupae to stimulate the development of the gonads (Paarmann 1976).

MEASUREMENTS AND DETERMINATION OF WING SIZE

Wing and body size (elytral size) were measured by means of a calibrated ocular under a binocular microscope. Wing size is expressed as an index that corrects for the allometric relationship between wing length and body size (den Boer 1980, Desender *et al.* 1986). More precisely, the relative wing size corrected for allometry expresses the percentage of the maximal realizable wing size (%MRWS). The relative wing size is wing length × width divided by elytral length × width. Relative wing size was expressed as a percentage of the maximal relative wing size for a beetle of a given size. This maximal realizable wing

size was derived from a regression of wing length and body size from Carabid species with always fully developed wings and functional flight muscles allowing comparisons of relative wing sizes of beetles with different body sizes (Desender *et al.* 1986).

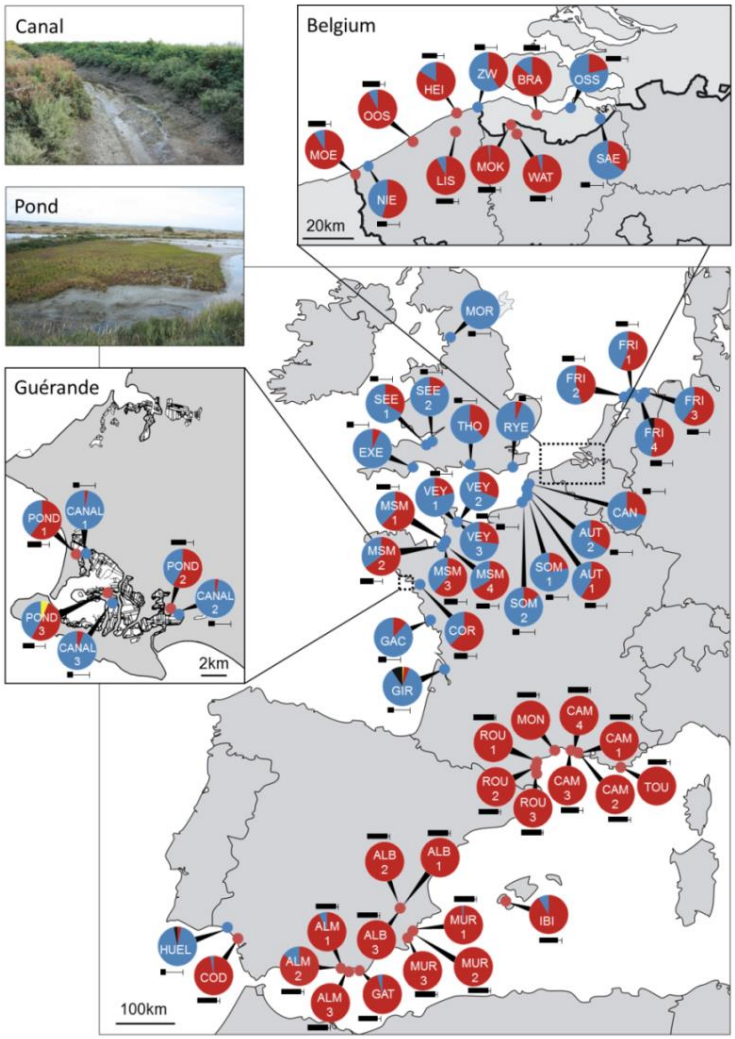


Figure 6. MtIDH allozyme distribution (mtIDH-A in yellow, mtIDH-B in red, mtIDH-D in blue, mtIDH-E, in black) of *P. chalceus* populations along the European coasts. Blue dots indicate tidal habitat, red dots indicate seasonal habitat. Horizontal bars give the mean %MRWS of each population. Pond 3 and Canal 3 were sampled for the crossbreeding experiment. Adopted from Dhuyvetter et al. (2004) with addition of unpublished data from Montpellier (MON), Gata (GAT), Coto Doñana (COD) and Huelva (HUE). For more information on other enzymes tested see Desender et al. (1998).

STATISTICAL ANALYSIS

Body size, absolute wing size and percentage of maximal realizable wing size (%MRWS) were first compared between the canal and pond populations by means of a general linear model including the factors population and sex (PROC GLM; SAS v9.1.2 Institute Inc). Heritability of wing and body size was determined by means of a parent-offspring analysis, wherein we regressed phenotypic traits of the F1 offspring against those of the F0 midparent and used the regression slope as an estimate of the narrow-sense heritability (h^2) (Falconer & Mackay 1996)

WING SIZE – mtIDH ASSOCIATION

SAMPLING

With the exception of the two samples from the Atlantic coast of Spain (Coto Doñana (COD) and Huelva (HUEL)) and two Mediterranean populations from France (Montpellier (MON) and Gata (GAT)), all population data on mtIDH allozyme frequencies and average wing size were published previously (Figure 6). The total dataset comprised 3,053 wing size measured and mtIDH genotyped individuals divided over 64 populations (see Appendix 1 for an overview of the original source of the data). Atlantic sample locations included ten sites from Belgium (including six seasonal populations), six sites from the United Kingdom, five sites from the Netherlands, twenty-one sites from France and two sites from Spain (one seasonal and one tidal populations). The sampled Mediterranean populations, which are all seasonally flooded (Paarmann 1976), included nine sites from southern France and eleven sites from Spain (including three populations from inland high elevation salt ponds near Albacete, at 600-800 m).

ASSOCIATION BETWEEN MTIDH AND WING SIZE ACROSS POPULATIONS

The across population association between mtIDH allozyme frequencies and average wing size has previously been tested by Dhuyvetter *et al.* (2004) ($r^2 = 0.95$; $P < 0.0001$). Here, we pooled all available data (Appendix 1) and included spatial autocorrelation in the model to test for the association between mtIDH allozyme frequencies and average wing size *across populations*. Frequencies of the mtIDH allozyme could also be spatially dependent due to higher levels of gene flow between more nearby populations. This could be particularly problematic if average wing size of the populations is also spatially structured across the region and could lead to a statistical association between mtIDH allele frequencies and wing size caused by neutral drift effects rather than resulting from

selection on mtIDH allozymes. We regressed the proportion of mtIDH-B alleles in a population against the average wing size by means of a generalized linear mixed model with a binomial distributed error and logit link. To account for spatial autocorrelation, we incorporated an exponential spatial variance structure based on the geographic coordinates of the populations. Interestingly, this model further allowed us to test if the geographic distance between populations significantly correlates with variation in mtIDH alleles given the average wing size. The model was constructed with the GLIMMIX procedure in SAS v9.3. using the proportion of mtIDH-B alleles to the total number of alleles (i.e. twice the number of individuals) as dependent variable, and population mean wing size as a fixed explanatory variable. To test for the significance of spatial autocorrelation, a model was run with the spatial correlation constrained to zero and the likelihood ratio of both models was tested against a χ^2 distribution. Individuals with the rare mtIDH-A, mtIDH-C and mtIDH-E allozymes, constituting 0.0020, 0.0016 and 0.0016 % of all sampled alleles respectively, were excluded from the analysis.

Association between mtIDH and wing size within populations

To investigate the degree of linkage disequilibrium between mtIDH and wing size alleles, we tested whether wing size differed significantly between individuals with different mtIDH genotypes *within* populations. For this analysis, only populations were considered which had at least 27 genotyped individuals and 29% of each mtIDH allozyme (Appendix 1). These cut-offs were chosen arbitrarily to include enough populations with a high number of genotyped individuals and with both the mtIDH-B and mtIDH-D allozymes presented in large frequencies, as it is difficult to assess the association between wing size and mtIDH alleles in populations which are nearly fixed for one of the mtIDH alleles. We first tested for a significant association between the proportion of mtIDH-B alleles in each individual (mtIDH-DD: 0.0, mtIDH-BD: 0.5 or mtIDH-BB: 1.0) and its wing size (%MRWS effect), and if this association differed between the different populations (%MRWS x population effect) by means of a generalized linear model with a binomial distribution error and logit link (PROC GENMOD in SAS v9.3). Next, we estimated the slope of the mtIDH – wing size association within each population by reformulating the previous model as a cell means model to estimate the slope for each population separately.

RESULTS

HERITABILITY ESTIMATES

COMPARISON OF SAMPLED CANAL AND POND POPULATION

To test the constancy of the distribution of the ecotypes and their differences found by Dhuyvetter *et al.* (2007), we first measured the wing and body size of the individuals used in this study. Body size (elytral length), absolute wing length and percentage of maximal realizable wing size (%MRWS) differed strongly between the two parental populations used in the study (Figure 7). Females were significantly larger than males in both canal and pond habitats ($F_{1,121} = 177.31$; $P < 0.0001$) and individuals from the pond population were significantly larger than the canal population ($F_{1,121} = 317.82$; $P < 0.0001$). Considering wing size, measured as both absolute wing length and %MRWS, differences were even more pronounced between the pond and canal population, with the canal population having much smaller wings compared to the pond population (Figure 7; $F_{1,121} = 1225.48$; $P < 0.0001$ and $F_{1,121} = 1058.04$; $P < 0.0001$ respectively). This pattern appeared consistent over both sampling dates ($F_{1,121} = 0.65$; $P = 0.42$ and $F_{1,121} = 3.02$; $P = 0.09$ for absolute wing size and %MRWS, respectively). Using pooled data for both sampling dates, females had larger absolute wing size compared to males ($F_{1,121} = 22.50$; $P < 0.0001$), but this reflects differences in adult size of both sexes as the sex difference could not be detected when using the body size corrected %MRWS ($F_{1,121} = 0.24$; $P = 0.63$).

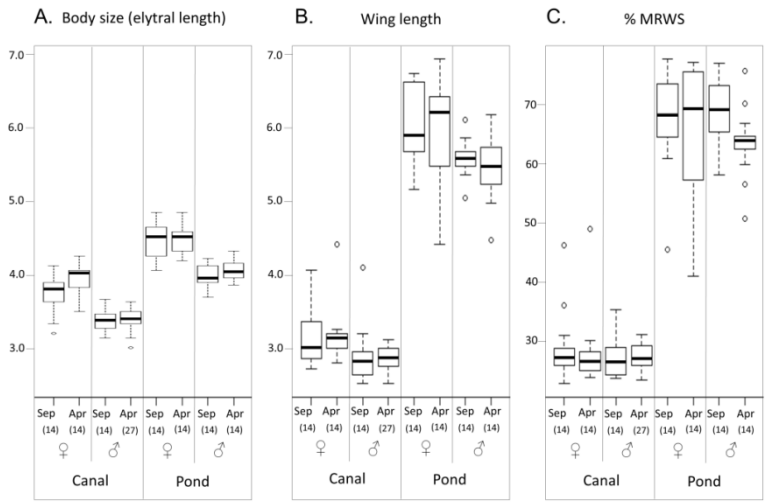


Figure 7. Box plots for body size (mm) (a), absolute wing length (mm) (b), and percentage of the maximal realizable wing size (%MRWS) (c) for the canal and pond population from which the individuals were sampled for breeding. Numbers between brackets indicate the number of individuals measured and used in the analysis. Sep = September 2010, Apr = April 2011

HERITABILITY ESTIMATES

In total, 16 parental pairs and their offspring (n = 223) were used to obtain heritability estimates. Absolute wing length and wing width showed a high heritability of $h^2 = 0.73 \pm 0.05$ and $h^2 = 0.65 \pm 0.05$, respectively (Table 1). Heritability of wing length was even closer to one when using the body size corrected %MRWS (Figure 8A; $h^2 = 0.90 \pm 0.05$). Although significant, variation in body size (elytral length) showed a significantly lower heritability (Figure 8B; $h^2 = 0.49 \pm 0.08$).

Table 1. Heritability estimates (slope of linear regression) for different traits between parents and offspring data. Heritability estimates of sympatric Guérande populations are compared to those of allopatric populations (data adopted from Desender 1989). Slope values are given with standard errors between brackets. Heritability values that significantly differed from zero ($P < 0.01$) are underlined.

Trait	Mean offspring - Midparent (h^2)	Mean offspring – Midparent (h^2) (Desender 1989)
Elytral length	<u>0.49</u> (0.08)	<u>0.68</u> (0.21)
Elytral width	0.19 (0.06)	0.12 (0.13)
Wing length	<u>0.73</u> (0.05)	<u>0.71</u> (0.08)
Wing width	<u>0.65</u> (0.05)	<u>0.65</u> (0.11)
Relative wing length	<u>0.94</u> (0.05)	<u>0.85</u> (0.06)
%MRWS	<u>0.90</u> (0.05)	<u>0.82</u> (0.07)

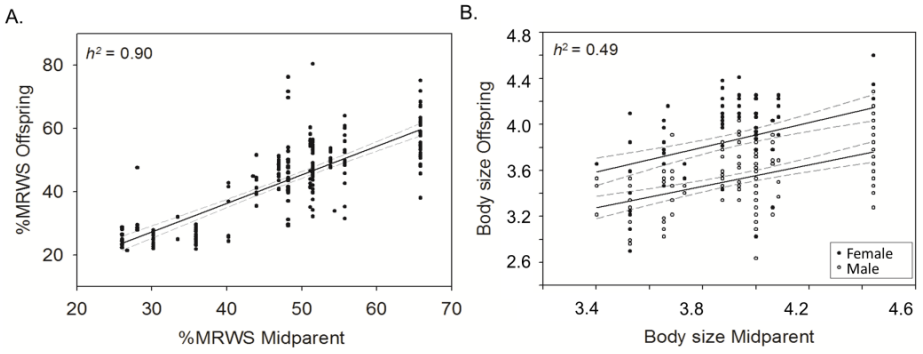


Figure 8. Midparent-offspring regressions in *P. chalceus* for the body size corrected wing size (%MRWS) (A) and body size (elytral length (mm)) (B). Heritability estimates for male and female beetles did not differ significantly.

WING SIZE – MTIDH ASSOCIATION

ASSOCIATION BETWEEN MTIDH AND WING SIZE ACROSS POPULATIONS

The association between the population frequency of the mtIDH-B allele and %MRWS, corrected for spatial autocorrelation, was very strong (Figure 9A; $F_{1,62} = 461$; $P < 0.0001$). As previously shown by Dhuyvetter *et al.* (2004), unstable Mediterranean and Atlantic populations all show high dispersal ability as well as high frequencies of the mtIDH-B allele, Atlantic populations with intermediate wing sizes had intermediate frequencies of the mtIDH-B and mtIDH-D allozymes, and Atlantic tidal populations with strongly reduced wing size had low frequencies of the mtIDH-B allele and high frequencies of the mtIDH-D allele. Besides the strong effect of %MRWS on mtIDH frequencies, we observed that a significant part of the variation in mtIDH frequencies was caused by spatial autocorrelation ($\chi^2 = 3.28$; $P = 0.035$), demonstrating that for a certain wing size, populations in closer proximity were more similar in mtIDH frequencies.

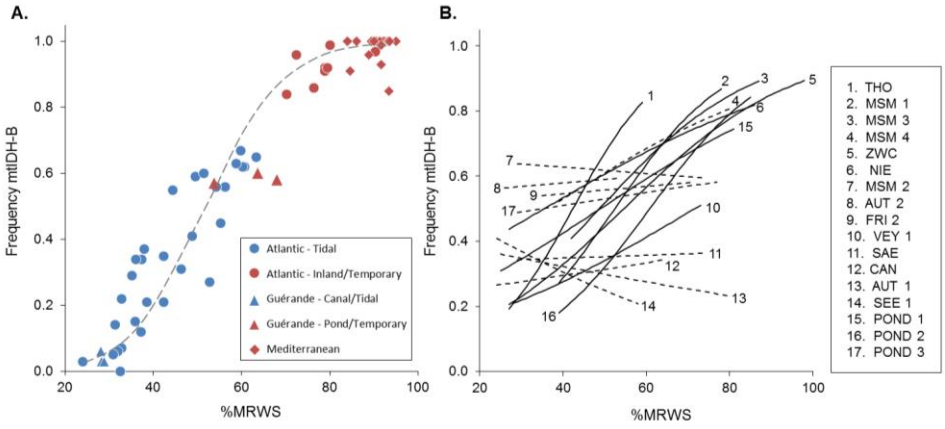


Figure 9. Comparison of the across and within population association of wing size (%MRWS) and mtIDH allozyme frequencies. (A.) Mean frequencies of the mtIDH-B allele for all studied populations compared to mean percentage of maximal realizable wing size (%MRWS) of the populations. The dashed line indicates the logistic regression of the data. (B.) Estimated logistic regression between individual mtIDH genotype and individual wing length within populations. The mtIDH – wing size association within each population was estimated by regressing the proportion of mtIDH-B alleles in each individual (mtIDH-DD: 0.0, mtIDH-BD: 0.5 or mtIDH-BB: 1.0) against its %MRWS. Logistic regression lines are plotted within the respective %MRWS range of each population. Solid black lines indicate populations in which the association is significant, dashed lines indicate non-significant associations. See Figure 6 and Appendix 1 for locations and sample sizes.

Association between mtIDH and wing size within populations

The mtIDH – wing size association within each population was estimated by regressing the proportion of mtIDH-B alleles in each individual (mtIDH-DD: 0.0, mtIDH-BD: 0.5 or mtIDH-BB: 1.0) against its %MRWS. Populations differed significantly in their association between mtIDH genotype and wing size (%MRSW x population effect; χ^2 = 36.38; P = 0.0026). Moreover, several of the investigated populations showed a strongly reduced, non-significant or even inverted relation between mtIDH genotype and wing size compared to the across population association (Figure 9B). This strongly suggests within population recombination between the mtIDH locus and genes involved in wing size.

Discussion

In this study, we present heritability estimates of a trait related to dispersal ability in a sympatric mosaic and argue that strong divergence exists in at least two genetically unlinked or weakly linked traits between both allopatric and sympatric populations. Previous work described the remarkable divergence of two ecotypes of the wing-polymorphic ground beetle *P. chalceus* in a sympatric mosaic. However, the genetic relationship of the involved traits and its implications for sympatric divergence has not been considered.

In the Guérande salterns in France, populations of the ground beetle *P. chalceus* are strongly diverged in wing size between tidally flooded (stable) canals and unpredictably flooded (temporary) pond habitats, despite the close proximity of these habitats and putatively high levels of gene flow. Our wing and body size measurements closely matched those from previous measurements of populations in the Guérande (Dhuyvetter *et al.* 2007) and showed that this distinction between the canal and pond populations has been maintained over several years and seasons.

When different phenotypes are associated with differences in environmental conditions of patches in a metapopulation, functionally adaptive phenotypic differentiation may be expected to result from a plastic response rather than through constitutively expressed genetic differences (Sultan & Spencer 2002). However, results of our study showed that for diverged populations occurring in a sympatric mosaic, variation in wing size was strongly genetically determined and significant differences did not result from phenotypic plasticity. Our heritability estimates correspond reasonably well with those found by Desender (1989) for other allopatric populations (Table 1), indicating that increased gene flow does not appear to increase plasticity of expression of variation in

the trait. Previous studies have shown that wing size does not show any plastic response towards food regime and temperature (Desender 1989a) and, therefore, genotype x environment interaction, i.e. when individual genotypes show a different plastic response towards the environment (Hoffmann & Merilä 1999), are expected to have minor effect on heritability estimates in this trait. Heritability of variation in body size was significant; however, environmental effects such as feeding conditions and temperature during larval development may have strong influences on variation in body size (Desender 1989a). Beetles with larger hind wings and functional flight muscles are found to be generally significantly larger (Desender 1985). However, the observation of lower heritability of body size compared to wing development suggests that the significant differences in body size between canal and pond populations may largely reflect the strongly different environmental conditions of these habitats as described in Dhuyvetter *et al.* (2007).

Previous work demonstrated a significant relationship between population dispersal ability and mtIDH allozyme frequencies (Dhuyvetter *et al.* 2004). However, whether these traits constitute physically linked variation or whether the association results from similar selection pressures working on different loci was not investigated. By studying the association of mtIDH allozymes and wing size within populations, we found that this correlation was strongly reduced within several populations compared to across populations (Figure 9). Such reduced association within several populations implies high recombination and, therefore, an unlinked genetic control of at least two traits involved in adaptive divergence.

However, several populations showed a significant relation between mtIDH genotype and wing size. Three factors may explain this marked association within these populations. First, population structure within the sampled populations and high gene flow between these alternatively selected patches can result in a significant association between wing size and mtIDH alleles within some of the populations. Theoretical models have shown that migration between genetically differentiated populations produces associations between alleles at different loci within populations, even when they are physically unlinked (Nei & Li 1973, Kirkpatrick *et al.* 2002). Moreover, these associations are expected to be proportional to the differences in allele frequencies between the contributing populations (Kirkpatrick *et al.* 2002). Accounting for spatial autocorrelation indeed showed that nearby populations were more similar in mtIDH allele frequencies, implying a role of gene flow and migration in the distribution of mtIDH frequencies, potentially resulting in slightly maladapted populations. The continuous distribution of wing sizes and mtIDH frequencies along the Atlantic coasts, may, therefore, result from migration and to a lesser extent from a gradient in selection pressures. Alternatively, natural selection might favor extremes and disfavor

intermediates. Hence, there might be selection for alternative combinations of traits within populations (e.g. long winged individuals with mtIDH-B and short winged individuals with mtIDH-D). Second, close genetic linkage may explain a tight association within populations. However, in this case we would not expect breakdown of this association in several other populations. Third, wing size is a polygenic trait and variation in different genes involved in wing development may result in variation in wing size in different populations. Hence, separate wing development genes may be selected to high frequency to obtain a certain wing size. If mtIDH was linked to one of these genes involved in wing development, this could explain a reduced association in some of the populations, but in this case we would also not expect to find the strong association across populations. The weak or nonexistent physical linkage of mtIDH and genes involved in wing size suggested by this study implies that similar selection pressures are affecting the *mtIdh* gene region and wing size. Whether selection is acting on the *mtIdh* gene itself or on a tightly linked locus is unclear.

Although we are not focusing on speciation and the evolution of reproductive barriers, multilocus evolution in sympatry and in the face of gene flow has long been a contentious issue (Slatkin 1987, Coyne & Orr 2004). Classic theory and empirical examples predict genomic clustering of divergent loci to reduce the blending effect of gene flow and recombination (Felsenstein 1981, Hawthorne & Via 2001, Via 2001, Rundle & Nosil 2005, Via & West 2008, Yeaman & Whitlock 2011). However, recent work has shown the possibility of moderate or weak genomic clustering of loci that are involved in adaptation when taxa diverge in the face of gene flow (Nosil *et al.* 2009a, Michel *et al.* 2010, Feder *et al.* 2012b, a). Altogether, only a few cases have been clearly identified in which genetic divergence has taken place in multiple characters despite the close proximity of differently selected environments (Nosil 2012). These study systems are extremely interesting as they allow study and identification of ecological and genetic mechanisms that drive divergence and ultimately speciation. From the allozyme level, we cannot infer the origin (i.e. single or multiple) of these alleles. Sequencing the *mtIdh* gene and further unraveling the genetic basis of wing development and the identification of genetic divergence at a genome wide scale and their evolutionary history will allow analysis of the importance of these factors in the evolution of the dispersal ecotypes found in *P. chalceus*.

ACKNOWLEDGMENTS

The present study would not have been possible without the extensive previous work of Hilde Dhuyvetter and the late Konjev Desender. Elena Dierick, Charlotte De Busschere, Bram Vanthournout, Annelies De Roissart and Viki Vandomme are thanked for help in gathering the specimens in the Guérande. Elena Dierick, Viki Vandomme and Lucien Shimirwa are gratefully acknowledged for their help in the laborious task of control and feeding of the beetles and larvae. Jonas Van Belleghem is thanked for his help with measuring of the beetles and their wings. Funding was received from the FWO-Flanders (PhD grant to Steven Van Belleghem) and the Belgian Science Policy (MO/36/025) and partly conducted within the framework of the Interuniversity Attraction Poles program IAP (SPEEDY)—Belgian Science Policy.

CHAPTER 2

EVOLUTIONARY HISTORY OF A DISPERSAL-ASSOCIATED LOCUS ACROSS SYMPATRIC AND ALLOPATRIC POPULATIONS OF A WING-POLYMORPHIC BEETLE ACROSS ATLANTIC EUROPE

Steven M. Van Belleghem [1,2]
Dick Roelofs [3]
Frederik Hendrickx [1,2]

[1] Terrestrial Ecology Unit, Biology Department, Ghent University, K. L. Ledeganckstraat 35, 9000 Gent, Belgium

[2] Department Entomology, Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000 Brussel, Belgium

[3] Department of Ecological Science, VU University Amsterdam, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands

ABSTRACT

Studying genetic variation at loci subjected to selection is necessary to understand the evolutionary mechanisms involved in adaptation. Here, we reconstruct the evolution of different alleles at the nuclear encoded mitochondrial NADP+-dependent isocitrate dehydrogenase (*mtIdh*) locus of the ground beetle *Pogonus chalceus* that are differentially and repeatedly selected in short- and long-winged populations at both allopatric and sympatric scales along the Atlantic European coasts. We sequenced 2,788 bp of the *mtIdh* locus spanning a ~7 kb genome region and compared its variation with that of two supposedly neutral genes. *mtIdh* sequences show (i) monophyletic clustering of the short-winged associated mtIDH-DE haplotypes within the long-winged associated mtIDH-AB haplotypes, (ii) a more than tenfold lower haplotype diversity associated with the mtIDH-DE alleles compared to the mtIDH-AB alleles, and (iii) a high number of fixed nucleotide differences between both mtIDH haplotype clusters and a divergence time estimated between 0.047 and 0.165 MY. Coalescent simulations further suggest that the observed sequencing variation in the *mtIdh* locus is most consistent with a relatively recent selective sweep and an origin in a large but partially isolated subpopulation. These results demonstrate that the adaptation associated with the *mtIdh* locus, which is found repeatedly in different populations, has evolved once and subsequently spread along the Atlantic coasts. Reuse of adaptive alleles, hence, plays an important role in the adaptive potential of populations when exposed to similar selection pressures and provides insights into the evolutionary history of ecologically important traits subjected to sympatric divergence.

INTRODUCTION

When organisms colonize new habitats or occupy new niches one fundamental question concerns the source of adaptive alleles, either as new or preexisting variation (Mitchell-olds *et al.* 2007, Barrett & Schluter 2008, Stern 2013, Messer & Petrov 2013). However, genetic variation is thought to be transient due to fixation by natural selection or neutral drift and, therefore, the importance of standing genetic variation in local adaptation concerns the persistence of genetic variation, either through balancing selection or local adaptation (i.e. much of the variation is maintained by natural selection) (Charlesworth 2006, Barrett & Schluter 2008) or through a mutation-selection balance (Kimura 1983, Turelli 1984). Hence, unraveling the link between DNA sequence variation and the evolutionary history of adaptive loci (including selection, migration, recombination and

demographic changes) is crucial to infer the dynamics and evolutionary processes of local adaptation (Barrett & Hoekstra 2011, Jones *et al.* 2012, Linnen *et al.* 2013, Savolainen *et al.* 2013). However, the multitude of studies inferring the ecological significance of adaptive traits still contrast strongly with the relatively few studies that thoroughly investigate patterns of variation in genes involved in local adaptation at different spatial scales. These studies have proven to be pivotal in our understanding of adaptive evolution and include a.o. nucleotide polymorphism at the alcohol dehydrogenase locus *of Drosophila melanogaster* (Hudson & Kaplan 1988, McDonald & Kreitman 1991) , spread of color-associated genes in deer mice *Peromyscus maniculatus* (Linnen *et al.* 2009, 2013) and moths (Van't Hof *et al.* 2011) and evolutionary history of the Ectodysplasin (*Eda*) gene related to lateral plate number in sticklebacks *Gasterosteus aculeatus* (Colosimo *et al.* 2005).

Finding and studying such genes can be a daunting task as it necessitates to clearly distinguish selective from demographic processes in generating genetic variation (Przeworski 2002, Li *et al.* 2012, Savolainen *et al.* 2013) . Therefore, when studying patterns of DNA sequence variation, promising study systems include those where (i) local adaptation along the same environmental gradient occurs recurrently, (ii) the geographical setting of the environmental gradient is spatially uncorrelated with the geographical distance among populations, (iii) there is a clear association between allele frequency and adaptive trait variation and (iv) sufficient gene flow exists among populations subjected to divergent selection to distinguish neutral from non-neutral processes. When these ecological and spatial settings are met, studying genetic variation at genes associated with locally adapted phenotypes enables to gain considerable insight into the process of local adaptation by distinguishing (i) a single evolution of the adaptive allele and subsequent colonization or introgression to similar pairs of habitats from, alternatively, repeated *de novo* divergence in multiple localities (Barrett & Schluter 2008, Messer & Petrov 2013), (ii) a recent selective sweep from long term (spatially heterogeneous) balancing selection (Charlesworth 2006) and (iii) sympatric (i.e. within population) versus an allopatric origin of the adaptive divergence .

One gene for which strong evidence of divergent selection has been provided is the nuclear encoded mitochondrial NADP+-dependent isocitrate dehydrogenase (*mtIdh*) locus in the wing-polymorphic salt marsh beetle *Pogonus chalceus* (Marsham 1802; Carabidae). Populations of *P. chalceus* are strongly differentiated in wing size as an adaptation to differences in habitat dynamics (Desender *et al.* 1998, Dhuyvetter *et al.* 2004). More precisely, two main habitats that select differently for dispersal ability are recognized; tidally inundated salt marshes being flooded regularly for short time periods (5-6 hours) and seasonally inundated inland salt marshes that are flooded unpredictably

for several months during winter, forcing the beetles to escape these inundations. In accordance with these habitat dynamics, populations inhabiting tidal and seasonal salt marshes have a low and high average wing size, respectively. Furthermore, across the European Atlantic and Mediterranean coast, mean population wing size in *P. chalceus* is tightly correlated with population frequencies of the mtIDH allozymes (Dhuyvetter *et al.* 2004, Van Belleghem & Hendrickx 2014) . Long-winged populations have high frequencies of the mtIDH-B allozyme and short-winged populations have high frequencies of the mtIDH-D allozyme, with wing size variation explaining up to 94.5 % of variation in mtIDH allozyme frequencies among populations (Dhuyvetter *et al.* 2004). Recent analysis showed that this association most likely results from similar selection pressures affecting mtIDH alleles and loci involved in wing size determination and not from strong genetic linkage (Van Belleghem & Hendrickx 2014). Moreover, this differentiation in at least two unlinked loci is also maintained at remarkably small geographical distances of a few meters only and, consequently, under ample opportunity for gene flow (Dhuyvetter *et al.* 2007). In contrast, neutral microsatellite and allozyme marker variation does not show any association with habitat dynamics or mean population wing sizes, indicating considerable gene exchange between both environments.
Although a biochemical link or causal association between the mtIDH allozymes and flight metabolism has not been confirmed and may be absent due to selection on a closely linked locus, the tight association of mtIDH allozymes with both dispersal ability and habitat dynamics at population level allows making inferences about the evolutionary history of the repeated evolution of populations differing strongly in dispersal capacity.

Here we analyze nucleotide polymorphism at the *mtIdh* locus among *P. chalceus* populations across the Atlantic European and the Mediterranean region to infer the evolutionary history of this selected locus and its contribution to the repeated evolution of dispersal related phenotypes. First, we report on the pattern and differentiation of *mtIdh* genetic variation and compare this with two supposedly neutral genes (cytoplasmic NADP+-dependent isocitrate dehydrogenase (*cytIdh*) gene and part of the *enolase* gene). Next, using coalescent simulations we estimate the probability of observing the *mtIdh* sequence variation pattern given different scenarios of gene flow, time since the origin of the derived alleles and population size. Finally, we calibrate divergence times between the different *mtIdh* alleles by constructing the phylogenetic relations of closely related species of both the genus *Pogonus* and *Pogonistes* using mitochondrial gene fragments with estimated substitution rates in other Coleoptera species.

These results allow us to (i) reconstruct the ancestral relationship among the different alleles, (ii) assess whether the repeated divergence is based on a singular mutational origin that subsequently spread across Atlantic Europe or evolved multiple times, and (iii) investigate the most likely scenarios of gene flow, historic population structure and time since divergence (i.e. selective sweep versus balancing selection) given the observed sequence variation patterns.

MATERIALS & METHOS

STUDY SPECIES AND SAMPLING

We selected *P. chalceus* samples from diverse geographical locations covering nearly the entire species range (Figure 10; Appendix 2). We used individuals from Atlantic populations from Belgium, England, France, the Netherlands, and Spain and Mediterranean populations from France and Spain, from which allozyme frequencies were previously obtained (Dhuyvetter *et al.* 2004, 2005b, 2007, Van Belleghem & Hendrickx 2014). Additionally, we used samples from two Portuguese populations. Samples used in previous studies were genotyped for both the mtIDH and cytIDH allozymes. Furthermore, we sequenced samples of related species belonging to the genera *Pogonus* (8, including *P. chalceus*) and *Pogonistes* (4) occurring in Europe to estimate the phylogenetic relations and age of divergence among these species (Appendix 2). This subsequently allowed calibrating the divergence time of the major branching events found within *mtIdh* locus.

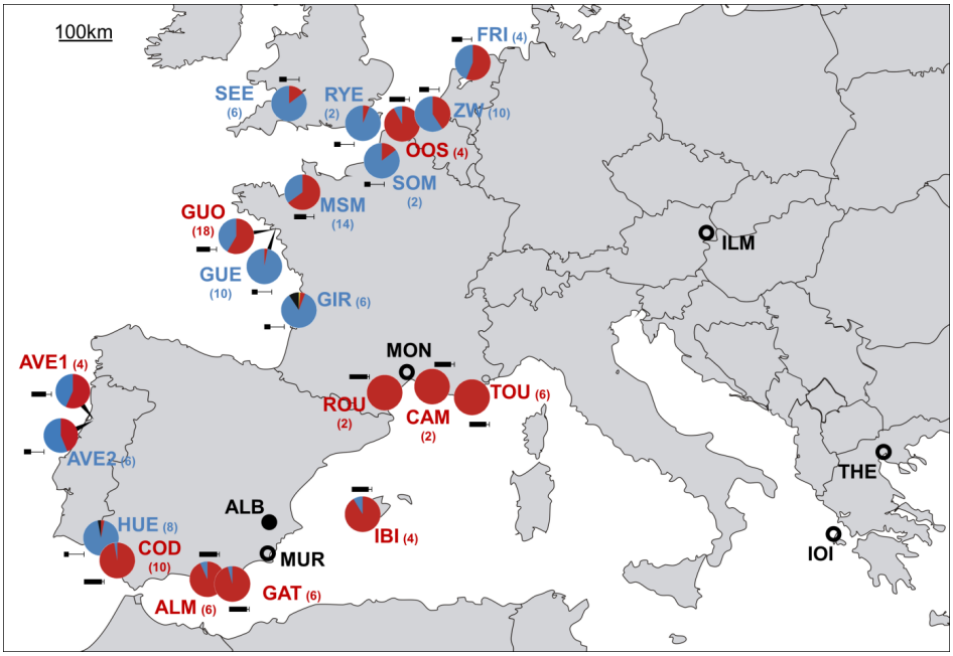


Figure 10. Sampling sites and mtIDH allozyme frequencies. Pie charts give the mtIDH allozyme distribution (mtIDH-A in yellow (only in GIR), mtIDH-B in red, mtIDH-D in blue, mtIDH-E, in black) of *P. chalceus* populations along the European coasts (see Van Belleghem and Hendrickx (2014) for more mtIDH genotyped populations and wing size distributions). Population codes in red or blue indicate seasonal or tidal habitat, respectively. Numbers indicate the amount of mtIDH haplotypes sequenced. For the ALB population no *P. chalceus* mtIDH was sequenced. Open circles indicate sites where no *P. chalceus*, but other *Pogonus* or *Pogonistes* species were sampled. See Appendix 1 for explanation of the codes and the number of sequences screened.

SEQUENCING

Two NADP+-dependent isocitrate dehydrogenases have been reported in eukaryotes, one of which is mitochondrial (mtIDH) and the other predominantly cytosolic (cytIDH) (Jenningss *et al.* 1994, Zhao & Mcalister-henns 1996). Both NADP+-dependent enzymes are homodimers encoded in the nuclear genome (Ceccarelli *et al.* 2002, Xu *et al.* 2004). According to the allozyme protocol of Hebert and Beaton (1993), the IDH protein associated with wing size in *P. chalceus* is most likely the mitochondrial IDH isozyme. However, we sequenced both the *mtIdh* and *cytIdh* gene to investigate whether sequencing variation indeed corresponds to the previously obtained allozyme variation as well as to contrast sequence variation of *mtIdh* with that of a gene with a comparable nucleotide composition but whose alleles are not associated with average population wing size. To sequence both genes, degenerate PCR primer pairs were developed by

aligning homologous sequences of other insect species from GenBank (See Appendix 3 for list of species and accession number). These primers allowed cloning and sequencing short fragments of the *mtIdh* and *cytIdh* gene. Based on the partial sequences, gene-specific primers were designed for the identification of 5′ and 3′ ends of the *mtIdh* and *cytIdh* mRNA by a RACE protocol (Roche, Inc.) (See Appendix 3 for detailed methods of DNA and RNA extraction and sequencing and Appendix 4 for primers).

We also sequenced part of the nuclear encoded *enolase* (633 bp) gene. Sequence variation analysis was performed on the genomic locus of selected *mtIdh* and the supposedly neutral *cytIdh* and *enolase* gene. A 673 bp *mtIdh* region (coding position 303 to 975 from start codon) and the complete coding sequence (1,224 bp) of the *cytIdh* gene were also sequenced for seven other *Pogonus* species and four species of the sister clade *Pogonistes* (Appendix 2). The primers did not amplify the *mtIdh* region for *Pogonus meridionalis*, probably due to sequence divergence. Subsequently, partial sequences were also obtained for the mitochondrial genes *cytochrome oxidase subunit one* (*cox1*; 1,130 bp), a mitochondrial region spanning the *NADH subunit I* (180 bp), the *tRNA-Leu* gene (64 bp) and the *16S rRNA* gene (111 bp) (*nad1*; 355 bp) and *cytochrome b* (*cob*; 468 bp) for the seven other *Pogonus* species and four *Pogonistes* species (Appendix 2). These sequences were used to estimate divergence times as substitution rate estimations are available for mitochondrial genes in beetles (Pons *et al.* 2010) . *Enolase* was also used to compare several *P. chalceus* populations. All sequences were uploaded in GenBank (*mtIdh*: KJ371353 - KJ371522; *cytIdh*: KJ371166 - KJ371315; *enolase*: KJ371316 - KJ371352; *cox1*: KJ371146 - KJ371165; *cob*: KJ371126 - KJ371145; *nad1*: KJ371523 - KJ371542).

SEQUENCE DATA ANALYSES

For all genes, we calculated the average number of pairwise differences between sequences (k) as a measure for haplotype diversity. Next, we calculated the nucleotide diversity (π) by averaging the number of nucleotide differences per site between two sequences and Watterson's θ_w as an estimate for the population mutation rate θ (= $4N_e\mu$, where N_e is the effective population size, and μ the mutation rate per sequence and per generation (Watterson 1975)). To test if the observed sequence variation deviated from the expectations of a standard coalescent process in a Wright-Fisher population (Hein *et al.* 2004) , the standardized difference between the nucleotide diversity (π) and Watterson's θ_W, known as Tajima's D was calculated (Tajima 1989) in DNAsp v5.0 (Librado & Rozas 2009). A negative Tajima's D signifies an excess of low frequency polymorphisms relative to the expectation under neutrality and is expected under population size expansion (e.g. after a bottleneck or a selective sweep) and/or purifying selection. A positive Tajima's D indicates low levels of both low and high frequency polymorphisms,

indicating a decrease in population size, population subdivision and/or balancing selection. For the *mtIdh* locus, sequences of the mtIDH-A and mtIDH-B allozymes (mtIDH-AB; associated with long-winged populations) as well as sequences from the mtIDH-D and mtIDH-E allozymes (mtIDH-DE; associated with short-winged populations) were pooled for calculating the nucleotide diversity both within and between these haplotype clusters and for counting fixed differences between these haplotype clusters. Indels occurring in the intron regions were not considered for the calculations of the sequencing statistics.

We analyzed the genealogical relations between the different haplotypes by constructing Median Joining haplotype networks using the program Network v4.6.1.1 (Bandelt *et al.* 1999) and Neighbor-Net networks using SplitsTree v4.13.1 (Huson & Bryant 2006) using default settings. Both methods allow incorporating reticulate events which indicate uncertainty of the genealogy (i.e. multiple plausible trees) and recombination.

RECOMBINATION

Recombination analyses were performed for the *mtIdh* locus on the complete dataset as well as for the mtIDH-AB allozymes to investigate differences in recombination and linkage disequilibrium (*LD*) within as well as among the differentially selected loci. We calculated the recombination parameter R (= $4N_e r$, with r the recombination rate per generation between the most distant sites) per gene and the minimum number (R_m) of recombination events based on the four gamete test (Hudson, 1987) in DNAsp v5.0 (Librado & Rozas 2009). The significance of intragenic recombination was assessed using the *ZZ* test (Rozas *et al.* 2001), which compares the average *LD* between adjacent sites with the average *LD* over all sites. In the case of recombination, *LD* is expected to decrease with distance. Wall's *Q* was used to test whether the observed *LD* deviates from neutral expectation, as excess linkage disequilibrium is expected under long-term balancing selection (Wall 1999, Charlesworth 2006). Significance of the *ZZ* test and Wall's *Q* was determined using 10,000 coalescent simulations in DNAsp v5.0 with observed values of θ_w and recombination.

COALESCENT SIMULATIONS

To gain insight into the evolutionary scenario that could have generated the observed sequence variation at the *mtIdh* locus, we performed coalescent simulations using MSMS v1.3 (Ewing & Hermisson 2010), followed by an Approximate Bayesian Computation (ABC) framework to sample combinations of likely parameter values (Beaumont 2010,

Csilléry *et al.* 2010). Simulations were performed under a model in which an allele is oppositely selected in two populations that are allowed to differ in size. These two populations represent the two different habitats that are generally inhabited by *P. chalceus* along the Atlantic coasts. No additional population structure was modeled as population structure among Atlantic populations is generally inferred to be small (F_{ST} = 0 - 0.12; Dhuyvetter et al. 2004). Given the strong support that the two alleles are differentially selected in both habitats (Dhuyvetter *et al.* 2004, 2007, Van Belleghem & Hendrickx 2014), we did not include a neutral scenario in which only demographic parameters are varied.

Coalescent simulations were implemented as follows (Figure 11). We assumed two populations, consisting of an arbitrary number of $N_e = 10^5$ diploid individuals each, that experience opposing selection at a locus of interest from a particular time S_t (scaled in $4N_e$ generations) onwards, but are exchanging the allele at a rate M (= $4N_em$, with m being the proportion of copies that are exchanged per generation). Mutation (θ = 13.2, with $\theta = 4N_e\mu$; $\mu = 3.3e^{-5}$) and recombination (ρ = 20.7, with $\rho = 4N_er$; $r = 5.3e^{-5}$) rates were estimated from the sequencing data of the mtIDH-AB haplotypes (Table 2), as this was inferred to be the ancestral allele from phylogenetic reconstructions. For the population migration rates M, a uniform prior was assumed ranging between 0 and 1,000 (m = 0 - 0.0025). Very large migration rate values were avoided as under these conditions the polymorphism was frequently lost by gene swamping (Lenormand 2002). Selection strength was implemented as S_s (= $2N_es$, with s the relative fitness of the derived allele) and discrete values of 1,000 (s = 1.005), 5,000 (s = 1.025), 10,000 (s = 1.05) and 50,000 (s = 1.25) were selected. For heterozygotes, selection strength was half compared to homozygotes. For selection time (S_t), i.e. start of selection past ward in time in units of $4N_e$ generations ago, a uniform prior was assumed between 0.01 and 2. Note that a selection time of 1 corresponds to selection acting during the average time course needed for all sequences in the sample to coalesce to one common ancestor. The selected allele was introduced by mutation at a rate θ_S ($4N_e\mu'$) of 0.01 ($\mu' = 2.5e^{-9}$). Finally, to investigate the effect of a different population size of the population experiencing positive selection on the sequencing variation for the derived allele, simulations were performed under relative sizes of the subpopulation experiencing positive selection for the derived allele (N_s) of 1, 0.5, and 0.1. See Appendix 5 for the complete MSMS code to implement this model.
We ran 12 million coalescent simulations (one million for each discrete parameter combination of S_s and N_s). From each simulation hundred sequences were sampled, 50 samples from each subpopulation. Only simulations where at least 30 of the sequences were of the derived and 30 were of the ancestral allele type were counted to the 12 million simulations and examined. Subsequently, the sequences with the derived allele

and the ancestral type were selected from these samples and the following statistics were calculated using a home-made Python script and DendroPy v3.12.0 (Sukumaran & Holder 2010): (i) average number of pairwise differences (k), Watterson's θ_w and Tajima's D among both the ancestral and derived allele sequences and (ii) and fixed differences between the derived and ancestral allele sequences. The number of fixed differences is defined as the number of sites at which all of the sequences in one sample are different from all of the sequences in a second sample.

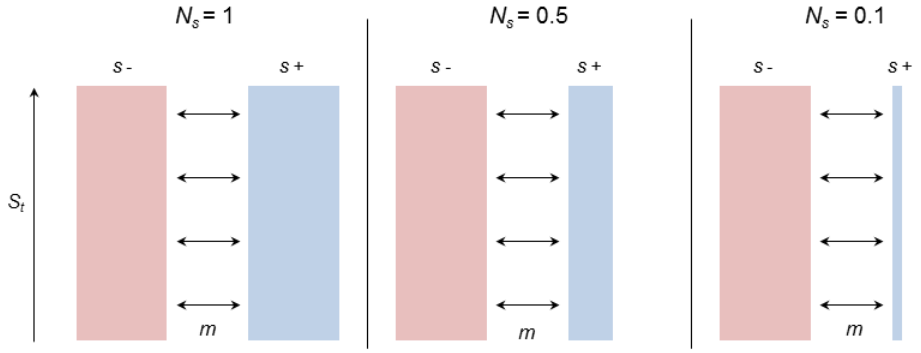


Figure 11. Demographic model used for coalescent simulations and Approximate Bayesian Computation (ABC). The parameters varied in the model are: m – proportion of gene exchange per generation, s – relative fitness of derived allele, S_t – Selection time and N_s – Subpopulation size. Blue shading indicates the population in which the derived allele is positively selected (s+). Red shading indicates the population in which the derived allele is selected against and the ancestral allele is preserved (s-).

APPROXIMATE BAYESIAN COMPUTATION

We first conducted exploratory coalescent simulations (100,000 for each discrete parameter combination of S_s and N_s) and used a principal component analysis (PCA) to test if the observed combination of summary statistics is within the range of the simulated combinations of summary statistics. This ensured that the evolutionary scenario's under which simulations were run are likely to generate the observed sequence variation. These summary statistics included (observed values between parentheses): number of fixed differences ($fixed_{DB}$ = 11), average pairwise differences among ancestral haplotypes (k_B = 15.58), average pairwise differences among the derived haplotypes (k_D = 1.19), average pairwise differences between ancestral and derived haplotypes (k_{DxB} = 25.83), average pairwise differences among all haplotypes (k_{D+B} =

18.01), Tajima's *D* among the ancestral haplotypes (0.61), Tajima's *D* among the derived haplotypes (-1.72) and Tajima's *D* among all haplotypes (0.51).
We used the R package abc (Csilléry *et al.* 2012) for sampling the closest 1 % of the complete simulation dataset to the observed summary statistics based on Euclidean distance and using the rejection method described in Pritchard et al. (1999). The parameter values of retained simulations were plotted for visualizing the parameter space for which we might expect to observe our *mtIdh* sequence variation.

DIVERGENCE TIME ESTIMATIONS

To estimate the divergence times of the *mtIdh* alleles, we first obtained estimates of the divergence time between the different *Pogonus* and *Pogonistes* species based on the mitochondrial genes (Appendix 2). The partition homogeneity test, as implemented in PAUP* 4.0 (Swafford 2003), did not show significant topological incongruence among the different markers (no *nad1*: P = 0.45, no *cob*: P = 0.71, no *cox1*: P = 0.91). Therefore, we concatenated these three gene fragments into a single alignment. Node ages were estimated by applying a nucleotide substitution rate of 0.0563 ± 0.00196 nucleotide substitutions/site/MY on the concatenated dataset. This substitution rate was calculated from rate estimates for these genes across 15 Coleoptera species (Pons *et al.* 2010), weighted by their respective sequence length. Based on the Akaike Information Criterion as implemented in MrModeltest v2 (Nylander 2004), a General Time Reversible model with estimated base frequencies, invariant sites and gamma distributed rate variation among sites was used as substitution model (GTR+I+G). As we could not reject a molecular clock when comparing likelihood ratio scores of a clock and non-clock tree obtained in PAUP*4 (χ^2 = 19.8, df = 14; P = 0.14) (Muse & Weir 1992), we used a strict clock as implemented in BEAST v1.7.5 (Bouckaert *et al.* 2014). The tree prior was set to the Yule process of speciation using standard priors. Two MCMC chains were run for 100 million generations, sampling every 1000 generations. The two chains were combined using Logcombiner v1.7.5 (Bouckaert *et al.* 2014). Convergence of the chains, appropriate burn-in (4000) and effective sample sizes of the parameters were checked using Tracer v1.5 (Drummond & Rambaut 2008). The phylogenetic analysis of the concatenated mitochondrial gene set and *mtIdh* revealed several well supported clades.
Next, the estimated divergence time between *P. chalceus* and its sister clade was used to calibrate the *mtIdh* tree and estimate the divergence time of the *mtIdh* alleles. Sequence data covering 673 bp (position 303 up to 975 from start codon) of the *mtIdh* gene for several species of the genus *Pogonus* and its sister genus *Pogonistes* (Appendix 2) were used to construct a phylogenetic tree of the *mtIdh* gene with BEAST v1.7.5. We calibrated

the *mtIdh* gene tree by using the divergence date of the most recent common ancestor between *P. chalceus* and its sister clade (including *P. littoralis, P. gilvipes, P. luridipennis, P. olivaceus* and *P. riparius*). This allowed calculating the substitution rate of the *mtIdh* gene and estimate the time of the most recent common ancestor (TMRCA) of the mtIDH-AB and mtIDH-DE alleles. As for the mitochondrial gene set, a strict molecular clock could not be rejected (χ^2 =29.1, df =29; P=0.46). BEAST v1.7.5 was run with equal parameters as for the mitochondrial gene set (GTR+I+G substitution model, strict clock, two chains of 100 million generations and burn-in of 4000).
The obtained species tree topology of the concatenated mitochondrial and 673 bp *mtIdh* dataset were checked using MrBayes 3.2 (Ronquist *et al.* 2012) and by constructing MrBayes species trees using the *enolase* and *cytIdh* gene sequences. Four simultaneous chains were run for twenty million generations using a GTR+I+G substitution model. The first 1000 trees were discarded and trees were sampled every 1000 generations.

RESULTS

MITOCHONDRIAL NADP$^+$-IDH (mtIDH) GENE STRUCTURE

The *mtIdh* gene and its putative promoter sequence span a total genomic region of about 7 kb (Figure 12A-B). cDNA cloning and genomic sequencing revealed 8 exons and 2 splice variants, resulting in an additional 14 amino acid residues at the C-terminal end when intron 7 is spliced. The two resulting proteins contain respectively 451 and 465 residues. Available RNAseq data (Van Belleghem *et al.* 2012) suggest low expression of the spliced variant in the larval (1%) and adult (2%) beetle stage and no expression in the pupal stage. Estimated from sequencing the spliced mRNA region from 14 samples (2 mtIDH-BB, 3 mtIDH-BD, 9 mtIDH-DD), the two splice variants seem to be present both in individuals genotyped as mtIDH-BB as well as mtIDH-DD.
Seventy bp and 251 bp of the 5′ and 3′ UTR, respectively, were obtained from *mtIdh* cDNA. The promoter sequence is estimated to be 572 bp in length and presumably also controls the expression of the upstream NADP$^+$-transhydrogenase (*Nnt*) that is transcribed on the reverse strand (see Appendix 3). The first 162 bp of the coding mRNA or 54 amino acids at the N-terminus of the mtIDH protein show little homology with the protein sequence of mtIDH of other eukaryotes. Potentially, these amino acids form a transit peptide involved in the translocation of the protein product to the mitochondria, as is found in e.g. porcine (UniProtKB: P33198) and human (UniProtKB: P48735) mtIDH. The complete sequence of intron 1 was not determined by sequencing; its length, about 4

kb, was determined from gel electrophoresis of an amplified fragment containing the intron.

mtIDH SEQUENCE VARIATION

A total of 104 complete *mtIdh* sequences and 128 sequences of the coding region were obtained from 22 different populations (Figure 10). For the *mtIdh* coding region, we found 24 segregating sites, of which 10 were at non-synonymous sites (Table 2; see Appendix 6 for a description of amino acid variation). Considering the full sequence length (2,788 bp; 104 sequences), we found 80 segregating sites. The estimated θ_w per sequence and average number of pairwise differences (k) for the total *mtIdh* locus was 15.39 and 18.01, respectively. The intron sequences were much more variable than exon sequences (Table 2). Variability in the promoter sequence was comparable to variability in the intron sequences.

The nucleotide diversity of the sequences coding for the mtIDH-DE allozymes, which are associated with the short-winged populations, was considerably reduced and about ten times lower (θ_W = 2.75 and k = 1.19) compared to the nucleotide diversity of sequences associated with the mtIDH-AB allozymes (θ_W = 13.2 and k = 15.58; Figure 12C). The differentially selected mtIDH-B and mtIDH-D allele are distinguished by only a single charge-changing amino acid substitution (Lys - Asn) at amino acid position 447. However, the mtIDH-AB and mtIDH-DE haplotypes showed a fixed difference at 11 sites. None of these fixed differences were located in the promoter region, 9 were located in the intron sequences and 2 in the exons (Figure 12B). Of the 55 haplotypes, 46 belonged to the mtIDH-AB sample and 9 to the mtIDH-DE sample. To obtain a statistic for expressing demographic changes in the mtIDH-AB and mtIDH-DE haplotypes, Tajima's D was calculated. Tajima's D for the *mtIdh* locus was 0.51. Considering the mtIDH-AB and mtIDH-DE haplotypes separately, Tajima's D was 0.61 and -1.72, respectively.

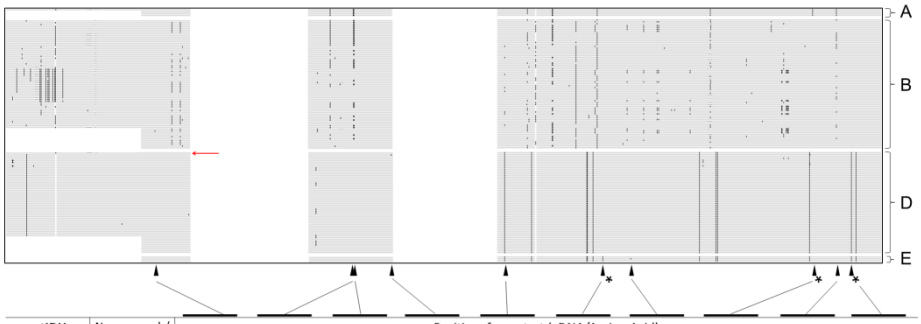
Haplotypes of the mtIDH-DE allozyme cluster monophyletically, which supports a singular mutational origin (Figure 13). Haplotype associations indicate extensive mixing along the Atlantic and Mediterranean European coasts (Figure 13 and Appendix 7 for the coding *mtIdh* sequence). Within the mtIDH-AB haplotypes, there are two main haplotype clusters, one of which is restricted to the Iberian Peninsula and the Mediterranean part of France. However, these clusters are not differentiated by any fixed differences. Further investigation of these two clusters indicates that the lack of fixed differences between these two haplotype clusters likely results from recombination between diverged haplotypes.

A. Genetic variation at the mitochondrial NADP⁺-dependent isocitrate dehydrogenase

mtIdh gene	Exon 1		Exon2		Exon3		Exon 4		Exon5		Exon6		Exon7		Exon8	
	5′ UTR	Intron1		Intron2		Intron3		Intron4		Intron5		Intron6		Intron7		3′ UTR
Transit peptide		1-85		1-77												
Splice variant 1	573	85	±4,000	187	488	107	63	202	58	237	286	252	59	282	74	43	139
Splice variant 2	573	85	±4,000	187	488	107	63	202	58	237	286	252	59	283	-	-	255

B.

C.



D.

mtIDH allozyme	N sequenced / frequency	Position from start (cDNA/Amino Acid) 65/22	104/35	106/36	242/81	292/98	599/200*	721/241	1,171/391*	1,276/426	1,339/447*
A	4 / 0.03	Ala	Thr	Gly	Cys	Val	Gly	Leu	Asn	Ile	Lys
B	10 / 0.08	.	.	.	.	.	.	.	Asp	.	.
	5 / 0.04	.	Ile	Ser	.	.	.	.	Asp	.	.
	3 / 0.02	.	Ile	Ser	.	Ile	.	.	Asp	.	.
	5 / 0.04	.	Ile	Ser	.	Ile	.	.	Asp	Val	.
	21 / 0.16	.	.	.	.	.	.	.	Asp	Val	.
	1 / 0.01	.	.	.	.	Ile	.	.	Asp	.	.
	22 / 0.16	.	Ile	Ser	.	.	.	.	Asp	Val	.
	1 / 0.01	Val	Ile	Ser	.	.	.	.	Asp	Val	.
D	52 / 0.41	.	Ile	Ser	.	Ile	.	.	Asp	Val	Asn
	1 / 0.01	.	Ile	Ser	Tyr	Ile	.	.	Asp	Val	Asn
E	2 / 0.02	.	Ile	Ser	.	Ile	Glu	.	Asp	Val	Asn
	1 / 0.01	.	Ile	Ser	.	Ile	Glu	Phe	Asp	Val	Asn

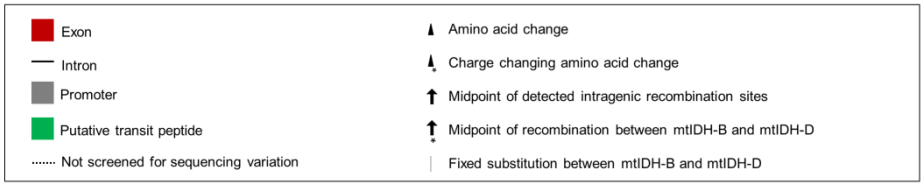


Figure 12. Genomic structure and genetic variation of the *P. chalceus mtIdh* gene. (A.) Length (bp) of the exons, introns, promoter and transit peptide for the *mtIdh* gene. (B.) Scaled diagram of the *P. chalceus mtIdh* gene, showing exons, introns, promoter, transit peptide and alternative splice variant. Arrows mark the midpoint of detected intragenic recombination. Recombination and fixed mutations between the divergently selected mtIDH-B and mtIDH-D allozyme haplotypes are also indicated. (C.) Variation in the sequenced *mtIdh* region. Rows are sequenced individuals; black dots mark variable positions; the red arrow indicates a recombinant haplotype between a mtIDH-B and mtIDH-D haplotype. (D.) Positions of non-synonymous nucleotide substitutions along the cDNA within the mtIdh gene. Amino acid names are according the IUPAC code. Charge-changing amino acid variants defining the EM classes are indicated with an asterisk.

