Local selection signals in the genome of blue tits emphasize regulatory and neuronal evolution

Jakob C. Mueller¹ | Esteban Botero-Delgadillo¹ | Pamela Espíndola-Hernández¹ | Carol Gilsenan¹ | Phil Ewels² | Joel Gruselius³ | Bart Kempenaers¹

¹Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, Seewiesen, Germany
²Science for Life Laboratory (SciLifeLab), Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden
³Science for Life Laboratory, Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden

Correspondence
Jakob C. Mueller, Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, Seewiesen, Germany. Email: mueller@orn.mpg.de

Present address
Joel Gruselius, Vanadis Diagnostics, PerkinElmer, Sollentuna, Sweden

Funding information
This work was supported by the Max Planck Society (to B.K.). The authors acknowledge support from the National Genomics Infrastructure in Genomics Application Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure.

Abstract
Understanding the genomic landscape of adaptation is central to understanding microevolution in wild populations. Genomic targets of selection and the underlying genomic mechanisms of adaptation can be elucidated by genome-wide scans for past selective sweeps or by scans for direct fitness associations. We sequenced and assembled 150 haplotypes of 75 blue tits (Cyanistes caeruleus) of a single Central European population by a linked-read technology. We used these genome data in combination with coalescent simulations (i) to estimate an historical effective population size of ~250,000, which recently declined to ~10,000, and (ii) to identify genome-wide distributed selective sweeps of beneficial variants probably originating from standing genetic variation (soft sweeps). The genes linked to these soft sweeps, but also those linked to hard sweeps based on new beneficial mutants, showed a significant enrichment for functions associated with gene expression and transcription regulation. This emphasizes the importance of regulatory evolution in the population’s adaptive history. Soft sweeps were further enriched for genes related to axon and synapse development, indicating the significance of neuronal connectivity changes in the brain potentially linked to behavioural adaptations. A previous scan of heterozygosity–fitness correlations revealed a consistent negative effect on arrival date at the breeding site for a single microsatellite in the MDGA2 gene. Here, we used the haplotype structure around this microsatellite to explain the effect as a local and direct outbreeding effect of a gene involved in synapse development.

KEYWORDS
3’UTR, alternative splicing, gene expression regulation, GO overrepresentation, HFC, neuron projection, RNA processing, synapse

1 | INTRODUCTION

Understanding the adaptive history of and the genomic mechanisms of selection in wild populations is at the core of evolutionary biology. Genome-wide scans for selection have become increasingly common, and a majority of the published studies report on comparisons between populations or species (Haasl & Payseur, 2016). Such comparisons have a high power to detect the genetic basis of the difference between the populations or species, but they do not allow assessment of the general background of the adaptive history
within a population, which includes both global and local adaptation. Arguably, the direct selection history of a population can better predict the population’s future evolutionary trajectory than a comparative history.

Large lists of specific protein-coding genes known to be affected by local selection have been compiled (e.g., for humans, Rees et al., 2020). Since the early 1960s, however, scientists have argued that much of biological diversity may have resulted from changes in gene expression rather than from changes at the protein sequence level (reviewed in Stern, 2016). Among the first evidence in favour of this idea was the observed weak correlation between the rate of protein sequence evolution and the rate of phenotypic evolution among different clades (Wilson et al., 1974). A multitude of studies have subsequently shown the importance of regulatory evolution by linking specific phenotypic trait variation to changes in specific gene expression networks (Stern & Orgogozo, 2008). Evidence has further accumulated showing that regulatory sequences that flank the protein-coding gene of interest (cis-acting), as well as specific regulatory proteins encoded elsewhere in the genome (trans-acting), control when expression of the gene of interest occurs, at what level it is expressed, under what environmental conditions, and in which cells or tissues (Wray et al., 2003). Most of these studies considered variation across species, while only a few investigated genomic mechanisms of adaptation within species (e.g., among ecotypes, Jones et al., 2012).

