belspo

BRAIN-be
Belgian Research Action through Interdisciplinary Networks

PIONEER PROJECTS

Extending the genomic toolbox to decipher lineage-wide parallel wing evolution

CONTRACT - BR/175/PI/PARAWINGS

FINAL REPORT

31/03/2020

Promotor

Frederik HENDRICKX
Koninklijk Belgisch Instituut voor Natuurwetenschappen, Vautierstraat 29, 1000 Brussel
Tel: 02/627.41.37 - E-mail: frederik.hendrickx@naturalsciences.be
https://publons.com/researcher/2642823/frederik-hendrickx/

Authors

Frederik HENDRICKX
Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000 Brussel

Zoë DE CORTE
Royal Belgian Institute of Natural Sciences, Vautierstraat 29, 1000 Brussel

museum

.be

Published in 2020 by the Belgian Science Policy
WTC III
Simon Bolivarlaan 30 bus 7
Boulevard Simon Bolivar 30 bte 7
B-1000 Brussels
Belgium
Tel: +32 (0)2 238 34 11
http://www.belspo.be

Contact person: Georges JAMART
+32 (0)2 238 36 90

Hendrickx, F., De Corte, Z. **Expanding the genomic toolbox to decipher lineage-wide parallel wing evolution.** Final Report. Brussels: Belgian Science Policy 2020 – 25 p. (BRAIN-be - (Belgian Research Action through Interdisciplinary Networks))

TABLE OF CONTENTS

SUMMARY

Context

The recurrent gain and loss of identical but complex traits during the time course of evolution has puzzled evolutionary biologists since Darwin. Until recently, the convergent evolution of traits across lineages was studied by the probabilistic reconstruction of ancestral states of characters in a phylogenetic framework. While such methods are pivotal in describing the evolutionary patterns of character evolution, they do not provide us with the necessary basis on the mechanism on *how* traits may repeatedly disappear or eventually re-evolve.

Recent developments in molecular and developmental biology are beginning to improve our insights into the different mechanisms underlying parallel trait evolution, but we still lack a comprehensive view on how and if similar mechanisms are playing over longer evolutionary timescales. The inability to trace back the genomic basis of the same phenotypic traits in distantly related species, as well as difficulties to reconstruct the entire sequence of the genomic region of interest strongly hampered further advancements in this research field.

Objectives

In the proposed project, we tackle these issues by investigating the genomic basis underlying repeated evolution of wing development in the insect linage of carabid beetles. Within this large beetle family, species can be found with either fully developed wings, reduced wings as well as wing-dimorphic species wherein only some individuals have reduced wings. Here, we take full advantage of these wing-dimorphic species as these allow us to target the genomic region responsible for wing development within species. In a first objective, we will associate genetic variation at a genome wide scale with wing development for multiple wing-dimorphic species to find genetic markers associated with wing development in each species. Second, we will attempt to fully reconstruct the genomic region underlying wing-dimorphism in a focal species (*B. properans*) as well as the sequences of the long- and short-winged allele in order to understand how wing-size reduction evolved. Here, we aim to understand what kind of mutations differentiate both alleles, which genes are involved and which genomic characteristics drove these mutations. Third, we will test if the same genomic region underlies wing development in the different species. More precisely, we aim to test if the same mutations/genes/genomic regions are responsible for the repeated evolution of wing size variation in the different species.

Conclusions

The genomic basis of four wing-dimorphic species was investigated and allowed us to identify several genetic markers associated with wing-dimorphism. These markers were, at least for the more distantly related species, not homologous and suggests that wing-dimorphism evolved by independent mutational events in the different species. Reconstruction of the short- and long-winged alleles in the focal species *B. properans* showed that both alleles did not evolve by simple point-mutations, but rather by large scale genomic rearrangements in genomic regions that are very rich in repetitive and transposable elements (“jumping genes”). Although this is a most interesting finding, these repetitive sequences strongly complicate an accurate reconstruction of short- and long-winged alleles. To deal with this challenge, we applied the most state-of-the-art long-read sequencing technologies to accurately reconstruct the associated DNA sequences. We preliminary conclude that wing-dimorphism most likely evolves by independent large-scale genomic rearrangements in genomic highly dynamic regions. Through this project, we gained experience in

the most state-of-the art sequencing and bioinformatic tools, which opens novel research avenues at the RBINS.

Keywords

Morphological evolution, parallel evolution, genomics, dispersal polymorphism

SAMENVATTING

Context

Het herhaaldelijk evolueren en verdwijnen van identieke maar complexe kenmerken is één van de grootste raadsels in de evolutiebiologie. Tot voor kort werd dergelijke convergente of parallelle evolutie van kenmerken bestudeerd door de voorouderlijke staat van kenmerken te reconstrueren op basis van fylogenetisch bomen. Hoewel dergelijke methoden cruciaal zijn voor het beschrijven van de evolutionaire patronen, leveren ze slechts beperkte mechanistische informatie over hoe kenmerken juist herhaaldelijke kunnen verdwijnen of opnieuw evolueren.

Recente ontwikkelingen in de moleculaire en ontwikkelingsbiologie beginnen ons inzicht te verschaffen in de verschillende mechanismen die aan de basis liggen van parallelle evolutie, maar het ontbreekt ons nog steeds aan inzicht of dezelfde mechanismen een rol spelen over langere evolutionaire tijdschalen. De moeilijkheid om de genomische basis van dezelfde fenotypische eigenschappen bij niet-nauwverwante soorten terug te vinden, alsook om de volledige sequentie van het betrokken genomische gebied te reconstrueren, hebben de verdere vooruitgang in dit onderzoeksdomein sterk belemmerd.

Doelstellingen

In het voorgestelde project pakken we dit probleem aan door de genomische basis te onderzoeken van herhaalde evolutie van vleugelontwikkeling bij loopkevers. Binnen deze grote keverfamilie kunnen soorten gevonden worden met ofwel volledig ontwikkelde vleugels, ofwel gereduceerde vleugels. Echter, een groot aantal soorten is vleugel-dimorf waarbij we zowel individuen met goed ontwikkelde als gereduceerde vleugels kunnen terugvinden. Dit verschaft ons de unieke gelegenheid om de genomische regio die verantwoordelijk is voor de ontwikkeling van de vleugels binnen een soort te achterhalen. In een eerste doelstelling associëren we voor een aantal soorten de genetische variatie op genoombrede schaal met de vleugelontwikkeling om zo genetische markers te vinden die geassocieerd zijn met de ontwikkeling van de vleugels in elke soort. Ten tweede reconstrueren we het genoomgebied dat ten grondslag ligt aan het vleugeldimorfisme in de loopkeversoort *Bembidion properans*, alsook de sequenties van de allelen die coderen voor lange en korte vleugels om te begrijpen hoe de vleugelafmeting is geëvolueerd. Op basis van deze reconstructie trachten we te begrijpen wat voor soort mutaties beide allelen onderscheiden, welke genen betrokken zijn en welke genomische kenmerken deze mutaties hebben veroorzaakt. Ten derde zullen we testen of dezelfde genomische regio ten grondslag ligt aan de ontwikkeling van de vleugels in de verschillende soorten. Meer in het bijzonder willen we testen of dezelfde mutaties/genen/genoomregio's verantwoordelijk zijn voor het herhaaldelijk evolueren van vleugelgrootte in de verschillende soorten.