Table 2. Sequencing statistics. Indels were not considered.

	length (bp)	Allozyme	N	S	h	π	k	F	θ_w	*D*	R	R_m
mtIdh	1,395	A and B	72	18	30	0.0030	4.20	-	3.71	0.35	62.30	3
coding		D and E	56	4	5	0.0002	0.25	-	0.87	-1.59	-	-
		ALL	128	24	35	0.0034	4.80	-	4.42	0.22	11.70	4
		Between	-	-	-	-	6.86	2	-	-	-	1
mtIdh	820	A and B	72	29	35	0.0078	6.03	-	5.98	0.02	3.50	6
intron		D and E	56	5	5	0.0007	0.55	-	1.09	-1.16	-	-
		ALL	128	40	40	0.0116	8.95	-	7.37	0.65	4.30	8
		Between	-	-	-	-	14.01	9	-	-	-	1
mtIdh	573	A and B	58	16	14	0.0104	5.48	-	3.46	1.77	0.70	3
promotor		D and E	44	5	4	0.0009	0.48	-	1.15	-1.45	-	-
		ALL	104	20	17	0.0084	4.38	-	3.85	0.24	0.001	2
		Between	-	-	-	-	5.11	0	-	-	-	-
mtIdh	2,788	A and B	58	61	46	0.0058	15.58	-	13.18	0.61	20.70	11
(total)		D and E	44	12	9	0.0004	1.19	-	2.76	-1.72	-	-
		ALL	104	80	55	0.0067	18.01	-	15.39	0.51	8.30	13
		Between	-	-	-	-	25.83	11	-	-	-	1
cytIdh	1,224		120	41	56	0.0071	8.58	-	7.65	0.29	27.80	11
Enolase	633		34	17	15	0.0077	4.87	-	4.16	0.57	13.90	3

N = Number sequenced
θ_w = Mutation rate (4Neμ) (per sequences, Watterson estimator)
S = Segregating sites
h = Haplotypes
π= Nucleotide diversity
D = Tajima's D
k = Average pairwise number of nucleotide differences
R = Recombination rate (4Ner) (Hudson 1987)
F = Number of fixed differences
R_m = Minimum number of recombination events

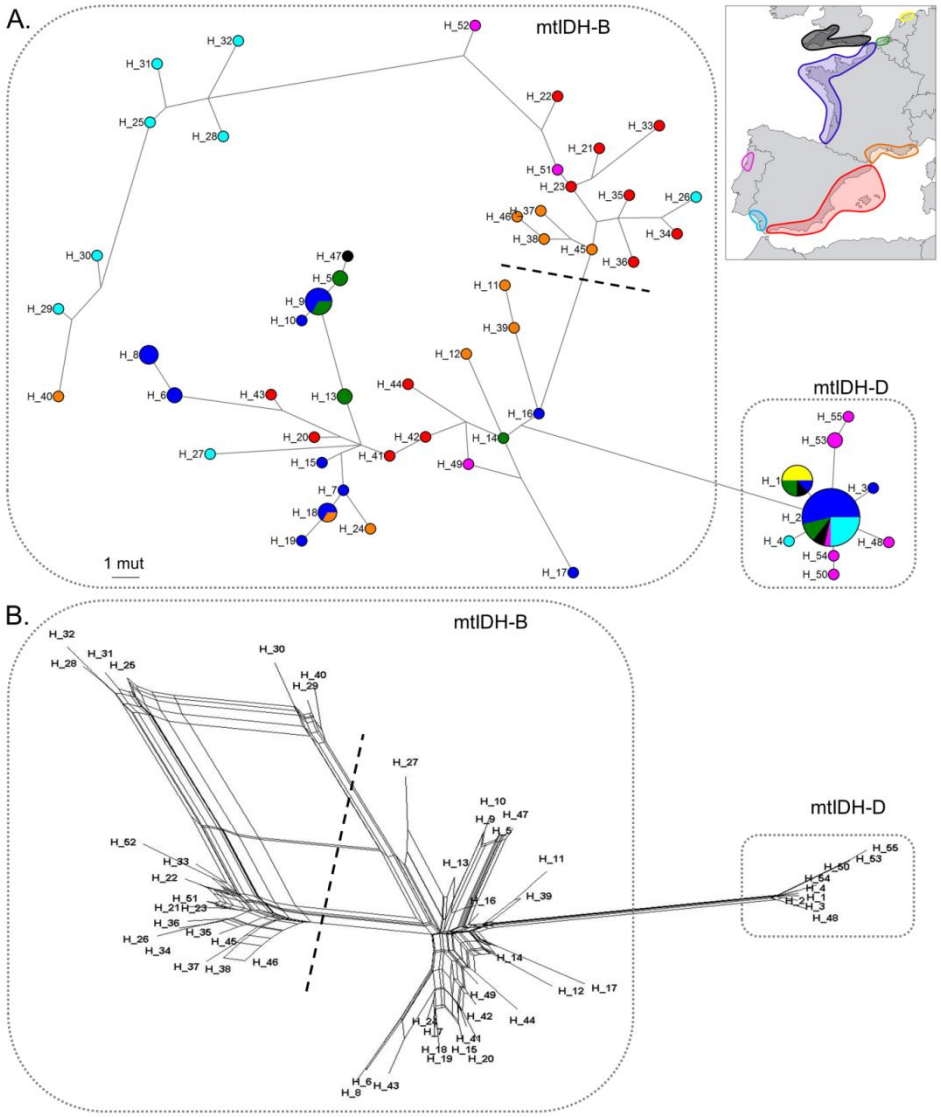


Figure 13 Median joining (A.) and Neighbor-Net (B.) network for 2,788 bp sequences of the *mtIdh* locus. The mtIDH-A and mtIDH-E haplotypes are not included in this figure as their promoter sequence was not obtained (see Appendix 7). Size of the pie charts indicates the relative frequency of the haplotypes. Colors in the network match with the shaded areas on the map of Europe. The dashed line indicates the split between two main haplotype clusters in the mtIDH-B allozyme.

cytIDH AND ENOLASE SEQUENCE VARIATION

We sequenced two supposedly neutral loci to compare nucleotide variation with the selected *mtIdh* locus. In the 120 coding sequences of the *cytIdh* gene, we found 41 segregating sites of which 7 were non-synonymous (Table 2; see Appendix 6 for a description of amino acid variation and gene structure; Appendix 8). The estimated θ_w per sequence and average number of pairwise differences (*k*) was 7.65 and 8.58, respectively. We found 56 haplotypes in the *cytIdh* sequence dataset. The Median Joining network indicates a high degree of haplotype interchange between the Atlantic and Mediterranean populations, however, one haplotype cluster is restricted to the Atlantic coasts (Appendix 9A). We found no clusters with a number of fixed differences larger than one and the Neighbor-Net network shows little indication of geographical structuring and relatively homogeneous recombination among haplotypes (Appendix 9B). We found 17 (2 non-synonymous) segregating sites in the *enolase* gene fragment, resulting in 15 haplotypes (Appendix 10). θ_w per sequence and average number of pairwise differences (*k*) was 4.12 and 4.87 for the *enolase* gene sequence. In the *enolase* median joining haplotype network, we found one haplotype cluster which is differentiated by two fixed differences and is restricted to the Atlantic coasts. Tajima's *D* for *cytIdh* and *enolase* was 0.29 and 0.57, respectively.

RECOMBINATION

The *ZZ* test statistic indicated a significant decay of linkage disequilibrium (*LD*) with physical distance due to recombination within the *mtIdh* locus spanning an approximately 7 kb genome region (*ZZ* = 0.20; P = 0.002). The estimate of the recombination rate ($R = 4N_e r$) is 8.3 for the total *mtIdh* dataset and 13 recombination events were detected by the four-gamete test. However, *R* is estimated much higher among the mtIDH-AB haplotypes (*R* = 20.7) among which 11 recombination events are detected. Due to low variability, *R* could not be estimated among the mtIDH-DE haplotypes. We found one clear recombinant haplotype between a mtIDH-B and mtIDH-D haplotype (GenBank Acc.: KJ371365). This recombination event included the promoter sequence and likely occurred near or in intron 1 which has a length of about 4 kb. In contrast, stronger breakdown of pairwise *LD* among the mtIDH-AB haplotypes compared to the total dataset including the mtIDH-DE haplotypes indicates reduced recombination between the mtIDH-AB and mtIDH-DE haplotypes (Figure 14). The estimated recombination rate for the *cytIdh* and *enolase* gene was 27.8 and 13.9 respectively.

Excess LD is expected under long-term balancing selection or when an allele has recently undergone a selective sweep. Using the total dataset including the mtIDH-AB and mtIDH-DE haplotypes, the Wall's *Q* statistic indicated that *LD* across the *mtIdh* locus is significantly higher than expected under the neutral model given the observed levels of polymorphism and recombination (Wall's $Q = 0.27$, $P = 0.016$). Using only the mtIDH-AB haplotypes, the significance of the Wall's *Q* statistic was reduced (Wall's $Q = 0.28$, $P = 0.054$). However, as a significant Wall's *Q* statistic may also be expected to result from population subdivision and expansion, we compared these statistics also for the other nuclear gene fragments. For all these fragments, the *ZZ* and Wall's *Q* statistic were both small and not significant for the *cytIdh* (*ZZ* = 0.04, P = 0.55; Wall's $Q = 0$, $P = 0.91$) and *enolase* (*ZZ* = 0.05, P = 0.18; Wall's $Q = 0.24$, $P = 0.35$) sequences.

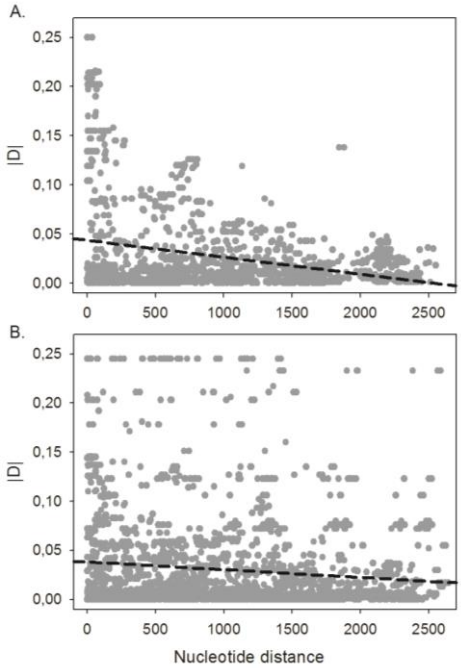


Figure 14. Linkage disequilibrium (*LD*) in the *mtIdh* locus. Each point represents pairwise *LD* (|D|) between two polymorphisms. The dashed line indicates the linear regression of *LD* and nucleotide distance. (A.) *LD* within the mtIDH-AB haplotypes. (B.) *LD* within all haplotypes (mtIDH-AB and mtIDH-DE). The stronger breakdown of *LD* among the mtIDH-AB haplotypes compared to the total dataset indicates reduced recombination between the mtIDH-AB and mtIDH-DE haplotypes.

COALESCENT SIMULATIONS

DNA sequences from oppositely selected alleles in two populations were simulated to estimate the parameters that were most likely to result in the observed sequencing variation both within and between the differentially selected mtIDH allozymes under different evolutionary scenarios. More precisely, we inferred under which evolutionary scenarios of gene flow (M), selection time (S_t), selection strength (S_s) and relative subpopulation size (N_s) we can expect the observed number of fixed and average pairwise differences between the ancestral and derived haplotypes and the observed average number of pairwise differences, haplotype diversity and Tajima's D within each haplotype cluster.

When assuming an equal size of the populations experiencing opposing selection, (Figure 15, left panels), only recent evolution of the derived allele was supported ($S_t \approx 0.2$) (scenario 1). Under these conditions, a selective sweep can explain the low average pairwise differences (k) among the derived haplotypes as there has been little time for mutation and recombination to increase the nucleotide diversity among the derived haplotypes. However, the simulations indicate that in this scenario the derived allele must have evolved under low levels of gene flow between both subpopulations, except when selection is very strong (Figure 15). Only under these conditions of low gene flow, the observed large number of fixed differences can be expected as reduced migration rates or gene flow between the populations reduces recombination between haplotypes from the different populations, resulting in a deeper split and strongly distinct haplotypes in which the derived mutation could evolve. Further, selection strength (S_s) strongly affects the plausible migration rates (M) for which we can observe our data, with higher selection strength generally allowing higher migration rates. This is most likely because a higher selection strength reduces effective migration and recombination between the alleles from the different populations.

Assuming a smaller size of the population experiencing positive selection for the derived allele (Figure 15, right panels), a more ancient evolution of the derived allele matched more closely with the observed sequence variation ($S_t > 0.5$) (scenario 2). Under these conditions, the longer selection times may explain the observed large number of fixed differences, as more mutations arise and build up in the haplotypes as time progresses. However, an increased selection time is also expected to result in a higher average number of pairwise differences (k) among the derived haplotypes. Therefore, in this scenario, a reduced subpopulation size (N_s), of the population in which the derived allele is selected, can explain the reduced average number of pairwise differences (k) among the derived haplotypes. Further, the selection time needed to observe our summary statistics is influenced by the migration rate, which, together with recombination counteracts the buildup of fixed differences. Additionally, this relation is affected by

selection strength (S_s). Again, selection strength (S_s) reduces effective migration and recombination and reduces the selection time needed to observe a large number of fixed differences. Hence, in contrast to the scenario of equal subpopulation sizes the small subpopulation size (N_s) of the population in which the derived allele is selected resulted in a low average number of pairwise differences (k) among the derived haplotypes. Subsequently, this allows longer selection times and, moreover, allows higher migration rates between the differently selected populations to explain the observed data.

Both scenarios are expected to result in a different distribution of Tajima's D values among the derived haplotypes. Negative Tajima's D values signify an excess of low frequency polymorphisms relative to expectation, which generally indicate population size expansion after a bottleneck or a selective sweep. When selection is acting in a large population (i.e. first scenario), negative Tajima's D values are only obtained when selection time is short as the derived allele is swept to high frequencies (Figure 16). In these simulations, the excess of low frequency polymorphisms disappears as selection time progresses and Tajima's D consequently approaches zero. Under this scenario, a short selection time ($0.1 < S_t < 0.3$) was most supported to explain the sequence variation. These selection times are expected to result in Tajima's D values that are within the range of the observed value (D = -1.72). In contrast, when the subpopulation size (N_s) of the population in which the derived allele is selected is small (i.e. second scenario), highly negative Tajima's D values are only obtained for very low selection times ($S_t <$ 0.1), followed by a wide range of negative values when selection time progresses (Figure 16). Given that the observed sequence variation under the second scenario matched more closely with longer selection times of 0.5 to more than 1.5 (Figure 15, right panels), the expected Tajima's D values under these conditions are less likely to result in the observed Tajima's D value of -1.72 (Figure 16).

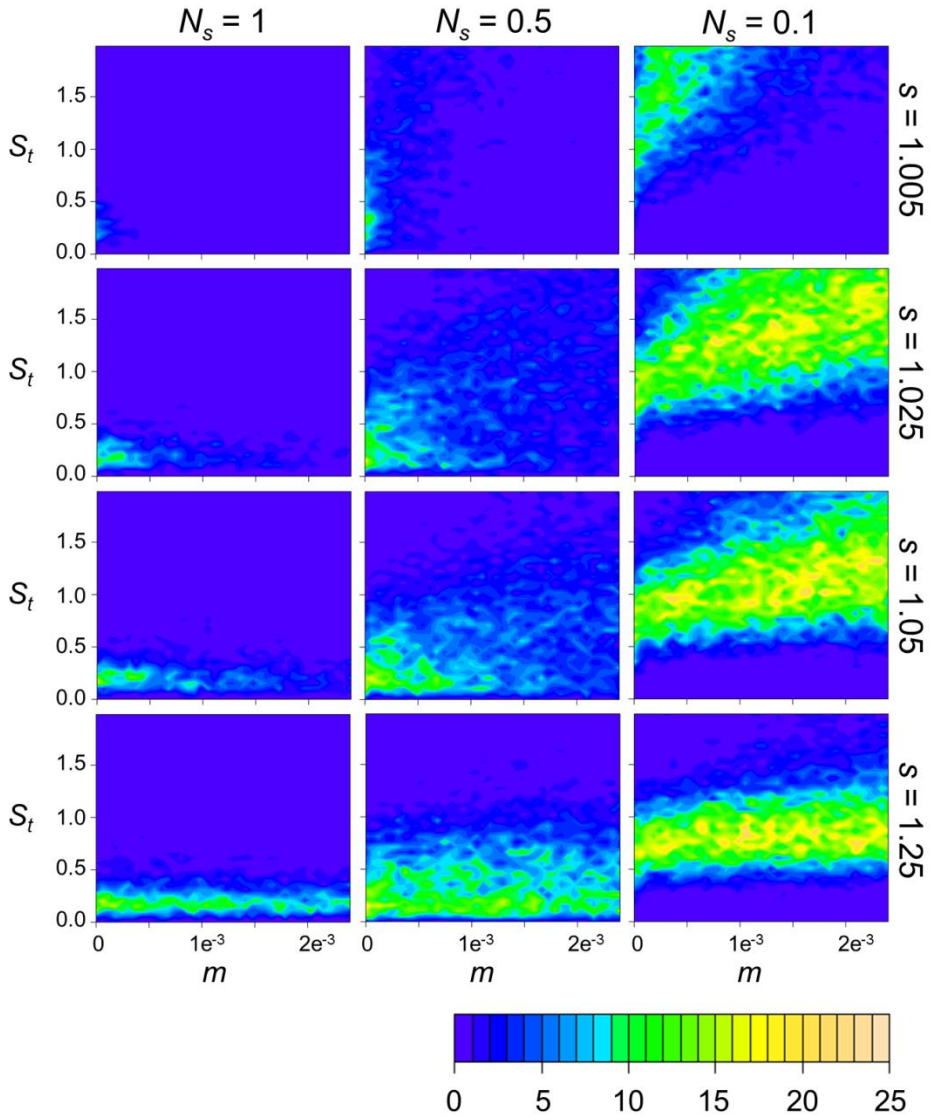


Figure 15. Frequency plots of the coalescent simulation parameters with sequencing statistics close to the observed sequencing statistics found in the *mtIdh* gene using Approximate Bayesian Computation (ABC). The parameters varied in the model are: m – proportion of gene exchange per generation, *s* – relative fitness of derived allele, S_t – Selection time and N_s – Subpopulation size. Colors indicate the sampling frequency. See Materials and Methods for the summary statistics used in the ABC.

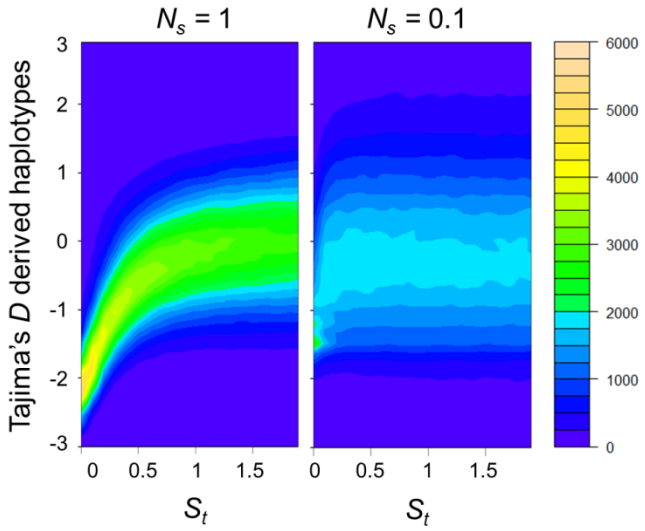


Figure 16. Comparison of Tajima's *D* values among the derived haplotypes between simulations in which the subpopulation size (N_s) of the population in which the derived allele was selected is 1 compared to 0.1. Data is shown for 1,000,000 simulations for which the relative fitness of the derived allele was s = 1.05. The proportion of gene exchange per generation (*m*) ranged between $2.5e^{-6}$ and $2.5e^{-3}$.

PHYLOGENY AND DIVERGENCE TIME ESTIMATIONS

The reconstructed phylogeny of species of the *Pogonus* and *Pogonistes* genera based on the concatenated mitochondrial dataset (*cox1*, *nad1*, *cob*) suggests that the split between these two genera has occurred between 0.95 and 1.38 MY ago (Figure 17A). Sequenced species of the *Pogonus* genus share a common ancestor between 0.83 and 1.22 MY. The major phylogenetic relations among the *Pogonus* and *Pogonistes* species were confirmed by means of trees generated by MrBayes using the concatenated mitochondrial, the 673 bp *mtIdh, enolase* and *cytIdh* gene sequence datasets (Appendix 11).

P. chalceus clusters within a well-supported clade that also contains the species *P. riparius*, *P. olivaceus*, *P. luridipennis*, *P. gilvipes* and *P. littoralis.* Calibrating this node in the *mtIdh* gene tree using the divergence date estimated from the concatenated mitochondrial gene tree (0.618 ± 0.06 MY) allowed estimating the divergence time of the *mtIdh* alleles. Based on this calibration point, the divergence between the mtIDH-AB and mtIDH-DE haplotypes was estimated between maximally 0.047 and 0.165 MY ago (Figure 17B).

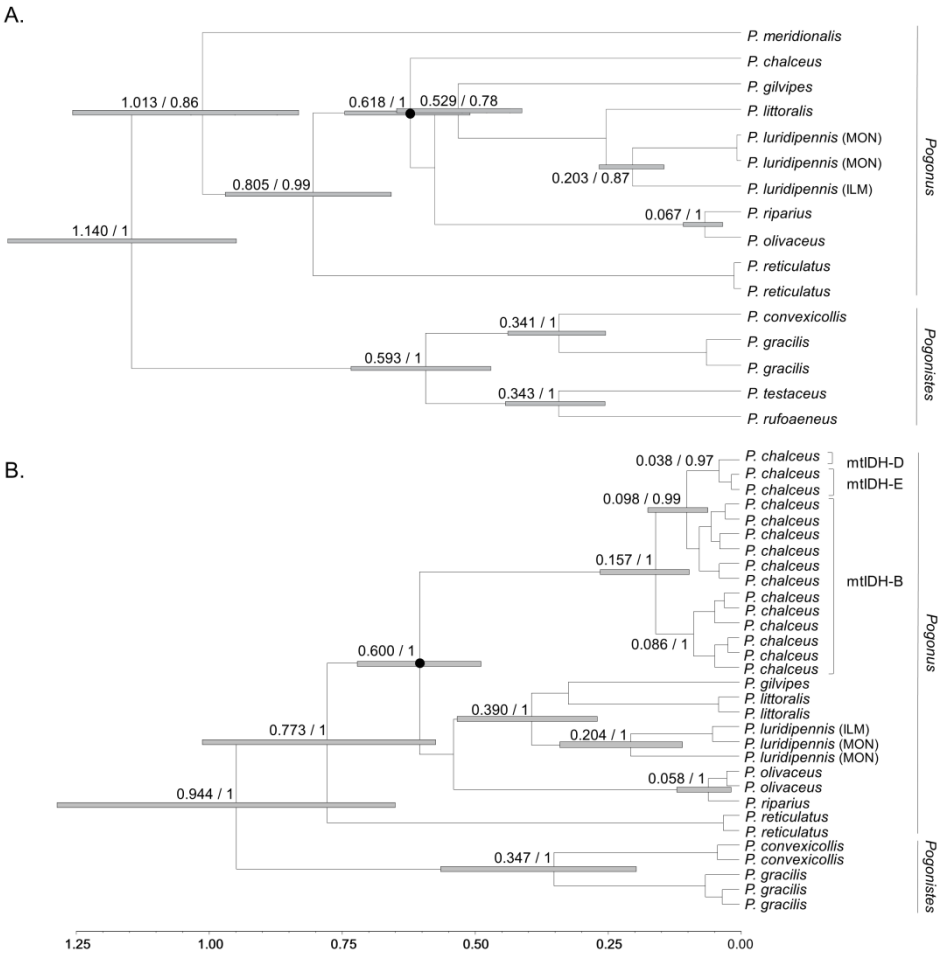


Figure 17. Divergence time estimates. (A.) Divergence time estimates (MY) between *Pogonus* and *Pogonistes* species based on the concatenated mitochondrial sequences. The black dot indicates the node that was used for calibrating the *mtIdh* gene trees. (B.) Divergence time estimates for part of the *mtIdh* gene (673bp). Gray bars, representing 95% confidence intervals of the estimated divergence times, are only shown for nodes with relatively high confidence based on MrBayes analysis (see Appendix 11). Node values represent estimated ages and Bayesian posterior probabilities, respectively. The x-axis shows time in million years (MY) before present.

DISCUSSION

The main findings reported in this study can be summarized as follows: (i) The mtIDH-DE haplotypes, associated with short-winged populations, are derived and monophyletic, indicating that these haplotypes have a singular mutational origin. The association between the mtIDH-DE allozymes and habitat stability and short wing size that is repeatedly found across Atlantic European coasts most likely results from the spread of this allele into other locations and populations. (ii) Coalescent simulations suggest that the observed sequence variation among the mtIDH allozymes, including a high number of fixed differences between the differently selected mtIDH allozymes and a low average number of pairwise differences among the derived haplotypes, indicates that the allele evolved most likely recently in a large population that is partially isolated from the ancestral population.

SINGLE ORIGIN OF REPEATED ADAPTATION

As all related species of *P. chalceus* studied are long-winged, it can be expected that the long-winged ecotype is the ancestral state. Furthermore, the long-winged ecotype can colonize new locations, which may then locally adapt to these environments. In this way, distinct populations of *P. chalceus* could have evolved reduced wing sizes along the Atlantic European coasts repeatedly as a response to differences in habitat dynamics. For instance, populations in the historical salt fields in the Guérande (France) strongly diverged in wing sizes, despite that distances between the habitats lie within a single generation walking distance (Dhuyvetter *et al.* 2007). Correspondingly, in this study, a long-winged population has been found in a temporary salt extraction pond in Aveiro (Portugal) surrounded by tidal marshes in which short-winged populations occur. Hence, local adaptation even occurs within sympatric mosaics with ample opportunity of gene flow. Furthermore, divergence between ecotypes based on microsatellite markers appeared to be negligible, indicating high levels of gene flow among ecotypes (Dhuyvetter *et al.* 2007). Allozyme frequencies of the mtIDH protein are strongly associated with this divergence in wing sizes and the monophyletic clustering of the mtIDH-DE haplotypes indicates a single origin of the adaptation associated with this locus, followed by a subsequent spread to other populations. This suggests that repeated local adaptation associated with the *mtIdh* locus is based on repeated colonization and/or introgression of genetic variation.

TRANSPORTER PROCESS

Although the mtIDH-D allozyme is associated with the low dispersal morph and shows little nucleotide diversity, it has a very wide distribution along the Atlantic European coasts. Therefore, it might be speculated that this allele is recessive, or at least codominant, compared to the allele that is frequent in the high dispersal morph. Codominance or recessiveness could easily allow the allele to be spread by long-winged individuals in analogy with the 'transporter process' hypothesized for the *Eda* locus in sticklebacks (Schluter & Conte 2009, Bell & Aguirre 2013). The similarities of the local adaptation dynamics of *P. chalceus* to that of sticklebacks emphasizes the importance of this transporter mechanism for the recurrence of similar ecotypic divergence that is facilitated by introgression of adaptive variation. Although the transporter hypothesis was not explicitly tested in *P. chalceus*, theoretical and simulation studies suggest that migration rates may be very small for adaptive alleles to spread to other populations (De Busschere *et al.* under review, Messer & Petrov 2013).

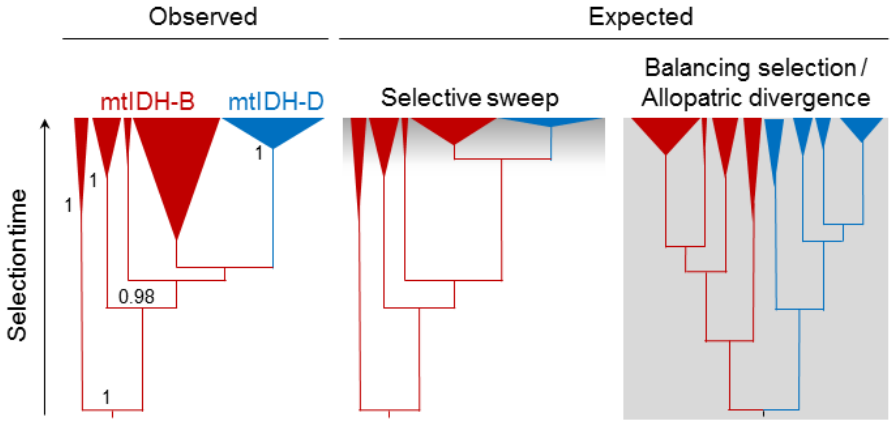
DEEP DIVERGENCE AND REDUCED RECOMBINATION BETWEEN MTIDH-AB AND MTIDH-DE HAPLOTYPES

Divergence between the mtIDH-AB and mtIDH-DE haplotypes was estimated at 0.047 to 0.165 MY ago. Node ages were estimated with a nucleotide substitution rate of 0.0563 ± 0.00196 nucleotide substitutions/site/MY (Pons *et al.* 2010), which is five times higher than the proposed standard rate of 2.3% divergence/MY (0.0115 nucleotide substitutions/site/my) for the insect mitochondrial genome (Brower 1994). Therefore, the estimated divergence time between the mtIDH-AB and mtIDH-DE haplotypes may be interpreted as a lower-bound estimation of the actual divergence time and indicates that this divergence may predate the end of the last glacial period which occurred approximately between 0.01 and 0.11 MY ago. Considering these time scales, if disruptive selection operates on different alleles within a single population, as for example under true balancing selection or high gene flow, recombination is expected between the positions linked to each allele and, therefore, a relatively low number of fixed differences is expected (Hey 1991). Given the estimated recombination rate among the ancestral mtIDH-AB haplotypes, significantly reduced levels of recombination were observed between the sequence clusters associated with each mtIDH allozyme which contributes to the large number of fixed differences. As suggested by the coalescent simulations discussed in the following section, these reduced recombination rates may result from selection as well as some degree of geographical isolation (i.e. reduced gene flow).

EVOLUTIONARY SCENARIOS

A high number of fixed differences and deep branches between the ancestral and derived haplotypes clusters suggests an old time since divergence (i.e. balancing selection), whereas a low average number of pairwise differences (*k*) among the derived haplotypes suggests a selective sweep (Figure 18). Accordingly, our coalescent simulations suggest that the observed sequence variation among the mtIDH allozymes is most consistent with a relatively recent selective sweep of haplotypes that evolved in a partially isolated subpopulation.

Analyzing relatively simple evolutionary scenarios, our simulations encompassed parameter combinations that resulted in summary statistics that were comparable to those observed for the *mtIdh* locus. Under the assumption that the allele evolved in a population of similar size as the population experiencing negative selection for the derived mutation (scenario 1), it is most likely that the origin of the derived mutation has been recent and that migration between both subpopulations is low (i.e. partial geographic isolation) (left panel in Figure 15). More precisely, in this scenario only a short selection time explains the low nucleotide diversity among the derived mtIDH-DE haplotypes because a mutation with strong selective advantage increasing rapidly in frequency will have little opportunity to incorporate variants by mutation and to recombine with variants in the surrounding region of the genome (Charlesworth 2006). On the other hand, reduced gene flow or geographical isolation allows for the buildup of nucleotide differences between populations, which may result in a high number of fixed differences when the derived allele evolves. Alternatively, the observed low nucleotide and haplotype diversity among the mtIDH-DE haplotypes might also be explained by a smaller size of the subpopulation experiencing positive selection for the derived allele (scenario 2). Under these conditions higher levels of gene flow between the subpopulations can also result in the observed sequence variation. A longer time since the derived allele arose and was selected is, however, needed to observe a high number of fixed differences.



k_B	High	High	High
k_D	Low	Low	High
$fixed_{BD}$	High	Low	High

Figure 18. Observed and expected coalescent pattern of *mtIdh* sequence variation under a stable allele polymorphism in which selection time ranges from short (i.e. selective sweep) to the total coalescent time (i.e. balancing selection or allopatric divergence). Numbers in the *mtIdh* gene tree (left) indicate posterior probability values. Triangles represent collapsed clades; triangle height represents coalescent time of the clades. The grey shaded area on the trees indicates the selection time. Average number of pairwise differences among ancestral allele sequences (k_B), among derived allele sequences (k_D) and fixed differences between derived and ancestral allele sequences ($fixed_{BD}$) for the *mtIdh* locus and for expected coalescent scenarios is also shown.

In further support of scenario 1, we found differences between Atlantic and Mediterranean coasts in haplotype composition for the mtIDH-AB haplotypes, the *cytIdh* gene and the *enolase* gene. This suggest geographical effects on variation which were also found when studying microsatellite and allozyme frequencies (Desender 2000, Dhuyvetter *et al.* 2004). Next, we found very low haplotype structure among the derived mtIDH-DE haplotypes, which suggests a relatively recent spread of these haplotypes along the Atlantic coasts. Furthermore, the majority of the Atlantic populations carry both alleles, which are often present in almost equal frequencies (heterozygosity within populations ranging between 0-50 %; Van Belleghem and Hendrickx 2013). This indicates current high rates of gene flow, low selection against heterozygotes and ample opportunity for recombination between the haplotypes, which support a recent evolution of the derived allele. Finally, considering the better fit of the simulated Tajima's *D* values to the observed values in the scenario in which the derived allele evolved recently in a geographically structured population, we might give more support

to this scenario compared to the scenario with smaller size of the subpopulation experiencing positive selection for the derived allele.

In contrast, in support of scenario 2, there are indications that disruptive selection maintains differentiation between the differently selected alleles despite high rates of gene exchange between populations. On relatively small distances (100 km), *P. chalceus* populations do not generally show a correlation between geographical and genetic distance in neutral markers (Dhuyvetter *et al.* 2005a). Altogether, this indicates that neutral genes are being exchanged but differentiation (at *mtIdh* locus) is maintained within the locations. Hence, the current distribution of *mtIdh* alleles is clearly influenced by spatially heterogeneous balancing selection (i.e. selection maintains the association between the alleles and the habitat) (Dhuyvetter *et al.* 2004, 2007, Van Belleghem & Hendrickx 2014). More precisely, a gene favorable in one given genomic context might be unfavorable in other genomic contexts (Wright 1931) and reduced recombination rates are likely caused by selection against these negative epistatic effects among closely linked genes. In accordance, selection on multiple alleles may also reduce recombination between the mtIDH-AB and mtIDH-DE haplotypes, because recombination of these adaptive allele combinations will be selected against (Feder *et al.* 2012a). Finally, an inverted chromosomal segment could also explain reduced recombination rates (Kirkpatrick 2010). However, this latter scenario is quite unlikely in the present system as a recombinant haplotype between the mtIDH alleles was observed.

Lack of empirical values of selection strength and migration rates render it difficult to make sound conclusions on the evolutionary scenario. Furthermore, intermediate and more complex evolutionary scenarios may be possible. For instance, the low haplotype structure among the mtIDH-DE haplotypes may also result from a recent bottleneck and subsequent population expansion. This could also result in the observed sequencing variation, despite long term balancing selection.

mtIDH ALLOZYMES AND SELECTION

All non-synonymous nucleotide variants found in this study, both in the mtIDH and cytIDH protein, occur along the enzyme's surface (Appendix 12 and Appendix 13). Watt and Dean (2000) argue that functional constraints shrink with distance from the active site because changes become less disruptive of function. This argumentation suggests that adaptive as well as neutral variation is expected to be found at the protein's surface. Whether the (Lys – Asn) amino acid substitution at amino acid position 447 that differentiates the mtIDH-AB and mtIDH-DE allozyme has a functional effect on the protein functioning is difficult to infer and necessitates functional analysis (Storz & Wheat 2010, Barrett & Hoekstra 2011). Alternatively, the amino acid polymorphism in

this position may have no functional effect, but rather be in close linkage with a selected target. For instance, the *mtIdh* gene is likely transcribed by the same (bidirectional) promoter as the $NADP^+$-transhydrogenase (*Nnt*) gene. Sequencing variation has not been studied in this gene, but the observed pattern in the *mtIdh* gene might be expected to extend into a far larger genomic region then currently studied.

ACKNOWLEDGEMENTS

This study could not have been conducted without access to the extensive collection of the late Konjev Desender. We gratefully thank José Serrano for supplying specimens of *P. meridionalis* and Alexandre Ramos for help in obtaining specimens from Portugal. Viki Vandomme, Elena Dierick and Andy Vierstraete are thanked for their help with sequencing. This work was supported by funding received from the FWO-Flanders (PhD grant to SVB) and the Belgian Science Policy (MO/36/025 to FH) and was partly conducted within the framework of the Interuniversity Attraction Poles program IAP (SPEEDY) – Belgian Science Policy. The simulations were carried out using the STEVIN Supercomputer Infrastructure at Ghent University, funded by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation and the Flemish Government – department EWI.

CHAPTER 3

DE NOVO TRANSCRIPTOME ASSEMBLY AND SNP DISCOVERY IN THE WING POLYMORPHIC SALT MARSH BEETLE *POGONUS CHALCEUS*

Steven M. Van Belleghem [1,2]
Dick Roelofs [3]
Jeroen Van Houdt [4]
Frederik Hendrickx [1,2]

Modified from: *PLoS ONE* 7(8): e42605.

[1] Terrestrial Ecology Unit, Biology Department, Ghent University, K. L. Ledeganckstraat 35, 9000 Gent, Belgium
[2] Department Entomology, Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000 Brussel, Belgium
[3] Department of Ecological Science, VU University Amsterdam, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands
[4] Laboratory of Cytogenetics and Genome Research, Leuven, Belgium

ABSTRACT

The salt marsh beetle *Pogonus chalceus* represents a unique opportunity to understand and study the origin and evolution of dispersal polymorphisms as remarkable inter-population divergence in dispersal related traits (e.g. wing development, body size and metabolism) has been shown to persist in the face of strong homogenizing gene flow. Sequencing and assembling the transcriptome of *P. chalceus* is a first step in developing large scale genetic information that will allow us to further study the recurrent phenotypic evolution in dispersal traits in these natural populations. We used the Illumina HiSeq2000 to sequence 37 Gb of the transcriptome reads and performed *de novo* transcriptome assembly with the Trinity short read assembler. This resulted in 65,766 contigs, clustering into 39,393 unique transcripts (unigenes). A subset of 12,987 show similarity (BLAST) to known proteins in the NCBI database and 7,589 are assigned Gene Ontology (GO). Using homology searches we identified all reported genes involved in wing development, juvenile- and ecdysteroid hormone pathways in *Tribolium castaneum*. About half (56.7%) of the unique assembled genes are shared among three life stages (third-instar larva, pupa, and imago). We identified 38,141 single nucleotide polymorphisms (SNPs) in these unigenes. Of these SNPs, 26,823 (70.3%) were found in a predicted open reading frame (ORF) and 6,998 (18.3%) were nonsynonymous. The assembled transcriptome and SNP data are essential genomic resources for further study of the developmental pathways, genetic mechanisms and metabolic consequences of adaptive divergence in dispersal power in natural populations.

INTRODUCTION

A vast number of insect species are characterized by remarkable and often discontinuous morphological variation in traits related to dispersal capacity (Roff 1986, Roff & Fairbairn 2007). As variation in such traits determines the ability of populations and species to persist in both patchy and changing landscapes (Denno *et al.* 1996, Dhuyvetter *et al.* 2004, Kokko & López-Sepulcre 2006, Hendrickx *et al.* 2009), research on the ultimate and proximate causes of dispersal is a central theme in both evolutionary ecology and conservation biology (Van Dyck & Matthysen 1999, Ronce 2007). Theoretical and empirical research on the ultimate cause of dispersal demonstrated that such dispersal polymorphisms are the result of disruptive selection in heterogeneous landscapes in response to habitat persistence (den Boer 1968, Roff 1994a, Denno *et al.* 1996) and fitness

homogenization under spatiotemporal population fluctuations (McPeek & Holt 1992, Holt & McPeek 1996, Doebeli & Ruxton 1997, Mathias *et al.* 2001, Hendrickx *et al.* 2013). Still, only little is known about the molecular basis of this profound phenotypic variation. For instance, it is unclear whether (i) divergence in dispersal traits is caused by a small set of genes that exert large effects or by many genes with moderate to small effect, and in which order they are involved in adaptive differentiation (Orr 2005, Hoekstra *et al.* 2006, Michel *et al.* 2010), (ii) whether adaptations and the evolution of distinct dispersal phenotypes are mainly the result of mutations in coding regions of the genome or rather due to differences in gene expression (i.e. regulatory changes) (West-Eberhard 2005, Steiner *et al.* 2007, Hoekstra & Coyne 2007), (iii) if the recurrent appearance of this trait is caused by independent mutations or rather by introgression of standing genetic variation (Arendt & Reznick 2008, Barrett & Schluter 2008) or the release of cryptic genetic variation by changes in epistatic interactions (Gibson & Dworkin 2004, Le Rouzic & Carlborg 2008), and (iv) how disruptive selection in dispersal traits affects metabolic pathways resulting in genetically correlated changes in other life history traits (Stevens *et al.* 2012). Such information is particularly crucial to link the proximate and ultimate mechanisms underlying the recurrent intra- and interspecific evolution of dispersal phenotypes.

The endangered halobiontic ground beetle *Pogonus chalceus* (Marsham, 1802) is a most suitable system to study the molecular mechanisms behind adaptive divergence in dispersal traits. The species exhibits a clear wing polymorphism with both short-winged individuals (brachypterous), long-winged individuals (macropterous), as well as intermediate forms (Desender 1985). These differences in dispersal power have been shown to be related to differences in habitat stability and persistence, with long-winged individuals occurring primarily in unstable and relatively recent salt marsh areas. The determination of wing size in this species is polygenic as crosses between brachy- and macropterous populations result in the production of individuals with intermediate wing sizes (Desender 1989a). Divergent selection on wing size likely results in simultaneous selection in other life history traits, as suggested by a strong correlation among populations between average wing size and frequencies of allozymes of the mitochondrial $NADP^+$-dependent isocitrate dehydrogenase (mtIDH) protein (Dhuyvetter *et al.* 2004, 2007). Moreover, within a salt marsh situated at the Atlantic coast in the Guérande region in France, individuals of *P. chalceus* occur chiefly in two habitat types interlaced at a very small scale, i.e. ponds and canals (Dhuyvetter *et al.* 2007). Salt extraction ponds are mostly occupied by long-winged individuals with larger body size and the mtIDH-B allozyme. The borders of tidal canals that lead sea water to these ponds are occupied by smaller short-winged individuals with the mtIDH-D allozyme.

While signals of strong divergent natural selection are observed between the ecotypes for the mtIDH allozymes, dispersal power and body size, no differentiation could be detected for neutral markers, suggesting high levels of gene flow among both ecotypes (Desender *et al.* 1998, Dhuyvetter *et al.* 2004, 2007). These findings and the incipient stage of divergence make the salt marsh beetle *P. chalceus* attractive for genetic studies of selection, adaptation, and gene flow.

It has been shown that portions of the wing development gene network are largely conserved among holometabolous insect orders (Weatherbee *et al.* 1999, Abouheif & Wray 2002). A number of genes involved in the patterning, growth and differentiation of the wing in *Drosophila* have been identified (Weihe *et al.* 2005) and characterized in *T. castaneum* (Richards *et al.* 2008). Furthermore, genes involved in the juvenile hormone (JH) and ecdysteroid (ECD) pathway have also been shown to be relevant for the study of insect polymorphisms, including wing polymorphisms (Zera & Denno 1997, Emlen & Nijhout 1999, Zera 2004, Ishikawa *et al.* 2012). However, little genomic resources are available to study the genetic architecture of dispersal polymorphisms in natural populations of ground beetles, in which intraspecific dispersal polymorphisms can be found abundantly (den Boer 1970, 1980, Desender 1988). Considering ground beetles (Carabidae), NCBI reports 306 ESTs from a study comparing seven coleopteran species (Theodorides *et al.* 2002) and a mitochondrial genome of a *Calosoma* species (Song *et al.* 2010). Other genomic resources comprise mostly single barcoding gene sequences, such as cytochrome oxidase and ribosomal RNA, used for phylogenetic studies. The only coleopteran species for which the genome has been sequenced is the red flour beetle *Tribolium castaneum* (Richards *et al.* 2008), belonging to the Polyphaga suborder. The evolutionary distance of this suborder to the Adephaga suborder, comprising Carabidae species, is estimated to be more than 200 MY (Hunt *et al.* 2007).
Short read *de novo* transcriptome analysis has proven to be a valuable first step to study genetic characteristics and allowed researchers to obtain sequence information and expression levels of genes involved in developmental and metabolic pathways, insecticide resistance, candidate transcripts for diapauses preparation based on homology with related organisms and to discover single nucleotide polymorphism (SNP) in all kinds of model and non-model organisms (Mittapalli *et al.* 2010, Xue *et al.* 2010, Poelchau *et al.* 2011, Sloan *et al.* 2012).
In this study, we used Illumina short read sequencing for *de novo* transcriptome assembly and analysis of the salt marsh beetle *P. chalceus.* We constructed three libraries covering three life stages, one third-instar larva, one pupa and one adult male beetle. We matched these sequences in a BLAST search to known proteins of the NCBI database and aligned the sequences to the genome of *T. castaneum*. Matches include a number of

genes relevant to the study of wing development and dispersal polymorphism. Furthermore, we screened the transcriptome for both conservative SNPs and SNPs resulting in amino acid changes, which will allow genome wide screening of variation between different ecotypes. The resulting assembled and annotated transcriptome sequences constitute comprehensive genomic resources, available for further studies and may provide a fast approach for identifying genes involved in developmental pathways (i.e. wing development, JH, and ECD) relevant to adaptive divergence in this species.

MATERIAL & METHODS

TISSUE MATERIAL AND NUCLEIC ACID ISOLATION

The geographical distribution of *P. chalceus* extends along the Atlantic coasts from Denmark up to and including the major part of the Mediterranean coasts (Turin 2000). Beetles were captured in the Guérande region, France. No specific permits were required for the described field study. Eggs were obtained from the canal ecotype (short-winged) and raised in a common environment. A larva (third-instar), pupa and imago (male) resulting from the same mother were frozen in liquid nitrogen and subsequently used for sequencing (Figure 19). The sex determination is probably of the XY type (Serrano 1981a). Total RNA was isolated from a complete larva (third-instar), pupa and newly emerged male imago. RNA was extracted using the SV Total RNA isolation System (Promega, Madison, USA) according to manufacturer's instructions and genomic DNA was removed by on-column digest with DNase I. RNA was quantified by measuring the absorbance at 260 nm using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Inc.). The purity of the RNA samples was assessed at an absorbance ratio of $OD_{260/280}$ and $OD_{260/230}$ and the integrity was confirmed on an Agilent 2100 Bioanalyzer (Agilent Technologies, Inc.).

Figure 19. *Pogonus chalceus* third-instar larva (left), pupa (middle) and adult beetle (right).

ILLUMINA PAIRED-END cDNA LIBRARY CONSTRUCTION AND SEQUENCING

The cDNA libraries were constructed for the larva, pupa and imago using the TruSeq™ RNA Sample Preparation Kit (Illumina, Inc.) according to the manufacturer's instructions. Poly-A containing mRNA was purified from 2 μg of total RNA using oligo(dT) magnetic beads and fragmented into 200-500 bp pieces using divalent cations at 94°C for 5 min. The cleaved RNA fragments were copied into first strand cDNA using SuperScript II reverse transcriptase (Life Technologies, Inc.) and random primers. After second strand cDNA synthesis, fragments were end repaired, a-tailed and indexed adapters were ligated. The products were purified and enriched with PCR to create the final cDNA library. The tagged cDNA libraries were pooled in equal ratios and used for 2 x 100 bp paired-end sequencing on a single lane of the Illumina HiSeq2000 (Genomics Core, UZ Leuven, Belgium). After sequencing, the samples were demultiplexed and the indexed adapter sequences were trimmed using the CASAVA v1.8.2 software (Illumina, Inc.).

DE NOVO TRANSCRIPTOME ASSEMBLY

The transcriptome reads were *de novo* assembled using Trinity (release 20111126) (Grabherr *et al.* 2011) on the STEVIN Supercomputer Infrastructure at Ghent University (48 cores, 350 G of memory). The three samples (i.e. larva, pupa, and imago) were assembled and analyzed as a pooled dataset. As the Trinity assembler discards low coverage *k*-mers, no quality trimming of the reads was performed prior to the assembly. Trinity was run on the paired-end sequences with the fixed default *k*-mer size of 25, minimum contig length of 200, paired fragment length of 500, 12 CPUs, and a butterfly HeapSpace of 25G (i.e. allocated memory). Prior to submission of the data to the Transcriptome Shotgun Assembly Sequence Database (TSA), assembled transcripts were blasted to NCBI's UniVec database (Cochrane & Galperin 2010) to identify segments with adapter contamination and trimmed when significant hits were found. This adapter contamination may result from sequencing into the 3' ligated adapter of small fragments (< 100 bp). Human and bacterial sequence contamination was investigated using the web-based version of DeconSeq (Schmieder & Edwards 2011), with a query coverage and sequence identity threshold of 90%.

FUNCTIONAL ANNOTATION

The assembled transcripts were subjected to similarity searches against NCBI's non-redundant (nr) database using the BLASTx algorithm (Altschul *et al.* 1990), with a cut-off E-value of ≤ 10^{-3} and a HSP (high-scoring segment pairs) length cut-off of 33. The publicly available platform independent java implementation of the Blast2GO software (Götz *et al.* 2008) was used for blasting and to retrieve associated gene ontology (GO) terms describing biological processes, molecular functions, and cellular components (Ashburner *et al.* 2000). Top 20 blast hits with a cut-off E-value of ≤ 10^{-6} and similarity cut-off of 55% were considered for GO annotation. Next, to get an idea of the amount of genes of the *T. castaneum* transcriptome that are covered by *P. chalceus* transcripts, assembled transcripts were aligned to the *Tribolium* Official Gene Set (Richards *et al.* 2008, Kim *et al.* 2010) using the PROmer pipeline of the MUMmer 3.0 software (Kurtz *et al.* 2004) with default parameters. The presence of open reading frames (ORFs) was investigated using the ORF-predictor server with an ORF cut-off length of 200 bp (Min *et al.* 2005).

GENES OF INTEREST

To guide our search for wing development genes, we used a previously generated list of *Tribolium castaneum* (Table S13b Richards *et al.* 2008 (Richards *et al.* 2008)). To find *P. chalceus* wing development orthologs, we used *T. castaneum* protein sequences in a local BLAST search (tBLASTn) querying the assembled *P. chalceus* transcriptome sequences. Hits with an E-value less than $1e^{-15}$ were examined. The most significant hit was considered to be the putative *P. chalceus* orthologue of the wing development gene in *T. castaneum*. Subsequently, the *P. chalceus* transcript sequence was used in a reciprocal blast to the NCBI nr database. If the BLAST and reciprocal BLAST matched, we assigned orthology to that sequence. For the *apterous* gene, we extracted sequences of *D. melanogaster, T. castaneum, A. mellifera* and *A.pisum* from GenBank and constructed a neighbor-joining tree of the protein sequences with MEGA 5.0 (Tamura *et al.* 2011), bootstrapped 1000 times. The methodology used is similar to that of Brisson *et al.* 2010 (Brisson *et al.* 2010).

Next, genes involved in the juvenile hormone (JH) (Bellés *et al.* 2005) and ecdysteroid (ECD) (Warren *et al.* 2004) pathway in *T. castaneum* were extracted from the KEGG pathway database (Kanehisa & Goto 2000) and the same procedure for orthologue discovery for wing development genes was followed. The assembled transcriptome was also investigated for the presence of the mitochondrial $NADP^+$-dependent isocitrate dehydrogenase (*mtIdh*) gene, which has been shown to be strongly correlated with

dispersal power in *P. chalceus* (Dhuyvetter *et al.* 2004, 2007). For this, the *T. castaneum* protein sequence of the gene homologues to *mtIdh* (XP_970446) was blasted to the *P. chalceus* transcript.

MAPPING READS TO REFERENCE TRANSCRIPTOME

To align the reads back to the assembled reference transcriptome the Burrows—Wheeler Aligner (BWA) program (Li & Durbin 2010) and the Bowtie aligner (Langmead *et al.* 2009) were used. BWA was used for variant analysis. Reads were mapped for each sample (i.e. larva, pupa, and imago) separately to the assembled transcriptome based on the pooled read data. The BWA default values for mapping were used, except for a maximum number of alignments (sampe -n) of 40. Under these settings, read pairs mapping to multiple equally best positions are placed randomly. Properly paired reads with a mapping quality of at least 25 (-q) were extracted from the resulting BAM file using SAMtools (Li *et al.* 2009) for further analyses. Properly paired is defined as both left and right reads mapped in opposite directions on the same transcript at a distance compatible with the expected mean size of the fragments. The high mapping quality ensures reliable (unique) mapping of the reads, which is important for variant calling.
As reads can map to multiple genes or isoforms and we have no available reference genome, we used the RSEM software (Li & Dewey 2011) to assign reads to genes and isoforms and to count transcript abundances. RSEM requires gap-free alignments and therefore the Bowtie aligner (older version, not Bowtie 2) was used and properly paired reads were extracted. RSEM and Bowtie were used as implemented in the Trinity software package (Grabherr *et al.* 2011). Bowtie mapping parameters were set as follows: a maximum number of 2 mismatches allowed (-v) and a number of valid alignments per read pair (-k) of 40. Setting the –k parameter allows reads to align against up to 40 different locations. The old version of Bowtie does not report mapping quality and, hence, does not enable filtering on this parameter. We compared the three developmental stages for transcript composition. Uniquely expressed genes for each life stage were counted and investigated for Gene Ontology (GO) composition.

VARIANT ANALYSIS

Only reliable properly paired BWA mapped reads were considered for Single Nucleotide Polymorphism (SNP) calling. Indels were not considered because alternative splicing impedes reliable indel discovery. SNPs were called using the SAMtools software package (Li *et al.* 2009). Genotype likelihoods were computed using the SAMtools

utilities and variable positions in the aligned reads compared to the reference were called with the BCFtools utilities (Li 2011). Using the varFilter command, SNPs were called only for positions with a minimal mapping quality (-Q) and coverage (-d) of 25. The maximum read depth (-D) was set at 200. The reference is based on all three samples combined. Therefore, to compare the variational composition of the samples, we extracted only heterozygous SNP positions (i.e. Max-likelihood estimate of the site allele frequency ≈ 0.5) from each sample for the unigenes. Unique and shared SNPs were extracted with the VCFtools software (Danecek *et al.* 2011). SNPs located in an open reading frame (ORF) ≥ 200 bp were extracted. A custom perl script was used to test whether these SNPs resulted in an amino acid change in the predicted ORF.

RESULTS & DISCUSSION

SEQUENCING, TRANSCRIPTOME ASSEMBLY AND VALIDATION

Three developmental stages (one third-instar larva, pupa and male adult beetle) were barcode tagged and sequenced on one lane of an Illumina HiSeq2000 sequencer. Sequencing of cDNA libraries generated a total of 184,749,261 raw paired end reads with a length of 101 bp, resulting in a total of 37.32 giga bases. The raw sequence reads were of good quality (≥ 20 Phred score). A summary of sequencing, assembly and annotation results for the three samples and the pooled reads dataset is presented in Table 3. For the pupa sample, notably less reads were sequenced. Reads were assembled using the RNAseq *de novo* assembler Trinity (Grabherr *et al.* 2011). The complete read dataset assembled into 65,766 contigs, clustering into 39,393 isoform clusters (i.e. unigenes). We selected the longest transcript as the representative for each cluster. The size of the contigs ranged from 200 (minimum contig length) up to 19,606 bp, with a mean length of 1,046 bp and totaling 68,799,644 bp for all contigs (Figure 20) and a mean length of 869 bp totaling 34,249,556 bp for the unigenes. The top longest (> 16,000 bp) assembled sequences were inspected for correctness. Overall these extremely long transcripts matched long gene sequences present in NCBI's nr database, indicating that these sequences are not the result of chimerical assembly errors due to repeat regions in the genes. The longest transcript (19,606 bp) also matches the *D. melanogaster* dumpy gene, a gigantic extracellular protein required to maintain tension at epidermal cuticle attachment sites (Wilkin *et al.* 2000).

Bacterial and human transcriptome contamination was negligible. Fifty and fifty-seven unigenes were identified by DeconSeq (Schmieder & Edwards 2011) as bacterial and

human contaminant sequences, respectively. However, these sequences were short in length (289 bp (SD = 148) and 251 bp (SD = 60) for bacterial and human contaminants, respectively) and most likely represent conserved protein regions.

All sequencing reads were deposited into the Short Read Archive (SRA) of the National Center for Biotechnology Information (NCBI), and can be accessed under the accession number SRA050429. The assembled transcriptome was submitted to the Transcriptome Shotgun Assembly Sequence Database (TSA) and can be accessed through the GenBank accession numbers JU404687 - JU470452.

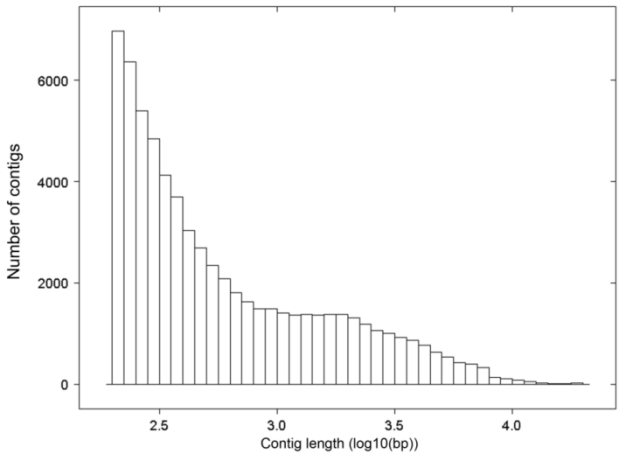


Figure 20. Contig length distribution of Trinity assembly for *Pogonus chalceus*. All assembled contigs were included.

Table 3. *Pogonus chalceus* transcriptome sequencing, assembly and annotation summary.