Here, we analysed historical and current selection signals in a large and stable natural bird population without a history of known specific environmental changes such as those experienced during island or urban colonizations (Johnson & Munshi-South, 2017). Nevertheless, the population probably experienced unknown adaptive events during its long history, presumably resulting from abiotic or biotic environmental changes (e.g., human-induced habitat change, climate change, disease epidemics). We therefore expect to detect signals of selection at genes involved in regulation, but also at genes related to other known hotspots of selection, such as immune system genes (Shultz & Sackton, 2019) or genes related to behavioural changes (Mueller et al., 2020). The latter is expected, because changes in behavioural traits may expose individuals to new environments and, vice versa, new environments may promote changes in behaviour (Zuk et al., 2014). Among the probable candidates for adaptive evolution of behaviour are genes that regulate neuronal connectivity in the brain (Mueller et al., 2020).

We scanned the genome for signatures of selection in a wild population of blue tits (Cyanistes caeruleus) breeding in a natural, oak-dominated forest in Central Europe. The blue tit has long been a model species for microevolutionary studies, and direct measures of fitness in terms of reproduction and survival are available from long-term field studies (Culina et al., 2020). In previous studies we tested for general and local inbreeding and outbreeding effects using a genome-wide set of 83–106 microsatellite loci (Botero-Delgadillo et al., 2020; Olano-Marín et al., 2010; Olano-Marín et al., 2011a, 2011b). In the population considered here, a single microsatellite locus (H095_TG05_053) showed a consistent negative correlation between individual heterozygosity and arrival date (Botero-Delgadillo et al., 2020). Even though this blue tit population is nonmigratory, individuals differ consistently in the date of yearly arrival at the breeding site (Gilsenan et al., 2020). Moreover, arrival date can be considered a key fitness-related trait, because it correlates with yearly reproductive success (Gilsenan et al., 2020). Here, we characterized this potential negative outbreeding effect using sequence data in addition to the de novo scan for selection signatures.

For this study, we analysed 150 haplotypes from 75 blue tits sampled in a single population. The haplotypes were sequenced and assembled using a linked-read technology and the sequence data were mapped against the blue tit reference (Mueller et al., 2016). We used these data to address the following questions: (i) What is the current and historical effective population size of the Central European blue tit population? The resultant demographic model will be used as the null model for the following selection tests. (ii) How abundant are genomic signatures of selective sweeps in the blue tit population? The number of different types of selective sweeps (hard and soft sweeps of de novo mutations or standing variation, respectively) will inform about the general background of adaptive events in this wild population. (iii) Are the detected selective sweeps located around genes with specific functions and is there evidence for functional hotspots of selection? (iv) Does the haplotype structure around the microsatellite locus H095_TG05_053 with a known negative heterozygosity–fitness correlation help to explain an outbreeding effect? This will provide information for an alternative mode of selection, in addition to positive selection as evidenced by selective sweeps.

2 METHODS

2.1 Study population and sampling

The forest Westerholz (48°08'26"N, 10°53'29"E) represents a high-quality breeding habitat for blue tits in southern Germany with an oak-dominated natural forest. In 2007, we installed 277 nestboxes and since then we have monitored the breeding biology of 60-170 breeding pairs each year. All birds were caught for measurement and blood sampling, either during the nestling period, during winter roosting in nestboxes or with mist-nets at feeders (Gilsenan et al., 2020).

We selected 75 adult birds (40 males, 35 females) from the breeding population of Westerholz in the years 2015–2017 (Table S1). Of the 75 birds, 35 were chosen randomly from among the 635 adult individuals for which genotype data at 83 microsatellite loci were available (Botero-Delgadillo et al., 2020). 20 were selected as those with the highest homozygosity-by-locus (HL) values and 20 as those with the lowest HL values (based on the 83 microsatellite loci; Botero-Delgadillo et al., 2020). We strived for a balanced sex ratio in each sampling category (Table S1). The 35 randomly selected birds were used to infer demography and selective sweeps, whereas all
75 birds were used to characterize the haplotype structure at the microsatellite with known heterozygosity–fitness effect.

