Likelihood-based inference of population history from low coverage *de novo* genome assemblies

Jack Hearn[1], Graham N. Stone[1], Lynsey Bunnefeld-McInnes[1], James A. Nicholls[1], Nick H. Barton[2], Konra

[1] Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

[2] Institute of Science and Technology, Am Campus 1, A-3400 Klosterneuburg, Austria

Keywords: Statistical phylogeography, population divergence, admixture, maximum likelihood

Proofs to be sent to:

Konrad Lohse

Institute of Evolutionary Biology,

University of Edinburgh,

Edinburgh EH9 3JT,

United Kingdom

Phone +44 (0)131 650 5508

Email: konrad.lohse@ed.ac.uk

Abstract

Short-read sequencing technologies have in principle made it feasible to draw detailed inferences about the recent history of any organism. In practice, however, such inferences remain challenging due to the difficulty of genome assembly in most organisms and the lack of statistical methods powerful enough to allow discrimination among recent, non-equilibrium histories. We address both the assembly and inference challenges. We develop a bioinformatic pipeline for generating outgroup-rooted alignments of orthologous sequence blocks from *de novo* low-coverage short-read data for a small number of genomes, and show how such sequence blocks can be used to fit explicit models of population divergence and admixture in a numerical likelihood framework. To illustrate our approach, we reconstruct the Pleistocene history of an oak-feeding insect (the oak gallwasp *Biorhiza pallida*) which, in common with many other taxa, was restricted during Pleistocene ice ages to a longitudinal series of southern refugia spanning the Western Palaearctic. Our analysis of blocks sampled from a single genome from each of three major glacial refugia reveals support for an unexpected history dominated by recent admixture. Despite the fact that 80% of lineages are affected by admixture during the last glacial cycle, we are able to infer the deeper divergence history of these populations. These inferences are robust to variation in block length, mutation model, and the sampling location of individual genomes within refugia. This combination of *de novo* assembly and numerical likelihood calculation provides a powerful framework for estimating recent population history that can be applied to any organism without the need for prior genetic resources.

Introduction

Short read sequencing technologies have made it affordable to sequence entire genomes and such data are now being used routinely to infer population history in humans (Gutenkunst *et al.*, 2009; Pickrell &

Pritchard, 2012; Durand *et al.*, 2011) and a small number of model species (Kulathinal *et al.*, 2009; Pool *et al.*, 2012). However, it has remained frustratingly difficult to use such data to infer population history in species for which no prior reference genome exists (Pool *et al.*, 2010; McCormack *et al.*, 2013). Historical inference not only matters for understanding the evolutionary past of a particular species or ecological community (Stone *et al.*, 2012), but is a fundamental pre-requisite for testing for past selection in sequence data (Sousa & Hey, 2013) and in analysing patterns of phenotypic evolution among populations (Stone *et al.*, 2011).

Short-read data pose several technical challenges for historical inference. First, we lack genomic resources for the great majority of organisms, necessitating some form of *de novo* assembly. Second, it is still not cost-effective to obtain genome-level data for many individuals – the general sampling design of population genomic analyses. Third, most methods available for inferring population history from genomic data are either based on allele frequency information (Gutenkunst *et al.*, 2009; Pickrell & Pritchard, 2012; Durand *et al.*, 2011) – which requires large samples and ignores the historical signal contained in the higher moments of branch length distributions – or simply do not scale up to genomic datasets (but see François *et al.*, 2008). And finally, given that population history is evolutionarily recent by definition, the information contained even in whole genomes is fundamentally limited by the time-scales of mutation and genetic drift (Sousa & Hey, 2013; Hey & Machado, 2003).

Given these difficulties, it is unsurprising that the few studies to have used high-throughput data in non-model organisms to date have all incorporated a "genomic reduction" step, resulting in sequencing of only a small proportion of the target genome (McCormack *et al.*, 2013; Arnold *et al.*, 2013). Restriction-site-associated DNA (RAD) sequencing approaches, whereby only a few hundred bases on either side of a particular set of restriction sites are sequenced (Davey & Blaxter, 2010), have been successfully used to detect population structure (Emerson *et al.*, 2010). Similarly, McCormack *et al.* (2012) have developed a

protocol that uses restriction digest to generate a reduced representation library of longer loci. However, while these methods drastically simplify the assembly challenge, they involve additional wet lab protocols for selecting loci and/or multiplexing of genomic libraries. Furthermore, the data generated are not necessarily ideal for inferring intra-specific history. For example, RAD data typically consist of large numbers of unlinked SNPs and so lack the much more detailed information about population history which is contained in the distribution of genealogical branch lengths and accessible only via longer sequences that span multiple, linked polymorphic sites.

An alternative strategy to "genomic reduction" is to work with whole genomes, but limit the analysis to just a few individuals. This has the great advantage that efficient likelihood methods able to deal with genome-scale data already exist (Wang & Hey, 2010; Li & Durbin, 2011; Lohse *et al.*, 2011; Gronau *et al.*, 2011). Discrete population models in particular, although less realistic than spatially continuous models (Barton *et al.*, 2010), have become a standard in population genomics (Harris & Nielsen, 2013; Li & Durbin, 2011; Green *et al.*, 2010; Lohse & Frantz, 2013) because they are tractable and easy to interpret. Li & Durbin (2011) have developed a hidden Markov approach for inferring past changes in effective population size from just a single diploid genome. Similarly, Harris & Nielsen (2013) use the length distribution of homozygous tracts in pairwise alignments to fit more complex histories of divergence and admixture between two populations. However, these methods are currently restricted to histories involving just one or two populations and rely on long, phased sequence blocks and hence near-complete assemblies, which remain challenging to obtain for most organisms. In contrast, multi-locus methods (Hey & Nielsen, 2004; Hey, 2010; Lohse *et al.*, 2011), which assume a set of loci or sequence blocks each of which is short enough to ignore recombination within it, but long enough to contain multiple polymorphic sites, are intuitively appropriate for the fragmented genome assemblies available at low cost using low coverage paired-end sequencing data.

Using such *de novo* assemblies in a multilocus framework requires matching and aligning orthologous sequences both between individuals and between in- and outgroup species. Perhaps more importantly, one needs to show that neither the assembly itself, nor the filtering steps involved in aligning sequences across individuals and species, lead to systematic biases that affect the population genetic analyses. Here we present a pipeline for generation of outgroup-rooted sequence blocks from small numbers of low coverage genomes, and show the resulting data to be representative of the mosaic of genealogies in the genome. We then use a numerical likelihood approach to illustrate the signal inherent in such data by reconstructing the Western Palearctic population history of an oak feeding insect, the oak apple gallwasp *Biorhiza pallida*.

A suite of detailed studies have addressed phylogeographic patterns in Western Palearctic oak gallwasp communities, both for the gall inducers (Stone & Sunnucks, 1993; Rokas *et al.*, 2001, 2003; Stone *et al.*, 2007; Challis *et al.*, 2007) and their parasitoid enemies (Hayward & Stone, 2006; Lohse *et al.*, 2010, 2012; Nicholls *et al.*, 2010a,b). Both groups show genetic structure compatible with three major Pleistocene refugial areas (Iberia; Italy and the Balkans; Asia Minor and Iran, Fig. 1) that broadly parallel those for deciduous oaks (Petit *et al.*, 2003). Most species in both groups show patterns compatible with westwards range expansion into Europe from Asia during or before the Pleistocene (the 'Out of Anatolia' hypothesis; Rokas *et al.* (2003); Challis *et al.* (2007); Stone *et al.* (2009); see also Connord *et al.*), a pattern also supported by a recent meta-analysis of 19 parasitoid and 12 gallwasp species (Stone *et al.*, 2012). The only exception to this pattern to date has been *Biorhiza pallida*, for which mitochondrial and ITS nuclear sequence data show evidence of a deep east-west divide (Rokas *et al.*, 2001). This raises the question of how general the 'Out of Anatolia' pattern is for all three trophic elements of this community (Stone *et al.*, 2009, 2012). Here we use *Biorhiza pallida* as a case study for phylogenomic inference, and ask whether genome-level data support the anomalous pattern for this species within the oak gallwasp community.

We focus on three refugial areas detailed above, referred to hereafter as the Western, Central and Eastern

refuge (Fig. 1). Modelling the relationship between these populations as a series of instantaneous divergence and admixture events (Fig. 2) enables us to test the longitudinal directionality of initial occupation of refugia and of admixture between them during subsequent periods of range expansion, whilst taking incomplete lineage sorting (which is expected) into account. Some recent studies (Stone *et al.*, 2012; Lohse *et al.*, 2010) have explicitly fitted a model of (E(C,W)) population divergence. However, given the recent timescale and the limited number of sequenced loci available in most species, it has so far rarely been possible to quantify the relative contributions of incomplete lineage sorting, divergence and admixture to genetic diversity currently present in refugial populations. We show that this can be achieved using low coverage data for only one individual from each refugial region.

