THE UNIVERSITY of EDINBURGH
SCHOOL OF BIOLOGICAL SCIENCES

Genome wide selection in a guild of hymenopteran parasitoids: Ten month report

William Walton

Principal supervisor: Konrad Lohse
Second supervisor: Graham Stone
Third supervisor: Mike Ritchie (St Andrews)
Committee chair: Andrew Rambaut

Word count: 5796

Introduction

The focus of my thesis is inferring selection from sequence data. The detection of natural selection has been an important part of population genetics since its foundation (Fisher, 1930; Hudson, Kreitman, & Aguade, 1987; Kimura, 1968; Tajima, 1989), however there is still much about selection that is not understood and many improvements to be made to tests for it. One such gap in our knowledge is the degree to which any particular ecological factor contributes to overall genomic selection (Egan, Nosil, & Funk, 2008). This question requires a comparative approach, and I will therefore attempt to answer this by comparing selective signatures across species. An outstanding problem with many selection tests is that they are vulnerable to the effects of demography, as it can have similar effects on the genome to selection (Li et al., 2012). I will therefore also work on explicitly incorporating demography in selection tests by first modelling demography separately. The action of selection on gene duplicates and the role of duplicates in adaptation is another area that requires further investigation. Although it is widely accepted that duplication is a major source of genetic variation (Zhang, 2003) there are few studies showing evidence of local adaptation by recent duplication (Qian & Zhang, 2014). I aim to improve our understanding by investigating selection on gene duplicates. Finally, existing methods to quantify selection typically either use haplotype or site frequency spectrum (SFS) information, and often require data from many individuals. The SFS is a summary of genetic data that counts the number of alleles in each frequency class across individuals in a sample. A test combining both of these types of information could be powerful using only a small sample size, and I will therefore attempt to develop a method to quantify selection based on the "blockwise" method of demographic inference described in Lohse et al (2011). This method uses both linkage and frequency information by cutting the sequence into short blocks and calculating a SFS for each block. I will answer these questions using whole genome data from two insect systems. Below I give an overview of my first thesis chapter which has been largely completed as well as a brief outline for chapters 2-4.

Chapter 1: Comparing population history in a guild of insect parasitoids

The detection and quantification of natural selection is becoming ever more advanced with improved data and analysis methods (Berg & Coop, 2015; Hudson, Kreitman, & Aguade, 1987; Manel et al., 2016; Nielsen et al., 2005). However, few studies have been able to link the action of selection with ecological factors in order to find out the causes of selection in an ecological context (Egan, Nosil, & Funk, 2008). This linking of ecology and selection requires a comparative approach with many species, and has only recently become possible due to the improvements in and reduction in cost of sequencing, allowing many species to be sampled and sequenced relatively inexpensively.

When inferring selection, it is necessary to control for demographic effects due to the similar effects on the genome of demography and selection. Rapid changes in population size are amongst the strongest factors known to confound tests of selection. For example, a strong population bottleneck can produce a star shaped genealogy with an excess of rare variants reminiscent of a strong selective sweep (Bunnefeld, Frantz, & Lohse, 2015). Likewise, simple rapid changes in population size can result in patterns of variation similar to those formed by the action of selection (Slatkin & Hudson, 1991; Williamson et al., 2005). Controlling for demography by explicit modelling of it ensures the greatest accuracy of the selection test (Huber, Nordborg, Hermisson, & Hellmann, 2014; Li et al., 2012). I therefore

aim to assess models of population bottlenecks and step changes in population size in order to use these as null models to infer selection.

However, not only can changes in population size affect genomic tests for natural selection, but selection can also skew the patterns of variation used for demographic inference (Schrider, Shanku, & Kern, 2016). To reduce the impact of selection, demographic tests can be performed using only intergenic genome regions as the effects of selection are often most pronounced around genes.

Bottlenecks and sudden changes in population size are important events in population histories that are often associated with introductions to new environments, range shifts and climate fluctuations. The assessment of demographic models can therefore be informative of the response of a species to geological events such as glaciations (eg. Moura et al., 2014). Assessing support for these changes in population size in species across a community can reveal the ways in which a species' ecology affects its response to these events, and the degree to which the effect is the same across species. Closely synchronised changes in population size across species could reflect the effect of climate or some other environmental factor on all species at once.

A comparative study of selection and demography requires a system in which potentially confounding factors can be controlled for. The system must be ecologically closed so that it can be considered in isolation of other communities, and species must be in taxa that are closely enough related to control for evolutionary history, and define panmictic populations (populations without structure) so that assumptions of population genetic models can be fulfilled. It is important for the population to be approximately panmictic because population structure creates linkage between alleles and makes patterns similar to bottlenecks and selective events, which disrupts tests for demography and selection (Huber et al., 2014). There must also be control for habitat variation and good existing data on background ecology and phylogeography so that variable factors of interest are known and all others are controlled for, and inferences on demography and selection can be seen in the context of the phylogeography.

Very few systems fulfilling these criteria exist, but one such system is the oak gall wasp parasitoid system which is present across the Western Palearctic region and consists of oak trees, herbivorous gallwasps, and parasitoids and inquilines (Hayward & Stone, 2005; Stone, Schönrogge, Atkinson, Bellido, & Pujade-Villar, 2002). These insects are abundant and provide a food source for higher trophic levels. Galls are induced on oak trees by gallwasps for the development of their larvae, and the galls are also inhabited by commensal inquilines and parasitoid enemies of the gallwasps. There is a wealth of interactions between these species, with each gallwasp being attacked by several species of parasitoid and each parasitoid also having a unique host range. The bioinformatics and population genetics are simplified by the wasps' haplodiploidy (Askew et al., 2013; Hayward & Stone, 2005). The system has been shown to contain at least three ice age refugia, where popualtions of each species were able to survive during glacial periods and from which they have subsequently spread in the last interglacial period. These refugia are in Iberia, the Balkans and Iran, and represent panmictic populations (Stone et al., 2012). The times at which these populations split from one another has been estimated for several parasitoid species (Bunnefeld et al., *unpubl*.). Furthermore, the reference genome of the chalcid wasp *Nasonia vitripennis,* in the same family Pteromalidae as some of the parasitoids, is available and some annotation has been performed (de Graaf et al., 2010).