### 2.2 Sequencing, reference mapping, SNP calling and phasing

Blood samples were stored in Queen’s lysis buffer and DNA was extracted with the Blood QuickPure kit (Macherey-Nagel). Sequencing libraries for each bird were individually prepared using the linked-read Chromium technology of 10x Genomics at the National Genomics Infrastructure (NGI), Sweden. Samples were then individually sequenced on a NovaSeq6000 device with a 2 × 151-bp setup and controlled for quality at the same service.

We used the **LongRanger** pipeline for barcode processing, mapping against the blue tit reference genome (Mueller et al., 2016), variant calling and phasing of each sample (Zheng et al., 2016). The pipeline makes use of the Burrows-Wheeler aligner (BWA; Li, 2013) and **Larar** aligner (Bishara et al., 2015) and the **GATK HaplotypeCaller** (van der Auwera et al., 2013). All samples had a final mean read coverage of 28.4x (range: 21.3–78.2x) and phase blocks with a mean N50 of 3.0 Mb (range: 1.8–3.7 Mb) (Table S1).

We merged all variants that passed the standard quality filter of the **LongRanger** across all samples in a single variant call format file (vcf) using **bcftools** (Daneczeek et al., 2021). This merging produced missing genotype codes at population-wide variants in samples in which these variants were homozygous for the reference allele and thus not called. We replaced these missing codes with a homozygous reference code (0/0) when read depth at the site was ≥6 using **FixVcfMissingGenotypes** (Lindenbaum, 2015). After this procedure, we quality-filtered the data selecting autosomal variants with fewer than seven missing genotype codes, a minor allele count >1 and a read depth smaller than the overall mean plus 5SD. In the final step, we phased the consecutive phase blocks produced by **LongRanger** (characterized by the PS tag in vcf files) within each scaffold using **Shaper4** (Delaneau et al., 2019). The final set of biallelic single nucleotide polymorphisms (SNPs) and indels within scaffolds larger than 110 kb (relevant for selective sweep analysis, see below) comprised 20,461,613 loci.

### 2.3 Demographic inference

We used the software package **popsizeABC** with an approximate Bayesian computation approach to estimate historical effective population sizes (N_e) (Boitard et al., 2016). First, we simulated 102,000 data sets of 70 15-Mb-long haplotypes using the coalescent-based simulator **scrm** (Staab et al., 2015) assuming a mutation rate of 4 × 10^-9 and a recombination rate of 2.5 × 10^-8. These values are estimates from a study on a closely related species, the great tit (Parus major) (Laine et al., 2016; van Oers et al., 2014). Effective population size is allowed to change randomly by a factor of 10 between 100 and ~400,000 individuals at 20 time points between three and 500 generations before the present. For each simulated data set, we then extracted the allele frequency spectrum (AFS) and allele allelic linkage-disequilibrium (LD) values of 20 distance bins ranging from 41.4 kb to 13.5 Mb. We included long-range LD values to estimate recent population size, because theory predicts that LD between markers of more than 10 Mb distance on chromosomes with a recombination rate of 2.5 × 10^-8 is already affected by population size two generations ago (Boitard et al., 2016; Hayes et al., 2003). Second, we extracted the observed AFS and LD values for the same distance bins from biallelic variants of the 23 scaffolds that were longer than 15 Mb, representing 51% of the assembled genome. Third, we compared the simulated summary statistics with the observed ones and identified the most similar simulation with the simple rejection method (acceptance rate 0.0001) using the R package **abc** (Csillery et al., 2012; R Core Team, 2020). We plotted the median N_e over time assuming an average generation time of 2 years (own observations, see also Laine et al., 2016 for the great tit). We tested the robustness of demographic inference using two alternative methods which are insensitive to potential phase-switch errors. The first one is based on the same procedure as described above (**popsizeABC**), but used genotypic LD instead of allelic LD as the summary statistic. The second is a sequentially Markovian coalescent method which includes information of sample frequency spectra to estimate more distant histories. Here, we inferred population size history between 100 and 100,000 years before the present using the program **smc++** (Terhorst et al., 2017). Estimates were performed for each of the 23 scaffolds larger than 15 Mb after excluding all stretches of unknown nucleotides (Ns).