Stage		**Larva**	**Pupa**	**Imago**	**ALL**
Sequencing	Sequencing reads (101 bp paired end)	66,595,267	48,251,298	69,902,696	184,749,261
	Bases (Gb)	13.45	9.75	14.12	37.32
Assembly	Trinity assembly (Transcripts)				65,766
	Unigenes (Isoform clusters)				39,393
	N50 length (bp) (Unigenes)*				1,904
	Max length (bp) (Transcripts)				19,606
	Max length (bp) (Unigenes)				19,606
	Mean length (bp) (Transcripts)				1,044
	Mean length (bp) (Unigenes)				868
	Median length (bp) (Transcripts)				422
	Median length (bp) (Unigenes)				365
Annotation	Transcripts with BLAST results				29,358
	Unigenes with BLAST results				12,987
	Transcripts annotated with GO terms				17,756
	Unigenes annotated with GO terms				7,589
Mapping (BWA)**	Read mappings (properly paired)	83,539,754	53,814,547	85,597,567	
	Properly paired reads (%)	92.6	90.4	93.1	
	Mean coverage (properly paired)	93.7	55.2	111.6	
	Median coverage (properly paired)	0.93	0.91	2.27	
Mapping (Bowtie)**	Read mappings	143,056,584	97,896,830	156,747,118	
	Properly paired reads (%)	86.8	87.2	87.7	
	Mean coverage (properly paired)	132.98	78.54	150.71	
	Median coverage (properly paired)	1.95	2.21	4.67	

*Contig length for which half of all bases in the assembled sequences are in a sequence equal or longer than this contig length

**Reads of each sample were mapped to the assembled transcriptome of the pooled data (ALL)

FUNCTIONAL ANNOTATION

From the assembled unigenes, 12,987 (33.0 %) showed significant similarity (E value < $1e^{-3}$) to proteins in NCBI's non-redundant (nr) database, with an average best-hit amino acid identity of 70.5% (SD=14.2). As expected, the majority of the sequences had top hits to *T. castaneum* proteins (54.5 %) (Figure 21), the only Coleoptera species for which a complete genome is available. Other insects resembling *P. chalceus* sequences are divided across different insect orders, the most relevant being Hymenoptera (*Nasonia vitripennis* (2.85%), *Camponotus floridanus* (2.41%), *Apis mellifera* (2.15%), *Harpagnathos saltator* (1.86%)), Lepidoptera (*Danaus plexippus* (2.48%)), Hemiptera (*Acyrthosiphon pisum* (2.24%)), and Diptera (*Aedes aegipty* (1.88%)). The only non-Arthropoda species with top blast hits worth mentioning is *Hydra magnipapillata* (0.53%). In total 7,589 (19.3 %) *P. chalceus* unigenes were assigned Gene Ontology (GO) terms based on BLAST matches to sequences with known function. The functional classification based on biological process, molecular function and cellular component is depicted in Figure 22. Among the biological process terms, a significant percentage of genes were assigned to cellular (22.1%) and metabolic (18.0%) processes. Molecular functions were for a high percentage assigned to binding (44.8%) and catalytic activity (36.4%), whereas many genes were assigned to cell part (48.2%) and organelle (27.5%) for the functional class cellular component. These observations are in accordance with observations of metabolic processes in other transcriptomic studies on insects (Mittapalli *et al.* 2010, Wang *et al.* 2010, Xue *et al.* 2010, Bai *et al.* 2011, Shen *et al.* 2011). Redundancy is expected in the assembled transcriptome due to the stochastic process of sequencing and the heuristic nature of the assembly process, which can result in the fragmented assembly of genes. To assess how many actual unique genes we have found in our data, we aligned the obtained unigenes to the 16,645 official genes reported for *T. castaneum.* Of these *Tribolium* genes, 6,883 were covered by *P. chalceus* transcripts based on the PROmer alignments (Kurtz *et al.* 2004), with a mean percent similarity of 76.2% (SD = 10.4). Next, mining the alignments shows that 764 of these *Tribolium* gene hits have more than one hit by unique *P. chalceus* transcripts (comprising 1,837 unigenes). For the transcripts with a PROmer alignment to a *Tribolium* gene this corresponds to a maximal redundancy of 15.6% ((1,837-764)/6,883). However, further investigating these multiple hits showed that most comprise genes that belong to the same gene family (i.e. paralogs). Only 272 *Tribolium* genes are matched by multiple non-overlapping *P. chalceus* contigs (comprising 649 unigenes) and align to different portions of the same gene. This reduces the redundancy to 5.5% ((649-272)/6,883). Hence, the contig sets that are different portions of the same gene do inflate the gene counts for *P. chalceus* to only a minor extent.

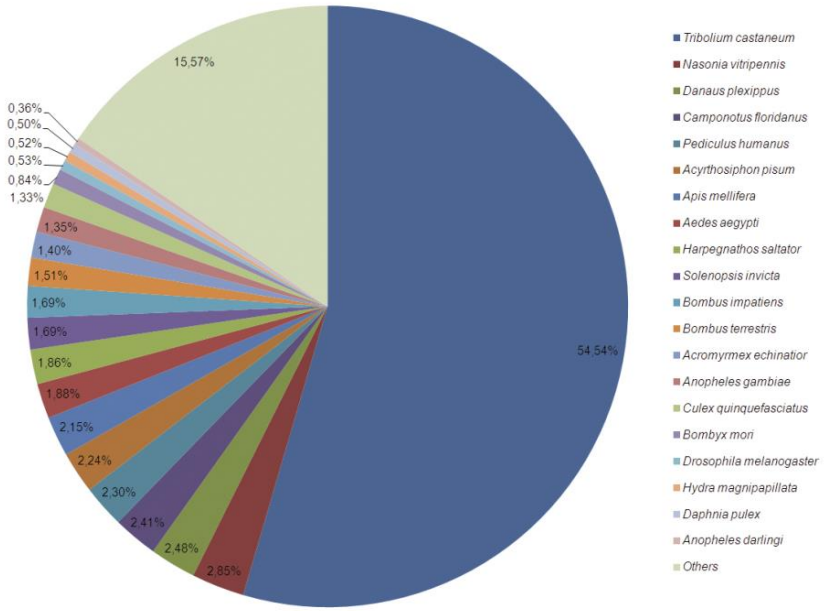


Figure 21. Species distribution of top BLASTx results. The pie chart shows the species distribution of unigenes top BLASTx results against the nr protein database with a cutoff E value < $1e^{-3}$.

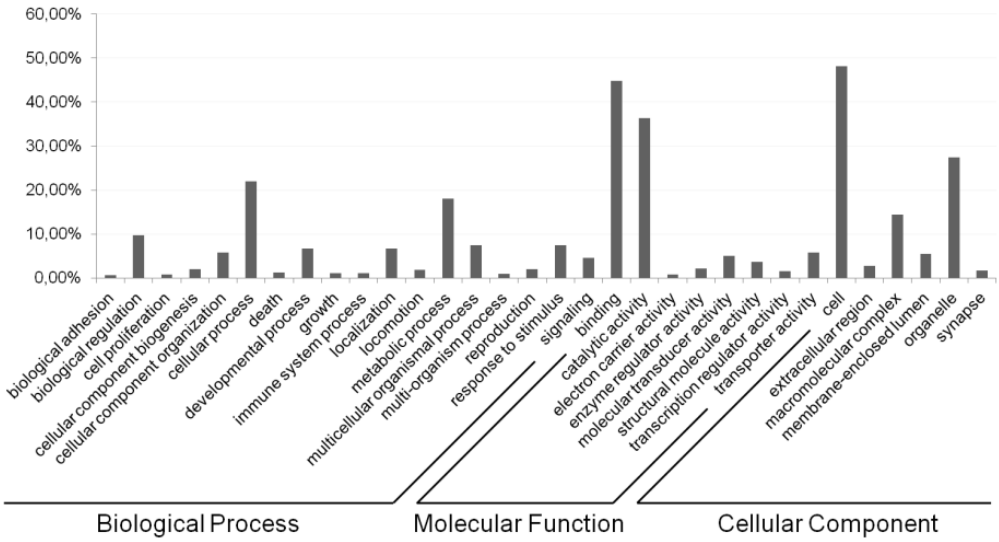


Figure 22. Gene Ontology (GO) categories of the unigenes. Distribution of the GO categories assigned to the *Pogonus chalceus* transcriptome. Unique transcripts (unigenes) were annotated in three categories: cellular components, molecular functions, biological process.

We calculated the "ortholog hit ratio" as described in O'Neil *et al.* 2010 (O'Neil *et al.* 2010) by dividing the length of the putative coding region of a unigene by the length of the ortholog found for that unigene. For this, each unigene and its best BLASTx hit were considered orthologs and the hit region in the unigene is considered to be a conservative estimator of the "putative coding region". In this way, the ortholog hit ratio gives an estimate on the amount of a transcript that is represented by each unigene. Ratios greater than 1.0 can indicate insertions in unigenes. Figure 23A shows that the completeness of the assembled transcripts decreases for very long genes. However, for genes with a length < 12,000 bp this relationship disappears, which shows that the sequencing design and Trinity assembler succeed well in assembling both short and long transcripts. The distribution of ortholog hit ratios is represented in Figure 23B. Overall, unigenes with BLASTx results have high ratios, indicating high completeness of these transcripts. Of the 12,987 transcripts with BLASTx results, 4,567 genes have a ratio ≥ 0.9 and 8,300 have a ratio ≥ 0.5.

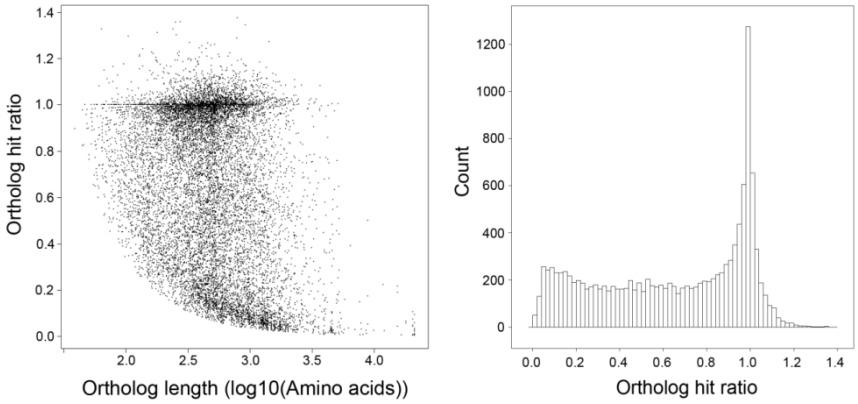


Figure 23. Relationship between ortholog hit ratio and ortholog length (left) and distribution of ortholog hit ratios (right). Ortholog hit ratios were calculated for contigs with BLASTx results. A ratio of 1.0 indicates the gene is likely fully assembled.

A high percentage of unigenes (31,804; 80.7%) could not be assigned a GO term. Examining the length and coverage distribution of these annotated and unannotated unique transcripts shows that most reads (68.8%) are, however, mapped to annotated transcripts. Furthermore, a major portion of the unannotated transcripts consist of assembled transcripts with very low coverage values and short length (Figure 24). For instance, 23,497 (59.6% of all unigenes) of these unannotated transcripts have a length shorter than 500 bp and only 3.1% of all reads map to these transcripts. These short low

coverage transcripts may represent chimeric sequences resulting from assembly errors, fragmented transcripts corresponding to lowly expressed genes, as well as untranslated regions. The remaining 8,427 unannotated sequences are more likely to represent true gene sequences, which may represent novel genes or less conserved genes for which no annotation is found. 15,765 (40.0%) of the unigenes had an ORF (open reading frame) ≥200 bp, with an average length of 1,040 bp and a median length of 659 bp. 7,203 (45.7%) of these unique sequences with ORFs were assigned GO annotations. The remaining sequences with an ORF ≥ 200 bp that lack annotation results might represent true gene sequences. From the *daphnia* genome sequence it was discovered that significant genomic regions without assigned open reading frames are actively transcribed (Colbourne *et al.* 2011). The functional significance of these regions remains to be elucidated, but such transcripts may also be present in the *Pogonus* transcriptome, which cannot be functionally analyzed. Furthermore, high numbers of unannotated contigs are frequently found in other transcriptome sequencing projects (Wang *et al.* 2010, Bai *et al.* 2011, Karatolos *et al.* 2011, Shen *et al.* 2011) and may give some indication of the limitation of inferring the relevant functions of transcripts assembled from sequence data from species with very limited genomic resources or with long evolutionary distances to model species. On the other hand, Trinity succeeds in assembling a reasonable set of annotated genes despite low coverage values (Figure 24).

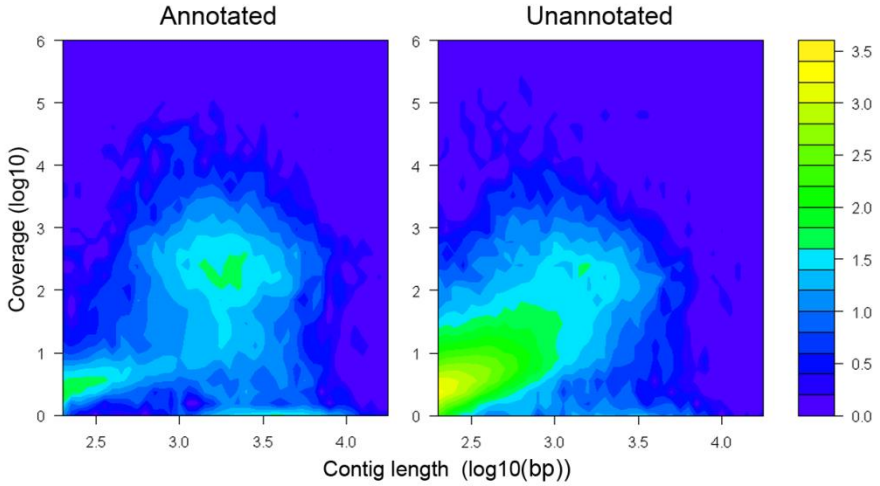


Figure 24. Contour plot of length and coverage distribution of annotated (left) and unannotated (right) unigenes. Transcripts were annotated using Blast2GO. Reads were mapped using BWA. For the annotated transcripts, mean length and coverage was 2,139 and 932, respectively. For the unannotated transcripts, mean length and coverage was 567 and 224, respectively. The color bar shows the log10 transformed count values.

GENES OF INTEREST

As we are interested in the adaptive divergence of wing length in populations of *P. chalceus,* we began our investigation by searching the assembled transcriptome for orthologous genes known to be involved in wing development in the fruit fly *Drosophila melanogaster*. In particular, we used a previously generated list of the wing development genes reported in the genome of the red flour beetle *Tribolium castaneum* (Table S13b of Richards *et al.* 2008 (Richards *et al.* 2008)), which was based on *Drosophila* wing development studies. We found orthologous genes for every wing development gene that we looked for in the assembled *P. chalceus* transcriptome with high confidence (Table 4). *Engrailed* (*en*) and *invected* (*inv*) blasted to the same *P. chalceus* transcript and reciprocal blast of this component returned engrailed. This is not surprising considering their similarity in sequences and function (Gustavson *et al.* 1996). Retrieving orthologous genes for the *apterous* (*ap*) gene was problematic as this gene exhibits a duplication in *T. castaneum* and *Acyrthosiphon pisum* (Brisson *et al.* 2010, Shigenobu *et al.* 2010). Therefore, we aligned the amino acid sequences of *apterous* genes from *D. melanogaster* (NP_724428), *T. castaneum* (*apA*: NP_001139341, *apB*: ACN43342), *Apis mellifera* (XP_392622) and *A. pisum* (*apA*: XP_001946004, *apB*: XP_001949543) with those retrieved from BLAST hits to the *P. chalceus* transcriptome (Figure 25). The *apterous* gene is a hox transcription factor and contains two conserved domains; the homeo domain and the LIM-containing region (Cohen *et al.* 1992). As we did not retrieve the homeo domain for *apB* of *P. chalceus*, we only compared the conserved LIM domain region of the *apterous* genes as reported in (Brisson *et al.* 2010). To root the tree, we added the closely related LIM-containing gene *tailup* (*tup*) of *A. pisum* (XP_001944557) and *T. castaneum* (XP_001815525). The phylogenetic inference indicates that *P. chalceus* exhibits both *apterous* paralogs that are present in *T. castaneum* and *A. pisum* genome, which were lost in the holometabolous insects *Drosophila* and *Apis*. The relationships are similar as the ones reported by (Brisson *et al.* 2010).

Subsequently, we performed similar similarity analyses for genes involved in the Juvenile hormone and ecdysteroid pathway. We found orthologous candidates with high certainty for each gene reported in the KEGG insect hormone biosynthesis pathway (Table 5). The length of the ORF of the *P. chalceus* match, compared to the ORF length in *T. castaneum* is also reported.

Finally, we identified the full coding sequence of the mitochondrial $NADP^+$-dependent isocitrate dehydrogenase (*mtIdh*) gene (Pc_comp1560_c0_seq1) based on homology to the *T. castaneum* protein sequence (EFA04299; E-value =0, bit score =760). The blast result also identified the cytoplasmic $NADP^+$-dependent isocitrate dehydrogenase (*cytIdh*) gene (Pc_comp296_c0_seq1), but with less support (E-value = e^{-172}, bit score = 602).

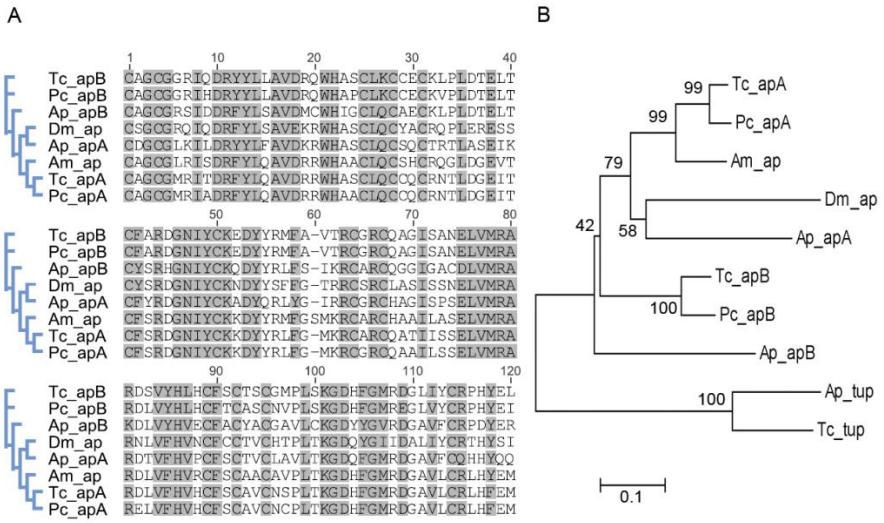


Figure 25. Phylogenetic analysis of the LIM domain of the *apterous* gene. (A.) Alignment of protein sequences of the LIM domain region of the *apterous* (*ap*) orthologs and paralogs of *Tribolium castaneum* (Tc), *Acirthosyphon pisum* (Ap), *Drosophila melanogaster* (Dm), *Apis mellifera* (Am) with the presumed paralogs found in the *Pogonus chalceus* (Pc_apA and Pc_apB) transcriptome. (B.) Neigbour-joining tree of *ap* protein sequences, rooted with *tailup* (*tup*). Bootstrap support values are given at each node.

Table 4. List of wing development genes found in *Pogonus chalceus* orthologous to *Tribolium castaneum*. %AAI = Amino Acid Identity (%). OHR= Orthologous Hit Ratio.

Function	**Gene**		**Accession *P. chalceus***	**%AAI**	**OHR**
Anterior/	*Engrailed*	(en)	Pc_comp5821_c0_seq1	62	1.27
Posterior	*Invected*	(inv)	Pc_comp5821_c0_seq1	56	1.31
	Hedgehog	(hh)	Pc_comp8905_c0_seq1	76	0.96
	Cubitus interruptus	(ci)	Pc_comp4719_c0_seq1	60	1.12
	Patched	(ptc)	Pc_comp7372_c1_seq1	78	0.62
	Decapentaplegic	(dpp)	Pc_comp8429_c0_seq2	64	0.85
	Daughters against	(dad)	Pc_comp5722_c0_seq1	63	1.08
	Brinker	(brk)	Pc_comp8966_c0_seq1	78	0.29
	Optomotor-blind-like	(omb)	Pc_comp6103_c0_seq1	77	0.68
	Spalt-like protein	(sal)	Pc_comp7794_c0_seq1	73	0.87
Dorsal/	*Apterous a*	(ap A)	Pc_comp9155_c1_seq1	77	0.76
Ventral	*Apterous b*	(ap B)	Pc_comp10531_c0_seq1	89	0.69
	Notch	(N)	Pc_comp3149_c0_seq1	81	1.02
	Serrate	(Ser)	Pc_comp6451_c0_seq1	80	1.00
	Wingless	(wg)	Pc_comp9580_c0_seq1	96	0.74
	Distal-less	(Dll)	Pc_comp7089_c0_seq1	77	1.08
Vein	*Serum response factor*	(srf)	Pc_comp3744_c0_seq2	96	0.36
and sensory	*Rhomboid*	(rho)	Pc_comp9713_c0_seq1	96	0.72
	Knirps	(kni)	Pc_comp8029_c0_seq2	74	0.83
	Knot transcription factor	(knot)	Pc_comp14479_c0_seq1	84	0.61
	Iiroquois	(iro)	Pc_comp4855_c0_seq2	74	1.04
	Abrupt	(ab)	Pc_comp3738_c0_seq3	85	1.00
	Noradrenaline transporter	(net)	Pc_comp9252_c0_seq1	85	0.94
	Delta	(Dl)	Pc_comp8811_c0_seq1	70	0.95
	Extramacrochaetae	(emc)	Pc_comp778_c0_seq1	86	1.04
	Achaete-scute	(ASH)	Pc_comp5966_c0_seq1	67	1.09
	Asense	(ase)	Pc_comp12489_c0_seq1	54	1.07
Bodywall/	*Teashirt*	(tsh)	Pc_comp7294_c0_seq1	69	1.13
wing	*Homothorax*	(hth)	Pc_comp2739_c0_seq1	87	1.04
	Nubbin	(nub)	Pc_comp7766_c0_seq1	93	0.36
	Ventral vein lacking	(vvl)	Pc_comp4049_c0_seq1	91	1.05
	Vestigial	(vg)	Pc_comp7899_c0_seq1	69	0.74
Hox	*Sex combs reduced Scr*	(Cx)	Pc_comp5657_c0_seq1	73	1.07
	Prothoraxless	(ptl)	Pc_comp8727_c0_seq1	100	0.31
	Ultrabithorax	(Ubx)	Pc_comp6090_c0_seq1	84	0.97

Table 5. List of insect hormone biosynthesis genes.

Function	**Gene**		**NCBI geneID *T. castaneum***	**Accession *P. chalceus***	**Amino acid identity (%)**	**Ortholog hit ratio**
Juvenile hormone	*juvenile-hormone esterase*	(JHE)	658208	Pc_comp7235_c0_seq1	62	0.97
	juvenile hormone acid methyltransferase	(JHAMT)	662961	Pc_comp8820_c0_seq1	65	1.01
	juvenile hormone epoxide hydrolase	(JHEH)	659305	Pc_comp841_c0_seq1	74	0.98
	cytochrome P450, family 15	(CYP15A1)	658858	Pc_comp2578_c2_seq2	77	0.95
Molting hormone (ecdysone)	*ecdysteroid 25-hydroxylase*	(PHM)	656884	Pc_comp6141_c0_seq1	72	0.98
	ecdysteroid 22-hydroxylase	(DIB)	663098	Pc_comp7215_c0_seq2	73	0.70
	ecdysteroid 2-hydroxylase	(SAD)	658665	Pc_comp5946_c0_seq1	64	0.75
	ecdysone 20-monooxygenase	(SHD)	661451	Pc_comp8625_c0_seq2	73	0.69
	cytochrome P450, family 307	(Spo/spok)	658081	Pc_comp9046_c0_seq1	79	0.93
	cytochrome P450, family 18	(CYP18A1)	656794	Pc_comp3811_c0_seq1	86	0.52

Note: Genes were extracted from *T. castaneum* through the KEGG pathway database.

MAPPING

Reads for each sample (i.e. larva, pupa, adult) were mapped back to the assembled reference transcriptome based on the pooled data and properly paired reads were extracted (Table 3; Figure 26). Based on the BWA mappings (Li & Durbin 2010), 92.6%, 90.4% and 93.1% of the mapped reads were aligned properly paired when aligning the reads of the larva, pupa and adult sample, respectively, to the assembled reference transcriptome. The mean coverage depth (reads covering each base pair) for the larva, pupa and adult sample is respectively 93.7, 55.2 and 111.6. The Bowtie aligner resulted in a higher mean coverage, owing to reads being mapped to multiple positions. The pupa sample has less mean coverage depth resulting from less sequenced reads.

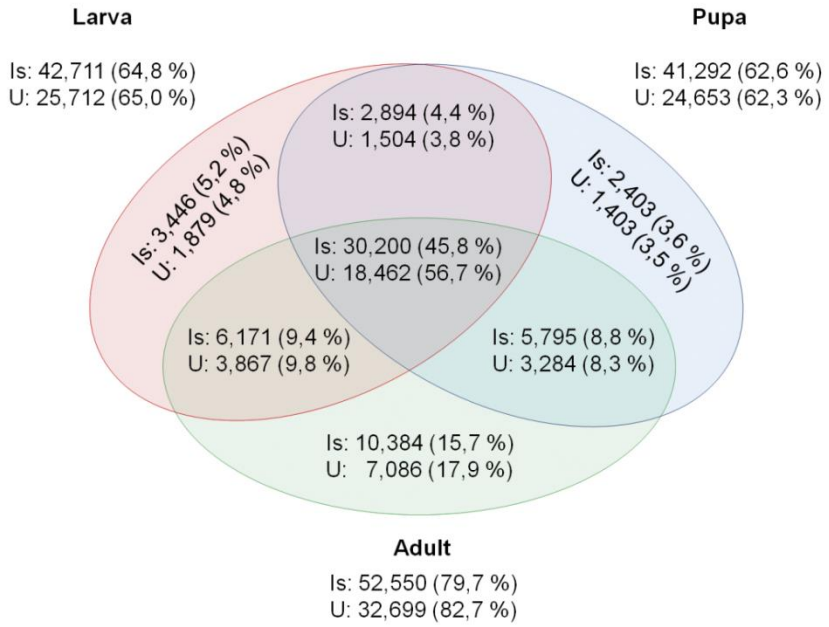


Figure 26. Unique and shared transcript presence of the three developmental stages. The venn diagram shows the unique and shared transcript presence of the three developmental stages (larva, pupa and adult), based on RSEM counts. Reads were assigned to isoforms (Is) or unigenes (U). When RSEM reported a count of at least one, the transcript was reported as present.

Some transcripts were represented by many reads. Moreover, 50% of the reads mapped to only 146 transcript sequences and 90% mapped to 2,971 transcripts. Mapping of the reads shows that read coverage is very high. However, the fact that only 149 transcripts

consume 50% of all reads may indicate that normalization can be useful for transcriptome assembling. The top twenty of these were investigated and are shown in Table 6. Amongst these transcripts, several are associated with energy metabolism (cytochrome c oxidase subunit II and III, succinate and NADH dehydrogenase and ADP/ATP translocase), locomotion (actin and myosin light chain), transcription (DNA topoisomerase 1) and translation (elongation factor 1 and 2). Ferritin is a protein that stores and buffers iron (Theil 1987) and its high abundance may resemble an accommodation to high reduced iron concentrations and high oxidative stress in salt marshes (Odum 1988, Orino *et al.* 2001) or a stress response.

Table 6. Top twenty transcripts with most reads assigned.

Accession *P. chalceus*	**Nr. reads**	**Length (bp)**	**Annotation**
Pc_comp0_c1_seq1	21905861	1,272	Unknown
Pc_comp5_c0_seq1	4116337	5,118	Succinate dehydrogenase*
Pc_comp18_c0_seq1	3016196	3,942	Melanization -related protein
Pc_comp23_c1_seq1	2836940	3,453	Unknown
Pc_comp7_c0_seq1	2585095	1,672	Myosin light chain 2**
Pc_comp32_c0_seq1	1912972	3,409	NADH dehydrogenase subunit 4*
Pc_comp4_c3_seq1	1842608	651	Unknown
Pc_comp30_c0_seq1	1823110	8,598	Alpha-tubulin
Pc_comp41_c0_seq1	1788846	1,961	Elongation factor 1-alpha***
Pc_comp1_c0_seq3	1511917	1,714	Actin**
Pc_comp39_c0_seq1	1501260	2,011	Unknown
Pc_comp14_c0_seq1	1505364	6,711	DNA topoisomerase 1***
Pc_comp16_c0_seq1	1501260	2,186	Muscular protein 20
Pc_comp58_c0_seq1	1419825	1,732	ADP/ATP translocase*
Pc_comp13_c0_seq1	1346169	759	Unknown
Pc_comp10_c4_seq1	1217481	1,679	Cytochrome c Oxidase subunit III (coxIII)*
Pc_comp26_c0_seq1	1178489	3,236	Elongation factor 2***
Pc_comp2_c0_seq1	1128159	634	Unknown
Pc_comp19_c1_seq1	1124751	821	Cytochrome c Oxidase subunit II (coxII)*
Pc_comp60_c0_seq1	1114040	2,504	Ferritin subunit

*Associated with mitochondria, energy metabolism and electron transport chain
**Associated with muscles and movement
***Associated with translation or transcription

COMPARISON OF THE SAMPLES

Reads were mapped with Bowtie (Langmead *et al.* 2009) and assigned to genes and isoforms with the RSEM software (Li & Dewey 2011). Shared and unique presence of genes and isoforms is shown in Figure 26. 30,200 (45.8%) and 18,462 (56.7%) of the isoforms and unigenes respectively were shared among life stages. 1,879 (4.8%), 1,403 (3.5%) and 7,086 (17.9%) of the unigenes are uniquely expressed in the larva, pupa and adult stage, respectively. Of these uniquely expressed unigenes, only 170, 106, and 243 respectively were assigned GO terms (Figure 27). Overall, the GO term composition of these uniquely expressed transcripts in each life stage corresponds well to the GO term composition of the complete transcriptome. No statistical differences in GO term composition were found between these sets of uniquely expressed genes. The higher amount of uniquely expressed genes in the adult stage most likely resulted from more short transcripts being assembled.

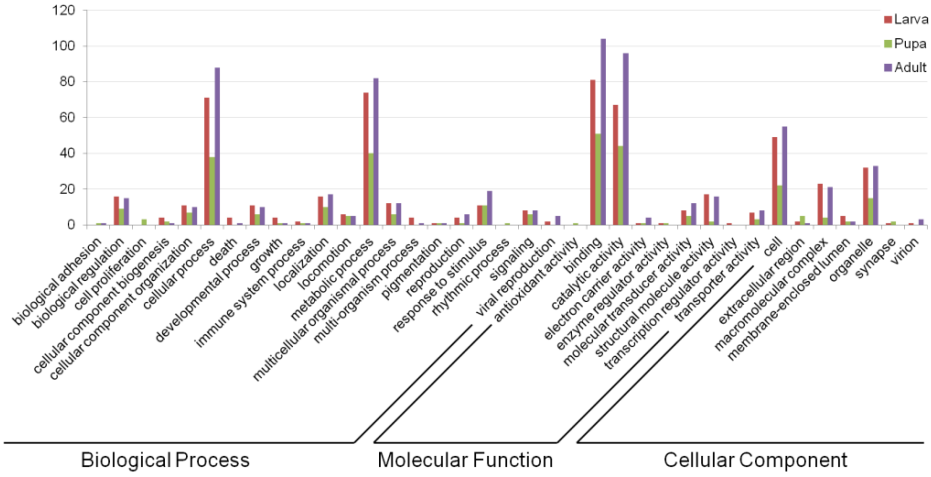


Figure 27. Gene Ontology (GO) distribution assigned to unigenes that are found uniquely in each life stage. Reads were mapped with Bowtie and assigned to genes and isoforms with the RSEM software.

VARIANT CALLING

For SNP calling, BWA was used to map the reads of each sample to the reference transcriptome. In total, SAMtools (Li *et al.* 2009) detected 38,141 different heterozygous SNP positions in unique transcript sequences using the stringent parameters (i.e. coverage and mapping quality of 25) (Figure 28). This is about one SNP per nine hundred bp of unique transcript sequence (1/898). Of these SNPs, 26,823 (70.3%) were found in a predicted open reading frame (ORF) ≥ 200 bp and 6,998 (18.3%) resulted in a amino acid change (nonsynonymous SNP (nsSNP)) and are found in 2,907 different unigenes. This results in a percentage of nonsynonymous changes in the coding region of 26.1%, which is lower compared to studies reporting up to 57.3% nsSNPs in coding regions in a single individual of Japanese native cattle (Kawahara-Miki *et al.* 2011) and 41 to 47% in human individual resequencing studies (Eck *et al.* 2009, Kim *et al.* 2009), but comparable to ratios found in other studies (Levy *et al.* 2007, Bentley *et al.* 2008).

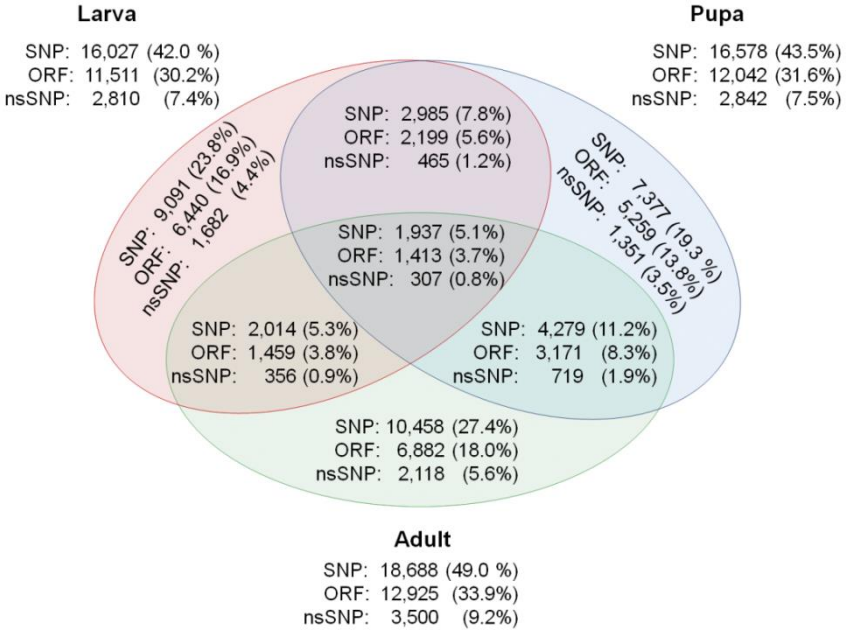


Figure 28. Shared and unique SNPs. Only Heterozygous SNPs are considered from unigenes. The total amount of heterozygous SNPs called in the three samples is 38,141. 70.3% (26,823) of these SNPs were found in an open reading frame (ORF) and 18.3% (6,998) resulted in an amino acid change (nsSNP).

CONCLUSION

In the present study, we sequenced and characterized the transcriptome in the wing polymorphic beetle *P. chalceus*. The assembled sequence data comprising 39,393 unique transcripts provides valuable resources to study wing polymorphism and the adaptive divergence in the face of strong gene flow found in *P. chalceus*. We characterized a large set of genes relevant to wing development and dispersal polymorphism with high significance, including paralogs, giving an indication of the integrity and completeness of the assembled *P. chalceus* transcriptome resulting from short read Illumina sequencing. We found a high number of putative SNPs (37,492). The combination of SNP calling with ORF prediction allowed us to infer that a large part of the SNPs located in a coding fragment (26,757) result in nonsynonymous nucleotide substitutions (23.2%). The results show that it is possible to combine transcriptome assembly and characterization with the discovery of both synonymous and nonsynonymous SNPs, providing a framework for further population genomic studies to identify the molecular basis underlying phenotypic variation of ecologically relevant traits in a non-model species.

ACKNOWLEDGMENTS

We thank Janine Mariën, affiliated to the Vrije Universiteit Amsterdam, for her help in preparation of the samples. This work was carried out using the Stevin Supercomputer Infrastructure at Ghent University, funded by Ghent University, the Hercules Foundation and the Flemish Government – department EWI and we are grateful to the ICT Department of Ghent University for assistance with our computations. Sequencing was performed by the Genomics Core of the University Hospital of Leuven, Belgium.

CHAPTER 4

THE DRAFT GENOME OF THE WING-POLYMORPHIC GROUND BEETLE *POGONUS CHALCEUS*

Steven M. Van Belleghem [1,2]
Frederik Hendrickx [1,2]

[1] Terrestrial Ecology Unit, Biology Department, Ghent University, K. L. Ledeganckstraat 35, 9000 Gent, Belgium
[2] Department Entomology, Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000 Brussel, Belgium

ABSTRACT

Ground beetles (Coleoptera, Carabidae) are a diverse group of beetle species, widely used in ecological and evolutionary studies. However, few genomic resources exist for Carabid beetles. Here, we report the draft genome sequence of the wing-polymorphic ground beetle *Pogonus chalceus*, which is of particular interest for studying the genetics of local adaptation in the face of gene flow. The draft genome was assembled from paired-end and mate-paired Illumina Hiseq2000 reads using SOAPdenovo2. We obtained 312.78 Mb of genome sequence comprising 109,580 unordered scaffolds covering 58.98 % of the estimated genome size (530.28 Mb). Repetitive and mobile elements comprised 18.60 % of the assembled genome. The intron and exon size distribution indicates intron definition as the major splice pathway in *P. chalceus*. Finally, alignment with the genome of *Tribolium castaneum* suggests a high rate of intra as well as interchromosomal rearrangements.

INTRODUCTION

Coleoptera (beetles) are by far the largest order of insects with more than 400,000 described species (Gaston 1991). Despite its species richness, to date, the only coleopteran species for which the genome has been sequenced is the red flour beetle *Tribolium castaneum* (superfamily Tenebrionoidea; Richards *et al.* 2008) and the mountain pine beetle *Dendroctonus ponderosae* (superfamily Curculionoidea (weevils); Keeling *et al.* 2013). Both species belong to the Polyphaga suborder which diverged from the Adephaga suborder around 280 MY ago (Hunt *et al.* 2007). For the Adephaga suborder, with more than 40,000 species and comprising ground beetles (Carabidae), no comprehensive genome sequences have been published. Apart from a few herbivorous species, most ground beetles are considered beneficial organisms as predators of invertebrates. Further, multiple species of ground beetles have lost their capability of flying or show wing dimorphism or polymorphism as seen by reduced wings and/or lacking functional flight muscles (den Boer 1980, Desender 1989b). Therefore, many ground beetles have been a popular object for studying the dynamics, ecology and evolution of dispersal (den Boer 1970, Desender *et al.* 1986, 2000, Aukema 1995, Hendrickx *et al.* 2013).

One carabid species which has gained a lot of attention is the wing-polymorphic ground beetle *Pogonus chalceus*. This ground beetle can be found in salt marshes along the Atlantic coasts from Denmark down to and including the entire Mediterranean region

(Turin 2000). Interestingly, *P. chalceus* shows remarkable local adaptation to habitats with different hydrological dynamics, with short-winged ecotypes in habitats that are frequently inundated, but for short periods, and long-winged ecotypes in habitats that are inundated irregularly for longer periods (Dhuyvetter *et al.* 2004, 2007, Van Belleghem & Hendrickx 2014). In some regions, these distinct habitats occur very close to each other (Dhuyvetter *et al.* 2007), on spatial scales in which geography alone cannot explain a reduction in gene flow to allow divergence. Therefore, *P. chalceus* is of particular interest for studying the genetics of local adaptation in the face of gene flow and the evolution of dispersal related traits.

Pogonus chalceus has a diploid genome with a karyotype of 11 pairs of chromosomes (Serrano 1981b). The sex-determination system is probably of the XY type, although the sex chromosomes have not been identified. The size of the haploid genome is unknown. Previously, the transcriptome of this species has been characterized (Van Belleghem *et al.* 2012). The construction of a draft genome sequence will provide an additional valuable reference for studying the genome wide signal of local adaptation and the genetic architecture of adaptive divergence. In particular when combined with molecular methods that reduce the complexity of the target genome, such as Restriction site Associated DNA (RAD) tag sequencing (Davey *et al.* 2011), the availability of genomic contigs allow (i) to estimate the degree of genetic linkage and, hence, the characterization of genomic islands of divergence and (ii) to characterize the genes associated with genomic sites that experience opposing selection. Here, we report the draft genome of *P. chalceus* and assess its completeness and quality.

MATERIALS & METHODS

ILLUMINA PAIRED-END AND MATE-PAIR LIBRARY CONSTRUCTION AND SEQUENCING

Total DNA was extracted from complete adult individuals using the DNA extraction NucleoSpin® Tissue kit (Macherey-Nagel GmBH). Four 200bp insert libraries, originating from genomic DNA from four males captured in the canal habitat in the Guérande region (France), were sequenced by the Genomics Core of the University Hospital of Leuven (Belgium). Additionally, 500 bp, 800 bp, 2 kb and 5 kb insert libraries, originating from two Guérande canal females, were sequenced by the Beijing Genomics Institute (BGI, China). Shortly, for the short-insert libraries (paired-end; 200 bp, 500 bp and 800 bp), genomic DNA was fragmented randomly, ends were repaired,

A-tailed, and ligated to paired-end adapters (Illumina, San Diego, CA, USA). After electrophoresis, DNA fragments of desired length were gel purified. Long-insert libraries (mate-pair; 2 kb and 5 kb) were constructed by shearing genomic DNA to the appropriate insert size. These fragments were end-repaired with biotinylated nucleotide analogues (Illumina, San Diego, CA, USA), and size-selected fragments (2 kb and 5 kb) were circularized via intramolecular ligation. Circular DNA was fragmented and biotin labels of the fragments (corresponding to the ends of the original DNA ligated together) were affinity purified. Purified fragments were end-repaired and ligated to Illumina paired-end sequencing adapters. Libraries were sequenced on a Hiseq2000 sequencing platform (Illumina, San Diego, CA, USA). Sequencing read length was 101 bp for both ends of the paired-end libraries with 200 bp inserts and 100 bp for the 500 bp and 800 bp insert libraries. The same procedure was used to sequence the mate-pair libraries, with a read length of 49 bp for both ends.

GENOME SIZE ESTIMATION

Genome size was estimated from the *k*-mer coverage estimate and the total number of non-error *k*-mers ($\sum_i d_i$) using all raw reads from the short insert libraries as follows:

$$Genome\ size = \frac{\sum_i d_i}{kmer\ coverage}$$

With d_i the depth (multiplicity) value of the i^{th} unique non-error *k*-mer. Non-error *k*-mers are considered those with a depth value larger than the *k*-mer valley (Figure 29). The *k*-mer coverage was estimated from the *k*-mer coverage peak. However, as we did not find a clear coverage peak in the *k*-mer species curve, *k*-mer coverage was estimated from the *k*-mer individuals curve (Figure 29) (Liu *et al.* 2013). The *k*-mer species curve represents the distribution of unique *k*-mers with a certain depth. The *k*-mer individuals curve is calculated from the product of *k*-mer species number (N_d ; number of unique *k*-mers found with a certain depth) and corresponding depth value (d).

$$kmer\ individual = d \times N_d$$

The *k*-mer individuals curve is a variation of a Poisson distribution, which has the same figure shape, but moves rightwards by one unit compared to the *k*-mer species curve (Liu *et al.* 2013) and allows calculating the *k*-mer coverage peak and valley.
Counting of *k*-mer frequency in the sequencing data was performed using Jellyfish v2.1.3. (Marçais & Kingsford 2011) using all short insert libraries combined. A *k*-mer size

of 17 was chosen so that most *k*-mers are expected to be unique in the genome based on initial genome size expectations. The peak of 17-mer frequency (*M*) in the reads is correlated with the real sequencing depth (*N*), read length (*L*), and *k*-mer length (*K*) and their relations can be expressed as follows (Li *et al.* 2010):

$$M = N \times \frac{(L - K + 1)}{L}$$

(*L – K +1*) gives the number of *k*-mers created per read.

GENOME ASSEMBLY

Adapter contamination in reads was deleted using Cutadapt v1.4 (Martin 2011) and reads that did not have a matching pair after adaptor filtering were removed. Reads were corrected for sequencing error with SOAPec v2.02 (Luo *et al.* 2012), using a *k*-mer size of 17 and a low frequency cutoff of consecutive *k*-mer of 3. Subsequently, reads were assembled using SOAPdenovo2 (Luo *et al.* 2012) using a *k*-mer parameter of 47, which was selected for producing the largest contig and scaffold N50 size after testing a range of *k*-mer settings between 19 and 71. The short insert libraries were used for both contig building and scaffolding. The long insert libraries were only used for scaffolding. The SOAPdenovo GapCloser v1.12 tool (Luo *et al.* 2012) was used with default settings to close gaps emerging during scaffolding. Transcript sequences were used to improve scaffolding using L_RNA_scaffolder (Xue *et al.* 2013). Finally, we used DeconSeq v0.4.3 (Schmieder & Edwards 2011) to identify and remove possible human, bacterial and viral contamination in the assembly. Completeness of the assembled genome was assessed by comparing the assembly with a highly conserved core gene dataset that occur in a wide range of eukaryotes using the CEGMA pipeline v2.5 (Parra *et al.* 2007). This dataset consists of 248 conserved genes representing different protein families from the eukaryotic orthologous groups (KOGs) database (Tatusov *et al.* 2003).

REPETITIVE ELEMENTS

Tandem repeats were predicted using Tandem Repeat Finder (TRF) v4.0.4 (Benson 1999) with recommended parameters (Match = 2; Mismatch = 7; Delta = 7; PM = 80; PI = 10; Minscore = 50; Maxperiod = 500). Next, we used RepeatMasker v4.0.5 and rmblastn v2.2.27 (Smit *et al.* 2014) using the RepBase Update Coleoptera library to identify known sequences representing repetitive DNA (Jurka *et al.* 2005). After these repetitive elements were masked in the assembly, novel repetitive elements were identified with

RepeatScout v1.0.5 (Price *et al.* 2005) using default parameters. Subsequently, these repeats were used as a repeat library in RepeatMasker v4.0.5 (Smit *et al.* 2014) and those occurring at least 10 times in the genome were counted.

ALIGNMENT OF TRANSCRIPTOME TO GENOME

The previously assembled *P. chalceus* transcriptome (Van Belleghem *et al.* 2012) was mapped to the assembled genome using Splign v1.39.8 (Kapustin *et al.* 2008) with default parameters. Only Splign alignments with an identity larger than 97 % (~ error rate) were considered. Average exon and intron sizes were calculated from the Splign results.

SYNTENY WITH THE *TRIBOLIUM CASTANEUM* GENOME

P. chalceus scaffolds longer than 600 Kb were mapped to the *Tribolium castaneum* genome v3.0 (09-06-2014) (Richards *et al.* 2008) using the PROmer pipeline of the MUMmer 3.0 software (Kurtz *et al.* 2004) with a minimum similarity of 80 % and alignment length of 100 bp. Alignments were visualized with Circos v0.66 (Krzywinski *et al.* 2009).

RESULTS & DISCUSSION

GENOME SIZE ESTIMATION

Sequencing resulted in approximately 56 Gb of sequencing data (Appendix 14). From the *k*-mer distribution, the *P. chalceus* genome size was estimated to be 530.28 Mb (Figure 29; Table 7). The *k*-mer species curve did not show a distinct frequency valley or peak. This may result from high heterozygosity in the data as individuals were not inbred and sequences from multiple individuals were pooled to obtain sufficient coverage (e.g. heterozygous *k*-mers will have half the frequency of homozygous *k*-mers). Therefore, the *k*-mer coverage valley and peak were estimated from the *k*-mer individual curve as 3 (= 4-1) and 25 (= 26-1), respectively (Figure 29). The estimated *k*-mer coverage valley corresponds to a 0.027 per base error rate in the raw sequencing reads. Sequencing coverage combining all the short insert size libraries was estimated to be 29.70.

Table 7. *k*-mer counts and estimation of sequencing coverage and genome size calculated from the short insert libraries. Non error *k*-mers indicates the number of *17*-mers with a frequency higher than the *k*-mer coverage valley.

Total *k*-mers (billion)	13.62
Distinct *k*-mers (billion)	0.50
k-mer coverage valley	3
Non error *k*-mers (billion)	13.26
k-mer coverage peak (*M*)	25
Sequencing coverage (*N*)	29.70
Genome size (Mb)	**530.28**

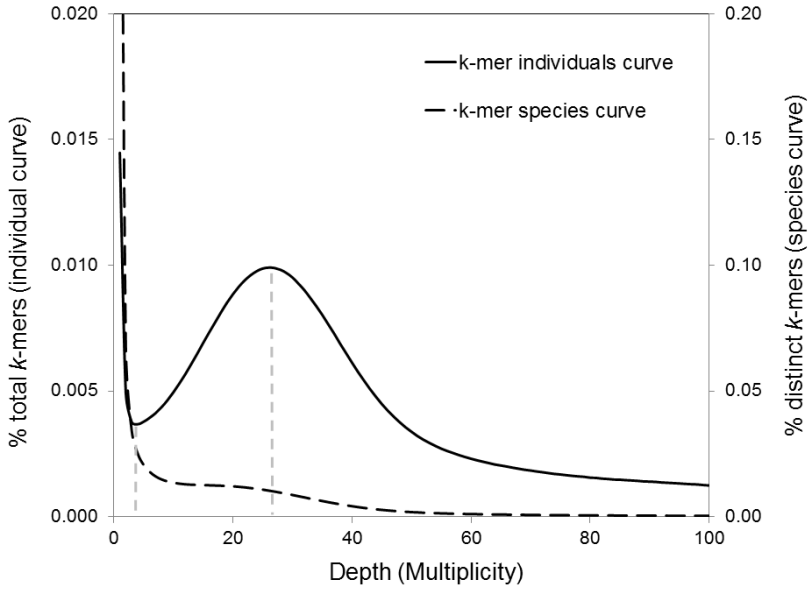


Figure 29. *k*-mer individual (left axis) and species (right axis) curve for *17*-mers. Grey dashed lines indicate the *k*-mer frequency valley (left) and the estimated *k*-mer coverage (right).

GENOME ASSEMBLY

Genomic sequencing reads were assembled into 110,093 scaffolds with a minimum size of 300 bp resulting in a total sequence length of 309 Mb and an N50 size of 55,203 bp (Table 8). 1.34 million scaffolds and contigs smaller than 300 bp with a total length of 207.26 Mb were excluded from the genome assembly used in further analysis. Gaps were closed by remapping the reads to the scaffolds reducing the percentage of Ns in the

scaffolds from 32.0 % to 15.8 %. Next, using transcripts the scaffolds were further linked increasing the N50 size to 72,845 bp and resulting in a total size of the assembly of about 313.08 Mb. Screening for bacterial, viral and human contamination removed 137 contigs covering 297,781 bp of the genome assembly. Most contamination was caused by endosymbiontic *Rickettsia sp.* (77.37 %), followed by *Wolbachia sp.* (10.22 %) (Appendix 15). Both these bacteria species are common endosymbionts in arthropod species (Duron *et al.* 2008). Viral contamination was about 4.38 % (*Staphylococcus* phage).

Mapping core eukaryotic genes (CEGs) to the assembled *P. chalceus* genome, using the CEGMA pipeline (Parra *et al.* 2007), identified 230 full-length and 9 partial CEGs out of the conserved set of 248 CEGs. The assembled genome size covers about 59 % of the estimated genome size when excluding contigs smaller than 300 bp. As about 93 % of the core genes are identified in full-length, the estimated missing 41 % of the genome likely comprises difficult to assemble repetitive regions.

Table 8. *Pogonus chalceus* genome assembly statistics. Statistics were calculated for scaffolds larger than 300 bp.

	Scaffolds	**+ Gaps closed**	**+ RNA scaffold**	**+ DeconSeq**
Assembled total size (bp)	309,290,483	309,346,765	313,080,367	312,782,586
Number	110,093	110,093	109,717	109,580
Largest (bp)	1,116,948	1,116,948	1,116,948	1,116,948
Average size (bp)	2,809	2,810	2,854	2,854
N50 size (bp)	55,203	55,231	72,845	73,053
Number included N50	1,227	1,224	916	914
N80 size (bp)	8,863	8,867	8,849	8,960
Number included N80	5,157	5,150	4,470	4,445
N	99,096,987	48,907,118	51,243,516	51,243,516
% N	32.04	15.81	16.37	16.38
% GC	26.86	26.87	26.43	26.25

REPETITIVE ELEMENTS

Repetitive elements occupied approximately 49.96 Mb or 18.60 % of the assembled genome (Appendix 16). A total of 7.12 Mb of tandem repeats were identified with TRF (Benson 1999), comprising 2.72 % of the assembled genome. Only 1.29 % of the *P. chalceus* genome assembly had similarity to the known Coleoptera repeats in RepBase. The remainder appeared to be unique to *P. chalceus*, with 2,414 novel elements appearing at least 10 times and comprising 14.78 % of the genome assembly. This percentage is in the range of 8 to 42% of repetitive elements found in other Coleoptera and insects (Wang *et al.* 2008, Keeling *et al.* 2013, Kocher *et al.* 2013).

ALIGNMENT OF TRANSCRIPTOME TO GENOME

Next, we assessed the completeness of the *P. chalceus* assembly by comparing it with an independently sequenced and assembled set of transcripts putatively representing 39,393 genes (Van Belleghem et al., 2012). 77.30 % of the putative gene transcripts mapped to the genome assembly (Identity > 97 %) covering a total of 36.72 Mb of the genome (Table 9). The average exon and intron size was estimated to be 332 bp (SD = 395.21) and 1,796 bp (SD = 7,656), respectively. Interestingly, the intron size distribution is strongly skewed towards small intron sizes, whereas the exon size distribution seems less constrained. This results in a median exon and intron size of 225 bp and 69 bp, respectively. This exon and intron size distribution found in *P. chalceus* (Figure 30A) closely resembles distributions found in other insects and lower eukaryotes (Collins & Penny 2006, McGuire *et al.* 2008). This distribution can be explained by the pathway of splice site recognition (Collins & Penny 2006, McGuire *et al.* 2008, Osella & Caselle 2009). Short introns are spliced away preferentially through a pathway called *intron definition* in which the spliceosomes interact with the ends of the intron (i.e. splice sites are recognized across introns). In contrast, the alternative *exon definition* pathway requires an initial interaction between the spliceosome factors, bound at the splice sites, across the exon (i.e. splice sites are recognized across exons). This latter pathway is thought to constrain the exon size distribution in vertebrates (Collins & Penny 2006). The exon size distribution in *P. chalceus* is most likely not constrained due the use of intron definition, whereas intron definition likely explains the high abundance of short introns. In particular, it has been suggested from analysis of the intron definition pathway that the threshold of intron length above which intron-defined splicing ceases almost completely is between 200 and 250 bp (Fox-Walsh *et al.* 2005). This threshold seems to be present in *P. chalceus* (Figure 30A). In contrast, the presence of long introns (up to 200 kb) indicates the activity of the exon definition pathway. Genes with a single coding exon appear in excess in *P. chalceus* (Figure 30B). However, this distribution may be slightly skewed to smaller values due to the fragmentation of the genome and transcripts. The maximum number of exons found in a gene was 120.

Table 9. Alignment results of the *P. chalceus* transcriptome to the assembled genome. Alignment results are shown for the unigenes (excluding splice variants) with identity > 97 % to the assembled genome (Van Belleghem *et al.* 2012).

		SD
Number of unigenes	39,393	
Genome covered by unigenes (Mb)	36.72	
Percentage of unigenes aligned to genome	77.30	
Mean/median exon length (bp)	332/225	395
Mean/median intron length (bp)	1,796/69	7,656
Number of genomic contigs with > 1 unigene mapped	5,546	
Number of genomic contigs with > 2 unigenes mapped	2,374	
Number of genomic contigs with > 10 unigenes mapped	849	

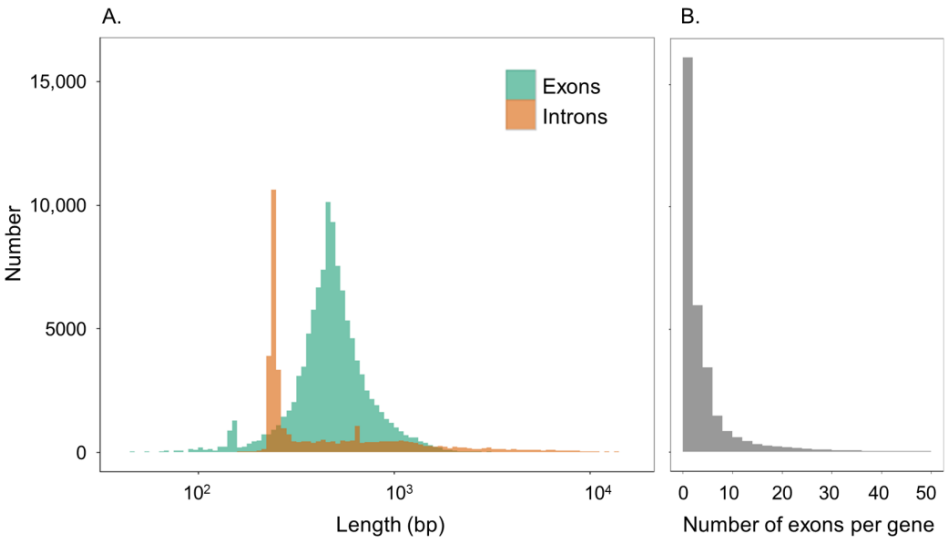


Figure 30. (A.) Exon and intron size distribution. Exon and intron sizes were calculated from Splign alignments of transcripts to the assembled genome. (B.) Distribution of the exon number per gene. The maximum number of exons found in a gene was 120 (not shown in the graph).

SYNTENY WITH THE *TRIBOLIUM CASTANEUM* GENOME

Sixteen *P. chalceus* genomic scaffolds had a length larger than 600 Kb and were aligned to the *Tribolium castaneum* genome (Figure 31). We only considered alignments with a nucleotide similarity higher than 80 %, giving relatively high confidence of homology. In comparison, amino acid similarities of a set of wing development genes identified as

homologous between *P. chalceus* and *T. castaneum* ranged between 60 % and 100 % (Van Belleghem *et al.* 2012).

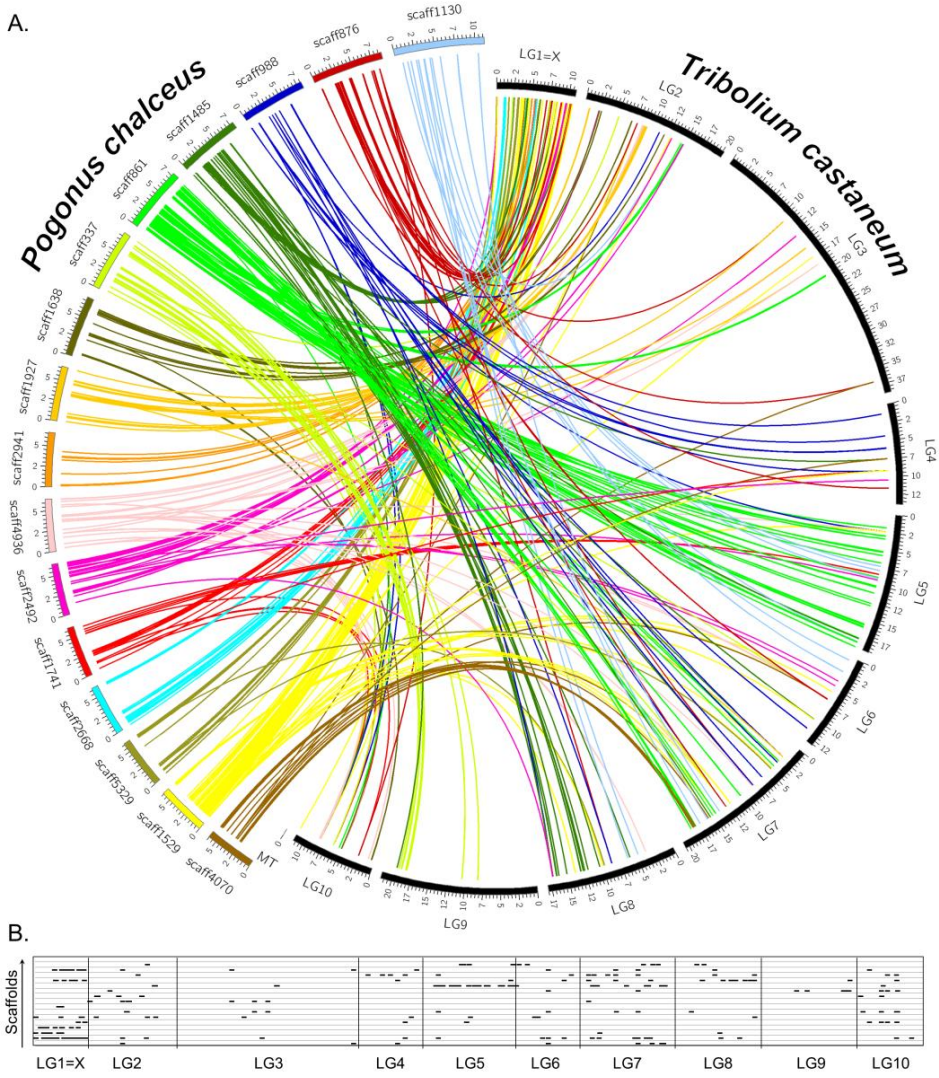


Figure 31. Shared synteny between *Pogonus chalceus* assembly scaffolds and *Tribolium castaneum* linkage groups. *P. chalceus* scaffolds longer than 600 Kb were aligned to the *T. castaneum* linkage groups with PROmer (similarity > 80%). (A.) Circos map depicting alignments and indicating intra - and interchromosomal rearrangements. *P. chalceus* scaffolds are drawn on a scale ten times larger compared to *T. castaneum* (100 Kb and 1 Mb, respectively). (B.) Dashes represent alignments of *P. chalceus* scaffolds to the *T. castaneum* linkage groups. Scaffolds were ordered in increasing length in the same order as in the Circos map. Scaffolds with hits on multiple *T. castaneum* linkage groups indicate interchromosomal rearrangements.