An inherent feature of model-based analyses is that it is necessary to limit the space of models to be explored. For example, it has been possible to fit a very specific divergence and admixture model in which admixture occurs only from the most anciently diverged population and after the most recent population split (see Fig. 2) to human and Neandertal genomic data (Green *et al.*, 2010; Durand *et al.*, 2011; Lohse & Frantz, 2013) because, in this case, the order of population divergence was known *a priori*. However, such prior information does not exist for *B. pallida* or indeed most species. Importantly, we have no reason to assume that population relationships are dominated by divergence (so are tree-like) rather than admixture in the first place (Pickrell & Pritchard, 2012). Thus, to be able to fit divergence and admixture in general, one needs to search model space more broadly. To this end, we have extended existing coalescent theory for the "Neandertal model" developed by Lohse & Frantz (2013) to all possible histories involving unidirectional admixture of a fraction f of lineages to or from the most anciently diverged population (Fig. 2). We feel our model set balances biologically realistic scenarios with computational tractability. It is important to note that unlike D statistics (Green *et al.*, 2010), which are defined relative to the majority topology (assumed to reflect population divergence), our framework can deal with histories that are dominated by admixture (i.e.

$f > 0.5$). This allowed us to use the *B. pallida* genomic data to compare the support for a large number of models ($n = 32$) and make detailed inferences about the history of this species. In particular we investigate i) how well the *B. pallida* data can be explained by an (E,(C,W)) divergence history as inferred previously for many species in the Western Palearctic, and ii) whether its history is dominated by deep population divergence or recent admixture.

Given that our likelihood scheme is restricted to minimal samples of individuals, it is crucial to test how representative individual genomes are of the long-term population relationships captured by discrete population models. Clearly, in spatially continuous populations, local genetic structure emerges as a consequence of the limited dispersal ability of individuals (Barton *et al.*, 2010), and so any model that approximates a population occupying a large area as a panmictic unit must break down over recent time-scales. To address this, we repeated our analyses using two different individuals from each refugium.

Materials and Methods

Sequencing and sampling

DNA was extracted from individual wasps using the Qiagen DNeasy kit. Like most Hymenoptera, *Biorhiza pallida* has haploid males and diploid females; haploid males were selected for genome sequencing because there is no need to phase alleles and SNP calling and estimation of sequencing error rates are greatly simplified. However, we stress that our method does not rely on haploid genomes and can easily be applied to unphased diploid data (Lohse & Frantz, 2013) (see Discussion). Illumina 50 and 100 base-pair paired-end libraries (Table S2) were prepared using the Illumina paired-end DNA sample preparation kit, the DNA sheared was using the Covaris S2 instrument and size selection was carried out on a 2% agarose TAE gel (fragments with an average insert size of 300bp were excised). These were then sequenced on the GAII,

GAIIX and HiSeq2000 platforms at the NERC GenePool facility in Edinburgh (CONFIRM THIS). Note that the different read lengths used are simply a result of technological improvements during the course of this work and not a necessary part of our sequencing strategy. Short read data are deposited at the ENA Sequence Read Archive (http://www.ebi.ac.uk/ena/data/view/ERP002280).

Each of five male in-group individuals (2 West, 2 Central and 1 Eastern see Table S1 and S2 and Fig. 1) was sequenced to a modal coverage of 1.5 fold per individual, yielding a total modal coverage of 7.45 for the *B. pallida* genome across all individuals. In each of the West and Central refuges we sampled replicate individuals (referred to as Wa/Wb and Ca/Cb respectively, see Fig. 1) from sites 400km apart. This separation is well above the dispersal ability of an individual gallwasp (Stone & Sunnucks, 1993) and was intended to incorporate any impact of within-refuge population genetic structure. Sampling of replicate individuals was not possible in Iran. To polarize mutations as ancestral or derived, we sequenced two diploid female individuals from a closely related outgroup species *Belizinella gibbera* (to a total coverage of 5.76 fold across individuals). A breakdown of reads per individual is given in tables S2 and S3.

Multi-locus inference methods (Hey & Nielsen, 2004; Lohse *et al.*, 2011) assume a large number of sequence blocks that are i) sampled at random from the genome, ii) short enough to ignore recombination within them and iii) in linkage equilibrium within populations. We developed a simple bioinformatic pipeline (Fig. 3) that generates out-group-rooted alignments meeting the above criteria for a small number of individuals. In short, the pipeline consists of three steps:

i) Assembly

Initial read quality was assessed using *FastQC* (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were quality trimmed at Q20 using *sickle* and adapter-trimmed using *scythe* and *cutadapt*. After quality filtering, reads from all individuals in each species were combined to create species-specific meta-

assemblies using the CLC *de novo* assembler v4.06 (Dryad repository doi: XXX). These were used as a reference for downstream bioinformatics processing (Table 1). For *B. pallida* we included a small amount of 50-base read data from a pool of six related (sib or half-sib) females that were sequenced as part of another project (Table S2).

Reads from each *B. pallida* individual were mapped back to the meta-assembly using *Stampy* (run with a substitution rate of 0.0025 and the -baq and -sensitive options) to create BAM files. Although a high proportion (92 – 97 % depending on the individual) of read-pairs mapped to the reference meta-assemblies, the percentage of "properly paired mappings", in which both read pairs align to the same contig, was much lower (37%) (see Table S3). This is expected, given the low coverage per individual and the short read length. Many read-mapping failures with *Stampy* are known to be due to reads that overlap the ends of contigs. This is related to assembly coverage, specifically of the reads used to generate the initial contigs (and contig trimming/minimum coverage filtering implemented by assemblers) rather than the reads used to call SNPs.

ii) Filtering

RepeatScout and *RepeatMasker* were used to *de novo* predict and mask repeat regions in both the *B. pallida* and *B. gibbera* meta-assemblies. This removed 51% and 34% of the *B. pallida* and *B. gibbera* meta-assemblies respectively. The data were further filtered according to the following three criteria. First, we searched for orthologous regions shared amongst in- and outgroup meta-assemblies using a discontiguous, reciprocal megaBlast search with an e-value cut-off of 10^{-20} to match contigs between the in- and outgroup meta-assemblies (Altschul *et al.*, 1990). We only kept the reciprocally-best hit if it was at least 100 bit scores better than the next best overlapping hit. To avoid penalising good unique hits with short overlaps, we allowed for 15-base overhangs between best hits. The coordinates of the retained reciprocal blast hits

were used to create a BED file for each species and reads for each individual overlapping these regions were extracted from the BAM alignments (Quinlan & Hall, 2010). VCF files were created for the ingroup and outgroup using the sub-sampled BAM files. Second, any contigs in these VCF files that matched putative contaminant (bacteria, fungi) or mitochondrial DNA were removed. Finally, we removed contigs with excessive coverage (>75 fold coverage for *B. pallida* and > 30 fold for *B. gibbera*, see Fig. S1) as they were likely to indicate remaining unfiltered collapsed repeats whose sequences had been amalgamated during assembly (Nagarajan & Pop, 2013). Together, these filtering steps reduced the number of contigs by 80.1% corresponding to 14.1% of the original *B. pallida* meta-assembly size (Table 1). Note that we are not attempting to distinguish between coding and non-coding sequence at this stage (see Sensitivity analyses). Raw variants (excluding indels) were called across individuals using *samtools mpileup* (Li *et al.*, 2009).

iii) Generating individual consensus sequences and triplet alignments

Consensus sequences were generated for each individual from the VCF files using a custom perl script (available on Dryad XXX). For the in-group; (1) the reference base was called if no variant was present or the variant did not reach a particular quality threshold, (2) an 'N' was coded if an individual had 0 coverage at that position or was called heterozygous by *samtools* (indicating a sequencing error because we know the individual to be haploid), or more than two alleles were present, violating the assumption of the infinite sites model assumed in our likelihood analysis; (3) a SNP was called for sites that differed from the reference. The script used to create consensus sequences from the VCF file has the option of specifying a quality score filter for SNP calling. We explored Q0, Q10, Q20, and Q30 to assess differences in SNP frequencies at different quality thresholds and selected Q0, as the frequencies did not change (see Fig. S2).

Because the outgroup was represented by two diploid individuals, '0/1' genotypes at a position could be true heterozygotes. To avoid any impact of ancestral polymorphisms between in- and outgroup, positions

that were variable between the outgroup individuals (0/0 versus 1/1) were coded 'N'. We generated a single outgroup consensus sequence, thus taking advantage of deeper sequencing by combining data from two individuals. Finally, we generated outgroup-rooted triplet alignments consisting of a single individual from each of the East (*E*), Center (*Ca* or *Cb*) and Western (*Wa* or *Wb*) refugia using *muscle* (Edgar, 2004). To avoid including linked sequences as separate blocks and to increase the size of blocks, non-overlapping alignments that mapped to the same contig in the *B. pallida* meta-assembly were concatenated into the same block. Alignments were generated for all four possible West/Center/East combinations of *B. pallida* individuals and the outgroup consensus sequence (Dryad repository doi: XXX).