### 2.4 Scanning for selective sweeps

We scanned the genome for selective sweeps in consecutive 10-kb windows using **diplos/hic** in haploid mode (Kern & Schrider, 2018). Two types of sweeps were distinguished: hard sweeps originating from de novo beneficial mutations, and soft sweeps originating from standing variation which becomes beneficial (Messer & Petrov, 2013). To characterize sweep evolution, the supervised machine-learning algorithm employs 12 major summary statistics (normalized values) of selective sweeps including diversity measures, allele frequency spectra, LD patterns and haplotype structure in the focal and 10 neighbouring 10-kb windows (Kern & Schrider, 2018). For the learning step we produced 46,000 genomic simulations of 110 kb (11 10-kb windows) for the neutral, hard sweep, soft sweep and linked sweep scenarios (neighbouring windows to hard or soft sweeps). The coalescent simulations were based on a simplified demographic of the blue tit population based on the analysis of changes in effective population size over time (Figure 1, orange line) using **discoal** (Kern & Schrider, 2016) with selection coefficient(s) between 0.01 and 0.001 and a starting minor allele frequency of 0.05 in the case of soft sweeps. Such weak to moderate selection coefficient(s) may be expected for soft sweeps and are within the estimated range for humans, but are still strong enough to
be effective (i.e. \( N_e > 1 \)) in our population with minimal effective population size of 10,000 (Eyre-Walker & Keightley, 2007; Kern & Schrider, 2018). We then extracted 2000 feature vectors of the 12 summary statistics in each window for each of the neutral and sweep scenarios and trained the model. After this learning step, the model was used to predict the selective sweep state of the 79,334 windows ranked according to their selective sweep probabilities. We are particularly interested in functions associated with higher mean ranks than expected by chance. Thus, we used the Wilcoxon rank-sum test as used for the simulations with selection (orange) to produce a Manhattan plot of hard sweep and soft sweep probabilities based on the model predictions.

2.5 | Functional enrichment analysis

We tested for enrichment of associated gene functions of all windows according to their selective sweep probabilities. We are particularly interested in functions associated with higher mean ranks than expected by chance. Thus, we used the Wilcoxon rank-sum test implemented in the R package \texttt{gofunck} (Grote, 2018) on all the windows ranked according to their sweep probabilities. Windows were annotated to gene ontology (GO) terms via overlapping (or lying within a maximum distance of 5 kb to) genes (whole regions including exons and introns) of our blue tit genome assembly (NCBI blue tit annotation release 100). We used the reduced list of 3336 GO-slim terms defined by \texttt{panther} (Huaiyu et al., 2021) as being informative of function and evolutionarily conserved. Using this method, we annotated 39,915 of the 79,334 windows (50.3%) to GO terms. We corrected for testing multiple GO terms within each of the three main GO families “biological process,” “cellular component” and “molecular function” by evaluating 10,000 simulated data sets with permuted rankings. The family-wise error rate (FWER) was defined as the fraction of random sets where the lowest \( p \)-value across all GO terms was lower than or equal to the raw \( p \)-value of the tested GO term. To reduce redundancy of significant GO sets, we simplified by grouping similar terms based on their semantic similarity according to the Relevance parameter (Schlicker et al., 2006) using the R package \texttt{rrvgo} (Sayols, 2020).

