The genetic basis of a regionally isolated sexual dimorphism involves *cortex*

Kalle Tunström[1,2,*], Ramprasad Neethiraj[1], Naomi L.P. Keehnen[3], Alyssa Woronik[4], Karl Gotthard[1], and Christopher W. Wheat[1]

[1]Department of Zoology, Stockholm University
[4]Sacred Heart University
[3]Swedish University of Agricultural Sciences
[2]Lund University
[*]kalle.tunstrom@gmail.com

January 25, 2024

Abstract

Sexual dimorphisms represent a source of phenotypic variation and result from differences in how natural and sexual selection act on males and females within a species. Identifying the genetic basis of dimorphism can be challenging, especially once it is fixed within a species. However, studying polymorphisms, even when fixed within a population, can provide insights into the genetic basis of sexual dimorphisms. In this study, we investigate the genetic basis of a regionally isolated sexual dimorphism in the wings of *Pieris napi adalvinda*, a subspecies of *P. napi* found in northernmost Scandinavia, where females exhibit heavily melanized wings. By using a combination of male and female informative crosses, genomic sequencing of melanic outliers, and a population genomic analysis with a new reference genome for the melanic morph, we demonstrate that the female-limited morph adalvinda is caused by a single dominant allele at an autosomal locus upstream of the gene *cortex*. This novel finding adds to the growing body of literature that connects repeated mutations in and near the cortex gene to the regulation of butterfly wing melanization, providing insights into the evolution of sexual dimorphisms and the recruitment of genes into monomorphic or sex-limited forms. This study thus highlights the significance of cortex as a basis for a female-limited trait and lays the foundation for future comparative analyses of dimorphism genetic underpinnings.

1 Introduction

Phenotypic variation within species exists at various levels, such as within populations as polymorphisms, between populations as local adaptation and between sexes as dimorphisms. Of course, these examples are not mutually exclusive, with some components combining for complex patterns where for example polymorphisms, or local adaptations simultaneously can exist as a sex-limited polymorphism (Vane-Wright, 1975). As we gain more understanding of the genotype to phenotype relation for such traits, we are able to ask questions regarding patterns in the genetic architecture. For example, if a similar phenotype evolves independently in multiple different taxa, understanding their genetic basis provides insights into the degree of evolutionary parallelism. Unfortunately, since many sexual dimorphisms are old, resolving their genetic basis and origin is difficult (Monteiro and Podlaha, 2009). However, if dimorphisms are the result of fixation single-sex polymorphisms, studying single-sex polymorphisms may supply insights in the genetic basis and evolution of sexual dimorphisms.

Across the order Lepidoptera, increased or decreased wing-melanization has evolved repeatedly in response to a range of selection pressures, such as thermo-regulation (Kingsolver and Wiernasz, 1991), camouflage (van't Hof et al., 2016; Kettlewell, 1973), sexual attraction (Ellers and Boggs, 2003), and immunity (Wittkopp and Beldade, 2009). Importantly, wing-melanization can often be dimorphic, e.g., the sexually dimorphic Asian pierid butterfly *Appias* nero where males are always bright orange but the females exhibit a range of locally isolated melanized morphs (Vane-Wright and Treadaway, 2011), or the classic Batesian mimic *Papilio glaucus* where females either have a male-like morph or a heavily melanized and female-limited form (Koch et al., 2000), to the more subtle variation in ventral dusting of melanin between males and females of many *Colias* butterflies. In other systems, such as *Biston betularia, Heliconius spp.* and *Vanessa cardui*, the genetic basis of variation in wing melanization is being revealed (Hanly et al., 2022; van't Hof et al., 2019; Koch et al., 2000; Nadeau et al., 2016; Zhang et al., 2017). Here, by focusing on the repeated evolution of changes in wing melanization among the Lepidoptera, we will begin to discuss how different sources of selection and different genetic architectures interact. We use this focus to frame our investigation into how evolution and selection may generate current patterns of sexual dimorphisms.

Melanin is a pigment found in a range of taxa, and in addition to its role in coloration it is involved in a range of important and diverse processes in insects, such as immune defense and neural development (Wittkopp and Beldade, 2009). As such, the melanin biosynthesis pathways have been extensively studied in a range of insects, and we have a good understanding of its genetic basis. In fact, *cis*-regulatory and coding mutations in genes directly involved in the melanin biosynthesis (e.g., *yellow, tan,* and *ebony*) are known to be responsible sex- and species-specific differences in body coloration of *Drosophila* (Jeong et al., 2008; Signor et al., 2016; Yassin et al., 2016). Surprisingly, while the existence and function of these genes is conserved in Lepidoptera (Matsuoka and Monteiro, 2017; Zhang et al., 2017), variation at these genetic loci is rarely associated with any species or sex-level differences in wing coloration (But see:(Martin et al., 2020)). In fact, early investigations of the industrial melanization in *Biston betularia* were unable to detect the causal locus using a candidate gene approach focused on known biosynthesis pathway genes (van't Hof and Saccheri, 2010). Instead, in Lepidoptera, nearly all of the wing color examples where genotype to phenotype connections have been made involve developmental patterning genes (Deshmukh et al., 2018; Nadeau, 2016). One gene that repeatedly has been associated with melanization is the cell-cycle regulator gene *cortex.* In *Biston betularia* (as well as other geometrid moths), once a genome-wide approach was applied, the causal locus was found to be a transposable insertion located in the intronic region of *cortex* (van't Hof et al., 2016; van't Hof et al., 2019). Independent genome wide studies of mimicry morphs in *Heliconius* butterflies found *cortex* also involved in a number of melanized morphs. In *H. erato,* a large number of SNPs, mostly located in the intronic regions *cortex* are associated with the formation of a yellow bar in the black region of the hind wing (Nadeau et al., 2016). In *H. elevatus* and *H. melpomene,* a similar yellow bar phenotype is associated with variants near the 5' UTR of *cortex* (Dasmahapatra et al., 2012; Nadeau et al., 2016). Finally, also in *H. melpomene,* a domesticated ivory morph, where melanin is absent, is caused by a large deletion of a previously unknown 5' UTR exon of *cortex* (Hanly et al., 2022). As such these repeated and independent cases of mutations in non-coding regions in and around the *cortex* locus suggest that it may be a hotspot locus for regulating wing patterning in butterflies. However, despite a number of butterfly species exhibiting sex-specific melanic morphs, whether these morphs are also controlled by *cortex* is unknown. If continued comparative analyses of the gene *cortex* reveals a ubiquitous role for the formation of melanic morphs among other insect species, comparison of studies with species in which it does or does not cause sexual dimorphisms, and more unbiased, could provide insights into how different sources of selection and different genetic architectures interact to generate of sexual dimorphisms.