The species to be investigated are parasitoid wasps from the superfamily Chalcidoidea. These are *Megastigmus dorsalis, M. stigmatizans* and *Torymus auratus* in the family Torymidae, *Ormyrus nitidulus* and *O. pomaceus* in the family Ormyridae, *Eurytoma brunniventris* in Eurytomidae and *Cecidostiba fungosa* in Pteromalidae. All of these seven

species under study have ranges across the Western Palearctic and have well studied historical phylogeography and background ecology (Stone et al., 2012, Bunnefeld et al., *unpubl.*). Existing data consist of fragmented assemblies from whole genome libraries of five individuals in each species sampled from the Iberian refugial population, as well as some individuals from the Balkan and Iranian refugial populations. The Iberian refugial population was selected as the focal population as it is well studied, panmictic and all species under study have overlapping ranges there (Askew et al., 2013; Stone et al., 2012). Whole genome data were used as they allow comparison of genic and intergenic regions, and provide the maximum number of loci for the greatest power when performing tests of demography and selection.

This first chapter aims to show, from the beginning of the Pleistocene epoch about two and a half million years ago up to the present date, how population history varies and the extent to which changes in population size are concurrent across co-distributed species in an Iberian community of parasitoid wasps. To do this, I will assess support for models of population bottlenecks and step changes in effective population size (N_e), and compare results across species. The impact of selection on these demographic inferences will be investigated by predicting genes and repeating the tests on genic and intergenic regions of the genome. The demographic models will then be used as null models when inferring selection in subsequent chapters.

Methods

Models of bottlenecks and step changes in population size were fitted using the blockwise composite likelihood method described by Lohse *et al.* (2011). This method is ideally suited to studies of non-model organisms with fragmented genome assemblies and small sample sizes such as these as it uses only short blocks of sequence from up to five individuals. Briefly, an alignment of the individual genomes is split into "blocks" of sequence, and a SFS is calculated for each block. This is a blockwise SFS (bSFS), and acts as a summary of the genealogy and mutational information for each block. The probability of each bSFS under a particular demographic model is calculated. Multiplying these probabilities across blocks in the data provides a composite likelihood of the model given the distribution of genealogies. It has been shown that this bSFS method for a sample of five individuals can be more powerful for detecting bottlenecks than using methods based on the SFS alone with a sample size of 20 (Bunnefeld et al., 2015; Lohse et al., 2011).

The bottleneck model is described in Bunnefeld et al. (2015), and describes a diploid, panmictic population. The bottleneck is considered to be instantaneous so that the parameters to describe it are a start time T_1 and a strength T_2, and its only effect is a burst of coalescence. The strength T_2 can be thought of as the time taken for a neutral model to accrue the same amount of coalescence as results from the bottleneck. Both parameters are measured in $2N_e$ generations. The step change model is the same except the event at T_1 is an instantaneous change in N_e, with strength parameter B_1 describing the N_e before the step change (looking backwards in time) as a proportion of the N_e after the step change. Therefore values of $B_1 < 1$ indicate a reduction in N_e (looking forward in time), values of $B_1 > 1$ indicate an increase, and $B_1 = 1$ indicates no change in N_e. Timings of inferred changes in population size represented by the time parameter T_1 were converted into years using the known generation times of the species and the spontaneous mutation rate for *Drosophila melanogaster* of $3.5*10^{-9}$ mutations per base per generation (Keightley et al., 2009).

To compare times of changes in population size across species and determine whether these differ significantly from one another, I will calculate the likelihood support for a shared time for every combination of species using the marginal likelihood curve of the time

parameter T_1. I will compare these to the likelihood support for the full model in which every species has its own time of population size change. If a shared time between species does not have a statistically reduced *lnL*, assuming *lnL* has a χ^2 distribution, the species in the cluster will be accepted as having changed in population size at the same time.

The difference in lnL between bottleneck and step change models was tested for significance by parametric bootstrapping. Simulations were performed using *msprime* (Kelleher, Etheridge, & McVean, 2016) with parameters inferred under the less supported model, and support for a bottleneck and a step change was assessed on each of 100 bootstrap replicates. 95% confidence intervals for the difference in *lnL* between the models were found by taking 2 standard deviations either side of the mean. The model with highest support in the real data was accepted as significantly better if the difference between its *lnL* and that of the less supported model was outside the 95% confidence interval calculated from the simulations.

The blockwise method of demographic inference was compared to an established method commonly used for inference of demographic history, the Pairwise Sequentially Markovian Coalescent (PSMC) (Li & Durbin, 2011), in order to check correspondence between the two methods. PSMC was also used to check the assumption of panmixia within an Iberian population, by running it on two different pairs of individuals. If there is a large difference in the population history depending on which individuals were used for inference, this indicates that there is population structure. PSMC uses haplotype information from one diploid (or two haploid) genomes to estimate population size at each time step, using the spatial distribution of homozygous and heterozygous loci along the genome to infer recombination events and the distribution of coalescence times of the sequences (Li & Durbin, 2011).

Gene prediction was performed *ab initio* using the gene prediction tool *Augustus* (Stanke & Morgenstern, 2005). *N. vitripennis* was used for the training set which teaches *Augustus* about the characteristics of the genomic features of the species for which genes are being predicted. *N. vitripennis* is a chalcid parasitoid in the same family Pteromalidae as *C. fungosa*, so was deemed a close enough relation to give reasonable gene predictions. Support for models of a bottleneck and a step change were assessed for genic and intergenic regions, and results were compared.

Results

Models of an instantaneous bottlneck and models of a step change are found to have a significantly greater *lnL* than null models of no change in population size in all parasitoid species (Tables 1 and 2). The step change model has a higher *lnL* than the bottleneck model in *T. auratus* and *M. stigmatizans*, but a lower *lnL* than the bottleneck model in all other species. The bottleneck model is found to fit significantly better than the step change model in *O. nitidulus* by parametric bootstrapping. The same bootstrap analysis has not yet been performed for the other species.