2.6 | Evaluation of a local heterozygosity–fitness effect

As mentioned above, individual heterozygosity at a single microsatellite locus \( \text{H095\_TG05\_053} \) correlated negatively with arrival date at the breeding site (a fitness-related trait) in the study population (Botero-Delgadillo et al., 2020). Following Gilsenan et al. (2020), we defined arrival date for each individual as the earliest record in a given season, starting in August and lasting until the start of the breeding season in March/April. The first record of an individual was detected by an automated radio-frequency identification (RFID)-based system installed at feeders and at all nestboxes in the study area. Individuals had previously been equipped with a passive integrated transponder (PIT-tag). All individuals included in this study were adults (i.e., at least in their second year), because they had been PIT-tagged as yearlings or adults in a previous season.

The \texttt{longRanger} pipeline (see above) identified four indels in the originally amplified region of the microsatellite locus \( \text{H095\_TG05\_053} \) in the 3’ untranslated region (UTR) of the \texttt{MDGA2} gene. Because the indels did not pass the quality filter, we checked and corrected the automatic calling by manually scoring the genotypes through direct inspection of the phased read alignments (phased bam-files). Using the same procedure we further phased the indels in relation to the surrounding already-phased variants (see above).

We assessed LD in the genomic region of interest by calculating pairwise allele and heterozygosity correlations between the neighbouring variants and the indels or direct scores of the microsatellite locus using the R package \texttt{pegas} (Paradis, 2010). We further investigated the genetic relationship among haplotypes in the microsatellite region by calculating simple sequence distances between haplotypes using the R packages \texttt{pegas} and \texttt{haplotypes} (Aktas, 2020). Each different allele of SNPs or indels was scored as one distance unit.

Coulson et al. (1999) suggested that within-individual allele distance reflects the degree of outbreeding better than simple heterozygosity. Thus, we repeated the single-locus analyses described in Botero-Delgadillo et al. (2020) for the larger data set (635 observations from 459 PIT-tagged adult individuals), but using the internal distance between the two alleles instead of heterozygosity as a predictor of arrival date. For each of the 83 microsatellite loci we fitted mixed-effects models with arrival date (standardized within each season) as the response variable and allele distance (absolute value), sex and whether the individual bred in the population in previous years (yes or no) as predictors using the R package \texttt{lmerTest} (Kuznetsova et al., 2017). Individual identity was entered as a random factor.
3 | RESULTS

3.1 | Demographic inference, selection signals and functional enrichment

The historical effective population size of our Central European blue tit population was estimated to be ~250,000 individuals with a declining trend starting about 180 years ago (Figure 1). The estimated current effective population size is ~10,000 individuals. These estimates are based on the summary statistics of fully phased haplotypes (\textit{popsizeABC} in haploid mode). Similar trajectories of demographic dynamics were estimated by two alternative methods which are independent of phasing information (Figures S1 and S2). This indicates an insignificant rate of phase switch errors in our data for haplotypic analyses.

We estimated sweep probabilities for 79,334 10-kb windows in 119 autosomal scaffolds of sufficient size and variant density. Of these, 310 windows showed evidence for hard sweeps and 1775 windows for soft sweeps with prediction probabilities >.8. The remaining windows were predicted to be neutral or linked to sweeps, or the evidence was inconclusive (low probabilities for all scenarios). The hard and soft sweep windows are distributed across the whole genome (Figure 2). There is also no clustering of hard or soft sweeps as shown by the distributions of intrachromosomal distances (Figure S3).

The higher ranking of hard sweep windows (higher probabilities) was significantly associated with genes that are functionally enriched for a total of 128 GO terms (Table S2). In other words, the windows associated with these GO terms showed on average higher sweep prediction probabilities than expected by chance. Most of the significant gene-rich GO terms within the major ontology "biological process" relate to the function of "gene expression regulation" (Figure 3). This includes functions of "transcription regulation," "RNA processing and splicing" and post-translational regulation ("ubiquination"). Other, less gene-rich groups of significant GO terms relate to "cell cycle/division," "protein complex assembly," "cell/cilium movement," "response to hypoxia," "calcium ion transport" and "reproduction/gamete generation."