For simplicity, the following analyses focus on two such triplet sets. One, referred to hereafter as dataset *a*, comprised individuals *Wa*, *Ca* and *E* (Table S1 and Fig. 1). The other, referred to hereafter as dataset *b*, comprised individuals *Wb*, *Cb* and *E*.

Counting mutation types

Given outgroup-rooting and assuming an infinite sites mutation model with only two allelic states per site, each mutation can be unambiguously placed onto a genealogical branch. This means that the polymorphism information can be condensed into a vector of mutation counts on branches (Patterson *et al.*, 2006). While these counts of mutation types within sequence blocks constitute the input for the likelihood analyses (see section below), their relative frequencies across all sites reveal the distribution of alternative genealogies across the genome and hence the types of history that are plausible. As a check, we counted the three types of shared derived mutations and the three singleton mutations before and after filtering alignments to contigs $> 2kb$. Note also that the current implementation of our model does not allow for back mutations; given the recent timescale of divergence the chance of a backmutation within the ingroup is slim. Back mutations on the outgroup branch remain a possibility, and if present would perhaps slightly the estimated divergence

times.

Maximum likelihood analyses of historical models

To conduct a broad search of model space, we took a strict divergence model between three populations as a starting point and considered all histories that involve a single unidirectional admixture event either to or from the oldest population. We did not include models with bidirectional or multiple admixture events because the additional parameters make the corresponding generating functions computationally intractable, but also because these models are biologically unexpected: expansion out of refugia is expected to be a unidirectional process. For a given order of population divergence, there are six possible models (Fig. 1), each with five parameters: the time of the older split (T_2); the time of the more recent split (T_1); the time of admixture, or gene flow, (T_{gf}) (all measured from the present); the admixture proportion (f) and the effective population size (N_e). Again, for the sake of computational tractability, we assumed a single N_e for both ancestral populations as well as the population receiving migrants. We assessed the support for all six admixture scenarios as well as simpler, nested models that assume no admixture and divergence between either three or two populations for each of the three possible orderings of population divergence (a total of 24 divergence and admixture models). We also quantified the support for a basal polytomy, a single panmictic population and for distinct ancestral N_e values in the strict divergence models (to test whether the additional parameter substantially improved model fit without the need to invoke admixture), giving 32 models in total.

The general method for calculating likelihoods is described in detail elsewhere (Lohse *et al.*, 2011; Lohse & Frantz, 2013). In short, the probability of observing a particular mutational configuration in a sequence block (which can be interpreted as the likelihood of the model) can be expressed in terms of a higher order derivative of the generating function (GF) of genealogical branch lengths (Lohse *et al.*, 2011, eq. 1). Thus, given the GF for a model, it is straightforward to tabulate the logarithm of the likelihood, lnL, given all

observed mutational configurations. Assuming that loci are unlinked and hence statistically independent, the joint lnL across loci is simply the sum of lnL. We used *Mathematica v.8* (Wolfram Research, 2010) to tabulate lnL values and maximise the joint likelihood numerically (Supporting *Mathematica* notebook). The GF conditional on a topology of a triplet genealogy has been previously derived for a divergence model with recent admixture from the population involved in the older divergence event (scenarios D and E in Fig. 2) (Lohse & Frantz, 2013). We used the general recursion for the GF (Lohse *et al.*, 2011, eq. 4) (and *Mathematica* to solve equations) to find analogous expressions for the other four admixture scenarios depicted in figure 1. Although their derivation is relatively straightforward, the resulting expressions are cumbersome and given in the Supporting *Mathematica* notebook.

The accuracy of the likelihood method to estimate particular model parameters can be quantified using the Fisher information (*I*), a measure of the sharpness of the lnL curve near the maximum (Edwards, 1972; Lohse & Frantz, 2013). The average information about a parameter contained in a sequence block is given by summing *I* over all possible mutational configurations weighted by their probability ((Lohse & Frantz, 2013), eqn. 3/6 check?). The expected information in a data set consisting of *n* sequence blocks is simply *n x E[I]*.

Sampling blocks

For each of the six models, we numerically computed the parameter values that maximized lnL across a large number of sequence blocks of fixed length.

Because this calculation ignores statistical associations between blocks due to linkage and we lack information about the relative position of contigs in the *B. pallida* genome, the number of blocks must be chosen such that the probability that two blocks are physically linked by chance can be ignored. Assuming a genome size of 1.75Gb for *B. pallida* (the average measured in oak gall wasps (Lima, 2012)) and sampling of blocks

by chance alone, the distance between neighbouring blocks is exponentially distributed with rate $n/1.75Gb$ (where n is the number of blocks). This implies that if we classify blocks separated from their nearest neighbour by 20kb or more as being in linkage equilibrium and want to ensure that less than 5% of all blocks fall below this threshold, we could in theory sample a maximum of $-(1.75Gb \times Log[0.95])/20kb \approx 4500$ blocks.

We chose a miminum length of 2kb for the inclusion of contigs in maximum likelihood analyses as this length represented a good trade-off for obtaining blocks long enough to include enough polymorphic sites for inference and short enough not to worry about linkage among contigs. Sub-sampling from the full set of contigs with this length cut-off gave between 2231 and 2640 blocks (depending on the combination of W/C/E individuals), roughly 10% of the contigs meeting the initial filtering requirements (Table 1). To be able to compare likelihoods across datasets, we fixed the number of blocks to 2231 in all analyses.

We initially used the first 1kb of sequence from each aligned contig in the 2kb-filtered data and explored the impact of block length by repeating the analysis with shorter (500bp) and longer (2kb) blocks.

We estimated the proportion of coding sequence in the filtered data by Blast-searching all aligned contigs against a preliminary *B. pallida* transcriptome assembly (Dryad repository doi XXX). To incorporate mutation rate heterogeneity, sequence blocks were partitioned according their predicted proportion of coding sequences into 10 equally spaced bins. We used the average divergence between *B. pallida* and *B. gibbera* to calibrate a relative mutation rate for blocks in each bin.

Results

Below, we first examine the counts of mutation types to draw qualitative inferences about the history of *B. pallida*. We then describe how maximum likelihood can be used to distinguish quantitatively between alternative historical scenarios. Finally, we assess the sensitivity of these inferences to the mutation model, length of sequence blocks, sampling location of individuals and our assumption of no intra-locus recombination.

Counting mutation types

The full *a* datasets comprised 84,822 aligned contigs > 300 bases long with an N50 value of 803 bases (see Table 1). We recovered a total of 171,694 polymorphic sites in the in-group, corresponding to an average per site diversity (as measured by Watterson's θ_W) of 0.188 % (Table 2). Average divergence between the outgroup and the Eastern individual was 4%. If population divergence were to take place in the order (E,(C,W)) without admixture, we expect derived sites shared by Central and Western individuals (C/W) to be more common than both derived sites shared by Central and Eastern individuals (C/E) and sites shared by Western and Eastern individuals (W/E). Likewise C/E and W/E sites, which correspond to internal branches of genealogies that are incongruent with the population history, are expected to occur at equal frequency (Hudson, 1983; Tajima, 1983). Analogously, under null models of a polytomic split or a single panmictic population, all three types of shared derived sites are equally likely. Contrary to these simple models, we found that C/E sites were more frequent (9.6 %) than W/E sites (5.1%), which in turn were more frequent than W/C sites (2.8 %) (see top two rows of Table 2, CHECK THESE FIGURES!). This double asymmetry suggests that simple divergence models without gene flow provide a poor fit to the data. If we assume that the majority class of informative sites corresponds to the order of population divergence, then these results imply that the Western population diverged from the common ancestor of the Central and Eastern populations before these in turn diverged. Under this model, the observed excess of W/E sites relative to W/C sites could arise as a consequence of gene flow between Western and Eastern refugia (Fig. 1) after the more recent C/E split (Durand *et al.*, 2011; Lohse & Frantz, 2013).

Maximum likelihood analyses of historical models

Comparing the three possible histories of strict divergence, a population tree topology (W,(C,E)) had highest support (lnL), as expected from the frequencies of shared derived sites. Allowing for different values of N_e

in the two ancestral populations did improve the fit of the strict divergence model. However, 8-9 of the 18 models involving admixture (depending on which of the a or b datasets is considered) had greater support than models of strict divergence (Table 3). The best supported history assumes a (W,(C,E)) population tree topology with substantial admixture ($f = 0.76 - 0.83$ across a and b datasets) (Table 4) from the Eastern into the Western refuge shortly after the split between Center and East populations (model B in Fig. 2). The observed number of blocks showing each mutational configuration fits the number expected (Table S4) under this best model well. Reassuringly, the alignments for the two sets of individuals *a* and *b* yielded the same ranking of models and gave very similar parameter estimates with broadly overlapping 95 % C.I. (see Tables 4 and S5). Interestingly, however, the estimated admixture proportion f was slightly higher in both triplet analyses involving the individual from southern rather than northern Spain (*Wb*, Fig. 1) (see Discussion).