Here we investigate the genetic basis of a sexual dimorphisms, in the form of a female-limited wing-melanic morph in *Pieris napi adalvinda,* a regionally isolated morph and subspecies of *P. napi* local in the northern most corner of Scandinavia (Petersen, 1949). Using a combination of long read sequencing, population genomics, and bulk segregant analysis of mapping populations, we construct a new reference genome for *P. n. adalvinda,* as well as associate female-wing melanization to a stretch of *adalvinda*-unique content upstream of the gene *cortex,* likely resulting from the insertion of a transposable element. We hypothesize that this TE-insertion contains tissue- and sex-specific transcription factors responsible for the female-limited expression. In closing, we note the similarity of the genetic architecture of this *adalvinda* locus and the *Alba* locus in *Colias* butterflies, discussing the implications for the evolution of sexual dimorphisms in butterflies in general.

2 Material and methods:

2.1 Rearing

Adult wild female *P. napi* butterflies were caught in Spain (Parc del Aiguamolls de l'Empordà, north-east of Barcelona, 42.23°N, 3.10°E) and norther Sweden (Abisko 68.36°N, 18.79°E) and brought to the laboratory at Stockholm University and allowed to oviposit on *Alliaria petiolata.* The offspring were reared separated by family on *A. petiolate* and *Armorancia rusticana* under long day conditions (Light: Dark 24:0 at 20°C) to promote direct development. These offspring were crossed to generate reciprocal F1 hybrids (female first: Abisko*Spain and Spain*Abisko) and pure populations (Abisko*Abisko and Spain*Spain), under short day conditions (L:D 8:16, 17°C) to induce diapause, which is known to follow the Spanish population in the hybrid crosses. In the next spring of 2014, these populations were used to generate three F2 backcrosses: SS*SA, SS*AS, SA*SS, under long day conditions L:D 23:0, 23°C. Four families were selected from SS*SA (10, 12, 21, 39), three from SS*AS (23, 31, 53), and one from SA*SS (206) for further experiments. For all crosses male and female identity were tracked, each female was allowed to mate only once, eggs were counted, each individual offspring was sexed, weighed at pupation, and recorded eclosion date and pupation date as well as hatching date

2.2 Image analysis

Images were taken with a Nikon D5100 DSLR camera using AF-5 Micro NIKKOR 60mm lens with a Sigma EM-140DG ring flash setup. In order to avoid any variation between the photographs, all photos were taken in a completely dark room against a blue background with the distance between the camera and the wings kept constant. Raw NEF files were converted to JPEG format using the convert function in ImageMagick (v7.0.0-0, https://www.imagemagick.org), and White balancing on the converted images was performed using batch-levels-stretch function in GIMP (v2.8, https://www.gimp.org/). We used the thresholder function of ImageJ (v1.49, (Schneider et al., 2012)) to capture the total number of yellow and black pixels on the wing, and the number of black pixels were divided by the total to obtain the ratio of the forewing covered by melanin (black pixels). Based on the proportion of melanin on the forewings, we selected the most-melanic and the least-melanic individuals from specific families to form our melanic and non-melanic groups for whole genome sequencing using a bulk segregant analysis (BSA) strategy.

2.3 Pooled sequencing

The genomic DNA (gDNA) for the BSA and the population comparisons was isolated either from the thorax or from the abdomen of adult butterflies using the E-Z 96 Tissue DNA kit (Omega Biotek, CA, USA). DNA integrity and quantity was quantified using 1% agarose gel electrophoresis and a fluorescence-based Qubit assay (Qubit, Thermo Scientific, MA, USA). Individual gDNA was pooled at equal concentrations, and if necessary, concentrated using Microcon centrifugal filters for DNA fast flow (Merck Milipore, Tullagreen, Ireland) and measured again using the Qubit assay. For the male- and female-informative crosses, gDNA was pooled by family and color-morph (melanic, non-melanic), and for the population pools DNA was pooled by their origin (Abisko, and Skåne).
The pools were sequenced using Illumina TruSeq 300bp insert libraries on an Illumina HiSeq 4000 with paired-end (PE) 101bp reads at Beijing Genome Institute (BGI). Family 23 high and low melanin pools generated 370 M reads (97% >= Q20) and 369 M reads (97% >= Q20), respectively. Family 21 high and low melanin pools generated 355 M reads (96% >= Q20) and 352 M reads (96% >= Q20), respectively. Family 206 high and low melanin pools generated 357 M reads (96% >= Q20) and 354 M reads (96% >= Q20), respectively.

2.4 Nanopore sequencing

Pieris napi adalvinda females were collected from Abisko Östra, Sweden (68.350, 18.835) and transported to Stockholm alive and allowed to oviposit on *Alliaria petiolata.* The offspring was then reared until pupation and diapause at 17C with light: dark-cycle of 12:12. High molecular weight genomic DNA was extracted from the middle third portion of one female pupa using a slightly modified protocol for paramagnetic nanodiscs (Nanobind Tissue Big DNA kit, Circulomics). Prior to extraction the pupal section was frozen in liquid nitrogen and ground into a fine powder using a ceramic pestle. The powdered tissue was then washed three times in 700μl cold buffer CT and HMW DNA subsequently isolated according to the manufacturer's recommended protocol for insect samples. The isolated DNA was then treated with Short Read Eliminator XL (SRE-XL, Circulomics) to selectively precipitate high molecular weight DNA (>20kb fragments). Isolated and size-selected DNA was sequenced on MinION platform using one R9.4.1 flow cell and ligation-based library prep LSK109. The library was split into three aliquots, each sequenced for 20h before the flow cell was washed using the flow cell wash kit (EXP-WSH003) and reloading the flow cell. Once sequencing finished, the raw reads were basecalled using Super High Accuracy basecalling mode in GUPPY v.5.0.2.

2.5 Genome assembly

From the base-called reads we assembled a draft genome assembly using Flye v2.9 using the default settings for nanopore reads basecalled with Super high accuracy mode (nano-hq) followed by two iterations of polishing with Flye v2.9 (Kolmogorov et al., 2019). Haplotype redundancies were identified and purged from the draft assembly using Purge_dups v1.2.5, default settings for nanopore data (Guan et al., 2020). Finally, we polished and separated two alternative haplotypes using HapDup v.0.7 (Kolmogorov et al., 2019; Shafin et al., 2021) All downstream subsequent analysis were done using haplotype 1, as determined by HapDup. Repetitive content was identified and soft masked from the genome using RED v.05/22/2015 (Girgis, 2015). In order to place our contigs in a Chromosome level framework we scaffolded the assembly against a chromosome level from the Darwin Tree of Life (DTOL) project using RagTag v2.0.1 (Alonge et al., 2021).