Soft sweep windows highly ranked by prediction probabilities were strongly enriched for genes associated with 28 GO terms (Table S3). Similar to hard sweeps, most blue tit genes annotated to GO terms within the ontology "biological process" were linked to "gene expression" and "transcription regulation" (Figure 4). Another prominent group of genes were related to "neuron projection/axonogenesis" and "synapse organization." Less gene-rich groups were associated with "cell/cilium movement" and "calcium ion transport," similar to hard sweeps.

3.2 | Evaluation of a local outbreeding effect

We summed the total length of each sequenced haplotype considering the four indels in the originally amplified microsatellite region of H095_TG05_053 (162 bp in the reference). These combined allele lengths were consistent with the previously determined microsatellite allele length scores for 87.3% of the 150 alleles in the 75 sequenced blue tit individuals. In terms of heterozygosity (yes/no), 90.7% of the 75 individuals were consistently scored between the two methods. The observed discrepancies between alleles scored by sequencing and by fragment amplification are probably due to some failed variants and mis-scoring of indels in the poorly mapped region with many sequence repeats and mononucleotide stretches.

Linkage-disequilibrium in the microsatellite region was generally low and the correlations between the microsatellite alleles and the alleles of the neighbouring SNPs and indels were generally weak (Figure 5a). This was also expressed in the low correlation of heterozygosity between the loci (Figure 5b). Haplotypes of 5 kb length across the microsatellite region (chr5_part1: 1366000–1371000) formed one diffuse cluster in terms of genetic relationship (Figure S4). We therefore conclude that the microsatellite locus stands isolated.

\textbf{FIGURE 2} Manhattan plot showing for each tested genomic window the selective sweep probability of hard sweeps (in red) and soft sweeps (in blue)
in terms of general genetic variation, but also in terms of potential rare genetic variation (i.e., there are no rare distantly related haplotypes; Figure S4).

This finding precludes a simplification of the microsatellite scoring by a specific haplotype grouping. We therefore tested for potential outbreeding effects in the larger sample (635 observations of 459 individuals) by using internal allele distance of each of 83 microsatellites as predictors for arrival date. H095_TG05_053 showed the strongest and the sole panel-wide significant association between allele distance and arrival date (corrected \( p = 0.026 \), Table S4). Individuals with higher allele distances arrived later in the season.

4 | DISCUSSION

Understanding the genomic landscape of signals of past selection is important for the assessment of genomic variation in microevolutionary and population genetic studies in model species such as the blue tit. To analyse the selection history from genomic variance data it is essential to simultaneously infer demographic history, because selection and demographic processes can both produce similar genomic patterns, and interact in the formation of such patterns (Parul et al., 2020). We applied a method with long-range LD measures (up to 13.5 Mb), which allowed us to estimate the recent demographic history of a blue tit population from a few generations ago to more than 1000 years ago. Our ancestral estimate of an effective population size of ~250,000 individuals lies within the range of estimates from a set of models of divergence history between mainland and island populations of blue tits (Perrier et al., 2020). These authors estimated a mean effective population size of 397,685 for the mainland in southern France. Our estimated decline in effective population size, starting about 180 years ago, could be a consequence of the increased replacement of natural mixed, oak-rich forests (and other suitable habitats for blue tits) by pure coniferous forests in the area (Freitag, 2012).
By comparing 12 major parameters as indicators of selection across the focal and flanking windows between empirical data and simulations based on the observed demographic history, we identified about 2000 windows with high probabilities of either a hard or a soft selective sweep. These selective sweep windows, both hard and soft, were widely distributed across all chromosomes and showed no clear clustering. Soft sweeps were more common than hard sweeps (85% vs. 15%). This is in line with recent empirical studies showing the predominance of soft selective sweeps in populations of *Drosophila* and humans (Garud et al., 2015; Schrider & Kern, 2017, but see also Jensen 2014). Abundant recent soft sweeps have also been described in a set of closely related species of southern capuchino seedeaters (*Sporophila* spp.), albeit in the context of sexual selection on pigmentation genes following speciation (Hejase et al., 2020). It has been argued that selection on standing genetic variation may be more likely than selection on new mutants, because established variants have already been filtered and shaped by selection in their ecological context (Hejase et al., 2020). Here, alleles under balancing selection have great potential to mediate quick genetic adaptations when the population encounters new environments, because they necessarily affect phenotype and fitness, and are already at intermediate frequencies at the onset of selection (De Filippo et al., 2016; Rees et al., 2020).