Because our models are not nested, we cannot use likelihood ratio tests to test for significance. Each admixture model also contains the same number of parameters, so comparisons based on Akaike information criterion (AIC) reduce to comparisons based on change in log likelihood (ΔlnL). To assess our confidence in our ability to identify the best model, we conducted a simulation study to quantify the power of our method. Briefly, we simulated 100 replicate 2231 loci datasets using Hudson's ms program (Hudson, 2002) and the ML parameter estimates for our 1kb *WaCaE* data (Table 4). Ninety-nine out of 100 replicates identified the same best model as obtained for our observed data (see Table 3). The second best models were 61% model A, 32% model C and 6% model E (Fig. 2, all (W,(C,E)) topology). Furthermore, the parameter estimates for the intervals between T_{gf} and T_1 were tiny, such that none predicted the observed asymmetry in the number of blocks specifying (C,(E,W)) or (E,(C,W)) topologies. This result underlines the fact that support for model B comes from both mutation counts and configurations.

To provide an order of magnitude calibration for the inferred history, we applied a direct, genome-wide estimate of the effective neutral mutation rate of 3.5×10^{-9} per site and generation as measured in

Drosophila melanogaster (Keightley *et al.*, 2009). To account for the bias towards conserved sequence in our 2kb filtered data, we scaled the *D. melanogaster* rate by the ratio of per site diversity in the filtered and unfiltered data (0.47 and 0.54 for *a* and *b* data respectively, see θ_W in Table 2). Assuming that *B. pallida* has two generations per year (Csóka *et al.*, 2005; Atkinson *et al.*, 2003) this calibration gives effective population sizes between 39,000 – 52,000 (Table 4). The time of admixture and the more recent split (t_{gf}, t_1) both date to the last glacial period (Weichselian, 12-110kya), whereas the maximum likelihood estimate for the oldest split (t_2) falls in the previous (Saalian, 130-200kya) glacial period (Table 4).

We also calculated ΔlnL as parameters move away from their maximum likelihood estimates (Fig. 4). and used Fisher Information to quantify how informative our data are about a particular model parameter, and hence how accurate one can expect parameter estimates to be. We found that, for our best supported model and the 1kb *WaCaE* dataset there is less information associated with T_2 than with the other three parameters (Table S6). With 2231 loci, we expect a standard deviation (SD) of 0.143 in estimates of T_2 but 0.0274 in estimates of T_1 (see also Table 4).

SNPs vs. blocks

To assess what information, if any, is gained by using sequence blocks instead of SNPs for inference, for each parameter we calculated the expected Fisher information in a single block as a function of θ (setting all other parameters to their maximum likelihood estimates from the best supported model for the 1kb *WaCaE* dataset). We also calculated the expected information in a single SNP for each parameter for the same model. We found that even blocks containing only one SNP on average (θ = 0.266) are more informative across all parameters than a single SNP (Fig. S3).

One can use the generating function framework to obtain expected branch lengths, the sum of which gives the expected total tree length. Our likelihood approach can then be used with the observed SNP frequencies

(Table 2, third row) to find maximum likelihood parameter values without recourse to any blocking scheme. Reassuringly, when applied to the generating function for model B, very similar parameter estimates are obtained, namely: T_{gf} = 1.095, T_1 = 1.095 and T_2 = 3.34. Note, that the interval between T_{gf} and T_1 is estimated at zero.

Sensitivity analyses

i) Length filtering

Filtering contigs by length could result in various biases that might affect inference. For example, more conserved and/or structurally complex regions of the genome are expected to assemble better and align with fewer errors, and so should be represented by longer contigs. To quantify this effect, we correlated contig length against per site divergence. As expected, longer contigs were on average less diverged between ingroup and outgroup (Fig. S4) (Kendall's $\tau = -0.0419$, $p < 10^{-6}$). Consistent with this, the average per site diversity (θ_W) in the 2kb filtered *a* data was about half of that in the unfiltered data (Table 2). This confirms that length filtering does indeed enrich for conserved sequences. However, for the purpose of estimating population history, any overall bias in absolute diversity can be incorporated by a simple rescaling of the mutation rate (see below). In contrast, in order to justify treating the 2kb-filtered data as a random sample of genealogies, we need to show that length filtering does not affect the relative frequencies of mutational types (i.e. the polarized allele frequency spectrum normalized by the proportion of polymorphic sites). To test this, we obtained a random sample of putatively unlinked SNPs before and after filtering the *a* data for contigs > 2kb by selecting one SNP at random from each sequence block. In the length-filtered data, all 2231 blocks were included. In the full data, SNPs were drawn from a random sample of 4500 sequence blocks to minimize linkage effects. Reassuringly, we found no significant difference between the filtered and unfiltered data in the relative frequencies of the three types of shared derived mutations (the

most informative site types) (Table 2) ($\chi^2 = 1.96$, $p = 0.38$). However, there was a significant (but slight) excess of singleton mutations compared to shared derived sites in the 2kb data ($\chi^2 = 9.3$, $p = 0.0023$). This may be either due to assembly or alignment bias or purifying selection (which is likely to be stronger in the 2kb-filtered data) (Charlesworth *et al.*, 1993).

ii) Mutational heterogeneity

The likelihood method ignores mutational heterogeneity between blocks. This assumption may be problematic given that the *B. pallida* data consists of a mix of coding and non-coding sequence. There was no hit to the transcriptome of *B. pallida* for 50% of all contigs and, across all sites, the proportion of coding sequence was 70%. This, together with the increased GC content in the filtered *a* data (Table 1), clearly showed that our filtering strategy enriched for coding sequence. To incorporate mutational heterogeneity, we partitioned blocks by their predicted proportion of coding sequence (see Methods) and scaled the effective neutral mutation rate of each bin using the within bin divergence (per site) relative to the total divergence across all sites. This drastically improved model fit (i.e. increased lnL) (see Table 4), but had no impact on the ranking of alternative models or parameter estimates under the best supported model. However, we did find that incorporating mutational heterogeneity led to a slight reduction in both divergence time and N_e estimates (see Table 4).

iii) Intra-locus recombination

To investigate the robustness of the maximum likelihood estimates to the assumption of no recombination within blocks, we repeated the analysis with shorter (500b) and longer (2kb) blocks, both sub-sampled from each contig in the 2kb-filtered data. In both cases, the relative ranking of models was unaffected and parameter estimates were similar to those obtained in the initial 1kb analysis (Tables S7 and S8). This suggests that undetected recombination within blocks has a minor effect on our results.

To investigate this further, we also conducted a simulation study to assess whether ignoring intra-locus recombination biases model choice or parameter values. As starting points, we took the ML parameter estimates from the 1kb *WaCaE* dataset (Table 4) and the estimated recombination rate for *Nasonia* of 1.5 cM/Mb (or crossovers/generation/bp x 10^{-8}) from Niehuis (2010). Using Hudson's *ms* (Hudson, 2002), we simulated triplet datasets using seven recombination rates (0, 0.015, 0.15, 0.3, 0.33, 1.5, 3.3, 7.5 crossovers/generation/bp x 10^{-8}). If the Keightley *et al.* (2009) mutation rate calibration is assumed, these correspond to r/μ ratios of 0, 0.0454, 0.454, 0.907, 1, 4.5, 10 and 22.7. For each parameter combination, we simulated 1,000,000 loci in order to obtain expected mutational configurations. We removed any loci that failed the four gamete test (removing between 0.0338 and 3.47 % of loci), although the remaining loci will still include undetected recombination events. We then parsed the polymorphism information into vectors of mutation counts. We found that across all r/μ ratios, the best supported model matches that recovered from the observed data. ΔlnL values between this and the second, third and fourth-ranked models are similar to that for the observed data (Fig. S6). As the r/μ ratio increases, parameter estimates become more biased: θ and f decrease, while the splitting and admixture times increase (Fig. S6).

Discussion

We show how outgroup-rooted alignments of thousands of orthologous sequence blocks can be generated for multiple individuals using low-coverage ($<$ 2 fold per individual) genomic data and standard *de novo* assembly tools. Although the requirement for orthologous sequences in in- and outgroup, the filtering against repetitive sequences and short contigs enrich for coding and otherwise selectively constrained sequence – in the case of *B. pallida* – the allele frequency spectrum is little affected. This suggests that the resulting data provide a representative sample of neutral variation in the genome which, if analysed in a multi-locus framework, is highly informative about recent history.

Admixture dominates the history of *Biorhiza pallida*

The model we fit to *B. pallida* of (W,(C,E)) population divergence with strong East to West admixture differs qualitatively from previous population genomic inferences of divergence with admixture (Green *et al.*, 2010; Lohse & Frantz, 2013) in two ways. Firstly, admixture is from the more recently diverged population (E) into the older population (W), and hence in the opposite direction to that observed in three-population analyses of our own Neandertal ancestry (Green *et al.*, 2010; Durand *et al.*, 2011). Secondly, the history of *B. pallida* is dominated by admixture rather than by divergence. Despite this, the majority class of shared derived sites is still C/E, and so concordant with the order of population divergence (W,(E,C)). This is a peculiar consequence of the direction of admixture: going backwards in time, W lineages that trace back to the E population via admixture only spend a short time in the E population before they trace back to the ancestral C/E population.