Table 1. Parameters inferred from bottleneck models in each of the seven species. $\theta=4N_e\mu$*block length, T_1= time in $2N_e$ generations, T_2=bottleneck strength

Species	*lnL* null	*lnL* bottleneck	θ	*Ne*	Time T_1	Strength T_2	years
C. fungosa	-26954.2	-26946.9	1.25267	267784	0.642868	0.234311	172149
E. brunniventris	-28518.1	-28510.6	1.82274	583837	0.500202	0.169463	292037
M. dorsalis	-26087.0	-26022.9	1.08152	138443	0	0.0859267	0
M. stigmatizans	-26768.0	-26768.0	1.293	58305	0	0.16399	0
O. nitidulus	-24636.4	-24120.9	1.51995	185269	0.0579716	0.384781	10740
O. pomaceus	-26482.1	-26427.7	1.48272	284929	0.527481	0.466117	150294
T. auratus	-29960.5	-29848.1	3.25691	709257	0.614549	0.509441	435873

Table 2. Parameters inferred from step change models in each of the seven species. $\theta=4N_e\mu$*block length, T_1= time in $2N_e$ generations, B_1= magnitude of step change as current N_e/ancestral N_e

Species	*lnL* null	*lnL* step change	θ	*Ne*	T_1	B_1	years
C. fungosa	-26954.2	-26950.6	1.08601	232156	4.00438	0.0050767	929643
E. brunniventris	-28518.1	-28514.6	1.59855	517947	3.19725	0.0005883	1656006
M. dorsalis	-26087.0	-26023.0	0.0259706	3363	0.0894048	0.0240534	301
M. stigmatizans	-26768.0	-26534.9	0.0133541	609	0.181098	0.0103384	220
O. nitidulus	-24636.4	-24142.1	0.380999	46441	0.68284	0.24511	31712
O. pomaceus	-26482.1	-26466.1	1.04849	201484	3.54446	0.0010467	714152
T. auratus	-29960.5	-29841.9	4.60327	1002454	0.124307	2.44719	124612

Evidence of clustering in time of changes in population size across species by visual inspection of the results is equivocal. When only the highest supported model in each species is considered, there is one possible cluster of three species with times of population size changes more recent than 11 thousand years ago (kya), another possible cluster of three species within 26 thousand years of 150 kya, and a single species close to 300 kya. (Figure 1). Whether these are statistically significant clusters will be tested by analysis of the marginal likelihood curves of the T_1 parameter (see methods). The most recent possible cluster, of two bottlenecks and one step change down, lies within the current interglacial period and these changes in population size could have occurred as a result of a common factor associated with this interglacial period. The other bottlenecks are estimated to have occurred during a glacial period more than 150 kya. The only increase in population size is seen in *T. auratus*, and appears to have occurred during the last interglacial period around 125 kya, which fits with the reasoning that population expansion is more likely during a period of warmer climate.

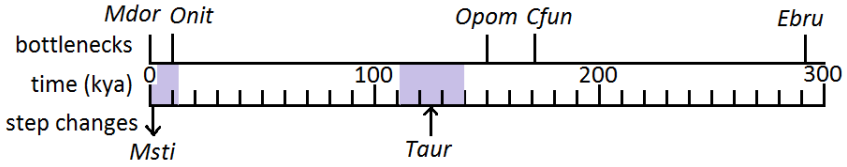


Figure 1. Point estimates of timings of changes in population size from the highest supported model in each species. The direction of step changes (forward in time) are indicated by arrows. Lilac bars indicate interglacial periods. *Mdor*: *M. dorsalis, Onit*: *O. nitidulus, Cfun*: *C. fungosa, Ebru*: *E. brunniventris, Msti*: *M. stigmatizans, Taur*: *T. auratus.*

The increase in population size seen in *T. auratus* is estimated to have occurred before the Iberian and Hungarian populations have been found to have split (Bunnefeld et al., *unpubl.*). If this is true, the same increase would be expected to be seen in the Hungarian population. To test this, I performed PSMC analysis on two Iberian individuals and on two Hungarian individuals, and the expected increase was observed in both populations (Figure 2). This indicates that the population size change did indeed occur before the two populations split, and is therefore not associated specifically with the Iberian population.

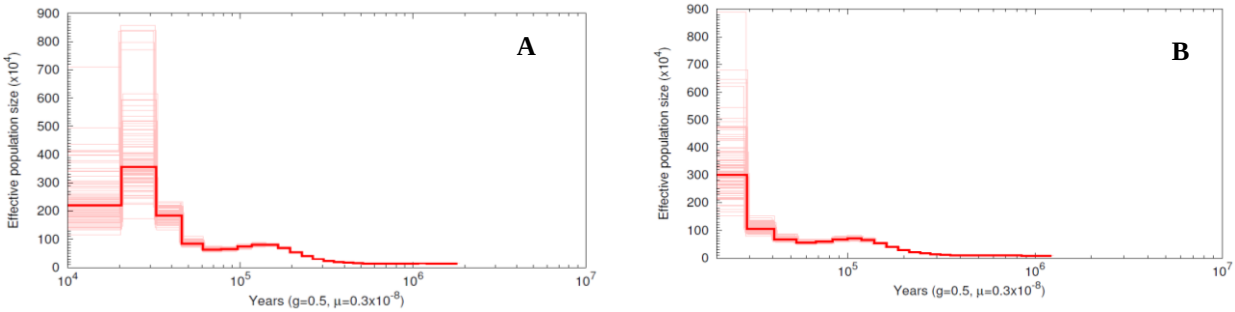


Figure 2. History of population size changes inferred by PSMC on **A**: Iberian and **B**: Hungarian individuals of *T. auratus*. The bold red line indicates the estimate, and the faint red lines are bootstrap replicates. The same increase in population size at just over $1x10^5$ years ago is observed in both populations, indicating that the increase inferred by the bSFS analysis occurred before the populations split.

For *O. nitidulus*, a decrease in population size was seen in the results of PSMC at approximately the same time as predicted by the bSFS step change model. However, the bottleneck model had better support for *O. nitidulus*. Parametric bootstrap simulations were performed using *msprime* (Kelleher et al., 2016) under the step change model to find confidence intervals for the estimated parameters. These were found to be relatively concordant with the PSMC results (Figure 3). The PSMC results for *O. pomaceus* showed a decrease followed by an increase in population size. The centre of the trough was at about 200 kya, close the time predicted by the bSFS analysis for a bottleneck, and PSMC and bSFS are therefore fairly concordant for this species as well. PSMC analysis has not been performed for the other four species as these three species were deemed enough for a simple visual test of agreement between the two methods.