Besluiten

De genomische basis van vier vleugeldimorfe soorten werd onderzocht en stelde ons in staat om verschillende genetische merkers te identificeren die geassocieerd zijn met vleugelontwikkeling. Deze merkers waren, in ieder geval voor de minder nauw verwante soorten, niet homoloog en suggereren dat vleugeldimorfisme is ontstaan door niet-identieke mutaties in de verschillende

soorten. Reconstructie van het lang- en kortvleugel allel in de focus-soort *B. properans* toonde aan dat beide allelen niet door eenvoudige puntmutaties evolueerden, maar door grootschalige genomische herschikkingen in genomische regio's die zeer rijk zijn aan repetitieve en transponeerbare elementen ("jumping genes"). Hoewel dit een zeer interessante bevinding is, bemoeilijken deze herhalende sequenties een nauwkeurige reconstructie van korte- en langevleugel-allel sterk. Om deze uitdaging het hoofd te bieden, hebben we de meest geavanceerde sequentietechnieken toegepast ('long-read sequencing') om de bijbehorende DNA-sequenties nauwkeurig te reconstrueren. We komen voorlopig tot de conclusie dat het vleugel-dimorfisme waarschijnlijk evolueert door grootschalige en onafhankelijke genoomherschikkingen in dynamische regios in het genoom. Het project liet tevens toe om ervaringen te ontwikkelen in de meest *state-of-the-art* genomische en bioinformatische tools die belangrijke toepassingen zullen vinden in het evolutionair en taxonomisc onderzoek aan het KBIN.

Trefwoorden

Morfologische evolutie, parallele evolutie, genomica, dispersie polymorfisme

RESUME

Contexte

Depuis Darwin, les biologistes de l'évolution ont toujours été intrigués par l'évolution et la perte récurrentes de caractéristiques identiques mais complexes. Jusqu'à récemment, ils ont étudié une telle évolution convergente ou parallèle par l'estimation de caractères ancestraux sur base d'arbres phylogénétiques. Bien que cruciales pour la description des schémas évolutifs, ces méthodes ne fournissent que peu d'informations mécanistes sur la manière dont les caractères peuvent disparaître ou réévoluer à plusieurs reprises.

Des développements récents dans la biologie moléculaire et celle du développement créent déjà une impression des différents mécanismes soutenant l'évolution parallèle, mais nous ne savons toujours pas si ces mêmes mécanismes jouent un rôle lors des plus longs délais d'évolution. L'incapacité de déterminer la base génomique des mêmes caractères phénotypiques chez des espèces qui ne sont pas étroitement apparentés, ainsi que la difficulté de reconstruire la séquence entière de la région génomique concernée, ont fortement entravé les avancées dans ce domaine de recherche.

Objectifs

Dans le projet proposé, nous abordons ces questions par l'étude de la base génomique de l'évolution répétée du développement des ailes chez les carabes. Au sein de cette grande famille de coléoptères, certaines espèces ont des ailes complètement développées, d'autres des ailes réduites. Mais un grand nombre d'espèces présente un dimorphisme alaire, avec des individus aux ailes complètements développées et des individus aux ailes réduites. Ceci nous donne une occasion unique de trouver la région génomique responsable du développement des ailes au sein d'une espèce.

Le premier objectif, c'est d'associer, pour un nombre d'espèces, la variation génétique au niveau du génome au développement des ailes, afin de trouver des marqueurs génétiques associés au développement des ailes chez chaque espèce. Ensuite, nous reconstruirons, chez une espèce focale (*Bembidion properans*), la région génomique qui est à la base au dimorphisme alaire, ainsi que les séquences des allèles qui déterminent des ailes longues ou courtes afin de comprendre comment la dimension des ailes a évolué. Nous nous appuyons sur cette reconstruction pour comprendre quels types de mutation distinguent les deux allèles, quels gènes sont concernés et quelles caractéristiques génomiques ont provoqué ces mutations. En troisième lieu, nous allons tester si la même région génomique est à la base du développement des ailes chez les différentes espèces. Plus particulièrement, nous voulons tester si les mêmes mutations/gènes/régions génomiques sont responsables de l'évolution répétée de la dimension des ailes chez les différentes espèces.

Conclusions

La base génomique de quatre espèces présentant un dimorphisme alaire a été étudiée, ce qui nous a permis d'identifier plusieurs marqueurs génétiques associés au développement des ailes. Ces marqueurs n'étaient pas homologues, certainement pas pour les espèces moins apparentées, et suggèrent que le dimorphisme alaire est provoqué dans plusieurs espèces par des mutations indépendantes. La reconstruction de l'allèle des ailes courtes et celui des ailes longues chez l'espèce focale *B. properans* a démontré que les deux allèles n'ont pas évolué par de simples mutations ponctuelles, mais par des réarrangements génomiques massifs dans des régions génomiques très riches en éléments répétitifs et transposables (« jumping genes »). Bien que cette observation soit intéressante, ces séquences répétitives compliquent fortement une reconstruction précise de l'allèle

des ailes courtes et celui des ailes longues. Pour relever ce défi, nous avons appliqué les technologies sophistiquées de séquençage « grande longueur » de pointe pour une reconstruction précise des séquences d'ADN associées. Notre conclusion provisoire est que dimorphisme alaire évolue probablement par des réarrangements génomiques massifs et indépendants dans des régions génomiques très dynamiques. Le projet a également permis de développer des expériences dans les outils génomiques et bioinformatiques les plus modernes qui trouveront des applications importantes dans la recherche évolutive et taxonomique a l'IRSNB.

Mots-clés

Évolution morphologique, évolution parallèle, génomique, polymorphisme de dispersion

1. INTRODUCTION

How identical but complex traits are recurrently gained and lost during the time course of evolution remains an evolutionary paradox [1]. With the advent of recent developments in molecular and developmental biology, evolutionary biologists are beginning to explore the genomic mechanisms underlying the repeated evolution of complex traits. These first studies demonstrated that diverse molecular mechanisms may underlie repeated trait evolution [2], but we still lack a comprehensive view on how these different mechanisms are involved over different evolutionary timescales.
To understand the genomic basis of parallel trait evolution, it is not only necessary to target the genomic basis of a complex trait, but also to compare this genomic basis in closely as well as distantly related species. This difficulty to target the genomic region of the same trait in species with different evolutionary distances has so far strongly hampered insights on parallel evolution along evolutionary pathways.
In the proposed project, we tackle these issues by investigating the repeated evolution of a wing development mechanism that evolved repeatedly across the insect lineage of carabid beetles. We take full advantage of their unique property to show profound intraspecific variation in wing development. Associating this phenotypic variation with genetic variation at a genome-wide scale should allow us to trace back the genomic region underlying this trait in multiple species distributed across an entire beetle family and is therefore expected to add an important layer of insight into the recurrent evolution of complex traits, and the genomic basis of adaptive evolution in general.

State of the art

Along evolutionary distinct pathways, organisms often find similar solutions to adapt to the same environmental challenges. This phenomenon wherein the same character evolves repeatedly in different lineages is referred to as convergent or parallel evolution. Historically, a distinction has been made between convergent and parallel evolution, wherein **convergent evolution** refers to the repeated evolution of traits in distantly related taxa through independent genetic or developmental pathways. **Parallel evolution** refers to the repeated evolution of homologous traits by the same genetic mechanism and is therefore assumed to be most common among closely related species and/or populations.