Although *P. chalceus* and *T. castaneum* diverged more than 280 MY, there is still evidence for shared synteny (Figure 31). However, the alignment of the *P. chalceus* scaffold and *T. castaneum* chromosomes also indicates extensive intrachromosomal as well as interchromosomal rearrangements (Figure 31). This is in contrast to genome alignments of the mountain pine beetle *Dendroctonus ponderosae* to the *T. castaneum* linkage groups, in which interchromosomal rearrangements seem less extensive (Keeling *et al.* 2013). Latter species diverged about 200 MY ago and both belong to the Polyphaga. Among the set of 16 largest assembled scaffolds, 8 showed homology to the LG1 *T. castaneum* linkage group (scaffold 1529, 5329, 2668, 1741, 2492, 2941, 1485, and 876), corresponding to the *T. castaneum* sex chromosome.

CONCLUSION

In the present study, we sequenced and assembled a large part of the genome of the ground beetle *P. chalceus*. The assembled sequence comprises 109,580 scaffolds spanning 312.78 Mb and covering approximately 58.98 % of the *P. chalceus* genome. Despite the fragmented assembly, repeat structure, exon-intron size distribution and synteny with *T. castaneum* could be well investigated. Future sequencing of larger insert size libraries and linkage mapping will aid in further ordering of the scaffolds and the development of a complete genome sequence for *P. chalceus*.

ACKNOWLEDGEMENTS

We gratefully thank Xiaobin Guo for his help in processing and sequencing of the samples at Beijing Genomics Institute (BGI, China) and genome assembly advice. Jeroen Van Houdt is thanked for processing and sequencing of the samples at the Genomics Core of the University Hospital of Leuven (Belgium). This work was carried out using the Stevin Supercomputer Infrastructure at Ghent University, funded by Ghent University, the Hercules Foundation and the Flemish Government – department EWI and we are grateful to the ICT Department of Ghent University for assistance with our computations. Funding was received from the FWO-Flanders (PhD grant to Steven Van Belleghem) and the Belgian Science Policy (MO/36/025) and partly conducted within the framework of the Interuniversity Attraction Poles program IAP (SPEEDY)—Belgian Science Policy.

CHAPTER 5

POPULATION GENOMICS OF PARALLEL ADAPTIVE DIVERGENCE IN SYMPATRY IN THE GROUND BEETLE *POGONUS CHALCEUS*

Steven M. Van Belleghem [1,2]
Carl Vangestel [1,2]
Frederik Hendrickx [1,2]

[1] Terrestrial Ecology Unit, Biology Department, Ghent University, K. L. Ledeganckstraat 35, 9000 Gent, Belgium

[2] Department Entomology, Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000 Brussel, Belgium

ABSTRACT

Understanding the genetic architecture and evolutionary history of divergence in the face of gene flow is essential to the study of adaptive divergence and, eventually, speciation processes. When rates of gene flow are high, genetic architectures underlying the oppositely selected traits are expected that reduce recombination, such as close genetic linkage or pleiotropy. Additionally, when replicated instances of adaptive divergence are found, a major question concerns whether adaptations evolved independently or whether genetic variation is rather shared through ancestral polymorphisms or introgression. Here, we use RAD tag sequencing to study population structure at a genome-wide scale of two ecotypes of the wing-polymorphic ground beetle *Pogonus chalceus* and identify loci associated with adaptive divergence in both sympatric and allopatric populations. Comparison of genome wide variation indicates low population divergence between sympatric as well as allopatric populations, suggesting high rates of gene flow and relatively recent separation. However, variation in multiple loci was strongly associated with adaptive divergence. The absence of genetic linkage between these loci indicates widespread genomic divergence even between sympatric populations. All the alleles identified as outlier loci have a singular mutational origin and are shared between repeatedly diverged populations. Moreover, most of these loci have a similar evolutionary history that suggests a recent increase of the alleles associated with the short-winged populations from tidal habitats. This shared evolutionary history suggests a singular evolutionary origin of the short-winged ecotypes in *P. chalceus* and a recent spread along the Atlantic coasts.

INTRODUCTION

Identifying both the genetic architecture of adaptive divergence as well as the evolutionary history of traits involved in adaptation allows understanding how populations adapt in response to the environment. When populations diverge in the face of gene flow, a major aspect concerns the amount of genetic divergence as well as the genetic architecture underlying oppositely selected traits. More precisely, when gene flow is ample, recombination will hamper the independent evolution of adaptive lineages (Felsenstein 1981) and, therefore, strong genetic linkage and/or pleiotropy is often suggested to play an important role in maintaining contrasting ecotypes in hybridizing populations (Via 2001, Griswold 2006, Yeaman & Whitlock 2011). Additionally, the origin of adaptive traits from new mutations or from preexisting

variation has important implications for the adaptation process (Barrett & Schluter 2008) and potentially the speciation process (e.g. Dasmahapatra et al. 2012; Feulner et al. 2013). First, introgression of adaptive gene variants into other populations and reuse of shared standing genetic variation may aid in rapid adaptation in different localities (Barrett & Schluter 2008, Arnold & Martin 2009, Jones *et al.* 2012). Secondly, it has been argued that reuse of standing genetic variation may have an important role in the maintenance of divergent ecotypes during early stages of reproductive isolation (Jones *et al.* 2012). For instance, when populations adapt from standing genetic variation, genetically unlinked traits may be in linkage disequilibrium. Moreover, the preexistence of multiple adaptive loci may allow for a rapid evolution of linkage disequilibrium between performance and assortative mating traits resulting in the evolution of reproductive isolation (Dieckmann & Doebeli 1999, Fry 2003, Nosil *et al.* 2012, Feder *et al.* 2012a).

To address both the origin adaptive alleles and the genetic architecture of differently selected traits, it is essential to identify the genomic regions that are involved in adaptation during sympatric divergence. However, complex interactions of selection, drift, migration, recombination, mutation, and ancestral polymorphism can lead to heterogeneity and noisy patterns in divergence (Noor & Bennett 2009, Martin *et al.* 2013). Therefore, occurrences of repeated adaptation to similar environmental gradients are of particular interest in evolutionary biology as they provide strong evidence for a role of natural selection and help in discriminating underlying evolutionary processes (e.g. Colosimo *et al.* 2005, Jones *et al.* 2012, Soria-Carrasco *et al.* 2014) and different historical sequences of events (Johannesson *et al.* 2010, Butlin *et al.* 2013). More precisely, when rates of gene flow between diverging populations are high, selection is expected to preserve the association of the genetic variation at the selected site (or closely linked to the selected site) with the environmental gradient. On the other hand neutral gene sequences not closely linked to the selected site can be freely exchanged and recombined into other genomic backgrounds (Maynard Smith & Haigh 1974, Hohenlohe *et al.* 2010b). Therefore, studying genome wide patterns of variation in these settings is expected to provide strong support for separating locus-specific effects that affect one or a few loci at a time (e.g. recombination, selection and mutation) from genome-wide demographic effects (e.g. population size increase, genetic bottlenecks, founder events and inbreeding) (Luikart *et al.* 2003, Stinchcombe & Hoekstra 2008).

The wing-polymorphic ground beetle *Pogonus chalceus* represents a situation of replicated adaptation in different spatial settings with different opportunities for gene flow. Populations of *P. chalceus* have repeatedly diverged in short-winged and long-winged populations as a response to different hydrological dynamics (Dhuyvetter *et al.* 2004, 2007, Van Belleghem & Hendrickx 2014). Short-winged ecotypes are found in *tidal*

habitats that are frequently inundated, but for short periods, whereas long-winged ecotypes are found in *seasonal* habitats that are inundated irregularly for longer periods. In some regions, such as the Guérande salterns in France, these distinct habitats occur very close to each other (Dhuyvetter *et al.* 2007), on spatial scales in which geography alone cannot explain a reduction in gene flow to allow divergence. In the Guérande salterns, both habitats are found in multiple replicates only 10-20 m apart. In these salterns, *ponds* are used to evaporate water and concentrate salt and resemble seasonal marshes in that they are flooded irregularly for extensive periods. *Canals*, on the other hand, are used to bring water to the ponds and are subject to the tides.

In this study, we explore the extent of genomic divergence among repeatedly adapted populations at different spatial scales. Establishing the demographic history and evolutionary relations of these populations is essential for reliable identification of loci under divergent selection (Crisci *et al.* 2012). Therefore, by using Restriction Associated DNA markers (RAD tags), representing randomly distributed but consistent genomic regions from multiple individuals, we first reconstruct population structure and demographic history of several diverged *P. chalceus* population pairs representing the nearly entire distribution of this species. Next, we identify markers that are linked to adaptation by using outlier analysis and identify variation that is shared among similarly adapted populations. By reconstructing phylogenies of these adaptive loci, we study their evolutionary history. Finally, by investigating genetic linkage of the adaptive variation, we discuss the genetic architecture of the adaptive differentiation. These results help us understand (i) to what extent the adaptive divergence evolved only once and colonized similar pairs of environments or occurred repeatedly in multiple localities (i.e. CHAPTER 2) and (ii) what the effect is of spatial proximity versus separation on genomic variation in *P. chalceus*.

MATERIALS & METHODS

SAMPLE DESIGN

Ecotypically diverged *P. chalceus* individuals were collected from both tidal and seasonal salt marshes representing nearly the entire species range (Figure 32). We sampled four geographically isolated population pairs (separated between approximately 450 km and 900 km) of a tidal and seasonally flooded inland population each, which were characterized by short and long wings respectively. Distances between tidal and

seasonal populations ranged from 20 m in France (GUO-GUE), 5 km in Portugal (AVE1-AVE2), 37 km in Belgium (DUD-NIE) and 50 km in Spain (HUE-COD). Additionally, allopatric populations were sampled from the tidal and short-winged population from the Severn Estuary population in the UK (SEE) and a seasonally flooded long-winged population from the Camargue (CAM). Twenty-four individuals of each ecotype were sampled in the Atlantic France and Belgian populations (GUO, GUE, NIE and DUD) and 8 individuals in the remaining populations (AVE1, AVE2, HUE, COD, SEE and CAM). The Atlantic France and Belgian populations were sampled and analyzed most extensively in this study. The remaining populations were used as a comparison and to study the genetic differentiation among these populations. More precisely, analyses aimed at quantifying genetic divergence between populations residing in contrasting environments in close allopatry (37-50 km), parapatry (5 km) or sympatry (20 m). We expected strong gene flow at these geographical scales. In addition, we evaluated the effect of strong geographical isolation by quantifying genetic divergence between populations separated by more than 1300 km.

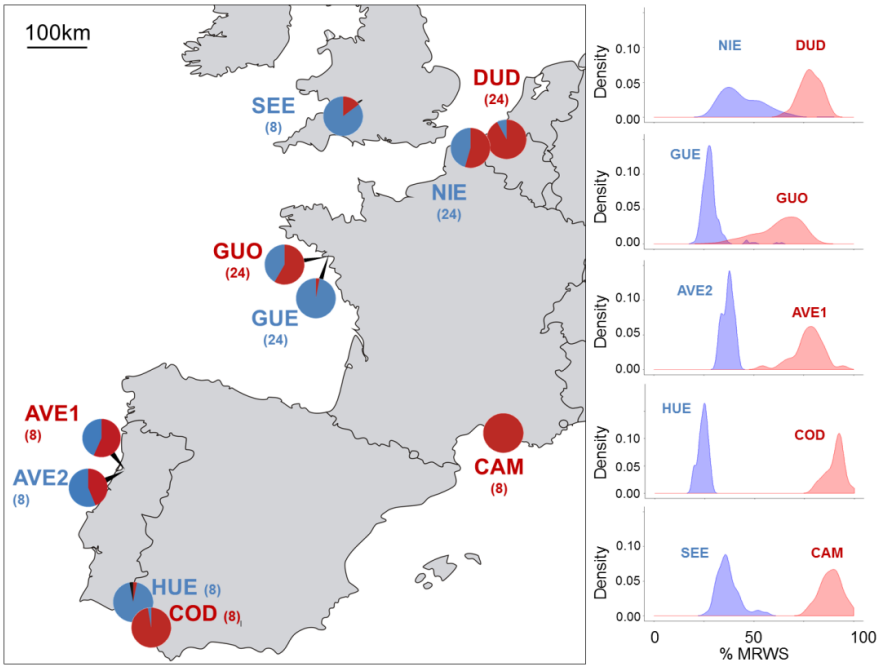


Figure 32. *Pogonus chalceus* sampling locations (left) and density plots of the wing size distribution of the sampled populations (right). Population names in blue letters indicate tidal habitat, red letters indicate seasonal habitat. Pie charts show the mtIDH allozyme distribution (mtIDH-B in red, mtIDH-D in blue; see Van Belleghem and Hendrickx 2013). Numbers between brackets indicate sample sizes. Wing size is expressed as percentage of maximal realizable wing size (%MRWS).

DNA EXTRACTION, RAD LIBRARY PREPARATION AND SEQUENCING

DNA was extracted using the DNA extraction NucleoSpin® Tissue kit (Macherey-Nagel GmBH). Extracted genomic DNA was normalized to a concentration of 7.14 ng/µl and processed into RAD libraries according to Etter et al. (2011), using the restriction enzyme SbfI-HF (NEB) and 16 PCR cycles for final enrichment. A total of nine RAD libraries including 16 individuals each and, hence, a total of 144 individuals were sequenced paired-end for 100 cycles (i.e. 100 bp) in a single lane of an Illumina HiSeq2000 platform according to manufacturer's instructions.

SEQUENCE ANALYSIS

The raw data was demultiplexed to recover individual samples from the Illumina libraries using the Stacks v1.20 software (Catchen *et al.* 2013). Reads were quality filtered when they contained 15 bp windows of mean Phred scores lower than 10. PCR duplicates were identified as almost (i.e. allowing for sequencing errors) identical reverse read sequences (which result from random shearing) and removed, using a custom Perl script (Kerth 2012). Loci were built *de novo* (denovo_map.pl) and using the *P. chalceus* genome assembly (ref_map.pl) with Stacks v1.20 (Catchen *et al.* 2013).

DE NOVO DATASET First, Stacks identifies exactly matching reads (i.e. stacks) within each individual. Next, loci are built within each individual by combining stacks that putatively represent alleles. Finally, loci from each individual are matched to determine which haplotype alleles are present at every locus in each individual (i.e. catalog loci). For the *de novo* building of loci we used a minimum depth of coverage (-m) of 5 for the exactly matching stacks. Distance allowed between stacks (-M) and distance allowed between catalog loci (-n) were set at 1 and 2, respectively. These latter parameters were chosen after testing a range of values for each parameter and the parameter combination that resulted in a minimum number of loci with more than two alleles and a maximum number of shared loci between individuals. This dataset is further referred to as the '*de novo* dataset' and is mostly used to study differentiation and evolutionary relations among populations.

REFMAP DATASET We mapped reads to the *P. chalceus* genome assembly v1.0 (CHAPTER 4) using BWA v.0.7.9a (Li & Durbin 2010). We used a maximum edit distance (-n) of 2 % of the read length and a maximum insert size (-a) of 1,000 bp. Uniquely mapped reads were filtered based on mapping quality using SAMtools (view –q

25) (Li *et al.* 2009). Next, loci of the forward read were built using Stacks v1.20 using the BWA alignments to the reference genome (Catchen *et al.* 2013) with a maximum number of two mismatches (-n) allowed between loci of different individuals. This dataset is further referred to as the 'refmap dataset' and was used to construct sequence alignments of the outlier loci and to help in assessing linkage between RAD tags.

We calculated the nucleotide diversity (π) by averaging the number of nucleotide differences per site between two sequences. Heterozygosity (H_z) was calculated as the average proportion of polymorphic sites within individuals, both within populations and across populations.

CLUSTER ANALYSIS

We used the Bayesian clustering method implemented in Structure v2.3.4 (Falush *et al.* 2007) to determine genetic clusters and to infer the number of clusters that best fit the data. We used the *de novo* dataset for this analysis and only loci that were present in at least 50 % of the individuals in all populations were retained for the analysis. Given a certain number of populations (K), we calculated the log-probability of the data ($LnP(D|K)$) and compared across a range of K values to determine which number of clusters best fits the data and whether these coincide with the geographical locations. A Monte Carlo Markov Chain (MCMC) was run for 100,000 iterations and a burn-in of 10,000 under the admixture model with correlated allele frequencies. We performed 3 replicate runs at each K from 1 to 5 for the Guérande and Belgian populations and 1 to 10 for all populations combined.

OUTLIER ANALYSIS

Outlier analysis was performed on both the *de novo* and the refmap dataset. Fst (Wright's fixation index) values were calculated between each pairwise population comparison for each SNP and all SNPs combined using Stacks v1.20 (Catchen *et al.* 2013). Loci showing extreme allele frequency differences across the tidal and seasonal environments were identified using Bayenv2 (Coop *et al.* 2010). To account for differences in sample sizes and neutral correlation of allele frequencies across populations due to shared history and gene flow, Bayenv2 implements a Bayesian method to estimate the empirical pattern of covariance in allele frequencies between populations from a set of markers, and then uses this as a null model for a test at individual SNPs. Population covariance matrices were calculated using all available SNP data and by averaging matrices from 500,000

MCMC iterations sampled every 500 iterations with a burn-in of 100,000. Next, correlations between the environment and SNPs were detected by estimating Bayes Factors (BF). For this, populations of the tidal and seasonal habitats were assigned an environmental value representing the mean wing size in the populations. BFs were estimated using 100,000 MCMCs and sampling every 500 iterations. When multiple SNPs within a RAD tag were present, the SNP with the highest BF value was selected.
Linkage Disequilibrium (*LD*) between each pair of outlier loci was tested within each population using Genepop v4.2 (Rousset 2008). Genepop v4.2 tests for an association between pairs of loci by constructing contingency tables of the genotypic counts and analyses them using a Markov chain method (Dememorization number = 1000; Number of batches = 100; Number of iteration per batch = 1000) to estimate exact *P* values (Raymond & Rousset 1995). P values were adjusted using Bonferroni correction.

From the BWA alignments to the *P. chalceus* reference assembly we obtained consensus sequences for genomic regions with a minimum coverage of 5 for each individual using SAMtools (Li *et al.* 2009) and Seqtk v1.0 (Li 2013). Sequence alignments were built for the outlier loci identified in the Canal-Pond-Nieuwpoort-Dudzele comparison using Bayenv2 with a minimum Bayes Factor (BF) of 15. These sequence alignments were then subdivided according to the genotype of the SNP that was identified as outlier (Figure 33). Only sequences were retained from individuals with a homozygous SNP genotype so that polymorphic positions associated with the outlier SNP could be correctly assigned to a subset. Nucleotide diversity (π) and Tajima's *D* values were calculated using a home-made Python script and DendroPy v3.12.0 (Sukumaran & Holder 2010). Differences in nucleotide diversity (π) and Tajima's *D* values between the subsets of sequences were tested using paired t-tests and their associations were tested using Proc GLM (SAS v9.4) and a Type 3 sum of squares analysis. Neighbor-Net networks were constructed using SplitsTree v4.13.1 (Huson and Bryant 2006) with default settings.

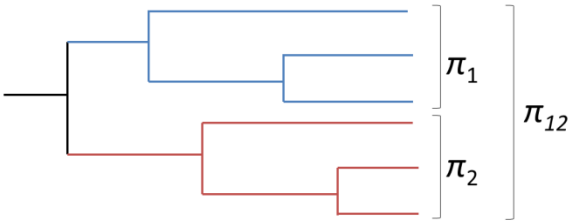


Figure 33. Graphical representation of calculation of nucleotide diversity (π) among sets of sequences. π_1 = π among sequences with SNP allele most frequent in the tidal canal and Nieuwpoort populations. π_2 = π among sequences with SNP allele most frequent in the seasonal pond and Dudzele populations. π_{12} = π among all sequences with homozygous SNP genotype.

Finally, previously identified gene sequences involved in wing development (Van Belleghem *et al.* 2012) were aligned to the *P. chalceus* genome assembly. Subsequently, RAD tags were identified in scaffolds associated with these wing development genes and *Fst* values and distance between the RAD tag and wing development gene were calculated.

RESULTS

READ MAPPING

Sequencing resulted in on average 229,519 (SD = 125,244) reads per individual, after removing PCR duplicates (Appendix 17). For the *de novo* dataset, we obtained on average 2,773 (SD = 922) loci per individual (Table 10). 128 of these loci were polymorphic and present in all individuals. 738 polymorphic loci were present in at least 50 % of the individuals in each population. Further, we obtained 1,325 polymorphic loci among the Guérande canal, Guérande pond, Nieuwpoort and Dudzele populations present in at least 50 % of the individuals.

On average 52.57 % (SD = 6.63 %) of individuals' reads mapped to the *P. chalceus* genome assembly. We estimated the genome size from the percentage of mapped reads for each individual onto the draft genome. As the assembled genome has a length of 312.78 Mb, this results in an expected genome size of 528.35 – 680.85 Mb. Although this comprises the estimated genome size of 530.28 Mb calculated using *k*-mer frequencies (Chapter 4), this previous estimation may have underestimated the genome size. This likely results from not incorporating sequence repetitiveness when using *k*-mer frequencies. Building loci using the reference genome resulted in on average 1,844 (SD = 459) loci in each individual (Table 10). 319 of these loci were polymorphic and present in all individuals. 814 polymorphic loci were present in at least 50 % of the individuals in each population. Further, we obtained 987 polymorphic loci among the Guérande canal, Guérande pond, Nieuwpoort and Dudzele populations present in at least 50 % of the individuals.

The higher number of polymorphic loci found when using reads aligned to the reference genome results from the alignment strategy in which BWA allows more mismatches when building loci. This allowed building loci for more strongly diverged haplotypes, but resulted in more loci with more than two alleles within a single individual (0.35 % in *de novo* loci building versus 5.37 % when using BWA alignments). These latter loci likely result from assembling paralogous or repetitive sequences.

GENOME-WIDE ESTIMATION OF GENETIC DIFFERENTIATION

From the loci and SNP genotype data we calculated genetic variation within and across populations (Table 10). Average nucleotide diversity (π) was 0.0031 (SD = 0.0006) *within* each population and 0.0039 (SD = 0.0006) across populations. Genetic diversity measures across populations increased only slightly, indicating recent population differentiation and/or high rates of gene flow. However, the conservative (and unbiased) nature with which loci were built *de novo* may have resulted in an underestimation of the nucleotide diversity. Conversely, the marked higher π and H_z among the loci built using the reference genome likely results from the high percentage of erroneous loci with more than two alleles.

Table 10. Comparisons of loci statistics built *de novo* and using the *P. chalceus* reference genome. CPND = Guérande canal, Guérande pond, Nieuwpoort and Dudzele.

	De novo		**Reference genome**	
		SD		**SD**
Average n loci/individual	2,773	922	1,844	459
Loci in all individuals	128	-	319	-
Loci in all CPND individuals	254	-	491	-
Loci in all individuals > 50 %	738	-	814	-
Loci in all CPND individuals > 50%	1,325	-	987	-
Polymorphic sites (Individual average)	281	41	125	61
Polymorphic sites (Population average)	1,882	439	401	58
π within individuals	0.0014	0.0002	0.0044	0.0005
π within populations	0.0031	0.0006	0.0054	0.0006
π across populations	0.0039	-	0.0065	-
H_z within individuals	0.0014	0.0002	0.0044	0.0005
H_z within populations	0.0017	0.0002	0.0050	0.0004
H_z across populations	0.0017	-	0.0045	-

As expected, *Fst* values significantly increased with geographical distance between each population ($F_{1,43}$ = 13.22, P = 0.0007; Figure 34). Distribution of *Fst* values between populations shows a clear L-shape for all population comparisons (Figure 35A). The L-shape is especially pronounced between the sympatric Guérande canal (GUE) and pond (GUO) population and the Nieuwpoort (NIE) and Dudzele (DUD) population. This likely indicates high rates of gene flow between sympatric or closely located pairs of populations because gene flow keeps most values low while selection increases divergence at a minority of loci. Between the Spanish Huelva (HUE) and Coto Doñana (COD) population *Fst* values are generally higher (Figure 35A). Moreover, the *Fst* distribution between the sympatric Portuguese populations (AVE1 versus AVE2) is even higher compared to allopatric populations (Figure 35A). Based on the *Fst* values, the Portuguese populations are more genetically diverged from all other populations (Figure 34).

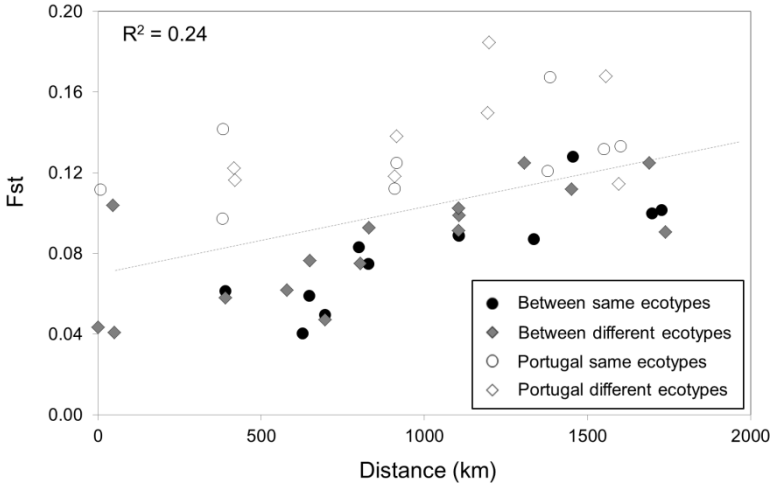


Figure 34. Relationship between average *Fst* between each population pair and distance between the populations. A distinction is made between *Fst* comparisons of populations from the same or different ecotypes. Comparisons including the Portuguese populations are indicated separately. The grey line indicates the regression line.

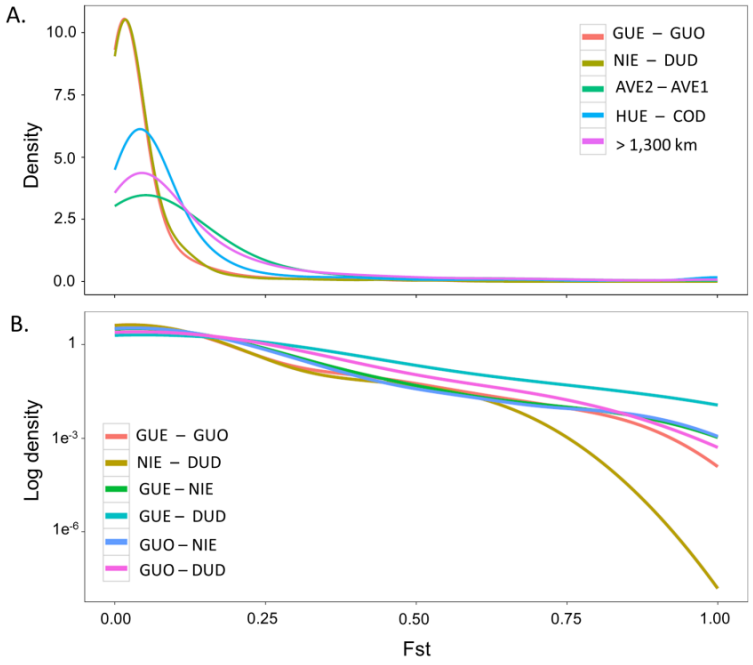


Figure 35. (A.) *Fst* density distribution of comparison between populations separated by less than 50 km (GUE-GUO, NIE-DUD, AVE2-AVE1 and HUE-COD) and of populations separated by more than 1,300 km. (B.) Logarithmic density distribution of *Fst* between GUE, GUO, NIE and DUD.

When only incorporating the Guérande canal (GUE), Guérande pond (GUO), Nieuwpoort (NIE) and Dudzele (DUD) populations, cluster analysis identified four genetic clusters corresponding to the four sampling locations (Figure 36A). Virtually all individuals were assigned to a genetic cluster that corresponds to the sampled population. When all populations were incorporated, highest likelihood was given to five genetic clusters (Figure 36B), wherein the Nieuwpoort and Dudzele as well as the Guérande canal and pond were assigned to the same cluster according to geographic proximity. In this latter case, only two clusters were recognized for the Guérande and Belgian populations according to location. The France Camargue (CAM) population was recognized as a separate genetic cluster. The Severn Estuary population from the UK (SEE) showed similarity to both the Belgian and France Atlantic populations. Among the Portuguese (AVE1 and AVE2) and Spanish populations (HUE and COD) two genetic clusters were recognized. AVE2 and COD seemed to be clearly distinct, whereas AVE1 and HUE have elements of both former clusters.

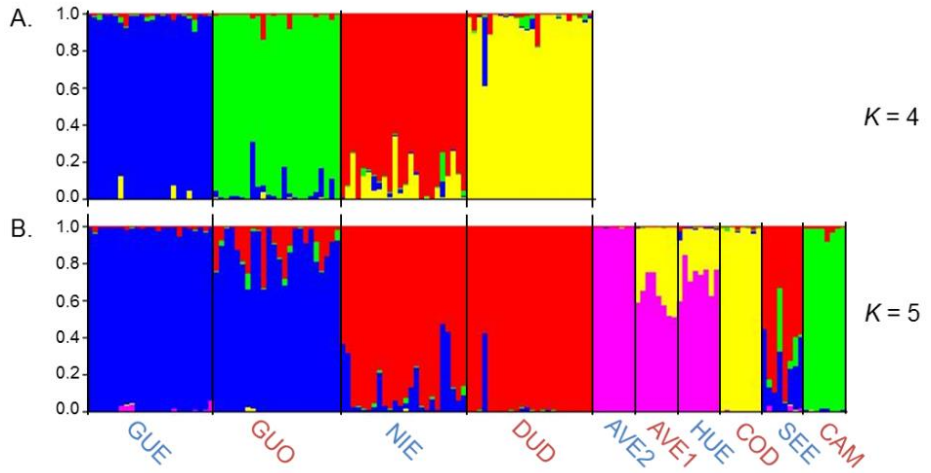


Figure 36. Population structure for (A.) the Guérande canal (GUE), Guérande pond (GUO), Nieuwpoort (NIE) and Dudzele (DUD) population and for (B.) all populations combined for a number of populations $K = 4$ and $K = 5$, respectively.

Principal Coordinate Analysis (PCoA) shows similar results as the genetic clustering analysis (Figure 37A). The Guérande canal (GUE) and Guérande pond (GUO) populations cluster together according to location. The Severn estuary (SEE) population clusters between the Belgian and France Atlantic populations. The Portuguese populations (AVE1 and AVE2) cluster separately from all the other populations. These results indicate a clear effect of geographical isolation on the genetic variation among

populations. However, the Coto Doñana (COD) population clusters closer together with the Mediterranean Camargue (CAM) population compared to the geographically close Huelva (HUE) population.

Performing a PCoA on only the Guérande canal (GUE), Guérande pond (GUO), Nieuwpoort (NIE) and Dudzele (DUD) populations indicates genetic variation resulting from geographical isolation as well as ecotypic divergence (Figure 37B). Moreover, in this analysis, the tidal populations from different localities (GUE and NIE) cluster closer together compared to the seasonal populations from the same localities. The distribution of genetic variation among the populations is comparable to the differences found in *Fst* distribution (Figure 35B).

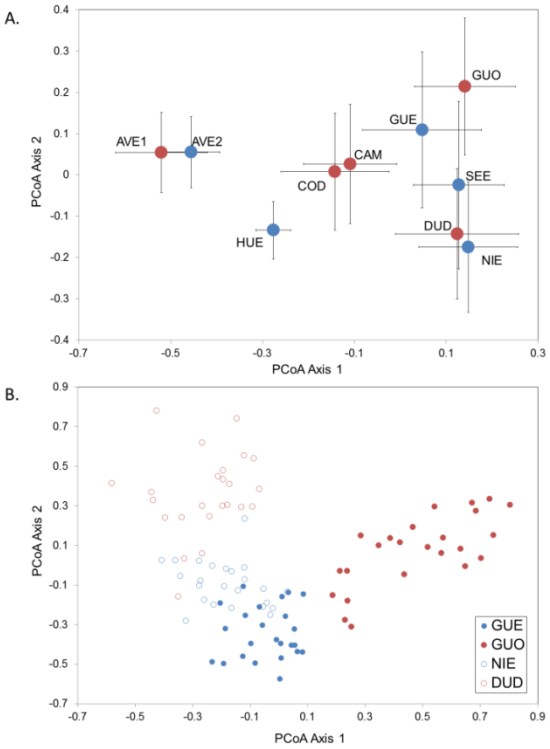


Figure 37. Principle Coordinate Analysis (PCoA). (A.) PCoA for all populations including 128 polymorphic markers. Points and error bars represent the population average and the standard deviation of the genetic variation along the first and second PCoA axis. The first and second axis explains 9.05 % and 7.03 %, respectively, of the variation contained in the dataset. Tidal and seasonal populations are indicated in blue and red, respectively. (B.) PCoA for the Guérande canal (GUE), Guérande pond (GUO), Nieuwpoort (NIE) and Dudzele (DUD) populations including 254 polymorphic markers. Each points represents a single individual. The first and second axis explains 7.75 % and 6.96 %, respectively, of the variation contained in the dataset. Genetic markers were used that were polymorphic, were sequenced in all individuals and had a minimum coverage of 5.

OUTLIER LOCI

SNPs showing strong association with the tidal and seasonal environments were identified while taking into account neutral population structure (Table 11). When using all populations, we found two strongly associated SNPs with the tidal and seasonal habitats with a Bayes Factor (BF) larger than 15. These were found among both the *de novo* built loci as well as among the reference built loci. Among the tidal Guérande canal (GUE) and Nieuwpoort (NIE) and seasonal Guérande pond (GUO) and Dudzele (DUD) populations we found 24 outlier loci among the *de novo* built loci. Only one pair of loci showed significant linkage disequilibrium. Among the reference built loci we found 40 outlier loci. However, after building alignments of the latter reference built loci and manual inspection, 17 loci were discarded as they resulted from incorrectly aligned reads. Further, four pairs of outlier loci were located on the same scaffold and additional testing for linkage disequilibrium found significant linkage for one more loci, reducing the total set of unlinked outlier loci to 18. Of the 24 outlier loci found among the *de novo* dataset, 15 were shared with the 23 outlier loci identified among the 'refmap dataset'.

Table 11. Number of identified outlier loci with Bayes factor larger than 15 and number of loci with significant pairwise linkage disequilibrium. CPND: Comparison of Guérande canal, Guérande pond, Nieuwpoort and Dudzele populations. ALL: Comparison among all populations.

	De novo		Ref map	
	CPND	ALL	CPND	ALL
Outlier loci	24	2	23	2
Pairs with LD	1	0	5	0
Unlinked outlier loci	23	2	18	2

Comparing *Fst* values between Guérande canal (GUE) and Guérande pond (GUO) populations and Nieuwpoort (NIE) and Dudzele (DUD) shows that most loci with a high Bayes Factor also have a high *Fst* value in both population comparisons (Figure 38). However, several loci have a high *Fst* value in only one of the population comparisons, indicating independent differentiation. Further, *Fst* values between the sympatric Guérande canal (GUE) and Guérande pond (GUO) populations are generally higher than *Fst* values between the Nieuwpoort (NIE) and Dudzele (DUD) population which are located approximately 37 km apart.

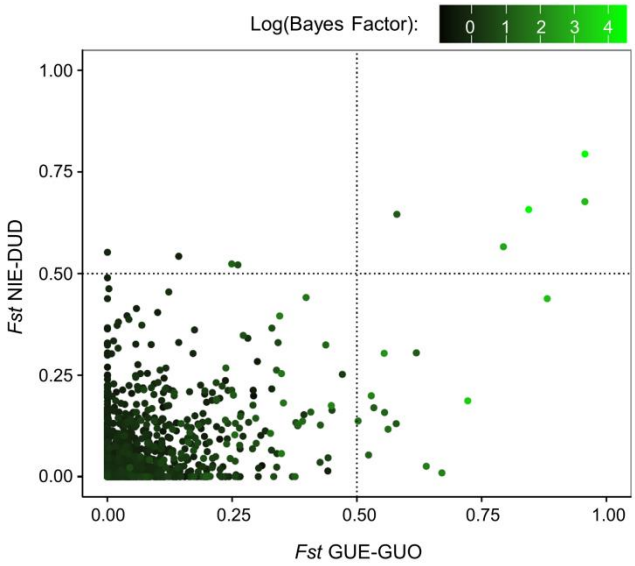


Figure 38. Plot of the *Fst* values between the Canal-Pond populations against the *Fst* values between the Nieuwpoort-Dudzele populations. Dashed line indicates *Fst* value of 0.5. Points are colored according to Bayes Factors (BF) estimated using Bayenv2. Points with high BF indicate SNPs that are strongly correlated with the environmental variable differentiating the tidal and seasonal populations.

For the 18 retained outlier loci, we constructed alignments from the paired-end RAD tag read mappings to the *P. chalceus* genome assembly. This was done for individuals from the Guérande canal (GUE), Guérande pond (GUO), Nieuwpoort (NIE) and Dudzele (DUD) populations. The obtained alignments had an average length of 858 bp (SD = 254). Next, these alignments were subdivided according to the outlier SNP allele and the nucleotide diversity and Tajima's *D* values among each subset of sequences was calculated for the following subsets: (i) nucleotide diversity and Tajima's *D* among the sequences associated with the allele most frequent in the *tidal* habitats (π_1 and D_1), (ii) nucleotide diversity and Tajima's *D* among the sequences associated with the allele most frequent in the *seasonal* habitats (π_2 and D_2) and (iii) nucleotide diversity and Tajima's *D* among all the sequences (π_{12} and D_{12}).

In general, nucleotide diversity among the sequences associated with the allele most frequent in the tidal habitats (π_1) was much lower compared to nucleotide diversity among the sequences associated with the allele most frequent in the seasonal habitats (π_2) (paired t(17) = -3.5, P = 0.003; Figure 39A). Fifteen out of the 18 unlinked outlier loci had nucleotide diversity lower among the sequences associated with the allele most frequent in the tidal habitats. Nucleotide diversity among the sequences associated with

the allele most frequent in the tidal habitats (π_1) did not markedly increase with the total nucleotide diversity (π_{12}) ($F_{1,16}$ = 1.87, P = 0.20). In contrast, nucleotide diversity among the sequences associated with the allele most frequent in the seasonal habitats (π_2) strongly increased with total nucleotide diversity (π_{12}) ($F_{1,16}$ = 23.88, P = 0.0002), indicating a different evolutionary history of these alleles (Figure 39A). In contrast, overall nucleotide diversity did not markedly differ between populations from tidal and seasonal habitats (Appendix 18), suggesting that the reduced nucleotide diversity among the sequences associated with the allele most frequent in the tidal habitats (π_1) is restricted to the outlier loci and not neutral gene sequences.

We did not find any significant differences in Tajima's *D* values between the sequence subsets associated with the allele most frequent in the seasonal and tidal habitats (paired t(15) = -1.14, P = 0.27). However, among the sequences associated with the allele most frequent in the tidal habitats (D_1) more extreme negative Tajima's *D* values were observed (Figure 39B). Negative Tajima's *D* values signify an excess of low frequency polymorphisms relative to the expectation under neutrality and are expected under population size expansion (e.g. after a bottleneck or a selective sweep) and/or purifying selection. Furthermore, we found that Tajima's *D* values significantly increase with increasing nucleotide diversity among the subsets of sequences ($F_{1,32}$ = 4.67, P = 0.04). This indicates a disappearing signal of the bottleneck or selective sweep as time progresses and more mutations arise in the sequences.

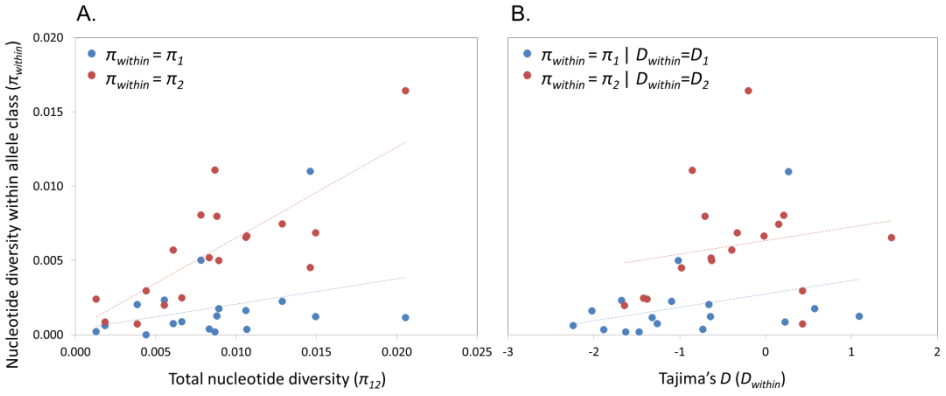
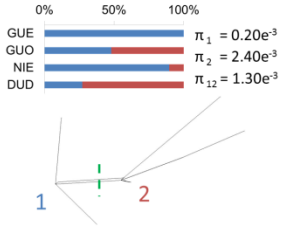


Figure 39. (A.) Plot of nucleotide diversity among the total set of sequences (π_{12}) against nucleotide diversity within each subset of sequences (π_{within}). π_1 = π among sequences with SNP allele most frequent in canal and Nieuwpoort populations. π_2 = π among sequences with SNP allele most frequent in pond and Dudzele populations. Dashed lines represent regression lines. (B.) Plot of Tajima's *D* values against the nucleotide diversity (π_{within}) within each subset of sequences.

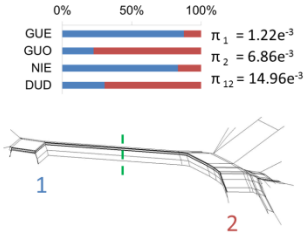
Haplotypes associated with each nucleotide that differentiated both ecotypes consistently clustered together, indicating that the shared genetic variation between populations of the same habitats has a singular mutational origin (Figure 40). Neighbor-Net networks of 13 loci showed resemblance with the network constructed for the *mtIdh* gene in *P. chalceus* with low haplotype diversity among the sequences associated with the tidal habitat and high haplotype diversity among sequences associated with the seasonal habitat (CHAPTER 2). Moreover, several networks showed strong differentiation between the subsets of sequences, indicating reduced recombination between these haplotype sets. Two loci (locus 917 and 1652) showed a notable opposite pattern of sequence variation with more nucleotide variation among the sequences associated with the tidal habitat.

We identified all wing development genes in the *P. chalceus* genome assembly, apart from one (Appendix 19). All these genes were located in a different scaffold. At least one RAD tag was present in ten of these scaffolds (Appendix 19). Three of these scaffolds had a RAD tag with a markedly higher *Fst* value. Two had a high *Fst* in both the Guérande canal versus pond and Nieuwpoort versus Dudzele population comparison (*wingless* (*Fst* = 0.45/0.17) and *Sex combs reduced* (*Fst* = 0.16/0.28)). One RAD tag associated with a wing development gene had a higher *F*st value only in the Guérande canal versus pond population comparison (*Nubbin* (*Fst* = 0.41/0.08)). All genes related to hormones were identified in the *P. chalceus* genome assembly. However, only two scaffolds associated with one of these genes (cytochrome P450, family 307 and ecdysteroid 22-hydroxylase) had a RAD tag which both had *Fst* values of 0. The *mtIdh* gene maps to scaffold 1175 in the *P. chalceus* genome assembly. The *Fst* value of the RAD tag positioned in this scaffold was 0.46 for the Guérande canal versus pond comparison and 0.38 for the Nieuwpoort versus Dudzele comparison. This RAD tag was located 3,627 bp downstream of the end of the *mtIdh* gene, indicating that the previously found differentiation in the *mtIdh* gene extends into a genomic island of differentiation and may be hitchhiking with a closely linked selected target. However, the RAD tag was not identified as an outlier in our analysis due to very stringent scoring conditions.

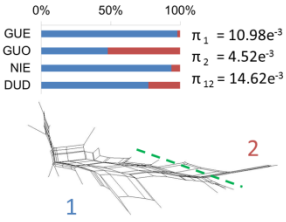
Locus 361



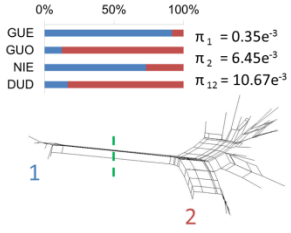
Locus 879



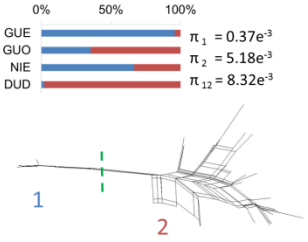
Locus 917



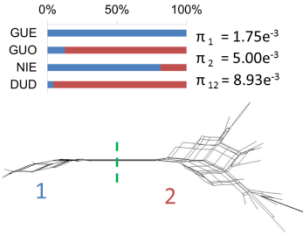
Locus 921



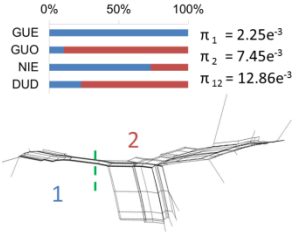
Locus 923



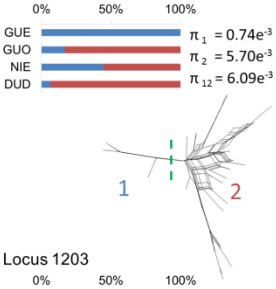
Locus 949



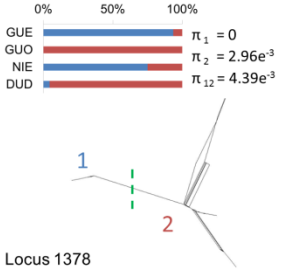
Locus 1081



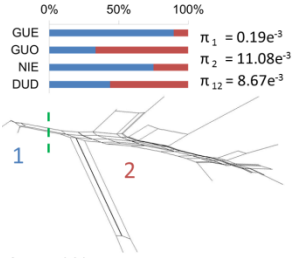
Locus 1093



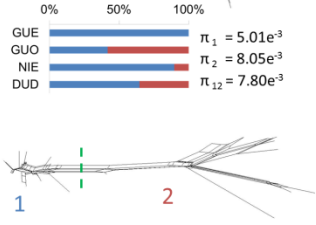
Locus 1104



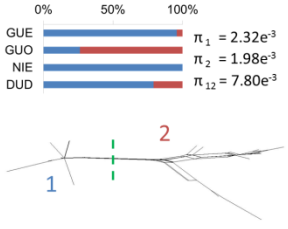
Locus 1108



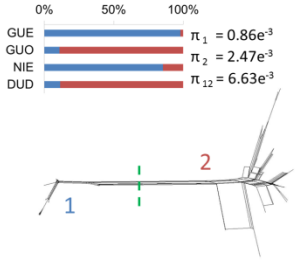
Locus 1203



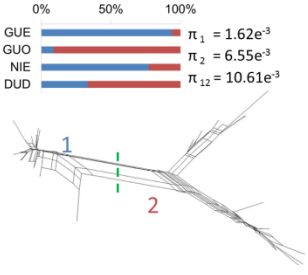
Locus 1378



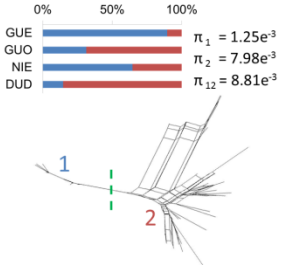
Locus 1415



Locus 1555



Locus 1617



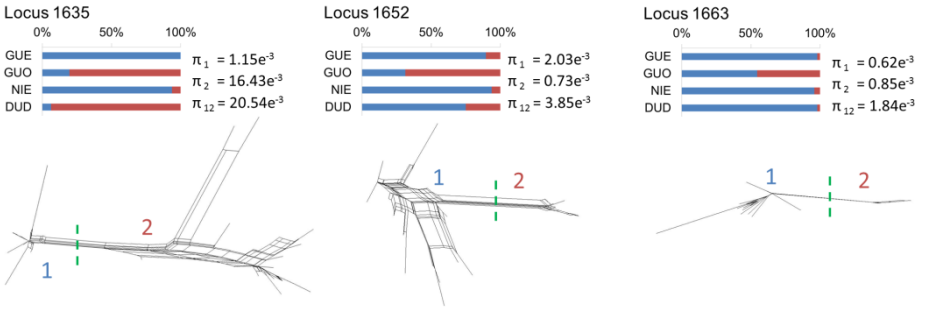


Figure 40. Neighbor-Net networks of unlinked outliers with Bayes Factor (BF) larger than 15. The green dashed line indicates the split between allele 1 and 2. Allele 1 represents the allele most frequent in the tidal populations (blue), whereas allele 2 represents the allele most frequent in the seasonal populations (red). The frequencies of allele 1 and 2 are given in each population; Guérande canal (GUE), Guérande pond (GUO), Nieuwpoort (NIE) and Dudzele (DUD). π_1 = nucleotide diversity among the haplotypes of allele 1. π_2 = nucleotide diversity among the haplotypes of allele 2. π_{12} = nucleotide diversity among all haplotypes.

DISCUSSION

It is generally argued that high rates of gene flow result in divergence in only a few regions that harbor genes under strong divergent selection (Via 2001, Savolainen *et al.* 2006). Hence, the 'L-shaped' distribution of *Fst* values (i.e. most loci have low *Fst* values and a pronounced tail of extreme values) between the sympatric *P. chalceus* populations is consistent with gene flow keeping most values low while selection increases divergence at only a minority of loci. Allopatric divergence, on the other hand, is expected to generate a more even distribution across the genome (e.g. Martin et al. 2013). However, in the early stages of allopatric divergence the 'L-shaped' distribution of *Fst* values is also expected and the long tail of extreme *Fst* values may be explained by selection in allopatry (Butlin 2010). This may explain the largely 'L-shaped' distribution found between allopatric population comparisons in *P. chalceus*. In addition, we found only a small increase of genetic diversity when combining populations. These estimated nucleotide diversity values in *P. chalceus* are in rough agreement with studies of genetic variation within and among stickleback populations using RAD tag sequencing (wihin: π = 0.0020 - 0.0027, across: π = 0.0034; Hohenlohe, Bassham, et al. 2010). In these sticklebacks, it is argued that this small increase in genetic diversity when combining populations is in agreement with the hypothesis that freshwater populations in the studied region have been derived post-glacially from oceanic populations. Similarly,

Atlantic coasts may have been inhabitable during the last glacial period for *P. chalceus* beetles and a recent spread may explain why genetic diversity increases only slightly when combining populations. Moreover, sticklebacks generally have a generation time of two years (Bell *et al.* 2004), whereas *P. chalceus* beetles have a generation time of one year, suggesting that the colonization of *P. chalceus* along the European coasts may be more recent.
Further, *Fst* values between the sympatric Guérande canal (GUE) and Guérande pond (GUO) populations were generally higher than *Fst* values between the Nieuwpoort (NIE) and Dudzele (DUD) populations that are located approximately 37 km apart. This may indicate more pronounced reproductive isolation between the tidal and seasonal populations in the Guérande. Possibly, this results from stronger selection against hybridization between the sympatric populations compared to the allopatric populations (i.e. reinforcement; Servedio and Noor 2003). Moreover, *Fst* values between the sympatric Portuguese populations (AVE1 and AVE2) were in the same range or even higher as very distant (> 1,300 km) allopatric populations.

The idea that speciation may occur in the presence of gene flow is being increasingly accepted (Pinho & Hey 2010, Smadja & Butlin 2011, Feder *et al.* 2012a). Theoretical models of sympatric speciation discuss the importance of genetic linkage versus the possibility of genome wide differentiation during early speciation (Via 2001, Fry 2003, Yeaman & Whitlock 2011, Feder *et al.* 2012a). Genetic architectures with fewer, larger and more tightly linked divergent alleles are expected to reduce the swamping of divergence at weakly selected alleles (Yeaman & Whitlock 2011) and the effect of recombination on the association between different adapted loci or traits (Fry 2003, Bolnick & Fitzpatrick 2007, Pinho & Hey 2010). Indeed, several studies have found traits defined by only a few quantitative trait loci (QTL) of large effect. These include pelvic girdle (Shapiro *et al.* 2004) and armor plating (Colosimo *et al.* 2004) in sticklebacks, coloration in beach mice (Steiner *et al.* 2007) and large effect QTLs explaining phenotypic variation in several *Timema cristinae* stick insect traits (Comeault *et al.* 2014). Moreover, fine mapping of skeletal traits in mice revealed that the large effect QTL associated with this trait actually contains several tightly linked genes contributing small individual effects (Christians & Senger 2007). Further, in lake whitefish large islands of divergence were found in the early stages of adaptive divergence (Renaut *et al.* 2012). Finally, in pea aphids genetic linkage has been found between performance and assortative mating traits which is argued to promote speciation in this species (Hawthorne & Via 2001). However, these examples are in contrast with recent studies demonstrating genome wide divergence between sympatrically differentiating populations in *Rhagoletis pomonella* (Michel *et al.* 2010) and *Anopheles gambiae* (Lawniczak *et al.* 2010) and

theoretical arguments that selection on many loci can generate widespread divergence even with gene flow (Feder & Nosil 2010, Feder *et al.* 2012b).
Indeed, we found 24 unlinked SNPs that are strongly associated with adaptive divergence between the tidal and seasonal habitats. A RAD tag located in a genomic scaffold comprising the *mtIdh* locus showed marked *Fst* values between both the tidal Guérande canal and seasonal pond populations as well as between the tidal Nieuwpoort and seasonal Dudzele populations. However, our outlier analysis did not retain this locus, emphasizing our conservative scoring and that many more loci may be differentiated between populations and be associated with the repeated adaptive divergence. The 24 retained outlier loci did not show significant linkage disequilibrium within populations, suggesting that selection acts on multiple unlinked loci and that linkage is not a main factor maintaining the ecotypic differences in sympatry. These findings are in agreement with a previously demonstration that genes involved in wing size and the strongly differentiated *mtIdh* alleles, that are associated with the tidal and seasonal habitats, are genetically unlinked (Van Belleghem & Hendrickx 2014). Furthermore, finding these multiple unlinked outlier loci between sympatric populations may be indicative that even very early stages of the speciation process may by characterized by genome wide adaptation. Possibly, this is driven by a reproductive isolating mechanism that reduces gene flow and assists natural selection in the evolution of distinct ecotypes that have diverged in multiple unlinked loci. Additionally, comparisons between populations within localities showed several loci with increased *Fst* values, indicating adaptation unique to the localities.

Some of the outlier loci were situated in scaffolds that contain genes that are important for wing development in *Drosophila melanogaster* (Weihe *et al.* 2005) and *Tribolium castaneum* (Richards *et al.* 2008), such as *wingless*, *Sex comb reduced* and *Nubbin*. *Wingless* is a segment polarity gene and has a role in the establishment of different cell fates. *Sex comb reduced* is required for labial and first thoracic segment development. The RAD tag associated with *Nubbin* only showed a marked *Fst* value between the Guérande canal and pond population. *Nubbin* is a regulatory protein implicated in early development. Studying the genetic variation associated with these genes in more detail should allow identifying whether these genes are actually targets of selection.

All outlier loci investigated were shared between the repeatedly adapted populations. Whether this shared genetic variation results from reuse of standing genetic variation or rather introgression of adaptations from one population to other populations is difficult to infer. Reuse of globally shared standing genetic variation has, for instance, been demonstrated to play an important role in repeated evolution of distinct marine and

freshwater sticklebacks (Jones *et al.* 2012). Introgression, on the other hand, has been shown to even allow transferring adaptations between infrequently hybridizing *Heliconius* butterfly species (Dasmahapatra *et al.* 2012). Thirteen out of the 18 retained outlier loci identified using the reference genome had a marked reduction in nucleotide diversity, indicative for a recent selective sweep or bottleneck. Furthermore, reconstructing the genealogical relationships between the haplotypes showed that about half of the outlier loci have a deep differentiation between the subsets of sequences associated with the tidal and seasonal habitats. This suggests reduced recombination between these haplotype sets. Moreover, the pattern of haplotype variation and structure is similar to the pattern found in the *mtIdh* gene in which the reduced recombination is suggested to be due to geographical isolation as suggested by coalescent simulations performed on the *mtIdh* gene (CHAPTER 2).

Finding a similar evolutionary history in the majority of the outlier loci suggests a largely singular evolution of the short-winged ecotypes. Possibly, as suggested by the deep divergence and reduced nucleotide diversity in several of the outlier loci associated with the tidal habitat, a large set of adapted loci associated with the short-winged ecotype from tidal marshes spread recently along the Atlantic coasts from a partially isolated subpopulation. Hence, the observed pattern of genetic variation among sympatric populations may be consistent with secondary contact and admixture after a period of geographical separation. Moreover, the reuse of adaptive genetic variation in the repeated occurrence of the ecotypes may be important for the repeated evolution and maintenance of the ecotypes in sympatry.

ACKNOWLEDGEMENTS

We gratefully thank Dr. Karim Gharbi from the Genepool Unit at the University of Edinburgh for his help in processing of the libraries and sequencing of the RAD tags. Peter Verhasselt and Peggy Van den Zegel from the Nucleomics Core (Leuven, Belgium) are thanked for their help in shearing of the RAD libraries. Alexandre Ramos is thanked for his help in obtaining specimens from Portugal.

