Local adaptation for bill length inferred from European great tit genomes

Genetic and phenotypic evidence for selection on bill length in British great tits

British gene variant affects bill length and fitness in European great tits

Genetics of great tits suggest bird feeding in the UK stimulated selection for longer bills

Rapid adaptation for increased bill length in British great tits

Mirte Bosse1,2\*, Lewis G. Spurgin3,4\*, Veronika N. Laine1, Martien A. M. Groenen2, Phillip Gienapp1, Kees van Oers1, Ella F. Cole3, Ben C. Sheldon3, Marcel E. Visser1,2, Jon Slate5

1 Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, the Netherlands

2 Animal Breeding and Genomics Center, Wageningen University, the Netherlands

3 Edward Grey Institute, University of Oxford, United Kingdom

4 School of Biological Sciences, University of East Anglia, Norwich Research Park, United Kingdom

5 Department of Animal and Plant Sciences, University of Sheffield, United Kingdom

\*These authors contributed equally to this manuscript

**Abstract**

A major aim of biology is to link how natural selection operates at the genetic and phenotypic levels. While there have been successful recent examples identifying the genetic basis of known adaptive traits1,2, few studies have used genetic data to identify previously unknown phenotypes subject to natural selection3. Here we show that genetic data reveal a clear signal of phenotypic adaptation across two European great tit (*Parus major major*) populations in the UK and Netherlands. Using a panel of ~500,000 SNP markers4 screened in ~2000 individuals, we found that genes under divergent selection were almost solely associated with morphology, with multiple candidates genes involved in craniofacial features in humans, or with bill morphology in Darwin’s finches. Association analyses between genetic and morphological data confirmed that bill length was under selection between populations. In particular, one of the genes under selection across populations, COL4A5, was associated with bill-length, and allele frequency comparisons suggest that the UK population has recently been under positive selection for longer bills. We then found that variation at COL4A5 was significantly associated with fitness, with individuals with the ‘long-billed’ allele producing more fledglings than those with the ‘short-billed’ allele. Finally, we found that variation at COL4A5 was associated with usage of bird feeders, potentially suggesting that the widespread use of supplementary feeding in the UK since the 19th century has been resulted in selection for longer bills. Together, our data provide a unique example of how genomes can reveal how natural selection operates on phenotypes.

**Main**

Identifying the genetic basis of natural selection in wild populations is becoming an increasingly feasible prospect. Most of the prominent examples using genomics to study adaptation have focused on identifying the genetic basis of known adaptive traits, which has in turn provided insight into the evolution of those traits. However, in many populations and species identifying the genetic basis of natural selection has not been possible. This is because in many species it is not immediately obvious which phenotypic traits are subject to natural selection within and among populations, and these traits may be multifarious and polygenic. Genomic approaches can also be used to identify novel phenotypes previously unknown to be under natural selection, but few clear examples of such studies exist.

The European great tit (*Parus major major*) is a widespread and abundant passerine bird, and a widely used model system for evolutionary and ecological studies. It has recently colonised a wide range of habitats, most likely from a single refugium in the Balkans after the last glacial period. Within and across population analyses suggest that European Despite the wide range of this species, genetic studies suggest that levels of differentiation are extremely low. The great tit genome has recently been sequenced, and a high-density chip of XX polymorphic single nucleotide polymorphism markers has been developed. These excellent genomic resources, combined with the wealth of phenotypic data available from long-term studies, make the great tit an excellent system for studying natural selection in the wild.

We analysed genomic variation within and among three long-term study populations of the European great tit from the UK (Wytham) and the Netherlands (Oosterhout and Veluwe; figure 1A). After filtering (Appendix XX), we had a sample of XX great tits at YY SNPs. Levels of genetic diversity were high and linkage disequilibrium (LD) decayed rapidly within all three sample sites (figure S1b), suggesting large effective population sizes (XX-YY; Table SXX). Admixture and principle component analyses both suggest that genetic structure is limited, but that what structure exists is explained by the variation across three sample sites (figure 1b,c). Genome-wide *F*ST was low across all pairwise comparisons, but highest between Wytham and Oosterhout (*F*ST = 0.0065) and lowest between the two Dutch populations (*F*ST =0.0030). Similar results were found in mitochondrial and microsatellite based studies of European great tits, which are all consistent with a recent and rapid expansion from southern refugia, with high levels of contemporary gene flow between populations.