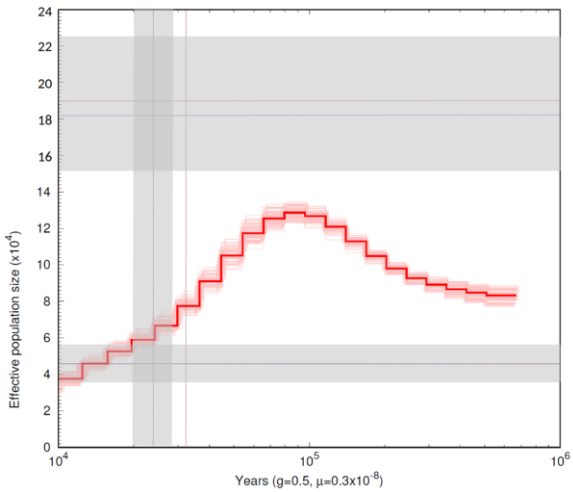


Figure 3. PSMC results for two *O. nitidulus* individuals (red line) with 100 bootstrap replicates (faint red lines), plotted with the parameters estimated from the bSFS step change model (red dotted lines) and 100 parametric bootstrap replicates (grey bars, blue dotted lines are the mean). The lower horizontal lines are the current N_e, the vertical lines are the time of the step change and the upper horizontal lines are the ancestral N_e.

To check the assumption of panmixia within Iberia, the same PSMC analysis was run using a different pair of Iberian *O. nitidulus* individuals as was used previously and the outputs were compared (Figure 4). Little difference was observed, so the assumption of panmixia appears to be upheld and further analysis of population structure was not pursued.

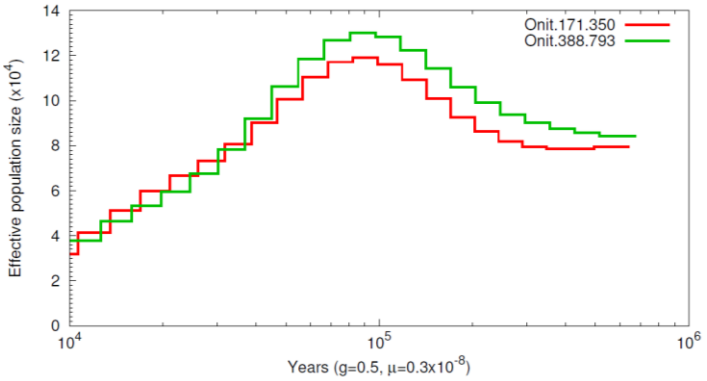


Figure 4. History of population size changes inferred by PSMC on two different pairs of *O. nitidulus* individuals from Iberia, showing little difference between them. Individuals 171 and 350 were from Portugal and Madrid, and 388 and 793 were from Madrid and Avila, respectively.

24,103 genes were predicted in *O. nitidulus*, comprising 61Mbp of the 260Mbp of sequence. This is more genes than in *N. vitripennis*, which was used for training the *Augustus* software and has about 19,000 predicted gene transcripts (Werren et al., 2010). The higher number of predicted genes could be due partially to the lower quality assembly of *O. nitidulus* resulting in some genes or parts of genes being present in multiple copies. Using only intergenic regions for the bSFS analysis, there was slightly greater diversity and a slightly more recent, smaller step change detected than when using only genic regions. The same was true for the bottleneck model. The greater diversity in intergenic regions was expected, as both positive and background selection act to reduce diversity around genes (Charlesworth, 2012; Smith & Haigh, 1974). The changes in population size are found to be smaller using only intergenic regions, possibly for the same reason, that selection on genic regions reduces diversity thus increasing the apparent drop in population size. These results show that there is an effect of selection on these demographic inferences, however the effect is small so does not greatly change the conclusions or the results on other species.

Components of this chapter remaining to be completed are testing for significance between the bottleneck and step change models in each species, finding confidence intervals for inferred parameters, and testing for clustering of the times of population size changes across species. When these are completed, I will have an accurate estimate of population history for each species and be better able to compare across species. It will improve our knowledge of how concordant demographic histories are across species within a system, and may suggest events such as glaciations and range expansions associated with these population

size changes. The demographic models will be used in further chapters as null models from which to infer selection.

Chapter 2: Gene duplication

Gene duplication has been known to be an important process in evolution since the mid 20th century (Bridges, 1936; Ohno, 1970), and the signature of positive selection has been found in gene duplicates in several species (Hughes, Green, Garbayo, & Roberts, 2000; Tanaka & Nei, 1989; Zhang, Rosenberg, & Nei, 1998). However, the role of positive selection over a short timescale in the fixation of recent gene duplicates is less well understood (Cardoso-Moreira et al., 2016). It is typically considered that duplicates are either initially neutral and fixed by drift, followed by functional divergence or subfunctionalization (when each duplicate assumes part of the function of the original gene), or that their fixation is a result of positive selection (Hughes, 1994; Zhang, 2003). In insects, habitat diversity including host range has been shown to correlate with the rate of evolution and loss of gene duplicates. This is suggested to be due to duplicated genes facilitating adaptation to diverse environments, thus resulting in generalist species having a higher proportion of duplicates than specialist species (Makino & Kawata, 2012; McBride, 2007). Investigating recently duplicated genes to identify their functions and the action of positive selection across species with different habitat diversities could therefore aid our understanding of the role of gene duplications in adaptation. Here I aim to test for an elevated rate of evolution in recently duplicated genes and show how this varies according to species ecology and gene function.

This question requires a comparative approach, and a system in which there is good existing knowledge on background ecology and phylogeography, there can be control for evolutionary history, and which is ecologically closed. There must be a population which is panmictic and has been diverged from other populations for long enough for some genes to have duplicated and be undergoing selection as a result of being in a different environment to other populations. I shall therefore use the same gall wasp parasitoid system and data as for chapter 1, again using the Iberian population of parasitoids as these are estimated to have split from the rest of Europe between 14 and 210 thousand years ago (Bunnefeld et al, *unpubl*.). The rate of gene duplication has been estimated in *Drosophila melanogaster* as 1.25*e*-7 complete duplications/gene/generation (Schrider et al. 2013), which if similar to that of the parasitoids gives plenty of time for duplications to have occurred in even the most recently split Iberian populations. Some parasitoid species have also been reared from hosts in Iberia that have not been shown to be used as hosts elsewhere and *vice versa* (Askew et al. 2013), providing a potential role for gene duplications in adaptation to a new host range.