The advent of high-throughput sequencing technologies [3,4] has revolutionized our insight in trait evolution and indicate that the previous distinction between convergent and parallel evolution is much less clear-cut as previously assumed [2,5]. First, a trait may evolve repeatedly by recurrent selection of identical alleles that were present as standing genetic variation in the ancestral population, or introgressed from another species [1]. This represents the clearest case of parallel evolution. Second, repeated evolution of the same trait can evolve by different and independent mutations in the same gene or even by a different genetic pathway resulting in exactly the same phenotype. For example, when the repeated loss of a trait is due to a deleterious mutation that silences the same gene, independently derived mutations may affect the same genetic pathway. This is an example of convergent evolution. Third, identical mutations in the same gene can occur independently in different lineages with the same phenotypic effect. In this latter case, the same molecular mechanism is involved (~ parallel evolution) but it evolved independently (~ convergent evolution)[5].

Distinguishing between these different mechanisms is clearly not just a semantic issue, but crucial if we want to understand the genetics behind adaptive evolution in general as it allows to distinguish between the respective roles **of ecological processes versus developmental and molecular mechanisms in adaptive evolution** [6]. If parallel evolution arises mainly by different *denovo* mutations in the same gene (~convergent), even in closely related species, this points into the direction that

mutations may easily allow organisms to adapt to ecological drivers. Conversely, if similar allelic variants underlie the repeated evolution of a character, even in more distantly related species (~parallel), this indicates that trait evolvability is strongly constraint by mutational rates and/or the availability of the standing genetic variation.

Investigating the molecular mechanism underlying convergent evolution of the same trait across species with different phylogenetic relationships would thus offer **unprecedented insight into the evolvability of ecologically important and complex traits**. This is not straightforward as it necessitates a system wherein a homologous trait evolved repeatedly in closely as well as distantly related species. Moreover, there should be enough intraspecific variation in the trait to trace back the genomic region associated with the trait of interest within each species.

Objectives

In the current project, we aim to get **a better understanding of the molecular mechanisms driving the repeated evolution of morphological traits** and test to what extent these **mechanisms differ among closely versus distantly related taxa**.

We will use the repeated evolution of **wing-dimorphism in (carabid) beetles as a model system**. Within the family of carabid beetles, and beetles in general, profound inter as well as intraspecific variation in the degree of wing development is present. At the interspecific level, species can either develop full wings and associated flight muscles that allow them to perform long distance dispersal by flight (*macropterous* species) or have strongly reduced or even a complete absence of wings (*brachypterous* species). While approximately half of the carabid species is either long-winged (*macropterous)* or short-winged (*brachypterous)*, a substantial number of species exhibits a remarkable **wing-dimorphism** (*wing-dimorphic* species) with some individuals developing full wings while others lack these flight structures completely. Empirical and theoretical work demonstrated that these dimorphisms evolved in response to local population fluctuations in landscapes with strong heterogeneity in habitat quality [7] and underlines the relevance for a proper understanding of the evolution of dispersal traits within the context of metapopulation and conservation research [8]. The evolution of macropterous, brachypterous and wing-dimorphic species are scattered throughout the entire phylogeny of carabids, which demonstrates that these highly distinct dispersal types evolved repeatedly within this beetle family.

Previously conducted breeding experiments in carabid beetles revealed that wing development in wing-dimorphic species is inherited according to the expectations of a single Mendelian element, with the allele coding for short wings being dominant over the allele coding for long wings . Remarkably, dispersive individuals do not only develop wings, but also a suite of other morphological adaptations such as flight muscles that enable them to disperse by flight. It remains currently less understood how an apparent single Mendelian element can result in a discrete switch of a suite of traits characterising each dispersal morph. This suggests that the locus determining wing development is either a gene that initiates a cascade of developmental processes associated with dispersal (a transcription factor) or a set of loci wherein the different genes coding for increased dispersal are in close physical linkage (“a supergene”) [9,10]. In this latter case, recombination between the different genes would, however, still result in recombinant offspring showing intermediate phenotypes. As these intermediate individuals are surprisingly never observed in wing-dimorphic species, it has been suggested that additional genomic features, such as chromosomal rearrangements, supress recombination in this region.

The first aim of the project is to fully characterize the wing-dimorphism locus in the focal species *Bembidion properans*. This will allow us to (i) determine the exact length of this locus; (ii) obtain the full sequence information of the haplotypes associated with the short- and long-winged allele and their variation, (iii) identify coding regions (genes) and transposable elements present within the locus and (iv) investigate the role of chromosomal rearrangements in the maintenance of the distinct haplotypes associated with both alleles and the exact location of the breakpoints of an inversion. This information is indispensable to reconstruct the evolutionary history of wing-dimorphism and the molecular mechanism that gave rise to these distinct phenotypes. **The second aim is to test in at least one closely related species and two more distantly related species to what extent wing-dimorphism in these species is based on the same mutations, genes, genomic region or genetic pathway compared to *B. properans*.**

2. METHODOLOGY AND RESULTS

General description of the workflow

The general workflow of the project is as follows (Fig. 1). First (WP1), species for which individuals either develop fully developed or reduced wings (i.e. wing-dimorphic species) will be selected and sampled. Those will be the species that are subject to the current study. Second (WP2), we aim at identifying genetic markers that are associated with wing development by means of RADtag sequencing. Once these markers are identified, we aim at reconstructing the full and complete genomic sequence that underlies the wing-dimorphism (i.e. the "wing locus") and aim to reconstruct the two alleles (haplotypes) that are responsible for either the long- and short-winged morph and identify nucleotide and structural variations and coding sequences between these two alleles (WP3). Finally, we want to test if the same mutations/genes/genomic regions underlying wing-dimorphism in the investigated species (WP4).

WP1: Species selection and sampling

We explored the tissue collection of the RBINS wherein an enormous collection of >83.000 freshly frozen carabid beetles (-80°C) is maintained. We selected candidate species based on the following criteria: (i) availability of large numbers of individuals and (ii) presence of both long- and short-winged individuals that occur in approximately equal proportions. The species *Bembidion properans*, *B. lampros*, *B. obtusum*, *Calathus melanocephalus*, *C. cinctus*, *Notiophilus palustris* and *N. biguttatus* fulfilled these criteria and were selected as candidate species. For comparative purposes, we also include data from a carabid species showing continuous, rather than dimorphic, variation in wing size i.e. the species *Pogonus chalceus*.

These species comprise an adequate set of both closely as well as distantly related species pairs and wing-dimorphic as well as a polymorphic species. Individuals were checked for the presence of wings, and a set of on average 20 long-winged and 20 short-winged individuals was selected for further downstream analysis.