The genes associated with the highly ranked selective sweeps (ranked by probability) showed strong functional enrichments. Functions related to gene expression were strongly overrepresented in many genes associated with hard and soft sweeps. The results suggest that the full range of functional modules related to the regulation of gene expression are overrepresented, that is from transcription initiation, elongation, termination, transport, RNA processing and differential alternative splicing to post-translational modification such as protein ubiquitination. Ubiquitination in
different forms and degrees (the so-called "ubiquitin code") has recently emerged as an omnipresent signalling mechanism that controls diverse physiological processes including gene expression via histone ubiquitination (Kliza & Husnjak, 2020; Swatek & Komander, 2016). A large set of over-represented genes is also specifically related to the transcription of mRNA by the typical RNA polymerase II (Sims III et al., 2004). As expected, functions related to the first steps in gene expression such as transcription initiation were often over represented, because regulation at these steps avoids the costs of follow-up processes such as translation (Wray et al., 2003). Overall, these results emphasize the importance of regulatory evolution in the history of this blue tit population.

At the between-species level, a direct relationship between when, where or how strong a gene is expressed and functionally significant phenotypes has been shown for many genes and gene regulatory networks (Wray et al., 2003). Regulatory changes also appear to predominate adaptive changes among populations or ecotypes of threespine sticklebacks (Gasterosteus aculeatus) that have repeatedly transitioned from marine to freshwater environments (Jones et al., 2012). Our method of GO annotations only allows us to detect trans-acting gene expression regulation, such as via transcription factors or other gene products. Interestingly, trans-regulatory variation contributes more strongly to expression differences within species than between species (Hill et al., 2021). Given the larger mutational target size, trans-regulatory factors might be less affected by the depletion of genetic variation through selective sweeps within populations than cis-regulatory regions. Overall, significant regulatory evolution might have been
triggered multiple times by changing environments. However, compensatory evolution in the extensive regulatory network triggered by single mutations could have produced similar patterns. Indeed, there is strong evidence for compensatory evolution of gene expression in many organisms, also among trans-acting regulatory factors (Signor & Nuzhdin, 2019).

High-ranked soft sweep windows also showed a strong association with many genes acting in transcription regulation. This further supports the importance of selection on regulatory functions in the evolutionary history of this population. In addition, functions related to neuron development, in particular axon and synapse development, were also over represented around soft sweeps. In general, both axon and synapse development can modify brain connectivity and signalling and thus can influence behavioural variation. There is abundant evidence in vertebrates that the rich variation in synapse and neurogenesis genes (and proteins) associates with variation in several behavioural and cognitive traits (Grant, 2016; Lam et al., 2017). A previous study using a genome scan showed that genes acting in synapses and neuron projections are early targets of selection during recent urban colonizations by burrowing owls (*Athene cunicularia*) (Mueller et al., 2020). Several other studies have reported allele frequency shifts in specific synaptic genes during colonization or urbanization in various bird species (van Dongen et al., 2015; Mueller et al., 2013, 2017). Selection tests across different habitats in blue tits and great tits also revealed an enrichment of neuronal functions among the significant associated genes (Laine et al., 2016; Perrier et al., 2020).