2.6 Genome annotation

Braker2 automated annotation pipeline was used to generate a comprehensive annotation of protein coding genes in the final assembly. We first ran Braker2 in the genome and protein mode, using reference proteins from the Arthropoda section of OrthoDB v.10 (Brůna et al., 2021, 2020; Buchfink et al., 2015; Gotoh, 2008; Hoff et al., 2016, 2019; Iwata and Gotoh, 2012; Lomsadze et al., 2005; Stanke et al., 2006, 2008) Filtering of genome annotation to the longest isoform used scripts from the AGAT suite of tools v.0.5.1 (Dainat et al., 2022), including agat_convert_sp_gxf2gxf.pl, agat_sp_keep_longest_isoform.pl, and agat_sp_extract_sequences.pl. The resulting annotation was assessed based upon number of complete genes and BUSCO scores, for both all proteins and longest isoforms per locus. We assigned gene names and function to our predicted genes using eggNOG-mapper v.2 (Cantalapiedra et al., 2021; Huerta-Cepas et al., 2019).

2.7 Short-read mapping and Variant calling

After trimming the paired reads from the pools for low quality bases and adapter sequences we aligned the reads to the new *P. napi* adalvinda reference genome using bwa-mem2 v2.2.1 (Vasimuddin et al., 2019). Unaligned reads were filtered out using Samtools v1.11 (Li et al., 2009). Samtools was additionally used to generate mpileup files for the melanic and non-melanic pool combinations from each family in the BSA cross, for the population comparisons between Abisko and the other Swedish populations and for each individual pool. Mpileup files were converted to sync files using mpileup2sync.jar from Popoolation2 keeping bases with a quality score higher than 20 (Kofler et al., 2011). Additionally, Indel regions were identified and masked using the identify-indel-regions.pl and filter-sync-by-gtf.pl scripts also from the Popoolation2 package.

2.8 Identifying the adalvinda contig

We identified regions of divergence between the dark and light morphs from each of our male- and female-informative crosses using two alternative approaches, 1) identifying regions where the three crosses shared a signal of elevated F_{ST}, and 2) using BayPass to identify genetic markers that are associated with melanism while simultaneously accounting for underlying relationship among the crosses and pools.

In *Pieris napi,* like other Lepidoptera, females, the heterogametic sex, produce gametes without recombination. As such, in our female informative we expect the melanic individuals to inherit the complete chromosome harboring the adalvinda locus, whereas in the male-informative cross, recombination will lead to only part of this region to be inherited. We therefore expect to see a much narrower locus of elevated F_{ST} in the case of the male-informative cross somewhere in the chromosome identified by the female-informative cross.

2.9 Long read alignment

HiFi-PacBio long read sequence data generated by the DTOL project (ERR6594498, ERR6594499) was aligned against the *P. n. adalvinda* reference genome using pbmm2 v 1.7. ONT long reads used for the genome assembly were aligned using mimimap2 v 2.24 (with the -ax map-ont setting) (Li, 2018).

2.10 BSA F_{ST}

We used F_{ST} to identify regions of divergence between melanic and non-melanic females of our F2 crosses. We generated paired mpileup files for each Family using Samtools mpileup (filtering sites reads and alignments for a mapQ and phred score of >20. Each mpileup was then further filtered for indels using the identify-genomic-indel-regions.pl and filter-pileup-by-gtf.pl respectively. The indel filtered mpileup files were subsequently converted into sync files using the mpileup2sync.jar script also included in popoolation2. We finally calculated F_{ST} for every SNP included in the sync file, as well as in non-overlapping windows of 10Kbp using the parallel_fstsliding.sh script including all sites with a coverage between 30 and 500 reads.

2.11 F_{ST} of population samples

In order to further identify the adalvinda locus, and to identify further signatures of local adaptation between the Abisko population and other populations of *P. napi* in Sweden and Europe, we calculated F_{ST} in non-overlapping windows of 100 and 10000bp between pooled sequence data from Abisko with pooled sequence data from Skåne. Mpileup files for each pair were generated using the same methods as for the BSA F_{ST} analysis. For the downstream analysis only windows containing more than 3 SNPs, read depth between 20 and 150 and at least 50% of the window covered with reads. Outlier windows were calculated as those belonging to the top 99$^{\text{th}}$ percentile (sex chromosomes and autosomes were each calculated independently of each other).

2.12 BayPass

BayPass (v2.3) uses a Bayesian approach to identify loci whose allele frequencies correlate with a phenotypic trait across populations, while controlling for the underlying population covariance structure. The input for this was the same Mpileup and sync files used by popoolation2 for the F_{ST} scan. To mitigate the effect of linkage disequilibrium we subset and thinned the SNP data into 10 groups, running each batch independently and later merging them while running the core model. To identify discrete peaks and stretches of elevated BF we calculated the average BF in sliding windows of 20 SNPs across Chromosome 17.

3 Results:

3.1 Heritability

In *P. bryoniae*, the dark female morph is known to be associated with a single dominant autosomal locus (Lorkovic, 1962), we therefor hypothesized that the same would be true for this highly similar and closely related morph. To test this, we generated one female and seven male-informative crosses involving hybrid individuals between *P. napi adalvinda* from Abisko and *P. napi napi* from Spain that were back crossed to *P. napi napi.* In the resulting F2 offspring we expected the melanic and non-melanic phenotypes to segregate at ˜50% frequency. In the female informative cross (family 206), there was a clear bimodal distribution of melanin levels. Since all females would get their W from their hybrid mother, and their Z from their *P. n. napi* father, the dark morph must have an autosomal locus controlling it. In the male informative crosses, only four of the seven families analyzed (21, 23, 53, and 206) showed a clear bimodal distribution in their melanin levels, four families (10, 12, 31, and 39) did not. Family 10 had too few individuals (n=14) for us to examine the distribution of the melanic phenotype among its offspring. As males do not express the dark phenotype, it was expected that we would be missing the adalvinda allele in some of the male informative baccrosses by chance. However, based upon the crosses that resulted in melanized female offspring, these results are consistent with the dark *P. n. adalvinda* morph being caused by a single dominant autosomal locus, similar to that of *P. bryoniae*. We hereafter refer to the dark morph allele either being the dominant adalvinda allele, or the light colored napi allele, at the melanic locus.