Gene duplicates will be identified by using sequencing coverage and heterozygous variant calls. For example, a gene with two recent copies may only be mapped as one gene, but will have on average twice the coverage as a single copy gene. I will therefore first predict genes using the software *Augustus* (Stanke & Morgenstern, 2005) *ab initio* with *N. vitripennis* for the training set, as I have already done for *O. nitidulus* (see Chapter 1). I will then find the average coverage for each gene and plot this distribution to look for additional peaks at multiples of the median depth. Duplicate genes may also have heterozygous variant calls due to divergence between the gene copies, whereas heterozygous calls in single copy genes are impossible in haploid individuals such as the male parasitoids which have been sequenced. The only exceptions to this are sequencing errors. To identify heterozygous calls that are not sequencing errors, I will perform variant calling again with a diploid variant caller such as *GATK* (DePristo et al., 2011). Genes that have both a higher coverage than expected and heterozygous variant calls will make up a candidate set of duplicated genes.

Comparison with individuals from other populations will identify duplicates specific to Iberia.

I will determine the functional categories of each of the candidate genes using Gene Ontology (http://www.geneontology.org) or BLAST to the annotated *N. vitripennis* genome. I will find out if duplications are over-represented in any functional categories, as well as in host generalists compared to specialists, and I will control for phylogeny using the independent comparisons method (Felsenstein, 1985). Functional categories expected to have a high proportion of duplicates include olfactory receptor genes and gustatory receptor genes as these are involved in chemoreception of host signals, and rate of loss of these has been shown to be higher in the specialist fly species *Drosophila sechellia* than its generalist sibling *D. simulans* (McBride, 2007). Generalist parasitoids are expected to have a higher proportion of gene duplicates than specialists due to their wider host range providing a higher habitat variability. This is hypothesised to create a greater selective advantage to gene duplicates which increase genetic variation and allow improved environmental adaptability (Makino & Kawata, 2012). I will also test for a higher rate of evolution in duplicated genes in specialist parasitoids compared to generalists. This will support the hypothesis that specialists are under stronger directional selection and/or lower constraint due to co-evolution with their hosts than generalists. This will also demonstrate that gene duplications are an important source of variation associated with adaptation to hosts.

Chapter 3: Quantifying selection with the bSFS

There are many and varied existing tests for selection. They can be broadly classified into rate-based, site frequency spectrum (SFS) -based and haplotype-based methods. Rate-based methods include the McDonald-Kreitman (MK) test, which compares the rate of evolution between synonymous and non-synonymous sites and the level of polymorphism and divergence to detect an elevated rate of evolution attributable to selection. This assumes that synonymous sites are selectively neutral, and that there are no weakly deleterious non-synonymous polymorphic sites. This type of test detects selection on long time scales and can be strongly affected by selection on linked sites, including selective sweeps and background selection (Messer & Petrov, 2013a). A selective sweep is when an adaptive allele rises swiftly from a low to a high frequency in a population as a result of positive selection, and drags surrounding variation to a high frequency along with it (Nielsen et al., 2005). This causes a skew in the SFS and strong linkage disequilibrium in the surrounding area of the genome, which are used by SFS-based and haplotype-based methods respectively to infer selection. SFS-based methods include Tajima's D, Fay & Wu's H, and the Composite Likelihood Ratio (CLR) test (Vitti, Grossman, & Sabeti, 2013). Although the data for these tests are relatively easily obtained using short-read sequencing technology, they do not utilise any linkage information (Bhaskar & Song, 2014). Haplotype-based tests however, such as iHS, nS_L and H_{12} (Ferrer-Admetlla, Liang, Korneliussen, & Nielsen, 2014; Garud, Messer, Buzbas, & Petrov, 2015; Messer & Petrov, 2013b; Vatsiou, Bazin, & Gaggiotti, 2016), do use long range linkage information but require well assembled genomes, and both SFS- and haplotype-based methods often require knowledge of which alleles are inherited together on the same chromosome (phasing).

Many of these tests can be useful for answering some questions using model species with large sample sizes and well assembled genomes. However, many interesting and important ecological questions are only answerable using non-model species for which data are currently scarce or non-existent, and these questions therefore require tests that are powerful using only small sample sizes and fragmented genome assemblies. There are currently no available tests for selection that use both linkage and SFS information, and do

not require phasing, high-quality genome assemblies, or large sample sizes. In contrast to this, the bSFS method of demographic inference is well suited to the kind of data often generated for non-model species, as it works well with fragmented genome assemblies and requires a maximum sample size of only five individuals (Bunnefeld et al., 2015).

In this chapter, I aim to further develop the bSFS maximum likelihood method of demographic inference in order to apply it to the detection of selective events. I will then demonstrate its use by identifying the action of selection in the genome of the fly *Drosophila mojavensis*. The application of a demographic test to the study of selection is possible due to the local genomic effects of a demographic event being very similar to that of selection, for example both a population bottleneck and a selective sweep create a star shaped coalescent history (Galtier, Depaulis, & Barton, 2000). The bSFS framework can be used to distinguish the effects of demography and selection by incorporating demographic parameters in the selection model and setting them to those previously inferred by the demographic model.

To show that the bSFS contains information to distinguish between demography and selection, I have plotted the proportions of each mutational configuration (arrangement of mutations on the branches of the genealogy in each block) in genic and intergenic blocks in the parasitoid wasp *O. nitidulus* (Figure 5). Genic blocks are defined as those within 1kb of genes predicted by the software *Augustus* (Stanke & Morgenstern, 2005), and intergenic blocks are all others. If the only difference between the categories is that intergenic blocks are less influenced by selective events than genic blocks, then the deviation from the line y=x must be a result of selection acting on genes. Genic regions are expected to have a higher proportion of invariant blocks as a result of both background and positive selection eliminating variants. Positive selection in the form of selective sweeps is expected to leave an excess of singletons in genic blocks, changing the proportions of configurations in Figure 5 when compared to the effects of background selection. To confirm that the deviation from the line y=x indicating neutrality is a true and consistent effect, I will use the SliM forward simulator (Messer, 2013) to simulate data containing regions evolving neutrally and affected by both background and positive selection, and again compare the counts of bSFS configurations. This will inform expectations of configuration counts under different scenarios, and aid the derivation of the GF for a model including parameters for the strength and timing of selective events.