Both the order of population divergence and the direction of admixture are unexpected. First, our inference of initial divergence of the Western refuge contrasts with a previous meta-analysis of 12 oak gallwasps (including *B. pallida*) and 19 associated parasitoid species (Stone *et al.*, 2012), as well as a multi-locus study that compared the history of four oak gall parasitoid species (Lohse *et al.*, 2012). Both studies found a general signature of (E,(C,W)) divergence on a community scale, but had insufficient power to resolve the order of population divergence in individual species (or to fit additional admixture parameters). Interestingly, however, the deep split of the Iberian population from other refugia we infer here for *B. pallida* is compatible with the mitochondrial genealogy reconstructed by Rokas *et al.* (2001). Second, the history of *B. pallida* involves substantial admixture from the Middle East into Iberia without affecting the Balkans. One plausible route for such admixture that would not pass through the central refuge is westwards migration into Iberia across North Africa, possibly via southern Italy and Sicily. Striking floristic links between Iberia and Asia Minor have been found across a range of plant taxa (Davis & Hedge, 1971), including oaks

(Lumaret *et al.*, 2002), and there is genetic evidence that Iberia was colonised from North Africa during the Pleistocene by some animal taxa (Griswold & Baker, 2002; Habel *et al.*, 2008). This scenario is also compatible with our finding of a higher admixture fraction for the sample from Southern Iberia (*Wb*) compared to Central (*Wa*) Iberia, since the Wb sample would be closer to the putative origin of North African immigrants. Similarly, the genetic similarity of extant populations of oak gallwasps (Rokas *et al.*, 2003) and their parasitoids (Nicholls *et al.*, 2010b) in Morocco and Spain suggests that the Strait of Gibraltar presents little or no barrier to gene flow. Given that we lack molecular calibrations for Hymenoptera in general and gallwasps in particular, our absolute time estimates are tentative at best. Nevertheless, it is clear that the divergence and admixture between refugial populations of *B. pallida* is recent, encompassing no more than two or three glacial cycles.

Sampling the genome and the limits of power

While in the past, most statistical analyses of phylogeographic scenarios were limited in power by the number of available loci (Carstens *et al.*, 2009; Lohse *et al.*, 2012), the massive replication of sequence blocks afforded by short-read sequencing overcomes this and – in the case of *B. pallida* – allowed us to reliably identify the best fitting history among a set of alternative divergence and admixture scenarios.

However, despite increasing the number of loci by several orders of magnitude, the difference in support we find for some alternative models (Table 3) is still relatively modest, suggesting that the power to distinguish more complex models is limited. For example, it would be hard to distinguish multiple admixture events from a single event or a model of continuous migration (Hey & Nielsen, 2004). It is worth reiterating that the lack of linkage information for the *B. pallida* assembly imposes a limitation on the number of blocks we were able to include in the maximum likelihood analyses. The final analysis only included 2.2Mb of sequence, a mere 0.13 % of the genome, and most of the assembled genome remained unused. If

one had complete linkage information, i.e. if the relative position of blocks was known, one could sample blocks at fixed intervals (Lohse & Frantz, 2013), which would increase the number of blocks that can safely be taken as unlinked by an order of magnitude. Alternatively, one could ignore linkage between blocks altogether when obtaining point estimates of parameters (which are unaffected) and use a simple scaling factor to adjust confidence intervals and ΔlnL (see Lohse & Frantz, 2013). However, the gain in power one could expect is limited. In general, increasing the number of independently segregating blocks by a factor k increases the accuracy of parameter estimates by $\sim \sqrt{k}$. Instead, it is the recent time-scale of the *B. pallida* history that sets an inherent limit to the complexity of models among which one can hope to discriminate using a multi-locus approach. If only a small number of mutations have occurred during the history of interest (as is the case in *B. pallida* where most 1kb blocks contain two or fewer mutations), there are only a few mutational configurations that are observed at appreciable frequency (Table S4).

Given this mutational limitation, it is clear that increasing the number of individuals sampled from within each population would also only slightly improve inference: most ancestral lineages would coalesce rapidly, i.e. the vast majority of genealogical branches added by larger samples would be unresolved, and so would not give much extra information. Very large samples of a long non-recombining sequence can be informative (Kong *et al.*, 2011), but mainly about even more recent population history than the timescale considered here. Sampling individuals a further distance apart would give extra information, but also requires more complex models, involving multiple parameters for separation times and admixture rates. In general, these considerations suggest that there will be an upper limit to the signal contained in even an extremely large number of short, unlinked sequence blocks. Nevertheless, even short blocks containing on average only one SNP contain more information than single SNPs (Fig. S4).

In contrast, we would have far more information if we could analyse the full linear sequence and explicitly use linkage information. In *B. pallida*, a total of 3.5% of the genome would be usable after filtering for

unique orthologous sequence, but allowing an arbitrary degree of linkage; ultimately, of course, we could use the whole genome in such an analysis. The gain does not come primarily from the sheer volume of data; rather, we gain extra information from the lengths of sequence blocks. For example, the length of block that shares the same genealogy within a population is inversely proportional to its coalescence time, and the length of unrecombined, introgressed blocks of genome decreases with the time since introgression. Thus, recombination gives an additional time-scale, beyond that provided by mutation, as used here. Barton *et al.* (2013) show that in a two-dimensional continuum, the distribution of block lengths shared between genomes allows inference of both dispersal rate and neighbourhood size, whereas samples of allele frequencies do not give information about dispersal rate. Li & Durbin (2011) use the distribution of heterozygous SNPs to infer ancestral population size through time, whilst Harris & Nielsen (2013) use this information to infer complex migration histories. However, a full statistical analysis that takes into account the linear structure of the genetic map not only remains extremely challenging analytically, but also requires much better assemblies or linkage maps than can currently be achieved for most organisms in practice. Nevertheless, even without such whole-genome data, correlations between linked loci can be informative and it will be interesting to see to what extent including this information in the present maximum likelihood framework improves inference.

In the meanwhile, the combination of *de novo* assembly and numerical likelihood computation we develop here provides a level of resolution far beyond that of traditional phylogeographic analyses of a few loci. The fact that our bioinformatic pipeline yielded sufficient data (and resolution to distinguish between models) in an oak gallwasp, the group with the largest known genomes in the Hymenoptera (Lima, 2012), should encourage those working on other non-model species species and ecological communities (Stone *et al.*, 2012). Furthermore, our sensitivity analyses suggest that population historical inferences based on large numbers of blocks and few individuals are robust in two fundamental ways. Firstly, and despite the fact that undetected recombination can bias multi-locus analyses (Strasburg & Rieseberg, 2009), neither

model selection nor parameter estimates are much affected by the length of sequence block. The slight biases observed at high r/μ ratios are in line with expectations: as recombination scrambles histories across blocks of sequence the variance in branch lengths across loci is artificially decreased, leading to underestimations of θ and overestimations of divergence times (Wall, 2003). Incidentally, the fact that small (500bp) blocks with less than two mutations on average are sufficient to distinguish these models also implies that the unresolved phase in diploid genomes would not be an issue when applying this framework (Lohse & Frantz, 2013). The chance of multiple heterozygous sites within such short blocks and within an individual is negligible. Secondly, the fact that we recover essentially the same population history using individuals sampled many dispersal distances apart highlights that simple, discrete population models can be a useful approximation to recent, intra specific histories.

Acknowledgements

We would like to thank Majide Tavakoli, Julja Ernst, Pablo Fuentes-Utrilla and Rachel Atkinson for collecting specimens, Marian Thomson for help with the DNA extractions, the NERC Biomolecular Analysis Facility (NBAF) node in Edinburgh (The GenePool) for library preparation and Illumina sequencing and John Davey and Mark Blaxter for advice on bioinformatic strategy. This work was funded by NERC grants to G Stone, J. Nicholls, K Lohse and N Barton (NE/J010499, NBAF375, NE/E014453/1 and NER/B/S2003/00856). K Lohse was supported by a UK NERC fellowship (NE/I020288/1).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.G. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403 – 410.

Arnold, B., Corbett-Detig, R.B., Hartl, D. & Bomblies, K. (2013). Radseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, 22(11), 3179–3190. ISSN 1365-294X. doi:10.1111/mec.12276.

Atkinson, R.J., Brown, G.S. & Stone, G.N. (2003). Skewed sex ratios and multiple founding in galls of the oak apple gall wasp *Biorhiza pallida*. *Ecological Entomology*, 28(1), 14–24.

Barton, N.H., Etheridge, A.M., Kelleher, J. & Véber, A. (2013). Inference for the spatial lambda-Fleming-Viot process. *Theoretical Population Biology*, page *in press*.