CHAPTER 6

DETERMINING THE LINK BETWEEN SELECTION AND ASSORTATIVE MATING IN SYMPATRIC *POGONUS CHALCEUS* POPULATIONS

Steven M. Van Belleghem [1,2]
Katrien De Wolf [1,2]
Frederik Hendrickx [1,2]

[1] Terrestrial Ecology Unit, Biology Department, Ghent University, K. L. Ledeganckstraat 35, 9000 Gent, Belgium

[2] Department Entomology, Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000 Brussel, Belgium

ABSTRACT

Understanding how disruptive selection can result in the evolution of reproductive isolation is of major interest in the study of sympatric speciation. One such mechanism proposes that traits subjected to disruptive selection directly result in spatial sorting between the diverging populations. This may result in a match between phenotype and habitat performance and decrease mating opportunities between the diverging ecotypes. Here we study habitat preference and performance in the salt marsh beetle *Pogonus chalceus*, which is found in two strongly contrasting habitats which differ in hydrological regime. Tidal habitats are inundated frequently for only a few hours, whereas seasonal habitats are usually inundated for longer periods. Based on choice experiments, we found no indication of assortative mate preference in sympatric *P. chalceus* populations. Alternatively, we demonstrate that short-winged populations from tidally inundated marshes show less reluctance to inundation compared to populations from seasonal marshes. We argue that these behavioral differences may result in spatial sorting and can as such provide a unique and simple explanation for the persistence of distinct ecotypes in sympatric mosaics.

INTRODUCTION

Natural selection has long been claimed to be a major force in the evolution of new species (Darwin 1859). Moreover, when speciation occurs in the face of homogenizing gene flow, it is argued that strong disruptive selection is necessary to cause the evolution of reproductive isolation (Coyne & Orr 2004). However, verifying the link between natural selection and the evolution of reproductive barriers remains a major challenge (Kirkpatrick & Ravigné 2002, Coyne & Orr 2004, Schluter & Conte 2009, Hendry 2009, Nosil 2012, Faria *et al.* 2014). Studying this link is important as it allows to understand if and how populations can adapt to divergent ecological conditions when there is ample opportunity for homogenizing gene exchange.

A common problem encountered in the joint evolution of two or more genetically unlinked traits is that random mating and recombination will break down the adaptive gene combination (Felsenstein 1981). Therefore, it is often argued that sympatric speciation requires the evolution of non-random or assortative mating (Thibert-Plante & Gavrilets 2013). One mechanism that is often considered is mate preference in which, for instance, females prefer certain display traits that are only present in males belonging to

the same ecotype (Jones & Ratterman 2009). Another important and often considered more plausible mechanism for the evolution of assortative mating is habitat preference or spatial sorting. Here, individuals adapted to the same environmental conditions will experience positive selection for preferring the habitat to which they adapted. This spatial sorting of ecotypes in their respective habitat will then result in lower gene flow levels between individuals belonging to the different ecotype. However, even if variation exists for genes underlying habitat preference, recombination will yield individuals that prefer the habitat in which they are unfit, which challenges the likelihood of both the evolution as well as persistence of sympatric divergence. Therefore, for speciation to occur, an association is expected between habitat preference and genes that are involved in performance in that habitat. This link may include closely linked genes in the genome, or one gene affecting preference and performance pleiotropically. Pleiotropy causing this link has been called an automatic magic trait, indicating the debated and rare occurrence and nature of these traits (Servedio *et al.* 2011).

When individuals experience fitness trade-offs across different environments, selection should favor mechanisms that allow individuals to select and use habitats that best suit their phenotype. The importance of habitat preference as a means to optimize an individuals' fitness has long been recognized (Mayr 1963). Moreover, this is expected to greatly facilitate phenotypic segregation as the non-random distribution of locally specialized phenotypes also results in assortative mating among individuals with a similar genetic constitution. Several theoretical studies have demonstrated that habitat choice may aid sympatric speciation (Rice 1984, Johnson *et al.* 1996b, Kawecki 1996, Fry 2003). However, empirical examples of habitat preference as a promoting factor in sympatric divergence and speciation are strongly underrepresented (Edelaar *et al.* 2008, Edelaar & Bolnick 2012). Despite the relative ease of the concept, this most likely results from difficulties to study non-random dispersal and its effect on restricting gene flow between diverging ecotypes in the field (Jaenike and Holt 1991). Scarce examples include habitat preference in lake and stream three-spined sticklebacks (Bolnick *et al.* 2009) and genetic linkage between performance and host plant preference in phytophagous insects (Feder *et al.* 1994, Hawthorne & Via 2001, Berlocher & Feder 2002).

In this study, we investigate mechanisms that may affect assortative mating and their effect on the persistence of ecological divergence in a sympatric mosaic of the wing-polymorphic beetle *Pogonus chalceus*. The distribution of this carabid beetle is restricted to salt marshes along the Atlantic and Mediterranean coast. Here, the species is found in two contrasting environments that strongly differ in hydrological regime. One habitat consists of *tidal marshes* being year-round flooded on a regular basis, but for short

periods of at maximum a few hours only. The second habitat consists of *seasonal marshes* that are disconnected from the sea and are permanently inundated from late autumn to early spring. These pronounced hydrodynamic differences are associated with strong differences in ecological traits, wherein populations from tidal marshes are characterized by a reduced wing size, a smaller body size and high frequencies of the D allele of the *mitochondrial NADP⁺-dependent isocitrate dehydrogenase* gene (*mtIdh*). Conversely, populations inhabiting seasonal marshes have fully developed wings, a larger body size and higher frequencies of the mtIDH-B allele. These differences in wing size, and evidently mtIDH allele frequency, have a strong genetic basis as confirmed from lab crosses showing high heritability estimates of $h^2 = 0.9$ for wing size and a very strong association between mtIDH allele frequencies across, but not within populations demonstrating that both traits are exposed to similar selection pressures but genetically unlinked (Van Belleghem & Hendrickx 2014). Furthermore, these genetically divergent populations often co-occur at a very small geographic scale, such as at the Guérande salterns in France where both ecotypes occur in a sympatric mosaic wherein both habitats are separated by distances of a few meters (Dhuyvetter *et al.* 2007). The lack of significant genetic differentiation based on microsatellite markers among ecologically divergent but geographically nearby populations demonstrated that there is considerable gene flow among the two ecotypes (Dhuyvetter *et al.* 2007).

The high correlation between habitat types with different hydrological regimes and wing size in *P. chalceus* strongly suggests that these habitats select differently for dispersal ability (Van Belleghem & Hendrickx 2014). More specifically, we hypothesize that resident behavior during inundations is selected in tidal habitats, but not in seasonal habitats as beetles can tolerate the short inundations. This adaptation to tidal habitats may then result in spatial sorting of the ecotypes which, subsequently, results in assortative mating if the beetles mate in the habitat which they prefer.

MATERIALS & METHODS

SAMPLING

P. chalceus populations were sampled in two different regions i.e. in Belgium at Nieuwpoort (NIE) and Dudzele (DUD) in August 2013 and in France in the canal and pond habitats of the Guérande salterns in June 2013 (Figure 41). Populations from Nieuwpoort and Guérande canal, located in tidal marshes, have high frequencies of

individuals with strongly reduced wing sizes, whereas populations from Dudzele and Guérande pond populations, located in seasonal marshes, have on average long wing sizes (Figure 41). Differences in hydrological dynamics are the major differences between tidal and seasonal habitats. Populations from Nieuwpoort are sampled in a tidal marsh that is inundated frequently but for a few hours only, whereas Dudzele habitats comprise an inland salt marsh that is separated from the influence of the tides but becomes largely inundated for extensive periods during winter. Guérande ponds are being used for over a millennium to concentrate salt by evaporate water and are inundated irregularly for extensive periods and thus strongly resemble the seasonally inundated Dudzele habitats in its hydrological dynamics as well as in its vegetation. Guérande canals, on the other hand, are used to bring ocean water into the ponds and are subjected to the tides. These canals strongly resemble the Nieuwpoort habitats in its hydrological dynamics as well as vegetation. The Guérande ponds and canals are separated by distances of a few meters only.

Sampling was performed by hand and beetles were kept individually in a climate chamber at 15°C with a 16:8h light:dark photoperiod in small plastic cups containing a plaster being saturated with brackish water. Beetles were fed with parts of mealworms (*Tenebrio molitor*) every 2 to 3 days.

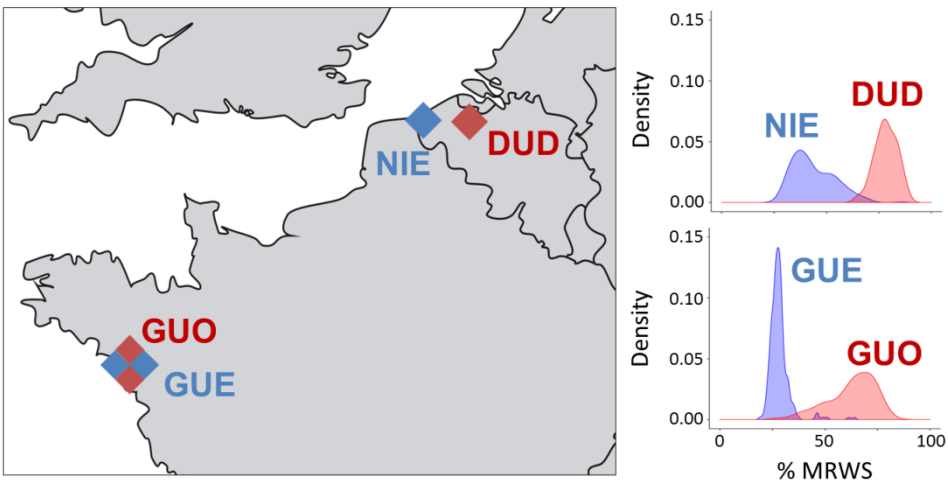


Figure 41. Sampling sites (left) and density plots of the wing size distribution of the sampled populations (right). Tidal habitats are indicated in blue, seasonal habitats are indicated in red. In the Guérande, tidal and seasonal habitats are found in hundreds of replicates and only separated 10-20 m. Wing size is expressed as percentage of maximal realizable wing size (%MRWS). GUO = Guérande pond habitat, GUE = Guérande canal habitat, NIE = tidal Nieuwpoort habitat, DUD = seasonal Dudzele habitat.

MATE PREFERENCE

P. chalceus beetles hibernate as adult and have their reproductive activity mainly during spring (Desender 1985). We tested if females of *P. chalceus* beetles show a higher acceptance rate towards males belonging to their respective ecotype. This was performed by subjecting beetles from the Guérande canal and pond habitats to a reciprocal no-choice mate preference trial. Females of each ecotype were exposed to a single male, which was placed in the container of the female. Couples were observed for 5 minutes and the occurrence of mating, the latency time until mating occurred and the duration of mating were recorded. A Generalized Linear Model (Proc GENMOD, SAS v9.4) assuming a binomial distribution and a logit link function was fitted to the data and significance of the effect of male and female ecotype and their interactions was tested by means of a Type 3 likelihood ratio (LR) test. A significant interaction between male and female ecotype indicates assortative mating (or negative assortative mating). To test significant differences in latency time until mating occurred and the duration of mating, a General Linear Model was fitted (Proc GLM, SAS v9.4) on log transformed time data and significance of the effects was tested by means of a Type 3 sum of squares analysis.

BEHAVIORAL RESPONSE TO INUNDATION

In this experiment, we tested whether both ecotypes differ in their response towards inundation of their habitat. To simulate inundation, beetles were kept in a sealed plastic cup containing plaster. The plaster was hollowed out at the bottom and this open space was accessible through a central corridor from the top to the bottom (Figure 42). At the side there was a small groove which allowed adding brackish water in a drop wise fashion to simulate inundation. Adding water through the groove ensured that no air bubbles would persist under the plaster in which the beetles could reside after the flooding. During the simulated inundation event, the behaviour of the beetles was observed and timed using a stopwatch. We noted the following behaviors: (i) presence or absence of escape behavior upon flooding and (ii) time until beetles reach the surface for taking air. For each test, three trials were performed, separated by approximately one week and using the same beetles.

Differences in behavioral response were compared between trials, habitats and regions. Habitats comprise the tidal versus seasonal habitats. Regions comprise the Guérande canal and pond populations versus the Belgian Nieuwpoort and Dudzele populations. We compared the proportion of individuals that expressed immediate escape behaviour

compared to the individuals that remained submerged upon inundation by means of a generalized linear model assuming a binomial distribution, a logit link function and significance of the fixed effects habitat, trial, region and their interactions were tested with a Type 3 likelihood ratio (LR) test (Proc GENMOD, SAS v 9.4). For individuals that stayed submerged, the average time they spent under water was compared using a General Linear Model (Proc GLM, SAS v9.4) on log transformed data with significance of factors habitat, trial, region and their interaction being tested by means of a Type 3 sum of squares analysis. Best statistical models were selected by stepwise simplifying the models and using the Akaike Information Criterion (AIC).

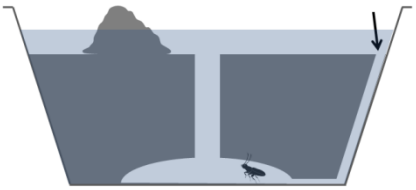


Figure 42. Experimental set-up of inundation experiment**.** Plaster was hollowed out at the bottom and this open space was accessible through a central corridor from the top to the bottom The arrow indicates a small groove for adding brackish water in a drop wise fashion to simulate inundation.

INUNDATION TOLERANCE

During the inundation experiment it was noted that in rare cases beetles entered a non-responsive state (i.e. hypoxic coma; Hoback and Stanley 2001; Pétillon et al. 2009). Therefore, a performance experiment was performed to compare (i) the beetles' entrance in a non-responsive comatose state when forced to long term inundations and (ii) their ability to recover from this state. We used *P. chalceus* beetles from the tidal Nieuwpoort and Guérande canal populations and from the seasonal Dudzele and Guérande pond populations. Individuals were placed separately in a test tube filled with brackish water. A parafilm piece was brought in the tube to prevent the beetles to reach the top of the tube to breathe and take air bubbles under their elytra. Tubes were incubated at a constant temperature of 10 °C. Individuals were monitored during a time span of one hour with intervals of 10 minutes. If the beetles did not respond by leg movement after tapping with the index finger on the tube, they were considered as non-responsive and the time of entering the hypoxic coma was noted. We compared the number of individuals that remained active during inundation by means of a generalized linear model assuming a binomial distribution, a logit link function and a Type 3 likelihood

ratio (LR) test (Proc GENMOD, SAS v 9.4) using time, habitat, region and their interactions as fixed effects.
The number of individuals that remained active during inundation was also compared per time frame (each 10 minutes) between the different populations from the Guérande (pond and canal) and Belgium (Dudzele and Nieuwpoort) using Fisher's exact test (SAS v9.4). After one hour the beetles were removed from the tubes and their recovery capacity was measured as the time until the beetles regained leg movements. The recovery time of individuals that had been in a hypoxic comatose state was related to the duration of the hypoxic comatose state, habitat, region and their interactions using Proc GLM (SAS v9.4) and a Type 3 sum of squares analysis. Best statistical models were selected using the Akaike Information Criterion (AIC). Additionally, it was investigated how long *P. chalceus* beetles can maximally endure a forced comatose state. For this experiment, several groups of Guérande canal and pond individuals were placed under water consecutively for 6, 12, 24, 48 and 72 hours and their recovery was recorded.

RESULTS

MATE PREFERENCE

In total, 74 mating trials were performed among beetles from the sympatric Guérande populations (Table 12). We found no significant interaction between male and female ecotype on the mating probability (Table 13). Female ecotype did also not significantly affect the probability of mating (Table 13). However, mating probability was significantly higher for males of the canal population compared to those sampled at the pond habitat (Table 13). We found no significant effect of male or female ecotype on mating time as well as no significant interaction between male and female ecotype (Table 13). Additionally, we found no significant effect of male ecotype or interaction between male and female ecotype on the time until mating occurred (Table 13). Females of the canal population accepted males more readily compared to those of the pond population (Table 12; Table 13).

Table 12. Proportion of mating occurrences, mean mating time and time until mating from the reciprocal mate preference experiment between Guérande canal and Guérande pond individuals.

Ecotype ♂	Ecotype ♀	Number of trials	Proportion of mating	Mean mating time (s)	Mean time until mating (s)
Canal	Canal	15	0.67	44.20 (± 49.35)	58.30 (± 70.79)
Canal	Pond	20	0.60	41.08 (± 21.33)	89.83 (± 52.11)
Pond	Canal	19	0.32	32.33 (± 16.50)	73.17 (± 41.83)
Pond	Pond	20	0.50	25.50 (± 15.77)	106.80 (± 60.50)

Table 13. Statistical comparison of the effect of male and female ecotype on mating occurrence, mating time and time until mating between Guérande canal and Guérande pond individuals.

	Mating occurrence		Mating time		Time until mating	
	X^2	P	$F_{1,34}$	P	$F_{1,34}$	P
Ecotype ♂	4.02	0.05	1.17	0.29	2.60	0.11
Ecotype ♀	0.56	0.45	0.00	1.00	4.53	0.04
Ecotype ♂ x Ecotype ♀	1.43	0.23	0.67	0.42	1.15	0.29

BEHAVIORAL RESPONSE TO INUNDATION

Beetles from Nieuwpoort, Dudzele, Guérande canal and pond habitats were subjected to inundation in three consecutive trials. The proportion of individuals that remained under water after inundation was significantly higher for the tidally inundated Nieuwpoort and Guérande canal populations compared to the seasonal Dudzele and Guérande pond populations (habitat: $\chi^2 = 66.49$; $P < 0.0001$; Figure 43). Additionally, the populations from the Guérande region had a significantly lower proportion of individuals that remained under water compared to the populations from Belgium, Nieuwpoort and Dudzele (region: $\chi^2 = 116.51$; $P < 0.0001$). We found no significant effect of trial on the proportion of individuals that remained under (trial: $\chi^2 = 1.71$; $P = 0.42$). However, trials had a significantly different effect between the two regions (region*trial: $\chi^2 = 11.57$; $P < 0.0031$).

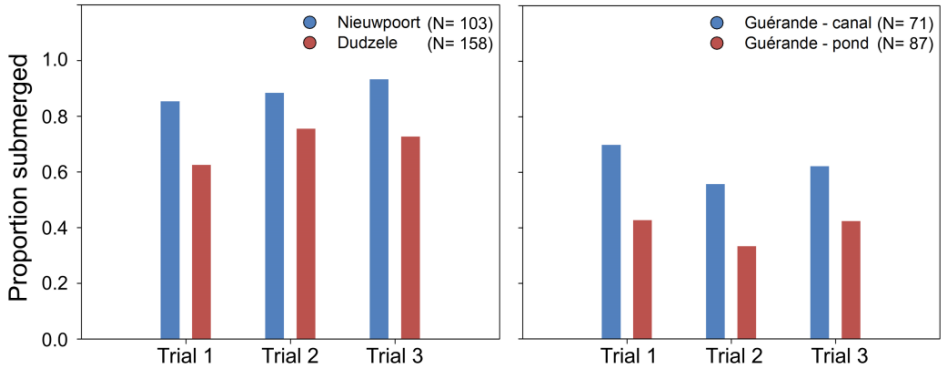


Figure 43. Proportion of *P. chalceus* beetles that remained submerged upon inundation for Nieuwpoort and Dudzele populations (left) and Guérande canal and pond populations (right).

Further, individuals from the tidal Nieuwpoort and Guérande canal populations that stayed submerged upon inundation, remained submerged significantly longer than individuals from seasonal Dudzele and Guérande pond populations (habitat: $F_{5,817}$ = 109.06; P < 0.0001; Figure 44). Again, the populations from the Guérande region had a significantly higher submergence time compared to populations from Belgium (region: $F_{5,817}$ = 5.56; P = 0.019). Moreover, differences in submergence time between habitats were significantly more pronounce between the Guérande canal and pond population compared to differences in submergence time the Nieuwpoort and Dudzele population (habitat*region: $F_{5,817}$ = 6.27; P = 0.013). Trial in which the time of submergence was measured did not significantly affect the time of submergence (trial: $F_{5,817}$ = 1.90; P = 0.15).

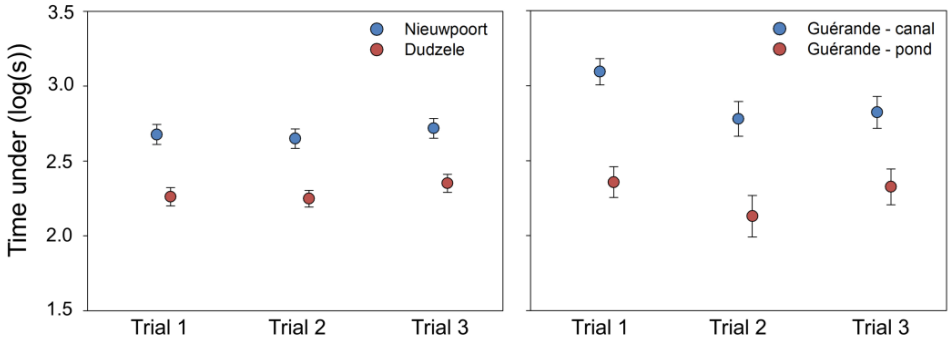


Figure 44. Duration of submergence for Nieuwpoort and Dudzele populations (left) and Guérande canal and pond populations (right). The tidal populations are indicated in blue, the seasonal populations in red. Error bars represent standard errors.

INUNDATION TOLERANCE

When *P. chalceus* beetles were forced to stay submerged, the average proportion of reactive individuals significantly differed between tidal and seasonal populations as submergence time progressed (habitat*time: $\chi^2 = 5.85$; $P < 0.016$; Figure 45). The average proportion of reactive individuals was significantly higher for individuals from tidal compared to individuals from seasonal habitats when both comparing Nieuwpoort and Dudzele and Guérande canal and pond populations (Figure 45). However, for the Guérande populations, significant differences were only observed after one hour of forced inundation. Hence, the average proportion of reactive individuals within ecotypes significantly differed between regions (habitat*region: $\chi^2 = 4.22$; $P < 0.04$). Further, we found a significantly different response on submergence time between regions (region*time: $\chi^2 = 7.11$; $P < 0.008$).

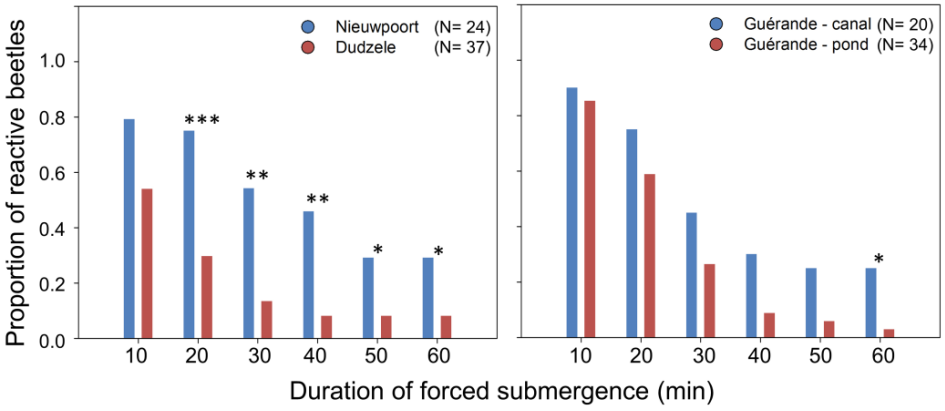


Figure 45. Comparison of *P. chalceus* inundation tolerance of Nieuwpoort, Dudzele (left) and Guérande canal and pond (right) populations. P-values were calculated with a Fisher's exact test. *: P-value < 0.05; **: P-value < 0.01; ***: P-value < 0.001.

The recovery of individuals that were in a hypoxic coma correlated strongly with the duration of the coma (Figure 46). Beetles that were in a hypoxic coma for a longer time, took longer to recover than beetles that had been in a shorter comatose state (coma time: $F_{3,89} = 26.61$; $P < 0.0001$). This relationship significantly differed between individuals from tidal and seasonal habitats (coma time*habitat: $F_{3,89} = 6.63$; $P = 0.012$). That is, individuals from seasonal habitats needed increasingly more time to recover compared to individuals from tidal habitats. Recovery times did not significantly differ between regions.

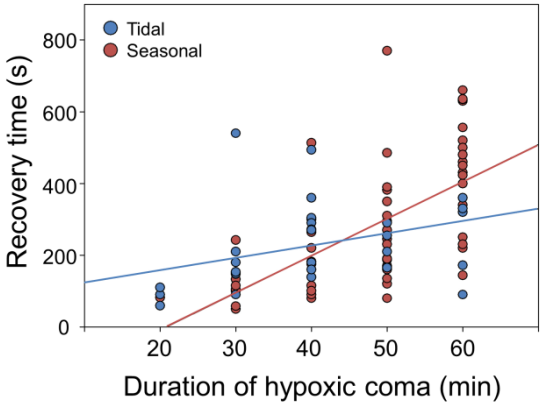


Figure 46. Recovery time compared to the duration of the hypoxic coma. The blue and red lines indicate the significantly different regression lines of the tidal and seasonal populations respectively.

Only after an extensive forced submergence time of 48 hours, some *P. chalceus* beetles did not recover from the hypoxic coma (Figure 47). The tested Guérande pond individuals seemed to be more resistent to forced inundation as seen by the higher proportion of recovery after 48 and 72 hours of inundation.

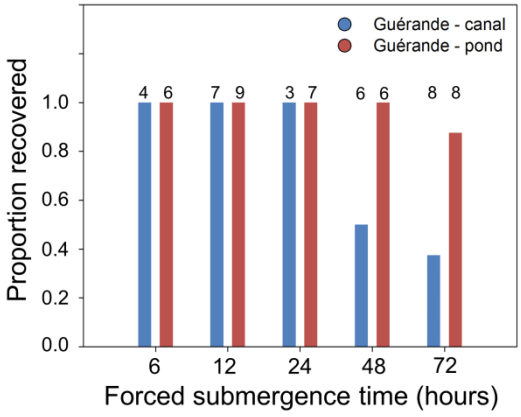


Figure 47. Proportion of *P. chalceus* beetles from the Guérande populations recovering after forced inundation. Numbers above bars indicate the number of individuals tested.

DISCUSSION

In this study, we show (i) little or no indication of assortative mating through mate preferences, but (ii) strong indications of a different response to environmental cues (i.e. inundation); beetles from seasonally flooded marshes showed significantly more escape behavior when exposed to inundations. This different behavioral response is likely to result in spatial sorting and promote the association of *P. chalceus* populations with their habitat. Hence, these latter findings provide a potential mechanism for reproductive isolation between *P. chalceus* populations.

Our mating experiments revealed that when individuals are exposed to the same environmental conditions, individuals belonging to the same ecotype did not prefer to mate with each other compared to individuals sampled at different habitats. Such patterns have been observed, for instance, in other arthropod species that mate assortatively by size (Crespi 1989). The long -and short-winged *P. chalceus* populations generally also significantly differ in body size (Desender 1989a, Dhuyvetter *et al.* 2007, Van Belleghem & Hendrickx 2014). However, we found no significant indication of mate preference based on ecotype in *P. chalceus*. Another mechanism that can lead to assortative mating is allochrony resulting from differences in developmental timing or phenology in response to different environments. We did find significant differences in the willingness to pair between Guérande canal and pond males. These differences might result from phenological differences as different environmental conditions between the habitats (Dhuyvetter *et al.* 2007) could result in a slight offset of the reproductive period. This offset has not been investigated in the field, but the reproductive period of the ecotypes clearly overlaps (personal observation). Further, physical difficulties or mechanical incompatibilities may result in a reduction of the probability or duration of mating. Possible mechanical incompatibilities are also thought to be negligible in *P. chalceus* as crosses of both ecotypes are easily obtained in the lab, which result in intermediate phenotypes when considering wing and body size (Van Belleghem & Hendrickx 2014). Hence, given the absence of beetles' discrimination according to ecotype, we argue that mate preferences and mechanical incompatibilities are unlikely to be sufficient for the persistence of these ecotypes in sympatric settings.

Alternatively, our results demonstrate that, apart from the previously demonstrated genetic divergence in wing size and *mtIdh* alleles, the investigated tidal and seasonal populations strongly differ in behavioral responses to the different hydrological dynamics present in both habitats. Short-winged *P. chalceus* populations from tidally

inundated habitats (i.e. Nieuwpoort and Guérande canals) are significantly less reluctant to inundation compared to beetles from seasonal habitats (i.e. Dudzele and Guérande ponds), which are usually inundated for longer periods. Oppositely, when confronted with inundation, beetles from the seasonal habitats showed a significantly higher probability to avoid submergence, both in the Dudzele as well as in the Guérande pond populations. Moreover, individuals from the seasonal Dudzele and Guérande pond populations spend significantly less time submerged than beetles from, respectively, the Nieuwpoort and Guérande canal populations. These effects were consistent over all consecutive trials. We did find significant differences in responses between the Belgian and France populations. However, differences between tidal and seasonal habitats were in the same direction in both regions.

After prolonged submergence, beetles go into a hypoxic coma. This kind of non-responsive state is interpreted as an adaptation in salt-marsh insects to cope with repeated and high frequency inundation (Pétillon *et al.* 2009). Beetles from the seasonal marshes seem to be more susceptive to a hypoxic coma as is indicated by the significantly shorter time needed for pond individuals to go into a non-responsive state. These differences in inundation tolerance possibly result from differences in metabolic activity between the long -and short-winged beetles from, respectively, seasonal and tidal marshes (Mueller & Diamond 2001, Haag *et al.* 2005). In contrast, beetles from seasonal habitats seemed to survive prolonged inundations to a higher rate compared to beetles from tidal marshes. Only after 72h a marked proportion of individuals did not recover from the hypoxic coma, probably due to the accumulation of anaerobic end products (e.g. lactate and alanine) (Hoback & Stanley 2001). This could be an adaptation to prolonged inundations. However, the higher susceptibility to a hypoxic coma of seasonal individuals may increase the necessity of these beetles to avoid submergence.

Most important, the different responses of both ecotypes can be interpreted as an adaptation towards the different hydrological regimes that simultaneously and directly acts as a reproductive barrier through spatial sorting of individuals in their respective habitat. This process wherein an adaptive trait results in non-random habitat use and, therefore, assortative mating among individuals adapted to the same habitat has been referred to as *matching habitat choice* (Edelaar *et al.* 2008, Edelaar & Bolnick 2012). Indeed, escape behavior during inundations can be considered to be an adaptive response in habitats that are unpredictably flooded for several months, as is the case for seasonal marshes (i.e. pond habitat in the Guérande). Hence, when these beetles end up in nearby tidal marshes, dispersal out of these habitats is expected during tidal floods. Alternatively, remaining submerged during floods is likely adaptive in tidal habitats (i.e. canal habitat in the Guérande) as beetles do not face predation risks by repeated escape behavior during these frequent and short inundations that can easily be survived as

shown by our experiments. This absence of escape behavior is in contrast highly detrimental during long term inundations.

This mechanism, wherein a single phenotypic trait (i.e. escape behavior) involved in adaptation towards these habitats can readily and directly result in spatial separation, could be a key mechanism promoting reproductive isolation in this system. Furthermore, this behavioral differences between individuals from seasonal and tidal marshes may result in similar patterns of habitat preference as described, for instance, in phytophagous insects (Fry 2003). In these phytophagous insects, different preference alleles are argued to be associated with alleles that are oppositely selected on different hosts resulting in reproductive isolation between these races (Hawthorne & Via 2001). However, in contrast to the mechanism presented in phytophagous insects, the mechanism presented here suggests a simple and direct link between divergent selection (hydrology) and spatial sorting (i.e. escape behavior). From an ecological perspective, this trait could, hence, be classified as an automatic magic trait in which a trait affects both performance and assortative mating (Servedio *et al.* 2011). However, the genetic basis of this trait is unknown and may be affected by maternal effects as well as habitat experience during larval or early life stages.
In conclusion, our study showed clear behavioral differences to inundation between *P. chalceus* populations. These behavioral differences are argued to play an important role as a mechanism resulting in spatial sorting and reproductive isolation, which provides an explanation for the persistence of distinct ecotypes in sympatric settings.

ACKNOWLEDGEMENTS

We gratefully thank Lut Van Nieuwenhuyse, Marc Van Kerckvoorde, Viki Vandomme and Carl Vangestel for their help in sampling of the beetles.

CHAPTER 7

SYMPATRIC SPECIATION BY MEANS OF NATAL HABITAT PREFERENCE?

Steven M. Van Belleghem [1,2]
Katrien De Wolf [1,2]
Frederik Hendrickx [1,2]

[1] Terrestrial Ecology Unit, Biology Department, Ghent University, K. L. Ledeganckstraat 35, 9000 Gent, Belgium

[2] Department Entomology, Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000 Brussel, Belgium

ABSTRACT

Genetic habitat preferences can play an important role as a reproductive barrier between sympatric ecotypes. However, a possibility that has received much less attention is that individuals may develop a preference for the habitat they experienced at a young age. Such natal habitat preference may easily result in spatial sorting of individuals that experienced different environmental conditions and, hence, quickly lead to assortative mating within habitats. Ecotypes of the ground beetle *Pogonus chalceus* likely mate assortatively through spatial sorting by habitat preference, which results from alternative dispersal responses to environmental (inundation) cues present in tidal and seasonal habitats. In this study, *P. chalceus* larvae and pupa from Guérande canal and pond populations were reared under different hydrological regimes, simulating environmental differences between tidal and seasonal habitats. Adults of both ecotypes that were exposed to frequent but short inundations showed a significantly lower response to escape these inundations in the adult stage. However, these effects changed after consecutive exposure to inundation. Besides, we also found strong indications that responses to inundation have a genetic component as seen by the consistently higher water reluctance of seasonal beetles. These indications of natal habitat experience may have important consequences for both the evolution and persistence of sympatric races in *P. chalceus*.

INTRODUCTION

When disruptive selection favors ecological specialization on two different habitats, habitat preference can be an important mechanism for the occurrence of sympatric speciation (i.e. speciation in the absence of a geographical barrier). If sympatric speciation occurs via the evolution of a genetic habitat preference, the genes involved in habitat preference need to become associated with genes that affect performance in the habitats (Dieckmann & Doebeli 1999, Fry 2003, Beltman & Metz 2005, Thibert-Plante & Gavrilets 2013). This association can result from the evolution of linkage disequilibrium of the assortative mating and performance genes (Dieckmann & Doebeli 1999), from physical linkage of the assortative mating and performance genes (Hawthorne & Via 2001) or through pleiotropy in which one trait affects both assortative mating and performance (Servedio *et al.* 2011). Alternatively, a largely neglected mechanism in the emergence of an association between performance and habitat choice is through the effect of *natal habitat preferences* (Davis & Stamps 2004, Beltman & Metz 2005). Here,

individuals develop a preference for the habitat that they have experienced as immatures, which most often constitutes the habitat wherein females oviposit. Accidental or forced niche shifts will then result in assortative mating of individuals raised in the same habitat as well as preference to produce offspring in this habitat. Two hypotheses address the possible adaptive significance of natal habitat preference (Stamps 2001). First, the *habitat cuing* hypothesis suggests that dispersing individuals can quickly and efficiently locate suitable habitat based on stimuli comparable to those in the natal habitat. Alternatively, the *preference-performance* hypothesis suggests that it is beneficial to select a habitat of the same type as the natal habitat to which the phenotypes are likely adapted.

Speciation through the evolution of a natal habitat preference is expected to be an effective mechanism because it may represent a one-allele mechanism (Felsenstein 1981). That is, alleles that affect learning have the same effect in both habitats, hence, resolving the problem of evolving an association between assortative mating and performance genes. In contrast, speciation through the evolution of a genetic habitat preference is a two-allele mechanism in which different alleles promote assortative mating among different populations (Felsenstein 1981). Simulation models have shown that natal habitat preferences can quickly and easily lead to the formation of sympatric races (Beltman *et al.* 2004, Beltman & Haccou 2005, Beltman & Metz 2005, Ravigné *et al.* 2009). Interestingly, this theoretical mechanism easily explains the colonization of new niches, and is thus of particular relevance in the context of eco-evolutionary mechanisms of adaptation to novel habitats. Despite that the ecological implications have long been recognized (Maynard Smith 1966, Immelmann 1975), relatively few empirical studies have observed natal habitat preferences (reviewed in Davis and Stamps 2004) and tested how effective and plausible this mechanism is for speciation in natural systems. One important example of natal habitat preference involved in speciation includes imprinting on hosts in brood parasitic indigobirds (Sorenson *et al.* 2003). As adults, male indigobirds mimic host song, whereas females imprint on these songs to choose both their mates and the nests they parasitize (Payne *et al.* 2000). It is found that these behavioral mechanisms promote the association of indigobird populations with a given host species, and provide a mechanism for reproductive isolation after a new host is colonized. Scarce other examples in which natal habitat learning has been found include odor learning in parasitoid wasps (Smith & Cornellt 1979, Cortesero *et al.* 1995, Storeck *et al.* 2000, Kaiser *et al.* 2003). However, in the latter examples it is unclear whether this mechanism may be involved in maintaining the association of races with their hosts (Smith & Cornellt 1979).

We previously demonstrated the existence and importance of spatial sorting resulting from disruptive selection on traits that affect dispersal upon inundation in the

sympatrically diverging beetle *Pogonus chalceus* (Chapter 6). This beetle inhabits two distinct types of marshes; tidal and seasonal marshes. *Tidal marshes* are year-round flooded on a regular basis, but for short periods of at maximum a few hours only. *Seasonal marshes* are disconnected from the sea and are permanently inundated for extensive time periods. Tidal and seasonal marshes are generally inhabited by, respectively, short- and long-winged *P. chalceus* beetles (Van Belleghem & Hendrickx 2014). In some areas, such as the Guérande salterns in France, both habitats are found in multiple replicates only 10-20 m apart (Dhuyvetter *et al.* 2007). In these salterns, *ponds* are used to evaporate water and concentrate salt and resemble seasonal marshes in that they are flooded irregularly for extensive periods. *Canals*, on the other hand, are used to bring water to the ponds and are subject to the tides. The different hydrological regimes present in the habitats likely serve as cues to which populations from tidal and seasonal marshes respond differently. More precisely, individuals from tidal marshes tend to stay submerged more frequently and longer compared to individuals from seasonal marshes which tend to be more reluctant to submergence (Chapter 6). These behavioral differences likely result from disruptive selection and, moreover, are expected to result in spatial sorting and assortative mating in the respective habitats and, hence, likely are an important mechanism for the preservation of distinct *P. chalceus* ecotypes in sympatric settings.

In this chapter, we shortly explore the possibility of natal habitat experience affecting beetles' response to inundation. If significant effects of natal habitat experience on traits involved in spatial sorting (and assortative mating) exist in *P. chalceus*, this may be an important factor influencing repeated colonization of new and distinct habitats and local adaptation.

MATERIALS & METHODS

REARING AND EXPOSURE TO DIFFERENT HYDROLOGICAL REGIMES

Eggs from adults sampled in the field in both tidal (*canal* habitat) and seasonal (*pond* habitat) marshes at the Guérande were raised in the lab and emerging larvae and pupae were reared in a common-garden set-up simulating the exposure to different hydrological regimes during development. One group of pond and one group of canal individuals were not exposed to inundation during their development, while a second set of pond and canal individuals experienced inundation for 30 min each two days

during larval and pupal development. If more than one egg was raised from the same female, the progeny was evenly divided over the different exposure experiments. After eclosion these adult beetles were tested for their response towards inundation.

INUNDATION EXPERIMENT

To simulate inundation, beetles were kept in a sealed plastic cup containing plaster. The plaster was hollowed out at the bottom and this open space was accessible through a central corridor from the top to the bottom (Figure 42 in Chapter 6). At the side there was a small groove which allowed adding brackish water in a drop wise fashion to simulate inundation. Adding water through the groove insured that no air bubbles would persist under the plaster in which the beetles could reside after the flooding. During the simulated inundation event, the behaviour of the beetles was observed and timed using a stopwatch. We noted the following behaviours: (i) presence or absence of escape behaviour upon flooding and (ii) time until beetles reach the surface for taking air. For each test, three trials were performed, separated by approximately one week and using the same beetles.

Differences in behavioural response were compared between Guérande canal and pond ecotypes that were subjected to different hydrological regimes during larval and pupal development. We compared the proportion of individuals that expressed immediate escape behavior compared to the individuals that remained submerged upon inundation by means of a generalized linear model assuming a binomial distribution and a logit link function (Proc GENMOD, SAS v 9.4) and tested the significance of the factors habitat (pond versus canal), larval environment (inundated versus dry), trial and their interactions as fixed effects by means of a likelihood ratio test. For individuals that stayed submerged, the average time they spent under water was compared using a General Linear Model (Proc GLM, SAS v9.4) on log transformed data with a Type 3 sums of squares analysis.

RESULTS

In total, we raised 15 (from 11 females) offspring from Guérande canal beetles under dry conditions and 13 (from 9 females) under frequently inundated conditions. Additionally, offspring from 19 (from 14 females) Guérande pond beetles were raised under dry conditions and 18 (from 14 females) under frequently inundated conditions. Ecotype significantly affected the proportion of beetles that remained submerged upon inundation (habitat: $\chi^2 = 10.35$; $P = 0.001$). More precisely, individuals from the tidal canal habitat consistently had a higher probability of remaining submerged during inundations (Figure 48). Considering all three trials, we did not find a consistent effect of treatment on the proportion of beetles that remained submerged upon inundation. In contrast, we found that the effect of treatment significantly differed between consecutive trials (treatment*trial: $\chi^2 = 9.39$; $P = 0.009$). Moreover, considering only the first trial, we did find a significant effect of treatment on the proportion of beetles that remained submerged (treatment trial 1: $\chi^2 = 5.73$; $P = 0.017$). More precisely, in the first trial, beetles raised in an inundated environment had a significantly lower propensity to escape inundations in the first trial. Also, in the first trial, no significant habitat effect was observed. In the second trial, the effect of treatment was also significant, but reversed (treatment trial 2: $\chi^2 = 7.85$; $P = 0.005$). Further, the canal and the pond ecotypes did significantly differ in their response in the second trial (habitat: $\chi^2 = 9.42$; $P = 0.002$). Moreover, the canal and pond populations responded significantly different to the treatment (treatment trial 2*habitat: $\chi^2 = 5.12$; $P = 0.024$). Finally, in the third trial, we only found a significant effect of habitat on the proportion of beetles that remained submerged (habitat: $\chi^2 = 5.12$; $P = 0.024$).

The time individuals remained submerged significantly differed according to the population of origin, with canal individuals remaining submerged for a significantly longer time (habitat: $F_{7,159} = 51.99$; $P < 0.0001$; Figure 49). Interestingly, treatment significantly affected submergence time, with individuals raised in a frequently inundated environment staying submerged for a longer time (treatment: $F_{7,159} = 4.02$; $P = 0.047$). This effect was consistent over all three trials (trial*treatment = NS). Further, treatment did not affect populations differently (treatment*habitat = NS). Finally, the submergence time of populations differed significantly between trials (trial*habitat: $F_{7,159} = 3.63$; $P = 0.029$).

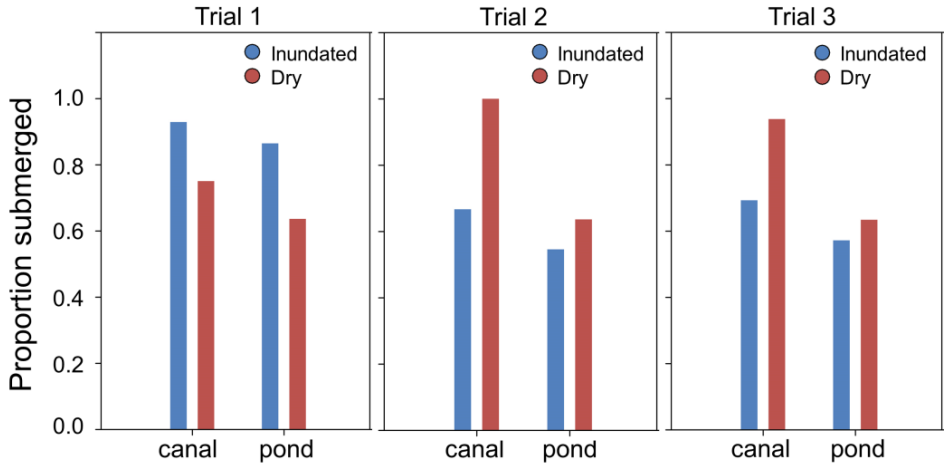


Figure 48. Proportion of *P. chalceus* beetles from Guérande canal (tidal) and pond (seasonal) habitat raised under different hydrological regimes that remained submerged upon inundation.

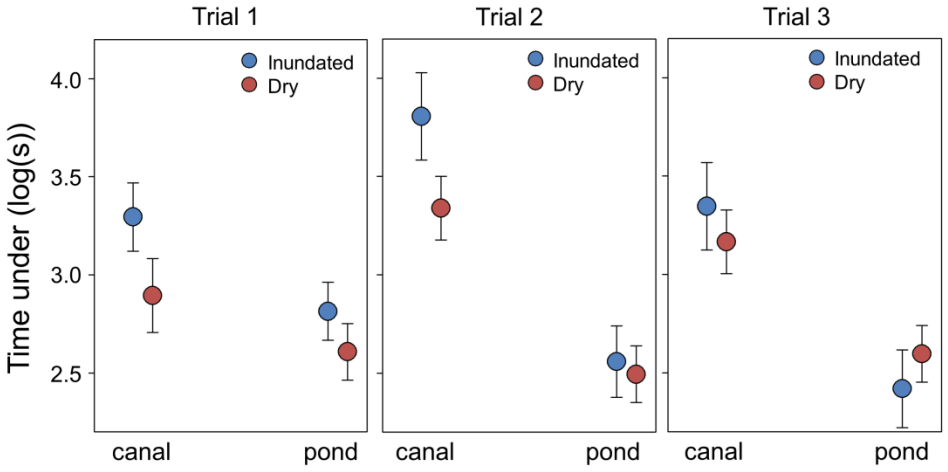


Figure 49. Duration of submergence of *P. chalceus* beetles raised under different hydrological regimes that remained submerged upon inundation. Error bars represent standard errors.

DISCUSSION

Our experiments established that (i) natal habitat experience may influence differences in response to inundation between tidal canal and seasonal pond populations, but that (ii) there is likely also a genetic component as seen by the consistently higher water reluctance of beetles from seasonal habitats.

Rearing the larvae and pupae in a frequently inundated versus a non-inundated environment significantly affected the adult beetles' response to inundation, both for individuals from tidal canal and seasonal pond habitats. Further, the effect of treatment significantly differed between consecutive trials. During the first exposure to inundation, beetles that were subjected to frequent inundation tended to stay submerged more frequently and for a longer time period, both for individuals from tidal canal and seasonal pond habitats. However, in the subsequent trial this pattern was reversed and disappeared in the final trial. Changes in the effect of natal experience have also been found in the parasitoid wasp *Leptopilina boulardi* (Kaiser *et al.* 2003). This species has been found to condition on odor when looking for *Drosophila* larvae to oviposit in. Furthermore, it has been shown that the conditioning of an association between odor and ovipositioning strongly depends on the number of conditioning trials and decreases during consecutive testing trials. Altogether, this results in highly plastic responses to cues in this wasp species. Additionally, the habitat of origin significantly affected both the proportion of *P. chalceus* beetles that stayed submerged upon inundation as well as the time they stayed submerged. As previously found, individuals from the seasonal pond habitat escape inundations with higher probability compared to tidal canal individuals (Chapter 6). This suggests a genetic component affecting response to inundation. However, maternal effects have not been controlled for in this study.

Determining the importance of natal habitat experience for the initial colonization and speciation event is a daunting task. When individuals initially colonize a new habitat, learning processes may be the most important factor keeping populations in their respective habitats. However, as time progresses and populations adapt, these learning processes may become accommodated by genetic differences which strengthen the association of the adapted phenotypes with their environments (West-eberhard 2003). This may partly explain why few examples exist that link this mechanism to speciation despite that natal habitat preference could strongly facilitate phenotype-habitat matching.

Moreover, it has been argued that the intensity of selection on learned preferences is lower than on genetic preferences (Beltman & Metz 2005). This can be understood by

considering individuals that are adapted to habitat A ending up in habitat B to which they are less adapted. In case of learned habitat preferences, the offspring of these dispersers, adapted to habitat A, will be more likely to stay and produce their offspring in habitat B (the 'wrong' habitat). However, the offspring of the dispersers would do best by producing their offspring in habitat A, but this is only achieved when they would have a genetic habitat preference. In contrast, genetic preferences necessitate the evolution of an association between preference and performance genes. Interestingly, in pea aphids, early experience is known not to influence habitat preference (Via 1991) and genes for habitat choice and ecological adaptation appeared to be closely linked or influenced by the same gene (Hawthorne & Via 2001).

In conclusion, finding significant effects of natal environment on response to inundation in *P. chalceus* despite the relatively low sample number implies that this effect is substantial and may be of particular importance in *P. chalceus* during colonization and subsequent differentiation.

ACKNOWLEDGEMENTS

We gratefully thank Viki Vandomme and Carl Vangestel for their help in sampling of the beetles in the Guérande salterns.

GENERAL DISCUSSION

Steven M. Van Belleghem

Investigating the genetic and behavioral aspects of population differentiation in the ground beetle *Pogonus chalceus* led to the identification of several factors and mechanism that have important implication for understanding the evolution of adaptation and ultimately the origin of reproductively isolated populations in this beetle species. The peculiarity of this study system resulted from finding differentiation in multiple traits (i.e. wing and body size and allozymes of the mtIDH protein) between populations that are not separated by any geographical barriers. Moreover, we found that wing size is strongly genetically determined in these sympatric populations and that genes involved in wing size are likely not genetically linked to the *mtIdh* locus (CHAPTER 1). Additionally, by studying the genome-wide pattern of differentiation (CHAPTER 5), we identified a genomically widespread set of unlinked outlier loci associated with ecotypic differentiation. This strongly invokes the selection-migration antagonism because recombination is expected to result in maladapted gene combinations when gene flow is high. Therefore, we looked for ecological and genetic factors that may facilitate sympatric divergence and/or reduce gene flow or recombination. First, we found that the differentiation in the *mtIdh* locus has a single origin (CHAPTER 2). This singular evolution may have provided building material for rapid and recurrent sympatric divergence. However, understanding the evolutionary history of this locus in high detail necessitates detailed knowledge of historic rates of gene flow, population size and selection strengths. Next, most of the identified outlier SNPs as well as the *mtIdh* allele associated with short-winged populations from tidal habitats have reduced nucleotide diversity (π) compared to SNPs associated with the long-winged populations form seasonal habitats, indicating a relatively recent increase to high frequencies and a similar evolutionary histories of the SNPs associated with the tidal habitats (CHAPTER 2, CHAPTER 5). Finally, by studying behavioral variation in the response to inundation, we identified a mechanism that may result in spatial sorting and, hence, assortative mating and reproductive isolation between individuals from differently selected populations (CHAPTER 6-7). Here, I discuss implications of these findings for the interpretation of the *P. chalceus* evolutionary system, delineate the evolutionary processes and mechanisms involved in the differentiation and persistence of *P. chalceus* ecotypes and, finally, discuss future prospects.

GENOMICALLY WIDESPREAD ADAPTATION IN *P. CHALCEUS*

Ecotypic divergence in *P. chalceus* is associated with local adaptation in multiple genetic traits. First, Wing size and allozymes of the mtIDH protein were previously found to be strongly correlated with habitat stability (Dhuyvetter *et al.* 2004). These traits are likely associated with hydrological dynamics in the habitats (i.e. tidal versus seasonal marshes). Additionally, we demonstrated that for a sympatric setting, wing size has a high heritability, comparable to heritability estimates found in allopatric populations, and that genes involved in wing size are likely not closely linked to the adaptation associated with the *mtIdh* locus (CHAPTER 1). Next, we demonstrated that local adaptation also resulted in behavioral differences in response to inundation between the habitats (CHAPTER 6). These differences likely have a genetic component as raising beetles in common environments still express behavioral differences according to their original habitat. Finally, using RAD tag sequencing, we identified multiple unlinked loci that are strongly differentiated between sympatric populations (CHAPTER 5). Therefore, altogether ecotypic divergence in *P. chalceus* seems to be genomically widespread, even between sympatric populations. However, it is argued that adaptive divergence in multiple loci between sympatric populations suffers from the antagonism between recombination and selection (Felsenstein 1981). Therefore, models of speciation in sympatry usually make stringent assumptions (Kirkpatrick & Ravigné 2002), which may potentially result in widespread genomic differentiation if they result in a reduction of gene flow and/or recombination. These assumptions include a source of disruptive selection, an isolating mechanism and a link between disruptive selection and the isolating mechanism. Furthermore, the genetic basis of the traits affected by disruptive selection and/or the traits involved in reproductive isolation has important implications for the evolution of reproductive isolation (GENERAL INTRODUCTION). In the following section, the possible effects of these assumptions are discussed and related to sympatric divergence in *P. chalceus*.

MECHANISMS MAINTAINING DIVERGENCE IN SYMPATRY

Here, I discuss mechanisms that may affect the evolution and persistence of population differentiation in *P. chalceus* despite the ample opportunity of gene flow between the ecotypically diverged populations (Figure 50).

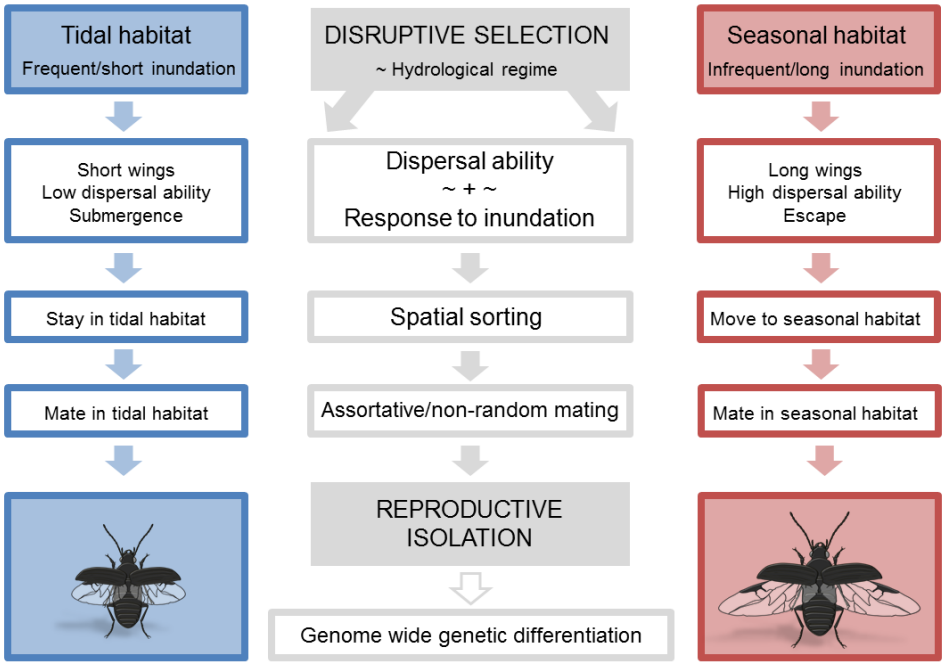


Figure 50. Overview of the evolutionary processes and mechanisms involved in the differentiation and persistence of *P. chalceus* ecotypes.

SOURCE OF DISRUPTIVE SELECTION Differences in hydrological dynamics most likely provide a major source of disruptive selection. Tidal and seasonal habitats differ strongly in the duration of inundation, with tidal habitats flooded frequently for short periods (5-6h) and seasonal habitats flooded for extensive periods (months). Remaining submerged during floods is likely adaptive in tidal habitats as beetles do not face predation and mortality risk by repeated escape behavior during these frequent short inundations that can easily be survived as shown by our experiments (CHAPTER 6). The absence of escape behavior is in contrast highly detrimental during long term inundation in seasonal marshes. Hence, these differences can be expected to exert strong opposing selection pressures, with seasonal habitats selecting for dispersal and a resident strategy selected in tidal marshes. Consequently, differences in hydrological regime have resulted in adaptive differences between beetles from tidal and seasonal marshes. One set of adaptations includes performance in the habitats, with a higher susceptibility to a hypoxic coma and longer recovery time in beetles from seasonal habitats (CHAPTER 6). Additionally, the differences in hydrological regime likely also select variation in

dispersal ability as beetles are forced to disperse from the seasonal habitats. Interestingly, this adaptation also includes different dispersal behavior to a similar cue (CHAPTER 6). More precisely, *P. chalceus* beetles from tidal marshes are less reluctant to inundation and can survive the inundations staying submerged. This is also observed in the field when tides rise and quickly flood large areas. The beetles stay buried in the sand and most likely use air bubbles under their elytra or stay in small air pockets (e.g. under shell particles) during the inundation (personal observation). Alternatively, inundation in seasonal marshes serves as a cue to escape these detrimental flooding conditions.

ISOLATING MECHANISM Reproductive isolation likely results from spatial sorting in *P. chalceus*. We demonstrated that beetles from seasonal habitats tend to escape inundation events. This is expected to result in spatial sorting as these beetles from seasonal habitats should avoid habitats that are frequently inundated whereas beetles from tidal habitats remain in these habitats. Consequently, the spatial sorting is expected to result in assortative mating and, consequently, in a reduction of gene flow (i.e. reproductive isolation). This reduction in gene flow is expected to be strong, as for instance all 48 individuals from the canal and pond population sampled for the RAD tag sequencing were assigned to a genetic cluster matching their habitat (CHAPTER 5). Furthermore, previous sampling in the Guérande canal and pond habitats resulted in only 1% – 1.5% out of 404 individuals found in the wrong habitat (based on wing size and mtIDH genotype; Dhuyvetter et al. 2007). Hence, this mechanism may be expected to aid disruptive selection in maintaining divergence in multiple genetically unlinked loci, while gene flow may be high enough to swamp or mix neutral variation. In contrast, assortative mating resulting from mate preference so that each morph mates preferentially with similar individuals (e.g. based on body size; Conde-Padín et al. 2008) has not been found in *P. chalceus*. Mating in *P. chalceus* seems little discriminative and most likely takes place within their habitat (CHAPTER 6).

LINK BETWEEN DISRUPTIVE SELECTION AND ISOLATING MECHANISM The link between disruptive selection and reproductive isolation is obvious in *P. chalceus* as selection for alternative behaviors results in spatial sorting and, consequently, assortative mating within the habitats and partial reproductive isolation. Hence, we argue that the effect of selection is related directly to the isolating mechanism. More precisely, genes that affect habitat choice also directly influence assortative mating in *P. chalceus.* Therefore, reproductive isolation may be evolving automatically as a result of selection to the divergent environments. Such traits are referred to as *automatic magic traits* (Servedio *et al.* 2011). Moreover, besides a genetic component, we also demonstrated significant

effects of the habitat experienced by the larval stages on the adult behavior (CHAPTER 7). Such natal habitat preferences may easily result in the evolution of reproductive isolation between differently selected populations (Davis & Stamps 2004, Beltman & Metz 2005).