I will use this bSFS method of selection inference to identify factors potentially important in adaptation in the cactophilic fly *Drosophila mojavensis*. *D. mojavensis* is a fly in the *D. repleta* group native to the southern United States and Mexico. Its phylogeography is well studied, and it is known to have evolved in Baja California and subsequently spread to the mainland. The Baja population uses a different species of host cactus to the mainland populations, and there is divergence between them as well as some pre- and post-mating reproductive isolation (Etges et al., 2015). Existing data consist of fragmented whole genome sequence, as well as microarray data of genes differentially expressed between populations, regions, sex, and host cactus (Etges et al., 2015). I will use the bSFS method to find maximum likelihood estimates of selection for different sets of candidate genes. One set will be a control set of genes taken randomly from the genome, length matched to the other candidate sets so as to provide a genome-wide estimate of selection. Duplicated genes will form another candidate set, and will be found using the method described in Chapter 2. Further sets will be genes differentially expressed between populations, sexes and host cacti as identified by Etges et al. (2015). This will help to identify the factors most affected by selection and therefore potentially the most important in adaptation in this species.

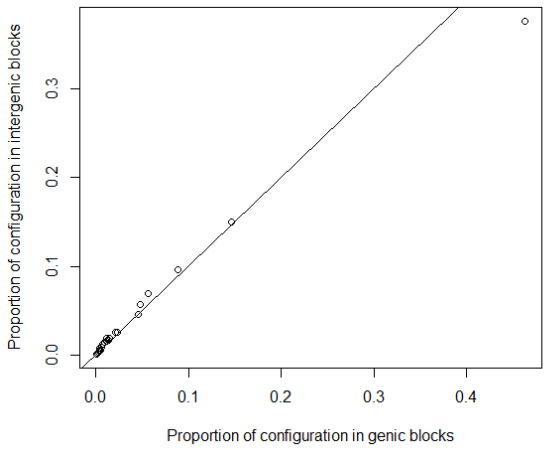


Figure 5. Comparison of the proportions of mutational configurations in genic and intergenic blocks in *O. nitidulus*. Deviations from the x=y line (black) indicate the effect of selection on the relative proportions of mutational configurations within each block. This can be investigated further by simulation. The most common configuration is that of an invariant block (0 singletons, 0 doubletons), followed by increasing numbers of singletons then doubletons.

Chapter 4: Comparative selection

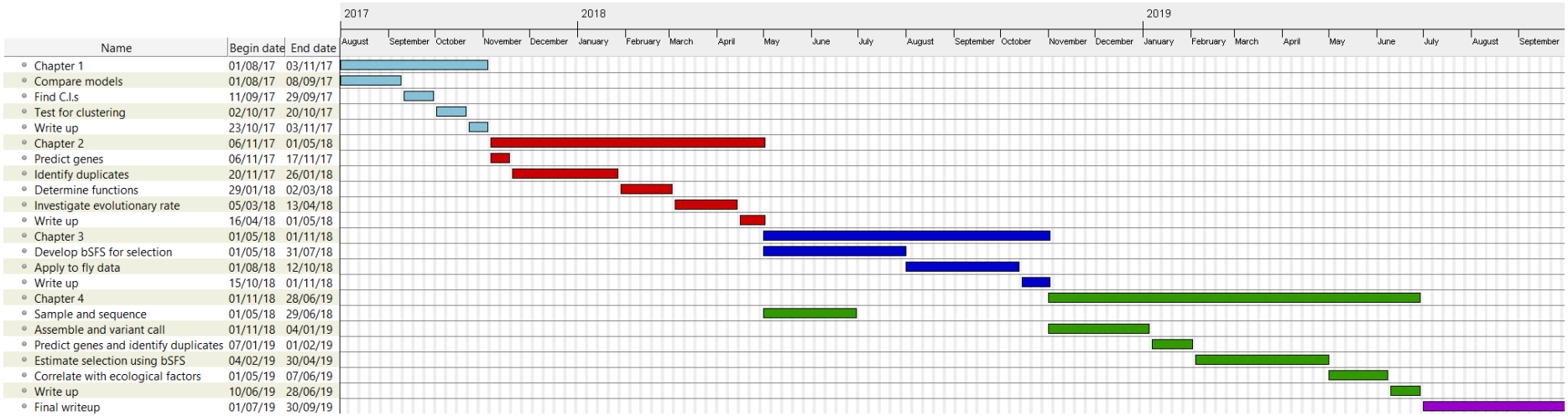
Although selection has been shown to have occurred in many species, it is largely unknown how selection is brought about by ecology (Egan et al., 2008). It has been suggested that specialist species with narrow ecological niches, such as insect herbivores and parasitoids with a very small number of host species, experience stronger selection on genes involved in species interactions than their more generalist counterparts (McBride & Arguello, 2007). However, a mechanism by which generalist species may be able to diversify and adapt to multiple hosts is gene duplication (Makino & Kawata, 2012). Therefore, in generalist species recently duplicated genes may be under stronger directional selection than both other genes in generalists and duplicated genes in specialist species. In contrast, other non-duplicated genes involved in host interactions such as host detection may be under stronger selection in specialist species. Here I aim to find out how strength of selection varies according to host range in duplicated genes and genes involved in host interactions.

Finding ecological factors that predict the action of selection requires a comparative approach to isolate the factors of interest and control for other confounding factors. For the reasons described in Chapter 1, I will use the gall wasp parasitoid system to investigate this problem, and the bSFS method of selection inference that I will develop in Chapter 3. I will control for demography explicitly using the models of population size changes fitted in Chapter 1 as null models. I will test for selection in candidate gene sets, including a set of duplicated genes identified in Chapter 2, olfactory and gustatory receptor genes thought to be involved in host detection, and a length matched set of control genes selected at random from the rest of the genome to estimate the genome-wide strength of selection. Strength of selection in each of these candidate sets will be tested for correlation with host range, and for association with timings of known population size changes, population splits (Bunnefeld et al., *unpubl.*), and glaciation events. Phylogeny will be controlled for with the independent comparisons method (Felsenstein, 1985).

Sequence data on several more species will also be gathered to increase the power of the comparative analysis. The type of sequencing and number of extra species sequenced depends on cost. Three species, *Eurytoma adleriae, Eupelmus annulatus* and *Eupelmus urozonus*, have two Spanish individuals already sequenced so would require only three more each to have the maximum number that the demographic bSFS method can use. It is expected that this will be the same for a selection test based on the bSFS. These species are present in Iberia and other European populations, and have contrasting sizes of host ranges with the other species sampled in their genera, providing high power for the independent comparisons.