Barton, N.H., Kelleher, J. & Etheridge, A.M. (2010). A new model for extinction and recolonisation in two dimensions: Quantifying phylogeography. *Evolution*, 64(9), 2701–2715.

Carstens, B.C., Stoute, H.N. & Reid, N.M. (2009). An information-theoretical approach to phylogeography. *Molecular Ecology*, 18(20), 4270–4282.

Challis, R.J., Mutun, S., Nieves-Aldrey, J.L., Preuss, S., Rokas, A., Aebi, A., Sadeghi, E., Tavakoli, M. & Stone, G.N. (2007). Longitudinal range expansion and cryptic eastern species in the western palaearctic oak gallwasp *Andricus coriarius*. *Molecular Ecology*, 16(10), 2003–2014.

Charlesworth, B., Morgan, M.T. & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4), 1289–303.

Connord, C., Gurevitch, J. & Fady, B. (????). Large-scale longitudinal gradients of genetic diversity: a meta-analysis across six phyla in the mediterranean basin. *Ecology and Evolution*, 2, 2600–2614.

Csóka, G., Stone, G.N. & Melika, G. (2005). The biology, ecology and evolution of gall-inducing Cynipidae. In C. Raman, W. Schaefer & T.M. Withers, editors, *Biology, ecology and evolution of gall inducing insects*, pages 573–642. Science Publisher, Enfield, New Hampshire.

554 Davey, J.W. & Blaxter, M.L. (2010). Radseq: next-generation population genetics. *Briefings in Functional*
555 *Genomics*, 9(5-6), 416–423. doi:10.1093/bfgp/elq031.

556 Davis, P.H. & Hedge, I. (1971). Floristic links between NW Africa and SW Asia. *Annales Naturhistorisches*
557 *Museum Wien*, 75(1), 43–57.

558 Durand, E.Y., Patterson, N., Reich, D. & M., S. (2011). Testing for ancient admixture between closely
559 related populations. *Mol Biol Ecol*, 28(8), 2239–2252.

560 Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic*
561 *Acids Res.*, 32(5), 1792–1797.

562 Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press, Cambridge.

563 Emerson, K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W.E. & Holzapfel,
564 C.M. (2010). Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of*
565 *the National Academy of Sciences*, 107(37), 16196–16200. doi:10.1073/pnas.1006538107.

566 François, O., Blum, M.G.B., Jakobsson, M. & Rosenberg, N.A. (2008). Demographic history of european
567 populations of *Arabidopsis thaliana*. *PLoS Genet*, 4(5), e1000075. doi:10.1371/journal.pgen.1000075.

568 Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W.,
569 Fritz, M.H.Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.S., Jensen, J.D., Marques-Bonet, T., Alkan, C.,
570 Prufer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B.,
571 Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit,
572 J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V.B., Golovanova,
573 L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler,
574 E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D.
575 & Paabo, S. (2010). A draft sequence of the Neanderthal genome. *Science*, 328(5979), 710–722.

576 Griswold, C.K. & Baker, A.J. (2002). Time to the most recent common ancestor and divergence times of
577 populations of common chaffinches (*Fringilla coelebs* in europe and north africa: Insights into pleistocene
578 refugia and curent levels of migration. *Evolution*, 56(1), 143–153.

579 Gronau, I., Hubisz, M., Gulko, B., Danko, C. & Siepel, A. (2011). Bayesian inference of ancient human
580 demography from individual genome sequences. *Nature Genetics*, page 43.

581 Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & D., B.C. (2009). Inferring the joint demo-
582 graphic history of multiple populations from multidimensional SNP frequency data. *PLoSGenetics*, 5(10),
583 e1000695.

584 Habel, J.C., Meyer, M., El Mousadik, A. & Schmitt, T. (2008). Africa goes europe: The complete phylo-
585 geography of the marbled white butterfly species complex *Melanargia galathea/M. lachesis* (lepidoptera:
586 Satyridae). *Organisms Diversity & Evolution*, 8(2), 121–129.

587 Harris, K. & Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths.
588 *PLoS Genetics*, page *in press*.

589 Hayward, A. & Stone, G.N. (2006). Comparative phylogeography across two trophic levels: the oak gall
590 wasp *Andricus kollari* and its chalcid parasitoid *Megastigmus stigmatizans*. *Molecular Ecology*, 15(2),
591 479–489.

592 Hey, J. (2010). Isolation with migration models for more than two populations. *Molecular Biology and*
593 *Evolution*, 27, 905–920.

594 Hey, J. & Machado, C.A. (2003). The study of structured populations - new hope for a difficult and divided
595 science. *Nature Reviews Genetics*, 4(7), 535–543.

596 Hey, J. & Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and
597 divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*.
598 *Genetics*, 167(2), 747–760.

599 Hudson, R.R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution*,
600 37, 203–217.

601 Hudson, R.R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioin-*
602 *formatics*, 18, 337–338.

603 Keightley, P.D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S. & Blaxter, M.L. (2009). Analysis of the
604 genome sequences of three drosophila melanogaster spontaneous mutation accumulation lines. *Genome*
605 *Research*, 19(7), 1195–1201.

606 Kong, Q.P., Sun, C., Wang, H.W., Zhaio, M., Wang, W.Z., Zhong, L., Hao, X.D., H. Pan, S.Y.W., Cheng,
607 Y.T., Zhu, C.L., Wu, S.F., Liu, L.N., Jin, J.Q., Yao, Y.G. & Zhang, Y.P. (2011). Large-scale mtDNA
608 screening reveals a surprising matrilineal complexity in East Asia and its implications to the peopling of
609 the region. *MBE*, 28, 513–522.

610 Kulathinal, R.J., Stevison, L.S. & Noor, M.A.F. (2009). The genomics of speciation in *Drosophila*: Diversity,
611 divergence, and introgression estimated using low- coverage genome sequencing. *PLoSGenetics*, 5(7),
612 e1000550.

613 Li, H., Wysoker A, Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. (2009). The
614 sequence alignment/map SAMformat. *Bioinformatics*, pages 2078–2079.

615 Li, H. & Durbin, R. (2011). Inference of human population history from individual whole-genome se-
616 quences. *Nature*, 475(7357), 493–6.

Lima, J. (2012). *Species Richness and Genome Size Diversity in Hymenoptera with Different Developmental Strategies: A DNA Barcoding Enabled Study*. Ph.D. thesis, University of Guelph.

Lohse, K., Barton, N.H., Melika, N. & Stone, G.N. (2012). A likelihood-based comparison of population histories in a parasitoid guild. *Molecular Ecology*, 49(3), 832–842.

Lohse, K. & Frantz, L. (2013). Maximum likelihood evidence for Neandertal admixture in Eurasian populations from three genomes. *MBE*, page *in review*.

Lohse, K., Harrison, R.J. & Barton, N.H. (2011). A general method for calculating likelihoods under the coalescent process. *Genetics*, 58(189), 977–987.

Lohse, K., Sharanowski, B. & Stone, G.N. (2010). Quantifying the population history of the oak gall parasitoid *C. fungosa*. *Evolution*, 58(4), 439–442.

Lumaret, R., Mir, C., Michaud, H. & Raynal, V. (2002). Phylogeographical variation of chloroplast DNA in holm oak (*Quercus ilex* l.). *Molecular Ecology*, 11(11), 2327–2336.

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C. & Brumfield, R.T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, 66, 526–538.

McCormack, J.E., Maley, J.M., Hird, S.M., Derryberry, E.P., Graves, G.R. & Brumfield, R.T. (2012). Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Molecular Phylogenetics and Evolution*, 62(1), 397–406.

Nagarajan, N. & Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics*, 14(157-167).

Nicholls, J.A., Preuss, S., Hayward, A., Melika, G., Csóka, G., Nieves-Aldrey, J.L., Askew, R.R., Tavakoli,

637 M., Schönrogge, K. & Stone, G.N. (2010a). Concordant phylogeography and cryptic speciation in two
638 western Palaearctic oak gall parasitoid species complexes. *Molecular Ecology*, 19, 592–609.

639 Nicholls, J., Fuentes-Utrilla, P., Hayward, A., Melika, G., Csoka, G., Nieves-Aldrey, J.L., Pujade-Villar, J.,
640 Tavakoli, M., Schonrogge, K. & Stone, G. (2010b). Community impacts of anthropogenic disturbance:
641 natural enemies exploit multiple routes in pursuit of invading herbivore hosts. *BMC Evolutionary Biology*,
642 10(1), 322.

643 Niehuis, Oliver Gibson, J.D.R.M.S.P.B.A.K.T.J.A.K.D.C.A.K.K.D.D.B.L.W.v.d.Z.L.S.D.M.W.J.H.G.J.
644 (2010). Recombination and its impact on the genome of the haplodiploid parasitoid wasp nasonia. *PLoS*
645 *ONE*, 5(1), e8597.