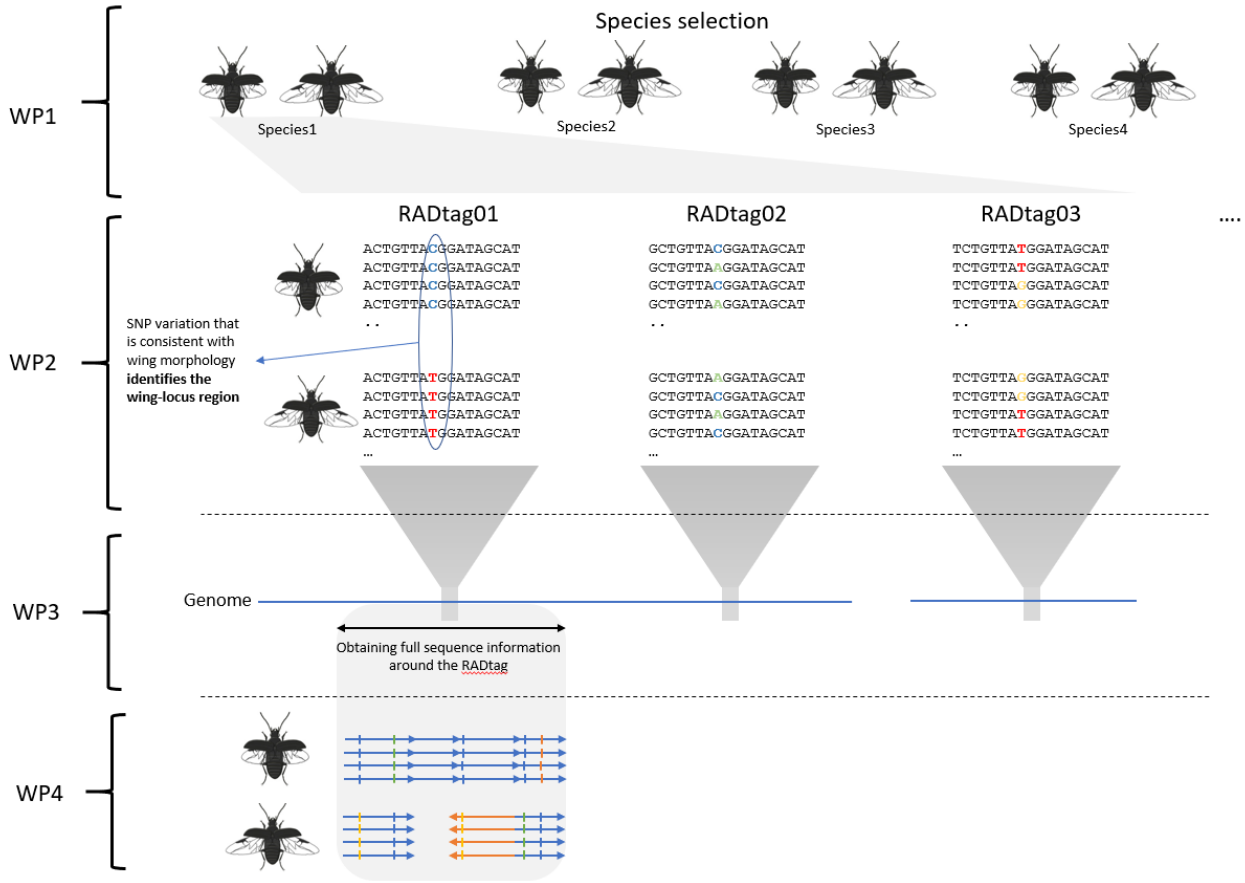


Fig 1. | Schematic representation of the general workflow to identify, characterize and compare the genomic region underlying wing-dimorphism in carabid beetles.

WP2: Identification of the genomic markers associated with wing-dimorphism in different species

In this work package, we identified the genomic region that is associated with wing-size variation in each species. As a general workflow, we conducted Restriction-site Associated DNA sequencing (RADseq), which is a method to obtain short sequence information for thousands of regions across the genome for multiple individuals. Given that we have no prior information on the length of the genomic region that is involved in wing development, we opted to perform RADseq with a restriction enzyme that uses a short sequence (6bp) as recognition site i.e. the enzyme PstI (recognition site GACGTC). Restriction enzymes with shorter recognition sequences will cut the genome more frequently, and therefore result in a higher number of RADseq tags per individual and thus a higher probability to capture the region of interest. This reduces the risk that the region of interest will not be targeted by a RADtag.

We currently generated RADseq libraries for three species i.e. *B. lampros* (13 long- and 16 short-winged), *B. properans* (10 long – and 22 short-winged) and *B. obtusum* (8 long- and 24 short-winged). We also integrate the previously available data of the wing polymorphic species *Pogonus chalceus*. RAD libraries were sequenced for a total of ~6.5M sequencing reads (150bp paired-end sequenced) and 2Gb per individual (May – September 2017). The sequencing reads were analyzed with the

STACKS [11] software package and resulted in approximately 70.000 RADtags and 200.000 single nucleotide polymorphisms (SNPs) that are present in at least 80% of the individuals for each species.

To identify the genomic region(s) associated with wing development for these species, two different approaches were used. First, measures of differentiation (*F*st) were calculated for each SNP, and SNPs showing the highest values are, thus, most differentiated between long- and short-winged individuals, and selected as likely candidates (Fig. 2). Second, we searched for SNPs and RADtags that show a genotype pattern across all individuals that is consistent with the wing phenotype. Because the short-winged allele is dominant, we looked for SNPs that are consistently homozygote in all long-winged individuals and heterozygote or homozygote for the alternative alleles in all short-winged individuals. For all four species, we could identify RADtags and SNPs that were highly divergent between long- and short-winged individuals and are therefore putatively associated with the genomic region that determines wing development of these species.

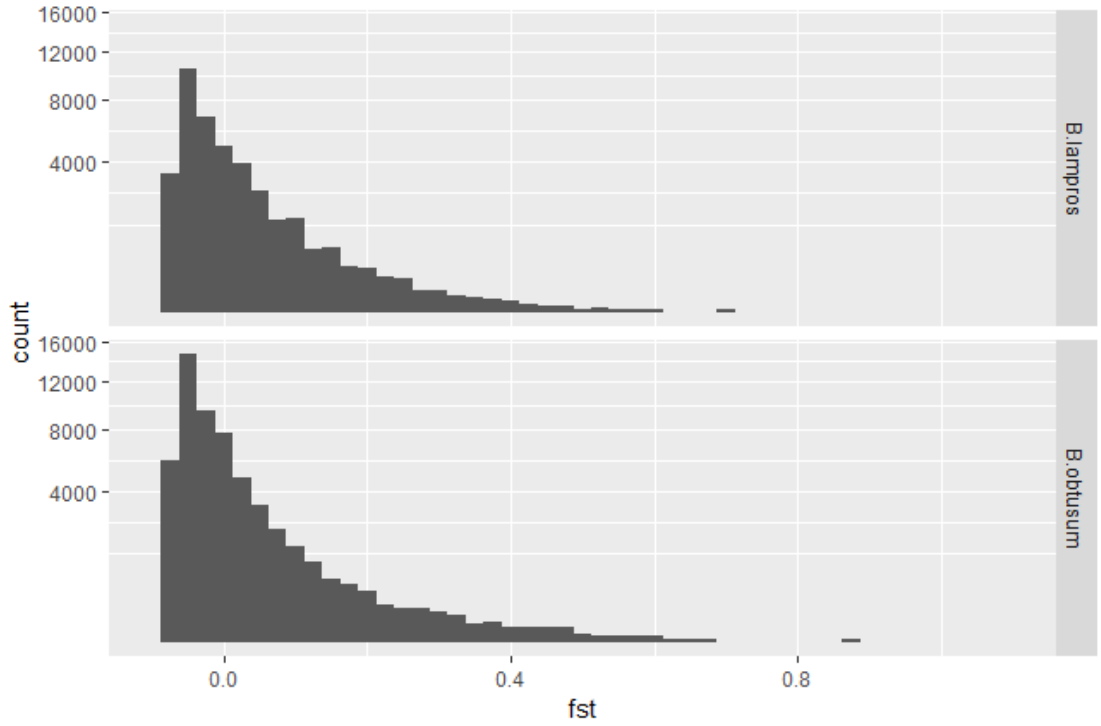


Fig. 2. | Distribution of Fst values between long- and short-winged individuals for *B. lampros* and *B. properans* as obtained by RADseq based on the PstI restriction enzyme. SNPs with high Fst value (situated on the right of the graphs) are putatively linked to the wing-development locus.