An alternative that is less expensive than whole genome sequencing is pooled RNAseq which requires fewer library preparations. However, RNAseq gives less accurate allele frequency estimation and does not sequence intergenic regions. Genes with low expression level will also have lower coverage, with expression level depending on many factors such as age, sex and environmental conditions (Konczal, Koteja, Stuglik, Radwan, & Babik, 2014). Many individuals are also needed per pool for pooled RNAseq, as well as replicate pools. Furthermore, whole genome sequence data would fit better with the existing data, allowing a more straightforward comparison between species.

Whole genome sequencing on the Illumina HiseqX at Edinburgh Genomics for the extra three species above at approximately the same coverage as the existing data is estimated to cost £2035. Preparing my own libraries would reduce the cost, potentially allowing another two species to be sequenced. With this additional sequencing effort, the estimated maximum number of species that could be used for the comparative analysis is 12, leaving it with quite low power to detect correlations. However, it would still be informative of strong trends and could inform further studies of the ecological determinants of selection.



Name	Begin date	End date
Chapter 1	01/08/17	03/11/17
Compare models	01/08/17	08/09/17
Find C.I.s	11/09/17	29/09/17
Test for clustering	02/10/17	20/10/17
Write up	23/10/17	03/11/17
Chapter 2	06/11/17	01/05/18
Predict genes	06/11/17	17/11/17
Identify duplicates	20/11/17	26/01/18
Determine functions	29/01/18	02/03/18
Investigate evolutionary rate	05/03/18	13/04/18
Write up	16/04/18	01/05/18
Chapter 3	01/05/18	01/11/18
Develop bSFS for selection	01/05/18	31/07/18
Apply to fly data	01/08/18	12/10/18
Write up	15/10/18	01/11/18
Chapter 4	01/11/18	28/06/19
Sample and sequence	01/05/18	29/06/18
Assemble and variant call	01/11/18	04/01/19
Predict genes and identify duplicates	07/01/19	01/02/19
Estimate selection using bSFS	04/02/19	30/04/19
Correlate with ecological factors	01/05/19	07/06/19
Write up	10/06/19	28/06/19
Final writeup	01/07/19	30/09/19

Figure 6. Gantt chart showing the expected timings of each stage of each chapter.

References

Askew, R. R., Melika, G., Pujade-Villar, J., Schönrogge, K., Stone, G. N., & Nieves-Aldrey, J. L. (2013). *Catalogue of parasitoids and inquilines in cynipid oak galls in the West Palaearctic*. *Zootaxa* (Vol. 3643). https://doi.org/10.11646/zootaxa.3643.1.1

Berg, J. J., & Coop, G. (2015). A Coalescent Model for a Sweep of a Unique Standing Variant. *Genetics*, *201*(2), 707–725. https://doi.org/10.1534/genetics.115.178962

Bhaskar, A., & Song, Y. S. (2014). Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *The Annals of Statistics*, *42*(6), 2469–2493. https://doi.org/10.1214/14-AOS1264

Bridges, C. B. (1936). The Bar "gene" a duplication. *Science*, *83*(2148), 210–211. Retrieved from http://science.sciencemag.org/content/sci/83/2148/210.full.pdf

Bunnefeld, L., Frantz, L. A. F., & Lohse, K. (2015). Inferring bottlenecks from genome-wide samples of short sequence blocks. *Genetics*, *201*(3), 1157–1169. https://doi.org/10.1534/genetics.115.179861

Cardoso-Moreira, M., Arguello, J. R., Gottipati, S., Harshman, L. G., Grenier, J. K., & Clark, A. G. (2016). Evidence for the fixation of gene duplications by positive selection in Drosophila. *Genome Research*, *26*(6), 787–98. https://doi.org/10.1101/gr.199323.115

Charlesworth, B. (2012). The effects of deleterious mutations on evolution at linked sites. *Genetics*, *190*(1), 5–22. https://doi.org/10.1534/genetics.111.134288

de Graaf, D. C., Aerts, M., Brunain, M., Desjardins, C. A., Jacobs, F. J., Werren, J. H., & Devreese, B. (2010). Insights into the venom composition of the ectoparasitoid wasp *Nasonia vitripennis* from bioinformatic and proteomic studies. *Insect Molecular Biology*, *19*(s1), 11–26. https://doi.org/10.1111/j.1365-2583.2009.00914.x

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Egan, S. P., Nosil, P., & Funk, D. J. (2008). SELECTION AND GENOMIC DIFFERENTIATION DURING ECOLOGICAL SPECIATION: ISOLATING THE CONTRIBUTIONS OF HOST ASSOCIATION VIA A COMPARATIVE GENOME SCAN OF NEOCHLAMISUS BEBBIANAE LEAF BEETLES. *Evolution*, *62*(5), 1162–1181. https://doi.org/10.1111/j.1558-5646.2008.00352.x

Etges, W. J., Trotter, M. V., de Oliveira, C. C., Rajpurohit, S., Gibbs, A. G., & Tuljapurkar, S. (2015). Deciphering life history transcriptomes in different environments. *Molecular Ecology*, *24*(1), 151–179. https://doi.org/10.1111/mec.13017

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, *125*(1), 1–15. Retrieved from http://www.indiana.edu/~kettlab/A501/Felsenstein1985.pdf

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, *31*(5), 1275–91. https://doi.org/10.1093/molbev/msu077

Galtier, N., Depaulis, F., & Barton, N. H. (2000). Detecting Bottlenecks and Selective Sweeps From DNA Sequence Polymorphism. *Genetics*, *155*, 981–987.