To identify loci under divergent selection between the three identified clusters, we ran a genome-wide association study (GWAS) on the first two eigenvectors from the PCA (EigenGWAS, REF). We identified highly significant regions of the genome under selection between the UK and Dutch populations (figure 2a), but not between the two Dutch populations (figure SXX). An *F*ST outlier approach yielded identical patterns to the EigenGWAS (supplementary Figure SXX). We extracted all genes linked to markers with EigenGWAS p-value <10-9 to test for an over-representation of genes belonging to specific pathways or ontologies. We found that almost all outlier regions contained candidate genes involved in skeletal development and morphogenesis, with XX of the XX most divergent SNPs within candidate genes for cranial/facial morphology (figure 2b; Table S1). Genes involved in the ontology palate development (GO:0060021) were most significantly overrepresented. The strongest EigenGWAS signal was found at the LRRIQ1 gene (figure 2b). This gene was in the 240kb region found to be associated with beak shape in Darwin’s finches, although the authors of this study suggested that the neighbouring, linked ALX1 gene was responsible for determining beak shape in this species, due to the known involvement of ALX1 in cranial morphology. In our study the LD between LRRIQ1 and ALX1 was low, and the LRRIQ1 locus showed strongest evidence of selection (figure SXX).Also notable are two interacting genes on chromosome 2, *GLI3* and *TRPS1,* which are known to affect facial morphology in humans and mice. Further candidates for genes involved in morphology include *CDHR1, HEPHL1, SATB2, ALX4, SRBD1, BMPR1A, VPS13B, NELL1, NRG3* and *COL4A5* (Table S1).

Because patterns of genetic differentiation suggested that natural selection on bill morphology may have occurred across the two great tit populations, we used morphometric data available for the two populations to test whether variation within regions under divergent selection explained significantly more variation in bill length than expected by chance. We found that... We also performed a GWAS on bill length, and found that multiple regions of the genome had both low GWAS and EigenGWAS p values (figure 2c). These regions can be considered as both under genetic and phenotypic selection across the two populations. Genes in these regions include SOX6, which is involved with xxx, and PTHrP, which is involved with yyy. One ~400kb region of chromosome 4A, containing the COL4A5 gene, contained the most significant GWAS peak for bill length and exactly corresponded to an EigenGWAS peak (figure 2c). Association analyses using haplotype data confirmed the association at the region surrounding COL4A5 (table SXX). Furthermore, examining haplotype variation surrounding the COL4A5 locus revealed a clear signature of recent selection in the Wytham population for the haplotype carrying the ‘C’ allele at the most significant SNP (figure 3a). This allele was associated with longer bill length, and was at high frequency in the UK compared to the Dutch populations (allele frequency = X and Y in the UK and Dutch populations, respectively), where overall bill-length was ca 10% longer (figure 3b). COL4A5 is a type IV collagen gene that has been associated with Alport’s syndrome in humans, but has not previously been associated with craniofacial morphology, or with bill length in birds.

To further elucidate how natural selection has shaped variation in bill length across the two populations, we tested how variation at the COL4A5 locus was related to annual reproductive success and survival. We found a contrasting relationship COL4A5 genotype and the number of fledglings produced between the two populations (GLMM, interaction between genotype and population ID: *n* = ; estimate = ; P = ). Birds harbouring the ‘long-billed’ genotype produced more offspring in the UK compared to the Netherlands, while heterozygotes and short-billed genotype birds did not differ in fitness between the two populations (figure 3c). Thus, we suggest that selection on the COL4A5 gene for longer bills confers a fitness advantage in the Wytham population. We found no evidence that variation at COL4A5 influenced survival.

Selection on bill-length has been documented multiple times in birds, perhaps most notably in Darwin’s finches. However such a finding in omnivorous great tits, not part of a group noted for variation in bill morphology, was unexpected. The UK great tit populations have been characterised as a different race (*P. major newtoni*) compared to the rest of mainland Europe based on bill length, but this classification has been disputed. There are no obvious differences in great tit habitat between the UK and the rest of mainland Europe, and the agent of selection for longer bills in the UK is therefore puzzling. One possibility is that garden bird feeders, which have been widely used in the UK for at least 200 years, has rave resulted in selection for longer bills. It has recently been suggested that garden bird feeders have resulted in natural selection for longer bills in blackcaps (Sylvia atricapilla). We tested this possibility using bird feeder visitation data available for three winter seasons from the Wytham population. We tested whether COL4A5 genotype was associated with variation in the amount of time birds spent on feeders, using seed consumption data estimated from three winter seasons (see supplementary material xx for details). We found that long-billed genotype birds consumed more seeds over the course of winter seasons compared to heterozygotes of short-billed homozygotes (figure 4c). This finding was not due to long-billed genotype birds consuming seeds at a higher rate, but because they visited over more days across the winter season.

Overall, we provide a unique example of how genetic information of wild populations can be used to reveal how selection acts at the genomic and phenotypic level.

**Methods**

Sampling

Sample sites: Wytham (UK); Oosterhout (Netherlands) and Veluwe (Netherlands). The Veluwe forest area is subdivided into six interconnected forests: Bennekomste Bos, Buunderkamp, Hoge Veluwe, Roekelse Bos, Warnsborn and Westerheide. All three sample locations are long-term study sites for great tit research.