3.2 Genome assembly

Using DNA from a single individual, 17.8 Gb of long read data, with an N50 of 58768, was generated. Using Flye v. 2.9 we assembled a contiguous, but highly duplicated genome (463 contigs, N50=6.7, assembly size 614Mb, Busco = C:99.2% [S:12.3%, D:86.9%], F:0.4%, M:0.4%, n:5286). The inflated genome size, in combination with the large amount of duplication in the BUSCO indicates that few haplotypes were collapsed, and that each haplotype is represented in the assembly. The assembly was reduced to a single haplocopy using purge_dups v1.2.5, resulting in an assembly consisting of 154 contigs, N50 = 7.3Mb, and an assembly size 313.4Mb, which is in line with other available *P. napi* assemblies. HapDup v0.7 was used to resolve potential haplotype switch errors, assembly errors, as well as polish the assembly, leaving us with a final assembly consisting of 313.421Mb across 150 contigs with an N50 of 7.6. Genome completeness was assessed using BUSCO, revealing that the final assembly contained 98.6% complete BUSCO genes (S:97.8%, D:0.8%, F:0.7%, M:0.7%) out of 5286 genes in the Lepidopteran_OD10 database. Finally, using Ragtag we were able to place 135 of the 150 contigs into a chromosome level framework, based upon the chromosome level *Pieris napi* genome from the DTOL project, leaving 15 contigs and 595491 bp as unplaced contigs.

3.3 Genome annotation

Our annotation using Braker2 resulted in 16410 genes and 17837 transcripts, which is in line with previously annotated lepidopteran genomes. BUSCO completeness assessment of the genome annotation revealed that we annotated 97.1% of BUSCO genes (C:97.1% [S:88.2%, D:8.9%], F:1.4%, M:1.5%, n:5286 BUSCO genes from Lepidopteran_OD10 database). Functional annotation of the assembly was performed using EggNOG mapper (v2.1.7) comparing it against the EggNOG database v5 and integrated into the annotation GFF, a total of 16203 genes were given a functional annotation.

3.4 Genetic basis of the adalvinda morph

To identify the genetic basis of the adalvinda allele, we performed a series of bulk segregate analyses on the family crosses exhibiting the bimodal distribution of melanic phenotypes. We chose 30, 15, and 30 of the most melanic individuals and 30, 17, and 30 of the least melanic individuals from family 206 (female informative), and families 21 and 23 (male informative), respectively for pooled sequencing. While there is a range of melanic morphs within the bimodal distributions of these families, we expect these to be due to alleles with minor effects or plasticity. However, by selecting the tails of the distribution, we expected these individuals to only consistently different for the adalvinda and napi alleles.

3.5 F_{ST} estimates of BSA crosses

Using the fst-sliding.pl script in the popoolation2 suite, F_{ST} was estimated between the melanic and non-melanic individuals of each family cross both at individual SNP level, and in non-overlapping windows of 10kb. The female-informative cross identified Chromosome 17 as being associated with female wing melanization (Fig. 1A), however due to the lack of recombination in females, we were unable to narrow down the locus further. The two male-informative crosses also pointed to Chromosome 17 and indicated that the locus is located between 8 and 12 Mb on contig_71_1 (the contig making up most of chromosome 17).

3.6 BayPass

To further identify genomic regions associated with female wing melanization, while also accounting for the underlying demographic relationships among the male- and female-informative crosses, we applied a genotype–phenotype association approach using BayPass (Gautier, 2015). While the underlying data for this approach is the same as that used in the F_{ST} analysis, it is able to integrate across all our families and reduce the background noise caused by lack of recombination. BayPass indicated that by a distinct concentration in elevated BF score upstream of the *Cortex* gene on Chromosome 17 (Fig. 2). To identify localized spikes in BF score against the background noise, we calculated a 20 SNP rolling average of BF-score. This narrowed down an 18kb (9530683-9548749) region 62kb upstream of *Cortex* (pos:9593536) on contig_71_1 (Chr 17) that was strongly associated (BF>20) with wing melanization.

3.7 F_{ST} – Genome wide

To identify regions of divergence between *P. n. adalvinda* and *P. n. napi,* and to narrow down our candidate region controlling female melanization, we estimated genome wide F_{ST} between *P. n. adalvinda* from Abisko, with population sequence data of *P. napi* collected in Skåne. F_{ST} was calculated both on a SNP level, as well as in windows of 100bp and 10kbp. On a genome wide level using 10kb windows we observed elevated F_{ST} across all chromosomes with distinct spikes on the Z chromosome, Chr 3, 5 and 24 but nothing near our candidate locus (Fig. 3a). However, these populations are very distant, and experience large differences in ecology, phenology, and selection in addition to wing color, making it unlikely that we would be able to pick up this locus using this analysis alone. However, when we use single SNP level F_{ST} and focus on the 8-12 Mb window on contig_71_1 (Chr. 17) we can see a clear spike in F_{ST} at the same locus indicated by the BayPass analysis.

3.8 Adalvinda locus characterization and annotation

Due to substantial amounts of repetitive content, it was not possible to clearly delineate the exact boundaries of the adalvinda locus using pooled short-read data alone. Instead, we used PacBio-hifi data generated by the Darwin tree of life project to delineate the borders. These alignments revealed a single 15kb large region of missing content in the UK population that overlaps with the locus identified with by the BayPass analysis. Within this ˜15kb window we see alternating spikes and gaps in coverage from all our non Abisko populations, indicating it being composed both of common repetitive content and unique adalvinda content (Fig.

4 Discussion:

Here we show that adalvinda, the female-limited regionally isolated morph of *P. napi*, had a single, is caused by a single autosomal dominant allele. This is reminiscent of the similar phenotype found in *P. bryoniae* (Lorkovic, 1962). Using male and female informative crosses we identified a 20 kb large locus at which the allele for adalvinda is located, which is 50 kb upstream from the gene cortex. While a population genomic comparison between individuals from adalvinda and P. n. napi populations identified a significant F_{ST} outlier in the locus region, this locus was not among the most predominant F_{ST} outlines in analysis. These findings add to the growing number of cases reporting a role for *cortex* in Lepidopteran wing color evolution, supporting earlier claims of it being a hot spot locus for evolution.

Our findings fit well with what is seen in other genotype to phenotype studies, highlighting the importance of cis-regulatory changes and co-option, over changes in the amino acid sequence of genes for evolution of novel traits. Additionally, our addition to the cortex literature adds to a growing divergence in the literature regarding melanic phenotypes among insects. While in Lepidoptera, a role for *cortex* is common, genes in the melanin biosynthesis pathways are notable absent. In contrast, studies of melanic morphs among species of Drosophila repeatedly identify genes in the biosynthesis pathway (Prud'homme et al., 2006; Signor et al., 2016; Yassin et al., 2016). Due to cortex not being involved in regulating coloration in *Drosophila* and other model systems, the molecular mechanism by which it is doing this in Lepidoptera remains unclear. However, it appears that cortex has a critical role in scale color identity of Lepidoptera (Nadeau et al., 2016).