In addition to signals of past positive selection, we also identified a different type of selection, namely a negative correlation between heterozygosity at a single microsatellite locus and the fitness-related trait “arrival date at the breeding ground” (Botero-Delgadillo et al., 2020). Here we show through haplotype analysis that this negative outbreeding effect of the specific microsatellite is local and potentially a direct cis-acting effect in the 3′UTR of the gene MDGA2 (MAM domain containing glycosylphosphatidylinositol anchor 2). Linkage with multiple rare partial recessive deleterious alleles—another potential cause for negative heterozygosities–fitness-correlations–appears unlikely in this case (Mueller et al., 2011). Interestingly, MDGA2 interacts with neurolignins and neurexins, which promote synapse development and validation by forming trans-synaptic bridges spanning the synaptic cleft (Gangwar et al., 2017). MDGA2 is overexpressed in the brain and testis, and variants have been associated with harm avoidance and personality disorders in humans (Heck et al., 2011). Of note, Heck et al. (2011) found a negative heterosis effect for one of the intronic SNPs in this gene. Thus, MDGA2 seems a good candidate gene for the study of local adaptation in behaviour.

In summary, a genome-wide scan for signals of selection revealed evidence for selection on gene regulation in the evolution of a blue tit population. Our data further support the idea that genes modulating neuronal connectivity in the brain are important targets of recent soft selective sweeps and outbreeding effects.

**ACKNOWLEDGEMENTS**

We thank A Türk, A Wittenzellner, P Loës, P. Skripsky and the many assistants for fieldwork, S. Kuhn, A. Girg, C. Baumgartner, F. Taborsak-Lines for laboratory work, R-A. Olsen for bioinformatic work, Y. Pei for the drawing of the blue tit in Figure 1 and two anonymous reviewers for constructive comments. Open access funding enabled and organized by ProjektDEAL.

**CONFLICT OF INTEREST**

We declare we have no competing interests.

**AUTHOR CONTRIBUTIONS**

J.C.M., E.B.-D. and B.K. contributed to study conception and design. C.G. collected the arrival data and blood samples with the help of field assistants. P.E. and J.G. coordinated and performed the linked-read sequencing. J.C.M. performed all data analyses with help from P.E.-H. J.C.M. and B.K. drafted the manuscript and all authors provided comments and approved the final manuscript.

**DATA AVAILABILITY STATEMENT**

The blue tit genome browser can be found at http://public-genomes-ngs.molgen.mpg.de. The NCBI Blue tit Annotation Release 100 from the NCBI eukaryotic genome annotation pipeline can be found in the NCBI genome db. The sequence data (fastq files) and the variant call format (VCF) file with all filtered variants and phased genotypes are deposited at the NCBI archive under bio-project number PRJNA759394. Benefits from this research accrue from the sharing of our data and results on public databases as described above.

**ORCID**

Jakob C. Mueller https://orcid.org/0000-0001-6676-7595
Esteban Botero-Delgadillo https://orcid.org/0000-0003-4653-7551
Phil Ewels https://orcid.org/0000-0003-4101-2502
Bart Kempenaers https://orcid.org/0000-0002-7505-5458

**REFERENCES**

Aktas. (2020). https://CRAN.R-project.org/package=haplotypes
Bishara, A., Liu, Y., Weng, Z., Kashif-Haghighi, D., Newburger, D. E., West, R., Sidow, A., & Batzoglou, S. (2015). Read clouds uncover variation in complex regions of the human genome. Genome Research, 25, 1570–1580. https://doi.org/10.1101/gr.191189.115
Boitard, S., Rodriguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genomewide molecular data – An approximate Bayesian computation approach. PLoS Genetics, 12(3), e1005877. https://doi.org/10.1371/journal.pgen.1005877
Botero-Delgadillo, E., Gilsenan, C., Mueller, J. C., & Kempenaers, B. (2020). Negative effects of individual heterozygosity on reproductive success in a wild bird population. Molecular Ecology, 29, 3196–3216. https://doi.org/10.1111/mec.15553
Coulson, T., Albon, S., Slate, J., & Pemberton, J. (1999). Microsatellite loci reveal sex-dependent responses to inbreeding and outbreeding in red deer calves. Evolution, 53, 1951–1960. https://doi.org/10.1111/j.1558-5646.1999.tb04575.x