646 Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S. & Reich, D. (2006). Genetic evidence for complex
647 speciation of humans and chimpanzees. *Nature*, 441(7097), 1103–1108.

648 Petit, R.J., Csaikl, U.M., Bordacs, S. & Burg, K. Coart, E.e.a. (2003). Chloroplast dna variation in european
649 white oaks phylogeography and patterns of diversity based on data from over 2600 populations. *Forest*
650 *Ecology and Management*, 176, 595–599.

651 Pickrell, J.K. & Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele
652 frequency data. *PLoS Genet*, 8(11), e1002967. doi:10.1371/journal.pgen.1002967.

653 Pool, J.E., Corbett-Detig, R.B., Sugino, R.P., Stevens, K.A., Cardeno, C.M., Crepeau, M.W., Duchen, P.,
654 Emerson, J.J., Saelao, P., Begun, D.J. & Langley, C.H. (2012). Population genomics of sub-saharan
655 *Drosophila melanogaster*: African diversity and non-african admixture. *PLoS Genet*, 8(12), e1003080.
656 doi:10.1371/journal.pgen.1003080.

657 Pool, J.E., Hellmann, I., Jensen, J.D. & Nielsen, R. (2010). Population genetic inference from genomic
658 sequence variation. *Genome Research*, 20(3), 291–300. doi:10.1101/gr.079509.108.

Quinlan, A.R. & Hall, I.M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842. doi:10.1093/bioinformatics/btq033.

Rokas, A., Atkinson, R., Brown, G., West, S.A. & Stone, G.N. (2001). Understanding patterns of genetic diversity in the oak gallwasp *Biorhiza pallida*: demographic history or a *Wolbachia* selective sweep? *Heredity*, 87, 294–305.

Rokas, A., Atkinson, R.J., Webster, L., Csóka, G. & Stone, G.N. (2003). Out of Anatolia: longitudinal gradients in genetic diversity support an eastern origin for a circum-mediterranean oak gallwasp *Andricus quercustozae*. *Molecular Ecology*, 12(8), 2153–2174.

Sousa, V. & Hey, J. (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nat Rev Genet*, 1(6), 404–414.

Stone, G.N., Challis, R.J., Atkinson, R.J., Csóka, G., Hayward, A., Melika, G., Mutun, S., Preuss, S., Rokas, A., Sadeghi, E. & Schönrogge, K. (2007). The phylogeographical clade trade: tracing the impact of human-mediated dispersal on the colonization of northern Europe by the oak gallwasp *Andricus kollari*. *Molecular Ecology*, 16, 2768–2781.

Stone, G.N., Hernandez-Lopez, A., Nicholls, J.A., di Pierro, E., Pujade-Villar, J., Melika, G., Cook, J.M. & Abbot, P. (2009). Extreme host plant conservatism during at least 20 million years of host plant pursuit by oak gallwasps. *Evolution*, 63(4), 854–869.

Stone, G.N., Lohse, K., Nicholls, J.A., Fuentes-Utrilla, P., Sinclair, F., Schönrogge, K., Csóka, G., Melika, G., Nieves-Aldrey, J.L., Pujade-Villar, J., Tavakoli, M., Askew, R.R. & Hickerson, M.J. (2012). Reconstructing community assembly in time and space reveals enemy escape in a western palaearctic insect community. *Current Biology*, 22(6), 531–537.

680 Stone, G.N. & Sunnucks, P. (1993). Genetic consequences of an invasion through a patchy environment - the
681 cynipid gallwasp *Andricus quercuscalicis* (Hymenoptera: Cynipidae). *Molecular Ecology*, 2(4), 251–268.

682 Stone, G., Nee, S. & Felsenstein, J. (2011). Controlling for non-independence in comparative analysis of
683 patterns across populations within species. *Phil. Trans. Roy. Soc. Series B Biol Sci.*, 366, 1410–1424.

684 Strasburg, J.L. & Rieseberg, L.H. (2009). How robust are isolation with migration analysies to violations of
685 the IM model? A simulation study. *Molecular Biology and Evolution*, 27(2), 297–310.

686 Tajima, F. (1983). Evolutionary relationships of DNA sequences in finite populations. *Genetics*, 105(2),
687 437–460.

688 Wall, J.D. (2003). Estimating ancestral population sizes and divergence times. *Genetics*, 163(1), 395–404.

689 Wang, Y. & Hey, J. (2010). Estimating divergence parameters with small samples from a large number of
690 loci. *Genetics*, 184, 363–373.

691 Wolfram Research, I. (2010). *Mathematica, Version 8.0*. Wolfram Research, Inc., Champaign, Illinois.

692 Data Accessibility

693 • Raw Illumina reads for *B. pallida* and *B. gibbera*: ENA Sequence Read Archive (http://www.ebi.ac.uk/ena/data/view/ERP002

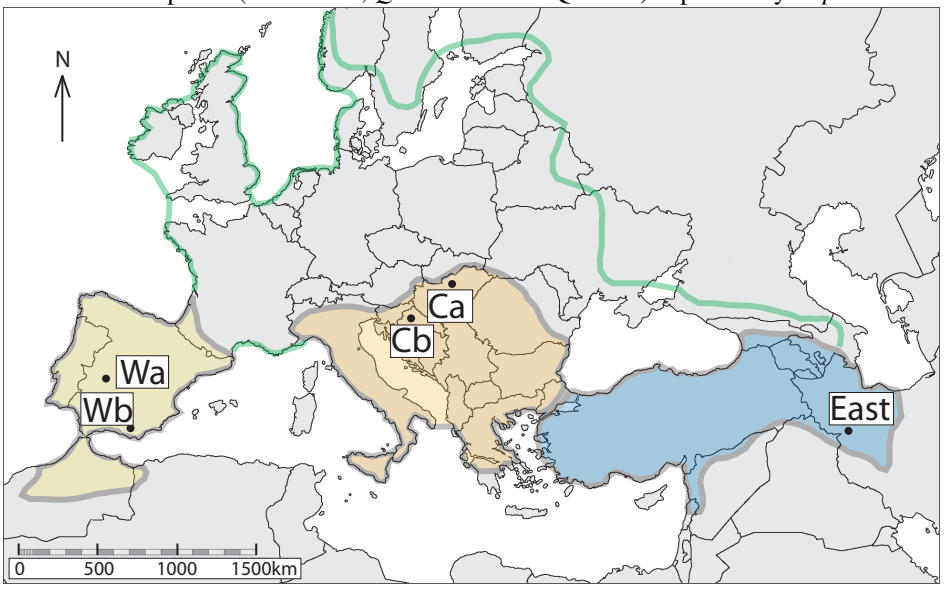
694 • Meta-assemblies for *B. pallida* and *B. gibbera*: Dryad doi: XXX

695 • *B. pallida* transcriptome: Dryad doi: XXX

696 • Final triplet alignments and summary files for Mathematica: Dryad doi: XXX.

Figures

Figure 1: Sampling locations of the five *B. pallida* individuals (Wa, Wb, Ca, Cb and E, see Table S1) used for genome sequencing and population genomic analyses. Refugial regions are colour coded as follows: W, West, in green; C, Centre, in orange; and E, East, in blue. The green line shows the current postglacial extent of the oak host plants (white oaks, *Quercus* section Quercus) exploited by *B. pallida*.



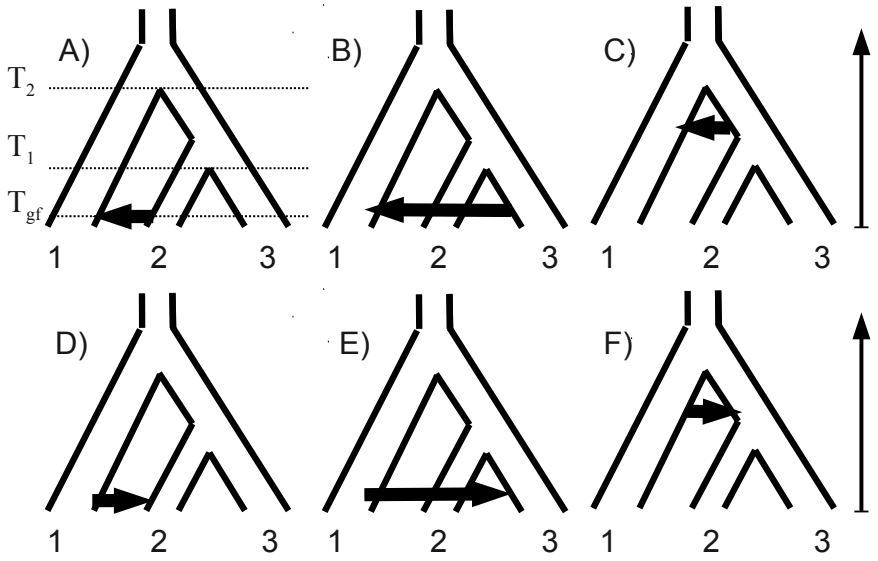


Figure 2: The six possible models for divergence histories involving unidirectional admixture to or from the older population (labelled: 1). Given the three possible orders of population divergence, there are 18 admixture models in total. Divergence times (T_1 and T_2) and the time of admixture (T_{gf}) are measured back in time from the present.