While adalvinda has been suggested to be a thermal adaptation to the cold and unstable climate in northern Scandinavia (Petersen, 1949), this has yet to be empirically tested. Numerous other studies of butterfly wing color document a clear role for wing melanization in thermoregulation, where greater absorption of solar radiation enable them to reach optimal flight temperature faster (Kingsolver, 1983; Watt, 1968). Certainly, considering the short season and variable climate in northern Scandinavia, this could be beneficial for the females. However, darker wings may also result in overheating in warm climates, which could potentially also explain the ansence of Adalvinda further south. Future field and laboratory studies of the thermal and behavioral impacts of adalvinda are needed to determine the fitness effects. However, even if there is a thermal ecology aspect, the sex-limited nature of adalvinda remains curious.

It is also not clear as to why the phenotype is limited to females. Adalvindas female-limited nature could either be due to divergent selection on males and females, either through natural or sexual selection, or due to the genetic architecture of the traits, where it through chance happened to evolve through the insertion of a sex-limited enhancer, similar to what is seen in *Colias* butterflies (Tunström et al., 2023). Among butterflies, wing color polymorphisms tend to either be found in both sexes, or limited to females (Vane-Wright, 1975). It could be that the *adalvinda* locus contains a female-limited modular enhancer similar to the *Alba* locus, and that it co-opts cortex and its downstream pathways, causing melanization. While more genome wide studies of similar polymorphisms and dimorphisms are needed, we hypothesize that these female-specific modular enhancers are abundant among lepidoptera, and likely to be found near loci generating dimorphic traits.

5 Acknowledgements

The authors would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure, NGI, and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure. Funding was provided by the Swedish Research Council, and the Knut and Alice Wallenberg Foundation.

References

Alonge, M., Lebeigle, L., Kirsche, M., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., and Soyk, S. (2021). Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing.

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., and Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics and Bioinformatics*, 3(1).

Brůna, T., Lomsadze, A., and Borodovsky, M. (2020). GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, 2(2):lqaa026.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12):5825–5829.

Dainat, J., Hereñú, D., LucileSol, and pascal-git (2022). NBISweden/AGAT: AGAT-v0.8.1. Zenodo.

Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., Nadeau, N. J., Zimin, A. V., Hughes, D. S. T., Ferguson, L. C., Martin, S. H., Salazar, C., Lewis, J. J., Adler, S., Ahn, S.-J., Baker, D. A., Baxter, S. W., Chamberlain, N. L., Chauhan, R., Counterman, B. A., Dalmay, T., Gilbert, L. E., Gordon, K., Heckel, D. G., Hines, H. M., Hoff, K. J., Holland, P. W. H., Jacquin-Joly, E., Jiggins, F. M., Jones, R. T., Kapan, D. D., Kersey, P., Lamas, G., Lawson, D., Mapleson, D., Maroja, L. S., Martin, A., Moxon, S., Palmer, W. J., Papa, R., Papanicolaou, A., Pauchet, Y., Ray, D. A., Rosser, N., Salzberg, S. L., Supple, M. A., Surridge, A., Tenger-Trolander, A., Vogel, H., Wilkinson, P. A., Wilson, D., Yorke, J. A., Yuan, F., Balmuth, A. L., Eland, C., Gharbi, K., Thomson, M., Gibbs, R. A., Han, Y., Jayaseelan, J. C., Kovar, C., Mathew, T., Muzny, D. M., Ongeri, F., Pu, L.-L., Qu, J., Thornton, R. L., Worley, K. C., Wu, Y.-Q., Linares, M., Blaxter, M. L., ffrench-Constant, R. H., Joron, M., Kronforst, M. R., Mullen, S. P., Reed, R. D., Scherer, S. E., Richards, S., Mallet, J., Owen McMillan, W., Jiggins, C. D., and The Heliconius Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405):94–98.

Deshmukh, R., Baral, S., Gandhimathi, A., Kuwalekar, M., and Kunte, K. (2018). Mimicry in butterflies: Co-option and a bag of magnificent developmental genetic tricks. *WIREs Developmental Biology*, 7(1):e291.

Ellers, J. and Boggs, C. L. (2003). The Evolution of Wing Color: Male Mate Choice Opposes Adaptive Wing Color Divergence in Colias Butterflies. *Evolution*, 57(5):1100–1106.

Girgis, H. Z. (2015). Red: An intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics*, 16(1):227.

Gotoh, O. (2008). A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Research*, 36(8):2630–2638.

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., and Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics (Oxford, England)*, 36(9):2896–2898.

Hanly, J. J., Livraghi, L., Heryanto, C., McMillan, W. O., Jiggins, C. D., Gilbert, L. E., and Martin, A. (2022). A large deletion at the cortex locus eliminates butterfly wing patterning. *G3 Genes—Genomes—Genetics*, 12(4):jkac021.

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, 32(5):767–769.

Hoff, K. J., Lomsadze, A., Borodovsky, M., and Stanke, M. (2019). Whole-Genome Annotation with BRAKER. *Methods in Molecular Biology (Clifton, N.J.)*, 1962:65–95.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., and Bork, P. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314.

Iwata, H. and Gotoh, O. (2012). Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Research*, 40(20):e161.

Jeong, S., Rebeiz, M., Andolfatto, P., Werner, T., True, J., and Carroll, S. B. (2008). The Evolution of Gene Regulation Underlies a Morphological Difference between Two Drosophila Sister Species. *Cell*, 132(5):783–793.

Kettlewell, HBD. (1973). The evolution of melanism. Clarendon.

Kingsolver, J. G. (1983). Thermoregulation and Flight in Colias Butterflies: Elevational Patterns and Mechanistic Limitations. *Ecology*, 64(3):534–545.

Kingsolver, J. G. and Wiernasz, D. C. (1991). Seasonal Polyphenism in Wing-Melanin Pattern and Thermoregulatory Adaptation in Pieris Butterflies. *The American Naturalist*, 137(6):816–830.

Koch, P. B., Behnecke, B., and ffrench-Constant, R. H. (2000). The molecular basis of melanism and mimicry in a swallowtail butterfly. *Current Biology*, 10(10):591–594.

Kofler, R., Pandey, R. V., and Schlötterer, C. (2011). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, 27(24):3435–3436.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5):540–546.

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, 33(20):6494–6506.

Lorkovic, z. (1962). The Genetics and Reproductive Isolating Mechanisms of the Piers napi - bryoniae group. *Journal of the Lepidoterists' Society*, 16(1):5–19.