Sampling: Blood was collected from a total of 949 specimens in Wytham, 254 in Oosterhout and 1812 in the Veluwe. Pedigree information has been recorded for all individuals. To avoid bias in further analyses due to family relationship, and to exclude any possible mistakes in the pedigree, all individuals with genomic relatedness >0.4 were removed, resulting in 605 birds in Wytham, 188 in Oosterhout and 1519 in the Veluwe.

Sample preparation, genotyping and filtering

DNA was extracted ..? All samples were genotyped on the recently developed high-density 650K affymetrix SNPchip for great tit (?).

Genetic diversity analyses

Inbreeding coefficients were calculated per population in PLINK v1.90b3x by comparing observed and expected homozygous sites with the inbreeding function -het. Genetic linkage within populations was calculated for all marker pairs up to 200kb apart in PLINK and plotted against distance. Pairwise relatedness was calculated in PLINK using the –genome identity by descent (IBD) function. For individual pairs with relatedness >0.4, one of the two was randomly removed to keep relatedness <0.4 in all further analyses, leaving us with 2312 birds. The unrelated dataset was pruned in PLINK (Variance Inflation Factor>2) which resulted in 375846 markers. Principle component analysis was performed on the filtered and pruned dataset using the gcta package. Chromosome 1A, as well as Chrom Z and the small linkage groups were excluded from PCA analysis. Admixture analysis was performed on the filtered pruned data using the software package ADMIXTURE v1.23 with K ranging from 2 to 5. Fst (pairwise) was inferred in PLINK with predefined clusters according to sampling sites.

Selection analyses

The populations were screened for between-population signatures of selection with two distinct measures. The EigenGWAS package was used to apply a GWAS framework to the first two eigenvectors of the principle component analysis. Chromosome 1A was excluded from the PCA but included in the EigenGWAS. Single-marker pairwise Fst was calculated in PLINK with predefined clusters according to sampling sites. We used Pearson's product-moment correlation in R to test for correlation between the EigenGWAS 'Chi' and Fst between Veluwe and Wytham.

We screened the populations for more recent selective sweeps at the COL4A5 locus with an extended haplotype homozygosity test. First, the full dataset was phased with Shapeit v2 with inclusion of genomic relatedness information. After filtering for related individuals, the integrated haplotype score (iHS) was inferred for the most significant GWAS marker at the COL4A5 locus, identifying long and frequent haplotypes as implemented in the R package rehh (Gautier and Vitalis 2012). Starting from the marker with highest iHS signal, a bifurcation diagram was created for both alternative alleles in R package rehh. The site-specific extended haplotype homozygosity (EHHS) and integrated EHHS were visualized with the calc\_ehh(s) function in rehh.

Gene ontology analysis

Regions under selection were tested for an overrepresentation of genes belonging to specific gene ontology terms. All maker annotations within selective sweeps were aligned to chicken genes and tested against the full chicken geneset for human GO-terms with the ClueGo package (?). -vero help.

Gene network/pathway analysis

Association analysis

A genome-wide association study (GWAS) was performed for bill length for both populations. Each trait was tested within-population as well as in a combined setting. …..?????

(Best if Jon adds this part?)

Ecological analyses

We tested how variation at COL4A5 was related to survival and reproduction in the UK using long-term data from the Wytham population. For all fitness analyses we used variation at the most significant SNP identified as being associated will bill length using GWAS as numerical explanatory term (with individuals coded as have 0, 1 or 2 copies of the ‘short billed’ allele). We estimated annual survival rates using capture-mark-recapture models implemented in the program MARK, using the RMark interface. Survival models were constructed as in Sepil et al. (2013), with COL4A5 genotype fitted as an additional explanatory term. We also tested how variation at COL4A5 was related to reproductive success, using generalised linear mixed models implemented in the R package glmmADMB. We used a zero inflated Poisson model, with the number of fledglings produced in an individual year as a response. COL4A5 genotype, age and sex were fitted as explanatory terms and year and individual ID were fitted as random effects.

We tested how variation at COL4A5 was related to usage of artificial bird feeders data from using Passive Integrated Transponder (PIT) tags fitted to birds, which are detected by Radio Frequency Identification (hereafter ‘RFID’) antennae fitted to bird feeders (see Crates et al. 2016 for details). We used data from three winter seasons between 2007 and 2010, for which we had reasonable sample sizes of genotyped birds (N = 167 for 2007-2008, 142 for 2008-2009 and 135 for 1009-2010). We estimated the number of seeds consumed per day by each bird following Crates et al., and included this as a response variable in a linear mixed model with COL4A5 genotype, month and sex as explanatory variables, and individual ID and winter season as random effects.

**References**