WP3: Obtaining full sequence information of the wing-locus for the focal species *B. properans*

We identified in WP2 a set of short sequences located near the locus that determines wing development. However, the short sequences of these tags can only be used as markers and do not allow to reconstruct the full sequence in this region. In the original project proposal, we proposed in to obtain full sequence information of the wing locus by means of the TLA method. This method uses

anchor sites (e.g. the RADtags associated with wing development) and generates a library of sequences that are crosslinked, and thus in close vicinity of the anchor sites. However, detailed inspection of the mapping of the RADtags (WP2) on the species *B. properans* strongly suggested the presence of a sequence that is paralogous to the "winglocus" (Fig. 3). Hence, the genomic region underlying the wing-locus is likely duplicated in the genome, which results in the presence of two highly similar sequences in the genome, but only one of them is likely involved in the development of the dimorphism. A consequence is that the anchor site will not bind uniquely to the genome, and that two different regions will be amplified simultaneously by the TLA method, resulting in a mixture of DNA sequences that do not allow to correctly assemble the true wing locus.

Fig. 3. | Graphical representation of the interpretation of nucleotide variation in a sequencer output. Variation in nucleotide composition as observed in the output of the genome assembler is generally interpreted as differences in the nucleotide sequence between the two alleles (haplotypes) in the diploid genome (a). However, when the sequence is present twice in the genome (paralog), but with small differences in the sequence (b), an identical output will be generated. Nucleotide variation at the wing-locus in the focal species *B. properans* is most likely due to the presence of paralogs.

Given the new developments and strong reduction in prices for genome sequencing, we opted to assemble a more accurate draft genome of the species, as this allows to characterize and separate the putative paralogous region. To achieve this, we currently applied two state-of-the-art genome sequencing and assembly methods on the focal species *B. properans* i.e. (i) single molecule real time (SMRT) sequencing on a Pacific Biosciences sequencing platform (https://www.pacb.com/) and (ii) Chromium 10x (https://www.10xgenomics.com/solutions/genome/). Pacific biosciences (PacBio) is currently one of the only two commercially available sequencing technologies that generate sequences of single and long DNA molecules. Previous NGS sequencing technologies, like Illumina, on which our previous assembly was based generates very high-quality reads of the genome, but their sequences are short (~100bp – 300bp). Although Illumina is currently the standard NGS sequencing technology, the short reads that the technology generates are less suitable to assemble genomic regions that contain long repeats and paralogous regions, as is likely the case for our focal species *B. properans*. PacBio sequencing in contrast produces reads reaching lengths of ~20.000bp and is therefore highly suited to bridge paralogous and repeat regions. The only drawback of this technology is that the proportion of errors in the sequencing reads are still relatively high compared to Illumina reads. A PacBio sequencing was outsourced to Macrogen Europe and a total of 5M reads

were generated, resulting in a total of 50Gb (~90x coverage, average read length 14kb). We sequenced two separate pools i.e one of short- and one of long-winged individuals.

Second, we made use of the 10x Chromium library preparation and sequenced a single heterozygous individual. Briefly, this method isolates single genomic DNA molecules, fragments them and provides each fragment with a unique barcode (Fig 4). Hence, sequences that have the same barcode are known to originate from the same single DNA molecule and allows for highly accurate assemblies and phasing (separation) of the diploid sequences. This technique was applied at the Leiden Genomics Technology Center, The Netherlands where 10x chromium was applied on a single heterozygous individual. The data were assembled with the Supernova assembler that is specifically designed to assemble and phase data generated by this methodology (https://support.10xgenomics.com/de-novo-assembly/software/pipelines/latest/using/running).

Given the recent developments in the assembly of genomes based on long-read sequencing data, we compared the performance of different genome assemblers (Table 1.). This clearly revealed the superiority of the PacBio data as this assembly was least fragmented (lowest number of contigs) and resulted in the longest contiguous sequences (highest average scaffold length, highest N50). However, the quality of the assembly depended strongly on the assembler algorithm that was used. For our data, comparable results were achieved with the Flye and wtdbg2 assembler [12]. Both assemblers have only very recently been developed (2019). Another important method to check the completeness of genome assemblies is to check the presence of gene sequences that are omnipresent in animals (e.g. insects, arthropods), called Benchmarking Universal Single-copy Orthologs (BUSCO[13]). If the majority of these genes are present in the assembled draft genome, this points towards a rather complete assembly. Also for this analyses, the assemblies based on Pacbio reads and assembled with Flye or wtdbg2 clearly scored best as up to 93% of the genes were recovered with this assembly method (Table 1). Based on the combination of these criteria, we selected the wtdbg2 assembly as the new draft assembly of the focal species *P. properans.* To target the contigs that underlie wing polymorphism for this species, we resequenced the entire genome of 10 long- and 10 short-winged individuals by short-read Illumina sequencing. These data were mapped with specific programs (BWA) [14] to the new reference genome and screened for the presence of SNPs with the Genome Analysis Toolkit GATK[15]. We then searched for the presence of single nucleotide polymorphisms (SNPs) or insertions/deletions that are consistently different between long- and short winged individuals. This analysis was based on both a home-made python script as well as statistical association method (PLINK[16]). Based on these data, we detected a total of ~35k positions that are mainly distributed over 7 candidate contigs.

Remarkably, most positions associated with the wing-dimorphism show a particular pattern of being absent in one of the two homozygote types rather than showing a segregating SNP pattern. This strongly suggests that the main difference between the two alleles mainly involves a large-scale structural variation rather than a simple nucleotide variation pattern (Fig. 6). This is in accordance with the increasing recognition that Mendelian elements often constitute large scale chromosomal rearrangements [10,17]. However, our finding that large scale deletions may underlie such changes was hitherto unknown and represents one of the most important findings of the current project.

Table 1. | Comparison of the different draft genomes based on different sequencing methodologies and different assembly algorithms. The N50 (and other Nxx values) corresponds to the length for which the collection of all scaffolds of that length or longer make up 50% (or xx) of the assembly length.