GENETIC BASIS It is a comprehensive task to determine the genetic basis that links selection to assortative mating. Most importantly, as reproductive isolation progresses, populations will diverge in multiple loci (Wu 2001b, Hendry *et al.* 2009). Hence, as populations differentiate, finding the loci that are directly involved in reproductive isolation becomes increasingly difficult from genomic comparisons. Indeed, we found multiple diverged loci in *P. chalceus*, but it is difficult to infer whether these directly influence reproductive isolation, or, conversely, whether reproductive isolation allows these multiple genomic regions to diverge. Quantitative Trait Loci (QTL) mapping using second generation (F2) crosses of individuals from contrasting habitats might help to identify the genetic basis of reproductive isolation (Kronforst *et al.* 2006, Nosil & Schluter 2011, Van Ooijen & Jansen 2013). Nevertheless, it is reasonable to suggest that genomic architectures that suppress recombination are expected to facilitate coupling of genetic adaptations and, hence, speciation in *P. chalceus*. If natal habitat experience has a major influence on adult habitat preferences, the same allele may increase isolation in both habitat types. This could then potentially be classified as a *one-locus, one-allele* mechanism, which very easily solves the selection-recombination antagonism (Felsenstein 1981). However, one-allele mechanisms are difficult to detect as they do not leave a population-specific genetic signature in the genome at the primary isolation locus (Seehausen *et al.* 2014). Only if they arise during speciation they would be detectable as sweeps that are shared by both diverging populations (Seehausen *et al.* 2014). To date, this has not been detected in any case to our knowledge. Alternatively, if habitat preference (~ reluctance to inundation) is strictly determined by different alleles in both populations (i.e. *one-locus, two-alleles* mechanism), recombination is expected to break up the association between the habitat preference locus and performance genes and, therefore, close genetic linkage might be expected. However, it may also be argued that when *multiple loci* are involved in habitat preference and adaptation, the chance of correlation between several of these traits increases, which can result in an association between habitat preference and habitat performance, leading to the onset of sympatric adaptation and speciation (Feder *et al.* 2012b, a).

P. CHALCEUS ALONG THE SPECIATION CONTINUUM

It is difficult to infer whether the distinct ecotypes represent a stable outcome versus an intermediate state that is bound to progress to complete reproductive isolation. Important elements for speciation in sympatry are present in *P. chalceus*. More precisely, incipient reproductive isolation is evident and likely allows populations to diverge in multiple genetically unlinked traits. However, given the incipient stage of speciation in *P. chalceus*, intrinsic postzygotic barriers (genetic incompatibilities that are independent of the environment) have likely not evolved as they usually evolve later in the speciation process (Seehausen *et al.* 2014). Therefore, isolating barriers depend mostly on disruptive selection (i.e. selecting against intermediate phenotypes) and a mechanism resulting in assortative mating in *P. chalceus* and the absence of intrinsic postzygotic barriers would allow high rates of hybridization in the absence of disruptive selection.

EVOLUTION OF *P. CHALCEUS* ECOTYPES

The evolution of distinct ecotypes despite gene flow has been found in multiple species (e.g. Johannesson et al. 2010; Cristescu et al. 2012; Drotz et al. 2012; Butlin et al. 2013). In these study systems, conflicting patterns in neutral and selected genes can either be explained by initial divergence in *allopatry followed by secondary overlap* and extensive introgression that homogenizes neutral differences evolved under allopatry, or by *repeated evolution in sympatry*, with the same ecotypes appearing in each local site (Faria *et al.* 2014). From a genetic perspective, repeated evolution of ecotypes can result from *separate parallel mutations* in the (i) same gene or (ii) different genes, or may have a *single origin* and result from (iii) shared ancestral polymorphisms (i.e. standing genetic variation) or (iv) introgression in which a new positive mutation spreads rapidly among populations inhabiting similar habitats (Johannesson *et al.* 2010). Considering the evolution of the long- and short-winged ecotypes found in *P. chalceus*, these scenarios can be described as follows: (i) a single and allopatric origin of the short-winged ecotype in tidal marshes that spread along the Atlantic coasts and hybridized with the long-winged ecotype (Figure 51A), (ii) different adaptive traits evolved in different short-winged tidal populations and rapidly introgressed into other tidal populations (Figure 51B), and (iii) the short-winged ecotype has evolved repeatedly in each location (Figure 51C). Notably, different genes and adaptations may correspond to different scenarios.

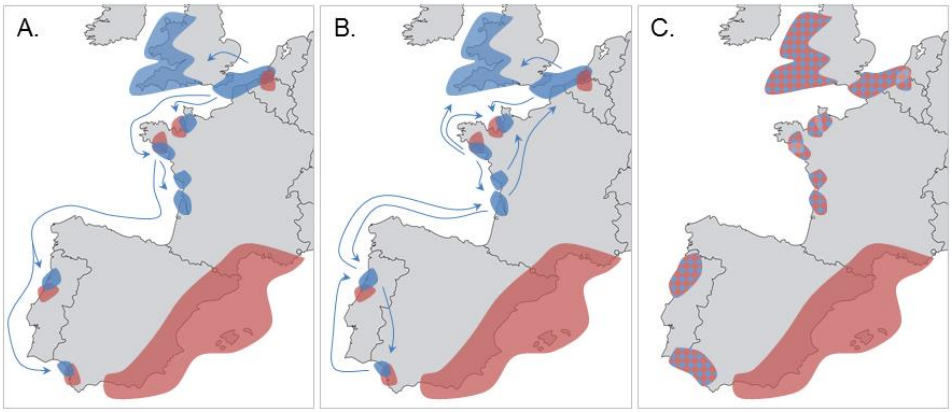


Figure 51. Different scenarios depicting the evolutionary history of the *P. chalceus* ecotypes. (A.) The short-winged ecotype inhabiting tidal marshes evolved only once and spread along the Atlantic coast. (B.) Different adaptive traits evolved in different tidal populations and introgressed into other tidal populations. (C.) Ecotypes have evolved *de novo* in each location and are being maintained by spatially heterogeneous balancing selection.

Considering the mitochondrial $NADP^+$-dependent isocitrate dehydrogenase (*mtIdh*) gene we found strong indications that the adaptation associated with this locus has a single origin (CHAPTER 2). Additionally, all outlier loci associated with the tidal and seasonal habitats identified from the RAD tag sequencing were shared between the repeatedly adapted populations (CHAPTER 5). This demonstrates that for these loci or traits, adaptations did not evolve *de novo* within each location, but evolved once and spread to other locations. Whether this adaptive variation results from standing genetic variation or introgression is difficult to distinguish and partly depends on the definition of these concepts and the proportion of individuals being exchanged between populations. For instance, if many individuals are being exchanged or colonize new locations, the spread of adaptive alleles could be considered as resulting from ancestral polymorphisms present in the populations (i.e. standing genetic variation). Oppositely, when low rates of migration are considered between established populations, the spread of adaptive alleles could be considered as resulting from introgression of adaptive alleles from one population into the other.

The *mtIdh* locus and most outlier loci had a strongly reduced genetic diversity among the haplotypes associated with the allele most frequent in the tidal habitats. This is in accordance with a recent spread or selective sweep of these loci along the Atlantic coasts. Moreover, as all these loci show a similar evolutionary history it is reasonable to suggest that the short-winged ecotype has (for a large part) spread recently and adaptation can

therefore be argued to largely result from ancestral polymorphism (Figure 51A), rather than from introgression of adaptive traits with different evolutionary histories (Figure 51B). Further, we found a deep divergence between the differently selected *mtIdh* alleles as well as between several outlier loci. As inferred from comparing coalescent simulations to *mtIdh* variation (CHAPTER 2), these deep divergences may result from evolution in a partially isolated subpopulation, suggesting a possible isolated origin of the short-winged ecotype. Additionally, looking at *Fst* values, we also found indications of adaptation unique to the localities, indicating either different selective pressures or unique *de novo* adaptation within these localities (Figure 51C). Furthermore, a few outlier loci showed a markedly different evolutionary history compared to the evolutionary history suggested by the recently spread outlier loci, which may emphasize the importance of introgression of adaptive traits from different populations.

The geographical origin of the ecotypes or adaptations has not been inferred. Divergence time estimation of the *mtIdh* alleles suggests an age between 0.047 and 0.165 MY ago. Hence, this divergence may predate the end of the last glacial period which occurred approximately 0.01 ago. However, it should be noted that this estimation corresponds to the age of the haplotypes and the adaptive mutation might have evolved more recently. Nevertheless, the preservation of relatively deep divergence between the *mtIdh* alleles (CHAPTER 2) likely suggests population subdivision which may predate the end of the last glacial period. This is strengthened by finding deep divergences between differently selected haplotypes in most identified outlier loci (CHAPTER 5). However, divergence times between haplotypes have not been estimated in these loci.

ALLOPATRY AS A DRIVER FOR SYMPATRIC DIVERGENCE

It can be argued that an initial evolution of complex adaptations in allopatry can facilitate evolution and persistence in sympatry after secondary contact. First, when populations diverge in allopatry, they are less constrained by high rates of potentially maladaptive gene flow (Lenormand 2002). Next, when the adapted ecotypes spread to other locations, adaptive variation is readily available (Barrett & Schluter 2008). Finally, genetically unlinked traits may be in linkage disequilibrium when they spread to other locations. For instance, in *P. chalceus*, the difference in behavioral response to inundation which is expected to result in spatial sorting and reproductive isolation between sympatric populations is also present between allopatric populations. When the ecotypes from tidal and seasonal habitats come into close contact, such as in the France Guérande or Portuguese Aveiro salterns, the genes responsible for these behavioral differences are likely in linkage disequilibrium with genes adapted to the habitat. This preexistence of

linkage disequilibrium between performance and assortative mating traits upon secondary contact is expected to promote the preservation of the distinct ecotypes in sympatry (see for instance the necessity for buildup of linkage disequilibrium for sympatric speciation; Fry 2003). Given the ample evidence of reuse of standing genetic variation and/or introgression and the possible isolated evolution of several adapted loci, this factor may be an important driver for the repeated occurrences of sympatrically diverged populations in *P. chalceus*.

P. CHALCEUS – COMPARISON WITH EVOLUTIONARY MODEL SYSTEMS

Studying individual evolutionary systems has provided pivotal insights to understand the effect of genetic and ecological factors on adaptation and speciation. However, comparing different groups of species is important as individual systems generally allow addressing only subsets of the ecological and genetic factors. Moreover, different species groups have shown to follow alternative evolutionary trajectories (e.g. parallel adaptation from ancestral polymorphism versus new mutations, genetic linkage versus genomically widespread adaptation, etc.) and it is important to obtain general patterns by comparing evolutionary study systems to fully understand the effect of ecological and genetic factors involved in adaptation and speciation.

The evolutionary trajectory we identified in *P. chalceus* suggests a largely singular and allopatric origin of the derived short-winged ecotype (i.e. deep divergence, singular origin and similar evolutionary history of adaptive loci), which subsequently spread along the Atlantic coasts and hybridized with the long-winged ecotype and obscuring the signatures of events at neutral loci. Both an allopatric origin as well as the existence of an isolating mechanism is argued to explain the observed genomically widespread differentiation in *P. chalceus*. In the following section, several evolutionary study systems are discussed to highlight similarities and differences of the insights gained from studying *P. chalceus*.

STICKLEBACKS Studies on threespine stickleback (*Gasterosteus aculeatus*) have demonstrated extensive reuse of standing genetic variation in the repeated evolution of distinct marine and freshwater stickleback (Colosimo *et al.* 2005, Schluter & Conte 2009, Jones *et al.* 2012). In respect to these finding, Schluter and Conte (2009) proposed the 'transporter hypothesis' in which marine threespine stickleback populations contain numerous freshwater-adapted alleles that have been acquired by introgressive hybridization with freshwater stickleback and reassembled when fresh water habitats

are colonized (Schluter & Conte 2009). These freshwater-adapted alleles are suggested to have been disaggregated into the genetic background of marine stickleback by backcrossing. As each member of the marine population carries only a very small number of freshwater alleles, these alleles are expected to have little effect on the total fitness of marine stickleback. When marine stickleback colonize fresh water habitats, the freshwater-adapted alleles will quickly rise in frequency and result in fast adaptation by reuse of standing genetic variation. Moreover, many freshwater alleles have been found to be recessive compared to marine alleles (Bell & Aguirre 2013). Recessiveness will reduce selection against freshwater-adapted alleles in oceanic populations and when these alleles are present at low frequencies, homozygotes for disadvantageous but recessive freshwater-adapted alleles will rarely be expressed and experience selection in marine populations. Moreover, many genes that have been repeatedly selected to high frequencies in freshwater habitats are clustered into several genomic regions (i.e. blocks of linked alleles) and sometimes contained within inversions (Hohenlohe *et al.* 2010a, Jones *et al.* 2012). Linkage of adaptive alleles has been suggested to be favored during the radiation of freshwater stickleback. This is because alleles that are physically linked to other freshwater-adapted alleles would covary among individuals and increase more rapidly in frequency due to direct selection as well as hitchhiking on linked alleles that are favored by selection. This selection for linkage among freshwater-adapted alleles is suggested to occur episodically after freshwater colonization (Bell & Aguirre 2013). Selection on freshwater adapted traits encoded by these blocks of linked alleles that are maintained in low frequency in marine populations through introgressive hybridization helps explain the repeated and predictable evolution of threespine stickleback (Bell & Aguirre 2013).

As the short-winged ecotype of *P. chalceus* is less dispersive, it might be speculated that alleles adapted to the tidal habitats may have been disaggregated into the genetic background of long-winged individuals from seasonal habitats in accordance to the transporter hypothesis. Partial disintegration (and possibly recessiveness) of short-winged or tidal adapted alleles into seasonal long-winged populations and subsequent repeated reassembly may have helped in the recent spread of the short-winged ecotypes along the Atlantic coasts. This may have happened after a singular origin of the short-winged tidal ecotype resulting in a largely similar evolutionary history of the short-wing-adapted alleles. In stickleback, the evolutionary origin of freshwater alleles remains at present little explored. However, invoking the transporter hypothesis suggests the high possibility of different evolutionary origins of freshwater alleles in stickleback. In further contrast to stickleback, we did not find any blocks of linked alleles in *P. chalceus*. However, these may have been missed due to sparse sampling of genomic variation.

Finally, several mechanisms have been suggested that result in reproductive isolation between stickleback populations. One set of mechanisms describes differences in female mate preference for size and shape between limnetic and benthic populations (Head *et al.* 2013). In *P. chalceus*, we did not find significant indications of mate preference between individuals from tidal and seasonal habitats (CHAPTER 6). Additionally, natal habitat preference has been shown to reduce migration between parapatric lake and stream stickleback by 76 % and increases the extent of adaptive divergence between populations (Bolnick *et al.* 2009). Similarly, natal habitat preference may play an important role in the evolution and preservation of distinct ecotype in *P. chalceus* as it provides a mechanism for rapid evolution of reproductive isolation (CHAPTER 7).

HELICONIUS BUTTERFLIES Passion-vine butterflies (*Heliconius*) have radiated into an extraordinary evolutionary continuum of divergent races and species at different stages of speciation (Mallet 2008, Supple *et al.* 2014). In this radiation, different stages during speciation show different degrees of genomic differentiation (Martin *et al.* 2013, Supple *et al.* 2014). Similarly to *P. chalceus*, very low genomic differentiation is found among sympatric population pairs (Martin *et al.* 2013). However, in contrast to *Heliconius*, allopatric populations of *P. chalceus* likely only established relatively recently or are subjected to high degrees of gene flow resulting in low genomic differentiation between allopatric *P. chalceus* populations compared to allopatric *Heliconius* populations. Further, *Heliconius* species represent a classic example of Müllerian mimicry, in which distantly related distasteful species have converged on the same warningly colored pattern. The convergence of these mimicry patterns results from the selective advantage from sharing the cost of educating predators (Jiggins 2008). Recently, genomic studies have shown that convergent evolution between races and species with even low levels of hybridization occurs by sharing uniquely derived color pattern alleles (Dasmahapatra *et al.* 2012). This has been demonstrated to result mainly from adaptive introgression after the different species established rather than reuse of ancestral polymorphism (Dasmahapatra *et al.* 2012). In contrast, the repeated occurrence of short-winged *P. chalceus* populations in tidal habitats likely has a largely singular origin and subsequently spread along the Atlantic coast (CHAPTER 5). Colonization of the short-winged ecotypes was likely followed by very high levels of gene exchange among the ecotypes. Hence, this pattern differs from adaptive introgression in which species or populations only exchange adaptive genetic variation. Additionally, continuous gene exchange between several *Heliconius* species has been demonstrated to result in up to 40 % of 100 kb genomic windows clustering according to geography rather than by species among sympatric species which indicates high rates of gene exchange between sympatric species (Martin *et al.* 2013). Similarly, hybridization and gene exchange is high

between the *P. chalceus* ecotypes which explain low genomic differentiation. Finally, in some *Heliconius* species, tight clusters of loci that facilitate the co-segregation of adaptive variation and determine wing-pattern morphs have been found. We did not identify any blocks of linked alleles in *P. chalceus*.

Ecological separation as well as mate preference have been suggested to play a major role in the evolution of reproductive isolation in *Heliconius* (Smith *et al.* 2001, Jiggins 2008). Moreover, male preference for color variation may provide a direct link between divergent selection on mimicry patterns and assortative mating and has been suggested to play a major role in the evolution of reproductive isolation among *Heliconius* species (Jiggins 2008). Similarly, in *P. chalceus* we determined a link between divergent selection and assortative mating through spatial sorting by different responses to inundation of individuals from tidal and seasonal habitats (CHAPTER 6 and CHAPTER 7). However, in *Heliconius*, the link between selection and assortative mating requires the evolution of both performance alleles (i.e. wing color pattern) as well as preference alleles (i.e. male preference allele for certain color pattern). In *Heliconius* butterflies it has been suggested that these traits are closely linked in the genome (Kronforst *et al.* 2006) and, thus, overcome the hampering effect of recombination on the evolution of reproductive isolation (Felsenstein 1981). In *P. chalceus*, the assortative mating mechanism may be pleiotropically linked to performance as escaping during inundation affects both spatial sorting and performance in the respective habitats.

BEACH MICE In *Peromyscus polionotus* beach mice a single mutation in the coding region of a pigmentation gene (melanocortin-1 receptor, *Mc1r*) has been found to play a major role in determining adaptive color pattern variation (Hoekstra *et al.* 2006). However, this association was absent from other similarly light-colored beach mouse populations suggesting that different molecular mechanisms are responsible for convergent phenotypic evolution in beach mice (Hoekstra *et al.* 2006). Additionally, cis-regulation of Agouti signaling protein (*Agouti*) expression also strongly affects coat color variation (Steiner *et al.* 2007, Linnen *et al.* 2009). This locus maps to an independent region and, together with *Mc1r*, is responsible for most of the differences in pigmentation between the subspecies. Moreover, the *Mc1r* and *Agouti* genes interact epistatically and the phenotypic effects of *Mc1r* are visible only in genetic backgrounds containing the derived *Agouti* allele (Steiner *et al.* 2007). Finally, it has been shown that at the *Agouti* locus, local adaptation is the result of independent selection on many mutations within a single locus and that a large-effect locus can fractionate into many small- to moderate-effect mutations (Linnen *et al.* 2013).

In *P. chalceus*, the derived *mtIdh* allele has been repeatedly selected to high frequencies in short-winged populations. Differences between the alternatively selected mtIDH

allozymes result from only one amino acid change (Lys – Asn at amino acid position 447). However, functional effects in the *mtIdh* gene of *P. chalceus* have not been studied and it is difficult to infer whether selection directly affects this gene. Marked differentiation at a locus 3,627 bp downstream of the end of the *mtIdh* gene suggests that selection affects genetic variation at an expanded region and that *mtIdh* may be linked to the target of selection. Also, the relative importance of coding variation and cis-regulatory changes has not been determined in *P. chalceus*.

STICK INSECTS Analysis of whole-genome divergence between replicate pairs of stick insect (*Timema cristinae*) populations that are adapted to different host plants revealed many modest-sized genomic regions of accentuated divergence (Soria-Carrasco *et al.* 2014). Most are inferred to be unique to individual population pairs. Similarly to *P. chalceus*, differentiation between populations is genomically widespread, even during the earliest stages of adaptation and speciation. In contrast to *P. chalceus*, reuse of adaptive variation is suggested to play a subordinate role in the parallel evolution of stick insect populations (83% of divergent SNPs are unique to population comparisons; Soria-Carrasco *et al.* 2014). Although, the exact ratio of shared and unique variation between repeatedly diverged *P. chalceus* populations has not been determined, the most strongly diverged loci are shared between parallel pairs of *P. chalceus* populations (CHAPTER 5).

GALÁPAGOS WOLF SPIDERS In the Galápagos, wolf spiders of the genus *Hogna* have repeatedly evolved into 'high elevation' and 'coastal dry' species on two islands (i.e. San Cristóbal and Santa Cruz) (De Busschere *et al.* 2012). Phylogenic reconstructions using several neutral genes suggests that these species specialized repeatedly and independently on the two islands (De Busschere *et al.* 2010). However, the genetic basis of adaptive variation is unknown and comparing rates of gene exchange between islands with coalescent simulations suggests the possibility of introgression of adaptive genetic variation between similarly adapted species of the different islands (De Busschere *et al.* under review). These speculations are strengthened by findings of introgression of adaptive variation and reuse of standing genetic variation in other evolutionary model systems such as *Heliconius* butterflies, sticklebacks and *P. chalceus* beetles.

LITTORINA SNAILS Ecotypes of the rocky-shore gastropod, *Littorina saxatilis*, have been demonstrated to have arisen in the face of continuous gene flow (Butlin *et al.* 2013). Moreover, studying genome wide neutral variation, it has been suggested that the ecotypes evolved repeatedly in different localities and do not have a single origin. However, Butlin *et al.* (2013) note that adaptive differentiation may not be fully

genetically independent and that alleles affected by selection may be shared from ancestral polymorphism or introgression. The study of *P. chalceus* indeed shows irreconcilable differences between studying neutral and adaptive variation which results from free interchange of neutral genetic variation but extensive reuse of adaptive variation among parallel population pairs. These results have important implications for the interpretation of the evolution of ecotypes in *P. chalceus* as well as *Littorina* snails.

MELITAEA AND *COLIAS* BUTTERFLIES The enzyme phosphoglucose isomerase (PGI) is associated with variation in flight metabolic rate, dispersal rate, fecundity and local population growth rate in both *Melitaea* (Orsini *et al.* 2009) and *Colias* (Wheat *et al.* 2006) butterflies from Finland. The observed polymorphism at this enzyme is suggested to result from balancing selection maintaining genetic variation within populations through heterozygote advantage (Watt *et al.* 1983, Wheat *et al.* 2006). Moreover, the split between the balanced polymorphism is estimated to predate the last common ancestor of a clade of five extant *Melitaea* species consistent with long-term balancing selection. In contrast, comparative structural analysis of *Pgi* polymorphism in the *Melitaea cinxia* and the *Colias eurytheme* butterfly suggests a similar but not identical target of balancing selection indicating independent convergent evolution between these species (Wheat *et al.* 2009).

In *P. chalceus*, polymorphism and old divergence of the *mtIdh* alleles most likely does not result from heterozygote advantage. Rather, the polymorphism is currently preserved by divergent selection in which the mtIDH-B allozyme is selected in seasonal habitats and the mtIDH-D allozyme is selected in tidal habitats. Furthermore, as suggested by coalescent simulations (CHAPTER 2) the old or deep divergence between the *mtIdh* haplotype clusters (which correspond with the mtIDH allozymes) results from a geographically separated evolution rather than from long-term balancing selection.

CICHLID FISHES In radiations of cichlid fishes from the Great Lakes of East Africa it is suggested that standing genetic variation present well before the start of the radiations has been important in facilitating evolutionary diversification (Brawand *et al.* 2014). Additionally, female preference for male coloration has been shown to be of major importance for the evolution of reproductive isolation between cichlid species (Maan *et al.* 2004). Moreover, heterogeneous light conditions in Lake Victoria has led to diversifying selection on *opsin* genes, which produce visual pigments, as a function of water depth (Seehausen *et al.* 2008). Divergence in *opsins* influences color perception and concomitantly affects female preference for male coloration. Hence, this mechanism results in a direct link between divergent selection and assortative mating through mate preferences. Similarly, in *P. chalceus* we determined a link between divergent selection

and assortative mating through spatial sorting by different responses to inundation of individuals from tidal and seasonal habitats (CHAPTER 6 and CHAPTER 7). In contrast, we found no significant indications of mate preference in *P. chalceus* (CHAPTER 6).

PHYTOPHAGOUS INSECTS Fruit flies belonging to the *Rhagoletis pomonella* species have shifted and adapted to new host plants in sympatry (Feder *et al.* 2003a). However, studying several neutral genes and an inversion polymorphism that affects key diapause traits showed that ancestral *R. pomonella* populations first became geographically isolated into a Mexican and North American population and that, subsequently, the inversion polymorphism introgressed from Mexico into the North American population. The inversion polymorphism affecting diapause is suggested to have aided North American flies in adapting to a variety of plants with differing fruiting times. The origin of adaptive genetic material in allopatry is suggested to have triggered the host shift and sympatric divergence (Feder *et al.* 2003a). In *P. chalceus*, we have not identified inversion polymorphisms. Although inversion polymorphisms may be present in *P. chalceus*, recombination within the *mtIdh* locus as well as between a large set (18) of adaptive loci indicates that inversions do not seem to play a major role in the evolution of the *P. chalceus* ecotypes. In agreement with *R. pomonella*, adaptive variation (i.e. *mtIdh* alleles (CHAPTER 2) and most loci identified as outlier loci (CHAPTER 5)) in *P. chalceus* is suggested to have evolved largely in geographical separation and subsequently spread along the Atlantic coast.

In specialized host races of the pea aphid *Acyrthosiphon pisum pisum*, strong genetic linkage or pleiotropy of performance (ecological specialization) and reproductive isolation through habitat preference has been demonstrated (Hawthorne & Via 2001). This mechanism is argued to strongly facilitated speciation between the pea aphid races. Similarly, in *P. chalceus*, selection may pleiotropically affect performance and assortative mating through spatial sorting. In *P. chalceus*, it may be speculated that for the link between performance and assortative mating pleiotropy is involved rather than close genetic linkage as escaping during inundation affects both spatial sorting and performance in the respective habitats and may arguably be determined by one genetic trait.

P. CHALCEUS – COMPARISON WITH EXISTING DISPERSAL POLYMORPHISM MODELS

Generally, theoretical models simulating the evolution of dispersal polymorphisms make assumptions about fluctuations in carrying capacity and/or persistence of habitats, trade-offs between reproduction and dispersal and the genetic architecture of the

polymorphism (GENERAL INTRODUCTION). These models have helped in understanding the range of evolutionary scenarios in which dispersal polymorphisms can evolve and be maintained (Johnson & Gaines 1990). One theoretical approach focuses on *trade-offs* between energy available for reproduction versus dispersal (i.e. oogenesis-flight syndrome) (Roff 1986, Roff & Fairbairn 2007). In these models, reduction in dispersal ability is favored because the resources that are otherwise allocated to dispersal can be used for reproduction, whereas the retention of a high dispersal morph is selected when habitats disappear and only dispersive individuals are able to colonize new habitats. Other models focus on *metapopulation dynamics* for the evolution of dispersal dimorphisms (McPeek & Holt 1992, Olivieri *et al.* 1995, Mathias *et al.* 2001, Cantrell *et al.* 2010, Hendrickx *et al.* 2013). For instance, it has been demonstrated that a dispersal dimorphism can be maintained by the interplay of within-population and between-population selection as individuals carrying dispersal genes tend to leave the local deme, resulting in a progressive decline in such genes within demes, whereas the high-dispersal genotypes will be overrepresented in newly colonized habitats (Olivieri *et al.* 1995). Another model, simulated by Mathias *et al.* (2001), investigated the evolution of dispersal in a landscape of many patches with fluctuating carrying capacities and spatial heterogeneity in temporal fluctuations. They assumed two kinds of habitats each consisting of many patches with patches of the 'good' habitat often having large carrying capacities and patches of the 'bad' habitats having large carrying capacities infrequently. No costs to dispersal were assumed (however, costs of dispersal did not change the general outcome of this model). In this kind of population structure they found that both the high-dispersal and the low-dispersal strategies are at an advantage when rare and coexist as a stable polymorphism. This is because when a high-dispersal strategy is established such that all individuals become redistributed each generation, the number of individuals becomes equal in each generation. In this situation, fitness will be higher in the 'good' patches (with high carrying capacity) and lower in the 'bad' patches due to crowding and this results in a source-sink structure in which a rare low-dispersal strategy becomes advantageous. Alternatively, if a low dispersal strategy is prevalent, environmental fluctuations within the patches will favor a high-dispersal strategy. Under these assumptions, the interplay of source-sink dynamics and patch fluctuations results in the evolution of a high and low dispersal phenotype which has been studied in a multiple models (Cohen & Levin 1991, McPeek & Holt 1992, Holt & McPeek 1996, Mathias *et al.* 2001, Hendrickx *et al.* 2013). The net outcome of the interacting selection forces depends on the frequency of the environmental changes and the size and quality of the habitats. Interestingly, in the model of Mathias *et al.* (2001), coexistence of the alternative strategies is made possible by partial spatial segregation of the phenotypes. The low dispersal phenotype will mainly occur in the 'good' habitat, while the high

dispersal phenotype occurs in both habitats. Furthermore, they argue that this spatial separation of the different phenotypes may facilitate local adaptation and speciation by reducing recombination and facilitating the development of assortative mating.

The model of Mathias *et al.* (2001) resembles the situation in *P. chalceus* in that tidal and seasonal or, respectively, Guérande canal and pond populations show contrasting dynamics that likely influence fluctuations in carrying capacity (or even extinction). Moreover, Desender (2000) was not able to provide support for the oogenesis-flight syndrome in *P. chalceus*, which is not a prerequisite in this model. However, it differs in that the frequency of the high dispersal phenotype is very low in the tidal habitat, whereas the metapopulation dynamics or trade-off models do not predict a strong spatial separation of the dispersal morphs. Furthermore, the models cited above describe the evolution of stable dispersal dimorphisms, whereas we find that, at least, wing size in *P. chalceus* is polymorphic. Alternatively, wing polymorphism likely does not result from metapopulation dynamics in *P. chalceus*, but rather from local adaptation to differing environmental conditions. Only local adaptation and the evolution of isolating mechanisms can explain both the repeatedly found polymorphic differences in dispersal ability and the spatial separation of ecotypes.

FUTURE PROSPECTS

The repeated ecotypic divergence of populations of the ground beetle *P. chalceus* offers to address a wide set of evolutionary questions. These questions range from identifying the genes and genetic architecture involved in polymorphic traits and local adaptation up to identifying genes involved in reproductive isolation and studying the process of speciation. Several answers have been put to forward, but much is still to learn. Several topics could be (relatively) readily addressed:

❖ LONG INSERT LIBRARIES AND LINKAGE MAPPING TO ORDER SCAFFOLDS

Constructing and sequencing very large insert mate libraries (>10 kb) will strongly improve the genome assembly. Additionally, single parents from tidal and seasonal populations (both allopatric and sympatric) or polymorphic individuals within populations can be crossed resulting in F1 offspring. Performing RAD tag sequencing on these F1 offspring will allow constructing a linkage map in which the available scaffolds are divided according to the chromosomes and correctly ordered. This will result in an improvement of the draft genome assembly and, consequently, allow studying the genetic architecture of divergence in more detail. Also, genomic structural variation, such as inversions, deletions end duplications, will be more easily identified. Furthermore, an F2 generation can be bred based on crosses within F1 families to allow for recombination between the different traits (Van Ooijen & Jansen 2013). When measuring phenotypic traits in the F2 offspring, Quantitative Trait Loci (QTL) mapping will allow linking certain genomic regions or markers to the phenotypic traits of interest.

❖ GENOMIC REGION SURROUNDING THE *mtIDH* GENE

Identifying and sequencing the genomic region flanking the *mtIdh* gene will (i) allow studying the genomic region affected by selection, (ii) help in determining the precise target of selection and (iii) help in discriminating alternative evolutionary scenarios responsible for the observed sequencing variation. For instance, a large genomic region of reduced nucleotide diversity would provide a strong indication of a recent selective sweep (e.g. Linnen et al. 2009).

❖ COMPLETE GENOME RESEQUENCING

RAD tag sequencing provides a valuable tool to compare genome wide mutational variation among multiple individuals and populations. However, from a genome wide perspective, the information obtained from the number of loci using a rare-cutting restriction enzyme (SbfI-HF) is limited (e.g. 2,800 RAD tags in a 530 Mb genome provides approximately one tag every 189 kb). Therefore, performing more elaborate sequencing using a frequent-cutting restriction enzyme or, given the construction of a

well finished genome sequence, complete genome resequencing will allow studying the genomic pattern of population differentiation in high detail (e.g. Ellegren et al. 2012; Jones et al. 2012; Soria-Carrasco et al. 2014). Moreover, it should be possible to construct a replicate mesocosmos experiment wherein tidal and seasonal hydrological dynamics are simulated to study the effect on genomic differentiation (see e.g. Soria-Carrasco *et al.* 2014, Arnegard *et al.* 2014).

❖ FUNCTIONAL DIFFERENTIATION

Differential expression analysis of (common garden raised) *P. chalceus* beetles from differentiated populations using high throughput transcriptome sequencing can help identifying the genetic pathways that affect the development of traits involved in local adaptation and reproductive isolation (e.g. Wheat *et al.* 2011).

❖ MORE EXTENSIVE STUDY OF (NATAL) HABITAT PREFERENCE

We identified different behavioral responses to inundation between differentiating populations and a possible effect of natal habitat experience. First, the constancy of this latter effect should be addressed by repeating the experiment. Subsequently, quantifying the effect size of these traits (i.e. the increase in total reproductive isolation caused by its divergence) can be addressed. Finally, the relative importance of genetic and non-genetic (i.e. natal habitat preference) causes for the evolution of reproductive isolation can be studied. For instance, gene flow between the two types of habitats could be quantified by studying migration and spatial sorting on a larger scale. By constructing a replicate mesocosmos setting, wherein the two habitats are spatially separated but connected, it can be tested if spatial sorting either occurs in response to population origin or natal experimental treatment and to what extant these traits reduce gene flow.

❖ UNRAVELING THE GENETIC BASIS OF THE TRAIT(S) INVOLVED IN REPRODUCTIVE ISOLATION

An ultimate goal in evolutionary biology is to identify genes involved in reproductive isolation (Orr *et al.* 2004, Nosil & Schluter 2011). By quantifying the behavioral responses to inundation in an F2 cross, it should be possible to relate this variation to variation in molecular markers (i.e. QTL mapping using markers from RAD tag sequencing or complete genome resequencing). Identifying and studying the genes or genomic regions directly involved in reproductive isolation will provide unprecedented insights into the factors that are thought to be representative of those underlying the origin of species (Orr *et al.* 2004). Again, constructing replicate mesocosmos experiments would allow studying the effect of different genes or genomic regions on adaptation and the evolution of reproductive isolation (Arnegard *et al.* 2014).

SUMMARY

In this thesis, ecological and genetic mechanisms are investigated that are involved in the ecotypic divergence of the wing-polymorphic ground beetle *Pogonus chalceus* in the absence of a geographical barrier (i.e. sympatry). Studying these mechanisms is essential to unravel how populations adapt to differing and changing environments and, ultimately, gain insights in speciation.

The ground beetle *P. chalceus* represents an interesting case of replicated adaptation to different hydrological regimes present in salt marshes along the Atlantic European coasts. Short-winged populations are found in *tidal marshes* that are inundated frequently but for short periods only (maximally 6 h), whereas long-winged populations are found in *seasonal marshes* separated from the tidal influence of the sea, but inundated irregularly for longer time periods. Developing a dispersive phenotype is thought to be energetically costly and the retention of long wings is, therefore, argued to be an adaptation to escape long term inundations and uninhabitability of the habitats. Additional to the correlation between wing size and habitat, there is a strong correlation between mean population wing size and frequency of different allozymes of the mitochondrial $NADP^+$-dependent isocitrate dehydrogenase (mtIDH) enzyme. Interestingly, in some locations, such as the Guérande salterns in France, long -and short-winged populations are found in the hydrological alternative habitats on very small distances from each other (i.e. canals and ponds; only 10-20 m). In CHAPTER 1, we find that wing size is strongly genetically determined in these sympatric populations and that genes involved in wing size are likely not genetically linked to the *mtIdh* locus. This strongly invokes the selection-migration antagonism because recombination is expected to result in maladapted gene combinations when gene flow is high. To gain more insights in this evolutionary system we attempt to answer the following questions: (i) Does the repeated occurrence of the locally adapted populations in distinct locations result from the single origin or did the adaptation evolve multiple times in different locations? (ii) How is adaptive divergence maintained despite the ample opportunity of gene flow?

Whether the *mtIdh* gene is the target of selection or rather closely linked to the target of selection is unknown. Nevertheless, in CHAPTER 2 we use the tight association of mtIDH allozymes with both habitat dynamics and dispersal ability (i.e. wing size) at population level to make inferences about the evolutionary history of the repeated evolution of the adaptation associated with this locus (research question i). By studying

sequences of the *mtIdh* gene, we find that the differentiation in the *mtIdh* locus has a single origin. Furthermore, comparing the observed pattern of sequencing variation with coalescent simulations suggests that the *mtIdh*-allele associated with the short-winged populations from the tidal habitats likely evolved in a partially isolated subpopulation and spread recently along the Atlantic coast.

To gain more insight into the genetic aspects of adaptive sympatric divergence in *P. chalceus*, we first expand the genomic resources available for *P. chalceus* by sequencing the transcriptome and genome in, respectively, CHAPTER 3 and CHAPTER 4. Transcriptome sequencing resulted in 65,766 contigs, clustering into 39,393 unique transcripts or genes (unigenes). Furthermore, using homology searches we identified all reported genes involved in wing development, juvenile- and ecdysteroid hormone pathways in *Tribolium castaneum*. The draft *P. chalceus* genome assembly consists of 312.78 Mb of genome sequence comprising 109,580 unordered scaffolds and covering about 58.98 % of the estimated genome size (530.28 Mb). Repetitive elements comprise about 18.60 % of the assembled genome. Finally, alignment with the genome of *T. castaneum* suggests a high rate of intra as well as interchromosomal rearrangements since their divergence.

Next, in CHAPTER 5, we use RAD (Restriction Associated DNA) tag sequencing to study population structure at a genome-wide scale and identify the genomic pattern of adaptive differentiation among repeatedly adapted sympatric and allopatric populations. Comparison of genome wide variation among populations covering nearly the entire species range indicates low population divergence between sympatric as well as allopatric populations, suggesting high rates of gene flow and relatively recent separation. Contrastingly, we find multiple unlinked loci that are strongly associated with adaptive divergence, indicating widespread genomic divergence even between sympatric populations. By using the assembled *P. chalceus* genome as a reference to construct sequence alignments of the RAD tags, we find that all the alleles identified as outlier loci have a singular mutational origin and are shared between repeatedly diverged populations. Moreover, most of these loci have a similar evolutionary history as the *mtIdh* locus, which suggests a recent increase of the alleles associated with the short-winged populations from tidal habitats. This shared evolutionary history suggests a largely singular evolutionary origin of the short-winged ecotypes in *P. chalceus* and a recent spread along the Atlantic coasts (research question i). Moreover, the rapid and recurrent sympatric divergence in *P. chalceus* may have been promoted by the singular evolution of the adapted traits and high rates of exchange of this genetic building material among populations (research question ii).

Finding multiple unlinked outlier loci between the differently selected sympatric *P. chalceus* populations indicates that even very early stages of the speciation process may by characterized by genome wide adaptation. Possibly, this is driven by a reproductive isolating mechanism that reduces gene flow and assists natural selection in the evolution of distinct ecotypes that diverge in multiple unlinked loci in sympatry (research question ii). Moreover, understanding how disruptive selection can result in the evolution of such a mechanism is of major interest in the study of sympatric speciation. Therefore, in CHAPTER 6 we study behavioral variation in the response to inundation in *P. chalceus*, which may provide a mechanism resulting in assortative mating between individuals from differently selected populations. We demonstrate that short-winged populations from tidally inundated marshes show less reluctance to inundation compared to long-winged populations from seasonal marshes. These behavioral differences may result in spatial sorting and can as such provide a unique and simple explanation for the persistence of distinct ecotypes in sympatric mosaics. The mechanism proposes a direct link between traits subjected to disruptive selection and habitat preference resulting in the evolution of assortative mating and, hence, reproductive isolation.

Finally, in CHAPTER 7, we find significant indications of natal habitat experience on habitat preference in *P. chalceus*. More precisely, we demonstrate that adults of both ecotypes that were exposed to frequent but short inundations during larval and pupal development have a significantly lower response to escape these inundations in the adult stage. Such natal habitat preference may easily result in spatial sorting of individuals that experienced different environmental conditions and, hence, quickly lead to assortative mating within habitats and, therefore, may have important consequences for both the evolution and persistence of sympatric races in *P. chalceus*. Besides, we also found strong indications that responses to inundation have a genetic component as seen by the consistently higher water reluctance of seasonal beetles, independent from the environment in which they are raised.

Altogether, we find widespread genomic divergence and extensive reuse of adaptive genetic variation in the sympatric and repeated evolution of *P. chalceus* ecotypes. Furthermore, we identify a trait that likely results in reproductive isolation between the sympatric ecotypes through spatial sorting. This trait provides a direct link between selection and reproductive isolation as adaptation of the individuals to the different habitats also causes reproductive isolation. Such traits are often called 'magic traits'.

SAMENVATTING

In dit proefschrift worden de ecologische en genetische mechanismen bestudeerd die betrokken zijn bij de ecotypische divergentie van de vleugel-polymorfe loopkever *Pogonus chalceus* in de afwezigheid van een geografische barrière (i.e. sympatrie). Het bestuderen van deze mechanismen is essentieel om te begrijpen hoe populaties zich aanpassen aan verschillende en veranderende omgevingen en uiteindelijk inzicht te krijgen in soortvorming.

De loopkever *P. chalceus* vertegenwoordigt een interessant voorbeeld van herhaalde ecotypische adaptatie aan verschillende hydrologische omstandigheden langsheen de Atlantische Europese kust. Kortvleugelige *P. chalceus* populaties komen voor in tidale schorren, terwijl langvleugelige populaties eerder voorkomen in seizoenale of binnendijkse zilte moerassen. De tidale schorren staan onder een sterke invloed van de getijden en worden frequent overspoeld, maar slechts voor korte periodes (maximaal 6u). De seizoenale zilte moerassen, daarentegen, staan niet onder invloed van de getijden en worden onregelmatig overstroomd voor langere perioden. Gezien het ontwikkelen van een dispersief fenotype verondersteld wordt energetisch kostelijk te zijn, wordt het behoud van lange vleugels gezien als een aanpassing om aan de lange termijn overstromingen en ongunstige omstandigheden te ontsnappen. Naast dit verband tussen vleugellengte en habitat is er ook een sterke correlatie tussen de gemiddelde vleugellengte van de populaties en de frequentie van verschillende allozymes van het mitochondriale $NADP^+$-afhankelijke isocitraatdehydrogenase (mtIDH) enzym. In sommige locaties, zoals in de zoutpannen in de Guérande te Frankrijk, kunnen zowel lang -als kortvleugelige populaties dicht bij elkaar worden gevonden in de alternatieve hydrologische habitats (i.e. kanalen en poelen; slechts 10-20 m). In HOOFDSTUK 1 tonen we voor de sympatrische populaties aan dat vleugellengte sterk genetisch wordt bepaald en dat de genen die betrokken zijn bij vleugellengte waarschijnlijk niet genetisch gekoppeld zijn aan het *mtIdh* locus. Aangezien een genetische link tussen deze kenmerken ontbreekt, wordt verwacht dat sterke genuitwisseling tussen deze nabijgelegen populaties zal leiden tot recombinatie tussen deze kenmerken en dus tot het ontstaan van minder gunstige gencombinaties (selectie-migratie antagonisme). Om meer inzicht te krijgen in dit intrigerend evolutionair systeem trachten we de volgende onderzoeksvragen te beantwoorden: (i) Heeft het herhaaldelijk voorkomen van gelijkaardige ecotypess in verschillende locaties een enkelvoudige oorsprong of is de aanpassing aan het habitat meerdere keren

onafhankelijk geevolueerd op verschillende locaties? (ii) Hoe wordt adaptieve divergentie in stand gehouden ondanks de ruime gelegenheid voor genuitwisseling?

Of het *mtIdh* gen het effectieve doelwit is van selectie of eerder nauw gekoppeld is aan het geselecteerde doelwit is onbekend. Desondanks gebruiken we in HOOFDSTUK 2 de sterke associatie van de mtIDH allozymes met zowel habitat dynamiek als gemiddeld populatie dispersievermogen (i.e. vleugellengte) om de evolutionaire geschiedenis te achterhalen van de herhaalde adaptieve evolutie geassocieerd met dit locus (onderzoeksvraag i). Door sequenties van het *mtIdh* gen te vergelijken tussen populaties, tonen we aan dat de differentiatie in het *mtIdh* locus een enkelvoudige oorsprong heeft. Bovendien vinden we, door het waargenomen patroon van sequentievariatie te vergelijken met coalescentie simulaties, dat het *mtIdh*-allel geassocieerd met de kortvleugelige populaties uit de tidale habitats waarschijnlijk geëvolueerd is in een gedeeltelijk geïsoleerde subpopulatie en zich recent heeft verspreid langsheen de Atlantische kust.

Om betere inzichten te verkrijgen in de genetische aspecten van adaptieve en sympatrische ecotypische divergentie *P. chalceus,* breiden we in HOOFDSTUK 3 en HOOFDSTUK 4 de genomische data uit die beschikbaar is voor *P. chalceus* door het transcriptoom en genoom te sequeneren. Transcriptoom sequenering resulteerde in 65.766 contigs die samen kunnen worden geclusterd tot 39.393 unieke transcripten of genen (unigenes). Bovendien vinden we homologe transcripten terug van alle genen die betrokken zijn bij de vleugelontwikkeling en deel uitmaken van de juveniele- en ecdysteroid hormoon pathways in *Tribolium castaneum*. Het geassembleerde *P. chalceus* genoom bestaat uit 312,78 Mb genomische sequenties bestaande uit 109,580 ongeordende scaffolds die ongeveer 58.98% van de geschatte genoomgrootte bedekken (530,28 Mb). Repetitieve elementen omvatten ongeveer 18.60% van het geassembleerde genoom. Tenslotte, vergelijking met het genoom van *T. castaneum* suggereert een hoge mate van intra evenals interchromosomale herschikkingen sinds hun divergentie.

Vervolgens wordt in HOOFDSTUK 5 gebruik gemaakt van RAD (Restriction Associated DNA) tag sequencing om de populatiestructuur te bestuderen op het genoomniveau en om het genomische patroon van adaptieve differentiatie te achterhalen tussen herhaaldelijk aangepaste sympatrische en allopatrische *P. chalceus* populaties. Genomische variatie tussen populaties die nagenoeg de gehele soortspreiding bestrijken toont lage divergentie tussen sympatrische evenals allopatrische populaties, wat suggereert dat er een hoge mate van genuitwisseling is tussen sympatrische populaties evenals een recente scheiding van de allopatrische populaties. Daarentegen vinden we

meerdere niet-gelinkte loci terug die sterk geassocieerd zijn met de adaptieve divergentie, wat wijdverspreide genomische verschillen aangeeft ook tussen sympatrische populaties. Door gebruik te maken van het geassembleerde *P. chalceus* genoom als referentie voor het construeren van sequentie alignementen van de RAD tags tonen we aan dat alle geïdentificeerde allelen die herhaaldelijk sterk geassocieerd zijn met de adaptieve divergentie een enkelvoudige oorsprong hebben. Bovendien hebben de meeste van deze loci een vergelijkbare evolutionaire geschiedenis met die van het *mtIdh* locus, welke een recente verspreiding van de allelen geassocieerd met het kortvleugelige ecotype uit tidale schorren suggereert. Deze gedeelde evolutionaire geschiedenis suggereert een grotendeels unieke evolutionaire oorsprong van het kortvleugelige ecotype en een recente verspreiding langs de Atlantische kusten (onderzoeksvraag i). Het is mogelijk dat snelle en herhaaldelijke sympatrische divergentie in *P. chalceus* wordt bevorderd door de enkelvoudige evolutie van de adaptieve eigenschappen in combinatie met sterke genuitwisseling van dit genetisch bouwmateriaal tussen de populaties (onderzoeksvraag ii).

Het vinden van meerdere ongelinkte loci die geassocieerd zijn met de adaptieve divergentie tussen sympatrische populaties geeft aan dat zelfs het zeer vroege stadium van soortvorming kan gekenmerkt worden door adaptatie langsheen het volledige genoom. Mogelijks wordt dit proces gedreven door een mechanisme dat zorgt voor reproductieve isolatie en dus genuitwisseling reduceert waardoor natuurlijke selectie wordt geholpen in de evolutie van divergente ecotypes die verschillen in meerdere niet gelinkte loci (onderzoeksvraag ii). Verklaren hoe disruptieve selectie kan leiden tot de evolutie van zo een mechanisme is van groot belang in de studie van sympatrische speciatie. Daarom bestuderen we in HOOFDSTUK 6 variatie in gedrag als reactie op inundatie in *P. chalceus*. Variatie in dit gedrag zou een mechanisme kunnen zijn dat leidt tot assortatief paargedrag en zodus tot een reductie in genuitwisseling tussen individuen van de verschillend geselecteerde populaties. Meer bepaald tonen we aan dat kort-gevleugelde populaties uit tidale schorren significant minder vluchtgedrag vertonen bij innundatie in vergelijking met lang-gevleugelde populaties uit seizoenale moerassen. Deze gedrags verschillen kunnen leiden tot ruimtelijke scheiding van de ecotypes en kan als zodanig een unieke en eenvoudige verklaring geven voor het voortbestaan van verschillende ecotypes in sympatrie. Daarenboven maakt dit mechanisme een directe link mogelijk tussen de eigenschappen onderworpen aan disruptieve selectie en habitat preferentie, wat resulteert in de evolutie van assortatief paargedrag en dus reproductieve isolatie.

Tenslotte vinden we in HOOFDSTUK 7 aanwijzingen dat het natale leefgebied in *P. chalceus* een significant effect heeft op habitatvoorkeur van adulte kevers. Meer bepaald, zien we dat volwassen kevers van beide ecotypen die werden onderworpen aan veelvuldig korte inundaties tijdens het larvaal en popstadium een significant lager vluchtgedrag vertonen als reactie op inundatie in het volwassen stadium. Dergelijke natale habitatvoorkeur kan theoretisch gemakkelijk leiden tot ruimtelijke scheiding van individuen die verschillende milieu-omstandigheden ervaren en kan dus snel leiden tot evolutie van assortatieve paring binnen habitats. Dit mechanisme kan belangrijke implicatie hebben voor zowel de evolutie als het voortbestaan van sympatrische ecotypes in *P. chalceus*. Daarnaast vonden we ook sterke aanwijzingen dat de reactie op overstroming een genetische component heeft zoals gezien door het consequent hogere vluchtgedrag van de langvleugelige kevers uit seizoenale moerassen onafhankelijk van de omgeving waarin ze worden opgegroeid.

Alles bij elkaar vinden we wijdverspreide genomische divergentie en extensief hergebruik van adaptieve genetische variatie in de sympatrische en herhaalde divergentie van *P. chalceus* ecotypes. Daarnaast identificeren we een eigenschap van de kevers die waarschijnlijk leidt tot reproductieve isolatie via ruimtelijke scheiding tussen de ecotypes in sympatrie. Deze eigenschap zorgt bovendien voor een directe link tussen de disruptieve selectie en reproductieve isolatie aangezien de aanpassing van de individuen aan de verschillende habitats ook reproductieve isolatie veroorzaakt. Dergelijke eigenschappen worden vaak 'magic traits' genoemd.

GLOSSARY

Allozyme

Allozymes are variant forms of an enzyme that are coded by different alleles at the same locus.

Balancing selection

A selective process by which multiple alleles are maintained in the gene pool. This may happen, for instance, when heterozygotes for the alleles have higher adaptive values than the homozygotes or by frequency dependent selection (Futuyma 2005).

Bayes Factor (BF)

In Bayesian statistics, a model choice decision can be performed using the so-called 'Bayes Factors' (Goodman 1999). Given two models M_1 and M_2 (for instance neutral versus selection) trying to explain a data set N, the Bayes factor BF for model M_2 is given by:

$$BF = \frac{P(N|M_2)}{P(N|M_1)}$$

Hence, the BF provides a scale of evidence in favor of one model versus another. For example, BF = 2 indicates that the data favors model M_2 over model M_1 at odds of two to one.

Cell means model

There are several ways to parameterize a general linear model (GLM). In its most standard form, the means of each factor level are not estimated directly, but the model is parameterized such that the mean of one single factor level is estimated as well as the difference between the means of the remaining levels. This model is useful for conducting Type III tests, and thus testing the significance of a particular factor of interest. However, this model does not allow obtaining the means and SE of all the group levels. These can be obtained by re-parameterizing the GLM such that the means of each level (and their SE) are estimated directly. This latter model is called the cell means model.

Cis-regulation

Regulation of gene transcription by nearby non-coding DNA which typically functions as a binding site for transcription factors (i.e. trans-regulatory elements).

Coalescent simulations

Coalescent theory attempts to estimate population parameters (e.g. population size, migration and selection) based on sequence variation. More precisely, coalescent theory provides a retrospective model that attempts to trace all alleles of a gene shared by all members of a population to a single ancestral copy, called the most recent common ancestor (MRCA). The coalescent is typically represented as a gene genealogy, similar to a phylogenetic tree. Apart from genetic drift, complex coalescent models allow incorporating recombination, natural selection, and gene flow or population structure to estimate population parameters. Oppositely, coalescent simulation programs, such as MSMS (Ewing & Hermisson 2010) use coalescent theory to simulate trees and DNA sequences given population parameters as input.

Contigs

Contiguous consensus sequences that are derived from collections of overlapping reads.

Dispersal syndrome

Refers to the association between dispersal and other behavioral and/or life-history traits (Stevens *et al.* 2013).

Disruptive selection

Special case of divergent selection in which selection favors extreme phenotypes over intermediate phenotypes (Rundle & Nosil 2005).

Divergent selection

Selection is divergent when it acts in contrasting directions or favors different phenotypes in the different populations (Nosil 2012).

Ecological speciation

The process by which barriers to gene flow evolve between populations as a result of ecologically based divergent selection between environments (Nosil 2012).

Ecotype

An ecotype is a genetically distinct variety, population or race within a species (i.e. are capable of interbreeding), which is adapted to specific environmental conditions.

Epistasis

Epistasis is when the effect of one gene depends on the presence of one or more other genes (genetic background) (Cordell 2002).

Fst (Wright's fixation index)

The fixation index is a measure of population differentiation. It is calculated as the fraction of the total genetic variation that is distributed among subpopulations in a subdivided population. One way to calculate Fst is (Holsinger & Weir 2009):

$$F_{st} = \frac{\sigma_S^2}{\sigma_T^2}$$

where σ_S^2 is the variance in frequency of alleles in different subpopulations, and σ_T^2 is the variance of allele frequencies in the total population.

Genetic markers

Heritable polymorphisms that can be measured in one or more populations of individuals.

Heritability

The proportion of observable differences in a trait between individuals within a population that is due to genetic differences (Falconer & Mackay 1996).

k-mer

String of nucleotides of length k in a sequence read or genomic sequences (Compeau *et al.* 2011).

Linkage disequilibrium (*LD*)

Linkage disequilibrium is the occurrence of some combinations of alleles or genetic markers in a population more often or less often than would be expected from a random formation of haplotypes from alleles based on their frequencies. It results from population substructure or from linkage, which is the presence of two or more loci on a chromosome with limited recombination between them.

Linkage disequilibrium (*LD*) mapping

Testing for a statistical association between genetic markers and particular phenotypes based on the premise that the marker(s) is in *LD* with the causal locus, or less likely, is in fact the causal mutation itself. (Stinchcombe & Hoekstra 2008)

Mapping Quality (*MQ*)

MQ gives an expression of the estimated probability of true alignment of a read to the reference (Li *et al.* 2008). Expressing this probability (*Pe*) in the Phred scale gives the *MQ*:

$$Pe = 10^{MQ/10}$$

Given 1000 read mappings, a *MQ* of 30 indicates that one will be wrong on average. *MQ* calculation considers repeat structure, base quality, sensitivity of mapping algorithm and whether reads are mapped in pair.

Monte Carlo Markov Chain (MCMC)

Algorithm that samples from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. Samples of the chain can be used as a sample of the desired distribution.

Maximal Realizable Wing Size (MRWS)

Wing size is expressed as an index that corrects for the allometric relationship between wing length and body size (den Boer 1980, Desender *et al.* 1986). More precisely, the relative wing size corrected for allometry expresses the percentage of the maximal realizable wing size (%MRWS). The relative wing size is wing length × width divided by elytral length × width. In %MRWS, relative wing size is expressed as a percentage of the maximal relative wing size for a beetle of a given size. This maximal realizable wing size was derived from a regression of wing length and body size from Carabid species with always fully developed wings and functional flight muscles allowing comparisons of relative wing sizes of beetles with different body sizes (Desender *et al.* 1986).

Median Joining network

Median networks are usually constructed for closely related sequences that have evolved without recombination. In a median network, every sequence of a given multiple sequence alignment is represented by a node and additional nodes are said to represent unobserved sequences. Two nodes are connected by an edge if they differ by exactly one mutation. The so called median is the median sequence between a set of sequences. To construct a network, the original sequences have to be converted into binary sequences. Therefore, so called quasi-median networks are constructed from the quasi median sequences. These quasi median sequences represent the set of possible median sequences whenever three different states occur at a certain position in the sequence set.

The number of nodes of the quasi-median network associated with a multiple sequence alignment can become very large, even for a small number of short sequences. Therefore, the Median Joining method applies two different algorithms (Bandelt *et al.* 1999). First, it repeatedly constructs so called

'minimum spanning networks' and, secondly, repeatedly uses the quasi-median calculation of three mutually close sequences at a time. In this way it constructs an informative subnetwork of the full quasi-median network, guided by the minimum spanning network. The minimum spanning network combines 'minimum spanning trees', which are trees found in a graph connecting sequences by edges weighted by their distance. Using both algorithms, the median-joining method attempts to provide a useful network of intermediate size. (Adopted from Huson *et al.* 2010)

Microsatellite

A class of repetitive DNA that is made up of repeats that are 2-8 nucleotides in length. They can be highly polymorphic and are frequently used as molecular markers in population genetics studies.

Mutation-order speciation

The evolution of reproductive isolation by the fixation of different advantageous mutations in separate populations experiencing similar selection pressures (Schluter 2009).

N50 size

By far the most widely used statistics for describing the quality of a genome assembly are its scaffold and contig N50s. A contig N50 is calculated by first ordering every contig by length from longest to shortest. Next, starting from the longest contig, the lengths of each contig are summed, until this running sum equals one-half of the total length of all contigs in the assembly. The contig N50 of the assembly is the length of the shortest contig in this list. The scaffold N50 is calculated in the same fashion but uses scaffolds rather than contigs. The longer the scaffold N50 is, the better the assembly is. However, it is important to keep in mind that a poor assembly that has forced unrelated reads and contigs into scaffolds can have an erroneously large N50. Note too that scaffolds and contigs that comprise only a single read or read pair — often termed 'singletons' — are frequently excluded from these calculations, as are contigs and scaffolds that are shorter than ~800 bp. The procedures used to calculate N50 may therefore vary between genome projects. (Yandell & Ence 2012).

Neighbor-Net network

Neighbor-Net is a distance based split method for constructing phylogenetic networks that is based on the Neighbor-Joining (NJ) algorithm (Bryant & Moulton 2004). The NJ algorithm takes a distance matrix to construct a fully resolved (bifurcating) phylogenetic tree (i.e. joining sets of sequences with consecutive larger distance values). Neighbor-Net proceeds by constructing a collection of weighted splits (bipartitions of the taxa set) based on a distance matrix and then represent these splits using a splits graph (a special type of

phylogenetic network that simultaneously represents both groupings in the data and evolutionary distances between taxa).