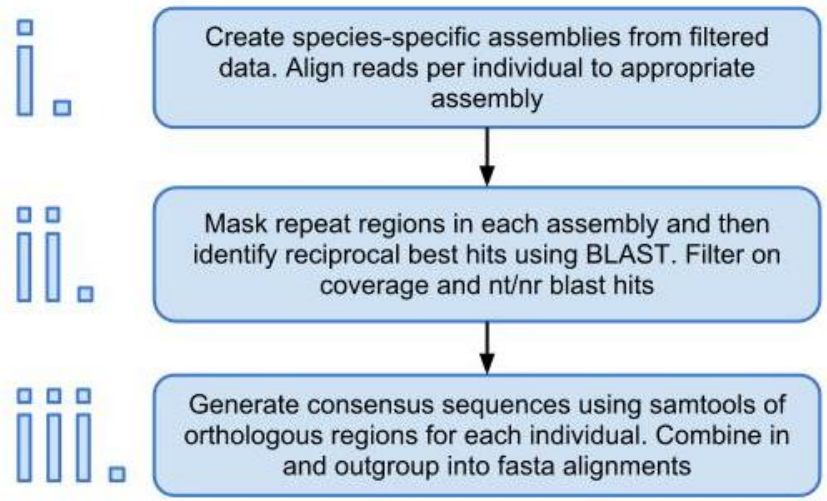


Figure 3: Assembly and filtering steps used to generate population genomic datasets in *B. pallida*.

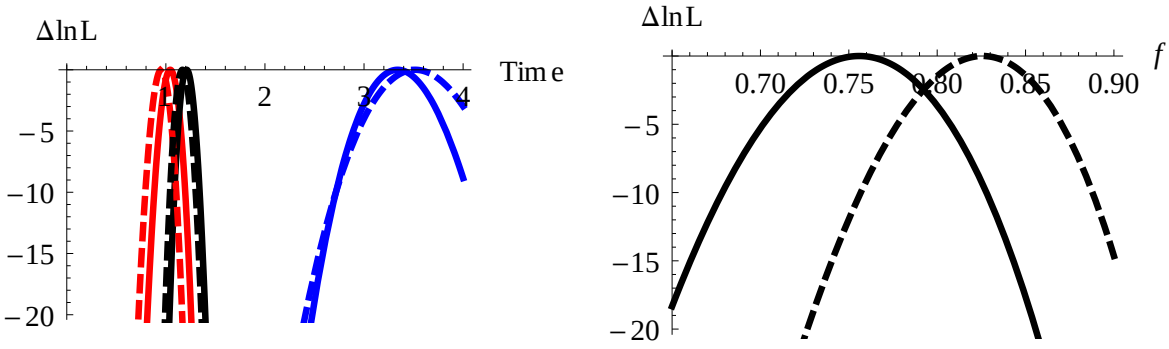


Figure 4: A) ΔlnL plots for the times of divergence (T_1 (black) and T_2 (blue)) and admixture T_{gf} (red). B) ΔlnL for the admixture proportion f. Estimates from the a data are shown as solid lines, those from the replicate data set b as dashed lines.

Tables

Table 1: **Summary statistics for meta-assemblies and filtered datasets**

Species	N50	# of contigs	Total bases	GC %
Biorhiza pallida	1,075	1,163,314	805,102,378	32.9
Belizinella gibbera	643	817,710	443,963,639	36.1
Biorhiza pallida, post Filter	734	232,097	113,583,710	36.7
Belizinella gibbera, post Filter	508	290,379	111,785,775	35.9
a data >300 bp	803	84,822	61,012,720	38.6
b data >300 bp	768	77, 752	54,117,641	39.0
a data 2 kb	2,000	2,640	5,280,000	40.1
b data 2 kb	2,000	2,231	4,462,000	40.2

Summaries are shown for in- and outgroup meta-assemblies (first four rows) and the *a* and *b* triplet data before and after length filtering. N50 is defined as the length N for which 50% of all sequenced bases are assembled in a contig of length $<$ N.

Table 2: **Genetic diversity and relative frequencies of mutational types in *B. pallida* alignments.**

Dataset	θ_W	W	C	E	W/C	W/E	C/E
a data $>$ 300 bp	0.00188	0.325	0.214	0.263	0.040	0.058	0.100
b data $>$ 300 bp	0.00147	0.269	0.244	0.283	0.044	0.060	0.100
a data 2 kb	0.00089	0.338	0.220	0.267	0.027	0.049	0.098
b data 2 kb	0.00079	0.276	0.250	0.287	0.035	0.054	0.099

Before (>300bp) and after (>2kb) length filtering.

Table 3: **Support for alternative scenarios of divergence and admixture in the oak gall wasp *B. pallida* (*WaCaE*, 1kb data)**

	Model	k			
	Panmixia	1		-589.3	
	Polytomy	2		-88.7	
	gene flow		$(W_1, (C_2, E_3))$	$(C_1, (E_2, W_3))$	$(E_1, (C_2, W_3))$
A)	$2 \rightarrow 1$	5	-9.1, (T_1)	-18.8	-18.2, (f^*)
B)	$3 \rightarrow 1$	5	**0**	-88.7, (T_1, T_2)	-88.7, (T_1, f^*)
C)	$2/3 \rightarrow 1$	5	-4.8	-88.7 (T_{gf}, f^*)	-88.7, (T_{gf}, T_2)
D)	$1 \rightarrow 2$	5	-25.7, (f)	-18.2, (T_1)	-18.2, (f^*)
E)	$1 \rightarrow 3$	5	-18.0	-88.7, (T_1, T_2)	-88.7, (T_1, T_2)
F)	$1 \rightarrow 2/3$	5	-25.7 (f^*)	-79.4	-33.4, (T_{gf})
	2 pop.	2	-260.8	-404.0	-474.5
	3 pop.	2	-25.7	-88.7, (T_2)	-88.7, (T_2)
	2 pop. N_e	3	-48.5	-90.1	-93.7
	3 pop. N_e	4	-20.8	-88.7, (T_2)	-88.7, (T_2)

Support (ΔlnL) relative to the best model for alternative histories of refugial populations of *B. pallida* estimated from the *a* dataset (Model B in Fig. 2 has highest support and is shown in bold). The labelling of populations (1–3) and of models (A–F) corresponds to that in Fig. 2; all scenarios involving unidirectional admixture were assessed for each of the three possible orders of population divergence (columns 1–3). Models of strict divergence without admixture between two (2 pop., i.e. $T_1 = 0$) or three (3 pop.) populations were fitted assuming either a single or two different N_e for ancestral populations. Parameters for which the maximum likelihood estimate is 0 (i.e. the model reduces to a simpler nested model) are indicated in brackets (f^* refers to complete admixture, i.e. $f = 1$).

Table 4: **Parameter estimates under the best supported model (see Table 3).**

dataset	μ het.	lnL	f	θ (N_e)	T_{gf} (t_{gf})	T_1 (t_1)	T_2 (t_2)
a, 1kb	no	-9269.3	0.76	0.69 (52,000)	1.04 (54KY)	1.21 (63KY)	3.34 (173KY)
			(0.72, 0.79)		(51–58KY)	(60 – 66 KY)	(158 – 189KY)
b, 1kb	no	-8815.1	0.83	0.64 (43,000)	0.95 (41KY)	1.17 (50KY)	3.51 (151KY)
			(0.80, 0.86)		(38–44KY)	(51 – 57 KY)	(135 – 168KY)
a, 1kb	yes	-8769.7	0.76	0.61 (45,900)	1.10 (50KY)	1.26 (58KY)	3.45 (158KY)
			(0.72, 0.79)		(47–54KY)	(55 – 60 KY)	(143 – 172KY)
b, 1kb	yes	-8444.0	0.82	0.58 (39,100)	0.97 (38KY)	1.17 (51KY)	3.47 (136KY)
			(0.79, 0.85)		(35–40KY)	(49 – 54 KY)	(121 – 151KY)

Maximum likelihood estimates are given for different triplet combinations and analyses with and without mutational heterogeneity (μ het.; see Methods). Both effective population size and divergence time parameters are scaled relative to the rate of coalescence, i.e. in $2N_e$ generations. Absolute values are given in brackets, calibrated using a direct, genome-wide mutation rate for *Drosophila* (Keightley *et al.*, 2009) and assuming two generations per year. 95 % confidence intervals of scaled parameter values are given in brackets below the point estimate. f is the admixture proportion, θ is the scaled mutation rate, N_e is the effective population size, T_{gf} is the time ago of admixture (t_{gf} is the calibrated estimate), T_1 and T_2 are the younger and older splitting times in the population topology, with t_1 and t_2 the absolute ages, respectively.