Assembly name	Bprop_platanus		Bprop_10x		Bprop_pacbio_canu		Bprop_pacbio_Flye		Bprop_pacbio_wtdbg2	
Library preparation method	170bp, 500bp and 800bp PE		Chromium 10x		20kb SMRTbell templates		20kb SMRTbell templates		20kb SMRTbell templates	
	2kb and 5kb mate-paired libraries									
Sequencing methodology	Illumina HiSeq		Illumina NovaSeq		PacBio Sequel		PacBio Sequel		PacBio Sequel	
Assembler	Platanus		Supernova		Canu		Flye		Wtdbg2	
Contiguity	**Size (bp)**	**Number**	**Size (bp)**	**Number**	**Size (bp)**	**Number**	**Size (bp)**	**Number**	**Size (bp)**	**Number**
Total size	382,308,441	809,071	362,837,618	60,485	905,528,555	24,612	362,613,467	6,887	379,237,563	7,316
% of estimated genome size1	68%		64%		161%		64%		67%	
Average scaffold length	473		5,998		36,792		52,651		51,837	
Largest scaffold	464,986		522,968		1,724,414		7,724,900		8,477,330	
N's			5,708,380		-		3,200		-	
Gaps			9340		-		32		-	
N50	3,048	10,630	14,989	4,616	45,892	4,732	698,429	86	827,019	81
N60	864	37,850	9,265	7,730	36,335	6,963	237,953	180	211,867	183
N70	466	100,118	5,736	12,774	29,309	9,742	127,744	395	74,554	514
N80	211	217,663	3,489	20,938	23,398	13,203	73,693	775	34,512	1,295
N90	126	467,521	2,007	34,628	17,688	17,639	36,107	1,472	16,581	2,896
Completeness										
BUSCO (n = 1658)	62%		87%		71%		93%		84%	
Single [C]	61.60%		83%		37.60%		91%		82%	
Duplicated [D]	0.80%		4%		33.10%		2%		2%	
Fragmented [F]	19.50%		7%		7.50%		4%		9%	
Missing [M]	18.10%		6%		21.80%		3%		8%	

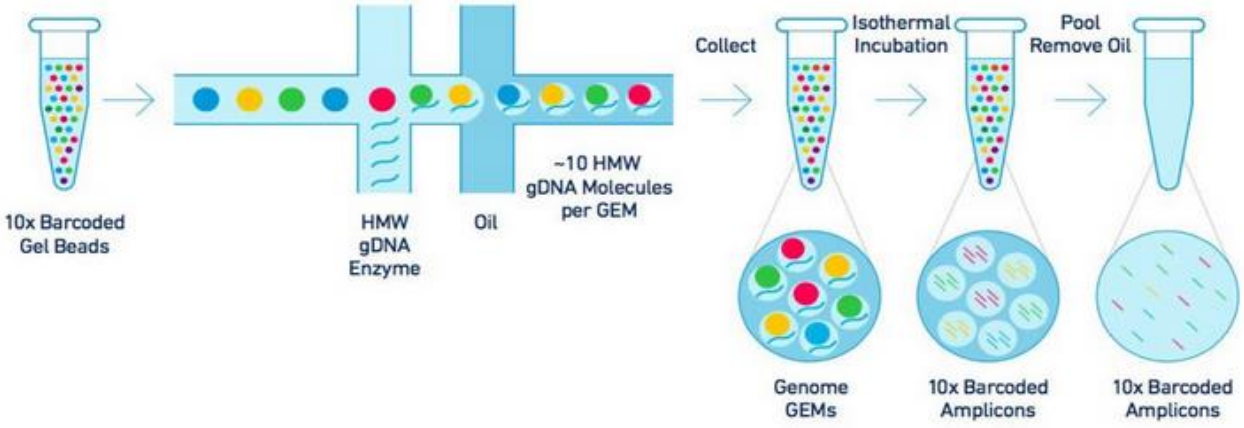


Fig. 4. | Overview of the 10x chromium methodology. Long genomic DNA molecules (HMW gDNA) are combined with 10x barcoded gel beads in aqueous droplets in an oil emulsion. The gDNA molecules are fragmented in the droplets and ligated with the barcodes, resulting in a library wherein all fragments originating from the same gDNA molecule have the same barcode. Sequencing of the small fragments with their barcode allows to link reads that originate from the same gDNA molecule (source https://www.10xgenomics.com/solutions/genome/).

We are currently assembling the two haplotypes by linking the detected contigs with 10x and Pacbio data. This region appears very difficult to assemble with more standard methods because of the following reasons. First, the genomic region that we identified previously as being the focal region appeared to show nucleotide variation that is consistent for a bi-allelic locus as more than two haplotypes (alleles) were identified for this region. This strongly indicates that the sequence is likely not unique, and that a highly similar sequence occurs twice in the genome (Fig. 5). Genes for which more than one copy are present in the genome are called paralogs, and strongly complicate the accurate assembly of genomic regions. Consequently, manual curation, assembly and scripting is required for the genome assembly, which is very time-consuming and labor intensive. Second, the region also appears to be characterized by a high density of repeat regions (i.e. short sequences that are scattered throughout the genome and often of retroviral origin)(Fig. 7). As for paralogous regions, these repeat regions complicate an accurate assembly because of the repeated occurrence of identical sequences. Third, the two alleles are likely very divergent, which makes it rather difficult to assign homology between the haplotypes and the sequence of each haplotype is likely present as a separate sequence in the genome assembly.

The fact that the wing locus region is very difficult to assemble is, however, indicative that it is a highly complex genomic region. Although this renders the assembly time consuming and challenging, it also promises that highly interesting molecular mechanisms underly the divergent evolution of these alleles.

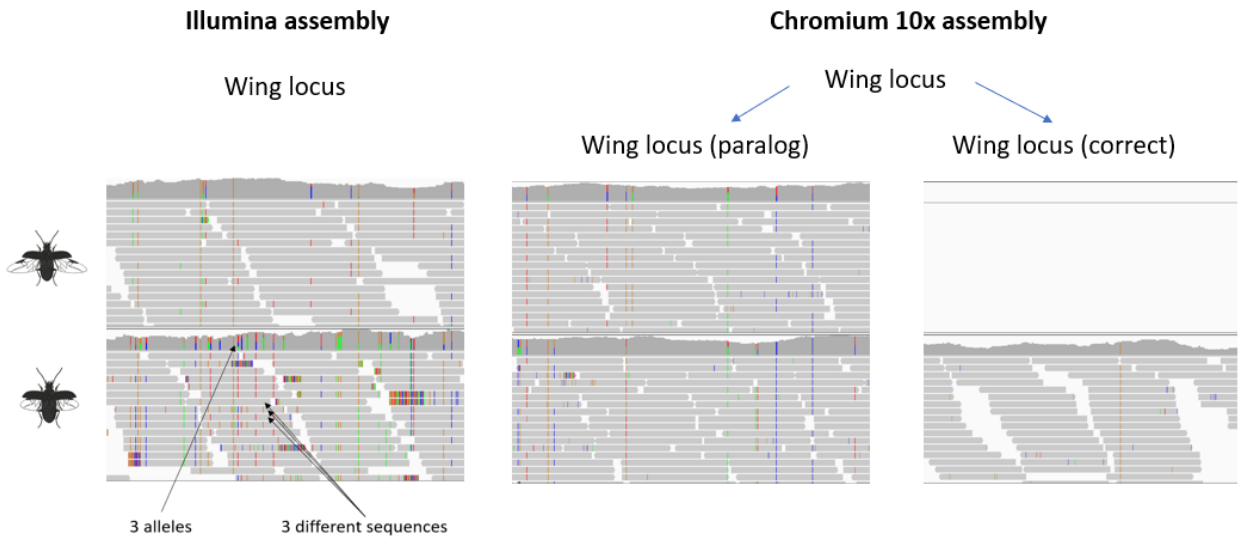


Fig. 5. | Results of the mapping of the resequencing data of a single long- (above) and sort- (below) winged individual within the target wing locus region. Left panel shows the mapping on the initial draft genome and reveals that three different haplotypes are present in this region, indicating that the sequence is paralogous. Mapping of the reads to the new assembly (10x) shows that the two paralogs are correctly assembled in two different sequences, wherein part of the allele determining the long-winged phenotype is deleted (no reads present).

WP4: Comparison of the dispersal alleles among the different species and reconstructing the evolutionary history of dispersal alleles.