Nucleotide diversity (π)

Nucleotide diversity measures the degree of polymorphism within a population. One common measure is defined as the average number of nucleotide differences per site between any two DNA sequences chosen randomly from the sample population (Nei & Li 1979):

$$\pi = \sum_{ij} x_i x_j \pi_{ij}$$

where x_i and x_j are the respective frequencies of the i^{th} and j^{th} sequences and π_{ij} is the number of nucleotide differences per nucleotide site between the i^{th} and j^{th} sequences.

Outlier locus

A genetic marker showing a degree of divergence statistically departing beyond background or neutral expectations. Outlier loci are often interpreted as being affected by divergent selection (Stinchcombe & Hoekstra 2008).

Parapatry

Parapatry is the geographic relationship between populations whose ranges do not significantly overlap, but are immediately adjacent to each other.

Phred quality (PQ)

Phred quality is defined on base calls and each base call is an estimate of the true nucleotide. The probability that a base call is wrong is called the error probability (Be) and PQ is calculated as (Ewing & Green 1998):

$$PQ = -10 \times \frac{\log(Be)}{\log(10)}$$

If the quality of a base call is 30, the probability that it is wrong is 0.001.

Pleiotropy

Pleiotropy occurs when one gene influences multiple phenotypic traits which may seem unrelated.

Polyphenism

Polyphenism is a special case of phenotypic plasticity in which discrete phenotypes arise from a single genotype as a result of differing environmental conditions.

Quantitative trait locus (QTL)

A locus that controls a quantitative phenotypic trait, identified by showing a statistical association between genetic markers surrounding the locus and phenotypic measurement (Van Ooijen & Jansen 2013).

Quantitative trait locus (QTL) mapping

Similar to LD mapping, but with use of controlled crosses (pedigree information). This approach necessitates constructing a genome-wide linkage map with the relative positions of markers. (Van Ooijen & Jansen 2013)

RAD tag sequencing

RAD tags are the DNA sequences that immediately flank each instance of a particular restriction site of a restriction enzyme throughout the genome. By labeling DNA from multiple individuals, RAD-tag sequencing results in sequencing randomly distributed but consistent genomic regions from multiple individuals. The density of RAD tags in a genome depends on the restriction enzyme used during the isolation process. Sequencing is performed using a high throughput sequencing platform. DNA sequence polymorphisms in the resulting RAD-tags can be used for association mapping, QTL-mapping and population genetics. (Davey *et al.* 2010)

RNA-mediated interference (RNAi)

Biological process by which RNA molecules inhibit gene expression. RNA interference has an important role in defending cells against parasitic nucleotide sequences (i.e. viruses and transposons) and developmental regulation. In experimental biology, this process can be used to knockdown expression of target genes by adding double stranded RNA complementary to the gene of interest.

Scaffolds

Ordered and orientated sets of contigs that are linked to one another by mate pairs of sequencing reads.

Segregating sites

Nucleotide sites which are polymorphic within a set of sequences.

Selective sweep

The reduction of nucleotide variation in loci neighboring the target of recent and strong positive selection. Selective sweeps can be 'hard', where a single adaptive allele sweeps through the population, or 'soft', where multiple adaptive alleles at the same locus sweep through the population at the same time. (Messer & Petrov 2013)

Spatially heterogeneous balancing selection

The conceptual definition of temporary or spatially heterogeneous selection strictly differs from balancing selection (i.e. heterozygous advantage and frequency-dependent selection), but the concepts are similar in the aspect that selection maintains diversity and long term temporary or spatially heterogeneous selection are, therefore, generally considered forms of balancing selection.

Spliceosome

Spliceosomes are complex molecular machines that remove introns from transcribed pre-mRNA. Spliceosomes are assembled from small nuclear RNAs (snRNAs) and associated protein complexes. The RNA-protein complexes are called snRNPs. Spliceosomes assemble on the pre-mRNA strands after recognizing specific sequence elements (i.e. 5′ GU and 3′ AG splice site, polypyrimidine (uracil rich) tract and the branch point sequence).

Sympatric speciation

Sympatric speciation is the process through which species evolve from an ancestral species while inhabiting the same geographic region. In its most extreme use, the term refers to populations with identical ranges or panmictic (random) mating. However, these situation are rare or even absent in nature and, therefore, the term is often used to indicate that species evolved without the existence of geographical barriers and in the face of putatively ample gene flow (Mallet *et al.* 2009).

Tajima's *D*

The Tajima's *D* test (Tajima 1989) compares the total number of segregating sites to the average number of mutations between pairs of samples. More precisely, the standardized difference between the nucleotide diversity (π) and Watterson's θ_W, known as Tajima's *D*.

$$D = \frac{\hat{\pi} - \hat{\theta}_W}{\sqrt{Var(\hat{\pi} - \hat{\theta}_W)}}$$

A negative Tajima's *D* signifies an excess of low frequency polymorphisms relative to the expectation under neutrality, indicating population size expansion (e.g., after a bottleneck or a selective sweep) and/or purifying selection. A positive Tajima's *D* indicates low levels of both low and high frequency polymorphisms, indicating a decrease in population size and/or balancing selection

Wall's tests

Wall's tests were developed to detect events that produce trees with relatively longer external branches, such as under balancing or population structure. Wall's B and Q are based on so called congruent sites (B'), which are pairs of adjacent segregating sites (S) in a set of sequences that, if taken as a subset, form only two possible haplotypes among sequences (Wall 1999). Wall's B is calculated as:

$$B = \frac{B'}{(S-1)}$$

Wall's Q also includes the number of different partitions defined by the congruent sites (A). A partition is a subset of congruent segregating sites that divide the sequence set into the same subgroup of sequences. Wall's Q is calculated as:

$$Q = \frac{B + |A|}{S}$$

Watterson's θ (θ_W)

The Watterson estimator is a method for estimating the population mutation rate, $\theta = 4N_e\mu$, where N_e is the effective population size and μ the per-generation mutation rate (Watterson 1975). The estimate is

$$\hat{\theta}_W = \frac{K}{a_n}$$

where K is the number of segregating sites in the sample and

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

is a correction factor ((n-1)th harmonic number) for the number of segregating sites that do not completely sum up if the number of sequences in the sample increases.

ZZ test

The *ZZ* statistic provides information about intragenic recombination and compares the average linkage disequilibrium (*LD*) between adjacent sites with the average *LD* over all sites (Rozas *et al.* 2001). *ZZ* is calculated as:

$$ZZ = Z_A - Z_{ns}$$

With Z_{ns} using information of the r^2 value between all pairs of polymorphic sites (*S*):

$$Z_{ns} = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} r_{ij}^2$$

r^2 (measure of linkage disequilibrium) among a pair of loci (i and j) is defined as:

$$r_{ij}^2 = \frac{D_{ij}^2}{p_i(1-p_i)p_j(1-p_j)}$$

p_i and p_j are the frequencies of the polymorphic alleles i and j and D_{ij} is the measure of LD between both loci:

$$D_{ij} = p_{ij} - p_i p_j$$

Z_A is very similar to Z_{ns} but only takes into account r^2 values between adjacent polymorphic pairs (*S*):

$$Z_A = \frac{1}{S-1} \sum_{i=1}^{S-1} r_{ij}^2$$

ZZ is expected to become increasingly positive as recombination increases.

Appendix

Appendix 1. Sampled populations, including code, habitat, number genotyped (N), frequency of the mtIDH-B allozyme, percentage of maximal realizable wing size (%MRWS), standard deviation of %MRWS and reference to the publication in which the data has been used.

	Sampling site	Code	Habitat	N	Freq. mtIDH-B	%MRWS	%MRWS SD	Ref.
Atlantic								
United Kingdom	Severn estuary	SEE 1*	Tidal	58	0.34	37.33	6.90	1
	Severn estuary	SEE 2	Tidal	41	0.15	35.94	4.67	1
	Rye harbor	RYE	Tidal	41	0.06	31.94	3.99	1
	Morecambe	MOR	Tidal	5	0.00	32.50	7.64	1,2
	Exe estuary	EXE	Tidal	35	0.07	32.82	2.99	1
	Thorney	THO*	Tidal	42	0.37	37.99	7.33	1
The Netherlands	Friesland	FRI 1	Tidal	10	0.45	55.23	12.18	1,2
	Friesland	FRI 2*	Tidal	32	0.56	56.22	8.46	1,2
	Friesland	FRI 3	Tidal	10	0.60	51.39	9.60	1,2
	Friesland	FRI 4	Tidal	9	0.56	54.22	8.69	1,2
	Ossenisse	OSS	Tidal	14	0.21	42.41	9.73	2
Belgium	Heist	HEI	Temp	47	0.84	70.20	12.73	3
	Nieuwpoort	NIE*	Tidal	254	0.55	44.45	10.27	1,2
	Saeftinghe	SAE*	Tidal	132	0.35	42.34	8.11	2,4
	Braakman	BRA	seasonal	70	0.86	76.32	6.18	4
	Lissewege	LIS	seasonal	76	0.92	78.88	5.54	4
	Moeren	MOE	seasonal	73	0.91	78.87	5.54	4
	Molenkreek	MOK	seasonal	34	0.99	80.01	4.44	2,4
	Watervliet	WAT	seasonal	37	0.96	72.39	5.69	4
	Zwin	ZW*	Tidal	255	0.41	48.80	12.02	1,2
	Oostende	OOS	seasonal	159	0.92	79.46	5.87	1,2

	Sampling site	Code	Habitat	N	Freq. mtIDH-B	%MRWS	%MRWS SD	Ref.
France	Authie	AUT 1*	Tidal	40	0.59	49.54	13.65	5
	Authie	AUT 2*	Tidal	37	0.34	36.11	6.36	5
	Les Sables d'olonne	GAC	Tidal	34	0.12	37.22	7.25	5
	Canches, Etaples	CAN*	Tidal	47	0.29	35.15	7.15	5
	Baie de Veys	VEY 1*	Tidal	145	0.31	46.32	10.12	5
	Baie de Veys	VEY 2	Tidal	37	0.27	52.84	13.39	5
	Baie de Veys	VEY 3	Tidal	33	0.21	38.61	16.08	5
	Somme estuary	SOM 1	Tidal	90	0.22	32.89	5.46	1,2
	Somme estuary	SOM 2	Tidal	66	0.14	31.38	4.75	1,2
	Mont St Michel	MSM 1*	Tidal	27	0.65	63.35	12.41	1,2
	Mont St Michel	MSM 2*	Tidal	60	0.62	60.70	9.75	1,2
	Mont St Michel	MSM 3*	Tidal	62	0.62	60.31	9.59	1,2
	Mont St Michel	MSM 4*	Tidal	63	0.67	59.86	8.44	1,2
	La Guérande-oeillet	POND 1*	seasonal	102	0.60	63.64	10.59	3,5
	La Guérande-oeillet	POND 2*	seasonal	101	0.58	68.00	8.97	3,5
	La Guérande-oeillet	POND 3*	seasonal	86	0.57	53.79	11.29	3,5
	La Guérande-ethier	CANAL 1	Tidal	68	0.03	28.35	5.63	3,5
	La Guérande-ethier	CANAL 2	Tidal	67	0.03	28.86	5.58	3,5
	La Guérande-ethier	CANAL 3	Tidal	52	0.06	28.09	4.46	3,5
	Corsept	COR	Tidal	8	0.63	58.83	5.97	1
	Gironde estuary	GIR	Tidal	40	0.05	30.87	6.51	1
Spain	Coto Doñana	COD	seasonal	18	0.97	90.38	4.48	6
	Huelva	HUE	Tidal	16	0.03	24.10	2.39	6

	Sampling site	Code	Habitat	N	Freq. mtIDH-B	%MRWS	%MRWS SD	Ref.
Mediterranean								
France	Roussillon	ROU 1	Temp	11	1.00	95.03	3.29	1
	Roussillon	ROU 2	Temp	13	1.00	90.51	4.14	1
	Roussillon	ROU 3	Temp	7	1.00	92.56	3.62	1
	Camargue	CAM 1	Temp	11	1.00	86.07	4.83	3
	Camargue	CAM 2	Temp	10	1.00	90.54	5.21	3
	Camargue	CAM 3	Temp	3	1.00	83.89	3.47	3
	Camargue	CAM 4	Temp	5	1.00	92.52	3.71	3
	Toulon	TOU	Temp	28	1.00	84.07	4.70	3
	Montpellier	MON	Temp	3	1.00	89.50	4.29	6
Spain	Murcia	MUR 1	Temp	8	1.00	93.22	1.74	1
	Murcia	MUR 2	Temp	8	1.00	93.47	2.73	1
	Murcia	MUR 3	Temp	35	0.99	91.65	11.66	1
	Gata	GAT	Temp	12	0.96	88.85	4.74	6
	Albacete	ALB 1	Temp	32	1.00	92.00	3.66	1
	Albacete	ALB 2	Temp	35	1.00	92.11	3.76	1
	Albacete	ALB 3	Temp	35	1.00	91.35	3.34	1
	Almería	ALM 1	Temp	10	0.85	93.45	5.98	3
	Almería	ALM 2	Temp	10	1.00	89.97	5.32	3
	Almería	ALM 3	Temp	15	0.93	91.65	4.14	3
	Ibiza	IBI	Temp	29	0.91	84.53	6.05	3

** Populations used for the within populations association analysis between wing size and mtIDH*

1 (Desende & Serrano 1999) *3 (Dhuyvetter et al. 2004)* *5 (Dhuyvetter et al. 2007b)*

2 (Desender et al. 1998) *4 (Dhuyvetter et al. 2005)* *6 This study*

Appendix 2. *Pogonus* and *Pogonistes* species and populations sequenced for the genes considered in this study.

Species	Sampling site		Code	*mtIdh*				*cytIdh*	*enolase*	*cox1*	*nad1*	*cob*
				promoter	coding	intron	303 - 975					
Pogonistes convexicollis	Greece	Thessaloniki	THE	-	-	-	2	-	4	1	1	1
Pogonistes gracilis	France	Toulon	TOU	-	-	-	4	2	2	2	2	2
Pogonistes rufoaeneus	Greece	Ioninan Islands -Zakynthos	IOI	-	-	-	-	-	4	1	1	1
Pogonistes testaceus	France	Toulon	TOU	-	-	-	-	-	4	1	1	1
Pogonus gilvipes	Spain	Almería	ALM	-	-	-	4	4	4	1	1	1
Pogonus littoralis	France	Rousillon	ROU	-	-	-	4	4	4	1	1	1
Pogonus luridipennis	Austria	Illmitz	ILM	-	-	-	4	4	4	2	2	2
	France	Montpellier	MON	-	-	-	4	-	2	2	2	2
Pogonus meridionalis	Spain	Murcia	MUR	-	-	-	-	2	2	1	1	1
Pogonus olivaceus	Greece	Ioninan Islands -Zakynthos	IOI	-	-	-	4	6	4	2	2	2
Pogonus reticulatus	Greece	Thessaloniki	THE	-	-	-	4	4	2	2	2	2
Pogonus riparius	France	Montpellier	MON	-	-	-	4	4	4	2	2	2
Pogonus chalceus	The Netherlands	Friesland	FRI	4	4	4	4	-	2	-	-	-
	UK	Severn estuary	SEE	4	6	6	6	-	-	-	-	-
		Rye harbor	RYE	-	2	2	2	-	-	-	-	-
	Belgium	Zwin	ZWC	10	10	10	10	10	4	-	-	-
		Oostende	OOS	4	4	4	4	2	-	-	-	-
	France	La Guérande-ethier (canal)	GUE	8	10	10	10	20	4	1	1	1
		La Guérande-œillet (pond)	GUO	12	18	18	18	22	-	-	-	-
		Mont St Michel	MSM	14	14	14	14	12	4	-	-	-
		Roussillon	ROU	2	2	2	2	2	-	-	-	-
		Toulon	TOU	6	6	6	6	8	4	-	-	-
		Camargue	CAM	2	2	2	2	8	2	-	-	-
		Somme estuary	SOM	-	2	2	2	-	2	-	-	-
		Gironde estuary	GIR	-	6	6	6	-	-	-	-	-
	Portugal	Aveiro - Pond	AVE 1	6	6	6	6	-	-	-	-	-
		Aveiro - Marsh	AVE 2	4	4	4	4	-	-	-	-	-
	Spain	Coto Doñana	COD	8	10	10	10	8	4	-	-	-
		Huelva	HUE	8	8	8	8	8	6	-	-	-
		Gata	GAT	4	6	6	6	6	-	-	-	-
		Almería	ALM	4	6	6	6	6	2	-	-	-
		Ibiza	IBI	4	4	4	4	4	-	-	-	-
		Albacete	ALB	-	-	-	-	4	-	-	-	-

GenBank accession numbers:(*mtIdh*: KJ371353 - KJ371522; *cytIdh*: KJ371166 - KJ371315; *enolase*: KJ371316 - KJ371352; *cox1*: KJ371146 - KJ371165; cob: KJ371126 - KJ371145; *nad1*: KJ371523 - KJ371542.

cDNA SEQUENCING

Total RNA was extracted from adult *P. chalceus* beetles using the RNeasy Plus Mini Kit (Qiagen Inc.) according to the manufacturer's instructions. cDNA synthesis used 2 μg of total RNA, oligo(dT)18 primer and RevertAid™ H Minus First Strand cDNA Synthesis Kit (Fermentas GmbH Inc.) according to the manufacturer's instructions. Initially we used RNA from one individual to clone and sequence the cDNA copy of both the mitochondrial *NADP+-Idh* (*mtIdh*) and cytoplasmic *NADP+-Idh* (*cytIdh*) gene.

To obtain the sequence of the *cytIdh* gene a degenerate PCR primer pair was obtained using Primo Degenerate 3.4 (Chang Bioscience Inc.) by aligning homologous sequences of nine other insect species from GenBank (*Acyrthosiphon pisum*; XP_001946553, *Aedes aegypti*; XP_001650675, *Anopheles gambiae*; XM_001688896, *Bombyx mori*; NP_001040134, *Culex quinquefasciatus*; XM_001841858, *Drosophila melanogaster*; NP_001137910, *Gryllus firmus*; ABI52605, *Nasonia vitripennis*; XP_001608101, *Tribolium castaneum*; XM_963757). The degenerate primer pair (Appendix 4) amplified a cDNA fragment of 733 bp (excluding primers). PCR was carried out for 40 cycles using the following conditions: denaturation at 95° for 1 min, annealing at 55° for 1 min, and extension at 72° for 1 min. Next, the amplified fragment was cloned into One Shot® TOP10 Chemically Competent *E. coli* cells (Invitrogen) and sequenced.

A degenerate PCR primer pair to amplify the *mtIdh* gene was obtained with the iCODEHOP program (Boyce *et al*., 2009) by aligning mitochondrial *NADP+-Idh* sequences of eight other arthropod species (*Acyrthosiphon pisum*; XM_001943663, *Anopheles gambiae*; XM_312860, *Bombyx mori*; NM_001099620, *Culex quinquefasciatus*; XM_001844978, *Dendroctonus ponderosae*; BT127538, *Drosophila melanogaster*; NM_001144438, *Ixodes scapularis*; XM_002409629, and *Tribolium castaneum*; XM_965353). The degenerate primer pair (Appendix 4) amplified a cDNA fragment of 600 bp (excluding primers), which was cloned into One Shot® TOP10 Chemically Competent *E. coli* cells (Invitrogen Inc.) and sequenced. The iCODEHOP primer pair also amplified the cytoplasmic *Idh* gene.

Based on the partial sequence of the *cytIdh* and *mtIdh* gene, gene-specific primers were designed allowing the identification of 5′ and 3′ ends of the *cytIdh* and *mtIdh* mRNA by a RACE protocol (Roche, Inc.). Gene specific primers for 5′RACE and 3′RACE can be found in Appendix 4. Cloned RACE products yielded the remaining coding sequence of the *cytIdh* and *mtIdh* gene. In both cases the 5′ end was obtained by three rounds of nested PCR. After cDNA synthesis using a first gene specific primer, purified cDNA was used for poly(A) tailing at the 5′end using terminal transferase. Next, a second (nested) PCR was performed with a anchored oligo d(T) primer (5′-GAC CAC GCG TAT CGA TGT CGA CTT TTT TTT TTT TTV-3′) and a second gene specific primer. The product of a third PCR round using the anchor primer (5′-GAC CAC GCG TAT CGA TGT CGA C-3′) and a third gene specific primer was cloned into One Shot® TOP10 Chemically Competent *E. coli* cells (Invitrogen Inc.) and sequenced.

The 3′ ends of the *cytIdh* and *mtIdh* genes were obtained by cDNA synthesis with the anchored oligo d(T) primer, followed by two consecutive rounds of PCR using the anchor primer and two nested gene specific primers.

GENOMIC SEQUENCING

Sequence variation analysis was performed on the genomic sequence of the *mtIdh* and *cytIdh* gene. DNA extractions were performed using the DNA extraction NucleoSpin® Tissue kit (Macherey-Nagel GmbH). PCR were run in a Tpersonal thermal cycler (Biometra®). Several gene specific primer sets were designed based on the full *mtIdh* and *cytIdh* mRNA and UTR sequences (Appendix 4). The genomic sequence of the whole *cytIdh* gene was amplified with one primer pair (F1-R3). Three internal primers were designed for sequencing.

A previous transcriptome sequencing project (Van Belleghem *et al.* 2012) identified a transcript that is homologous to the $NADP^+$-transhydrogenase (*Nnt*) gene and was found upstream to the *mtIdh* gene in the *Tribolium castaneum* genome (Richards *et al.* 2008). Moreover, in *T. castaneum* the *Nnt* and mitochondrial *$NADP^+$-Idh* gene appear in a head-to-head arrangement, i.e. facing away from one another, are separated less than 1,000 bp and are transcribed from opposite strands of DNA. Therefore, the *Nnt* and mitochondrial *$NADP^+$-Idh* gene are most likely transcribed by a bidirectional promoter and, therefore, their position may be conserved in *P. chalceus* (Adachi & Lieber 2002). Using a primer in the 3′ coding region of the *Nnt* gene and the 5′ coding region of the mitochondrial *$NADP^+$-Idh* gene allowed amplifying and sequencing both the 5′ coding region and the promoter region of the *mtIdh* gene from genomic DNA.

For all PCR amplification reactions, Platinum Taq DNA polymerase (Invitrogen) was used according to manufacturer's instructions. All amplification reactions were carried out for 40 cycles using the following conditions: denaturation at 95° for 1 min, annealing at 60°C for 1 min, and extension at 72° for 1 min. The PCR products were purified and sequenced in both directions using an ABI Prism® BigDye® V 1.1 Terminator Cycle Sequencing kit. All sites were scored at least twice and heterozygous positions were only scored when both forward and reverse direction sequencing reads were consistent. Nucleotide sequences were aligned directly on the sequencer output files and edited if necessary by using the sequence analysis and alignment software SEQSCAPE (Version 2.5; Applied Biosystems, Inc.). Haplotypes were unphased with the PHASE algorithm (Stephens *et al.* 2001) implemented in DNAsp v5.0 software (Librado & Rozas 2009).

STRUCTURAL ANALYSIS

To localize the amino acid changes in the protein structure, we used the crystal structure of porcine mitochondrial $NADP^+$-IDH (PDB: 1lwd; 73.71% sequence identity for both long and short splice variant) (Ceccarelli *et al.* 2002) and mouse cytoplasmic $NADP^+$-IDH (PDB: 2cmj; 77.89% sequence identity) to build a model for *P. chalceus* mitochondrial and cytoplasmic $NADP^+$-IDH respectively using the web based Swiss-Model program (Bordoli *et al.* 2009).

Appendix 4. Sequencing primers.

	Forward (5′- 3′)	Reverse (5′-3′)	Region	Ref.
mtIdh				
Degenerate primer pair (iCODEHOP)	*GCCCATGGCGACCARTAYAARGC*	*GCGAAGATGGAGGCGAYNGGRTTNGT*		
Gene specific primers 5′RACE		*1:AGGCCAACGTTTTTGAAGTGCCA* *2:AACACCACCAGCTTTGTAAGTG* *3:CGAGTTCCACCTTTCCTGGATTGGT*		
Gene specific primers 3′RACE	*1:TCGCTCAGGGTTATGGGTC* *2:CAGAAGCTGCACATGGCAC*			
Gene specific primers (for mRNA amplification)	*1:TGCTGTGTGTCATTTGTAATAG* *2:ACAATAGATGCTGCTCATGC* *3:CAATCCAGGAAAGGTGGAACTCG* *4:TTGATGATATGGTGGCACAAGC*	*1:TCTACGGCTTTGTATTGATCAC* *2:GACAAGTACAAAGGCCAACG* *3:TCTAAACCACGGGTCCATGC* *4:CATTCATAAACCATAGCAAATTTTCG*		
Gene specific primers (for genomic amplification)	*1:CGCTGTGCAAGTTTAGCGTA* *2:ATTAAGACTGATGCACTCTATG* *3:TTCCGGCTATTTACGTTCTTG* *4:AGACCAAGAAGTACTTCTACAC* *5:ACAATAGATGCTGCTCATGC* *6:GCTGGTGGTGTTGCAATGGGC* *7:TTGATGATATGGTGGCACAAGC*	*1:TTTTTGTAAAAGCACAGGAGCA* *2:CTTAAATTGCTGACAGAAAATTG* *3:TGTTAAGCGAATAAGTCTCG* *4:TCTACGGCTTTGTATTGATCAC* *5:GACAAGTACAAAGGCCAACG* *6:CCCAATGACCCATAACCCTGAGCG* *7:CATTCATAAACCATAGCAAATTTTCG*	Promoter E1 E2 I2/E3/I3/E4 I3/E4/I4/E5 E5/I5/E6 E6/I6/E7/I7/E8	
cytIdh				
Degenerate primer pair (Primo)	*RTCYTNGGNGAYGARATGAC*	*YTTRCANGCCCANACRAANC*		
Gene specific primers 5′RACE		*1:TCCCATAGCAACACCGGGTCCT* *2:TGTGGGCTCGCCATTTTCAGGA* *3:TCAACCCGATTCTCGTCTGGTGT*		
Gene specific primers 3′RACE	*1:TCGCTCATTCATCATTCCAATAC* *2:ACTGCCTTCGAGGCTAAGAAAATCTG*			
Gene specific primers (for genomic amplification)	*1:TCCATCGTTAATCCATCGCAAC* *2:GGGCACAGTATTCCGTGAAGCA* *3:TGGCGGTTTTGTCTGGGCTTG*	*1:TCCCATAGCAACACCGGGTCCT* *2:TGGCGGGTAACAGTACCATGAGCA* *3:TTGCGTGAACCCGTTCACC*		
enolase	*GACTCTCGTGGNAAYCCNACNGTNGAGGT*	*CTTGTAGAACTCNGANGCNGCNACRTCCAT*		1
cox1	*1:GGTCAACAAATCATAAAGATATTGG* *2:GAGCTCCTGATATAGCTTTTCC*	*1:TAAACTTCAGGGTGACCAAAAAATCA* *2:GGATAATCAGAATATCGTCGAGG*		2 3
nad1	*GCATCACAAAAGGCTGAGGA*	*ACATGATCTGAGTTGAAACC*		4
cob	*TATGTACTACCATGAGGACAAATATC*	*ATTACACCTCCTAATTTATTAGGAAT*		4

1 (Wild & Maddison 2008)
2 (Folmer *et al.* 1994)
3 (Simon 1994)
4 (Clarke *et al.* 2001)

Appendix 5. Coalescent simulations.

MSMS was run as follows:

msms 100 1 -N 100000 -t [*mutation rate*] -r [*recombination rate*] -I 2 50 50 -n 1 [*subpopulation size*] -ma x [*migration rate population 1 to 2*] [*migration rate population 2 to 1*] x -Sc 0 1 [*selection coefficient*] -Sc 0 2 -[*selection coefficient*] -Smu 0.01 -SI [*selection time*] 2 0 0 –s 80 –Smark

Parameters were varied as follows:

Mutation rate = (13.2)
Recombination rate (R) = (20.7)
Subpopulation size = (1, 0.5, 0.1)
Population migration rates (M) = random.uniform(1, 1000)
Selection coefficient (S_s) = (1000, 5000, 10000, 50000)
Selection time (S_t) = random.uniform(0.01,2)

mtIDH ELECTROMORPH (EM) CLASSES

Non-synonymous substitutions in the *mtIdh* sequencing dataset corresponded to the allozymes found of the mtIDH protein by gel electrophoresis (Desender & Serrano 1999, Dhuyvetter *et al.* 2004). The differentially selected mtIDH-B and mtIDH-D allele are distinguished by only a single charge-changing amino acid substitution (Lys - Asn) at amino acid position 447. The mtIDH-A allozyme differentiates from the mtIDH-B allozyme by a charge changing amino acid substitution (Asp - Asn) at amino acid position 391. The mtIDH-E allozyme differentiates from the mtIDH-D allozyme by a charge changing amino acid mutation (Glu - Gly) at amino acid position 200. Within the mtIDH-B allozyme class we found seven amino acid substitutions, two of which are frequent and each comprises 31% of the mtIDH-B samples. In both the mtIDH-D (Cys - Tyr) and mtIDH-E (Leu - Phe) class we found one rare amino acid variant that did not result in a charge change (Figure 3D in main article). Haplotypes associated with the mtIDH-C allozyme class were identical to haplotypes found within one of the previous classes, and therefore likely constitute errors in allozyme scoring. Individuals with the rare mtIDH-A, mtIDH-C and mtIDH-E allozymes, constitute 0.0020, 0.0016 and 0.0016 % of all sampled alleles respectively (Van Belleghem & Hendrickx 2014). Sequence conservation of this enzyme across diverse taxa allowed constructing an approximate structural model for the *P. chalceus* mtIDH protein based on homology with porcine *mtIdh* for which high-resolution structures are known (Appendix 12). The protein model shows that all of the identified amino acid polymorphisms are found at or near the enzymes surface. The amino acid changes at position 22, 35 and 36 are located in the putative transit peptide and were not included in the protein model.

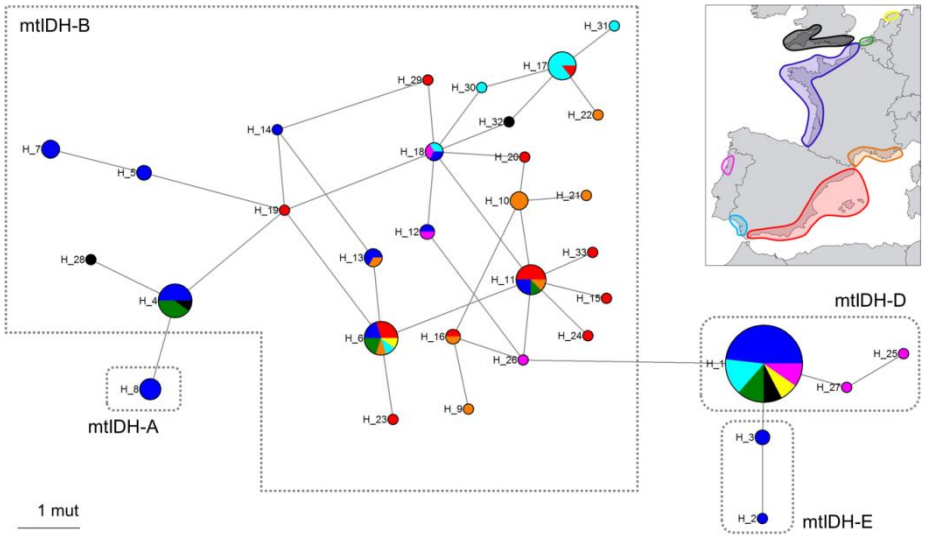
CYTOPLASMIC NADP$^+$-IDH (cytIDH) GENE STRUCTURE

The coding sequence of the cytoplasmic *NADP$^+$-Idh* gene has a length of 1,227 bp (Appendix 8A-B). 77 bp of the 5' untranslated region (UTR) and 87 bp of the 3'UTR up to the poly-A tail were also obtained. Sequencing genomic DNA identified one intron in the 5'UTR of 147 bp. The resulting protein contains 408 residues. As for the mtIDH protein, the first 48 bp of the coding mRNA or 16 amino acids at the N-terminus of the protein of the *cytIdh* sequence show little homology with the protein sequence of *cytIdh* of other eukaryotes and potentially form a transit peptide.

cytIDH ELECTROMORPH (EM) CLASSES

The rare cytIDH-A allozyme differentiates from the cytIDH-B allozyme by a charge changing amino acid change (Gly – Arg) at amino acid position 367 (Appendix 8C). In the cytIDH-B allozyme class we found four amino acid substitutions, one of which is frequent (96%). Two different amino acid substitution result in two different proteins characterized as the cytIDH-C allozyme, one at amino acid position 342 (Asn – Lys) and one at amino acid position 404 (Lys – Gln). As for the mtIDH protein, the protein model of cytIDH shows that all segregating amino acid sites found occur near the enzymes surface (Appendix 13).

Appendix 7. Median joining haplotype network for the coding *mtIdh* sequences. The haplotypes belonging to the mtIDH-A, mtIDH-B, mtIDH-D and mtIDH-E allozymes are indicated. Size of the pie charts indicates the relative frequency of the haplotypes. Colors in the network match with the shaded areas on the map of Europe.

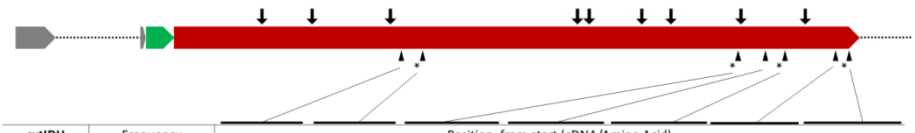


Appendix 8. Genomic structure and amino acid variation of *P. chalceus cytIdh* gene. (A.) Length (bp) of the exons, introns, promoter and transit peptide for the *cytIdh* gene. (B.) Scaled diagram of the *P. chalceus cytIdh* gene, showing exons, introns, promoter and transit peptide. Arrows mark the midpoint of detected intragenic recombination. (C.) Positions of non-synonymous nucleotide substitutions along the cDNA within the *cytIdh* gene. Amino acid names are according the IUPAC code. Charge-changing amino acid variants defining the EM classes are indicated with an asterisk.

A. Cytoplasmic NADP⁺-dependent isocitrate dehydrogenase

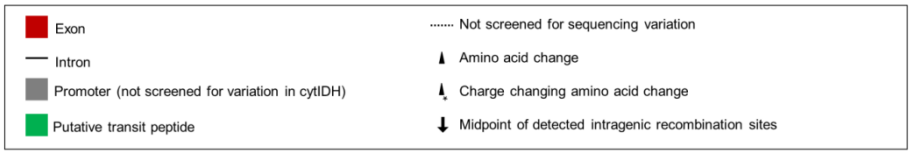
***cytIdh* gene**	5' UTR1	5'UTR inton	5'UTR2	**Exon 1**	3'UTR
Transit peptide				1-48	
Length (bp)	68	147	9	1,224	90

B.

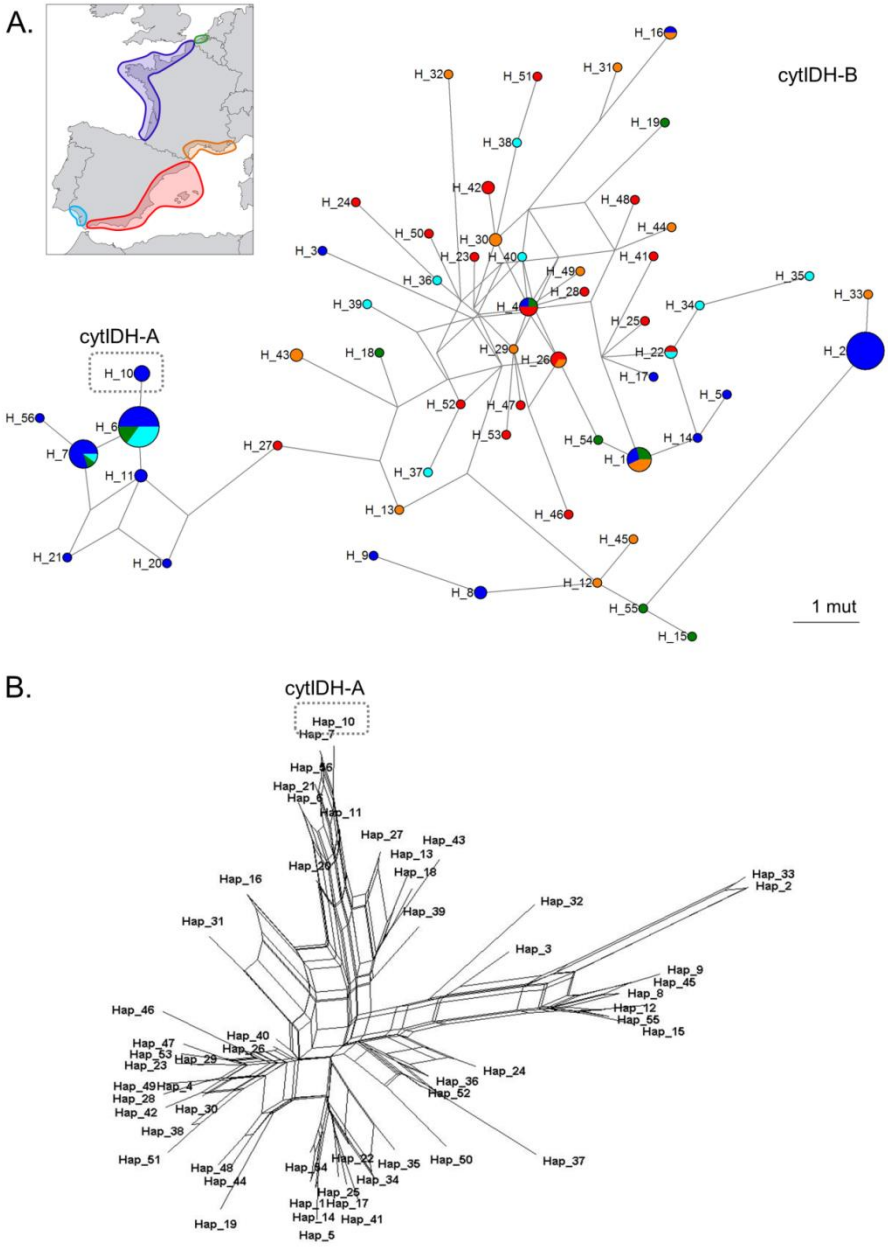


C.

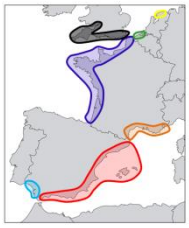
cytIDH	Frequency	Position from start (cDNA/Amino Acid)						
allozyme	Total (within)	436/146	475/159	1,026/342	1,072/358	1,099/367	1195/399	1,210/404
A	3 / 0.03	Pro	Asn	Lys	Ala	Arg	Ala	Lys
B	109 / 0.91	.	.	.	.	Gly	.	.
	3 / 0.03	.	His	.	.	.	.	.
	2 / 0.02	.	.	.	Thr	.	.	.
	1 / 0.01	Ser	.	.	.	.	Ser	.
C	1 / 0.01	.	.	Asn	.	.		.
	1 / 0.01	.	.	.	.	.		Gln

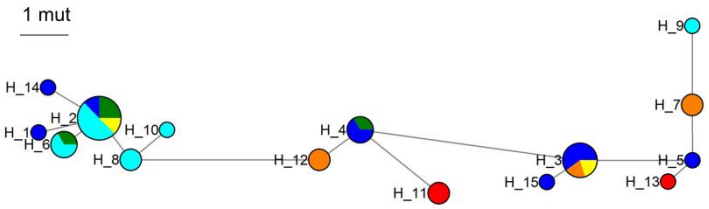


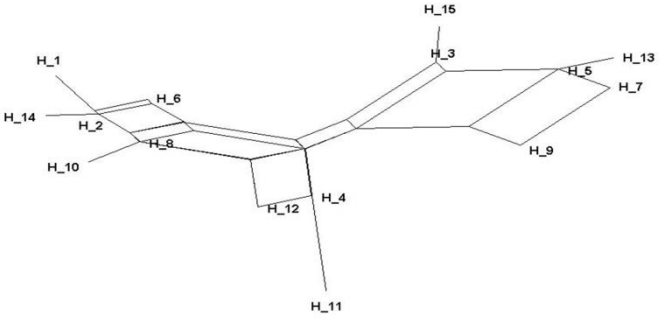
Appendix 9. Median joining (A.) and Neighbor-Net (B.) network for the *cytIdh* gene. The haplotypes belonging to the cytIDH-A and cytIDH-B are indicated. The cytIDH-C haplotypes (H_47 and H_50) are not highlighted in the graph as this EM class does not form a monophyletic group. Size of the pie charts indicates the relative frequency of the haplotypes. Colors in the network match with the shaded areas on the map of Europe.



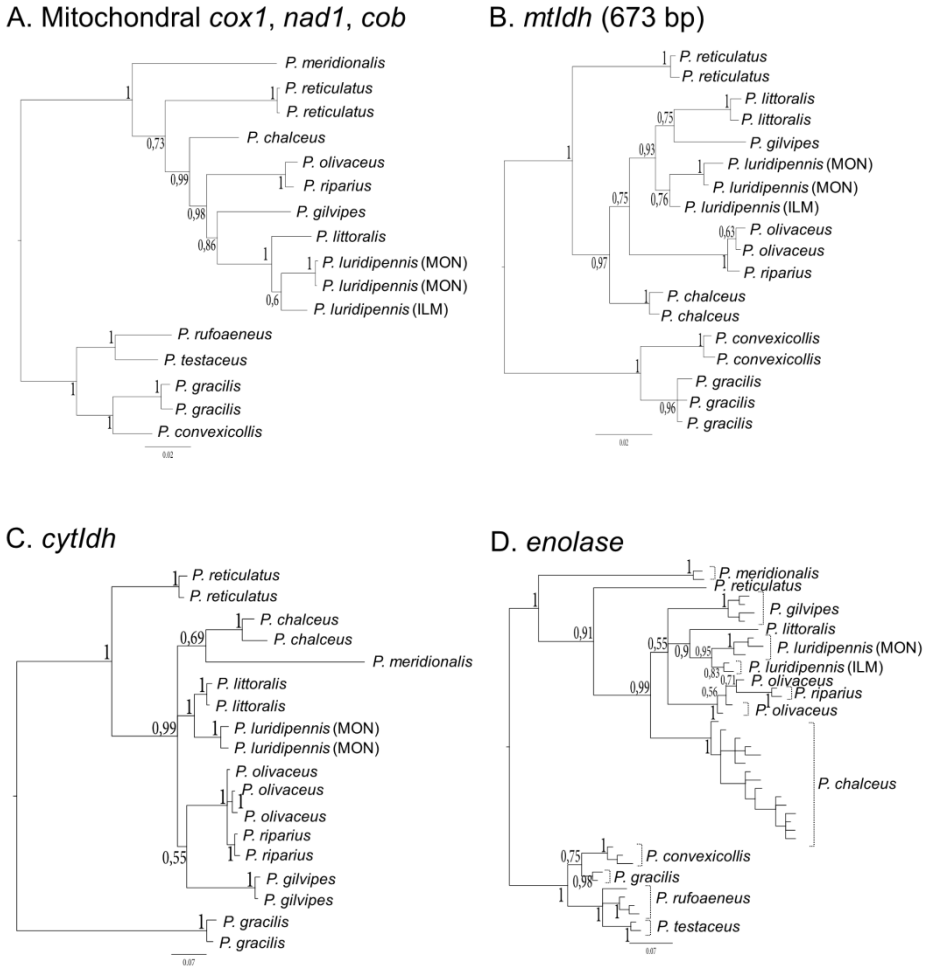
Appendix 10. Median joining (above) and Neighbor-Net (below) network for part of the *enolase* gene. Size of the pie charts indicates the relative frequency of the haplotypes. Colors in the network match with the shaded areas on the map of Europe.



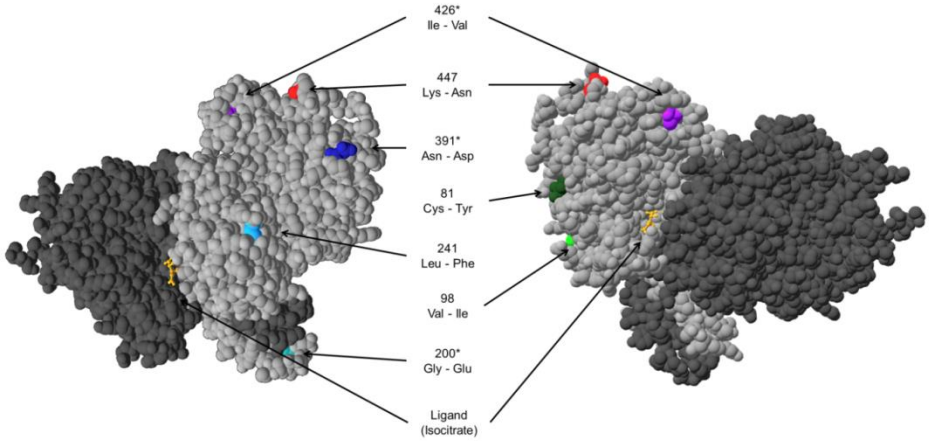




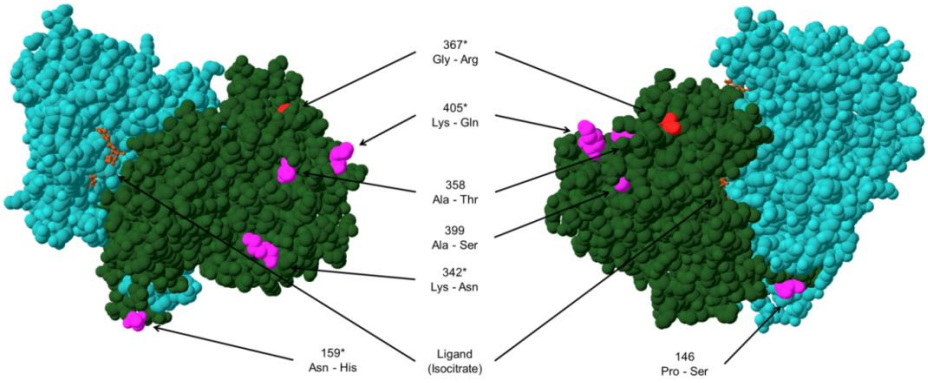
Appendix 11. Species trees. MrBayes trees were constructed to confirm the phylogenetic relations among several *Pogonus* and *Pogonistes* species. Numbers indicate posterior probability values.



Appendix 12. Structure of the mtIDH enzyme in *Pogonus chalceus* and location of amino acid substitutions. Amino acid substitution at residue 22, 35 and 36 of the protein are located in the putative transit peptide and were not included in the homology based protein model. All segregating amino acid sites found occur near the enzyme surface.



Appendix 13. Structure of the cytIDH enzyme in *Pogonus chalceus* and location of amino acid substitutions. All segregating amino acid sites found occur near the enzyme surface.



Appendix 14. *Pogonus chalceus* sequencing read statistics. Error corrected reads were used for SOAPdenovo2 assembly.

					Raw reads			Cleaned (and error corrected)			
Individual	sex	Library	Insert (bp)	GC %	Length (bp)	Reads (M)	Bases (Gb)	Length (bp)	Reads (M)	Bases (Gb)	Corrected bases (Mb)
GC3b_01	m	Paired-end	200	28	101	107.24	10.83	100.04	107.06	10.71	2.18
GC3b_02	m	Paired-end	200	28	101	72.84	7.36	99.92	72.66	7.26	1.89
GC3b_03	m	Paired-end	200	28	101	53.92	5.46	99.70	53.76	5.36	1.59
GC3b_04	m	Paired-end	200	28	101	101.3	10.23	99.97	101.13	10.11	2.40
GC3b_055L10	f	Paired-end	500	28	100	79.46	7.95	99.17	79.26	7.86	2.29
GC3b_055L10	f	Paired-end	800	28	100	61.74	6.17	99.10	61.54	6.10	2.44
GC3b_055L10	f	Mate-pair	2,000	29	49	112.18	5.50	47.34	103.26	4.89	2.57
GC3b_003	f	Mate-pair	5,000	29	49	87.14	4.27	47.24	79.78	3.77	1.77
Total						675.82	57.77		658.45	56.06	17.13

Appendix 15. Bacterial and viral genome assembly contamination.

Species	**Number (%)**
Rickettsia sp.	106 (77.37)
Wolbachia sp.	14 (10.22)
Staphylococcus aureus	6 (4.38)
Staphylococcus phage	6 (4.38)
Other bacterial species	6 (4.38)
Total	**137**

Appendix 16. Repeat analysis.

	Number of elements	Length occupied (bp)	Percentage of sequence
RepeatMasker (RepBase)	202,288	11,251,189	3.818
Retroelements	491	158,654	0.061
SINEs:	0	0	0.000
Penelope	0	0	0.000
LINEs:	46	9,532	0.004
CRE/SLACS	0	0	0.000
L2/CR1/Rex	10	879	0.000
R1/LOA/Jockey	33	8,457	0.003
R2/R4/NeSL	3	196	0.000
RTE/Bov-B	0	0	0.000
L1/CIN4	0	0	0.000
LTR elements:	445	149,122	0.057
BEL/Pao	156	62,423	0.024
Ty1/Copia	31	10,344	0.004
Gypsy/DIRS1	258	76,355	0.029
Retroviral	0	0	0.000
DNA transposons	1,324	330,253	0.126
hobo-Activator	9	1,953	0.001
Tc1-IS630-Pogo	108	21,913	0.008
En-Spm	0	0	0.000
MuDR-IS905	0	0	0.000
PiggyBac	0	0	0.000
Tourist/Harbinger	0	0	0.000
Other (Mirage, P-element, Transib)	0	0	0.000
Rolling-circles	0	0	0.000
Unclassified:	1	60	0.000
Total interspersed repeats:		488,967	0.187
Small RNA:	311	76,572	0.029
Satellites:	3	594	0.000
Simple repeats (TRF):	150,413	7,116,659	2.718
Low complexity:	45,593	2,317,299	0.885
RepeatScout	2,414	38,707,718	14.784
Total	204,702	49,958,907	18.602

Appendix 17. Mean number of reads after demultiplexing and quality filtering (demultiplexed) and reads after removing PCR duplicates (purged) for each individual grouped per location.

Population		N	Demultiplexed		Purged				
			Mean reads	SD	Mean reads	SD	Min reads	Max reads	Mean % recov
Belgium - Dudzele	DUD	24	524,627	555,103	194,354	85,618	78,098	342,773	37
Belgium - Nieuwpoort	NIE	24	547,920	269,601	259,084	144,943	66,232	547,737	47
France - Guérande - Canal	GUE	24	753,231	206,938	282,459	100,688	128,326	486,432	37
France - Guérande -Pond	GUO	24	659,047	335,820	133,526	49,333	64,238	231,941	20
Portugal - Aveiro-Pond	AVE1	8	661,608	499,593	287,701	148,589	112,884	578,683	43
Portugal - Aveiro-Schor	AVE2	8	909,661	291,523	389,655	96,498	254,176	254,176	43
Spain - CotoDonana	COD	8	610,113	281,236	158,505	58,810	71,523	226,805	26
Spain - Huelva	HUE	8	644,734	360,228	168,222	86,878	36,619	272,148	26
France - Camargue	CAM	8	1,176,254	567,447	364,114	114,637	230,003	569,458	31
UK - SevernEstuary	SEE	8	482,976	465,222	154,872	110,702	61,163	408,085	32
Total		144	95,518,562		33,050,689				

Appendix 18. Nucleotide diversity (π) and heterozygosity (H_z) within each population from loci built de novo and using the *P. chalceus* genome assembly.

	De novo		**Reference genome**	
Population	π	H_z	π	H_z
Belgium - Dudzele	0.0022	0.0017	0.0050	0.0048
Belgium - Nieuwpoort	0.0024	0.0017	0.0050	0.0046
France - Guérande - Canal	0.0022	0.0015	0.0046	0.0043
France - Guérande -Pond	0.0023	0.0016	0.0050	0.0046
Portugal - Aveiro-Pond	0.0025	0.0019	0.0058	0.0054
Portugal - Aveiro-Schor	0.0019	0.0015	0.0048	0.0049
Spain - CotoDonana	0.0029	0.0020	0.0063	0.0055
Spain - Huelva	0.0026	0.0019	0.0057	0.0054
France - Camargue	0.0029	0.0020	0.0062	0.0050
UK - SevernEstuary	0.0024	0.0018	0.0053	0.0054

Appendix 19. List of wing development genes found in the *P. chalceus* transcriptome and genome. Fst values are given for scaffolds in which we found a RAD tag (stack ID) for the comparison between the canal and pond (CP) population and between the Nieuwpoort and Dudzele (ND) population. High Fst values are indicated in bold. Dist. = distance between wing development gene and RAD tag.

Gene	**Accession *P. chalceus***	**Genome**	**Stack ID**	**Dist. (kb)**	**Fst (CP/ND)**
Engrailed/ Invected	Pc_comp5821_c0_seq1	scaffold5175	-	-	-
Hedgehog	Pc_comp8905_c0_seq1	scaffold1402	-	-	-
Cubitus interruptus	Pc_comp4719_c0_seq1	scaffold3209	1181	69.35	-/0.02
Patched	Pc_comp7372_c1_seq1	scaffold1529	932	337.87	0.01/0
Decapentaplegic	Pc_comp8429_c0_seq2	scaffold1908	-	-	-
Daughters against	Pc_comp5722_c0_seq1	scaffold1054	-	-	-
Brinker	Pc_comp8966_c0_seq1	scaffold4806	-	-	-
Optomotor-blind-like	Pc_comp6103_c0_seq1	scaffold4718	-	-	-
Spalt-like protein	Pc_comp7794_c0_seq1	scaffold3155	-	-	-
Apterous a	Pc_comp9155_c1_seq1	scaffold3204	-	-	-
Apterous b	Pc_comp10531_c0_seq1	scaffold1186	-	-	-
Notch	Pc_comp3149_c0_seq1	scaffold68	710	241.11	0.04/0.06
Serrate	Pc_comp6451_c0_seq1	scaffold352	-	-	-
Wingless	Pc_comp9580_c0_seq1	scaffold288	390	174.40	**0.45/0.17**
Distal-less	Pc_comp7089_c0_seq1	scaffold4162	-	-	-
Serum response factor	Pc_comp3744_c0_seq2	scaffold1599	-	-	-
Rhomboid	Pc_comp9713_c0_seq1	scaffold1552	-	-	-
Knirps	Pc_comp8029_c0_seq2	scaffold1878	-	-	-
Knot transcription factor	Pc_comp14479_c0_seq1	scaffold3686	1260	1.28	0.02/0.05
Iiroquois	Pc_comp4855_c0_seq2	scaffold4868	1533	139.55	0/0.01
Abrupt	Pc_comp3738_c0_seq3	scaffold5117	1548	41.59	0.07/0
Noradrenaline transporter	Pc_comp9252_c0_seq1	scaffold1582	-	-	-
Delta	Pc_comp8811_c0_seq1	scaffold2089	-	-	-
Extramacrochaetae	Pc_comp778_c0_seq1	scaffold54	-	-	-
Achaete-scute	Pc_comp5966_c0_seq1	scaffold6174	-	-	-
Asense	Pc_comp12489_c0_seq1	scaffold3902	-	-	-
Teashirt	Pc_comp7294_c0_seq1	scaffold2744	-	-	-
Homothorax	Pc_comp2739_c0_seq1	scaffold4679	-	-	-
Nubbin	Pc_comp7766_c0_seq1	scaffold3320	1203	67.88	**0.41**/0.08
Ventral vein lacking	Pc_comp4049_c0_seq1	scaffold1757	-	-	-
Vestigial	Pc_comp7899_c0_seq1	scaffold794	-	-	-
Sex combs reduced	Pc_comp5657_c0_seq1	scaffold1990	1948	104.41	**0.16/0.28**
Prothoraxless	Pc_comp8727_c0_seq1	-	-	-	-
Ultrabithorax	Pc_comp6090_c0_seq1	Scaffold518	1561	60.50	0.03/0.02

REFERENCES

Abouheif E, Wray GA (2002) Evolution of the gene network underlying wing polyphenism in ants. *Science* 297: 249–252.

Adachi N, Lieber MR (2002) Bidirectional gene organization: A common architectural feature of the human genome. *Cell* 109: 807–809.

Allen Orr H (2001) The genetics of species differences. *Trends in ecology & evolution* 16: 343–350.

Alp PR, Newsholme EA, Zammit VA (1976) Activities of citrate synthase and NAD+-linked and NADP+-linked isocitrate dehydrogenase in muscle from vertebrates and invertebrates. *The Biochemical journal* 154: 689–700.

Altschul SF, Gish W, Miller W, Myers EW, Lipman. DJ (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215: 403–410.

Arendt J, Reznick D (2008) Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends in ecology & evolution* 23: 26–32.

Arnegard ME, McGee MD, Matthews B *et al.* (2014) Genetics of ecological divergence during speciation. *Nature* 511: 307–311.

Arnold ML, Martin NH (2009) Adaptation by introgression. *Journal of biology* 8: 82.

Ashburner M, Ball CA, Blake JA *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature genetics* 25: 25–29.

Aukema B (1990) Wing-length determination in two wing-dimorphic *Calathus* species (Coleoptera: Carabidae). *Hereditas* 113: 189–202.

Aukema B (1995) The evolutionary significance of wing dimorphism in carabid beetles (Coleoptera: Carabidae). *Research on Population Ecology* 37: 105–110.

Bai X, Mamidala P, Rajarapu SP, Jones SC, Mittapalli O (2011) Transcriptomics of the bed bug (*Cimex lectularius*). *PloS one* 6: e16336.

Bandelt HJ, Forster P, Ro A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution* 16: 37–48.

Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nature reviews. Genetics* 12: 767–80.

Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. *Trends in ecology & evolution* 23: 38–44.

Bateson W (1922) Evolutionary faith and modern doubts. *Science* 55: 55–61.

Beaumont MA (2010) Approximate Bayesian Computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics* 41: 379–406.

Bell MA, Aguirre WE (2013) Contemporary evolution, allelic recycling, and adaptive radiation of the threespine stickleback. *Evolutionary Ecology Research* 15: 377–411.

Bell MA, Aguirre WE, Buck NJ (2004) Twelve years of contemporary armor evolution in a threespine stickleback population. *Evolution* 58: 814–824.

Bellés X, Martín D, Piulachs M-D (2005) The mevalonate pathway and the synthesis of juvenile hormone in insects. *Annual review of entomology* 50: 181–199.

Beltman JB, Haccou P (2005) Speciation through the learning of habitat features. *Theoretical population biology* 67: 189–202.

Beltman JB, Haccou P, ten Cate C (2004) Learning and colonization of new niches: a first step toward speciation. *Evolution* 58: 35–46.

Beltman JB, Metz JAJ (2005) Speciation: more likely through a genetic or through a learned habitat preference? *Proceedings of the Royal Society B* 272: 1455–1463.

Bengtsson BO (1978) Avoiding inbreeding: at what cost? *Journal of theoretical biology* 73: 439–444.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* 27: 573–580.

Bentley DR, Balasubramanian S, Swerdlow HP *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53–59.

Berlocher SH, Feder JL (2002) Sympatric speciation in phytophagous insects: moving beyond controversy? *Annual review of entomology* 47: 773–815.