Martin, S. H., Singh, K. S., Gordon, I. J., Omufwoko, K. S., Collins, S., Warren, I. A., Munby, H., Brattström, O., Traut, W., Martins, D. J., Smith, D. A. S., Jiggins, C. D., Bass, C., and ffrench-Constant, R. H. (2020). Whole-chromosome hitchhiking driven by a male-killing endosymbiont. *PLOS Biology*, 18(2):e3000610.

Matsuoka, Y. and Monteiro, A. (2017). Melanin pathway genes regulate color and morphology of butterfly wing scales.

Monteiro, A. and Podlaha, O. (2009). Wings, Horns, and Butterfly Eyespots: How Do Complex Traits Evolve? *PLOS Biology*, 7(2):e1000037.

Nadeau, N. J. (2016). Genes controlling mimetic colour pattern variation in butterflies. *Current Opinion in Insect Science*, 17:24–31.

Nadeau, N. J., Pardo-Diaz, C., Whibley, A., Supple, M. A., Saenko, S. V., Wallbank, R. W. R., Wu, G. C., Maroja, L., Ferguson, L., Hanly, J. J., Hines, H., Salazar, C., Merrill, R. M., Dowling, A. J., ffrench-Constant, R. H., Llaurens, V., Joron, M., McMillan, W. O., and Jiggins, C. D. (2016). The gene cortex controls mimicry and crypsis in butterflies and moths. *Nature*, 534(7605):106–110.

Petersen, B. (1949). On the Evolution of Pieris napi L. *Evolution*, 3(4):269–278.

Prud'homme, B., Gompel, N., Rokas, A., Kassner, V. A., Williams, T. M., Yeh, S.-D., True, J. R., and Carroll, S. B. (2006). Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, 440(7087):1050–1053.

Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, 9(7):671–675.

Shafin, K., Pesout, T., Chang, P.-C., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Kolmogorov, M., Eizenga, J. M., Miga, K. H., Carnevali, P., Jain, M., Carroll, A., and Paten, B. (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods*, 18(11):1322–1332.

Signor, S. A., Liu, Y., Rebeiz, M., and Kopp, A. (2016). Genetic Convergence in the Evolution of Male-Specific Color Patterns in Drosophila. *Current Biology*, 26(18):2423–2433.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24(5):637–644.

Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, 7(1):62.

Tunström, K., Woronik, A., Hanly, J. J., Rastas, P., Chichvarkhin, A., Warren, A. D., Kawahara, A. Y., Schoville, S. D., Ficarrotta, V., Porter, A. H., Watt, W. B., Martin, A., and Wheat, C. W. (2023). Evidence for a single, ancient origin of a genus-wide alternative life history strategy. *Science Advances*, 9(12):eabq3713.

Vane-Wright, R. and Treadaway, C. (2011). Female-limited polymorphism and its significance in *Appias* (*Catophaga*) *nero corazonae* Schröder and Treadaway, 1989 (Lepidoptera: Pieridae). *Journal of Natural History*, 45(37-38):2355–2362.

Vane-Wright, R. I. (1975). An integrated classification for polymorphism and sexual dimorphism in butterflies. *Journal of Zoology*, 177(3):329–337.

van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C., and Saccheri, I. J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, 534(7605):102–105.

van't Hof, A. E., Reynolds, L. A., Yung, C. J., Cook, L. M., and Saccheri, I. J. (2019). Genetic convergence of industrial melanism in three geometrid moths. *Biology Letters*, 15(10):20190582.

van't Hof, A. E. and Saccheri, I. J. (2010). Industrial Melanism in the Peppered Moth Is Not Associated with Genetic Variation in Canonical Melanisation Gene Candidates. *PLOS ONE*, 5(5):e10889.

Vasimuddin, Md., Misra, S., Li, H., and Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 314–324.

Watt, W. B. (1968). Adaptive Significance of Pigment Polymorphisms in Colias Butterflies. I. Variation of Melanin Pigment in Relation to Thermoregulation. *Evolution*, 22(3):437–458.

Wittkopp, P. J. and Beldade, P. (2009). Development and evolution of insect pigmentation: Genetic mechanisms and the potential consequences of pleiotropy. *Seminars in Cell & Developmental Biology*, 20(1):65–71.

Yassin, A., Bastide, H., Chung, H., Veuille, M., David, J. R., and Pool, J. E. (2016). Ancient balancing selection at tan underlies female colour dimorphism in Drosophila erecta. *Nature Communications*, 7(1):10400.

Zhang, L., Martin, A., Perry, M. W., van der Burg, K. R. L., Matsuoka, Y., Monteiro, A., and Reed, R. D. (2017). Genetic Basis of Melanin Pigmentation in Butterfly Wings. *Genetics*, 205(4):1537–1550.

6 Figures

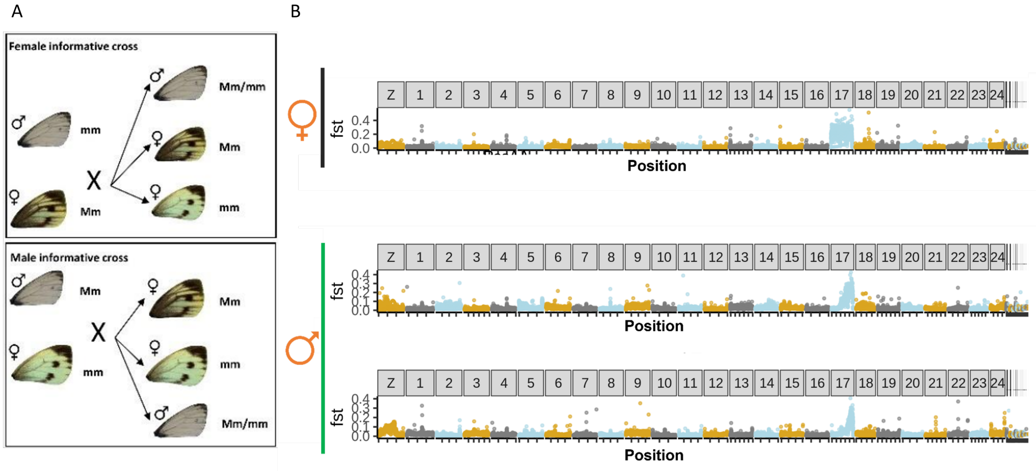


Figure 1. Bulk seggregant analysis of male and female informative crosses to identify the adalvinda locus. Males and females from Abisko were crossed with *P. napi* individuals from Spain, resulting in heterozygote offspring. from Abisko and Spain were crossed using male and female informative crosses (**A**) between individuals from Abisko and Spain to produce F1 crosses. These hybrid individuals were subsequently backcrossed with pure Spanish individuals in order to isolate the adalvinda allele in a Spanish background. The darkest and lightest females in each cross were selected for sequencing (Sex information of the cross is indicated with male or female symbol). Fst scan between melanic and non-melanic offspring in female- and male-informative crosses between Abisko and Spain (**B**). Elevated Fst indicates chromosomal location of the locus controlling female wing melanization. As female butterflies do not recombine the region of increased Fst covers the entire chromosome 17 in the female informative cross.