Reconstructing the evolutionary history of both dispersal alleles can only be initiated when the haplotypes of both alleles are reconstructed with high accuracy. As we are still within the phase of assembling the dispersal locus and both alleles, this is WP is planned for the forthcoming months. We already obtained the raw data to start this analysis as single individuals of the related species *B. lampros*, *B. obtusum* and *P. chalceus* have already been resequenced at whole genome level (Macrogen Europe, June 2018) and should therefore allow for a swift completion of this workpackage.

For the wing-polymorphic species *P. chalceus*, RADseq data were used to infer the evolutionary history of alleles that are involved in the contemporary evolution of short- and long-winged populations[18]. We reveal that alleles selected in short-winged populations are spread across the genome and evolved during a singular and, likely, geographically isolated divergence event, within the last 190 Kya. Hence, these alleles are much older than the divergence times between the current evolution of the short-winged populations. Due to subsequent admixture after this initial and old divergence, the ancient and differentially selected alleles are currently polymorphic in most populations across its range, which could potentially allow for the fast evolution of one ecotype from a small number of random individuals, as low as 5 to 15, from a population of the other ecotype. An important implication of this finding is that cases of fast parallel ecological divergence can be the result of evolution at two different time frames: divergence in the past, followed by repeated selection on the same divergently evolved alleles after admixture. These findings highlight the importance of an ancient and, likely, allopatric divergence event for driving the rate and direction of contemporary fast evolution under gene flow. This mechanism is potentially driven by periods of geographic isolation imposed by large-scale environmental changes such as glacial cycles.

Fig. 6. | Mappings of Pacbio data from long- and short-winged individuals on two contigs that are associated with wing-dimorphism in *B. properans*. These contigs are clearly only present in the short-winged individuals (multiple mappings), but not in the long-winged individuals, demonstrating that complex structural variations underlie both alleles.

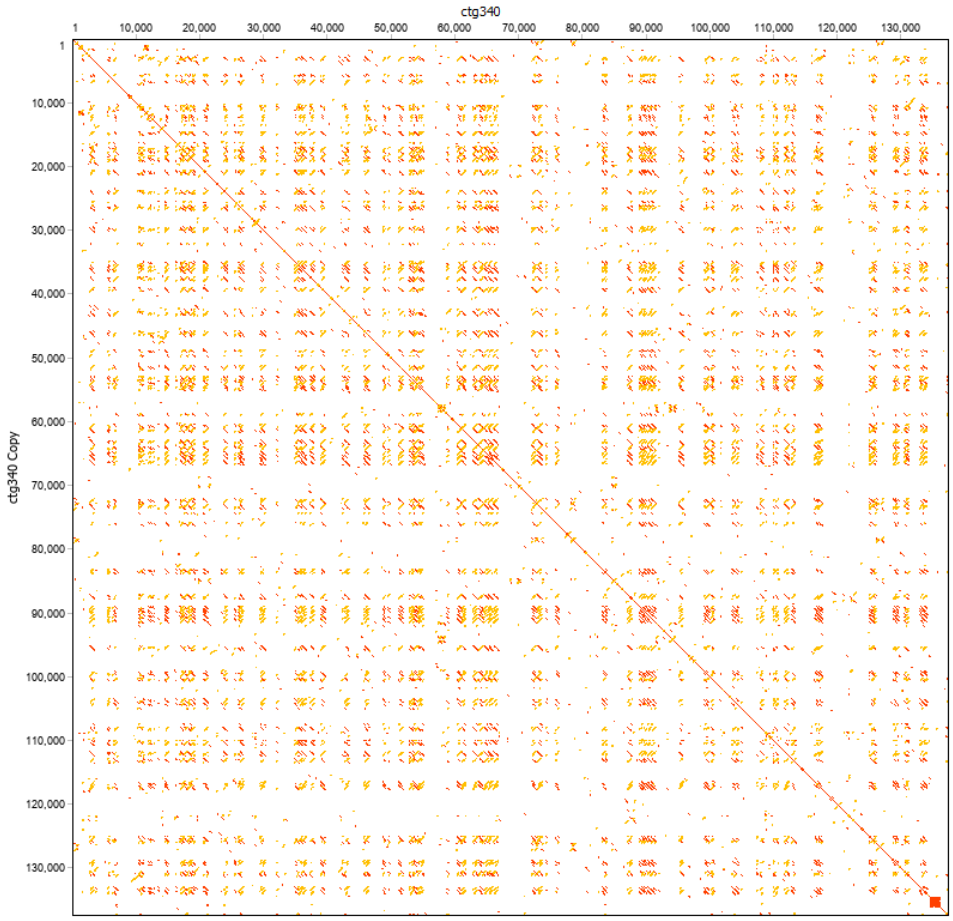


Fig. 7. Self-alignment of the sequence of a major contig associated with wing development in *B. properans* (ctg340) reveals that these contigs consist of short sequences that are repeated multiple times throughout the entire sequence (yellow and red line fragments).

List of used abbreviations:

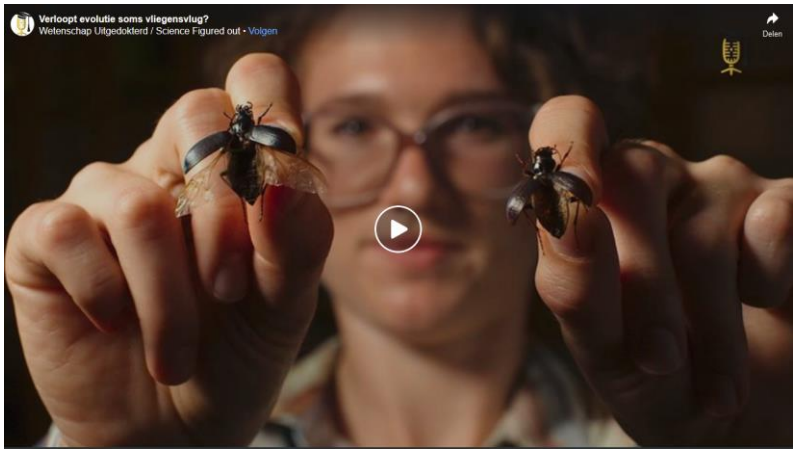
- NGS: next-generation sequencing
- bp: base pair in the DNA sequence
- TLA: targeted locus amplification
- RAD: Restriction site associated DNA sequencing

3. DISSEMINATION AND VALORISATION

Preliminary results of the work have been presented at international symposia and seminars in France, The Netherlands, Italy, Germany and the US. Most notably is the presentation of the work during an Invited Plenary Lecture at the European Carabidologist meeting in Rennes, France.

- Presentation of preliminary results during a key-note lecture (invited) at the 18th European Carabidologist Meeting (ECM), Rennes, Brittany, France, 24th - 29th September 2017.
- Invited seminar at the University of Torino (Host: Prof Dr. M. Isaia), Turin, Italy 2017.
- Invited seminar at the PhD course "Principles of Ecological and Evolutionary Genomics" within the framework of SENSE ("Research School for Socio-economic and Natural Sciences of the Environment"), Wageningen, the Netherlands, 28th September 2018.
- Invited seminar at University of Greifswald, Germany (host: Prof. Dr. Gabriele Uhl) "Rapid adaptation in dispersal traits" within the framework of "PlanetErde" seminars. 27th November 2018.
- Presentation at the "Genome assembly and annotation" workshop, 11th – 15th February 2019, Berlin, Germany.
- Presentation at 'Evolution 2019' meeting, Providende (US). 21th-25th June 2019. " Ecological and genomic drivers of the repeated evolution of wing dimorphism in carabids" (by Zoë De Corte).
- Presentation of preliminary results in the department of biology at the University of Rochester 27th February 2020 (by Zoë De Corte)

We also presented the work for a wide community through the "Wetenschap uitgedokterd" forum wherein PhD student Zoë De Corte presented the project in a video presentation (https://www.wetenschapuitgedokterd.be/verloopt-evolutie-soms-vliegensvlug) intended to reach a very wide community of non-experts.