Bierne N, Roze D, Welch JJ (2013) Pervasive selection or is it…? Why are FST outliers sometimes so frequent? *Molecular ecology* 22: 2061–2064.

Bitume E V, Bonte D, Ronce O *et al.* (2013) Density and genetic relatedness increase dispersal distance in a subsocial organism. *Ecology letters* 16: 430–437.

Bolnick DI, Fitzpatrick BM (2007) Sympatric speciation: Models and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics* 38: 459–487.

Bolnick DI, Near TJ (2005) Tempo of hybrid inviability in centrarchid fishes (Teleostei: Centrarchidae). *Evolution* 59: 1754–1767.

Bolnick DI, Snowberg LK, Patenia C *et al.* (2009) Phenotype-dependent native habitat preference facilitates divergence between parapatric lake and stream stickleback. *Evolution; international journal of organic evolution* 63: 2004–2016.

Bonte D, Van Dyck H, Bullock JM *et al.* (2012) Costs of dispersal. *Biological reviews of the Cambridge Philosophical Society* 87: 290–312.

Bordoli L, Kiefer F, Arnold K *et al.* (2009) Protein structure homology modeling using SWISS-MODEL workspace. *Nature protocols* 4: 1–13.

Bouckaert R, Heled J, Kühnert D *et al.* (2014) BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS computational biology* 10: e1003537.

Braendle C, Davis GK, Brisson J a, Stern DL (2006) Wing dimorphism in aphids. *Heredity* 97: 192–199.

Brawand D, Wagner CE, Li YI *et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature*.

Brisson J a (2010) Aphid wing dimorphisms: linking environmental and genetic control of trait variation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 365: 605–616.

Brisson J a, Ishikawa A, Miura T (2010) Wing development genes of the pea aphid and differential gene expression between winged and unwinged morphs. *Insect molecular biology* 19: 63–73.

Brower AVZ (1994) Rapid morphological radiation and convergence among races of the butterfly *Heliconius erato* inferred from patterns of mitochondrial DNA evolution. *Proceedings of the National Academy of Sciences of the United States of America* 91: 6491–6495.

Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution* 21: 255–265.

Buerkle CA, Lexer C (2008) Admixture as the basis for genetic mapping. *Trends in ecology & evolution* 23: 686–694.

Burns CM (1991) *Will there ever be a rainbow?* The Simpsons, Episode: Blood Fued, Season 2:22.

Butlin RK (2010) Population genomics and speciation. *Genetica* 138: 409–418.

Butlin RK, Saura M, Charrier G *et al.* (2013) Parallel Evolution of Local Adaptation and Reproductive Isolation in the Face of Gene Flow. *Evolution*: 1–15.

Calleja M, Moreno E, Pelaz S, Morata G (1996) Visualization of Gene Expression in Living Adult *Drosophila*. *Science* 274: 252–255.

Cantrell RS, Cosner C, Lou Y (2010) Evolution of dispersal and the ideal free distribution. *Mathematical Biosciences and Engineering* 7: 17–36.

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

Ceccarelli C, Grodsky NB, Ariyaratne N, Colman RF, Bahnson BJ (2002) Crystal structure of porcine mitochondrial NADP+-dependent isocitrate dehydrogenase complexed with Mn2+ and isocitrate. Insights into the enzyme mechanism. *The Journal of biological chemistry* 277: 43454–43462.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS genetics* 2: e64.

Christians JK, Senger LK (2007) Fine mapping dissects pleiotropic growth quantitative trait locus into linked loci. *Mammalian genome* 18: 240–245.

Clarke TE, Levin DB, Kavanaugh DH, Reimchen TE (2001) Rapid evolution in the Nebria gregaria group (Coleoptera: Carabidae) and the paleogeography of the Queen Charlotte Islands. *Evolution* 55: 1408–1418.

Cochrane GR, Galperin MY (2010) The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic acids research* 38: D1–4.

Cohen D, Levin SA (1991) Dispersal in patchy environments: The effects of temporal and spatial structure. *Theoretical population biology* 39: 63–99.

Cohen B, McGuffin ME, Pfeifle C, Segal D, Cohen SM (1992) apterous, a gene required for imaginal disc development in Drosophila encodes a member of the LIM family of developmental regulatory proteins. *Genes & Development* 6: 715–729.

Colbourne JK, Pfrender ME, Gilbert D *et al.* (2011) The ecoresponsive genome of Daphnia pulex. *Science* 331: 555–561.

Collins L, Penny D (2006) Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Investigating the intron recognition mechanism in eukaryotes. *Molecular biology and evolution* 23: 901–910.

Colosimo PF, Hosemann KE, Balabhadra S *et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* 307: 1928–1933.

Colosimo PF, Peichel CL, Nereng K *et al.* (2004) The genetic architecture of parallel armor plate reduction in threespine sticklebacks. *PLoS biology* 2: E109.

Comeault A a, Soria-Carrasco V, Gompert Z *et al.* (2014) Genome-wide association mapping of phenotypic traits subject to a range of intensities of natural selection in *Timema cristinae*. *The American naturalist* 183: 711–727.

Compeau PEC, Pevzner P a, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. *Nature biotechnology* 29: 987–991.

Conde-Padín P, Cruz R, Hollander J, Rolán-Alvarez E (2008) Revealing the mechanisms of sexual isolation in a case of sympatric and parallel ecological divergence. *Biological Journal of the Linnean Society* 94: 513–526.

Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B* 279: 5039–5047.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185: 1411–1423.

Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* 11: 2463–2468.

Cortesero AM, Monge JP, Huignard J (1995) Influence of two successive learning processes on the response of *Eupelmus vuilleti* Crw (Hymenoptera: Eupelmidae) to volatile stimuli from hosts and host plants. *Journal of insect behavior* 8: 751–762.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates, Sunderland, MA, USA.

Crespi BJ (1989) Causes of assortative mating in arthropods. *Animal Behavior* 38: 980–1000.

Crisci JL, Poh Y-P, Bean A, Simkin A, Jensen JD (2012) Recent progress in polymorphism-based population genetic inference. *The Journal of heredity* 103: 287–296.
Cristescu ME, Constantin A, Bock DG, Cáceres CE, Crease TJ (2012) Speciation with gene flow and the genetics of habitat transitions. *Molecular ecology* 21: 1411–1422.
Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends in ecology & evolution* 25: 410–418.
Csilléry K, François O, Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution* 3: 475–479.
Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics* 27: 2156–2158.
Darwin C (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, Albemarle street, London, UK.
Dasmahapatra KK, Walters JR, Briscoe AD *et al.* (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487: 94–98.
Davey JW, Davey JL, Blaxter ML, Blaxter MW (2010) RADSeq: next-generation population genetics. *Briefings in functional genomics* 9: 416–23.
Davey JW, Hohenlohe P a, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature reviews. Genetics* 12: 499–510.
Davis JM, Stamps JA (2004) The effect of natal experience on habitat preferences. *Trends in ecology & evolution* 19: 411–416.
De Busschere C, Baert L, Van Belleghem SM, Dekoninck W, Hendrickx F (2012) Parallel phenotypic evolution in a wolf spider radiation on Galápagos.
De Busschere C, Hendrickx F, Van Belleghem SM *et al.* (2010) Parallel habitat specialization within the wolf spider genus Hogna from the Galápagos. *Molecular ecology* 19: 4029–4045.
De Busschere C, Van Belleghem SM, Hendrickx F (under review) Inter and intra island introgression in a wolf spider radiation from the galápagos, and its implications for parallel evolution. *Molecular phylogenetics and evolution*.
den Boer PJ (1968) Spreading of risk and the stabilization of animal numbers. *Acta Biotheoretica* 18: 165–194.
den Boer PJ (1970) On the significance of dispersal power for populations of carabid-beetles. *Oecologia* 4: 1–28.
den Boer PJ (1980) Wing polymorphism and dimorphism in ground beetles as stages in an evolutionary process (Coleoptera, Carabidae). *Entomologia Generalis* 6: 107–134.
Denno RF, Roderick GK, Peterson MA *et al.* (1996) Habitat persistence underlies intraspecific variation in the dispersal strategies of planthoppers. *Ecological Monographs* 66: 389–408.
Desender K (1985) Wing polymorphism and reproductive biology in the halobiont carabid beetle *Pogonus chalceus*. *Biologisch Jaarboek Dodonaea* 53: 89–100.
Desender K (1988) Flight-muscle development and dispersal in the life-cycle of Carabid beetles. *Annales De La Societe Royale Zoologique De Belgique* 118: 78–79.
Desender K (1989a) Heritability of wing development and body size in a carabid beetle, *Pogonus chalceus* Marsham, and its evolutionary significance. *Oecologia* 78: 513–520.
Desender K (1989b) *Dispersievermogen en ecologie van loopkevers (Coleoptera, Carabidae) in Belgie: een evolutionaire benadering*. Institut royal des sciences naturelles de Belgique.
Desender K (2000) Flight muscle development and dispersal in the life cycle of carabid beetles: Patterns and processes. *Bulletin de l'Institut Royal des Sciences Naturelles de Belgique* 70: 13–31.
Desender K, Backeljau T, Delahaye K, De Meester L (1998) Age and size of European saltmarshes and the population genetic consequences for ground beetles. *Oecologia* 114: 503–513.

Desender K, Maelfait J-P (1999) Diversity and conservation of terrestrial arthropods in tidal marshes along the River Schelde: a gradient analysis. *Biological Conservation* 87: 221–229.

Desender K, Maelfait J-P, Vaneechoutte M (1986) Allometry and evolution of hind wing development in macropterous Carabid beetles. In: *Carabid Beetles, their Adaptations, Dynamics and Evolution* (eds den Boer PJ, Luff ML, Mossakowski D, Weber F), pp. 101–112. Stuttgart, Germany.

Desender K, Maes D, Maelfait J, Van Kerckvoorde M (1995) *Een gedocumenteerde Rode lijst van de zandloopkevers en loopkevers van Vlaanderen*. Instituut voor Natuurbehoud.

Desender K, Serrano J (1999) A genetic comparison of Atlantic and Mediterranean populations of a saltmarsh beetle. *Belgian journal of zoology* 129: 83–94.

Desender K, Serrano J, Verdyck P (2000) Genetic diversity and wing polymorphism in the saltmarsh beetle *Pogonus chalceus*: an Atlantic-Mediterranean comparison. In: *Natural History and Applied Ecology of Carabid Beetles* (eds Brandmayr P, Lövei GL, Brandmayr ZT, Casala A, Taglianti VA), pp. 35–43. Pensoft, Sofia.

Dhuyvetter H, Gaublomme E, Desender K (2004) Genetic differentiation and local adaptation in the salt-marsh beetle *Pogonus chalceus*: a comparison between allozyme and microsatellite loci. *Molecular ecology* 13: 1065–1074.

Dhuyvetter H, Gaublomme E, Desender K (2005a) Bottlenecks, drift and differentiation: the fragmented population structureof the saltmarsh beetle *Pogonus chalceus*. *Genetica* 124: 167–177.

Dhuyvetter H, Gaublomme E, Verdyck P, Desender K (2005b) Genetic differentiation among populations of the salt marsh beetle Pogonus littoralis (Coleoptera: Carabidae): a comparison between Atlantic and Mediterranean populations. *The Journal of heredity* 96: 381–7.

Dhuyvetter H, Hendrickx F, Gaublomme E, Desender K (2007) Differentiation between two salt marsh beetle ecotypes: evidence for ongoing speciation. *Evolution; international journal of organic evolution* 61: 184–193.

Dieckmann U, Doebeli M (1999) On the origin of species by sympatric speciation. *Nature* 400: 354–357.

Dieckmann U, Doebeli M, Metz JAJ, Tautz D (2004) *Adaptive speciation*. Cambridge University Press, Cambridhe, UK.

Dobzhansky T (1935) A critique of the species concept in biology. *Philosophy of Science* 2: 344–355.

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

Doebeli M, Ruxton GD (1997) Evolution of dispersal rates in metapopulation models: Branching and cyclic dynamics in phenotype space. *Evolution* 51: 1730–1741.

Drotz MK, Brodin T, Saura A, Giles BE (2012) Ecotype differentiation in the face of gene flow within the diving beetle *Agabus bipustulatus* (Linnaeus, 1767) in northern Scandinavia. *PloS one* 7: e31381.

Drummond AJ, Rambaut A (2008) Tracer v1.5. Available from: http://tree.bio.ed.ac.uk/software/tracer.

Duron O, Bouchon D, Boutin S *et al.* (2008) The diversity of reproductive parasites among arthropods: *Wolbachia* do not walk alone. *BMC biology* 6: 27.

Eanes WF (1999) Analysis of selection on enzyme polymorphisms. *Annual Review of Ecology and Systematics* 30: 301–326.

Eck SH, Benet-Pagès A, Flisikowski K *et al.* (2009) Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. *Genome biology* 10: R82.

Edelaar P, Bolnick DI (2012) Non-random gene flow: an underappreciated force in evolution and ecology. *Trends in ecology & evolution* 27: 659–665.

Edelaar P, Siepielski AM, Clobert J (2008) Matching habitat choice causes directed gene flow: a neglected dimension in evolution and ecology. *Evolution* 62: 2462–2472.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* 491: 756–760.

Emlen DJ, Nijhout HF (1999) Hormonal control of male horn length dimorphism in the dung beetle *Onthophagus taurus* (Coleoptera: Scarabaeidae). *Journal of insect physiology* 45: 45–53.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing (V Orgogozo, M V. Rockman, Eds,). *Methods in molecular biology* 772: 1–19.

Ewing B, Green P (1998) Base-Calling of Automated Sequencer Traces Using Phred . II . Error Probabilities. *Genome research* 8: 186–194.

Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics (Oxford, England)* 26: 2064–2065.

Excoffier L, Ray N (2008) Surfing during population expansions promotes genetic revolutions and structuration. *Trends in ecology & evolution* 23: 347–351.

Fairbairn DJ, Roff DA (1991) Wing dimorphisms and the evolution of migratory polymorphisms among the insecta. *Integrative and comparative biology* 31: 243–251.

Falconer DS, Mackay TFC (1996) *Introduction to quantitative genetics*. Longman, Essex, U.K.

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular ecology notes* 7: 574–578.

Faria R, Renaut S, Galindo J *et al.* (2014) Advances in Ecological Speciation: an integrative approach. *Molecular ecology* 23: 513–521.

Feder JL, Berlocher SH, Roethele JB *et al.* (2003a) Allopatric genetic origins for sympatric host-plant shifts and race formation in *Rhagoletis*. *Proceedings of the National Academy of Sciences of the United States of America* 100: 10314–10319.

Feder JL, Egan SP, Nosil P (2012a) The genomics of speciation-with-gene-flow. *Trends in genetics* 28: 342–350.

Feder JL, Gejji R, Yeaman S, Nosil P (2012b) Establishment of new mutations under divergence and genome hitchhiking. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367: 461–474.

Feder ME, Micthell-Olds T (2003) Evolutionary and ecological functional genomics. *Nature Reviews Genetics* 4: 649–655.

Feder JL, Nosil P (2010) The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution* 64: 1729–1747.

Feder JL, Opp SB, Wlazlo B *et al.* (1994) Host fidelity is an effective premating barrier between sympatric races of the apple maggot fly. *Proceedings of the National Academy of Sciences of the United States of America* 91: 7990–7994.

Feder JL, Roethele JB, Filchak K, Niedbalski J, Romero-severson J (2003b) Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly , Rhagoletis pomonella. 953: 939–953.

Feder JL, Xie X, Rull J *et al.* (2005) Mayr, Dobzhansky, and Bush and the complexities of sympatric speciation in *Rhagoletis*. 102: 6573–6580.

Felsenstein J (1981) Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution; international journal of organic evolution* 35: 124–138.

Feulner PGD, Chain FJJ, Panchal M *et al.* (2013) Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. *Molecular ecology* 22: 635–649.

Fitzpatrick BM, Fordyce J a, Gavrilets S (2008) What, if anything, is sympatric speciation? *Journal of evolutionary biology* 21: 1452–1459.

Flaxman SM, Feder JL, Nosil P (2013) Genetic Hitchhiking and the Dynamic Buildup of Genomic Divergence During Speciation With Gene Flow. *Evolution* 67: 2577–2591.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular marine biology and biotechnology* 3: 294–299.

Fourcade Y, Chaput-Bardy A, Secondi J, Fleurant C, Lemaire C (2013) Is local selection so widespread in river organisms? Fractal geometry of river networks leads to high bias in outlier detection. *Molecular ecology* 22: 2065–2073.

Fox-Walsh KL, Dou Y, Lam BJ *et al.* (2005) The architecture of pre-mRNAs affects mechanisms of splice-site pairing. *Proceedings of the National Academy of Sciences of the United States of America* 102: 16176–16181.

Friesen VL, Smith a L, Gómez-Díaz E *et al.* (2007) Sympatric speciation by allochrony in a seabird. *Proceedings of the National Academy of Sciences of the United States of America* 104: 18589–18594.

Fry JD (2003) Multilocus models of sympatric speciation: Bush versus Rice versus Felsenstein. *Evolution; international journal of organic evolution* 57: 1735–1746.

Futuyma D (2005) *Evolution*. Sinauer Associates, Sunderland, Massachusetts.

Futuyma D, Mayer GC (1980) Non-allopatric speciation in animals. *Systematic Zoology* 29: 254–271.

Gaston KJ (1991) The Magnitude of Global Insect Species Richness. *Conservation Biology* 5: 283–296.

Gavrilets S (2003) Perspective: models of speciation: what have we learned in 40 years? *Evolution* 57: 2197–215.

Gibson G, Dworkin I (2004) Uncovering cryptic genetic variation. *Nature reviews. Genetics* 5: 681–690.

Gompert Z, Buerkle CA (2009) A powerful regression-based method for admixture mapping of isolation across the genome of hybrids. *Molecular Ecology* 18: 1207–1224.

Goodman SN (1999) Toward evidence-based medical statistics. 2: The Bayes factor. *American College of Physicians–American Society of Internal Medicine* 130: 1005–1013.

Götz S, García-Gómez JM, Terol J *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research* 36: 3420–3435.

Gourbière S, Mallet J (2010) Are species real? The shape of the species boundary with exponential failure, reinforcement, and the "missing snowball". *Evolution* 64: 1–24.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29: 644–652.

Grant BS (2004) Allelic Melanism in American and British Peppered Moths. *Journal of Heredity* 95: 97–102.

Griswold CK (2006) Gene flow's effect on the genetic architecture of a local adaptation and its consequences for QTL analyses. *Heredity* 96: 445–453.

Gustavson E, Goldsborough AS, Ali Z, Kornberg TB (1996) The Drosophila *engrailed* and *invected* genes: Partners in regulation, expression and function. *Genetics Society of America* 142: 893–906.

Haag CR, Saastamoinen M, Marden JH, Hanski I (2005) A candidate locus for variation in dispersal rate in a butterfly metapopulation. *Proceedings of the Royal Society B* 272: 2449–2456.

Harrison RG (1980) Dispersal polymorphisms in insects. *Annual Review of Ecology and Systematics* 11: 95–118.

Hawthorne DJ, Via S (2001) Genetic linkage of ecological specialization and reproductive isolation in pea aphids. *Nature* 412: 904–907.

Head ML, Kozak GM, Boughman JW (2013) Female mate preferences for male body size and shape promote sexual isolation in threespine sticklebacks. *Ecology and evolution* 3: 2183–2196.

Hebert PDN, Beaton MJ (1993) *Methodologies for allozyme analysis using cellulose acetate electrophoresis: A practical handbook*. University of Guelph, Ontario.

Hein J, Schierup MH, Wiuf C (2004) *Gene genealogies, variation and evolution*. Oxford University Press, USA.

Hendrickx F, Maelfait J-P, Desender K *et al.* (2009) Pervasive effects of dispersal limitation on within- and among-community species richness in agricultural landscapes. *Global Ecology and Biogeography* 18: 607–616.

Hendrickx F, Palmer SCF, Travis JMJ (2013) Ideal free distribution of fixed dispersal phenotypes in a wing dimorphic beetle in heterogeneous landscapes. *Ecology* 94: 2487–2497.

Hendry AP (2009) Ecological speciation! Or the lack thereof? *Canadian Journal of Fisheries and Aquatic Sciences* 66: 1383–1398.

Hendry AP, Bolnick DI, Berner D, Peichel CL (2009) Along the speciation continuum in sticklebacks. *Journal of Fish Biology* 75: 2000–2036.

Hey J (1991) The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. *Bioinformatics* 128: 831–840.

Hoback WW, Stanley DW (2001) Insects in hypoxia. *Journal of insect physiology* 47: 533–542.

Hoekstra HE, Coyne J a (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61: 995–1016.

Hoekstra HE, Hirschmann RJ, Bundey R a, Insel P a, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* 313: 101–104.

Hoffmann AA, Merilä J (1999) Heritable variation and evolution under favourable and unfavourable conditions. *Trends in Ecology & Evolution* 14: 96–101.

Hoffmann AA, Rieseberg LH (2008) Revisiting the impact of inversions in evolution: From population genetic markers to drivers of adaptive shifts and speciation? *Annual Review of Ecology, Evolution, and Systematics* 39: 21–42.

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

Hohenlohe P a, Phillips PC, Cresko W a (2010b) Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *International journal of plant sciences* 171: 1059–1071.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting Fst. *Nature reviews. Genetics* 10: 639–650.

Holt RD, McPeek MA (1996) Chaotic population dynamics favors the evolution of dispersal. *The American naturalist* 148: 709–718.

Huber SK, De León LF, Hendry AP, Bermingham E, Podos J (2007) Reproductive isolation of sympatric morphs in a population of Darwin's finches. *Proceedings of the Royal Society B. Biological sciences* 274: 1709–1214.

Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. *Genetics* 120: 831–840.

Huestis DL, Marshall JL (2006) Is natural selection a plausible explanation for the distribution of Idh-1 alleles in the cricket *Allonemobius socius*? *Ecological Entomology* 31: 91–98.

Huestis DL, Oppert B, Marshall JL (2009) Geographic distributions of Idh-1 alleles in a cricket are linked to differential enzyme kinetic performance across thermal environments. *BMC evolutionary biology* 9: 113.

Hunt T, Bergsten J, Levkanicova Z *et al.* (2007) A comprehensive phylogeny of beetles reveeals the evolutionary origins of a superradiation. *Science* 318: 1913–1916.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* 23: 254–67.

Huson DH, Rupp R, Scornavacca C (2010) *Phylogenetic networks: Concepts, algorithms and applications*. University Press, Cambridge, UK.

Immelmann K (1975) Ecological significance of imprinting and early learning. *Annual Review of Ecology and Systematics* 6: 15–37.

Ishikawa A, Ogawa K, Gotoh H *et al.* (2012) Juvenile hormone titre and related gene expression during the change of reproductive modes in the pea aphid. *Insect molecular biology* 21: 49–60.

Jaenike J, Holt RD (1991) Genetic variation for habitat preference: Evidence and explanations. *America Naturalist* 137: S67–S90.

Järvinen O, Vepsäläinen K (1976) Wing dimorphism as an adaptive strategy in water-striders (*Gerris*). *Hereditas* 84: 61–68.

Jenningss GT, Sechinll S, Stevenson PM *et al.* (1994) Cytosolic NADP + -dependent Isocitrate Dehydrogenase. *The Journal of biological chemistry* 269: 23128–23134.

Jiggins C (2008) Ecological speciation in mimetic butterflies. *BioOne* 58: 541–548.

Jo SH, Son MK, Koh HJ *et al.* (2001) Control of mitochondrial redox balance and cellular defense against oxidative damage by mitochondrial NADP+-dependent isocitrate dehydrogenase. *The Journal of biological chemistry* 276: 16168–16176.

Johannesson K, Panova M, Kemppainen P *et al.* (2010) Repeated evolution of reproductive isolation in a marine snail: unveiling mechanisms of speciation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 365: 1735–1747.

Johnson ML, Gaines MS (1990) Evolution of dispersal: Theoretical models and empirical tests using birds and mammals. *Annual Review of Ecology and Systematics* 21: 449–480.

Johnson PA, Hoppensteadt FC, Smith JJ, Bush GL (1996a) Conditions for sympatric speciation - diploid model incorporating habitat fidelity and non-habitat assortative mating. *Evolutionary Ecology* 10: 187–205.

Johnson PA, Hoppensteadt FC, Smith J., Bush GL (1996b) Conditions for sympatric speciation: a diploid model incorporating habitat fidelity and non-habitat assortative mating. *Evolutionary Ecology* 10: 187–205.

Johnson FM, Schaffer HE (1973) Isozyme variability in species of the genus *Drosophila*. VII. Genotype-environment relationships in populations of *D. melanogaster* from the eastern United States. *Biochemical genetics1* 10: 149–163.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484: 55–61.

Jones AG, Ratterman NL (2009) Mate choice and sexual selection: what have we learned since Darwin? *Proceedings of the National Academy of Sciences of the United States of America* 106: 10001–10008.

Joron M, Frezal L, Jones RT *et al.* (2011) Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* 477: 203–206.

Jurka J, Kapitonov V V, Pavlicek A *et al.* (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* 110: 462–467.

Kaiser L, Pérez-Maluf R, Sandoz JC, Pham-Delègue MH (2003) Dynamics of odour learning in Leptopilina boulardi, a hymenopterous parasitoid. *Animal Behaviour* 66: 1077–1084.

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28: 27–30.

Kapustin Y, Souvorov A, Tatusova T, Lipman D (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biology direct* 3: 20.

Karatolos N, Pauchet Y, Wilkinson P *et al.* (2011) Pyrosequencing the transcriptome of the greenhouse whitefly, Trialeurodes vaporariorum reveals multiple transcripts encoding insecticide targets and detoxifying enzymes. *BMC genomics* 12: 56.

Kawahara-Miki R, Tsuda K, Shiwa Y *et al.* (2011) Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle *Kuchinoshima-Ushi*. *BMC genomics* 12: 103.

Kawecki TJ (1996) Sympatric speciation driven by beneficial mutations. *Proceedings of the Royal Society B* 263: 1515–1520.

Keeling CI, Yuen MM, Liao NY *et al.* (2013) Draft genome of the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major forest pest. *Genome biology* 14: R27.

Keller I, Wagner CE, Greuter L *et al.* (2013) Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Molecular ecology* 22: 2848–2863.

Kennerdell JR, Carthew RW (1998) Use of dsRNA-mediated genetic interference to demonstrate that *frizzled* and *frizzled 2* act in the wingless pathway. *Cell* 95: 1017–1026.

Kerth C (2012) Scripts for RAD. https://github.com/claudiuskerth/scripts_for_RAD.

Kim J-I, Ju YS, Park H *et al.* (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature* 460: 1011–1015.

Kim HJ, Kang BS, Park JW (2005) Cellular defense against heat shock-induced oxidative damage by mitochondrial NADP+-dependent isocitrate dehydrogenase. *Free Radical Research* 39: 441–448.

Kim HS, Murphy T, Xia J *et al.* (2010) BeetleBase in 2010: revisions to provide comprehensive genomic information for Tribolium castaneum. *Nucleic acids research* 38: D437–442.

Kimura M (1983) *The neutral theory of molecular evolution.* Cambridge University Press.

Kirkpatrick M (2010) How and why chromosome inversions evolve. *PLoS biology* 8: e1000501.

Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. *Genetics* 173: 419–434.

Kirkpatrick M, Johnson T, Barton N (2002) General models of multilocus evolution. *Genetics* 1750: 1727–1750.

Kirkpatrick M, Ravigné V (2002) Speciation by natural and sexual selection: models and experiments. *The American naturalist* 159: S22–35.

Kocher SD, Li C, Yang W *et al.* (2013) The draft genome of a socially polymorphic halictid bee, *Lasioglossum albipes*. *Genome biology* 14: R142.

Kokko H, López-Sepulcre A (2006) From individual dispersal to species ranges: perspectives for a changing world. *Science* 313: 789–791.

Kondrashov AS, Mina MV (1986) Sympatric speciation: when is it possible? *Biological Journal of the Linnean Society* 27: 201–223.

Kreitman M (1983) Nucleotide polymorphism at the alcohol-dehydrogenase locus of *Drosophila melanogaster*. *Nature* 304: 412–417.

Kronforst MR, Young LG, Kapan DD *et al.* (2006) Linkage of butterfly mate preference and wing color preference cue at the genomic location of wingless. *Proceedings of the National Academy of Sciences of the United States of America* 103: 6575–6580.

Krzywinski M, Schein J, Birol I *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome research* 19: 1639–1645.

Kurtz S, Phillippy A, Delcher AL *et al.* (2004) Versatile and open software for comparing large genomes. *Genome biology* 5: R12.

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

Lawniczak MKN, Emrich SJ, Holloway a K *et al.* (2010) Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science (New York, N.Y.)* 330: 512–514.

Lee SM, Koh HJ (2002) Cytosolic NADP(+)-dependent isocitrate dehydrogenase status modulates oxidative damage to cells. *Free Radical Biology and Medicine* 32: 1185–1196.

Lenormand T (2002) Gene flow and the limits to natural selection. *Trends in Ecology & Evolution* 17: 183–189.

Levy S, Sutton G, Ng PC *et al.* (2007) The diploid genome sequence of an individual human. *PLoS biology* 5: e254.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987–2993.

Li H (2013) Seqtk: Toolkit for processing sequences in FASTA/Q formats.

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 12: 323.

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

Li R, Fan W, Tian G *et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature* 463: 311–317.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment / Map (SAM) Format and SAMtools 1000 Genome Project Data Processing Subgroup. *Bioinformatics* 25: 2078–2079.

Li J, Li H, Jakobsson M *et al.* (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular ecology* 21: 28–44.

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* 18: 1851–1858.

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

Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE (2009) On the origin and spread of an adaptive allele in deer mice. *Science* 325: 1095–1098.

Linnen CR, Poh Y-P, Peterson BK *et al.* (2013) Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* 339: 1312–1316.

Liu B, Shi Y, Yuan J *et al.* (2013) Estimation of genomic characteristics by analyzing k- mer frequency in de novo genome projects Keywords. *arXiv*: 1–47.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature reviews. Genetics* 4: 981–994.

Luo R, Liu B, Xie Y *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1: 18.

M'Gonigle LK, Mazzucco R, Otto SP, Dieckmann U (2012) Sexual selection enables long-term coexistence despite ecological equivalence. *Nature* 484: 506–509.

Maan ME, Seehausen O, Söderberg L *et al.* (2004) Intraspecific sexual selection on a speciation trait, male coloration, in the Lake Victoria cichlid Pundamilia nyererei. *Proceedings. Biological sciences / The Royal Society* 271: 2445–2452.

Machado C a, Kliman RM, Markert J a, Hey J (2002) Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Molecular biology and evolution* 19: 472–488.

Mackay TFC, Langley CH (1990) Molecular and phenotypic variation in the *achaete-scute* region of *Drosophila melanogaster*. *Nature* 348: 64–66.
Mackay TFC, Lyman RF (2005) Drosophila bristles and the nature of quantitative genetic variation. : 1513–1527.
Mackay TFC, Stone E a, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. *Nature reviews. Genetics* 10: 565–577.
Malek TB, Boughman JW, Dworkin I, Peichel CL (2012) Admixture mapping of male nuptial colour and body shape in a recently formed hybrid population of threespine stickleback. *Molecular ecology* 21: 5265–5279.
Mallet J (2008) Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 363: 2971–2986.
Mallet J, Beltrán M, Neukirchen W, Linares M (2007) Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC evolutionary biology* 7: 28.
Mallet J, Meyer A, Nosil P, Feder JL (2009) Space, sympatry and speciation. *Journal of evolutionary biology* 22: 2332–2341.
Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27: 764–770.
Marden JH (2013) Nature's inordinate fondness for metabolic enzymes: why metabolic enzyme loci are so frequently targets of selection. *Molecular* 22: 5743–5764.
Marden JH, Fescemyer HW, Schilder RJ *et al.* (2012) Genetic variation in HIF signaling underlies quantitative variation in physiological and life-history traits within lowland butterfly populations. *Evolution* 67: 1105–1115.
Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet* 17: 10–12.
Martin SH, Dasmahapatra KK, Nadeau NJ *et al.* (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome research* 23: 1817–1828.
Matessi C, Gimelfarb A, Gavrilets S (2001) Long term buildup of reproductive isolation promoted by disruptive selection: how far does it go ? *Selection* 2: 41–64.
Mathias A, Kisdi E, Olivieri I (2001) Divergent evolution of dispersal in a heterogeneous landscape. *Evolution* 55: 246–259.
Maynard Smith SJ (1966) Sympatric speciation. *The American naturalist* 100: 637–650.
Maynard Smith SJ, Haigh J (1974) The hitchhiking effect of a favorable gene. *Genetic Research* 23: 23–35.
Mayr E (1942) *Systematics and the origin of species*. Columbia University Press, New York.
Mayr E (1963) *Animal Species and Evolution*. Harvard University Press, Cambridge.
McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351: 652–654.
McGuire AM, Pearson MD, Neafsey DE, Galagan JE (2008) Cross-kingdom patterns of alternative splicing and splice recognition. *Genome biology* 9: R50.
McKinnon JS, Pierotti MER (2010) Colour polymorphism and correlated characters: genetic mechanisms and evolution. *Molecular ecology* 19: 5101–5125.
McPeek MA, Holt RD (1992) The evolution of dispersal in spatially and temporally varying environments. *The American naturalist* 140: 1010–1027.
McPhail JD (1994) Speciation and the evolution of reproductive isolation in the sticklebacks (*Gasterosteus*) of southwestern British Columbia. In: *The evolutionary biology of the threespine stickleback* (eds Bell MA, Foster SA), pp. 399–437. Oxford University Press, Oxford, UK.

Messer PW, Petrov D a (2013) Population genomics of rapid adaptation by soft selective sweeps. *Trends in ecology & evolution* 28: 659–669.

Metzker ML (2010) Sequencing technologies - the next generation. *Nature reviews. Genetics* 11: 31–46.

Michel AP, Sim S, Powell THQ *et al.* (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences of the United States of America* 107: 9724–9729.

Min XJ, Butler G, Storms R, Tsang A (2005) OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic acids research* 33: W677–680.

Mitchell-olds T, Willis JH, Goldstein DB (2007) Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics* 8: 845–856.

Mittapalli O, Bai X, Mamidala P *et al.* (2010) Tissue-specific transcriptomics of the exotic invasive insect pest emerald ash borer (*Agrilus planipennis*). *PloS one* 5: e13708.

Mole S, Zera AJ (1993) Differential allocation of resources underlies the dispersal-reproduction trade-off in the wing-dimorphic cricket, *Gryllus rubens*. *Oecologia* 93: 121–127.

Mueller P, Diamond J (2001) Metabolic rate and environmental productivity: well-provisioned animals evolved to run and idle fast. *Proceedings of the National Academy of Sciences of the United States of America* 98: 12550–12554.

Muse S V, Weir BS (1992) Testing for equality of evolutionary rates. *Genetics* 132: 269–276.

Nachman MW, Hoekstra HE, D'Agostino SL (2003) The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences of the United States of America* 100: 5268–5273.

Nadeau NJ, Martin SH, Kozak KM *et al.* (2013) Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular ecology* 22: 814–826.

Nagel L, Schluter D (1998) Body size, natural selection, and speciation in sticklebacks. *Evolution* 52: 209–218.

Nei M, Li W (1973) Linkage disequilibrium in subdivided populations. *Genetics* 75: 213–219.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* 76: 5269–5273.

Noor M a F, Bennett SM (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity* 103: 439–444.

Noor M a F, Garfield D a, Schaeffer SW, Machado C a (2007) Divergence between the *Drosophila pseudoobscura* and *D. persimilis* genome sequences in relation to chromosomal inversions. *Genetics* 177: 1417–1428.

Nosil P (2012) *Ecological Speciation*. Oxford University Press, Oxford, UK.

Nosil P, Feder JL, B PTRS (2012) Genomic divergence during speciation: causes and consequences. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367: 332–342.

Nosil P, Funk DJ, Ortíz-Barrientos D (2009a) Divergent selection and heterogeneous genomic divergence. *Molecular ecology* 18: 375–402.

Nosil P, Harmon LJ, Seehausen O (2009b) Ecological explanations for (incomplete) speciation. *Trends in ecology & evolution* 24: 145–156.

Nosil P, Schluter D (2011) The genes underlying the process of speciation. *Trends in Ecology & Evolution* 26: 160–167.

Nüsslein-volhard C, Wieschaus E (1980) Mutations affacting segment number and polarity in *Drosophila*. *Nature* 287: 795–801.

Nylander JAA (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

O'Neil ST, Dzurisin JDK, Carmichael RD *et al.* (2010) Population-level transcriptome sequencing of nonmodel organisms Erynnis propertius and Papilio zelicaon. *BMC genomics* 11: 310.
Odum WE (1988) Comparative Ecology of Tidal Freshwater and Salt Marshes. *Annual Review of Ecology and Systematics* 19: 147–176.
Oliveira GA, Baptista DL, Guimarães-motta H, Atella GC, Almeida IC (2006) Flight-oogenesis syndrome in a blood-sucking bug : biochemical aspects of lipid metabolism. *Archives of Insect Biochemistry and Physiology* 62: 164–175.
Olivieri I, Michalakis Y, Gouyon P (1995) Metapopulation genetics and the evolution of dispersal. *The American naturalist* 146: 202–228.
Orino K, Lehman L, Tsuji Y *et al.* (2001) Ferritin and the response to oxidative stress. *The Biochemical journal* 357: 241–247.
Orr HA (2005) The genetic theory of adaptation: a brief history. *Nature reviews. Genetics* 6: 119–127.
Orr HA, Coyne JA (1992) The genetics of adaptation: a reassessment. *American Naturalist* 140: 725–742.
Orr HA, Masly JP, Presgraves DC (2004) Speciation genes. *Current Opinion in Genetics & Development* 14: 675–679.
Orsini L, Wheat CW, Haag CR *et al.* (2009) Fitness differences associated with Pgi SNP genotypes in the Glanville fritillary butterfly (Melitaea cinxia). *Journal of evolutionary biology* 22: 367–375.
Ortíz-Barrientos D, Noor M a F (2005) Evidence for a one-allele assortative mating locus. *Science (New York, N.Y.)* 310: 1467.
Osella M, Caselle M (2009) Entropic contributions to the splicing process. *Physical biology* 6: 046018.
Paarmann W (1976) Annual periodicity of polyvoltine ground beetle *Pogonus chalceus* Marsham (Coleoptera: Carabidae) and its control by environmental-factors. *Zoologischer Anzeiger* 196: 150–160.
Parchman TL, Gompert Z, Braun MJ *et al.* (2013) The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Molecular ecology* 22: 3304–3317.
Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23: 1061–1067.
Payne R, Payne L, Woods J, Sorenson M (2000) Imprinting and the origin of parasite-host species associations in brood-parasitic indigobirds, Vidua chalybeata. *Animal behaviour* 59: 69–81.
Peccoud J, Ollivier A, Plantegenest M, Simon J-C (2009) A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. *Proceedings of the National Academy of Sciences of the United States of America* 106: 7495–500.
Pétillon J, Montaigne W, Renault D (2009) Hypoxic coma as a strategy to survive inundation in a salt-marsh inhabiting spider. *Biology letters* 5: 442–445.
Pinho C, Hey J (2010) Divergence with Gene Flow: Models and Data. *Annual Review of Ecology, Evolution, and Systematics* 41: 215–230.
Poelchau MF, Reynolds J a, Denlinger DL, Elsik CG, Armbruster P a (2011) A *de novo* transcriptome of the Asian tiger mosquito, *Aedes albopictus*, to identify candidate transcripts for diapause preparation. *BMC genomics* 12: 619.
Pons J, Ribera I, Bertranpetit J, Balke M (2010) Nucleotide substitution rates for the full set of mitochondrial protein-coding genes in Coleoptera. *Molecular phylogenetics and evolution* 56: 796–807.

Poulton EB (1904) What is a species? *Transactions Royal Entomological Society London* 1903: 77–116.

Price AL, Jones NC, Pevzner P a (2005) De novo identification of repeat families in large genomes. *Bioinformatics (Oxford, England)* 21 Suppl 1: i351–358.

Pritchard JK, Di Rienzo A (2010) Adaptation - not by sweeps alone. *Nature reviews. Genetics* 11: 665–667.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution* 16: 1791–1798.

Przeworski M (2002) The signature of positive selection at randomly chosen loci. *Genetics* 160: 1179–1189.

Ramsey J, Bradshaw HD, Schemske DW (2003) Components of reproductive isolation between the monkeyflowers *Mimulus lewisii* and *M. cardinalis* (Phrymaceae). *Evolution* 57: 1520–1534.

Rankin MA, McAnelly ML, Bodenhamer JE (1986) The oogenesis-flight syndrome revisited. In: *Insect flight: dispersal and migration* , p. 27–48. Springer-Verlag, Berlin.

Räsänen K, Hendry AP (2008) Disentangling interactions between adaptive divergence and gene flow when ecology drives diversification. *Ecology letters* 11: 624–636.

Ravigné V, Dieckmann U, Olivieri I (2009) Live where you thrive: joint evolution of habitat choice and local adaptation facilitates specialization and promotes diversity. *The American naturalist* 174: E141–169.

Raymond M, Rousset F (1995) GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. *Journal of Heredity* 86: 248–249.

Reed RD, Papa R, Martin A *et al.* (2011) *optix* drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science* 333: 1137–1141.

Renaut S, Maillet N, Normandeau E *et al.* (2012) Genome-wide patterns of divergence during speciation: the lake whitefish case study. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367: 354–363.

Rice WR (1984) Disruptive selection on habitat preference and the evolution of reproductive isolation: a simulation study. *Evolution* 38: 1251–1260.

Rice WR (1985) Disruptive selection on habitat preference and the evolution of reproductive isolation: a simulation study. *Evolution* 39: 645–656.

Rice WR (1987) Speciation via habitat specialization: the evolution of reproductive isolation as a correlated character. *Evolutionary Ecology* 1: 301–314.

Rice WR, Hostert EE (1993) Laboratory experiments on speciation: what have we learned in forty years? *Evolution* 47: 1637–1653.

Rice AM, Rudh A, Ellegren H, Qvarnström A (2011) A guide to the genomics of ecological speciation in natural animal populations. *Ecology letters* 14: 9–18.

Richards S, Gibbs RA, Weinstock GM *et al.* (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452: 949–955.

Rieseberg LH, Willis JH (2007) Plant speciation. *Science* 317: 910–914.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake – stream stickleback population pairs. *Molecular ecology* 21: 2852–2862.

Roff DA (1986) The evolution of wing dimorphism in insects. *Evolution* 40: 1009–1020.

Roff DA (1994a) Habitat persistence and the evolution of wing dimorphism in insects. *The American naturalist* 144: 772–798.

Roff DA (1994b) Why is there so much genetic variation for wing dimorphism? *Research on Population Ecology* 36: 145–150.

Roff DA, Fairbairn DJ (2007) The evolution and genetics of migration in insects. *BioScience* 57: 155–164.

Rogers SM, Bernatchez L (2005) Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). *Molecular ecology* 14: 351–361.

Rogers SM, Bernatchez L (2007) The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (*Coregonus* sp. Salmonidae) species pairs. *Molecular biology and evolution* 24: 1423–1438.

Ronce O (2007) How does it feel to be like a rolling stone? Ten questions about dispersal evolution. *Annual Review of Ecology, Evolution, and Systematics* 38: 231–253.

Ronquist F, Teslenko M, van der Mark P *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* 61: 539–542.

Rørth P, Szabo K, Bailey A *et al.* (1998) Systematic gain-of-function genetics in *Drosophila*. *Development* 125: 1049–1057.

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular ecology resources* 8: 103–106.

Le Rouzic A, Carlborg O (2008) Evolutionary potential of hidden genetic variation. *Trends in Ecology & Evolution* 23: 33–37.

Rozas J, Gullaud M, Blandin G, Aguadé M (2001) DNA variation at the rp49 gene region of Drosophila simulans: evolutionary inferences from an unusual haplotype structure. *Genetics* 158: 1147–1155.

Rundle HD, Nagel L, Boughman JW, Schluter D (2000) Natural Selection and Parallel Speciation in Sympatric Sticklebacks. *Science* 287: 306–308.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters* 8: 336–352.

Saccheri IJ, Rousset F, Watts PC, Brakefield PM, Cook LM (2008) Selection and gene flow on a diminishing cline of melanic peppered moths. *Proceedings of the National Academy of Sciences of the United States of America* 105: 16212–1617.

Sauer J, Hausdorf B (2009) Sexual selection is involved in speciation in a land snail radiation on crete. *Evolution* 63: 2535–2546.

Savolainen V, Anstett M-C, Lexer C *et al.* (2006) Sympatric speciation in palms on an oceanic island. *Nature* 441: 210–213.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature reviews. Genetics* 14: 807–280.

Schilder RJ, Zera AJ, Black C, Hoidal M, Wehrkamp C (2011) The biochemical basis of life history adaptation: molecular and enzymological causes of NADP(+)-isocitrate dehydrogenase activity differences between morphs of *Gryllus firmus* that differ in lipid biosynthesis and life history. *Molecular biology and evolution* 28: 3381–3393.

Schlötterer C (2003) Hitchhiking mapping--functional genomics from the population genetics perspective. *Trends in genetics : TIG* 19: 32–38.

Schluter D (2001) Ecology and the origin of species. *Trends in Ecology & Evolution* 16: 372–380.

Schluter D (2009) Evidence for ecological speciation and its alternative. *Science* 323: 737–741.

Schluter D, Conte GL (2009) Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America* 106 Suppl: 9955–9962.

Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one* 6: e17288.

Seehausen O (2009) Progressive levels of trait divergence along a " speciation transect " in the Lake Victoria cichlid fish Pundamilia. In: *Speciation and patterns of diversity* (eds Butlin R, Bridle J, Schluter D), pp. 155–176. Cambridge University Press, Cambridhe, UK.

Seehausen O, Butlin RK, Keller I *et al.* (2014) Genomics and the origin of species. *Nature reviews. Genetics* 15: 176–192.

Seehausen O, Terai Y, Magalhaes IS *et al.* (2008) Speciation through sensory drive in cichlid fish. *Nature* 455: 620–626.

Serrano J (1981a) A chromosome study of spanish Bembidiidae and other Caraboidea (Coleoptera, Adephaga). *Genetica* 57: 119–129.

Serrano J (1981b) Chromosome number and karyotypic evolution of Caraboidea. *Genetica* 55: 51–60.

Servedio MR, Noor MAF (2003) The role of reinforcement in speciation: Theory and Data. *Annual Review of Ecology, Evolution, and Systematics* 34: 339–364.

Servedio MR, Van Doorn GS, Kopp M, Frame AM, Nosil P (2011) Magic traits in speciation: "magic" but not rare? *Trends in ecology & evolution* 26: 389–397.

Shapiro MD, Marks ME, Peichel CL *et al.* (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. 428: 717–724.

Shen G-M, Dou W, Niu J-Z *et al.* (2011) Transcriptome analysis of the oriental fruit fly (Bactrocera dorsalis). *PloS one* 6: e29127.

Shigenobu S, Bickel RD, Brisson J a *et al.* (2010) Comprehensive survey of developmental genes in the pea aphid, *Acyrthosiphon pisum*: frequent lineage-specific duplications and losses of developmental genes. *Insect molecular biology* 19 Suppl 2: 47–62.

Siepielski AM, DiBattista JD, Carlson SM (2009) It's about time: the temporal dynamics of phenotypic selection in the wild. *Ecology letters* 12: 1261–1276.

Simon C (1994) Evolution, Weighting, and Phylogenetic Utility of Mitochondrial Gene Sequences and a Compilation of Conserved Polymerase Chain Reaction Primers.pdf. *Entomological Society of America* 87: 651–701.

Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science* 236: 787–792.

Sloan DB, Keller SR, Berardi AE, Sanderson BJ (2012) De novo transcriptome assembly and polymorphism detection in the flowering plant Silene vulgaris (Caryophyllaceae). *Molecular ecology resources* 12: 333–343.

Smadja CM, Butlin RK (2011) A framework for comparing processes of speciation in the presence of gene flow. *Molecular ecology* 20: 5123–5140.

Smit AFA, Hubley R, Green P (2014) RepeatMasker. http://www.repeatmasker.org/.

Smith A, Cornellt V (1979) Hopkins host-selection in *Nasonia vitripennis* and its implications for sympatric speciation. *Animal Behavior* 27: 365–370.

Smith M, Jiggins CD, Naisbit RE, Coe RL, Mallet J (2001) Reproductive isolation caused by colour pattern mimicry. *Nature* 411: 302–305.

Song H, Sheffield NC, Cameron SL, Miller KB, Whiting MF (2010) When phylogenetic assumptions are violated : base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. *Systematic Entomology* 35: 429–448.

Sorenson MD, Sefc KM, Payne RB (2003) Speciation by host switch in brood parasitic indigobirds. *Nature* 424: 928–931.

Soria-Carrasco V, Gompert Z, Comeault A a *et al.* (2014) Stick insect genomes reveal natural selection's role in parallel speciation. *Science* 344: 738–742.

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

Stamps JA (2001) Habitat selection by dispersers: integrating proximate and ultimate approaches. In: *Dispersal* (eds Clobert J, Danchin E, Dhondt AA, Nichols JD), pp. 230–242. Oxford University Press.

Stapley J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics : the next generation. *Trends in Ecology & Evolution* 25: 705–712.
Steiner CC, Weber JN, Hoekstra HE (2007) Adaptive variation in beach mice produced by two interacting pigmentation genes. *PLoS biology* 5: 1880–1889.
Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American journal of human genetics* 68: 978–989.
Stern DL (2013) The genetic causes of convergent evolution. *Nature reviews. Genetics* 14: 751–764.
Stern DL, Orgogozo V (2008) The loci of evolution: how predictable is genetic evolution? *Evolution* 62: 2155–2177.
Stevens VM, Trochet A, Blanchet S *et al.* (2013) Dispersal syndromes and the use of life-histories to predict dispersal. *Evolutionary applications* 6: 630–642.
Stevens MV, Trochet A, Van Dyck H, Clobert J, Baguette M (2012) How is dispersal integrated in life histories: a quantitative analysis using butterflies. *Ecology letters* 15: 74–86.
Stevens VM, Whitmee S, Le Galliard J-F *et al.* (2014) A comparative analysis of dispersal syndromes in terrestrial and semi-terrestrial animals (J Chase, Ed,). *Ecology Letters* 17: 1039–1052.
Stevison LS, Hoehn KB, Noor M a F (2011) Effects of inversions on within- and between-species recombination and divergence. *Genome biology and evolution* 3: 830–841.
Stinchcombe JR, Hoekstra HE (2008) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* 100: 158–170.
Storeck A, Poppy GM, Emden HF, Powell W (2000) The role of plant chemical cues in determining host preference in the generalist aphid parasitoid *Aphidius colemani*. *Entomologia Experimentalis et Applicata* 97: 41–46.
Storz JAYF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. : 671–688.
Storz JF, Wheat CW (2010) Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution* 64: 2489–2509.
Sukumaran J, Holder MT (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics* 26: 1569–1571.
Sulloway FJ (1979) Geographic isolation in Darwin's thinking: the vicissitudes of a crucial idea. *Studies of History of Biology* 3: 23–65.
Sultan SE, Spencer HG (2002) Metapopulation structure favors plasticity over local adaptation. *The American naturalist* 160: 271–283.
Supple M, Papa R, Counterman B, Mcmillan WO (2014) The genomics of an adaptive radiation: Insights across the Heliconius speciation continuum. In: *Ecological Genomics: Ecology and the evolution of genes and genomes* Advances in Experimental Medicine and Biology. (eds Landry CR, Aubin-Horth N), pp. 249–271. Springer Netherlands, Dordrecht.
Swafford D (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.
Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.
Tamura K, Peterson D, Peterson N *et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution* 28: 2731–2739.
Tatusov RL, Fedorova ND, Jackson JD *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC bioinformatics* 4: 41.
Theil EC (1987) Ferritin - Structure, Gene-Regulation, and Cellular Function in Animals, Plants, and Microorganisms. *Annual Review of Biochemistry* 56: 289–315.

Theodorides K, Riva AD, Foster PG, Vogler AP (2002) Comparison of EST libraries from seven beetle species : towards a framework for phylogenomics of the Coleoptera. *Insect molecular biology* 11: 467–475.

Thibert-Plante X, Gavrilets S (2013) Evolution of mate choice and the so-called magic traits in ecological speciation. *Ecology letters* 16: 1004–1013.

Turelli M (1984) Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. *Theoretical population theory* 25: 138–93.

Turin H (2000) *De Nederlandse loopkevers, verspreiding en oecologie (Coleoptera. Carabidae), Nederlandse Fauna3.* Nationaal Natuurhistorisch Museum Naturalis, KNNV Uitgeverij & EIS, Leiden, Nederland.

Van Belleghem SM, Hendrickx F (2014) A tight association in two genetically unlinked dispersal related traits in sympatric and allopatric salt marsh beetle populations. *Genetica* 142: 1–9.

Van Belleghem SM, Roelofs D, Van Houdt J, Hendrickx F (2012) De novo transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle *Pogonus chalceus* (Coleoptera, Carabidae). *PloS one* 7: e42605.

Van Dyck H, Matthysen E (1999) Habitat fragmentation and insect flight: a changing "design" in a changing landscape? *Trends in ecology & evolution* 14: 172–174.

Van Ooijen JW, Jansen J (2013) *Genetic mapping in experimental populations*. Cambridge University Press, New York.

Van't Hof AE, Edmonds N, Dalíková M, Saccheri IJ (2011) Industrial melanism in British peppered moths has a singular and recent mutational origin. *Science (New York, N.Y.)* 332: 958–960.

Vellichirammal NN, Zera AJ, Schilder RJ *et al.* (2014) De Novo Transcriptome Assembly from Fat Body and Flight Muscles Transcripts to Identify Morph-Specific Gene Expression Profiles in *Gryllus firmus*. *PloS one* 9: e82129.

Via S (1991) Specialized host plant performance of pea Aahid clones is not altered by experience. *Ecology* 72: 1420–1427.

Via S (2001) Sympatric speciation in animals: the ugly duckling grows up. *Trends in ecology & evolution* 16: 381–390.

Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Sciences of the United States of America* 106: 9939–9946.

Via S (2012) Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367: 451–460.

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular ecology* 17: 4334–4345.

Wall JD (1999) Recombination and the power of statistical tests of neutrality. *Genetic Research* 74: 65–79.

Wang S, Lorenzen MD, Beeman RW, Brown SJ (2008) Analysis of repetitive DNA distribution patterns in the Tribolium castaneum genome. *Genome biology* 9: R61.

Wang X-W, Luan J-B, Li J-M *et al.* (2010) De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC genomics* 11: 400.

Warren JT, Petryk A, Marqués G *et al.* (2004) Phantom encodes the 25-hydroxylase of *Drosophila melanogaster* and *Bombyx mori*: a P450 enzyme critical in ecdysone biosynthesis. *Insect biochemistry and molecular biology* 34: 991–1010.

Watt WB, Cassin RC, Swan MS (1983) Adaptation at specific loci. III. Field behavior and survivorship differences among *colias* PGI genotypes are predictable from *in vitro* biochemistry. *Genetics*: 725–739.

Watt WB, Wheat CW, Meyer EH, Martin J-F (2003) Adaptation at specific loci. VII. Natural selection, dispersal and the diversity of molecular–functional variation patterns among butterfly species complexes (*Colias*: Lepidoptera, Pieridae). *Molecular ecology* 12: 1265–1275.

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

Weatherbee SD, Nijhout HF, Grunert LW *et al.* (1999) *Ultrabithorax* function in butterfly wings and the evolution of insect wing patterns. *Current biology* 9: 109–115.

Weihe U, Milán M, Cohen SM (2005) Drosophila limb development. In: *Insect Development: Morphogenesis, Molting and Metamorphosis* (ed Gilbert LI), p. 730. Elsevier, London, UK.

West-eberhard MJ (2003) *Developmental plasticity and evolution*. Oxford University Press, New York.

West-Eberhard MJ (2005) Developmental plasticity and the origin of species differences. *Proceedings of the National Academy of Sciences of the United States of America* 102: 6543–6549.

Wheat CW, Fescemyer HW, Kvist J *et al.* (2011) Functional genomics of life history variation in a butterfly metapopulation. *Molecular ecology* 20: 1813–1828.

Wheat CW, Haag CR, Marden JH, Hanski I, Frilander MJ (2009) Nucleotide polymorphism at a gene (Pgi) under balancing selection in a butterfly metapopulation. *Molecular biology and evolution* 27: 267–281.

Wheat CW, Watt WB, Pollock DD, Schulte PM (2006) From DNA to fitness differences: sequences and structures of adaptive variants of Colias Phosphoglucose Isomerase (PGI). *Molecular biology and evolution* 23: 499–512.

Wild AL, Maddison DR (2008) Evaluating nuclear protein-coding genes for phylogenetic utility in beetles. *Molecular phylogenetics and evolution* 48: 877–891.

Wilkin MB, Becker MN, Mulvey D *et al.* (2000) *Drosophila* Dumpy is a gigantic extracellular protein required to maintain tension at epidermal – cuticle attachment sites. *Current biology* 10: 559–567.

Wright S (1931) Evolution in mendelian populations. *Genetics* 16: 97–159.

Wu C (2001a) Genes and speciation. *Journal of evolutionary biology* 14: 889–891.

Wu C (2001b) The genic view of the process of speciation. *Journal of evolutionary biology* 14: 851–865.

Xu X, Zhao J, Xu Z *et al.* (2004) Structures of human cytosolic NADP-dependent isocitrate dehydrogenase reveal a novel self-regulatory mechanism of activity. *The Journal of biological chemistry* 279: 33946–33957.

Xue J, Bao Y-Y, Li B-L *et al.* (2010) Transcriptome analysis of the brown planthopper Nilaparvata lugens. *PloS one* 5: e14233.

Xue W, Li J-T, Zhu Y-P *et al.* (2013) L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC genomics* 14: 604.

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. *Nature reviews. Genetics* 13: 329–342.

Yeaman S (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences of the United States of America* 110: E1743–1751.

Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection balance. *Evolution; international journal of organic evolution* 65: 1897–1911.

Zera AJ (2004) The endocrine regulation of wing polymorphism in insects: state of the art, recent surprises, and future directions. *Integrative and comparative biology* 43: 607–616.

Zera AJ, Denno RF (1997) Physiology and ecology of dispersal polymorphism in insects. *Annual review of entomology* 42: 207–230.

Zera A., Larsen A (2001) The metabolic basis of life history variation: genetic and phenotypic differences in lipid reserves among life history morphs of the wing-polymorphic cricket, Gryllus firmus. *Journal of Insect Physiology* 47: 1147–1160.

Zera AJ, Zhao Z (2003) Morph-dependent fatty acid oxidation in a wing-polymorphic cricket: implications for the trade-off between dispersal and reproduction. *Journal of Insect Physiology* 49: 933–943.

Zhan S, Zhang W, Niitepõld K *et al.* (2014) The genetics of monarch butterfly migration and warning colouration. *Nature*.

Zhao WN, Mcalister-henns L (1996) Expression and gene disruption analysis of the isocitrate dehydrogenase family in yeast. *Biochemistry* 35: 7873–7878.

Zhao Z, Zera AJ (2006) Biochemical basis of specialization for dispersal vs. reproduction in a wing-polymorphic cricket: morph-specific metabolism of amino acids. *Journal of insect physiology* 52: 646–658.