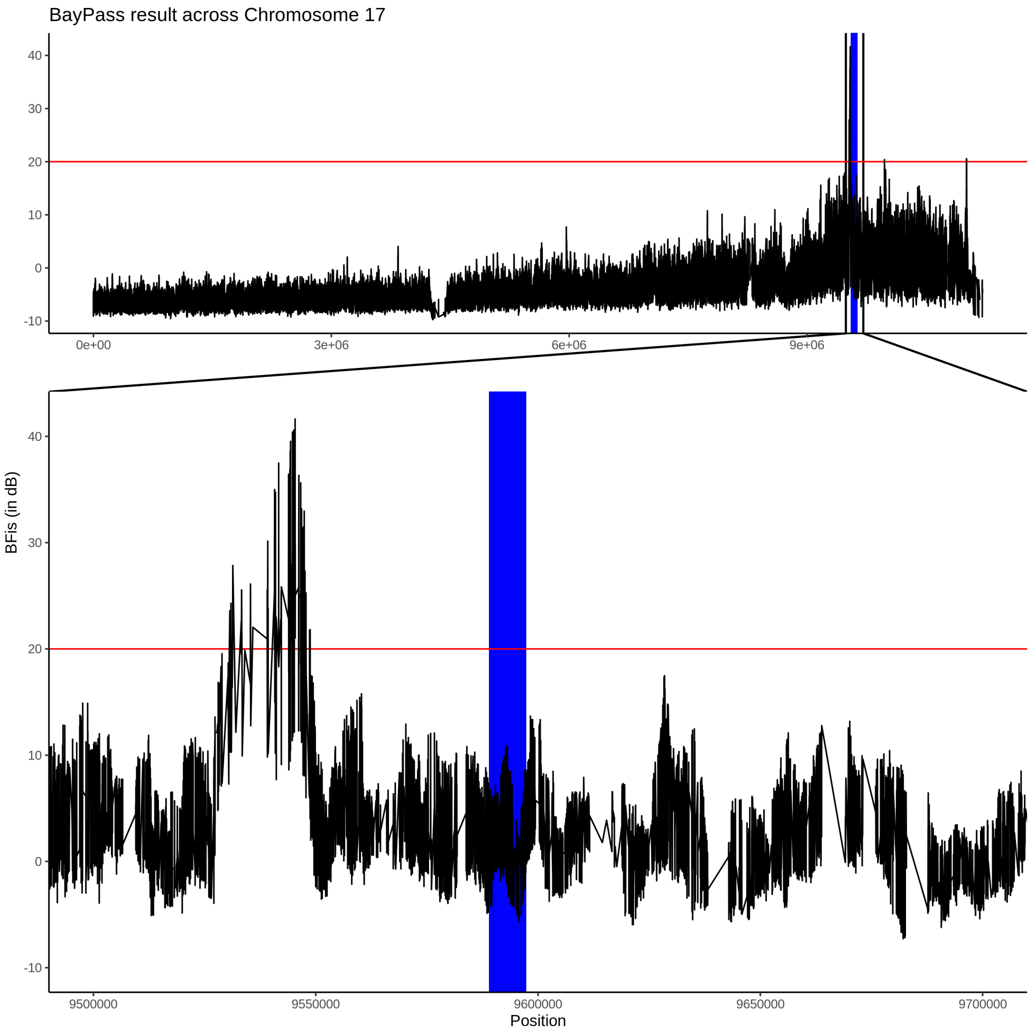


Figure 2. BayPass results. Bayes factor (BF) plotted against as a rolling mean of 20 SNPs across a) Chromosome 17, and b) 200Kb surrounding the gene *Cortex* (highlighted in blue). The horizontal red line represents a BF score threshold of 20.