Importantly, the project allowed us to gain ample experience in the application of highly novel sequencing technologies (10x genomics, PacBio sequencing) that was disseminated to other researchers and research groups in evolutionary biology at RBINS. This dissemination took place through a tight collaboration with the Joint Experimental Molecular Unit (JEMU) as well through monthly organised 'EVOLUNCHES' held within the RBINS.

4. PERSPECTIVES

As described above, the complexity of the genomic sequence underlying wing-dimorphism in the focal species *B. properans* renders the assembly of the wing locus highly challenging and time-consuming. Although we anticipated on this problem by using the most state-of-the-art genomic techniques to tackle these problems, the timing of the finalization and publication of the results has been delayed. Yet, several initiatives have been initiated to continuate this promising research. We particularly anticipated this by exploring opportunities for the continuation of the project. In concreto, Zoë De Corte (who is currently working as a researcher at the project) applied for (i) a Fulbright scholarship to continue the research in collaboration with the Brisson lab (University of Rochester, US, http://www.sas.rochester.edu/bio/people/faculty/brisson_jennifer/index.html) and received this prestigious grant. She continued the research from 09/2019 till 01/04/2020 at the Brisson Lab (https://www.brissonlab.org/lab-members) where she made considerable progress in the assembly of the wing-locus. We also applied for a FNRS as well as FRIA PhD grant, to continue the research in close collaboration with Prof. K. Van Doninck at the University of Namur (http://www.lege-unamur.be/contact.htm). Although the project proposal was rated very good to excellent, the grant was not funded because of financial constraints.

Lastly, Zoë De Corte recently obtained the 2020 Evolution, Ecological and Conservation Genomics (EECG) Research Award from the American Genetics Association (https://www.theaga.org/news-detail.php?news=80)($7650) to continue this work. More precisely, the grant will be used to conduct RNA sequencing in order to functionally annotate the entire genome and particularly the genes that are located in the wing-locus.

In sum, the pioneer project PARAWINGS allowed us to set up a new line of research that already resulted in strong national and international collaborations that will likely be extended in the future.

5. PUBLICATIONS

Van Belleghem, S.M., Vangestel, C., De Wolf, K., De Corte, Z., Rastas, P., Möst, M., De Meester, L. & F. Hendrickx (2018). Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLoS Genetics* 14 (11): e1007796 [most recent I.F. = 6.1]

. This publication was rated as very good by F1000.

. This publication received considerable attention in the media, and was covered on the VRT Newssite, Gazet van Antwerpen, Nieuwsblad and Belga and had an estimated reach of about 1.5 people.

WETENSCHAP

Een schorreloopkever (Pogonus chalceus). De vleugels zitten onder de dekschilden.

Snelle evolutie door "ontdooide" genen uit de voorlaatste ijstijd

Luc De Roy

Organismen kunnen verbazend snel evolueren door oeroude genvarianten, die ooit nuttig waren, opnieuw in te schakelen. Dat hebben onderzoekers van het Koninklijk Belgisch Instituut voor Natuurwetenschappen (KBIN) ontdekt bij kevers. Begrijpen hoe soorten erin slagen zich snel aan te passen, is belangrijk in deze tijden van plotse veranderingen in het klimaat en de omgeving.

Evolutie is een langzaam proces, en planten en dieren evolueren doorgaans ontzettend traag. Ze hebben nieuwe genvarianten nodig om te evolueren, en die ontstaan alleen door zeldzame mutaties in het DNA.

Toch zien biologen dat sommige populaties zich razendsnel aanpassen aan een nieuwe omgeving. Dat is onder meer zo bij de schorreloopkever (*Pogonus chalceus*): individuen met lange vleugels, die in een moeras leven dat één keer per jaar onder water staat, evolueren in amper twintig generaties - in dit geval ook twintig jaar - tot een kleiner en kortvleugelig type als ze een moeras koloniseren dat elke dag blank komt te staan. Evolutionair gezien is twintig jaar een oogwenk.

Voorlaatste ijstijd, 200.000 jaar geleden

Onderzoekers van het Koninklijk Belgisch Instituut voor Natuurwetenschappen wilden het mechanisme achter die razendsnelle evolutie ontrafelen, en screenden het volledige genoom van verschillende populaties schorreloopkevers.

. The study was selected as cover image for the issue of *PLoS Genetics*

6. ACKNOWLEDGEMENTS

The genomic analyses 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.

7. REFERENCES

1. Rosenblum EB, Parent CE, Brandt EE. 2014. The Molecular Basis of Phenotypic Convergence. *Annu Rev Ecol Evol Syst.* 45(1):203-226.

2. Stern DL. 2013. The genetic causes of convergent evolution. *Nat Rev Genet*. 14(11):751-764.

3. Metzker ML. Sequencing technologies - the next generation. 2010. *Nat Rev Genet*. 11(1):31-46.

4. Goodwin S, Mcpherson JD, Mccombie WR. 2016. Coming of age : ten years of next- generation sequencing technologies. *Nat Rev Genet*. 17(6):333-351.

5. Arendt J, Reznick D. 2008. Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends Ecol Evol*. 23(1):26-32.

6. Reznick DN, Losos JB, Travis J. 2018. From low to high gear : there has been a paradigm shift in our understanding of evolution. *Ecol Lett*. 22(2):233-244.

7. Hendrickx F, Palmer SCF, Travis JMJ. 2013. Ideal free distribution of fixed dispersal phenotypes in a wing dimorphic beetle in heterogeneous landscapes. *Ecology*. 94(11):2487-2497.

8. Fountain T, Nieminen M, Sirén J, Chong S, Lehtonen R, Hanski I. 2016. Predictable allele frequency changes due to habitat fragmentation in the Glanville fritillary butterfly. *PNAS*. 113(10).

9. Thompson MJ, Jiggins CD. 2014. Supergenes and their role in evolution. *Heredity.* 113(1):1-8.

10. Schwander T, Libbrecht R, Keller L. 2014. Supergenes and complex phenotypes. *Curr Biol.* 24(7):R288-R294.

11. Rochette NC, Catchen JM. 2017. Deriving genotypes from RAD-seq short-read data using Stacks. *Nat Protoc.* 12(12):2640-2659.

12. Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*. 17:155-158.

13. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva E V, Zdobnov EM. 2015. Genome analysis BUSCO : assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31:3210-3212.

14. Li H, Handsaker B, Wysoker A, et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25:2078-2079.

15. McKenna A, Hanna M, Banks E, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 20(9):1297-1303.

16. Purcell S, Neale B, Todd-brown K, et al. 2007. PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*. 81:559-575.

17. Joron M, Frezal L, Jones RT, et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*. 477:203-206.

18. Van Belleghem SM, Vangestel C, Wolf K De, et al. 2018. Evolution at two time frames : Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLoS Genet*. (e1007796):1-26.

ANNEXES

NA