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genetics*, *11*(2), e1005004. https://doi.org/10.1371/journal.pgen.1005004

Hayward, A., & Stone, G. N. (2005). Oak gall wasp communities: Evolution and ecology. *Basic and Applied Ecology*, *6*(5), 435–443. https://doi.org/10.1016/j.baae.2005.07.003

Huber, C. D., Nordborg, M., Hermisson, J., & Hellmann, I. (2014). Keeping it local: evidence for positive selection in Swedish Arabidopsis thaliana. *Molecular Biology and*

Evolution, *31*(11), 3026–39. https://doi.org/10.1093/molbev/msu247
Hudson, R. R., Kreitman, M., & Aguade, M. (1987). A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics*, *116*(1), 153–159. Retrieved from http://www.genetics.org/content/genetics/116/1/153.full.pdf
Hughes, A. L. (1994). The Evolution of Functionally Novel Proteins after Gene Duplication. *Proceedings of the Royal Society of London B: Biological Sciences*, *256*(1346). Retrieved from http://rspb.royalsocietypublishing.org/content/256/1346/119?hwshib2=authn%3A1501925909%3A20170804%253A56fb887e-e680-43f4-8b27-c319d7cb787f%3A0%3A0%3A0%3AHqqwWocrUp1%2FyIfYQ702LA%3D%3D
Hughes, A. L., Green, J. A., Garbayo, J. M., & Roberts, R. M. (2000). Adaptive diversification within a large family of recently duplicated, placentally expressed genes. *Proceedings of the National Academy of Sciences*, *97*(7), 3319–3323. https://doi.org/10.1073/pnas.97.7.3319
Keightley, P. D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., & Blaxter, M. L. (2009). Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Research*, *19*(7), 1195–201. https://doi.org/10.1101/gr.091231.109
Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*, *12*(5), 1–22. https://doi.org/10.1371/journal.pcbi.1004842
Konczal, M., Koteja, P., Stuglik, M. T., Radwan, J., & Babik, W. (2014). Accuracy of allele frequency estimation using pooled RNA-Seq. *Molecular Ecology Resources*, *14*(2), 381–392. https://doi.org/10.1111/1755-0998.12186
Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231
LI, J., LI, H., JAKOBSSON, M., LI, S., SJÖDIN, P., & LASCOUX, M. (2012). Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Molecular Ecology*, *21*(1), 28–44. https://doi.org/10.1111/j.1365-294X.2011.05308.x
Lohse, K., Harrison, R. J., & Barton, N. H. (2011). A general method for calculating likelihoods under the coalescent process. *Genetics*, *189*(3), 977–87. https://doi.org/10.1534/genetics.111.129569
Makino, T., & Kawata, M. (2012). Habitat Variability Correlates with Duplicate Content of Drosophila Genomes. *Molecular Biology and Evolution*, *29*(10), 3169–3179. https://doi.org/10.1093/molbev/mss133
Manel, S., Perrier, C., Pratlong, M., Abi-Rached, L., Paganini, J., Pontarotti, P., & Aurelle, D. (2016). Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Molecular Ecology*, *25*(1), 170–184. https://doi.org/10.1111/mec.13468
McBride, C. S. (2007). Rapid evolution of smell and taste receptor genes during host specialization in Drosophila sechellia. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(12), 4996–5001. https://doi.org/10.1073/pnas.0608424104
McBride, C. S., & Arguello, J. R. (2007). Five Drosophila genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics*, *177*(3), 1395–1416. https://doi.org/10.1534/genetics.107.078683
Messer, P. W. (2013). SLiM: Simulating Evolution with Selection and Linkage. *Genetics*, *194*(4).
Messer, P. W., & Petrov, D. A. (2013a). Frequent adaptation and the McDonald-Kreitman

test. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(21), 8615–20. https://doi.org/10.1073/pnas.1220835110

Messer, P. W., & Petrov, D. A. (2013b). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology & Evolution*, *28*(11), 659–669. https://doi.org/10.1016/j.tree.2013.08.003

Moura, A. E., Janse van Rensburg, C., Pilot, M., Tehrani, A., Best, P. B., Thornton, M., … Hoelzel, A. R. (2014). Killer Whale Nuclear Genome and mtDNA Reveal Widespread Population Bottleneck during the Last Glacial Maximum. *Molecular Biology and Evolution*, *31*(5), 1121–1131. https://doi.org/10.1093/molbev/msu058

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research*, *15*(11), 1566–75. https://doi.org/10.1101/gr.4252305

Ohno, S. (1970). *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-86659-3

Schrider, D. R., Shanku, A. G., & Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, *204*(3), 1207–1223. https://doi.org/10.1534/genetics.116.190223

Slatkin, M., & Hudson, R. R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, *129*(2). Retrieved from http://www.genetics.org/content/129/2/555?ijkey=f6ca6a169e69064e27930d166d28591c0661bef5&keytype2=tf_ipsecsha

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res., Camb*, *23*, 23–35.

Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*, *33*(Web Server issue), W465-7. https://doi.org/10.1093/nar/gki458

Stone, G. N., Lohse, K., Nicholls, J. A., Fuentes-Utrilla, P., Sinclair, F., Schönrogge, K., … Hickerson, M. J. (2012). Reconstructing Community Assembly in Time and Space Reveals Enemy Escape in a Western Palearctic Insect Community. *Current Biology*, *22*(6), 532–537. https://doi.org/10.1016/j.cub.2012.01.059

Stone, G. N., Schönrogge, K., Atkinson, R. J., Bellido, D., & Pujade-Villar, J. (2002). The population biology of oak gall wasps (Hymenoptera : Cynipidae). *Annual Review of Entomology*, *47*(1), 633–668. https://doi.org/10.1146/annurev.ento.47.091201.145247

Tanaka, T., & Nei, M. (1989). Positive Darwinian Selection Observed at the Variable-Region Genes of Immunoglobulins ’. *Molecular Biology and Evolution*, *6*, 447–459. Retrieved from https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/mbe/6/5/10.1093_oxfordjournals.molbev.a040569/1/1tana.pdf?Expires=1501930378&Signature=Be75jqBNk6lVjxGyc37-y8VGyjUFPuxM7E3ovPcR0CyfWgPqbwvSupraGAGsp74S-h0~oVlU5~iUZOHp3yMkc2WrLcUHUzp9CCNu

Vatsiou, A. I., Bazin, E., & Gaggiotti, O. E. (2016). Detection of selective sweeps in structured populations: A comparison of recent methods. *Molecular Ecology*, *25*(1), 89–103. https://doi.org/10.1111/mec.13360

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, *47*(1), 97–120. https://doi.org/10.1146/annurev-genet-111212-133526

Werren, J. H., Richards, S., Desjardins, C. A., Niehuis, O., Gadau, J., Colbourne, J. K., … Gibbs. (2010). Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science (New York, N.Y.)*, *327*(5963), 343–8. https://doi.org/10.1126/science.1178028

Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., & Bustamante, C. D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(22), 7882–7. https://doi.org/10.1073/pnas.0502300102

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, *18*(6), 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

Zhang, J., Rosenberg, H. F., & Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *PNAS*, *95*, 3708–3713. Retrieved from http://www.pnas.org/content/95/7/3708.full.pdf