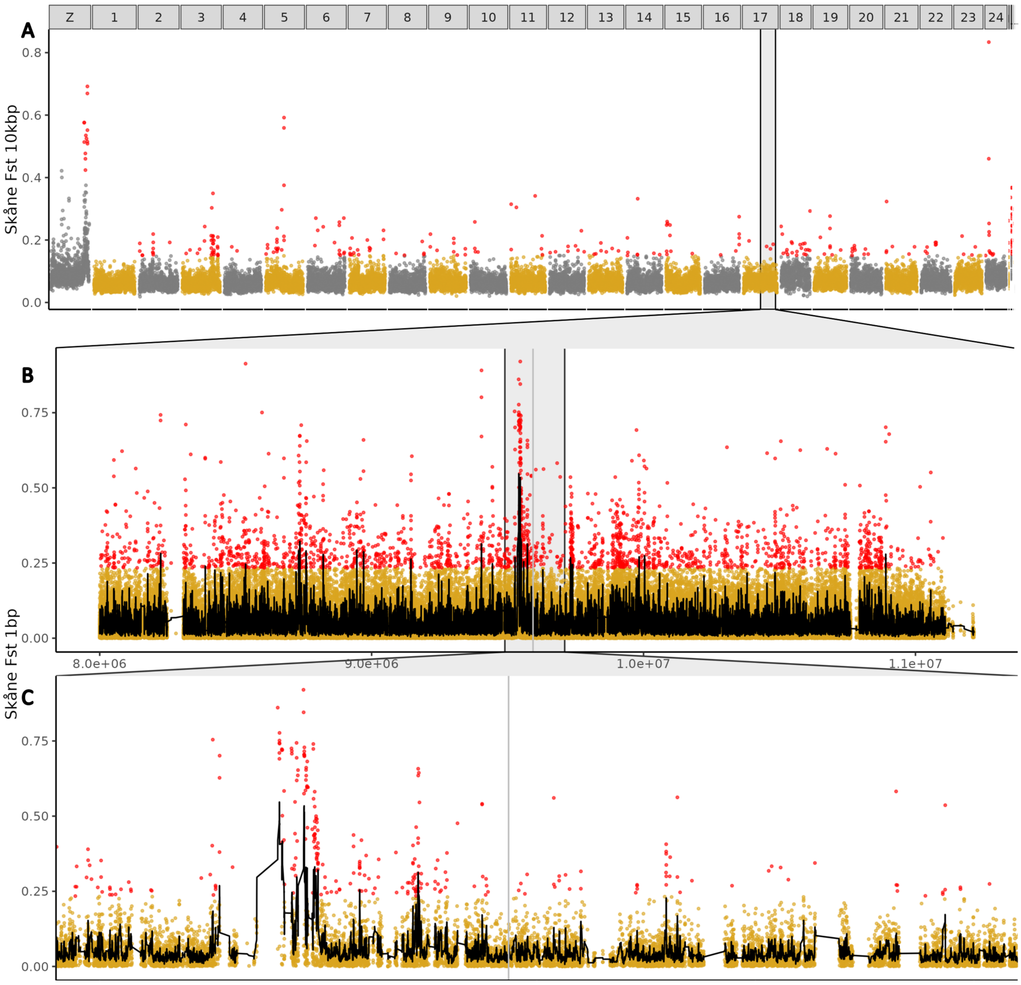


Figure 3. F_{ST} **comparison between Abisko and Skåne at three different scales of analysis. A**) Top row 10kb windows genome wide, with chromosomes colored in alternating colors and outlier windows highlighted in red. **B**) F_{ST} at single SNP level across the 4Mb region of Chr 17 identified by the BayPass analysis., and in **C**) an inset focusing at the adalvinda locus and the location of the cortex gene (grey vertical bar). The black line represents average F_{ST} in 20 SNP sliding windows.

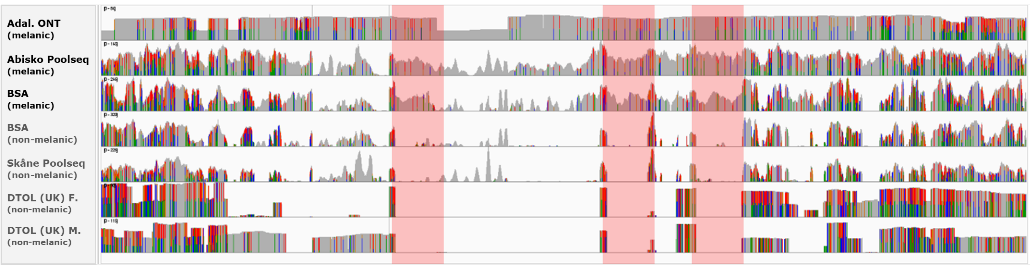
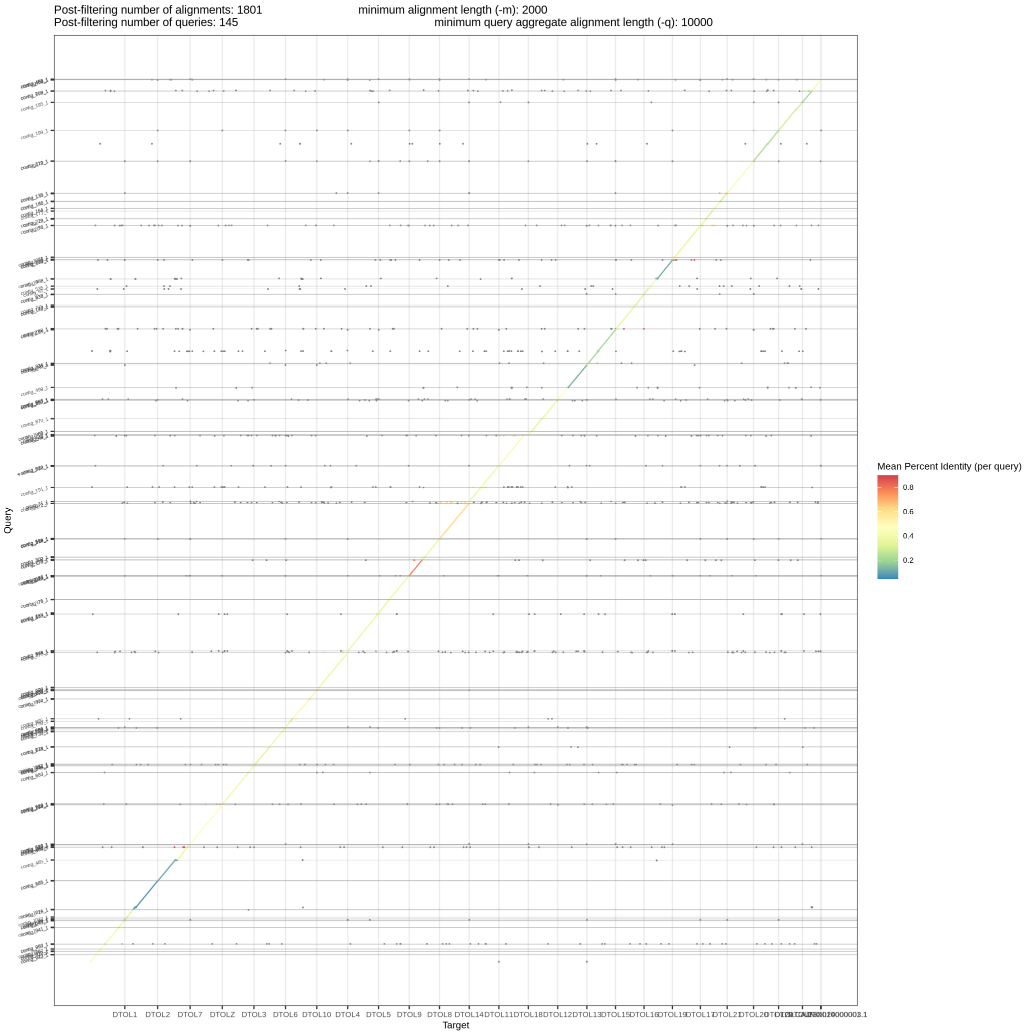


Figure 4. Illustration of the presence-absence of genomic content in the adalvinda candidate locus to highlight the likely boundaries. Read depth comparison across the adalvinda candidate region illustrating presence of coverage in the top three rows (adalvinda ONT, Abisko Pool, and BSA_f21_dark) and lack of coverage in the remaining datasets (BSA_f21_light, Skåne Pool and two HiFi-datasets form the UK. While high levels of repetitive and low complexity sequence in the region makes it hard to identify the region in the Pooled NGS datasets as seen by small spikes in coverage that are absent in the long read datasets, but also hard to identify unique content. However, some regions (highlighted in pink) exhibit consistently high and even coverage in all melanized samples and low to no coverage the other populations. All data has been filtered for reads with a mapQ score of 20.

7 Supplementary Figures



Supplementary figure 1. Synteny assessment of final *P. napi adalvinda* assembly with DTOL *P. napi* genome assembly revealing high synteny and a substantial proportion of almost chromosome length contigs.